Daniel Sanchez

BAN 525

Dr. Cetin Ciner

Module 7 Final Project

June 27th, 2021

Introduction

In the 1992 Disney adaptation of Aladdin, the hero (precariously balancing atop his magic carpet) offers his hand to the princess: "*Do you trust me?*" After some hesitation, the princess accepts his hand, and is whisked away on an exhilarating flight through a whole new world: shining, shimmering, splendid. This experience helps build a strong foundation to their long-lasting relationship.

The credit and lending market relies similarly on this concept of trust. Credit-worthy individuals are often showered with invitations to apply for new accounts. Meanwhile, those with low credit scores can be saddled with high interest rates and more stringent repayment terms, if not being denied outright. It is understandable for a lender to protect itself from delinquent borrowers, especially in cases where the stakes are high.

Home equity loans, for example, carry the additional burden of being tied to a property. If a borrower falls into delinquency, they could lose their home. However, the lender would also be forced to sell or transfer the property to new buyers. This is a lose-lose situation, best to be avoided. The challenge is knowing which borrowers are most deserving of a long-term financial relationship: "*Do you trust me?*"

Analyzing borrower data can provide some clues as to who might pose more or less of a credit risk. The data set in this analysis consists of 13 variables taken from 5,960 individuals. Not all of the variables contain complete information from each respondent. The missing value counts for each column are displayed in the table on the right. This will be taken into consideration during our predictive modeling. The variable of interest is BAD, a binary feature referring to the credit risk of each individual ("Good Risk" or "Bad Risk"). We are hoping to determine and better understand the predictors of Bad Risk.

Column	Number Missing
MORTDUE	518
VALUE	112
REASON	252
JOB	279
YOJ	515
DEROG	708
DELINQ	580
CLAGE	308
NINQ	510
CLNO	222
DEBTINC	1267

The analysis in this report has been performed in JMP Pro 16 using a wide array of methods, listed below:

Nominal Logistic Regression
Stepwise Forward Regression
Stepwise Backward Regression
Adaptive Lasso Regression
Adaptive Elastic Net Regression
Standard Decision Tree (with and without Informative Missing)
Random Forest (with and without Informative Missing)
Boosted Neural Nets, with various configurations
- One Layer, Three Nodes (3-3-3), 40 Models, 20 Tours, Squared Penalty, Informative Missing
- One Layer, Three Nodes (3-3-3), 40 Models, 20 Tours, Squared Penalty, No Informative Missing
- One Layer, One Node (3-0-0) 40 Models, 20 Tours, Squared Penalty, Informative Missing
- One Layer, One Node (3-0-0) 40 Models, 20 Tours, Squared Penalty, No Informative Missing
- One Layer, One Node (3-0-0) 40 Models, 20 Tours, Absolute Penalty, Informative Missing
- One Layer, One Node (1-0-0) 40 Models, 20 Tours, Squared Penalty, Informative Missing
- One Layer, One Node (1-0-0) 40 Models, 20 Tours, Squared Penalty, No Informative Missing
- One Layer, One Node (1-0-0) 40 Models, 20 Tours, Absolute Penalty, Informative Missing
- One Layer, One Node (1-0-0) 40 Models, 20 Tours, Absolute Penalty, No Informative Missing
Boosted Tree (with and without Informative Missing)

Analysis and Model Comparison

Ordinary Logistic Regression: This is often used as a benchmark for analysis. It is not expected to outperform the other models in the group. By retaining all of the data set variables in its results, this method tends to overfit models to the training set.

Stepwise Forward and Backward: These alternative approaches are helpful in determining dependent variables. They each take incremental approaches to improving the final model. One concern with these methods is multicollinearity, since individually removing/adding variables might impact the inclusion/exclusion of other variables further down the chain.

Adaptive Lasso: Using the absolute value of a determined penalty, this method shrinks the variables that contain little information, so that those are not included in the final model. This method also tends to eliminate redundant variables that are highly correlated with others. Being adaptive, it takes into consideration the results of Ordinary Logistic Regression before applying its penalization factor to unimportant variables.

Adaptive Elastic Net: This also considers the results of Ordinary Logistic Regression before applying its penalized regression approach. However, this method applies both an absolute (Lasso Regression) and squared (Ridge Regression) penalization value to uninformative

variables. While unimportant variables are eliminated in the resulting model, highly correlated variables are normally retained.

Standard Decision Tree (with and without Informative Missing): This method creates a tree by separating the variables at each node, according to the largest disparity. This greedy approach does not always provide an optimal solution. This method is simply included in the analysis to be compared with the Random Forest model.

