A Comparison of Feature Selection Methodology for Solving Classification Problems in Finance

Xia Xiaomao\textsuperscript{1,*}, Zhang Xudong\textsuperscript{1} and Wang Yuanfang\textsuperscript{1}

\textsuperscript{1}JD Digits

*baiyouxxm@163.com

Abstract. Classification is one of the most important tasks in real world with the intention of finding the underlying patterns of the data and making use of the found patterns. Other than considering lots of parameters that may influence the classification model accuracy, the influence of the feature selections has been rarely discussed. SHAP (SHapley Additive exPlanation) value is a novel ensemble learning measurement that is the unique consistent and locally accurate attribution value. Its applications to feature selection are not mentioned in the literature. In this paper, we compare the prediction accuracy of several feature selection techniques on a set of Polish companies. We show that new feature selection method based on the SHAP values is superior (not worse) to other widely used prediction methods.

1. Introduction

There are two famous special methods of dimensionality reduction. The first one is feature extraction where the input data is transformed into a reduced representation set of features, so new attributes are generated from the initial ones. The second category is feature selection. In this category a subset of existing features is selected without a transformation. Generally, feature selection is preferred over feature extraction since it keeps all information about the importance of each single feature while in feature extraction obtained variables are usually not interpretable [1].

Conserving the information of each feature provides much simplicity and interpretability to financial data processing. Hence, feature selection is more appropriate to our study. Feature selection is an important framework in knowledge discovery and specially in financial applications, not only for the insight gained from determining predictive modelling features but also for the improved performance, understandability and accuracy of the financial models.

Feature selection is a process of finding an optimal subset which is given by a feature subset selection algorithm that provides the highest possible accuracy. The following definition is presented by Kohavi and John (1997): "Given an inducer \( \mathcal{I} \) and a dataset \( \Delta \) with features \( X_1, X_2, \ldots, X_n \) from a distribution \( D \) over the labeled instance space, an optimal feature subset, \( X_{opt} \), is a subset of the features such that the accuracy of the induced classifier \( X = \mathcal{I}(\Delta) \) is maximal (p. 276)” [2]. An optimal feature subset is not necessarily unique because when one feature can be replaced by another feature since they are perfectly correlated to each other, the accuracy derived by different combination of features is the same.

Until recently the dominating methods for predicting corporate bankruptcies have been based on statistical modelling, however, lately models based on machine learning have been proposed [3].
Recently, machine learning models have successfully been used for many classification and regression problems and these models have often outperform traditional classification methods [4]. The purpose of bankruptcy prediction is to assess the financial health status and future perspectives of a company. For a given period of time this problem can be modelled as a two-class classification problem. Companies either survive the given time period, or go bankrupt during it. The problem is to predict which of these two possible outcomes is the more probable one.

Financial reporting laws and public expectations on transparency means that there is a lot of data available concerning companies’ financial status [5]. The large amount of data makes the area well suited for sophisticated data intensive computation methods [6].

Estimating the risk of corporate bankruptcies is of large importance to creditors and investors. There are large indirect and direct costs associated with financial distress and bankruptcies [7] and for this reason bankruptcy prediction has for a long period of time constituted an extensive area of research [3]. Corporate bankruptcies can have serious effects both locally and globally [8], employees, investors, customers, suppliers and their financiers are all affected when a company disappears [9]. In some cases, a corporate bankruptcy can cause an entire industry to suffer [10].

According to Rodriguez et al. (2010) a successful feature selection: a) reduces the feature space’s dimensionality, b) speeds up and reduces the cost of a learning algorithm and c) obtains the feature subset which is the most relevant to classification [11]. Feature selection algorithms are typically composed of the following four components: search direction, search strategy, evaluation function and the stopping point.

Random forest is an ensemble learning method. The algorithm is proposed by Ho [12]. The training algorithm is based on bagging, which is short for bootstrap aggregating. Each decision tree is trained on a data set randomly sampled from the training data with replacement. Such an approach means that each tree learner is shown a different subset of the training data and that same observation can be chosen more than once in a sample [13].

Extreme Gradient Boosting (XGBoost) is similar to gradient boosting framework but more efficient. It has both linear model solver and tree learning algorithms. So, what makes it fast is its capacity to do parallel computation on a single machine. This makes XGBoost at least 10 times faster than existing gradient boosting implementations. It supports various objective functions, including regression, classification and ranking. It also has additional features for doing cross validation and finding important variables [14].

