2017

A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit

Jessica M. Rudd MPH, GStat
Kennesaw State University

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

Recommended Citation

Rudd, Jessica M. MPH, GStat and Priestley, Jennifer L., "A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit" (2017). *Grey Literature from PhD Candidates*. 5.
http://digitalcommons.kennesaw.edu/dataphdgreylit/5

A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit

Jessica M. Rudd, MPH, GStat
Department of Statistics and Analytical Sciences
College of Science and Mathematics
Kennesaw State University
jrudd1@students.kennesaw.edu

Jennifer Lewis Priestley, PhD
Department of Statistics and Analytical Sciences
College of Science and Mathematics
Kennesaw State University
jpriestl@kennesaw.edu

***Abstract*—Credit risk prediction is an important problem in the financial services domain. While machine learning techniques such as Support Vector Machines and Neural Networks have been used for improved predictive modeling, the outcomes of such models are not readily explainable and, therefore, difficult to apply within financial regulations. In contrast, Decision Trees are easy to explain, and provide an easy to interpret visualization of model decisions. The aim of this paper is to predict worst non-financial payment status among businesses, and evaluate decision tree model performance against traditional Logistic Regression model for this task. The dataset for analysis is provided by Equifax and includes over 300 potential predictors from more than 11 million unique businesses. After a data discovery phase, including imputation, cleaning, and transforming potential predictors, Decision Tree and Logistic Regression models were built on the same finalized analysis dataset. Evaluating the models based on ROC index, and Kolmogorov-Smirnov statistic, Decision Tree performed as well as the Logistic Regression model.**

Keywords—Logistic Regression; Decision Tree; Credit Risk; Commercial Credit

I. Introduction

Credit risk analysis is an important aspect of the financial services domain. Logistic Regression has traditionally been used to model credit risk because it models a binary outcome (e.g. default/no default), the outcome is between 0 and 1 and is readily interpreted as probability of default, and the variable coefficients can be interpreted separately to assess importance of each variable in the credit decision[1]. The latter aspect of Logistic Regression is important for applying credit risk models within financial regulations, i.e. reason codes.

Considering credit information and the financial market is constantly changing, building predictive models is time consuming and computationally expensive. When building predictive models, we must contend with extremely large datasets and high dimensionality of the data, as well as unknown relationships between various data characteristics. Machine Learning techniques, such as Neural Networks, have been proposed for their advanced computational speed and applicability to large, high dimensional data with unknown characteristics[2]. However, the outcomes of Neural Networks are not easily explainable and, therefore, they are difficult to apply in a heavily regulated industry such as consumer and commercial credit. In contrast, Decision Tree Models are a machine learning technique that are readily explainable, present a visual representation of the model choices, and require minimal knowledge of underlying data relationships[3].

In this paper, we build Logistic Regression and Decision Tree models to predict commercial credit risk by way of worst non-financial payment status in Equifax provided datasets. After a review of previous research, we describe the process of imputing, transforming, and selecting model variables from a large but sparse dataset. We then describe the Logistic Regression and Decision Tree methodologies. Finally, we discuss results and compare the models using ROC index, and Kolmogorov-Smirnov statistic.

II. Literature Review

Credit decisioning is an important financial problem that requires substantial amounts of decisioning variables in a constantly changing market. Khandani, et al.[2] point out that many institutions build their own internal models for credit decisioning, and these models only change slowly over time, whereas market conditions affecting credit change much more rapidly. They advocate for various machine learning techniques, including Support Vector Machines and Decision Trees, to build credit risk models, because these methods can tackle computationally intensive analyses with large, complex datasets with improved speed.

Decision Trees have been used for disease classification problems as well. W.J. Long, et al.[4] presents a comparison of Decision Tree and Logistic Regression for classifying patients with heart disease. They point out that Decision Trees are adjustable for “noisy” data, including missing values. In contrast, Logistic Regression cannot include missing data. Satchidananda, et al.[5] apply the comparison of Decision Tree and Logistic Regression to credit risk data in India. They found that Decision Tree produced more precise and parsimonious models than Logistic Regression.

