Artificial Intelligence Methods For Financial Technologies

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Electronics, Computer And Electrical Engineering

by

Golnoosh Babaei

Matriculation no. 495359

Supervisor

Prof. Paolo Giudici

University of Pavia

DECLARATION

I declare that the thesis entitled “Artificial Intelligence Methods For Financial Technologies” submitted by me, for the award of the degree of *Doctor of Philosophy* to the University of Pavia is a record of the work carried out by me under the supervision of Prof. Paolo Giudici, professor of statistics and data science, University of Pavia.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Pavia, Italy

2024

TABLE OF CONTENTS

ABSTRACT

Financial technologies (FinTechs) have changed the way people invest their money in the last years. Artificial Intelligence (AI) based algorithms can detect complex patterns, assess risks, and make decisions from real-time data. This enables Fintech companies to provide personalized services and optimize operations effectively. Therefore, it is necessary to evaluate and study Artificial Intelligence (AI) and Machine Learning (ML) models to improve available methods not only to contribute to the literature but also to help investors improve their investments in the FinTech market. ML models are complex approaches that can lead to highly accurate predictions. Hence, decision-making models that are powered by ML methods can help investors to make more accurate decisions. However, accuracy is not the only important aspect of making decisions but a reliable approach should also be explainable and fair. For explainability, explainable AI (XAI) has emerged to find the contribution of available variables in a model. In terms of fairness, ML models could have some bias against a special group of data. Consequently, the decisions made by these biased models are not fair to be used for investment problems. The contributions of this thesis are developed in six self-contained chapters. Chapter 1 proposes a novel decision-making framework as a combination of lazy learning and eager learning methods. The proposed hybrid instance-based learning (HIBL) model can predict the return and risk of new loans and help investors to find an optimal portfolio. Chapter 2 compares the performance of three different classifiers, Random Forest, Support Vector Machines, and Naive Bayes in combination with a portfolio optimization approach. The proposed methodology can evaluate loans from the two important aspects of an investment: risk and return. Chapter 3 mentions the lack of explainability of accurate ML models and proposes a model selection approach using an explainable artificial intelligence (XAI) method. Then, in Chapter 4, XAI is utilized for a portfolio optimization problem in the cryptocurrency market. To this end, contributions of cryptocurrencies are found using Shapley values. Chapter 5 aims to improve Shapley values which are commonly used in the literature. For this purpose, an instance-based method is applied to the estimation phase of the SHAP concept. Since, in addition to the explainability, it is vital to check fairness of the model, Chapter 6 discusses this issue and checks if a credit scoring model is fair towards loan applicants in the United States.

ACKNOWLEDGEMENT

At the same time I started my PhD, I started a new life. I can say that I was born for the second time. My passion for learning and education made me leave everything behind in my country and immigrate to Italy to pursue my PhD degree. In this journey, I met many people and from each of them I have learned something that made up what I am right now. I have to thank all of them for this.

First of all, I should thank my supervisor, for being the present guide who drove me throughout my PhD journey. I have learned so many things from him, not only in academics but also many things in life. From the first seconds in Italy, he helped me to have a happy life here. So, in addition to research, he taught me lots of life things. Thank you very much prof. Giudici! As I have always told him, I owe my success to him. I also thank all my colleagues for being always around. I was really lucky to have these colleagues in this journey because they always helped me and supported me in difficult days of life.

I also thank the two reviewers, Dr. Shahrooz Bamdad and Dr. Parvati Neelakantan, who spent their valuable time on helping me to improve my thesis with their constructive comments. Dr. Bamdad was the first person who believed in me in the academic world and helped me to be confident. I remember when I was supposed to decide about leaving my country I was near to give up several times but he gave me power and helped me to achieve my goals.

Last but not least, I thank my family especially my fiance for always supporting me and their constant encouragement. Being far from me was not easy for them but they sacrificed themselves and ignored their feeling just to give me this chance to grow. Special thanks go to my mother, who gave me all the opportunities in life. She has been a real teacher for me and she has always tried to make a strong and independent woman from me. I should say that I am proud of you MAMAN and I wish I could be a great mom like you in the future.

Pavia, Italy

2024

GENERAL INTRODUCTION

This thesis aims to contribute to the existing literature from a methodological and empirical viewpoint, i.e. providing new methodological tools and extending the use of existing models, and uncovering insights on the FinTech market. The thesis mainly contributes to the literature on the Machine Learning (ML) models for FinTech, explainable Artificial Intelligence (AI) methods to improve decision-making models, and fairness of the provided decisions.

It should be noted that since the number of stored data of transactions, borrowers, etc has increased in the last years, it is necessary to have powerful models that can make accurate decisions and also automate decision-making in the booming financial markets. Therefore, AI is important in finance and one of its benefits is automation. In other words, AI possesses the capability to streamline workflows and procedures, operate independently while adhering to ethical standards, and facilitate informed decision-making and service provision. In addition, AI can assist financial service entities in mitigating human errors associated with data processing, analytics, onboarding procedures, customer interactions, and various other tasks. This is achieved through the implementation of algorithms that consistently adhere to predefined processes on each occurrence.

However, considering the new regulations in financial markets, clarification of the decisions and the reasons behind them should be provided for the stakeholders of the models. Hence, it has been announced that the accuracy and speed of highly capable AI models in FinTech, are not the only significant factors. The explainability of these models is essential when they want to be used in financial decision-making problems. For instance, if an ML model is used to decide if a lending platform should issue an applicant's loan request or not, not only an accurate decision is needed in this case but also it is needed that the platform provide reasons for the decisions received by the applicant.

Another ethics in AI that has been considered recently in the literature is fairness. It has been claimed recently that some AI-based decision-making models are not fair so consequently they are sometimes biased against a special group of observations. Therefore, it is obliged to check the fairness of the models in FinTech and become sure that the model performs fairly towards whole the population.

Finally, we can say that AI is an essential tool and the basis of the FinTech market.

Consequently, it is needed to improve models based on different aspects of accuracy, explainability, and fairness. Contributions of this thesis are presented in six self-contained chapters. Since AI can be applied to different types of datasets for various purposes, I utilize different kinds of financial data in the following six chapters. Therefore, readers of this thesis can recognize how we use different ML models in different settings for various purposes using different datasets.

Chapter 1 proposes a novel decision-making framework in which lazy learning methods (e.g. instance-based learning) and eager learning approaches (e.g. artificial neural networks) are integrated. The proposed hybrid instance-based learning (HIBL) model can predict the return and risk of new loans and help investors to find the optimal portfolio. In order to check the effectiveness of the proposed model, I use the LendingClub dataset, a real-world dataset from one of the most popular P2P lending marketplaces. Finally, a comparison between my proposed model and a rating-based method is done and the results show that the proposed HIBL model can improve investments in P2P lending. In Chapter 2, an investment decision model is proposed based on the non-default loans predicted using three different classifiers, including random forest (RF) that is a multitude of decision trees, support vector machines, and naive Bayes. In particular, each of these classifiers is integrated with a portfolio optimization problem to understand which combination leads to the best portfolio concerning risk and return. To find the best integration, numerical studies are conducted based on the P2P lending data. The results indicate that combining the RF algorithm and portfolio optimization problem leads to the least level of risk. Moreover, it is concluded that RF is the best classifier in terms of performance analysis.

In the proposed methods in the previous chapters, I use machine learning models that are highly accurate but however, in addition to accuracy, the explainability of machine learning models is important. Therefore, chapter 3 utilizes Shapley values to propose a model selection approach. In particular, I use Random forest (RF) to predict the credit risk and expected return of credit lending to a set of Italian small and medium enterprises (SMEs). Therefore, the approach in this chapter includes two different strategies, classification and regression for credit scoring and the expected return prediction respectively. Then, the contribution of each variable is found using the Shapley values and they are sorted based on the overall explanation scores. My model selection algorithm is proposed using these scores in a way that variables are removed one by one starting from the least explainable variable. Comparing my model with a simple Logistic regression model, I show that my approach performs better. The reason I used Shapley values to propose a model selection is that this explainable approach is widely used in the literature but it is usually applied to complex ML models to make their outputs more understandable. However, as a novelty of my study and also considering the importance of the trade-off between accuracy and explainability of the models

I utilized Shapley values as a tool to propose a model selection approach.

Chapter 4 uses XAI for a portfolio optimization problem in the cryptocurrency market. The output of an asset allocation problem is a set of weights corresponding to the including assets in the portfolio to optimize the purpose of investments. This allocation problem could be considered as a black box model because it cannot explain the outputs and show the general contributions of the assets to the output. To fill this gap and provide explanations for a portfolio optimization problem, I apply Shapley values to this problem. For this purpose, I consider daily information on 8 cryptos from 2017 to 2019. In this strategy, for each day, a portfolio is created based on the historical information of the last 30 days before the day on which the portfolio is created. Therefore, the portfolio and weights, are updated on a daily basis. After finding portfolios, their risk and return are calculated to be used in the calculation of the z score for each portfolio. In addition, the z score of each cryptocurrency is calculated. The reason that I use the Z score is to focus on both risk and return. The Z scores of cryptocurrencies and Z score of the portfolios are fed to a Random forest model as the predictors and target variable respectively. The output of this model is explained by Shapley values. Hence, the general contribution of Z score of each crypto to the Z score of the portfolios is found using the Shapley values. As a result, I can understand which cryptocurrencies are the most influential assets. The idea of this study is novel so there are many opportunities for future research such as the application of other more complex portfolio optimization and XAI methods. After using Shapley values in the previous chapters, I go in more depth and study how Shapley values can be improved. Therefore, chapter 5 improves the estimation of Shapley values using an instance-based method. Theoretically, to calculate Shapley values, different combinations of variables are considered to find how adding the variable of interest to the combination changes the outcome of the model. Then, the weighted average of these contributions among all the possible combinations will show the global Shapley value of the variable. When a variable is missing in the combination, a value should be set for it in the model because the machine learning model cannot predict the outcome with less number of variables compared with that of the training phase. The value that is set for the missing variable is usually selected randomly but in this paper, I use instance-based learning to find similar training observations to the test observation for which I find the corresponding contribution to the output of the model. Through the empirical analysis, applying the approach to the Lendingclub dataset, it is found that the proposed instance-based model can distinguish the variable contributions better than a simple SHAP approach in which value zero is set for the missing variables.

In addition to the explainability of the machine learning models, it is vital to check if these models are fair toward the observations in different groups based on the protected variable such as gender. In the last chapter, chapter 6, I discuss the Fairness of credit lending to individuals who asked for a loan from different states in the US. I check

how machine learning models are fair towards the applicants based on the place from which they applied for credit. For the purpose of decreasing the scale of the protected variable, United States, I group applicants to the four census regions using the information of their states. Therefore, finally, fairness is evaluated among “West”, “South”, “Midwest” and “Northeast” protected groups. In addition to checking fairness using statistical metrics such as confusion-based fairness measures, I check how models are fair based on Shapley values. Finally, considering the Gini index and p-values of the Kolmogorov-Smirnov (K-test), I show that models are not fair among the applicants from different regions. I also apply the propensity score matching (PSM) to the privileged and unprivileged groups to match the most similar observations from the opposite group to finally balance the train data in terms of the protected variable. Numerical results claim that this matching-based balancing method improves the fairness of the decisions of the model. This chapter, mentions the possibility of the application of the explainable AI methods to be used as fairness measures.

CHAPTER 1

A New Hybrid Instance-Based Learning Model for Decision-Making in the P2P Lending Market

Babaei, G., Bamdad, S. (2021). A new hybrid instance-based learning model for decision-making in the P2P lending market. Computational Economics, 57, 419-432.

1.1 Abstract

Peer-to-Peer (P2P) lending has grown rapidly in the past years. Therefore, borrowers and lenders are provided with the opportunity of lending and borrowing independently of the banks. Lenders in the P2P lending market can share their total investment amount among different loans, so making a decision may be difficult for inexpert lenders. The aim of this study is to propose a novel decision-making framework in which instance-based learning as a lazy learning method and artificial neural networks as an eager learning approach are integrated. The proposed hybrid instance-based learning (HIBL) model has the ability to predict the return and risk of new loans and help investors to find the optimal portfolio. In order to check the effectiveness of our model, we use a real-world dataset from one of the most popular P2P lending marketplaces, namely Lending Club. Moreover, a comparison among our proposed model and a rating-based method reveals that the proposed HIBL model can improve investments in P2P lending.

1.2 Introduction

In the last years, peer-to-peer (P2P) lending as a booming market has grown rapidly in many countries all around the world. P2P lending business originates in the United Kingdom, but nowadays there are P2P online platforms in a lot of countries even in the middle-east. In 2005, Zopa as the first P2P lending company was founded in the UK (Xia et al. (2017), Zhang and Chen (2017)).

Generally, P2P lending is an online service which lends money to individuals or businesses, so traditional banks do not have an intermediary role between lenders and borrowers (Lee and Lee 2012). Although P2P lending is a novel phenomenon, there is an increasing amount of literature concerning this topic (Chen et al. (2016), Gao et al. (2018), Ma et al. (2017), Wang et al. (2019), Yum et al. (2012)). Generally, the popular scopes that researchers have paid considerable attention are credit scoring (Guo et al. (2016), Malekipirbazari and Aksakalli (2015), Wang et al. (2018)), default risk (Emekter et al. (2015), Galindo and Tamayo (2000), Li et al. (2019), Ma et al. (2018)), and portfolio selection (Tan et al. (2017), Zhao et al. (2016)). When borrowers want to apply for a loan in the platform, they provide information about the loan and themselves. P2P lending platforms take into account this information and provide a risk rating for requests (Ye et al. 2018). The assumption of these scoring systems is that loans in the same level (with similar credit scores) bear the same level of risk. Such a grading system enables investors to build a portfolio based on their risk aversion. Machine learning is a subset of artificial intelligence and has been used by many researchers for different purposes (Chevallier et al. (2021), Mittal et al. (2019)) especially credit scoring and decision-making. Instance-based learning is a "lazy learning" method in which the training examples are stored and when a new instance is given to the model, similar instances are detected from the memory and the new sample is predicted by the use of these similar data. In contrast, there are many machine learning algorithms called "eager learning" methods; e.g. artificial neural networks (ANNs) and Decision Tree (DT). In these models, the target function is generally constructed when the training instances are available. In most studies, eager learners or model-based ones are utilized to provide new models to help investors to make a decision about their investments in P2P lending. Many researchers have indicated that combining lazy and eager learning methods produces better predictions than each single method (Quinlan (1993), Solomatine et al. (2008)). Therefore, our study that proposes a new decision-making model for P2P investors by combining an instance-based learning method with an eager approach can contribute to the current literature. In general, the focus of this article is combining the instance-based learning as a lazy learner with ANNs as an eager learner to propose a novel investment decision- making model in the P2P lending market. Specifically, in the proposed hybrid instance-based learning model, we first check the similarity of loans based on the difference between their probabilities of default, which are computed using a logistic regression model. Then, the optimal weights between training samples (closed loans) and new loans (listing) are derived from the kernel regression method. After that, a subset of closed loans with the biggest weights are extracted as the most similar loans to each listing to predict the return of listings using ANNs. The risk of each listing is predicted as a weighted variance of similar loans. Finally, we formulate investment decision-making in P2P lending as a portfolio optimization problem. To validate the proposed HIBL model, we apply real-world data from a pioneer P2P lending marketplace. Through the Empirical results, we conclude that the proposed HIBL model outperforms a current rating-based model.

The remainder of this paper is organized as follows: Sect. 2 presents the literature review. In Sect. 3 the description of the methods used in our study and the proposed algorithm are explained. Section 4 describes the selected data for our study and finally the results and discussion are provided in Sect. 5. Section 6 concludes this study.

1.3 Literature Review

Proposing decision-making models for P2P lending has been one of the popular topics among researchers in the last years. Guo et al. (2016) developed an investment decision-making model as a portfolio optimization problem and provided an instance-based credit risk assessment model for investors. The results of this study showed that this model can improve investment performances in P2P lending. Many researchers have utilized the idea of their study to improve P2P decision-making since then. Cho et al. (2019) utilized the instance-based entropy fuzzy support vector machine (IEFSVM) to create a decision-making model for P2P lending. Their proposed model predicts which loans would be fully paid and builds an investment portfolio with the predicted most profitable fully paid loans. The results of their empirical study reveal that the proposed IEFSVM has a better performance than current state-of-the-art classifiers. Babaei and Bamdad (2020*a*) also introduced a multi-objective instance-based decision-making model for the P2P investors that can estimate the return and risk of new loan requests using the closed loans in the past and optimize an investment portfolio based on two purpose of risk minimization and return maximization. Through their empirical study, it was concluded that their proposed algorithm leads to a better decision in comparison with two other profit scoring and single objective models considering the return and risk of the portfolio. In terms of combining the instance-based learning method with other models, we can mention the study done by (Quinlan 1993) as one of the first studies that integrated instance-based learning with other methods. He focused on the prediction of continuous values by applying both instance-based and model-based learning such as linear regression, model trees, and ANNs. Three composite models were compared with other single ones based on average and relative error amounts and finally concluded that in most cases the composite models performed better than other proposed methods. Cheng and Hüllermeier (2009) introduced a novel method for multi-label classification in which each instance can have multiple labels. For this purpose, they unified instance-based learning and logistic regression by considering the labels of the similar cases as the features of the logistic regression to predict the labels of the unseen cases. Through analyzing seven data sets from different domains, they claimed that their approach is able to improve predictive accuracy based on several evaluation criteria. The topic that has been paid attention to by few researchers is the evaluation of combining instance-based learning with other eager learners for decision-making especially in the P2P lending market.

1.4 Methodology

1.4.1 Instance-Based Learning

The basis of the instance-based learning method is the storage of data in the past and predict a future instance using these stored data (Hüllermeier 2003). One of the advantages of the instance-based learning is its simplicity. In fact, this method applies the neighborhood of a new instance and make a local approximation of that so it considers simple local approximations for complex target functions (Solomatine et al. 2008). The reason that this method is called lazy learning is that this approach learns by the simply storage of the observed samples and predictions are derived from the combination of the provided information by the stored data (Hüllermeier 2003). In contrast, eager learning methods require less storage and generate predictions indirectly. It means that the observed data are utilized to fit a model; e.g. regression models. Then, the target function is predicted using this fitted model. Generally, eager learners in comparison with lazy ones have higher computational costs during the process of training (Hüllermeier (2003), Quinlan (1993)).

1.4.2 Artificial Neural Networks

An artificial neural network (ANN) is an eager learner which is one of the most popular data-driven models. It mimics the human brain system and is consisted of a number of elements called neurons that are connected to each other by weight vectors. Feed-forward neural networks are the common networks that have been utilized successfully in many different works (Paliwal and Kumar (2011), Yang and Ma (2019)). Neurons are represented by circles and the weights vectors are depicted by arrows. Input neurons receive data and spread them through the layers of the network so the output values are the weighted sum of input values. In general, ANNs are eager learners that sometimes suffer from the problem of being encapsulated in software codes so they are not transparent enough. Instance-based learning is one of the methods that can solve the problem of transparency when predictions are based on the historical samples that are similar to the new inputs. As a result, integrating ANNs with an instancebased learning approach can be effective.

1.4.3 Hybrid Instance-Based Learning (HIBL) Model

First, in order to show how instance-based learning and ANNs are utilized together to propose a hybrid decision-making model for P2P lending, we explain the general process of this study in Fig. 1.

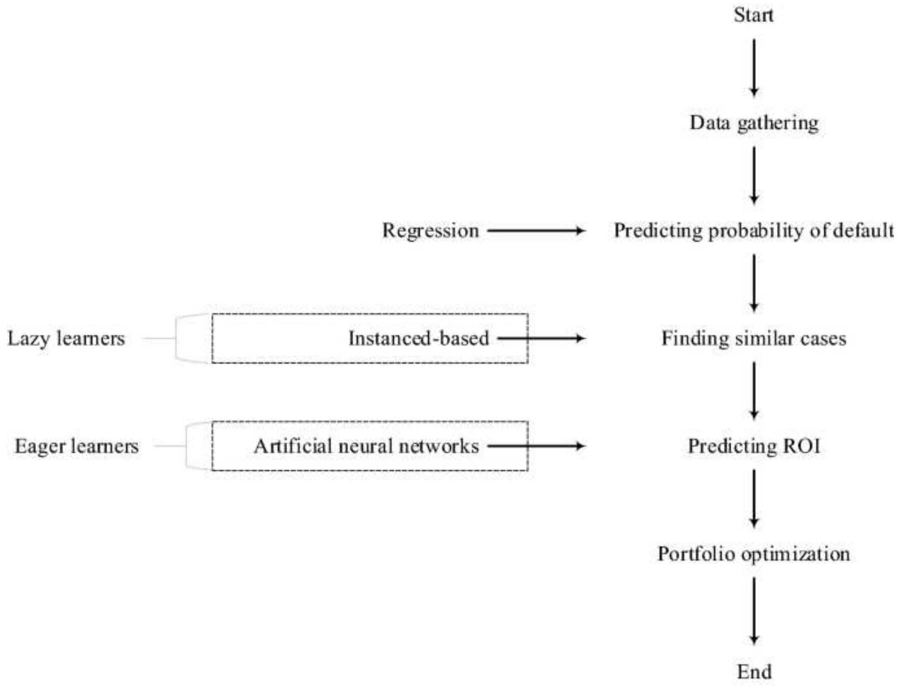


Fig. 1.1 Global Shapley values of the four census regions. These values show the contribution of each region to the prediction of our XGBoost classifier.

In order to create a portfolio, we need to predict the return and risk of listings so we use closed loans as the historical data and try to find similar loans and the relation between listings and closed loans. Each issued loan has an actual return $R_j(j = 1, \ldots, n)$ which is considered the return on investment (ROI) in our model. ROI is an effective performance measure that tries to indicate that if an investment project is really worth or not. The similarity of loans is computed based on the difference of their probability of default (PD) which is calculated using a logistic regression model as:

$$logit(\hat{PD}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_f X_f$$

(1.1)

Here, f indicates the number of independent features that is 17 in our study. After the calculation of PD for all loans, the distance between them is computed as follows:

$$d_{ij} = |PD_i - PD_j|$$

(1.2)

The less amount of d_{ij} shows that loans i and j are more similar and have a stronger relation and bigger weight W_{ij}. To find optimal weights using d_{ij}, we use kernel regression that is a statistical method to investigate the nonlinear relation between random variables as:

$$W_{ij} = \frac{K(\frac{d_{ij}}{h})}{\sum_{j=1}^{n} K(\frac{d_{ij}}{h})} \tag{1.3}$$

$K(.)$ is a kernel function. Considering the convenient mathematical properties of the Gaussian kernel function, we use it as:

$$K(Q) = (\frac{1}{\sqrt{2\pi}})e^{-\frac{Q^2}{2}} \tag{1.4}$$

In order to fit this regression model, we should find the optimal h that minimizes the following cross-validation error:

$$CV(h) = \frac{1}{n}\sum_{j=1}^{n}(\frac{\sum_{z=1,z\neq j}^{n} K(\frac{P_j - P_z}{h})R_z}{\sum_{z=1,z\neq j}^{n} K(\frac{P_j - P_z}{h})} - R_j)^2 \tag{1.5}$$

After finding the optimal h, W_{ij} can be calculated using Equation 2. For the purpose of building a hybrid model, after finding the weight matrix, closed loans with the biggest weights are chosen as the inputs for ANNs. Therefore, ROI_i of listings are predicted using an eager learner. For risk prediction, we predict the risk of a new loan i, σ_i^2, as the weighted variance among the closed loans in the past as:

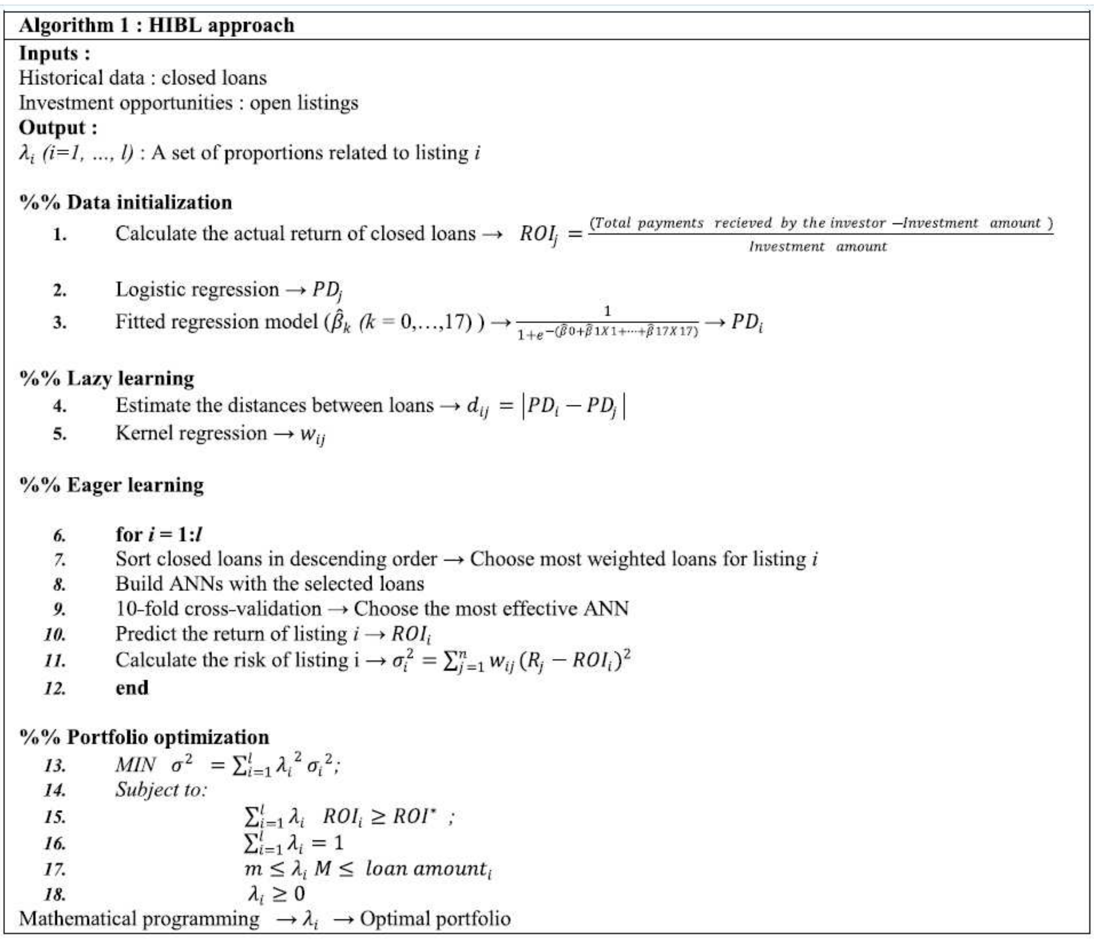
$$\sigma_i^2 = \sum_{j=1}^{n} W_{ij}(R_j - ROI_i)^2 \tag{1.6}$$

Here, ROI_i is the output of ANNs that shows the return of listing i. The last part of our HIBL model is the portfolio optimization which results in an investment recommendation based on the evaluated listings.

The basis of the portfolio optimization problem in our study is the modern portfolio theory developed by Harry Markowitz (1994). In this problem, we seek to minimize the risk of a portfolio for a given minimum expected return (ROI^*). In order to clarify the steps of the proposed model, we explain it in Algorithm 1. There are two input datasets: Closed loans whose pay-back statuses are known and open listings with unknown pay-back statuses. As the selected issued loans are closed we can compute the actual ROI of them. In addition, in terms of default, a logistic regression model is fitted based on closed loans with known statuses to find the PD_j. After fitting the regression model, the PD of open listings can be computed by the use of the extracted coefficients using the equation in the 3rd line of the model. For lazy learning process, distances between closed loans and listings are calculated based on the amounts of their PDs. Then, following the introduced steps in Sect. 4, we find out the amounts of W_{ij}. For each listing i, closed loans are sorted in descending order based on W_{ij} and most weighted loans are selected as inputs for ANNs; i.e. eager learning. For this purpose, we use a feed-forward ANN. One of the most important factors that has effects on the performance of an ANN is the number of hidden neurons so we use the ten-fold cross-validation method to find the optimal number of hidden neurons of ANNs for the prediction of ROI of each listing. Once the return of listings is computed, the risk of listing i is estimated as a weighted variance based on the mentioned equation in the 11th line of the model. After finding the return and risk of listings, we can solve the portfolio optimization problem to help investors to realize how much money they should allocate to each open listing. In this optimization problem, the risk of the portfolio is minimized for an expected level of return (ROI^*). λ_i is the decision variable which represents the optimal proportion of the total investment amount (M) allocated to the ith listing. In our study, M and ROI^* are 5000 and 0.2 respectively. While investors create their own portfolios in the P2P lending market, they are faced with some constraints on investment amounts. P2P lending platforms set a minimum investment amount (m) for their issued loans. For example, in Lending Club, m is equal to 25$. On the other hand, when an investor wants to select the ith loan, he or she can only lend an amount less than the ith loan amount.

1.5 Dataset Description

The information of loans issued by Lending Club is available in the Lending Club public dataset that is the world's most popular P2P lending network. Since we need to have finished loans to predict their return, we utilize 36-month loans issued in 2017 that consists 34,679 loans. Table 1 presents 17 independent variables in our study.

