Data Mining

Final Report

Group-16

Forecasting Financial Insolvency

Team Members:

Sreekar Thanda	sreekarthanda@my.unt.edu	11636306
Divya Lakshmi Thota	divyalakshmithota@my.unt.edu	11605208
Pranavi Itharaj	pranaviitharaj@my.unt.edu	11581907
Sindhuvyshnavi Kodakandla	sindhuvyshnavikodakandla@my.unt.edu	11598286
Sai Meghana Reddy Chukka	saimeghanareddychukka@my.unt.edu	11602012

Abstract: The project focuses on predicting bankruptcy risk in companies by analyzing past financial data. Various financial indicators like current assets, liabilities, sales, and return on investments are considered. The goal is to build a model that can accurately predict the risk of bankruptcy in the future based on historical financial information.

Key Machine Learning Models Used: Decision Trees, Random Forests, and K-Nearest Neighbors (K-NN) are the main machine learning models employed in the project. Decision Trees are used to create a sequence of financial factors leading to a risk conclusion. Random Forests improve Decision Trees by combining multiple trees for better predictions. K-NN, on the other hand, is based on the similarity of data points for classification or regression problems.

Model Evaluation and Recommendation: The models are evaluated based on accuracy, with Random Forests showing the best performance with 98% accuracy and the lowest RMSE. Therefore, the recommendation is to use the Random Forest model for predicting bankruptcy risk in the dataset. The models help identify warning signs of financial distress, aiding businesses in making informed decisions at different stages.

Introduction: Predicting Bankruptcy has been one of the most important parts of financial management for many corporations. The risk of bankruptcy has been gradually increasing. The conventional way of handling it must be done only after it takes place. But that might not be a solution to the problem. Predicting it might not only help in future bankruptcy but also help in estimating how the company gets financially impacted. However, in every given scenario, preventing bankruptcy by taking the required precautions beforehand is always preferable to dealing with it after it has already occurred. Therefore, a company's ability to predict its likelihood of bankruptcy is increasingly critical to the success of any enterprise. The art of anticipating bankruptcy also involves gauging the firm's level of financial hardship using a variety of metrics.

So, for this project, we got a dataset that has the financial details of a company as a main feature. The dataset includes features such as Net income, sales, and debts. Therefore, we can analyze the data to see if any patterns emerge, and draw a conclusion based on the facts.

The main goal of this project is to build a model that can forecast the likelihood of bankruptcy in the future by using the information from the previous scenarios. And whether there is a risk or not, a binary classification model would be our fundamental use case. Thus, the anticipated result would be either 0 or 1 indicating no risk as 0 and possible risk as 1.

Our comprehensive pre-processing and visualization methods in this project will be useful in gaining an understanding of the companies' financial situation and performance over time, as bankruptcy is not an overnight event and often begins other warning signs that a company exhibits.

Therefore, the warning signs would help us in identifying the risk and help in making the right decision at various stages. To prevent some variables from dominating the other, feature scaling is being implemented. Upon undergoing such extensive data training, the model progressively gains knowledge. Next, we can assess the model we developed and determine its accuracy—that is, whether it is correctly predicting the risk—using the test data. If so, to what degree of accuracy? We conclude based on the model's performance.

Background: Financial insolvency, or a company's inability to pay its debts, is an existential risk that may result in bankruptcy and have a negative impact on the economy. This strategy not only puts stakeholders in danger of large losses, but it can also spark wider economic unrest that could disrupt entire industries or even bring down recessions.

Predictive analytics offers a proactive solution by examining historical financial data to identify patterns and early warning signs of insolvency. Common indicators include net income, sales, debts, and cash flow. Machine learning models like Decision Trees, logistic regression, and random forests can analyze these features to classify companies based on their insolvency risk.

