2017

Binary Classification on Past Due of Service Accounts using Logistic Regression and Decision Tree

Yan Wang
Kennesaw State University

Jennifer L. Priestley
Kennesaw State University, jpriestl@kennesaw.edu

Binary Classification on Past Due of Service Accounts using Logistic Regression and Decision Tree

Yan Wang
Department of Statistics and Analytical Sciences
College of Science and Mathematics
Kennesaw State University

Jennifer Lewis Priestley
Department of Statistics and Analytical Sciences
College of Science and Mathematics
Kennesaw State University

***Abstract*—This paper aims at predicting businesses' past due in service accounts as well as determining the variables that impact the likelihood of repayment. Two binary classification approaches, logistic regression and the decision tree, were conducted and compared. Both approaches have very good performances with respect to the accuracy. However, the decision tree only uses 10 predictors and reaches an accuracy of 96.69% on the validation set while logistic regression includes 14 predictors and reaches an accuracy of 94.58%. Due to the large concern of false negatives in financial industry, the decision tree technique is a better option than logistic regression on the given dataset in terms of its relative lower false negative. Accuracy, false positive and false negative are all very important criteria in model selection and evaluation. Decision making should rely more on the research purpose, rather than on the exact values of these criteria.**

Keywords—Past Due, Binary Classification, Logistic Regression, Decision Tree

I. INTRODUCTION

One of the biggest concerns for the financial industry when lending money is the likelihood of repayment. As for a lender institution, businesses are considered as "good" with respect to credit risk if they don't have any past due activities in the accounts while those are considered as "bad" if they have ever had the past due histories. Therefore, an accurate prediction on default activities can provide support to the decision making in financial institutions when small businesses apply for a loan.

The goal of this paper is to find a reliable method that can predict business' past due activities in their service accounts. Since "good" and "bad" businesses are mutually exclusive groups, binary classification methods can be used to solve this problem.

Logistic regression is a traditional technique that is commonly used for binary classification in the financial services domain[1]. In this paper, the probability of the past due is predicted using the logistic regression, then a decision of "good" or "bad" on a certain business is made based on an established cutoff value[2]. However, due to the large number of variables in the given dataset, highly nonlinear relationships between variables may decrease the power of logistic regression. Decision tree, which is becoming more active since Quilan introduced ID3 in 1986[3], is used as an alternative approach for the binary classification problem in this paper.

II. LITERATURE REVIEW

The volume of applications for loans from small businesses grows very fast in recent years. Therefore, decision makers in financial institutions need help to decide whether to approve or disprove a business' application for a loan. A good and effective tool for financial institutions is to use credit-scoring models to predict businesses' possibilities on default. Small businesses that have a high possibility on default may not be approved in the applications for loans.

Crook and Edelman have summarized numerous credit-scoring methods that have been proposed to evaluate the loans performance in the last few years[4]. These methods are either parametric or non-parametric statistical approaches. Logistic regression is a representative parametric statistical approaches and was proposed by Henley[5]. However, this method usually has low prediction accuracy as they cannot capture the nonlinear relationships among the variables, especially analyzing noisy and complex datasets[6].

For non-parametric statistical approaches, decision tree and support vector machine are generally regarded as the most efficient single scoring models to tackle the credit-scoring tasks[7]. The largest advantage for decision tree method is that it can better capture the nonlinear relationships in the dataset without affecting the tree performance. One of the concerns in using decision tree is that its greedy characteristic may lead to the over-sensitivity to the training set, to irrelevant attributes, and to noise. Therefore, some two-stage scoring models have been presented recently to overcome the shortcoming of the single scoring models[8].

III. DATA DISCOVERY

The dataset used in this paper was contributed by Equifax. A total of 36 separate datasets were used, with each dataset representing quarterly financial information from 2006 to 2014. Each dataset contains 11,787,287 observations coming from unique companies. The 305 explanatory variables include

companies' information such as region, zip code as well as the consumer's information such as account activities, liabilities, and liens. Among these explanatory variables, commercial account activities are in the majority. Furthermore, those account variables fall into five categories: non-financial account, telecommunication account, utility account, service account, and industry account. This paper emphasized on predicting businesses' default activities in their service accounts.