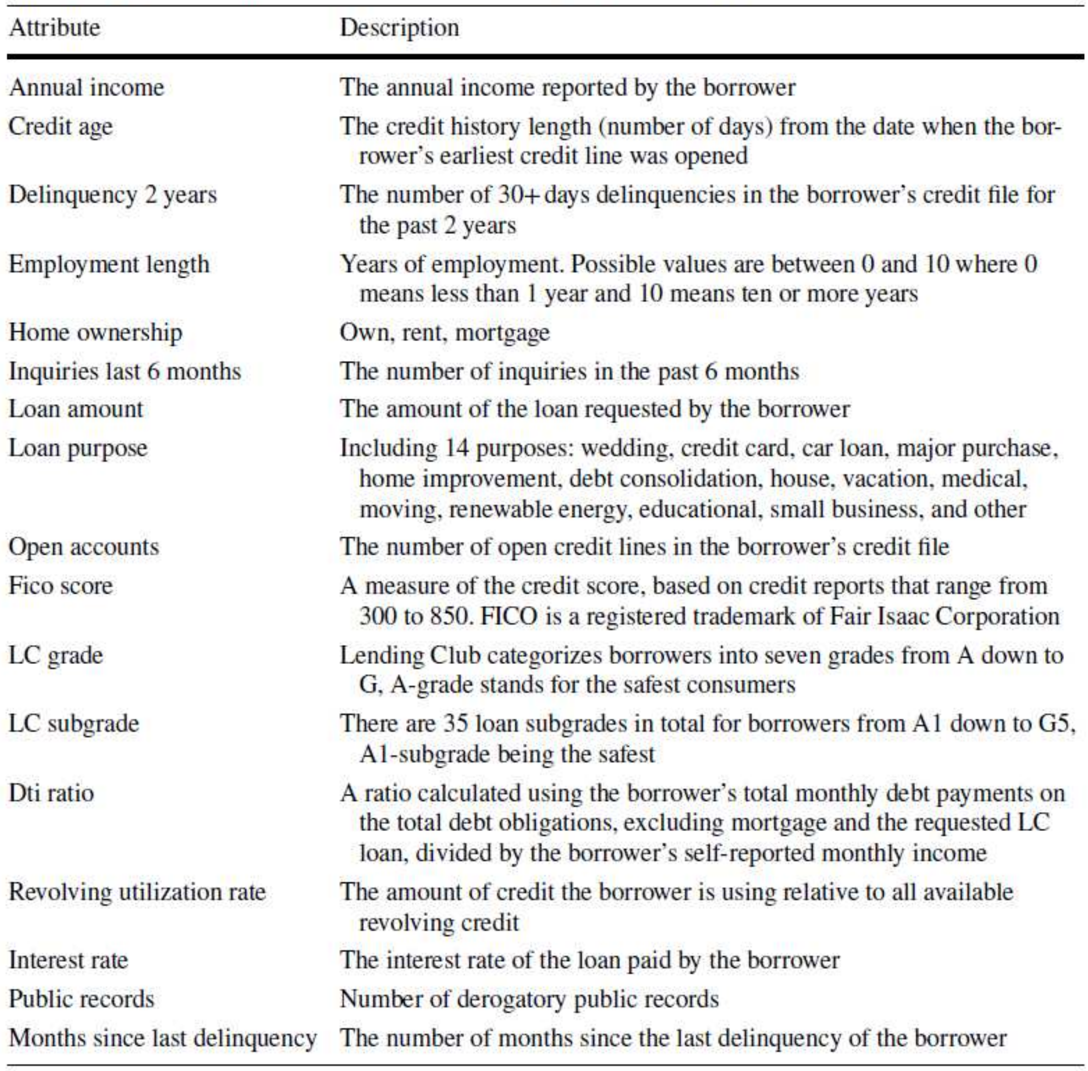


```
Algorithm 1 : HIBL approach
Inputs :
Historical data : closed loans
Investment opportunities : open listings
Output :
$\lambda_i$ (i=1, ..., l) : A set of proportions related to listing i

%% Data initialization
    Calculate the actual return of closed loans → $ROI_j = \frac{(Total\ payments\ recieved\ by\ the\ investor\ -Investment\ amount\ )}{Investment\ amount}$

    Logistic regression → $PD_j$
    Fitted regression model ($\hat{\beta}_k$ (k = 0,…,17) ) → $\frac{1}{1+e^{-(\hat{\beta}0+\hat{\beta}1X1+\cdots+\hat{\beta}17X17)}}$ → $PD_i$

%% Lazy learning
    Estimate the distances between loans → $d_{ij} = |PD_i - PD_j|$
    Kernel regression → $w_{ij}$

%% Eager learning

    for $i$ = 1:$l$
    Sort closed loans in descending order → Choose most weighted loans for listing i
    Build ANNs with the selected loans
    10-fold cross-validation → Choose the most effective ANN
    Predict the return of listing i → $ROI_i$
    Calculate the risk of listing i → $\sigma_i^2 = \sum_{j=1}^{n} w_{ij} (R_j - ROI_i)^2$
    end

%% Portfolio optimization
    $MIN\ \sigma^2 = \sum_{i=1}^{l} \lambda_i^{\ 2} \sigma_i^{\ 2};$
    Subject to:
                $\sum_{i=1}^{l} \lambda_i \ ROI_i \geq ROI^*$ ;
                $\sum_{i=1}^{l} \lambda_i = 1$
                $m \leq \lambda_i M \leq loan\ amount_i$
                $\lambda_i \geq 0$
Mathematical programming → $\lambda_i$ → Optimal portfolio
```

1.6 Results and Discussion

In order to evaluate the performance of the proposed HIBL model, we apply it to the selected dataset introduced in Sect. 4. 70% of the dataset is selected as the train data and the remaining for testing. First of all, the actual $ROI_j (j = 1, \ldots, n)$ and PD_j of closed loans are calculated using ANNs and a logistic regression model respectively. Then $PD_i (i = 1, \ldots, l)$ of test data can be predicted by the use of the extracted coefficients $\hat{\beta}_f (f = 0, \ldots, 17)$ from the fitted regression model in the Eq. 1. Using Equation 2, we understand the similarity between closed loans and open listings, and then find the weights among them. For this purpose, we find the optimal h that minimizes $CV(h)$. Then, we use these weights to detect much more similar loans to each listing and sort closed loans based on their weights decreasingly. Top loans are selected as the inputs for ANNs to predict the return of new requests. The performance of ANNs is related to the number of hidden neurons so we use the ten-fold cross-validation method. For this purpose, different ANNs with different amounts of neurons are compared based on the mean squared error (MSE). The results of this comparison are shown in Table 2. The best ANN for each listing is presented in the bold face. Finally, the selected top closed loans are fed to these best ANNs to predict ROI_i. Moreover, the risk of listings

Table 1 Independent variables (Adopted from Lending Club dictionary)



Attribute	Description
Annual income	The annual income reported by the borrower
Credit age	The credit history length (number of days) from the date when the borrower's earliest credit line was opened
Delinquency 2 years	The number of 30+ days delinquencies in the borrower's credit file for the past 2 years
Employment length	Years of employment. Possible values are between 0 and 10 where 0 means less than 1 year and 10 means ten or more years
Home ownership	Own, rent, mortgage
Inquiries last 6 months	The number of inquiries in the past 6 months
Loan amount	The amount of the loan requested by the borrower
Loan purpose	Including 14 purposes: wedding, credit card, car loan, major purchase, home improvement, debt consolidation, house, vacation, medical, moving, renewable energy, educational, small business, and other
Open accounts	The number of open credit lines in the borrower's credit file
Fico score	A measure of the credit score, based on credit reports that range from 300 to 850. FICO is a registered trademark of Fair Isaac Corporation
LC grade	Lending Club categorizes borrowers into seven grades from A down to G, A-grade stands for the safest consumers
LC subgrade	There are 35 loan subgrades in total for borrowers from A1 down to G5, A1-subgrade being the safest
Dti ratio	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income
Revolving utilization rate	The amount of credit the borrower is using relative to all available revolving credit
Interest rate	The interest rate of the loan paid by the borrower
Public records	Number of derogatory public records
Months since last delinquency	The number of months since the last delinquency of the borrower

is predicted based on Equation 6. In order to check the effectiveness of the proposed HIBL model, we compare it with a practical rating-based model explained in the study done by Guo et al. (2016) that is based on the credit grades set by the Lending Club platform. In this model, all loans are grouped into seven grades based on their credit risk amounts. First of all, the mean and standard deviation of the return of closed loans in each rating grade are calculated. Then, the mean and standard deviation from each credit grade will be used as the predicted return and risk of listings with the same grade, respectively. These estimated return and risk are used for the portfolio optimization problem.

Finally, the return and risk of the recommended portfolios by each model are compared. Table 3 shows the return, risk and the Sharpe ratio (Cho et al. 2019) of recommended portfolios by both models. It can be seen that our HIBL model can increase the portfolio return by more than 131.33% while the risk increases 50%. It means that by using HIBL, the return will increase significantly, but the investor should tolerate more risk. Also, the Sharpe ratio in our model is bigger than in the rating-based model. In general, the proposed HIBL model can improve the investment in the P2P lending market concerning return and bearing more risk of the portfolio. Figure 2 shows the results. In addition, in order to show the contribution of combining IBL as a lazy learner with eager learning, we remove IBL from our algorithm and use all instances for prediction and estimating the return and risk of the listings. Finally we conclude that by removing IBL the return of portfolio decreases by 36% while the risk of investment increases by 10% simultaneously. For the purpose of the performance evaluation of the model for different amounts of assumed variables (i.e. M and R^*), we run a sensitivity analysis that depicted in Table 4. It can be concluded that, by increasing M and keeping R^* as constant, the return and risk will increase. On the other hand, by rising of R^* and keeping M as constant, the return and risk will rise. But it can be seen that the increased value of return and risk are much more when R^* is constant. Based on this fact, it is obvious that the portfolio risk and return are much sensitive to different amounts of M rather than R^*.

Table 2 *MSE* of ANNs with different number of neurons

	Hidden neurons									
	5	10	15	20	25	30	35	40	45	50
MSE	0.00568	0.00525	0.00568	0.00533	0.00537	0.00539	0.00557	0.00549	0.00544	0.00575

Table 3 Return and risk of the recommended portfolios by models

Models	Portfolio return	Portfolio risk	Sharpe ratio
HIBL	0.4629	0.000015	119.52
Rating-based	0.2001	0.000010	63.28

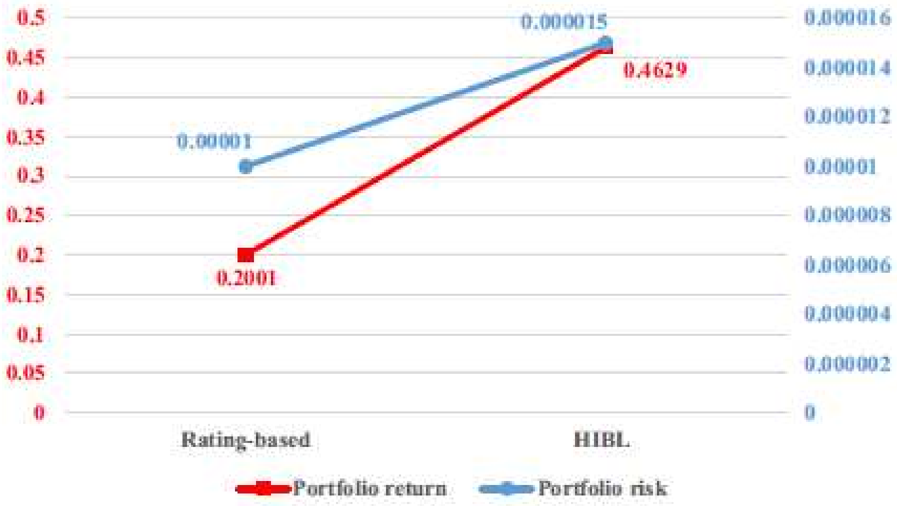


Fig. 2 Comparison of the HIBL and rating-based models

Table 4 Sensitivity analysis results

R* M	0.05	0.1	0.15	0.2
5000				
Return	0.3905	0.4132	0.4374	0.4629
Risk	0.000009	0.000011	0.000012	0.000015
100,00				
Return	0.5880	0.6111	0.6350	0.6598
Risk	0.000029	0.000032	0.000035	0.000039
15,000				
Return	0.4962	0.5189	0.5427	0.5675
Risk	0.000018	0.000020	0.000023	0.000026
20,000				
Return	0.4404	0.4462	0.4524	0.4590
Risk	0.000755	0.000757	0.000758	0.000760
25,000				
Return	0.5200	0.5248	0.5298	0.5350
Risk	0.001098	0.001098	0.001098	0.001098

1.7 Conclusion

The aim of this paper was to integrate instance-based learning as a lazy learner with artificial neural networks as eager learners to propose a hybrid investment decision-making model in the P2P lending market. In fact, the instance-based learning method was utilized to understand the relation between new loans (listings) and the issued ones in the past and select the most relevant instances to predict the future of the open loans for investment. Then, ANNs were applied to forecast the return of listings that was ROI in our study. The risk of each listing was computed as a weighted variance by the generated weights using kernel regression in the instance-based learning process. As investors not only need to choose loans, but also decide how much money to allocate, so we solved a portfolio optimization problem to help investors to find their optimal portfolio. For the evaluation of our study, we used Lending Club data set in 2017 and compared our proposed hybrid instance-based learning (HIBL) model with a practical rating-based method. The results revealed that our proposed hybrid model can improve investment decision-making in comparison with an available practical model in the P2P lending market. In addition, we showed the effectiveness of combining lazy with eager learning methods by removing IBL from our algorithm. Also, the sensitivity analysis showed a robust performance of our model in the P2P investment decision-making problem.

1.7.1 Investors' Insight

In terms of the real-world implication of the proposed HIBL model, investors with different amounts of money can improve their investments based on risk and return.

1.7.2 Future Research

For further research, other risk and return criteria can be used in the decision-making process for P2P lending. Also, applying the proposed model to datasets of other P2P platforms like Prosper and Zopa to examine the performance of the proposed model could be considered as a suggestion for future studies.

CHAPTER 2

Application of Credit-Scoring Methods in a Decision Support System of Investment for Peer-to-Peer Lending

Babaei, G., Bamdad, S. (2023). Application of credit-scoring methods in a decision support system of investment for peer-to-peer lending. International Transactions in Operational Research, 30(5), 2359-2373.

2.1 Abstract

Peer-to-peer lending, as a novel lending model, has challenged investors to make effective investment decisions. Issued loans are grouped into default and non-default. Therefore, different classification methods can be utilized to predict the status of loans in the future. Our study aim is to propose an investment decision model based on the non-default loans predicted using three different classifiers, including random forest (RF) that is a multitude of decision trees, support vector machine, and naive Bayes. In fact, we combine each of these classifiers with the portfolio optimization problem to understand which combination leads to the best portfolio concerning the risk and return. In order to find the best combined model, numerical studies are conducted based on real-world data. In addition, the performances of these classifiers are evaluated based on different performance measures using the 10-fold cross-validation method. The results indicate that combining the RF algorithm and portfolio optimization problem leads to the least level of risk. Moreover, it is concluded that RF is the best classifier in terms of performance analysis.

2.2 Introduction

The peer-to-peer (P2P) lending market contains individuals who lend to and borrow from each other using online platforms. Lenders review open loans for investment (listings) and may decide to fund many of them partially (Bastani et al. (2019), Zhang et al.

(2018)). P2P lending has recently become popular because not only borrowers can get loans for lower interest rates but also lenders receive a higher return in comparison with their investment in the traditional lending systems (Malekipirbazari and Aksakalli 2015). P2P loans are unsecured, and investors should bear all the risk of default if borrowers do not pay off loans. The level of mentioned risk is known as the probability of default (PD) and typically defined as how likely the borrower defaults on the loan. In order to help lenders to control PD, it is vital to assess the level of risk associated with each loan (Bastani et al. 2019). In general, loans are separated into two groups, namely default and non-default. Online P2P platforms not only decide to approve a listing but also utilize credit-scoring models to find the credit rate of borrowers. Therefore, loans are classified into predefined groups based on the attributes of the loan. However, these models do not only meet the needs of lenders in the P2P lending market because investors in this marketplace do not only decide which loans to fund but also compute the amount of money to allocate to each loan, which minimizes risk for a given expected return (Guo et al. 2016). In fact, investors face a portfolio optimization problem while investing in the P2P lending market. The portfolio problem is the process of computing the proportions of the initial asset that should be allocated to loans. Markowitz (1994) gave an initial answer to this problem and proposed the mean–variance model, which is the basis of the modern portfolio theory (Sawik 2012). In this study, we propose an investment decision model for investors in P2P lending, which in addition to select appropriate loans (i.e., non-default loans), has the ability to evaluate both risk and return of available listings and recommends corresponding investment amounts for each open listing. As classifiers perform differently, we utilize three classifiers that is, support vector machines (SVM), naive Bayes (NB), and random forest (RF) that is a multitude of decision tree (DT) algorithms. At first, we address the problem of imbalanced data in P2P lending. This problem occurs when the number of instances in the major class is more than the instances in the minor (Namvar et al. 2018). In the P2P lending platforms, non-default loans form the majority class, while the prediction of default loans is much more important for investors. Then, we apply the three mentioned classifiers to detect which listings would be non-default. After that, the risk and return of non-default listings are estimated. Finally, we try to find the optimal portfolio based on the predicted non-default loans. To the best of our knowledge, this study is one of the few studies that compares different classifiers concerning the return and risk of the portfolios that they can create. To validate the effectiveness of the proposed approach, extensive experiments are conducted using data from one of the largest P2P lending marketplaces in the United States, namely Lending Club (LC). The remainder of the paper is organized as follows. Literature review is presented in Section 2. Section 3 illustrates the credit-scoring method used in our model. Section 4 introduces the methodology of portfolio optimization. The procedure of our proposed algorithm is presented in Section 5.

Section 6 shows the empirical study. Finally, Section 7 concludes this work.

2.3 Literature review

The class imbalance problems are common in classification problems in the P2P lending context. Class imbalance occurs when the number of records in one class is very different from the records in another (Veganzones and Séverin 2018). Training classifiers based on such a data set leads to the ignorance of the minority class while the model has a bias toward the majority class. One of the most important methods for solving this issue is resampling (Haixiang et al. 2017). In a resampling approach, a balanced training data set is generated prior to building the classification model. The three types of resampling methods are oversampling, undersampling, and a hybrid of the two. The problem of the imbalanced data set in the P2P lending market has been mentioned in many studies in the past years (Namvar et al. 2018). Considering the computational time, the undersampling method is the best method for big data sets. An undersampling approach eliminates the records of the majority class randomly (Bastani et al. 2019). Namvar et al. (2018) combined various classifiers with resampling techniques. The credit from each combination was predicted. They concluded that the combination of RF and random undersampling is an effective strategy for calculating the credit risk of loans. In terms of credit-scoring, different statistical models such as artificial neural networks (ANNs) (West (2000), Babaei and Bamdad (2020*b*)), SVM (Lessmann et al. 2015), and K-nearest neighbors (KNN) (Abdelmoula 2015) have been used in many studies. Many researchers have compared classification methods. For example, Teply and Polena (2020) made a comparison among logistic regression (LR), linear discriminant analysis, SVM, ANNs, KNN, NB, Bayesian network, classification and regression tree (CART), and RF. According to their ranking, LR was the best. Malekipirbazari and Aksakalli (2015) compared different machine learning algorithms and identified the RF-based classification method best for predicting a borrower's status. In addition, the empirical results indicated that the proposed model based on RF outperformed the Fair Isaac Corporation (FICO) credit scores and LC grades for identifying good borrowers. Chang et al. (2015) compared the performance of different NB distributions and kernel methods for an SVM. They found that NB with Gaussian distribution and an SVM with the linear kernel are the best based on their performance. Tsai et al. (2014) used machine learning algorithms to classify and optimize the lending risk of each borrower. They used a modified version of LR, SVM, NB, and RF. It was concluded that logistic LR was the best classifier. We compare three common classification methods, namely, RF, SVM, and NB in a decision-making algorithm. All the above studies compared classifiers using many performance measurements. It means that they examined which classification method is the most accurate in order to classify loans into default and non-default. Therefore, the final results of the investors' portfolios were not evaluated. However, investors, in addition to selecting loans, need to decide how much money to be allocated to each listing, which presents a portfolio optimization problem. This feature of the P2P lending market has attracted many investors (Babaei and Bamdad 2021) in the past years. Guo et al. (2016) optimized a lender's decision in this financial market using a portfolio optimization problem with boundary constraints. They proposed an instance-based model that had the ability to measure the risk and return of listings based on the issued loans in the past. In order to help investors with finding the amount of investment in each listing, they minimized the portfolio risk for a given level of return. Xia et al. (2017) proposed a cost-sensitive boosted tree for loan evaluation, which considers the annualized rate of return (ARR) as the expected profitability for the assessment. Through their analysis, it was concluded that to evaluate models based on profitability, the area under the ARR curve is not a suitable metric. In the study done by Guo et al. (2016), LR was utilized to estimate the PD and then based on an instance-based approach, the P2P lending investment was considered as a portfolio optimization problem and solved to find the suitable amounts of investment of each available listing (open loan to be selected for the investment). In another paper by Xia et al. (2017), a cost-sensitive XGBoost method is used for the credit-scoring purpose and a portfolio optimization problem is solved using linear programming. The proposed model is evaluated based on the ARR. Therefore, the performance of the proposed algorithms in these studies is not compared using different machine learning methods to understand which machine learning method can improve the decisions made by the model. To fill this gap, we apply different machine learning models in a decision-making algorithm and examine their performance through the model to understand by using which of them better investments are proposed. NB has been widely used in many classification problems recently. For example, Pattekari and Parveen (2012) utilized NB to diagnose heart disease and make intelligent clinical decisions. Vedala and Kumar (2012) used soft and hard customer information in a P2P lending platform to propose a credit-scoring model based on a multirelational Bayesian classification method. In the study done by Shen, Wang and Shen (2020), different classifiers, including DT, were used for credit-scoring purpose. The empirical results showed the acceptable performance of DT in comparison with other applied methods. Considering these papers, we use RF (instead of DT) due to its highly accurate performance shown by many studies, NB, and SVR in a decision-making model. A grid search approach is utilized for hyperparameter tuning, and after finding the best estimators, we find the predicted non-default loans for the credit-scoring part of the paper. To make this procedure of this paper clearer, we should say that, first of all, three classifiers are utilized to predict non-estimated loans. Then these extracted records are transferred to the following stages of the study and are analyzed based on both risk and return aspects. Finally, a portfolio optimization problem is solved using these selected loans as the inputs and available investment opportunities. As mentioned above, we combine different classifiers with portfolio optimization to come up with all needs of investors (i.e., selecting appropriate loans and find the suitable investment amount that should be allocated to each selected loan).

2.4 Credit-scoring methods

A vital part of long-term success for the P2P lending market as a booming business is a proper credit-scoring model. In general, this helps to determine whether credit should be given to a borrower. There are different classification techniques for creating credit-scoring models. In this study, we utilize SVM, NB, and RF algorithms.

2.4.1 Support vector machines

In SVM-based models, binary classified data are separated by a hyperplane such that the margin width between the hyperplane and the examples is maximized. It is revealed in statistical learning theory that maximizing the margin width leads to reducing the complexity of the model, consequently reducing the expected general risk of error. The SVM optimization problem can be defined algebraically as a dual form mathematical programming problem as follows:

$$max(\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j k(x_i, x_j) \tag{2.1}$$

Here α_i is a Lagrange multiplier for each training sample i. C controls the cost of the goal. In addition, class label $y_i \in \{-1, +1\}$, in which $i = 1, 2, \ldots, n$. The kernel function k can be used to find the nonlinear relation of the data.

2.4.2 Naive Bayes

In the Bayesian classifiers, the most likely class is assigned to a given example described by its feature vector. In order to simplify the learning of such classifiers, it can be assumed that features are an independent given class, that is, $P(XC) = \prod_{i=1}^{n} P(X_i C)$, where $X = (X_1, \ldots, X_n)$ is a feature vector and C is a class. Despite this unrealistic assumption, NB is remarkably successful in practice. Let $X = x_1, \ldots, x_n$ be a vector of features, where each feature takes value from its domain, D_i. Also, $\Omega = D_1 \ldots D_n$ denotes the set of all feature vectors. Let C be an unobserved feature indicating the class of an example and C is in the range of $\{0, \ldots, m-1\}$. A function $g : \Omega \rightarrow \{0, \ldots, m-1\}$ is defined, where $g(X) = C$ indicates a concept to be learned. A classifier is explained by function $h : \Omega \rightarrow \{0, \ldots, m-1\}$ that allocates a class to any given example. The Bayes classifier $h^*(x)$ utilizes as discriminant functions the posterior class probabilities given a feature vector, that is, $f_i^*(x) = P(C = i, X = x)$. Using Bayes rule, $P(C = i, X = x) = \frac{P(X=x,C=i)P(C=i)}{P(X=x)}$, where $P(X = x)$ is equal in all classes, and hence can be ignored. Therefore, the Bayes classifier $h^*(x) = argmax_i P(X = x, C = i)P(C = i)$ detects the maximum posterior probability hypothesis regarding example x. By simplifying the assumption that features are independent, the NB classifier is explained by discriminant functions $f_i(x) = \prod_{j=1}^{n} P(X_j = x_j, C = i)P(C = i)$.

2.4.3 Decision tree and random forest

Through the classification procedure of the DT method, first, in order to make the homogeneity of default risk in the subset higher than original sets, the data are divided into subsets. This division process continues until the new subsets meet the requirements of the end node. Three main construction parts exist in a DT called bifurcation, stopping, and deciding rules. Bifurcation rules are for the division of the new subsets. Moreover, stopping rules indicate that the subset is an end node or not. A DT model includes a target variable, Y, and continuous variables, x_i. The primary portions of a DT model are nodes and branches and the steps in building a model are splitting, stopping, and pruning. There are three types of nodes including (i) root nodes, which represents a choice that will lead to the division of all records into two or more unique subsets; (ii) internal nodes that show one of the possible options at that point in a tree structure; (iii) leaf nodes show the end result of a combination of decisions. Branches show chance outcomes that originate from root nodes and internal nodes. The DT model is formed using a hierarchy of branches. Each path from the root node through the internal nodes to the leaf node represents a classification decision rule. Only input variables related to the target variable are used to divide the main nodes into child nodes that are purer than the target variable. Properties associated with the degree of purity of child-derived nodes (e.g., relative to target conditions) are used to select between different potential input variables. These features include entropy, Gini index, classification error, information acquisition, profit ratio, and doubling criteria. To avoid overfitting, stop rules must be applied when building the DT. Common parameters used in stop laws are (i) the minimum number of records per leaf; (ii) the minimum number of records in a node before division; and (iii) the depth (e.g., number of steps) of each leaf of the root node. A common method for selecting the best possible subcategory from multiple candidates is to consider the proportion of records with error prediction. There are two types of pruning: prepruning and postpruning. Prepruning uses chi square tests or multiple comparison adjustment methods to prevent the formation of non significant branches. After producing a complete DT, postpruning is used to remove the branches in a way that improves the overall classification accuracy when applied to the validation data set. There are several statistical algorithms for building DTs, including CART, C4.5, chi-squared automatic interaction detection, and quick, unbiased, efficient, statistical tree. RF is a learning method that operates by constructing a multitude of DTs in training phase. In classification, the RF's output is a class selected by most trees.

2.5 Portfolio optimization

The traditional banking sector can only approve or reject a request. Therefore, in such a system, loan evaluation predicts only a label for each loan. However, investing process in P2P lending can be regarded as a portfolio selection problem (Xia et al. 2017). According to modern portfolio theory (Markowitz 1994), investors aim to achieve a portfolio that the expected risk is minimized for a given expected return. The portfolio optimization problem in this study is as follows:

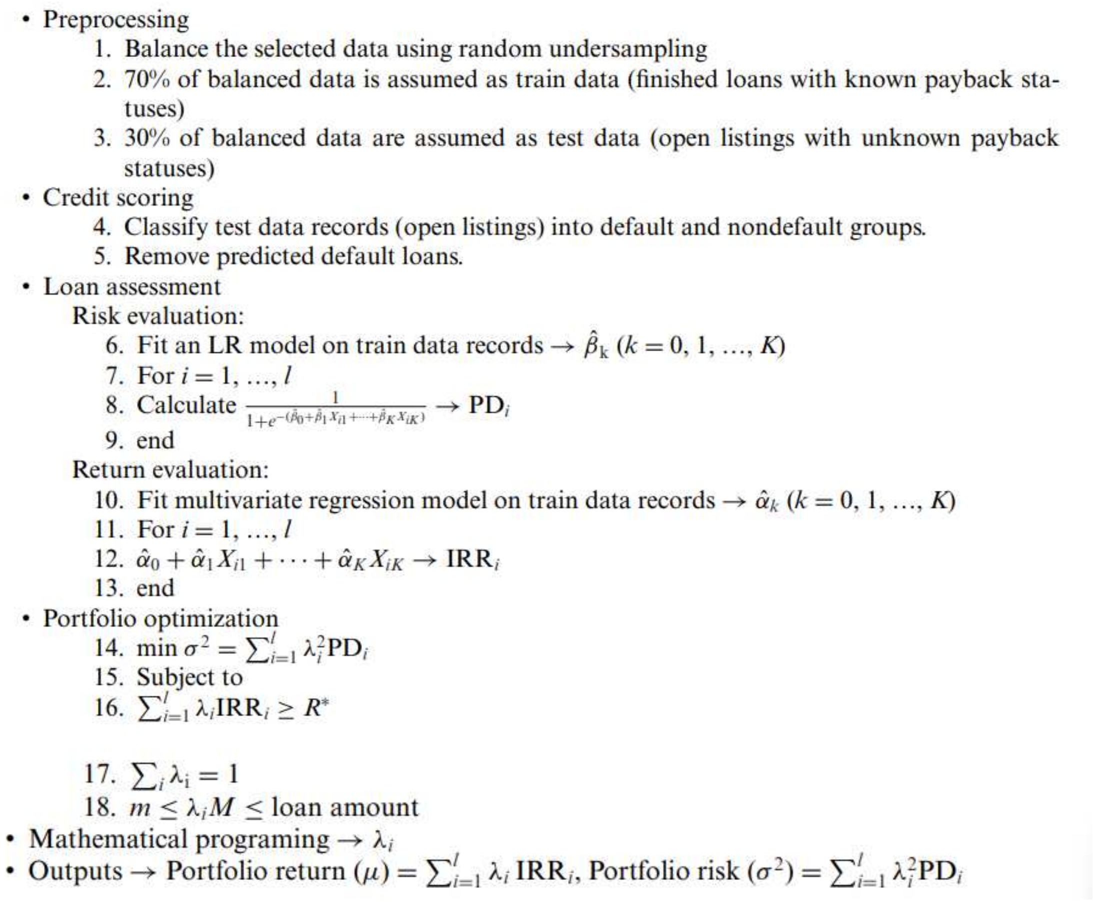
$$\min \sigma^2 = \sum_{i=1}^{l} PD_i subjectto \sum_{i=1}^{l} \lambda_i IRR_i \tag{2.2}$$

$$\begin{aligned}
\text{min} \qquad & \sigma^2 = \sum_{i=1}^{l} PD_i \\
\text{subject to} \quad & \sum_{i=1}^{l} \lambda_i IRR_i \succeq R^*, \\
& \sum_{i} \lambda_i = 1, \\
& m \leq \lambda_i M \leq loanamount.
\end{aligned}$$

Here, λ_i is the decision variable and shows the corresponding investment amount that investors should allocate to the ith listing $(i = 1, \ldots, l)$. PD, which has been used by many researchers as the risk metric (Guo et al. (2016), Bastani et al. (2019)), represents the predicted risk of each available listing. Equation 2.2 indicates that the investors have an expected return (R^*) and wants to find an optimal portfolio with a minimum return level equal to R^*. IRR_i shows the expected return of the ith listing. This equation also ensures the investment of all funds. Most P2P lending platforms have a minimum investment amount (m) for every loan (e.g., \$25 in the LC). Moreover, it is not possible that a lender who has a total investment amount (M) devotes much more money than the required money by the borrower (loan amount).

2.6 Proposed algorithm

The purpose of this research is to propose an algorithm that, in addition to improving the lenders' decisions, investigates the performance of different classification methods in a P2P lending decision algorithm. For this purpose, SVM, NB, and RF are applied. In the end, the return and risk of portfolios built by each model are compared to understand which classifier can improve an investor's decision. The inputs of our model contain the selected data set, the minimum investment amount (m), which is equal to \$25 in our study, and a total investment amount (M), assumed to be equal to \$15,000. The proposed algorithm is as follows:



```
• Preprocessing
        1. Balance the selected data using random undersampling
        2. 70% of balanced data is assumed as train data (finished loans with known payback statuses)
        3. 30% of balanced data are assumed as test data (open listings with unknown payback statuses)
• Credit scoring
        4. Classify test data records (open listings) into default and nondefault groups.
        5. Remove predicted default loans.
• Loan assessment
    Risk evaluation:
        6. Fit an LR model on train data records → $\hat{\beta}_{\mathrm{k}}$ ($k = 0, 1, \ldots, K$)
        7. For $i = 1, \ldots, l$
        8. Calculate $\frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X_{i1}+\cdots+\hat{\beta}_K X_{iK})}} \rightarrow \mathrm{PD}_i$
        9. end
    Return evaluation:
        10. Fit multivariate regression model on train data records → $\hat{\alpha}_k$ ($k = 0, 1, \ldots, K$)
        11. For $i = 1, \ldots, l$
        12. $\hat{\alpha}_0 + \hat{\alpha}_1 X_{i1} + \cdots + \hat{\alpha}_K X_{iK} \rightarrow \mathrm{IRR}_i$
        13. end
• Portfolio optimization
        14. min $\sigma^2 = \sum_{i=1}^{l} \lambda_i^2 \mathrm{PD}_i$
        15. Subject to
        16. $\sum_{i=1}^{l} \lambda_i \mathrm{IRR}_i \geq R^*$

        17. $\sum_i \lambda_\mathrm{i} = 1$
        18. $m \leq \lambda_i M \leq$ loan amount
• Mathematical programing → $\lambda_i$
• Outputs → Portfolio return ($\mu$) = $\sum_{i=1}^{l} \lambda_i$ IRR$_i$, Portfolio risk ($\sigma^2$) = $\sum_{i=1}^{l} \lambda_i^2 \mathrm{PD}_i$
```

In the preprocessing stage of our algorithm, the imbalanced data are balanced using the random undersampling method. It means that some records in the majority class are removed to make a balance between the minority and majority classes. The balanced data are separated into train and test data set with a ratio of 70:30, respectively. Then, data are grouped into default and non-default using the classifiers to remove the predicted default loans in the credit-scoring stage. To evaluate the risk of available listings that are predicted to be non-default, an LR model is fitted on finished loans because their loan statuses are clear. As a result, the PD of open listings can be computed using the equation in line 7. On the other hand, for return evaluation, we fit a multivariate regression model. The response variable is the actual IRR of loans in the train data set computed using the IRR function in Excel software. After that, the returns of predicted non-default listings are estimated. The next stage of our proposed algorithm is portfolio optimization. In this stage, the mentioned model in Section 4 is implemented. R^* is equal to 0.03 in our numerical study. Through mathematical programming, λ_i is found. Therefore, investors can estimate the risk and return of their selected portfolios.