Random Forest (with and without Informative Missing): This method creates many decision trees, using a subset of variables to enforce an element of randomness. The resulting array of uncorrelated decision trees make predictions, and the one that makes the most accurate predictions is the model that is chosen.

Boosted Neural Nets (with and without Informative Missing): This method is helpful for modeling very complex non-linear relationships. Designed to simulate how the human brain processes information, this method sends inputs through nodes, where a transformation function is applied. The assorted configurations differ in complexity, varying the number of nodes, selected activation function(s), type of error penalty, and their use of informative missing data.

Boosted Tree (with and without Informative Missing): This method applies a similar boosting approach to a decision tree framework. This incremental approach collects the residuals, or errors, as each model is estimated. Those errors are then factored into the estimation of subsequent model estimates, after which an average of all models is used.

Since the home equity loan data set in this analysis is cross-sectional in nature (as opposed to being time series), the cross-validation approach for this analysis incorporates randomness. The data set has been split into three parts. The *Make Validation Column* facility in JMP Pro 16 was selected, applying a random seed of 123. Implementing a 60-20-20 split of the 5960 total rows, 3576 rows were used to train the models, 1192 rows were tied to validation, and 1192 rows were set aside for testing.

The “Bad Risk” factor-level of the BAD feature was selected in JMP Pro 16 as the Y variable. All of the other variables in the data set were included as candidate predictors. The Validation Column was implemented, and the analysis was conducted using all methods. The random seed of 123 was used where applicable (e.g., configuring Neural Net models). As

predictions from each model were generated, they were saved and stored in new columns in the data set. The figure below shows the comparison of how all models performed with the test set.

Model Comparison
Measures of Fit for BAD

Creator	Misclassification Rate	N	AUC
Ordinary Logistic Regression	0.0817	673	0.7861
Stepwise Forward	0.0833	708	0.7715
Stepwise Backward	0.0832	673	0.7843
Adaptive Lasso	0.0817	673	0.7862
Adaptive Elastic Net	0.0817	673	0.7862
Standard Decision Tree, Informative Missing	0.1258	1192	0.8672
Standard Decision Tree, No Informative Missing	0.1669	1192	0.7698
Random Forest, Informative Missing	0.1065	1192	0.9355
Random Forest, No Informative Missing	0.1124	1192	0.9075
Neural Net: 3-3-3, 40(40)/20, SQ, Informative Missing	***0.0780***	***1192***	***0.9502***
Neural Net: 3-3-3, 40/20, SQ, No Inf Missing	0.0490	673	0.8967
Neural Net: 3-0-0, 40(37)/20, SQ, Informative Missing	***0.0847***	***1192***	***0.9454***
Neural Net: 3-0-0, 40(30)/20, SQ, No Informative Missing	0.0550	673	0.8790
Neural Net: 3-0-0, 40(19)/20, ABS, Informative Missing	0.0956	1192	0.9241
Neural Net: 1-0-0, 40(26)/20, SQ, Informative Missing	0.1116	1192	0.9021
Neural Net: 1-0-0, 40(27)/20, SQ, No Informative Missing	0.0594	673	0.8373
Neural Net: 1-0-0, 40(27)/20, ABS, Informative Missing	0.1166	1192	0.8997
Neural Net: 1-0-0, 40(16)/20, ABS, No Informative Missing	0.0728	673	0.8323
Boosted Tree, Informative Missing	***0.0805***	***1192***	***0.9474***
Boosted Tree, No Informative Missing	0.1435	1192	0.8607

The ideal models in this comparison should have the highest AUC (Area Under the ROC Curve) and lowest Misclassification Rate. The AUC metric was given priority consideration, since none of the models held both the highest AUC *and* lowest Misclassification Rate. Of the top three models in this comparison, two were Boosted Neural Nets (3 Nodes with 3 Activation Functions, and 3 Nodes with 1 Hyperbolic Tangent Activation Function). All of these winning models included missing information from the data set to inform its predictions. This implies that missing data is a contributing factor in determining the credit risk of individuals. Those who leave certain fields blank may have a dubious reason for doing so.

It is interesting to find that a less complex Neural Net model is able to compete with a more complex one. The Complex Neural Net had an AUC of 0.9502 and a Misclassification Rate of 0.0780, not far off from the Simple Neural Net's AUC of 0.9474 and Misclassification Rate of 0.0847. Another model that performed in the top three was a Boosted Tree, with its AUC of

0.9474 and Misclassification Rate of 0.0805. These three models will be reviewed in further detail for the rest of this analysis.

