BANKRUPTCY PREDICTION

Group 20

Contents

Business Problems and Motivations

Research Questions

Dataset and Data Visualization + Data Split

Feature Selection

Model selection and Training Steps

Logistics Regression

Decision Tree

SVM

Model Comparison

Conclusions

Limitations and Future Work

The business problem and motivation

The prediction of bankruptcy in companies is a problem that has concerned entrepreneurs, researchers and even governments for years, since detecting early signs that a company is going to enter bankruptcy involuntarily and being able to save it from that process can help reduce the economic losses that bankruptcy entails, both in quantitative and qualitative terms.

The ability to predict corporate bankruptcy would be beneficial for several stakeholders, including: investors, creditors, financial service institutions such as banks, insurance companies, etc, corporation themselves so that they can change their financial structure, employees.

The aim of the study is to narrow down the variables that could potentially predict a company's bankruptcy and to know the value of these variables at which it is likely that the company will bankrupt.

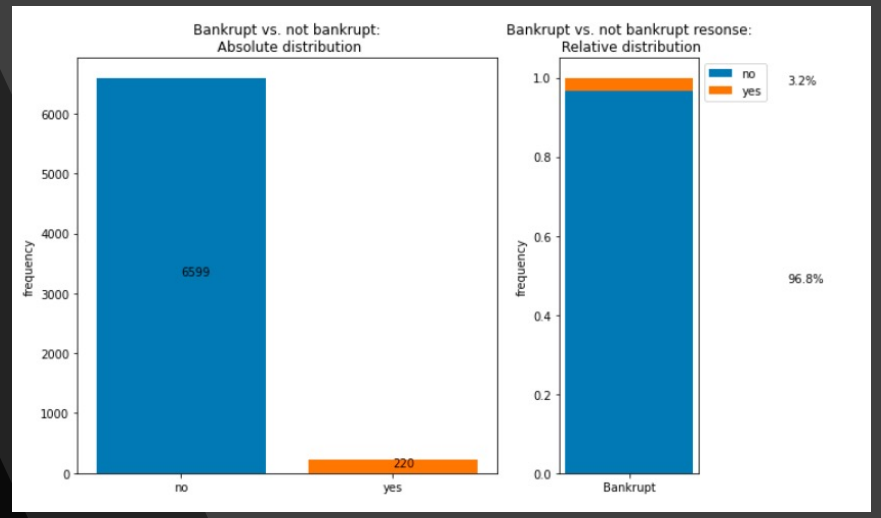
Research questions

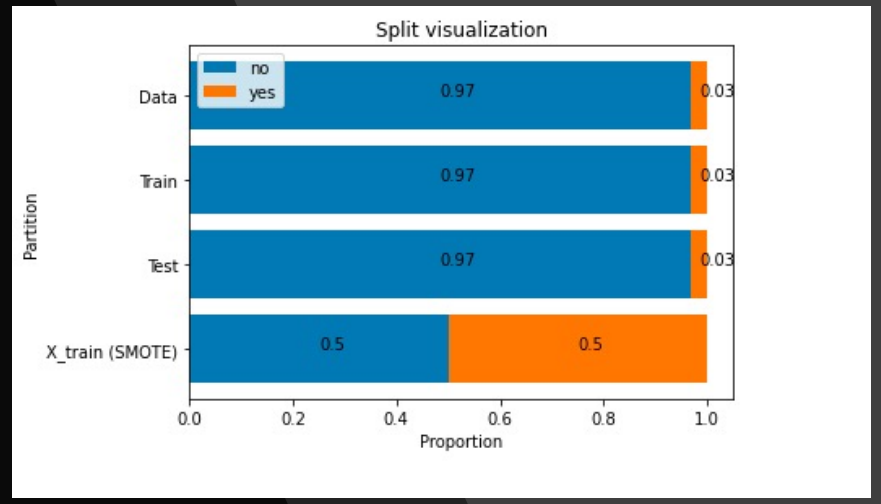
This project is an exploratory study that identifies the relevant indicators for the classification of corporate bankruptcy in Taiwan from 1999 to 2000. The research questions are determined as follows:

- Which variables play the most important roles in company bankruptcy?
- How unbalanced data affect the prediction result?
- Which model is the best for predicting bankruptcy? Why?

