Project Summary

Batch details	DSE November 2020 - Online
Team members	Apoorva Garg , Pranavi Krishnamsetty, Vishal Pandey, Yash Vahi
Domain of Project	BFSI
Proposed project title	Bankruptcy Prediction
Group Number	3
Team Leader	Yash Vahi
Mentor Name	Mr. Animesh Tiwari

Date: 10th July 2021

Signature of the Mentor

Signature of the Team Leader

Table of Contents

Final Report

Summary

Our goal was to predict whether or not a company would go bankrupt based on various financial ratios. The dataset we were working on contained financial information about companies in Taiwan and their health status(Bankrupt or not). Based on this information, we applied various machine learning algorithms to construct a generalised model that would predict bankruptcy across the board, regardless of the scale of the firm. We worked on this problem with the perspective of a bank. For a bank, giving loans to companies that are not healthy is extremely costly and results in unwanted losses. Naturally, our goal was to predict bankruptcy with accuracy as high as practicable and our focus was mainly on maximising recall at a reasonable rate so that other metrics like precision and f1-score don't get too low. There was a severe imbalance in our dataset with bankrupt companies occupying approximately 3% of the dataset. Using various machine learning algorithms we didn't just predict bankruptcy at a high recall but we also found important financial ratios that have the heaviest impact on whether or not a company will go bankrupt. In this report, we will explain the steps we took to build the models and how we interpreted the results.

Overview of the final process

The final stage included building predictive models using the cleaned data we had after our EDA was complete. The only algorithm for which we used raw data was Decision Tree, as Decision Trees are not sensitive to outliers, multicollinearity and skewed data. We also tried extreme gradient boosting but we did not use raw data for it even though, in practice, the algorithm can handle raw data. The reason we did this is because we wanted control over how the troublesome factors like high outliers, null values and multicollinearity were handled. If we used raw data, our control over these factors would be limited. Other than Logistic Regression, we followed the same steps for nearly all of the models that we built. The reason for this is that we used the Maximum Likelihood

approach for Logistic Regression which is a statistical approach, we took the MLE route because we wanted to interpret the coefficients for the financial ratios. This would help us understand the degree of statistical impact that different financial ratios would have on bankruptcy. In the next section, we will one by one explain all the steps we took in the final process of modelling.

Model Construction

To predict bankruptcy, we tried every supervised machine learning technique we knew of. Here are the list of algorithms that gave us interesting results.

1. Logistic Regression(MLE)
2. Decision Tree Classifier
3. Random Forest Classifier
4. Boosting : Adaptive, Gradient and extreme gradient
5. K-Nearest Neighbours
6. Gaussian Naive Bayes

Logistic Regression

Logistic regression is a classification algorithm built on top of linear regression. We will be going for the Maximum likelihood estimation approach which will help us to interpret the coefficients easily and allow us to establish relationships between the independent and target variable. Our aim is to maximise recall as letting unhealthy companies slip by is more costly than classifying some of the healthy companies as unhealthy.

We followed the following steps for Logistic Regression:

1. Build a full Logistic Regression model using Maximum Likelihood estimation.
2. Plot Receiver Operating Characteristic(ROC) and get Area under Curve(AUC)
3. Build scorecard to see how different metrics like Precision, Recall, F1-score and kappa are performing under different thresholds.
4. Perform recursive feature elimination(RFE).
5. Build a model with features that survived RFE.

6. Repeat steps 2 and 3 for RFE model.
7. Compare both models.

Full model Performance:

1. Our full Logistic Regression(MLE) model gave us an ROC of 0.9555.(*Figure 1.1*)
2. Maximum precision was 0.833 at thresholds 0.9.
3. Maximum Recall was 0.948718 at threshold 0.020431 through Youdens' index.
4. Maximum F1 was 0.483516 at threshold 0.2.
5. Maximum kappa was 0.459989 at threshold 0.2.