A. *The Dependent Variable*

The variable named "totSPDAmt3mon", denoting the total number of past due days reported in service accounts in the last 3 months, is used as the dependent variable in this paper. For each of the 305 variables, the values are "real" if they fall inside the range of 0 to the variable's upper bound subtracting 7. For instance, a variable with values from 0 to 99 has a true value from 0 to 92 while 99 represents values larger than 92 or even missing values. Therefore, 99 is considered as the coded value for this variable and doesn't have meaning in the context of the scale.

Observations with either coded or missing value for totSPDAmt3mon were removed from each of the given 36 datasets. Then, all the 36 datasets were merged together to create a larger dataset that contains 29,691,317 observations.

Since the goal of this paper is to predict whether the businesses have past due or not in their service accounts, it is necessary to transform the values of the dependent variable totSPDAmt3mon into binary values, i.e., 0 and 1. The frequency table of totSPDAmt3mon shows that over 70% observations have the value 0, denoting that businesses don't have any past due in the service accounts, while less than 30% observations have the value equal to or larger than 1, denoting at least 1 day passing the deadline. Therefore, a new binary dependent variable GOODBAD was defined where observations with totSPDAmt3mon valued 0 would be assigned a value of 0 for GOODBAD, denoting that they are potential "good", while those who have totSPDAmt3mon valued at least 1 would be assigned a value of 1 for GOODBAD, denoting that they are potential "bad". Table. 1 shows that distribution of GOODBAD in the merged dataset.

Table 1: Distribution of the Binary Dependent Variable GOODBAD in the Merged Dataset

GOODBAD	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21355550	71.93	21355550	71.93
1	8335767	28.07	29691317	100.00

B. *Independent Variables*

- *Dimensionality Reduction by Removing Variables with High ratio of Coded or Missing values:* Figure 1 shows the distribution of one example variable with high ratio of coded values. It can be seen that HstNFACmt3mon (denoting Highest Non-Financial Account Limit Reported in Last 3 Months) has over 80% of the coded values. Therefore, this variable couldn't provide comprehensive and reliable information in the prediction of the dependent variable. Due to the same reason, predictors with a high ratio (>70%) of coded or missing values were removed from the sampled dataset. Based on this criterion, 114 variables (96 numeric and 18 categorical) were removed.

- *Median Imputation on the Missing or Coded Values*: In order to maintain as much as information provided by the given dataset, an imputation strategy was used to replace the coded or missing values. Regression imputation works well in most cases. However, this strategy becomes exponentially more complex as the variables with missing values increases. Instead, median imputation strategy was used in this paper since most predictors have a right-skewed distribution. Figures 2 and 3 show the distributions of the example variable NoSasNFA (denoting Number of Satisfactory Non-Financial Accounts) before and after median imputation, respectively. In Figure 2, around 10% of the observations have the coded value "99" for NoSasNFA. These observations were imputed with the median value of NoSasNFA (median value was calculated based on all known valid values only) in Figure 3.

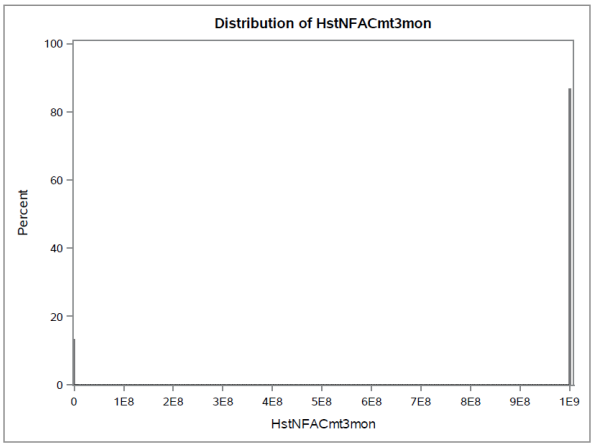

Fig. 1: Distribution of the Example Variable HstNFACmt3mon with over 80% Coded Values

