ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics
Master Thesis Business & Economics: Financial Economics

Efficacy of Public Guarantees in Preventing Potential Bankruptcy during the Pandemic in Italy: A Machine Learning Comparative Analysis

Sruthi Ramesh
618851sr

Supervisor: Fabrizio Core, Dr.
Second assessor: Adriana Breaban, Dr.

07-08-2023

ABSTRACT

This paper examines the effect of public credit guarantees in reducing bankruptcy in Italian companies during the pandemic. The analysis is performed by employing machine learning models of Random forest and XGBoost. This paper also examines whether the probabilities of bankruptcy have reduced for those companies that received these loans. I find evidence that while the companies that received the loans have marginally higher bankruptcy probabilities than the ones that did, their individual bankruptcy scores are lowered once they receive the loans.

Table of Contents

1. Introduction:

1.1. Background and relevance:

The COVID-19 pandemic has wrought unprecedented challenges upon global economies, posing a significant threat to the survival of businesses, particularly small and medium-sized enterprises (SMEs). As countries struggled to contain the virus and mitigate its impact, governments worldwide took various measures to support struggling businesses and prevent widespread bankruptcies. One such measure that gained prominence was the provision of public guarantees to SMEs, aiming to alleviate their financial burden and enhance their survival prospects during the crisis.

This thesis undertakes a comprehensive comparative analysis to shed light on the potential bankruptcy levels of SMEs with public guarantees versus non-guaranteed firms in the face of the pandemic. By exploring this critical aspect of government intervention, the aim is to gain valuable insights into the effectiveness of public guarantees in safeguarding the survival of SMEs, which are the backbone of many economies, and draw valuable lessons for future policy-making. The rationale for this study stems from the urgent need to understand the true impact of public guarantees on SMEs' financial resilience during times of crisis. Historically, SMEs have faced higher bankruptcy risks than larger enterprises, mainly due to limited access to credit, thin profit margins, and limited resources to weather economic downturns. The pandemic, with its far-reaching consequences, exacerbated these vulnerabilities and amplified the challenges faced by SMEs, making them particularly susceptible to bankruptcy.

To conduct the comparative analysis, relevant financial data will be analysed, such as revenue trends, profitability ratios, liquidity positions, and leverage levels, for SMEs that had historically gone bankrupt. These inputs will then be used in machine learning models to predict the performance of firms that received the public guarantees and their non-guaranteed counterparts in the short term.

By systematically examining the bankruptcy levels of SMEs with and without public guarantees, this study aspires to identify patterns, correlations, and potential causal relationships that can inform policy design and improve future crisis management strategies. These findings may assist governments, financial institutions, and policymakers in better targeting their efforts and resources to safeguard the survival and growth of SMEs, not only during pandemics but also in times of economic uncertainty.

1.2. Bankruptcy analysis:

Ever since there were companies, there has been an inherent curiosity and even a necessity to analyse their potential collapse. Investors, governments, customers, stock traders world-wide have vested interest in understanding the performance and taking proactive decisions based on their expectations. Throughout history, financial crises and economic recessions have repeatedly highlighted the significance of bankruptcy analysis as a means of assessing and addressing the stability of businesses and institutions. Scholars and researchers have employed various methodologies to investigate bankruptcy risks, from traditional statistical approaches (Altman et al., 1968) to more contemporary machine learning techniques (Perboli et al 2021., Son et al., 2019).

1.3. Public guarantee system:

During the COVID-19 pandemic, governmental authorities employed public guarantees as a fiscal policy tool to ease the economic predicaments confronted by enterprises lacking sufficient scale to autonomously safeguard their operations or endure the disadvantageous impacts of the prevailing global crisis. By offering such guarantees, governments sought to bolster market confidence and mitigate the adverse repercussions on vulnerable businesses, ultimately aiming to bolster overall economic stability during this unprecedented period of uncertainty.

The implementation of public guarantees during the pandemic can be traced back to the valuable insights gained from past financial crises. The urgency to avoid potential bank runs and systemic risks was paramount, especially considering the ongoing turmoil caused by pervasive food, health, and safety issues. To confront this multifaceted challenge, governments and banks collaborated to fortify their resilience, bolstered by an augmented regulatory framework exemplified by the Basel III norms. The concerted efforts of these stakeholders resulted in a notably robust global banking system that demonstrated its ability to withstand the unprecedented storm of the pandemic.

However, while the immediate crisis may have been effectively managed, the long-term ramifications remain to be ascertained. As of May 2023, the World Health Organization has deemed the pandemic no longer a global health emergency, signifying a notable milestone in the trajectory of the crisis. Nonetheless, the lasting impact on economies, industries, and societies necessitates further examination, given the complexity and scale of the pandemic's disruption. It is imperative for researchers, policymakers, and stakeholders to diligently evaluate the enduring consequences to glean meaningful insights and inform future preparedness strategies.

Government aids during the pandemic were a much-needed breath of life that most struggling companies, especially small and medium enterprises that did not have easy access to credit desperately needed. But was this enough? Was there any real difference in the growth of firms and their avoidance of bankruptcy, or was it just a delayed pulling of the plug? While these state aid programs were much appreciated, the question remains, how effective were they in the overall growth of the firms that received them? Some studies indicate that while bankruptcy filings in the European union are at an all-time high and the collapse of "zombie" firms radical, Italy's bankruptcy filings have actually declined (Arnold et al., 2023).

1.4. Fondo di Garanzia – the credit guarantee scheme:

To answer these questions, I aim to focus on the public guarantees provided in Italy, which had some of the biggest public guarantee programs for SMEs. In particular, the DL Liquidità– a public credit guarantee specifically for alleviating the distresses of SMEs affected by the pandemic. Being the first country in the European Union to be hit by the pandemic and thereafter facing one of the roughest economic hits, Italy presents a compelling subject for examination due to its notable concentration of small and medium-sized enterprises (SMEs), which accounts for a significant 95% of the country's total business landscape (OECD, 2021).

The decree called "Decreto Liquidità" established a dedicated credit guarantee fund (Fondo di Garanzia). The supported loans provide under this scheme was largely to finance working capital needs of the companies and help them recover from the impact of the crisis. The scheme also included simplified application processes, favourable repayment terms, interest rates and durations of the loans to ease the burden on these organisations. The eligibility requirements were strictly for those of small and medium enterprises with less than 50 million euros worth of assets, 43 million euros of Revenues, or 500 employees.

1.5. Research question and main contribution:

The implementation and effectiveness of this decree are subject to continued scrutiny and also forms the motivation for this research. This research will hence centre around the following question and hypothesis:

What was the efficacy of Public Guarantees in Preventing Potential Bankruptcy during the Pandemic in Italy?

Hypothesis: The provision of public credit guarantees during the pandemic effectively prevented or lowered bankruptcy rates among Italian firms.

This study question investigates whether the public credit guarantees provided by the Fondo di Garanzia during the pandemic truly helped prevent or lower bankruptcy among Italian firms. Machine learning techniques of Random Forest and XGBoost, will be utilized to examine the relationship between the utilization of public guarantees and the likelihood of bankruptcy. These techniques offer robust predictive capabilities and will help shed light on the effectiveness of the program in safeguarding firms from insolvency.

Through rigorous analysis and innovative methodologies, this study seeks to contribute to the existing literature on public credit guarantees and their implications for economic recovery. The findings hope to inform policymakers and assist firms in their decision-making processes, enabling them to navigate future crises effectively. By contributing to the existing body of knowledge in this crucial area, the study aspires to support evidence-based decision-making and help build more resilient and adaptable economies for the future.

1.6. Content overview:

The rest of this paper is organised as follows. Section 2 provides a brief overview of the literature on the evolution of bankruptcy predictions and analyses of the public guarantee system. It focuses briefly on the history of bankruptcy indicators and models used, and discusses the current literature on machine learning languages in more detail. Afterwards, Section 3 sets up and explains the data extraction process, the financial information collected or computed along with the data cleaning process. Following this Section 4 talks about the different machine learning tools employed on this cleaned data, their operative processes along with their advantages and disadvantages. Following the presentation of both models, their respective outcomes are thoroughly examined and interpreted in Section 5. Subsequently, Section 6 provides a comprehensive overview of the primary discoveries, concluding the paper succinctly. Finally, in Section 7, the study's principal constraints are elucidated, accompanied by suggestions for potential avenues of further investigation.

2. Literature review

2.1. Evolution of Bankruptcy Prediction Models and Rationale for Model Selection:

The inquiry into the fundamental reasons behind corporate bankruptcies has captivated researchers for decades. This area of study has undergone continuous advancements, with scholars introducing a wide range of models spanning from simple financial ratio-based approaches to intricate multivariate regressions utilizing cutting-edge machine learning mechanisms.

(Altman et al., 1960) pioneered the study of bankruptcy by bringing about the concept of a z-score which consisted of five predominant ratios with which to analyse the financial health and potential bankruptcy of a firm. Furthering the above thought (Ohlson et al., 1980), invented an O-score bankruptcy prediction model, embracing wider applicability (Altman's was primarily designed for Manufacturing companies). Where Altman's model focused on a single point in time, Ohlson's proved to be more dynamic encouraging updates over time to reflect the company's changes.

Hazard models became increasingly popular in the early 2000's. Models like Shumway's offered certain advantages over the predominant O-score and Z-score models. One notable advantage of this model was its incorporation of time-varying covariates, which consequently enhanced our comprehension of the evolving impact of financial variables on bankruptcy probabilities over time.

Although traditional models continue to be relevant in the present day, the emergence of machine learning techniques has revolutionized this field. This was not enough and existing models of default risk needed to be upgraded as suggested by multiple researchers (Begley, Ming & Watts, 1996). Researchers have embarked on a quest to enhance prediction accuracy and gain deeper insights into the intricacies of bankruptcy indicators.

Early ML techniques addressed several limitations of the traditional models and surpassed them. The first researches using Machine learning started with Artificial Neural networks (ANNs) (1990s) and Support Vector Machines (early 2000s). One of the main advantages that these models have over the previous non-ML models is their ability to capture complex non-linear relationships among variables, their robustness to outliers and noise in the data and introduction of automated feature selection methods.

Despite the progress made in the aforementioned methods, certain crucial aspects remained lacking. Researchers continued to seek solutions to enhance prediction accuracy and develop advanced models capable of effectively capturing complex relationships. Numerous studies conducted across various fields have compared the performance of different machine learning (ML) techniques. The majority of these studies concur that ensemble model techniques exhibit significantly superior prediction accuracy when compared to simple logistic regressions. Empirical findings have found that machine learning models are approximately 10% more accurate compared to the traditional models (Acharya et al., 2017).

Ensemble models, specifically the classification models can be divided into two main categories: bagging and boosting. Bagging is a homogenous model that learns from each iteration parallelly and then uses this knowledge to arrive at the model average. Random Forest is one of the most popular bagging techniques (Brieman., 2001). It is very precise (Kruppa, Schwarz, Arminger, & Ziegler, 2013) and usually performs much better than the previous SVMs or logistic regressions (Choi, Son, & Kim, 2018).

In the recent past, another ensemble method that has become increasingly popular and is widely used in statistical research is the boosting technique - Gradient boosting (Perboli et al., 2021). Gradient Boosted Machines (GBMs) and their various iterations, provided by multiple communities, have garnered significant traction in recent years. This increased interest can be attributed to the enhanced performance offered by decision trees when compared to alternative machine learning algorithms especially in its real-world applications.

There are three main types of gradient boosting – CATboost, LightGBM and XGBoost. CATBoost is specifically designed to handle categorical variables better than the other models. LightGBM is known specifically for its superior speed. It uses a technique called "Gradient-based One-Side Sampling" (GOSS). GOSS is a sampling technique specifically used with datasets that contain too many samples. XGBoost is useful to create a strong prediction model. It does this by leveraging on the weak learners and using regularization techniques. It is also superior in speed but not as much as LightGBM. There have been studies that analyse how useful each of these models are in financial and economic research during the pandemic (Papík et al., 2023). The studies show that XGboost outperforms other classification models significantly. This led to the choice of this ML as one of the models to be used in this thesis.

A number of these studies as mentioned above, have recommended both random forest and gradient boosting, specifically XGboost for bankruptcy prediction. In this vein, both these models are used in our analysis to gain advantages of the comparable predictions as well as increased interpretability.