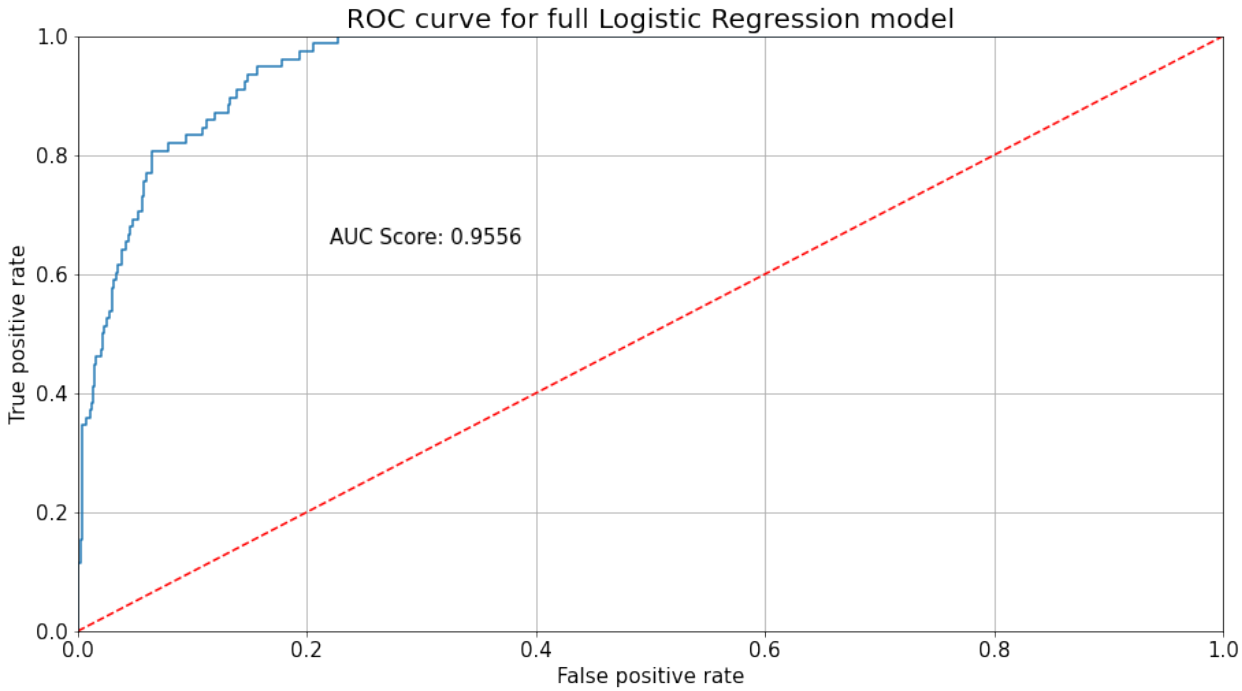


Figure 1.1

RFE model performance:-

1. Our Logistic Regression model after RFE gave us an ROC of 0.9593, which is barely above our full model.(*Figure 1.2*)
2. Maximum precision was 0.8333 at threshold 0.9.
3. Maximum Recall was 0.974359 at threshold 0.020815 obtained through Youdens' index.
4. Maximum F1 was 0.5 at threshold 0.3.