Dataset and Data Visualization + Data Split

- The data were 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.
- The original dataset has 6819 rows and 96 columns which make up 1 dependent and 95 exploratory variables. After removing one categorical variable that has no variation in values, there are 95 columns remaining.
- There is no missing value.
- We observe that the non-bankrupt cases account for about 96.8% of all observations. This means that the data is unbalanced and rebalancing should be made to avoid the result completely ignoring the minority class, which is 'Bankruptcy' in this case.
- The rebalancing is done using SMOTE.
- As the distribution of bankrupt to non-bankrupt cases is almost identical to the distribution of the original, whole dataset (though there will always be a slight deviation), we conclude that the split was successful and move on to next step.





Feature selection

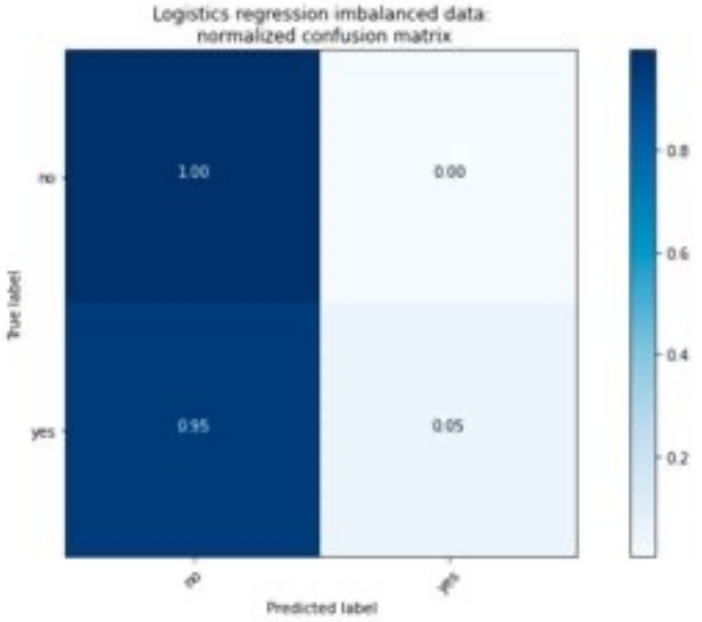
- We conducted feature selection by removing some of the highly correlated independent variables (those that have correlation with each other > 0.7 or < -0.7).
- This allows us to narrow down the features from 94 to 48.
- The goal of us in choosing this method is to minimize the effect of multicollinearity on the results.
- Particularly, the independent variables suffer from severe multicollinearity which could increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. Another consequence is that the coefficient estimates are unstable and difficult to interpret.

Model selections and training steps

- To highlight the positive effect of rebalancing, we first train a model on the original, unbalanced data set.
- Next, we will use **SMOTE** (Synthetic Minority Oversampling Technique) to rebalance the training data.
- Finally, we train the actual model (that we expect to perform better) using the rebalanced data.
- In order to compare models, we apply these steps to train three different models: Logistics Regression, Decision Tree and SVM
- To evaluate models, we will use the following evaluation methods:
  - Confusion matrix
  - AUC (Area Under the Curve) and ROC curve

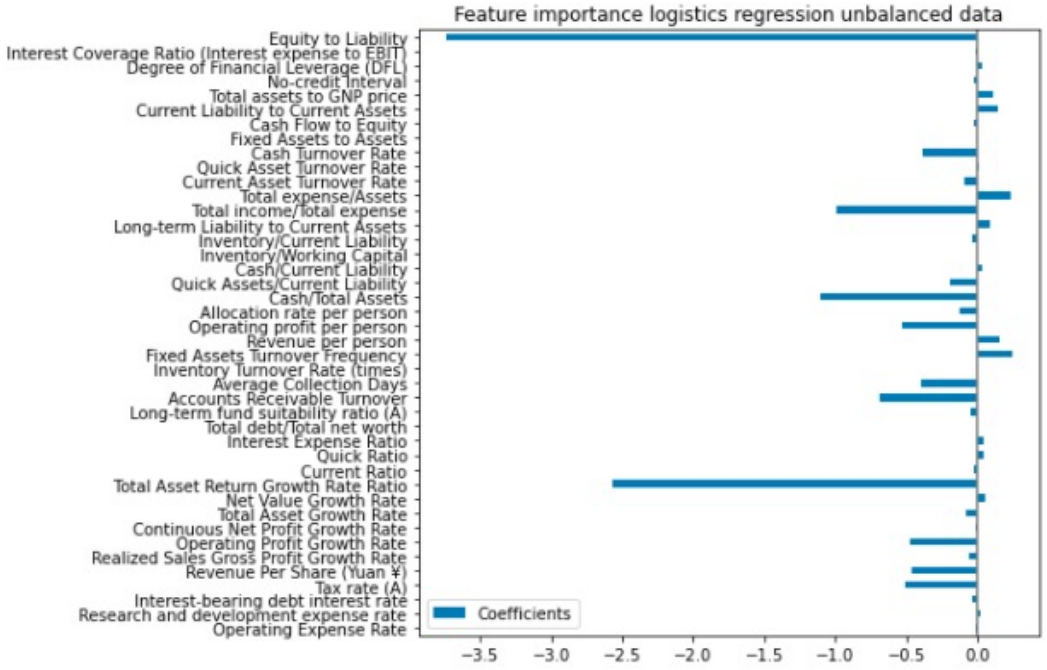
Logistics Regression Model with unbalanced data