III. DATA DISCOVERY

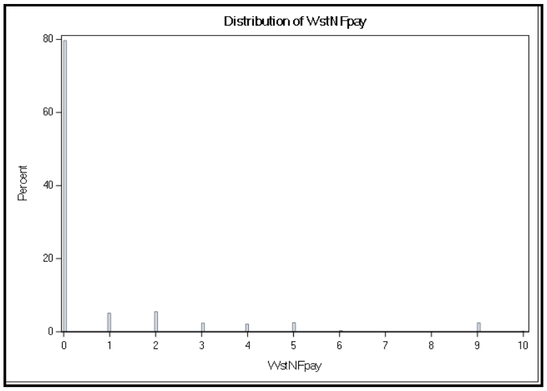
The data for this analysis was provided by Equifax and included 36 quarterly datasets between 2006 and 2014. Each individual dataset was comprised of over 300 potential predictors, including business demographic information, and account activities for over 11 million unique businesses.

A. *Assignment of Dependent Variable*

In this project, we examined the prediction of worst non-financial payment status (WSTNFPay), assumed to be a proxy here for business credit risk. Since this variable captures the worst payment status to date, we used the last quarterly dataset for the analysis, capturing the worst business activity over the study period. The worst non-financial payment status for each business was chosen as a conservative approach for assigning credit default risk. Businesses who have experienced higher number of consecutive delinquent months are more likely to fall behind on payments and default. Even though some of these businesses may pay back their debt and not default (potentially good businesses that would be rejected for credit), it is costlier to extend credit to a potential business who will default. In this case, using worst non-financial payment status for each business maximizes the financial output while minimizes the likelihood of default. In addition, using the last quarterly dataset allowed us to maintain a large sample size for analysis but also consider computational efficiency.

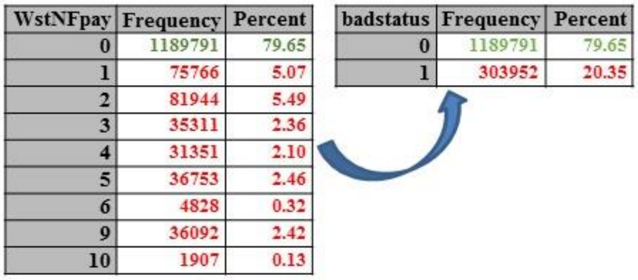
Another factor we considered when choosing a dependent variable is that most variables in the dataset have over 50% missing or coded values. We cannot impute the values of our dependent variable, so choosing a variable which represents the proxy for default with as many valid values as possible was important for maintaining a large valid dataset. WSTNFPay had the most valid values of the potential dependent variables. Once filtering the dataset on valid values of WSTNFPay, our dataset included 305 variables and 1,493,743 observations. Fig. 1 shows the distribution of valid values for WSTNFPay.

FIGURE 1. DISTRIBUTION OF WSTNFPAY



For the purpose of building a binary Logistic Regression, it was necessary to transform the dependent variable into a binary outcome, where 0 denotes the business had no delinquency and 1 denotes any delinquency during the study period. Fig. 2 shows the frequency distribution of WSTNFPay and illustrates the appropriateness of assigning “good” status to businesses with no delinquency and “bad” status to businesses with any delinquency.

FIGURE 2. FREQUENCY DISTRIBUTION AND TRANSFORMATION OF DEPENDENT VARIABLE



WstNFpay	Frequency	Percent
0	1189791	79.65
1	75766	5.07
2	81944	5.49
3	35311	2.36
4	31351	2.10
5	36753	2.46
6	4828	0.32
9	36092	2.42
10	1907	0.13

badstatus	Frequency	Percent
0	1189791	79.65
1	303952	20.35

B. *Imputation*

Considering the high variability and sparseness of the data, variables with greater than 30% missing or coded values were removed from analysis using SAS® macro code. Missing and coded values for the remaining variables were imputed using the median of the valid values, since the median is more stable in highly skewed data. To avoid skewing the coefficients in the Logistic Regression model, outlier values, greater than 4 standard deviations (SD) from the mean, were imputed to the value equal to the 4th SD cut-point. Post-imputation, 102 predictors remained. Fig. 3a and Fig. 3b show an example of predictor variable pre-and post-imputation, respectively.