5. Maximum kappa was 0.480183 at threshold 0.3.

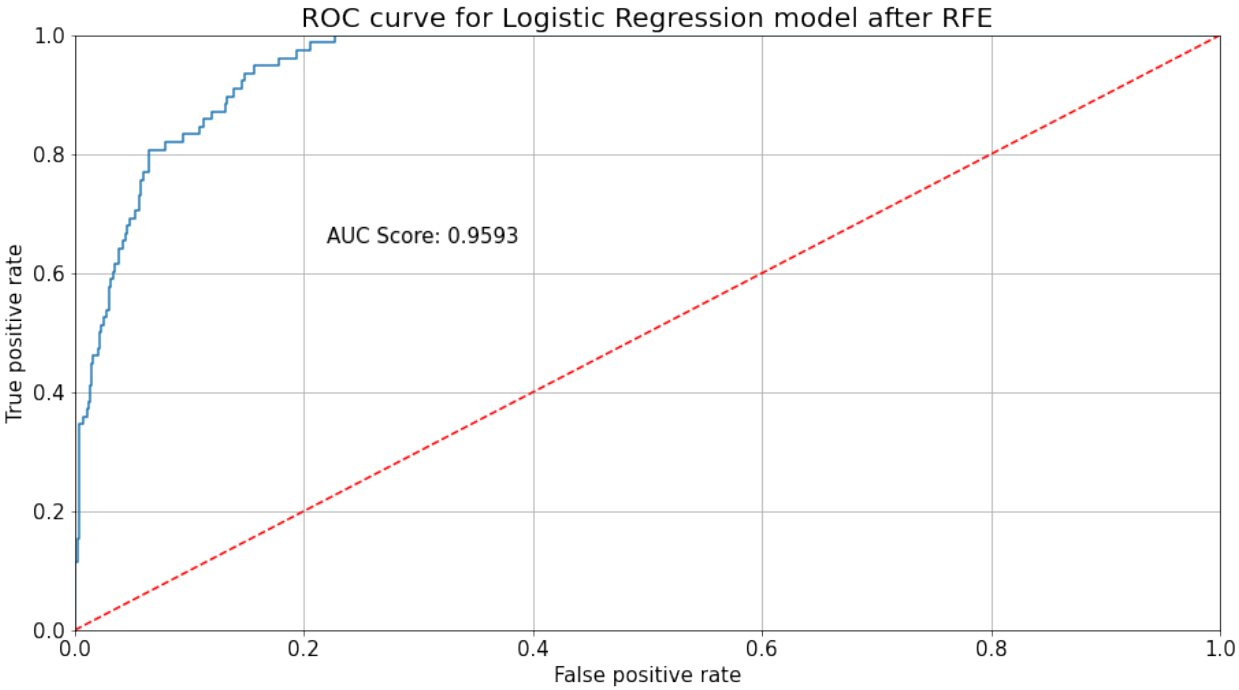
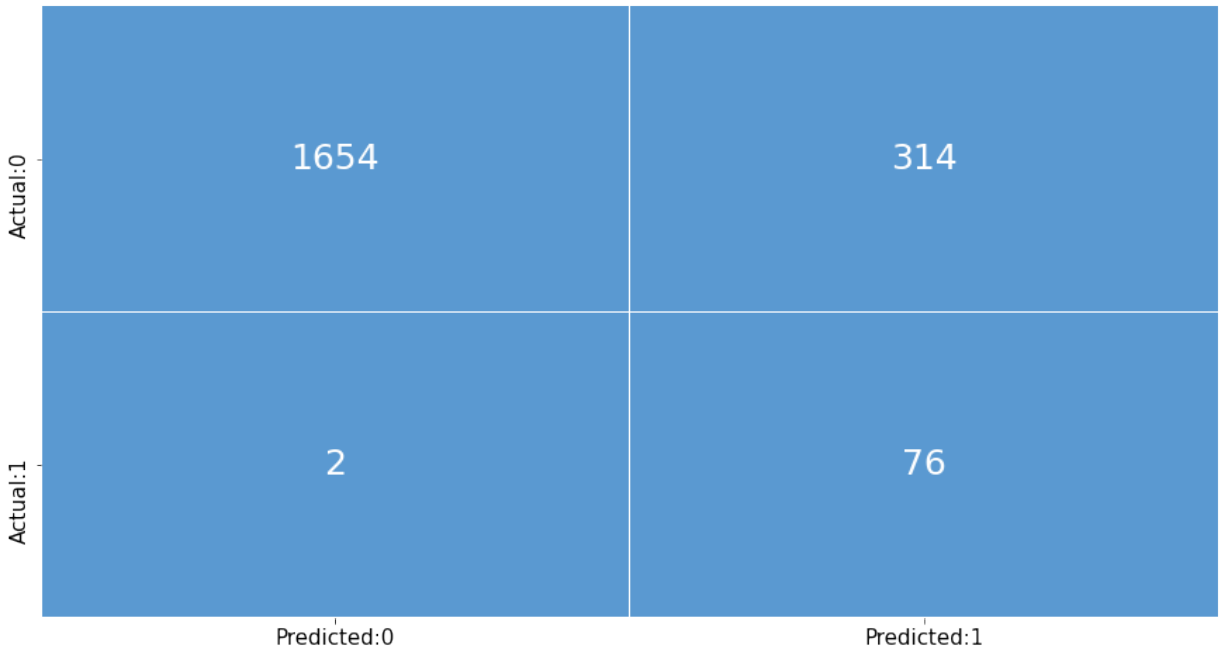


Figure 1.2

Model Comparison:-

As the purpose of our project is to maximise recall and detect as many bankrupt firms as possible, we will select the RFE model as it maximises our recall at 0.974359 at threshold 0.020815. Here's the confusion matrix for it,



Decision Tree, Random Forrest, Boosting, KNN and Gaussian Naive Bayes

After building our logit model and getting great results, we proceeded with the rest of the algorithms. For the rest of the models, we followed the same steps.

