Corporate Bankruptcy Project Final Report

Jiaxing Wang

May 2, 2019

1 Problem description

The possibility of bankruptcy is always the main concern for a enterprise. Not only because it reflects the current financial condition of a company, but also influence the financial players to make decision for the company in the future. Moreover, for others companies, it can be treated as a sign of whether or not to cooperate with it. Finally, for the governments, they could take measures to prevent financial crisis if too many companies have high possibility of bankruptcy.

Generally, deciding whether the company would go bankrupt or not in a short of time for the given data set is what we want to solve for this project. According to Wagenmans[1], financial environments in which the company operates and decision makers of a company are the key factors for causing the bankruptcy. In the economy domain, there are many factors influencing the possibility of bankruptcy such as ROA (a financial ratio that shows the percentage of profit a company earns in relation to its overall resources), COGS (the total revenue minus the direct costs of producing that good or service). We would choose representative factors to analyze and predict the company's bankrupt possibility.

There is an increasingly interests for researchers to apply different machine learning models for predicting the possibility of bankruptcy due to its importance to financial area. Sonam Guptahas[2] uses bagging and Adaboost classifier to train the model and the precision rate is up to 91.64% According to its paper" Systematic review of bankruptcy prediction models: towards a framework for tool selection " published in 2017.

2 Data Description

The chosen data set contains bankruptcy information about Polish companies, which can be obtained from UCI Machine Learning Repository[3][3]. The data set has five files corresponding to different forecasting periods, which enables us to analyze the bankruptcy different periods of time . Later, we would compare the five data sets and choose a best one to train

the model afterwards. The 2 year data set contains 10173 data sample, which is the largest dataset among the five data sets. Each dataset has 64 features, which are calculated ratios related to companys financial annual statement. The 65 feature is the target output, 0 represents non-bankrupt and 1 represents bankrupt. According to the figure 1, we can see that all the five data sets are unbalanced, class 0 (non-bankrupt) has much more sample points compared with class 1(bankrupt). Besides, all the dataset has missing data and the year 2 data set has the largest missing data rate according to the calculation.

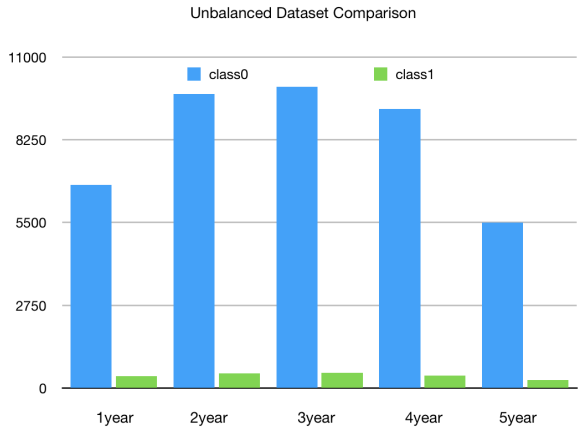

Figure 1: Unbalanced data set comparison

To better visualize the data, I used linear dimensionality reduction (t-SNE) to project the 64 feature data. As some of data are missing in the data set, we just delete the whole row if that row contains missing data (AKA NaN) and figure 2 shows the five data sets result. From the figure 2, we can clearly see that class 1 has fewer data samples compared with class 0. As a result, when we pre-process the dataset, we need to take unbalanced data and missing data issue into consideration. Unbalanced data issue would affect our training result. For example, the majority class would be biased when we using methods like KNN especially for bigger K. Missing data issue also needs to be addressed. According to Korean Anesthesiol [4]statistic strength would be reduced if data are missing and bias could exists when estimation the parameter for classifier.