2.2. Machine learning models recent studies:

The data collected for the purpose of this research will be comparable to data used in the paper Core et al 2019. This study aims to apply in part, the methodology used in Perboli et al., (2021) for bankruptcy prediction using machine learning analysis. The deviations primarily occur in the time period under analysis. While Perboli et al., (2021) use a mid to long term prediction period (60 months), this study aims to find the short run effects of the pandemic on the firms that had received the public guarantees. The primary reason for choosing a shorter timeline was that wanted the dataset that used for training has to be unaffected by financial crises which may bias the dataset and thus project companies to be underperforming as compared to a normal period also leading to skewed bankruptcy predictors in the model. Hence, it was crucial to identify companies that would go bankrupt in a normally functioning period of time, as only then this can this comparably be applied on the dataset post pandemic to extract meaningful predictions.

3. Data:

The data was primarily collected from ORBIS Bureau van Dijk. ORBIS is an extensive database that contains information on the financials and company specific information on organisations across the world. The database contains historical information on the financial statements, company structure, size, years of operation, status (bankrupt/ active) and other such relevant company-specific information.

3.1. First level data collection:

The dataset for the training data had to have characteristics that were similar to the final data on which we were using the trained model. The loan recipients under the DL Liquidita scheme had to be those Italian companies that had utmost 500 employees or 43 million Euros of total assets or 50 million in revenue in order to qualify for a loan. Hence these exact filters were applied to the search operator on the ORBIS database.

At the first stage we wanted to train the model to understand the features of bankrupt companies. Hence financial information of those 40 ratios mentioned above were extracted for companies that were inactive – bankrupt. This data was extracted for a period of two years before the company actually went bankrupt so the model could understand the trend leading up to bankruptcy for these companies. Each set was treated as individual firm year observations.

Financial information for 40 variables (appendix 1) containing ratios falling in 3 main brackets: Profitability, operational, and structure ratios. Ratios that were still relevant but not fitting within the scope of these brackets were classified as a separate 'others' list. As mentioned before, using financial ratios has had the most success in predicting bankruptcy across different methodologies over the years.

At a bird's eye view, the idea is to train the model on bankrupt and active companies during the years 2015-2018. For example, for companies that were bankrupt during the year 2019, financial data for the two years before bankruptcy (2018 and 2017) were extracted from ORBIS and the ratios were calculated. Likewise for companies bankrupt during the year 2018, information with respect to 2017 and 2016 were collected and ratios were calculated. The same was performed with companies bankrupt during 2016. All these data were then combined, and the data with respect to one year before bankruptcy were named "Yr1" and data two years before bankruptcy were named “Yr2”. Now, the complete file has data of companies that went bankrupt from 2016-2018 with financial information pertaining to one year before and two years before bankruptcy.

To ensure that the companies classified as bankrupt in the above were bankrupt due to legitimate causes, certain filters were applied to ensure that the companies had at least 5 official years of financial information. After this, a separate extract of active companies was made from ORBIS. The filters were the same as before, except now the status filter was specified as “active companies”. The time period was also comparable to the previous extract. There were 197401 companies in this set.

XGBoost can handle unbalanced data, but with 3000 bankrupt companies and close to 200,000 active companies the data appears to be highly unbalanced. Random Forest is not as good at handling unbalanced data, and if the model was executed on this dataset the results would be spurious. Hence as suggested in Perboli et al., (2021), 6000 companies were randomly sampled from the larger population of 200000 companies. While this does sacrifice precision, it helps with recall. This step is also important because it is essential for the model to understand what the characteristics of a bankrupt company are, so that it can accurately help make predictions.

Now the final data set has 8906 companies.

Year of bankruptcy	T-1 year from bankruptcy (Y1)	T-2 year from bankruptcy (Y2)	Total firm-year observations
2016	855	855	1710
2017	1206	1206	2412
2018	845	845	1690
Total	2906	2906	5812

Table 1: summary of bankrupt companies extracted year-wise.

3.2. Second level data collection:

Then at the second level, data for the years 2021-2022 was extracted, the trained model from the previous data set was applied and then the results were evaluated. At the end of the year 2019 was when the pandemic started accelerating world-wide. Hence this was not an appropriate year to use for training the data. The entire analysis at all levels was in the short term – 2 years. Here we deviate from Perboli et al, 2021, that performed a mid to long run prediction since this analysis aims to predict primarily the short-term effects of the pandemic. The subsequent paragraphs explain the collection processes at each level in detail.

At the second level we needed data from ORBIS post the DL liquidita scheme (post April 8th 2020). Our main requirement is to check if the companies that received the loans had a lesser potential to go bankrupt than similar companies that did not receive the loans. Hence a random sample of companies' financial information for the years 2020, 2021 and 2022 was picked from ORBIS. The tax Identification number (TIN) was also extracted as part of this information.

The Fondo di Garanzia website provides month-wise information on all the companies that had received loans under various schemes. This information was extracted from the month April 2020 till the month of June 2022 (till the end of the program). The data for the companies that received loan specifically under the DL scheme were filtered and combined into one single file for ease of use. After collecting and cleaning the two sets of data above at the second level, the data from ORBIS was then matched to the data from FG through the Tax identification number to analyse which of the companies in the random sample had received these DL loans and which had not. Once this data was complete, we could now run the model on these companies to further analyse the impact of these loans by comparing the performance of the companies that had received them with the companies that did not.

3.2.1. Cleaning and normalising the data:

Data from ORBIS is not complete, hence has some missing values. As suggested in Perboli et al., 2021 the missing values were replaced with zero and standard scaling was applied. Additionally, to handle the sensitivity of the variables to outliers, all data was winsorised at 1% which is commonly accepted practice.

Each data set, once extracted was cleaned. ORBIS gave outputs labelled n.s. to identify those values that were not significant or close to zero. These data points were all filtered and changed to zero to enable further calculations of the financial ratios. Each financial ratio was then individually calculated using these datapoints (please see appendix for the ratios and their formulae), for each of the firm-year observations. Multiple rounds of check were employed to ensure accuracy in calculations with respect to the formula and the year for which the calculations were made, as these would form the base for all further analyses. The formulae employed were generally accepted formulae and in common use.

The database often limited the extraction of data to a certain number of companies due to size restrictions for data export. Hence parts of the data were extracted in multiple runs, and then finally combined together in one file for ease of analysis.

3.3. Creation of a “superior” dataset:

The main problem with the current set up is that the active companies list that has been compiled could ideally also contain certain companies that could potentially go bankrupt in the near future. To negate this effect on the accuracy of our prediction, we used a 2-step procedure.

Once the first model was trained and tested on the set of active and bankrupt companies mentioned above, the classifier also provided probabilities for each individual company to go bankrupt. The study needed better classification probabilities for prediction, as the normal 50% benchmark left too much scope for ambiguity. Hence, a threshold of 60% (Perboli et al., 2021) was taken. Companies that had a greater than 60% threshold of being active across both years were picked deemed to be the “superior” active companies that had better financial statements and much less likely to go bankrupt in the near future.

This threshold yielded 100000 companies out of the 197401 companies to be active.
6000 companies were again sampled from this set and the machine learning models of XGBoost and Random Forest were re-run on this set combined with the set of bankrupt companies.

4. Methodology:

Once the cleaning process was complete a machine learning model was employed on the cleaned dataset to, in essence, classify the companies into the categories - bankrupt or not bankrupt. The financial ratios will be used as the main features of the companies. The following paragraphs seek to describe the models, and explain their advantages and disadvantages.

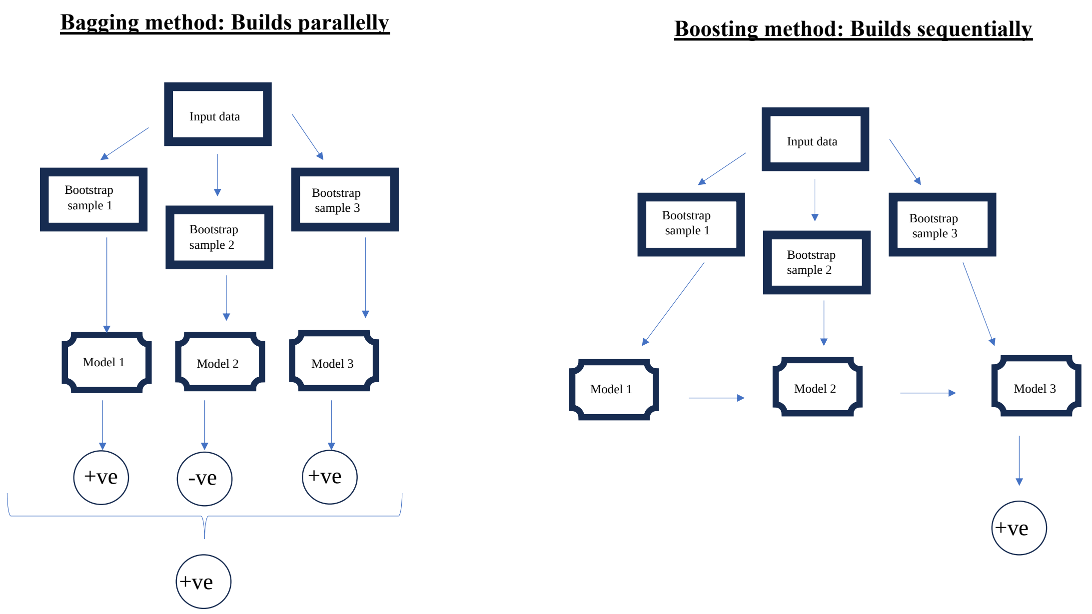


Figure 1: Comparing random forest model process with XGBoost model process.

4.1. Random Forest model:

The Random Forest algorithm is a well-known ensemble bagging technique that aims to improve predictive accuracy by combining the insights from multiple decision trees. Each decision tree is trained on different subsets of the dataset, allowing for a diverse collection of models (Brieman, 2001).

Ensemble Bagging Technique: Random Forest is a method that integrates multiple decision trees to fine tune the forecasting power of any individual tree. By aggregating the insights from these diverse decision trees, the algorithm aims to make more accurate predictions. In this study, the classifier would be used to predict companies in a binary classification of 1 (bankrupt) and 0 (non-bankrupt).

Decision Trees: The Random Forest algorithm utilizes decision trees as its base models. By using multiple decision trees, the algorithm can capture different aspects of the data and reduce the risk of bias. Using hyperparameters (such as max depth), the length of each tree will be fine-tuned to improve prediction accuracy (See appendix 2 for sample random forest decision tree).

Majority Voting Mechanism: After training the individual decision trees, the Random Forest algorithm combines their predictions through a majority voting mechanism (Brieman, 2001). Each tree in the forest produces its own prediction, and the final output is determined by the majority vote among all the trees. If a majority of the trees classify a company as bankrupt then the final output will predict the company to go bankrupt. This bolsters the robustness of the overall process and improves the quality of output.

Overfitting Mitigation: Random Forest addresses the common concern of overfitting. This is particularly advantageous in this study as the model is fed 39 features, all of which would not be relevant for the final predictive power of the model. By incorporating a substantial number of constituent trees in the forest, the algorithm reduces the risk of overfitting. The diversity in the training subsets and the aggregation of predictions helps to avoid the risk that the model may overly rely on its ability to predict the training data thus reducing its power to predict the test dataset or other similar datasets.

4.2. XGBoost the model:

XGBoost (Chen & Guestrin, 2016) simply expanded means extreme gradient boosting. Boosting is a technique that combines weak learners (poorly predicting variables) to create a strong learner with superior predictive capabilities. Boosting is an iterative ensemble learning technique. In each iteration, the algorithm gives more attention to data instances that were previously misclassified or had higher errors. These challenging instances are assigned higher weights, making the subsequent weak learner focus on correcting the mistakes made by its predecessors. This iterative process continues until the models collectively generate a strong learner with improved predictive accuracy.

XGBoost is a decision-tree based ensemble machine learning algorithm (Chen & Guestrin, 2016). The algorithm performs this process using a gradient boosting framework.

At its core XGBoost operates by employing numerous decision trees for training. By leveraging this ensemble of trees, the predictions generated by each tree are effectively amalgamated to produce the ultimate and more accurate prediction. Initiated earlier in the decade (Chen, 2014), now it has contributions from various developers.

XGBoost has 3 main advantages.

4.2.1. Missing values:

ORBIS database is not complete in the data it provides. There are missing values that have to be considered when downloading financial information. Only variables such as a firm's status do not contain missing values (owing to our filters), the rest of the 78 firm-year observations that we extract to make our computations for the ratios will have missing values. The usual methodology is to set the missing values to zero and include the variables in the sample dataset instead of completely excluding them. However, XGBoost improves this.

