Explainable AI in Fintech and Insurtech

Thesis by

Alex Gramegna

In Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

XXXV cycle

Electronic, Computer Science and Electrical Engineering

University of Pavia

Author: Alex Gramegna

Supervisor: Paolo Giudici

ABSTRACT

Explainable AI in Fintech and Insurtech

Alex Gramegna

The growing application of black-box Artificial Intelligence algorithms in many real-world application is raising the importance of understanding how models make their decision. The research field that aims to look into the inner workings of the black-box and to make predictions more interpretable is referred to as eXplainable Artificial Intelligence (XAI). Over the recent years, the research domain of XAI has seen important contributions and continuous developments, achieving great results with theoretically sound applied methodologies. These achievements enable both industry and regulators to improve on existing models and their supervision; this is done in term of explainability, which is the main purpose of these models, but it also brings new possibilities, namely the employment of eXplainable AI models and their outputs as an intermediate step to new applications, greatly expanding their usefulness beyond explainability of model decisions.

This thesis is composed of six chapters: an introduction and a conclusion plus four self contained sections reporting the corresponding papers. Chapter 1 proposes the use of Shapley values in similarity networks and clustering models in order to bring out new pieces of information, useful for classification and analysis of the customer base, in an insurtech setting. In chapter 2 a comparison between SHAP and LIME, two of the most important XAI models, evaluating their parameters attribution methodologies and the information they are capable of include thereof, in italian Small and Medium Enterprises' Probability of Default (PD) estimation, with balance sheet data as inputs. Chapter 3 introduces the use of Shapley values in feature selection techniques, with the analysis of wrapper

and embedded feature selection algorithms and their ability to select relevant features with both raw data and their Shapley values, again in the setting of SME PD estimation. In chapter 4, a new methodology of model selection based on Lorenz Zoonoid is introduced, highlighting similarities with the game-theoretical concept of Shapley values and their variability decomposition attribution to independent variables as well as some advantages in terms of model comparability and standardization. These properties are explored through both a simulated example and the application to a real world dataset, provided by EU-certified rating agency Modefinance.

Contents

LIST OF FIGURES

LIST OF TABLES

‘‘‘

To my family and the people who are with me along the journey of life, without whom I would be nothing.

Chapter 1

General Introduction

Artificial Intelligence, in its broad definition and meaning, is becoming an integral part of many real-world application. The main reasons for its spread are the exponential increase in availability and amount of data to be processed, the improvements in computing resources (e.g. GPUs, TPUs, cloud computing, etc) and the development of more complex algorithms. Nowadays, application of Artificial Intelligence affects numerous decision making processes, ranging from finance, medicine, robotics, agriculture, security and many more. In this ever-evolving environment and with the expansion of applications to new fields, the understanding of how these so called "black-box" models make their decision becomes crucial role. The research field that aims to "open" the black-box and to make the predictions more interpretable is referred to as eXplainable Artificial Intelligence (XAI). From a legal point of view, the introduction of regulation such as the European General Data Protection Regulation (GDPR) and the American Algorithmic Accountability Act, raised the concern of having a set of mandatory tools to make the models as transparent as possible to the customers, clearly stating any possible drawback and excluding any possibility of bias. From an ethical point of view, applications such as medical screening or security raised the problem of understanding the drivers of models' predictions so to avoid any kind of discrimination and possible social inequalities. Finally, aside from legal and ethic issues, a better knowledge of how models make their decision clearly has the added value for any users to leverage the information and to increase the performances. Explainability techniques can be classified according to several criteria:

- Intrinsic or post-hoc: distinction whether interpretability is achieved by restricting the complexity of the model (intrinsic) or by applying methods that analyze the model after training (post hoc). Example of intrinsic models are machine learning algorithms that are considered interpretable due to their simple structure, such as linear regression family (OLS, regularized, GLM) or decision trees. Post hoc techniques examples are Permutation Features Importance or Shapley values.

- Feature summary statistic: methods which provide summary statistics for each variable. Some methods return a single number per feature, such as feature importance, or a more complex result, such as the pairwise feature interaction strengths, which consist of a number for each feature pair.

- Feature summary visualization: feature summary statistics can be visualized. Some feature summaries are actually only meaningful if they are visualized and a table would be a wrong choice, such as for the partial dependence plot which are curves that show a feature and the average predicted outcome.

- Data point: this category includes all methods that return data points (already existing or newly created) to make a model interpretable. For example, counterfactual explanations explains the prediction of a data instance finding a similar data point by changing some of the features for which the predicted outcome changes in a relevant way.

- Approximation: black-box models can be approximated (globally or locally) with a more interpretable model. For example LIME locally approximates data points by fitting a regularized linear model such as LASSO.

- Model-specific or model-agnostic: model-specific interpretation techniques are limited to specific model classes. For example, the interpretation of regression coefficients in a linear model. Model-agnostic methods can be

used with regards to any model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs only. Shapley values are model- agnostic tools.

- Local or global:this class entails the concept whether the interpretation method explain an individual prediction or the entire model behaviour. For a more complete overview please refer to (Islam et al., 2021, Molnar, 2019).

The importance of having solid and well-researched XAI models will become paramount in the years to come, as the development and the spreading of Machine Learning and AI application looks increasingly clear. Other than this very desirable point, it is worth noting that eXplainable AI models themselves and their outputs can sometimes gauge different informations with respect to the ones picked up by regular models, and there is room to exploit this feature to improve the understanding of a specific problem or the modeling phase itself, much in the way GANs are used. This opens up many possibilities and spur us on developing better methodologies to solve old problems, improving in understandability, robustness and performance.

Chapter 2

Why to buy insurance? An explainable artificial intelligence approach

2.1 Introduction

The performance of the insurance sector is undergoing a transformation. While life insurance products are performing well in term of market penetration, non-life products are lagging behind. This may be detrimental to society, as the aim of the insurance industry is, in its essence, a protective one, serving as an hedge against the risk of contingent or uncertain losses, generating efficiency.

The gap of the non-life insurance sector may be the manifestation of the inability of traditional insurance companies to successfully complete the so-called "last mile": the effective communication to the final users of the importance of covering risks, either because they are not using the right tools or simply because they can not offer the protection the customers need. To close the gap, customers need to be understood, and effective communication is needed.

Technology based insurance (Insurtech), dependent on the application of Artificial Intelligence methods to data retrieved from users' engagement electronic devices, can close the gap between non-life insurance providers and customers, thereby improving the protection and the resilience of our societies. The advantage of using AI applications are, in a nutshell, the capability for insurance companies to better understand customer needs, listening to their preferences, as expressed by interaction generated data; and the possibility for insurance subscribers to receive an insurance coverage that well fit their needs.

The application of Artificial Intelligence to insurance is relatively recent.

Bernardino[1] provides an up-to-date review of the application of AI to the insurance sector, and of the related opportunities. Being the insurance sector highly regulated, Artificial intelligence applications, to be trustworthy, must be accurate and explainable: see, for example *European Commission (2020)* [2].

We propose to apply to the non-life insurance industry an accurate and explainable machine learning algorithm, based on Shapley values (see [3] and [4]), which helps us turn "black box" unexplainable algorithms into something closer to a white box. The application of Shapley values can shift perspective and gain insights into customers' needs and behavior, building relevant profiles and going more towards prescriptive analytics.

We show the advantages of our proposal within two case-studies, the first aimed at estimating the probability of buying, the second the probability of churning, a specific non-life insurance product. We then show the utility of the proposed model to highlight customers who are at risk of churn. In both cases we are able to estimate the amount of opportunity/risk both at the individual and at the overall level, while analysing the factors that are responsible for it.

2.2 Methodology

2.2.1 Building a predictive classifier

The first step of our proposal is to select a highly accurate predictive model. The research literature shows that ensemble methods, consisting in the combination of several different learners to obtain low variation and low bias predictors are particularly suited for this kind of problems (see e.g. *Breiman (2000)* [5]). Ensembles made up of classification trees, which natively capture interactions and non linearities, are particularly suited for predictive classification problems. Among the family of ensemble tree learners, we employ Extreme Gradient Boosting. This algorithm consistently scores better against its peers, and implements a gradient boosting algorithm which penalises trees with a proportional shrinking of the leaf nodes (*Chen and Guestrin, (2016)*[6]).

However, algorithms like the Extreme Gradient Boosting (XGBoost), which aggregate a series of learner into one output, are hardly interpretable, particularly by customers and regulators: the most it can be gained in terms of interpretability are scores about variables' importance, often extrapolated from aggregated calculations. That is why these algorithms are usually classified as "black boxes". This limitation counterbalances some of the advantages of being a better classifier. To overcome the issue of interpretability, we propose the use of explainable AI models for the output of Extreme Gradient Boosting, in the next subsection.

2.2.2 Explaining model predictions

In line with the request that AI applications must be trustworthy, researchers have recently proposed explainable machine learning models (for a review see e.g. *Guidotti (2018)* [7] and *Molnar (2019)* [8]).

Among explainable models, the Shapley value approach, proposed in *Shapley (1952)* [9] and operationalised by *Lundberg (2017)* [10] and *Strumbelj (2010)*[11], has many attractive properties. In particular, in the Shapley framework, the

variability of the predictions is divided among the available covariates. In this way, the contribution of each explanatory variable to each point prediction can be assessed regardless of the underlying model (*Joseph (2019)*[4]), in a model agnostic manner.

From a computational perspective, the SHAP framework (short for SHapley Additive exPlanation) returns Shapley values expressing model predictions as linear combinations of binary variables that describe whether each covariate is present in the model or not. With a specific implementation developed for tree-based algorithms, it is possible to overcome some limitations encountered with kernel-based SHAP estimation, due to long computing time (*Lundberg, (2018)*[12]).

More formally, the SHAP algorithm approximates each prediction $f(x)$ with $g(x')$, a linear function of the binary variables $z' \in \{0,1\}^M$ and of the quantities $\phi_i \in \mathbb{R}$, defined as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{2.1}$$

where M is the number of explanatory variables.

Lundberg, (2018) [12] has shown that the only additive method that satisfies the properties of *local accuracy*, *missingness* and *consistency* is obtained attributing to each variable x'_i an effect ϕ_i (the Shapley value), defined by:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{2.2}$$

where f is the model, x are the available variables, and x' are the selected variables. The quantity $f_x(z') - f_x(z' \setminus i)$ expresses, for each single prediction, the deviation of Shapley values from their mean: the contribution of the i-th variable.

Intuitively, Shapley values are an explanatory model that locally approximate the original model, for a given variable value x (*local accuracy*); with the property that, whenever a variable is equal to zero, so is the Shapley value (*missingness*);

and that if in a different model the contribution of a variable is higher, so will be the corresponding Shapley value (*consistency*).

2.2.3 Clustering the explained predictions

On top of being able to interpret and compare any model with the same framework, the Shapley values can be employed for further elaborations, fostering a new range of possibilities and perspectives to understand and communicate the characteristics of customers and their interaction with insurance products.

From a statistical viewpoint, this means we can search for patterns and regularities by putting in relation feature vectors with similar Shapley values, for example explaining similarity between customers in their determinants, with respect to the target variable. To this end, we employ similarity networks, to understand similarity between customers based on the standardized Euclidean distance between each pair $(\mathbf{x}_i, \mathbf{x}_j)$ of predictors. More formally, we define the pairwise distance $d_{i,j}$ as:

$$d_{i,j} = (\mathbf{x}_i - \mathbf{x}_j)\mathbf{\Delta}^{-1}(\mathbf{x}_i - \mathbf{x}_j)' \tag{2.3}$$

where $\mathbf{\Delta}$ is a diagonal matrix whose i-th diagonal element contains the standard deviation. The distances can be represented by a $N \times N$ dissimilarity matrix $\mathbf{D}$ such that the closer two customers i, j are in the Euclidean space, the lower the entry $d_{i,j}$. The matrix D may be highly dimensional, and consequently difficult to deal with. To simplify its structure, we employ K-means clustering, defined by *MacQueen (1967)* [13], to find whether consumers can be merged into groups that represent common behavioral characteristics.

2.3 Application

2.3.1 Data

The data with which we test our proposal is provided by the insurtech company Neosurance, based in Italy, and concern the purchasing of instant and micro-policies in the sports and travel domain. We will investigate two different user behaviours: the propensity to buy and customer's churn. Even though the data is the same, the actual dimensionality of the dataset is different as the propensity to buy includes users who became customers as well as users that have not purchased anything yet, while the definition of churn requires the existence of a purchasing history. We therefore have **3778 users** to estimate the propensity to buy, and **1689 users** to estimate customer churn. As explanatory variables we have some demographic information (mostly gender, age, approximative location, device used) and information regarding purchasing history and behavior, use of the application, user experience.

The target variable is a binary variable: the "buy" event in the propensity to buy case and the "leave" event in the churn case. The proportion of positive class for the propensity study is 27.5%, while for the churn study is 53.3%.

2.3.2 Results

The propensity study dataset is split in a 80% training and a 20% testing set. After adequate optimization of the hyperparameters, the XGBoost model on the training set is tested to obtain the relevant curves and metrics. In Figure 2.1 below we compare the performance of the XGBoost method with a benchmark logistic regression, obtained from a classic stepwise model selection.
Figure 1 shows the better predictive performance of the XGBoost method over the logistic regression. Indeed, the Area Under the Curve is 0.7715 for the logistic regression models and 0.9018 for the XGBoost model.

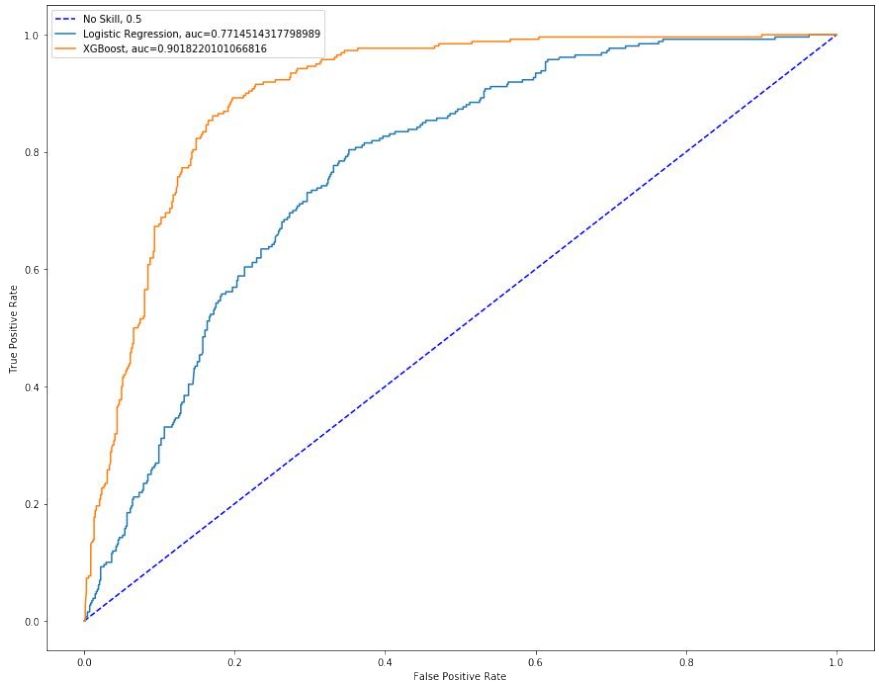


Figure 2.1: ROC curves comparison

We now interpret the output of the XGBoost method by means of the **SHAP values** approach, for each explanatory feature available. This can be done with the TreeSHAP implementation, whose computational complexity reduces from $O(T*L*2^M)$ to $O(T*L*D^2)$, where T is the number of trees, L is the maximum number of leaves in a tree and D the maximal depth of a tree. Figure 2 below contains the SHAP summary plot from TreeSHAP, which shows the contribution of each variable by representing its Shapley value averaged across all customers. In the figure, all observations are plotted row wise, separately for each explanatory variable. In each row, the color indicates the magnitude of each observation in terms of that variable: from low (colour blue), to high (colour red).