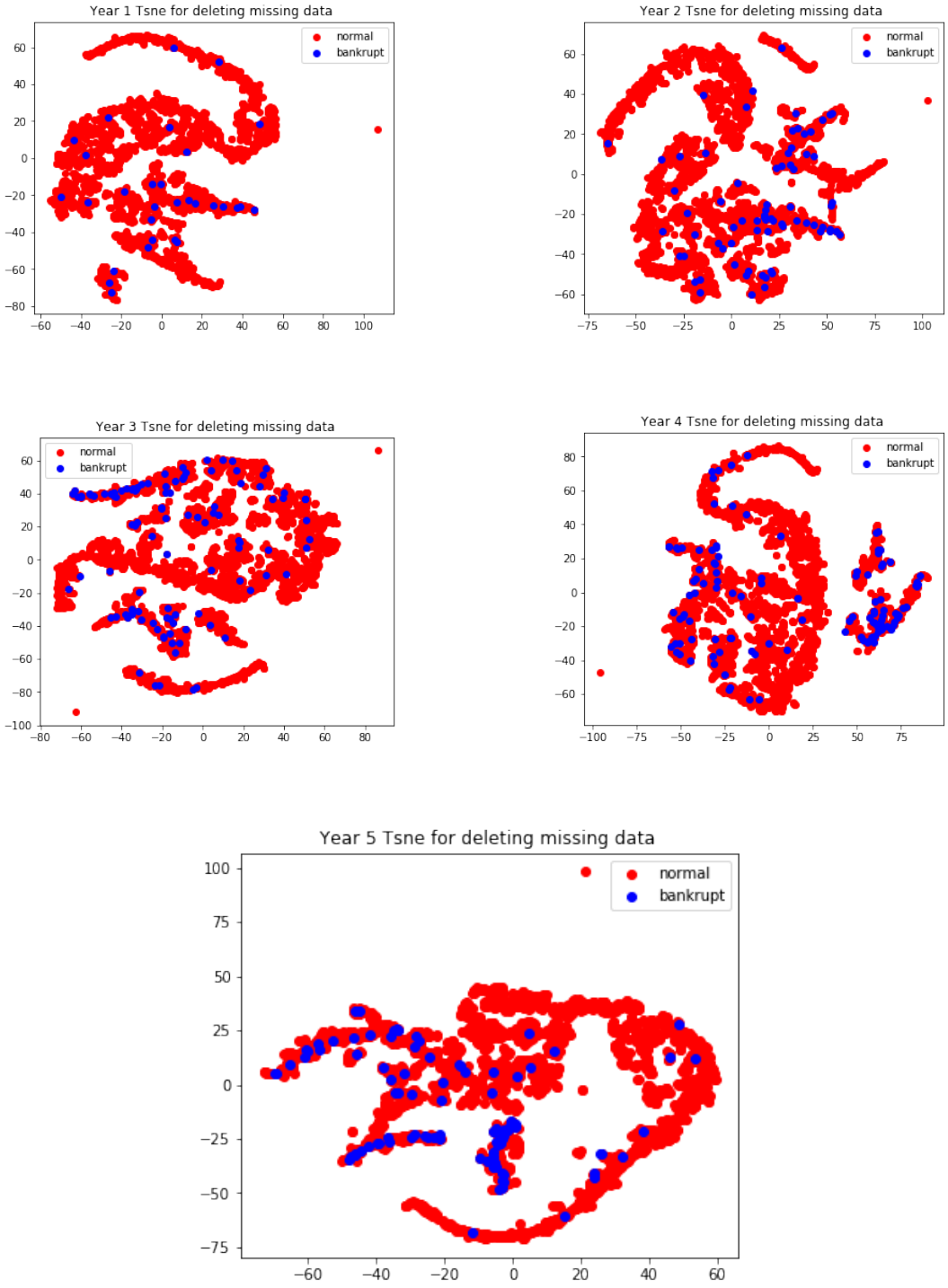

Figure 2: Five dataset comparison

3 Approaches/Methods

3.1 Pre-processing

To begin with, it is vital to focus on one data set because the five data sets belong to different periods of time, if simply combining them together, the time factor would be ignored. Taking both missing rate and sample amount into consideration, year 3 data set has been chosen because it has relatively low missing rate and larger amount of sample points. we need to apply some data cleaning methods to process the data set.

The data set misses lots of samples. if we simply dropping the row (company), there are only 4885 valid sample in the given data set, which is under-sampling. According to Grigorios Papageorgiou[5], The popular methods for handing missing data are complete case analysis and single imputation, which means we can use one particular value to replace the missing data. Here, using each features mean to replace the missing data within the feature is used because it is the most common and easy way to deal with the issue. The formula for calculating each features mean is:

- **Mean calculation**. A method to replacing the missing data sample in data set

$$\bar{X} = (\sum X_i)/n$$

  where $\bar{X}$ stands for the mean of each feature, n means the number of sample points in the dataset. $\sum X_i$ is the sum of the points for each feature.

