Bankruptcy Predictive Analysis

Kalyani Nikure, Bhagyashree Uttareshwar, Manasa Pendyala
Sai Kiran Reddy Lingampally, Akshay Ragi
Department of Computing and Engineering
University of Missouri-Kansas City
Kansas City, Missouri, USA
{kmn6bg, buwfc, mmvmq, slpgk, arrmw}@umkc.edu

ABSTRACT

Bankruptcy prediction, according to Wikipedia, is the practice of forecasting bankruptcy and various indicators of financial distress in public companies. For a long time, bankruptcy prediction has been a fascinating topic in the financial world. Lenders and investors are paying close attention to corporate bankruptcy as one of the main drivers of credit risk. The financial implications of a business's failure cannot be greatly exaggerated. Bankruptcy prediction, on the other hand, has piqued interest for more than a decade and remains one of the most widely discussed topics in economics. The goal is to develop a method for analyzing business bankruptcy that is both reliable and efficient.

The results of the Project Bankruptcy Predictive Analysis are summarized in this report. The objective of this project is to assess the accuracy, performance, and scalability of statistical techniques such as Logistic Regression, Decision Tree, K-nearest neighbors (KNN), Random Forest Classifier, Classification And Regression Trees (CART), and Gradient Boosting. We have used Best Subset Selection to highlight the dataset's most significant features or financial ratios, which are addressed in depth.

The paper further presents the results achieved from the various statistical techniques. We used well known resampling procedure like K-fold Cross Validation to evaluate the model performance over testing data. These results were also used to compare model performance of different techniques. We analyzed that tree-based classification techniques outperformed compared to regression models.

KEYWORDS

Bankruptcy Prediction, Gradient Boosting, Logistic Regression, Random Forest Classifier, Decision Tree, CART, KNN, XGBoost

1 Introduction

Bankruptcy or business loss may have a negative impact on both the company and the global economy. Business professionals, investors, governments, and academic researchers have long looked at ways to predict the likelihood of a company failing, with the aim of reducing the financial losses associated with bankruptcy. The financial crisis of 2007 illustrated the importance of market stability. Predictive techniques are needed to help forecast such incidents in order to avoid a repeat of the disaster.
Bankruptcy forecasting is a form of evaluating bankruptcy that is a different predictor of government entities' financial difficulties. This is an excellent area in which to research finance and accounting. The field's importance is partly due to founders' and investors' leniency in assessing the possibility of a business going bankrupt. The amount of research done is also a feature of data availability; for public bodies that went bankrupt or did not, different accounting ratios that may suggest danger can be measured, as well as a variety of other possible explanatory variables [2]. As a result, the field is well-suited to analyzing increasingly complex, data-intensive forecasting methods. In other words, deciding whether or not a financial institution is linked to a bankruptcy is crucial.

By combining various financial steps, the aim of bankruptcy prediction is to construct an assessment model that can predict an organization's financial condition [1]. In terms of long-term business operations, determining a company's financial status and future opportunities is also beneficial. In order to make wise lending decisions, financial institutions must be able to reliably predict bankruptcy. Many potential lenders use the credit scoring models to help them choose the factors that allow them to differentiate between good and bad credit risks, such as default or bankruptcy, in order to assess the creditworthiness of the prospective borrowers.

This paper starts with a review of previous work in the field of bankruptcy prediction (§ 2). Following that, we analyze the feature selection and dataset used to train various models (§3), as well as the methodologies used to implement predictive models (§4), and the representation of experimental findings (§5). Finally, we examine the contributions of the authors and draw a conclusion.

2 Related work

Bankruptcy prediction has been an issue of research since 1932, beginning with Fitzpatrick employing a feature set of date, size, and business [3]. His work arranged the inspiration of the contemporary statistical approach within the money prognostication domain.

In 1967, William Beaver applied statistical techniques like t-tests to judge the performance of prediction [4] followed by Edward Altman who effectively applied multiple discriminant analysis and

the use of Altman Z-score that continues to be utilized in today's model [5] that was more developed by others. Incongruity, a great interest was paid to generalized linear models that can be used both in decision making as well as for forecasting.

An increasing number of different predictive methods have been used to develop a more reliable bankruptcy prediction model in recent years as computational techniques and information technology have progressed. One of the most successful models was Support Vector Machine by Shin [6]. Lately, Neural network models and other sophisticated models have been tested on bankruptcy prediction with evolved features like age, bad press, payment incidents from creditors as well.

