Zackery Bradley

BAN 525

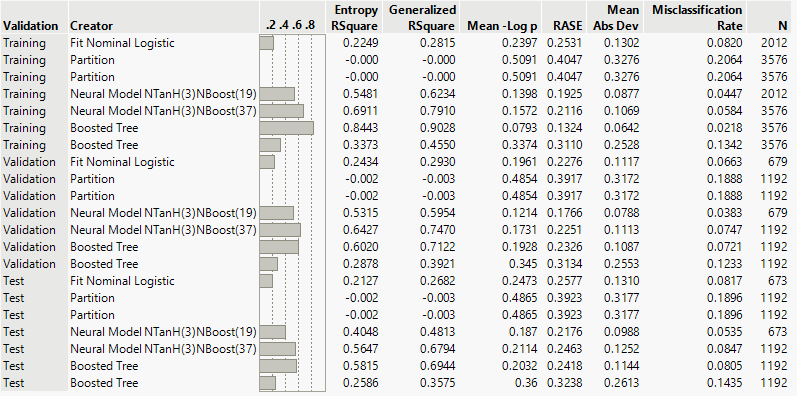
20220306

Model 7 Final Assignment

In our final assignment, we have been tasked with creating models displaying the plethora of knowledge we have gained over the entirety of the semester. Our purpose in this project is to analyze differing variables and how they contribute to negative credit risk. In addition, our provided dataset is missing several data entries, so we will be analyzing methods with auto filling those missing cells, as well as leaving them as is. Traditionally, there are several factors that have been known to weigh heavily on credit scores. Some of these include delinquencies, number of recent inquiries, and the reasons for the requested loan. Having substantial experience as a loan officer, it is my hypothesis that the reason for the loan, along with the number of delinquencies will have the most detrimental effect on calculating credit risk. Throughout our analysis, we chose to strategically employ four different modeling methods. These methods include our standard OLS method, in addition to Random Forest, Neural Network, and Boosted Tree models. OLS is typically used as a method for estimating the parameters that are unknown in the model. This method helps us to find the relationships between independent and dependent variables. Random Forest methods are usually highly desired in the statistic world and its creation is considered a modern innovation. Random Forest operates by selecting distinct relevant variables for prediction, without the need to rely on any functional or distributional assumptions Random Forest functions by forming a multitude of decision trees, all of which are independent of each other. Each node in the tree is then split using the best among a subset of predictors that have been randomly chosen at that node. By calculating each tree in random fashion, the Random Forest eliminates the criticisms of decision tree methodology in which individual trees have been known to be high correlated. This functionality gives Random Forest methods immense versatility and flexibility, as it is considered to work well with both categorical and continuous variables. NN models are considered very flexible algorithms that are able to model extraordinarily complex relations between model inputs in outputs. Each NN consists of an unput layer, in addition to one or more hidden layers followed by an output layer. The hidden layers are made up of one or more nodes, in which a linear combination of the input variables is transformed. In JMP Pro, there are three types of functions that can be utilized in a NN model consisting of TanH, Linear, and Gaussian. TanH, a hyperbolic tangent function, is similar to the shape of a logistic function. Linear is similar to a linear regression model. Gaussian is a bell-shaped function, which is similar to the normal distribution density function. Boosted Tree models are built sequentially, where at each step of the process the model learns from its previous errors. They begin by estimating a model and obtaining residuals in the case of a regression, and misclassification rate in case of classification. This model gives observations with the biggest errors additional weight as opposed to others. The goal for this particular model is to continuously improve model fit by correcting the bad fits given from the previous steps. After eliminating bad fits, Boosted Trees are then averaged to produce a prediction or observation from each observation. By averaging, it is implied that predictions or classifications will not be unstable.