- *Dimensionality Reduction by Variable Clustering*: To reduce the probability of multicollinearity and to further reduce the data dimensionality, variable clustering strategy was performed. Given the criterion that 93% of the total variation of the dependent variable within the sampled dataset was explained, the number of clusters was chosen as 70 (Figure 4). Within each cluster, the variable with smallest ratio of $1\text{-}R^2$ will be selected. Doing so yielded 70 variables and they will be used to build logistic regression as well as the decision tree.

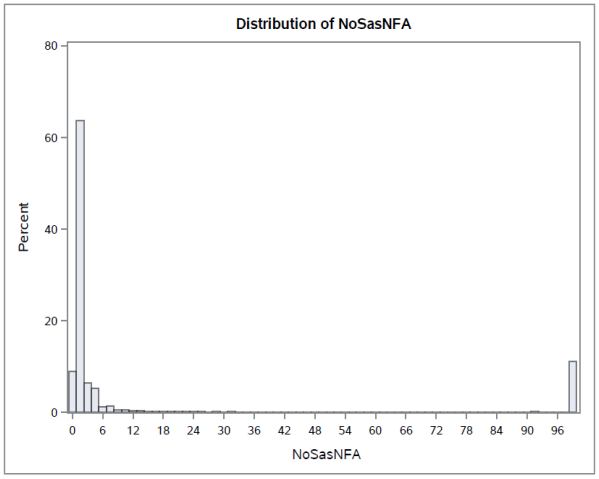

Fig. 2: Distribution of the Example Variable NoSasNFA before Median Imputation

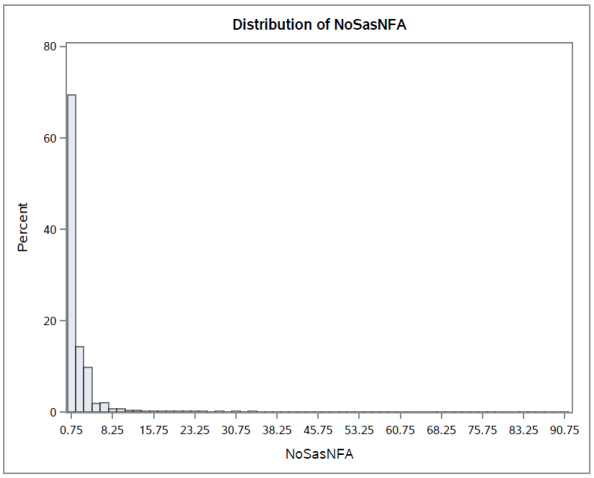

Fig. 3: Distribution of the Example Variable NoSasNFA after Median Imputation

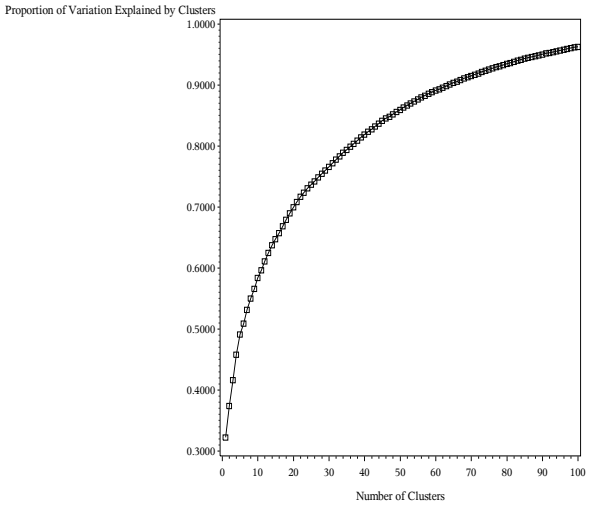

Fig. 4: Variable Clustering Analysis Curve

IV. METHODLOGY

A simple random sampling procedure was used to obtain the sampled dataset of size 100,000. Before entering the model development step, the sampled data was split into a training set (60%) and a validation set (40%). Both the logistic regression and the decision tree approaches were constructed on the training set. The performances of these two approaches were evaluated on the validation set.