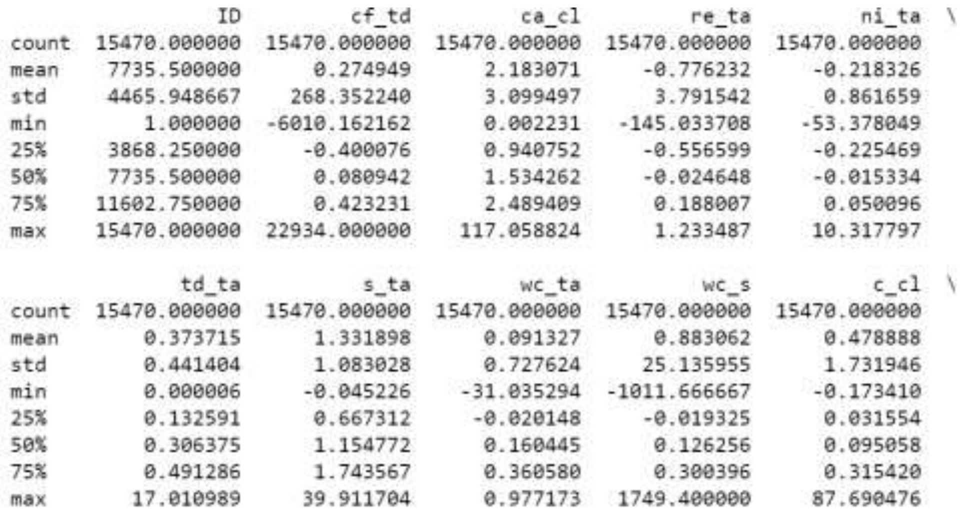
Techniques such as standardization or normalization can prevent certain features from dominating the model's outcomes. When creating these models, feature scaling guarantees that all variables are treated fairly, and test data is used to evaluate the model's accuracy and predictability. By empowering companies to take proactive steps, this predictive method lowers the chance of bankruptcy and promotes economic resilience.

Experiment Methodology: We carefully specified and implemented a cascade of steps that rigorously predict bankruptcy, taking this financial forecasting to an entirely new level. Let’s walk through each step:

Datasets: Our data procurement strategy focused on collecting the historical financial data of various companies from Kaggle. Our data set contained more than 20,000 records with various financial features. We thought it could be used for the prediction of insolvency., such as:

- CF_TD (Cashflow to Total Debts): Reflects the company's ability to generate cash flow relative to its total debts.

- CA_Cl (Current Assets to Current Liabilities): Indicates the ratio of current assets to current liabilities, providing insights into short-term liquidity.
- RE_TA (Retained Earnings to Total Assets): Represents the proportion of earnings retained by the company compared to its total assets.
- NI_TA (Net Income to Total Assets): Measures the company's profitability relative to its total assets.
- TD_TA (Total Debt to Total Assets): Shows the extent to which a company's assets are financed by debt.
- S_TA (Sales to Total Assets): Reflects the company's efficiency in generating sales relative to its total assets.
- WC_TA (Working Capital to Total Assets): Indicates the proportion of a company's total assets tied up in working capital.
- WC_S (Working Capital to Sales): Measures the efficiency of a company's working capital management relative to its sales.
- C_CI (Cash to Current Liabilities): Reflects the company's ability to cover its short-term liabilities with available cash.
- CL_E (Current Liabilities to Equity): Indicates the proportion of current liabilities financed by equity.
- IN_S (Interest Rate to Net Sales): Represents the company's interest expense relative to its net sales.
- MVE_TD (Market Value to Total Debts): Indicates the market value of the company relative to its total debts.

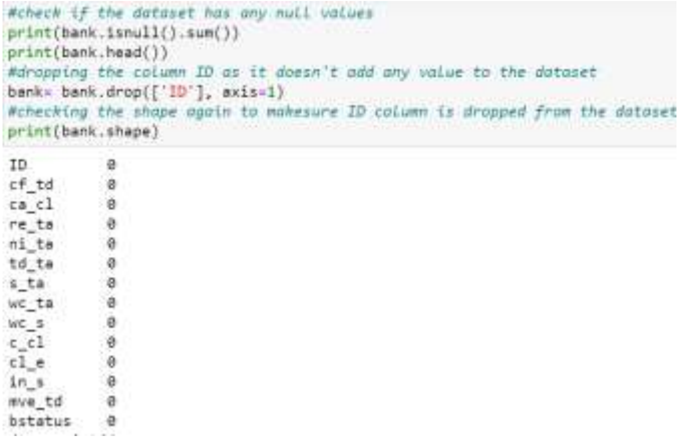


```
                 ID         cf_td         ca_cl         re_ta         ni_ta  \
count  15470.000000  15470.000000  15470.000000  15470.000000  15470.000000
mean    7735.500000      0.274949      2.183071     -0.776232     -0.218326
std     4465.948667    268.352240      3.099497      3.791542      0.861659
min        1.000000  -6010.162162      0.002231   -145.033708    -53.378049
25%     3868.250000     -0.400076      0.940752     -0.556599     -0.225469
50%     7735.500000      0.080942      1.534262     -0.024648     -0.015334
75%    11602.750000      0.423231      2.489409      0.188007      0.050096
max    15470.000000  22934.000000    117.058824      1.233487     10.317797

              td_ta          s_ta         wc_ta          wc_s          c_cl  \
count  15470.000000  15470.000000  15470.000000  15470.000000  15470.000000
mean       0.373715      1.331898      0.091327      0.883062      0.478888
std        0.441404      1.083028      0.727624     25.135955      1.731946
min        0.000006     -0.045226    -31.035294  -1011.666667     -0.173410
25%        0.132591      0.667312     -0.020148     -0.019325      0.031554
50%        0.306375      1.154772      0.160445      0.126256      0.095058
75%        0.491286      1.743567      0.360580      0.300396      0.315420
max       17.010989     39.911704      0.977173   1749.400000     87.690476
```