The latest research involves comparison of various approaches and modeling techniques to ascertain whether any one technique is superior to its counterpart. A novel rule-based system was introduced by Zhang, Wei, and Chan [8] to obtain compressible models in terms of first-order logic with an easy-to-understand knowledge representation. Recently, it has been shown that the ensemble classifier and automatic feature extraction can be successfully applied to the bankruptcy prediction [1] and it significantly beats other methods.

3 Data and Features

This section offers a glimpse into the dataset that was used to test different methodological approaches. After that, we go through the different pre-processing measures that must be completed before we can train models on the dataset.

3.1 About dataset

To evaluate the performance and accuracy of various techniques we are relying on dataset [11] which is hosted on UCI Machine Learning Repository. The selection of data depends on the selection of the database, the research period, the number of companies and the number of financial indicators to be analyzed. For this project, we used data collected from the Emerging Markets Information Service (EMIS)[15], an information database about emerging markets worldwide. The dataset is about bankruptcy forecast of Polish companies. The bankrupt companies were analyzed in the period 2000-2012, while the still operating companies were evaluated from 2007 to 2013.

The dataset consists of 43,400 Instances and it has the 64 financial indicators to be analyzed as features. These 64 attributes consist of Profitability Ratios, Liquidity Ratios, Contribution and Efficiency Ratios, Leverage Ratios, and Other Financial Ratios which are demonstrated in Table 1.

Table 1. Summary of feature in the Polish bankruptcy data

Attr1	net profit / total assets
Attr2	total liabilities / total assets
Attr3	working capital / total assets
Attr4	current assets / short-term liabilities
Attr5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
Attr6	retained earnings / total assets
Attr7	EBIT / total assets
Attr8	book value of equity / total liabilities
Attr9	sales / total assets
Attr10	equity / total assets
Attr11	(gross profit + extraordinary items + financial expenses) / total assets
Attr12	gross profit / short-term liabilities
Attr13	(gross profit + depreciation) / sales
Attr14	(gross profit + interest) / total assets
Attr15	(total liabilities * 365) / (gross profit + depreciation)
Attr16	(gross profit + depreciation) / total liabilities
Attr17	total assets / total liabilities
Attr18	gross profit / total assets
Attr19	gross profit / sales
Attr20	(inventory * 365) / sales
Attr21	sales (n) / sales (n-1)
Attr22	profit on operating activities / total assets
Attr23	net profit / sales
Attr24	gross profit (in 3 years) / total assets
Attr25	(equity - share capital) / total assets
Attr26	(net profit + depreciation) / total liabilities
Attr27	profit on operating activities / financial expenses
Attr28	working capital / fixed assets
Attr29	logarithm of total assets
Attr30	(total liabilities - cash) / sales
Attr31	(gross profit + interest) / sales
Attr32	(current liabilities * 365) / cost of products sold
Attr33	operating expenses / short-term liabilities
Attr34	operating expenses / total liabilities
Attr35	profit on sales / total assets
Attr36	total sales / total assets
Attr37	(current assets - inventories) / long-term liabilities
Attr38	constant capital / total assets
Attr39	profit on sales / sales
Attr40	(current assets - inventory - receivables)/short-term liabilities
Attr41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
Attr42	profit on operating activities / sales
Attr43	rotation receivables + inventory turnover in days
Attr44	(receivables * 365) / sales
Attr45	net profit / inventory
Attr46	(current assets - inventory) / short-term liabilities
Attr47	(inventory * 365) / cost of products sold
Attr48	EBITDA (profit on operating activities - depreciation) / total assets
Attr49	EBITDA (profit on operating activities - depreciation) / sales
Attr50	current assets / total liabilities
Attr51	short-term liabilities / total assets
Attr52	(short-term liabilities * 365) / cost of products sold)
Attr53	equity / fixed assets
Attr54	constant capital / fixed assets
Attr55	working capital
Attr56	(sales - cost of products sold) / sales
Attr57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
Attr58	total costs /total sales
Attr59	long-term liabilities / equity
Attr60	sales / inventory
Attr61	sales / receivables
Attr62	(short-term liabilities *365) / sales
Attr63	sales / short-term liabilities
Attr64	sales / fixed asset

3.2 Data Preparation

The aim of the experiment was to identify the best classification model for each bankruptcy prediction case represented by the dataset described in the previous subsection. The Jupyter Notebook tool was used to try and model different techniques using python language. We also used RStudio with R Programming Language to implement tree-based classification models. The implementation was done in following phases as discussed:

3.2.1 Pre-Processing. In very first steps after reading the dataset which was in csv format to our local code, we identified the null or unnamed attributes in the dataset which we cleaned as part of pre-processing activities. Another set of steps involved converting string data into float values.