In addition, as we can see from figure 1, the data set is unbalanced. The normal company (class 0) has much more sample points than the bankrupt company (class 1) for the given data set. We can simply use over-sampling method-SMOTE to increase the number of minor class data sample. According to Nitesh V. Chawla [6], the minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of thek minority class nearest neighbors. In this way, the decision region is more general for the minority class.

- **SMOTE**. A method to over-sampling the minority class

$$x_{i1} = x_i + \zeta_1 \cdot (x_{i(nn)} x_i)$$

  where x_{i1} is the new generated data sample, ζ_1 is the a random number between 0,1

After over-sampling, the data set has balance data, the number of class 0 and class 1 are the same and now the data set has 20016 data points as figure 3 shows. Last but not the least, the data set has 64 features, many features are quite related by calculating the correlation of features. feature selection method based on Extremely randomized trees is used to choose the top 2 feature. According to Pierre Geurts[7], compared with other tree-based methods,

extremely randomized tree reduces variance strongly by randomly choosing the cur-point and attribute with ensemble average. Figure 4 shows the top 10 feature.

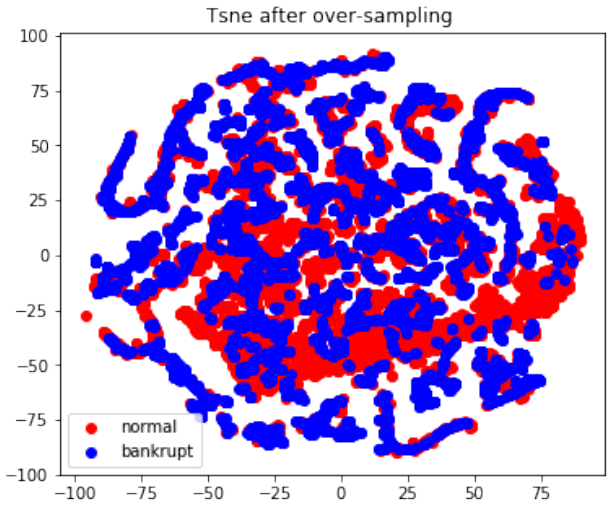

Figure 3: after pre-processing

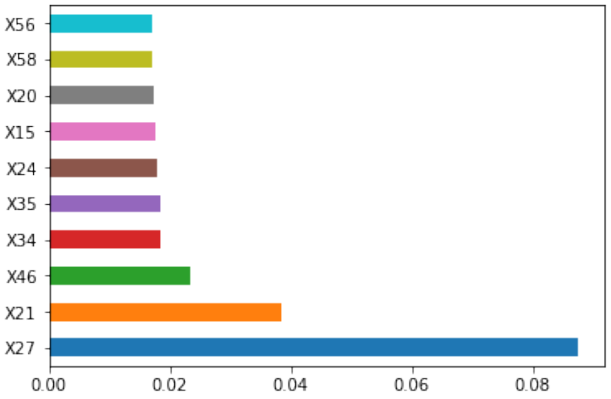

Figure 4: Feature selection using tree-based algorithm

3.2 Classifiers

It is not very suitable to use linear model like linear regression, because as we can see from the figure 3, it is hard to draw a linear decision boundary to separate the two classes. As the paragraph mentioned above, two features are selected from the 64 feature, which means we have a relatively small model size. Besides, the two feature selected are independence. By calculating the two classes covariance matrix in the data set, we found that for class 0, the

two features covariance is 0.014 and for class 1, the two features covariance is 0.012, which is very close to 0, which means they are independent. Bayes classifier has been chosen.

$$\sigma_1 = \begin{bmatrix} 0.39 & 0.014 \\ 0.014 & 0.707 \end{bmatrix}$$

$$\sigma_2 = \begin{bmatrix} 0.445 & 0.012 \\ 0.012 & 0.581 \end{bmatrix}$$

- **Naive Bayes**. The Naive Bayes algorithm[8]is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set.

$$P(X) = \sum_k P(X|Y = Y_k)P(Y_k), where \sum_k P(Y_k) = 1$$

$$P(Y_k|X) = \frac{P(X|Y_k)P(Y_k)}{\sum_k P(X|Y = Y_k)P(Y_k)}$$

  Here, $P(Y_k|X)$ is posterior probability, while $P(Y_k)$ is the prior probability associated with hypothesis Y_k.