Figure 1: Feature Statistics

Preprocessing: Before building the model, we preprocessed the dataset carefully so that it is fully cleaned and ready for learning. This includes a few very important steps: Treatment of Missing Values: We used the standard approach for dealing with missing data to mitigate the interference to this model.



```
#check if the dataset has any null values
print(bank.isnull().sum())
print(bank.head())
#dropping the column ID as it doesn't add any value to the dataset
bank= bank.drop(['ID'], axis=1)
#checking the shape again to makesure ID column is dropped from the dataset
print(bank.shape)
```

```
ID         0
cf_td      0
ca_cl      0
re_ta      0
ni_ta      0
td_ta      0
s_ta       0
wc_ta      0
wc_s       0
c_cl       0
cl_e       0
in_s       0
mve_td     0
bstatus    0
```

Figure 2: Identifying Missing values in features

Feature Selection and Engineering: We first identified or carefully engineered features with high correlations to financial insolvency, and then created new features or transformed existing features to further improve prediction.

Robust Outlier Detection and Treatment: Robust outlier detection methods helped detect and mitigate the impact of outliers on model training, an important step to make the training of our models much more stable.

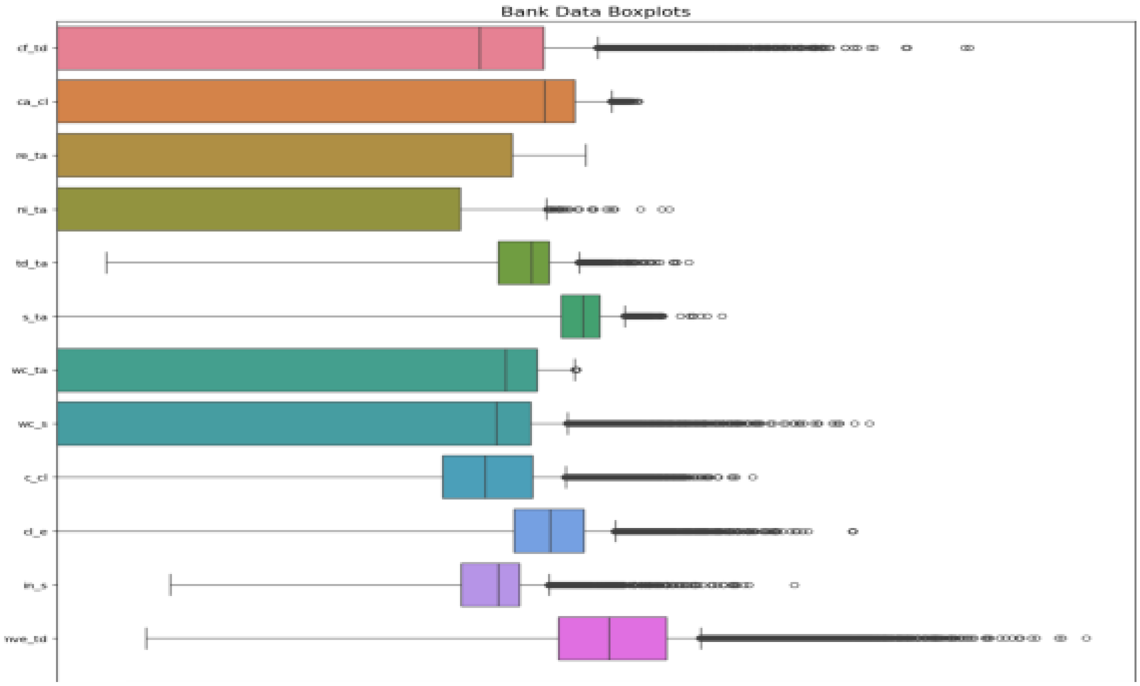


Figure 3: BoxPlots of all features to understand the Outliers

Exploratory Data Analysis (EDA): EDA was essential in finding the hidden signals in how the different variables were related. By using many different visuals, such as charts, histograms and scatter plots, we were able to detect trends, outliers and correlations. This was instrumental in deciding how to proceed with feature selection and model training.