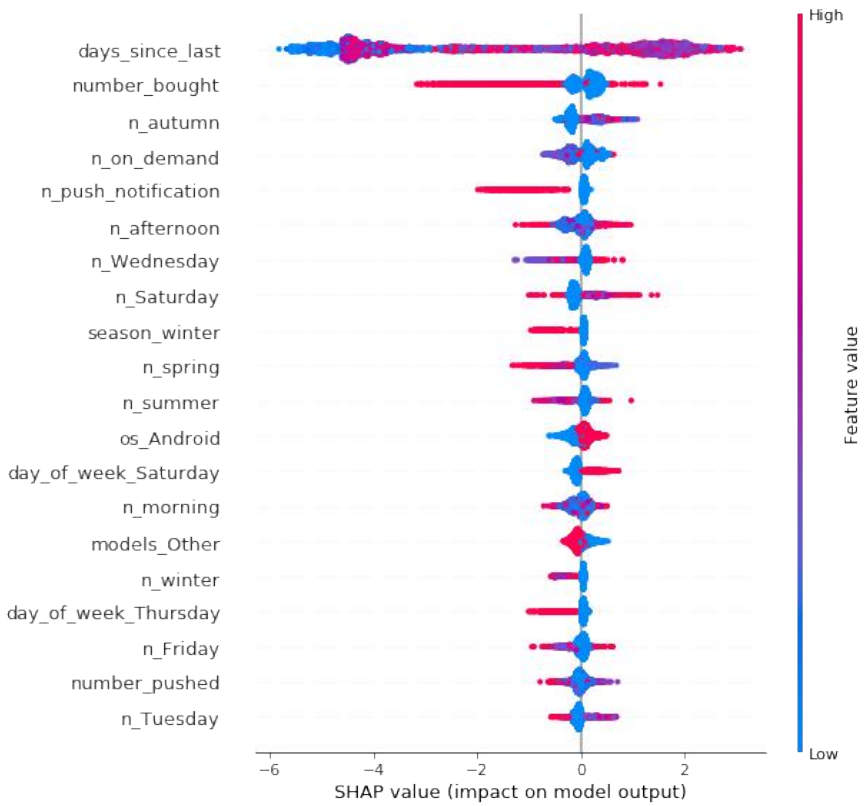


Figure 2.2: SHAP summary plot

From Figure 2.2 note that the most important variable to predict propensity to buy is the number of days since the last buy, followed by the number of bought items. In both cases, the impact on model output varies considerably among all observations (days since last) and especially for those with large values (number of bought items. Note also the effects of seasonality, in terms of weekdays and seasons.

The third part of the analysis involves using the shap values vectors corresponding to each user, calculated from the classification model, and look for the presence of structures which cluster together similar potential buyers. To this aim we employ a K-means clustering algorithm [14]. By plotting the within sum of squares against the number of clusters, we obtain that the optimal number of clusters is four.

We can thus plot, in Figure 3, the scatterplot of the first two principal components of the SHAP values, assigning each customers to one of the four clusters. In

the Figure, the four cluster means are indicated with bolder nodes, and positive events (customers that buy) are coloured in red.

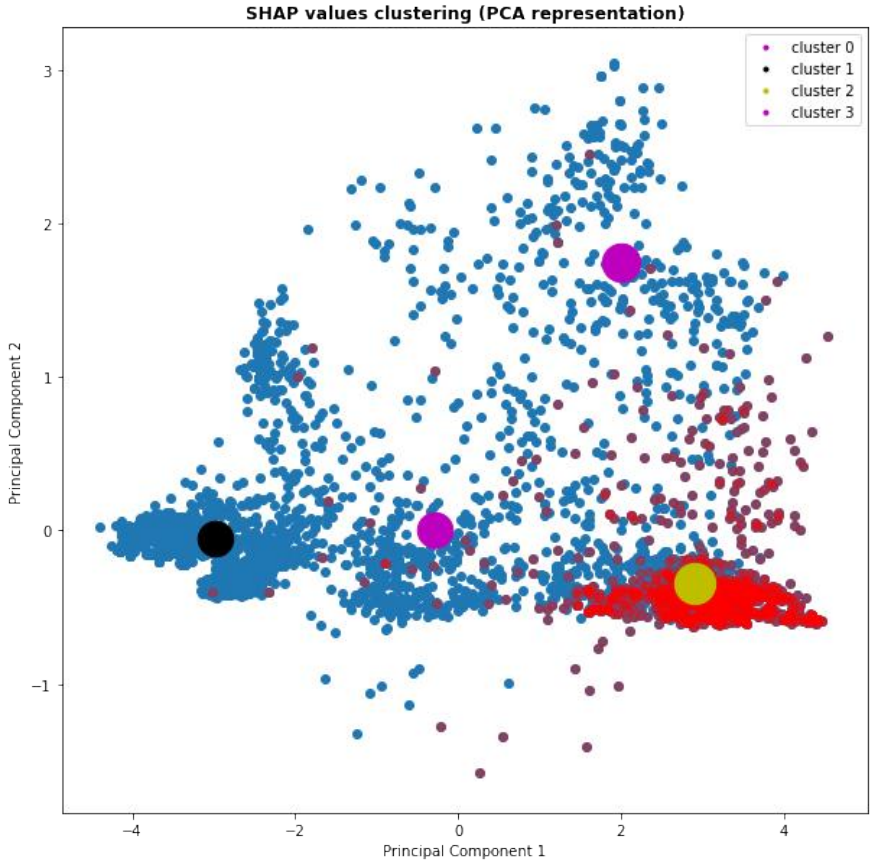


Figure 2.3: Clustering of Shapley values

From Figure 2.3 it can be noticed that one cluster is positioned in an area with virtually no red points (the black centroid), the two purple centroids are somewhat in-between and the cluster denoted by the yellow centroid is in an area where many users have high propensity to buy. Checking the proportion of positive classes with respect to each cluster, it turns out that the black cluster scores a 0.002 proportion (among the 1518 units contained in the cluster), the two purple clusters 0.09 and 0.093 (with 314 and 546 units in the clusters, respectively), while the yellow one shows a much larger 0.701, with 1400 units in the cluster.

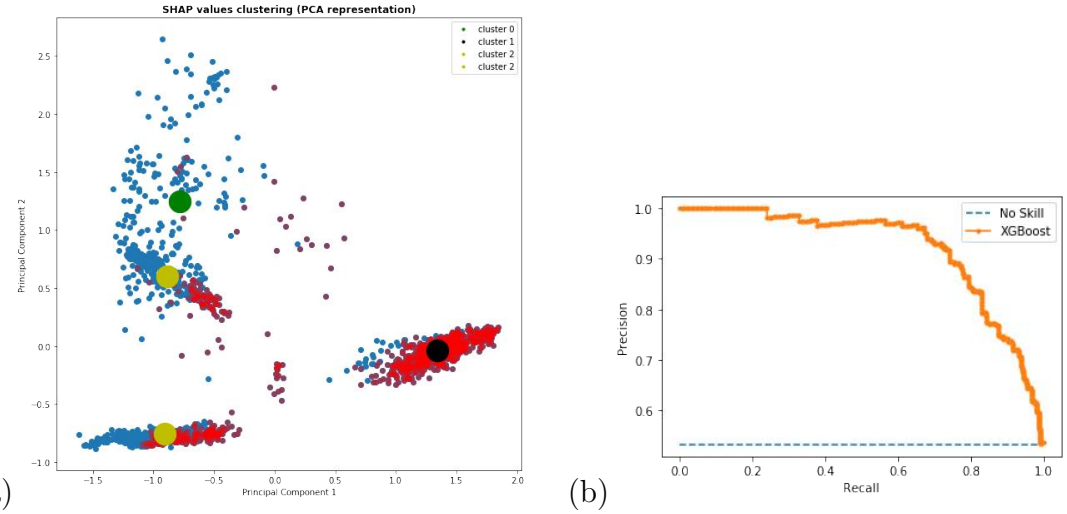
It seems reasonable to group the two intermediate clusters into a new one, leaving us with three final clusters. Thhis way, we operate an effective segmenta-

tion among users, with probability of buying ranging from 0.02% to 9% to 70%. The three clusters can be labeled, respectively, "unlikely", "less likely" and "very likely".

The obtained results are roughly consistent with what could be obtained applying the K-means algorithm directly on the data, before XGboost and SHAP. In this case, the three probabilities, for the same clusters of individuals, are: 6%, 34% and 70%. This reveals, as expected, the improved discriminatory capacity of the SHAP-XGBoost model over a pure empirical model, which does not filter any noise.

In addition, it can be shown that the three clusters that are obtained from the application of our proposal are well balanced, as we have 1495 users in the "unlikely" cluster, 866 users in the "less likely" cluster and 1417 in the "very likely" one. Conversely, if we apply the K-mean clustering to the raw data, we obtain a cluster of 951 units, with a 0.0641 proportion of events; two similar clusters with cumulatively 2807 units and a proportion of events of about 0.34 and a cluster with only 20 units and a 0.7 proportion: a rather unbalanced result. This shows further the advanage of our proposal, not only in terms of predictive accuracy and interpretability, but also in terms of profiling.

In a similar fashion, we can apply our proposal for the customer churn problem. The AUROC value is equal to 0.91 against 0.75 for the selected stepwise logistic regression model. The application of the K-means clustering to the SHAP values leads to clusters that better separated then in the buying behaviour case, as Figure 2.4 below shows.



(a) (b)

Figure 2.4: (**a**) Clustering of Shapley values with K = 4; (**b**) Propensity to buy: Precision/Recall curve (Area Under the ROC Curve = 0.91)

Figure 2.4 shows a clear separation in four clusters which can be again reduced to three, combining clusters 1 and 2. This leads to 222 users in the "unlikely" cluster, 803 in the "less likely" and 664 in the "very likely" one. We summarize the three clusters, reporting the proportion of y and mean probability of churn for each cluster in Table 2.1.

Table 2.1: Mean by cluster

cluster	**mean y**	**mean churn probability**
unlikely	0.117117	0.104915
less likely	0.313823	0.317958
very likely	0.936747	0.933060

We finally remark that also for this case we have compared the K-means results with SHAP vaues and the K-means results with the raw data and, again, the obtained clusters are better differentiated and balanced in the former case, confirming the advantage of using our proposed method.

2.4 Conclusions

To improve the understanding of consumers' needs with respect to non-life insurance products, we have proposed a novel methodology that can be embedded within a technological insurance service (Insurtech). The methodology, based on the combination of a highly accurate predictive method (XGBoost) with a model agnostic interpretability tool (Shapley Values), leads to a powerful segmentation of customers' profiles, both in terms of purchasing and churning behaviours.

Our approach brings several advantages and, in particular, the ability to perform behavioral segmentation based on the behavioural similarity existing between customers.The research suggests that explainable machine learning models can effectively improve our understanding of customers' behaviour. To further investigate this claim, future research may involve the application of the model to other situations arising in the insurance industry, which may gain from the application of artificial intelligence technologies, such as underwriting and claims management.

Our approach can also be extended to other financial technology applications, such as peer to peer lending (*Bussmann et al. (2020)*[15]) and financial pricing (*Giudici and Raffinetti (2020)*[16]).

Another line of research would be to extend our approach considering the Mean Absolute Shapley Values instead of the SHAP values, as in *Lundberg et al. (2020)*[17].

Chapter 3

SHAP & LIME: an evaluation of discriminative power in credit risk

3.1 Introduction

Probability of default (PD) estimation is an issue which banks and other financial institutions have been confronting with since the dawn of credit. Systems and methodologies evolved as knowledge and technology did, but it wasn't until recently that the incredible steps forward made by IT gave a real shake to the way it was performed through the industry. At first, incumbents institutions resisted the application of new paradigms, which favored the emergence of a growing number of Fintech startups which purpose is to provide an estimation of the creditworthiness of people and firms alike, and make it so that this estimation is as precise as possible.

To be able to give such estimation, these firms of course leverage new and diverse sources of data, take advantage of innovations in regulatory framework concerning financial data (e.g. European PSD2 [18]) and exploit the far higher predictive power that some of the newly implemented algorithms offer with respect to traditional methods. The increase in prediction power of new algorithms, though, takes a toll on explainability, since the models are now so complex that it is close to impossible to establish clear links between the inner workings of the model and the given output. This surely represents a problem and hinders their diffusion, other than raising a series of ethical and regulamentary problems, which are starting to be addressed (see, for example *European Commission (2020)* [2].

To solve this trade-off, the concept of eXplainable AI (XAI) emerged, introducing a suite of machine learning (ML) techniques that produce models that offer an acceptable trade-off between explainability as well as predictive utility and enables humans to understand, trust and manage the emerging generations of AI models. Among the emerging techniques, two frameworks have been widely recognized as the state-of-the-art in eXplainable AI and those are:

- the LIME framework, introduced by Ribeiro et al. in 2016 ([19])
- SHAP framework, introduced by Lundberg et al. in 2017 ([10]).

In finance, interpretability is especially important because the reliance of the model on the correct features must be guaranteed; yet, there aren't many studies focusing on the application of XAI in this specific context. *Bussman et al. (2020)* [20] propose a XAI model based on Shapley values applied in the context of loan decisions regarding SME seeking for financing through P2P platforms, whereas the research by *Ariza-Garzòn et al. (2020)* [21] aims to assess the predictive capacity of several ML models in the context of P2P lending platforms' credit scoring, then applying the Shapley algorithm to provide explainability to the prediction. The most interesting precedent is perhaps the research of *Misheva et al. (2021)* [22], where the authors explore the utility of both SHAP and LIME frameworks in the context of credit risk management, outlining the practical hurdles in applying these techniques to several different kinds of ML algorithms, as well as proposing solutions to the challenges faced.

Our study aims to compare the SHAP and LIME frameworks by evaluating their ability to define distinct groups of observations, employing the weights assigned to features by their local interpretability algorithm as input space for an unsupervised approach and a supervised one. We do this building our approach on one of the best performing, and more complex, supervised learning algorithm, XGBoost, [6] employed to predict the probability of default of italian Small and Medium Enterprises.

3.2 Methodology

3.2.1 LIME

Locally Interpretable Model Agnostic Explanations is a post-hoc model-agnostic explanation technique which aims to approximate any black box machine learning model with a local, interpretable model to explain each individual prediction[19]. The authors suggest the model can be used for explaining any classifier, irrespective of the algorithm used for predictions, as LIME is independent from the original classifier. Ultimately, LIME works locally which means that it's observation specific and, just like SHAP, it will provide explanations for the prediction relative to each observation. What LIME does is trying to fit a local model using sample data points that are similar to the observation being explained. The local model can be selected from the class of interpretable models such as linear models, decision trees, etc. The explanations provided by LIME for each observation x is obtained as follows:

$$\Phi(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{3.1}$$

where G is the class of potentially interpretable models such as linear models and decision trees,

$g \in G$: An explanation considered as a model

$f : \mathbb{R}^d \rightarrow \mathbb{R}$.

$\pi_x(z)$: Proximity measure of an instance z from x

$\Omega(g)$: A measure of complexity of the explanation $g \in G$

The goal is to minimize the locality aware loss L without making any assumptions about f, since a key property of LIME is that it is model agnostic. L is the measure of how unfaithful g is in approximating f in the locality defined by $\pi(x)$.

3.2.2 SHAP

The SHAP framework, proposed by *Lundberg (2017)* [10] adapting a concept coming from game theory *Shapley (1952)* [9], has many attractive properties. In this framework, the variability of the predictions is divided among the available covariates; this way, the contribution of each explanatory variable to each point prediction can be assessed regardless of the underlying model (*Joseph (2019)*[4]). From a computational perspective, SHAP (short for SHapley Additive exPlanation) returns Shapley values expressing model predictions as linear combinations of binary variables that describe whether each covariate is present in the model or not.

More formally, the SHAP algorithm approximates each prediction $f(x)$ with $g(x')$, a linear function of the binary variables $z' \in \{0,1\}^M$ and of the quantities $\phi_i \in \mathbb{R}$, defined as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.2}$$

where M is the number of explanatory variables.
Lundberg, (2018) [12] has shown that the only additive method that satisfies the properties of *local accuracy*, *missingness* and *consistency* is obtained attributing to each variable x'_i an effect ϕ_i (the Shapley value), defined by:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!} \left[f_x(z') - f_x(z' \setminus i)\right] \tag{3.3}$$

where f is the model, x are the available variables, and x' are the selected variables. The quantity $f_x(z') - f_x(z' \setminus i)$ expresses, for each single prediction, the deviation of Shapley values from their mean: the contribution of the i-th variable.

Intuitively, Shapley values are an explanatory model that locally approximate the original model, for a given variable value x (*local accuracy*); with the property that, whenever a variable is equal to zero, so is the Shapley value (*missingness*); and that if in a different model the contribution of a variable is higher, so will be

the corresponding Shapley value (*consistency*).

3.2.3 Evaluation approaches

While LIME and SHAP have similar behaviour in that they both obtain parameters for feature contribution to the prediction at the level of the observation (local explanation), they do differ in the algorithm which leads to such outcome. In order to see which approach is better in detecting variables' contribution at the local level, we attempt an unsupervised approach and verify if it is possible to cluster observation employing a dissimilarity matrix built on LIME weights and SHAP values, employing standardized Euclidean distance as the basis for clustering.

More formally, we define the pairwise distance $d_{i,j}$ as:

$$d_{i,j} = (\mathbf{x}_i - \mathbf{x}_j)\mathbf{\Delta}^{-1}(\mathbf{x}_i - \mathbf{x}_j)' \tag{3.4}$$

where $\mathbf{\Delta}$ is a diagonal matrix whose i-th diagonal element contains the standard deviation. The distances can be represented by a $N \times N$ dissimilarity matrix $\mathbf{D}$ such that the closer two observations i, j are in the Euclidean space, the lower the entry $d_{i,j}$.