4 Result

This section mainly includes two parts: pre-processing result and model prediction result. For the pre-processing, feature selection based on extra tree is used. According to figure 4, we can see that the top 2 features, which mostly related to the truth value are feature 27 and feature 21. The correlation between feature 21 and feature 27 is 0.0015, which is very close to 0 meaning that the two feature is independent. By partitioning the correlation map of each two features, feature 13 and feature 27s correlation factor is the smallest, which is 0.00017. Figure 5 shows the sample points after feature selection ((feature 21,27),(feature 13,27)). In both figures, the class 1 distribution is intensive because we over-sampling the class 1 which means we duplicate points in class 1 to increase its weight.

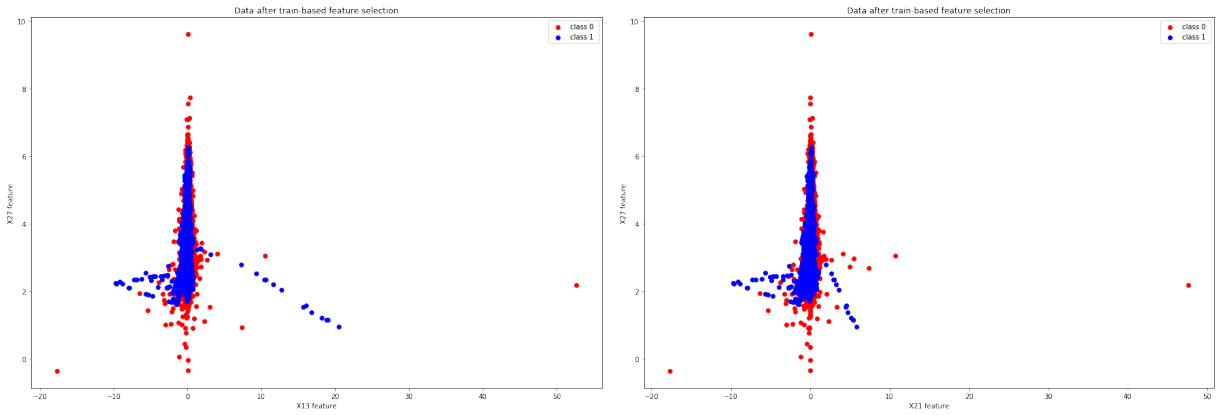

Figure 5: Comparison of Feature selection

For model prediction part, we use all the points in the data set as training data and using 10000 data (100*100) as testing data to draw the decision boundary showed in figure 6. From the decision boundary for each graph, it is hard to tell which feature pair has better performance.

After that, we use AUC-ROC curve performance measurement to tell the ability of the classifier to distinguish between classes. The bigger the AUC is, the better prediction ability for classifier. From figure 7, we can see X13 and X27 features has slightly bigger AUC than feature 21,27.