A. Logistic Regression

It is reported that logistic regression is the foundational model for credit industries[9]. It is a traditional statistical technique that is used when the dependent variable is assumed to be *Bernoulli* (0-1) distributed, with probability of success (*P*) (in this paper, *P* is the probability of being past due and higher *P* value denotes higher probability of being a "bad" consumer) being modeled as some linear combination of the explanatory variables.

SAS's PROC LOGISTIC was used to conduct the logistic regression and stepwise selection method was used for variables screening at the significant level of 0.05. Based on the classification table from the training set, a cutoff value of *P* was selected for prediction purpose. When making predictions on the validation set, observations with fitted value of *P* larger than the cutoff value are predicted as "bad" while those with fitted value of *P* smaller than the cutoff value are predicted as "good".

Since stepwise selection method for variable screening has the risk of increasing type *I* error and the probability is very high that one or more errors have been made in including and excluding variables[10], the following strategy was conducted for further variable screening. For the variables that were retained after stepwise selection method, those with smallest Wald Chi-Square values (meaning least significant) would be gradually removed and the model performance on the validation set will be compared. In the case that the removal of the variable doesn't change too much on the model performance on the validation set, a parsimonious model was used by deleting this variable.

B. Decision Trees

Decision trees have been widely used in the field of classification since 1960s and are becoming more popular in machine learning area[11]. Decision trees use a top-down recursive method. In the tree structure, the leaf nodes denote classifications while the inner nodes represent the current attributes. The branches denote the conjunctions of attributions and a path from the root to the leaf node can lead to the final classifications[5].

Compared with logistic regression method, the decision tree approach doesn't require the user to possess much domain knowledge. Furthermore, decision trees are likely to perform better when the number of attributes are relatively small (<100) while the sample size is relatively large (>100,000)[12]. Due to the relatively small size of the predictors after dimensionality reduction (70) while the large size of the sampled dataset (100,000) in this paper, it is reasonable to consider the decision tree approach as the candidate classification method.

SAS Enterprise Miner 14 was used to conduct the decision tree approach[13]. Gini index, which is an impurity-based criterion that measures the divergences between the probability distributions of the target attribute's values, was used as the univariate splitting criteria at each node[14]. A value of 6 for the maximum tree depth was used for the stopping criteria. Furthermore, the maximum number of branches at each node was set as 2 in order to avoid the possible overfitting problem.

V. Model Development and Comparison

A. Logistic Regression

After performing stepwise selection procedure in SAS, 26 variables were retained in the logistic regression. The cutoff value of *P* was selected as 0.35, since this value can reach a relative high accuracy (94.7%) and low false negative (4.0%) on the training set (shown in Figure 5). This cutoff value was then used to make predictions on the validation set.