3.2.2 Modelling. This phase was conducted after importing all the required libraries like `pandas` to perform dataset operations, `plotly.express` to plot graphs and `sklearn` library functions were used to implement various models and also to split data into train and test.

3.3 Missing Data and Class Imbalance

We discovered that our dataset had a large number of missing values across all functions. Missing data will inject a lot of prejudice into our learning when we're studying. It can also have an impact on the precision and performance of our models. Dropping such information is also dangerous for the reasons mentioned above. Mean, Median, Nearest Neighbors, and Multivariate Imputation were all options. We considered using Mean Imputation for our problem as it reduces the correlations involving the feature variable been imputated and replaces NA values with central statistics of the respective columns. We used the Imputer class from the `scikitlearn` library to achieve mean imputation.

Figure 1: Class imbalance in existing dataset

The figure 1 above depicts the class imbalance in the dataset which has we have total 43,401 instances(firms), out of which 2,091(4.82%) represents bankrupted companies, 41,310 (95.18%) firms that did not bankrupt in the forecasting period.
The figure 2 below shows the balanced class distribution in the dataset where bankrupt (Class 1) and non-bankrupt (Class 0) data have equal distribution of up to 50%. This is the result we received after performing random over sampling [19] which involves randomly selecting examples from the minority class, with replacement, and adding them to the training dataset.

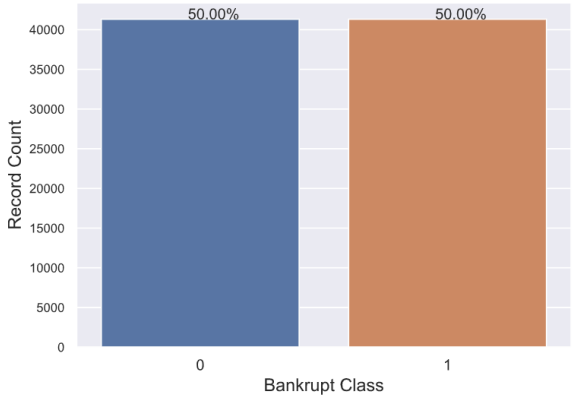

Figure 2: Balanced Class distribution after performing central imputation (Random Over Sampling)

3.4 Training and Testing Data

We choose and split the 5-year forecasting data randomly into 70% for training and hold out 30% data for testing. Each test data instance represents 64 economic indicators and bankruptcy status of various firms. In training data, we have total 30,381 instances(firms) whereas in test data, we have 13,020 instances.

3.5 Understanding data

We plotted various scatterplots of important attributes from the dataset to understand more about the data distribution for the Bankrupt versus non bankrupt data. The figure 3. below shows the distribution of sales vs equity for bankrupt and non-bankrupt companies. We see that very few bankrupt companies have stable values of these factors. Mostly they have low values of equity ratios.

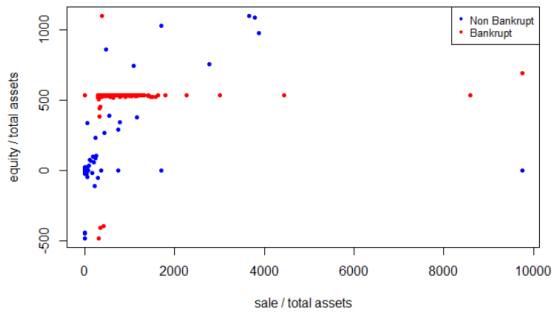

Figure 3: Sales and Equity

The figure 4. shows the distribution of gross profit now and in 3 years for bankrupt and non-bankrupt companies. We see that bankrupt companies have so many outliers compared to non-bankrupt companies. Though bankrupt companies have good gross profit ratios they may have gone bankrupt due to other factors like their total assets are very low compared to the gross profit.

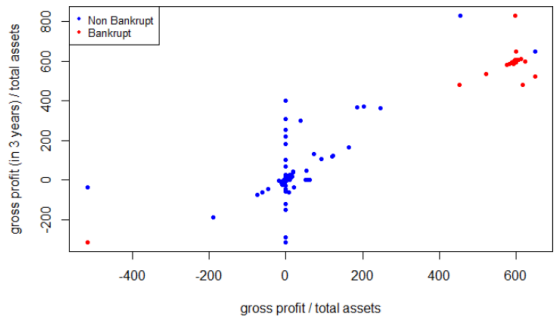

Figure 4: Gross profit now and in 3 years

The figure 5. below shows the distribution of EBIT (Earnings Before Interest and Taxes) and profit on operating activities for bankrupt and non-bankrupt companies. We see that bankrupt companies have high ratios of profit and EBIT compared to non-bankrupt companies. One of the possible reasons is that they have very fewer total assets which lowers a firm's stability compared to the operating profit.