On the similarity matrix we perform a classical K-means clustering (as defined by *MacQueen (1967)* [13]) and, to represent a different clustering approach and not limit ourselves to the convex clusters originated by K-means, we also run a spectral clustering algorithm, as outlined in *Ng et al. (2001)* [23]. This is done for dissimilarity matrices computed on both LIME weights and SHAP values. We then look for the best number of clusters K using measures that assess clusters' internal cohesion and external separation, namely the Silhouette [24] and the Davies–Bouldin index [25]. Other than using unsupervised tool to devise groups out of XAI models parameters, we run a supervised learning algorithm (Random Forest, as in *Breiman (2001)* [5]) on XAI parameters to see how they perform as input in predicting default, which was the problem we started the analysis

with. We compare the two predictive models, one for LIME weights and one for SHAP values, through AUROC (*Bradley, 1997*) [26]. This way, we have a thorough perspective on the discriminative power of eXplainable AI-assigned feature weights.

3.3 Application

3.3.1 Data

Data on italian SME is obtained through the Bureau van Dijk database, which sources data directly from Italian chamber of commerce. We employed some techniques to deal with the strongly unbalanced classes (e.g. *Lin et al. (2017)* approach) [27] and to remove time-specific factors. More specifically, we worked on data encompassing the last 6 years, comprising more than 2 millions SME observations, we kept all the defaulted cases and, for the not-defaulted ones, we randomly sampled a group of observation to maintain as they were (about 10000 for each year), while with the remaining we built 5000 clusters per year and employed the medoids as input observations. This brought down class imbalance from about $100:1$ to $5:1$, allowing the model to better frame risk patterns and give more amplitude to probability estimation.

The above procedure led us to a dataset with about 139000 observations, with 27200 defaults. We split the dataset assigning 70% of observation to the trainig set and 30% to the test set using stratified partitioning, run the chosen supervised algorithm (*XGBoost*), then apply LIME and SHAP to the test set to get the respective parameters; these are extracted for both methods as linear combinations of variables contributions', therefore are similar in magnitude and behaviour and thus comparable through our methodology.

3.3.2 Results

To select the number of clusters K we examine the silhouette plot [24] of both generated dataset, for K from 2 to 9. Either for SHAP or LIME, the number of

clusters which maximizes the silhouette score is two, coherently with the problem at hand (default prediction); we can see this by looking at the silhouette scores represented by the vertical red dashed line, which is higher for the plot with two clusters, and also from the part of the clusters who enter the X axis negative score, which increase as we increase the number of clusters. We show in figure 3.1 the silhouette graph for LIME data clustering in the graph below, with the one for SHAP being basically identical, albeit with a higher average silhouette score, as we show in the following lines.

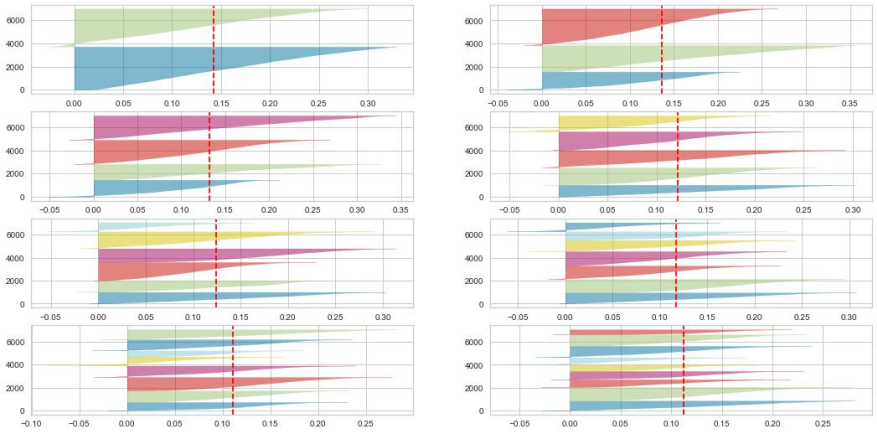


Figure 3.1: Silhouette plot for LIME data clustering

We then perform K-means clustering and Spectral clustering on the two sets of values, with the aim of evaluating the goodness of fit of the clustering approach on XAI parameters through Silhouette score and Davies–Bouldin index (DBI). Here, the higher the Silhouette score, the better externally separated and internally cohese are the clusters, while the reverse is true for Davies-Bouldin index.

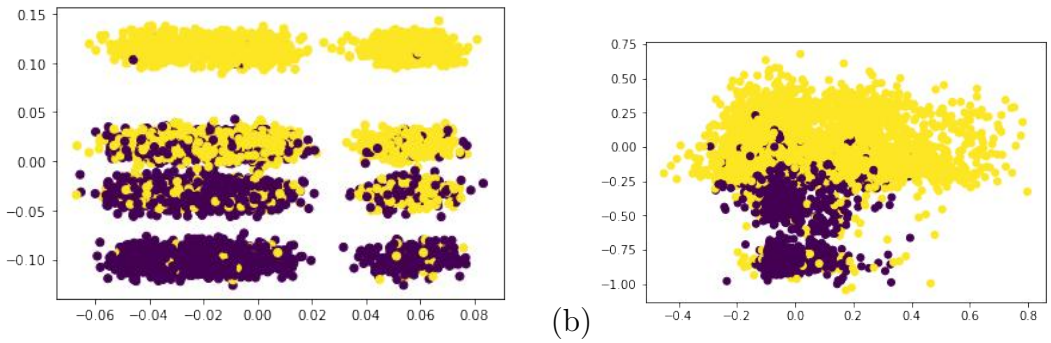
In the table below, we can see the results of both tests on each of the clustering techniques, for LIME weights and SHAP values respectively. Both techniques assign a score to represent internal clusters cohesion and external distance from one another: the silhouette scores tells us the clusters are better defined as its value increases, whereas for the Davies-Bouldin index dispersion is lower (and therefore clusters are better) the lower is the score.

Table 3.1: Clustering evaluation results

Method	LIME	SHAP
K-means Silhouette	0.143	0.370
Spectral clustering Silhouette	0.141	0.370
K-means DBI	2.325	1.126
Spectral Clustering DBI	2.329	1.106

As it turns out, SHAP values seem to constitute an input space more suitable to be divided into clusters, with a clear advantage in discriminative power in this unsupervised setting. The measures we employed for this evaluation take into consideration the entire numerosity of dimensions, which in this case is 46 since we have one parameter for each of the original feature, whereas with a scatterplot we can only evaluate two dimensions at a time.

For reference, we report one bidimensional plots for each case, where we can see how spectral clustering assigned each data point to the respective clusters by looking at the different colors; here, of course, we can only see this division across two dimensions, but we can already notice how SHAP value clustering seem to better divide the two clusters in space.



(a) (b)

Figure 3.2: (**a**) LIME Spectral Clustering ; (**b**) SHAP Spectral Clustering

Having established the superiority of SHAP values in the unsupervised environment, we can now test the predictive power of both families of parameters. To this end, we run several Random Forest algorithms [5] with optimized hyperparameters and compare the means of the Area under the Curve (AUC) [26]. We

employ Random Forests to evaluate parameters' preditive power because it has less hyperparameters to optimize with respect to other ensemble algorithms, it better handles multicollinearity and it's better parallelizable, thus allowing us to increase the number of runs significantly. Furthermore, we don't need to select a specific supervised learning algorithm to evaluate our problem, we just need it to be the same for both the sets of XAI parameters.

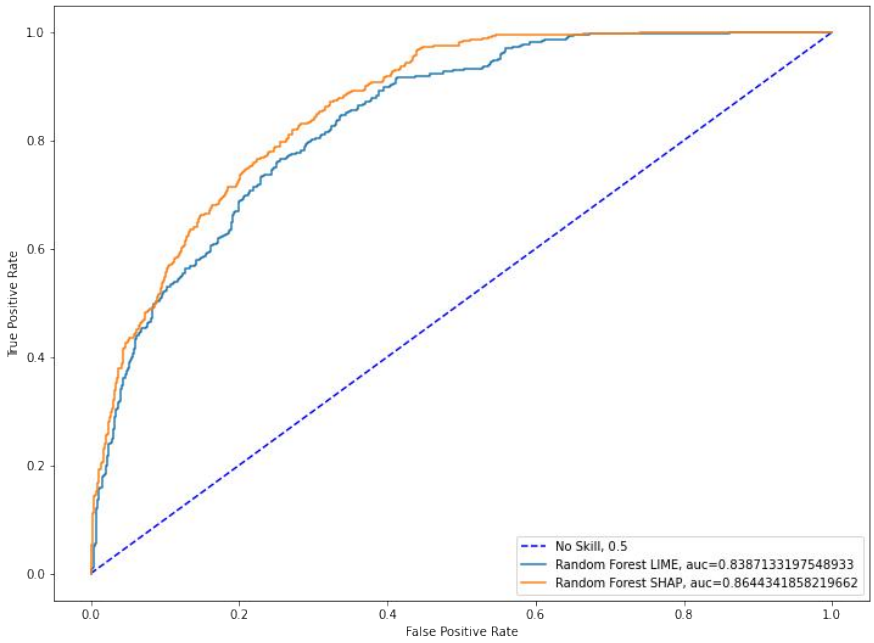


Figure 3.3: LIME and SHAP ROC curves

With a mean AUC of 0.864 for SHAP versus one of 0.839 for LIME and 50 repetitions, we find that the difference in means is statistically significant with a p-value of 0.0035.

Therefore, SHAP values appear to be better than LIME weights in assignign explanatory values to the dynamics of credit default as they are picked up by the XGBoost algorithm, through which we looked for discriminative power, that is the purpose of this paper.

3.4 Conclusions

The estimation of Probability of Default is a key element in the economic life of modern societies, and we now have the instruments and technologies to improve

it significantly and lead away from the simplistic assumptions we used to follow in order to avoid undetected risks. This concretizes in an improved adherence to reality, were we have more dimensions available regarding the entity we want to evaluate and at the same time we are more capable and correct in such evaluation. We have already seen in the aforementioned works that the methodology based on a highly accurate predictive model combined with an interpretability tool allows us to reap the benefit of this improved precision without sacrificing explainability; our approach shows that some XAI models may be better than others and, furthermore, that elements coming from eXplainable AI models can be used to further improve methodologies and add value to data.

Some other works are already moving in this direction: see for instance *Bussman et al. (2021)* [20] or *Gramegna and Giudici (2020)* [28] on the use of Shapley values to enrich the analysis and improve methods, but also *Giudici et al. (2020)* [29] and *Giudici and Raffinetti (2020)* [16], with some innovative methodologies that combine well with XAI models.

Further research could find new ways to leverage the power of explanatory parameters and use them to deal with other issues concerning the Machine Learning pipeline, as well as extend the approach to other domains.

Chapter 4

Shapley Feature Selection

4.1 Introduction

Feature selection is an area of research of great importance in machine learning. At the end of the last century, when a special issue on relevance including several papers on variable and feature selection was published [30], very few domains used more than 40 features in their models ([31]). The situation has changed drastically over the years, due to the increased capability to collect more data and to process multidimensional data. The problem with these developments is that, with so many dimensions, we also introduce many irrelevant or redundant features and often we have comparably few training examples. This hinders the ability of the model to generalize predictions [32] and, also, it increases its complexity, therefore its cost. Furthermore, there are many potential advantages in performing an effective feature selection: easier data visualization and explanation, lower requirements for measuring and storing data, lower training and utilization time, more easily performed sensitivity analysis. Moreover, feature selection helps to reduce the risk of incurring in overfitting due to the curse of dimensionality, and this increases performance and robustness.

In the available literature, there are a variety of methods which perform feature selection but there is no single method which is appropriate for all types of problems. The main directions that have been taken to tackle the issue originally divide into wrappers, filters and embedded methods (see *Stanczyk, 2015* [33]), up to more innovative approaches like Swarm Intelligence (see for instance *Brezočnik et al.* [34]) and similarity classifier used in combination with a new

fuzzy entropy measure in signal processing (see *Tran, Elsisi, Liu*[35]). Wrappers utilize the chosen machine learning model to score many different subsets of variables according to their predictive power, in an often greedy and computationally intensive approach; filters select the variables of interest as a pre-processing step, independently of the chosen predictor; embedded methods are peculiar to certain kinds of models which perform variable selection in the process of training. Each of these approaches has its strengths and weaknesses, which make them more or less suitable for a specific problem. Many algorithms have been developed, especially in the wrapper field, to improve selection robustness and relevance, but the increase in complexity of such algorithms limits the effectiveness of proposals. What would be interesting, instead, is looking for an integrated approach, and see if a contamination of methods can actually bring some benefit and improvements to feature selection.

Our contribution with this paper is to try and improve feature selection by combining the Shapley value framework (SHAP for brevity) with different feature selection approaches. We do this in order to take advantage of SHAP's many desirable properties (*Lundberg et al., 2017* [10]): local accuracy, missingness and consistency in a linear space, in their role of indicators of variables importance, adding to the literature that employs Shapley values a post-processing phase (see for instance *Bussman et al., 2020* [20] and *Gramegna and Giudici, 2020* [28]). We show that SHAP can indeed extract futher information from the nexus data-predictive model, and that such information can be useful in selecting relevant features.

Our proposal will be tested on a real dataset, provided by the fintech company MonAI, which among other things provides eXplainable credit scores to SMEs and professionals, using both traditional and alternative data. The aim of the application is to be able to select an adequate number of features, so to have a model which is both well explainable and performing, in the setting of probability of default prediction for the considered companies.

The remainder of the paper is organized as follows. The "Data and methods" section provides an overview of the data employed and of the tested methods. The "Results" section presents our results. Finally, conclusions and future research directions are indicated in "Conclusion and future work" section.

4.2 Methods

4.2.1 Data

The data we use to test our methodology is quite traditional, being balance sheet data from the last six years belonging to italian SMEs. We can find all the classical balance sheet entries, together with some composite indexes (e.g., leverage, Return on Sales - these will be masked with the letter V and numbers from 1 to 30). In a pre processing step, we have eliminated some *linear combinations* inherently present in the data (as done in *filter* methods). As a result, we have 49 features.

We then employed statistical tools to deal with strongly unbalanced classes (e.g., *Lin et al. (2017)* [27]), since defaults were slightly more than 1%, and to remove time-specific factors. More specifically, we employed data encompassing five years (2015 to 2019), comprising more than two millions of SME observations, keeping all the defaulted cases and randomly sampling a group of non-defaulted firms equal to about 10000 for each year. With the remaining observations we build 5000 clusters per year and employ the cluster medoids as input observations (the same way we did it in *Gramegna and Giudici, 2021*[36]). This brought down class imbalance from about 100:1 to 5:1, allowing the model to better frame risk patterns and give more amplitude to probability estimation. The above procedure has led us to a *training* dataset of about 139,000 observations, with 27,200 defaults; the entire year 2020 was left out from pre processing in order to have a clean, *validation* set to test our proposed methods. We performed stratified sampling for training and testing set to maintain the balance for the positive and negative class of the Y variable.

4.2.2 Models

LightGBM

To learn the default pattern from the data and be able to provide a probabilistic estimate for each observation, we use an improved implementation of XGBoost (*Chen and Guestrin, 2016* [6]), called LightGBM. This is a gradient boosted tree model very similar to XGBoost, which features the suggestions of *Ke G., Finley T., et al.* [37] which strongly increase efficiency and scalability, greatly improving the standard gradient boosting tree model (by about 20 times). LightGBM, on the top of featuring a light and fast implementation, differs from other gradient boosted tree models in that while other algorithms grow trees horizontally, LightGBM algorithm grows them vertically, meaning it grows leaf-wise while other algorithms of the family grow level-wise. LightGBM chooses the leaf with the largest loss as start to grow the next tree: in doing so, it can lower loss more than a level wise algorithm, since it originates less redundant leaves. It also employs binarization of continuous variables, which reduces computation time a lot because there is no need to evaluate the entire range of the continuous variables and to run dispendious sorting algorithms. This way of working also makes it less suitable for small datasets, where it can easily overfit due to its sensitivity. All the above elements make lightGBM suitable to us, as we need to run the model multiple times, and on a rather big dataset.

SHAP

The SHAP framework has been proposed by *Lundberg, 2017* [10] adapting a concept from game theory (*Shapley (1952)* [9]), and has many attractive properties. With SHAP, the variability of the predictions is divided among the available features; in this way, the contribution of each explanatory variable to each point prediction can be assessed regardless of the underlying model (*Joseph (2019)* [4]). From a computational perspective, the SHAP framework returns Shapley values which express model predictions as linear combinations of binary variables that

describe whether each covariate is present in the model or not.