2.7 Empirical study

The following chapters elaborate on the data set used to validate of the proposed model and how we run the mentioned algorithm for real data.

2.7.1 Data description

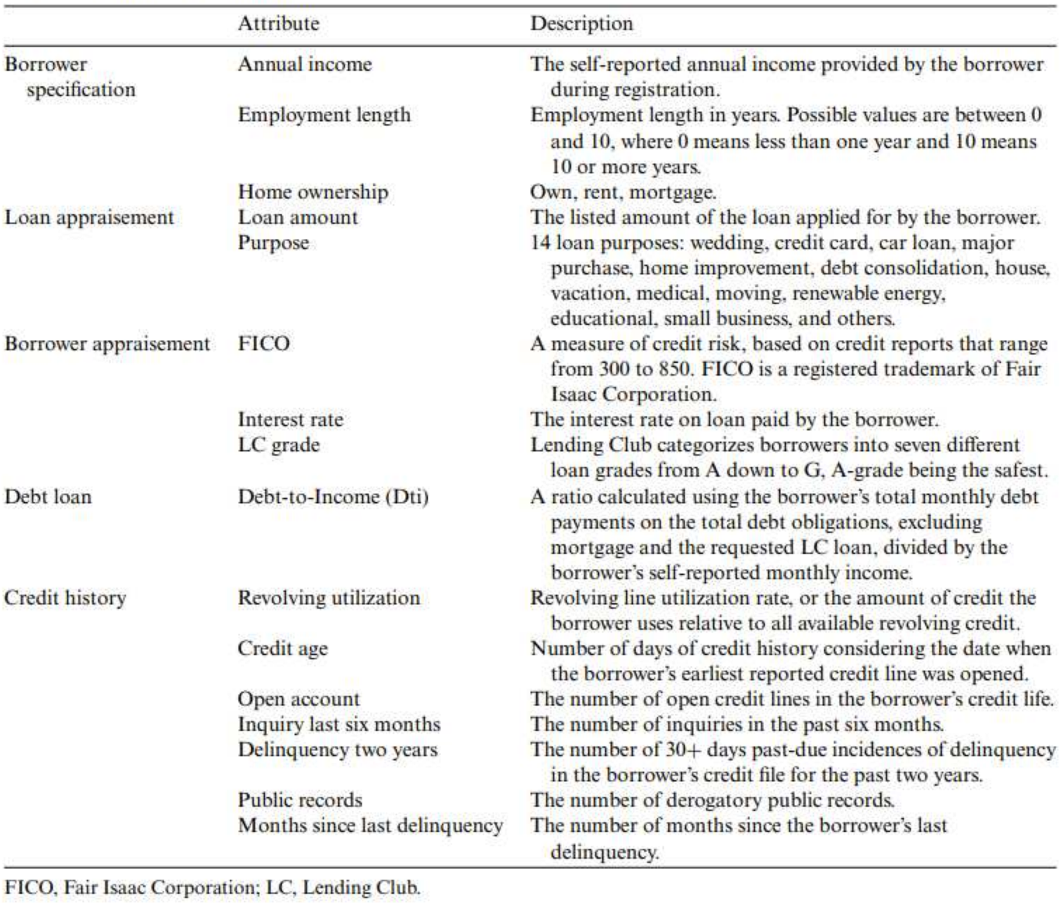
In this paper, the data set is adopted from the LC. Selected loans for this study contain 36-month loans issued during 2007–2018. We chose this time horizon because, in one part of our algorithm, we need to compute the actual IRR of loans, so we should just put finished loans in our selected data set. Table 1 presents the selected features, which are grouped into five categories: (1) borrower specification, (2) loan appraisement, (3) borrower appraisement, (4) debt loan, and (5) credit history. These 16 features form the independent variables.

2.7.2 Data analysis

Table 2 provides descriptive statistics of continuous variables. As expected, the mean interest rate in the default class is higher than in the non-default class. In the LC platform, the interest rate of listings is set based on the grades of the borrowers. Applicants with higher LC grades have lower interest rates, while riskier borrowers with lower grades have bigger interest rates. The distribution of LC grades is presented in Fig. 1. The most common grade is “B” in our selected data set.

The discrete variables, namely “Purpose,” “Grade,” and “Home ownership” are transformed into dummy variables. Also, to detect the multicollinearity problem, we utilize the variance inflation factor (VIF), which shows the collinearity between the independent variables with the others in the model. In our study, VIFs greater than 10 are assumed as the indicators of high collinearity and the associated variables are removed from the model.

Table 1
The selected features for our model



	Attribute	Description
Borrower specification	Annual income	The self-reported annual income provided by the borrower during registration.
	Employment length	Employment length in years. Possible values are between 0 and 10, where 0 means less than one year and 10 means 10 or more years.
	Home ownership	Own, rent, mortgage.
Loan appraisement	Loan amount	The listed amount of the loan applied for by the borrower.
	Purpose	14 loan purposes: wedding, credit card, car loan, major purchase, home improvement, debt consolidation, house, vacation, medical, moving, renewable energy, educational, small business, and others.
Borrower appraisement	FICO	A measure of credit risk, based on credit reports that range from 300 to 850. FICO is a registered trademark of Fair Isaac Corporation.
	Interest rate	The interest rate on loan paid by the borrower.
	LC grade	Lending Club categorizes borrowers into seven different loan grades from A down to G, A-grade being the safest.
Debt loan	Debt-to-Income (Dti)	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
Credit history	Revolving utilization	Revolving line utilization rate, or the amount of credit the borrower uses relative to all available revolving credit.
	Credit age	Number of days of credit history considering the date when the borrower's earliest reported credit line was opened.
	Open account	The number of open credit lines in the borrower's credit life.
	Inquiry last six months	The number of inquiries in the past six months.
	Delinquency two years	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past two years.
	Public records	The number of derogatory public records.
	Months since last delinquency	The number of months since the borrower's last delinquency.

FICO, Fair Isaac Corporation; LC, Lending Club.

2.7.3 Investment decision

After preparing the data set for training the model, we apply different classifiers, including Gaussian NB, SVM, and RF, to group observations to default and non-default classes. As the hyperparameter tuning has significant effects on the performance of most machine learning algorithms, we applied the grid search approach to find the best estimators of the classifiers. After training the models with optimal hyperparameters, these classifiers are compared based on the accuracy measure to better understand their performance. The results are shown in Table 3.

Despite many studies in which different machine learning models are compared only based on the performance measures, we aim to compare these classifiers in a decision-making framework and compare the decisions generated by different classifiers. As mentioned in Section 5, we apply each of these classification methods to our data set to predict which observations would be non-default and use them for the rest of the algorithm. After detecting the non-default samples, their return and risk are predicted using multivariate and LR models, respectively. The empirical results of this part of our proposed algorithm are presented in Table 4.

Considering p-values, just significant variables are used to fit regression models.

Table 2
Descriptive statistics on continuous variables

	Nondefault				Default			
Variable	Mean	SD	min	max	Mean	SD	min	max
Loan amount	9585.468	5175.177	1000	35,000	10,097.47	6258.547	1600	35,000
Interest rate	0.150364	0.032199	0.06	0.2499	0.161699	0.031575	0.079	0.234
Annual income	70,954.67	41,243.45	13,500	440,000	74,817.8	62,782.09	17,000	400,000
Debt to income	15.0659	7.627291	0.86	34.99	16.12926	7.46078	2.27	32.12
Delinquency two years	0.414778	1.056761	0	10	0.608108	1.763713	0	18
Credit age	6861.226	2645.358	2434	19,511	6440.426	2409.12	2314	14,428
FICO score	673.3448	12.48231	662	737	674.3649	14.95397	662	737
Inquiry last six months	0.990148	1.103116	0	6	1.175676	1.221776	0	6
Months since last delinquency	43.07094	21.00124	0	149	41.08108	22.29228	3	83
Open account	10.85714	4.160522	2	35	11.5473	4.435991	5	27
Public records	1.250246	0.723266	1	8	1.324324	0.682295	1	5
Revolving utilization	0.489749	0.204633	0	0.969	0.493459	0.212439	0	0.93

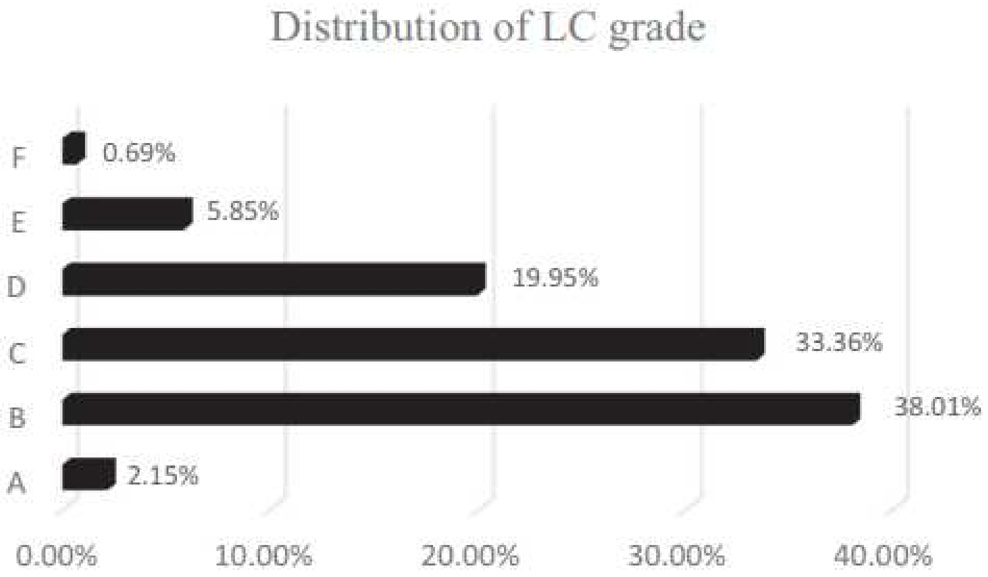


Fig. 1. Distribution of loans in different LC credit groups.

Table 3
Accuracy of classifiers

Method	Accuracy
Gaussian NB	0.851367
SVM	0.872241
RF	0.875922

RF, SVM, and NB approaches in line with the classical model are utilized for the credit-scoring part. In the classical model, statuses of listings are not predicted and all of them are selected for the portfolio. The results of each scenario are shown in Table 5.

The portfolio with the lowest level of risk is achieved in the RF scenario. In this scenario, the statuses of open listings for investment are predicted using the RF algorithm. After that, the predicted non-default loans are selected for the investor's portfolio. To validate the effects of the classifiers on our algorithm, we run the proposed model without them (i.e., classical model). It means that all available listings are chosen for the creation of the portfolio. The results show that the risk of the best portfolio in the RF scenario is 41.6% lower than the portfolio risk in the classical model.

In general, it can be concluded that applying classifiers in our proposed algorithm leads to safer investments.

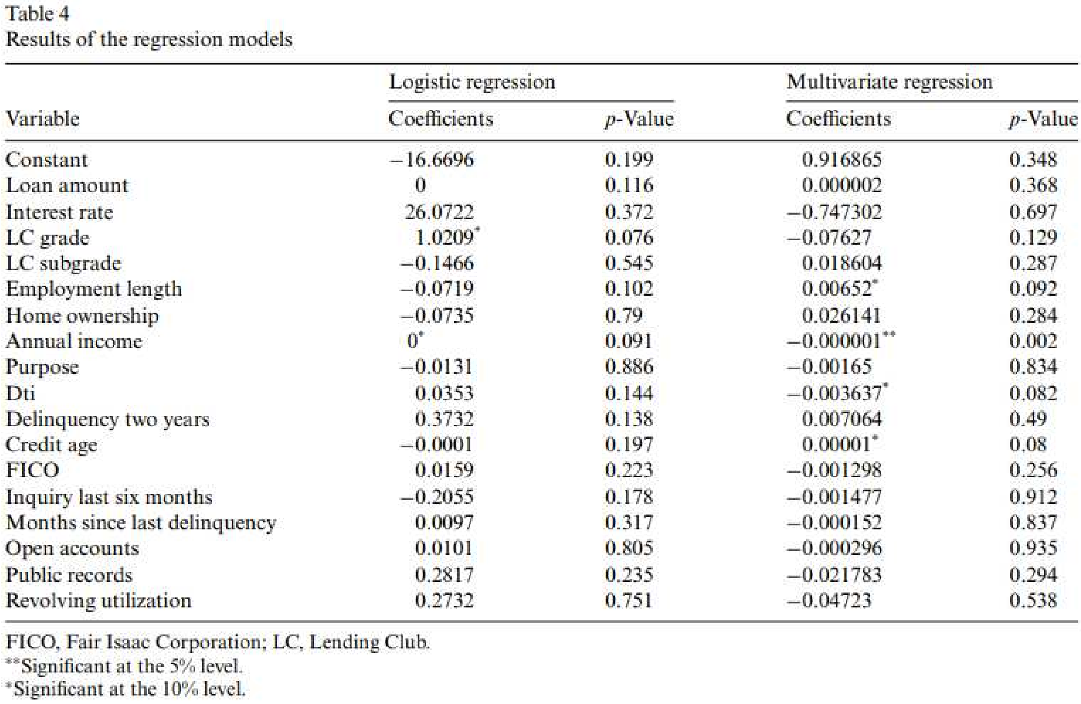


Table 4
Results of the regression models

	Logistic regression		Multivariate regression	
Variable	Coefficients	*p*-Value	Coefficients	*p*-Value
Constant	−16.6696	0.199	0.916865	0.348
Loan amount	0	0.116	0.000002	0.368
Interest rate	26.0722	0.372	−0.747302	0.697
LC grade	1.0209*	0.076	−0.07627	0.129
LC subgrade	−0.1466	0.545	0.018604	0.287
Employment length	−0.0719	0.102	0.00652*	0.092
Home ownership	−0.0735	0.79	0.026141	0.284
Annual income	0*	0.091	−0.000001**	0.002
Purpose	−0.0131	0.886	−0.00165	0.834
Dti	0.0353	0.144	−0.003637*	0.082
Delinquency two years	0.3732	0.138	0.007064	0.49
Credit age	−0.0001	0.197	0.00001*	0.08
FICO	0.0159	0.223	−0.001298	0.256
Inquiry last six months	−0.2055	0.178	−0.001477	0.912
Months since last delinquency	0.0097	0.317	−0.000152	0.837
Open accounts	0.0101	0.805	−0.000296	0.935
Public records	0.2817	0.235	−0.021783	0.294
Revolving utilization	0.2732	0.751	−0.04723	0.538

FICO, Fair Isaac Corporation; LC, Lending Club.
**Significant at the 5% level.
*Significant at the 10% level.

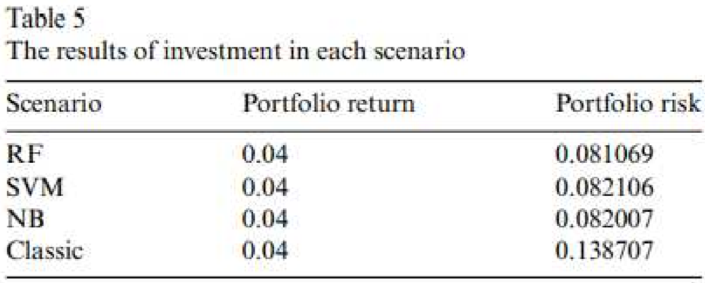


Table 5
The results of investment in each scenario

Scenario	Portfolio return	Portfolio risk
RF	0.04	0.081069
SVM	0.04	0.082106
NB	0.04	0.082007
Classic	0.04	0.138707

2.7.4 Evaluation metrics

It was concluded in Section 6.3 that the application of classification methods could result in safe investments. This section intends to compare the performance of all utilized classifiers using the 10-fold cross-validation. Table 6 is a confusion matrix that describes the performance of a classification model. We consider seven performance indicators in our study: accuracy, sensitivity, specificity, G-mean, the area under the receiver operating characteristic curve, precision, and F-measure.

We chose these indicators because they are common performance indicators that have been used in many studies (Lessmann et al. (2015), Cho et al. (2019)). Table 7 presents the performance measurements of the three classification methods. We use boldface to highlight the best performing classifier per performance measure. It is clear that in terms of five of seven indicators, RF shows the best performance.

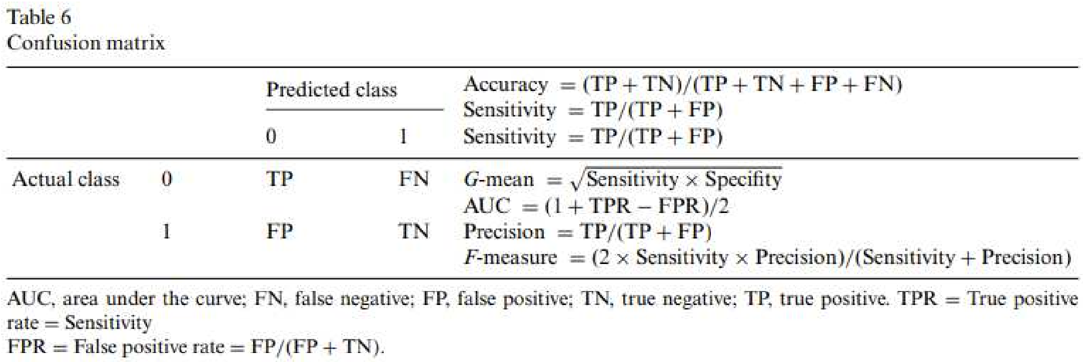


Table 6
Confusion matrix

		Predicted class		Accuracy $= (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN})$
				Sensitivity $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
		0	1	Sensitivity $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
Actual class	0	TP	FN	G-mean $= \sqrt{\text{Sensitivity} \times \text{Specifity}}$
				AUC $= (1+\mathrm{TPR}-\mathrm{FPR})/2$
	1	FP	TN	Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
				F-measure $= (2 \times \text{Sensitivity} \times \text{Precision})/(\text{Sensitivity}+\text{Precision})$

AUC, area under the curve; FN, false negative; FP, false positive; TN, true negative; TP, true positive. TPR = True positive rate = Sensitivity
FPR = False positive rate = FP/(FP + TN).

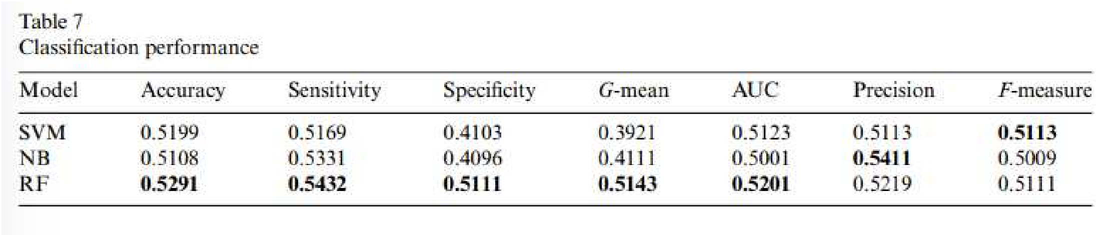


Table 7
Classification performance

Model	Accuracy	Sensitivity	Specificity	G-mean	AUC	Precision	F-measure
SVM	0.5199	0.5169	0.4103	0.3921	0.5123	0.5113	**0.5113**
NB	0.5108	0.5331	0.4096	0.4111	0.5001	**0.5411**	0.5009
RF	**0.5291**	**0.5432**	**0.5111**	**0.5143**	**0.5201**	0.5219	0.5111

2.8 Conclusion

This paper aimed to evaluate the effects of the classification approaches on decision making in P2P lending. First of all, we addressed the problem of imbalanced data using the undersampling method. Three different classification methods, that is, RF, SVM, and NB, were utilized to predict the class of available listings. After the categorization, a portfolio selection problem was optimized for those loans with nondefault predicted labels. In this optimization problem, the portfolio risk was minimized for a minimum expected return. To evaluate our algorithm, we used the LC data set, one of the most popular platforms in the world. The results of the numerical study demonstrated that applying credit-scoring methods in the decision-making algorithm reduced the risk of investment. Specifically, using the RF in combination with the portfolio optimization problem led to the least level of investment risk. Considering the evaluation metrics, our proposed algorithm was robust for different amounts of predefined variables. To direct future studies, we can point to the uncertainty in the proposed algorithm. In the real world, many parameters are not fixed and have a degree of uncertainty. Therefore, the algorithm presented in this research can be developed by considering different approaches to uncertainty, such as fuzzy theory.

CHAPTER 3

Explainable Artificial Intelligence for Crypto Asset Allocation

Babaei, G., Giudici, P., Raffinetti, E. (2022). Explainable artificial intelligence for crypto asset allocation. Finance Research Letters, 47, 102941.

3.1 Abstract

Many investors have been attracted by Crypto assets in the last few years. However, despite the possibility of gaining high returns, investors bear high risks in crypto markets. To help investors and make the markets more reliable, Robot advisory services are rapidly expanding in the field of crypto asset allocation. Robot advisors not only reduce costs but also improve the quality of the service by involving investors and make the market more transparent. However, the reason behind the given solutions is not clear and users face a black-box model that is complex. The aim of this paper is to improve trustworthiness of robot advisors, to facilitate their adoption. For this purpose, we apply Shapley values to the predictions generated by a machine learning model based on the results of a dynamic Markowitz portfolio optimization model and provide explanations for what is behind the selected portfolio weights.

3.2 Introduction

Cryptocurrency markets are a recent choice for investors. An important reasons for the emerging of this new financial market is the collapse of the banking sector in 2008 and the following low interest rate environment. After the launch of the first cryptocurrency, the Bitcoin, several other cryptoassets have been introduced by improving the existing technology (Chokor and Alfieri 2021).

Recently, many scholars and investors have been attracted by cryptocurrencies. From a research point of view, the application of machine learning (ML) models (Akyildirim et al. (2021), Alessandretti et al. (2018)), network models (Giudici and Polinesi

2021), and portfolio optimization models (Jiang and Liang 2017) for cryptocurrencies are new established topics.

In particular, Derbentsev et al. (2020) utilized some ML algorithms such as support vector machines, logistic regression, artificial neural networks, and random forests to analyze the predictability of cryptocurrencies at different frequencies, such as daily and minute levels. Comparing the results and performance of these algorithms they concluded that support vector machines can achieve the most accurate results compared to the logistic regression, artificial neural networks and random forest classification algorithms. Another related paper is Valencia et al. (2019) who used available social media data to predict the price movements of cryptocurrencies including Bitcoin, Ethereum, Ripple and Bitcoin using machine learning models. After comparing neural networks, support vector machines and random forest, they showed that Twitter data could indeed be used for cryptocurrency prediction by means of neural network modelling.

In terms of crypto asset allocation management, the Markowitz mean-variance framework for cryptocurrency portfolio was utilized recently by (Kurosaki and Kim 2022) and Brauneis and Mestel (2019). The latter authors used data on cryptocurrency prices from 1/1/2015 to 12/31/2017 and compared the risk and return of different mean-variance portfolios for cryptocurrency investments against two benchmarks: a naively diversified portfolio and the CRIX. Their Empirical results showed that portfolios containing several cryptocurrencies have the lowest level of risk and perform better than single cryptocurrencies in terms of Sharpe ratio and certainty equivalent returns. Similarly, (Jiang and Liang 2017) applied deep learning and convolutional neural network (CNN), to cryptocurrency exchange data for a portfolio management problem. The proposed model was reinforcement learning in which the maximization of the accumulated return is a reward function of the network. Their proposed approach was compared with 3 benchmarks and 3 other portfolio management algorithms, and led to the conclusion that deep reinforcement was the best performing algorithm.

Despite the high accuracy of machine learning models in portfolio optimisation, they are not explainable, as the reasons of their decisions are not clear (see e.g. Giudici and Raffinetti (2021)). They are like black boxes, that can predict accurately but cannot explain why a decision is made (Bussmann et al. 2021). For example, why the proposed optimal portfolio has a large weight on a certain asset? and why another one is highly diversified among a certain set of assets?

The previous lack of explainability has prevented a wider acceptance and diffusion of machine learning models for portfolio optimisations and, more generally, in finance, particularly because financial regulators are not akin to validate this type of models.

Trying to overcome this problem, explainable artificial intelligence (XAI) have been recently proposed.

While some preliminary study apply XAI to cryptocurrencies, none of them have focused on the application of explainable AI for portfolio management (for more information see Valencia et al. (2019) and Ahelegbey et al. (2022)). Therefore, in this study, we fill this gap and propose an explainable portfolio management approach for cryptocurrencies that aims to strike a balance between predictive accuracy and explainability.

The rest of the paper is organized as follows. In Section II, the methodological background is explained. Section III presents the data employed in the paper. In Section IV, we discuss and present our empirical results. Section V concludes by summarizing the findings and giving some future research directions.

3.3 Methodology

In this section, we explain the theoretical concepts and methods utilized in this paper.

3.3.1 Portfolio Allocation

Portfolio asset allocation is a popular topic in finance that indicates the investment amount which should be shared among different available investment opportunities to diversify the risk of a portfolio and simultaneously receive an acceptable return.

The mean-variance model was proposed by the pioneer of modern portfolio theory, (Markowitz 1994), that quantitatively express portfolio selection problem by treating it as a quadratic optimization problem. This model assumes that future asset prices can be accurately reflected by market historical prices, which are defined by a vector of expected returns and a covariance matrix based on mean of return, standard deviation of return, and correlation with other assets. An investor can achieve good investment performance by selecting the right combination of assets to invest in. Tracing the efficient frontier, a continuous curve of mean return and risk intersections that indicates the best investment strategy, allows for higher returns with the same risk rate. In essence, this model seeks a trade-off between risk and return (Kalayci et al. 2017).

Financial technology (FinTech) innovations are driving the banking sector, with more changes expected in the next few years than in the previous two centuries. One of the most significant changes is the use of advanced technologies, such as machine learning algorithms, to facilitate security trading and advisory services for investors, which are referred to collectively as robo-advisors (Leow et al. 2021).

Most robo-advisors perform and daily update asset allocation on the basis of Markowitz's model, based on a mean–variance analysis, or a variant of it. However, the same robo-advisors rarely disclose information on how they choose their asset class investment universe or how they estimate variances and correlations between asset classes. They even more rarely disclose their expected return and risk parameters explicitly (Phoon and Koh 2017).

In this paper, we aim to fill this gap. Before doing so we briefly revise Markowitz model, using a sample portfolio made of crypto assets to optimize daily crypto asset allocation for minimum volatility.

For a given portfolio p, the return is calculated as follows:

$$E[R_p] = \sum_{i=1}^{n} w_i E[R_i], \tag{3.1}$$

Where w_i is a weight in $[0, 1]$ and $E[R_i]$ represents the daily return of a set of cryptoassets, for $i = 1, \ldots, n$.

To calculate the risk of a portfolio, the following equation can be used:

$$\sigma_p^2 = \sum_{i=1}^{n} w_i^2 \sigma^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_i w_j cov(R_i, R_j), \tag{3.2}$$

Where cov(R_i,R_j) indicates the covariance between the return of the ith and the jth cryptoasset.

We assume that a robo-advisory updates daily the proposed portfolio, calculating the portfolio weights on the basis of previous information. We also assume, without loss of generality, that a relevant past time window corresponds to a month.

Let X_i ($i = 1, 2, \ldots, n$) indicate the price of an asset, in our case a crypto, on the i^{th} day of the available time horizon. We will use the information of the 30 days before the p^{th} day to calculate the portfolio weights for the p^{th} day. Consequently, if $p = 1$ indicates the first day of available data, the first portfolio is generated at day p=31.

For a given day p, with $p > 30$, let $E[R_i]$ indicate the mean daily returns of each crypto asset during the last 30 days. Equation 3.1 for the first portfolio would then become:

$$E[R_{p=1}] = w_1 E[R_1] + w_2 E[R_2] + \ldots + w_{30} E[R_{30}], \tag{3.3}$$

Whereas, in terms of portfolio risk, Equation 3.2 becomes:

$$\sigma_{p=1}^{\,2} = w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + \ldots + w_n^2\sigma_n^2 + w_1 w_2 cov(R_1, R_2) + \ldots + w_{n-1} w_n cov(R_{n-1}, R_n), \tag{3.4}$$

Where σ_1^2 is the variance of the first crypto asset, calculated using the previous 30^{th} day. $cov(R_i, R_j)$ refers to the covariance between the ith and jth cryptocurrency within the same thirty days. Hence, $cov(R_1, R_2)$ shows the covariance between the first and second cryptocurrencies during the mentioned time period.

Once operational, on day 31st, the robot advisor updates its portfolio weights every day. For example, it calculates a second optimal portfolio on the 32^{th} using Equations 3.3 and 3.4 with data from day 2 to day 31.

Therefore, we assume that for any given day, after the 31^{st} day in the available data, the robot advisor recalculates the optimal portfolio weights, using the previous 30 days.

A question that naturally arises for the user of a robot advisory platform is which crypto assets are most influential in the determination of the portfolio weights. From this viewpoint, Markowitz' model can be seen as a "black box" which produces an optimal set of weights, without explaining which are the most important determinants of such result. Explainable AI methods can be used for this purpose, to explain the most important factors that affect portfolio weights.

To achieve this goal, we first provide Z scores of cryptocurrencies in each portfolio, Z_i, to provide a measure that considers both volatility and return of these assets as follows:

$$Z_i = \frac{R_i - E(R_i)}{\sigma_i}, \tag{3.5}$$

where R_i represents the return of cryptos in the $p^t h$ day. $E(R_i)$ and σ_i are the average return of the corresponding crypto and the standard deviation of the asset return during the previous 30 days of the portfolio generation date.

Furthermore, to have a general measure that presents the overall performance of each portfolio, for each daily portfolio, the Z score, that combines the return and the risk, is calculated as:

$$Z_p = \frac{R_p - E(R_p)}{\sigma_p}, \tag{3.6}$$

Where $E(R_p)$ indicates the mean of the portfolio returns, and σ_p the standard deviation of the portfolio returns, in the previous 30 days.

