A Comparison of Machine Learning Algorithms for Prediction of Past Due Service in Commercial Credit

Liyuan Liu M.A, M.S.
Analytics and Data Science, lliyuan@students.kennesaw.edu

Jennifer Lewis Priestley Ph.D.
Analytics and Data Science, jpriestl@kennesaw.edu

Follow this and additional works at: https://digitalcommons.kennesaw.edu/dataphdgreylit

Part of the Business Analytics Commons, Finance and Financial Management Commons, and the Statistics and Probability Commons

Recommended Citation
https://digitalcommons.kennesaw.edu/dataphdgreylit/8

This Article is brought to you for free and open access by the Ph.D. in Analytics and Data Science Research Collections at DigitalCommons@Kennesaw State University. It has been accepted for inclusion in Grey Literature from PhD Candidates by an authorized administrator of DigitalCommons@Kennesaw State University. For more information, please contact digitalcommons@kennesaw.edu.
A Comparison of Machine Learning Algorithms for Prediction of Past Due Service in Commercial Credit

Liyuan Liu
Analytics and Data Science
Kennesaw State University
1000 Chastain Road
Kennesaw, Georgia 30144
Email: lliyuan@students.kennesaw.edu

Jennifer Lewis Priestley
Analytics and Data Science
Kennesaw State University
1000 Chastain Road
Kennesaw, Georgia 30144
Email: jpriestl@kennesaw.edu

Abstract—Credit risk modeling has carried a variety of research interest in previous literature, and recent studies have shown that machine learning methods achieved better performance than conventional statistical ones. This study applies decision tree which is a robust advanced credit risk model to predict the commercial non-financial past-due problem with better critical power and accuracy. In addition, we examine the performance with logistic regression analysis, decision trees, and neural networks. The experimenting results confirm that decision trees improve upon other methods. Also, we find some interesting factors that impact the commercials' non-financial past-due payment.

Index Terms—Past Due; Credit Risk Model; Logistic Regression; Decision Trees; Neural Networks;

I. INTRODUCTION

Commercial loan is a phrase commonly used to indicate a loan not ordinarily maintained by either the real estate or consumer loan departments. In asset distribution, commercial or business loans comprise one of the most critical assets of a bank [1]. The volume of the commercial loan is exceptionally high; according to Real Capital Analytics (RCA) report, the total global amount of commercial loan was $826 billion in 2016. Nowadays, the loan default still happened usually in the commercial lender. Based on the Federal Reserve senior loan officer opinion survey report, oil and gas companies defaulted on $39 billion in 2016, and the high yield bond default rate for the energy sector peaked at 18.8% during the year. In fact, the loan defaults happened in every industry. Therefore, the commercial credit risk prediction is a critical research part that helps to protect the economic environment.

In this study, we used a real-world dataset provided by Equifax and compare different machine learning algorithms including logistic regression, decision trees, and neural networks. The modeling process is described in Figure 1. At first, we described the process of data imputation, transforming, and selecting model variables from messy and sparse data. Secondly, we illustrated the logistic regression, decision tree, and neural network algorithms. Finally, we discussed the results, compared the models using ROC index and Kolmogorov-Smirnov statistic and found the most important factors that impact commercial credit.

II. RELATED WORK

There are many related works in credit risk modeling research. Nargundkar and Priestley examined and compared the most prevalent modeling techniques in the credit industry [2]. Laitinen predicted corporate credit analyst's risk estimate by the weighted logistic and linear regression analyses [3]. The datasets included 35 variables from 3200 observations. Atiya developed a neural network model to predict bankruptcy [4]. First, the author reviewed the topic of bankruptcy prediction, with emphasis on neural-network (NN) models. Second, the author developed an NN bankruptcy prediction model. Huang et al. employed support vector machines and neural networks to credit rating analysis [5]. The authors used backpropagation neural network (BNN) as a benchmark and obtained prediction accuracy around 80% for both BNN and SVM methods for the United States and Taiwan markets. The dataset included 74 cases with bank credit rating and 21 financial variables, which covered 25 financial institutes from 1998 to 2002. Wang et al. used the fuzzy support vector machine to evaluate credit risk [6]. Authors provided a new fuzzy SVM to evaluate
the credit risk of consumer lending. The dataset contains 30 failed and 30 non-failed firms. Twelve variables are used as the firms’ characteristics. Lin et al. published a survey that reviewed 130 related journal papers from the period between 1995 and 2010, focusing on the development of state-of-the-art machine-learning techniques, including hybrid and ensemble classifiers in financial crisis prediction research [7]. Pinches et al. utilized multiple discriminant analysis (MDA) to bear a linear discriminant function relating a set of independent variables to a dependent variable to better suit the ordinal nature of bond-rating data and increase classification accuracy. Other researchers also utilized logistic regression analysis and probit analysis [8], [9], [10]. Recently, artificial intelligence techniques and machine learning techniques such as neural networks and decision trees have been used to support such analysis. Neural networks emerged in Hagan et al.’s research and are widely used in many different domains, including credit risk modeling [11].

