Chart, histogram

Description automatically generatedThis is a data set with 5960 data records. The variable of interest is categorical: the likelihood of someone being a bad (or good) credit risk. I will examine 12 variables in relation to this y variable. Our data has 80% of the people identified as a good risk and 20% identified as a bad credit risk. I am interested in variables which might forecast the 20% “bad” credit risks.

Chart, box and whisker chart

Description automatically generated

Is it possible that the number of delinquent trade lines each person carries is highly predictive of if they are a good or bad credit risk? Possibly, I will know after I build my models, but this is an interesting graph of the data points. As logic might dictate, it appears that carrying more delinquent lines of credit predicts that a person will be a bad credit risk.

A picture containing graphical user interface

Description automatically generated

Table

Description automatically generated

I chose to look at a combination of linear and non-linear models. As we have learned, non-linear models have increased prediction power over linear models, but the interpretation of the non-linear model can be more challenging. However, JMP software has some unique tools that have been developed specifically to ease the interpretation.

For the linear models, I used the benchmark OLS and Adaptive Lasso. Lasso is able to do variable selection and estimation simultaneously. Lasso will shrink the co-efficient of an uninformative variable all the way to zero, thus the weaker variables are no longer entering the model. This is an example of a penalized logistic regression. This style of modeling handles highly correlated variables by assigning zero to an uninformative or redundant predictor variable(s). The adaptive Lasso model performed very well on the test data with a .0817 misclassification rate; the area under the ROC is .7862 which at a quick glance, looks to be a poorer response than the non-linear models.

Chart, line chart

Description automatically generated

The area under the curve is an overall measure of how well the model sorts the data (1 is perfect, the model misses nothing). The AUC and misclassification rate are the appropriate measures of a model with a nominal response variable.

AUC for Nominal Logistic Benchmark on the test data:

|  | **BAD** | **Area** |
| --- | --- | --- |
|  | Bad Risk | 0.7861 |
|  | Good Risk | 0.7861 |

The AUC for the OLS model is very similar to the Adaptive Lasso having virtually the same AUC. Linear models can be restrictive in that they cannot model highly complex relationships between variables.

Non-Linear Models:

I see the Boosted Neural Network has the next lowest area under the curve and this is the first look at a non-linear model. Neural network with TanH is supposed to account for data with non-linear and linear relationships. Based on my graph fitting response by of y variable with delinq, I wonder if we have some non-linearity in the variables. Also, I wanted to compare Boosted NN with both squared error and absolute error. Squared:

|  | **BAD** | **Area** |
| --- | --- | --- |
|  | Good Risk | 0.8761 |
|  | Bad Risk | 0.8761 |

Boosted Neural Network with 1 layer TanH 3 nodes, number of models is 40 and squared penalty method, AUC for “Bad” on test data is .876. I ran this twice and in the first chart it is .8749, both are slightly better than the absolute penalty method on test data for Boosted NN:

|  | **BAD** | **Area** |
| --- | --- | --- |
|  | Good Risk | 0.8508 |
|  | Bad Risk | 0.8508 |

Non linear models allow use of the informative missing option: Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated The fit of the bootstrapping improves with the informative missing box checked: .876 vs .936. The missing values are replaced with means of the non-missing values. This may not be a reasonable replacement in our dataset, however. I guess that the missing data represents data points that the loan applicant ommited because they believed them to be unfavorable. That would be directly correlated to our y variable. However, of all the linear and non-linear models I compared, boosted trees made the best model.

Boosted Trees model is our best model based on the .081 missclassification rate and .947 AUC. Chart, line chart

Description automatically generatedRandom Forest and Boosted Tree models are both ensemble methods and their model results are often similar. I will look more closely at the column contribtions for the Boosted Tree model as it shows the best predictive power with our data.

Chart, line chart

Description automatically generatedThe fit of the boosted tree model without the informative missing option is greatly reduced. The AUC drops to .701 and the misclassification rate is .19.

Boosted tree Column Contributions informative missing is used:Table

Description automatically generated

Surprisingly, the number one predictor variable in this model is the reason for the loan: debt consolidation or home improvement. The next four highly predictive variables in order are: number of recent credit inquiries, number of delinquent trade lines, and years on the job. Those five account for 67% of the variability in the good vs bad credit risk. The next three variables: number of derogatory credit reports, job category and number of credit trade lines open also are correlated. This data set appears to have quite a number of correlated variables with the y variable. All of the variables show non-linearity in their predictor profiles.Diagram

Description automatically generated

Now that we see the relationship of the variables to the predictor, it makes sense that the OLS regressions were poor modeling this data because OLS finds average relationships between the x and y variables. But if the data has outliers and an atypical distribution, averages are not meaningful. Thus, the partition models are better with this dataset.

Column Contributions ***without*** the informative missing option:

Table

Description automatically generatedThere is only one variable that contributes to the classification of “good” or ‘bad” credit risk and that is debt to income ratio (as a percentage). If we think that the likelyhood of missing data is intentional on the part of the loan applicant, or a tool applicants used to hide factors that might clasify them as bad credit risks, this is the model that applies best for modeling our y variable.

Calendar

Description automatically generated with low confidence

The profilers indicate a non-linearity between the debt to income variable and the good vs bad credit risk rating. This debt to income percentage has quite a large range of .52 – 203. The direction of the impact of debt to income on the y falls at the 50% mark and stays a 93% chance of being a bad credit risk from 50% to 203% debt ratio. Obviously, higher debt and lower income makes you a more likely bad credit risk.

Generally, the boosted trees model learns from it’s mistakes on former trees in the model. In boosting, the errors of the trees in the model are taken into account while new trees are created. So, there is a sequence or order of learning that the trees use while modeling. The advantage of this model is that it learns from the former trees.