Steps:

1. According to model requirement, proceed with the suitable dataset. For all models other than Decision Tree classifier we used the cleaned dataset.
2. Tune hyper-parameters using Random Search.
3. Build model with best hyper-parameters obtained through tuning.
4. Build thresh scorecard to check how metrics like precision, recall, f1, kappa are performing under different thresholds.
5. Calculate best thresh from youdens' index.
6. Pick best thresh and add it model to final scorecard.

After finishing all the models, we had a composite scorecard to compare all of them. (*Figure 2.1*)

	Model Name	Threshold	ROC-AUC	Recall	Precision	Kappa
0	Logistic Regression with RFE	0.020815	0.959311	0.974359	0.194872	0.278973
1	Decision Tree	0.074830	0.906276	0.756410	0.317204	0.415575
2	Random Forest	0.050021	0.946236	0.897436	0.201729	0.284894
3	Adaptive Boosting	0.336667	0.939995	0.974359	0.116386	0.150043
4	Gradient Boosting	0.020000	0.952357	0.876712	0.226148	0.321035
5	Extreme Gradient Boosting(XGBoost)	0.015142	0.950119	0.961538	0.159236	0.222356
6	K-Nearest Neighbors	0.044444	0.936568	0.910256	0.197222	0.279018
7	Gaussian Naive Bayes	0.350000	0.936568	0.910256	0.205202	0.290778

Figure 2.1

Model Selection and Interpretation

After building all the models and their respective thresholds to maximise recall, it was time to decide the best model out of all of them, for our situation. We visualised 3 thresholds for all the models(*Figure 3.1*), recall, precision and kappa, we choose recall because identifying bankrupt firms was our priority, precision because we wanted to minimise loosing good borrowers, and lastly, kappa because we wanted to analyse how better our predictions were compared random classifications.