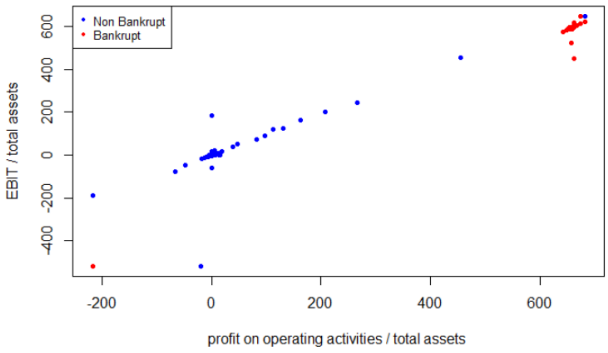

Figure 5: EBIT and profit on operating activities

The figure 6. below shows the distribution of working capital and total liabilities for bankrupt and non-bankrupt companies. We observe that bankrupt companies have very less working capital which lowers down their stability.

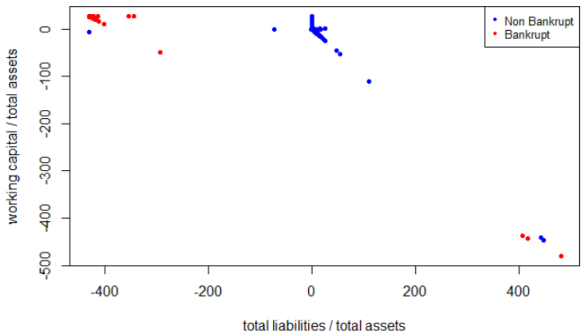

Figure 6: Working capital and total liabilities

A correlation matrix in Figure 7. is simply a table which displays the correlation coefficients for all the features from the dataset. The matrix depicts the correlation between the attributes where the correlation coefficient has a value between -1 and 1 where: -1 indicates a perfectly negative linear correlation between two variables. 0 indicates no linear correlation between two variables. 1 indicates a perfectly positive linear correlation between two variables. This is a nice summarization of a large dataset and to identify and visualize patterns in the given data.

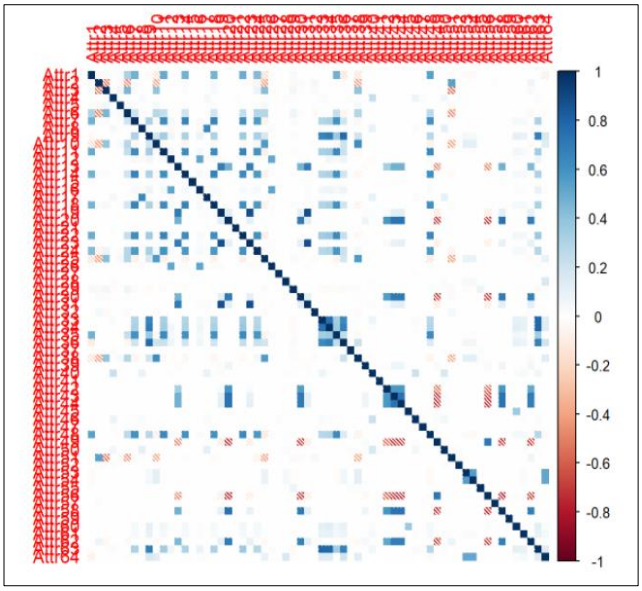

Figure 7: Correlation Matrix between all the 64 attributes of the dataset

4 Methodology

This section provides a detailed view of methodological implementation. We first define techniques used in the project (ȣ4.1) and introduce their functionality in (ȣ4.1.1) and subsections. We further discussed more on feature selection approach in (ȣ4.2).

4.1 Techniques Used

Different models are adopted to accurately predict the bankruptcy. The right choice of models affects the accuracy which can be decided by testing various techniques.[9] We implemented below stated models as part of bankruptcy prediction.

4.1.1 Logistic Regression. Logistic Regression is regression analysis to conduct when the dependent variable that is output data is categorical or binary. Logistic regression is a predictive analysis used to describe data and to explain the relationship between one dependent variable and one or more independent variables. Also called a generalized linear model.

4.1.2 K-nearest neighbors (KNN). K-Nearest Neighbors is the classification and regression algorithm based on the clustering. It's an example of instance-based learning in that it doesn't attempt to build a general internal model, instead storing instances of the training data. Prediction is based on a simple majority vote of each point's nearest neighbors: a question point is allocated to the information class with the highest recurrence.