FIGURE 3A. PRE-IMPUTATION VARIABLE DISTRIBUTION

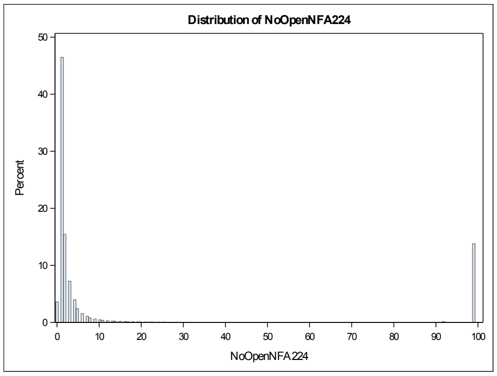
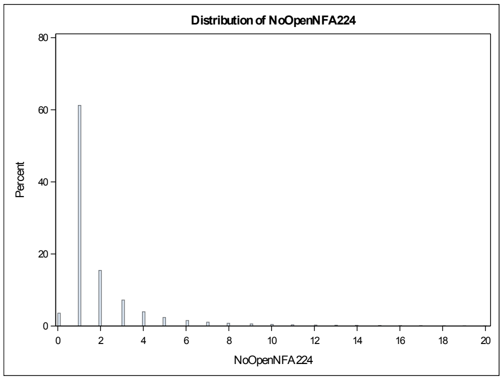


FIGURE 3B. POST-IMPUTATION VARIABLE DISTRIBUTION



C. *Variable Selection via Clustering*

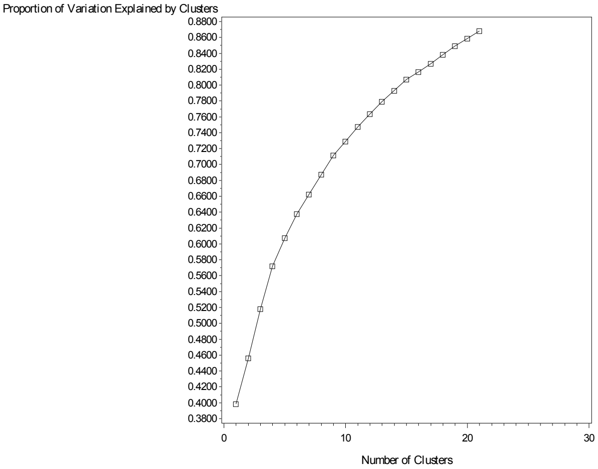
When building a model with many variables it's difficult to establish the "correct" relationships between the independent and dependent variables due to redundancy. High dimensionality increases the risk of overfitting due to correlations between redundant variables, increases computation time, increases the cost and practicality of data collection for the model, and makes interpretation of the model difficult. Variable clustering was used to reduce variables considered for the model by eliminating redundancy.

The VARCLUS procedure in SAS® was used to find groups of variables that are as correlated as possible among themselves and as uncorrelated as possible with variables in other groups. All variables start in one cluster, an algorithm similar to principal component analysis is performed, and the cluster is split if the second eigenvalue is greater than the specified threshold. This process is repeated until the second eigenvalue falls below the threshold. A threshold value of 0.7 is chosen as this is the accepted industry standard and accounts for sampling variability (as opposed to using the average eigenvalue of 1). Variable reduction was achieved by choosing the representative variable in each cluster that has high correlation with its own cluster and low correlation with other clusters; the lowest 1-R^2 ratio value in each cluster (1).

$$\text{1-R**2 ratio} = \frac{1\text{-}R^2 \text{ own cluster}}{1\text{-}R^2 \text{ next closest}} = \frac{1 - \uparrow}{1 - \downarrow} \Rightarrow \frac{\downarrow}{\uparrow} \Rightarrow \downarrow \qquad (1)$$

Fig. 4 shows that 21 clusters explained approximately 87% of the variability in the data, so we chose the best 21 variables to consider as predictors in our model.