3.3.2 Explainable Artificial Intelligence

Several efforts have been made in recent years to create intelligent agents capable of explaining their decisions (Townsend et al. 2019). According to (Adadi and Berrada 2018), interpretability is concerned with the understanding of a given model at two different granular levels: global (analyzing the whole model behavior) and local (focusing on a specific prediction) (Molnar et al. 2023). In addition, interpretability techniques can be divided into two types based on their application: model agnostic, which are independent of the model, and specific, which are designed for specific problems. In many applications, machine learning models that are complex black box methods are utilized and provide accurate predictions but these outputs need explanations (for more details see (Guidotti et al. 2018).

In this study we utilize a model agnostic XAI approach, the SHapley Additive exPlanations (SHAP) model, introduced by Lundberg and Lee (2017) that is based on the game theory. In SHAP approach, to understand the contribution of each featured explanatory variable, different feature coalitions are considered and the difference in each prediction made by adding a selected feature is its ”local” shapley value (Shapely 1953). The local explanations can be combined in an overall global Shapley value leading to the general contribution of each variable.

Specifically, the effect of each predictor X_k, for $k = 1, \ldots, K$, is calculated as follows:

$$\phi(f(\hat{X_i})) = \sum_{X' \subseteq \mathcal{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [\hat{f}(X' \cup X_k)_i - \hat{f}(X')_i] \tag{3.7}$$

Here K represents the number of predictors, X' is a subset of $|X'|$ predictors, $\hat{f}(X' \cup X_k)_i$ and $\hat{f}(X')_i$ are the predictions of the i-th observation obtained with all possible subset configurations, respectively including variable X_k and not including variable X_k.

Once Shapley values are calculated, the ”global” contribution of each predictor can be calculated averaging the ”local” contributions over all predicted values.

There are no reasons to assume that Markowitz asset allocation is a linear function of the returns and, therefore, we assume that the relationship between Z_p and the Z_i of the single crypto assets is non linear. To estimate such a non linear relationship we can employ a machine learning models, for example a Random Forest model. In such a model, we consider the daily returns of each crypto asset as the explanatory variables. On the other hand, the Z score of each portfolio is utilized as the response to be predicted.

We will apply Shapley values to the daily series of returns and risks generated by Equations 3.3 and 3.4.

3.4 Data and Descriptive Analysis

To illustrate our approach, , we consider the daily price of 8 cryptocurrencies: Bitcoin (BTC), Ethereum (ETH), Ripple (XRP), Bitcoin Cash (BCH), Litecoin (LTC), Binance Coin (BNB), Eos (EOS) and Stellar (XLM) over the period, September 2017 to October 2019 (771 daily observations). Figure 3.1 shows how the price of these cryptocurrencies is distributed over the time horizon.

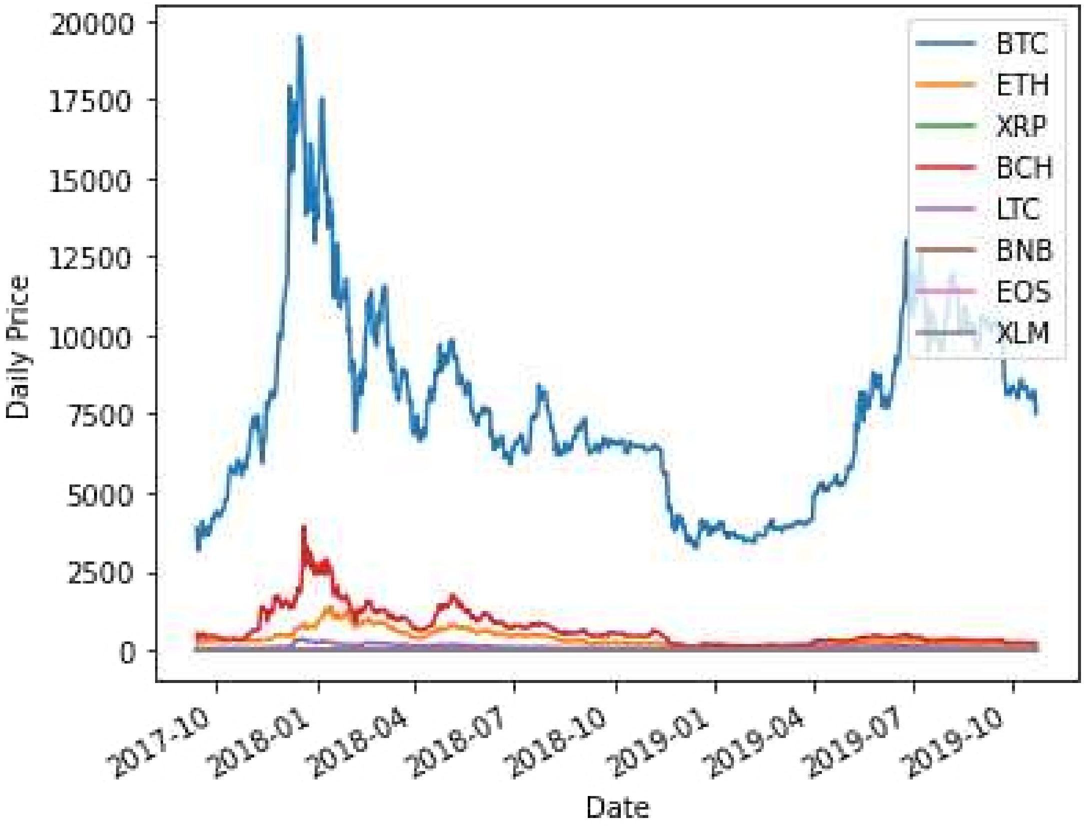


Fig. 3.1 Price distribution of cryptocurrencies

Figure 3.1 confirms that BTC has the highest price over the whole period and fluctuates between about 2500 and 20000. In addition, descriptive statistics of the observations are presented in Table 3.1.

Table 3.1 Descriptive statistics of the selected 8 cryptocurrencies

Cryptocurrency	Mean	STD	Min	Max
BTC	7597.92	3048.73	3154.95	19497.4
ETH	359.19	258.32	84.31	1396.42
XRP	0.49	0.40	0.16	3.38
BCH	667.82	597.61	77.37	3923.07
LTC	93.92	59.28	23.46	358.34
BNB	13.63	8.81	0.67	38.82
EOS	5.78	3.70	0.49	21.54
XLM	0.18	0.13	0.01	0.89

From Table 3.1, it can be claimed that BTC,ETH,BCH and LTC are the most volatile cryptocurrencies during the considered time period. Also, the lowest price is attributed to XLM whereas BTC shows the highest price.

3.5 Empirical Finding

In this section, we explain our empirical study and present the results we obtained from applying the proposed model to the selected data.

As explained in Section 5.4, the main goal of this paper is to propose an explainable portfolio management approach that has the ability to explain the weights assigned by a robot advisor that daily applied Markowitz' asset allocation model to available cryptocurrencies. For this purpose, we generate portfolios on a daily basis using the historical data of the last 30 days. Therefore, the first day for which we create a portfolio in our data set is 14-10-2017, using the daily price of the eight cryprocurrencies between 14-09-2017 and 13-10-2017. We then repeat portfolio construction for the 739 consequent available days, thereby generating a total of 740 portfolios. The portfolio weights can be stored in a matrix which includes 740 rows and 8 columns, representing the sets of weights for each crypto in each built daily portfolio. We recall that the covariance matrix is updated daily, for each portfolio, using the information of the 30 days prior to the date of the portfolio creation.

Once the portfolio weights are calculated, applying equations 3.1 and 3.2 to the observed data, we can obtain the actual return and risk of the 740 portfolios constructed by the robot advisor. To present the characteristics of the generated portfolios by our model, Figure 3.2 provides the distribution of portfolio risk for the 740 robot advised portfolios, in comparison with the portfolios generated using constant and equal weights.

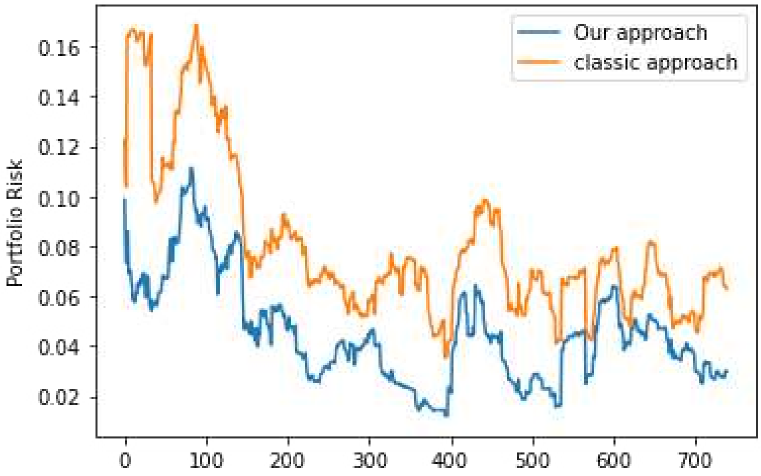


Fig. 3.2 Comparison of distribution of portfolio risk in our approach and for an equally weighted portfolio.

Figure 3.2, shows that the robot advisor allocation strategy, with weights that change over time, leads to portfolios with lower risk levels with respect to a static portfolio, with constant equal weights.

As we explained in Section 3.3.1, to consider both aspects of a portfolio, we focus on Z score of the portfolios and included cryptos. Therefore, after calculation of risk and return for each portfolio, we find the Z scores using Equations 3.5 and 3.6.

In the next step, to develop a machine learning model, Z scores of the cryptocurrencies are used as explanatory variables to predict the Z score of the portfolio. For this purpose, to build a random forest model we split the data into train and test using the train ratio of 70 percent. It should be noted that as the focus of this paper is not the prediction accuracy or selecting the best ML model to predict Z score of portfolios, we do not discuss about the model goodness of fit.

Finally, Shapley values corresponding to cryptocurrencies are computed to explain the prediction of our complex ML model. Specifically, using shapley values, we can understand which cryptocurrencies in a portfolio most contribute to explain the Z-score of each portfolio,which is a function of both its return and risk. Figure 3.3 presents the overall contributions of each cryptocurrency Z-score to the portfolio Z-score shapley values, averaged across different samples (portfolios).

From Figure 3.3 is obvious that Stellar (XLM), followed by, Bitcoin Cash (BCH) and Ripple (XRP) in comparison with other cryprocurrencies are the most explainable variables, that is, those that explain the largest part of the variations of the portfolio Z-scores (functions of both the return and the risk). From an asset allocation viewpoint, this means that the variations in the weights of the optimal portfolio over time are mainly explained by variation in the returns and/or in the risk of these three assets.

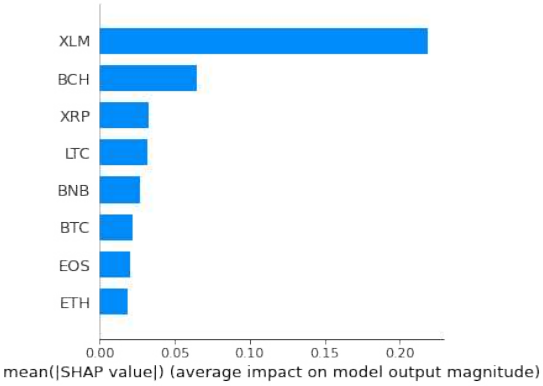


Fig. 3.3 Overall contribution of cryptocurrencies to the Z scores of portfolios

Figure 3.4 provides a better insight to understand the contribution of cryptocurrencies to portfolio Z score. It provides, for each crypto, and each time point, a color that indicates the magnitude of the corresponding z-score (blue for low values and red for high values) and a location on a horizontal line that indicates the Shapley value of each crypto with respect to the portfolio Z-score.

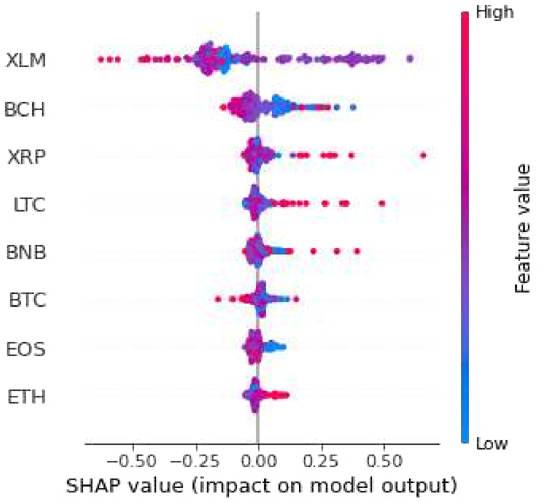


Fig. 3.4 Contribution of cryptocurrencies to portfolio risk

The plot in figure 3.4 allows a detailed analysis of the local (pointwise) explanation of each crypto, differently from Figure 3.3 which presents a global (averaged) explanation. The x-axis represents the shapley values of the 8 cryptocurrencies available on the y-axis. Colorful points on the left show those which shift the response (portfolio risk) in a negative direction while points on the right are the observations that contribute to the predicted values in a positive way. Figure 3.4 indicates that the contribution of XLM is highly variable, leading to high weights for some days and low weights for others. In contrast, less volatile assets such as BTC and ETH provide rather stable weights.

To better explain the ”local” explainability of portfolios in Figure 3.5 we present the Shapley values of the eight crypto in four portfolios, randomly selected among the 740 proposed allocations.

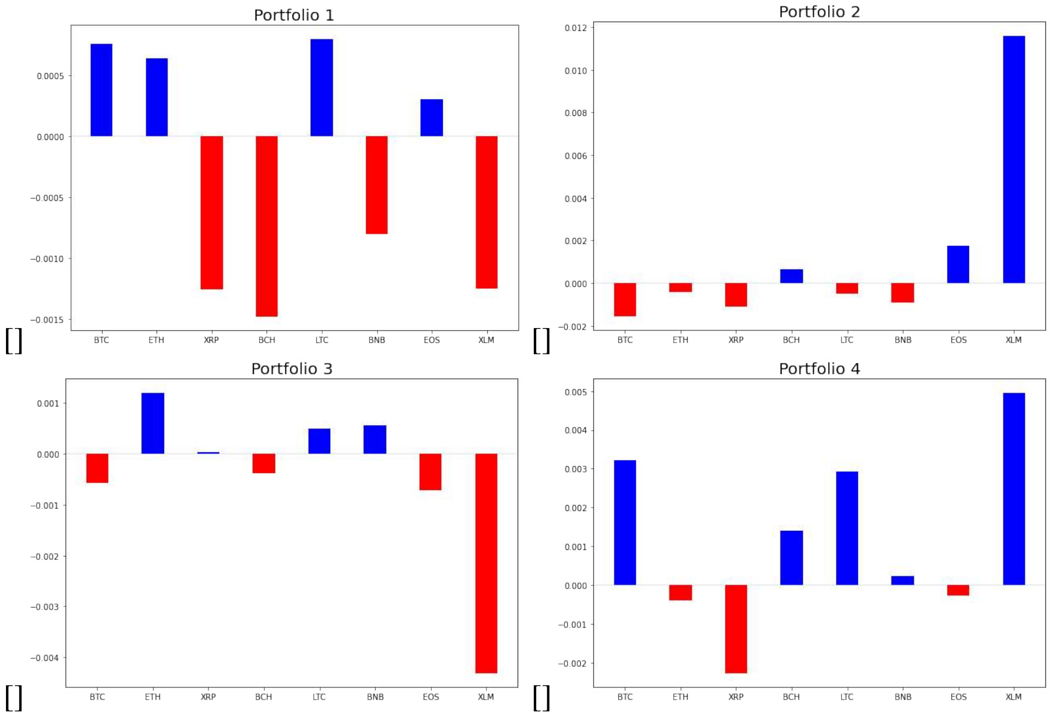


Fig. 3.5 Contribution of each explanatory variable to the Shapley’s decomposition of four generated portfolios. Red color and blue color show negative and positive distance from the mean, respectively.

Figure 3.5 clearly shows the advantage of our explainable model. It can indicate which variables contribute more to the prediction of each portfolio’s Z-score, based on both returns and risks. Figure 3.5 clearly shows how the explanations are different (“personalised”) for each of the four considered portfolios. For example, Portfolio (b) assigns a high value to XLM whereas Porfolio (c) a low value. Generally, it can be said that XLM is a crypto which has the highest importance score in almost all four portfolios, as already discussed.

3.6 Conclusions

We have presented a methodology that can explain the ”automatic” choices of a robot advisor which operates according to Markowitz’ optimal asset allocation. The methodology has been applied to a time series of portfolios calculated on a set of crypto assets. The obtained results suggest to implement the method as a useful tool for supervisors, which assess the compliance of robot advisories to financial regulations. But also for robot advisory FinTech companies, which can self-assess the same compliance. It may also be made disposable to investors, through appropriate application service interfaces.

Further research involves the application of the proposed methodology to different portfolio optimization methods, more complex than Markowitz's, such as those that consider systemic risk and/or tail risk. Also, other explainable AI methods, such as Shapley-Lorenz decompositions (Giudici and Raffinetti (2020)), can be applied to find the most explainable determinant of portfolio returns in terms of their predictive accuracy, rather than in terms of their pointwise predictions.

Finally, the proposed methods should also be applied to measure the operational risks that can heavily impact crypto markets and, in particular, cyber risks (see e.g. Giudici and Raffinetti (2021)

CHAPTER 4

Explainable FinTech Lending

Babaei, G., Giudici, P., Raffinetti, E. (2023). Explainable fintech lending. Journal of Economics and Business, 106126.

4.1 Abstract

Lending activities, especially for small and medium enterprises (SMEs), are increasingly based on financial technologies, facilitated by the availability of advanced machine learning (ML) methods that can accurately predict the financial performance of a company from the available data sources. However, despite their high predictive accuracy, ML models may not give users sufficient interpretation of the results. Therefore, it may not be adequate for informed decision-making, as stated, for example, in the recently proposed artificial intelligence (AI) regulations.

To fill the gap, we employed Shapley values in the context of model selection. Thus, we propose a model selection method based on predictive accuracy that can be employed for all types of ML models, those with a probabilistic background, as in the current state-of-the-art. We applied our proposal to a credit-scoring database with more than 100,000 SMEs. The empirical findings indicate that the risk of investing in a specific SME can be predicted and interpreted well using a machine-learning model which is both predictively accurate and explainable.

4.2 Introduction

The advent of financial technologies (fintech) has led to the emergence of several new companies in financial markets (Fasano and Cappa 2022). Thanks to technologies such as Artificial Intelligence (AI), Blockchain, and Cloud computing, these companies offer additional services to compete with traditional banks (Kendall 2017, Temelkov 2018). In fact, in fintech platforms, such as crowdfunding, peer-to-peer lending, and robot advisory, based on digitalised business models, finance and technology meet (Ayadi et al. 2021), improving customer experience, lowering costs and increasing transparency (Romānova and Kudinska 2016).

Fintech platforms have substantially changed several financial services. Among them, online lending platforms such as the Lending Club, one of the largest peer-to-peer lending organisations, have increased financial inclusion, allowing credit allocation to borrowers typically not funded by traditional banks while providing highly attractive returns for investors.

However, lending platforms bear high risks for investors, as borrowers are typically small and medium enterprises (SMEs) or low-income individuals (Correia et al. 2022), which, in addition, are largely interconnected. It follows that future perspectives of lending platforms largely depend on assessing credit risk and determining the causal drivers of such risks (Milne and Parboteeah 2016).

This study contributes to fintech credit risk assessment for lending to SMEs. Providing credit to SMEs is a key research topic for policymakers (Berger and Udell 2006, Ferri and Murro 2015). Fintech lending facilitates credit services and financial inclusion, directly connecting individual lenders with company borrowers through a credit assessment platform that analyses all available data on borrowers to learn and continuously update their scores and classes of scores (ratings). Compared to traditional bank lending, fintech lending improves customer experience and provides more credit to companies. However, they can suffer from information asymmetry between borrowers and lenders (Giudici et al. 2020, Bracke et al. 2019, Kumari et al. 2021), possibly leading to inaccurate creditworthiness estimates.

To solve this problem, ML models, a combination of statistical models and computational algorithms able to learn from large databases regularities and relationships between a large number of variables, can be applied to lending data to obtain estimates of creditworthiness more accurate than those obtained with classical credit scoring models.

ML models have been widely used in many financial studies, such as credit scoring (Bastani et al. 2019, Lee and Chen 2005, Shen, Zhao and Kou 2020), portfolio optimization (Guo et al. 2016), and profit scoring (Serrano-Cinca and Gutiérrez-Nieto 2016, Babaei and Bamdad 2020*a*). These studies demonstrate that ML models perform well in terms of predictive accuracy. However, their predictions are not easily interpretable because the underlying model is a nontransparent black box.

To solve this problem, explainable Artificial Intelligence (XAI) methods, in which humans can understand the results of the solution, have recently been proposed (Lundberg and Lee 2017, Ribeiro et al. 2016*b*, Bussmann et al. 2021, Sachan et al. 2020, Giudici and Raffinetti 2021).

The XAI models achieve a good trade-off between explainability and predictive accuracy. However, massive computation may be involved when the number of explanatory variables is large. To reduce the computational burden, we propose to apply XAI models and, specifically, Shapley values to interpret ex-post the predictive power of each variable and as an ex-ante variable selection criterion. This leads to more parsimonious models, which, while maintaining a good predictive accuracy, can be better interpreted by users while maintaining good predictive accuracy.

Our proposed model can thus support banks and fintechs in developing an AI-based credit scoring model which is 'trustworthy', accurate, and explainable, with the potential of being validated by supervisory authorities and regulators.

From a statistical viewpoint, the proposed variable selection method combines predictive accuracy with explainability. Variable selection methods are well known in the literature and relate to Occam's Razor principle: 'Among competing hypotheses, the ones with the fewest assumptions should be selected.' When many explanatory variables are available, applying this principle leads to the diffusion of stepwise variable selection algorithms that compare models comprising different sets of variables in terms of their statistical likelihood.

However, most ML models are non-probabilistic, and likelihoods are unavailable, preventing the use of stepwise selection algorithms. Alternative ML models, such as neural networks with different layers and hidden nodes or random forests with different input variables, can be compared in terms of predictive accuracy. This suggests that a different stepwise procedure is currently unavailable.

We propose to fill the gap by employing the Shapley value associated with each explanatory variable as the underlying metric to perform stepwise variable selection. Specifically, the variables that contributed the least to the predictions regarding their Shapley values were removed from the model. Variables were removed if the predictive accuracy of the model was not significantly reduced. Thus, both explainability and predictive accuracy were achieved.

We applied our methodological proposal to compare alternatives; we used random forest models that aimed to build credit scores that accurately predicted the probability of default (PD) for a set of companies. This helped to leverage the value of the data and the nonlinear relationships present in the data, leading to a more accurate and transparent credit scoring model that can be employed in fintech lending platforms.

From a managerial viewpoint, our proposal allows us to identify the variables that explain the credit risk of lending investments in SMEs.

To the best of our knowledge, this is the first methodological study to employ the XAI as a variable selection tool within the credit scoring context.

The rest of the paper is organised as follows. Section 2 contains a literature review on applying ML and AI to credit scoring. Section 3 introduces the proposed method. Section 4 presents the data and the main empirical findings. Finally, Section 5 concludes the study.

4.3 Literature Review

Credit scoring is a research topic which has attracted many researchers, who have employed different statistical learning models to measure it (Bücker et al. 2022, Liu et al. 2022, Dushimimana et al. 2020). The modern ML methods have found one of their first fields of application to economics in credit scoring: among the first studies, we can mention (Srinivasan and Kim 1987) in which decision trees are used; (Henley and Hand 1996) in which k-nearest neighbours are employed; (West 2000, Yobas et al. 2000), in which neural networks and support vector machines are applied; (Djeundje and Hamid 2021), in which a range of ML models are applied to both traditional and alternative credit scoring data. See (Hand et al. 2001) for a review of the data mining methods for credit scoring.

In the last few years, the emergence of ensemble methods, which aggregate results from different models, has substantially improved the performance of scoring models based on ML (Finlay 2011, Lessmann et al. 2015). In this respect, Li and Chen (2020) provides a comparison of different ensemble methods: random forests, adaptive boosting, gradient boosting, and light gradient boosting, applied to five alternative credit scoring models: neural networks, classification trees, logistic regression, naïve Bayes, and support vector machines. Our study shows that the performance of ensemble credit scores was better than that of individual scores. We also show that the ensemble random forest model achieved the best accuracy metrics, such as the Area Under the ROC Curve, the Kolmogorov–Smirnov statistic, and the Brier score. In another study, Chopra and Bhilare (2018) compared random forests and gradient boosting using a credit-scoring model based on a classification tree. They showed that for their credit scoring application, ensemble methods (gradient boosting and random forest) outperformed individual classification tree models, thereby adding further evidence to the higher accuracy of ensemble methods. A third study, Tripathi et al. (2022), undertook a comparative analysis of nine ensemble methods applied to different scoring models, such as logistic regression, naïve Bayes, and classification trees. As in previous studies, they found that ensemble scoring methods improve the performance of single-credit scoring methods.

The previous discussion indicates a consensus on the superior predictive accuracy of ensemble credit scoring models concerning single models. However, the increased accuracy comes with a cost: while most single scoring models, such as logistic regression, tree models and naïve Bayes are 'explainable', as they can identify the contribution of each explanatory variable to the credit scores, ensemble methods are 'black boxes', and cannot explain the determinants of credit scores to their users (Bracke et al. 2019, Giudici et al. 2020). This is a problem from a regulatory viewpoint because the application of ML AI to credit scoring, a high-risk application, must be accurate and explainable, as stated in the recently proposed European Artificial Intelligence Act (`https://artificialintelligenceact.eu`).

To overcome this problem, ensemble methods should be complemented with explainable AI methods that are to be applied a posteriori on the obtained credit scores. Explainable AI methods can be classified into model-specific and model-agnostic (Adadi and Berrada 2018). In contrast to model-specific methods, model-agnostic methods can be applied to any ML model. Local methods such as Local Interpretable Model agnostic explanations (LIME) (Ribeiro et al. 2016*a*) and Shapley values (Lundberg and Lee 2017) are of particular interest, both explaining each specific credit score based on the additional contribution of each explanatory variable to their values.

Local methods have been recently applied to explain credit scores based on ML. For instance, Bussmann et al. (2021) propose a methodology based on Shapley values as a post-processing analysis to explain the credit scores obtained from ensemble models applied to data that concern a sample of Italian SMEs, which apply for peer-to-peer lending. Their empirical results demonstrated the capability of explainable AI methods to achieve predictive accuracy and explainability. A related study, Moscato et al. (2021), proposed a credit scoring model to predict whether a loan will be repaid on a P2P platform. It compared different ML models and explainability methods, including LIME and SHAP, showing their advantages. Similarly, (Xia et al. 2021) showed how credit scores obtained with gradient boosting can be interpreted using Shapley values, and (Tyagi 2022) compared various ML models for credit scoring in terms of Shapley values to develop new investment models and portfolio strategies. All these studies provide evidence of the advantage of using explainable AI methods in combination with ML models in credit scoring. Our study falls into this research stream. It proposes a credit-scoring model based on an ensemble machine-learning method that can be explained using the Shapley value approach. Our original contribution is that we propose achieving explainability, not a posteriori, by applying the Shapley value to the obtained credit scores but ex-ante as a variable selection criterion.

Our proposal is inspired by the acknowledged advantage of variable selection in improving the predictive accuracy of ML models. For example, Laborda and Ryoo (2021) discussed the performance of three variable selection models: a filter method (based on statistical tests) and two wrapper methods (based on stepwise model selection) to obtain a more parsimonious credit scoring model based on logistic regression, support vector machine, K-nearest neighbours, or random forest. They concluded that stepwise selection yielded a superior predictive performance for all models. A related study, Trivedi (2020) employed chi-square testing as a filter method for ML classifiers, such as naïve Bayes, random forest, classification trees, and support vector machines, to improve credit scoring predictions. They found that chi-square testing with credit scoring improved the predictive accuracy of all classifiers.

In this study, we combined variable selection with explainability, proposing a variable selection model that chooses the most explainable variables as model predictors. Thus, variable selection improves the predictive accuracy and interpretability of the credit scores obtained with an ML model. It does so before reaching a final model (ex-ante perspective) rather than after a model has been selected (as in the available applications of XAI models), thereby reducing the computational burden.

To achieve explainability, we considered traditional Shapley values (Shapely 1953), implemented by following the Conditional Expectations approach (Lundberg and Lee 2017). However, what is presented can be extended, without loss of generality, to more advanced approaches, such as the Integrated Gradients Shapley (Sundararajan et al. 2017) and the Baseline Shapley value (Sundararajan and Najmi 2020). The integrated Gradients Shapley method extends Shapley values to the continuous setting. It can be applied to credit lending problems in which the response variable is continuous (such as when the loss-given default is considered the target variable). The Baseline Shapley value overcomes some counterintuitive results of the traditional Shapley approach, such as the assignment of nonzero values to features not used by the model, with a more general approach in which a missing feature for an observation is modelled randomly by drawing it from the sample feature distribution.

4.4 Methodology

4.4.1 Credit Risk Assessment

The evaluation of a company's credit risk depends mainly on its estimated probability of default (PD); that is, the probability that a company will fail to repay its financial obligations. This problem is usually addressed by estimating each company's credit score and setting a threshold to classify it predictively into two main classes: non-default and default. Imagine that information from T explanatory variables of N firms (usually balance sheet indicators) is available. For each firm, we also have a response variable Y, which indicates whether the company has defaulted or is still active (usually in the following period); that is, $Y = 1$ in the case of default and $Y = 0$ otherwise. In the credit scoring model, we aim to find a model that can describe the relationship between T explanatory variables and the response variable Y.

Credit scoring models can be classified into two main categories: black and white boxes. In the former, the relationship between the explanatory variables and the response is not transparent, and only the final classification is observed. Complex ML models such as neural networks, random forests or gradient boosting belong to this category, providing high predictive accuracy at the expense of explainability. In contrast, statistical learning models such as linear and logistic regression are transparent and considered white-box models. These simple models explain how they behave and how predictions are obtained.

4.4.1.1 Logistic Regression

The most commonly used method for credit scoring is logistic regression, a 'white-box' statistical learning method that finds application in many studies (Murdoch et al. 2019). Logistic regression models classify the response variable into two groups characterised by different statuses (default vs active). More formally, the logistic regression is specified as follows:

$$ln((p_n)/(1-p_n)) = \alpha + \sum_{t=1}^{T} \beta_t x_{n_t}, \tag{4.1}$$

where p_n is the probability of default for the nth firm, x_n = $(x_{n1}, \ldots, x_{nT})$ is the T-dimensional vector of the borrower-specific explanatory variables, parameter α is the model intercept, and β_t is the tth regression coefficient. It follows that the probability of default can be found as

$$p_n = exp\big(\alpha + \sum_{t=1}^{T} \beta_t x_{nt}\big)\big(1 + exp(\alpha + \sum_{t=1}^{T} \beta_t x_{nt})\big)^{-1} \tag{4.2}$$

Although the high interpretability of a logistic regression model follows from its explicit functional form, which is linear in the logarithm of odds, its predictive accuracy may be low because of its linear nature. When the available data are large and complex, the predictive accuracy of logistic regression may be inferior to that of a more complex ML model.

4.4.1.2 Random Forests

ML models are increasingly used in complex credit risk assessments (Bussmann et al. 2021). Among them, the random forest classifier, an ensemble of classification trees (Breiman 2001), performs well in many credit risk classification problems. Like logistic regression, in a classification random forest model, each observation (for example, a company with its corresponding vector of explanatory variables x_n) is mapped to a default response variable. A random forest classifier merges the rules obtained from a set of classification trees, each based on a training data sample and explanatory variables.