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	16826	0	43175	0	28.0	100.0	0.0	72.0	.
0.050	16629	33805	9370	197	84.1	98.8	78.3	36.0	0.6
0.100	16119	37999	5176	707	90.2	95.8	88.0	24.3	1.8
0.150	15843	39990	3185	983	93.1	94.2	92.6	16.7	2.4
0.200	15686	40748	2427	1140	94.1	93.2	94.4	13.4	2.7
0.250	15491	41196	1979	1335	94.5	92.1	95.4	11.3	3.1
0.300	15355	41584	1591	1471	94.9	91.3	96.3	9.4	3.4
0.350	15070	41766	1409	1756	94.7	89.6	96.7	8.6	4.0
0.400	14918	42002	1173	1908	94.9	88.7	97.3	7.3	4.3
0.450	14757	42132	1043	2069	94.8	87.7	97.6	6.6	4.7
0.500	14662	42237	938	2164	94.8	87.1	97.8	6.0	4.9
0.550	14548	42356	819	2278	94.8	86.5	98.1	5.3	5.1
0.600	14286	42416	759	2540	94.5	84.9	98.2	5.0	5.6
0.650	14136	42465	710	2690	94.3	84.0	98.4	4.8	6.0
0.700	13797	42509	666	3029	93.8	82.0	98.5	4.6	6.7
0.750	13565	42623	552	3261	93.6	80.6	98.7	3.9	7.1
0.800	13278	42673	502	3548	93.3	78.9	98.8	3.6	7.7
0.850	12928	42828	347	3898	92.9	76.8	99.2	2.6	8.3
0.900	12579	42967	208	4247	92.6	74.8	99.5	1.6	9.0
0.950	11808	43017	158	5018	91.4	70.2	99.6	1.3	10.4
1.000	0	43175	0	16826	72.0	0.0	100.0	.	28.0

Fig. 5: Classification Table for Logistic Regression

Table 2 shows the result of the model performances on the validation set after removing different number of the predictors according to their Wald Chi-Square values from logistic regression after stepwise selection. Decreasing the number of predictors demonstrates a minimal decrease in accuracy. The false positive rate has a slight increase when more predictors were removed. When the number of predictors was 14, the model reached a lowest false negative value (2.38%). Considering that model with 14 predictors reaches a similar accuracy (94.58%) as the model with all the 26 predictors selected by stepwise selection procedure (94.93%), the logistic regression with 14 predictors was selected as the preferred parsimonious model. The coefficient estimations as well as the *p* value for these 14 variables are shown in Figure 6. The variable with larger Wald Chi-Square value is considered to be more important in making predictions.

Table 2: Performances of Logistic Regression on Validation Set using Different Number of Predictors

Number of Predictors	Accuracy	False Positive	False Negative
26	94.93%	2.70%	2.37%
20	94.86%	2.72%	2.42%
14	94.58%	3.05%	2.38%
10	93.99%	3.53%	2.48%

Fig. 7 displays the ROC result for the logistic regression based on the validation set. The Area Under the Curve (AUC) is 0.9821, meaning that the selected model performs well in classifying the past due on service accounts.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.7293	0.0734	555.0761	<.0001
totTA1CPDC3mon	1	6.5398	0.1220	2874.8413	<.0001
NFA3monCurRate	1	-3.2615	0.0628	2698.9426	<.0001
NoNFChgAcc	1	-3.5724	0.1129	1000.3727	<.0001
pctNFChgAccAcc24mon	1	-0.0362	0.00130	774.5205	<.0001
totTA1CPDCCrly	1	3.8600	0.1375	787.9107	<.0001
totNFA1CPDC12mon	1	0.8700	0.0231	1417.2977	<.0001
totTA2CPDC12mon	1	-1.4329	0.0538	709.5527	<.0001
pctSasNFA12mon	1	-0.0145	0.000613	558.2815	<.0001
NoIAcbalance3mon	1	-0.8098	0.0249	1055.8020	<.0001
WstNFpay24mon	1	0.2663	0.0111	579.7699	<.0001
totTA1CPDcrly	1	-5.3058	0.2910	332.4156	<.0001
pctNFPDAmtst3mon	1	-0.00274	0.000077	1251.7903	<.0001
WstIpay24mon	1	-0.4662	0.0190	599.4132	<.0001
WstTpay	1	0.2239	0.0148	230.2803	<.0001

Fig. 6: Parameter Estimations for the Logistic Regression

B. Decision Trees

Figure 8 shows the confusion matrix based on the validation set for the decision tree from SAS Enterprise Miner. It can be calculated that decision tree approach can reach the accuracy of the value 96.69%, which is higher than the result from logistic regression. False positive (2.92%) is slightly lower than the result from logistic regression (3.05%) while

false negative is much lower (0.39%) than that in logistic regression (2.38%).