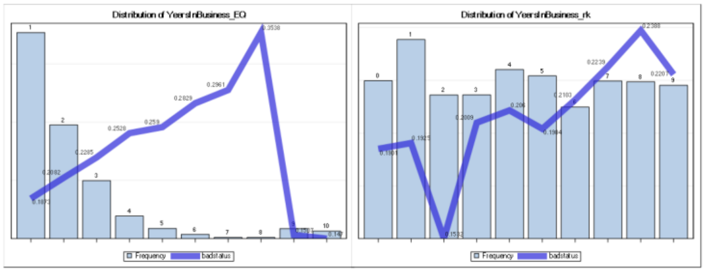
FIGURE 4. PROPORTION OF VARIATION EXPLAINED BY CLUSTERS



D. *Discretization and Transformation*

To account for various relationships between the predictors and the dependent variable, we conducted several discretization and transformation processes. The 21 potential predictor variables were discretized using 1) user-defined equal width based on distribution, and 2) SAS® PROC RANK for equal frequencies. The new discrete variables were then transformed using odds and log odds. The analysis dataset included 147 potential predictors. Fig. 5 shows an example of variable discretization.

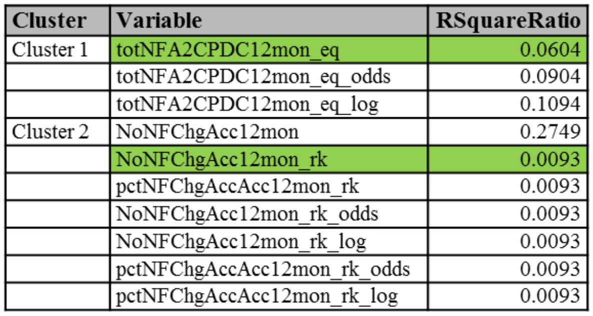
FIGURE 5. USER-DEFINED AND SAS DEFINED DISCRETIZATION



IV. METHODOLOGY

SAS® PROC VARCLUS was utilized post-discretization to reduce multicollinearity and eliminate redundancy. 38 clusters explained 94% of variation in dataset, and the variable with the lowest 1-R^2 ratio was selected from each cluster, Fig. 6. The final analysis dataset included 38 potential predictor variables and approx. 1.5 million records. For improved processing time, a 20% simple random sample was used. This sample was then split into 60% training data for building the models, and 40% validation.

FIGURE 6. SELECTION OF VARIABLES FROM CLUSTERING



Cluster	Variable	RSquareRatio
Cluster 1	totNFA2CPDC12mon_eq	0.0604
	totNFA2CPDC12mon_eq_odds	0.0904
	totNFA2CPDC12mon_eq_log	0.1094
Cluster 2	NoNFChgAcc12mon	0.2749
	NoNFChgAcc12mon_rk	0.0093
	pctNFChgAccAcc12mon_rk	0.0093
	NoNFChgAcc12mon_rk_odds	0.0093
	NoNFChgAcc12mon_rk_log	0.0093
	pctNFChgAccAcc12mon_rk_odds	0.0093
	pctNFChgAccAcc12mon_rk_log	0.0093

A. *Decision Tree*

A Decision Tree is a classification technique that assigns each object in a dataset (in this case, each business) into a predicted class (e.g. good/bad risk of default) based on each objects' attributes. The algorithm uses Information Gain (2) to find the best attribute for classifying the data, where *P* and *n* are the 0 and 1 values of a binary outcomes for the *i-th* object. Then, for each value defined for the decision values of the best attribute, the algorithm repeats the process with additional attributes[4].

$$G(A) = I(p,n) - \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

where

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \qquad (2)$$

We used SAS® Enterprise Miner™ to build and prune the Decision Tree with a maximum node depth of 5 and minimum of 30 observations per leaf. 10-fold cross validation was used for model evaluation. Although we did extensive pre-processing of our data to fit the Logistic model, it should be noted that Decision Tree models are less sensitive to outliers and missing data, and they do not require data to be transformed or normalized[6].