Although each classification tree has explicit rules of construction and allows us to understand how different credit scores are generated, a random-forest model aggregates the scores from each tree on a single average, thereby losing interpretability. A random forest is a black box model which cannot meet the need for explainability in the finance sector (Murdoch et al. 2019). To overcome this limitation, explainable AI models that provide details or reasons to make the functioning of AI clear or easy to understand can be employed.

4.4.2 Explainable Artificial Intelligence

Financial institutions and markets are subject to many regulations to maintain the stability of the financial system and protect consumers and investors. An important aspect of financial regulation concerns the supervision of risk management models, particularly credit risk models, for which regulators may seek assurance on the key drivers (Giudici et al. 2020). This suggests that black-box AI is unsuitable for credit risk measurement, which motivated the development of XAI models.

The most commonly employed explainable AI model is the Shapley values approach, a model-agnostic post-processing tool used to explain and interpret ML predictions. The Shapley value approach was introduced by Shapely (1953), who leveraged concepts from game theory to map predictive inferences to a linear space.

Specifically, we assumed a game exists for predicting each observation (row). For each game, the players were model predictors (explanatory variables), and the total gain is equal to the predicted value, obtained as the sum of the contributions of each predictor.

Following these assumptions, the Shapley value algorithm calculates the contribution of each variable to each prediction by considering its additional effect on all possible coalitions (groups) of other variables. Specifically, the effect of each variable X_k, for each credit score $i = 1, \ldots, n$, is calculated as follows:

$$\phi(\hat{f}(X_i)) = \sum_{X' \subseteq \mathcal{C}(X) \setminus X_k} \frac{|X'|!(K - |X'| - 1)!}{K!} [\hat{f}(X' \cup X_k)_i - \hat{f}(X')_i], \tag{4.3}$$

where K represents the number of predictors, X' is a subset that contains $|X'|$ predictors, $\hat{f}(X' \cup X_k)_i$ and $\hat{f}(X')_i$ are the predictions of the i-th observation obtained with all possible subset configurations, respectively including variable X_k and not including variable X_k.

Once Shapley values are calculated for each observation to be predicted, the overall contribution of each predictor, the 'global' Shapley value, is obtained as their sum.

4.4.3 Proposal

We propose employing the global Shapley values of each explanatory variable as the basis of a stepwise variable selection algorithm valid for all models, whether white box or black box.

The algorithm begins with a complete model containing all the available variables. It then removed the variable with the lowest global explainability from the model and evaluated whether the removal significantly decreased predictive accuracy. If so, it stops; otherwise, it proceeds with the deletion of variables until it stops.

A key element of our proposal is a significance test that compares the Area Under the Curve (AUC) of two alternative models differing in the presence of one variable. We recall that the AUC of a model is the most employed predictive accuracy measure for binary variables and is obtained as the area underlying the Receiver Operating Curve (ROC) of a model. The ROC curve was obtained by joining a set of coordinates which represented, for a given set of cutoff points (percentiles), the True Positive Rate against the False Positive Rate. While an ideal model should always have TPR=1, FPR=0, and an AUC equal to 1, the higher the AUC, the better the model.

The significance test for the AUC was based on DeLong's test (DeLong et al. 1988). It calculated the Area Under the ROC Curve for each pair of models compared: to model M_k $(k = 1, \dots, K)$ against model M_{k-1}. The test statistic was based on the difference between two AUC values.

More formally, the null hypothesis of the statistical tests is the equivalence of models M_k and M_{k-1}. If the $p$value is more significant than a threshold significance level, such as 5%, we fail to reject the null hypothesis; therefore, model simplification is accepted: variable k is not statistically significant in predicting the response variable.

The stepwise variable selection process continues until the p-value exceeds the set threshold significance level (e.g. 5%). When this occurs, H_0 is rejected such that M_k cannot be simplified to M_{k-1}. Consequently, the procedure stops, and M_k is selected as the final model. Note that the outlined procedure fully aligns with Occam's razor parsimony principle. If two models have similar predictive accuracy, we choose the simplest of the two (i.e. the one with the lowest number of predictors).

Finally, the choice of the AUC or other test statistics depended on the response variable. For a binary response, AUC is the most commonly employed measure. We employed the Mean Squared Error (MSE) and the corresponding Diebold–Mariano test (Diebold and Mariano 2002) for a continuous response.

4.5 Application

4.5.1 Data

We illustrate the application of our proposal to a large data sample which contains the balance sheet data for over 100,000 SMEs, referred to as the 2020 reporting year. The data were supplied by Modefinance (modefinance.com), a rating agency in a European

Credit Assessment Institution (ECAI) supervised by the European Securities and Markets Authority (ESMA), specialising in credit scores for P2P platforms focused on SME commercial lending. The presence of SMEs is a common trait in many countries; therefore, the data can be considered an instance of a more general situation. The companies in the available sample are headquartered in the largest European Union (EU) countries: Italy, France, Spain, and Germany. Their distribution across countries for 2020 is described in Table 4.1.

Table 4.1 Distribution of Small and Medium Enterprises in the sample by Countries.

Country	No.of.Companies	Percentage
Italy	59,864	49.37
Spain	25,949	21.40
France	33,865	27.93
Germany	1,575	1.30

From Table 4.1, note that most companies (49.37%) are located in Italy, with several SMEs. Italy is followed by France, where 27.93% of the companies are headquartered. Germany has the least number of companies in the table, containing only 1.30% of the companies. This is consistent with the fact that although Germany has a larger population than other countries, it does not require public deposits on company balance sheets. Although Germany has a limited number of companies in the sample, limiting its contribution, we prefer not to alter the supplied sample and keep all companies.

Examining the distribution of companies in the sample by ‘Industry Sector’, which shows to which industrial sector each SME belongs, is interesting. Table 4.2 lists the five most populated industries.

Table 4.2 Small and Medium Enterprises distribution in the sample by Industry sectors.

Industry Sector	No. Of Companies	Percentage
Retailing	30,201	24.91
Capital Goods	17,536	14.46
Materials	11,969	9.87
Commercial and professional services	10,861	8.96
Food and Staples Retailing	8,844	7.29

From Table 4.2 note that ‘Retailing’ is the most populated industry, followed by ‘Capital Goods’, ‘Materials’, and ‘Commercial and Professional Services’.

To estimate a credit scoring model from the data, we need a binary response variable that describes whether a company is in distress (indicating a likely default); and a set of explanatory variables, which may be considered likely causes (or not) of such distress. In the available data, such response variables can be obtained from the variable ‘MScore’, which is the rating assigned to each company by the rating agency modefinance. MScore can assume a set of ordered values that correspond to ratings of A, AA, AAA, B, BB, BBB, C, CC, CCC, and D, in which 'A' is assigned to companies with the lowest level of credit risk (lowest probability of default), whereas 'D' to those with the highest level (highest probability of default).

To convert the variable 'Mscore' into a binary default variable, as in the credit scoring context described in Section 4.4.1, we associated each company's rating to one of two alternative classes. On the one hand, we associated rating levels C, CC, CCC, and D with a perceived state of default (class 1); on the other, we associated rating levels A, AA, AAA, B, BB, and BBB with a perceived state of non-default (class 0). The resulting percentage of defaulting companies in the available SME sample was 14%.

The distributions of the sample default variable for any given country and industry sector are presented in Figures 4.1 and 4.2. In both figures, the total height of each bar is proportional to the number of companies in each group (country or industry sector), and at the top of each bar, we report the observed default percentages.

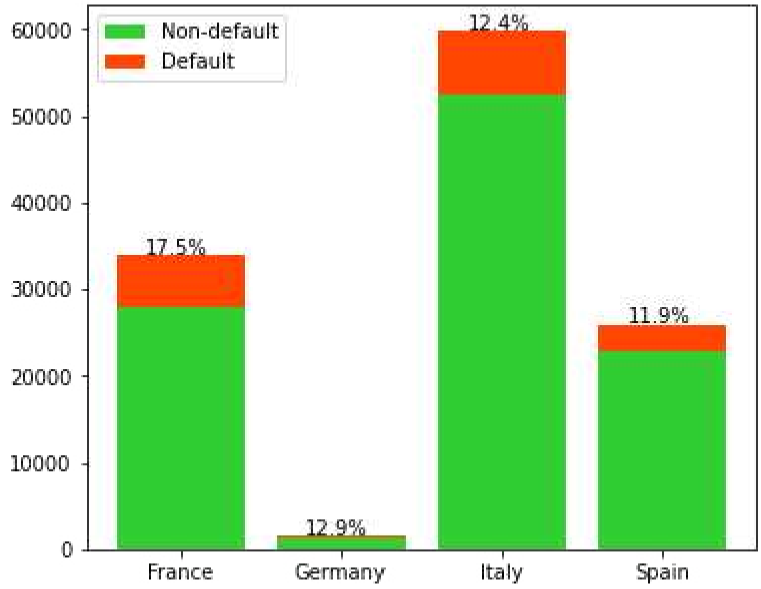


Fig. 4.1 Distribution of default and non-default SME by country.

From Figure 4.1, we can conclude that France is the riskiest country, with the highest default probability of approximately 17.5%. This is followed by Germany, with a 12.9% probability of default; however, its impact on the system is limited, as its frequency is low compared to those of more populated countries such as Italy and Spain. Similar conclusions can be obtained from Figure 4.2: 'Consumer services', 'Diversified financials', and 'Media and entertainment' are the riskiest industries, but the impact of the 'Commercial and professional services' sector is higher, being much more populated.

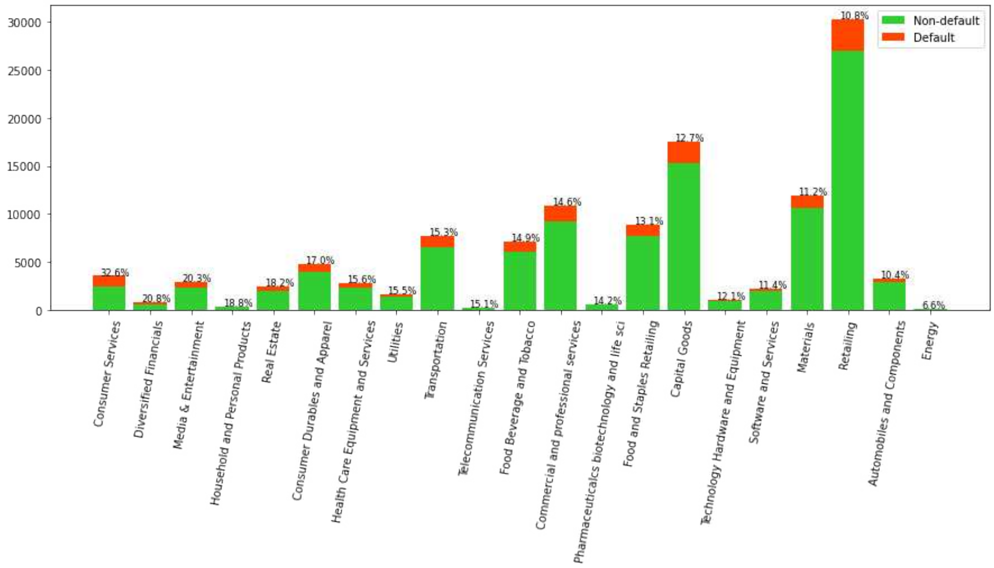


Fig. 4.2 Distribution of default and non-default SME by industry sector.

To complete the description of the variables in the sample data, Table 4.3 shows the considered explanatory variables, which are all financial ratios calculated by modefinance from the 2020 balance sheets of the available companies.

Table 4.3 Description of the Explanatory Variables.

Variable	Description
Turnover	Operating revenues in Thousands of Euro
Leverage	Leverage (ratio)
PLTax	Profit/Loss after tax in Thousands of Euro
TAsset	Total assets in Thousands of Euro
EBIT	Earnings Before Income Tax and Depreciation in Thousands of Euro
ROE	Return on Equity (percentage)

Table 4.3 indicates that the available explanatory variables are six financial variables which measure, respectively: the operating revenues (Turnover); the financial structure (Leverage); the size (Total Assets), and the profitability (EBIT, Profit and Losses after Tax, Return on Equity) of each considered company, based on the 2020 balance sheets.

Table 4.4 provides, as summary statistics, the mean of each explanatory variable, separately for the defaulted and the non-defaulted companies.

Table 4.4 Conditional means of the financial variables.

Class	Turnover.2020	EBIT.2020	PLTax.2020	Leverage.2020	ROE.2020	TAsset.2020
Non-Default (Class Zero)	10,950.948104	717.148144	521.539610	4.617414	13.895101	12,560.865164
Default (Class One)	10,261.416051	-828.175527	-1,003.877095	994.987954	-4.896900	15,836.216495

Comparing the conditional means of each variable in Table 4.4 EBIT, PLTax, and Leverage present the largest difference between defaulted and non-defaulted companies: they are likely to be the most impactful on the credit scores. Conversely, Turnover and Total Assets show a small difference between the conditional means.

4.5.2 Results

We first build a 'classic' credit scoring model based on the logistic regression model in Section 3.1.1. Therefore, we randomly split the data into training (70% of the data) and validation samples (the remaining 30% of the data). For comparison, the same data partitioning was used when applying the random forest model.

Initially, we considered, as explanatory variables, a full model, with all the six financial ratios described in Table 3, along with the Country and Industry sector classes. Applying a logistic regression model to predict a company's default and a full logistic regression model to the training data led to the estimated coefficients shown in Table 4.5, along with their corresponding Z and $p$values.

Table 4.5 Estimated coefficients using a full logistic regression credit scoring model.

Variable	Coefficient	Z-value	p-value
Turnover	-0.001231	-64.568443	0.000001
Leverage	0.000163	3.945933	0.000074
EBIT	-0.001479	-33.018932	0.000001
PLTax	-0.001799	-35.065625	0.000001
ROE	0.000012	2.972022	0.002958
Country	0.130159	-5.213115	0.000001
Industry	-0.552089	-25.116592	0.000001
TAsset	-0.000001	-8.50125	0.003952

From Table 4.5, we see that all variables are significant, as may be expected, given a large amount of considered training data (more than seventy thousand), which leads to high goodness of fit. Note that Country and Industry have the highest coefficients, but this does not mean they mostly impact the predictions because the variable scales differ. To understand the effect of each variable on the predictions, we employed the estimated model to predict the scores of the companies in the validation sample (30% of all data) and then calculated the Global Shapley values for each variable, summing the Shapley values for the observations in the test set (30% of the observations). The results are presented in Figure 4.3.

Figure 4.3 shows that 'PLTax' has the largest Global Shapley value: it is the variable that mostly impacts the predictions, followed by 'EBIT'. This result partially aligns with that observed in Table 4.4, as it consistently indicates the two profitability variables that present the highest difference in conditional means (PLTax and EBIT) but give low importance to financial leverage.

We built an ML credit scoring model based on the random forest model in Section

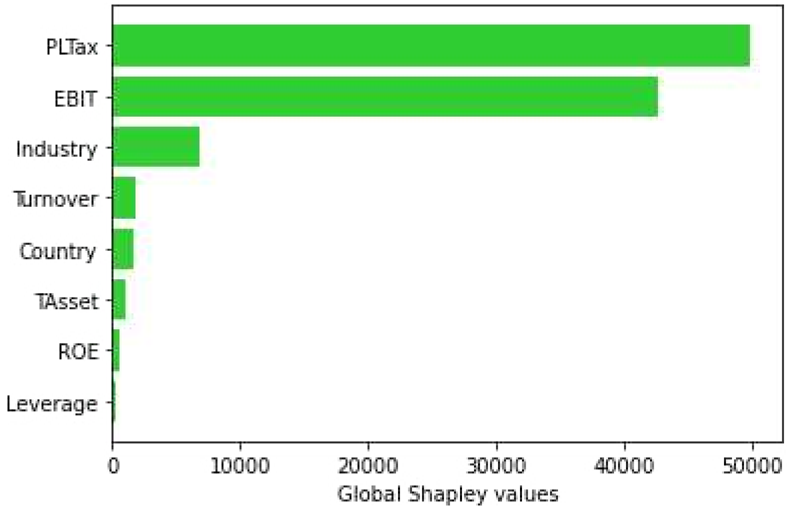


Fig. 4.3 Global Shapley values based on the predictions generated by the full logistic regression model.

3.1.2. After splitting data into training (70$\%$ of the observations) and validation samples (30$\%$ of the observations), with the same partitioning for the logistic regression, we applied the random forest GridSearch CV algorithm of Python to the training sample and used the estimated model to calculate the credit scores of the companies in the validation sample. Each company in the validation sample was then predicted: to default or not to default, comparing the model scores to a set threshold of 0.5. The performance of the full random forest model, which employs all six explanatory variables, is shown in Table 4.6 in comparison with the logistic regression model previously described.

Table 4.6 Comparison of Logistic Regression and Random Forest full models, regarding predictive accuracy measures.

Measure	Accuracy	Sensitivity	Specificity	F1 Score	AUC
Random Forest	0.97066	0.98687	0.86884	0.89051	0.92785
Logistic Regression	0.89646	0.98748	0.32459	0.46261	0.65603

Table 4.6 shows that, as expected, the accuracy of the random forest model is higher. The joint consideration of Sensitivity, Specificity, and F1 Score further indicates that the random forest model performs better because it balances sensitivity and specificity better. Consistent with this result, when the set threshold varies from 0.5, as in the AUC metrics, the random forest model strongly outperforms the logistic regression, with $AUC = 0.93$ versus $AUC = 0.63$. In conclusion, the available data indicate clear superiority in the predictive accuracy of the random forest credit scoring model over the logistic regression model.

However, although the random forest model is highly accurate, it does not produce a set of estimated coefficients, as shown in Table 4.5, indicating the relative impact of each explanatory variable. From a managerial perspective, we can predict whether to invest in an SME; however, we do not know why. From an SME perspective, it is unclear which variables improve credit scores.

To overcome this problem, we can post-process the predicted scores obtained with the random forest model using a ‘feature importance plot’. For each variable in the model, the plot represents the decrease in the Gini variability measure determined by each split of the tree induced by a given explanatory variable averaged over all tree models in the random forest built on the training data. Recall that, for a given split induced by an explanatory variable, the higher the reduction in variability, the more important the variable. The feature importance plot for our considered training data is shown in Figure 4.4.

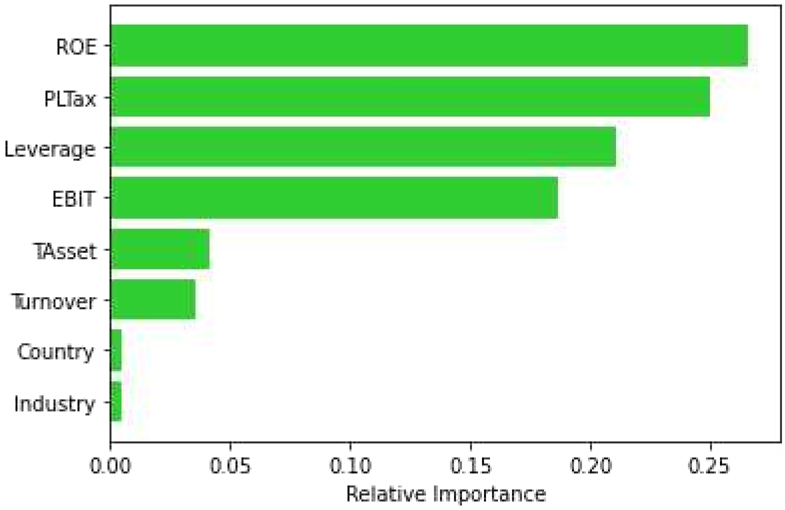


Fig. 4.4 Random Forest Feature Importance.

Figure 4.4 shows that ROE, PLTax, Leverage, and EBIT lead to the highest reduction of the Gini measure and are thus individuated as the most important variables. However, the lowest importance is related to Industry and Country, the only two variables in the sample that are not derived from the balance sheet.

Although the feature importance plot addresses, to some extent, explainability, it is not model-agnostic; it cannot be obtained for models different from random forests, such as logistic regression. Consequently, it does not allow a comparison of model explainability. Thus, we resort to a model-agnostic tool, Shapley values, calculated from the predicted credit scores in their validation sample.

Figure 4.5 shows the overall contribution of each variable, as described by the global Shapley values: the Shapley values for each variable, summed across all observations.

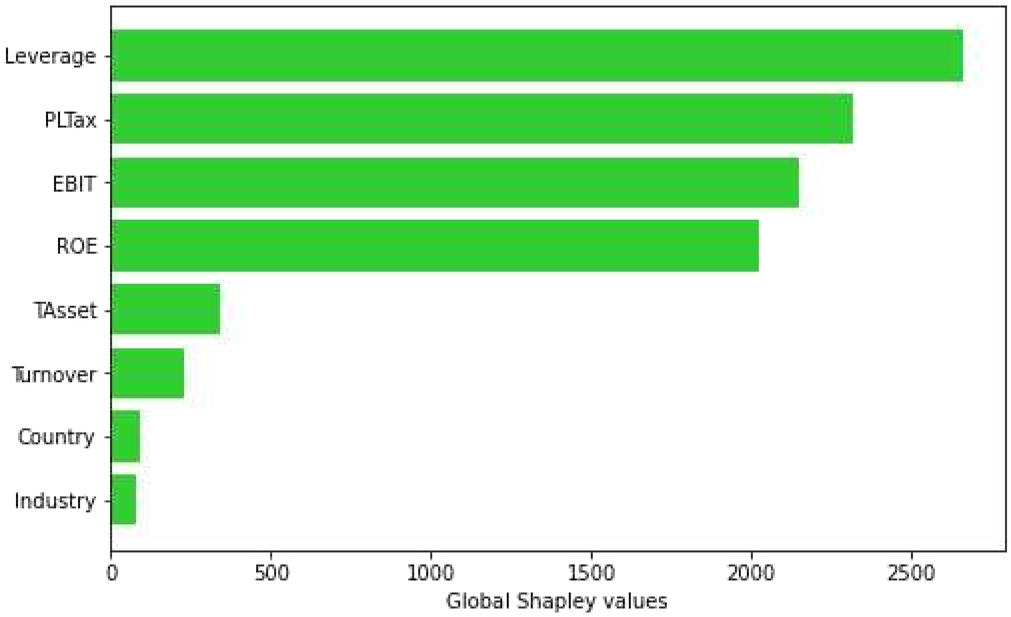


Fig. 4.5 Global Shapley values importance.

From Figure 4.5 note that the most explainable variable is ‘Leverage’, followed by ‘PLTax’ and ‘EBIT’, consistently with the difference in conditional means, and with the feature importance plot in Figure 4.4, but differently from what obtained applying Shapley values to logistic regression. Here, ‘Leverage’ is the least important variable. The same figure shows that the global Shapley values for ‘Country’ and ‘Industry’ are small, consistent with the feature importance plot but different to what occurs for the logistic regression in Figure. 4.3.

From a financial viewpoint, the Shapley values of the random forest credit scores indicate that the probability of default of an SME is mainly determined by its probability (as measured by ROE, PLTax, and ROE) and by its financial leverage; less affected by its size and operating revenues (as measured by Tasset and Turnover); and little affected by its corresponding Country or Industry, differently from what occurs using a logistic regression model.

To understand which variables are statistically significant to explain the probability of default, we applied our proposed selection procedure: a stepwise variable selection based on the comparison of the AUC and the ordering established by the Global Shapley values in Figure 4.5. Our procedure differed from classical stepwise procedures that compare models in terms of their likelihood. Instead, we compared the models in terms of their predictive accuracy. The advantage of doing so is generality: we can compare models with an underlying probabilistic model, such as logistic regression and all ML models, such as random forest models.

More precisely, we employed a backward selection procedure, which progressively eliminated variables from the full model, following the order determined by the global

Shapley values in Figure 4.5: from the least explainable ('Industry') to the most explainable ('ROE'). Each variable was removed from the least explainable to the most explainable variable. Specifically, a variable is removed from the model when its additional contribution to predictive accuracy, as measured by the Area Under the Receiver Operating Characteristics (AUC), is not statistically significant; that is, it leads to a DeLong test with $p$value larger than a threshold (e.g. 5%). The procedure was stopped when p was lower than the set threshold. As a result of our proposed procedure, the selected model is highly predictive and explainable using a parsimonious set of predictors.

Table 4.7 lists the results of this stepwise procedure. It contains the AUC values and the p-values of the DeLong test, corresponding to comparing subsequent pairs of AUCs.

Table 4.7 DeLong Tests of the considered pairs of Random Forest models.

Removed Variable	Number of Variables in model	AUC	P-value	H Measure
-	8	0.927856	-	0.818273
Industry	7	0.927604	0.167723	0.817766
Country	6	0.928045	0.258918	0.818519
Turnover	5	0.924413	0.000009	0.808577

Table 4.7 shows that 'Industry' is the least explainable variable: the first candidate for removal. The comparison of the AUC of the entire model, when all predictors are used in the random forest model, against the model without 'Industry', leads to a p-value of the DeLong test equal to 0.16772, leading to select the simpler model, without 'Industry'. The next variable candidate for removal is 'Country'. Comparing the AUC of the model without 'Industry' and 'Country' against the model which excludes only 'Industry', we found that the p-value of the DeLong test equals 0.25892, so the model can be further simplified. The procedure continues until a variable whose exclusion leads to a significant decrease in predictive accuracy is identified. In our case, this is the third most explainable variable, 'Turnover', for which the p-value is smaller than the threshold, leading to a rejection of model simplification and stopping the variable removal procedure. In conclusion, from Table 4.7, we obtain that the best trade-off between explainability and predictive accuracy is provided by a model that includes all available financial indicators but not the variables which describe the 'Industry' and the 'Country' of the companies.

From a financial viewpoint, this result indicates that the binarised ratings assigned by the rating agency are 'fair' across countries and sectors, with no bias in terms of financial inclusion.

To verify the results obtained in Table 4.7, we provide the results from applying Hand's H statistics Hand (2009) to our models. The results are consistent with those of the AUC: the values for the H measure are similar, approximately 0.818, for the first three models, before removing 'Turnover' and, when 'Turnover' is removed, H drops down to 0.808577, showing that the model should not be simplified any further, consistent with the results obtained applying DeLong's test to the AUCs.

Thus, our proposed selection procedure leads to a simpler random forest credit scoring model than the entire model, with six variables instead of eight. However, this does not result in a significant loss of predictive accuracy. For completeness and comparison, we should apply our proposed stepwise procedure also to the logistic regression scoring model, following the variable ordering determined by Figure 4.3. In this case, removing 'Leverage', the least explainable variable in Figure 4.3, leads to a p-value smaller than 0.05. Hence, the null hypothesis is rejected, and the full model cannot be simplified without a significant loss of accuracy. Thus, the selected random forest model with six variables is more parsimonious than the selected logistic regression model, a full model with eight variables.

Therefore, we conclude that the random forest model selected by our proposed procedure is more accurate and parsimonious than the selected logistic regression model.

4.6 Conclusions

Ensemble ML models, such as random forests, can improve the accuracy of credit scoring models but are not explainable. Explainable AI methods such as Shapley values can be employed to post-process credit scores to achieve explainability. This study employed Shapley values to achieve explainability and guide variable selection, leading to a parsimonious model that is a good trade-off between predictive accuracy and explainability.

To achieve this goal, we proposed a model selection strategy in which global Shapley values ordered the candidate explanatory variables in terms of their predictive importance, and a backward stepwise selection procedure, based on the comparison of predictive accuracy, was implemented to select a 'statistically optimal' subset of variables. Our proposal is applied to a database containing credit ratings for a large set of European SMEs, the values of six financial ratios from their 2020 balance sheets, and their country and sector of belonging. These results indicated that the nonlinear random forest credit scoring model was more accurate than the logistic regression. The application of our procedure also showed that the selected random forest model was more parsimonious than the selected logistic regression model because it depended only on balance sheet ratios and not on the country or industry sector of a company, with no bias in terms of financial inclusion.

From a methodological viewpoint, our proposed method: i) fills a gap, as it provides a model comparison procedure based on both accuracy and explainability, which can be equally applied to all types of ML models; and ii) leads to a credit scoring model which is a good trade-off between predictive accuracy and explainability.

From a managerial viewpoint, our model can support banks, fintech companies, and regulators in developing and supervising ML models for credit scoring compliant with regulatory requirements, particularly those concerning AI.

Our proposal makes three main contributions to literature. For research scholars, it proposes a novel model comparison approach, which combines explainability with predictive accuracy; for financial and fintech managers, it proposes a way to make AI applications explainable and, therefore, acceptable; for policymakers and regulators, it provides a methodology able to check whether a specific AI application for credit scoring is compliant with the existing regulations.

Our study is built on the standard axiomatisation of the Shapley value, which is only suitable for binary responses such as credit default. However, when continuous response variables such as loss given default or exposure at default are considered, the proposed method can be extended by considering the Baseline Shapley value (Bshap) or the Random Baseline Shapley value (Sundararajan and Najmi 2020) when implementing random forest or other ML approaches.

Further research is needed to experiment with the proposed model selection procedure using alternative Shapley axiomatisations. An interesting avenue of research would be to understand the impact of balance sheet variables on companies' financial exposure by extending the linear regression analysis of Fasano and Cappa (2022) to an ML context. Further research is also needed to consider the interpretation of the predictions in terms of their 'fairness', that is, to establish how independent the credit scores from country and industry sector or, for consumer credit applications, from gender, race or other types of stratifications.

CHAPTER 5

InstanceSHAP: An instance-based estimation approach for Shapley values

Babaeic, G., Giudicid, P. (2023). InstanceSHAP: an instance-based estimation approach for Shapley values. Behaviormetrika, 1-15.

5.1 Abstract

The growth of artificial intelligence applications requires to find out which explanatory variables mostly contribute to the predictions. Model-agnostic methods, such as SHapley Additive exPlanations (SHAP) can solve this problem: they can determine the contribution of each variable to the predictions of any machine learning model. The SHAP approach requires a background dataset, which usually consists of random instances sampled from the train data. In this paper we aim to understand the insofar unexplored effect of the background dataset on SHAP and, to this end, we propose a variant of SHAP, InstanceSHAP, that uses instance-based learning to produce a more effective background dataset for binary classification. We exemplify our proposed methods on an application that concerns Peer-to-Peer lending credit risk assessment. Our experimental results reveal that the proposed model can effectively improve the ordinary SHAP method, leading to Shapley values for the variables that are more concentrated on fewer variables, leading to simpler explanations.

5.2 Introduction

The majority of machine learning (ML) algorithms are complex, and it is very difficult to understand how they exactly process the data. Indeed, in addition to accuracy, explainability is an important characteristic required to a machine learning model. Machine learning is often criticized, because of its complexity, that can lead to a non transparent ”black box” model. From another point of view, the trade-off between model complexity and interpretability makes it difficult to understand why sophisticated ML models perform so well. This makes their users not to rely on the decisions made by the ML-based models (Ribeiro et al. 2016*a*).

A key solution to this issue is the recent introduction of model-agnostic explanation approaches to understand the contribution of each predictor towards the overall prediction (Burkart and Huber 2021). These model-agnostic methods play an important role in generalizing the use of ML models in different decision-making problems such as finance (Dahooie et al. 2021), and medicine (Law et al. 2022).