More formally, the SHAP algorithm approximates each prediction $f(x)$ with $g(x')$, a linear function of the binary variables $z' \in \{0,1\}^M$ and of the quantities $\phi_i \in \mathbb{R}$, defined as follows:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{4.1}$$

where M is the number of explanatory variables.

Lundberg, (2018) [12] has shown that the only additive method that satisfies the properties of *local accuracy*, *missingness* and *consistency* is obtained attributing to each variable x'_i an effect ϕ_i (the Shapley value), defined by:

$$\phi_i(f, x) = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{4.2}$$

where f is the model, x are the available variables, and x' are the selected variables. The quantity $f_x(z') - f_x(z' \setminus i)$ expresses, for each single prediction, the deviation of Shapley values from their mean: the contribution of the i-th variable.

We will use SHAP values as transformed data input to feed to feature selection methods and see how it compares with feature selections made on the original data values.

4.2.3 Feature Selection

Stepwise Feature Selection

It is a wrapper algorithm well-known in the statistical and data science communities. It performs a classical greedy approach which tests the predictive performance of different subsets of variables in a stepwise fashion, using some metric to sort variables and add or remove them from the subset to test. This because optimisation methods based on *best subset selection* quickly become intractable and prone to overfitting when p is large. Unlike best subset selection, which involves

fitting 2^p models, forward stepwise selection involves fitting one null model, along with $p - k$ models in the kth iteration, for $k = 0,..., p - 1$. This amounts to a total of $1 + p(p + 1)/2$ models. This is a substantial difference: when $p = 20$, best subset selection requires fitting 1,048,576 models, whereas forward stepwise selection requires fitting only 211 models.

Here we use an adaptation of the stepwise approach which evaluates the feature covariates in terms of predictive performance and is allowed to go in both directions when sequentially adding the variables. Basically, after adding each new variable, the method may also remove any variables that no longer provide an improvement to the model fit. Such an approach attempts to more closely mimic best subset selection while retaining the computational advantages of forward and backward stepwise selection. The details of the algorithm are described in *Introduction to Statistical Learning*, by *James, Witten, Hastie and Tibshirani* [38]. This feature selection method, being a wrapper, has the benefit of being "supervised", in the sense that we evaluate the performance of the variables directly on the output, so it is generally quite effective in identifying the most important variables. The cons are of course the computation cost which, though not as high as for best subset selection, is still something to consider; another downside of the method is that, differently from best subset selection, you are not guaranteed to select the best possible variables in term of predictive power, since this depends on the starting point and gradual inclusion of the variables, though parallel progressive exclusion of newly redundant variables does help in minimizing the problem.

LASSO

The Lasso method, short for Least Absolute Shrinkage and Selection Operator, is a linear model proposed by *Tibshirani* in 1995. It minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant; because of the nature of this constraint it tends to produce some

coefficients that are set exactly to 0 and therefore gives interpretable models. More formally, lasso regression adds "absolute value of magnitude" (*L1 penalty*) of coefficient as penalty term to the loss function, as we can see at the end of the below equation

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4.3}$$

On the contrary, ridge regression, as a shrinkage method, adds "squared magnitude" of coefficients as a penalty term to the loss function (*L2 penalty*)

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{4.4}$$

Tibshirani's simulation studies suggest that the lasso enjoys some of the favourable properties of both ridge regression and subset selection [39], making it an important example of *embedded* feature selection, which we will use in our application. LASSO has the advantage of having a low computation cost, since it is the cost of estimating regression parameters subject to penalty term; furthermore, it builds parsimonious models. On the other hand, it doesn't necessarilty select the most informative features and sometimes the variables selected are just too few.

BORUTA

The last feature selection methods we will employ is of particular interest because, in the implementation we will use, it already uses SHAP values within its algorithm to evaluate the importance of features. Boruta is a wrapper built around the random forest algorithm [5], based on two main ideas. The first is to generate "shadow features" by perturbing the original features to create a randomized version of them. These shadow features are then added to the model as further covariates and the threshold for significant variable importance becomes the highest ranked of these shadow features, according to the intuition that a feature is useful only if it is capable of doing better than the best randomized

feature. The second idea is to iterate the outlined process n times and use the binomial distribution to evaluate the significance of the feature in a probabilistic manner; that is, a *t-test* (see *Kursa et al., 2010* [40]).

Using SHAP instead of classical metrics of feature importance, such as gain, split count and permutation, can be e a nice improvement because SHAP values have proprieties, as we have seen, that allow to assess variable importance in a more thorough and consistent way.

Thanks to the above processes, Boruta is a wrapper with a somewhat different flavour with respect to other wrapper feature selection algorithms. The idea of the shadow features, combined with SHAP importance as feature score, allows it to be very effective in selecting relevant variables. Nevertheless, it's computation cost is high, since it is the cost of running the base model, estimating SHAP values and then iterate n times to perform the t-test. It also does not build parsimonious models, since it will select variables which deliver every bit of information. It is great if your goal is to eliminate noise.

4.3 Results

To compare the three proposed feature selection methods, we have applied them both to the regular dataset, with actual observations for each variable, and to a dataset made up of the SHAP values corresponding to each observation. The dimension of the data is the same in both cases.

We have then obtained the selected features under each of the three methods and for both versions of the dataset, then used a LightGBM model to assess the predictive power of the subsets.

Before comparing the performance of the 3×2 considered feature selection models, we remark that the performance of the LightGBM model with all the available 49 features, as measured by the Area Under the Curve (AUC) calculated on the test set, is equal to AUC $= 0.8706$, with an F1 score of 0.5451. We highlight that we used the mean default probability as cutoff value to binarize the target

variable for simplicity of comparison; there may be some other threshold that better maximise F1 score.

In the Table 4.1 we can see the Area Under the Curve (AUC) values for each of our proposed feature selection algorithms.

Table 4.1: Predictive performance of the compared feature selection methods

Method	**n. of Features**	**AUC**	**F1 Score**
LASSO Regular	7	0.8047	0.5156
LASSO SHAP	15	0.8625	0.5571
Bi-directional feature selection Regular	27	0.8674	0.5496
Bi-directional feature selection SHAP	33	0.8689	0.5569
Boruta Regular	26	0.8699	0.5581
Boruta SHAP	45	0.8721	0.5589

The above table empirically shows the well known trade off between explainability (better for models with fewer features) and predictive accuracy (better for models with more features). For instance, the model with only seven features selected is easier to explain but performs worse than the others. Indeed, LASSO applied to the original data is the model that leads to only seven features. This is due to the fact that it is the most parsimonious and that the marginal improvement in performance of adding a feature is lower than with other methods, as features that contribute the most are selected first.

The Lasso method applied to the SHAP dataset looks more appealing: it selects fifteen features, much less than the original 49, and with a performance that is almost as good as that obtained with the full dataset. In addition, we can take advantage of its computational speed, due to its belonging to the embedded family.

Figure 4.1 compares the ROC curves obtained with the two LASSO methods: that on the original data, and that based on Shapley values.

ROC curve comparison in figure 1 reinforce the previous comment: the LASSO on the SHAP values performs better than that on the original data.

We now move to stepwise feature selection, which compares the performances of adding/deleting a variable feature, trying to reduce the search space. The

previous table shows that the selected number of features is quite similar with both datasets, with number of variables selected from the SHAP data being slightly higher than those selected on the original data (33 against 27). The performances are also quite similar, as we expect dealing with this kind of data and with a relatively high number of selected variables.

Figure 4.2 compares the AUC performance of the stepwise method, for either datasets (original or SHAP), as the number of variables increase.

From Figure 4.2 we can see that the stepwise method on the SHAP data is quicker in achieving a high performance.

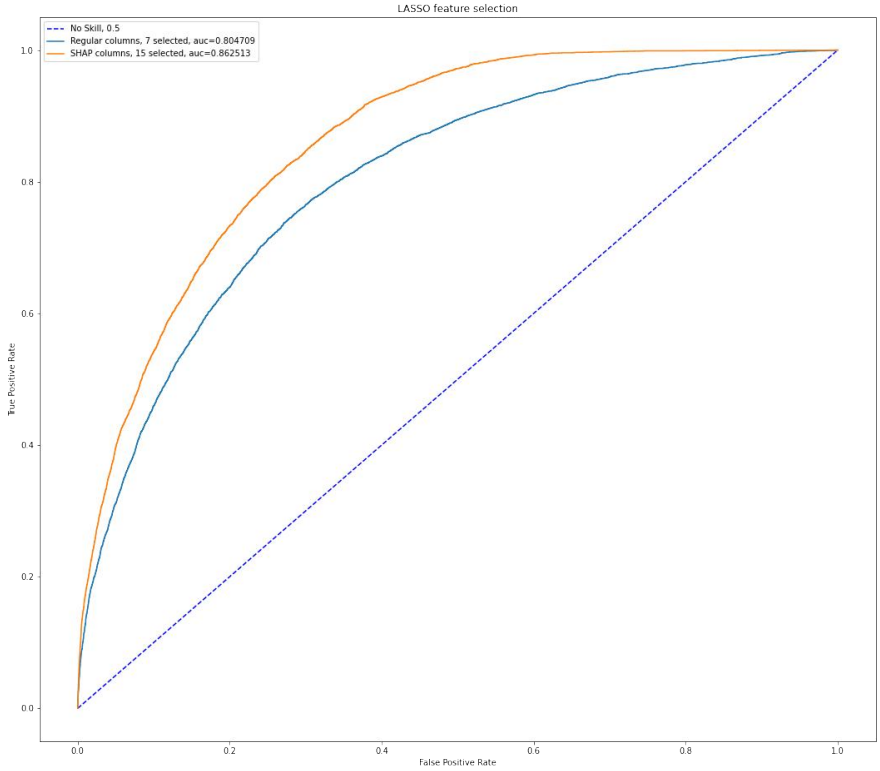
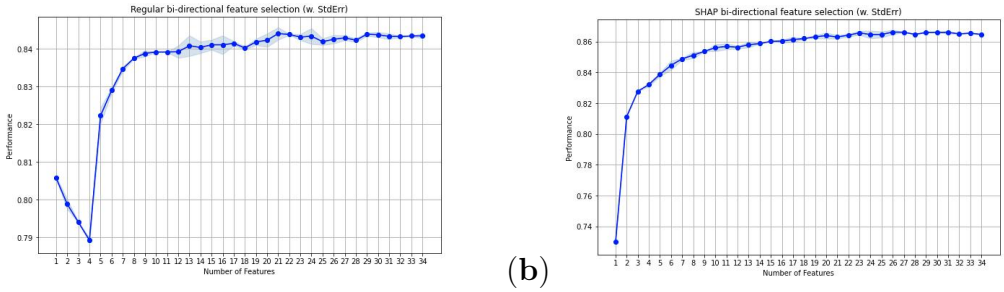


Figure 4.1: Performance of columns selected with LASSO.

We now consider the third feature selection method, Boruta. From the previous table, the number of features selected is quite different for the two considered datasets (original and SHAP). In particular, the feature selection made on the SHAP dataset leads to a model which is almost the same as the full set. Precisely, it has four less variables (45 vs. 49) and, therefore, it manages to remove some noise; nevertheless, it performs very little selection. The same cannot be said for

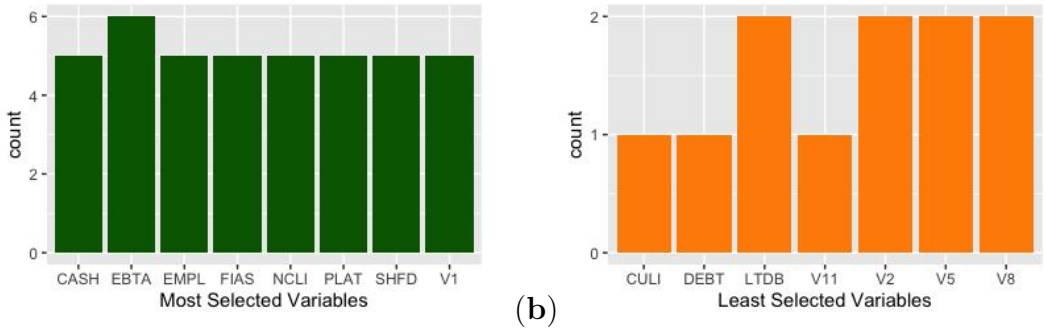
the selection made on the regular original set, where we see basically the same predictive performance as with all variables with just twenty six features. The apparent disadvantage of using SHAP values, which appears only for the Boruta method, can be explained by the fact that Boruta already uses SHAP and, therefore, using it twice is not of advantage since, as we said in the description of Boruta feature selection algorithm, this method is capable of picking up even the slightest bit of information from a variable, and the transformed SHAP dataset is made in a way that virtually every variable carry some more information with respect to base dataset.

We also consider the variables which were effectively selected by different methods by making a comparison between the overall most selected variables and overall least selected ones in figure 4.3. The methods were generally consistent in their selection, meaning that the more parsimonious methods chose variables that were also selected by more permissive algorithms, thus without overthrowing the underlying logic.



(**a**) (**b**)

Figure 4.2: (**a**) Feature selection on regular set; (**b**) feature selection on SHAP set



(**a**) (**b**)

Figure 4.3: (**a**) Frequently selected variables; (**b**) Less considered variables

We find some expected variables among the most selected ones, such as CASH (availability of liquid resources), EBTA (EBIDTA) and PLAT (Profit and Loss After Taxes); we may have expected a more involved role for debt variables (DEBT, LTDB and CULI), but this could be explained by the fact that we plugged in many balance sheet variables, together with some indexes (V1 to V12 variables), and relevant information is probably provided by complementary measures and/or ratios. This makes sense since overall debt taken as stand-alone measure does not necessarily imply a bad situation; it only has meaning when compared to other entries such as turnover, current assets and so on.

We finally remark that so far we have been comparing models on the test set, which comes from the same data preprocessing we used for the training set. To fully assess the usefulness of SHAP as a contributor to feature selection methods, we should compare the performance of all feature selection models, against that of the full model, on new, unprocessed data. We present this comparison in the next Table 4.2, using the clean data from 2020.

Table 4.2: Predictive performance on unseen data

Method	**n. of Features**	**AUC**	**F1 Score**
Full model	49	0.8137	0.5167
LASSO Regular	7	0.8012	0.5088
LASSO SHAP	15	0.8466	0.5364
Bi-directional feature selection Regular	27	0.8294	0.5188
Bi-directional feature selection SHAP	33	0.8519	0.5407
Boruta Regular	26	0.8480	0.5413
Boruta SHAP	45	0.8447	0.5430

From Table 4.2 we can see that relative performances on the 2020 data change with respect to what obtained in Table 4.1. The advantage of SHAP feature selection, for both LASSO and stepwise methods, is more evident than before. Whereas the previous caveats for the Boruta method continue to apply.

4.4 Conclusions and Future Works

In the paper, we have suggested to apply feature selection methods on the data transformed into Shapley values. Our findings show that this does improve model performance and can also reduce computational costs.

The findings also show that the best trade-off between parsimony and predictive power is obtained with a LASSO feature selection method applied to the SHAP-transformed dataset.

Future works should continue to build and investigate the possibility of integrating Shapley values with statistical model selection, as recently seen in the literature of network models (*Giudici et al., 2020* [29]), and on stochastic ordering (*Giudici and Raffinetti, 2020* [16]). It would be interesting to test the approach in different domains and within other feature selection algorithms, for instance in the medical domain (see *Baysal et al., 2020* [41]) or in remote sensing (see *Janowski et al, 2022* [42]).

Chapter 5

Machine learning classification model comparison

5.1 Introduction

Machine learning models are boosting Artificial Intelligence (AI) applications in many domains, such as automotive, finance and health care. This is mainly due to their advantage, in terms of predictive accuracy, with respect to "classic" statistical models. However, while complex machine learning models can reach high predictive performance, they have an intrinsic black-box nature.

This is a problem in regulated industries, as authorities aimed at monitoring the risks arising from the application of Artificial Intelligence (AI) methods may not validate them (see, e.g. *Joseph, A.* [43]). For example, the application of AI to finance may lead to automated decisions that can, for example, classify a company at risk of default, without explaining the underlying rationale and, therefore, impeding remedial actions.

The need to "explain" AI has become very important in recent years, following the increasing application of AI methods that impact the daily life of individuals and societies. At the institutional level, explanations can answer different kinds of questions about a model's operations depending on the stakeholder they are addressed to (see, e.g. [44]): developers, managers, model checkers, regulators. In general, to be explainable, AI methods have to provide details or reasons clarifying their functioning.

The explainability requirement is fulfilled "by design" when classic statistical models, such as logistic and linear regression, are employed within AI applications. However, in complex data analysis problems, classical statistical models may be

improved by using "black-box" machine learning models, such as neural networks and random forests.

From the previous discussion, it emerges the need to empower highly predictive machine learning models with statistical tools that can "explain" them.