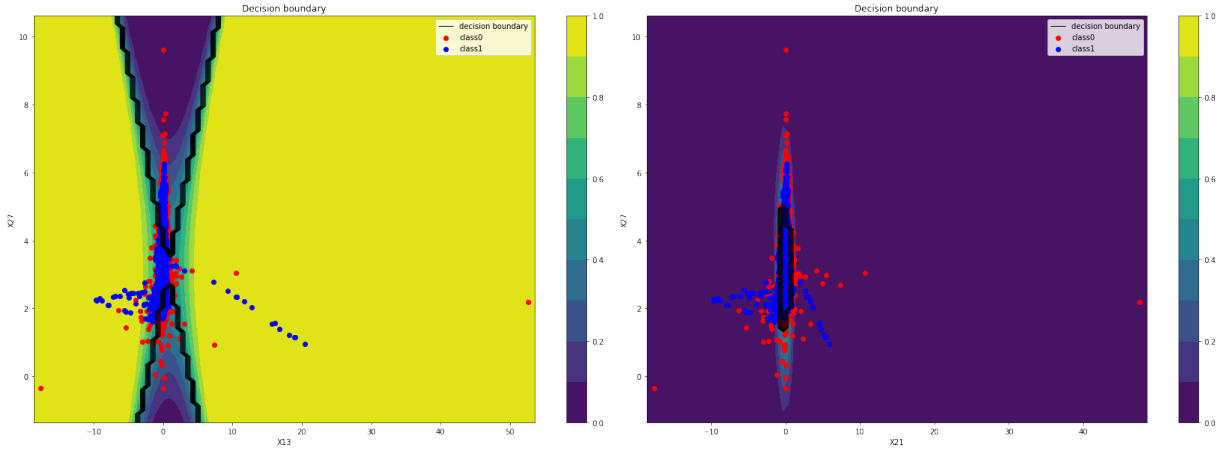

Figure 6: Comparison of Decision Boundary

To validate the stability of machine learning model is important to results. The AUC-ROC curve only specifies the performance of model working on the training dataset. It is possible that the model is over-fitting or under-fitting, which means that it is hard to tell whether the model can be generalized to an unseen data set. Here we use K-fold cross validation to

evaluate the model. The data set would be partitioned into K folds.

Each time, our of the K fold would be used for testing while the other K-1 folds are used for training. The process would be repeated K times and K times result would be averaged. As figure 8 shows, for 5-fold validation, X13 and X27 features still have better performance than feature 21,27.