The k-Nearest-Neighbors (kNN) method of classification may be a simple but efficient one. The most significant disadvantages of

relevancy kNN are its low performance - being a lazy learning process, it is not suitable for many applications such as dynamic web mining for a large repository - and its reliance on the selection of a "reasonable value" for k.

4.1.3 Decision Tree. A decision tree is a decision-making method that employs a tree-like model of choices and potential outcomes, such as natural disaster outcomes, resource costs, and utility. It's a method to display an algorithm that only contains conditional control statements which consists of three styles of nodes which are Decision nodes, Chance nodes, and End nodes.

4.1.4 Random Forest Classifier. It is a type of ensemble algorithm. Instead of relying on 1 decision tree. Random Forest Classifier combines multiple decision trees from a randomly selected subset of the training set. It then combines votes from different decision trees to decide the final class of the test object.

4.1.5 Gradient Boosting. It is one of the best possible models and a very high-performing model or algorithm. It is robust, very easy to use and it improves the performance of the algorithm by reducing over-fitting. This model also minimizes the prediction error compare to logistic regression and the random forest classifier model.

4.1.6 Classification And Regression Tree (CART). Decision Trees are commonly employed in data processing with the target of making a model that predicts the worth of a target (or dependent variable) supported the values of several input (or independent variables). The CART or Classification & Regression Trees methodology is as an umbrella term to sit down with the subsequent forms of decision trees as shown in Figure 8:

- Classification Trees: The target variable is categorical and also the tree is employed to spot the "class" within which a target variable would likely make up.
- Regression Trees: The target variable is continuous and tree is employed to predict it's value.

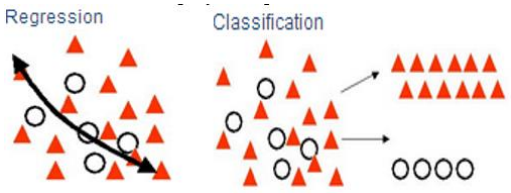

Figure 8: Classification and Regression Trees

4.1.7 XGBoost. It is an optimized distributed gradient boosting library designed to be highly reliable, efficient, flexible and portable, distributed machine learning system. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also called GBDT, GBM) that solves many data science problems in a fast and accurate way.

4.2 Feature Selection

Prediction, like any other machine learning problem, is heavily reliant on the availability of the right amount of data to train it accurately. For forecasting, we use data curated by domain experts with economic indicators/features, and we use machine learning methodologies to assess their accuracy and efficiency.

We performed Best Subset Selection on dataset to identify the most significant features that are contributing the most to the decision making. The best subset selection model produced a result of 8 attributes which are as mentioned in Table 2. The table provides a list of all 8 important attributes and their financial ratio term [16]. Description of those terms will educate to understand the impact of those ratios on the company.

Table 2: Result achieved from Best subset selection of size 8 containing important features from the dataset

Attribute	Meaning	Financial Ratio Term	Description
Attr29	logarithm of total assets	Firm Size	Firm size is measured using the logarithm of total assets.
Attr3	working capital / total assets	Working Capital to Assets ratio	Working Capital is the difference between current assets and current liabilities, so the Working Capital to Total Assets ratio determines the short-term company's solvency.
Attr1	net profit / total assets	Return On Assets (ROA)	Return on assets (ROA) is an indicator of how profitable a company is relative to its total assets. ROA gives a manager, investor, or analyst an idea as to how efficient a company's management is at using its assets to generate earnings.
Attr46	(current assets - inventory) / short-term liabilities	Quick ratio	The quick ratio measures a company's capacity to pay its current liabilities without needing to sell its inventory or obtain additional financing.
Attr33	operating expenses / short-term liabilities	Operating Cash Flow Ratio	The operating cash flow ratio is a measure of how readily current liabilities are covered by the cash flows generated from a company's operations. This ratio can help gauge a company's liquidity in the short term.
Attr39	profit on sales / sales	Profit Margin Ratio	It is called as the return on sales ratio or gross profit ratio, is a profitability ratio that measures the amount of net income earned with

			each dollar of sales generated by comparing the net income and net sales of a company. Profit margin ratio shows what percentage of sales are left over after all expenses are paid by the business.
Attr63	sales / short-term liabilities	Current Asset-to-Short-Term Debt Ratio	The current asset-to-short-term debt ratio provides a measure of whether a company would be capable of making payments on its short-term debt using only the value of its current assets. Ratios greater than one reflect favorably on the company; ratios less than one suggest that the company may be insolvent
Attr4	current assets / short-term liabilities	Current Ratio	The current ratio is a liquidity ratio that measures a company's ability to pay short-term obligations or those due within one year. It tells investors and analysts how a company can maximize the current assets on its balance sheet to satisfy its current debt and other payables.