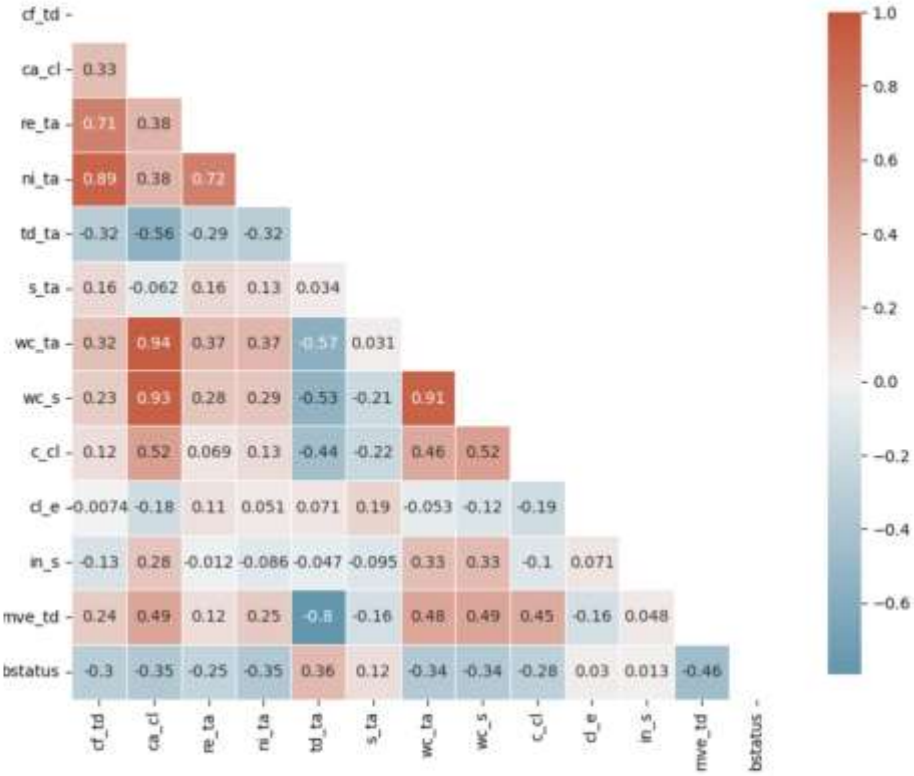


Figure 4: Heatmap displaying correlations between different features

Model Selection and Training: We evaluated different data mining algorithms that are suitable for classification tasks. The main ones used were Decision Trees, Random Forests and K-Nearest Neighbors (KNN). Each algorithm was trained and optimized to achieve the best performance and generalization ability.

Evaluation Metrics: Model performance was evaluated using a set of evaluation metrics appropriate for classification tasks, eg, accuracy, precision, recall, F1 score and area under the ROC curve (AUC-ROC). Confusion matrices detail true positives, true negatives, false positives and false negatives which give a lot of insight to the overall performance of the models. Parameters Tunning: To get the most out of each of these algorithms, we performed parameter tuning, where we tuned parameters like max_depth, min_samples_split, n_estimators for Decision Trees and Random Forests.

Results: The outcome of the experiment gave us striking insights and also provided the strong predictive models for identifying the financial insolvency case. Detailed exploration of finding and inference for the results:

Decision Tree Classifier:

Inference: The Decision Tree model provided good performance with an accuracy of around 83.48 per cent. It was also interpretable, which means that we were able to understand the factors that affect financial distress of firms.

Output Results: The model decision rules, and branching structures, gave actionable insights into the key predictors of insolvency, so that stakeholders could decide on which risk mitigation actions to prioritise.