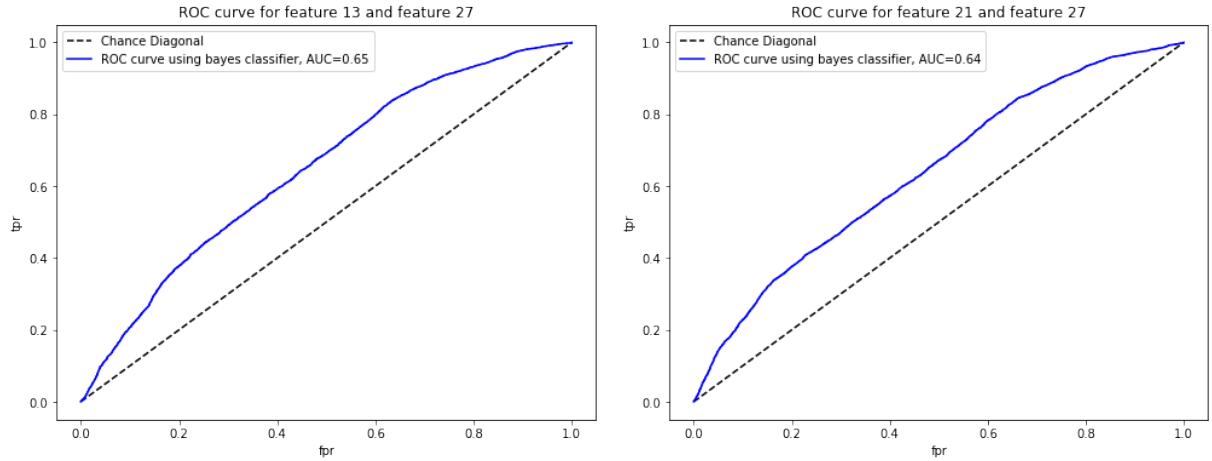

Figure 7: ROC comparison

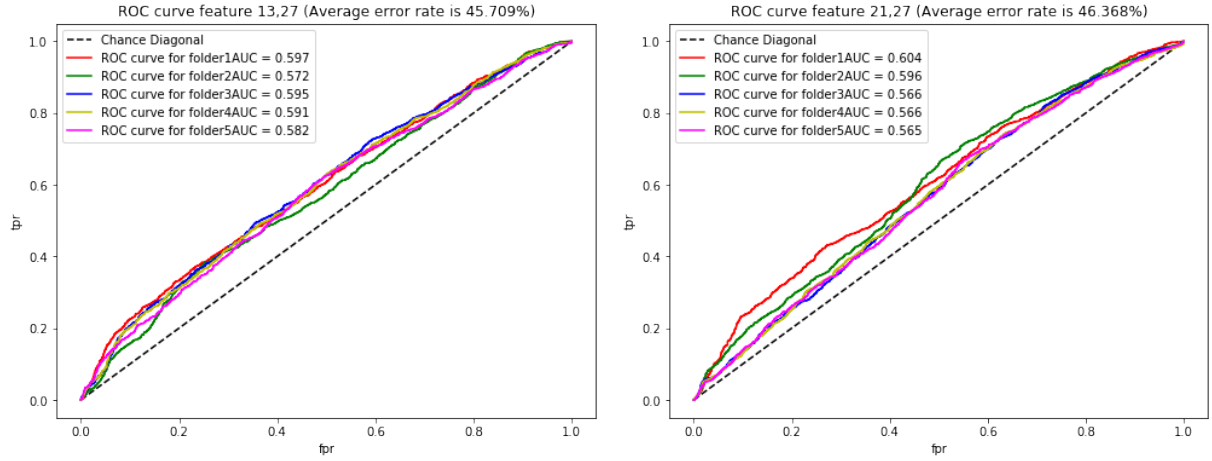

Figure 8: Cross validation comparison

5 Conclusion

- **Feature evaluation**
  It is true that we can also use feature reduction methods like PCA to pre-process

the dataset reducing the dimension. Figure 8 shows a curve for correlation between number of dimensions we use and the accuracy rate. It is obvious that even we choose the max point of the point (36 features), the accuracy only achieve 12 percent, which is quite small. In this case, instead of using feature reduction, we use tree-based feature selection algorithm to handle the multi-dimension problem.

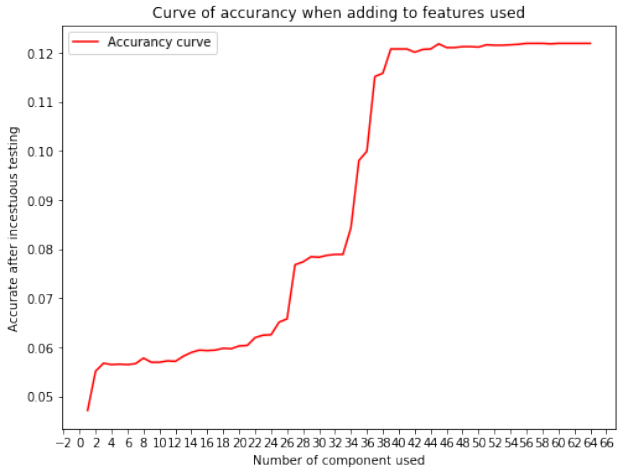

Figure 9: PCA feature reduction and accuracy

- **Key factor to strong performance**
  There are two factors to make a pipeline to a strong performance. The first one is that instead of deleting all the missing data point, we use mean calculation to replace the missing data. In this case, the data set has large training sample points. The second one is that instead of copying the minority of data points directly, SMOTE over-sampling method is used so that we take sample data points around the target data sample into consideration.

- **Factors may lead to degrade of performance**
  To begin with, even though we use SMOTE to increase the weight of class 1(minority class), the two classes still have large area of intersection and it is hard to separate them. Besides, we only select two features to train our model for visualization and computation. This may lead to performance degradation. lastly,even though we compare the result of PCA (a representative method of feature reduction) with extra-tree based feature selection method, it is not validate to say that feature reduction is not suitable for this data set.

References

[1] Frank Wagenmans. Machine learning in bankruptcy prediction. Master's thesis, 2017.

[2] Hafiz A Alaka, Lukumon O Oyedele, Hakeem A Owolabi, Vikas Kumar, Saheed O Ajayi, Olugbenga O Akinade, and Muhammad Bilal. Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, 94:164–184, 2018.

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

[4] Nadzurah Zainal Abidin, Amelia Ritahani Ismail, and Nurul A Emran. Performance analysis of machine learning algorithms for missing value imputation. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 9(6):442–447, 2018.

[5] Grigorios Papageorgiou, Stuart W Grant, Johanna JM Takkenberg, and Mostafa M Mokhles. Statistical primer: how to deal with missing data in scientific research? *Interactive cardiovascular and thoracic surgery*, 27(2):153–158, 2018.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[7] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.