III. DATA DISCOVERY

This large-scale study examined 36 separate datasets, with each dataset drawing quarterly commercial loan related information from 2006 to 2014. Each dataset contains 11,787,287 observations and 305 explanatory attributes from different commercials. These explanatory variables include five categories: non-financial account, telecommunication account, utility account, service account, and industry account. This study focuses on predicting commercial past due activities in their service accounts using SAS9.4.

A. Assignment of Dependent Variable

We assigned “totNFPDAmt12mon” as target variable of interested which represent the total non-financial past due amount in last 12 months. There are two main reseasons why we chose this variable as the target variable. Our first goal was to predict if the commerce would have non-financial past due in the future. Usually, the financial loan is the most prioritizes by the commerce because of the significant cost of the late payment. However, sometimes, even though the commerce doesn’t have financial past dues, it can still have the non-financial past dues. Therefore, predicting commerces non-financial past dues is more accurate than predicting commerces financial past dues to estimate if the commerce will default. Second, we chose a dependent variable that did not contain more than 80% coded values. We did this to ensure our variable had enough information for accurate prediction. Figure 2 shows the distribution of the selected target variable; the variable has 60% coded value which fits our rules. Besides, based on the previous research, the typical performance window of credit risk model is 12-24 months [12]. Ultimately, we decided to choose the total non-financial past due amount in last 12 months as the dependent variable. We created a new variable “GOODBAD” as the target variable. Figure 3 shows the example of the new target variable creation step. The value of “totNFPDAmt12mon” represents the commercial non-financial past due amount in last 12 months. If the “totNFPDAmt12mon” value is 0, we define the “GOODBAD” value as 0. Otherwise, “GOODBAD” is defined as 1.

B. Independent Variables Cleansing and Imputation

There were three steps of the cleansing and imputation of explanatory variables.

(i) The first step was dimensionality reduction by removing variables with the high ratio of coded or missing values. There are many pieces of research about how to remove the variables that including missing values. Based on the book *Discovering Knowledge in Data: An Introduction*
to Data Mining, variables that contain more than 90% missing or coded values should be removed [13]. In addition, based on Allison’s research, variables should be deleted when they include 70% missing values or higher [14]. In this case, we employed Allison’s research by deleting the independent variables containing more than 70% missing or coded values. We removed the categorical variables which only contain one level. We also removed the categorical variables that indicate specific information such as zip code and state name.

(ii) The second step was imputing the missing or coded values by median and most frequency. Median imputation is a method replacing any missing or coded value with the median of that variable for all other cases, which has the benefit of not changing the sample median for that variable. Most frequency imputation is an imputation method of the categorical variable that is replacing missing or coded value with the most frequency. In Bennett’s research, he points out some different methods to impute the missing value such as last value carried forward, mean substitution, regression methods, hot-deck imputation and cold-deck imputation [15]. In this study, since the numeric data are continuous and the dataset is extensive, we chose the median imputation for numeric variables and the most frequency imputation for categorical variables. Median imputation is a method in which the median of the available cases replaces the missing value on a particular variable. This method maintains the sample size and is easy to use [16]. This imputation method will help the variables avoid a skewed distribution. Figure 4 shows the example of a numeric variable’s distribution pre-imputation and post-imputation. From Figure 4, we can see the median imputation of the valid values is more stable in highly skewed data. After the first and second data cleaning steps, there are 109 variables, 96 numeric variables, and 13 categorical variables remaining. Before going to the next step, we ran the c-test and found that there was no significant improvement of models between including and excluding the categorical variables. Therefore, before going to the next step, we removed the categorical variables.

(iii) The third step was dimensionality reduction by variable clustering. Dimension reduction is a process of reducing
the number of random variables under consideration by obtaining a set of essential variables. It is a critical data preprocessing technique for large-scale and streaming data classification tasks. In order to speed up the modeling process, the predictor variables should be grouped into similar clusters. A few variables can then be selected from each cluster-this way the analyst can quickly reduce the number of variables and speed up the modeling process [17]. In this study, we found the cluster of variables that were highly correlated among themselves and not correlated with variables in other groups. Chose the variable which had the highest $R^2$ ratio in each cluster, thus reducing the dimension of the dataset. The $R^2$ ratio can be calculated by:

$$R^2 Ratio = \frac{R^2_{OwnCluster}}{R^2_{NextClosest}}$$ \hspace{1cm} (1)

Next, the cluster representatives are put into the predictive model. Figure 5 shows the example of how to choose the representative variable in each cluster. Figure 6 shows that 40 clusters explained approximately 81% of the variability in the data, so we chose the best 40 variables to consider as predictors in our model.