Analysis and Model Comparison

Regarding our predominant methods as previously stated, OLS can be described as the “workhorse” method and is a standard operating tool in statistics analysis. On the contrary, OLS tends to produce estimates with a great deal of variance regarding big data. In all, OLS typically generates poor forecasts. Random Forest has many additional advantages. One advantage is that Random Forest can automatically manage missing values in the data, as well as being uniquely robust to outliers, as the method oversees them autonomously. Random Forest methods also come with some disadvantages as well. One disadvantage worth nothing is that Random Forest usually is a tedious process, as computing the data often takes more time as compared to other methods. Our third method, NN also comes with its own list of advantages and shortcomings. As previously mentioned, NN typically has good predictive ability and is capable of capturing complex relationships. However, NN has no variable selection mechanism, so utilizing JMP’s variable importance is vital when identifying the most important variables in the data. Typically, NN is referred to as a “black box” method, meaning that findings are often difficult to interpret. Lastly, Boosted Trees are considered highly efficient on both classification and regression models, and typically provide more accurate predictions than that of a Random Forest. On the contrary, Boosted Trees are overly sensitive to outliers, and also tend to overfit data is too many trees are used. Procedurally, we will be employing a cross-validation analysis. Cross-validation techniques are built on the concept of leaving out a part of the data out of the estimation process as a buffer. As the predictions of the model stop improving as the data holdout process occurs, model growth then stops, and estimates are obtained. This creates a ‘split’ in the data. These splits are then broken down into three parts in the terms of our model. Those parts include training, validation, and testing splits respectfully. Training data is the portion that is used to estimate our model. Validation data is not used in estimate directly, but instead works in the background to determine the optimal point in which the model stops. Furthermore, the test data is never actually used in the model estimation. Test data is used to represent “new observations” and assists in providing and unbiased analysis of the predictive ability of the model. With the validation method only being utilized indirectly, we choose our model that displays the best performance on the test data. Upon examination of our analysis, we can readily see that one model outperformed the rest. Our Boosted Tree model outperformed our other models by a convincing margin, recording an RSquare of 0.6944, followed by a RASE of 0.2418 and misclassification rate of of 0.0905 with missing information auto filled, and an RSquare of 0.3575 followed by a RASE and misclassification rate of 0.3238 and 0.1435 respectfully, with the data being left as is. Close behind is our NN model which calculated a RSquare of 0.6794 followed by a RASE if 0.2463, and misclassification rate of 0.0847 with missing values auto filled, in addition to providing an RSquare of 0.4813 and RASE of 0.2176, in addition to a misclassification rate of 0.0535. with missing values left as is. After carefully analyzing our numbers, we have determined that our Boosted Tree model with missing values auto filled is by far our superior methodology.



After determining our top performing model, we conducted further analysis by employing an in-depth examination of the importance of each variable in the dataset. From our results, our previous hypothesis was deemed to be true as credit delinquencies accounted for an astronomical 0.52 of total effect on the dataset, and reasoning accounting for an additional 0.423 of total effect as seen in the visual below. Second, we chose to further analyze our data by examining the results of our confusion matrix. By examining predicted versus actual results, we can visualize how accurately our model performed with the missing values that were filled automatically.

| **Column** | **Main Effect** | **Total Effect** |  |
| --- | --- | --- | --- |
| DELINQ | 0.244 | 0.52 |  |
| REASON | 0.038 | 0.423 |  |
| DEROG | 0.069 | 0.34 |  |
| JOB | 0.039 | 0.204 |  |
| CLNO | 0.022 | 0.156 |  |
| NINQ | 0.022 | 0.107 |  |
| YOJ | 0.019 | 0.093 |  |
| CLAGE | 0.023 | 0.084 |  |
| DEBTINC | 0.023 | 0.068 |  |
| LOAN | 0.007 | 0.018 |  |
| VALUE | 0.006 | 0.017 |  |
| MORTDUE | 0.004 | 0.012 |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test   | **Actual** | **Predicted Count** | | | --- | --- | --- | | **BAD** | **Good Risk** | **Bad Risk** | | Good Risk | 941 | 25 | | Bad Risk | 71 | 155 | |
| | **Actual** | **Predicted Rate** | | | --- | --- | --- | | **BAD** | **Good Risk** | **Bad Risk** | | Good Risk | 0.974 | 0.026 | | Bad Risk | 0.314 | 0.686 | |

Lastly, we performed experimentation with our prediction profiler visual to show the effect reason had on the rest of the variables in order to provide a mental picture of how important reasoning is in credit risk assessment. Firstly, we set our reasoning on debt consolidation, and then placed it on home improvements in the second image. We can see that there is a sizeable variance in how our variables are impacted just by simply changing the need for the loan.

Variable factors with reasoning listed as “Debtcon”

Chart

Description automatically generated with medium confidence

Variable factors with reasoning listed as “HomeImp”

Graphical user interface

Description automatically generated with low confidence

In closing of our analysis, we can say with certain confidence that Boosted Tree methods performs adequately as opposed other methods such as Random Forest and Neural Network in the discussion of this particular dataset. This can be seen on our overall analysis of the other performance methods Boosted Tree has proven to be superior statistically speaking. Given the competition in the loan industry, it is vital that companies have access to the best modeling analysis methods available in order to stay relevant. By seeing the significance of each variable associated with credit assessments, banks are able to make better economic decisions on each assessment, potentially saving the banks millions in lost revenue while mitigating the risk of loss.