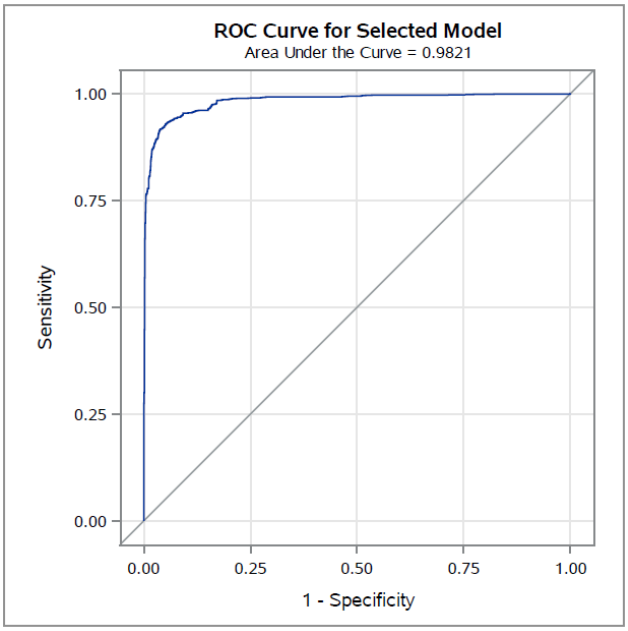

Fig. 7: ROC Curve for Logistic Regression based on the Validation Set

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	99.4383	95.9387	27615	69.0340
1	0	0.5617	1.3906	156	0.3900
0	1	9.5577	4.0613	1169	2.9224
1	1	90.4423	98.6094	11062	27.6536

Fig. 8: Confusion Matrix for the Decision Tree

Figure 9 displays the variable importance for the decision tree. While logistic regression has 14 predictors, the decision tree only uses 10 predictors. When comparing Figures 6 and 9, it can be found that 5 variables (NFA3monCurRate, pctNFPDAmtst3mon, NoIAcbalance3mon, WstIpay24mon, and WstTpay) were selected by both logistic regression and the decision tree.

Figure 10 shows the ROC result for the decision tree based on the validation set. The Area Under the Curve (AUC) is 0.9960, which is another evidence that the decision tree is a good strategy for the classification problem in this paper.

VI. Conclusion and Discussion

The goal in this paper is to predict whether small businesses would have past due or not in their service account, therefore, accuracy is not the only concern when evaluating the model performance. The most conservative model for the company is the one that can reach the lowest false negative since this can prevent the company from lending a lot of money to the businesses who are actually "bad" but predicted to be "good".

Variable Name	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
WstTpay3mon	1	1.0000	1.0000	1.0000
NFA3monCurRate	1	0.5234	0.5062	0.9671
pctNFPDAmtst3mon	4	0.3479	0.3402	0.9780
WstTpay	2	0.2736	0.2740	1.0013
totNFPDAmt12mon	3	0.2637	0.2554	0.9683
NoIAcbalance3mon	1	0.2626	0.2733	1.0408
totonSA3mon	3	0.2346	0.2249	0.9587
totNFPD	2	0.1042	0.0979	0.9392
WstIpay24mon	1	0.0819	0.0855	1.0437
NoNFChgAcc24mon	1	0.0309	0.0358	1.1602

