Bankruptcy Prediction Using Financial Ratios: A Machine Learning Analysis

Authors:

Anshika Sharma (ID: 305488635)
Alexander Ramos (ID: 605657325)
Cristian Martinez(ID: 205642760)

Abstract

Recently, machine learning based methods have been widely proposed to solve the problem of bankruptcy prediction. This study aims at predicting bankruptcies of companies in Taiwan. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange. The data for the period 1999 to 2009 was used for the study from UCI repository. It is not enough to rely on a single predictive model due to the vast number of factors responsible for bankruptcy and thus the challenge lies in identifying only the key factors that are more responsible. Hence, the Boruta Algorithm and Learning Vector Quantization (LVQ) were used as feature selection methods. Another major hurdle faced was the high-class imbalance within the data which hindered the model’s performance. The various Supervised Machine Learning models used were: Quadratic Discriminant Analysis (QDA), Logistic Regression, K-nearest neighbor classifiers (KNN), Support Vector Machines (SVM). The results indicated that the Quadratic Discriminant Analysis model exhibited higher overall accuracy using the features that LVQ deemed important.

Contents

1. Introduction

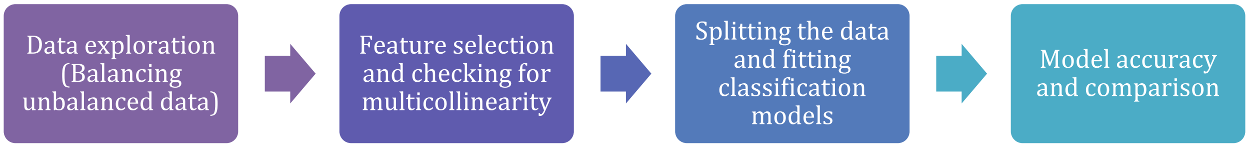
Major economic crises in the past have repeatedly initiated conversations regarding market sustainability and emphasized on determining the appropriate tools required to predict it. The need for accurate predictive and classification models thus becomes apparent to prevent or prepare for such disasters in the future. Within a larger context, bankruptcy of companies and enterprises directly affects financial markets at multiple fronts, and can have a negative impact on both the enterprise itself and the global economy. Predicting the likelihood that a firm may go bankrupt is critical for financial institutions to make appropriate lending decisions. Hence, the need to better understand what factors contribute to bankruptcy and to what extent, takes on an added significance.

In the prior literature, bankruptcy prediction models have been developed using various statistical and machine learning techniques. It is seen that machine learning techniques tend to outperform statistical techniques. Recent studies focus on the effect of financial ratios (FRs) and corporate governance indicators (CGIs) to predict the likelihood of bankruptcy. Financial ratios are usually classified into seven categories: solvency, profitability, cash flow ratios, capital structure ratios, turnover ratios, growth, and others. As for corporate governance indicators, they are classified into five categories: board structure, ownership structure, cash flow rights, key person retained, and others. One such prominent study was done by Liang et al. (2016) to examine the discriminatory power of CGIs combined with FRs for bankruptcy prediction with an aim to determine the best combination of FRs and CGIs to be used. In the process, various combinations of features from all categories of FRs and CGIs were used to train an optimal model.

The following study is rooted in the work of Liang et al (2016), given that it makes use of Financial Ratios variables from the same database, consisting of 6819 Taiwanese Companies. In the data, company bankruptcy is defined on the basis of the business regulations of the Taiwan Stock Exchange prevailing at the time of data collection. George Box, the renowned statistician once stated, "*All models are wrong, but some are useful*". With that in mind, we undertook the task of building different supervised-machine learning algorithms, along with a comparative analysis of each model, in order to identify those that are better suited for predicting bankruptcy of businesses and enterprises using their financial data. The aim of the study is two-fold:

- To identify key predictors (Financial ratios) which contribute the most to bankruptcy using prominent feature selection methods
- To use these predictors to fit classification models and determine the model with the highest prediction accuracy

2. Methodology


Data exploration (Balancing unbalanced data)
Feature selection and checking for multicollinearity
Splitting the data and fitting classification models
Model accuracy and comparison

The data (sourced from UCI Machine Learning Repository) were collected from the Taiwan Economic Journal for the years 1999 to 2009. It contains 95 features and 6819 observations, where the Bankruptcy variable is a factor variable where 1 implies "Bankrupt" and 0 implies "Not Bankrupt". The workflow diagram above summarizes our analysis process. The same has been described in-depth below.

a. Classification Issue

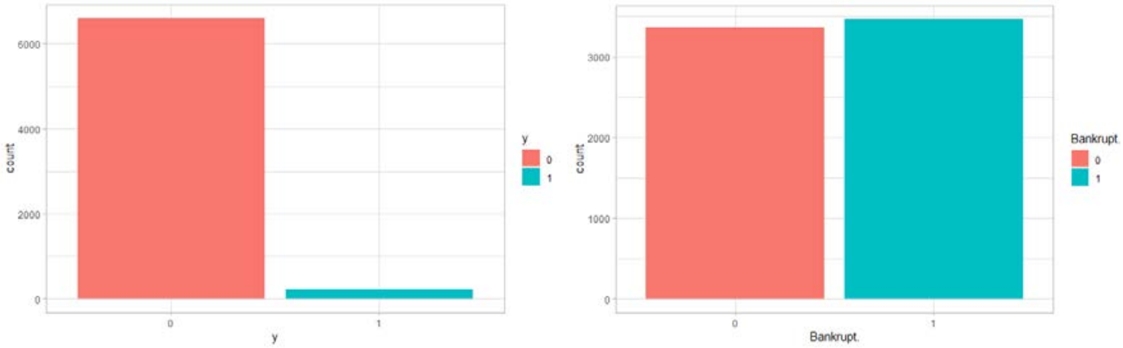


Figure 1.a) Unbalanced dependent variables. b) Balanced dependent variable after using ROSE

After gathering the data, we immediately observed that our Bankruptcy dependent variable's distribution is significantly skewed towards firms that did not declare bankruptcy. To put into perspective, we have 6819 observations and 6599 companies did not declare bankruptcy and 220 did declare bankruptcy. Based on these results, it is evident that we have an imbalanced classification problem and we determined that a correction must be made because we are concerned that our model's accuracy metrics would be deemed unreliable and would generate poor predictive performance. To correct for the imbalance classification problem, we generated synthetic data by enlarging the features space of minority and majority by drawing from a conditional kernel density of the two classes to increase the observation account for the minority class. This process is called ROSE, which is short for Randomly Over Sampling Examples and after the implementation was completed, we derived 3359 observations for firms that did not declare bankruptcy and 3460 observations that did declare bankruptcy. Liang, Lu, Tsai, and Shih corrected the imbalance dataset by performing stratified sampling and reduced their dataset from

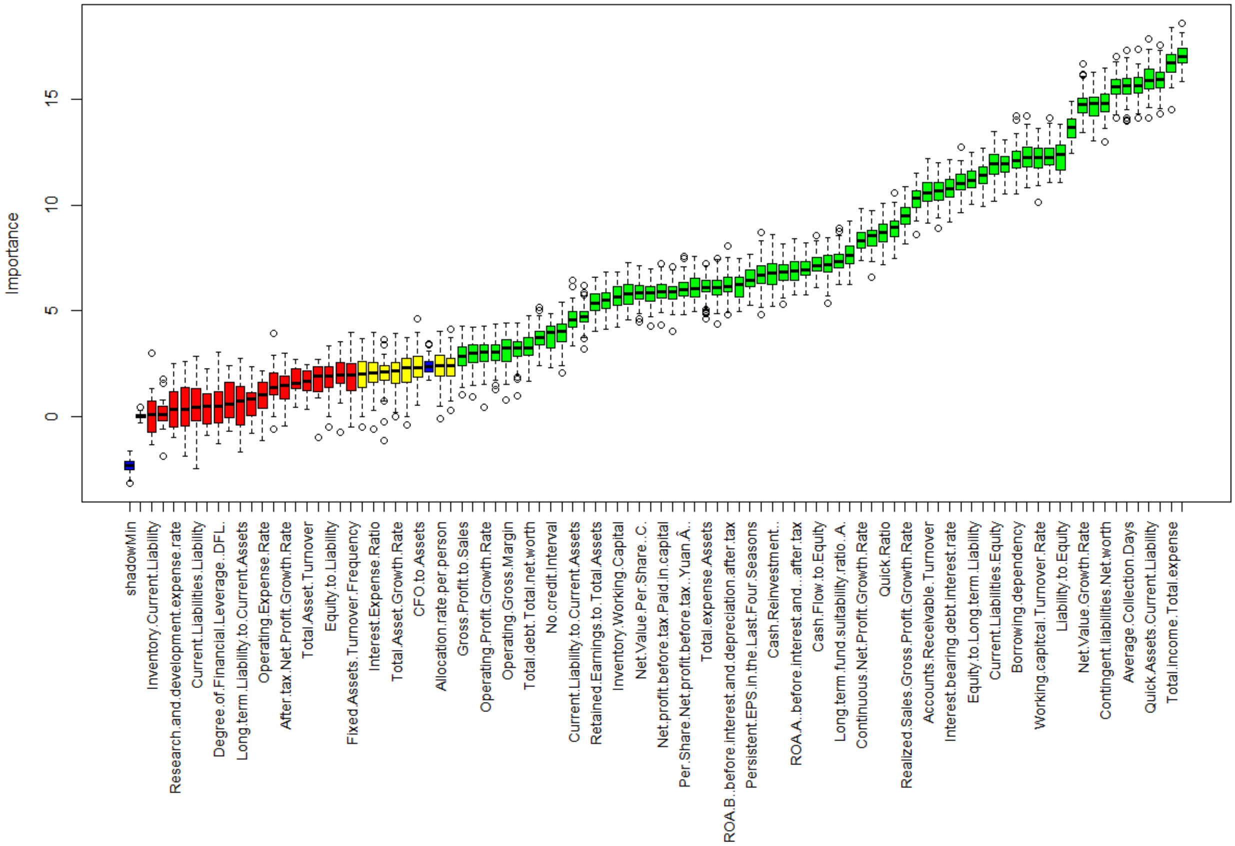
6819 to 468 observations. For purposes of our analysis, we will be evaluating our models using the entire dataset.

2.2 Feature Selection

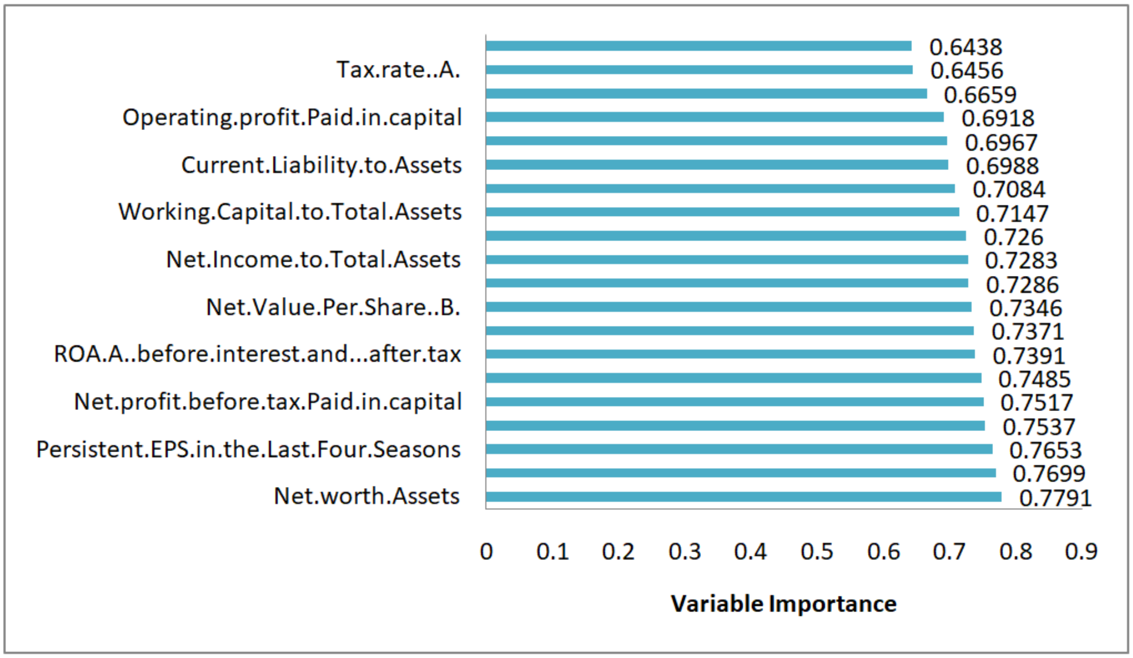
After correcting the imbalance dataset, we now focused on determining which features to use in our models. Since our data has 95 features, we were concerned with overfitting and also including features that did not provide any relevant explanatory power to our models. To assist in our determination of what features to include in our models, we performed two feature selection algorithms: Boruta and Learning Vector Quantization (LVQ). We opted to use these two feature selection models because the work performed by Liang, Lu, Tsai, and Shih (2016) did not use these two feature selection algorithms, thus, we wanted to observe if we would derive different accuracy metrics from our models. The feature selection methods that they used were Stepwise Discriminant Analysis (SDA), Stepwise Logistic Regression (SLR), t-testing, Genetic Algorithm (GA), and Recursive Feature Elimination (RFE).

Before transitioning to speaking about the results derived from the feature selection methods, we wanted to speak about what Boruta and Learning Vector Quantization are doing exactly. The Boruta algorithm is a wrapper-based method that is built around the Random Forest classification algorithm. It attempts to capture all the important features in the dataset with respect to an outcome variable. The algorithm first shuffles the data and creates shadow features and then trains a Random Forest Classifier on the dataset. The algorithm then checks each real feature by the level of importance by a metric like Mean Decrease Impurity or Mean Decrease Accuracy. Thus, the higher the score, the more important the feature is. The algorithm checks for importance by observing the calculated Z-Score and if it is higher than the shadow feature Z-score, then it is deemed important. This process is continued until every feature has been analyzed. Essentially, the algorithm is performing a cross-validation on each variable to observe if each of those features are important and it will report features that are important and features that the algorithm rejected.

The Learning Vector Quantization method is an artificial neural network algorithm that is used for classification that helps reduce the dimensionality of input features. The algorithm separates the features into training sets and observes how the models performed and repeats this process until every feature has been evaluated. The algorithm uses a process called codebook vectors that are a list of numbers that have the same input and output attributes as the training data. The algorithm starts with a pool of random codebook vectors and is evaluated one at a time until it determines which training set of features are the best in terms of classification. Once the analysis is completed, we are able to extract the features that the algorithm believed was the best for classification.



Features selected by Boruta Algortihm



Features selected by Learning Vector Quantization

After these feature selection algorithms were completed, we opted to only use the top 22 features from the Boruta Algorithm[1], after correcting for multicollinearity and further removing variables that would cause problems in our analysis (67 variables were deemed important). The Learning Vector Quantization algorithm selected 20 variables that were deemed important and we also checked for multicollinearity and no features were removed. We collectively agreed to use 22 features from the Boruta Algorithm because we wanted to ensure that our models can generate results that did not suffer from overfitting or convergence issues that would render the models useless in terms of accuracy. The work done by Liang, Lu, Tsai, and Shih (2016) did not specify how many features they used; thus, we cannot compare how our feature selection algorithms greatly differ in their identification approaches compared to the methods the authors used.

2.3 Data Processing

After correcting the imbalance dataset and identifying our features to be included in our models based on the feature selection methods used, we randomly split the data into training set (60%) and testing set (40%). Both the features and the target variable were normalized using the min-max scaler. The train and test features were normalized after splitting them in order to avoid "leaking" any knowledge of testing data while training the model. Feature scaling helps speed up the algorithm's convergence process. This process of pre-processing the data was replicated for both the Boruta dataset and the LVQ dataset.

2.4 Prediction Models

We opted to use 4 separate Machine Learning Models: Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Logistic Regression, and Quadratic Discriminant Analysis (QDA). In contrast, Liang, Lu, Tsai, and Shih (2016) used Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Classification and Regression Tree (CART), and Multilayer Perceptron (MLP). Our predictive models are slightly different and the authors reported average accuracies and it is unclear how many features they used for their analysis, and also used a reduced dataset, thus, we cannot compare how our models performed against the same models they used.

Before we report our results, we collectively agreed to provide theoretical background on exactly what our models are doing fundamentally. We first begin with Support Vector Machines (SVM). SVM models, a supervised learning model, are used to analyze data for classification and regression analysis. For our purposes, once we give a training set, with distinct classifications, the SVM algorithm builds a model that assigns new examples to each category

[1] See table 1 in Appendix for a list of the features selected from Boruta Algorithm and Learning Vector Quantization

and then maps the training samples to points in space with the purpose of maximizing the width of the gap between the two categories. With new observations given, the new observations are also mapped into the same space and are placed or predicted into a side the algorithm believes it belongs to.

The K-Nearest Neighbors (KNN) is a non-parametric method that is used for classification. How the KNN is performed is it uses nearby observations to classify if a new observation would be identified near similar observed observations. This is done by designating one parameter, K, which represents the number of nearby observations or "neighbors" that will be used to classify a new record or in other words observation. This approach of classifying the observation is data-driven and does not require assumptions about the data, which is the non-parametric element in place. Typically, the measured distance that is used when using the KNN is the Euclidean distance, thus, we will normalize our data to ensure that we are in compliance with the model specification requirement. We perform this normalization of the data because we do not want certain predictor variables to dominate the others.

Quadratic Discriminant Analysis (QDA) estimates separate variances/covariances for each class in our dataset. Since QDA allows for different variances among the classes in our dataset, the resulting boundaries become quadratic. In other words, the predictor variables are not assumed to have common variance across each of the k levels in Y QDA works best for datasets that have different variances for each class and it is assumed we have enough observations to accurately estimate the results. QDA is particularly useful if there is prior knowledge that individual classes exhibit distinct covariances. We opted to use this model compared to the Linear Discriminant Analysis (LDA) model because of the relaxed assumption of having different variances compared to the LDA that assumes that the covariance of every class is the same.

The Logistic Regression model is a statistical model that uses a logistic function to model a binary dependent variable that contains the values of “1” or “0”. The coefficients are estimated using maximum likelihood estimation and careful consideration is focused on this property because of the model’s convergence properties, we did not want to fit an abundant number of features that would render the Logistic Model’s coefficients useless and we also had to ensure that the features included did not have evidence of multicollinearity.

2.4 Econometric Approach

Since we are analyzing bankruptcy data, an economic and financial foundation is implied in the analysis. After performing the feature selection methods on our data, we carefully reviewed the variables that were deemed important. We wanted to understand why the variables from the methods were selected. As you can see from the table below, the variables selected have a strong economic and financial explanation since they depict the health of the company and if these metrics are poor, then there is a high probability of default. We simply didn't want to perform the feature selection and fit whichever variables were deemed important into the model and report the accuracy metrics because no economic/financial insight would be derived from our analysis. We wanted our models to have a strong economic and financial foundation so that helpful insight can be derived and shared among academics and practitioners. The models we used are Machine Learning techniques in nature, but the features we used were driven not only from a statistical approach, but also from an economic/financial approach. We also ensured to address any econometric issues like Multicollinearity that we might encounter and how was previously mentioned before, we corrected the data, by removing them entirely.

3. Results

3.1 Support Vector Machines(SVM)

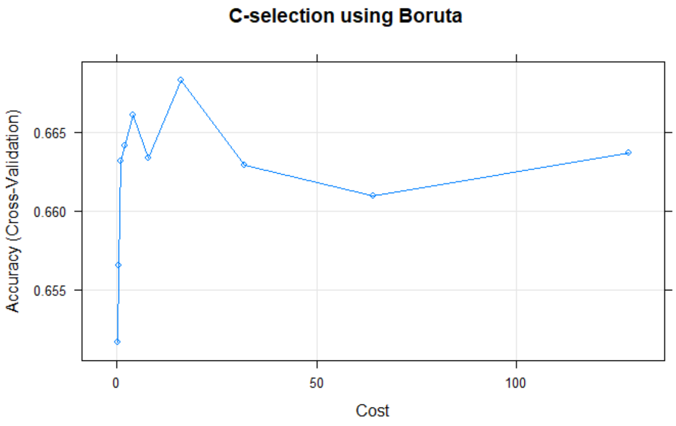
a. Using features from Boruta

For n features,SMV plots a hyperplane in n-dimensional space, in a manner that best differentiates the two classes. For computational purposes, support vector machines use Kernel Functions to systematically find support vector classifiers in higher dimensions. For our data we chose the radial kernel function[2] to train the model, given the non-linearity of the target data. The radial kernel finds support vector classifiers in infinite dimensions, and thus cannot be visualized.

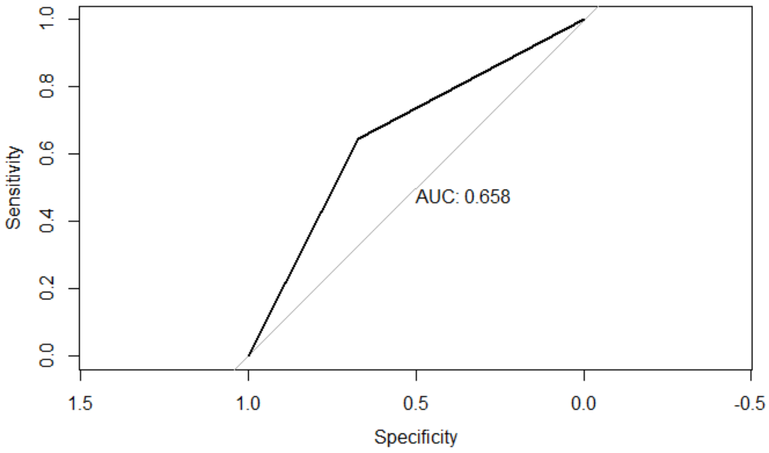
Maximum margin classifiers, like SVM, are very sensitive to outliers. To form a threshold that is not so sensitive to outliers, we can allow for some amount of misclassification using the C-parameter. SVM generally have low bias and high variance. The bias-variance trade-off in SVM can be controlled by changing the C-parameter.

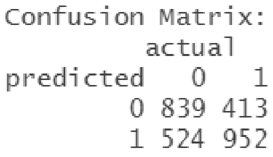
Ten-fold cross validation was used for tuning the hyper-parameter (C) of predictive models. Training data was partitioned into 10 subsets, and cross-validation errors were computed using this split error for SVM classifiers using different values of C. The optimal C value of 16 resulted in the lowest CV error and was used to train an SVM on the training set. When c is large, SVM tries to minimize the number of misclassified examples due to high penalty which results in a decision boundary with a smaller margin. This may lower the bias, but would result in higher variance. For features selected by Boruta, various C values and their accuracies are plotted below:

[2]"svmRadial" kernel from the caret package was used



When the training data for SVM is fitted onto the variables selected by Boruta(after checking for multicollinearity) with C=16 and fit on testing data,an accuracy of 65.65% was obtained along with a misclassification rate of 34.34%. The following confusion matrix and ROC curve were obtained:



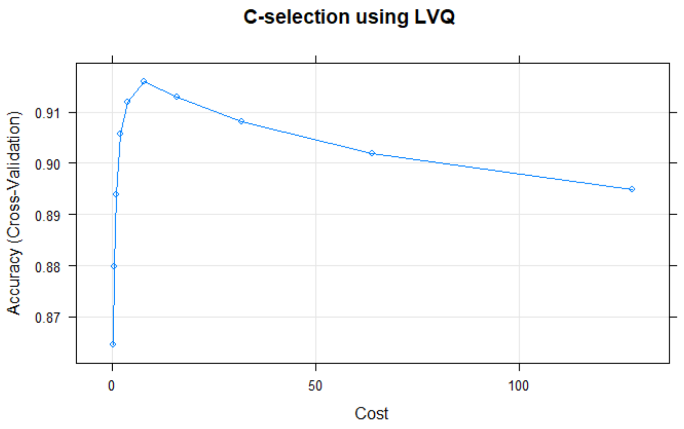


```
Confusion Matrix:
         actual
predicted   0   1
        0 839 413
        1 524 952
```

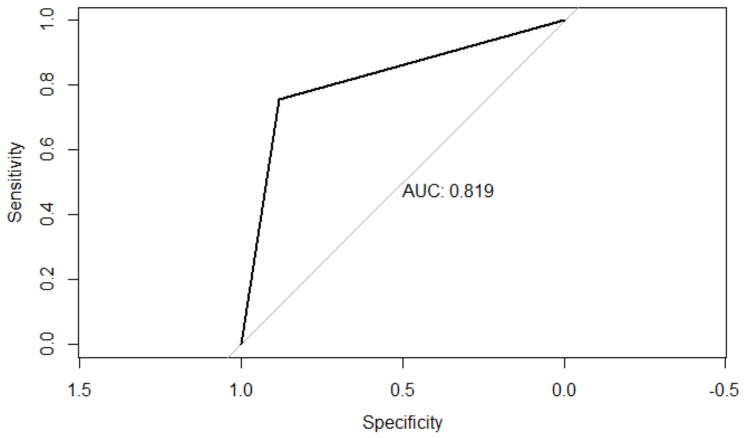
The ROC curve is a trade-off between sensitivity (true positive rate) and specificity (1-false positive rate). The larger the AUC (area-under-curve) is, the higher the accuracy of the test is. For SVM using Boruta features, we get an AUC of 0.658.

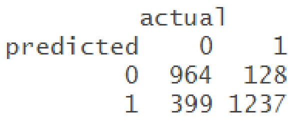
b. Using features from LVQ

A similar process to the one described above was adopted in fitting an SVM classifier to the features selected by LVQ. Using the 10-fold cross-validation, an optimal C-value of 8 was selected to give the highest accuracy while minimizing the cost. Various C-parameter values along with their accuracies have been plotted below. We see that after 8, the accuracy falls as C increases. A smaller value of C (compared to C=16 used for Boruta data) implies a smoother decision boundary. This may increase the bias by reduces the variance of the model.



When the training data for SVM is fitted onto the variables selected by LVQ with C=8 and fit on testing data, an accuracy of 80.68% was obtained along with a misclassification rate of 19.3%. The following confusion matrix and ROC curve were obtained:





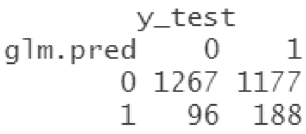
```
          actual
predicted    0    1
        0  964  128
        1  399 1237
```

From the ROC curve plotted above, we get an AUC of 0.819. We see that based on the higher accuracy, higher AUC and lower misclassification rate, the SVM model trained using LVQ features performs better on new (testing) data.

3.2 Logistic Regression

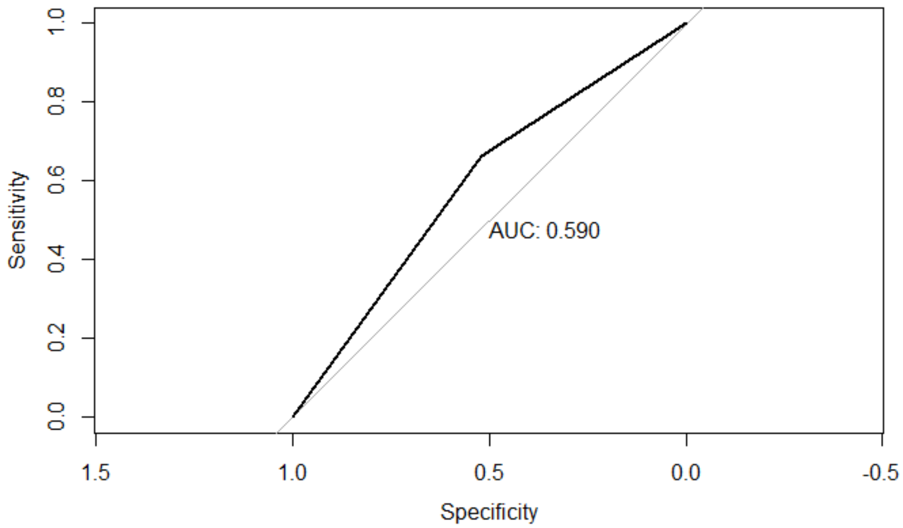
a. *Using features from Boruta*

After fitting the Binomial Logistic model to our training data, we wanted to observe the accuracy of the model. Thus, we used our trained model to predict the bankruptcy of the firms on the testing data. We obtain the following confusion matrix with an accuracy of 53.33%.



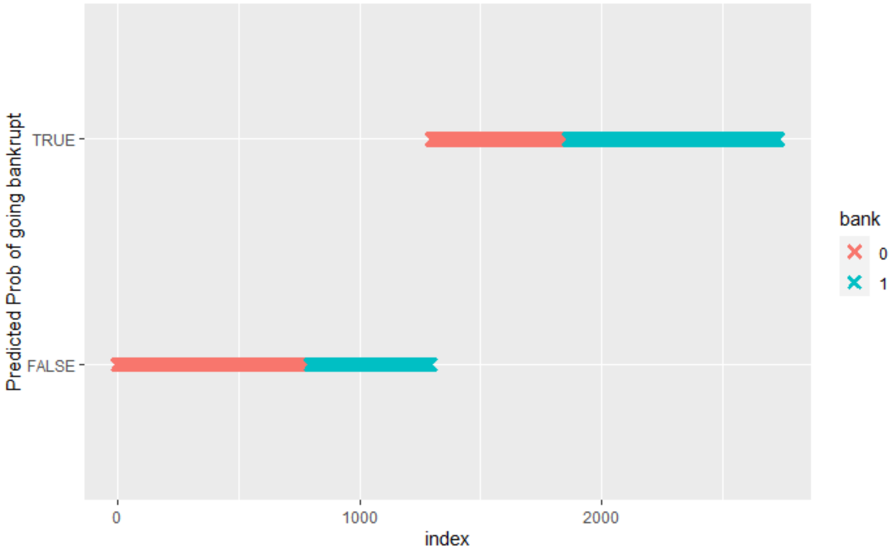
```
          y_test
glm.pred    0    1
       0 1267 1177
       1   96  188
```

The Receiver Operating Characteristics (ROC) curve indicates how well the probabilities from the positive classes are separated from the negative classes. The ROC curve tells us how good the model can distinguish between two classes, in our case declaring bankruptcy or not declaring bankruptcy. Thus, we want the ROC value to be large and to assist us in determining how good the model is, the Area Under the Curve (AUC) is calculated. The larger the value for AUC, the better the model is at classifying. Based on the results derived above, it gives an AUC of 59%.



As can be seen from the graph above, the line at the top is the region for category 1 which implies that the firm went bankrupt and the line at the bottom is the region for category 0 which implies the firm did not go bankrupt. The graph reports the precision rates for class 1 (Bankrupcy) and 0 (No Bankruptcy).

Precision rate for class 1 is number of correct predictions divided by total predictions of class 1 which is roughly 66% (Turquoise portion in the "true" region). Precision rate for class 0 is number of correct predictions divided by total predictions of class 0 which is roughly 51.8% (Red portion in the "false" region).

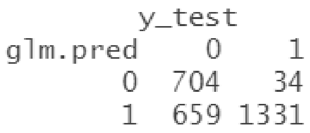


We will now transition to testing our Binomial Logistic model using a 10-Fold Cross Validation approach because we are concerned about introducing bias due to the data partition that was carried out.

After performing the 10-Fold Cross Validation technique on the Binomial Logistic Model, we find that the accuracy derived was 58.95%. This is an increase compared to 53.33% accuracy that was derived from our prior analysis which seems to indicate a decent accuracy rate for our model.

b. Using features from LVQ

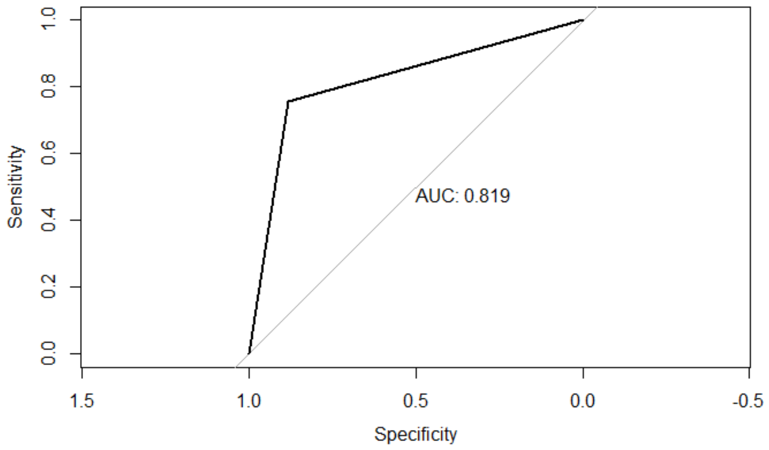
We now carry out logistic regression using the variables that were suggested by the LVQ algorithm. The following is the confusion matrix and accuracy that we obtain:



```
        y_test
glm.pred    0    1
       0  704   34
       1  659 1331
```

There is a tremendous improvement in our accuracy rate from 53.3% when we use Boruta algorithm to select features to 74.5%. This suggests that that the logistic model we derived based on the features we selected from the Learning Vector Quantization algorithm is stronger with regard to accuracy compared to the logistic model using the features Boruta deemed important.

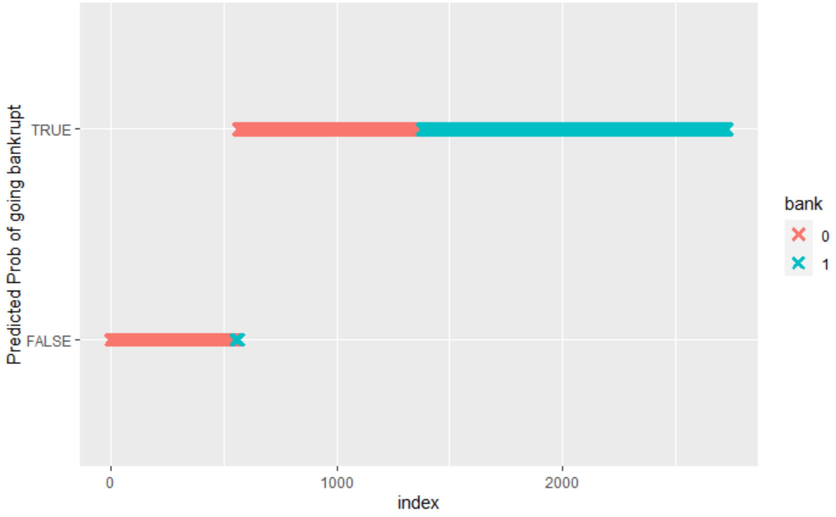
Below is the graph for the ROC curve. Our AUC also improved significantly to 81.1% (as compared to 59 % when we used Boruta algorithm for feature selection) which further supports that the features selected based on the Learning Vector Quantization algorithm have a stronger predictability compared to the features selected from the Boruta Algorithm.



The following is the visual representation of the precision rate logistic regression:

As can be seen from the graph above, the line at the top is the region for category 1 which implies that the firm went bankrupt and the line at the bottom is the region for category 0 which implies the firm did not go bankrupt. The graph reports the precision rates for class 1 (Bankruptcy) and 0 (No Bankruptcy).

Precision rate for class 1 is the number of correct predictions divided by total predictions of class 1 which is roughly 66% (Turquoise portion in the "true" region). Precision rate for class 0 is the number of correct predictions divided by total predictions of class 0 which is roughly 95% (Red portion in the "false" region).

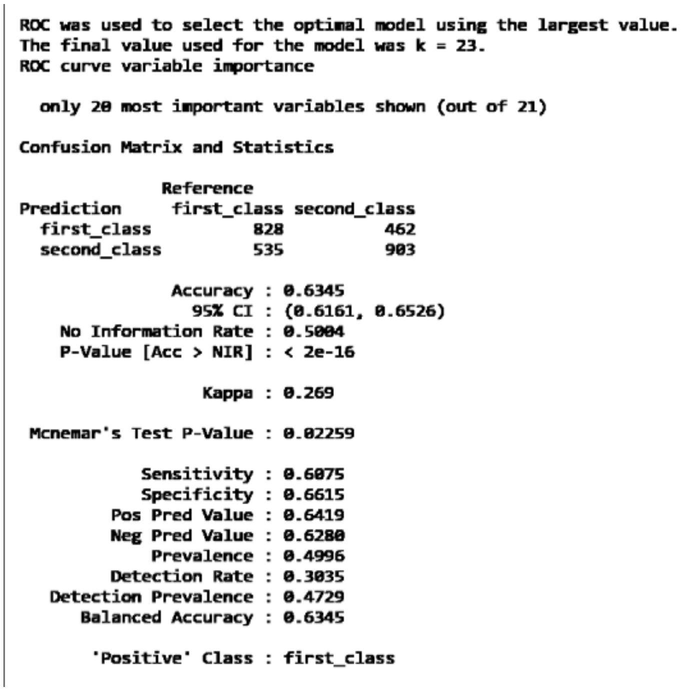


We now transition to carrying out cross validation to our logistic regression. There is a significant jump in the accuracy rate for our cross validation results suggesting that the variables selected by LVQ feature selection method are more robust. It is an increase from 74% accuracy derived earlier to 82.13% suggestive of the robustness in the model.

Based on the analysis performed using the logistic regression model using the features derived from the different feature selection methods, it is evident that the features from the Learning Vector Quantization created a more robust model that provides a stronger degree of accuracy.

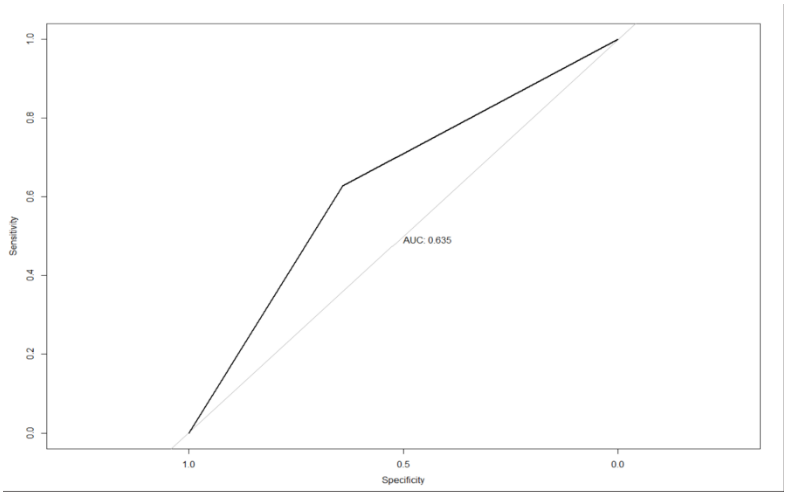
3.3 K-NN

a. Using features from Boruta



```
ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 23.
ROC curve variable importance

  only 20 most important variables shown (out of 21)

Confusion Matrix and Statistics

              Reference
Prediction     first_class second_class
  first_class          828          462
  second_class         535          903

               Accuracy : 0.6345
                 95% CI : (0.6161, 0.6526)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.269

 Mcnemar's Test P-Value : 0.02259

            Sensitivity : 0.6075
            Specificity : 0.6615
         Pos Pred Value : 0.6419
         Neg Pred Value : 0.6280
             Prevalence : 0.4996
         Detection Rate : 0.3035
   Detection Prevalence : 0.4729
      Balanced Accuracy : 0.6345

       'Positive' Class : first_class
```

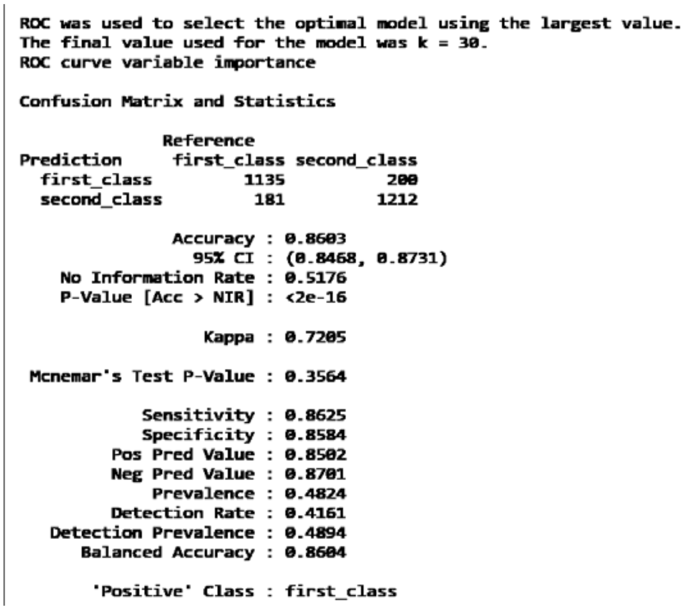
After running KNN algorithm based on ROC measure, we have obtained the optimal K to be 23. For the model with the optimal k, we acquired an accuracy of 63.45%, which is about 13.5% improvement from our model compared the random process to our dataset. Also, it is noteworthy that the sensitivity and specificity are 0.6075 and 0.6615 respectively. It means that the model holds nearly equal accuracy in detecting whether the company would go bankrupt or not. Precisely, for the detection rate of the case where a company would go bankrupt by KNN model is 60.75%, and similarly not going to bankrupt is 66.15%. Overall, we have solid improvement on our prediction accuracy from the KNN model.



The area under the curve, AUC, shows the data separability. It scales 0 to 1, and 1 means that the distribution of the two classes are not utterly overlapping so that the capability of a machine learning model to distinguish the two different classes such as whether a company would go bankrupt or not results in 100% accuracy. With our KNN model to the selected features, we obtained AUC of 0.635. It is larger than the value of 0.5. It means that the information has been gained throughout the training process, and the model is now capable of separating the two different classes. To elaborate the meaning of AUC to be 0.635, we must mention that our AUC value suggests that our KNN can correctly distinguish between the two classes at 63.5% chance.

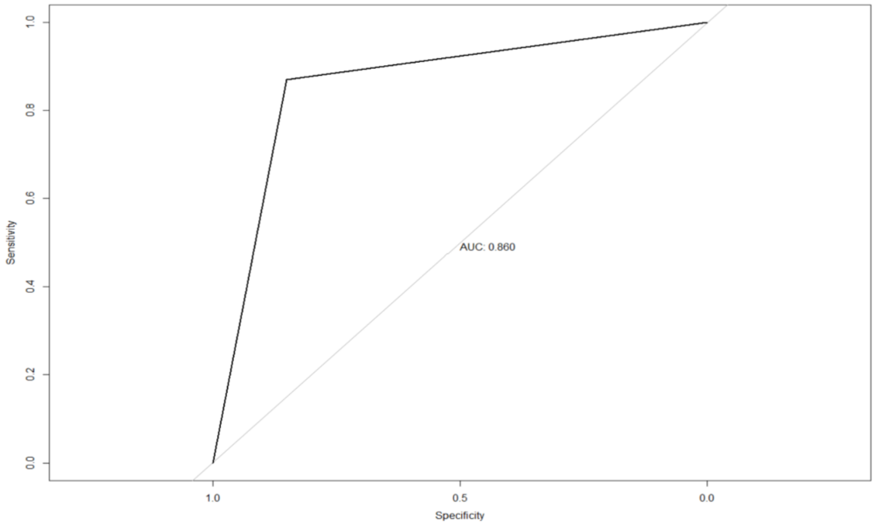
b. *Using features from LVQ*

We will be now conducting KNN with our features selected by LVQ.



```
ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 30.
ROC curve variable importance

Confusion Matrix and Statistics

              Reference
Prediction     first_class second_class
  first_class         1135          200
  second_class         181         1212

               Accuracy : 0.8603
                 95% CI : (0.8468, 0.8731)
    No Information Rate : 0.5176
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7205

 Mcnemar's Test P-Value : 0.3564

            Sensitivity : 0.8625
            Specificity : 0.8584
         Pos Pred Value : 0.8502
         Neg Pred Value : 0.8701
             Prevalence : 0.4824
         Detection Rate : 0.4161
   Detection Prevalence : 0.4894
      Balanced Accuracy : 0.8604

       'Positive' Class : first_class
```

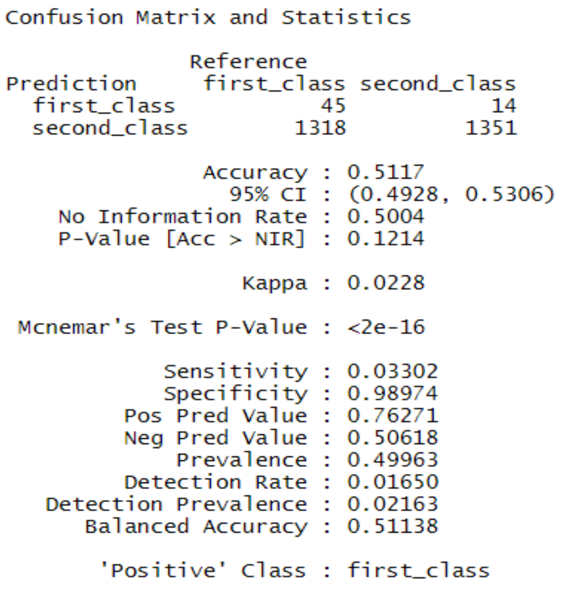
While the KNN with the features selected from Boruta have obtained the optimal K to be 23, the KNN model with the features selected by LVQ returned the optimal K to be 30. For the model with the optimal k = 30, we acquired an accuracy of 86.03%, which is significant and about 36% improvement from our model compared the random process to our original data. Also, it is noteworthy that the sensitivity and specificity are now 0.8625 and 0.8584 respectively. It means that the model holds nearly equal accuracy in detecting whether the company would go bankrupt or not. Precisely, for the detection rate of the case where a company would go bankrupt by KNN model is 86.25%, and similarly not going to bankrupt is 85.84%. Overall, we have solid improvement on our prediction accuracy from the KNN model.



The area under the curve, AUC, shows the data separability. It scales 0 to 1, and 1 means that the distribution of the two classes are not utterly overlapping so that the capability of a machine learning model to distinguish the two different classes such as whether a company would go bankrupt or not results in 100% accuracy. With our KNN model to the selected features, we obtained an AUC of 0.860. It is larger than the value of 0.5. It means that the information has been gained throughout the training process, and the model is now capable of separating the two different classes. To elaborate the meaning of AUC to be 0.860, we must mention that our AUC value suggests that our KNN can correctly distinguish between the two classes at 86.0% chance.

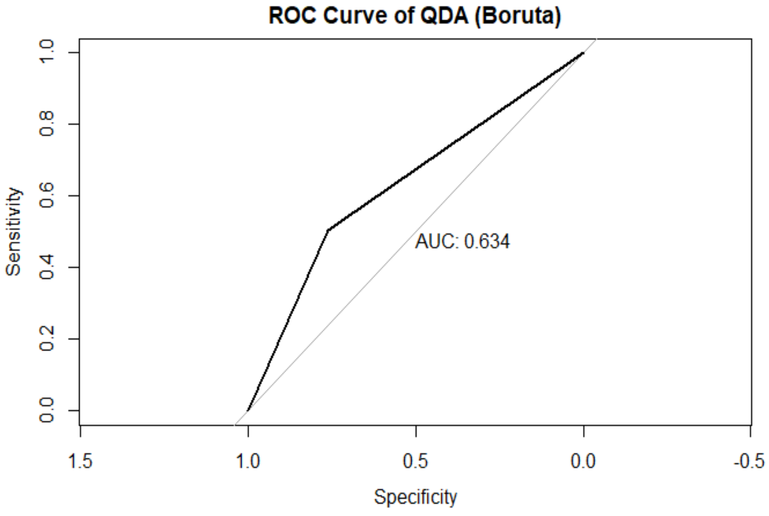
3.4 Quadratic Discriminant Analysis (QDA)

a. Using features from Boruta

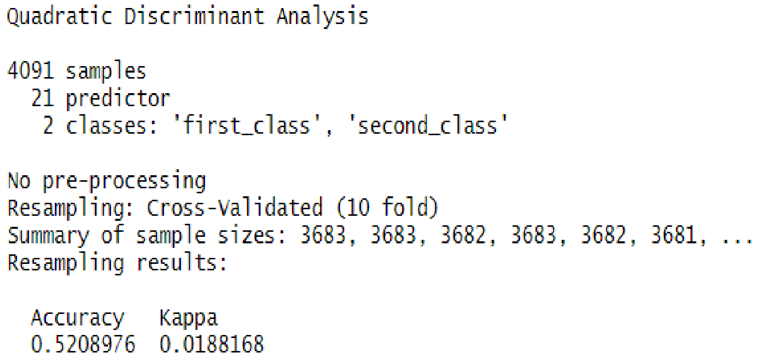


```
Confusion Matrix and Statistics

               Reference
Prediction     first_class second_class
  first_class           45           14
  second_class        1318         1351

               Accuracy : 0.5117
                 95% CI : (0.4928, 0.5306)
    No Information Rate : 0.5004
    P-Value [Acc > NIR] : 0.1214

                  Kappa : 0.0228

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.03302
            Specificity : 0.98974
         Pos Pred Value : 0.76271
         Neg Pred Value : 0.50618
             Prevalence : 0.49963
         Detection Rate : 0.01650
   Detection Prevalence : 0.02163
      Balanced Accuracy : 0.51138

       'Positive' Class : first_class
```

After we fit an QDA model to our training data, we wanted to observe the accuracy that the model would derive based on using the testing dataset that we created earlier in the analysis. We derived an accuracy of 51.17%, which is marginally better than random (Note: “first_class represents “yes” for bankruptcy and “second_class” represents “no” for not declaring bankruptcy). This statement is supported by observing the Kappa statistic of 0.0228. As a reminder, a Kappa statistic that is closer to 1 means there is a perfect agreement in terms of predictability and a Kappa statistic closer to 0 means that the model can predict bankruptcies equivalent to chance. Based on these results, it appears that the QDA model we derived based on the features we selected from the Boruta algorithm is comparable to purely guessing whether a firm given the features we included will declare bankruptcy. Another metric we can check is the AUC. The plot below has the ROC and the calculated AUC.



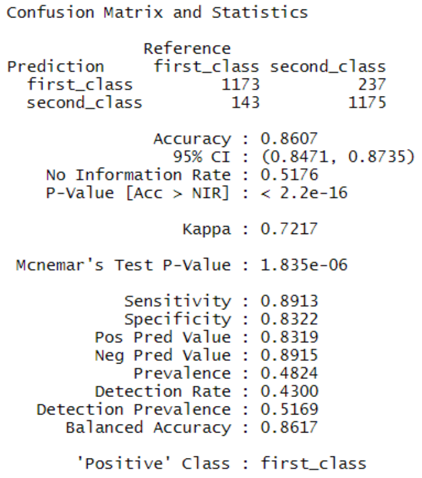
The ROC/AUC indicates how well the probabilities from the positive classes are separated from the negative classes. The ROC short for Receiver Operating Characteristics tells us how good the model can distinguish between two objects, thus, in our case declaring bankruptcy or not declaring bankruptcy. Thus, we want the ROC value to be large and to assist us in determining how good the model is, the AUC (short for Area Under the Curve) is calculated. The larger the value for AUC, the better the model is at classifying. Based on the results derived above, the QDA model derives a value of 0.634, which is better than 50%, which indicates pure randomness. We will now observe how the model does when we perform a 10-Fold Cross Validation.



```
Quadratic Discriminant Analysis

4091 samples
  21 predictor
   2 classes: 'first_class', 'second_class'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3683, 3683, 3682, 3683, 3682, 3681, ...
Resampling results:

  Accuracy   Kappa
  0.5208976  0.0188168
```

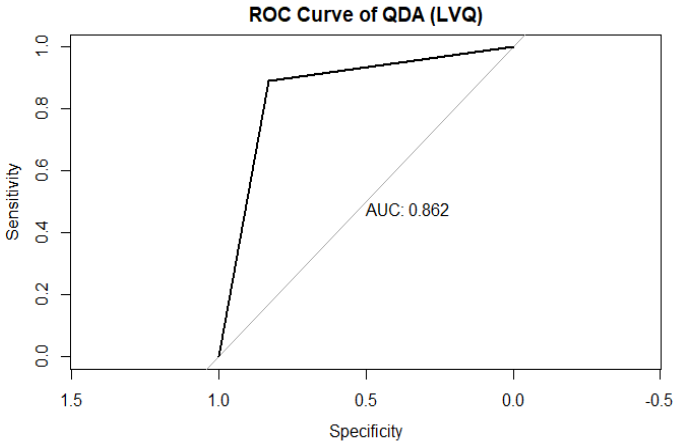
We wanted to perform a 10-Fold Cross Validation because we wanted to address any bias that we introduced when we partitioned the data before we performed our analysis. After performing the 10-Fold Cross Validation, an accuracy of 52.09% was derived, which is marginally better than the accuracy we derived from training the data. Again, the Kappa statistic is essentially 0, which means that the model's predictability is close to being pure randomness.

b. *Using features from LVQ*

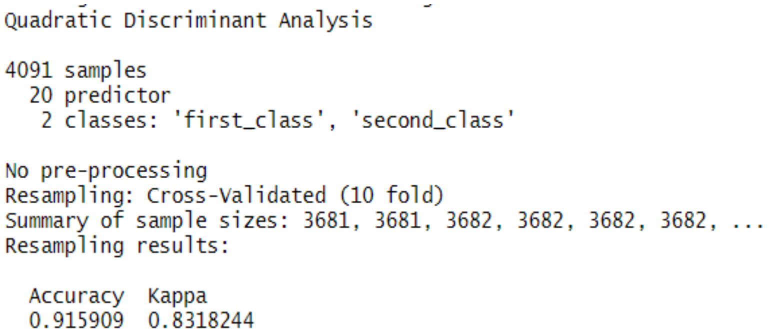


```
Confusion Matrix and Statistics

              Reference
Prediction     first_class second_class
  first_class         1173          237
  second_class         143         1175

               Accuracy : 0.8607
                 95% CI : (0.8471, 0.8735)
    No Information Rate : 0.5176
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7217

 Mcnemar's Test P-Value : 1.835e-06

            Sensitivity : 0.8913
            Specificity : 0.8322
         Pos Pred Value : 0.8319
         Neg Pred Value : 0.8915
             Prevalence : 0.4824
         Detection Rate : 0.4300
   Detection Prevalence : 0.5169
      Balanced Accuracy : 0.8617

       'Positive' Class : first_class
```

After we fit an QDA model to our training data, we wanted to observe the accuracy that the model would derive based on using the testing dataset that we created earlier in the analysis. We derived an accuracy of 86.07%, which is significantly higher than 51.17% derived from using the features selected from the Boruta Algorithm. Looking at the Kappa statistic, this QDA model using the features selected from the Learning Vector Quantization is higher at 0.7217 compared to a statistic of 0.0228 that was derived from the Boruta feature selection. Based on these results, it appears that the QDA model we derived based on the features we selected from the Learning Vector Quantization algorithm is stronger in regards to accuracy compared to the QDA model using the features Boruta deemed important. We will now observe the ROC/AUC plot.



A higher AUC of 0.862 was derived compared to 0.634 from the features selected from the Boruta algorithm, which further supports that the features selected based on the Learning Vector Quantization algorithm have a stronger predictability compared to the features selected from the Boruta Algorithm. As a reminder, we want the AUC value to be higher because it means that the model does a better job of correctly classifying the likelihood of declaring bankruptcy or not. We will now observe how the model does when we perform a 10-Fold Cross Validation



```
Quadratic Discriminant Analysis

4091 samples
  20 predictor
   2 classes: 'first_class', 'second_class'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3681, 3681, 3682, 3682, 3682, 3682, ...
Resampling results:

  Accuracy  Kappa
  0.915909  0.8318244
```

Again, we wanted to perform a 10-Fold Cross Validation because we wanted to address any bias that we introduced when we partitioned the data before we performed our analysis. After performing the 10-Fold Cross Validation, an accuracy of 91.59% was derived which was higher than the training accuracy of 86.07% derived when we trained the data. The Kappa statistic also increased from 0.7217 to 0.8318. In comparison, the cross validated accuracy for this QDA model using the results from the Learning Vector Quantization was higher at 91.59% compared to 52.09% accuracy that the QDA model using the features selected from the Boruta Algorithm derived. The Kappa statistic is significantly larger at 0.8318244 compared to 0.0188168 that the cross validated QDA model using the features derived from the Boruta algorithm.

Based on the analysis performed using the QDA model using the features derived from the different feature selection methods, it is evident that the features from the Learning Vector Quantization created a more robust model that provides a stronger degree of accuracy.

4. Summary

4.1 Actionable Insights and Lessons Learned

After the completion of our analysis, we were able to create classification models that demonstrated strong accuracies in regards to classifying whether a corporation would declare bankruptcy based on the features particularly selected from the Learning Vector Quantization method. We sought to not only implement machine learning techniques into our model creation, but we also followed econometric techniques in the attempt to combine both machine learning and econometrics into a field that has been growing ever since the introduction of using rigorous machine learning techniques to explain economic phenomena. Many have researched and attempted to create models that could determine whether a firm would declare bankruptcy based on financial and corporate governance metrics and collectively we felt that we could further contribute to the research. The Liang, Lu, Tsai, and Shih (2016) paper was the inspiration behind our decision to explore this topic, but we wanted to apply an entirely different approach. We did not use the same feature selection methods as them and instead opted to use both Boruta and

Learning Vector Quantization for our feature selection methods. We addressed the imbalance dependent variable classification problem by using the ROSE approach instead of using stratified sampling that the authors used. More importantly, we opted to use the entire dataset instead of using a portion of the data, which was the approach taken by the authors.

Our analysis produced strong results, but careful consideration needed to be used when selecting the features that would be included in our model because of concerns of overfitting and econometric issues like multicollinearity. With the vast number of features available to us, we knew we had to perform a feature selection method to help narrow down which features would create a model with high accuracies and also remove variables that were considered to be unimportant. The challenge was that Boruta deemed 64 out of 95 features as important, thus, we went through the process of further eliminating features because of econometric issues like multicollinearity. After this task was completed, we were left with 22 features from Boruta. Unfortunately, the models we used to fit the features created models that were only slightly better than random, meaning they had accuracies slightly above 50%, except for the Support Vector Machine model that reported an accuracy of 65.65%. We then focused on the results from the Learning Vector Quantization method and stronger accuracies were derived based on the features selected. We also noted that Boruta selected diverse features across the seven financial ratio categories described in the introduction, but most LVQ features fell under the Profitability and Solvency categories. Given the better performance of LVQ features, it may also be concluded that profitability and solvency ratios are better predictors of bankruptcy than the other categories.

Collectively, we learned that testing different feature selection methods and also incorporating economic theory into our analysis was the best approach. Once the feature selection methods were completed, we wanted to see which variables the algorithms deemed important and critically think on why those features would be important. We learned that certain models are better constructed for managing large datasets and input of features and these models have the capacity to extend further with the addition of more features and observations. The ability to read, understand, and implement an entirely different approach based on a credible academic paper peaked our intellectual curiosity and allowed us to not only learn, but perhaps explore other possibilities that the authors hadn’t done since the academic paper was written in 2016.

One possiblelimitation of our analysis could stem from the fact that the period of analysis (1999-2009) included a business cycle downturn (the 2007-08 recession) which might have skewed the results. We understand that bankruptcy prediction can be simplistic to think about in regards to what causes them, but predicting when they are going to base on a specific feature or features as leading indicator or indicators can provide valuable insight. We aimed to not only further understand the study of bankruptcy predictions, but perhaps contribute to the study itself. Perhaps our models of features selected can serve as a model that banks and other lending institutions can use to anticipate in advance when their borrower is leaning towards bankruptcy.

By having the ability to anticipate bankruptcies in advance, economic downturns could be mitigated or lessened in magnitude by lending institutions involving themselves in corporations before further problems arise.

4.2 Future work and Potential Additional Applications

Given that we used the Financial Ratios data from Liang, our analysis is in some ways rooted in their work. As we reflect on the completion of our analysis, there are a few alternatives that could be used that could derive vastly different results and can serve as future work. In the future, we could consider using both Financial Ratios and Corporate Governance Indicators. We could replicate the analysis process implemented in this study to the entire dataset and compare the results with that of Liang et al (2016).

One entirely different approach that can be taken is to evaluate the models strictly using financial metrics features and also models that use corporate governance features. Another approach is perhaps focusing on financial rations within specific categories like solvency, profitability, turnover, and capital structure ratios. By doing this, the models could derive stronger accuracy metrics and could be implemented in a real-life application.

Lastly, we could use a different definition of bankruptcy, as defined by another country because other countries may not necessarily have a standard definition of what exactly is a distressed company (For an example, China). If the definition of distressed companies is not clear, then it is may be difficult to assess the performance of prediction models.

Regardless of the approaches presented for future work, we do know that the analysis of bankruptcy is not finished yet, but in regards to our work, we are optimistic that perhaps we marginally contributed to this much needed study of bankruptcy prediction.

Appendix

Table 1: Variables selected from the Boruta feature selection process

Boruta Variables	Financial Ratio Category
Total Income Total Expense	Profitability
Quick Assets to Current Liability	Solvency
Revenue per Share Yuan	Others
Average Collection Days	Turnover Ratios
Current Ratio	Solvency
Operating Profit Rate	Profitability
Contingent Liabilities Net Worth	Solvency
Net Value Growth Rate	Growth
Fixed Assets to Assets	Capital Structure Ratios
Net Income to Stockholders Equity	Profitability
Total Asset Return Growth Rate Ratio	Growth
Working Capital Turnover Rate	Turnover
After tax net Interest Rate	Others
Equity to Long term Liability	Solvency
Interest bearing debt interest rate	Solvency
Revenue per person	Others
Accounts Receivable Turnover	Turnover
Realized Sales Gross Profit Growth Rate	Growth
Inventory and accounts receivable Net value	Turnover
Quick Ratio	Solvency
Net Income to Total Assets	Profitability
Continuous Net Profit Growth Rate	Growth

Table 2: Variables selected from the LVQ feature selection process

LVQ Variables	Financial Ratio category
Net worth Assets	Capital Structure Ratios
Debt ratio	Solvency
Persistent EPS in the Last Four Seasons	Profitability
ROA.C. before interest and depreciation before interest	Profitability
Net profit before tax Paid in capital	Profitability
Per Share Net profit before tax Yuan	Profitability
ROA.A before interest and after tax	Profitability
ROA.B before interest and depreciation after tax	Profitability
Net Value Per Share B	Others
Net Value Per Share A	Others
Net Income to Total Assets	Profitability
Net Value Per Share C	Others
Working Capital to Total Assets	Solvency
Retained Earnings to Total Assets	Profitability

Current Liability to Assets	Solvency
Operating Profit Per Share Yuan	Others
Operating profit Paid in capital	Solvency
Current Liability to Current Assets	Solvency

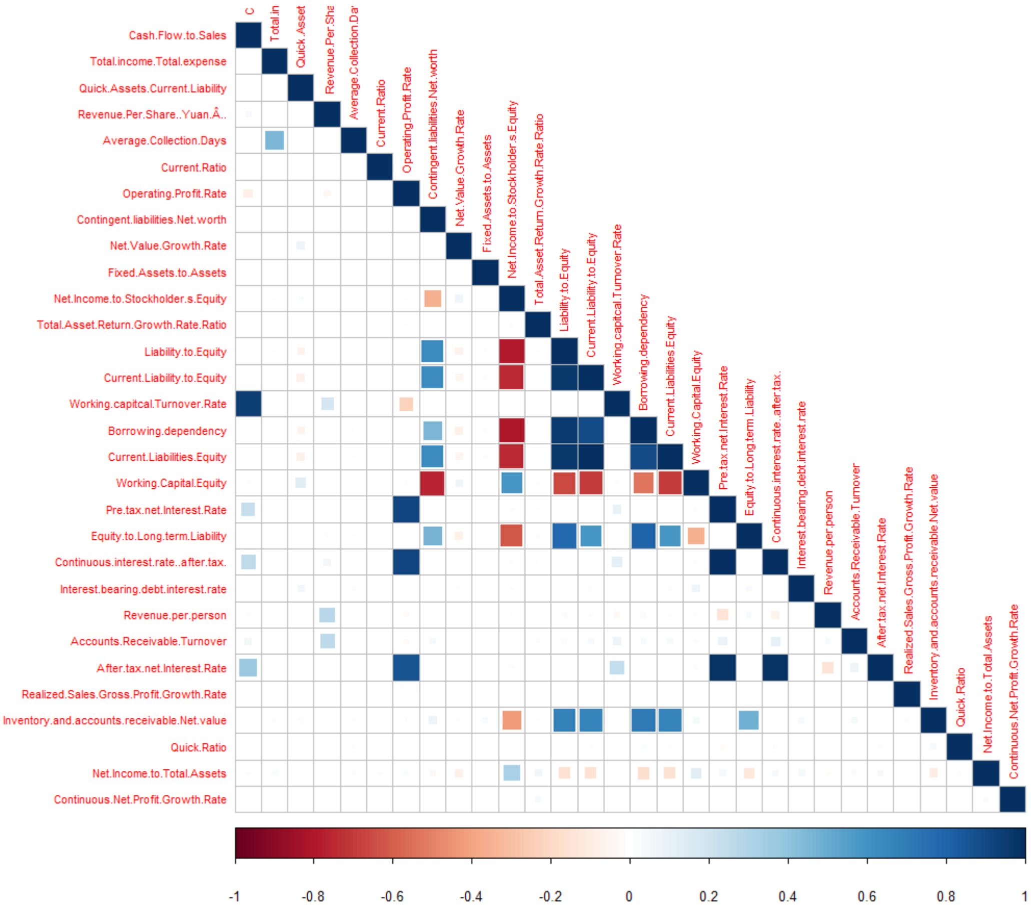


Figure: Plot for multicollinearity in the variables selected from Boruta

Multicollinearity variablesfrom Boruta **(with Correlation > 0.75: Removed)**:

1. LiabilitytoEquity
2. CurrentLiabilitytoEquity
3. CurrentLiabilitiesEquity-
4. Borrowingdependency-
5. WorkingCapitalEquity- Solvency
6. AftertaxnetInterestRate
7. Continuousinterestrateaftertax
8. Pre-taxnetInterestRate-
9. CashFlowtoSales- Turnover

Data Source

Deron Liang and Chih-Fong Tsai, deronliang@gmail.com; cftsai'@mgt.ncu.edu.tw, National Central University, Taiwan. The data was obtained from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction

References

Liang, D., Lu, C.-C., Tsai, C.-F., and Shih, G.-A. (2016) Financial Ratios and Corporate Governance Indicators in Bankruptcy Prediction: A Comprehensive Study. European Journal of Operational Research, vol. 252, no. 2, pp. 561-572.

https://www.sciencedirect.com/science/article/pii/S0377221716000412