5 Results

The experiments were conducted to compare the model performance for the dataset containing financial information of polish companies

5.1 Cross Validation

Cross-Validation could be a statistical procedure of evaluating and comparing learning algorithms by dividing data into two segments: one accustomed learn or train a model and the other used to validate the model. In cross-validation, the training and validation sets must cross-over in successive rounds such each data point encompasses a chance of being validated against.

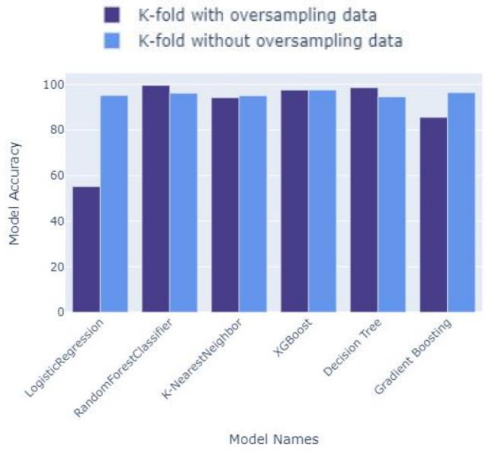

Figure 9: K-fold Cross validation accuracies over models

We have used k-fold cross-validation [17] with k value of 10 for performance estimation, model selection, and tuning learning model parameters used. Figure 9 shows the comparison of results achieved for various models with K-fold cross validation with/without oversampling the data.

5.2 Confusion Matrix

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is that the number of target classes. The matrix compares the particular target values with those predicted by the machine learning model. This provides us a holistic view of how well our classification model is performing and what styles of errors it's making. For a binary classification problem, we'd have a 2 x 2 matrix as shown below with 4 values:

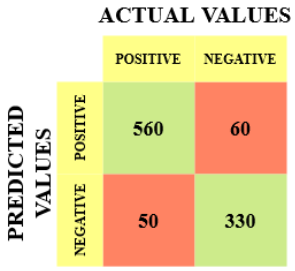

Figure 10: Confusion Matrix Classification Parameters

The different values of the Confusion matrix would be as follows:
True Positive (TP) = 560; 560 positive class data points were correctly classified by the model
True Negative (TN) = 330; 330 negative class data points were correctly classified by the model
False Positive (FP) = 60; 60 negative class data points were incorrectly classified as belonging to the positive class by the model
False Negative (FN) = 50; 50 positive class data points were incorrectly classified as belonging to the negative class by the model.

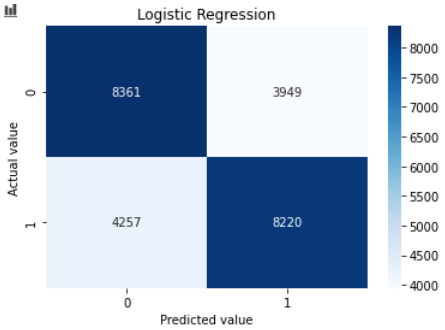

Figure 11: Confusion Matrix result for Logistic Regression

$$\frac{8361 + 8220}{8361 + 8220 + 3949 + 4257} * 100 = 66.89\%$$

Figure 11 shows the Confusion matrix [18] for the Logistic regression model with the accuracy rate of **66.89%** which is calculated as in above equation.

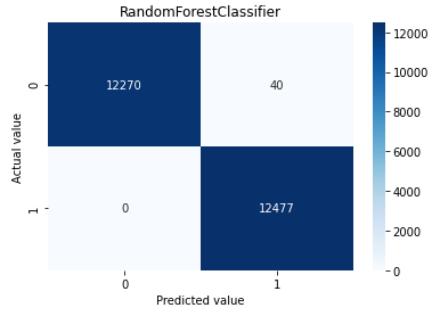

Figure 12: Confusion Matrix result for Random Forest Classifier

Figure 12 shows the Confusion matrix for the Random Forest Classifier model with the accuracy rate of **99.86%.**

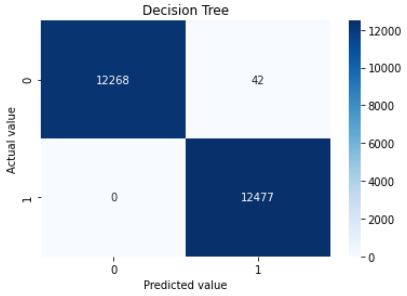

Figure 13: Confusion Matrix result for Decision tree

Figure 13 shows the Confusion matrix for the Decision tree model with the accuracy rate of **99.83%.**