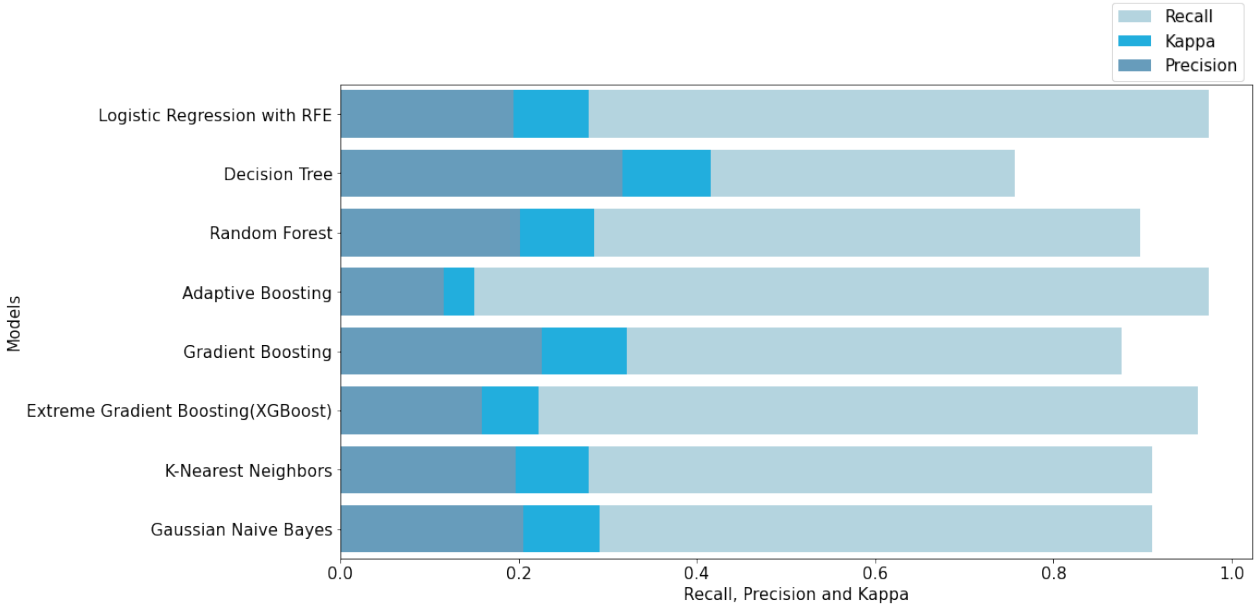


Figure 3.1

As we can see in the figure above us, Logistic Regression with RFE and Adaptive Boosting have the best recall but Logistic Regression beats AdaBoost in metrics like precision and kappa. Another interesting model is XGBoost which is slightly worse than AdaBoost and Logistic Regression in terms of recall(1 extra misclassification of bankrupt companies at healthy) but beats AdaBoost in terms of precision and kappa. Decision Tree classifier has the lowest recall but the highest precision and recall, pointing to the obvious fact that there is a tradeoff between recall and precision. Since recall was our priority, we went ahead with the Logistic regression model and started analysing its coefficients.

Interpretation:

To interpret the coefficients we first visualised them(*Figure 3.2*). The red bars signify coefficients that might've had an impact on the decision making but it statistically significant enough for us to make conclusions.

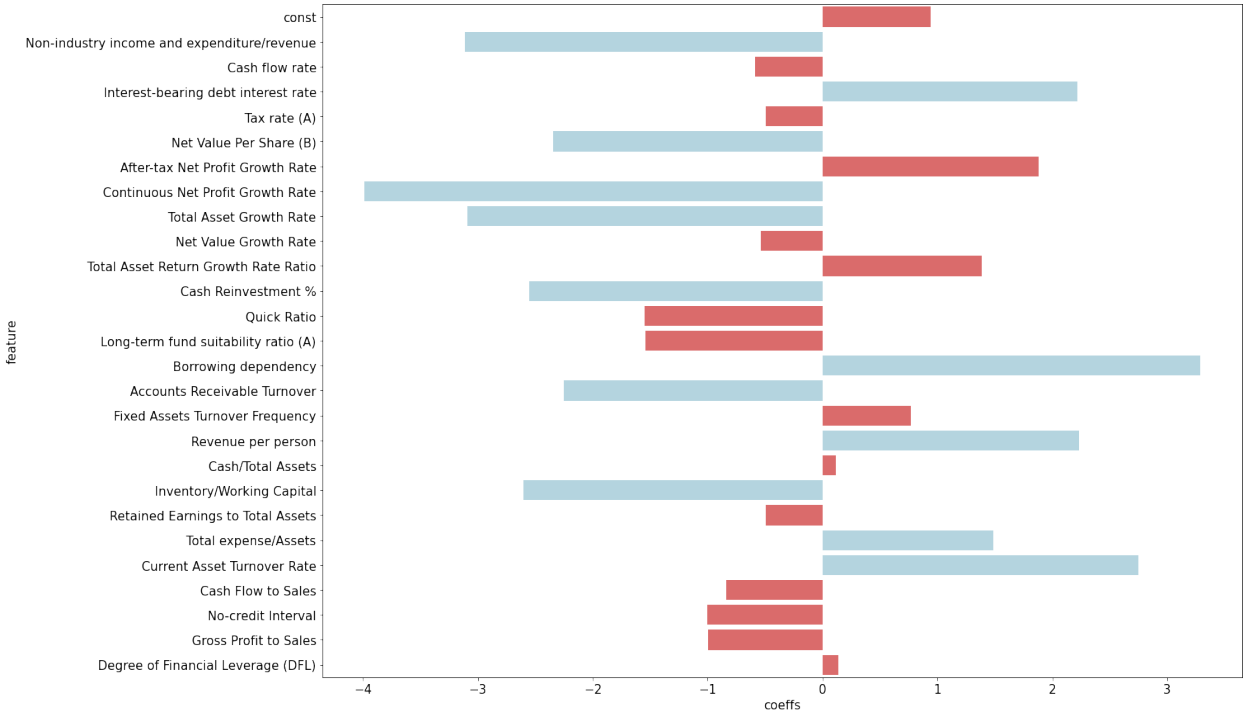


Figure 3.2

The coefficients in the logistic regression model signify the log(odds) of a company going bankrupt. We can say that if the coefficient for a feature is positive, the increase in that feature will lead to increase in probability that a company will go bankrupt and if the coefficient is negative then an increase in the feature will lead to a decrease in the probability that a company will go bankrupt.

The features that have a significant impact on bankruptcy are:-

1. An increase in Non-industry income and expenditure/revenue decreases the probability of a company going bankrupt
2. An increase in continuous net profit growth rate decreases the probability of a company going bankrupt.