To shine a better light on this, for example, if there is a parent node A (100 inputs) which has 10 missing variables, splitting into child nodes B and C. The 90 inputs (excluding missing values are split into nodes B and C accordingly. The treatment of the other 10 values is based on the combination that gives the highest gain score, that is, the model imputes these missing values (based on the 90 existing values) and then allocates them to B and C in a manner that provides the best predictive power. This is especially useful and more accurate than replacing missing values with zero.

4.2.2. Superior processing speed:

While the number of rows were not considerably large (maximum 20,000 samples), the model had to process 100+ financial ratios for each of these firm-year observations. Other machine learning models take long processing times at different stages of the process.

For example, support vector machines (SVMs) take a significantly long time to train, while K-nearest neighbours (KNNs) don't take a long time to train but requires more time to execute.

As compared to these other MLLs XGBoost handles large datasets with ease, uses advanced regularisation and can be parallelized to exploit computational capabilities of the hardware.

XGBoost employs tree boosting operations through the implementation of a Regularized Learning Objective while utilizing a Shrinkage Factor (also referred to as the learning rate or eta). The application of Shrinkage Factor serves to diminish the influence of each individual tree, thereby creating room for subsequent trees to exert a more substantial impact on the overall model performance. This in turn speeds up the processing capabilities of the model.

4.2.3. Feature reduction:

In this analysis 39 variables/ ratios have been used, that may be helpful in the prediction of bankruptcy of companies. But not all these variables would be effective, some may even be misleading. In a simpler ML analysis this would mean lower prediction accuracy, but XGBoost handles this with the help of its dimensionality reduction techniques.

There are multiple dimensionality reduction techniques that can be used, and each depends on the attributes of the dataset in use.

The first such technique that was to be applied to this study was Principal component analysis. This is used in principle for those datasets where the variables have a lot of variation within them. We found that most of the features used did not have much variation, as is expected since they are all supposed to be similar SME companies within the same financial constraints. Principal component analysis is also best used to describe a dataset, rather than to interpret it. Hence this was not considered to be the most appropriate for this study.

The second dimensionality technique applicable for this study was feature selection. This refers to a technique of selecting the most relevant features for a dataset. The model intuitively searches for those features that provide the largest predictive power. This results in two advantages. Firstly, it reduces the noise/ misleading variables that contribute little to zero additive improvements to the model. Secondly as less important features are trimmed off, the model simplifies thus reducing the risk of overfitting.

There are two main types of feature selection methods. The first is a wrapper-style selection method where the algorithm wraps the existing algorithm and then assists in selecting the best features. The second is a filter-style selection method where each feature is scored, and then the features with the largest or the smallest score (depending on the study) are chosen or filtered.

There are multiple such feature selection algorithms that can be employed. Recursive feature elimination was deemed to be most suitable as it combines the wrapper-style selection methodology with the filter-style and hence is an efficient, commonly used approach to remove undesirable features from the training dataset.

Feature reduction is one of the most important advantages of this model as it increases interpretability of the model, aids in understanding the real factors that affect the dependent variable and helps in arriving at conclusions that probe more into the 'why' of an issue, rather than acting as a simple classification of a yes or a no.

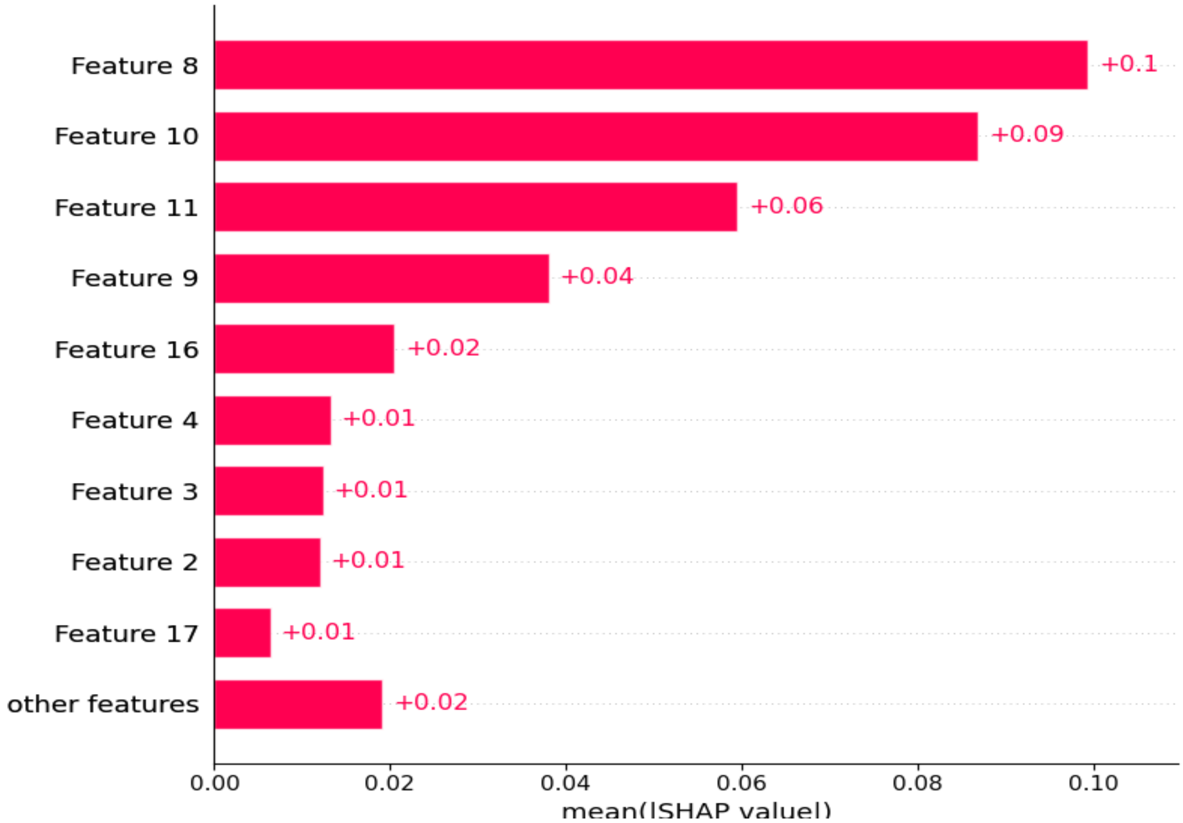


Figure 2: Feature importance – evaluation of features to select the top features and thus reduce irrelevant ones.

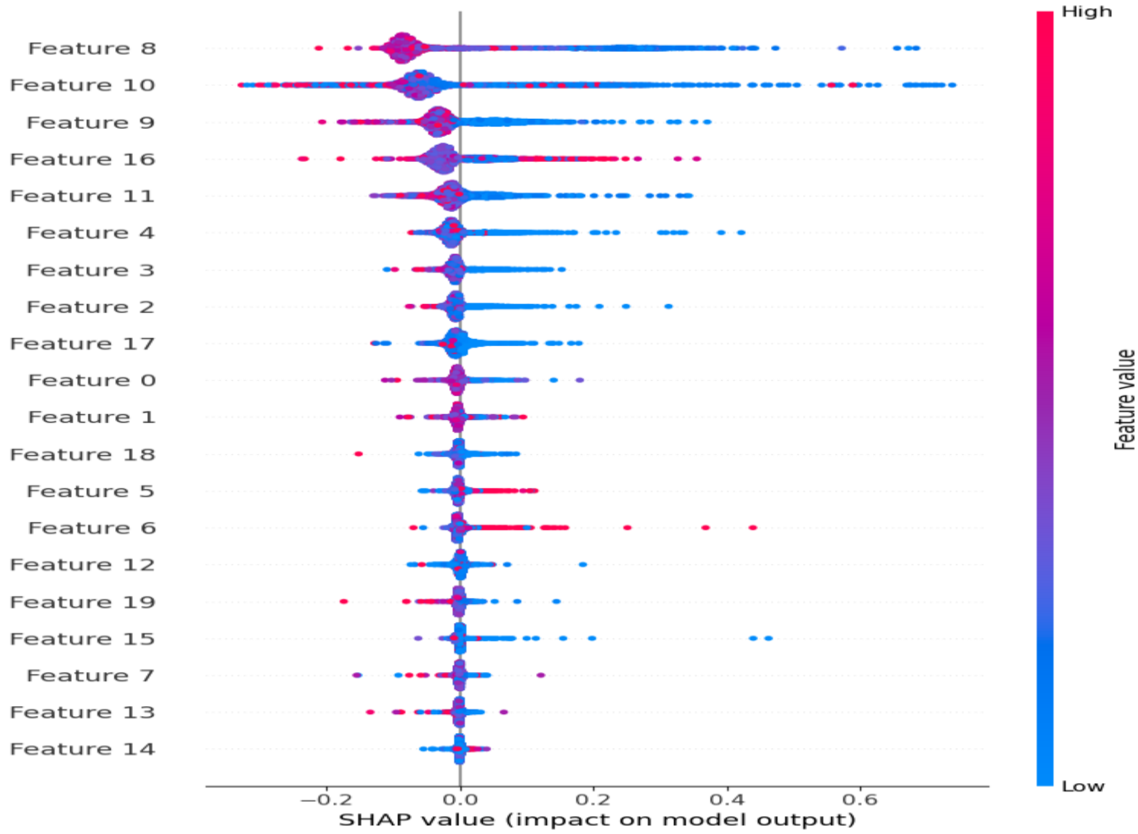


Figure 3: SHAP output – From the SHAP output it is noticed that feature 8 has the highest predictive power amongst all features.

While XGBoost appears to be a compact robust model, it does face some disadvantages:

4.2.4. Black box models:

A general definition for a black box model is one where useful information is obtained as output, but the process itself is not transparent and these workings cannot be revealed.

It is often argued that XGBoost is not entirely a blackbox model. Feature reduction provides valuable insights and good interpretations of how this learning rate works. There have been studies that have recently proven that this ML is completely not a black box model and explains how the predictions of XGBoost works (Carmona et al., 2022)

4.2.5. Sensitivity to outliers:

XGBoost's sensitivity to outliers arises from the inherent obligation of each classifier to correct the errors made by its predecessors, which consequently renders the model excessively reliant on the presence and influence of outliers. This study seeks to solve this problem through the technique of winsorising. The data provided as input is winsorised at 1% before it is fed into the XGBoost, thereby limiting the extreme values within this threshold. This successfully reduces the impact of these outliers on the model and allows for a more accurate prediction.

4.3. Hyperparameters:

Hyperparameters refer to the manually adjustable settings or configurations of a machine learning algorithm or model, and their optimization is performed externally to the algorithm itself. Furthermore, as the efficacy of a machine learning algorithm or model increases, the number of hyperparameters requiring manual tuning also tends to grow correspondingly.

The hyperparameters used to regularise the Random forest model are:

- **Max depth:**

Since in both models we are dealing with decision trees, there arises a question as to how big a tree can be allowed to grow. This is where Max depth assists. This is the maximum number of child nodes that a parent node is allowed to grow, until the tree is cut off.

- **Min impurities decrease:**

There should be checks on the quality of the splits being made by the machine learning models. If left unchecked a decision tree could keep building with nodes unnecessarily, and often irrelevant to the prediction accuracy. Hence one such check is the hyperparameter that checks the minimum impurity decrease that is achieved by a node. Once preset, a node is allowed to split only if the impurity decrease levels specified are met.

Example: In both Random Forest and XGBoost, the hyperparameter "min_impurity_decrease" can be set to control the threshold for node splitting or further partitioning. For instance, "min_impurity_decrease = 0.05" means that a node will only be split if the impurity decrease resulting from the split is greater than or equal to 0.05.

- **Min sample:**

The ‘min sample’ hyperparameter is a good check to ensure that a node does not hastily compute based on lesser number of samples thereby skewing the output. Instead, the minimum samples can be specified for a node to meet, post which only can it be split.

For example, "min_samples_split = 5" means that a node can only be split if it contains at least 5 samples.

- **N estimators:**

  Random Forest: The count of decision trees to create in the ensemble.

  XGBoost: The count of boosting rounds, i.e., the number of weak learners to combine in the boosting process.

Example: As mentioned earlier, "n_estimators" is the hyperparameter that controls the count of trees in Random Forest and the count of boosting rounds in XGBoost. For instance, setting "n_estimators = 100" means that the Random Forest will consist of 100 decision trees, and XGBoost will perform boosting with 100 rounds, combining 100 weak learners.

For the XGBoost model Max dept and N estimators were used. Apart from these, the following additional hyperparameters were used as well:

- **Gamma:**

  The minimum loss reduction required to make a further partition on a leaf in a weak learner (usually decision tree) during the boosting process.