B. Logistic Regression

Logistic Regression examines the non-linear relationship between a binary outcome and categorical or continuous predictor variables. The logistic model outputs a probability of an event between 0 and 1 as the log of the odds ratio (3)

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{3}$$

where β is the parameter coefficient and *x* is the value of the independent variable.

We used SAS® PROC LOGISTIC to build the model from the 38 predictor variables. Backwards elimination with α = 0.05 was used to eliminate redundancy and keep the strongest predictors in the model. 10-fold cross validation was used for model evaluation.

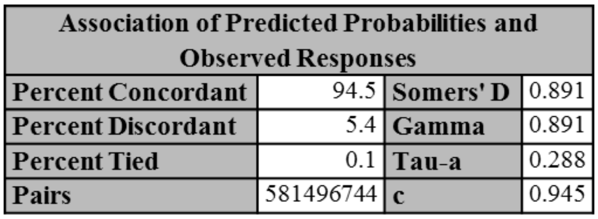
V. RESULTS

Using Logistic Regression, the following 11 variables remained as significant in the model, and pctSasNFA12mon_eq_log most strongly predictive of default, with odds ratio 2.35:

- **HstNFB12mon_rk_log**
- **NoClosedNFA226_eq_log**
- **NoNewNFAcc3mon26**
- **NoNFA3mon_rk_log**
- **NoOpenNFA224_rk_log**
- **pctNFChgAccAcc12mon_eq**
- **pctSasNFA12mon_eq_log**
- **totC1NFPDAmt12mon_rk**
- **totLAllLiens_rk**
- **totNFA4CPDC12mon_rk_log**
- **YearsinBusiness_rk_log**

Fig. 7 shows a strong Logistic Model performance, with C-statistic of 0.95, indicating high concordance of predicted with actual default. In addition, the Fig. 8 confusion matrix from validation data shows misclassification of approximately 9% at a cutpoint of 0.1.

FIGURE 7. LOGISTIC REGRESSION MODEL PERFORMANCE



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	94.5	Somers' D	0.891
Percent Discordant	5.4	Gamma	0.891
Percent Tied	0.1	Tau-a	0.288
Pairs	581496744	c	0.945

FIGURE 8. LOGISITIC REGRESSION CONFUSION MATRIX

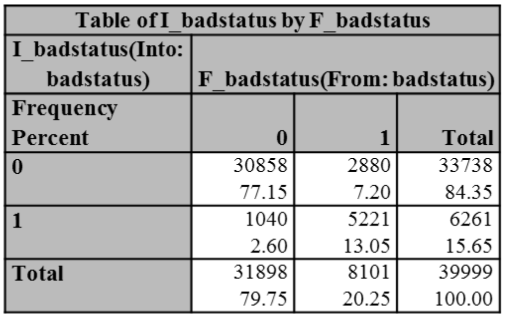
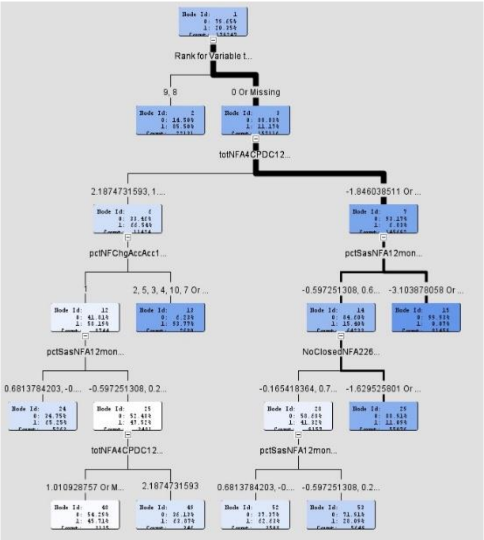


Table of I_badstatus by F_badstatus			
I_badstatus(Into: badstatus)	F_badstatus(From: badstatus)		
Frequency Percent	0	1	Total
0	30858 77.15	2880 7.20	33738 84.35
1	1040 2.60	5221 13.05	6261 15.65
Total	31898 79.75	8101 20.25	39999 100.00

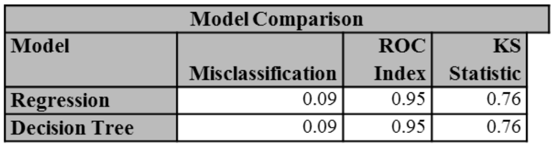
The Decision Tree model, Fig. 9, produced the same overall performance as Logistic Regression, however totC1NFPDAmt12mon_rk was found to be the most important predictor based on the ratio of Information Gain in the training versus validation datasets.