3. An increase in Borrowing dependency increases the probability of a company going bankrupt.
4. An increase in Net value per share decreases the probability of a company going bankrupt.
5. As Total expense/Assets increases probability of bankruptcy increases
6. As Current Asset Turnover Rate increases probability of bankruptcy increases
7. As Inventory/Working Capital increases probability of bankruptcy decreases
8. As Revenue per person increases probability of bankruptcy increases
9. As Interest-bearing debt interest rate increases probability of bankruptcy increases

Other Interesting Findings

Other than analysing the coefficients and interpretations from the logistic regression model, we also looked at the other models that we made. We looked at the important features in all the tree based models and we also analysed the common false positives from all the models. Lets first talk about important features from all the tree based models.

Important Features

Feature importance in tree based algorithms is calculated in 2 ways. First is the number of splits a feature is involved in divided by the total splits. Second is the increase in gain due a feature divided by total gain. We found feature importance for all the tree based models(DT,RFC,ADA,GB,XGB), and took the top five out of every model. After that, we visualised the features that show up in the top five of every model most commonly(*Figure 4.1*). We found that Net value per share and borrowing dependency were in the top 5 of 4 out of the 5 algorithms pointing to the fact that they are good deciders of bankruptcy as they are clearly causing a lot of splits in the trees, they were also had a significant impact on classification in the logistic regression model. Another interesting feature was Retained earnings to total assets which showed up in the top 5 of 3

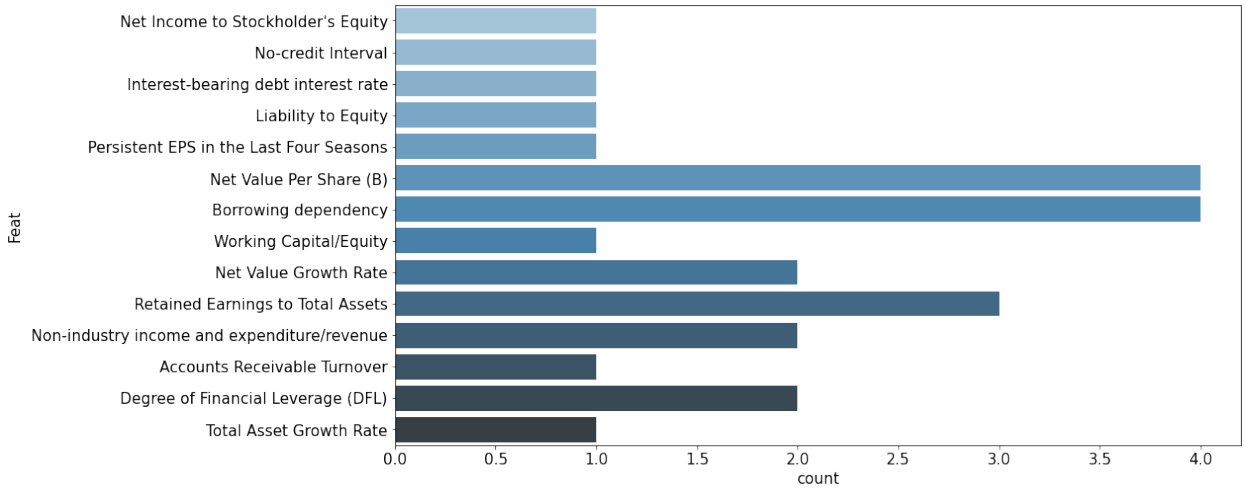


Figure 4.1

out of the 5 algorithms. Net value growth rate, Non-industry income and expenditure and degree of financial all were present in the top 5 of 2 out of the 5 algorithms.

Misclassification Analysis

Our final analysis of the project was looking into false positives of every model and see if there were any patterns on display. We wanted to find a point or range for the features that we could classify as a grey area, as we didn't have enough evidence to classify those companies as healthy. We found out all the false positives and true negatives for all the models, and then we filtered out all the common ones. We analysed the 5 point summary for both the datasets. We first visualised how different the medians of misclassified data were from the correctly classified data(*Figure 4.2*). We found that for a lot of statistically significant and important features like borrowing dependency the medians of misclassified data was really far from their median from correctly classified data. The features for which the medians of misclassified data was really close to correctly classified data are most probably not causing the misclassifications but we still can't rule them out. Those features in conjunction with other features together could be causing those misclassifications.