Recent attempts in this direction are based on the work of Shapley (see *Shapley, 1952*[9]) who proposed to assign a score to each candidate explanatory variable based on its additional contribution to each prediction. The application of Shapley's work has led to the development of a very promising research, especially in the field of computer science (see, e.g. [43] and [8]). One of the first applications of Shapley's work to finance is due by [15], who proposed to apply correlation networks (see, e.g. [45]) to the Shapley scores and, then, cluster them into rating classes.

Shapley values have the advantage of being agnostic: independent on the underlying model with which classifications and predictions are computed; but have the disadvantage of not being normalised and, therefore, difficult to be used in comparisons outside the specific application.

Interpretability and explainability appear more relevant in complex applications, where model comparison is necessary to select a model which, maintaining accuracy, becomes parsimonious and understandable. In the traditional paradigm, a statistical model is chosen through a sequence of pairwise comparisons, based on the ratio of the likelihoods (or of the posterior probabilities) of the models being compared. Unfortunately, these criteria are generally not applicable to machine learning models such as neural networks and random forests, which do not necessarily have an underlying probabilistic model.

The previous consideration explains why the last few years have witnessed the growing importance of model selection methods based on the comparison between the predicted and the actually observed cases. In these methods, the data is split in two sets, with a "training" set used to fit a model and a "validation" set used to compare the predictions made by the fitted model with the actual observed

values.

In this paper, we contribute to the literature on model selection for machine learning models with a model comparison criterion based on the extension of Shapley values. Specifically, rather than evaluating the additional contribution of each variable to the point values of the predictions (as in the Shapley's approach), we propose to evaluate the additional contribution of each variable to the predictive accuracy of the predictions. To achieve this aim, we implement the Lorenz Zonoid tool, introduced by [46] for all types of response variables, for a binary response, exploiting its general decomposition property to derive specific criteria to compare classification models.

Indeed, to develop our approach, we extend the available likelihood model comparison procedures, applicable only to machine learning models that have a probabilistic background, to a predictive accuracy comparison framework, applicable to all possible machine learning models. To achieve this goal, we also propose a statistical test to assess the significance of the additional contribution to predictive accuracy deriving from the inclusion of an extra explanatory variable in the model. This allows to overcome the main drawbacks of the BIC and the AIC, which require a probabilistic model specification to derive the likelihood of the data. When this is missing, as in complex machine learning models, model selection needs to be reformulated in terms of descriptive statistics of the distributions of the residuals (see, e.g. [47] for a discussion), for which statistical tests for variable importance can be derived only under specific conditions. This is the case for the Diebold-Mariano test, based on the Mean Squared Error of the residuals (see [48]).

To derive our proposed model comparison procedure, we will adapt to the binary response case the work of [49], who have shown the advantage of combining Lorenz Zonoids with Shapley values to select machine learning models. We will show how to build a model comparison methodology which can be used to order variables in terms of their contribution to predictive accuracy. Doing

so, we provide a methodology that is able to simultaneously achieve the goals of predictive accuracy and explainability, rather than one after the other, as done in the explainable AI literature (see, e.g. [15]).

We will test our methodology in two different contexts: a simulated study, aimed at assessing the comparative properties of our method; and a real study, that concerns the prediction of financial default by means of a large set of highly correlated company performance variables, taken from balance sheets.

The paper is organized as follows: the next section illustrates the methodology: its background, the notion of Lorenz Zonoid predictive accuracy, and the proposed Lorenz Zonoid model comparison test; Section 5.3 introduces a simulation study to assess the performance of the methodology in the model selection context; Section 5.4 discusses the empirical findings obtained applying our proposal to the available financial data; finally, Section 5.5 contains some concluding remarks.

5.2 Methodology

To meet the requirement of a reliable risk measurement, in this section we specialise the Lorenz Zonoid decomposition approach illustrated by [**?**] to the binary classification context. Our proposal derives from the combination of two research streams. The first one concerns the development of predictive machine learning methods for classification problems. The second one concerns the development of explainable methods to understand the contribution of the explanatory variables to the predictive accuracy of machine learning models. The result is a new metric which is, at the same time, predictively accurate and interpretable.

5.2.1 Background

Let Y be a binary response variable, which can, for example and without loss of generality, express whether a company defaults ($Y = 1$) or not ($Y = 0$), as in a typical credit scoring problem. A popular model for credit scoring is the logistic

regression model (see, e.g. [**?**]).

Given K explanatory variables $X_1, \ldots, X_K$, a linear logistic regression model for Y can be specified as follows:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki} = \eta_i,$$

where $i = 1, \ldots, n$; $\eta_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}$; π_i represents the probability of the event for the i-th observation (company); $\mathbf{x}_i = (x_{1i}, \ldots, x_{Ki})$ points out the K-dimensional vector reporting the values taken by the K explanatory variables referred to the i-th observation; β_0 and β_k $(k = 1, \ldots, K)$ are the parameters representing the intercept and the k-th regression coefficient, respectively.

By means of the maximum likelihood estimation method, the parameters β_0 and β_k can be estimated leading to derive the predicted probability of default as:

$$\hat{\pi}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}},$$

which can be employed to attach to the i-th observation a "score": a number between zero and one which can be interpreted to signal, for example, the creditworthiness of a company: the higher the score the lower the trust. A classification of each company can then follow, comparing the score with an appropriate threshold, chosen on an experiential basis.

On one hand, resorting to logistic regression models for the analysis of credit scoring seems appropriate, as logistic regression models belong to a class of models which appear highly interpretable by default. On the other hand, these models sometimes provide a limited predictive accuracy. To improve predictive accuracy, more complex machine learning models may be considered, such as neural network models and XGBoost models. The requirement of high predictive accuracy is fundamental, particularly in the field of credit risk classification (see, e.g., [50], [51] and [52], among others). A wide literature review on the use of AI methods in credit risk can be found in [53].

Neural network models were developed to mimic the structure of the human brain. The idea is to treat the brain as made up of highly interconnected elements (neurons) that work together to solve specific problems. Neural network models can be described by a graph organised according to different levels: the input, the hidden and the output layers, as displayed in Figure 5.1.

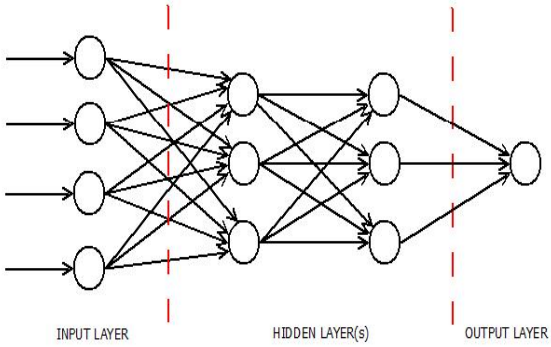


Figure 5.1: Structure of a neural network model

While the input layers receive information from the external environment and each neuron in it usually corresponds to a predictor, the output layers provides the final result to be sent outside of the system. The hidden layers define the complexity of the neural network as they contain some intermediate computational neurons, whose role is to increase the model fit. Data allow to learn the weights of the different connections between the neurons of the network.

More formally, a generic neuron j receives n input signals $x = [x_1, x_2, \ldots, x_n]$ from the neurons it is connected to in the previous layer. Each signal has an importance weight: $w_j = [w_{1j}, w_{2j}, \ldots, w_{nj}]$. Then, the same neuron elaborates the input signals through a combination function which gives rise to a value, called "potential", computed as:

$$P_j = \sum_{j=1}^{n} (x_i w_{ij} - \theta_j),$$

where θ_j is a threshold which is activated only above a certain value: a cut-off point. The output of the j-th neuron, denoted with y_j, derives from the application of a function, called activation function, to the potential P_j:

$$y_j = f(x, w_j) = f(P_j) = f\left(\sum_{j=1}^{n} x_i w_{ij} - \theta_j\right).$$

Ensemble models aim to combine the predictions derived from alternative machine learning models, thereby improving generalisation and robustness (see, e.g. [54] and the references therein).

The eXtreme Gradient Boosting (XGBoost) is one of the most popular ensemble models, particularly in credit scoring, as discussed by [**?**]. The XGBoost is a supervised model involving the combination of tree models with a Gradient Boosting Machine (GBM), which combines distinct decision trees' predictions to obtain "average" final predictions. In each decision tree, the nodes are built on a different subset of the features, implying that the trees are all different from each other and can catch distinct information from the data. At each step of the procedure, a new tree is built, learning from the errors generated by the previous trees. The XGBoost method shares the same functioning of the GBM, but it is faster and more advanced, in the sense that it provides specific regularization techniques that reduce under-fitting and over-fitting of the model, increasing its performance. A mathematical formalisation of the XGBoost is illustrated in [55].

5.2.2 Lorenz Zonoid predictive accuracy

Lorenz Zonoids were introduced in [56] as a generalisation of the ROC curve in a multidimensional setting. They were further developed by [**?**] who proposed a Lorenz Zonoid decomposition approach that can be employed for model comparison purposes. The Lorenz Zonoid is based on a measure of mutual variability and can be exploited to develop partial dependence measures that allow to detect the additional contribution of a new predictor into an existing model.

The key benefit related to the employment of the Lorenz Zonoid tool is the possibility of evaluating the contribution associated with any additional explanatory variable to the whole model prediction with a normalised measure that can be used to assess the importance of each variable.

Given a variable Y and n observations, the Lorenz Zonoid can be defined by: the Lorenz and the dual Lorenz curves (see, e.g. [57]).

The Lorenz curve for a variable Y, denoted with L_Y and obtained by re-ordering the Y values in non-decreasing sense, has points whose coordinates can be specified as $(i/n, \sum_{j=1}^{i} y_{r_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where r and $\bar{y}$ indicate the (non-decreasing) ranks of Y and the Y mean value, respectively. Similarly, the dual Lorenz curve of Y, indicated as $L_Y^{'}$ and obtained by re-ordering the Y values in a non-increasing sense, has points with coordinates $(i/n, \sum_{j=1}^{i} y_{d_j}/(n\bar{y}))$, for $i = 1, \ldots, n$, where d indicates the (non-increasing) ranks of Y. The area lying between the L_Y and $L_Y^{'}$ curves corresponds to the Lorenz Zonoid, whose graphical representation in the case of a binary response variable $Y = \{0, 1\}$ is displayed in Figure 5.2 (a).

It is worth mentioning that the Lorenz Zonoid fulfills some relevant properties. An important one is the "inclusion" of the Lorenz Zonoid built on the predicted values $\hat{Y}$ into the Lorenz Zonoid of the response variable Y, graphically depicted in Figure 5.2 (b).

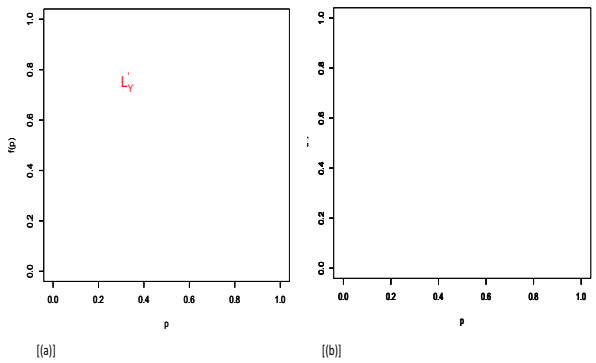


Figure 5.2: [(a)] The Lorenz curve (L_Y) and the dual Lorenz curve ($L_Y^{'}$) in the binary case; [(b)] The inclusion property $LZ(\hat{Y}) \subset LZ(Y)$ in the binary case

As shown in [**?**], given a set of K explanatory variables, and denoting with $\hat{Y}_{X' \cup X_k}$ and $\hat{Y}_{X'}$, respectively, the predicted values obtained from a model which includes a covariate X_k, and the predicted values provided by a reduced model (which excludes covariate X_k), the additional contribution related to the inclusion

of a covariate X_k can be expressed in terms of the Partial Gini Contribution (PGC) measure as:

$$PGC_{Y,X_k|X'} = \frac{LZ(\hat{Y}_{X'\cup X_k}) - LZ(\hat{Y}_{X'})}{LZ(Y) - LZ(\hat{Y}_{X'})}, \tag{5.1}$$

where $LZ(\hat{Y}_{X'\cup X_k})$, $LZ(\hat{Y}_{X'})$ and $LZ(Y)$ define: the Lorenz Zonoids computed on the predicted values provided by the model, including also covariate X_k; the Lorenz Zonoids computed on the predicted values provided by the model, including the $X^{'}$ covariates but excluding covariate X_k; the Lorenz Zonoid computed on the Y target variable values.

Note that the PGC measure can be interpreted within a game theoretical context, expressing the pay-off function in terms of the numerator of the PGC measure in equation (5.1). More precisely, for a set of statistical units ($i = 1, \ldots, n$), the pay-off in terms of the Lorenz Zonoids ($LZ(\cdot)$) is given by:

$$\textit{pay-off}\,(X_k) = LZ(\hat{Y}_{X'\cup X_k}) - LZ(\hat{Y}_{X'}), \tag{5.2}$$

where $LZ(\hat{Y}_{X'\cup X_k})$ and $LZ(\hat{Y}_{X'})$ describe the (mutual) variability of the response variable Y explained by the models which, respectively, include the $X^{'}\cup X_k$ predictors and only the $X^{'}$ predictors.

When the response variable is binary, $Y = \{0, 1\}$, the terms $\hat{Y}_{X'\cup X_k}$ and $\hat{Y}_{X'}$, in equations (5.1) and (5.2) can be re-written as the predicted probabilities of default $\hat{\pi}_{X'\cup X_k}$ and $\hat{\pi}_{X'}$, using a model that includes also the explanatory variable X_k, or a model that does not include the explanatory variable X_k. Thus, equations in (5.1) and (5.2) become

$$PGC_{Y,X_k|X'} = \frac{LZ(\hat{\pi}_{X'\cup X_k}) - LZ(\hat{\pi}_{X'})}{LZ(Y) - LZ(\hat{\pi}_{X'})} \tag{5.3}$$

and

$$\textit{pay-off}\,(X_k) = LZ(\hat{\pi}_{X'\cup X_k}) - LZ(\hat{\pi}_{X'}). \tag{5.4}$$

The pay-off in equation (5.4) measures a predictive gain, that is, the contribution to the explanation of the response variable due to each additional predictor included into the model. This result derives from the decomposition of the Lorenz Zonoid, which can be expressed as the sum of a component related to the explanatory variables X', and of a further component, function of the additional explanatory variable X_k.

The previously mentioned decomposition specialises what proved by [46], in the case of a continuous response, to the binary case. More precisely, in [46] the Authors prove that the overall contribution provided by K covariates to the explanation of a continuous response variable depends on the single contributions according to the following formula:

$$MGC_{(Y|X_1,\ldots,X_K)} = \sum_{j=1}^{K} PGC_{Y,X_j|X_{i<j}}(1 - MGC_{Y|X_1,\ldots,X_{j-1}}), \tag{5.5}$$

where $MGC_{(Y|X_1,\ldots,X_K)}$ denotes the overall response variable variability explained by all the explanatory variables (i.e., $LZ(\hat{Y}_{X_1,\ldots,X_K})$); $PGC_{Y,X_j|X_{i<j}}$ is the contribution associated with the j-th explanatory variable included into the model and $MGC_{Y|X_1,\ldots,X_{j-1}}$ is the overall contribution provided by the remaining $(j-1)$-th explanatory variables (i.e., $LZ(\hat{Y}_{X_1,\ldots,X_{j-1}})$), with $j = 1,\ldots,K$.

Note that the previous decomposition parallels the well known decomposition of the goodness of fit coefficient R^2 for linear models:

$$R^2_{Y,X_1,\ldots,X_K} = \sum_{j=1}^{K} r^2_{Y,X_j|X_{i<j}}(1 - R^2_{Y,X_1,\ldots,X_{j-1}}), \tag{5.6}$$

where $R^2_{Y,X_1,\ldots,X_K}$ represents the determination coefficient of the linear model built on the K explanatory variables, $R^2_{Y,X_1,\ldots,X_{j-1}}$ denotes the coefficient of multiple correlation between Y and the fitted plane determined by the explanatory variables $X_1,\ldots,X_{j-1}$, and $r_{Y,X_j|X_{i<j}}$ denotes the coefficient of partial correlation between Y and X_j, conditional on the explanatory variables previously included into the model.

The analogy with the R^2 decomposition can be exploited to derive a decomposition of the Lorenz Zonoid for binary response variables. To achieve this goal, we need to define goodness of fit for a binary response variable. A contribution in this direction can be found in [58], which shows that, in the binary case

$$R^2 = \frac{Var(\hat{\pi})}{Var(\hat{\pi}) + \sum_{i=1}^{n} \hat{\pi}_i(1-\hat{\pi}_i)/n}, \tag{5.7}$$

where $Var(\hat{\pi}_i)$ is the sample variance (see, e.g. [59]).