FIGURE 9. DECISION TREE



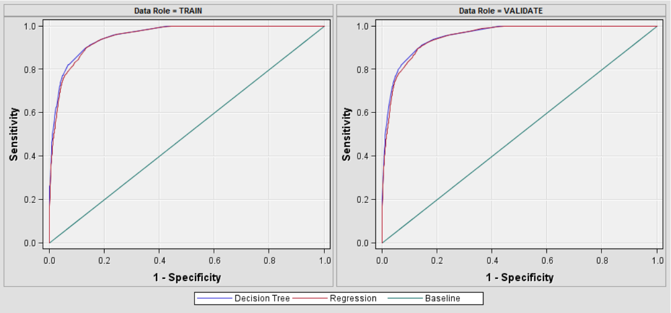
From Fig. 10 and Fig. 11 we can see that misclassification, ROC Index area under the curve (AUC), and KS-Statistic are effectively the same for both models. For credit risk, the KS-statistic measures how well the model distinguishes between good and bad credit risk businesses. For both models, the KS-statistic is maximized within the 2nd decile of our dataset.

FIGURE 10. MODEL PERFORMANCE COMPARISON



Model Comparison			
Model	Misclassification	ROC Index	KS Statistic
Regression	0.09	0.95	0.76
Decision Tree	0.09	0.95	0.76

FIGURE 11. COMPARISON OF MODEL AUC



VI. DISCUSSION

Considering the Decision Tree model performs nearly as well as the Logistic Regression, it presents a useful alternative for credit risk analytics. Decision Trees are advantageous for predictive modeling due to:

- Implicit variable screening and selection – the top nodes of the tree are the most important variables in the dataset!
- Less data prep – data does not need to be normalized, and decision trees are less sensitive to missing data and outliers
- Decision trees do not require assumptions of linearity
- Decision tree output is graphical and easy to explain – decision based on cut points

While the Logistic model is considered the gold standard for credit risk prediction, we advocate for implementation of the Decision Tree where possible due to the simplicity of data preparation and interpretation required in comparison to the Logistic procedures. The Decision Tree itself can be applied to a web application, for example, where an employee can input financial values for a business and output credit risk within minutes, without requiring extensive training. Decision Trees themselves are visually comparable to human decision making and can be readily applied to industry regulation for credit reasoning. Future research applying Decision Trees to consumer credit risk portfolios will be valuable for predictive modeling of additional consumer segments with notoriously sparse data, such as new immigrants and other customers without pre-established credit.

ACKNOWLEDGMENT

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

REFERENCES

[1] Altman, E.I., G. Sabato, “Modelling Credit Risk for SMEs: Evidence from the U.S. Market,” Abacus, vol. 43-3, Blackwell Publishing Asia, 2007.

[2] Khandani, A.E., A. J. Kim, and A. W. Lo, “Consumer credit-risk models via machine-learning algorithms,” Journal of Banking & Finance, vol. 34, 2010, pp. 2767-2787.

[3] How Decision Tree Algorithm Works [Web log post]. Retrieved April 15, 2017 from https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

[4] Long, W.J., J.L Griffith, H.P. Selker, and R.B D’Agostino, “A Comparison of Logistic Regression to Decision Tree Induction in a Medical Domain,” Computers in Biomedical Research, vol 26, 1993, pp. 74-97.

[5] Satchidananda, S.S., and J. B. Simha, “Comparing decision trees with logistic regression for credit risk analysis,” SAS APAUGC MUMBAI, 2006.

[6] Ryza, S., U. Laserson, S. Owen, and J. Wills (2015). *Advanced Analytics with Spark.* Sebastopol, CA: O’Reilly Media.