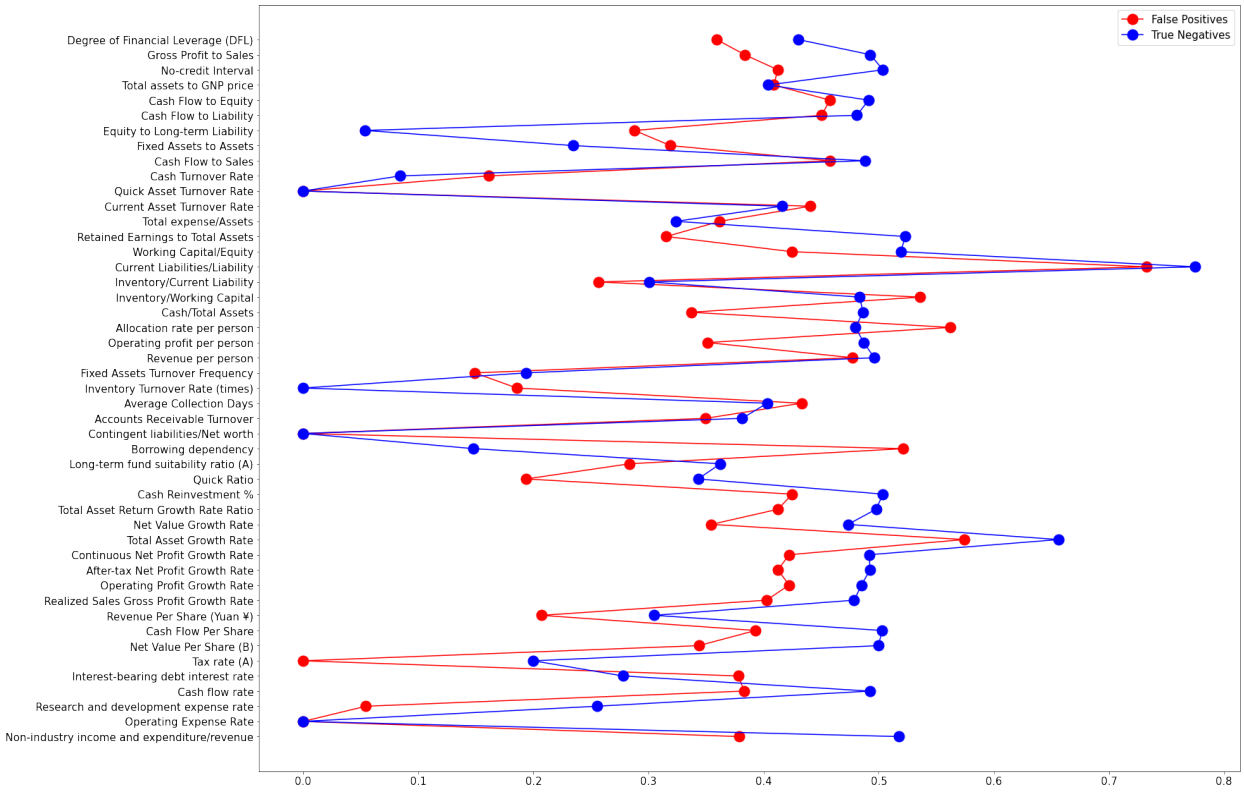


Figure 4.2

To analyse these relationships further we plotted box plots for misclassified data to visualise the IQR for the features(*Figure 4.3*).

We did this to get an idea about the range in which the features in conjunction with other features are causing those misclassifications. Further analysis could conclude that the companies for which these misclassifications are happening because might not be unhealthy now but could be in the future or are heading in the direction of bankruptcy. For now, the only thing we can say is that when certain financial ratios fall in a specific range together, they cause misclassifications.