Suppose to consider, for the sake of simplicity, only two explanatory variables X_1 and X_2 (i.e., $K = 2$). Equation (5.6) can then be expressed as:

$$\begin{aligned} R^2_{X_1,X_2} &= \frac{Var(\hat{\pi}_{X_1})}{Var(\hat{\pi}_{X_1}) + \sum_{i=1}^{n} \hat{\pi}_{X_{1i}}(1-\hat{\pi}_{X_{1i}})/n} \\ &+ \frac{Var(\hat{\pi}_{X_1 \cup X_2}) - Var(\hat{\pi}_{X_1})}{\sum_{i=1}^{n} \hat{\pi}_{X_{1i}}(1-\hat{\pi}_{X_{1i}})/n} \\ &+ \left[1 - \frac{Var(\hat{\pi}_{X_1})}{Var(\hat{\pi}_{X_1}) + \sum_{i=1}^{n} \hat{\pi}_{X_{1i}}(1-\hat{\pi}_{X_{1i}})/n}\right]. \end{aligned} \tag{5.8}$$

And the decomposition in equation (5.5) can be expressed as:

$$\begin{aligned} &MGC_{(Y|X_1,X_2)} = MGC_{Y|X_1} + PGC_{Y,X_2|X_1} \cdot \\ &(1 - MGC_{Y|X_1}) = \frac{LZ(\hat{\pi}_{X_1})}{LZ(Y)} \\ &+ \frac{LZ(\hat{\pi}_{X_1 \cup X_2}) - LZ(\hat{\pi}_{X_1})}{LZ(Y) - LZ(\hat{\pi}_{X_1})} \left[1 - \frac{LZ(\hat{\pi}_{X_1})}{LZ(Y)}\right], \end{aligned} \tag{5.9}$$

where $\frac{LZ(\hat{\pi}_{X_1 \cup X_2})}{LZ(Y)} = MGC_{(Y|X_1,X_2)}$ represents the response variability share explained by the two jointly considered explanatory variables X_1 and X_2; $\frac{LZ(\hat{\pi}_{X_1 \cup X_2}) - LZ(\hat{\pi}_{X_1})}{LZ(Y) - LZ(\hat{\pi}_{X_1})} = PGC_{Y,X_2|X_1}$ measures the partial contribution provided by the inclusion of the explanatory variable X_2 in the model; $\left[1 - \frac{LZ(\hat{\pi}_{X_1})}{LZ(Y)}\right]$ denotes the variability not explained by X_1.

We remark that the relation in equation (5.9) can be derived using the proof of Result 5 in [**?**]. It can also be shown that, when used in a stepwise model

selection procedure, the path selected by the Lorenz Zonoid has a monotonicity property. More precisely, following the inclusion property, The Lorenz Zonoids of the predictions generated by a more complex model is an area which is greater than that associated with a simpler model, implying that the explained variation of Y monotonically increases with the number of predictors included into the model.

The Lorenz Zonoids $LZ(\hat{\pi}_{X' \cup X_k})$ and $LZ(\hat{\pi}_{X'})$ can also be expressed using ordinary covariance operators (see, e.g. [60]), i.e.,

$$LZ(\hat{\pi}_{X' \cup X_k}) = \frac{2Cov(\hat{\pi}_{X' \cup X_k}, r(\hat{\pi}_{X' \cup X_k}))}{nE(\hat{\pi}_{X' \cup X_k})} \quad \text{and}$$
$$LZ(\hat{\pi}_{X'}) = \frac{2Cov(\hat{\pi}_{X'}, r(\hat{\pi}_{X'}))}{nE(\hat{\pi}_{X'})}, \tag{5.10}$$

where $r(\hat{\pi}_{X' \cup X_k})$ and $\mathrm{r}(\hat{\pi}_{X'})$ are the rank scores of $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$; n is the sample size; $E(\hat{\pi}_{X' \cup X_k})$ and $E(\hat{\pi}_{X'})$ are the expected values of $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$.

5.2.3 Lorenz Zonoid model comparison

We now move to the model comparison framework.

A stepwise model comparison procedure can be implemented considering the Lorenz Zonoid tool and, more precisely, the term $LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})$ in equation (5.4). The procedure starts building K models, each one depending on one of the K predictors, and then by computing the Lorenz Zonoids of the predicted values derived from any single model.

When resorting to a forward stepwise algorithm the predictor providing the highest Lorenz Zonoid value has to be chosen as the first variable to be included into the model. Otherwise, if a backward stepwise algorithm is applied, the predictor with the lowest Lorenz Zonoid value has to be chosen as the first variable to be removed from the full model.

In the former case, the procedure continues by fitting, at each step, a more complex model by including the predictor which provides the highest contribution

measured by the difference in equation (5.4). In the latter case, the procedure continues by fitting, at each step, a simpler model by deleting the predictor characterised by the lowest contribution, which is measured by the same difference in equation (5.4).

To evaluate the statistical contribution of a single variable, we need to derive the distribution of the difference $LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'})$, where $\hat{\pi}_{X' \cup X_k}$ are the predicted values generated by the most complex model (involving the additional X_k variable) and $\hat{\pi}_{X'}$ are the predicted values generated by the simplest model (without the X_k variable), has to be derived.

To this aim, based on equation (5.10), the difference in equation (5.4) can be expressed as:

$$\begin{aligned} LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'}) = & \frac{Cov(\hat{\pi}_{X' \cup X_k}, r(\hat{\pi}_{X' \cup X_k}))}{nE(\hat{\pi}_{X' \cup X_k})} \\ & - \frac{Cov(\hat{\pi}_{X'}, r(\hat{\pi}_{X'}))}{nE(\hat{\pi}_{X'})}. \end{aligned} \tag{5.11}$$

As $r(\cdot)/n$ is the empirical transformation of the cumulative distribution function $F(\cdot)$, the terms in equation (5.11) can be re-expressed as:

$$\begin{aligned} LZ(\hat{\pi}_{X' \cup X_k}) - LZ(\hat{\pi}_{X'}) = & \frac{Cov(\hat{\pi}_{X' \cup X_k}, F(\hat{\pi}_{X' \cup X_k}))}{E(\hat{\pi}_{X' \cup X_k})} \\ & - \frac{Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'}))}{E(\hat{\pi}_{X'})}, \end{aligned} \tag{5.12}$$

where $F(\hat{\pi}_{X' \cup X_k})$ and $F(\hat{\pi}_{X'})$ are the cumulative distribution functions of $\hat{\pi}_{X' \cup X_k}$ and $\hat{\pi}_{X'}$, respectively.

In the case of linear regression, the mean of the predicted response values is always equal to the mean of the original target values, implying that $E(Y) = E(\hat{Y})$. For more general models, the aforementioned condition does not fully hold, implying that $E(\hat{\pi}_{X' \cup X_k}) = E(\hat{\pi}_{X'}) = \mu$ becomes a reasonable approximation.

Assuming such approximation, equation (5.12), which describes the marginal contribution (MC) provided by X_k, can be simplified as follows:

$$MC = \frac{Cov(\hat{\pi}_{X' \cup X_k}, F(\hat{\pi}_{X' \cup X_k}))}{\mu} - \frac{Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'}))}{\mu}. \tag{5.13}$$

In line with the previous mathematical derivations, we propose γ as an adjusted version of equation (5.13), i.e.

$$\gamma = \mu \cdot MC = Cov(\hat{\pi}_{X' \cup X_k}, F(\hat{\pi}_{X' \cup X_k})) - Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'})). \tag{5.14}$$

By denoting the covariances $Cov(\hat{\pi}_{X' \cup X_k}, F(\hat{\pi}_{X' \cup X_k})) = \xi(\hat{\pi}_{X' \cup X_k})$ and $Cov(\hat{\pi}_{X'}, F(\hat{\pi}_{X'})) = \xi(\hat{\pi}_{X'})$, γ in (5.14) can be re-written as:

$$\gamma = \xi(\hat{\pi}_{X' \cup X_k}) - \xi(\hat{\pi}_{X'}). \tag{5.15}$$

A test for the equality of the two Lorenz Zonoids, assuming the continuity of the $\hat{\pi}$ distribution, can thus be developed by setting the following hypotheses

$$H_0 : \xi(\hat{\pi}_{X' \cup X_k}) = \xi(\hat{\pi}_{X'}) \quad \text{vs} \quad H_1 : \xi(\hat{\pi}_{X' \cup X_k}) \neq \xi(\hat{\pi}_{X'}).$$

To proceed with the test, $\xi(\hat{\pi}_{X' \cup X_k})$ can be derived in terms of a U-statistic, U_1, which estimates $Cov(\hat{\pi}_{X' \cup X_k}$,
$F(\hat{\pi}_{X' \cup X_k}))$. The estimator is defined as:

$$\hat{\xi}(\hat{\pi}_{X' \cup X_k}) = U_1 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n) \hat{\pi}_{X' \cup X_{k(i)}},$$

where $\hat{\pi}_{X' \cup X_{k(i)}}$ is the i-th order statistic of $\hat{\pi}_{X' \cup X_{k1}}, \ldots,$
$\hat{\pi}_{X' \cup X_{kn}}$.

Similarly, the estimator of $\xi(\hat{\pi}_{X'})$ is U_2, specified as:

$$\hat{\xi}(\hat{\pi}_{X'}) = U_2 = \frac{1}{4\binom{n}{2}} \sum_{i=1}^{n} (2i - 1 - n)\hat{\pi}_{X'_{(i)}},$$

where $\hat{\pi}_{X'_{(i)}}$ is the i-th order statistic of $\hat{\pi}_{X'_1}, \ldots,$

$\hat{\pi}_{X'_n}$.

An estimator of $\gamma = \xi(\hat{\pi}_{X' \cup X_k}) - \xi(\hat{\pi}_{X'})$ can then be provided as a function of two dependent U-statistics:

$$\hat{\gamma} = \hat{\xi}(\hat{\pi}_{X' \cup X_k}) - \hat{\xi}(\hat{\pi}_{X'}) = U_1 - U_2. \tag{5.16}$$

Based on [61], a function of several dependent U-statistics has, after appropriate normalisation, an asymptotically normal distribution. As suggested by [62], a way to estimate the variance is to resort to the jackknife method. Specifically, the n values of $\hat{\gamma}$, pointed out with $\hat{\gamma}_{(-i)}$ (where $i = 1, \ldots, n$), are calculated by omitting one pair $(\hat{\pi}_{X' \cup X_k}, \hat{\pi}_{X'})$ at a time and the estimated variance is

$$Var(\hat{\gamma}) = \frac{n-1}{n} \sum_{i=1}^{n} (\hat{\gamma}_{(-i)} - \bar{\gamma}),$$

where $\bar{\gamma}$ is the average of $\hat{\gamma}_{(-i)}$, for $i = 1, \ldots, n$.

Following the previous derivations, the null hypothesis $H_0 : \xi(\hat{\pi}_{X' \cup X_k}) = \xi(\hat{\pi}_{X'})$ can be tested by the test statistic:

$$Z = \frac{\hat{\gamma}}{\sqrt{Var(\hat{\gamma})}} \rightarrow N(0, 1) \tag{5.17}$$

and, for a given selected significance level α, a rejection region for the null hypothesis H_0 can be defined as $|Z| \geq z_{\frac{\alpha}{2}}$.

5.3 Simulation study

In this section we present a simulation study aimed at examining the performance of the proposed model comparison procedure based on the Lorenz Zonoids.

The simulation design is illustrated in Subsection 5.3.1, whereas the corre-

sponding results are reported and commented in Subsection 5.3.2.

5.3.1 Simulation design

We consider a vector of seven random variables, including a response variable Y and six explanatory variables $X_1, \ldots, X_6$. We then generate two samples, one with 1,000 and one with 10,000 observations, both from a seven-dimensional normal distribution, with a correlation matrix specified as in Table 5.1.

Table 5.1: Correlation matrix

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1	0.8	0.5	0.3	0.1	0	0
X_1		1	0.2	0.7	0.3	0	0
X_2			1	0.05	0.1	0	0
X_3				1	0.5	0	0
X_4					1	0	0
X_5						1	0
X_6							1

Table 5.1 assumes that:

- Y is highly correlated with X_1: $\rho = 0.8$;
- Y is correlated with X_2: $\rho = 0.5$;
- Y has a low correlation with X_3: $\rho = 0.3$;
- Y has a very low correlation with X_4: $\rho = 0.1$;
- Y is not correlated with X_5 and X_6: $\rho = 0$.
- Variables X_5 and X_6 are not correlated with the other four explanatory variables X_1, X_2, X_3 and X_4.

In agreement with the rest of the paper, the response Y variable is binarised assigning values equal to 1 and 0 when the Y values are, respectively, greater or equal than the average. We apply our procedure to compare different logistic regression models, explainable by design and, consequently, simpler to be understood.

The whole dataset is then split into a training set, composed of 80% of the observations, and a test set, composed of the remaining 20% of the observations. A forward stepwise procedure is implemented by first fitting a logistic regression model on the training set and, then, including the explanatory variables which progressively provide the highest marginal contribution, as measured by the pay-off based on the Lorenz Zonoids, computed on the test set. For comparison purposes, we also consider stepwise model selections based on the AUROC and the AIC, recalling that the latter is, differently from the others, calculated on the training set. The procedure stops when the additional contribution provided by a new included predictor is not significant, using the proposed test to compare Lorenz Zonoids and the DeLong test to compare ROC curves (see, e..g. [63]).

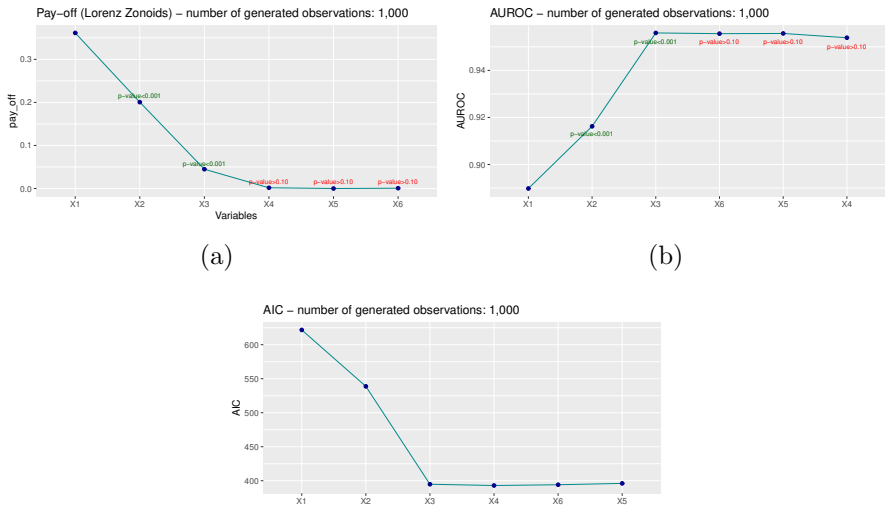
5.3.2 Simulation results

The results for model comparison are displayed in Figures 5.3 and 5.4. Figure 5.3 refers to the generating data process with 1,000 observations, while Figure 5.4 refers to the case of 10,000 observations.

At each step of the stepwise procedure, the significance of the contribution given by an additional explanatory variable is assessed through the Lorenz Zonoid and DeLong tests, whose results are reported in Figures 5.3 and 5.4 (a) and (b) in terms of the corresponding p-values.

Figure 5.3 orders the six considred explanatory variables in terms of their marginal Lorenz Zonoids, AUROC and AIC. When the marginal Lorenz Zonoid are used (Figure 5.3 (a)), the ordering is consistent with the assumed correlations between the X variables and the response variable. When the AUROC is applied (Figure 5.3 (b)), the ordering changes with X_6 (not correlated with Y) taking the place of variable X_4 (correlated with Y). Finally, the application of the AIC measure (Figure 5.3 (c)) reveals a behaviour similar to that of the marginal Lorenz Zonoids.