  Example: The hyperparameter "gamma" can be set in XGBoost to control the threshold for making further partitions on leaves. For example, "gamma=0.1" means that a further partition will only be made if it leads to a loss reduction of at least 0.1.

- **Minimum child weight:**

  This hyperparameter adds the weights for the child node and then checks what is the smallest value required for the child node to further split during the boosting process.

  Example: In XGBoost, hyperparameter "min_child_weight" can be set to determine the lowest total of weights needed for child node to perform further partitioning. For example, "min_child_weight = 3" means that a child node must have an accumulated weight of at least 3 to be eligible for further splitting.

- **Learning rate:**

In XGBoost, the hyperparameter "learning_rate" can be set to control the step size at each boosting round. For instance, a smaller learning rate, like "learning_rate = 0.1," means that each weak learner's contribution will have a smaller impact, resulting in a more conservative learning process.

4.4. Log loss Function:

The risk of a company being predicted as not bankrupt, when it is factually bankrupt (a false classification), threatens the accuracy of the predictive power. To negate this effect, log loss functions are used. By levying extra penalties on the mistakes, that is, the wrong classifications, the log loss function helps develop the model better. In order to execute this, the log loss function reworks the existing prediction with an extra weight. These functions are extremely popular when used for binary classification such as the bankruptcy predictions of 0 or 1 used in this study. The log loss function is expected to decrease with increased accuracy (Perboli et al., 2021). In this study, the closer the firm-year observation is to bankruptcy, the more the decrease in the log loss function. For example: the Year 1 observations which are expected to have more accuracy as compared to Year 2 observations (further from bankruptcy), should have reduced log loss functions as compared to Year 2.

4.5. Comparison and choice of model:

The first major difference between random forest and XGBoost is in their approach to model optimization. XGBoost places a higher emphasis on refining the functional space to reduce the cost of the model, whereas Random Forest prioritizes the optimization of hyperparameters.

The second difference lies in the fact that Random forests are more amenable to distributed computing than XGBoost. What this means is that to train a Random Forest, the dataset can be divided into smaller subsets, and individual decision trees can be trained on each subset simultaneously across different computational nodes. Once all the trees are trained, the final prediction for each image can be efficiently calculated by aggregating the results of the trees.

However, boosting algorithms would train weak learners iteratively, and each learner's training heavily depends on the outcome of the previous one. This dependency limits the scope for parallel processing, as each iteration needs to be completed sequentially before moving on to the next one.

Thirdly, XGBoost performs much better than random forest for unbalanced datasets such as ours. When our sample of active firms is much higher than that of bankrupt companies in the training dataset, this is where XGBoost assists better than random forest.

In XGBoost, the model addresses difficulties in predicting anomalies by giving them greater attention and importance during subsequent iterations. This iterative emphasis enhances the model's capability to predict classes with low representation. On the other hand, random forest does not guarantee a systematic approach to handling class imbalances. This means that it may not effectively handle imbalanced datasets where certain classes have significantly fewer instances.

From the above we see that XGboost and Random forest yield their unique advantages to the study and it may not be appropriate to choose one single model to analyse our data. Hence this study utilises both machine learning classification techniques to observe which model gives better precision and recall. It is also commonly felt that the unique findings from each model could enhance the interpretability of the data than by using only one of them.

5. Results:

5.1. Results of Round 1 and Round 2:

Round 1:

Classifier	Brier loss	Log loss	Roc auc	Precision	Recall	F1
RF-Y2	0.052	0.183	0.974	0.930	0.755	0.834
RF-Y1	0.039	0.148	0.982	0.961	0.800	0.873
XGB-Y2	0.0640	0.250	0.967	0.926	0.747	0.827
XGB-Y1	0.0531	0.220	0.974	0.949	0.795	0.865

Table 2: Metrics across all models for round 1.

Round 2:

Classifier	Brier loss	Log loss	Roc auc	Precision	Recall	F1
RF-Y2	0.020	0.085	0.979	0.979	0.923	0.950
XGB-Y2	0.027	0.141	0.987	0.978	0.902	0.939
RF-Y1	0.019	0.079	0.985	0.977	0.909	0.942
XGB-Y1	0.027	0.143	0.985	0.981	0.900	0.938

Table 2: Metrics across all models for round 2.

In this section the results of the model will be discussed. The model results and accuracy will also be compared to that of Perboli et al., (2021). The performance is measured with the help of metrics Brier score, AUC, F1, precision and recall.

In this study, the XGBoost model achieved accuracies comparable (but slightly better) to the Random Forest model. Its AUC is higher than random forest across all models by 1-2%. These results are comparable to that of Perboli et al., (2021). The increasing AUC, F1, precision and recall scores year on year are also comparable to Perboli et al., (2021) that shows similar trends for its metrics suggesting that the models are able to better predict one year to bankruptcy than 2 years to bankruptcy. The longer the duration the lesser the scores and accuracy thus adding to the general consensus that it is more difficult to predict in the longer term due to the presence of omitted variables.

The entire reason for implementing the two-rounds data set creation is because the majority class of active companies also contain some companies that may go bankrupt in the future. To avoid these challenges, the two-rounds data set creation was introduced. From the results its clearly shown that implementing this procedure, improves the accuracy score by 1-2%. This is also comparable to Perboli et al., (2021) where the AUC metric increased by 7%. Our metrics already had a higher accuracy than the paper compared by 6-7%. Hence the improvement from the first round to the second is smaller as the accuracy increases, but the increase itself is still comparable. The same results as the AUC are confirmed in F1, precision and recall with a general trend of increased scores even when individual years are compared.

5.2. Application of the models on a real dataset of Italian companies post 2020:

The intent of this entire exercise was to accurately predict the future of those companies that received the public guarantee loans and compare them to those that did not receive these guarantees. Now that the models concerned have an average accuracy greater than 95%, the models are deemed fit enough to be applied to the dataset of Italian companies post 2020 for bankruptcy prediction. This study has applied the XGBoost round 2 model on this dataset, as the model has better metrics than the random forest models.

Additionally, the study ensures that both the models of XGBoost are considered, the one year to bankruptcy model as well as 2 years to bankruptcy model. This is done by performing a simple average of the probabilities for bankruptcy for the results enumerated below.

Risk of bankruptcy	Prob <=50%	50% <Prob < =70%	Prob> 70%
Did not receive loan Y1	80.37%	11.42%	8.21%
Did receive loan Y1	78.93%	13.14%	7.93%
Did not receive loan Y2	77.24%	8.43%	14.33%
Did receive loan Y2	72.81%	12.93%	14.26%

Table 3: Italian companies bankruptcy segregated into probabilities below 50%, between 50% and 70% and greater than 70% for year1 and year 2 XGBoost models. This is then bifurcated into companies that received loan and didn't receive loans.

From the above we see that at very strong probability of 70% and above of the companies going bankrupt, the firms that did receive the loans did marginally better. But when we look at the whole picture, and check for a greater than 50% probability of firm bankruptcy, the firms that did receive the loans appear to be worse off than the firms that did not.

For the model evaluating T-1 to bankruptcy, we see that the firms that did receive the loans have a 21.07% probability of going bankrupt, whereas the firms that did not receive the loans have 19.63% probability.

Likewise, for the model evaluating T-1 to bankruptcy, we see that the firms that did receive the loans have a 27.19% probability of going bankrupt, whereas the firms that did not receive the loans have 22.76% probability, which is significantly lower.

We also see that the trends shown in Perboli et al., (2021), where the closer the firm-year observation is to the actual year of bankruptcy, the higher is the accuracy percentage, thus confirming that the predictions are better and more accurate in the short term.

Industry	Received loan	Prob <=50%	50% <Prob < =70%	Prob> 70%
	Not received	67.80%	16.71%	15.49%
Agriculture	Received	68.75%	15.63%	15.63%
	Not received	70.12%	12.46%	17.42%
Construction	Received	81.48%	10.82%	7.70%
	Not received	87.43%	6.04%	6.53%
Manufacturing	Received	77.20%	12.07%	10.72%
	Not received	84.00%	8.39%	7.60%
Wholesale	Received	79.97%	12.50%	7.53%
	Not received	76.40%	11.21%	12.39%
Retail	Received	70.92%	14.73%	14.36%
	Not received	78.26%	10.42%	11.31%
Services	Received	77.15%	12.55%	10.30%
	Not received	71.29%	11.78%	16.93%
Transport	Received	67.78%	15.01%	17.21%
	Not received	83.97%	8.26%	7.77%
Others	Received	75.89%	14.61%	9.50%

Table 4: Probabilities of firm bankruptcy across sectors classified based on companies that received the loans and companies that did not receive the loans.

In Table 4, we see the split of the results based on the industry that the companies operate in. Here it is clearer to see that there is a marked change from industry to industry. The industries of construction and services see a positive effect in the short term for companies that received the loan. Evaluating based on the ‘companies>70%’ bankruptcy column, the companies in the Construction industry, that did receive the loan have a significantly lower percentage bankruptcy of 7.7% as compared to 17.42% for the companies that didn't. Similarly, in the services industry the companies that did receive the loan have a lower percentage bankruptcy of 10.3% as compared to 11.31% for the companies that didn't.

For Agriculture, manufacturing, retail and transport industries we see that this does not hold. Companies that received the loans are disadvantaged as compared to the companies that didn't in these sectors. The biggest differences can be seen in the Manufacturing and retail sectors where the companies that received the Public guarantees were significantly more at risk of bankruptcy by 4.19% and 2% respectively.

Others is a miscellaneous classification of all remaining industries and here too it is seen that public guarantee firms did not do better. If the >50% probabilities were considered too, for brevities' sake, it is seen that apart from the construction industry all other industries did worse.

Name of region	Prob <=50%	50% <Prob < =70%	Prob> 70%
Abruzzo	73.24%	12.89%	13.87%
Basilicata	82.26%	9.68%	8.06%
Calabria	76.33%	13.00%	10.67%
Campania	77.35%	10.89%	11.76%
Emilia-Romagna	78.71%	10.53%	10.76%
Friuli-Venezia Giulia	84.80%	7.55%	7.65%
Lazio	71.90%	13.72%	14.38%
Liguria	72.20%	13.17%	14.63%
Lombardia	80.05%	9.98%	9.97%
Marche	78.01%	9.07%	12.92%
Molise	74.14%	13.79%	12.07%
Piemonte	79.15%	9.54%	11.31%
Puglia	72.31%	12.31%	15.38%
Sardegna	80.00%	7.78%	12.22%
Sicilia	73.41%	13.21%	13.38%
Toscana	77.30%	9.89%	12.81%
Trentino-Alto Adige	80.10%	9.24%	10.67%
Umbria	75.32%	13.29%	11.39%
Valle D'Aosta	70.59%	20.59%	8.82%
Veneto	81.21%	9.66%	9.14%

Table 5: Probabilities of firm bankruptcy across regions classified based on companies that received the loans and companies that did not receive the loans.

Table 5 provides a comprehensive overview of the results of the bankruptcy prediction across regions. In the short term, we see that Puglia, Lazio and Abruzzo have the highest expected bankruptcy rates (probability >70%), while Valle D'Aosta, Friuli-Venezia Giulia and Basilicata have some of the lowest predictions for bankruptcy.

5.3.Comparison with Altman's Z score:

Edward Altman, the pioneer of the Z-score model, recommended a z-score that could be applied when evaluating bankruptcy of companies (Altman et al., 2012). The z-score has been calculated to verify and to serve as a further check if the bankruptcy predictions computed via the models are in line with the industry accepted norms.

Risk of bankruptcy	XGBoost>50% prediction	Altman Z score
Did not receive loan Y1	19.63%	25.54%
Did receive loan Y1	21.07%	23.27%
Did not receive loan Y2	22.76%	25.09%
Did receive loan Y2	27.19%	25.51%

Table 6: Comparison of z-score predictions with predictions from the XGBoost model for both one year to bankruptcy and two years to bankruptcy.

From the above it is surmised that the Altman z-score predictions are comparable to that of the machine learning model. There is a marginal difference between the companies that received the loans and the companies that did not receive the loans favouring the companies that did receive the loans. While the z-score deviates here from XGBoost, the deviations are marginal. Furthermore, the data is unbalanced, and skewed more towards the active companies. While XGBoost allows for such unbalanced data, this being one of its primary advantages, Altman's z-score is a simple weighted average predictor and will have to be adjusted for such discrepancies. On the whole, the average values across year 1 and year 2 for companies that did receive the loan are similar across both predictors, thus bolstering the predictions of the XGBoost model.