C. Multicollinearity Assessment

Multicollinearity may have several opposing impacts on estimated coefficients in classification regression analysis. Consequently, it is essential that researchers should focus on detecting its existence. Analyzing latent roots and latent vectors of the correlation matrix and the variance inflation factors (VIF) is necessary for the analysis process. A VIF quantifies how much the variance is inflated. The VIF of $K^h$ predictor can be calculated by:

$$VIF_K = \frac{1}{1 - R^2_K}$$ \hspace{1cm} (2)

The VIF of 1 indicates that there is no correlation between the $K^h$ predictor and the remaining predictor variables. The general rule of thumb is that VIFs exceeding 4 warrant further investigation, while VIFs exceeding 10 are signs of severe multicollinearity requiring correction [18]. In this study, we tested VIFs of the 40 variables. Figure 7 shows the example of VIFs of a part of variables. From the results, there are two variables in which the VIF values are greater than 10, so we removed these two variables because it will cause the multicollinearity problem. After checking the multicollinearity of the 40 variables, we found some variables represent same meanings with the target variables. Based on the financial knowledge and the data description of the codebook, we removed these variables. Finally, there were 16 independent variables that remained in the dataset.

IV. METHOD

A. Logistic Regression

Logistic regression is an important machine learning algorithm. The goal is to model the probability of a random variable $Y$ is 0 or 1 given experimental data. The brief explanation of logistic regression concepts is below [19]. The generalized linear model function of logistic regression can be defined with the parameter $\theta$.

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$ \hspace{1cm} (3)

The likelihood function is below, that assuming all the samples are independent.

$$L(\theta|x) = Pr(Y|X; \theta) = \prod_i Pr(y_i|x_i; \theta) = \prod_i h_\theta(x_i)^{y_i}(1 - h_\theta(x_i))^{1-y_i}$$ \hspace{1cm} (4)

Typically, the log likelihood is maximized with a normalizing factor $N^{-1}$. This maximized is used in gradient descent.

$$N^{-1}logL(\theta|x) = N^{-1} \sum_{i=1}^{N} logPr(y_i|x_i; \theta)$$ \hspace{1cm} (5)

Assuming the $(x,y)$ pairs are drawn uniformly from the underlying distribution, then in the limit of large $N$, The $H(Y|X)$ is the conditional entropy and the $D_K L$ is the Kullback-Leibler divergence.

$$\lim_{N \to +\infty} N^{-1} \sum_{i=1}^{N} logPr(y_i|x_i; \theta) = \sum_{x \in X, y \in Y} Pr(X = x, Y = y)logPr(Y = y|X = x; \theta)$$

$$= \sum_{x \in X, y \in Y} Pr(X = x, Y = y)Pr(X = x, Y = y)$$

$$= \sum_{x \in X, y \in Y} (\log Pr(X = y|X = x)) + \log Pr(Y = y|X = x)$$

$$= -D_K L(Y||Y_\theta) - H(Y|X)$$

We used stepwise selection method for variables screening at the significant level of 0.05. After we generated the classification table, a cutoff value of $P$ was selected for prediction purpose. $P$ is calculated by the probability that each observation will belong to one of the two classes. The math formula of this regression, where $B$ denotes the models parameters and $X$ refers to the input parameters:

$$P = \frac{1}{1 + e^{-(B_0+B_1X_1+...+B_nX_n)}}$$ \hspace{1cm} (7)
When making predictions on the validation set, observations with the appropriate P value exceed the cutoff value are predicted as 1 while those with the fitted value of P smaller than the cutoff value are predicted as 0.

B. Decision Tree

The decision tree is currently one of the most interesting supervised learning algorithms [20]. In this algorithm, an empirical tree expresses a classification of the data that is created by applying a series of simple rules. These algorithms produce sets of rules which can be employed for prediction through the repeated process of splitting. The chi-squared automatic interaction detection (CHAID), classification and regression trees (CART), C4.5 and C5.0 are some of the most common tree methods. Information gain and entropy are used to create the trees [21]. Information gain estimates how well a given attribute separates the training examples according to their dependent variable. The measure is used to choose among the candidate variables at each step while growing the tree. Information gain \( (S, X) \) of a variable \( X \) relative to a collection of examples \( S \), is defined as:

\[
Gain(S, X) = Entropy(S) - \sum_{v \in value(X)} \left| S_v \right| \frac{Entropy(S_v)}{|S|} \tag{8}
\]

Entropy is a measure of homogeneity that can be defined as below, \( c \) denotes to the number of classes, \( P_i \) denotes to the proportion of sample \( S \) that belong to class \( i \).

\[
Entropy(S) = \sum_{i=1}^{c} -P_i \log_2 P_i \tag{9}
\]

In this study, the class \( i=2 \) because our dependent variable is binary.