Figures 5.3 (a) and (b) also report the p-values that correspond to the progres-



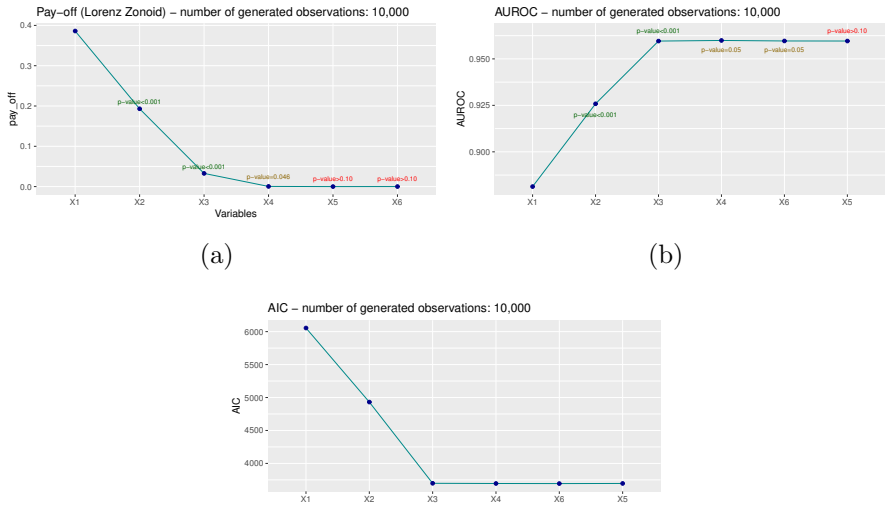
(a) (b)

(c)

Figure 5.3: Variables' selection with different methods, n = 1000

sive tests of variable inclusion. Figure 5.3 (a) indicates that a stepwise selection based on the Lorenz Zonoid tests stops with a model that contains (X_1, X_2, X_3), the most correlated variables. Figure 5.3 (b) indicate similar results when the stepwise selection is based on the DeLong tests for the AUROC.

Figure 5.4 replicates the previous analysis using a larger sample of 10,000 observations.



(a) (b)

(c)

Figure 5.4: Variables' selection with different methods, n = 10000

The p-values in Figure 5.4 (a) indicate that, with the Lorenz Zonoid procedure,

variable X_4 becomes significant, reflecting the large sample size, which allows to recognise all assumed non zero correlations. On the other hand, the procedure based on the AUROC fails to recognise the correct model, as it selects, besides X_1, X_2, X_3, X_4 also variable X_6. Last, the AIC procedure confirms the model selected with $1,000$ variables. In summary, it seems that our proposal is the best performer, as it recognises the correct correlation structure, taking sample size into account.

5.4 Application

5.4.1 Data

In this section we apply our proposed method to data supplied by Modefinance, a European Credit Assessment Institution (ECAI) that specializes in credit scoring for P2P platforms focused on SME commercial lending. The whole dataset is described by [29] to which we refer for further details. Here we focus on the twelve explanatory variables selected by the Authors: Total Assets/Total Liabilities (X_1); Current Assets/Current Liabilities (X_2); (Profit or Loss before tax+Interest paid)/Total Assets (X_3); Return on Equity (X_4); Operating Revenues/Total Assets (X_5); Interest paid/(Profit before taxes+Interest paid) (X_6); EBITDA/Interest paid (X_7); EBITDA/Operating Revenues (X_8); EBITDA/Sales (X_9); Trade Receivables/Operating Revenues (X_{10}); Inventories/Operating Revenues (X_{11}); Turnover (X_{12}).

The data on the above mentioned explanatory variables is extracted from the balance-sheets of 15,045 SMEs, mostly based in Southern Europe, for the year 2015. The data on the response variable is obtained from information about the status (0 = active, 1 = defaulted) of each SME one year later (2016), as collected from the official registers of bankruptcy. Note that the observed proportion of defaulted companies is equal to 10.9%.

5.4.2 Results

With the same data, [**?**] have constructed logistic regression scoring models that aim at estimating the probability of default of each company, using the available explanatory data and, in addition, network centrality measures that are obtained from similarity networks.

To improve the predictive performance of the model, [15] have applied the Gradient Boosting (XGBoost) tree algorithm, and obtained a substantial increase in predictive performance: the Area Under the ROC Curve (AUROC) increases from a value of 0.81 obtained with the application of the logistic regression, to a value of 0.93, obtained with the Gradient Boosting method.

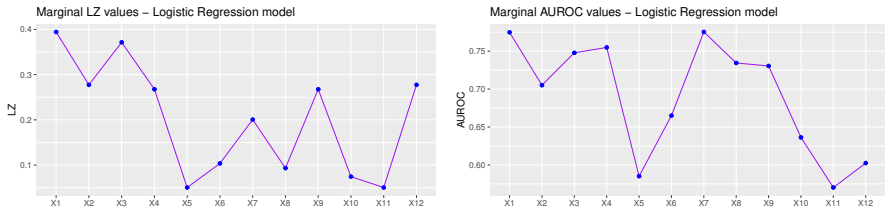
The same Authors identify the variables X_1 and X_3 as the variables that rank highest in terms of the Shapley value explanation of the probability of default, a result that is quite consistent with most credit scoring models, that typically include, among the explanatory variables of credit default, a measure of financial leverage (such as variable X_1) and a measure of profitability (such as variable X_3).

We consider the same data, and the same twelve explanatory variables as in [15], on which we apply a logistic regression model after the data is randomly split in a training set (80%) and a test set (20%). We then calculate, on the test set, the contribution of each of the explanatory variables to the estimate of the probability of default, using our proposed Lorenz Zonoid based approach. Additionally to what Shapley values can do, we provide contributions that are normalised in the $[0, 1]$ interval, and whose additional value can be assessed in terms of its statistical significance. Doing so, we show how a model comparison procedure based on the Lorenz Zonoids can improve the explainability of a machine learning model, choosing a parsimonious set of explanatory variables while maintaining a high predictive accuracy.

The implementation of our proposed model comparison procedure starts by computing the marginal contribution of each single explanatory variable X_j, for

$j = 1, \ldots, 12$, to the explanation of the probability of default. The marginal contributions are determined by building twelve simple logistic regression models, each of them involving only one of the twelve predictors, and calculating the Lorenz Zonoid value $LZ(X_j)$ for each of them. This leads to a ranking of the explanatory variables, to be used in the stepwise procedure. In the forward perspective, the variable with the highest $LZ(X_j)$ value is selected as the first variable to be included in the model. Then, progressively, more complex models are implemented by introducing at each step an additional variable, according to the obtained variable ranking. Conversely, in the backward perspective, the variable with the lowest $LZ(X_j)$ value is selected as the first variable to be removed from the full model and, then, progressively, simpler models are implemented by deleting at each step according to the reversed variable ranking.

The marginal contributions of each considered explanatory variable, measured in terms of $LZ(X_j)$, along with the corresponding value of the AUROC, for comparison purposes, are displayed in Figure 5.5 (a) and (b), respectively.



(a) (b)

Figure 5.5: Variables' marginal contribution - Logistic regression model

From Figure 5.5 (a), the variables that contribute the most are variables X_1 and X_3, as in [**?**], followed by X_9 and, then, the others. The least important results to be X_{11}. Differently, from Figure 5.5 (b), the most important variable is X_7, followed by X_1, X_4 and the others. The least important results to be X_{11}.

We have then implemented a Lorenz Zonoid and an AUROC forward stepwise procedure starting from X_1 and, then, progressively adding the other variables, up to the full model. At each step, the additional contribution of the new added variable is measured by *pay-off* (X_k). For the sake of completeness, we also report

the F_1 accuracy index, a standard practice as the AUROC, in the seventh column of Table 5.2. To decide when to stop the procedure at a certain step, we apply the statistical test proposed in Subsection 5.2.3 and, thus, continue the process until the additional contribution is significantly different from zero. In this way the selected model represents a good trade-off between predictive accuracy (which increases with model complexity) and explainability (which decreases with model complexity).

The results of the procedure, based on the Lorenz Zonoid pay-offs, are illustrated in Table 5.2.

Table 5.2: Logistic regression model (forward stepwise) - Marginal contributions ($LZ(\hat{Y}_{X_j})$); additional contributions ($pay\text{-}off\,(X_k)$); significance (p-value) of the additional contributions; F_1 metric. Legend: TA/TL=Total assets/Total Liabilities; (PLBT+IP)/TA=(Profit or Loss before tax+Interest paid)/Total Assets; EBITDA/S=EBITDA/Sales; TO=Turnover.

ID	Variable	$LZ(\hat{Y}_{X_j})$	ID of the included variables	$pay\text{-}off\,(X_k)$	p-value	F_1
1	TA/TL	0.3943	1	-	-	-
3	(PLBT+IP)/TA	0.3714	1, 3	0.0544	$<$ 0.001	0.3844
9	EBITDA/S	0.3244	1, 3, 9	0.0081	$<$ 0.001	0.3865
12	TO	0.3061	1, 3, 9, 12	0.0002	0.2069	0.3865

Looking at Table 5.2 and, in particular, at the p-values of the test, reported in the sixth column, we obtain that the best model includes three explanatory variables: X_1, X_3, as in the reference literature (see, e.g. [**?**]), and also variable X_9. For comparison purposes, Table 5.3 highlights the results of the procedure based on the AUROC differences.

Table 5.3: Logistic regression model (forward stepwise) - Marginal contributions (AUROC); additional contributions (difference of AUROC); significance (p-value) of the additional contributions; F_1 metric. Legend: EBITDA/IP=EBITDA/Interest paid; TA/TL=Total assets/Total Liabilities.

ID	Variable	$AUROC_{X_j}$	ID of the included variables	$pay\text{-}off\,(X_k)$	p-value	F_1
7	EBITDA/IP	0.7753	7	-	-	-
1	TA/TL	0.4113	7, 1	0.0016	0.9050	0.2942

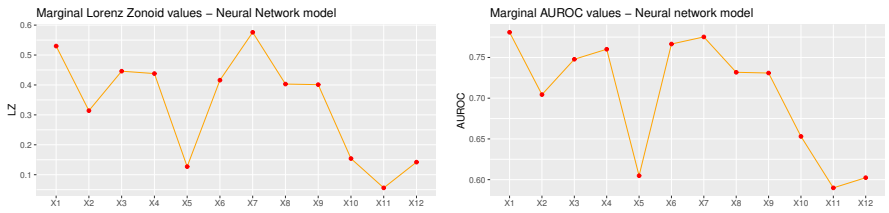
In agreement with Figure 5.5 (b), Table 5.3 shows that the best model contains

variable X_7 (EBITDA/Interest paid). In addition, the DeLong test indicates to stop at that point, leading to a very parsimonious model, with only one variable. We remark that the result of the AUROC based procedure is not in line with the literature, as it includes in the model a measure of profitabilty but not a measure of financial leverage.

We also remark that, for robustness purposes, we have implemented a backward stepwise procedure, for both the Lorenz Zonoid pay-off and the AUROC. The results have confirmed the significance of the variables contained in the models selected with the forward procedure.

We also remark that a very important aspect of our proposal is its generality: it allows to extend the same model comparison procedure to more complex frameworks, which do not necessarily have a probabilistic background, such as those based on the employment of neural networks and on tree models such as XGBoost models.

We now report the results of model comparison, for a neural network model built (without loss of generality) with five neurons in the hidden layer. The behaviour of the $LZ(X_j)$ and of the AUROC for each explanatory variable is shown in Figures 5.6 (a) and (b), respectively.



(a) (b)

Figure 5.6: Variables' marginal contribution - Neural network model

From Figure 5.6 (a), the variables that contribute the most are variable X_7 and X_1, and similarly in Figure 5.6 (b), although in a reversed order. Additionally, Figure 5.6 (b) indicates a high importance also for variable X_6. In both cases, the least important results to be X_{11}.

The results of the stepwise procedure for the neural network models are re-

ported, respectively, in Table 5.4, for the $LZ(X_j)$ measure; and in Table 5.5, for the AUROC measure.

Table 5.4: Neural network model (forward stepwise) - Marginal contributions ($LZ(\hat{Y}_{X_j})$); additional contributions ($pay\text{-}off\,(X_k)$); significance (p-value) of the additional contributions; F_1 metric. Legend: TA/TL=Total assets/Total Liabilities; IP/(PBT+IP)=Interest paid/(Profit before taxes+Interest paid); EBITDA/IP=EBITDA/Interest paid.

ID	Variable	$LZ(\hat{Y}_{X_j})$	ID of the included variables	$pay\text{-}off\,(X_k)$	p-value	F_1
1	TA/TL	0.5343	1	-	-	-
6	IP/(PBT+IP)	0.4684	1, 6	0.0212	< 0.001	0.4154
7	EBITDA/IP	0.4574	1, 6, 7	0.0009	0.7806	0.400

Table 5.5: Neural network model (forward stepwise) - Marginal contributions (AUROC); additional contributions in terms of AUROC difference; significance (p-value) of the additional contribution; F_1 metric. Legend: TA/TL=Total assets/Total Liabilities; EBITDA/IP=EBITDA/Interest paid; IP/(PBT+IP)=Interest paid/(Profit before taxes+Interest paid).

ID	Variable	$AUROC_{X_j}$	ID of the included variables	$pay\text{-}off\,(X_k)$	p-value	F_1
1	TA/TL	0.7809	1	-	-	-
7	EBITDA/IP	0.7752	1, 7	0.0219	0.0426	0.4366
6	IP/(PBT+IP)	0.7665	1, 7, 6	0.0013	0.8348	0.4000

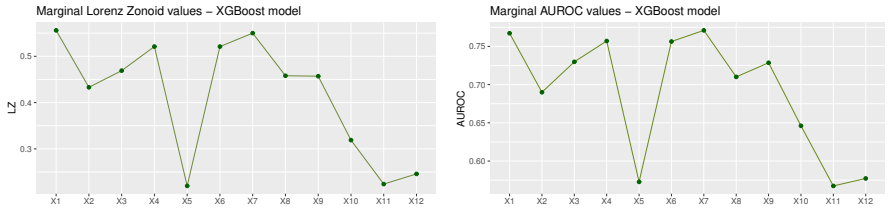
From Table 5.4 we obtain that, similarly from what occurs for logistic regression models, the neural network procedure selects two variables, and one is X_1. However, the second variable is X_6 and not X_3. From a financial viewpoint, the results are indeed similar, as both X_3 and X_6 measure profitability, whereas X_1 indicates financial leverage.

Similar conclusions can be derived when the AUROC metric is employed in place of the Lorenz Zonoid pay-off. Table 5.5 shows that, again, two explanatory variables are included in the selected model. While the first one is confirmed to be X_1, the second is X_7, instead of X_6: another function of the profitability. These results are confirmed when a backward selection procedure is implemented, for robustness.

In summary, the application of the procedure to neural networks shows that

both the Lorenz Zonoid and the AUROC model selection lead to choose a model with two variables (one measuring leverage and one measuring profitability), which represents a very good trade-off between explainability and accuracy. On one hand, the model is more explainable than the full model, as the response depends significantly only on two variables, and we know which ones (whereas a full neural network model is a black box); on the other hand, the model is accurate as its predictive accuracy is not significantly improved making it more complex (adding more variables).

We can apply our procedure, in the same way, to another type of machine learning model: the XGBoost, which belongs to the class of tree models. The results are illustrated, from a graphical view point, in Figure 5.7 (a) and (b); and are specified with numerical details in Tables 5.6 and 5.7.



(a) (b)

Figure 5.7: Variables' marginal contribution - Extreme gradient boosting model

Table 5.6: XGBoost model (forward stepwise)- Marginal contribution in terms of each single explanatory variable ($LZ(\hat{Y}_{X_j})$); marginal contribution in terms of any additional explanatory variable ($pay\text{-}off\,(X_k)$); the marginal contribution significance (p-value); F_1 metric. Legend: TA/TL=Total assets/Total Liabilities; EBITDA/IP=EBITDA/Interest paid; IP/(PBT+IP)=Interest paid/(Profit before taxes+Interest paid); ROE=Return on Equity.

ID	Variable	$LZ(\hat{Y}_{X_j})$	ID of the included variables	$pay\text{-}off\,(X_k)$	p-value	F_1
1	TA/TL	0.5565	1	-	-	-
7	EBITDA/IP	0.5496	1, 7	0.0747	<0.001	0.4170
6	IP/(PBT+IP)	0.5212	1, 7, 6	0.0052	<0.001	0.4386
4	ROE	0.5210	1, 7, 6, 4	0.0035	0.0758	0.4390

Figure 5.7 (a) shows that variables X_1 and X_7, followed by X_6, are the factors with the highest impact on the probability of default. Figure 5.6 (b) shows a

Table 5.7: XGBoost model (forward stepwise) - Marginal contributions (AUROC); additional contributions in terms of AUROC difference; significance (p-value) of the additional contribution; F_1 metric. Legend: EBITDA/IP=EBITDA/Interest paid; TA/TL=Total assets/Total Liabilities; ROE=Return on Equity.

ID	Variable	$LZ(\hat{Y}_{X_j})$	ID of the included variables	*pay-off* (X_k)	p-value	F_1
7	EBITDA/IP	0.7710	7	-	-	-
1	TA/TL	0.7672	7, 1	0.0362	< 0.001	0.4170
4	ROE	0.5210	7, 1, 4	0.0068	0.1282	0.4105

similar results, swapping X_1 with X_7 and replacing X_6 with X_4.

In terms of model selection, both procedures lead to select a model that contains X_1 and X_7. Additionally, the Lorenz Zonoid based procedure includes also X_6, leading to a more complex model, with three significant contributions. We remark that also in this case, the backward model search confirms the selected variables.