- The confusion matrix reveals how the unbalanced Logistic Regression classifier is nearly always predicting the majority class "0, no". Due to the highly imbalanced class distribution, this strategy results in a high testing accuracy of 96.53%. This means if the data has >96.8% "no" observations then always predicting "no" will result in about 96.8% correct "predictions". However, such a classifier has clearly not learned anything and is utterly useless in practice.
- For the unbalanced Logistics Regression model, the strongest features are Equity to Liability and Total Asset Return Growth Rate Ratio



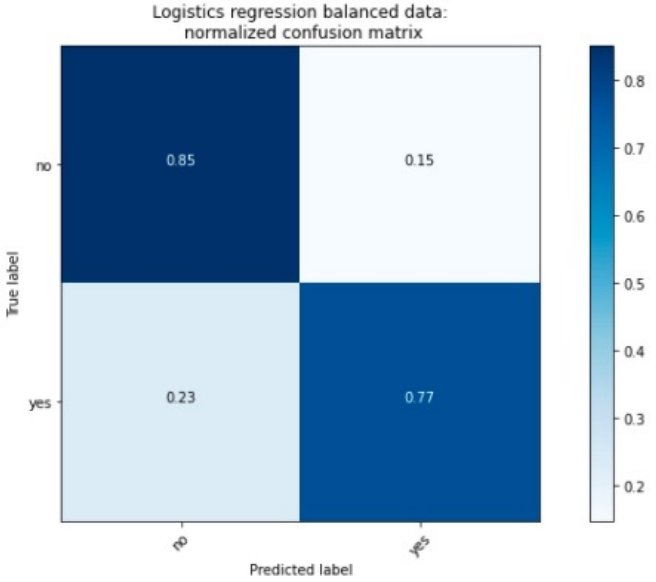
Unbalanced Logistics Regression Model

Accuracy is: 96.53



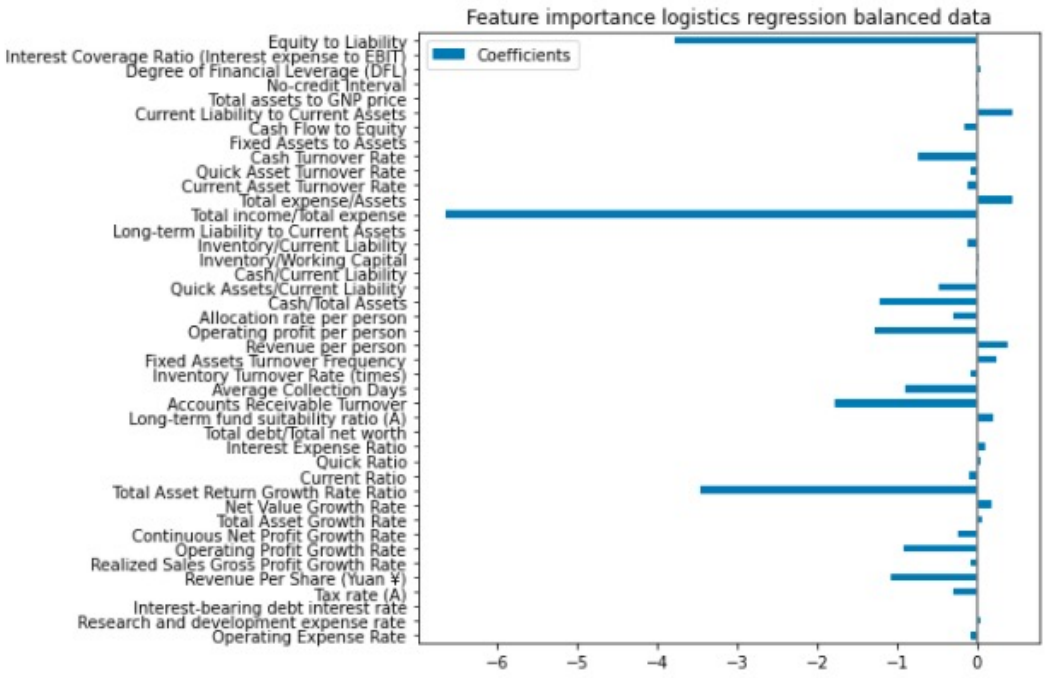
Logistic Regression Model with balanced data

