AUGUST 31, 2021

PREDICTING INDIAN CONSUMER LOAN DEFAULTS

Minimum Viable Product

Prepared for: Metis Staff Instructors(Kimberly Fessel, Brian McGarry)

Prepared by: Rahul Raju

August 31, 2021

An initial proposal was made to aid Indian financial institutions in more effectively evaluating consumer loan applicants. The profile and default data of more than 240,000 past retail loan applicants were used to build several models. Below is a table that provides the features used in the analysis as well the data type. The target will be represented as a 1 for applicants that have defaulted on loans(positive) and a 0 for applicants that have not defaulted(negative)

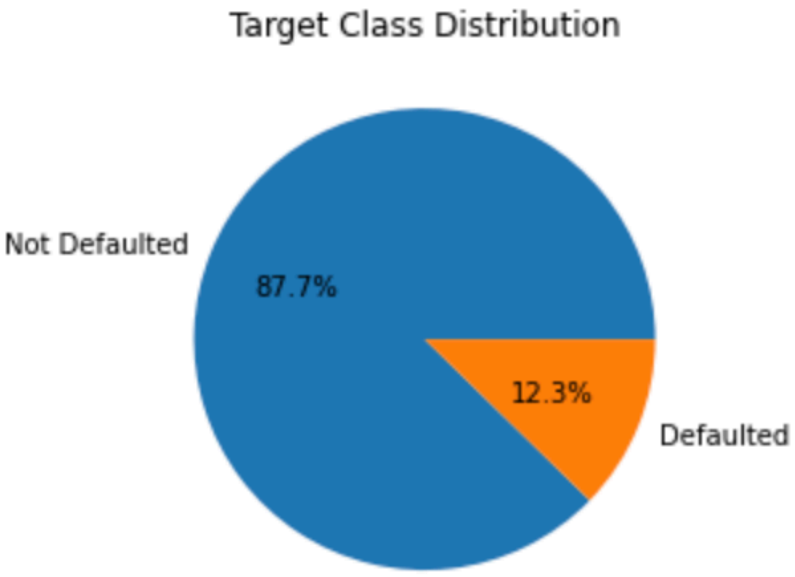
INPUT VARIABLES	
TYPE	**FEATURE**
Quantitative	Income
Quantitative	Age
Quantitative	Experience(years worked)
Qualitative	Marital Status
Qualitative	Home Ownership
Qualitative	Car Ownership
Qualitative	Profession
Qualitative	State of Residence
Quantitative	Current Job Years
Quantitative	Current House Years

Evaluation Metrics

Prior to the modeling phase, metrics were established that be would used to evaluate model effectiveness. This was done first, as a choice of metric often informs modeling parameters. In this case, a focus was placed on the "F beta score." This was chosen due to its functionality in changing the relative importance of 'recall' and 'precision' through adjustments of a "beta" parameter. Generally, during times of economic decline, financial institutions tend to tighten their lending standards as they are more concerned with the loss of principle. Therefore the focus will be placed on limiting false negatives or not approving loan applicants that will default. This would result in placing more importance on recall(false negatives) by adjusting the beta to greater than 1. When the economy improves lenders loosen their standards and aggressively compete for new business. During these time the focus could be more on precision through an adjustment of the beta to below 1.

Model Selection

Numerous baseline models were created across a variety of algorithms. Each model has its advantages and disadvantages but an initial focus was placed on logistic regression. This model predicts the probability of an observation being positive(will default) but also allows for hard classification by the setting of a threshold. Adjusting this threshold also allows for a trade-off between precision and recall. The data was standardized prior modeling, and a grid search was performed to optimize F1 via the tuning of regularization strength. Due to the significant class imbalance of the target, as depicted in the pie chart below, the *class weight parameter* was set to *'balanced.'*



The table below summarizes the results of the logistic model as well as supplemental metrics for informational purposes.

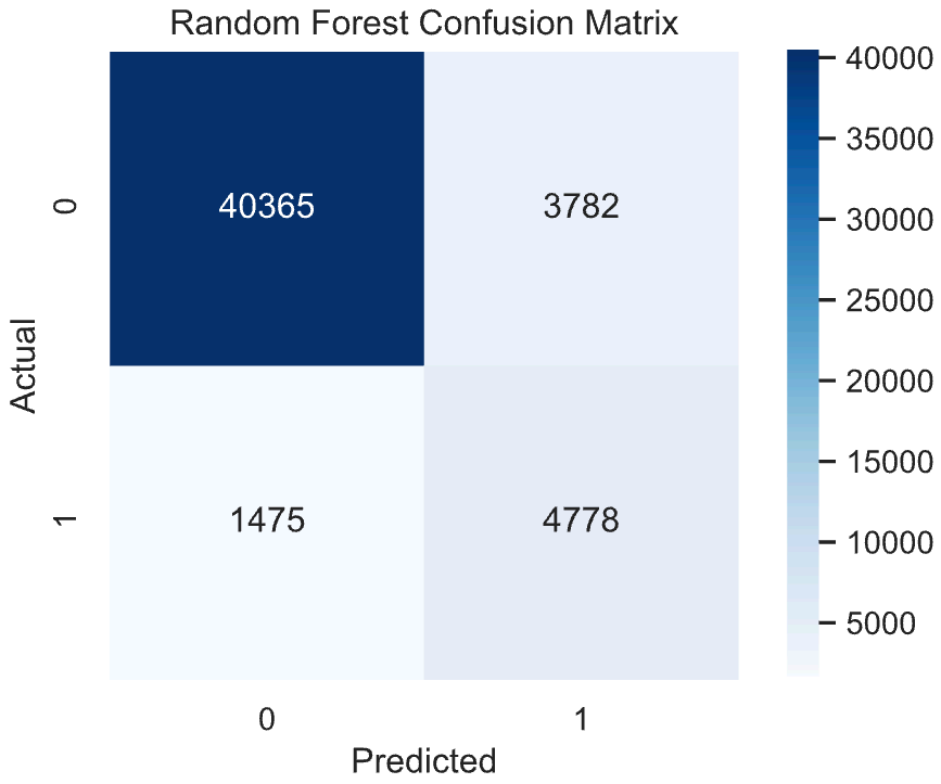
METRIC	SCORE
F 2(beta=3)	0.44
Accuracy	0.54
Precision	0.14
Recall	0.56
F 1	0.23
ROC AUC	0.58
Log Loss	0.68

The beta of the F 2 was adjusted to give 3 times more importance to recall but did not yield a high score.

More baseline models were built using algorithms such as k-nearest neighbors, decision trees and random forest. Of all the 4 models evaluated, random forest has thus far has seemed to provide the best results as outlined in the table below.

METRIC	SCORE
F 2(beta=3)	0.74
Accuracy	0.90
Precision	0.56
Recall	0.76
F 1	0.65

This confusion matrix displays inputs used to calculated many of the metrics shown thus far.



Further Work

- A “Grid Search CV” will be performed to test and optimize hyper-parameters in relation to F 1
- An *ensemble* tree method such as XGboost will be used to enhance the final algorithm.