Lundberg and Lee (2017) proposed a model-agnostic attribution method for machine learning, the SHapley Additive exPlanations (SHAP) method, based on the classic Shapley values from game theory. Shapley values are based on coalition game theory. Coalition game theory involves a group of players collaborating to generate value, which can be linked to a group of people forming a company to make a profit. The Shapley value is a method used to distribute this profit equitably among the players, taking into account their individual contributions. The Shapley value represents the average marginal contribution of a player, across all potential coalitions. To calculate the marginal contribution, the difference in predictions is measured with and without a particular player. The Shapley values are then derived through a weighted average of these contributions, and this process is repeated for all players. This approach forms the fundamental basis for using Shapley values to interpret model predictions in terms of feature variable contributions, although there may be some variations in how the process is implemented. The SHAP value technique aims to implement the Shapley values theory to explain the prediction of any machine learning model in terms of the contributions of each feature to the predictions.

SHAP has been a popular method in the feature attribution literature (Janzing et al. 2020) and many efforts have been devoted to improving it. For example, Aas et al. (2021) address the problem of dependent variables in the SHAP algorithm. Covert and Lee (2021) mention the heavy computational costs of calculating the SHAP values and develop a new version of KernelSHAP that yields fast new estimators. Also in another study by Kwon and Zou (2022), a rigorous analysis is performed to find where Shapley values are mathematically suboptimal. For this purpose, they propose WeightedSHAP that is a simple modification of SHAP and show their proposed model is better able to identify influential features compared to a standard SHAP.

SHAP values proposed by Lundberg and Lee (2017) are the Shapley values of a conditional expectation function of the original model. The exact computation of SHAP values is challenging, especially when the dimension of the data increases and, therefore, SHAP becomes computationally expensive. To mitigate this problem, Lundberg and Lee (2017) propose to approximate SHAP values with KernelSHAP. An alternative approximate explainer is the TreeSHAP (Lundberg et al. 2019).

A prerequisite of these approximate explainers is a background dataset, a set of data with the same predictors as the training sample. This because, when a feature is removed from a coalition to calculate its corresponding Shapley value contribution, its values become as they were missing. To obtain the predictions necessary to calculate the Shapley contributions, we have to replace the missing feature values with values that are functions of the background data, such as the means of the predictors (Merrick and Taly 2020).

In this paper, we try to improve SHAP explanations for binary classification problems by focusing on the background data used to estimate SHAP values. In particular, we consider a credit scoring model for peer-to-peer (P2P) lending, a new method of lending to individuals without intermediaries of traditional banks which happens through online lending platforms. While some recent works, such as the study done by Aas et al. (2021), improve SHAP contributions exploiting the correlation between variables and using conditional sampling for dependent features, we utilize the similarities among observations proposing an instance-based approach to provide a conditional selection for the background data.

The rest of this paper is organized as follows. Section 6.3 discusses the related work on P2P lending and SHAP values. Section 5.4 presents the methods we use in this paper. An experimental study, including description of the selected data to validate the proposed model, and the obtained results, is presented in Section 5.5. Finally, Section 5.6 concludes the study and contains some remarks for future research.

5.3 Related Work

5.3.1 SHAP Values

Relevant methods that have been used in the literature to explain the predictions of the ML models are LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al. 2016*a*) and SHAP (Lundberg and Lee 2017). The former is a local interpretability framework that provides a number of explainable models (for example linear regression) that approximates locally each observation. The predictions of these models are reliable only in the neighbourhood of the corresponding observation.

In this study, we focus on the SHAP approach, which is based on the Shapley game theory concept. SHAP is a method proposed by Lundberg and Lee (2017) to explain individual predictions, that borrows concepts of both LIME and Shapley value theory, representing Shapley value as an additive feature attribution method: a linear model. To implement it, in the same paper, Lundberg and Lee (2017) introduce KernelSHAP, a kernel-based estimation approach for Shapley values. In another paper, Lundberg et al. (2019) propose TreeSHAP for tree-based models, such as random forests. In our experimental study, we use TreeSHAP to provide explanations. Here, explanations refer to the contributions of the predictors to the output of a model's predictions. Therefore, using TreeSHAP, we help the users of an ML model to understand how each of predictors contribute to predictions.

SHAP values have been used widely in different areas of research. For example, in the study done by Giudici and Raffinetti (2022), a new explainable model to assess cyber risks is proposed. Their proposed model combines SHAP values with a Lorenz Zonoid based statistical normalization to assess cyber risks. This explainable model can identify the drivers of cyber risk, which should be controlled. Through the experimental analysis, they show that the proposed model has accurate, explainable and robust predictions for cyber risk management. Walambe et al. (2021) considers the secure Blockchain nature for ML-based credit scoring but they also enable explanations for customers in a secure manner. Their results represent the trustworthiness of end-users in an explained model. Buckmann et al. (2022) provide inferences on the SHAP values and check if the contribution of a feature is significant or not. Kwon and Zou (2022) find the contribution of each feature to the Receiver-operating Characteristics (ROC) curve and the Area under the ROC curve (AUC) of a model as a robustness measures. In this study, SHAP values are calculated to find how adding a feature to a coalition of available predictors changes the robustness of the model compared to a random classifier for which AUC is equal to 0.5. In a random classifier, the True Positive Rates (TPRs) are equal to False Positive Rates (FPRs). If the random classifier is assumed to be the baseline for the TPRs, the difference between a TPR and the associated FPR can be explained as the contribution of features when they are added to a coalition.

Despite the wide use of SHAP in the literature, the ordinary SHAP values have some weak points that have been the motivation of many studies. For example, Aas et al. (2021) criticize the assumption of independency between predictors and propose four approaches to take dependency among variables into account. For instance, they suggest considering a conditional distribution to approximate SHAP values. Another paper by Molnar et al. (2023) mentions the problem of dependent features for finding the contribution of the predictors. The authors focus on partial dependence plot (PDP) and permutation feature importance (PFI), two model-agnostic ML interpretation methods, that are based on perturbing features. They address the problem of dependent features by sampling a feature conditional on other features. In other words, they propose conditional subgroup PDP and PFI called csPDP and csPFI respectively, based on the conditional subgroup permutation. In order to incorporate dependencies among predictors to SHAP, Li et al. (2021) propose a novel concept-based neighbor Shapley approach to find the contribution of each concept (predictors as the players in the game theory) by considering its neighbors, and interpret model knowledge with both instance-wise and class-wise explanations. As a result of this approach, the interactions among predictors are fully considered.

Another factor that affects the explainability of SHAP is the background data, as recently studied by a few authors. Kwon and Zou (2022) study the effect of the different background data sizes on explanations by doing bootstrap sampling. Also, Yuan et al. (2022) show the importance of exploring the effect of the sampling size on SHAP. Their results indicate that SHAP values and variable rankings fluctuate when using different background datasets but it is mentioned that fluctuations decrease with enlarging the background dataset. Albini et al. (2022) mention the importance of background datasets when using SHAP values and propose Counterfactual SHAP (CF-SHAP) which is a variant of SHAP. In their approach, counterfactual information is utilized to produce a background dataset. Finally, they show the superiority of CF-SHAP with respect to existing methods by using public datasets and ensemble tree models.

Considering the importance of the background dataset in calculating the contributions of features, we contribute to the literature on SHAP values by proposing the InstanceSHAP approach, which is based on conditional sampling. Here we apply conditional sampling to the background data to obtain better SHAP values.

5.3.2 Peer-to-Peer Lending

Since the proposed InstanceSHAP algorithm is applicable to binary classification problems, we consider a credit scoring problem in P2P lending to validate our model. In this subsection, we review the literature on peer-to-peer lending market.

P2P lending platforms such as ”LendingClub” in the last decades have led to a boom in the online lending market. Borrowers, who usually have low credit scores from traditional lending institutions, are connected to the individual lenders through these online platforms (Babaei and Bamdad 2021). Hence, there has been a significant rise in research on the credit risks analysis for the P2P lending platforms. Ariza-Garzon et al. (2021) provide a good review of the academic literature on P2P lending. Based on the business problem categories of the published documents described in this paper, the most popular topic in the P2P lending research has been default classification (i.e. credit scoring).

Guo et al. (2016) propose an instance-based credit risk management that not only predicts the class (zero and one for fully paid and failed loans respectively) but also can find the optimal portfolio constructed of available loans considering the risk minimization objective for a predefined level of return. We utilize their proposed instance-based method in our explainable approach, more information are provided in Subsection 5.4.2. Serrano-Cinca and Gutiérrez-Nieto (2016) consider another significant objective of investments and proposes a profit-scoring model.

Bastani et al. (2019) considers both credit scoring and profit scoring in a two-stage scoring approach to help lenders in the P2P lending. In the first stage of the proposed model, non-default loans are identified and then moved to the second stage to be evaluated based on the profitability. In both stages, deep learning is used to build the predictive models. The numerical results indicate that the two-stage scoring approach outperforms the existing credit scoring and profit scoring approaches.

In almost all research papers, the proposed ML-based models outperform the simpler linear models. However, these accurate credit risk models lack explanatory power while individuals demand effective and clear explanations. Bastani et al. (2019) propose a machine learning credit scoring model that is both accurate and transparent. More precisely, they compare ML algorithms such as decision tree, random forest and XGBoost, with logistic regression, which is a well-established method in credit risk management. Regarding the explainability, they apply Shapley values and they also extend the notion of SHAP values to logistic regression. Their results show that, in addition to performance, the ML approach can detect complex relationships that cannot be easily found by logistic regression.

A key challenge for ML models in finance is providing explanations on decisions for investors. In our study, we apply SHAP values to an ML-based credit scoring problem for the P2P lending and we improve the robustness of the explanations using Instance-SHAP.

5.4 Methodology

5.4.1 Shapley values as a feature attribution method

SHAP values are based on the definition of Shapley values (Shapely 1953), which calculate the marginal contribution of each variable towards the predictions. From a game theory point of view, features of an instance in the data set behave as players in a group (coalition), and Shapley values represent the distribution of the prediction among the features according to their contribution. Following these assumptions, the SHAP algorithm calculates the contribution of each variable to each prediction considering its additional effect to all possible coalitions (groups) of the other variables.

The Shapley value of variable i for $i = 1, \ldots, n$, is calculated as follows:

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus i} \frac{|S|!(n-1-|S|)!}{n!}[f(S \cup i) - f(S)] \tag{5.1}$$

In the above expression, i is a feature, S is a subset of features, $|S|$ is the number of variables in the subset S, n is the number of features in the model,and $\mathcal{F}$ is the set of all possible coalitions. This because the effect of a feature depends on the coalition it is added to and, therefore, it is necessary to consider all possible coalitions. Indeed, the difference between the output of the model with and without the feature is computed for all possible subsets without feature i. In other words, to compute the effect of each feature, the difference between the prediction of the model including that feature, $f(S \cup i)$, and the prediction of the model excluding the feature, $f(S)$, is calculated. It should be noted that the value of the features that are not available in the coalition needs to be estimated and, for this purpose, a background dataset is employed.

We remark that, for n variables, we will have these calculations for 2^n coalitions: the number of coalitions increases exponentially while n increases. This is the main disadvantage of Shapley values that has led to the use of explainers such as KernelSHAP (Lundberg and Lee 2017) and TreeSHAP (Lundberg et al. 2019) to approximate Shapley values.These explainers are provided in data analysis languages, such as Python. In our application, since we employ random forest, we utilize a non-conditional version of TreeExplainer [1]. TreeExplainer is a built-in explainer in the shap python library that implements the TreeSHAP algorithm, particularly useful for tree-based models such as Decision Trees, Random Forests and gradient boosting.

5.4.2 An Instance-based approach

The background dataset which is needed to estimate SHAP values is usually selected randomly from the train data. We propose an alternative approach to select background data that looks more similar to the test data. Indeed, a model (that is built on train data) needs to be applied to the test data and, therefore, we can assume that if background data is similar to test data explanations will be improved. To achieve this goal, a feasible approach is to use an instance-based method that has the ability to find the train observations that are the most similar instances to the test data that we pass through the explainers to calculate SHAP values. Therefore, instead of random sampling from train data, we can find background data using the instance-based method, which provides a data distribution more similar to that of the test data for the estimation of SHAP values.

The similarity between train and test observations is reflected by weights that are found as a function of distance among instances. This means that a weight depends on the similarity between train and test observation. Intuitively, more similar items should have greater weight, and vice versa.

In other words, the weights for an instance-based model can be calculated using the distance between train and test instances. In binary classification problems, such as the credit risk management problem in this study, we can define this distance measure as a default likelihood distance between the ith instance in train data ($i = 1, \dots, n$) and the jth instance in test data ($j = 1, \dots, m$), as follows:

$$d_{ij} = |PD_i - PD_j| \tag{5.2}$$

[1] https://shap-lrjball.readthedocs.io/en/latest/generated/shap.TreeExplainer.html

Here PD_i and PD_j are the probabilities of default for loans i and j, respectively. It is obvious that we expect a higher amount of weight when instances are near to each other. However, this is not an optimal weighing approach and, thus, as suggested by Guo et al. (2016) we use kernel regression to find the weights between train and test data. Indeed, as the authors mention, kernel weights for instance-based modeling is a technique to find the relation between a pair of random variables.

More formally, the kernel weights between the ith train observation and the jth test observation as defined by Guo et al. (2016) are calculated using the following equation:

$$W_{ij} = \frac{K(\frac{d_{ij}}{h})}{\sum_{j=1}^{m} K(\frac{d_{ij}}{h})} \tag{5.3}$$

Here, $K()$ is the Gaussian kernel function whose expression is:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \tag{5.4}$$

where h (h¿0) is the bandwidth which follows the definition given by Guo et al. (2016). After finding the optimal h, the weights between train and test observations are calculated. These weights are then used to find the background distribution for SHAP values.

5.4.3 InstanceSHAP setting

Our proposed algorithm, InstanceSHAP, is based on the assumption that if a more similar distribution is selected for estimation of the variable in interest, finding the contribution of the variable will improve. Considering this assumption, similar observations to test data for which we want to find the contribution of variables are selected among the train observations using an instance-based learning approach which is explained in Section 5.4.2.

For a better understanding of our proposed approach, the InstanceSHAP is illustrated in Algorithm 1.

From the above algorithm note that, first of all, different subsamples of the same size are selected from the dataset. All steps of the algorithm are applied to each of these subsamples. In particular, observations in each subsample are grouped into train and test data. A machine learning model which, in our case, is the random forest, is trained on the train data to predict the probability of default for the train and test observations, PD_i and PD_j respectively. Then, using Equation 5.2, d_{ij} is calculated. d_{ij} shows how far train and test observations are from each other. After that, using this distance measure and the optimal bandwidth that is found based on the description in (Guo et al. 2016), kernel weights are calculated regarding Equation 5.3.

It should be mentioned that for the process of optimizing the bandwidth suggested

Algorithm 1 InstanceSHAP

```
Input:
Data
ML model
R = Iteration times

Output:
Background dataset with a more similar distribution to the data for which we find the contribution of variables
SHAP values

while $r \leq R$ do
    Subsample $(r)$                          ▷ 1000 random observations from Data
    Traindata($r$), $i = (1, \ldots, n)$     ▷ 70% of data
    Testdata($r$), $j = (1, \ldots, m)$      ▷ 30% of data
    Find $d_{ij}$ (Equation 5.2), distance between train and test observations
    Calculate Kernel weights:
    $W_{ij} = \frac{K(\frac{d_{ij}}{h})}{\sum_{j=1}^{n} K(\frac{d_{ij}}{h})}$
    Threshold = $\overline{W}$
    Background data($r$) = Train observations for which weight values
    are higher than the threshold
    if Number of background data rows $\neq 0$ then
        Set TreeExplainer with the trained ML model and
        the background data
    else if Number of background data rows $= 0$ then
        Assume zero value for all missing variables
    end if
    Find SHAP values through the explainer
end while
```

by Guo et al. (2016), we consider the range [0.25,1.5), with a step equal to 0.1. Next, we select observations from train data for which the corresponding weight values are higher than a threshold that in our study is set to be equal to the average of the calculated weights. In this case, if no observations meet this condition, zero values are considered as the background data to estimate the missing features. Finally, the instance-based background distribution is used in a TreeExplainer to find SHAP values. We note that we used the interventional (a.k.a., non-conditional) version of SHAP (default setting) in the shap Python library.

To check the superiority of our proposed InstanceSHAP, we also calculate the ordinary SHAP values. This means that we find the contribution of variables for all subsamples but this time the background dataset includes all train observations. Finally, to compare the SHAP values generated by InstanceSHAP with the ordinary SHAP, we consider the variability of SHAP values using the Gini heterogeneity index. For a better understanding of the procedure of InstanceSHAP, all the steps in Algorithm 1 are further visualized in Figure 5.1.

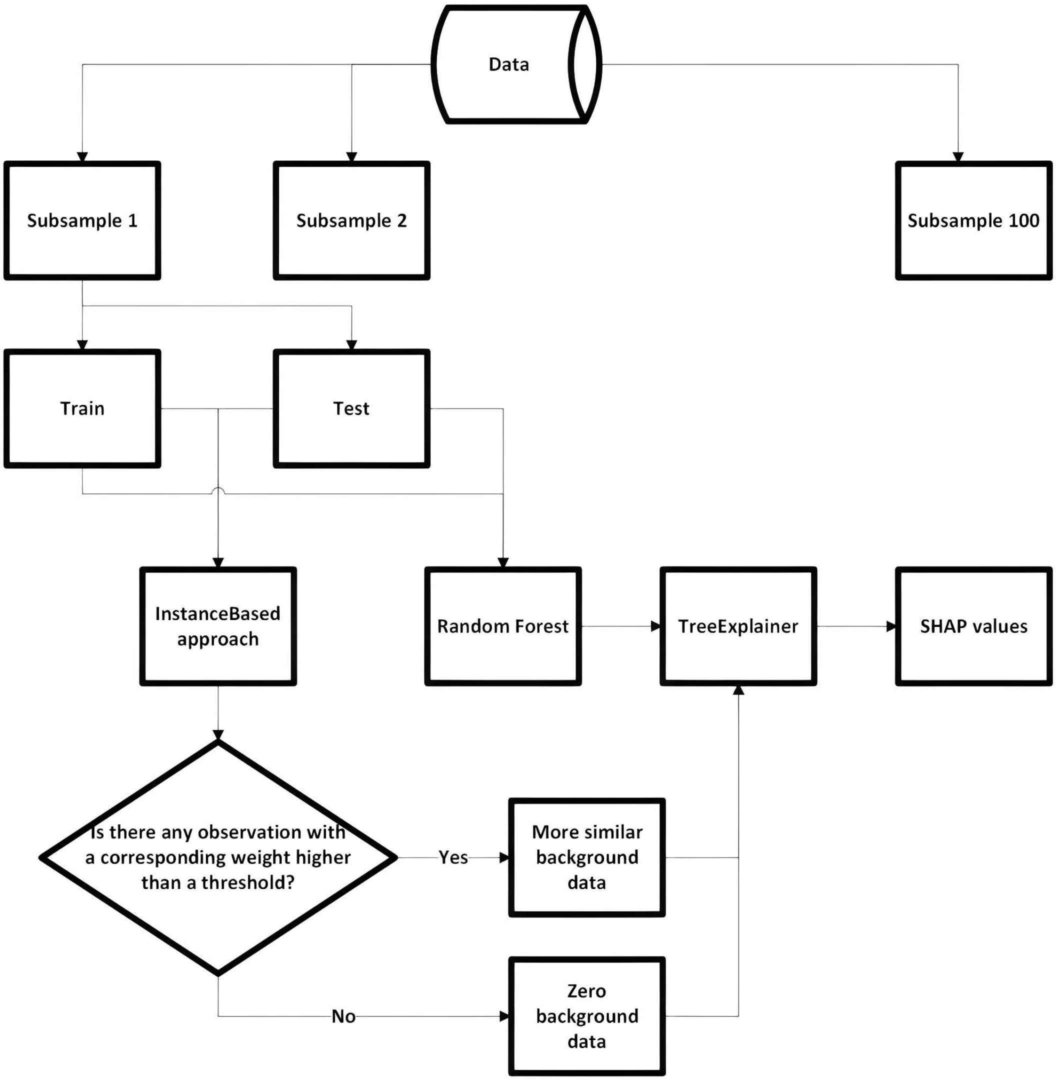


Fig. 5.1 InstanceSHAP Flowchart

5.5 Application

5.5.1 Data Description

To validate our proposed InstanceSHAP algorithm, we consider an example of a credit scoring problem using a commonly used dataset in the P2P Lending Credit Scoring literature. This dataset [2] includes information of individuals who applied for loans on the LendingClub P2P lending platform. The time horizon we consider for our paper is from 2007 to 2018. The dataset contains loans with different statuses such as Current, Late, Fully Paid, and other. However, We consider only observations for which ”Loan Status” is equal to ”Fully Paid” (1) or ”Charged Off” (0). This binary variable is the response variable in the credit scoring model.

Concerning the predictors, LendingClub dataset is a large dataset which includes more than 100 variables. However, we select only the most common used variables in the literature, represented in Table 5.1.

Table 5.1 Variable description. The table shows the description of the selected explanatory variables related to the LendingClub dataset. More information on the data is available at: `https://www.kaggle.com/datasets/jonchan2003/lending-club-data-dictionar`

Variable	Description
Annual income ($)	The self-reported annual income provided by the borrower
Dti	Ratio of the borrower’s total monthly debt to the borrower’s self-reported monthly income
Interest rate	Interest Rate on the loan
Grade	Grade assigned by LendingClub platform to the borrowers which shows riskiness of them. From A to G, A is the safest and G is the riskiest
Loan amount ($)	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value
Loan status	Current status of the loan
Open accounts	The number of open credit lines in the borrower’s credit file
Public records	Number of derogatory public records
Delinquencies in 2 years	The number of 30+ days delinquencies for the past two years
Loan purpose	The purpose of borrowers to get the loan
Inquiries in the last 6 months	The number of inquiries in the past 6 months
Home ownership	The ownership of home clarified by the borrower

After preprocessing the data, including dealing with missing values and encoding, explanatory analysis of the selected variables are as presented in Table 5.2. We should mention that for dealing with missing values, mode and mean values are utilized for the categorical and numerical variables, respectively. In addition, for encoding the categorical variables, one hot encoding is applied to our dataset.

[2] https://www.kaggle.com/datasets/wordsforthewise/lending-club

Table 5.2 Summary statistics for the LendingClub data.

Type	Variable	Mean	STD	Min	Max
Numerical	Loan amount	14588.273	8970.471	500	40000
	Annual income	77369.565	117821.800	0	110000000
	Dti	18.567	13.087	-1	999
	Open accounts	11.606	5.575	0	90
	Inquiries in the last 6 months	0.613	0.902	0	8
	Delinquencies in 2 years	0.313	0.876	0	42
	Public records	0.208	0.590	0	86
	Interest rate	13.170	4.828	5.310	30.990
Categorical	Grade	7 Grades, From A(Safest) to G(Riskiest)			
	Loan purpose	8 Categories, medical, car and etc.			
	Home ownership	4 categories, Rent, Own, Mortgage and Any			
Binary	Loan status	1=Failed Loans, 0=Fully Paid Loans			

To have a better understanding of the riskiness of the groups of the categorical variables, we utilize the distributions in Figure 5.2.

From Figure 5.2, as we expect, G is represented as the riskiest grade of borrowers, that means most of the the loans categorized as G failed on their payments. In terms of Home ownership, Mortgage shows the lowest share of failures so we can mention it as the safest home category. In terms of purpose, most loans issued for the small business purpose failed to payback completely to the lender.

5.5.2 Results

In this section, we present the results we achieved applying our proposed InstanceSHAP method.

The original data is a huge dataset and contains over 1 million rows. To decrease computational time, we sample 1000 observations from the original data. Then, we select 70% of the dataset for training, and use the remaining, 30%, for testing. Using the Algorithm 1 explained in Section 5.4.3, we find the most similar train observations to the test data and use them as the background data for finding SHAP values. Therefore, when we want to approximate the contribution of variables to the predictions of ML models for the test data, we have a background dataset with a more similar distribution to the test data.

Considering Algorithm 1, after finding SHAP values as an output of the algorithm, the whole process is repeated for another set of train and test data. The iteration times of the loop in our study is set to be equal to 100, without loss of generality.

For the purpose of comparison and validation of the InstanceSHAP algorithm, we also computed the ordinary SHAP values. Here, simple ordinary SHAP values mean the contribution of variables found by the TreeExplainer and the whole train data as the background data.

In each round of the calculation, we calculate the Gini concentration index of SHAP

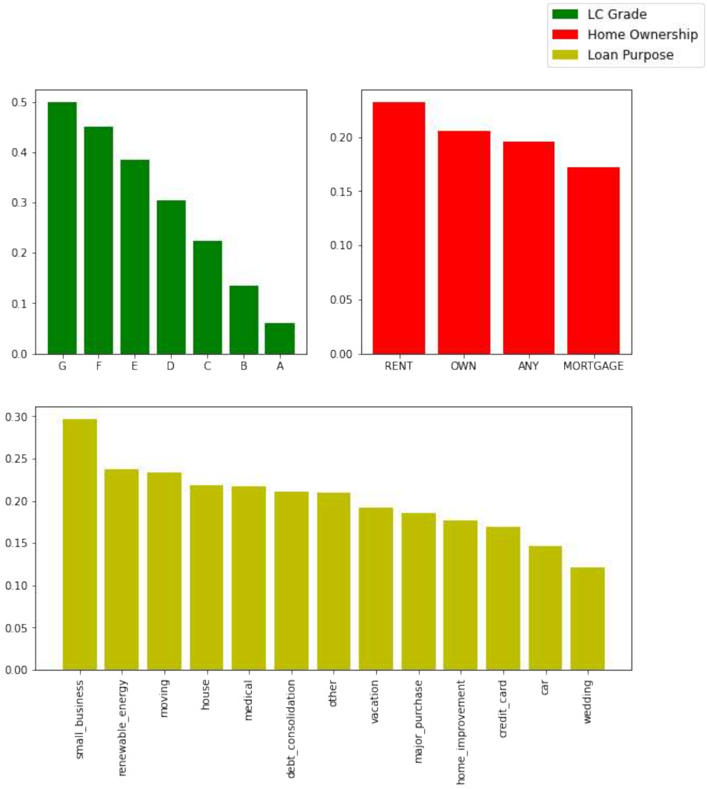


Fig. 5.2 Distribution of class 1 among LC grade, Home ownership and Purpose categories; Each portion represents the percentage of failed loans in the corresponding category.

values found by both explanation methods: InstanceSHAP and OrdinarySHAP. After having obtained all 100 Gini indices, we find their mean for each method. The results are represented in Table 5.3.

Table 5.3 Comparison of the explanation models

Method	InstanceSHAP	OrdinarySHAP
Gini Index	0.54	0.51

From Table 5.3 it can be seen that the value of Gini for the proposed InstanceSHAP method is slightly higher than that for OrdinarySHAP. This shows that, even for a relatively simple Random Forest model as the one we employ, our proposed model has is more concentrated and, therefore, is more selective, leading to more parsimonious models. We remark that we have used a simple Random Forest model without any special settings for the hyper parameters. By appropriately tuning the hyperparameters, it is very likely that our propsoed model will reach a higher superiority. We also remark that we are considering a very large dataset, with almost 2 million observations. A small difference in performance, as the one we found, is relevant, and shows that our proposal is promising.

We also point out that our proposed algorithm is not related to resampling method for predictive methods Austin and Tu (2004), Murtaugh (1998) In resampling approaches used to improve predictions or for the selection of predictors it is the train data that is usually resampled, while test data do not play a role. Instead, in our approach, we aim to find a subsample from train data that has a similar distribution to the test data, by means of our proposed instance-based learning approach.

5.6 Conclusion

We have proposed InstanceSHAP, a variant of OrdinarySHAP to better explain predictions in binary classification problems, such as the credit scoring utilized in this study. SHAP explanations vary when background data changes. In the paper we have shown that providing a background data with more similar distribution to the test data leads to better explanations. The comparison of the Gini concentration index for our proposed InstanceSHAP, versus the OrdinarySHAP, in a credit scoring application, leads to a more concentrated distribution, which indicates a better discriminatory power of the explanations.

Our results suggest that the variability of the explanations caused by the choice of the background dataset should not be ignored, and more work should be dedicated to the topic of choosing appropriate background data. In this respect, future work should include: (i) extending the work to the comparison of the explanations obtained with different background data selections; and (ii) extending the work to the comparison of the explanations obtained with different machine learning models.

CHAPTER 6

How fair is machine learning in credit lending?

Babaei, G., Giudici, P. (2023). How fair is machine learning in credit lending?

6.1 Abstract

Machine learning models are widely used to decide whether to accept or reject credit loan applications. However, similarly to human-based decisions, they may discriminate between special groups of applicants, for instance, based on age, gender, and race. In this paper, we aim to understand whether machine learning credit lending models are biased in a real case study, that concerns borrowers asking for credits in different regions of the United States. We show how to measure model fairness using different metrics, and we explore the capability of explainable machine learning to add further insights. From a constructive viewpoint, we propose a propensity matching approach that can improve fairness.

6.2 Introduction

”Equal opportunity by design” across different demographic groups is an important requirement for decision-making processes. For this reason, data driven decisions should pay special attention to ”protected variables”, containing information on individuals from different demographic groups (Mitchell et al. 2021).

Currently, a vast number of decisions in our daily life are increasingly being taken by machine learning (ML) algorithms. Hence, the need to propose ML algorithms that are not only accurate but also fair (Pessach and Shmueli 2022) becomes necessary.

From a statistical viewpoint, a decision-making model is considered fair when it behaves equally towards the individuals from different groups of a protected variable. In other words, the value taken by a protected (sensitive) variable must not impact the results of a statistical model.

From a machine learning viewpoint, the principle of fairness ensures that there is no discriminatory or unjust impact across different demographic groups, such as gender when algorithms make an automated decision.

Fairness measures are generally classified into three main groups: 1. Group-based measures; 2. Individual-based measures; 3. Counterfactual fairness measures. The first category compares the output of a classification model for the groups of the sensitive variable. The groups created when splitting individuals by a sensitive attribute are referred to as sensitive or protected groups. Among group-based measures, the simplest is demographic parity, which considers a model fair if the proportion of the protected variable is the same across all class categories. For example, an equal number of males and females should belong to the ”rejected” class for a loan request. Other group-based measures can be derived from the confusion matrix of a model, as we shall see later.

In contrast to group-based metrics, which focus on comparing two or more groups, individual-based measures consider the outcome for each individual (observation). In other words, they consider a model fair if similar individuals are treated similarly by the model. The similarity between individuals is usually defined in terms of distance metrics which represents how similar individuals are ”close” to each other, in terms of their corresponding features. Finally, counterfactual-based metrics use causal inference models to measure fairness. Counterfactual fairness considers a decision to be fair for an individual if it aligns with the decision that would have been made if the sensitive variable(s) had assumed different values.

In this paper, we focus on group-based measures of fairness and contribute to the related literature in two main ways. We first show that measures of explainability, such as Shapley values, which assign importance measures to the available variables can be usefully utilized as measures of fairness. Such explainability measures can show how much each protected variable contributes to the output of a model. In terms of fairness, Shapley values can indicate that a sensitive feature has a larger effect than it should have (Cesaro and Gagliardi Cozman 2019). Our paper can be considered as one of the first studies that evaluates feature importance measures in assessing fairness in the presence of an ordinal (and not only binary) protected variable.

The second main contribution of our paper is a constructive way to improve the fairness of an ML model, based on the findings obtained on variable importance. More specifically, we propose a matching approach, based on the variable explanations of the protected variable, which can improve model fairness. In particular, we utilize a Propensity Score Matching (PSM) method, to match the observations in the privileged and unprivileged groups, to obtain an experimental setting capable of evaluating the fairness of a model. Consequently, thanks to the matching, data is balanced in terms of the protected variable, for a better evaluation of the fairness condition. A further contribution of our paper is the practical application to a real credit lending decision-making context that concerns the predictive classification of borrowers in a peer-to-peer lending platform: Lending Club, one of the biggest digital lending platforms in the United States. In the context of LendingClub, an ML model predicts whether a loan request will be financed by the lenders available on the platform or not. The ”rejected” loan observations relate to the loan requests that had been approved by LendingClub and assigned a credit score but were not funded by lenders. To evaluate the fairness of the model, we check whether the proposed model treats equally towards applicants from different regions of the United States: an ordinal protected variable.