Model Comparison

Comparing our Logit model to a system in a bank that randomly classifies companies as unhealthy or healthy or does so by looking at just a few things like losses or number of employees. The random or less accurate system would have 3%-50% chance to accurately classify bankrupt companies, our Logistic Regression model classifies

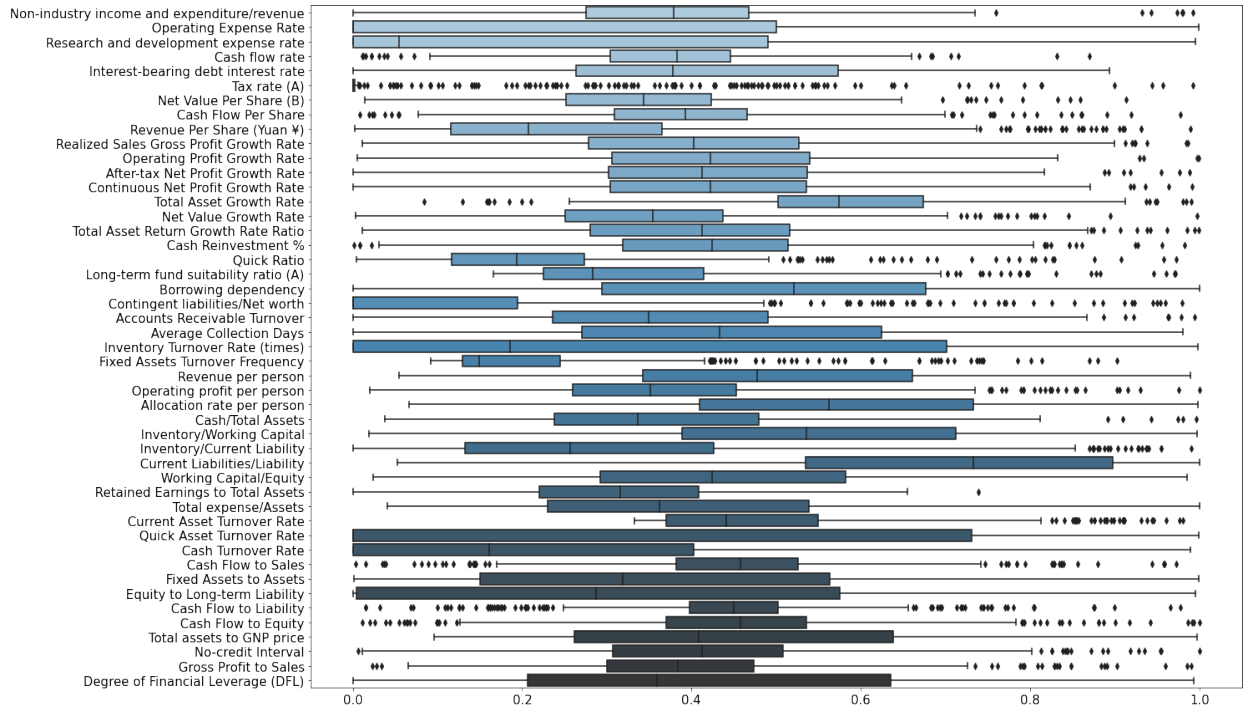


Figure 4.3

bankrupt companies with an accuracy of roughly 97% which is a huge improvement from some less accurate, too simplistic ways of classifications. Even though our precision is on the lower side, banks are usually hurt most by defaulters and identifying them at a high rate reduces their losses significantly.

Closing Reflections

We would have liked to analyse the false positives further as it looked like there was a lot of potential for analysis there but we still came up with quite useful information.

We did try techniques like smote and stacking but the results were not interesting enough to mention in this report.

We would've liked to spend more time looking for ways to increase our precision. Nonetheless, we're certain that the reason precision is so low is because our dataset was highly imbalanced and our false positives are not extremely high compared to thr true negatives.

Closing remarks from some of the group members:

"This project was a great learning(haha) experience for us, we learned how to work as a team and how to compartmentalise different tasks and make the process more efficient."

-Yash Vahi(Team Leader)

"I'm so glad that like minded people came together and started this project. Even though most of being working professionals ,we pushed our limits to met every other day and worked on this project with our hearts and souls. I feel if this project wasn't there, we'd have missed the essence of this course. I personally have evolved my interpersonal skills to new extends. Thanks my teammates and mentor for being cooperative, supportive and so nice to me."

-Vishal Pandey

"Working on this project has been a wonderful experience in my learning path. From data-preprocessing to building different predictive models, it depicts how much I have evolved as a data scientist"

-Apoorva Garg

Notes For Project Team

Original owner of data	Taiwan Economic Journal
Data set information	The data was collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.
Any past relevant articles using the dataset	-
Reference	Kaggle
Link to web page	https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

****