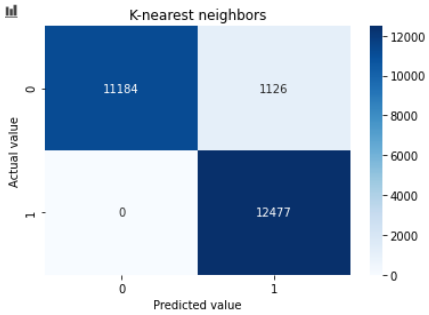

Figure 14: Confusion Matrix result for K-nearest neighbors

Figure 14 shows the Confusion matrix for the K-nearest neighbors' model with the accuracy rate of **95.45%.**

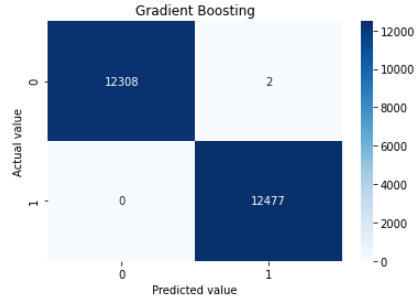

Figure 15: Confusion Matrix result for Gradient Boosting

Figure 15 shows the Confusion matrix for the Gradient boosting model with the accuracy rate of **99.99%.**

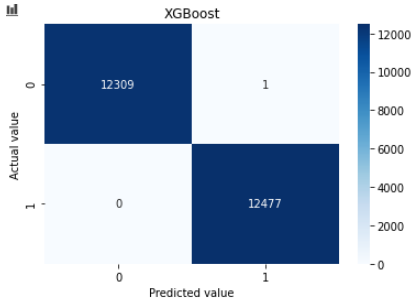

Figure 16: Confusion Matrix result for XGBoost

Figure 16 shows the Confusion matrix for the Gradient boosting model with the accuracy rate of **99.99%.**

5.3 F1 Score

The F1 score is a better measurement than accuracy since it is the harmonic mean of precision and recall values. It is also known as the F Score or the F Measure in this country. To put it another way, the F1 score conveys the balance between the precision and the recall.

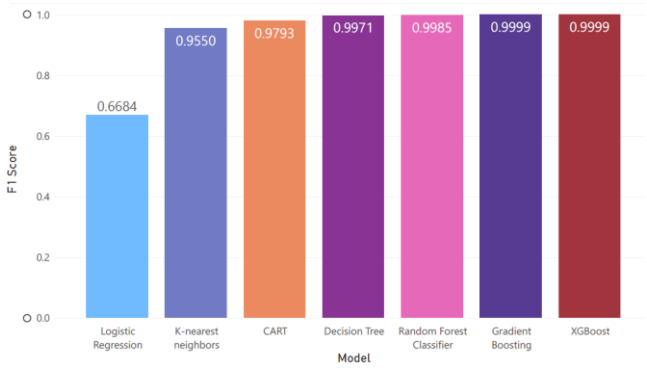

Figure 17: Comparison of F1 Score of models under evaluation

Figure 17 above shows the F scores for the models under evaluation which also stands for how fit the model is. However, Random Forest Classifier, Gradient Boosting, and XGBoost performs the best with higher f1 scores compared to other models.

5.4 Results Comparison

After performing series of experiments over different approaches, we have achieved a good level of testing and training accuracy for the models under evaluation.

Figure 18 depicts the percentage of results achieved for each test and training dataset for respective model. We can conclude from the figure that testing accuracies dropped compared to the training accuracies but the loss was not significant and also statistically it is a correct behavior. Hence, we can say that Decision Tree, Random Forest Classifier, and XGBoost seems to be a better fit for bankruptcy prediction.

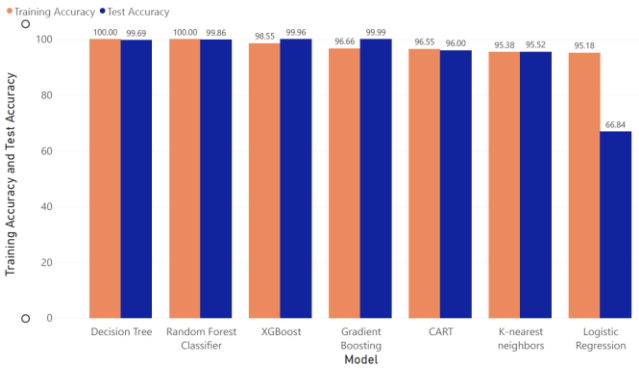

Figure 18: Comparison of Training and Test Accuracies of models under evaluation

6 Conclusion

The project evaluated approaches for the problem of predicting the bankruptcy based on the accuracy achieved for various models. We took under consideration the financial ratios/features which are contributing the most towards bankruptcy prediction. To solve the classification problem, we applied different models like KNN, Gradient Boosting, and Random Forest Classifier. The results gained best accuracy for the Random Forest Classifier, XGBoost, and Gradient Boosting were significantly better than the results gained by Logistic Regression and others. Furthermore, we can say that mostly the classification tree-based models performed well compared to the regression models. Hence, companies can well rely on the proposed novel models to forecast bankruptcy.