The remainder of the paper is organized as follows. Section 2 provides an overview of the literature on fairness in machine learning problems. Section 3 describes how fairness can be operationally measured. Section 4 describes our considered data. Section 5 illustrates our proposal and the corresponding empirical results. Finally, the study is concluded in Section 6.7.

6.3 Literature Review

The concept of fairness in machine learning has attracted much interest recently. Despite the vast number of studies on fairness, there are no universal measures to calculate it, and several notions of fairness have been considered in the literature. This section provides an overview of the measurement methods employed in the literature on fairness.

Teodorescu and Yao (2021) uses statistical fairness measures and relates fairness of ML models with computational complexity employing an imbalanced large credit scoring data. They show that when more than one protected variable is considered in a classification problem, the typical ML models do not satisfy more than one fairness criterion. This shows the difficulty of calculating fairness in decision-making processes using large databases such as credit scoring. Teodorescu and Yao (2021) discusses the problem of selecting suitable algorithms that satisfy specific ethical criteria to classify real data. In particular, they focus on the statistical fairness measures to show that typical classification algorithms are not always fair towards different protected attributes. The results found by analyzing a huge credit scoring dataset show the need for human input in fairness decisions, especially when deciding tradeoffs between fairness criteria. Shui et al. (2022) formulates a fair predictor for the available subgroups in data as a two-level objective. In the first level, considering a small part of the data, the subgroup-specific predictors are learned. Then, in the second level, the learned predictor is updated to be close to all subgroup-specific predictors. Through the empirical analysis of a real-world dataset, it is proved that their approach leads to the improvement of group sufficiency and error reduction.

In another study (Kim et al. 2023), it is mentioned that the training loss discrepancy between the groups of the protected feature is a consequence of different numbers of training samples included in each group. Hence, the classification model focuses more on learning toward the larger groups. To address this problem, they consider the target and protected attribute labels to define a target-protected group. With the aim of balancing the training losses, they propose a fairness-aware batch sampling scheme that updates batch sampling probability (BSP). Their model is validated in different classification settings and finally, it is shown that the proposed approach has the best trade-off between fairness and classification performance compared with other methods. Despite many studies on fairness in binary classification, d'Aloisio et al. (2022) refers to the bias in multi-class classification, and proposes a debiaser that not only improves fairness in multi-class classification but also in binary classification. Le Quy et al. (2022) analyze different real-world datasets used for the fairness evaluation of ML models. They find relationships between the different attributes, particularly with respect to protected attributes and the class attribute, using a Bayesian network. In (Zhang et al. 2019), it is mentioned that considering the dynamic nature of the relationship between machine learning predictions and the underlying groups corresponding to the protected variable is a vital task. Through their empirical analysis, they show that disparity worsens over time so fairness criteria should be defined based on the user dynamics. For this purpose, they develop a method of selecting a proper fairness criterion using prior knowledge of user dynamics.

The study done by Horesh et al. (2020), deals with the problem of fairness for individuals in the case that there are no only categories for which fairness should be guaranteed. In this case, there is no explicit protected variable but however, there are other correlated demographic features that lead to discrimination and bias in the model predictions. The proposed approach in this paper has the following two general aspects: despite most fairness methods which rely on the presence of at least a protected variable, this proposed approach evaluates individual-based fairness using an implicit idea of a potentially discriminatory variable, that does not have to be directly measurable. Secondly, it is applicable to a wide range of ML models. In particular, they define a paired-consistency score, which measures the similarity of a model's predictions, both for regression and classification, for paired members.

As represented by reports devoted to the governance of AI in finance, making sure that ML-based credit scoring models treat individuals (borrowers), in a fair way is a top priority for regulators. Hurlin et al. (2022) discusses the lack of fairness in credit scoring and evaluates fairness using statistical tests, and identifies the cause of unfairness. In particular, their study presents a fairness evaluation framework that assesses gender discrimination in a lending database. As a result of their analysis, direct discrimination when gender is used as a predictor in credit scoring is detected. In addition, they define two types of candidate variables, the features originating the lack of fairness: those that correlate with gender and/or default, target variable, and those that exhibit weak correlations with gender and default. Their findings indicate that neutralizing a single variable can improve fairness while maintaining overall predictive performance, surpassing alternative mitigation methods based on model re-estimation in terms of the fairness/performance trade-off. Goethals et al. (2022) distinguish between Explicit bias and Implicit bias and use counterfactual explanations to detect discrimination between the protected groups. Explicit bias occurs when a protected variable is available in the dataset, and Implicit bias is the case when the protected variable is removed but discrimination still happens because of the proxy variables correlated with the protected feature. Another study in the credit scoring literature, Kozodoi et al. (2022) mentions a lack of literature on fair ML models in credit scoring and tries to fill this gap. They use statistical fairness criteria and evaluate their suitability for credit scoring. Interestingly, different fairness processors in profit-oriented credit scoring are empirically compared using real-world data. The empirical results find that multiple fairness criteria can be met simultaneously and recommend separation as a suitable measure of scorecard fairness.

We finally remark that the requirement that an ML model is fair is related to the requirement of explainability. A recent work, Grabowicz et al. (2022) combines the two requirements to obtain models which are both fair and interpretable. Another paper, Stevens et al. (2020) proposes a fair and explainable recommendation system using data from Kiva Platform, a non-profit microfinance lending platform that lends money to entrepreneurs. Cesaro and Gagliardi Cozman (2019) apply SHAP values as feature importance to find how importance values vary with and without bias. To do this, they utilize a reweighing technique that assigns lowers weights to the privileged observations. The limitation of their model is that their proposed approach is only applicable to a binary classification with a binary sensitive variable. In another study (Hickey et al. 2021), they also utilize some equivalent measures to capture fairness but in their case, they use the external auditor model to aggregate the auditor's explanations with the measures. They also mention the missing relation between these measures and the interpretability of the models. In order to fill this gap, using statistical fairness measures they propose an approach in which external actors evaluate the output of the model. Therefore, using a model trained by the auditor, the contribution of the protected variable is calculated using SHAP concept. Then, the fairness of the model is evaluated finding the difference between the average contribution of the protected variable in the privileged and unprivileged groups. In this study, we try to improve their proposed approach using a multiple protected variable in a binary classification problem and also by testing the difference between the protected groups using suitable statistical tests.

Considering the matching methods, Karimi et al. (2022) propose the "FairMatch" algorithm which introduces a new approach to pairing similar and dissimilar individuals using the PSM method. In this paper, they focus on individual fairness and propose a novel metric to evaluate individual fairness that captures the notion of similar treatment in probabilistic classifiers in a better way. Using four real-world datasets, they show the superiority of FairMatch to the existing approaches. We also use the PSM method not only for matching similar individuals but also for improving the group-based fairness.

6.4 Proposal

To better understand how fairness can be operationally assessed, in this section, we explain the measurement of fairness and importance features using a real case study. Consider LendingClub, a P2P lending platform that uses an ML model to predict whether a loan request will be funded or rejected. The response variable in this model is a binary label, $Y = \{1, 0\}$ where the value 0 corresponds to the unfunded request: the loan requests that had been approved by LendingClub and assigned a credit score but not funded by lenders; and the value 1 refers to the funded request (issued loans).

6.4.1 Group-based Fairness Measures

Two main possible sources of unfairness can arise in our context: historical bias and proxy variables. To evaluate historical bias, we can consider prevalence: if the fraction of borrowers classified as accepted ($Y = 1$) is the same in each protected group, there is no prevalence. More formally, considering that a borrower is predicted to be accepted when its estimated probability of acceptance (PA) is higher than a certain threshold α, the non-prevalence (non-bias) condition can be expressed as follows:

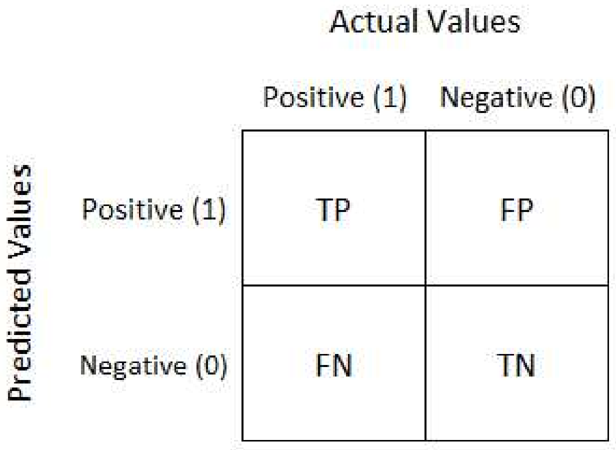
$$P(PA_i > \alpha, x_a = 0) = P(PA_i > \alpha, x_a = 1) \tag{6.1}$$

Here, PA_i is the probability that applicant i will be accepted and α is a threshold probability that can be, for example, equal to 0.5. In the above equation, we assumed that x_a, the protected variable, is binary. However, the definition can be extended, without loss of generality, when there are more than two protected groups similar to our paper. In the fairness literature, prevalence is also known as demographic or statistical parity (Chouldechova 2017), and also as independence condition (Kozodoi et al. 2022).

Another source of bias that could cause unfairness derive from the so-called proxy variables. These variables are the explanatory variables that are highly correlated with the protected feature. Finding proxies for the protected variables is important because they may lead to a lack of fairness in decision-making. There is a very limited literature on the relationship between proxy variables and fairness.

As mentioned in Section 6.2, fairness can be measured at both individual and group levels. In this paper, we focus on the group level, but we also explore how Shapley values, that explain the importance of each variable for each individual, can shed further insights.

We first check group-based fairness and check whether the considered ML model behaves similarly toward the borrowers corresponding to different protected groups, deriving from different regions. We therefore initially focus on the confusion matrix, which has already been used to assess the fairness of ML models. Figure 6.1 shows the structure of a confusion matrix. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Fig. 6.1 Confusion matrix.

In Figure 6.1, TP represents the number of positive observations that are classified as positives. On the other hand, FP shows how many positive observations are given the label zero and classified as negative observations. From Figure 6.1, we can derive ”Accuracy”, the most popular classification metric, which measures the ratio of the corrected classified observations over the total available observations:

$$ACC = \frac{TP+TN}{N} \tag{6.2}$$

Accuracy can be considered as a fairness metric. To evaluate whether fairness is achieved, one can check whether the accuracy is equal for the protected groups referring to the sensitive variable. The confusion matrix can also be employed to calculate the True Positive Rate (TPR):

$$TPR = \frac{TP}{TP+FN} \tag{6.3}$$

The TPR can be employed to assess ”Equal Opportunity”, which answers the question: ”Is TPR the same across the different groups of the protected variable?”. The question can be answered by checking whether the ratio of the correctly classified as positive cases out of all actual positive cases is the same for different groups. Another fairness assessment that can be based on the confusion matrix is the "Equalized Odds" Kozodoi et al. (2022). The equalized odds criteria not only takes the True Positive Rate (TPR) into account but it also considers the False Positive Rate (FPR). It can answer the question: "Are the TPR and FPR the same across different groups of the protected variable?". The False Positive Rate (FPR) is defined as follows:

$$FPR = \frac{FP}{FP + TN} \tag{6.4}$$

A final fairness metric to be mentioned is the Positive predictive value (PPV), which is the same in the protected groups, when fairness holds. PPV can be calculated as follows:

$$PPV = \frac{TP}{TP + FP}. \tag{6.5}$$

6.4.2 Feature Importance Measure

In this subsection we connect fairness with feature importance, leading to propensity socre matching. Feature importance is determined using SHAP values, which are based on the definition of Shapley values (Shapely 1953), which calculate the marginal contribution of each variable towards the predictions. From a game theory point of view, features of an instance in the data set behave as players in a group (coalition), and Shapley values represent the distribution of the prediction among the features according to their contribution. Following these assumptions, the SHAP algorithm calculates the contribution of each variable to each prediction considering its additional effect to all possible coalitions (groups) of the other variables. The Shapley value of variable i for $i = 1, \dots, n$, is calculated as follows:

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus i} \frac{|S|!(n-1-|S|)!}{n!}[f(S \cup i) - f(S)] \tag{6.6}$$

In the above expression, i is a feature, S is a subset of features, $|S|$ is the number of variables in the subset S, n is the number of features in the model, and $\mathcal{F}$ is the set of all possible coalitions. This is because the effect of a feature depends on the coalition it is added to and, therefore, it is necessary to consider all possible coalitions. Indeed, the difference between the output of the model with and without the feature is computed for all possible subsets without feature i. In other words, to compute the effect of each feature, the difference between the prediction of the model including that feature, $f(S \cup i)$, and the prediction of the model excluding the feature, $f(S)$, is calculated.

To make a connection between a feature importance measure such as SHAP and the fairness of a model, we can find the global impact of each feature and then using the following measure, to understand how the contribution of variables changes among the protected groups:

$$D_j = (\frac{1}{N}\sum_{n=1}^{N} abs(\phi_j^n)) - (\frac{1}{M}\sum_{m=1}^{M} abs(\phi_j^m)) \tag{6.7}$$

Here, N and M represent two different categories of the protected variable. In our case, using this equation, the distance of the global Shapley values of the variables $(j = 1, ..., theJ)$ between each pair of regions is calculated. A value of D_j close to zero shows fairness between the two corresponding groups. In the case that D_j is negative, it favors the second category (M) while a positive D_j favors the first category (N) in the equation.

6.4.3 Propensity Score Matching (PSM)

In statistical analysis, to find out the relation between a treatment and the outcome, observations are divided into a treatment group and a control group. Then, through further investigations, individuals in the same strata are compared to each other. This concept has been also used in fairness measurement. For this purpose, while the protected and unprotected groups are considered as the treatment and control groups, respectively, the protected variable (e.g., sex or race) is considered as the outcome. In this paper, we use PSM technique to perform stratification. Then it is assumed that individuals whose propensity scores are close should belong to the same strata. In our paper, to calculate propensity scores, we encode ”Region”, our protected variable, as a binary variable and then calculate propensity scores as the conditional probability of being in the protected group ($S = 1$) given the rest of the variables X. Therefore, it can be estimated by logistic regression. We explain the application of PSM in our paper in section 6.6 in detail.

6.5 Data

Our empirical experiment is based on the LendingClub dataset. LendingClub is a P2P lending platform in the United States that provides loans for individuals through an online platform (Babaei and Bamdad 2021). This data set includes a set of features describing a loan applicant and characteristics of the requested loan, for more understanding of the variables used in our experiment see Table 6.1.

The target variable y is a binary indicator of whether the applicant has been rejected (y = 0) or not (y = 1). The sensitive attribute x in this dataset is ”State” which indicates

Table 6.1 LendingClub variables. After preprocessing and joining two separate rejected and accepted loan datasets, these 7 variables remain in our experiment.

Feature	Definition
Loan Amount	The amount of money asked by the borrower.
Dti	Ratio of the borrower's total monthly debt to the borrower's self-reported monthly income.
Employment Length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
Loan Title	The loan title provided by the applicant.
Date	Date of loan application and issuance for the rejected and accepted loans, respectively. Starting from 2007 to 2020.
State	Abbreviations of the 51 United states
Status	Status of the loan, including ”Rejected” and ”Accepted”

in which state the applicant's request is located. In a fair lending system, borrowers from different places should be provided with equal credit opportunities. To simplify the sensitive variable (in terms of the number of categories) and to consider a higher level of the borrowers' location, we consider the four census regions in the US. Therefore, applicants are categorized into ”South”, ”West”, ”Midwest”, and ”Northeast” based on Figure 6.2. Other variables are the explanatory variables used to predict the target variable. ”Loan Title” is a categorical variable that includes more than 1000 unique values. ”Debt consolidation”, ”Credit card refinancing”, ”Home Improvement”. and ”Other” are the most relevant categories that are assigned to the loan requests. To generalize categories of the ”Loan Title” and because most categories appear only once, all the other titles are put in the ”Other” category. In addition, it should be mentioned that in the preprocessing of the dataset, missing values are dropped. Finally, there are 28788509 observations.

This dataset is unbalanced, the imbalance ratio (the ratio of class one over class zero) is equal to 0.07842, showing a high imbalance rate between the minority and majority observations. We continue our analysis with this imbalanced data because, for problems involving ethics, data manipulation could be unacceptable. The main goal of our paper will be evaluation of the fairness over the four regions described in 6.2. In our dataset, borrowers are mostly distributed in the ”South” region with 40.71% of the borrowers. This is followed by ”West” showing a percentage of 22.62%, and by ”Northeast”, with a percentage of 18.79%. ”Midwest” has the lowest percentage and only 17.88% of the borrowers applied for credit from the states in this region.

6.6 Results and Discussion

As described in Section 6.4.1, a possible source of unfairness arises from prevalence across different groups in the data. Here prevalence concerns ”loan_status”, and it can

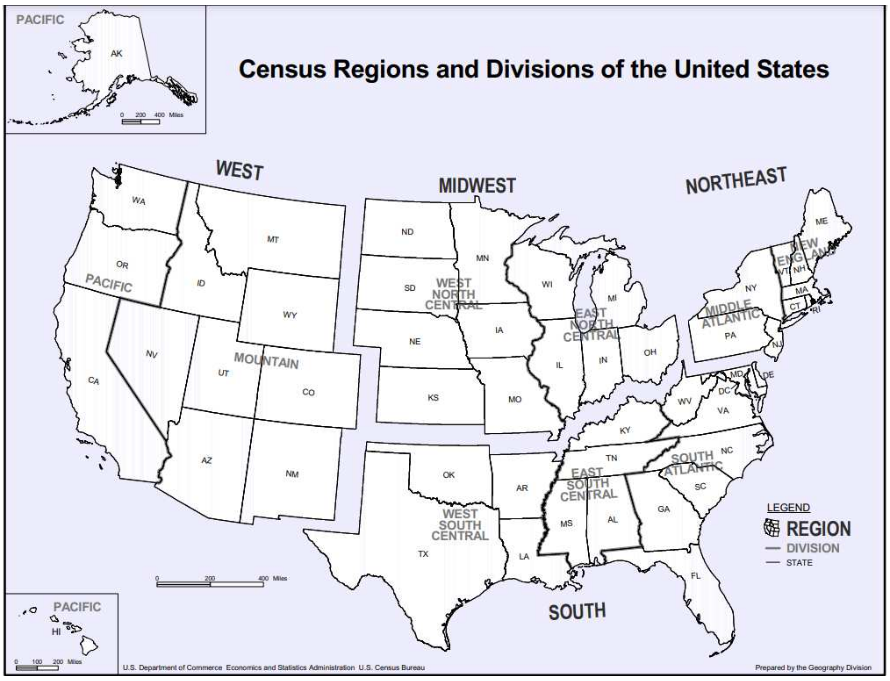


Fig. 6.2 Divisions of the United States to the census regions, namely South, West, Midwest, and Northeast. This figure shows that 51 states of the US are categorized into four regions.

be measured as the proportion of accepted loans. In our dataset, the overall proportion of acceptance is 7.27%. Figure 6.3 shows almost the same proportion in the different Census regions.

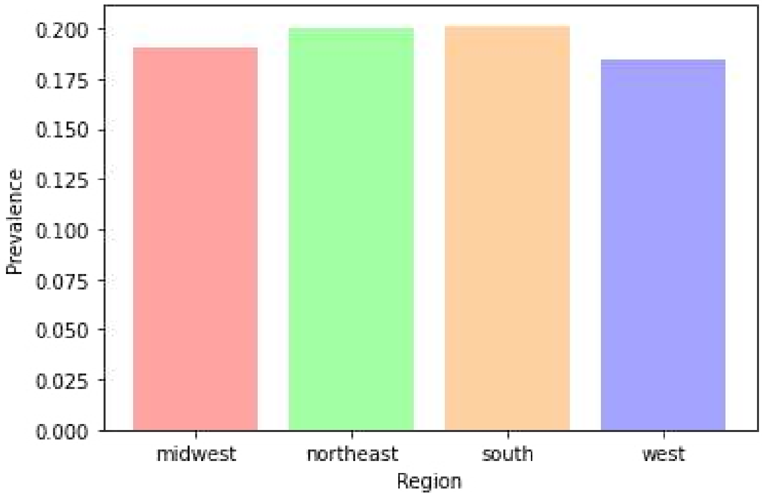


Fig. 6.3 Prevalence values: proportion of accepted loans, in each region.

Figure 6.3 shows that prevalence is almost the same in all four census regions and, therefore, it is unlikely to be a potential reason for unfairness. We, therefore, continue searching for the second potential source of bias, arising from proxy variables, highly correlated with the protected feature, in our case ”Region”.

We propose to find proxy variables by means of an explainable AI method, SHAP, applied to an ML classification model for the protected variable. More precisely, we use a XGBoost algorithm in which ”Region” is the response variable and all the other explanatory variables are the potential predictors. We then apply the SHAP algorithm Lundberg and Lee (2017) to estimate the global Shapley values of each feature. The resulting values are presented in Figure 6.4.

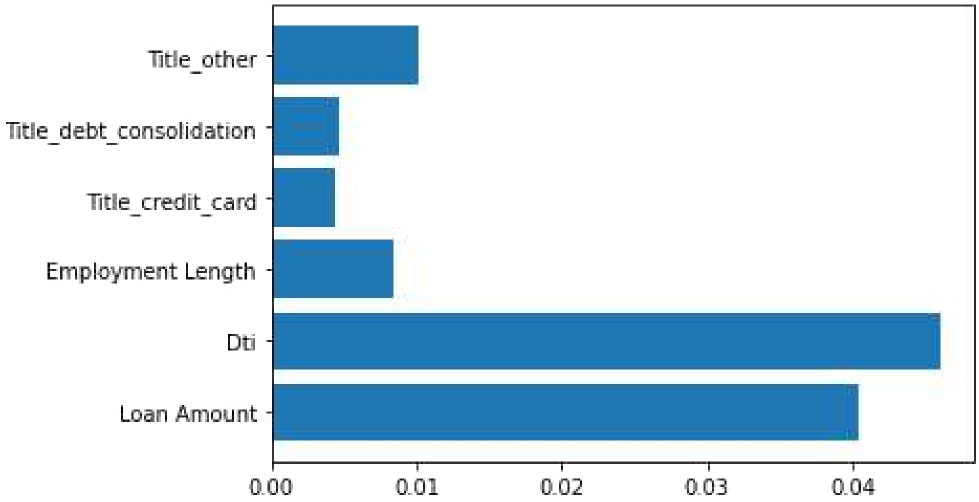


Fig. 6.4 Global Shapley values of the explanatory variables from an XGBoost classifier to predict ”Regions”.

From Figure 6.4, note the high values for ”Dti” and ”Loan Amount”, which suggest a possible strong relationship between these variables and ”Region”, a potential source of unfairness.

Fairness of a model can be also assessed by applying different definitions of fairness, as seen in the methodological section. Table 6.2 reports the group-based fairness metrics calculated on the groups of the protected variable (Region):

Table 6.2 Fairness metrics. The confusionbased fairness metrics are presented for the protected groups.

Fairness Metric	South	West	Midwest	Northeast
Accuracy	0.96305079	0.95290099	0.96276944	0.95530114
TPR	0.68506227	0.70469935	0.73947001	0.71608872
FPR	0.01798814	0.02467965	0.01980869	0.02387526
PPV	0.72204042	0.72060728	0.74441162	0.72306102

Considering the values of the accuracy in the four census regions in Table 6.2, we can say that the model is fair towards the regions because trained classifiers using data of the different regions perform almost equally. Regarding TPR, we see that values vary slightly. The smallest TPR, 0.685, is obtained on ”South” data while that is equal to 0.739 for the ”Midwest” group. Therefore, regarding the ”Equal Opportunity” fairness definition, the credit scoring model is not completely fair. Considering FPR, it can be said that all values are almost equal to 0.02. Referring to the ”Equalized odds” fairness metric explained in Section 6.4.1, our model is unfair towards borrowers of different regions because TPR and FPR are not the same across the protected groups: while FPR has almost the same values for all the regions, TPR fluctuates. Finally, PPV values vary slightly across the regions and, therefore, the model is fair according to this metric. To have a better understanding of the changes in the fairness metrics across the census regions, we propose to calculate the Gini concentration measure on the four fairness metrics, as shown in Table 6.3.

Table 6.3 Gini concentration index of Fairness measures and the p-values found by the Kolmogorov-Smirnov test. The represented P-values compare the underlying distribution of fairness measures against a given uniform distribution.

Fairness Metric	Gini Index	P-value
Accuracy	0.0024724	0.00795969
TPR	0.01534208	0.00907862
FPR	0.06989172	0.01384801
PPV	0.00622256	0.00848954

From Table 6.3, the smallest Gini index is obtained for the Accuracy, followed by the PPV, which can be considered ”Fair”; whereas for TPR and FPR the Gini index is higher, indicating a potential unfairness, as already commented.

In order to find if the model behaves equally towards the four regions, we found the global Shapley value of the four regions contributed to the prediction of the probability of a loan request being accepted, shown in Figure 6.5. From Figure 6.5 we can claim that the four regions contribute differently to the probability of acceptance, i.e. output of the classifier. For example, while the global Shapley value for ”South” is -0.005520, the contribution of ”Midwest” is lower than -0.001. Therefore, we can say that this classifier is not fair towards applicants from different regions of the US.

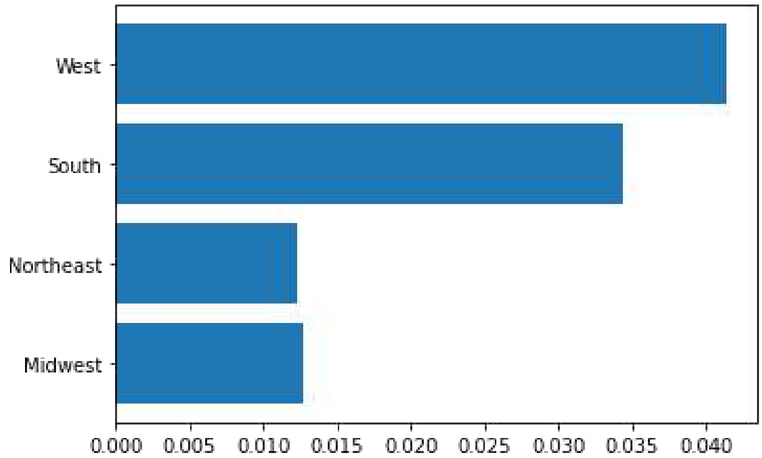


Fig. 6.5 Global Shapley values of the four census regions. These values show the contribution of each region to the prediction of our XGBoost classifier.

In addition, we can compare the global Shapley values of the explanatory variables in the four different regions represented in Table 6.4. It can be seen that ”Employment Length” has the highest contribution to the prediction of loan status compared with the remaining variables. We propose to use Shapley values also to assess the fairness of the model, indirectly capturing the effect of proxy variables. To this end, we provide in Table 6.5 the Gini index calculated over the rows of Table 6.4, along with the related p-values obtained by means of a Kolmogorov-Smirnov test.

Table 6.4 Global Shapley values of the explanatory variables in our classifier.

Feature	South	West	Midwest	Northeast
Loan Amount	0.6659745	0.6050389	0.6925815	0.6520588
Dti	0.9523098	0.9316295	1.2025243	1.276348
Employment Length	1.8416198	1.7988015	1.8481079	1.8522105
Loan Title _ credit card	0.06848664	0.06530799	0.08204002	0.06652775
Loan Title _ debt consolidation	0.01730897	0.01886303	0.01514196	0.02030704
Loan Title _ other	0.13235654	0.10679261	0.10664613	0.09622573

Table 6.5 Gini concentration index of Shapley values and the p-values for a Kolmogorov-Smirnov test. The second column shows the Gini concentration index calculated based on the Shapley values for each explanatory variable in the different regions. In addition, the P-values of a k-test which compares the underlying distribution of Shapley values against a given uniform distribution are presented in the third column.

Feature	Gini index	P-value
Loan Amount	0.02643162	0.06341286
Dti	0.07359761	0.00043702
Employment Length	0.00567770	0.00000000
Loan Title _ credit card	0.04617734	0.00009060
Loan Title _ debt consolidation	0.05951229	0.00000034
Loan Title _ other	0.06138787	0.00061378

Table 6.5 indicates for which variables we have unfairness. It can be said that for those variables which have the p-value of the test lower than 5%, there is unfairness. Therefore, the model is unfair for all variables except ”Loan Amount”.

The results from the explainable machine learning model indicate that it is important to balance the data appropriately, to remove the unfairness that can derive from unrepresented data groups. To this aim, we utilize the PSM approach explained in section 6.4.3 to propose an approach that improves fairness by balancing data matching observations from the protected group to the most similar unprotected instances. In particular, by fitting a logistic regression on the training data, we find propensity scores of the train data, then training observations located in the protected group are connected to the closest training point from the unprotected group. Based on the demographic parity concept in group-based fairness, we show in Table 6.6 how the PSM-based approach can improve fairness towards the applicants from different regions.

Table 6.6 Demographic imparity between the protected and unprotected groups for the two applied models. While the PSM-based model uses only the balanced train data (only matched observations) the Full model includes all the train observations.

Approach	Demographic imparity
PSM-based model	0.00430835
Base model	0.00785428

Considering the difference between the values of demographic parity, we can claim that balancing train data by matching the protected observations to the most similar unprotected observations decreases demographic dipsarity. In our context, words, a balanced train data leads to the detection of a small amount of fairness, quantifiable in a difference of 4 basis points.

6.7 Conclusion

We presented a framework to detect fairness of machine learning models. The approach allows to identify fairness using group based feirness measures, in line with the extant literature. It then innovates the existing methods by making a connection between fairness and explainability, which allows to understand the existence of an indirect source of fairness, coming from an uneven distribution of the explanatory variables in the data. Shapley values can understand if this is the case. We also proposed how to balance the sample, in the case of an indirect source of fairness, found by means of Shapley values. Our proposal is based on propensity score matching at the individual level.

Our proposal has been applied to a real use case that concerns credit lending, The empirical findings indicate that feature importance measures can identify group fairness of the model; that Shapley values suggest balancing of the data; and that new experimental data, based on propensity score matching, can reduce unfairness. We finally remark that our work concerns ordinal sensitive features and two-class classification problems. In future research, such restrictions could be lifted.

Another promising extension of this work would be to evaluate the visualization techniques that should be used to present the results. SHAP graphs speed up the perception of feature importance, but additional insights for fairness would be very useful.

CONCLUDING REMARKS

The motivation of this thesis is grounded on improving the technologies in the financial markets. The rapid development of financial technologies (FinTechs) has paved the way to a strand of new methods for decision-making in the FinTech market. This thesis contributes to the literature on decision-making powered by Artificial Intelligence and Machine Learning models by developing new techniques capable of improving investments in the new leading financial markets.