5.4. Comparison with pre-loans data:

It is important to compare the results of the model with data from before the receipt of the loans. This helps establish the situation that the companies that received the loans were in before the said receipt. For this, data for the same companies taken for 2021 and 2022 that received the loans were used. The same models used for year 2 (slightly longer term than year

1, hence better for data comparison) were then used to compare data for the years 2019 (the years prior to the receipt of the loans). This was done to verify the position of the companies after the impact of the pandemic and before the receipt of loans. It was identified that 32.7% of the companies that received the loans (in 2020) would have gone bankrupt in the year 2019. This bankruptcy rate came down to 27.2% in the year 2021. This goes to show that the hypothesis holds. If the companies had not received these loans in the year 2020, they would have been more susceptible to bankruptcy. From the above data, it is plausible to surmise that the public guarantees reduced the probability of bankruptcy for these companies.

6. Conclusions:

Overall, the findings and model outcomes have many similarities to the findings of Perboli et al., (2021). The trend of the metrics, the trends year on year as well as the applications to the real case study are all comparable.

In conclusion, the findings of this study highlight a cause for concern regarding the performance of companies that received public guarantees from Fondo di Garanzia, as they have demonstrated inferior performance and higher susceptibility to bankruptcy compared to companies that obtained loans. The possibility that companies opting for loans were inherently in a weaker financial state, resulting in extended recovery periods and bankruptcy-like characteristics, offers a potential explanation for these outcomes.

It is important to emphasize that while the companies receiving loans were indeed considered worse off, the disparity between them and the non-recipients was only marginal. Consequently, there is a plausible belief that, over the long term, these firms may have the capacity to recover. This suggests that the public guarantee system should not be regarded as an abject failure, but rather as a mechanism that may have prevented more severe financial distress for the recipient companies, enabling them to continue functioning.

Nevertheless, a comprehensive understanding of the long-term effects of these guarantees on the recipient companies necessitates further investigation. Subsequent research and analysis are imperative to provide a holistic assessment of the overall impact and implications of the public guarantee system. Only through such a thorough examination can a more informed evaluation of the system's effectiveness be achieved, and potential areas for enhancement be identified.

7. Limitations of this research:

7.1. Timing:

This analysis of the effects of the guarantees during the pandemic, is taking place just 3 years after the crisis. This may not be completely indicative of the effects of these credit guarantees, and further research may be needed in a few years to test if the results of this analysis remain valid.

7.2. Overfitting:

Machine learning techniques do have a tendency of overfitting the data. This was mitigated farily in this research by ensuring that the hyper parameters, especially for XGBoost which has a tendency to quickly overfit the model, are set rather conservatively. Despite these checks in place, overfitting is a common problem with respect to machine learning models and a potential disadvantage of using them.

8. References:

Zecchini, S., & Ventura, M. (2007). The impact of public guarantees on credit to SMEs. *Small Business Economics*, 29(1-2), 83-99. https://doi.org/10.1007/s11187-007-9077-7

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. Retrieved from https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf

Bertoni, F., Colombo, M. G., & Quas, A. (2013). The long-term effects of loan guarantees on SME performance. *Journal of Banking & Finance*, 37(5), 1865-1881. https://doi.org/10.1016/j.jbankfin.2013.01.023

Gonzalez-Uribe, J., & Wang, S. (2020). The Effects of Small-Firm Loan Guarantees in the UK: Insights for the COVID-19 Pandemic Crisis (FMG Discussion Paper No. DP795). *Financial Markets Group, London School of Economics*. Retrieved from https://www.fmg.ac.uk/sites/default/files/publications/DP795.pdf

Core, F., & De Marco, F. (2021). Public Guarantees for Small Businesses in Italy During COVID-19. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3604114

Edward Altman (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance,* 23(4), 589-609. doi:10.1111/j.1540-6261.1968.tb00843.x

H. Son, C. Hyun, D. Phan, H. J. Hwang (2019). Data analytic approach for bankruptcy prediction. *Expert Systems with Applications*, 131, 120-131. doi:10.1016/j.eswa.2019.07.033

Mario Papík, Lenka Papíková (2023). Impacts of crisis on SME bankruptcy prediction models' performance. *Expert Systems with Applications,* 185, 119072. doi:10.1016/j.eswa.2022.119072

Flavio Barboza, Herbert Kimura, Edward Altman (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 77, 259-271. doi:10.1016/j.eswa.2017.04.006

Elena Fedorova, Svetlana Ledyaeva, Pavel Drogovoz, Alexandr Nevredinov (2022). Economic policy uncertainty and bankruptcy filings. *International Review of Financial Analysis,* 85, 102174. doi:10.1016/j.irfa.2022.102174

James A. Ohlson (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1), 109-131. Retrieved from https://www.jstor.org/stable/2490395

Joy Begley, Jin Ming, Susan G. Watts (1996). Bankruptcy classification errors in the 1980s: An empirical analysis of Altman's and Ohlson's models. *Journal of Accounting, Auditing & Finance*, 11(3), 291-308.

Viral V. Acharya, Allen N. Berger, Raluca A. Roman (2017). Lending Implications of U.S. Bank Stress Tests: Costs or Benefits? Retrieved from SSRN: http://dx.doi.org/10.2139/ssrn.2972919

Hyunchul Choi, Hyojoo Son, Changwan Kim (2018). Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Systems with Applications,* 106, 61-71. doi:10.1016/j.eswa.2018.05.026

Jochen Kruppa, Alexandra Schwarz, Gerhard Arminger, Andreas Ziegler (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications,* 40(18), 7562-7571. doi:10.1016/j.eswa.2013.03.019

Perboli, G., & Arabnezhad, E. (2021). A machine learning-based DSS for mid and long-term company crisis prediction. *Expert Systems with Applications*, 186, 115537. https://doi.org/10.1016/j.eswa.2021.115537

Radovanovic, J., & Haas, C. (2023). The evaluation of bankruptcy prediction models. *Expert Systems with Applications*, 101, 115-128. https://doi.org/10.1016/j.eswa.2023.03.016

European Commission. (2020). State aid: Commission approves Italian guarantee scheme to support economy in coronavirus outbreak. Retrieved from European *Commission Press Corner website*: https://ec.europa.eu/commission/presscorner/detail/pt/ip_20_655

Croft, A. (2021). EU bankruptcy filings jump to 8-year high as pandemic aid ends. *Financial Times*. Retrieved from https://www-ft-com.eur.idm.oclc.org/content/c90c3556-3218-47ff-aeda-3f23af217c11

Faria, A., Gupta, R., & Zampolli, F. (2023). Evaluating the Costs of Government Credit Support Programs during COVID-19: *International Evidence. IMF Working Papers*, 23(16). Retrieved from https://www.elibrary.imf.org/view/journals/001/2023/016/article-A001-en.xml

European Central Bank. (2020). Public loan guarantees and bank lending in the COVID-19 period. *ECB Economic Bulletin, Issue 7/2020*. Retrieved from https://www.ecb.europa.eu/pub/economic-bulletin/focus/2020/html/ecb.ebbox202006_07~5a3b3d1f8f.en.html#:~:text=The%20overall%20size%20of%20the,%E2%82%AC300%20billion%20for%20Italy.&text=SMEs%20in%20the%20sectors%20that,from%20public%20loan%20guarantee%20schemes.

OECD. (2020). Italian regional SME policy responses. Retrieved from *OECD Policy Responses to Coronavirus (COVID-19) website*: https://www.oecd.org/coronavirus/policy-responses/italian-regional-sme-policy-responses-aa0eebbc/

OECD. (n.d.). 22. Italy. Retrieved from *OECD iLibrary* website: https://www.oecdilibrary.org/sites/82245334en/index.html?itemId=/content/component/82245334-en

Bogatyy, A. (2021, February 3). CatBoost vs LightGBM vs XGBoost. Towards Data Science. https://towardsdatascience.com/catboost-vs-lightgbm-vs-xgboost-c80f40662924

Sharma, A. (2023, January 3). Ensemble Learning Methods: Bagging, Boosting, and Stacking. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/

GeeksforGeeks. (n.d.). Bagging vs Boosting in Machine Learning. https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/

Giudice, M. D., & Romano, F. (2023). Economic and Political Crisis in Italy. In F. G. Guimarães (Ed.), The European Union: Progress, Tensions, and Future Directions (pp. 145-162). Springer. https://doi.org/10.1007/978-3-030-54895-7_10

Gazzetta Ufficiale della Repubblica Italiana. (2020, April 8). Official Gazette Notification (Issue No. 43). https://www.gazzettaufficiale.it/eli/id/2020/04/08/20G00043/s

Rashidi, A., & Balakrishnan, S. (2018). Hyperparameter Tuning. In Machine Learning with XGBoost (pp. 149-162). O'Reilly. https://www.oreilly.com/library/view/machine-learning-with/9781789343700/e3b1b1fd-ddff-435e-aff8-44eb01608b91.xhtml

Prashant. (n.d.). A Guide on XGBoost Hyperparameters Tuning. Kaggle. https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning

Singh, K. (2022, November 15). XGBoost vs Random Forest - Adv/Disadv. Medium. https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30

Stack Overflow. (2013, April 27). Difference between PCA (Principal Component Analysis) and Feature Selection. https://stackoverflow.com/questions/16249625/difference-between-pca-principal-component-analysis-and-feature-selection#:~:text=The%20difference%20is%20that%20PCA,takes%20the%20target%20into%20consideration.

Brownlee, J. (2022, June 6). Feature Reduction. Machine Learning Mastery. https://machinelearningmastery.com/dimensionality-reduction-algorithms-with-python/

Brownlee, J. (2021, March 31). Recursive Feature Elimination. Machine Learning Mastery. https://machinelearningmastery.com/rfe-feature-selection-in-python/

Nilesh, T. (n.d.). XGBoost Algorithm Explained in Less Than 5 Minutes. Medium. https://medium.com/@techynilesh/xgboost-algorithm-explained-in-less-than-5-minutes-b561dcc1ccee

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754.

Jabeur, S. B., Stef, N., & Carmona, P. (2022). Bankruptcy Prediction using the XGBoost Algorithm and Variable Importance Feature Engineering. *Computational Economics*. https://doi.org/10.1007/s10614-021-10227-1

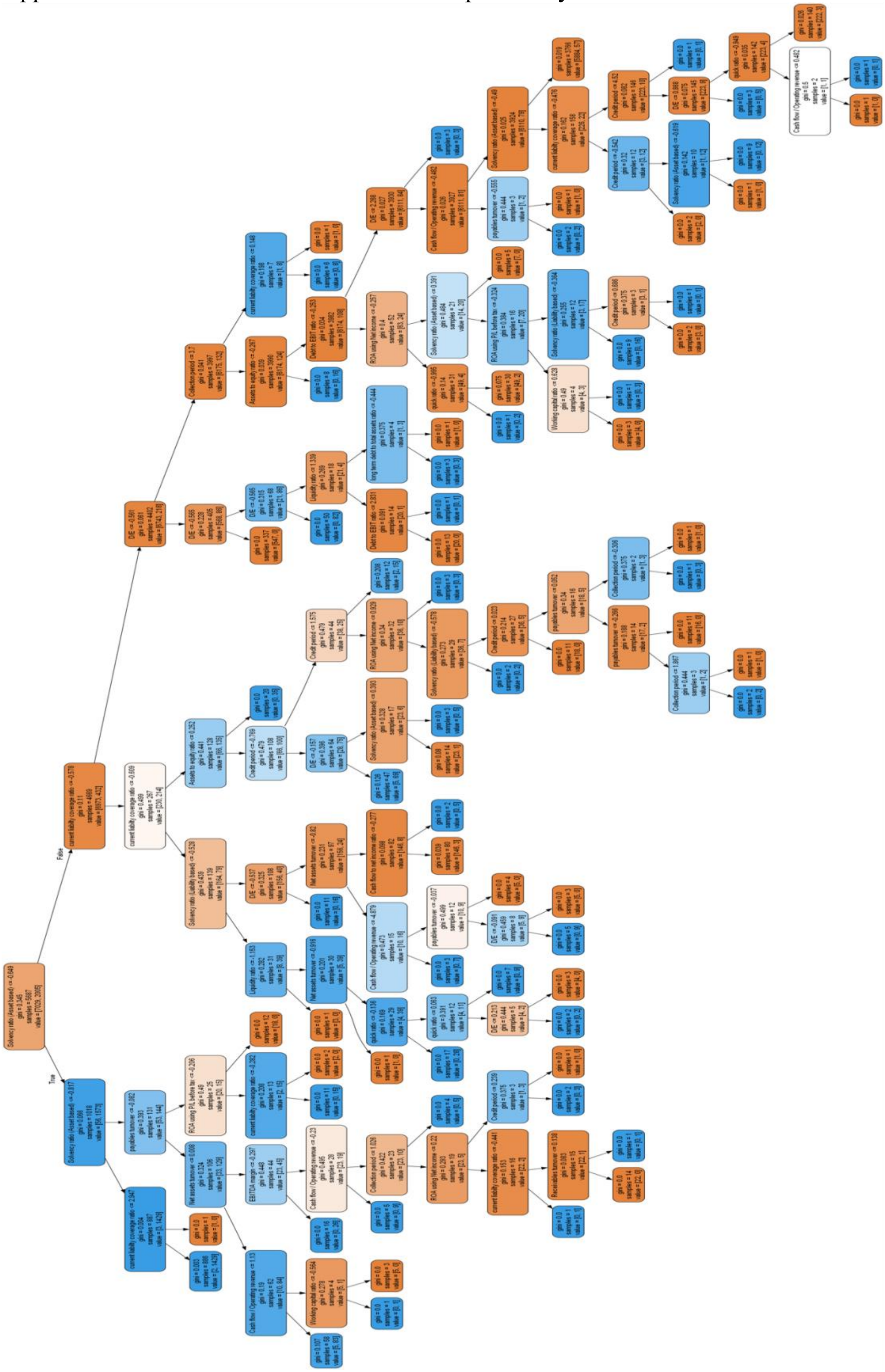
9. Appendix:

Appendix 1: The table displays all 39 features used in both the Random forest and XGBoost models, along with their formula and classification.

