Prediction of Borrowers’ Credit Risk

**Introduction**

The current analysis applies supervised classification models to the prediction of borrowers’ credit risk and the identification of the primary predictors of credit risk. The dataset used for the analysis consists of the credit risk, credit history, indebtedness, and occupational information for home equity loan customers. The response variable, credit risk (BAD), has two levels – the target level, bad risk, and its complement good risk. Of the 5,960 borrowers included in the dataset, most customers had good credit risk with bad risk customers corresponding to 20% of the sample.

The dataset consists of ten numeric and two categorical predictors. All predictors except for requested loan amount (LOAN) had missing values. For the numeric predictors, the number of missing values ranged from 112 for the value of current property (VALUE) to 708 for number of major derogatory reports from lenders (DEROG). For the categorical variables, reason for borrowing (REASON) had 252 missing values while job had 279 missing values. Whether or not a record has missing observations could be informational: in the present scenario, the absence of critical self-reported credit information could be indicative of bad risk as individuals are more likely to underreport negative information. For this reason, extracting information from missing values could improve the performance of the selected algorithms. The analyst conducted initial analyses to predict credit risk using adaptive lasso logistic regression, boosted trees, random forest, and boosted neural networks, without extracting information from missing records. The latter three analyses were then repeated with the informative missing option active.

**Analysis and Method Comparison**

The logistic regression model was selected for its simplicity and the information provided on the effect of predictors on credit risk as indicated by the variables’ odds ratios. As the default statistical model for the prediction of binary variables, the logistic regression model also forms a baseline for the comparison of the more advanced ML and neural network (NN) models. The analyst used two tree-based models – boosted trees and bootstrap/random models – and the boosted NN. The boosted approach, as applied to both decision trees and neural networks, fits a large set of weak/simple learners with each subsequent learner trained on the residual error from the previous simple model (Grayson, Gardner & Phillips, 2015; JMP Statistical Discovery LLC, 2022). Boosting has the advantage of efficiently determining the optimal number of layers to maximize the model’s predictive power when compared to the reliance on the user to specify the number of layers and evaluate performance across the fitted algorithms (Grayson et al., 2015). However, to maintain computational efficiency, the base model should be sufficiently simple with limited number of leaves for boosted trees and nodes for NNs (JMP Statistical Discovery LLC, 2022).

As with boosted trees, random forest models use the decision tree (DT) format and, therefore, also benefit from the unbiased prediction arising from the non-distributional nature of DTs or the need for a linear relationship between the response and predictor variables. In addition, random forest (RF) models resolve for the unstable predictions and overemphasis of predictors that are strongly correlated with the target, common to simple DTs, through the bagging and random sampling methods (Grayson et al., 2015). At each split in the decision tree, RFs use a bootstrap sample of drawn from the training data and a random sample of n predictors, less than the total number of predictors. Consequently, variables with little direct correlation to the target are more likely to be included as decision variables. Predictions are made from all trees and the final prediction is determined through a majority vote.

Other than the logistic regression model for which the informative missing option is unavailable in JMP, the predictive models were fitted to the data, with and without the option.

The treatment of missing values when the informative missing option is left unselected varies depending on the algorithm. The rows with missing values are ignored for logistic regression and neural networks while the records with missing values on a splitting variable are randomly assigned to either side of the split and missing values are filled by mean/mode imputation of the corresponding subset (JMP Statistical Discovery LLC, 2022). In contrast, the informative missing option extracts new information from missing records by adding an additional level, labelled missing, for categorical predictors with missing values and creating a binary missing value indicator for numeric columns (JMP Statistical Discovery LLC, 2022).

The analyst created as a validation column as the first step in the analysis with a 60/20/20 training, validation and test split, stratified by credit risk, and using a random seed (123) for reproducibility. For the logistic regression model, the analyst compared the test set R-Squared for different penalization approaches and with/without the adaptive alternative. The adaptive lasso method was selected as it provided the highest explanatory power of the logistic regression models. For the random forest and boosted tree models, the analyst retained the default parameters. For the neural network, the analyst specified 500 boosts and added an additional Gaussian layer in the base model. The model parameters were retained in the analysis with the informative missing option selected, facilitating the evaluation of the change in predictive performance arising solely from extracting information from missing records.