- In contrast, the balanced Logistic Regression classifier performs much better: while it "only" identifies 85% of the "no" cases correctly, it is able to get 77% of the "yes" cases right.
- Similar to the unbalanced classifier, for the balanced logistics regression model, Equity to Liability and Total Asset Return Growth Rate Ratio , and Total income/total expense are considered to be strongest features



Balanced Logistics Regression Model

Accuracy is: 84.85



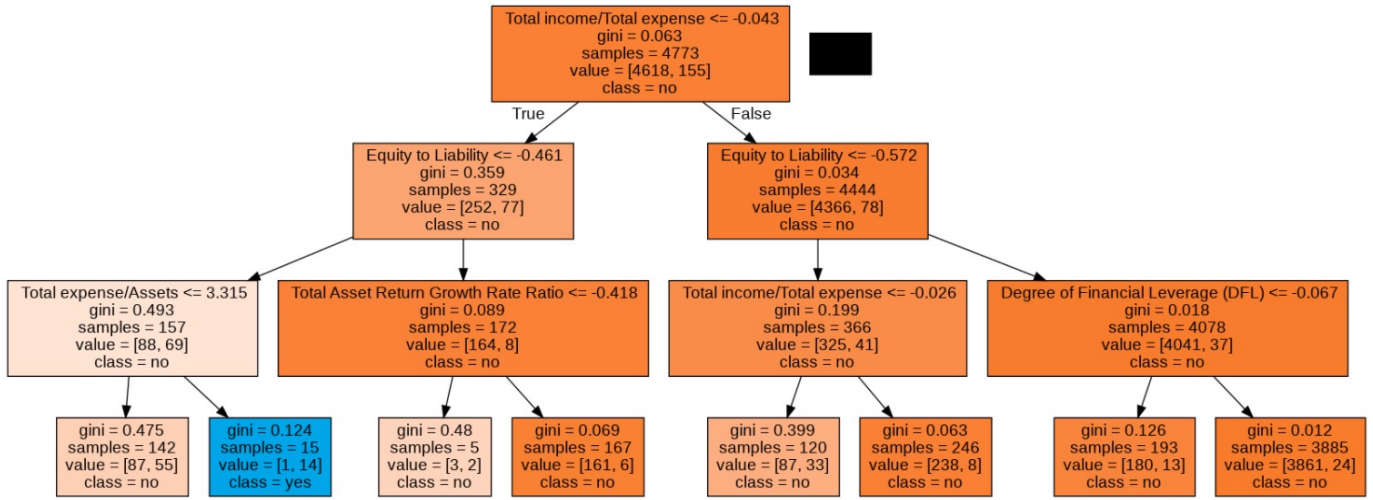
Decision Tree Model with imbalanced data

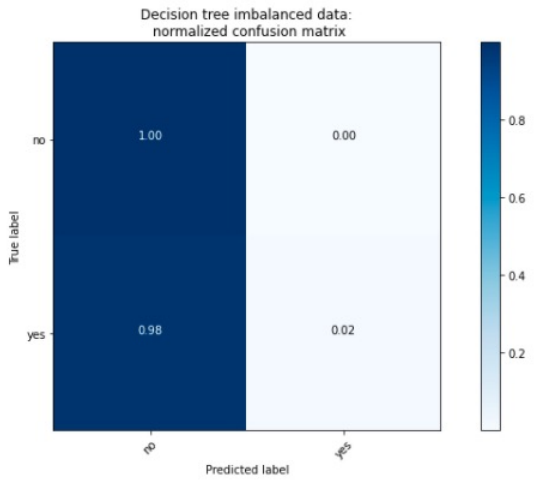
While our classifier labels 100% of cases correctly as non-bankrupt, it only labels 2% correctly as bankrupt. It seems that the classifier predicts "non-bankrupt" most of the time. But since the data is imbalanced (96.8% not-bankrupt vs. 3.2% bankrupt), guessing "non-bankrupt" still means labeling a high proportion of cases correctly.