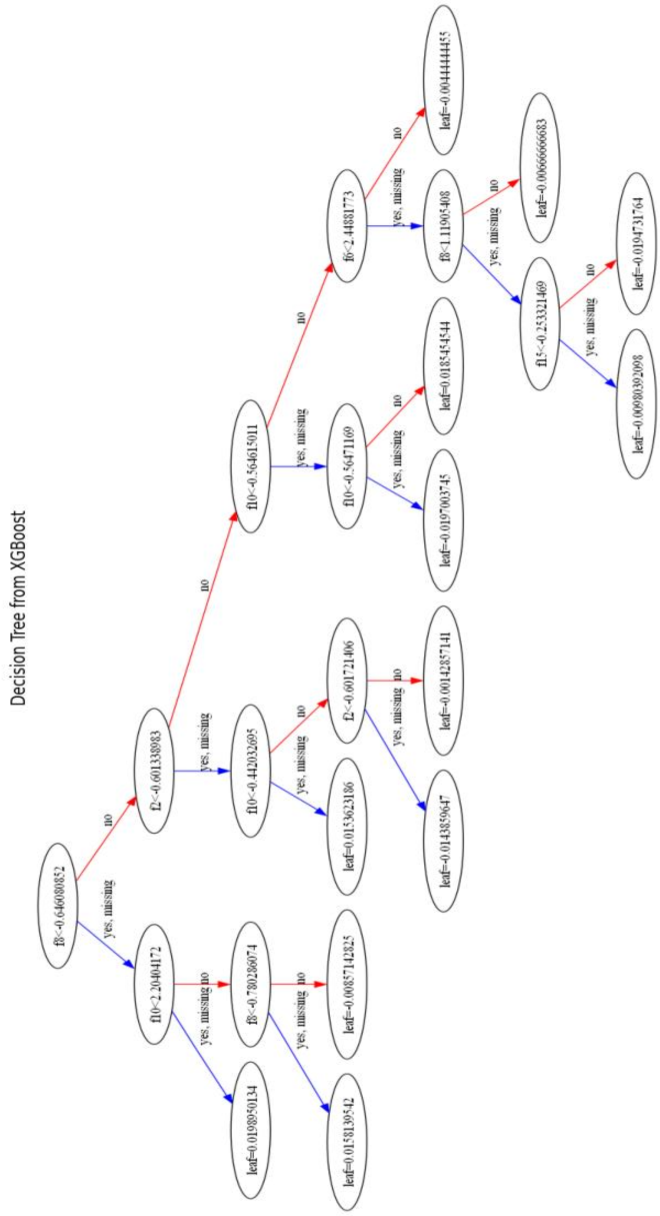
Ratio	Category	Formula
ROE using P/L before tax	Profitability Ratio	(Net Income Before Tax / Total Equity) * 100
ROCE using P/L before tax	Profitability Ratio	(Net Income Before Tax / (Total Equity + Total Debt)) * 100
ROA using P/L before tax	Profitability Ratio	(Net Income Before Tax / Total Assets) * 100
ROCE using Net income	Profitability Ratio	(Net Income / (Total Equity + Total Debt)) * 100
ROA using Net income	Profitability Ratio	(Net Income / Total Assets) * 100
Profit Margin	Profitability Ratio	(Net Income / Total Revenue) * 100
EBITDA Margin	Profitability Ratio	(EBITDA / Total Revenue) * 100
EBIT Margin	Profitability Ratio	(EBIT / Total Revenue) * 100
Cash Flow / Operating Revenue	Cash Flow Ratio	Cash Flow from Operations / Total Revenue
Net Assets Turnover	Operational Ratio	Total Revenue / Average Net Assets
Interest Cover	Operational Ratio	EBIT / Interest Expense
Stock Turnover	Operational Ratio	Cost of Goods Sold / Average Inventory
Credit Period	Operational Ratio	(Accounts Payable / Total Credit Purchases) * Number of Days
Collection Period	Operational Ratio	(Accounts Receivable / Total Credit Sales) * Number of Days
Current Ratio	Liquidity Ratio	Current Assets / Current Liabilities
Liquidity Ratio	Liquidity Ratio	(Current Assets - Inventory) / Current Liabilities
Shareholders Liquidity Ratio	Liquidity Ratio	(Total Equity + Reserves) / Total Assets
Solvency Ratio (Asset-based)	Solvency Ratio	Total Assets / Total Liabilities

Solvency Ratio (Liability-based)	Solvency Ratio	Total Liabilities / Total Assets
Gearing	Solvency Ratio	Total Debt / Total Equity
Short-term Gearing	Solvency Ratio	Short-term Debt / Total Equity
D/E (Debt-to-Equity)	Solvency Ratio	Total Debt / Total Equity
Cash Flow Coverage Ratio	Cash Flow Ratio	Cash Flow from Operations / Total Debt
Cash Flow Margin Ratio	Profitability Ratio	Cash Flow from Operations / Total Revenue
Current Liability Coverage Ratio	Cash Flow Ratio	Operating Cash Flow / Current Liabilities
Cash Flow to Net Income Ratio	Profitability Ratio	Cash Flow from Operations / Net Income
Cash Interest Coverage Ratio	Operational Ratio	Operating Cash Flow / Cash Interest Expense
Revenue Growth	Profitability Ratio	(Current Year Revenue - Previous Year Revenue) / Previous Year Revenue
Margin Growth	Profitability Ratio	(Current Year Margin - Previous Year Margin) / Previous Year Margin
Quick Ratio	Liquidity Ratio	(Current Assets - Inventory) / Current Liabilities
Long-term Debt to Total Assets Ratio	Solvency Ratio	Long-term Debt / Total Assets
Long-term Debt to Capital Ratio	Solvency Ratio	Long-term Debt / (Long-term Debt + Total Equity)
Assets to Equity Ratio	Solvency Ratio	Total Assets / Total Equity
Operating Margin	Profitability Ratio	Operating Income / Total Revenue
Pretax Margin	Profitability Ratio	Pre-Tax Income / Total Revenue
Net Profit Margin	Profitability Ratio	Net Income / Total Revenue
Receivables Turnover	Operational Ratio	Net Credit Sales / Average Accounts Receivable
Payables Turnover	Operational Ratio	Cost of Goods Sold / Average Accounts Payable
Working Capital Ratio	Liquidity Ratio	Current Assets / Current Liabilities

Appendix 2: Random forest decision tree sample from year1 model.



Appendix 3: XGBoost decision tree sample from Year 1 model.



Appendix 4: SHAP outputs

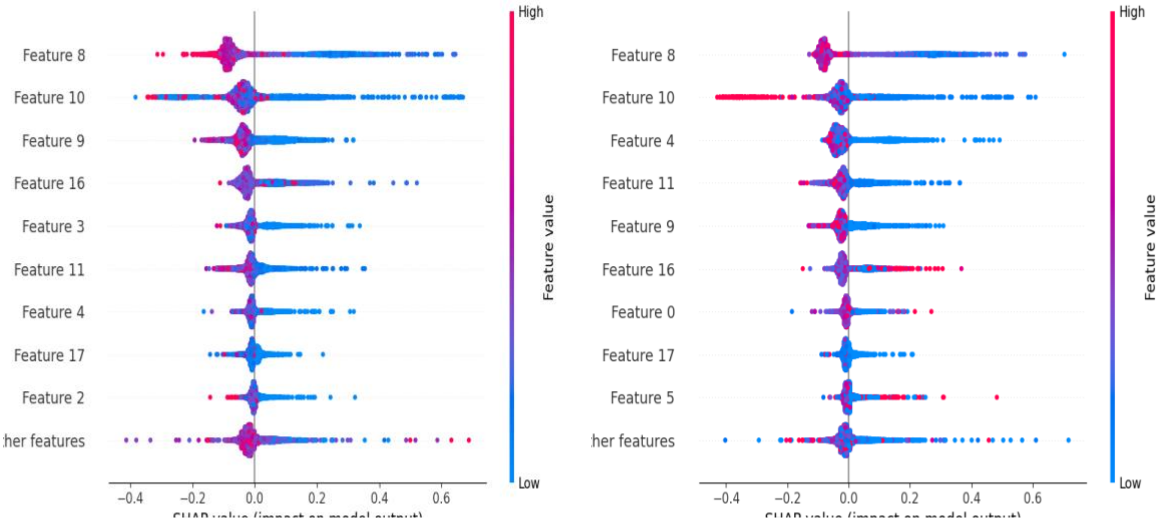


Figure shows SHAP outputs for year 1 and year 2 for the random forest models. Figure identifies features 8 and 10 to have the most predictive power amongst all features.

