The joined dataset was further filtered to remove the variables that were not relevant (ex: time of day of the application) or identified as unethical (ex: gender). 19 variables were shortlisted and used to model and predict the target variable. Box plots were made to identify any outliers and to discuss if any datapoints should be removed. Next the data set was normalized to bring all the variables to the scale of 0 to 1. The data was split into two parts with 70:30 split for training and testing datasets. Logistic regression model was created and variables that were insignificant were identified (p<0.05) and removed. Variables with high VIF (variable inflation factors) were also identified and removed and the logistic regression model was refitted. As an initial attempt, probability threshold value of 0.5 was used to predict the target variable. The confusion matrix, accuracy, sensitivity and specificity are shown below.

Confusion Matrix and accuracy of the model for p=0.5 probability cut-off threshold

|  |  |  |  |
| --- | --- | --- | --- |
| 0= Can pay  1= Can’t pay | | Reference | |
| 0 | 1 |
| Prediction | 0 | 59840 | 5211 |
| 1 | 46 | 31 |

|  |  |
| --- | --- |
| Accuracy | 91.93 % |
| Sensitivity | 99.92 % |
| Specificity | 5.91% |

Since the cost (or damage) of giving a loan to a customer who cannot re-pay is much more costlier than not giving a loan to an eligible customer (cost of loosing business), a cost analysis was done by making the cost of approving a bad loan to be five times more costly than not approving a loan for a customer who would repay. A probability threshold value of 0.16 gave the minimum cost as shown in the below figure.

A graph with a line graph

Description automatically generated

Confusion Matrix and accuracy of the model for p=0.16 probability cut-off threshold after optimizing for cost are shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| 0= Can pay  1= Can’t pay | | Reference | |
| 0 | 1 |
| Prediction | 0 | 53735 | 3394 |
| 1 | 6151 | 1848 |

|  |  |
| --- | --- |
| Accuracy | 85.34 % |
| Sensitivity | 89.73 % |
| Specificity | 35.25% |

AUC was 0.738 for the optimum probability cut-off threshold of 0.16

A graph with a curve

Description automatically generated