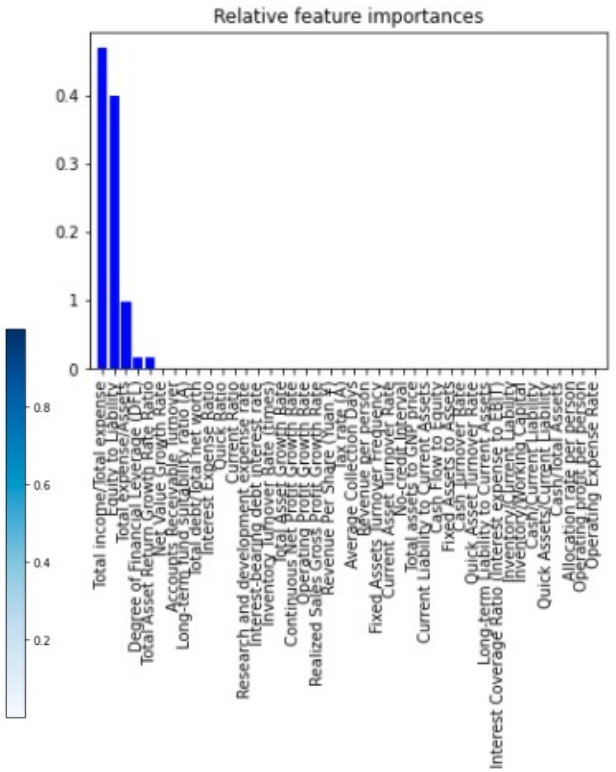
As can be seen below the unbalanced classifier use: Total income/Total expense', ' Equity to Liability' for predictions.

Unbalanced Decision Tree

Accuracy: 96.77







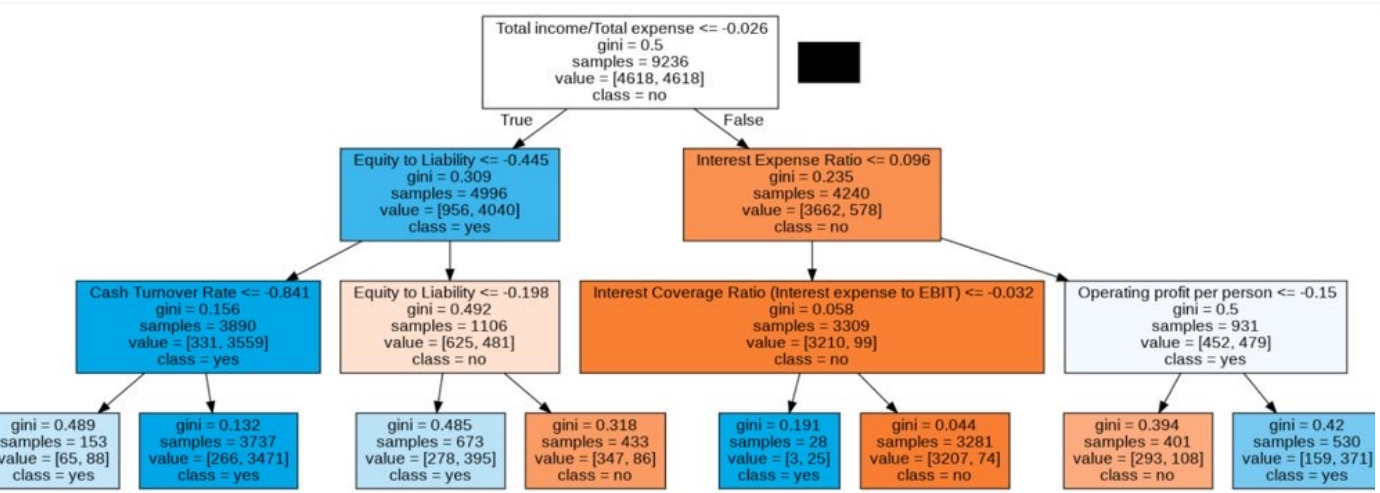
Decision Tree Model with balanced data

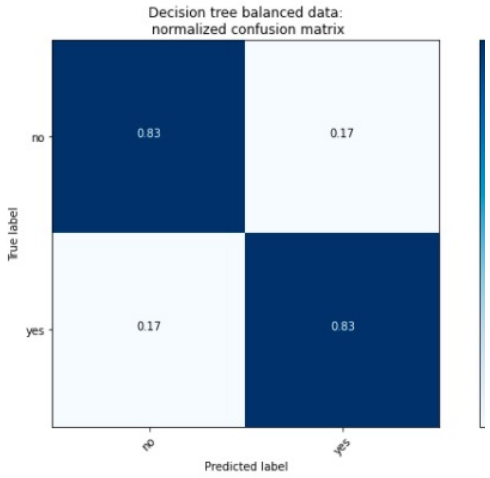
On a balanced data set, this would lead to a lower accuracy score. However, the model seems to perform better since it identifies 83% of the "no" cases correctly while it is able to get 83% of the "yes" cases right.

