Home Equity Loan Customers

BAN 525

Khayrayyah Haamid-Day

Financial institutions around the world provide some type of credit to their clients. This credit can range from a simple credit card to mortgage loans and a vast majority of bank clients will need a line of credit at some point. Many of these financial institutions will also have a massive number of clients who all come from various backgrounds and with this, comes different potential risks that must be considered. Therefore, financial institutions implement methods to determine potential credit risk to ensure their clients can manage their loans. Mainly the concern is making sure loans are not given to those with high default risk. While it is impossible to know for certain who will default, this analysis will seek to understand what factors may increase the likelihood of being a “bad risk”. The dataset that will be used consist of 5,960 rows and 13 columns, the response variable will be “BAD”, and each model will focus on predicting “bad risk”. The predictor variables will be loan amount, mortgage amount due, assessed valuation, reason for the loan, their job and years working at job, number of derogatory reports, number of delinquencies, oldest line of loan, number of credit inquiries, number of trade lines, and debt to income ratio.

The statistical methods that will be used are OLS as a baseline model, Boosted Tree, Bootstrap Forest, and Boosted NN. The last three models consist of two variations, one that account for informative missing and another that does not account for informative missing. All models will be constructed with random seed 123 and a validation column. This OLS method will be a logistic regression as this is a classification analysis. This method is known for its simplicity which makes it easy to understand and highly interpretable. Though, this method can be sensitive to outliers and may produce overfit models. The two method that will be Boosted work very similarly to each other. They begin with an estimating model that will learn from its errors at each step and observations with the largest error are given additional weight. Aside from this, Boosted Tree is built in a similar fashion that Bootstrap Forest is. These methods split the data and randomly construct decision trees from only a part of the sample of the data. This is supposed to address the issue with decision trees where individuals would be highly correlated with each other. Downside to these methods is they can be computationally intensive, especially with larger datasets. Boosted Tree will be built with the default settings and random seed while Bootstrap Forest will be built with 500 trees and 3 terms per split due to the size of the dataset. The Boosted NN will have one layer and three nodes in TanH and will use 40 models. This is a great ability of the NN, the number of inputs and layers that can be implemented The NN method is also very efficient as well as flexible as it allows for complex model building. The downside to this method is that they are black boxes which makes it difficult to interpret. The validation column that will be used to build each model has a 60/20/20 split into training, validation, and test set with a random seed of 123.

Text

Description automatically generated with medium confidenceA picture containing text

Description automatically generated

A picture containing text

Description automatically generatedGraphical user interface

Description automatically generated with low confidence

Text

Description automatically generated

Table

Description automatically generated with medium confidence

Graphical user interface, text, application, table

Description automatically generated with medium confidence

These results are organized beginning with OLS, Boosted Tree with informative missing and without informative missing. Bootstrap Forest with informative missing and without informative missing. While the last two are the Boosted NN with informative missing and without informative missing. A quick glance at methods with and without informative missing gives a good understanding of the impact of missing data on model performance. It was expected that OLS would lag in performance compared to the other methods. Most of the methods also share similar RSquared values to each other, other than Boosted Tree without informative missing being used, this model performed the worst out of all.

Table

Description automatically generated

The first thing to note is that model comparison shows entirely different RSquare values for the methods that were run without the informative missing ticked. For instance, Boosted Trees without informative missing has a RSquare of 0.25 and 0.35 but in the initial run it was 0.09 and 0.14. The same is true for Bootstrap Forest but is not for the Boosted NN. The model comparison does highlight a distinct winner in performance. The first Boosted Tree, built using informative missing, has a RSquare of 0.58 and 0.69, or 0.70. This is the highest amongst the other models. Though the final Boosted NN method does have the lowest errors and misclassification rate. Despite this, it is not by much and the Boosted Tree has a very high AUC of 0.94 which is very near perfect. Due to these factors, the first Boosted Tree is the best performing model in this comparison.

Table

Description automatically generated

The first thing to investigate is the confusion matrix, which is relevant due to this being a classification analysis. Confusion matrix are an excellent resource for determining the quality of the predictive analysis. A good model will have predicted response levels that are the same as the actual response level. The confusion matrix for this model shows great values for predicted values and actual response and these values from the training to the test set remain fairly consistent.

Table

Description automatically generated

The next important consideration, and perhaps the most important, is understanding what factors predict a person as a “Bad Risk”. Column contributions works similarly to variable importance, in this case, the largest contribution is from the predictor variable “Reason”. This variable accounts for the reason why an individual is applying for a loan. In this instance there are two reasons listed, debt consolidation and home improvement. The first is simply debt refinancing and the latter is for home renovations. Typically, those who are looking to consolidate loans are doing so because they have multiple debts that they wish to have under a single payment plan. It is understandable that a financial institution would see this as a bad risk, it’s a clear implication that this person has a high amount of debt. For home improvement, this is a more frivolous loan type that is categorized as personal debt. These types of loans require a higher credit score but do tend to get approved quickly. The next contribution is number of credit inquiries, too many inquires will elevate the level of risk a borrower poses and can negatively impact credit scores. The third contribution is number of delinquencies, this variable refers to how many accounts have not been paid in 30 days or more. This would be a clear indication of a bad risk that a financial institution would not want to have.

Diagram

Description automatically generated

The prediction profile provides a great way to elaborate on how each factor can impact the likelihood of being considered a bad or good risk. For instance, this profile has been determined to be a good risk at 0.95. If the number of delinquencies is increased to just two, this number drastically drops and now this profile is considered a bad risk at 0.81. If it is the other way around and the only variable changed is reason, from home improvement to debt consolidation, the good risk decreases to 0.47 and bad risk increases to 0.52.

In conclusion, the chosen predictive model appears to accurately predict what factors will increase the likelihood of a person being deemed a “bad risk” by a financial institution. Further research shows that financial institutions do in fact look at variables such as these in order to ensure that loans are only given to those with very low, to non-existent, default risk.