Interpretation

Reviewing the parameters of both Neural Net models, 40 models were selected with 20 tours. The more complex model ran through all 40 models, while the simpler model stopped at 37. The Boosted Tree model contained 189 layers, with 15 splits per tree.

Variable importance analysis reveals some divergence in the predicted results across the three models. The Complex Neural Net model and Boosted Tree both give the DELINQ (Number of Delinquent Trade Lines) variable the highest total effect (77.6% and 52%, respectively). Meanwhile, the Simple Neural Net model lists the total effects of DEBTINC (Debt to Income Ratio, 62.6%) and CLAGE (Age of Oldest Trade Line, 43.5%) ahead of DELINQ (31.1%).

One trait all of these models share is a large disparity between the main effects and total effects for the important variables. This indicates that there is a substantial amount of interaction between the variable effects, and that not just one feature is important. This may be one reason for the different ordering structures resulting from all three models, shown below.

Variable Importance: Independent Uniform Inputs

Neural Net: 3-3-3, 40(40)/20 Informative Missing

Column	Main Effect	Total Effect
DELINQ	0.056	0.776
DEBTINC	0.045	0.770
DEROG	0.011	0.299
VALUE	0.009	0.263
JOB	0.009	0.227
INQ	0.007	0.218
CLAGE	0.007	0.171
LOAN	0.004	0.125
MORTDUE	0.003	0.105
YOJ	0.004	0.084
REASON	0.004	0.083
CLNO	0.003	0.068

Neural Net: 3-0-0, 40(37)/20 Informative Missing

Column	Main Effect	Total Effect
DEBTINC	0.269	0.626
CLAGE	0.104	0.435
DELINQ	0.081	0.311
DEROG	0.031	0.176
JOB	0.026	0.159
VALUE	0.023	0.149
NINQ	0.021	0.136
MORTDUE	0.021	0.132
CLNO	0.020	0.082
REASON	0.018	0.078
LOAN	0.017	0.068
YOJ	0.018	0.067

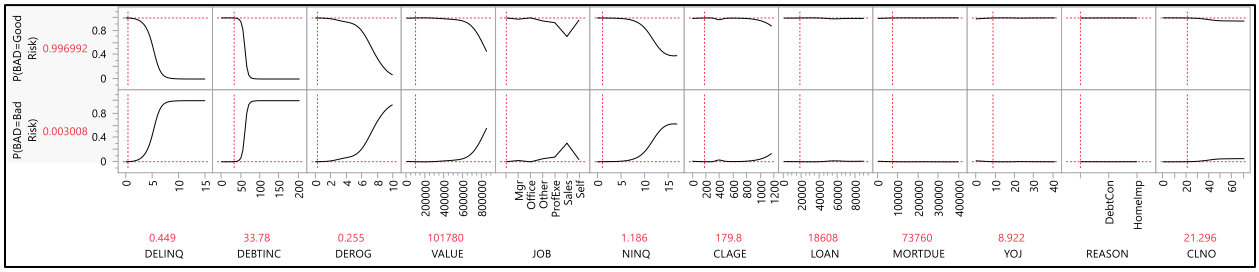
Boosted Tree Informative Missing

Column	Main Effect	Total Effect
DELINQ	0.222	0.520
REASON	0.040	0.434
DEROG	0.075	0.335
JOB	0.031	0.204
CLNO	0.021	0.136
NINQ	0.019	0.107
YOJ	0.019	0.092
DEBTINC	0.024	0.078
CLAGE	0.020	0.071
VALUE	0.006	0.015
LOAN	0.004	0.014
MORTDUE	0.004	0.011

Interactions between variables, as well as their effects on the Bad Risk response variable can be examined through the model profilers. Each of these profilers will be reviewed individually.

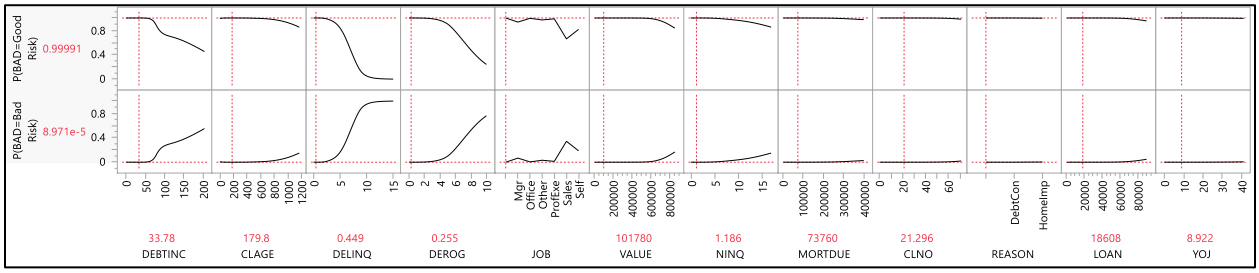
Complex Neural Net: 3-3-3, 40(40) Models/20 Tours, Squared Penalty, Informative Missing
Prediction Profiler