Fig. 9: Result of Variable Importance for the Decision Tree

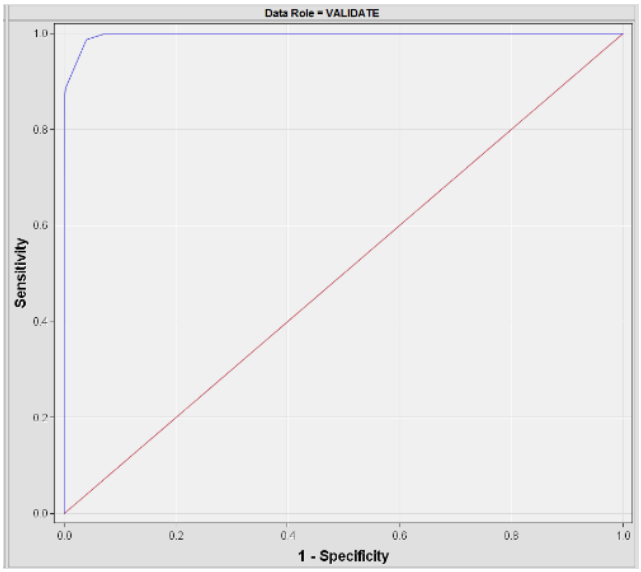

Fig. 10: ROC Curve for Decision Tree based on the Validation Set

As expected, the decision tree approach has a very good performance on the given data since it is a common machine learning strategy in analyzing big and complex dataset[15]. It can reach a very high accuracy (96.69%) on the validation set. False positive and false negative are both lower in decision tree than in logistic regression. Moreover, the decision tree use 4 less variables than logistic regression (shown in Figures 6 and 9). For the above reasons, the decision tree is more preferred than logistic regression for the given dataset.

The decision tree is a non-parametric machine learning approach while logistic regression is a traditional parametric strategy. It is difficult to decide which approach is "better"

when these two methods are applied to future dataset. The decision of method selection depends on the biggest concern of the research. Under the condition that both logistic regression and decision tree can achieve similar and high accuracy, more attention should be put on false negative or false positive when comparing the two approaches. Moreover, other machine learning techniques such as support vector machine and random forest are also good candidate methods for binary classification problems.

REFERENCES

[1] H. Kim and S. Gu. A Logistic Regression Analysis for Predicting Bankruptcy in the Hospitality Industry. *Journal of Hospitality Financial Management*. 12(1), 2010.

[2] T. Verbraken and C. Bravo. Development and Application of Consumer Credit Scoring Models using Profit-based Classification Measures. *European Journal of Operational Research*: 505-513, 2014.

[3] J.R. Quinlan. Induction of Decision Trees. *Machine Learning*, 1:81-106, 1986.

[4] J. Crook and D. Edelman. Recent Developments in Consumer Credit Risk Assessment. *European Journal of Operational Research*, 183(3):1447-1465, 2007.

[5] W. Henley. Statistical Aspects of Credit Scoring. *Dissertation Milton Keynes*, UK:Springer.

[6] J. Cruz and D. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Informatics*, 2:59-77,2006.

[7] D. Zhang and X. Zhou. Vertical Bagging Decision Trees Model for Credit Scoring. *Expert Systems with Applications*, 37(12), 2010.

[8] C. Chuang and R. Lin. Constructing a reassigning credit scoring model, Part 1. Expert Systems with Applications, 36(2): 1685-1694, 2009.

[9] S. Nargundkar and J. Priestley. Assessment of Model Development Techniques and Evaluation Methods for Binary Classification in the Credit Industry. *DSI Conference*, 2013.

[10] W. Mendenhall and T. Sincich. A Second Course in Statistics: Regression Analysis. Seventh Edition: 326-338.

[11] Y. Jiang. Credit Scoring Model Based on the Decision Tree and the Simulated Annealing Algorithm. *World Congress on Computer Science and Information Engineering*, 2009.

[12] X. Niuniu and L. Yuxun. Review of Decision Trees. *IEEE*: 105-109, 2010.

[13] M. Abdullah and A. Ghoson. Decision Tree Induction & Clustering Techniques in SAS Enterprise Miner, SPSS Clementine, and IBM Intelligent Miner – A Comparative Analysis. International Journal of Management & Information Systems, 14(3):57-70, 2010.

[14] L. Rokach and O.Maimon. Data Mining with Decision Trees: Theory and Applications. Second Edition: 165-170.

[15] L. Hall and N. Chawla. Decision Tree Learning on Very Large Data Sets. IEEE: 2579-2584, 1998.