In Chapter 1, following the findings of (Solomatine et al. 2008, Quinlan 1993) indicating that integration of lazy and eager learning could lead to better results, I proposed the Hybrid Instance-Based Learning (HIBL) algorithm in which an instance-based approach is integrated with artificial neural networks. This algorithm allows to decide about the loans in P2P lending. In particular, similar instances are detected using the instance-based learning approach in which the probability of default (PD) is utilized as the similarity measure. Then, Return On Investment (ROI) is predicted by artificial neural networks. Considering risk and return of the investments, applying a portfolio optimization method, the optimal portfolio is found for the investor. For the evaluation of the proposed algorithm, I used the LendingClub dataset in 2017 and compared the HIBL model with a practical rating-based method. The results revealed that the proposed hybrid model can improve investment decision-making in comparison with an available practical model in the P2P lending market. In addition, the effectiveness of combining lazy with eager learning methods was shown by removing the instance-based learning from the algorithm. Chapter 2 aimed to evaluate the effects of the classification methods on the decisions in P2P lending. For this purpose, three different classification methods, Random Forest (RF), Support Vector Machines (SVM), and Naive Bayes (NB), were utilized to predict the status of loans (status is equal to one for the default loans and zero for the non-default). Then, a portfolio selection problem was optimized for those loans with non-default predicted labels. In this optimization problem, the portfolio risk was minimized for a minimum expected return. To evaluate the proposed algorithm, I again used the LendingClub dataset. Results of the numerical study demonstrated that applying credit-scoring methods in the decision-making algorithm reduced the risk of investment. Specifically, using the RF in combination with the portfolio optimization problem led to the least level of investment risk. In Chapter 3, I presented a methodology that can explain the ”automatic” choices of a robot advisor based on Markowitz's optimal asset allocation. The methodology was applied to a time series of portfolios calculated on a set of crypto assets. Particularly, Shapley values were applied to the predictions generated by a machine learning model based on the results of a dynamic Markowitz portfolio optimization model to provide explanations for what is behind the selected portfolio weights. The obtained results suggested to implement the method as a useful tool for supervisors, which assesses the compliance of robot advisories to financial regulations.

In Chapter 4, I extended the application of an explainable AI method, SHAP, utilized in Section 3. I employed Shapley values to achieve explainability and guide variable selection, leading to a parsimonious model that is a good trade-off between predictive accuracy and explainability. To achieve this goal, I proposed a model selection strategy in which global Shapley values ordered the candidate explanatory variables in terms of their predictive importance, and a backward stepwise selection procedure, based on the comparison of predictive accuracy was implemented to select a 'statistically optimal' subset of variables. This proposal was applied to a database including credit ratings for a large set of European Small and Medium Enterprises (SMEs). The obtained results indicated that the nonlinear random forest credit scoring model was more accurate than the logistic regression. The proposed method in this section provided a model comparison procedure based on both accuracy and explainability, which can be equally applied to all types of ML models. It also led to a credit scoring model which is a good trade-off between predictive accuracy and explainability. In Chapter 5, a variant of OrdinarySHAP, utilized in chapters 3 and 4, named InstanceSHAP was proposed to better explain predictions in binary classification problems, such as credit scoring. SHAP explanations vary when background data changes. Background data is the data used for the estimation of the Shapley values. I particularly showed that providing background data with a more similar distribution to the test data leads to better explanations. The comparison of the Gini concentration index for the proposed InstanceSHAP, versus the OrdinarySHAP, in a credit scoring application, showed a more concentrated distribution for my proposed method, which indicates a better discriminatory power of the explanations. In Chapter 6, I presented a framework to detect the fairness of machine learning models. The approach allows identifying fairness using group-based fairness measures. It then innovates the existing methods by making a connection between fairness and explainability, which allows us to understand the existence of an indirect source of fairness, coming from an uneven distribution of the explanatory variables in the data. I also evaluated the effect of balancing data based on the protected variable using propensity score matching at the individual level. Empirical findings indicated that feature importance measures can identify the group fairness of the model. Numerical results also showed that balancing data based on propensity score matching can reduce imparity.

Overall, in this thesis, I have analyzed many different data to evaluate and consequently improve Artificial Intelligence and Machine Learning models in the FinTech market. This has been accomplished throughout the development and usage of a variety of techniques related to different mainstream domains, such as statistics, machine learning, and data science. The lessons learned from the current thesis are manifold. Firstly, complex machine learning models lead to accurate investment decisions in financial problems such as credit scoring. It was also shown that integration of different models could improve the results. However, stakeholders of these models need to know the reasons behind the decisions made by the automated models. Therefore, as the second lesson from this thesis, it can be said that to have a responsible and reliable method, it is needed to explain accurate results found by the models using explainable artificial intelligence methods. Thirdly, machine learning models are sometimes biased against a special group of instances. Hence, it is vital to check the fairness of the models before utilizing them for different applications. The methodological tools presented in this thesis have a variety of real world applications, especially in the finance domains. Future research might extend and improve the methodological frameworks presented so far, as well as provide further interesting domains of application for the methodologies proposed.

REFERENCES

Aas, K., Jullum, M. and Løland, A. (2021), ‘Explaining individual predictions when features are dependent: More accurate approximations to Shapley values’, *Artificial Intelligence* **298**, 103502.

Abdelmoula, A. K. (2015), ‘Bank credit risk analysis with k-nearest-neighbor classifier: Case of tunisian banks’, *Accounting and Management Information Systems* **14**(1), 79.

Adadi, A. and Berrada, M. (2018), ‘Peeking Inside the Black-Box: a Survey on Explainable Artificial Intelligence (XAI)’, *IEEE access* **6**, 52138–52160.

Ahelegbey, D. F., Giudici, P. and Mojtahedi, F. (2022), ‘Crypto asset portfolio selection’, *FinTech* **1**(1), 63–71.

Akyildirim, E., Goncu, A. and Sensoy, A. (2021), ‘Prediction of cryptocurrency returns using machine learning’, *Annals of Operations Research* **297**, 3–36.

Albini, E., Long, J., Dervovic, D. and Magazzeni, D. (2022), Counterfactual Shapley Additive Explanations, *in* ‘2022 ACM Conference on Fairness, Accountability, and Transparency’, pp. 1054–1070.

Alessandretti, L., ElBahrawy, A., Aiello, L. M. and Baronchelli, A. (2018), ‘Machine learning the cryptocurrency market’, *Complexity* **2018**.

Ariza-Garzon, M.-J., Segovia-Vargas, M.-J., Arroyo, J. et al. (2021), ‘Risk-return modelling in the P2P lending market: Trends, gaps, recommendations and future directions’, *Electronic Commerce Research and Applications* **49**, 101079.

Austin, P. and Tu, J. (2004), ‘Bootstrap methods for developing predictive models’, *The American Statistician* **58(2)**, 131–137.

Ayadi, R., Bongini, P., Casu, B. and Cucinelli, D. (2021), ‘Bank Business Model Migrations in Europe: Determinants and Effects’, *British Journal of Management* **32**(4), 1007–1026.

Babaei, G. and Bamdad, S. (2020*a*), ‘A multi-objective instance-based decision support system for investment recommendation in peer-to-peer lending’, *Expert Systems with Applications* **150**, 113278.

Babaei, G. and Bamdad, S. (2020*b*), ‘A neural-network-based decision-making model in the peer-to-peer lending market’, *Intelligent Systems in Accounting, Finance and Management* **27**(3), 142–150.

Babaei, G. and Bamdad, S. (2021), ‘Application of credit-scoring methods in a decision support system of investment for peer-to-peer lending’, *International Transactions in*

Operational Research .

Bastani, K., Asgari, E. and Namavari, H. (2019), 'Wide and deep learning for peer-to-peer lending', *Expert Systems with Applications* **134**, 209–224.

Berger, A. N. and Udell, G. F. (2006), 'A more complete conceptual framework for SME finance', *Journal of Banking and Finance* **30**(11), 2945–2966.

Bracke, P., Datta, A., Jung, C. and Sen, S. (2019), 'Machine learning explainability in finance: an application to default risk analysis'.

Brauneis, A. and Mestel, R. (2019), 'Cryptocurrency-portfolios in a mean-variance framework', *Finance Research Letters* **28**, 259–264.

Breiman, L. (2001), 'Random Forests', *Machine learning* **45**, 5–32.

Bücker, M., Szepannek, G., Gosiewska, A. and Biecek, P. (2022), 'Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring', *Journal of the Operational Research Society* **73**(1), 70–90.

Buckmann, M., Joseph, A. and Robertson, H. (2022), An interpretable machine learning workflow with an application to economic forecasting, Technical report, Bank of England.

Burkart, N. and Huber, M. F. (2021), 'A survey on the explainability of supervised machine learning', *Journal of Artificial Intelligence Research* **70**, 245–317.

Bussmann, N., Giudici, P., Marinelli, D. and Papenbrock, J. (2021), 'Explainable Machine Learning in Credit Risk Management', *Computational Economics* **57**, 203–216.

Cesaro, J. and Gagliardi Cozman, F. (2019), Measuring unfairness through game-theoretic interpretability, *in* 'Joint European Conference on Machine Learning and Knowledge Discovery in Databases', Springer, pp. 253–264.

Chang, S., Kim, S. D. and Kondo, G. (2015), 'Predicting default risk of lending club loans', *Machine Learning* pp. 1–5.

Chen, X., Zhou, L. and Wan, D. (2016), 'Group social capital and lending outcomes in the financial credit market: An empirical study of online peer-to-peer lending', *Electronic Commerce Research and Applications* **15**, 1–13.

Cheng, W. and Hüllermeier, E. (2009), 'Combining instance-based learning and logistic regression for multilabel classification', *Machine Learning* **76**, 211–225.

Chevallier, J., Zhu, B. and Zhang, L. (2021), 'Forecasting inflection points: Hybrid methods with multiscale machine learning algorithms', *Computational Economics* **57**, 537–575.

Cho, P., Chang, W. and Song, J. W. (2019), 'Application of instance-based entropy fuzzy support vector machine in peer-to-peer lending investment decision', *IEEE Access* **7**, 16925–16939.

Chokor, A. and Alfieri, E. (2021), 'Long and short-term impacts of regulation in the cryptocurrency market', *The Quarterly Review of Economics and Finance* **81**, 157–173.

Chopra, A. and Bhilare, P. (2018), 'Application of Ensemble Models in Credit Scoring

Models', *Business Perspectives and Research* **6**(2), 129–141.

Chouldechova, A. (2017), 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big data* **5**(2), 153–163.

Correia, F., Martins, A. and Waikel, A. (2022), 'Online financing without FinTech: Evidence from online informal loans', *Journal of Economics and Business* **121**, 106080.

Covert, I. and Lee, S.-I. (2021), Improving KernelSHAP: Practical Shapley value estimation using linear regression, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 3457–3465.

Dahooie, J. H., Hajiagha, S. H. R., Farazmehr, S., Zavadskas, E. K. and Antucheviciene, J. (2021), 'A novel dynamic credit risk evaluation method using data envelopment analysis with common weights and combination of multi-attribute decision-making methods', *Computers and Operations Research* **129**, 105223.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988), 'Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach', *Biometrics* **44**(3), 837–845.

Derbentsev, V., Matviychuk, A. and Soloviev, V. N. (2020), 'Forecasting of cryptocurrency prices using machine learning', *Advanced studies of financial technologies and cryptocurrency markets* pp. 211–231.

Diebold, F. X. and Mariano, R. S. (2002), 'Comparing Predictive Accuracy', *Journal of Business and Economic Statistics* **20**(1), 134–144.

Djeundje, VB, C. J. C. R. and Hamid, M. (2021), 'Enhancing credit scoring with alternative data', *Expert Systems with applications* **163**.

Dushimimana, B., Wambui, Y., Lubega, T. and McSharry, P. E. (2020), 'Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans', *Journal of Risk and Financial Management* **13**(8), 180.

d'Aloisio, G., Stilo, G., Di Marco, A. and D'Angelo, A. (2022), Enhancing fairness in classification tasks with multiple variables: A data-and model-agnostic approach, *in* 'Advances in Bias and Fairness in Information Retrieval: Third International Workshop, BIAS 2022, Stavanger, Norway, April 10, 2022, Revised Selected Papers', Springer, pp. 117–129.

Emekter, R., Tu, Y., Jirasakuldech, B. and Lu, M. (2015), 'Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending', *Applied Economics* **47**(1), 54–70.

Fasano, F. and Cappa, F. (2022), 'How do banking fintech services affect SME debt?', *Journal of Economics and Business* **121**, 106070.

Ferri, G. and Murro, P. (2015), 'Do firm–bank 'odd couples' exacerbate credit rationing?', *Journal of Financial Intermediation* **24**(2), 231–251.

Finlay, S. (2011), 'Multiple classifier architectures and their application to credit risk assessment', *European Journal of Operational Research* **210**(2), 368–378.

Galindo, J. and Tamayo, P. (2000), ‘Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications’, *Computational economics* **15**, 107–143.

Gao, Y., Yu, S.-H. and Shiue, Y.-C. (2018), ‘The performance of the p2p finance industry in china’, *Electronic Commerce Research and Applications* **30**, 138–148.

Giudici, P., Hadji-Misheva, B. and Spelta, A. (2020), ‘Network based credit risk models’, *Quality Engineering* **32**(2), 199–211.

Giudici, P. and Polinesi, G. (2021), ‘Crypto price discovery through correlation networks’, *Annals of Operations Research* **299**, 443–457.

Giudici, P. and Raffinetti, E. (2020), ‘Lorenz model selection’, *Journal of Classification* **37**(3), 754–768.

Giudici, P. and Raffinetti, E. (2021), ‘Shapley lorenz explainable artificial intelligence’, *Expert systems with applications* **114104**(167).

Giudici, P. and Raffinetti, E. (2022), ‘Explainable AI methods in cyber risk management’, *Quality and Reliability Engineering International* **38**(3), 1318–1326.

Goethals, S., Martens, D. and Calders, T. (2022), ‘Explainability methods to detect and measure discrimination in machine learning models’.

Grabowicz, P. A., Perello, N. and Mishra, A. (2022), Marrying fairness and explainability in supervised learning, *in* ‘2022 ACM Conference on Fairness, Accountability, and Transparency’, pp. 1905–1916.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. and Pedreschi, D. (2018), ‘A survey of methods for explaining black box models’, *ACM computing surveys (CSUR)* **51**(5), 1–42.

Guo, Y., Zhou, W., Luo, C., Liu, C. and Xiong, H. (2016), ‘Instance-based credit risk assessment for investment decisions in P2P lending’, *European Journal of Operational Research* **249**(2), 417–426.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017), ‘Learning from class-imbalanced data: Review of methods and applications’, *Expert systems with applications* **73**, 220–239.

Hand, D. J. (2009), ‘Measuring classifier performance: a coherent alternative to the area under the ROC curve’, *Machine learning* **77**(1), 103–123.

Hand, D., Mannila, H. and Smyth, P. (2001), *Principles of data mining*, MIT Press.

Henley, W. and Hand, D. J. (1996), ‘A k-nearest-neighbour classifier for assessing consumer credit risk’, *Journal of the Royal Statistical Society: Series D (The Statistician)* **45**(1), 77–95.

Hickey, J. M., Di Stefano, P. G. and Vasileiou, V. (2021), Fairness by explicability and adversarial shap learning, *in* ‘Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III’, Springer, pp. 174–190.

Horesh, Y., Haas, N., Mishraky, E., Resheff, Y. S. and Meir Lador, S. (2020), Paired-consistency: An example-based model-agnostic approach to fairness regularization in machine learning, *in* 'Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I', Springer, pp. 590–604.

Hüllermeier, E. (2003), 'Possibilistic instance-based learning', *Artificial Intelligence* **148**(1-2), 335–383.

Hurlin, C., Pérignon, C. and Saurin, S. (2022), 'The fairness of credit scoring models', *arXiv preprint arXiv:2205.10200* .

Janzing, D., Minorics, L. and Blöbaum, P. (2020), Feature relevance quantification in explainable AI: A causal problem, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 2907–2916.

Jiang, Z. and Liang, J. (2017), Cryptocurrency portfolio management with deep reinforcement learning, *in* '2017 Intelligent systems conference (IntelliSys)', IEEE, pp. 905–913.

Kalayci, C. B., Ertenlice, O., Akyer, H. and Aygoren, H. (2017), 'An artificial bee colony algorithm with feasibility enforcement and infeasibility toleration procedures for cardinality constrained portfolio optimization', *Expert Systems with Applications* **85**, 61–75.

Karimi, H., Khan, M. F. A., Liu, H., Derr, T. and Liu, H. (2022), Enhancing individual fairness through propensity score matching, *in* '2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA)', IEEE, pp. 1–10.

Kendall, J. (2017), 'Fintech Companies Could Give Billions of People More Banking Options', *Harvard Business Review* **1**.

Kim, D., Park, S., Hwang, S. and Byun, H. (2023), 'Fair classification by loss balancing via fairness-aware batch sampling', *Neurocomputing* **518**, 231–241.

Kozodoi, N., Jacob, J. and Lessmann, S. (2022), 'Fairness in credit scoring: Assessment, implementation and profit implications', *European Journal of Operational Research* **297**(3), 1083–1094.

Kumari, B., Kaur, J. and Swami, S. (2021), System Dynamics Approach for Adoption of Artificial Intelligence in Finance, *in* 'Advances in Systems Engineering: Select Proceedings of NSC 2019', Springer, pp. 555–575.

Kurosaki, T. and Kim, Y. S. (2022), 'Cryptocurrency portfolio optimization with multivariate normal tempered stable processes and foster-hart risk', *Finance Research Letters* **45**, 102143.

Kwon, Y. and Zou, J. (2022), 'Weightedshap: analyzing and improving shapley based feature attributions', *arXiv preprint arXiv:2209.13429* .

Laborda, J. and Ryoo, S. (2021), 'Feature Selection in a Credit Scoring Model', *Mathematics* **9**(7), 746.

Law, W. K., Yaremych, H. E., Ferrer, R. A., Richardson, E., Wu, Y. P. and Turbitt, E. (2022), 'Decision-making about genetic health information among family dyads: a systematic literature review', *Health Psychology Review* **16**(3), 412–429.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W. and Ntoutsi, E. (2022), 'A survey on datasets for fairness-aware machine learning', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **12**(3), e1452.

Lee, E. and Lee, B. (2012), 'Herding behavior in online p2p lending: An empirical investigation', *Electronic commerce research and applications* **11**(5), 495–503.

Lee, T.-S. and Chen, I.-F. (2005), 'A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines', *Expert Systems with applications* **28**(4), 743–752.

Leow, E. K. W., Nguyen, B. P. and Chua, M. C. H. (2021), 'Robo-advisor using genetic algorithm and bert sentiments from tweets for hybrid portfolio optimisation', *Expert Systems with Applications* **179**, 115060.

Lessmann, S., Baesens, B., Seow, H.-V. and Thomas, L. C. (2015), 'Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research', *European Journal of Operational Research* **247**(1), 124–136.

Li, G., Shi, Y. and Zhang, Z. (2019), P2p default risk prediction based on xgboost, svm and rf fusion model, *in* '1st International Conference on Business, Economics, Management Science (BEMS 2019)', Atlantis Press, pp. 470–475.

Li, J., Kuang, K., Li, L., Chen, L., Zhang, S., Shao, J. and Xiao, J. (2021), Instance-wise or class-wise? a tale of neighbor Shapley for concept-based explanation, *in* 'Proceedings of the 29th ACM International Conference on Multimedia', pp. 3664–3672.

Li, Y. and Chen, W. (2020), 'A Comparative Performance Assessment of Ensemble Learning for Credit Scoring', *Mathematics* **8**(10), 1756.

Liu, W., Fan, H. and Xia, M. (2022), 'Credit scoring based on tree-enhanced gradient boosting decision trees', *Expert Systems with Applications* **189**, 116034.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. and Lee, S.-I. (2019), 'Explainable AI for trees: From local explanations to global understanding', *arXiv preprint arXiv:1905.04610* .

Lundberg, S. M. and Lee, S.-I. (2017), 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems* **30**.

Ma, B.-j., Zhou, Z.-l. and Hu, F.-y. (2017), 'Pricing mechanisms in the online peer-to-peer lending market', *Electronic Commerce Research and Applications* **26**, 119–130.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018), 'Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning', *Electronic Commerce Research and Applications* **31**, 24–39.

Malekipirbazari, M. and Aksakalli, V. (2015), 'Risk assessment in social lending via random forests', *Expert Systems with Applications* **42**(10), 4621–4631.

Markowitz, H. M. (1994), 'The general mean-variance portfolio selection problem', *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* **347**(1684), 543–549.

Merrick, L. and Taly, A. (2020), The explanation game: Explaining machine learning models using Shapley values, *in* 'Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4', Springer, pp. 17–38.

Milne, A. and Parboteeah, P. (2016), 'The Business Models and Economics of Peer-to-Peer Lending', *ECRI Research Report* .

Mitchell, S., Potash, E., Barocas, S., D'Amour, A. and Lum, K. (2021), 'Algorithmic fairness: Choices, assumptions, and definitions', *Annual Review of Statistics and Its Application* **8**, 141–163.

Mittal, M., Goyal, L. M., Sethi, J. K. and Hemanth, D. J. (2019), 'Monitoring the impact of economic crisis on crime in india using machine learning', *Computational Economics* **53**, 1467–1485.

Molnar, C., König, G., Bischl, B. and Casalicchio, G. (2023), 'Model-agnostic feature importance and effects with dependent features: a conditional subgroup approach', *Data Mining and Knowledge Discovery* pp. 1–39.

Moscato, V., Picariello, A. and Sperlí, G. (2021), 'A benchmark of machine learning approaches for credit score prediction', *Expert Systems with Applications* **165**, 113986.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. and Yu, B. (2019), 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences* **116**(44), 22071–22080.

Murtaugh, P. A. (1998), 'Methods of variable selection in regression modeling,', *Communications in Statistics, Simulation and Computation* **27**, 711–734.

Namvar, A., Siami, M., Rabhi, F. and Naderpour, M. (2018), 'Credit risk prediction in an imbalanced social lending environment', *arXiv preprint arXiv:1805.00801* .

Paliwal, M. and Kumar, U. A. (2011), 'Assessing the contribution of variables in feed forward neural network', *Applied Soft Computing* **11**(4), 3690–3696.

Pattekari, S. A. and Parveen, A. (2012), 'Prediction system for heart disease using naïve bayes', *International journal of advanced computer and mathematical sciences* **3**(3), 290–294.

Pessach, D. and Shmueli, E. (2022), 'A review on fairness in machine learning', *ACM Computing Surveys (CSUR)* **55**(3), 1–44.

Phoon, K. and Koh, F. (2017), 'Robo-advisors and wealth management', *The Journal of Alternative Investments* **20**(3), 79–94.

Quinlan, J. R. (1993), Combining instance-based and model-based learning, *in* 'Proceedings of the tenth international conference on machine learning', pp. 236–243.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016*a*), 'Model-agnostic interpretability of machine learning', *arXiv preprint arXiv:1606.05386* .

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016*b*), "Why Should I Trust You?" Explaining the Predictions of Any Classifier, *in* 'Proceedings of the 22nd ACM SIGKDD international Conference on Knowledge Discovery and Data Mining', pp. 1135–1144.

Romānova, I. and Kudinska, M. (2016), Banking and Fintech: A Challenge or Opportunity?, *in* 'Contemporary issues in finance: Current challenges from across Europe', Vol. 98, Emerald Group Publishing Limited, pp. 21–35.

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E. and Li, Y. (2020), 'An explainable ai decision-support-system to automate loan underwriting', *Expert Systems with Applications* **144**, 113100.

Sawik, B. (2012), 'Bi-criteria portfolio optimization models with percentile and symmetric risk measures by mathematical programming', *Przeglad Elektrotechniczny* **88**(10B), 176–180.

Serrano-Cinca, C. and Gutiérrez-Nieto, B. (2016), 'The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending', *Decision Support Systems* **89**, 113–122.

Shapely, L. (1953), A value for n-person games, *in* 'Contributions to the Theory of Games', Vol. II, Princeton University Press, pp. 307–317.

Shen, F., Wang, R. and Shen, Y. (2020), 'A cost-sensitive logistic regression credit scoring model based on multi-objective optimization approach', *Technological and Economic Development of Economy* **26**(2), 405–429.

Shen, F., Zhao, X. and Kou, G. (2020), 'Three-stage reject inference learning framework for credit scoring using unsupervised transfer learning and three-way decision theory', *Decision Support Systems* **137**, 113366.

Shui, C., Xu, G., Chen, Q., Li, J., Ling, C. X., Arbel, T., Wang, B. and Gagné, C. (2022), 'On learning fairness and accuracy on multiple subgroups', *Advances in Neural Information Processing Systems* **35**, 34121–34135.

Solomatine, D. P., Maskey, M. and Shrestha, D. L. (2008), 'Instance-based learning compared to other data-driven methods in hydrological forecasting', *Hydrological Processes: An International Journal* **22**(2), 275–287.

Srinivasan, V. and Kim, Y. H. (1987), 'Credit Granting: A Comparative Analysis of Classification Procedures', *The Journal of Finance* **42**(3), 665–681.

Stevens, A., Deruyck, P., Van Veldhoven, Z. and Vanthienen, J. (2020), Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva, *in* '2020 IEEE Symposium Series on Computational Intelligence (SSCI)', IEEE, pp. 1241–1248.

Sundararajan, M. and Najmi, A. (2020), The many Shapley values for model explanation, *in* 'International conference on Machine Learning', PMLR, pp. 9269–9278.

Sundararajan, M., Taly, A. and Yan, Q. (2017), Axiomatic Attribution for Deep Networks, *in* 'International Conference on Machine Learning', PMLR, pp. 3319–3328.

Tan, Y., Zheng, X., Zhu, M., Wang, C., Zhu, Z. and Yu, L. (2017), Investment recommendation with total capital value maximization in online p2p lending, *in* '2017 IEEE 14th international conference on E-Business engineering (ICEBE)', IEEE, pp. 159–165.

Temelkov, Z. (2018), 'Fintech firms opportunity or threat for banks?', *International journal of Information, Business and Management* **10**(1), 137–143.

Teodorescu, M. H. and Yao, X. (2021), Machine learning fairness is computationally difficult and algorithmically unsatisfactorily solved, *in* '2021 IEEE High Performance Extreme Computing Conference (HPEC)', IEEE, pp. 1–8.

Teply, P. and Polena, M. (2020), 'Best classification algorithms in peer-to-peer lending', *The North American Journal of Economics and Finance* **51**, 100904.

Townsend, J., Chaton, T. and Monteiro, J. M. (2019), 'Extracting relational explanations from deep neural networks: A survey from a neural-symbolic perspective', *IEEE transactions on neural networks and learning systems* **31**(9), 3456–3470.

Tripathi, D., Shukla, A. K., Reddy, B. R., Bopche, G. S. and Chandramohan, D. (2022), 'Credit Scoring Models Using Ensemble Learning and Classification Approaches: A Comprehensive Survey', *Wireless Personal Communications* pp. 1–28.

Trivedi, S. K. (2020), 'A study on credit scoring modeling with different feature selection and machine learning approaches', *Technology in Society* **63**, 101413.

Tsai, K., Ramiah, S. and Singh, S. (2014), 'Peer lending risk predictor', *CS229 Autumn* .

Tyagi, S. (2022), 'Analyzing machine learning models for credit scoring with explainable ai and optimizing investment decisions', *American International Journal of Business Management* **5**(1), 1–161.

Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B. (2019), 'Price movement prediction of cryptocurrencies using sentiment analysis and machine learning', *Entropy* **21**(6), 589.

Vedala, R. and Kumar, B. R. (2012), An application of naive bayes classification for credit scoring in e-lending platform, *in* '2012 International Conference on Data Science and Engineering (ICDSE)', IEEE, pp. 81–84.

Veganzones, D. and Séverin, E. (2018), 'An investigation of bankruptcy prediction in imbalanced datasets', *Decision Support Systems* **112**, 111–124.

Walambe, R., Kolhatkar, A., Ojha, M., Kademani, A., Pandya, M., Kathote, S. and Kotecha, K. (2021), Integration of explainable AI and blockchain for secure storage of human readable justifications for credit risk assessment, *in* 'Advanced Computing: 10th International Conference, IACC 2020, Panaji, Goa, India, December 5–6, 2020, Revised Selected Papers, Part II 10', Springer, pp. 55–72.

Wang, C., Zhang, W., Zhao, X. and Wang, J. (2019), 'Soft information in online peer-to-peer lending: Evidence from a leading platform in china', *Electronic Commerce Research and Applications* **36**, 100873.

Wang, Z., Jiang, C., Ding, Y., Lyu, X. and Liu, Y. (2018), 'A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending', *Electronic Commerce Research and Applications* **27**, 74–82.

West, D. (2000), 'Neural network credit scoring models', *Computers and operations research* **27**(11-12), 1131–1152.

Xia, Y., Liu, C. and Liu, N. (2017), 'Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending', *Electronic Commerce Research and Applications* **24**, 30–49.

Xia, Y., Yinguo, L., Lingyun, H., Yixin, H. and Yigun, M. (2021), 'Incorporating multilevel macroeconomic variables into credit scoring for online consumer lending.', *Electronic commerce research and applications)* (10195).

Yang, J. and Ma, J. (2019), 'Feed-forward neural network training using sparse representation', *Expert Systems with Applications* **116**, 255–264.

Ye, X., Dong, L.-a. and Ma, D. (2018), 'Loan evaluation in p2p lending based on random forest optimized by genetic algorithm with profit score', *Electronic Commerce Research and Applications* **32**, 23–36.

Yobas, M. B., Crook, J. N. and Ross, P. (2000), 'Credit scoring using neural and evolutionary techniques', *IMA Journal of Management Mathematics* **11**(2), 111–125.

Yuan, H., Liu, M., Krauthammer, M., Kang, L., Miao, C. and Wu, Y. (2022), 'An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models', *arXiv preprint arXiv:2204.11351* .

Yum, H., Lee, B. and Chae, M. (2012), 'From the wisdom of crowds to my own judgment in microfinance through online peer-to-peer lending platforms', *Electronic Commerce Research and Applications* **11**(5), 469–483.

Zhang, H., Zhao, H., Liu, Q., Xu, T., Chen, E. and Huang, X. (2018), 'Finding potential lenders in p2p lending: A hybrid random walk approach', *Information Sciences* **432**, 376–391.

Zhang, K. and Chen, X. (2017), 'Herding in a p2p lending market: Rational inference or irrational trust?', *Electronic Commerce Research and Applications* **23**, 45–53.

Zhang, X., Khaliligarekani, M., Tekin, C. et al. (2019), 'Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness', *Advances in neural information processing systems* **32**.

Zhao, H., Liu, Q., Wang, G., Ge, Y. and Chen, E. (2016), Portfolio selections in p2p lending: A multi-objective perspective, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining', pp. 2075–2084.

LIST OF PUBLICATIONS

1. Babaei, G., Giudici, P. (2023), How fair is machine learning in credit lending? (Under Review)
2. Babaei, G., Giudici, P. (2023), InstanceSHAP: An Instance-Based Estimation Approach for Shapley Values. Behaviormetrika, 1-15.
   (https://doi.org/10.1007/s41237-023-00208-z)
3. Babaei, G., Giudici, P. Raffinetti, E (2023), Explainable fintech lending. Journal of Economics and Business, 106126.
   (https://doi.org/10.1016/j.jeconbus.2023.106126)
4. Babaei, G., Giudici, P. Raffinetti, E (2022), Explainable Artificial Intelligence for Crypto Asset Allocation, Finance Research Letters, 47, 102941.
   (https://doi.org/10.1016/j.frl.2022.102941)
5. Babaei, G., Bamdad, Sh (2023), Application of credit scoring methods in a decision support system of investment for peer-to-peer lending, International Transactions in Operational Research, 30(5), 2359-2373.
   (https://doi.org/10.1111/itor.13064)
6. Babaei, G., Bamdad, Sh (2021), A New Hybrid Instance-Based Learning Model for Decision-Making in the P2P Lending Market. Computational Economics 57, 419–432.
   (https://doi.org/10.1007/s10614-020-10085-3)