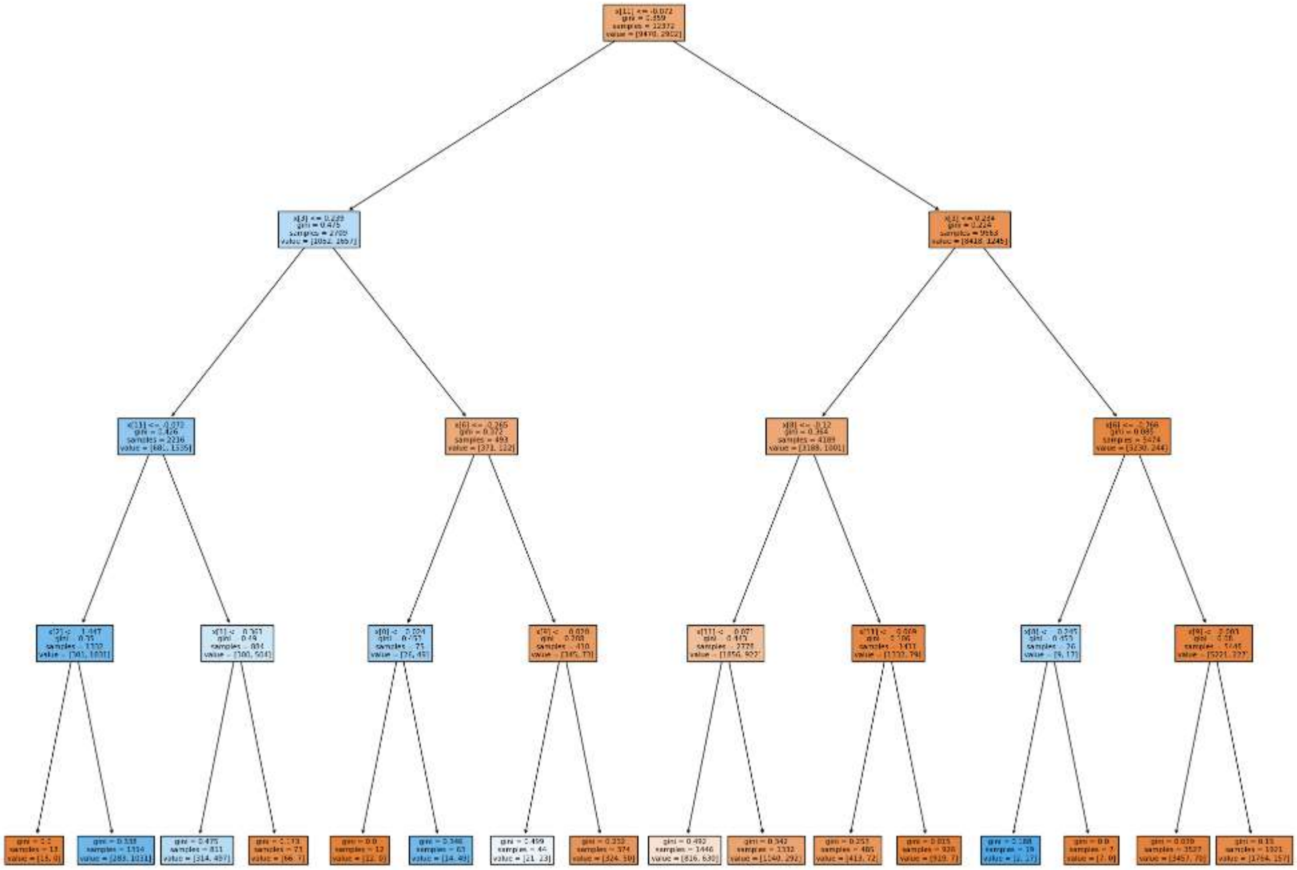
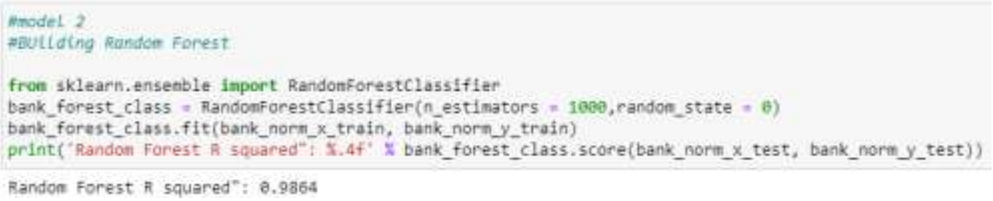


Figure 5: Decision Tree Graph

Random Forest Classifier:

Inference: Random Forest had the highest score of around 98.64 per cent, showing an excellent performance, likely due to its ensemble approach that is less prone to overfitting compared with the other models, both simple and ensemble. This also enabled the Random Forest model to fit the data well in the presence of highly nonlinear patterns.