SHapley Additive exPlanation (SHAP) value, a unified measure of feature importance was first proposed by Lundberg and Lee [15]. They are based on ideas from game theory and local explanations. It can be proved that SHAP values are the only consistent and locally accurate individualized feature attributions.

The paper is organized as below. In section 1, we describe the concepts of SHAP values, the importance of feature selections when we solve classification problems and issues on bankruptcy predictions. In section 2, we describe the dataset and samples of our research. Then, in section 3, we discuss our outcomes of different feature selection methods. In section 4 of conclusion and discussion part, we make a conclusion of the study and discuss other potential applications.

2. Dataset
The dataset used in this paper consisted of financial condition about polish companies in the manufacturing sector. Many Polish manufacturing companies went bankrupt in this sector since 2004 [16]. The dataset contained information for both bankruptcy companies and still operating ones. The financial information describing the health of the 700 bankrupt companies (almost 2,400 financial statements) was gathered during the period 2007-2013 and more than 10,000 (more than 65,000 financial statements) operating companies was collected during the period 2000-2012.

The data is divided into five different subsets, which are described in Table 1. The subsets are created to allow for different lengths of forecasting period. The task in each of them is to predict whether or not a company goes bankrupt within five, four, three, two and one year respectively based
on information that could be retrieved from 64 different financial features (regressor variables or indicators). All five data sets are, as expected also in real life, heavily imbalanced. There are much fewer bankrupt companies compared to still operating ones. 64 different financial indicators are used to describe the financial status of the companies. Almost all of the features are financial ratios constructed from information found in the companies’ economic statements. The use of ratios should mean that the predictors are not too heavily correlated with the size of the companies. A complete list of the features descriptions can be found in UCI Machine Learning Repository.

### Table 1 Description of the data pertaining to the five different classification tasks

| Dataset | Features from | Bankruptcy after | No. of bankruptcy | No. of not bankruptcy | Total |
|---------|---------------|------------------|-------------------|-----------------------|-------|
| 1st year | 1st year | 5 years | 271 | 6,756 | 7,027 |
| 2nd year | 2nd year | 4 years | 400 | 9,773 | 10,173 |
| 3rd year | 3rd year | 3 years | 495 | 10,008 | 10,503 |
| 4th year | 4th year | 2 years | 515 | 9,277 | 9,792 |
| 5th year | 5th year | 1 year | 410 | 5,500 | 5,910 |

### 3. Methods, applications and results

#### 3.1. Methods

This paper considers only supervised learning. Corporate bankruptcy prediction is treated as a two-category classification problem. A company either belongs to the class of companies that faces bankruptcy or to the class that still operates after a given time period.

Here we consider several methods to do feature selection. The focus is on the importance ranking of different features. (1) Information Value ranking (IV) (2) Random Forest with the default mean decrease in impurity (or Gini importance) ranking (RF) (3) Random Forest with SHAP values ranking (RFS) (4) Extreme Gradient Boosting with split count importance ranking (XGBSC) (5) Extreme Gradient Boosting with gain importance ranking (XGBG) (6) Extreme Gradient Boosting with coverage importance ranking (XGBC) (7) Extreme Gradient Boosting with SHAP values ranking (XGBS).

To compare results from different methods, cross validation is implemented in this study. The dataset is divided into $n$ equally large subsets, $n - 1$ of those are applied to the training set and the last subset is applied to testing. We repeat the procedure $n$ times, with every of the $n$ sets acting as test subset once each. In this way we get $n$ measurements of the scoring metric that we are concerned with, and we can calculate both the mean and standard deviation of this scoring metric.