The conclusions that can be drawn from the XGBoost model selection procedure are in line with those from the neural network model. Overall, the empirical findings from our analysis can be summarised with the conclusion that the proposed model selection procedure, based on the Lorenz Zonoids, is able to simplify a black box machine learning model into an explainable model.

From a financial viewpoint, all models indicate that the most important variables for credit scoring are: a measure of financial leverage and a measure of profitability, confirming the previous analysis of [15] and [22] on the same data.

A natural question that arises is: which of the three model champions is the best model overall, both in absolute terms (predictive accuracy) and in relative terms, with respect to the full model (explainability)? To answer this question, the logistic regression, neural network and XGBoost models selected with the Lorenz Zonoid approach are compared in terms of the predictive accuracy of their full model and selected model. To achieve an "external" evaluation, predictive accuracy is evaluated using the AUROC measure. The results can be found in

Table 5.8.

Table 5.8: Predictive accuracy of the selected and full models

	AUROC selected model	AUROC full model
Logistic regression model	0.8037	0.8045
Neural network model	0.7800	0.7810
XGBoost	0.8110	0.8557

From Table 5.8 note that the best machine learning model, in terms of predictive accuracy, is the XGBoost model, with an AUROC of 0.8110; whereas the neural network model is the worst one, with an AUROC of 0.78. On the other hand, the XGboost model is the least explainable model: differently from what occurs for the logistic regression and neural networks, the AUROC of the full model reduces substantially and in a significant way (p-value greater than 0.05) moving to the reduced model.

5.5 Concluding remarks

The paper proposes to improve machine learning models by means of a model selection methodology, based on the Lorenz Zonoids, which allows to maintain a high predictive accuracy, explaining the predictions with a parsimonious set of explanatory variables.

The proposal is quite general and can be applied to any machine learning model, whether based on a probabilistic framework or not. In the case of a binary response, the approach is also consistent with the results that can be obtained applying the well known AUROC accuracy measure.

Further advantages of our proposed procedure are: its generality (in the paper we have considered a binary response, but the same tool can be applied for ordinal or continuous response, differently from what occurs for the AUROC); its computational efficiency (we do not need to calculate the Lorenz Zonoids of all models, but only of those considered in the stepwise path, differently from what occurs with the Shapley value approach to explainability).

The application of the proposal to a simulated data has shown that it is

capable to select the correct model, and to take into account the sample size. The application of the proposal to a real credit scoring database has shown its capability to identify, as relevant variables, those that concern the profitability and the financial leverage of the companies asking for credit.

We believe that the proposed method could be employed as a use case to improve the compliance of Artificial Intelligence applications in finance to principles such as Sustainability, Accuracy, Fairness and Explainability, leading to a S.A.F.E. approach to AI which can be desumed, for example, from the European AI act (`artificialintelligenceact.eu`).

Further research may focus on the application of the methodology to other machine learning applications, that involve different type of variables: ordinal or continuous. The generality of the proposed measure allows to do so, differently from what occurs with available metrics such as the AUROC and the MSE.

Chapter 6

Concluding Remarks

6.1 Summary

The motivation of this thesis is due to the growing attention to explainability in Machine Learning and Artificial Intelligence applications. In these contexts, the understanding of how these so called "black-box" models make their decision has a crucial role. Other than regulatory and ethical issues, exploring what causes a model to output a particular prediction, as well as which are the most important global predictors for a specific problem, can increase model performances and give useful insights on which features should be leveraged. In this thesis, these concepts are brought one step further: we research the ability of eXplainable AI models of framing and exploiting new pieces of information, information which is already there, within the predictive models and the data matrix, but emerges from another perspective and makes a useful addendum to traditional data modeling. In fact, we show how to use XAI model outputs not only for analysis, which is still a primary feature of these models, but as transformed inputs as well, employing them in predictive models, clustering, feature and model selection.

In chapter 2, we explore the behaviour of customer interacting with insurance policies in a new, digital way. We modeled behaviours of customer churn and propensity to buy with state-of-the-art, powerful classifiers, in order to focus then on interpretability, which we used to build a model for customer segmentation which is different from traditional ones. Results suggest that explainable machine learning models can effectively improve our understanding of customers' behaviour, and that further investigation may involve the application of the model

to other situations arising in the insurance industry, which may gain from the application of artificial intelligence technologies, as well as extension to other industries and case studies.

In chapter 3 we focus on the analysis of two of the most accredited XAI models, specifically Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). These models are compared through the case study of Probability of Default (PD) estimation for italian Small and Medium Enterprises. Needless to say, it is important to understand which factors impact complex models predictions and are relevant for increasing or lowering the predicted PD. We therefore look at these XAI models in order to see how consistent their parameters attribution are across the whole set of data, with SHAP coming out as the best comparable method, due to its very desirable properties.

In chapter 4, we look for ways to leverage values attributed to the feature-observation pair by explainable model. We use a dataset made up of these values to test how different feature selection models fare with respect to their application to original data. Our findings confirm the validity of the approach, obtaining a balanced model capable of satisfying both parsimony and predictive accuracy, beating traditional feature selection methodologies. This finding is important because in an environment ever-growing with features and data, appropriate feature selection serves the purpose of increasing robustness and generalization of models while keeping computational cost low.

In chapter 5 we propose a methodology to improve machine learning models through a model selection procedure based on the Lorenz Zonoids. This is based on the idea of calculating the marginal contribution of variables to the model as a whole, together with the provision of a path to perform model selection and a statistical test of significance. We show that, in case of binary target variable, the approach is consistent with the results provided by the well known AUROC accuracy measure. Further advantages are generality (the tool can can be applied

for ordinal or continuous response contrary to what happens with AUROC; plus, it can be applied to any machine learning model, whether probabilistic or not), computational efficiency (it is easy to calculate the Lorenz Zonoids of the models considered in the stepwise path, differently from what happens with the Shapley value approach to explainability) and statistical validity. We applied the proposed methodology to both simulated data, to show its behaviour in a controlled environment, and on real data provided by a rating agency.

All the outlined works show it is possible to fruitfully employ this new layer of models, the eXplainable ones, synergistically with predictive and clustering models, to achieve a higher level of understanding of problems and ultimately improve the benefits that Machine Learning and Artifical Intelligence can bring to the world.

REFERENCES

[1] G. Bernardino, "Chalenges and opportunities for the insurance sector," *Annales des Mines*, 2020.

[2] E. Commission, "On artificial intelligence - a european approach to excellence and trust," 2020. [Online]. Available: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

[3] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[4] A. Joseph, "Shapley regressions: A framework for statistical inference on machine learning models," 03 2019. [Online]. Available: https://www.kcl.ac.uk/business/assets/pdf/dafm-working-papers/2019-papers/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.pdf

[5] L. Breiman, "Bias, variance , and arcing classifiers," *Technical Report 460, Statistics Department, University of California*, 11 2000. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.7931&rep=rep1&type=pdf

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," p. 785–794, 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[7] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *CoRR*, vol. abs/1802.01933, 2018. [Online]. Available: http://arxiv.org/abs/1802.01933

[8] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[9] L. S. Shapley, "A value for n-person games." *Defense Technical Information Center*, 8 1952. [Online]. Available: https://www.rand.org/pubs/papers/P0295.html

[10] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NIPS*, 2017. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[11] E.

Strumbelj and I. Kononenko, "An efficient explanation of individual classifications using game theory," *J. Mach. Learn. Res.*, vol. 11, pp. 1–18, 2010. [Online]. Available: https://dl.acm.org/doi/10.5555/1756006.1756007

[12] S. Lundberg, G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," 02 2018. [Online]. Available: https://arxiv.org/pdf/1802.03888.pdf

[13] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297. [Online]. Available: https://pdfs.semanticscholar.org/a718/b85520bea702533ca9a5954c33576fd162b0.pdf

[14] K. Bindra and A. Mishra, "A detailed study of clustering algorithms," in *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, 2017, pp. 371–376.

[15] M. Bussmann, Giudici and Papenbrock, "Explainable machine learning in credit risk mamagement," *Computational Economics*, 2020.

[16] P. Giudici and E. Raffinetti, "Shapley lorenz zonoids," *Journal of Classification*, 2020.

[17] S. e. a. Lundberg, "From local explanations to global understanding with explainable ai for trees," *Nature machine learning*, 2020. [Online]. Available: https://www.nature.com/articles/s42256-019-0138-9

[18] E. Commission, "Directive (eu) 2015/2366 of the european parliament and of the council of 25 november 2015 on payment services in the internal market," 2015. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2015/2366/oj/eng

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," p. 1135–1144, 2016. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[20] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable ai in fintech risk management," *Frontiers in Artificial Intelligence*, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frai.2020.00026

[21] A. C. M. J. Ariza-Garzón, J. Arroyo and M. Segovia-Vargas, "Explainability of a machine learning granting scoring model in peer-to-peer lending," *IEEE Access, vol. 8, pp. 64873-64890*, 2020. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9050779

[22] B. H. Misheva, J. Osterrieder, A. Hirsa, O. Kulkarni, and S. F. Lin, "Explainable ai in credit risk management," 2021.

[23] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," p. 849–856, 2001.

[24] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[25] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[26] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320396001422

[27] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409-410, pp. 17–26, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025517307235

[28] A. Gramegna and P. Giudici, "Why to buy insurance? an explainable artificial intelligence approach," *Risks*, vol. 8, no. 4, 2020. [Online]. Available: https://www.mdpi.com/2227-9091/8/4/137

[29] P. Giudici, B. Hadji-Misheva, and A. Spelta, "Network based credit risk models," *Quality Engineering*, vol. 32, no. 2, pp. 199–211, 2020. [Online]. Available: https://doi.org/10.1080/08982112.2019.1655159

[30] P. J. Subramanian D., Greiner R., Ed., *Land Economics*, vol. 97, no. 1-2, 1997. [Online]. Available: https://www.sciencedirect.com/journal/artificial-intelligence/vol/97/issue/1

[31] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1157–1182, mar 2003.

[32] X.-w. Chen and M. Wasikowski, "Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems," p. 124–132, 2008. [Online]. Available: https://doi.org/10.1145/1401890.1401910

[33] U. Stanczyk, "Feature evaluation by filter, wrapper, and embedded approaches," *Studies in Computational Intelligence*, vol. 584, pp. 29–44, 12 2015.

[34] L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," *Applied Sciences*, vol. 8, no. 9, 2018. [Online]. Available: https://www.mdpi.com/2076-3417/8/9/1521

[35] M.-Q. Tran, M. Elsisi, and M.-K. Liu, "Effective feature selection with fuzzy entropy and similarity classifier for chatter vibration diagnosis," *Measurement*, vol. 184, p. 109962, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224121008964

[36] A. Gramegna and P. Giudici, "Shap and lime: An evaluation of discriminative power in credit risk," *Frontiers in Artificial Intelligence*, vol. 4, p. 140, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frai.2021.752558

[37] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[38] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013. [Online]. Available: https://faculty.marshall.usc.edu/gareth-james/ISL/

[39] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society (Series B)*, vol. 58, pp. 267–288, 1996.

[40] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software*, vol. 36, no. 11, p. 1–13, 2010. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v036i11

[41] Y. A. Baysal, S. Ketenci, I. H. Altas, and T. Kayikcioglu, "Multi-objective symbiotic organism search algorithm for optimal feature selection in brain computer interfaces," *Expert Systems with Applications*, vol. 165, p. 113907, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420307028

[42] L. Janowski, K. Tylmann, K. Trzcinska, J. Tegowski, and S. Rudowski, "Exploration of glacial landforms by object-based image analysis and spectral parameters of digital elevation model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[43] A. Joseph, "Parametric inference with universal function approximators," *Bank of England working papers*, vol. 784, 2020.

[44] C. J. Philippe Bracke, Anupam Datta and S. Sen, "Machine learning explainability in finance: An application to default risk analysis," *Bank of England working papers*, vol. 816, 2019.

[45] R. Mantegna and H. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance*, 12 2000, vol. 53.

[46] P. Giudici and E. Raffinetti, "Lorenz model selection," *Journal of Classification*, vol. 37, 01 2020.

[47] R. Rossi, A. Murari, P. Gaudio, and M. Gelfusa, "Upgrading model selection criteria with goodness of fit tests for practical applications," *Entropy*, vol. 22, no. 4, 2020. [Online]. Available: https://www.mdpi.com/1099-4300/22/4/447

[48] F. Diebold and R. Mariano, "Comparing predictive accuracy," *Journal of Business Economic Statistics*, vol. 20, pp. 134–44, 02 2002.

[49] P. Giudici and E. Raffinetti, "Shapley-lorenz explainable artificial intelligence," *Expert Systems with Applications*, vol. 167, p. 114104, 10 2020.

[50] V. Pacelli and M. Azzollini, "An artificial neural network approach for credit risk management," *JILSA*, vol. 3, pp. 103–112, 01 2011.

[51] N.-C. Hsieh and L.-P. Hung, "A data driven ensemble classifier for credit scoring analysis," *Expert Systems with Applications*, vol. 37, pp. 534–545, 01 2010.

[52] P. Kumar and V. Ravi, "Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review," *European Journal of Operational Research*, vol. 180, pp. 1–28, 02 2007.

[53] V. Pacelli and M. Azzollini, "An artificial neural network approach for credit risk management," *JILSA*, vol. 3, pp. 103–112, 01 2011.

[54] C. Zhang and Y. Ma, *Ensemble machine learning: Methods and applications*, 01 2012.

[55] J. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, 11 2000.

[56] G. Koshevoy and K. Mosler, "The lorenz zonoid of a multivariate distribution," *Journal of the American Statistical Association*, vol. 91, no. 434, pp. 873–882, 1996.

[57] M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publications of the American Statistical Association*, vol. 9, no. 70, pp. 209–219, 1905.

[58] J. T. Kent and J. O'Quigley, "Measures of dependence for censored survival data," *Biometrika*, vol. 75, no. 3, pp. 525–534, 09 1988. [Online]. Available: https://doi.org/10.1093/biomet/75.3.525

[59] M. Schemper, "Predictive accuracy and explained variation," *Statistics in medicine*, vol. 22, pp. 2299–308, 07 2003.

[60] R. I. Lerman and S. Yitzhaki, "A note on the calculation and interpretation of the gini index," *Economics Letters*, vol. 15, no. 3, pp. 363–368, 1984. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165176584901265

[61] W. Hoeffding, "A Class of Statistics with Asymptotically Normal Distribution," *The Annals of Mathematical Statistics*, vol. 19, no. 3, pp. 293 – 325, 1948. [Online]. Available: https://doi.org/10.1214/aoms/1177730196

[62] A. Y. Schechtman E, Yitzhaki S, "The similarity between mean-variance and mean gini: Testing for equality of gini correlations," *Advances in Investment Analysis and Portfolio Management.*, pp. 97–122, 2008.

[63] C.-P. D. DeLong ER, DeLong DM, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, pp. 837–845, 1988.

APPENDICES

A Appendix A Title

Table A.1: Variables' description

Variable Name	Description
day_of_week	The weekday of the event
days_since_last	Days passed since last interaction with user
device	Type of device
models	Device model
models_Other	Catch-all label for low frequency device models
month	Month where the event occurs
n_afternoon	Cumulative number of interactions occurred this moment of the day
n_autumn	Cumulative number of interactions occurred this season
n_Friday	Cumulative number of interactions occurred this day
n_morning	Cumulative number of interactions occurred this moment of the day
n_on_demand	Cumulative number of requested policy quotes
n_push_notification	Cumulative number of notification pushed on device
n_Saturday	Cumulative number of interactions occurred this day
n_spring	Cumulative number of interactions occurred this season
n_summer	Cumulative number of interactions occurred this season
n_Tuesady	Cumulative number of interactions occurred this day
n_Wednesday	Cumulative number of interactions occurred this day
n_winter	Cumulative number of interactions occurred this season
number_bought	Number of bought policies
number_pushed	Number of times the insurance quote has been sent
os_Android	Flag to represent device OS Android
os_iOS	Flag to represent device OS iOS
season	Season where the event occurs
time_of_day	Moment of the day where the event occurs

B Published and Submitted Papers

• Alex Gramegna and Paolo Giudici, "Why to Buy Insurance? An Explainable Artificial Intelligence Approach ", *Risks*, December 2020, doi: 10.3390/risks8040137

• Alex Gramegna and Paolo Giudici, "SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk", *Frontiers in Artificial Intelligence*, September 2021, doi: 10.3389/frai.2021.752558.

• Alex Gramegna and Paolo Giudici, "Shapley Feature Selection", *Fintech*, February 2022, doi: 10.3390/fintech1010006.

• Paolo Giudici, Alex Gramegna and Emanuela Raffinetti "Machine learning classification model comparison", submitted to *Socio-Economic Planning Sciences*, October 2022.