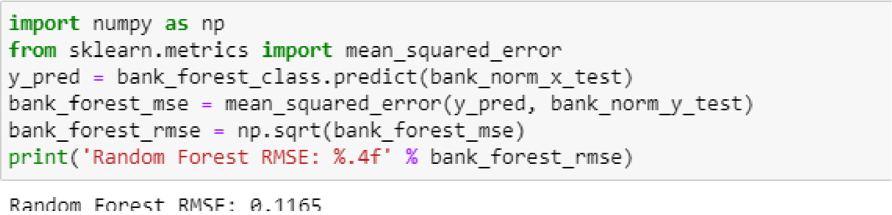
```
#model 2
#Building Random Forest

from sklearn.ensemble import RandomForestClassifier
bank_forest_class = RandomForestClassifier(n_estimators = 1000,random_state = 0)
bank_forest_class.fit(bank_norm_x_train, bank_norm_y_train)
print('Random Forest R squared": %.4f' % bank_forest_class.score(bank_norm_x_test, bank_norm_y_test))

Random Forest R squared": 0.9864
```

Figure 6: Random Forest Model Training

Output Results: In turn, these scores revealed the variables that were most important in determining financial insolvency, allowing stakeholders to focus their mitigation efforts on those critical problems.



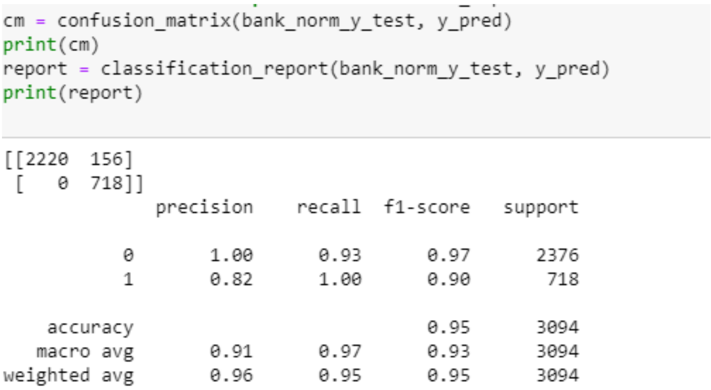
```
import numpy as np
from sklearn.metrics import mean_squared_error
y_pred = bank_forest_class.predict(bank_norm_x_test)
bank_forest_mse = mean_squared_error(y_pred, bank_norm_y_test)
bank_forest_rmse = np.sqrt(bank_forest_mse)
print('Random Forest RMSE: %.4f' % bank_forest_rmse)

Random Forest RMSE: 0.1165
```

Figure 7: Random Forest Model Results

K-Nearest Neighbors (KNN) Classifier:

Inference: The KNN model, while a little less accurate than Random Forest, performed quite competitively with different parameter settings, and because it relies on local similarity measures, it’s responsive to local patterns, patterns with very local details.



```
cm = confusion_matrix(bank_norm_y_test, y_pred)
print(cm)
report = classification_report(bank_norm_y_test, y_pred)
print(report)

[[2220  156]
 [   0  718]]
              precision    recall  f1-score   support

           0       1.00      0.93      0.97      2376
           1       0.82      1.00      0.90       718

    accuracy                           0.95      3094
   macro avg       0.91      0.97      0.93      3094
weighted avg       0.96      0.95      0.95      3094
```

Figure 8: KNN Classifier Model Metrics

Output Results: By exploring different values of k (the number of neighbors), we could find a configuration with the best predictive performance. This insight into parameter sensitivity gave us valuable feedback for future model iterations.

Visualization and Interpretation:

Inference: These techniques enabled us to understand which potential relationships amongst the variables were visible in the data without having to know anything about the context of the data in advance. Box plots, scatter plots, and heatmaps were very helpful for visualizing the underlying pattern and provided clues for understanding the intuitive interpretation of what the model is saying to us, or how the data is distributed.

Output Results: Transparent visualizations of feature distributions, outliers and feature correlations e.g. box plots showing the difference in feature distributions between solvent and insolvent companies, and actionable insights for risk assessment and decision making.

