Predictive Analytics Final Project

Aditya Gidda - 108750868

Jaswanth Das Gunturu - 108673927

Venkatesh Gaallee - 108634231

Fake Name : 3G

University of Colorado Denver

Part-A

* Missing Values in Utilization were dealt with Multivariate Normal Imputation.
* Missing values in the Months column were imputed as 24, from the definition of Delinquencies, Months, we concluded that the no. of delinquencies in last 24 months is zero meant that the no. of months since the last delinquency would be at least 24 or more. So, we went for 24. We ignored one row where the months and delinquencies were both given zero.
* As there were only 2 rows for none in HOME column and there were no none rows in the NEW Applications file, we recoded NONE as ANY in the Personal Loans file.
* As the Income was left skewed, log transformation was applied, log10 transformation was applied over open(LoC) too as it was left skewed and Balance was transformed using square root transformation twice.
* When checking the new applications .jmp file we found that Amt Funded, Monthly Pmt, Pmts to Date were all zeros and Standing was N/A, which meant that they were not to be considered int the prediction and we didn’t consider them into the model, as they were completely new applicants who don’t have an of them.
* We changed the variable type of Income to continuous, FICO to ordinal, Employ to nominal.
* We found some relations like, as income increases interest rate decreases and as amount funded increases interest rate increases, FICO score increases interest rate decreases.

Part-B

* The variables we decided to use in the final model were Amt Requested, Length, Log(Income), Log10(openLoC), Employ, Purpose, FICO, Home(Recoded), Utilization(Imputed), D/I, √√(Balance), Delinquent, Months(Recoded).
* We arrived with these variables after many trial and error methods and finding the contributions of the variables to the model over different models and ended up using the boosted tree model. We truncated the variables to just 13, with some facts we found in the Domain Knowledge.
* Model Comparison:

|  |  |
| --- | --- |
| Model Type | Validation R2 |
| General Regression | 78.59% |
| Stepwise Regression(Forward & Backward) | 78.62% |
| KNN | 72.78% |
| Boosted Tree | 83.9% |
| Boosted Tree-1 | 81.16% |
| Bootstrap Forest | 81.42% |
| Neural Nets | 82.89% |

These were the best models in all the models we’ve tried.

* The Boosted tree which we arrive as the best model has a Training R2 of 83.7%, Validation R2 of 83.9%, Test R2 of 83.6%, Splits of 3, Learning Rate 0.4, RMSE 0.0152, Random Seed of 123.