C. Neural Networks

The neural network is another general method of regression and classification. Neural networks were initially developed by researchers trying to simulate the neurophysiology of the human brain. The feedforward neural network is the most manageable and most popular application. Training a neural network is the process of setting the best weights on the inputs of each of the units, and backpropagation (backdrop) is the most common method for computing the error gradient for a feedforward network [22]. Back-propagation is a supervised learning technique used in neural networks and is most suitable for diagnostic and predictive problems. Backpropagation neural network involves multi-layer topology that includes an input layer, a hidden layer, and an output layer [23]. In this study, we used linear combination functions in the hidden layer and sigmoid function in the output layer. The formula for the sigmoid is below:

\[
sigmoid(x) = \frac{1}{1 + e^{-x}} \tag{10}
\]

The number of hidden layers applied is defined by evaluating the generalization error of each network.

V. Model Development and Comparison

A. Logistic Regression

Before testing the model, we split the dataset with the 16 explanatory variable as 70% training data and 30% validation data. We employed the stepwise selection and generated a classification table based on the training data. Figure 8 shows the classification table, and it suggests the cutoff value of \( P=0.5 \) should be selected since this value can reach a relatively high accuracy (84.6%) and relative low false negative (15.5%) on the training dataset. False negative is a test result indicates a condition does not hold, while in fact it does. The false negative rate and accuracy calculated by:

\[
ACC = \frac{TP + TN}{FN + FP + TP + TN} \tag{11}
\]

\[
FNR = \frac{FN}{TP + FN} \tag{12}
\]

This cutoff value was then used to make predictions on the validation set. Figure 9 demonstrates the ROC curve of logistic regression based on the validation dataset. The ROC AUC of logistic regression algorithm is 88.55%. The coefficient estimations and the P value for these 16 variables are shown in Figure 10. The variable with larger Wald Chi-Square values.
are considered to be more important in making predictions. The Kolmogorov-Smirnov statistic is a method to evaluate the models' performance. The Kolmogorov-Smirnov statistic quantifies a distance between the experimental distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples [24]. Figure 11 shows the KS test result of logistic regression. The D value of the KS test is 0.634 which shows the model has a good performance.

**B. Decision Trees and Neural Networks**

Figure 12 presents the variable importance of the decision tree. Figure 13 shows the ROC curve of the decision tree and neural networks. The neural networks get the highest accuracy, and both neural network and decision tree’s ROC AUC are higher than logistic regression.
Table I shows the accuracy and D value of KS test for each model. From the table, we can find that neural networks have the best accuracy and the decision tree has the best D value in KS test. It shows the new machine learning algorithms have better performance compare with the logistic regression.

VI. CONCLUSION

In this study, we used the real world data provided by Equifax and compared the different machine learning algorithms including logistic regression, decision tree, and neural network. We found that, even though neural networks returned the best accuracy, it is harder to explain due to the calculations hidden layers and variety nodes, especially for the risk models. The decision tree is our recommended model since it has superior performance and it is easy to interpret. The most critical factors that affect the commerce's past due are: “NoNFABalance3mon”, the Number of Non-Financial Accounts with Balance Reported in the Last 3 Months; “TotUtilNFA3mon”, the Total Utilization on Non-Financial Accounts; “totCuronSA”, the Total Current Balance on Service Accounts and “YearStarted”, the Commerce Year.

In summary, this study first uses the real world large and unique datasets to build models. Secondly, this study’s explored how to predict if enterprises will have past due in next 12 months of their non-financial account which researchers have seldom focused. Thirdly, in this study, we compared three different models: logistic regression, decision tree, and neural networks. We found decision tree is the best model since it has better performance and is easy to interpret. The last but not the least, we discovered some interesting factors that can impact the enterprises’ credit score.

VII. LIMITATIONS

Our study provides an in-depth illustration of the potential benefits that machine-learning techniques can bring to commercial credit prediction. However, neural networks are not interpretable since they contain hidden layers and nodes. Thus, how to interpret neural networks in commercial credit prediction will be our future research.

VIII. ACKNOWLEDGMENT

I would like to thank Equifax for providing the datasets and Dr. Jennifer Lewis Priestley for providing the classroom framework and fantastic class instruction for completing this study.

REFERENCES