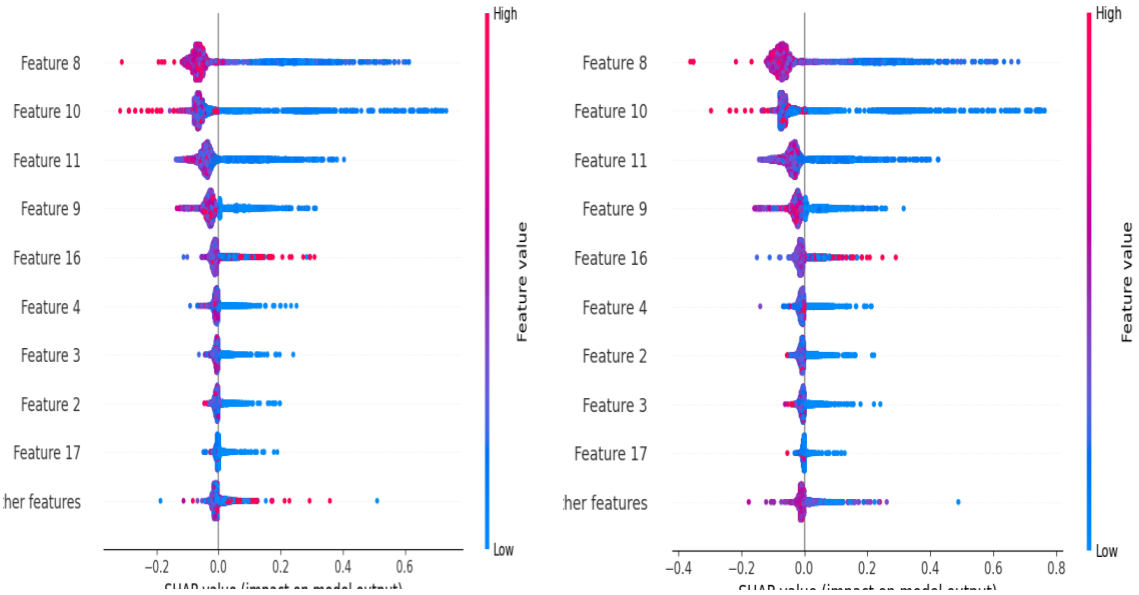
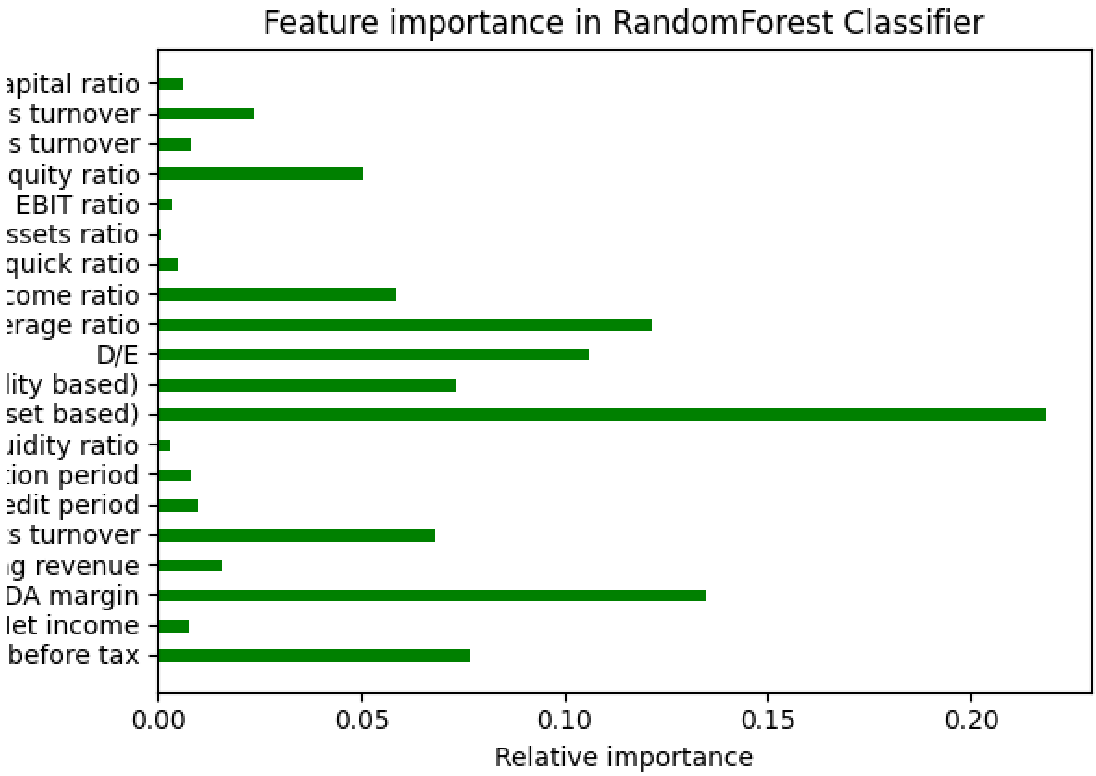
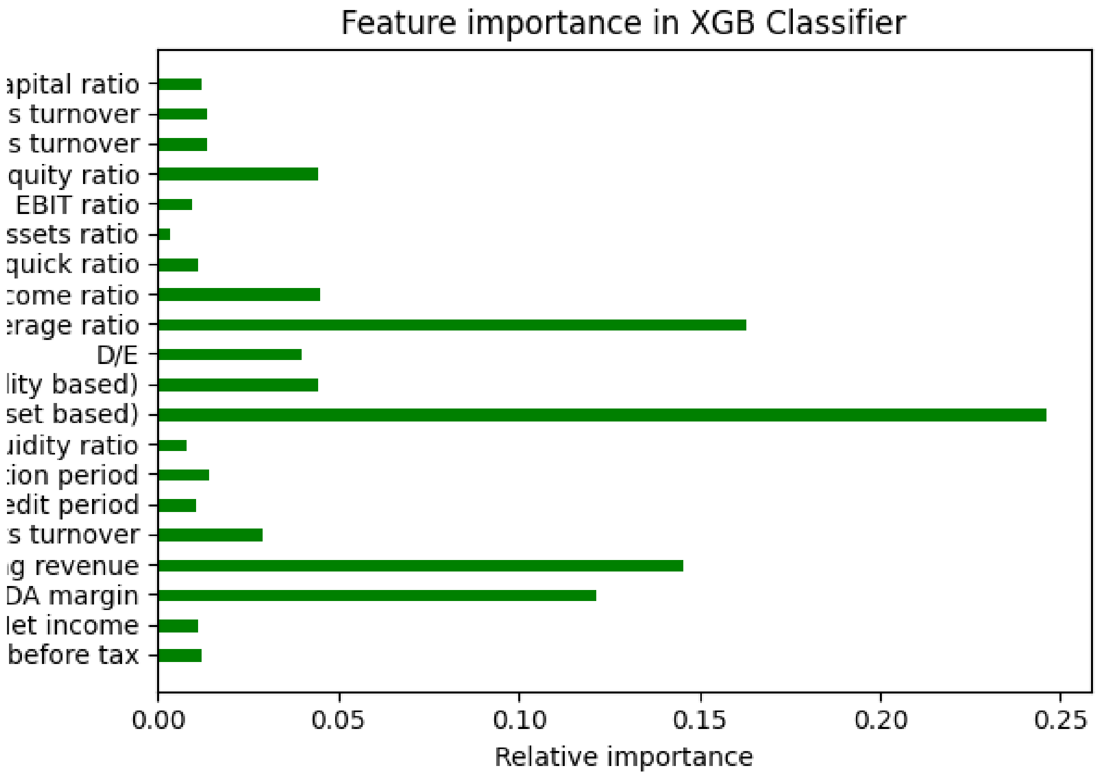


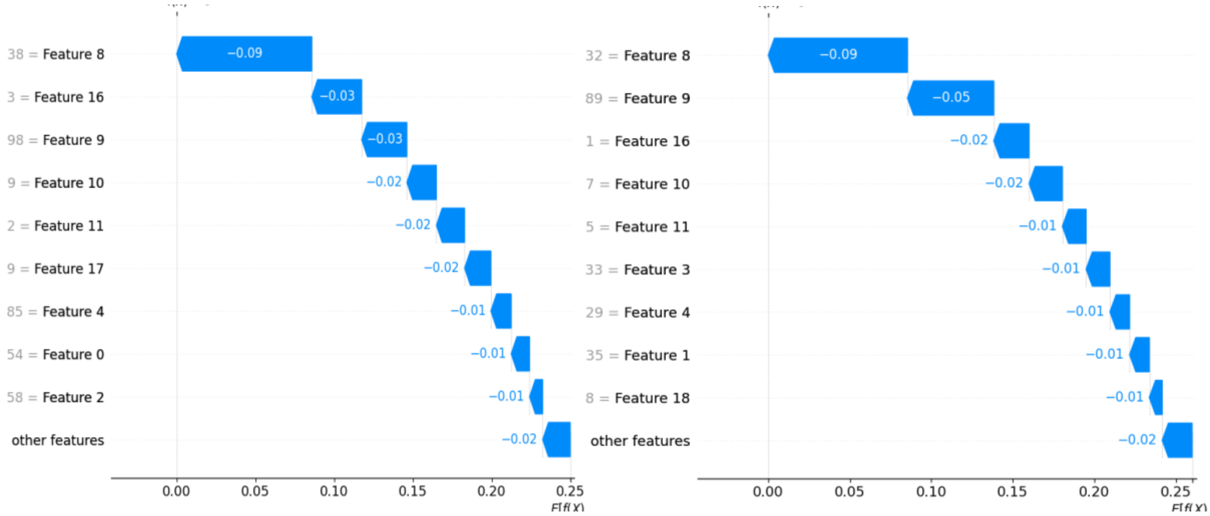
Figure shows SHAP outputs for year 1 and year 2 for the XGBoost models. Figure identifies features 8 and 10 to have the most predictive power amongst all features.

Appendix 5:
Feature importance for both Random Forest and XGBoost models. Solvency ratio (asset based has the highest predictive power, followed by current liability coverage ratio and EBITDA margin across both models.

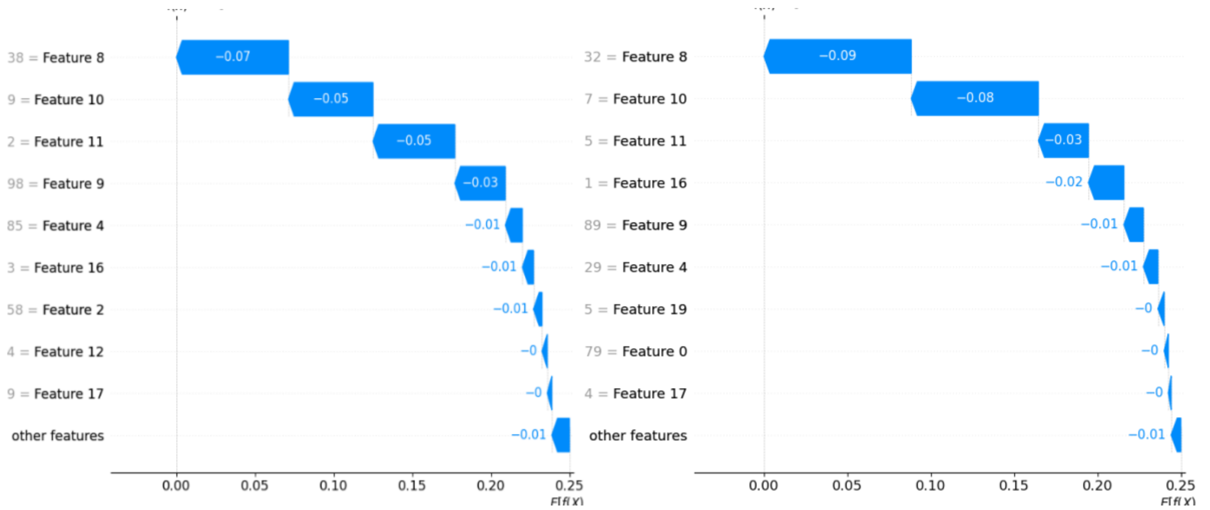




Appendix 6: Waterfall graphs

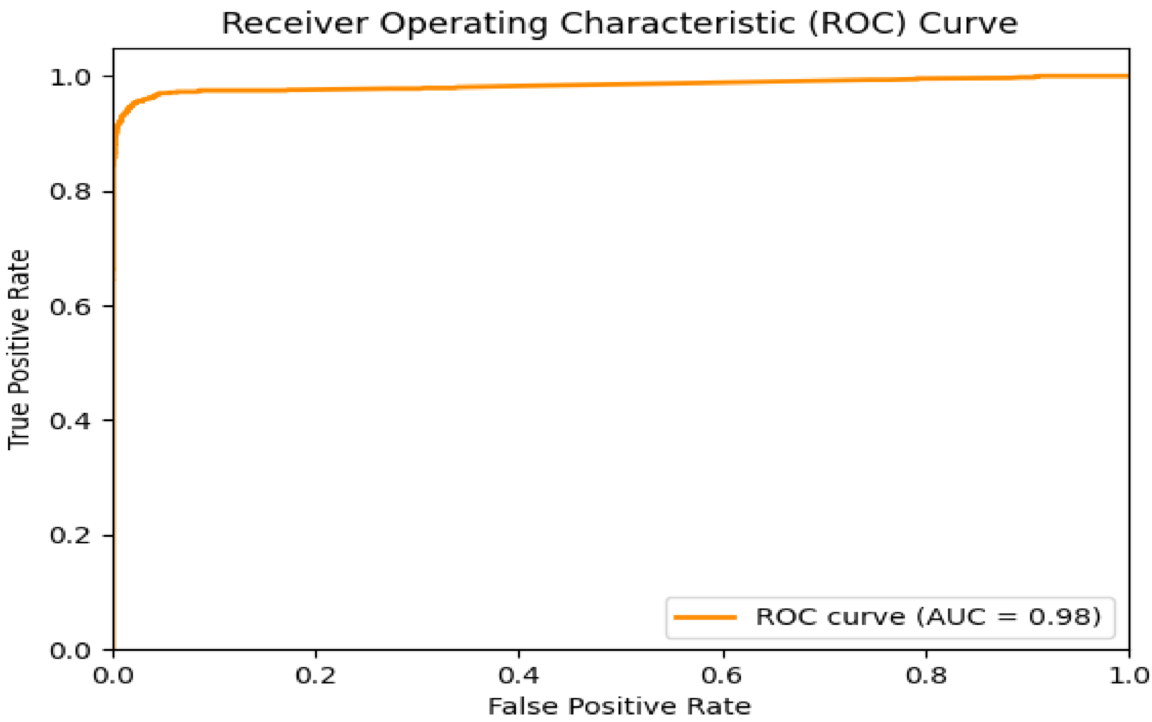


Random forest waterfall graphs for years 1 and 2. Displays feature importance.