The Complex Neural Net model profiler exposes a handful of variable ranges that contribute to a bad credit risk. There is a sharp incline in Bad Risk between three and seven delinquent lines of credit (DELINQ). This really should come as no surprise. Similarly, there is a steep slope for debt-to-income ratio (DEBTINC) between 50 and 65. The number of derogatory credit reports (DEROG) lowers the threshold for DELINQ and DEBTINC to impact Bad Risk.



Simple Neural Net: 3-0-0, 40(37) Models/20 Tours, Squared Penalty, Informative Missing
Prediction Profiler

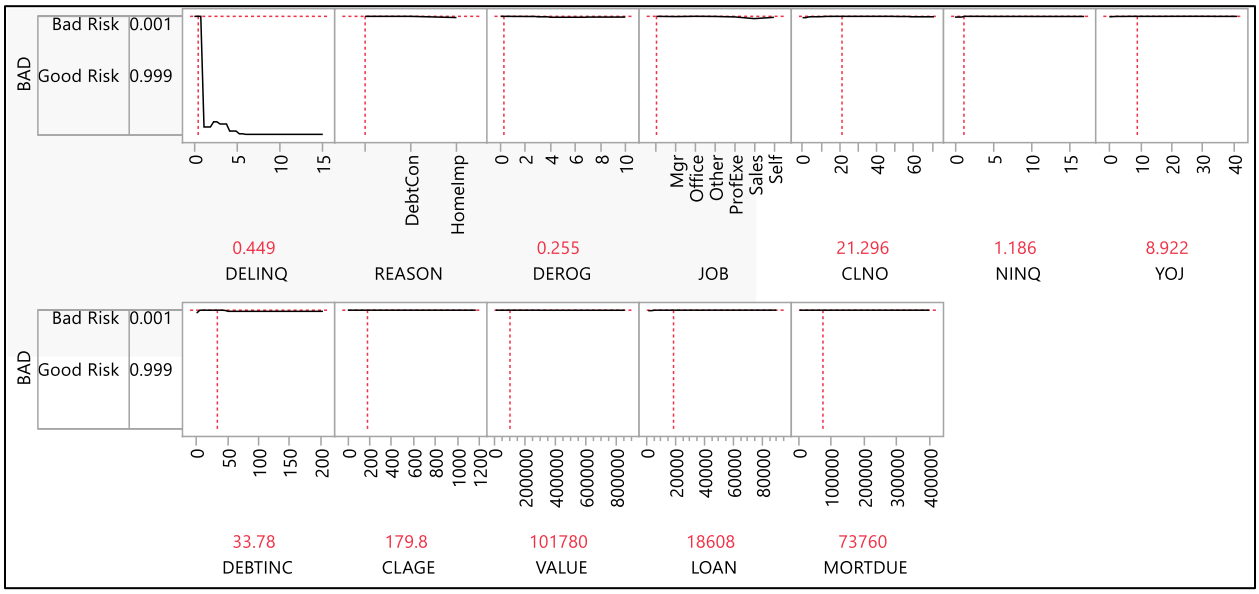
The Simple Neural Net model highlights a more intriguing set of important variables, with debt-to-income ratio (DEBTINC) more clearly impacting the other variable effects. With this model, an increase to DEBTINC lowers the threshold for DEROG and DELINQ to predict a bad credit risk. Stratifying the data by job type (JOB) shows a change to the effect pattern of DEBTINC, with most job types revealing a dip in bad credit risk around the 0.25 level.



Boosted Tree, Informative Missing
Prediction Profiler

The Boosted Tree profiler shows the dramatic impact that delinquencies have on a borrower's Bad Risk categorization. In fact, keeping all other values at their mean position and adjusting the value of DELINQ to just 1 brings the probability of Bad Risk to 0.932. However, disaggregating the REASON variable between DebtCon (Debt Consolidation) and HomeImp

(Home Improvement) reveals some intriguing trend distinctions in other features in the data set. For example, JOB (Job Type) and YOJ (Years On the Job) begin to show more variability.



The predictive models in this analysis show that not only one feature is solely predictive of bad credit risk. Multiple forces are at play, and certain features can influence others. This interaction between variable effects would necessitate a holistic approach to predicting bad credit risk. Also, taking into consideration any missing data in a borrower's profile or application would be critical for determining their overall risk.