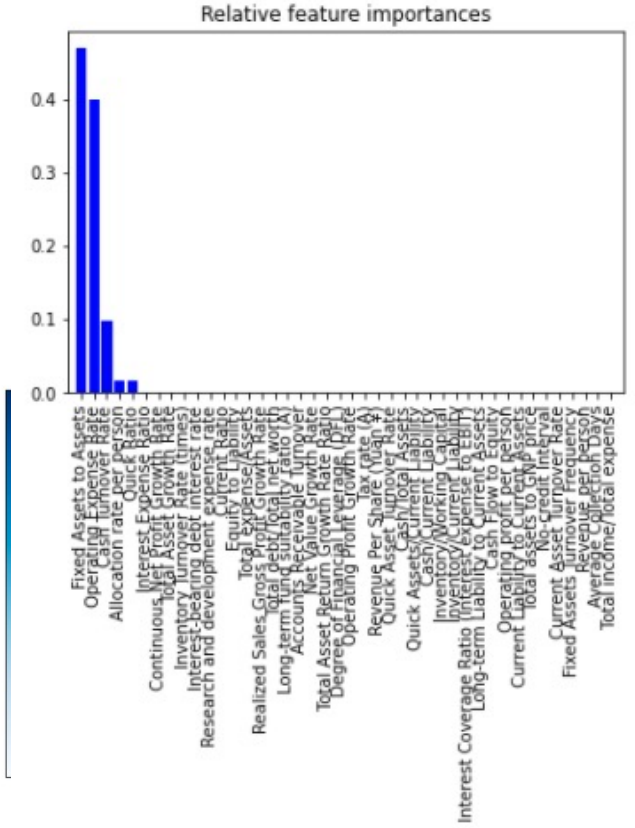
Different form Unbalanced Decision Tree model, balanced classifier use Fixed Assets to Assets', ' Operating Expense Rate' for predictions.

Unbalanced Decision Tree

Accuracy: 83.24







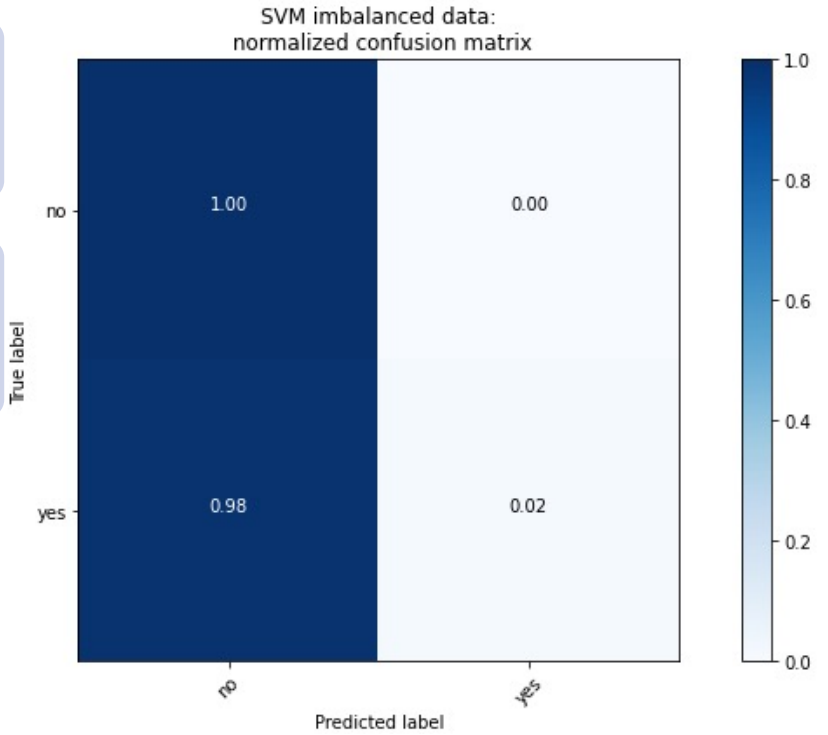
SVM with imbalanced data

SVM imbalanced data

Accuracy: 96.87

With imbalanced data the model predict correctly all non-bankruptcy cases, however almost impossible to predict bankrupt cases (only 2%). This shows the model nearly does not predict anything and just regard almost every case as non-bankrupt.