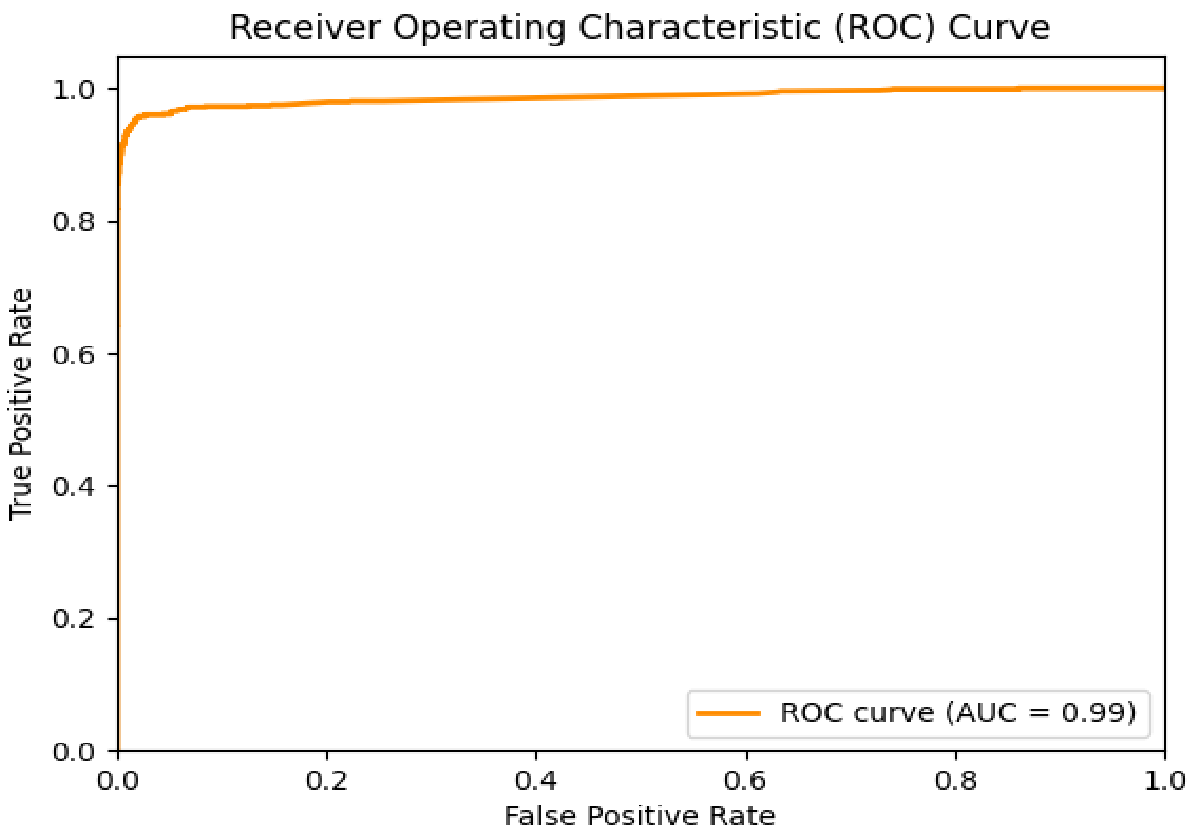


XGBoost waterfall graphs for years 1 and 2. Displays feature importance.

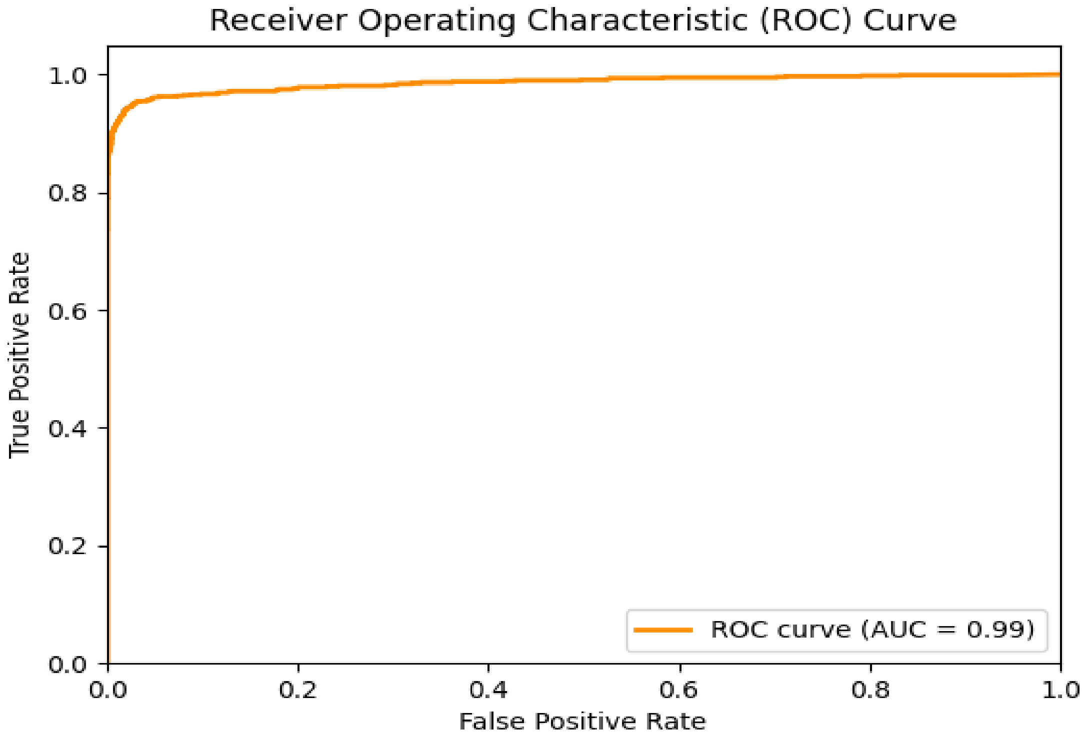
Appendix 7: Receiver operating characteristic (ROC) curve. These curves display the predictive accuracy of the models.



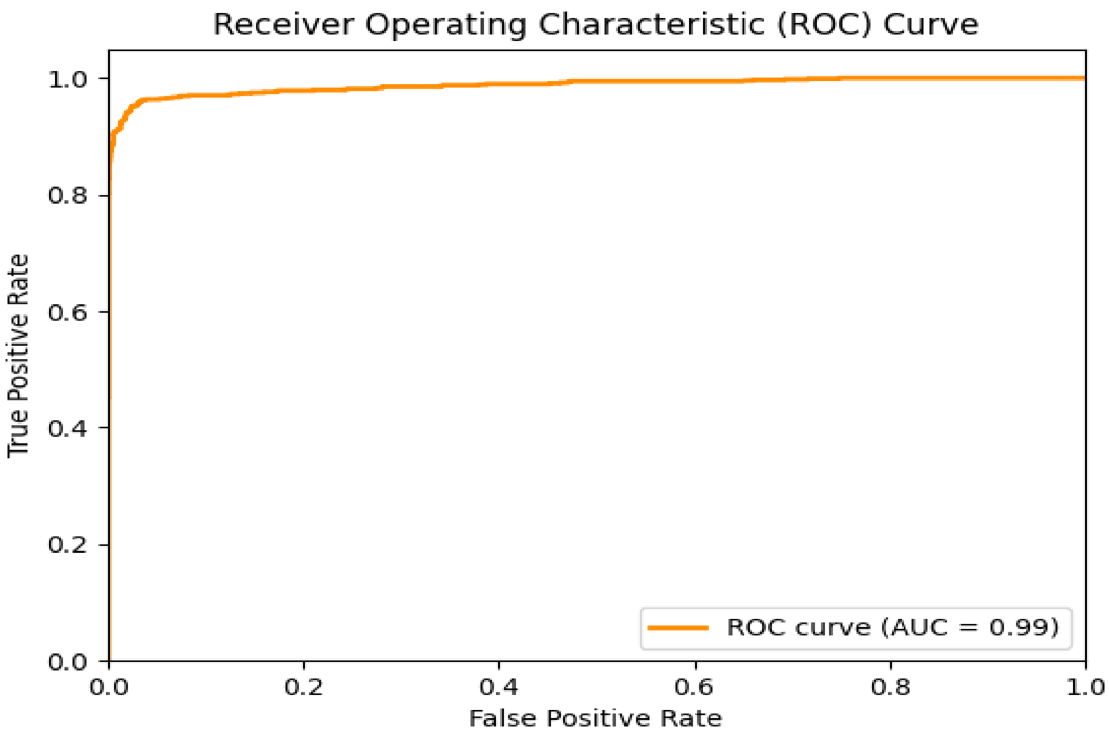
Random forest year 1



Random forest year 2

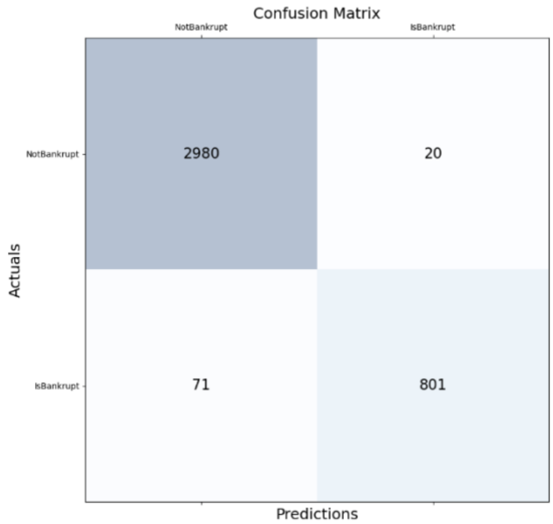


XGBoost year 1

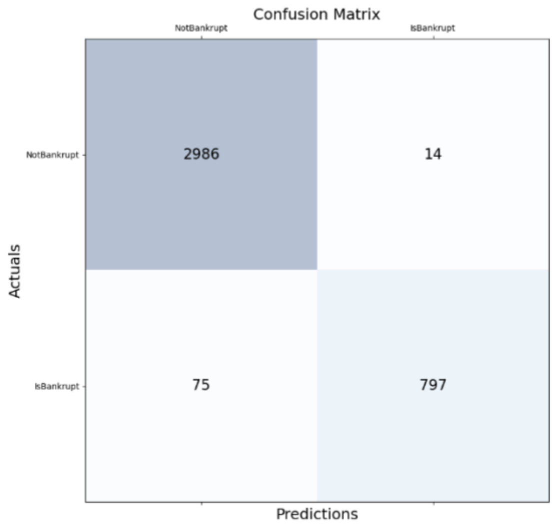


XGBoost year 2

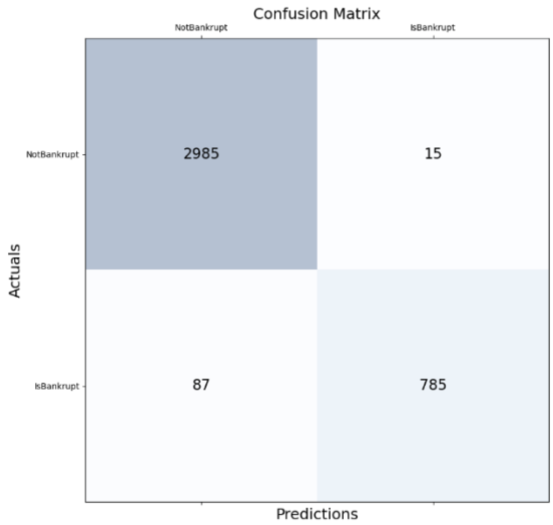
Appendix 8: Confusion matrices. These graphs display the true positive, true negative, false positive and false negative rates for each model.



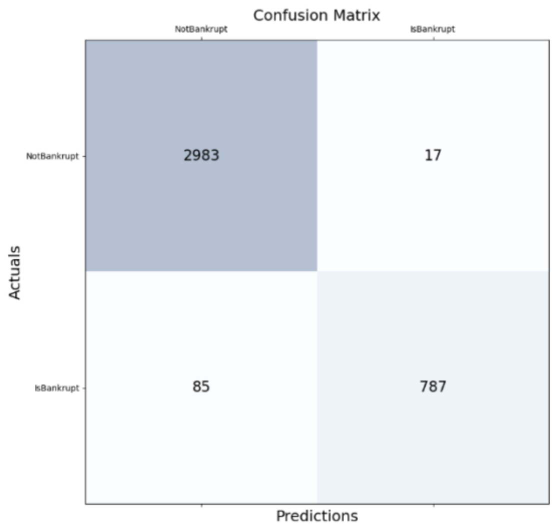
Random forest year 1



Random forest year 2



XGBoost year 1



XGBoost year 2