The predictive performance of the adaptive lasso logistic regression, bootstrap forest, boosted trees, and boosted neural network models in the test set when the informative missing functionality is not used is compared in the table below. Using the generalized R-square value, boosted trees explained the least variation in credit risk (17.31%) while the boosted neural network explained the highest proportion of variance at 6.59%. The random forest model had the second-best performance while adaptive lasso logistic regression had the second worst performance. However, when considering the misclassification rate, the logistic regression model had the second-best performance after the boosted neural network with the latter inaccurately classifying 6.59% of the observations compared to 8.05% for the adaptive lasso logistic regression model. The inaccuracy rate was much higher for boosted trees and random forest models which misclassified 18.5% and 17.5% of the test set observations, respectively. The performance of all four models was dismal in the prediction of the target level, Bad Risk as indicated by the low sensitivity. Boosted NN and adaptive lasso LR had the highest sensitivity but still inaccurately predicted 63% and 53% of respondents with high credit risk as having low risk.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Logistic Regression - Adaptive Lasso | Bootstrap/Random Forest | Boosted Trees | Boosted Neural Network |
| Generalized R-Square | 29.20% | 40.63% | 17.31% | 48.03% |
| Misclassification rate | 8.05% | 17.45% | 18.54% | 6.59% |
| Sensitivity | 27.10% | 16.40% | 9.70% | 37.10% |
| Specificity | 99.30% | 99.70% | 99.40% | 99.80% |

Including information from missing values improved the performance of all three models, as indicated in the table below. The improvement in performance was highest for boosted trees as its generalized R-Square increased by slightly more than 50%. The random forest model’s explanatory value increased by 34% while the boosted NN showcased the least improvement with the proportion of credit risk explained increased by roughly 10%. Boosted trees had the best performance across board, except for sensitivity where random forest was superior. The boosted tree model misclassified 8.47% of all observations with a lower misclassification rate for good risk clients (1.7% misclassified) compared to bad risk clients where 35.7% of clients were misclassified. The RF model had the highest accuracy in the prediction of bad risk, inaccurately prediction 34.5% of bad risk clients, compared to 35.7% for boosted trees and 37.4% for boosted neural networks.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Bootstrap Forest | Boosted Trees | Boosted Neural Network |
| Generalized R-Square | 62.86% | 67.62% | 58.01% |
| Misclassification rate | 10.32% | 8.47% | 10.98% |
| Sensitivity | 65.50% | 64.30% | 62.60% |
| Specificity | 95.70% | 98.30% | 95.60% |

Overall, the boosted neural network without the information missing option had the lowest misclassification rate (6.59%) across board. However, the model heavily sacrificed the accuracy in the prediction of bad credit risk, which is the primary aim from the lender’s perspective. Targeting the accuracy of bad risk predictions, i.e., the model’s sensitivity, in the validation set rather than the R-Squared statistic could have resulted in a model whose performance is aligned with the institution’s objective of identifying borrowers who are likely to default, although at the expense of an increase false default rate. The boosted tree model with active informative missing strikes a better balance between misclassification rate and sensitivity: the algorithm misclassified 8.47% of the test set observations, only 1.88% higher than the aforementioned boosted NN model, while accurately predicting 64.3% of bad test set credit risk, 27.2% higher than the comparative model. Thus, the analyst selected the boosted tree model with the active informative missing option.

**Interpretation**

Variable importance was assessed from the column contributions to the informative missing boosted trees algorithm. Reason for seeking a loan was the most important predictor, contributing 35.43% of the model’s explanatory power. Other variables that contributed more than 10% of the model’s explanatory power, in descending order of importance are the number of recent credit inquiries (NINQ), number of credit lines (CLNO) and the number of delinquent credit lines (DELINQ). Overall, the four variables accounted for 70.21% of the explained proportion of variance. From the prediction profiler, home improvement loans had higher risk of default when compared to debt consolidation loans. The least important variables with less than 2% explanatory power for which less than 100 splits were performed, in ascending order of importance include, amount due for existing mortgage, value of current property, age of oldest credit line in months, and the borrower’s debt to equity ratio.

The risk of default for records with missing values for REASON was as high as the risk for home improvement loans, illustrating the usefulness of extracting missing values information when predicting credit risk. The risk of default increased sharply for recent credit inquiries exceeding 1.186. The risk of default decreases for more than 0.86 delinquent credit lines while the risk for CLNO increases for more than 21 credit lines. However, the predictors interaction with the response as indicated by the cross-sectional profiler is of little use for decision tree models given the interaction of the predictors to influence credit risk. For instance, when the reason for the loan is set at home improvement, CLNO and DELINQ are much more useful predictors with more consistent increase and decrease in credit risk with the respective increase in CLNO and DELINQ. The described borrower is categorized as Bad Risk, with his/her probability of defaulting on the loan predicted as a near certainty (p = 0.9997)

References

Grayson, J., Gardner, S., & Stephens, M. (2015). *Building better models with JMP® Pro*.

SAS Institute Inc.

JMP Statistical Discovery LLC (2022). *JMP ® 17 predictive and specialized modeling*. JMP

Statistical Discovery LL