We cannot find which features are most important in this model as the SVM seems to choose non-linear kernel.



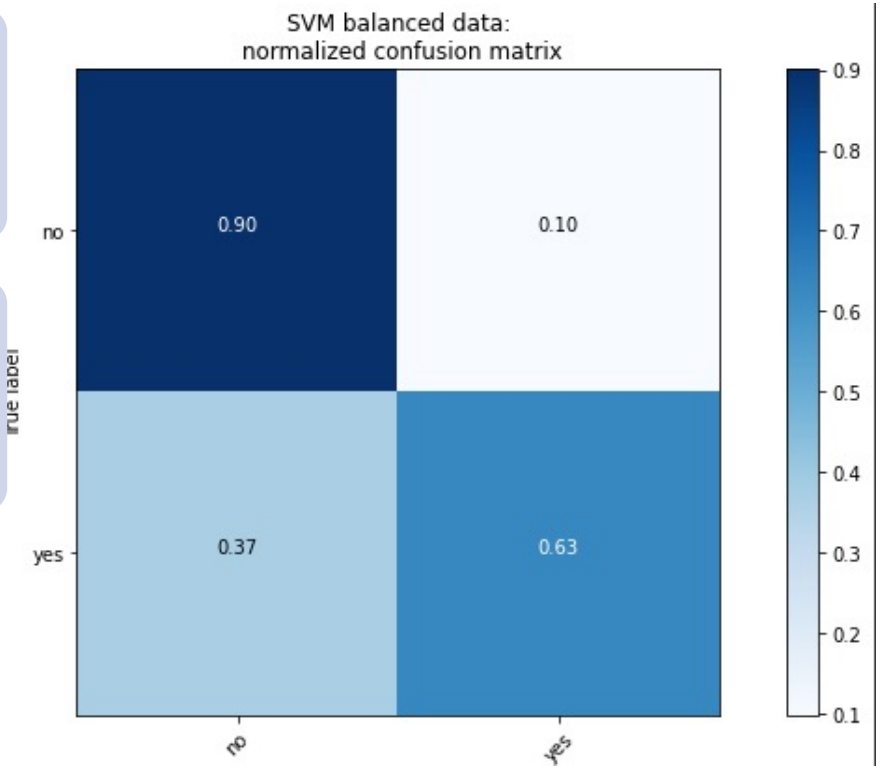
SVM with balanced data

Compared with SVM imbalanced data, SVM with balanced data predict bankrupt cases much better with 623% of bankrupt cases predicted correctly (compared to 2% of SVM imbalanced), while in trade-off there are only 10% of non-bankrupt cases predict incorrectly (compared to 0% of SVM imbalanced data).

We cannot find which features are most important in this model as the SVM seems to choose non-linear kernel.