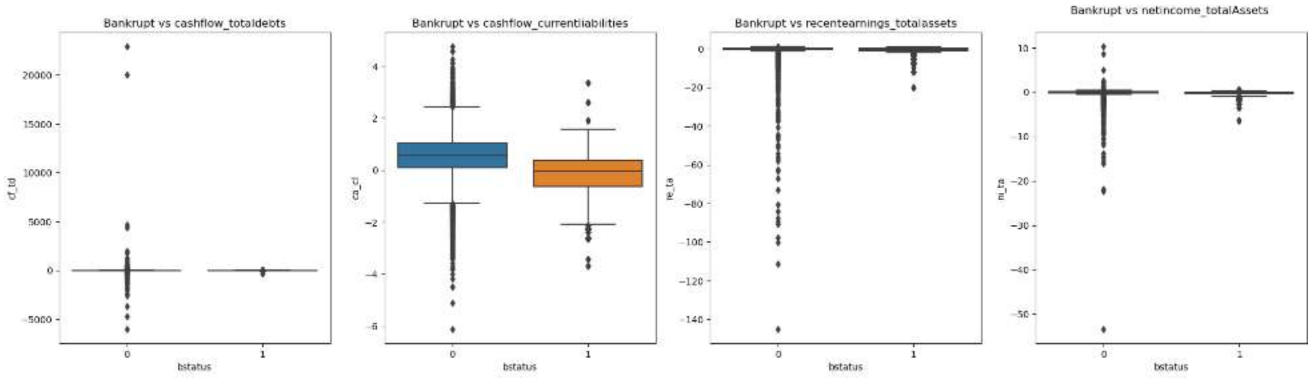


Figure 9: Box Plots of Features displaying Outliers (1)

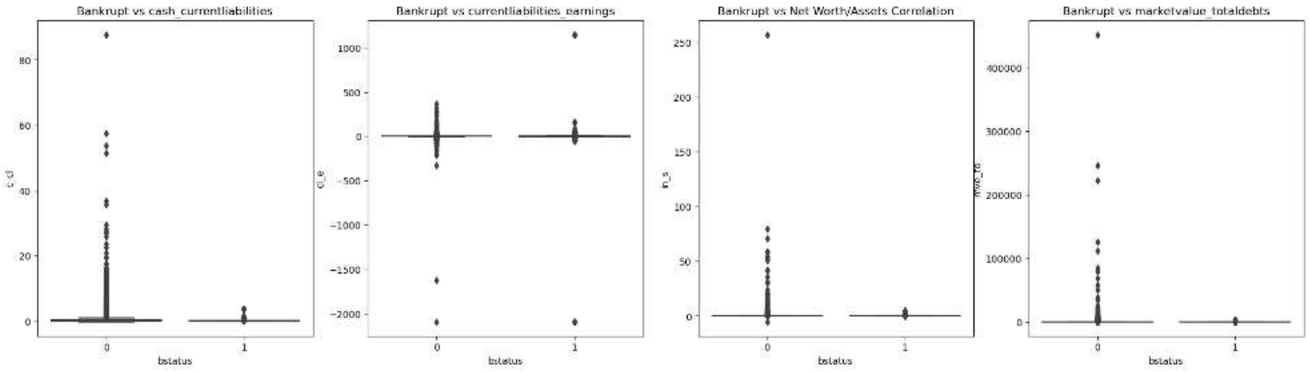


Figure 10: Box Plots of Features displaying Outliers (2)

In conclusion, our experiment proved not only that data mining methods can be successfully used to detect financial insolvency, but also provided new actionable information. The potential of using an integrated approach, based on a specific algorithm, rigorous evaluation metrics and relevant visualizations, is to allow firms to be better informed on how to manage their business, going beyond traditional risk assessment, and prepare for new and unforeseen financial and business challenges.

Related Work:

Financial Insolvency is about the situation when an entity or an individual gets hard in financial status like unable to fulfill their financial bills. We as a team, have studied thoroughly and discussed below papers and decided to bring out an easy model for predicting/forecasting the financial Insolvency. The related work of these research papers is below,

[1]. Explainability of Machine Learning Models for Bankruptcy Prediction:

This study is about the explainability of machine learning models for bankruptcy prediction. The authors used the Local Interpretable Model-Agnostic Explanations (LIME) algorithm to find feature selection importance and compared it with tree-based models. They found that the feature importance measured through LIME is more consistent and effective with that of tree-based models and so could be used for finding appropriate information for credit rating standards. The authors used a dataset of Korean companies in which they found that a random forest model is best performed, with 93.5% accuracy. To explain the feature importance of the model, with cash flow from total debt, income worth to total assets as the most important features for predicting bankruptcy. This research paper suggests that ML models are effective analytical tools for predicting bankruptcy, giving both high accuracy and interpretability.

[2]. Forecasting Bankruptcy More Accurately: A Simple Hazard Model

This paper explains how to use a simple hazard model in place of the static models that researchers and practitioners employ when they forecast bankruptcy. The hazard model described uses all of the available information to determine the relative risk of bankruptcy today for each firm at each point in time, eliminating the potential for selection bias inherent in the static models. We find that it is quite simple to use the hazard model to generate out-of-sample bankruptcy forecasts that are more accurate than those provided by the industry standard, Altman's Z-score model.