Most machine learning algorithms does not return a binary prediction (e.g. still operating or bankrupt), but rather a probability. To calculate true/false positive rate from the returned probabilities we have to choose a threshold, which we use to decide which class we consider a sample to belong to. For example, all companies that the machine identifies will face bankruptcy by more than 70% probability we will consider as bankrupt. As we increase this threshold both the true/false positive rate will decrease. If we set the threshold to 100% we will have TPR (true positive rate) = FPR (false positive rate) = 0 and if it is set the threshold to 0% we will get TPR = FPR = 1. In between these two endpoints we have a space of opportunity upon which we can decide how we want our classifier to perform. The curve which connects these two values are called a Receiver operating characteristic (ROC) curve. Values closer to the upper left corner are better since they mean a higher true positive rate as well as a lower false positive rate [17]. From just looking at a graph it is hard to measure the performance of two or more types of algorithms. To overcome this a metric called Area Under Curve (AUC) has been derived. It is, as the name suggests, defined as the area under the Receiver operating characteristic (ROC) curve. An algorithm that guesses randomly will obtain a value of 0.5, since the
area under a triangle with legs of length one is precisely 0.5. AUC can be interpreted as how much opportunities our algorithms give us for different scenarios, where potentially different measures are of importance.

3.2. Applications and results

The first 30 most important features are chosen for its own method except for IV rank. Particularly for IV rank, the first 30 chosen features are used for training random forest model. All the seven models are fit and computed in Python.

Hyperparameters values typically have large influence on the performance and the behaviour of a machine learning algorithms. For each method, the hyperparameters are explored and tuned as below. For information value (IV), random forest (RF) and random forest with SHAP values (RFS) method, the following numbers are assigned. The maximum depth of tree is 6. The number of trees in the forest is 300. The minimum number of samples required to be at a leaf node is 100. The L1 regularization term on weights is 0.05. L2 regularization term on weights is 0.005.

There is a large discrepancy in the distribution of missing values between the two different classes. In other words, information about the missing values should be very helpful in the classification. We try to (1) impute missing values as the average (2) impute missing values as very large or small values (3) add the total number of missing values as a synthetic feature (4) add a dummy variable for each possible missing value (1-hot). There is no significant difference between these four methods in AUC and KS values. In the following tables of the study, we choose to report the results from the first method.

Table 2 Experimental results for considered classification models. The numbers are the area under curve (AUC) of the ROC curve. STD is short for the standard deviation.

|         | 1st Year | 2nd Year | 3rd Year | 4th Year | 5th Year |
|---------|----------|----------|----------|----------|----------|
| XGBSC   | MEAN     | STD      | MEAN     | STD      | MEAN     | STD      |
|         | 0.953    | 0.023    | 0.918    | 0.037    | 0.930    | 0.020    | 0.934    | 0.024    | 0.952    | 0.015    |
| XGBG    | 0.955    | 0.024    | 0.931    | 0.028    | 0.923    | 0.022    | 0.934    | 0.020    | 0.954    | 0.015    |
| XGBC    | 0.830    | 0.044    | 0.724    | 0.027    | 0.746    | 0.037    | 0.823    | 0.042    | 0.934    | 0.015    |
| XGBS    | 0.951    | 0.016    | 0.930    | 0.025    | 0.924    | 0.026    | 0.936    | 0.021    | 0.953    | 0.016    |
| IV      | 0.872    | 0.044    | 0.833    | 0.058    | 0.842    | 0.043    | 0.832    | 0.025    | 0.893    | 0.018    |
| RF      | 0.899    | 0.045    | 0.865    | 0.050    | 0.861    | 0.039    | 0.847    | 0.025    | 0.887    | 0.018    |
| RFS     | 0.896    | 0.041    | 0.856    | 0.053    | 0.857    | 0.041    | 0.843    | 0.027    | 0.891    | 0.018    |

Table 3 Experimental results for considered classification models. The numbers are the Kolmogorov-Smirnov (KS) statistic. STD is short for the standard deviation.

|         | 1st Year | 2nd Year | 3rd Year | 4th Year | 5th Year |
|---------|----------|----------|----------|----------|----------|
| XGBSC   | MEAN     | STD      | MEAN     | STD      | MEAN     | STD      |
|         | 0.792    | 0.076    | 0.703    | 0.095    | 0.729    | 0.067    | 0.745    | 0.063    | 0.793    | 0.043    |
| XGBG    | 0.798    | 0.086    | 0.733    | 0.074    | 0.721    | 0.062    | 0.735    | 0.048    | 0.812    | 0.029    |
| XGBC    | 0.553    | 0.080    | 0.366    | 0.049    | 0.416    | 0.068    | 0.526    | 0.084    | 0.743    | 0.042    |
| XGBS    | 0.794    | 0.060    | 0.728    | 0.073    | 0.723    | 0.061    | 0.741    | 0.051    | 0.811    | 0.053    |
| IV      | 0.633    | 0.097    | 0.564    | 0.091    | 0.566    | 0.058    | 0.547    | 0.047    | 0.666    | 0.036    |
| RF      | 0.686    | 0.106    | 0.606    | 0.086    | 0.620    | 0.062    | 0.566    | 0.044    | 0.675    | 0.039    |
| RFS     | 0.693    | 0.103    | 0.593    | 0.089    | 0.602    | 0.062    | 0.558    | 0.047    | 0.678    | 0.040    |
Table 4 The p-values of signed rank Wilcoxon test of the KS/AUC of XGBS vs the best model