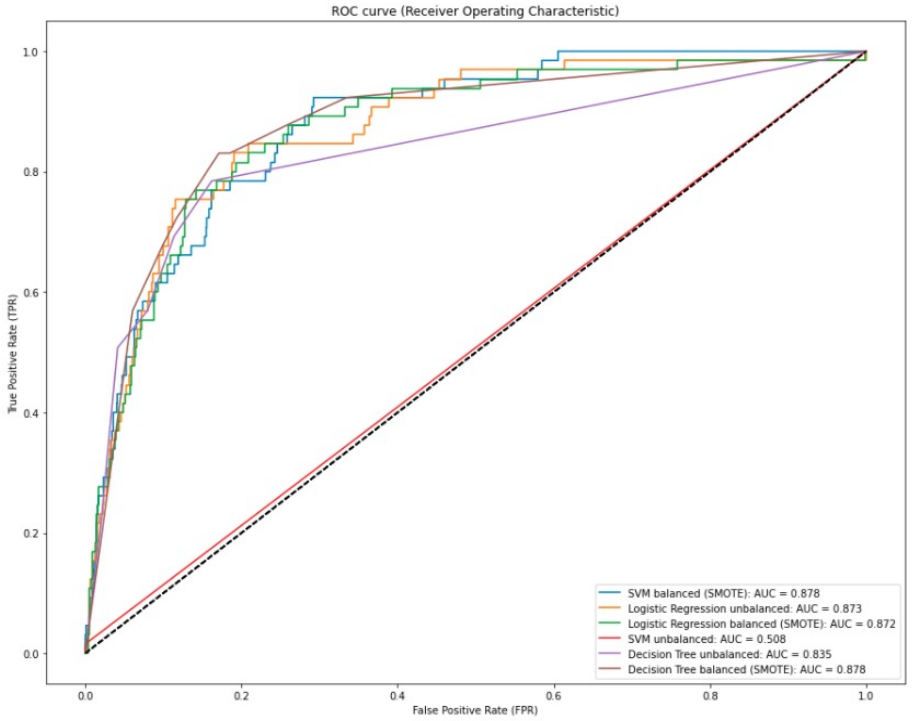
SVM balanced data

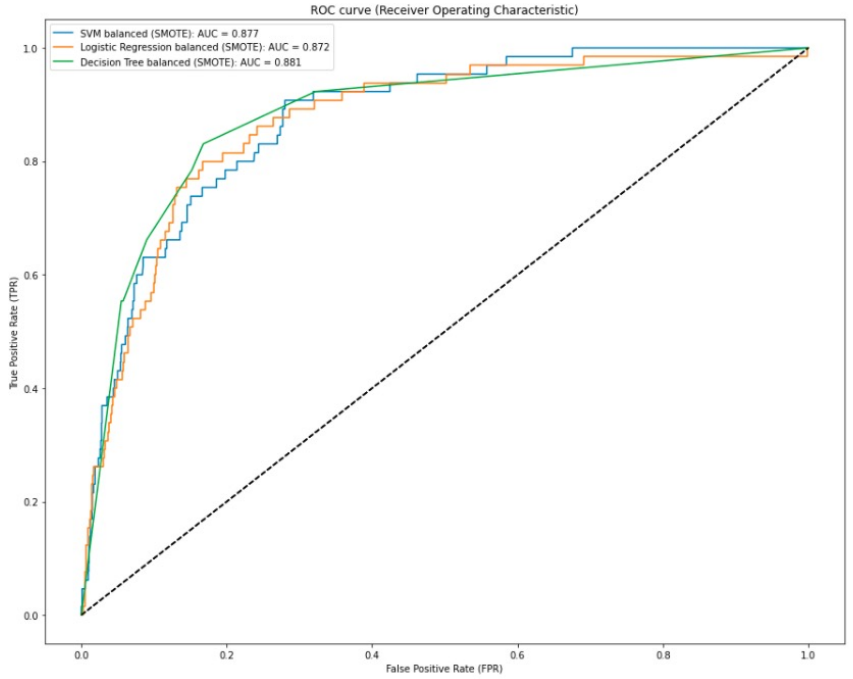
Accuracy: 89.3



Model Comparison

- Rebalancing significantly improve model's prediction capabilities, especially with regard to SVM model. Therefore, we mostly focus on balanced models for prediction.
- The best performing model would be able to predict bankruptcy with minimal amount of false positives and false negatives. While a high rate of false positives would lead to losses on investments, a high rate of false negatives would lead to huge economic losses from that bankruptcy entails, both in quantitative and qualitative terms for many stakeholders
- Examining the ROC curve, in the area of interest for the problem (the left side of the chart - high true positives and low true negative rate), the balanced decision tree model performs the best. Moreover, when examining the confusion matrix with our area of interest, the balanced decision tree model performs the best with the true positive rate of 82%.
- Logistics Regression model and Decision Tree model with balanced data have different important features for prediction while with regard to SVM model, we could not select which features are most important because SVM seems to choose non-linear kernel.





Conclusions

- The project focused on decision trees, SVM, and logistic regression. The full potential of other models for predictions in the data set at hand should be explored in future research endeavors.
- Rebalancing significantly improves the results
- Features like Fixed Assets to Assets, Operating Expense Rate could provide policy implications for all of the involved stakeholders in early bankruptcy detection.
- As for bankruptcy prediction, the balanced decision tree model is able to predict a few bankruptcy, albeit with large amount of false positive.

Limitations and future work

- The current max depth of Decision Tree model is 3, which could make the model underfitting. Therefore, we will try to find the best max depth
- We could try different methods of feature selection in order to deal with multicollinearity such as PCA or partial least squares regression
- We could also explore other models for prediction in the dataset such as Logistics regression model with L1 and L2
- With model evaluation, we can use K-fold and Gain Charts to aid model evaluation process. In addition, we will consider using expected benefit from confusion matrix to further understand the true implications of each model especially regarding the cost/benefit of false positives and false negatives.