AUTHOR CONTRIBUTIONS

Bhagyashree executed most of the models and tested the model's accuracy using k-fold cross validation. Additionally, Kalyani implemented few other models like CART, Random Forest Classifier to the existing codebase and created few data understanding plots. Manasa added XGBoost model in the codebase. Sai Kiran worked on data pre-processing like random oversampling. Akshay pitched in to calculate F1 Scores for models and collected information of all the related works. We maintained the code on GIT Repository [12] in order to have integration platform. All the team members collaborated to create the final report of the project.

ACKNOWLEDGMENTS

A special thanks to our Professor Dr. Uddin for his valuable guidance and critical feedback to improve on the model's accuracy which benefitted us for the selection of right approach. His good suggestions on data visualization, and identification of a few research challenges provided us opportunity to improve on the project report. We also tried to amend the review comments received for the Mid-report and enhanced the improvement areas pointed by the reviewers.

REFERENCES

[1] Zieba, M., S. K. Tomczak, and J. K. Tomczak. 2016. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. In *Proceedings of the Expert Systems with Applications Volume 58, 1 October 2016, Pages 58:93–101.* DOI: https://dl.acm.org/doi/10.1016/j.eswa.2016.04.001

[2] Deron Liang, Chia-Chi Lu, Chih-Fong Tsai and Guan-An Shih. Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. In *Proceedings of the European Journal of Operational Research Volume 252, Issue 2, 16 July 2016, Pages 561-572.* DOI: https://doi.org/10.1016/j.ejor.2016.01.012

[3] FitzPatrick, P. J. (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies. Washington.

[4] Beaver, W. (1966). Financial Ratios As Predictors of Failure. In *Proceedings of the Journal of Accounting Research, 4, 71-111.* DOI: https://doi.org/10.2307/2490171

[5] Edward I. Altman, 1968. Financial Ratios, Discriminant Analysis and The Prediction of Corporate Bankruptcy. In *Proceedings of The Journal of Finance, Volume 23, Issue 4, September, 1968.* DOI: https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[6] Shin, K.S., Lee, T.S., Kim, H.j., 2005. An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, Volume 28, Issue 1, January 2005, Pages 127-135.* DOI: https://doi.org/10.1016/j.eswa.2004.08.009

[7] A. Fan and M. Palaniswami, 2000. Selecting bankruptcy predictors using a support vector machine approach. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, pp. 354-359 vol.6.* DOI: 10.1109/IJCNN.2000.859421

[8] W. Fan, S. Stolfo, J. Zhang, P. Chan, May 1999. AdaCost: Misclassification Cost-Sensitive Boosting. *ICML '99: Proceedings of the Sixteenth International Conference on Machine LearningJune 1999 Pages 97–105.*

[9] Mu-Yen Chen, 2011. Bankruptcy prediction in firms with statistical and intelligent techniques and a comparison of evolutionary computation approaches. In Proceedings of the Computers & Mathematics with Applications Volume 62, Issue 12, December 2011, Pages 4514-4524. DOI: https://doi.org/10.1016/j.camwa.2011.10.030

[10] Shi, Y., & Li, X. (2019). An overview of bankruptcy prediction models for corporate firms: A systematic literature review. *Intangible Capital, Volume 15, Issue 2, April 2019, Pages 114-127.* DOI: https://doi.org/10.3926/ic.1354

[11] The dataset source: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[12] GIT Code repository: https://github.com/Big-data-mgmt/Bankruptcy-Prediction-Analysis

[13] XGBoost Documentation: https://xgboost.readthedocs.org/en/latest/

[14] Random Forest Classifier: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

[15] Emerging Markets Information Service: https://www.emis.com/

[16] Financial Ratios: https://www.investopedia.com/terms/o/operatingratio.asp

[17] k-fold cross validation: https://medium.datadriveninvestor.com/k-fold-cross-validation-6b8518070833

[18] Confusion Matrix Visualization: https://medium.com/@dtuk81/confusion-matrix-visualization-fc31e3f30fea

[19] Resampling strategies: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets