ST537: Project Report

Bankruptcy Prediction of a company based on its econometric measures

Bhavana Lalwani, Mallika Sinha, Vikram Patil

North Carolina State University, Raleigh.

Abstract:

Financial measures of a company can provide valuable information about the company's future. The current financial measures can help predict the bankruptcy of a company in the near future. However, many complex econometric measure play role in the bankruptcy of a company, therefore, statistical learning based prediction model could be useful for bankruptcy prediction. The data for Polish companies with information about econometric measures of several companies and their bankruptcy status after one year is considered for predictive model development. This real data contains many missing values and bankruptcy status of a company is a rare event, therefore the data is highly imbalanced. Additionally, many of the econometric measures of the company are highly correlated. The principal component analysis method is used for imputing missing data and dimensionality reduction. Furthermore, predictive models considering various supervised classification approaches like the decision trees, support vector machine (SVM), logistic regression, neural network, and k-nearest neighbor (KNN) are explored to estimate the accurate class of bankruptcy of a company. It is found that the KNN leads to a best predictive model based on statistical criteria like high accuracy, minimum apparent error, sensitivity, and high specificity.

Keywords:

Supervised learning, Regression Trees, Logistic Regression, KNN model, Neural Network

1. Introduction

Bankruptcy prediction of companies is a complex problem. Prediction of the possible bankruptcy of a company help guide investors, policy makers and company management to take make necessary precautionary measures to avoid unforeseen financial distress. The current financial attributes of the company play an important role in deciding the bankruptcy status in the near future. It involves many variables like net sales, liabilities, gross profit, working capital, etc.

The dataset considered is for Polish companies in the year 2007. This data contains information about 64 financial attributes of 7027 companies. The bankruptcy status for each of the companies after a year is also known. The dataset has very imbalanced data where a bankruptcy status is a rare event. From 7027 companies, only 271 belong to bankrupted companies and rest 6756 companies did not bankrupt. Also, there are many missing values for financial attributes. If the companies with at least one missing attribute are removed, then there are only 3194 companies with information for all attributes of which only 30 are bankrupted and remaining 3164 are not bankrupt. Therefore, the data becomes extremely unbalanced with the removal of companies with at least one missing value.

The objectives of this project are to achieve maximum accuracy of bankruptcy prediction using various classification models. Specifically, the answers to the following questions are explored through statistical data analysis for the dataset for the Polish companies.

1. Can we predict the bankruptcy status a company from its econometric measures?
2. If yes, which model is best for classification of imbalanced data?
3. Can the data reduction be done and used for modelling without compromising prediction accuracy?

2. Methodology

The data contains various financial attributes of the company. These financial attributes of the companies are expected to be highly correlated among themselves. Therefore, multivariate analysis by accounting correlation between predictors is considered for modeling of this data for classification.

The data contains many missing values (NAs). There are 5835 missing entries in a total of 449728 entries which is about 1.3% of total data. So, firstly missing values was replaced with suitable parameter as omitting only increased the imbalance in the data.

1. The imputation of missing values is performed first using the method of Principal Component Analysis.
2. The updated data by replacing missing values is then used principal component analysis to reduce the dimensionality of original data without significant loss of information from the data.

The different prediction models are explored for better accuracy of bankruptcy prediction using various modeling approaches. Models based on regression trees, support vector machine (SVM), logistic regression, neural network, and k-nearest neighbor (KNN) are considered in this study.

1. Each predictive model is analysed on how they perform with complete data and PCA reduced data.
2. Analysis of better predictive model is performed based on minimum apparent error (APER) or high accuracy, high specificity, and high sensitivity obtained from the confusion matrix from the model prediction.

3. Results and Discussion

3.1. Imputation of missing values

The missing values in the dataset are replaced by prediction based on the similarity between covariates and the relationship between them. Principal components are considered to identify the similarity between covariates. The criteria of Mean Square Error of Prediction (MSEP) was considered to evaluate the number of principal components needed for imputation. Figure 1 shows a plot of MSEP values as a function of the number of principal components retained for imputation of missing values. It is observed that the first 2 principal components result in significantly lower MSEP and additional principal components just provide a marginal improvement in MSEP. Therefore, missing values are predicted using the first 2 principal components. The completed data with missing values replaced with corresponding predicted values are used for further analysis.

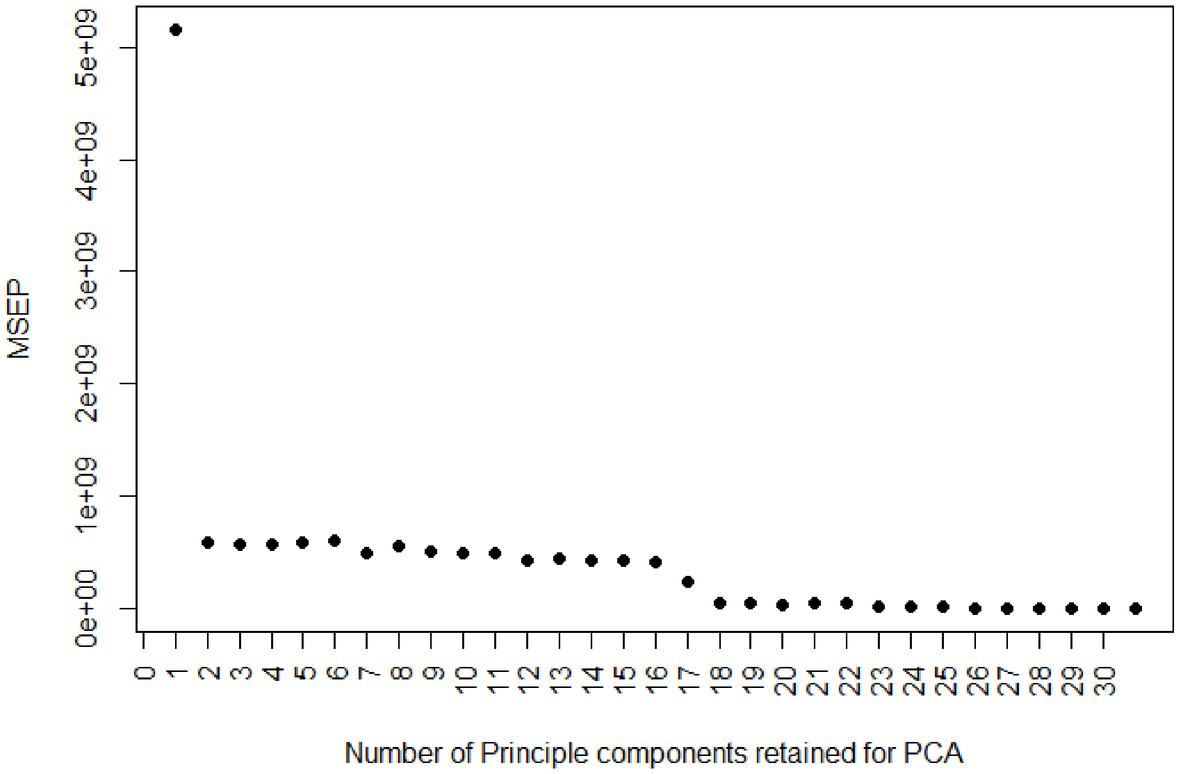


Figure 1: Mean Square Error of Prediction (MSEP) for number of principal components retained for missing value imputation.

3.2. Principal Component Analysis

Many of the financial metrics of a company are expected to have a strong correlation with each other. Figure 2 shows the correlation plot for all 64 predictors. It is observed there are many clusters of variables with strong positive correlation (dark blue circles) and strong negative correlation (dark brown circles). This strong correlation indicates that many of the predictors add little information to the data. Therefore, principal component analysis is considered to attain mutually uncorrelated covariates and dimensionality reduction.

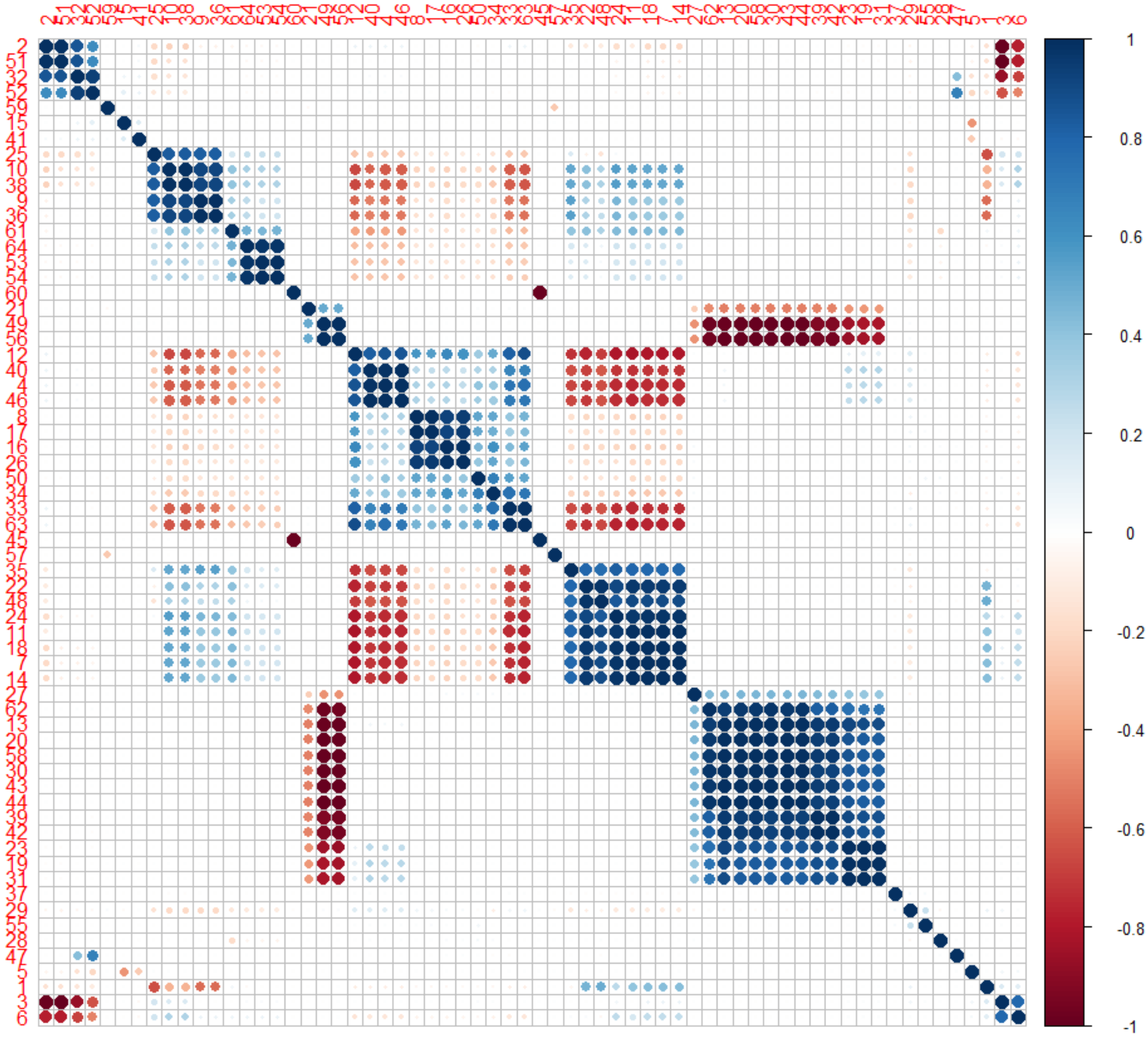


Figure 2: Correlation plot for all 64 predictor variables (Information about labels 1-64 is available in reference [1])

The principal component analysis is performed using eigenvalue decomposition of the correlation matrix (covariance matrix of normalized data) on the predictors using complete dataset. Figure 3 shows the cumulative percentage variance explained as a function of the number of principal components. Clearly, a significant rise in percentage variance explained in the initial part of the plot indicate that a few principal components (PCs) can explain much of the information in the predictors. The first 21 principal components explain about 95% of the variation in the data whereas the first 26 principal components capture about 99% of the variation in the data. The flat profile for the number of PCs higher than 33 indicates that additional PCs do not add any meaningful information from the data. As 95% of variation from the data would be good enough to provide meaningful information for the prediction purposes, the first 21 principal components are considered for model building.

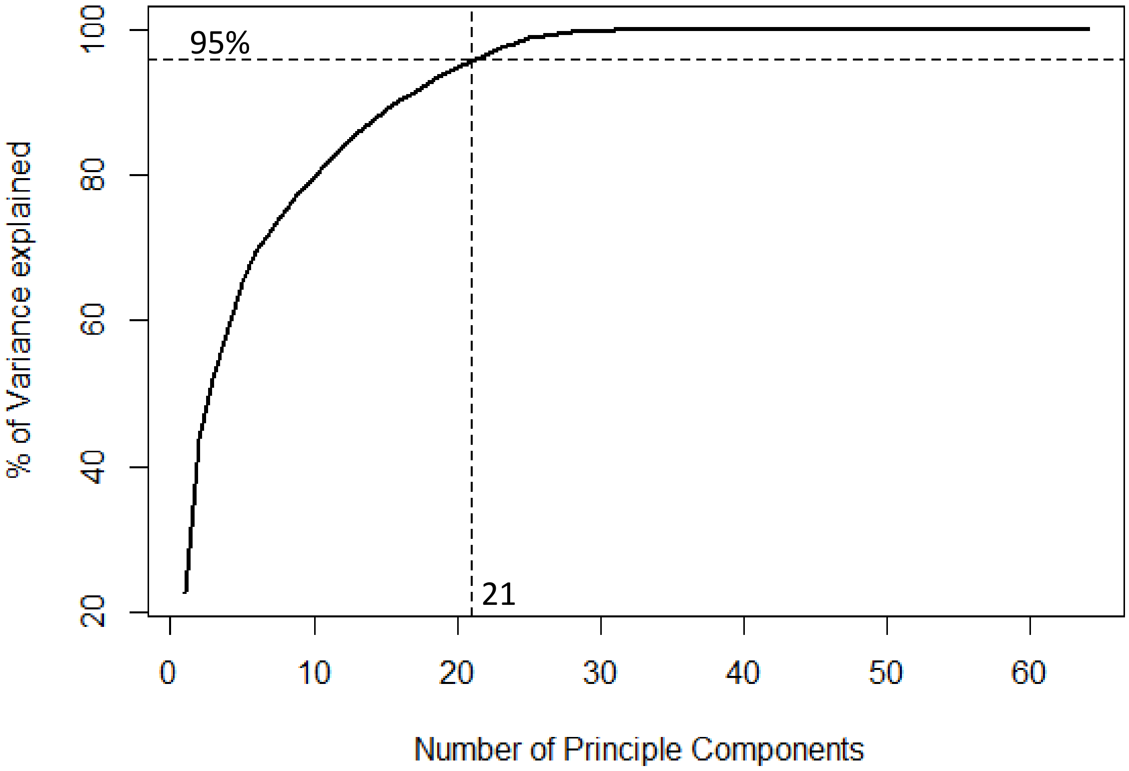
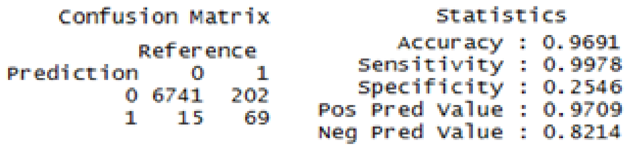


Figure 3: Cumulative % of variation explained for various number of principal components.

3.3. Regression Tree Model

The regression tree model is considered for classification. As regression trees can handle missing data, original data with all 64 attributes (predictors) are considered in regression tree modeling. A regression tree is built based on decision boundary, it is easy to interpret. However, regression trees can perform badly for prediction if there are not clear and distinct decision boundaries based on attributes. Figure 4 shows the regression tree model for this classification using all the observations in the dataset. The confusion matrix and various accuracy statistics of the regression tree model are shown in Table 1. Overall, the regression tree model has good accuracy but lower prediction accuracy for bankruptcy observation.

Table 1: Results of Regression Tree model



```
  Confusion Matrix                      Statistics

           Reference                Accuracy : 0.9691
Prediction    0    1             Sensitivity : 0.9978
         0 6741  202             Specificity : 0.2546
         1   15   69          Pos Pred Value : 0.9709
                              Neg Pred Value : 0.8214
```

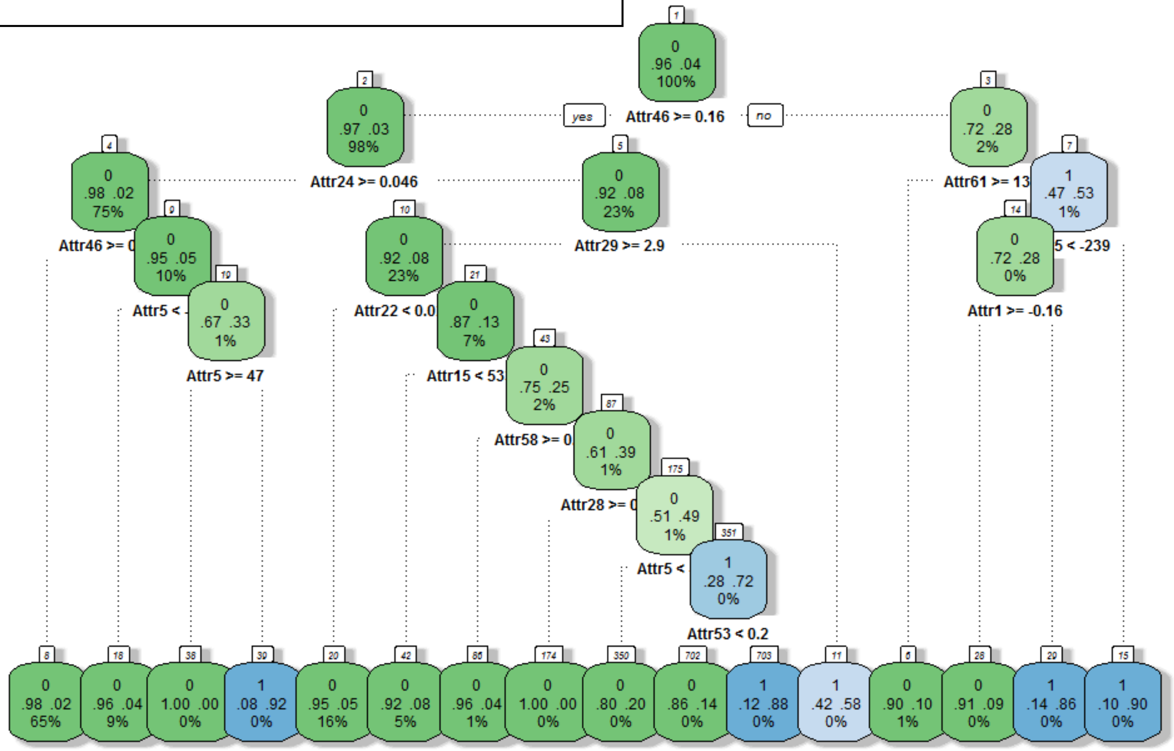


Figure 4: Classification decision tree using original data with missing values
(0- Not bankrupt, 1- Bankrupt) (Information about Attr1-Attr64 is available in reference [1])

3.4. Support Vector Machine Model

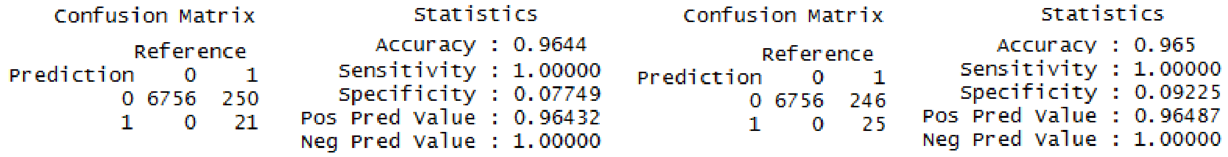
A model based on Support Vector Machine (SVM) is considered for prediction. It is observed that the SVM model is not robust enough to handle missing data, therefore, updated complete data (all 64 predictors) after PCA based imputation is first considered. Furthermore, reduced data with 21 covariates based on PCA are considered to see the change in accuracy with the smaller dimension data. Various kernels functions with SVM models are explored and it is observed that Radial Basis Gaussian kernel shows the best overall accuracy. Table 2 shows confusion matrix and accuracy statistics for the SVM models. Table 2-(a) is for the model using the original 64 attributes as predictors whereas Table 2-(b) is for reduced data using PCA. Both the SVM models could classify non-bankrupt companies with high accuracy but shows a very small prediction accuracy for bankrupt companies (low specificity). Both the SVM models predict non-bankrupt companies with 100% accuracy. An interesting observation is that the reduced dimension could improve the accuracy of bankruptcy prediction marginally. This indicates that principal components are relatively easy to create support vectors without compromising much of the information from the data for prediction.

In comparison to regression trees, SVM models show lower accuracy and lower specificity but have higher sensitivity. However, regression trees could use data with missing values whereas SVM required imputation of missing values.

Table 2: Summary Results of Support Vector Machine Models

(a) Using Original 64 attributes as predictors

b) Using 21 PCs as predictors



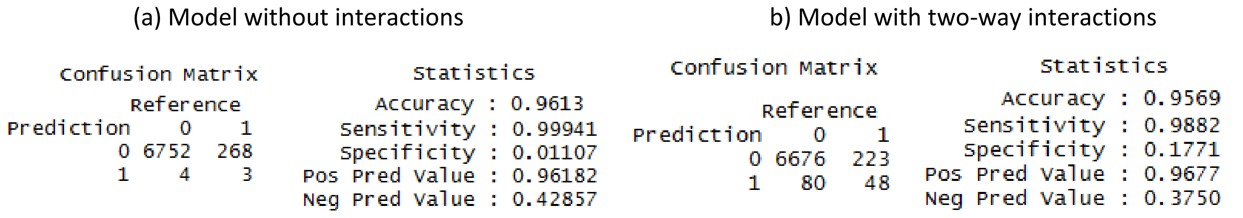
```
   Confusion Matrix                     Statistics
                                 Accuracy : 0.9644
          Reference           Sensitivity : 1.00000
Prediction    0    1          Specificity : 0.07749
         0 6756  250       Pos Pred Value : 0.96432
         1    0   21       Neg Pred Value : 1.00000
```

```
   Confusion Matrix                     Statistics
                                 Accuracy : 0.965
          Reference           Sensitivity : 1.00000
Prediction    0    1          Specificity : 0.09225
         0 6756  246       Pos Pred Value : 0.96487
         1    0   25       Neg Pred Value : 1.00000
```

3.4. Logistic Regression Model

The logistic regression model is a generalized linear model for binary response. As the response variable for bankruptcy status is binary, it can be used to build a prediction model. First, a model without any interactions between covariates is considered with 21 PCs as covariates. Such a model can capture non-bankrupt companies with good accuracy (high sensitivity), however, the prediction is not good for bankrupt companies as observed from the results shown in Table 3-(a). The plot of predicted probabilities indicates that many observations lies in between 0 and 1 suggests the need for additional factor for the better prediction. Moreover, a model with all two-way interactions is fitted and it is observed that such a model shows predicted probabilities in the extreme end as needed for classification. Prediction of bankrupt status is better with two-way interaction model but slightly smaller overall accuracy of prediction.

Table 3: Summary Results of Logistic Regression Models



(a) Model without interactions

```
         Confusion Matrix

              Reference
Prediction     0     1
         0  6752   268
         1     4     3
```

```
          Statistics
     Accuracy : 0.9613
  Sensitivity : 0.99941
  Specificity : 0.01107
Pos Pred Value : 0.96182
Neg Pred Value : 0.42857
```

b) Model with two-way interactions

```
   Confusion Matrix

          Reference
Prediction    0    1
         0 6676  223
         1   80   48
```

```
           Statistics
      Accuracy : 0.9569
   Sensitivity : 0.9882
   Specificity : 0.1771
Pos Pred Value : 0.9677
Neg Pred Value : 0.3750
```

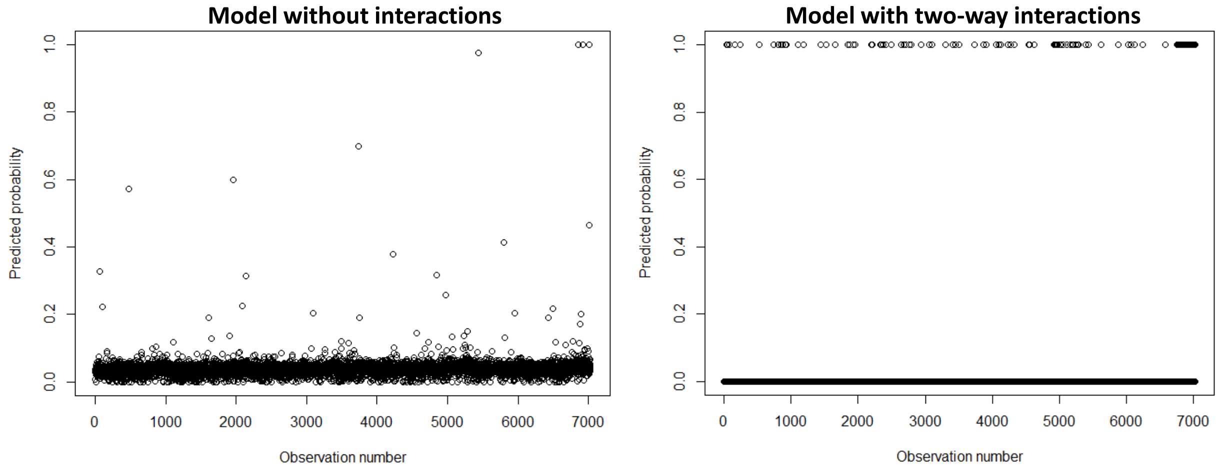


Figure 5: Predicted probabilities for Logistic Regression models (left plot is for model without interactions and right plot is for model with two-way interactions)

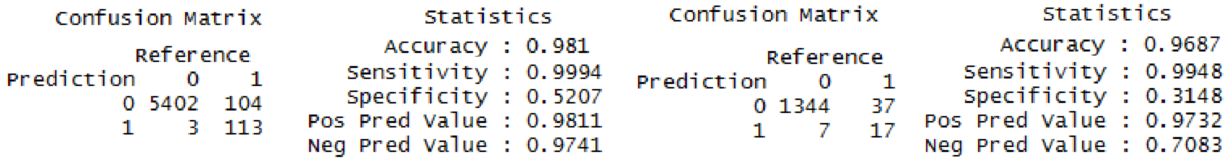
3.5. Neural Network Model

The neural network model is an extension of logistic regression with multiple layers of logistic regression. Neural network model can consider the non-linear relationship between predictors and higher order interactions. However, training of neural net and finding an optimal network structure is a computationally intensive process. The reduced dimensional data of PCs is divided into training and testing with 80:20 proportion for model training and testing respectively. A neural network with a single hidden layer of 5 nodes is considered as shown in Figure 6. The model is built using training data and predictions are performed on testing data. Table 4 shows the results of the neural network model on training and testing data. It can be observed that prediction accuracy is better on training data but relatively less on test data. The neural network model shows a better prediction of bankrupt companies (higher specificity) than logistic regression.

Table 4: Summary Results of Neural Network Model

(a) For Training data

b) For Testing data



(a) For Training data

```
   Confusion Matrix                           Statistics
                                        Accuracy : 0.981
              Reference              Sensitivity : 0.9994
Prediction     0     1               Specificity : 0.5207
          0 5402   104            Pos Pred Value : 0.9811
          1    3   113            Neg Pred Value : 0.9741
```

b) For Testing data

```
   Confusion Matrix                           Statistics
                                        Accuracy : 0.9687
              Reference              Sensitivity : 0.9948
Prediction     0     1               Specificity : 0.3148
          0 1344    37            Pos Pred Value : 0.9732
          1    7    17            Neg Pred Value : 0.7083
```

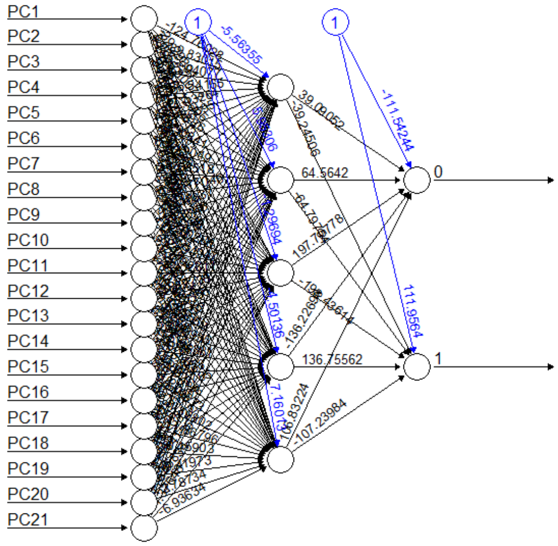
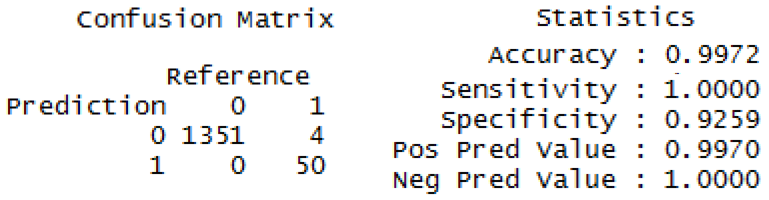


Figure 6: Trained Neural Network using single hidden layer of 5 nodes.

3.6. K-Nearest Neighbor (KNN) Model

The KNN is a non-parametric model for classification which uses majority votes among k-neighbors identify the class of the observation. The number of neighbors is tuning parameter which is determined to evaluate repeated cross-validation accuracy. Figure 7 shows accuracy for the model with a different number of neighbors. It was observed that cross-validation accuracy is optimal for k=7 neighbors. Therefore, a model with 7 neighbors is fitted on training dataset (80% data) and prediction on test data (20% data) is performed. Table 5 shows the confusion matrix and various accuracy statistics for this model. The performance of the KNN model is very good for predicting both bankruptcy and non-bankruptcy of companies.

Table 5: Summary Results of KNN model with 7 neighbors



```
     Confusion Matrix                    Statistics
                                   Accuracy : 0.9972
           Reference            Sensitivity : 1.0000
Prediction    0    1            Specificity : 0.9259
         0 1351    4         Pos Pred Value : 0.9970
         1    0   50         Neg Pred Value : 1.0000
```

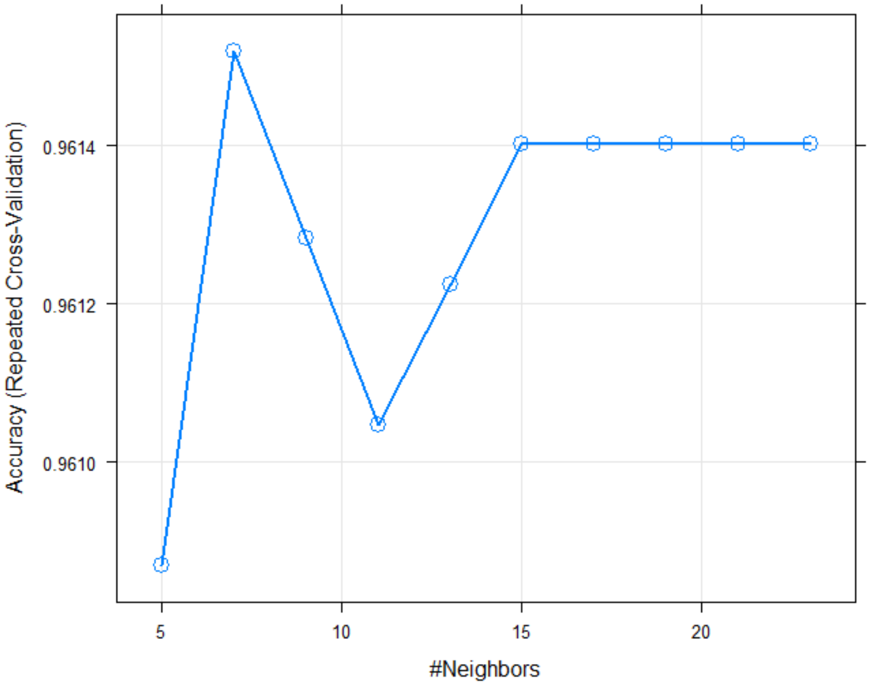


Figure 7: Accuracy of KNN model with number of neighbors (k).

4. Conclusions

Classification models for bankruptcy prediction of a company using its econometric measures are explored using multivariate supervised learning approaches. This study addresses the handling of missing financial attributes, dimensionality reduction, and predictive model development with unbalanced data. As simply omitting the missing values increases the imbalance in the data, imputation of missing values is performed using the method of Principal Component Analysis. Additionally, dimensionally reduction from 64 financial attributes to 21 features is performed using the principal component analysis. Predictive models using various approaches including regression tree, support vector machine (SVM), logistic regression, neural network, and k-nearest neighbors (KNN) is studied. The apparent error associated with the regression tree model is 3.08%, SMV model is 3.5%, logistic regression model is 4.31%, neural network model is 3.13%, and KNN model is 0.28%. Many models show high accuracy of prediction of non-bankruptcy status but show poor performance on bankruptcy status due to unbalanced data. The KNN model with 7 neighbors is the best predictive model with all the statistical criteria of maximum accuracy, high specificity, and high sensitivity.

References

[1] Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58, 93-101. Webpage: https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

[2] Maher Maalouf, Dirar Homouz, Theodore B. Trafalis. Logistic regression in large rare events and imbalanced data: A performance comparison of prior correction and weighting methods. Webpage: https://onlinelibrary.wiley.com/doi/full/10.1111/coin.12123

[3] David A. Cieslak Nitesh V. Chawla. Learning Decision Trees for Unbalanced Data. Webpage: https://link.springer.com/chapter/10.1007/978-3-540-87479-9_34