[3]. An overview of bankruptcy prediction models for corporate firms: A Systematic literature review.

The study is based on corporate bankruptcy prediction using different models. For this prediction, Logistic regression and Neural Networks are popular models used in the field, followed by discriminant analysis and

Support Vector Machines. As increasing interest in more advanced machine learning models like Adaboost, Case-based reasoning, Particle Swarm Optimization, and K-nearest neighbor which can also be used when Big Data come into picture. These models suggest a growing interest in advanced ML techniques in corporate bankruptcy prediction. The authors investigated about the failures of some business firms and their solutions. Also, authors have included the papers/ studies related to corporate bankruptcy prediction between 1968 and 2017 for review and understand Further collaboration among researchers in this area is suggested to promote advancements in prediction models. The authors concluded that bankruptcy prediction issues is one of growing interest, especially after 2008 global financial crisis and many researches having been going on by reputed organizations.

Our Project is designed in such a way that of predicting the long-term Insolvency- Balance Sheet Insolvency based on analysis of previous data. Our project makes it unique and diverse, since we have used a large dataset (have 15000+ records and 14 features) such that this project principle can be deployed to bigdata with minor integrations and less changes for code. We have deployed the Decision Trees, Random Forest and KNN-Classifier to predict the Insolvency. For this we have worked on various reference works(some of them are given in references)

Conclusion: To predict bankruptcy might be important to the enterprises. They will be able to see the financial disasters at first sight in order to take measures early enough. The admin of a corporate company can collect past data from the group of companies in the same field. We can see common patterns between different companies and predict the possibility of bankruptcy from them in future. Now in our project, we have a task to model an algorithm that can figure out the possibility of a future bankruptcy based on the indicators of a company's financial numbers. Ultimately, by conducting intensive analysis of the structured data (pre-processing, visualization etc. we will deliver companies' financial situations and backgrounds along with their performance. From the models developed such as Decision Tree, Random Forest, KNN, and Nueral networks we can state with confidence that certain models have outperformed the rest for prediction.

Example Random Future score accuracy = 98%, RMSE=0.11

The best model in my opinion is Random future over one test case, Random future model at 98% accuracy and the lowest RMSE of 0.11. I recommend this model for prediction of bankruptcy with respect to the dataset we provided. However, scaling the features an intensive training on them, it provides the model with an opportunity to learn and make prediction at an advanced rate which in return support business with a formidable decision-making in the various stages of the business event. All in all, our project aims to help in establishing warning signs for bankruptcy, allowing for potentially forewarning business enterprises about potential financial pitfalls. Therefore, by utilizing the beneficial features that machine learning implies, as well as examining financial data, it is possible to provide a tool for bankruptcy risk prediction and support to global corporative decision making.

References:

[1] M. Park, H. Son, C. Hyun and H. J. Hwang, "Explainability of Machine Learning Models for Bankruptcy Prediction," *IEEE Access,* p. 13, 2021.

[2] T. Shumway, "Forecasting Bankruptcy More Accurately: A Simple Hazard Model," *The Journal Of Business,* vol. 74, pp. 101-124, 2001.

[3] Y. Shi and X. Li, "An overview of bankruptcy prediction models for corporate firms: A Systematic literature review," *Omnia Science,* pp. 114-127, October 2019.

[4] E. Sfakianakis, "Bankruptcy prediction model for listed companies in Greece," *Investment Management and Financial Innovations,* vol. 18, no. bankruptcy prediction in distressed economies, pp. 166-180, 2021.

[5] A. D. Voda, G. Dobrota, D. M. Tirca, D. D. Dumitrascu and D. Dobrota, "Corporate bankruptcy and insolvency prediction model," *Technological and Economic Development of Economy,* vol. 27, pp. 1039-1056, August 2021.

[6] y. Wu, C. Gaunt and S. Gray, "A comparion of alternative bankruptcy prediction models," *Journal of Contemporary Accounting & Economics,* vol. 6, no. 1, pp. 33-45, 2010.

[7] K. D, V. B, M. S, P. K, T. T and J. D, "Early Insolvency Prediction as a Key for Sustainable Business Growth," *Economic and Business Aspects of Sustainability,* 2023.

[8] A. A. Khundhur and A. I. Al-Alawai, "The use of Machine Learning to Forecast Financial Performance: A Literature Review," *International Conference in Emerging Technologies for Sustainability and Intelligent Systems,* pp. 1-6, 2024.