| XGBS vs Best | 1st Year AUC | 2nd Year AUC | 3rd Year AUC | 4th Year AUC | 5th Year AUC |
|--------------|--------------|--------------|--------------|--------------|--------------|
| P-value      | 0.545        | 0.880        | 0.496        | NA           | 0.880        |

From Table 2 and Table 3, the mean of the best model and XGBS are very similar. To further explore the difference significance between results obtained by XGBS vs Best model, signed rank Wilcoxon test (also known as ‘Mann-Whitney’ test) is applied to AUC and KS on the test dataset from 10-fold cross validation. The p-values of signed rank Wilcoxon test of the KS/AUC of XGBS vs the best model are in Table 4. Assuming the significance level equal to 0.05, for the Wilcoxon test, the null distribution difference hypotheses cannot be rejected. In conclusion, XGBS performs as good as the best models in terms of AUC or KS values. It is summarized that the performance of XGBS is very close to the best model which is in bold if XGBS is not the best feature selection model. XGBS is proven to be a very effective feature selection method.

4. Conclusions and discussions
The predictive performance achieved on the dataset is very good. One possible reason for the superb performance is the information contained in the missing values. Future interesting research topics in the development of algorithms for classification problems in finance could be to incorporate new types of data. For example, unstructured data, such as texts in annual reports, could be of interest. Also, apply described algorithms to another dataset would also be of interest to better understand the validity of the results.

In conclusion, we introduced a new feature selection method based on SHAP values for identification of informative variables. The SHAP framework identifies the class of additive feature importance methods and shows there is a unique solution in this class that adheres to desirable properties. SHAP values consistently attribute feature importance, better align with human intuition, and better recover influential features. When applied to the Polish bankruptcy data, the method not only discovered all the important features but also produced the correct classifier. With better interpretability, the SHAP framework performs not worse than any other common feature selection methods in our empirical financial classification study.

References
[1] Giudici, P. (2003). Applied Data Mining: Statistical Methods for Business and Industry. West Sussex PO19 8SQ, England: John Wiley & Sons Ltd. The Atrium, Southern Gate, Chichester.
[2] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1–2), 273-324.
[3] Heikki Linden et al. Synthesis of research studies examining prediction of bankruptcy. 2015.
[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
[5] Xavier Bredart. Bankruptcy prediction model using neural networks. Accounting and Finance Research, 3(2):124, 2014.
[6] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1):1–16, 2016.
[7] Edward I Altman. A further empirical investigation of the bankruptcy cost question. The Journal of Finance, 39(4):1067–1089, 1984.
[8] Ralph De Haas and Neeltje Van Horen. International shock transmission after the lehman brothers collapse: Evidence from syndicated lending. The American Economic Review, 102(3):231–237, 2012.
[9] Stefan Engström. Kan nyckeltal påvisa framtida betalningsförmåga? Balans, 3:36–45, 2002.
[10] S Alex Yang, John R Birge, and Rodney P Parker. The supply chain effects of bankruptcy. *Management Science*, 61(10):2320–2338, 2015.

[11] Rodriguez, I., R. Huerta, C. Elkan, and C. S. Cruz. (2010). Quadratic programming feature selection. *Journal of Machine Learning Research* 11 (April): 1491 – 1516.

[12] Tin Kam Ho. Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, volume 1, pages 278–282. IEEE, 1995.

[13] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[14] Chen, T., and Guestrin, C. (2016). “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16* (New York, NY: ACM Press), 785–794.

[15] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4768–4777.

[16] Maciej Zieba, Sebastian K Tomczak, and JakubMTomczak. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 58:93–101, 2016.

[17] Christopher M Bishop. Pattern recognition and machine learning. Springer, 2006.