CredX:
Acquisition and Operation Risk Analytics:
Final Submission

PGDDS: December-2018

By: Pravin Pawar and Aashish Sharma

Objective

Problem Statement

- CredX is a leading credit card provider that gets thousands of credit card applications every year. But in the past few years, it has experienced an increase in credit loss.
- The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

Objective

- The objective is to help CredX identify the right customers using predictive models.
- To build an application scorecard and identify the cut-off score below which one would not grant credit cards to applicants.
- We need to determine the factors affecting credit risk and create strategies to mitigate the acquisition risk and assess the financial benefit of the project.

Methodology

Analytics Framework	**CRISP-DM Framework** *(Cross-industry standard process for data mining)*
Objective	**Identify Customer who could default?**
Available Data	Customer Demographic Data and Customer Credit Bureau Data
Variable Identification Techniques	**WOE and IV** *(Weight Of Evidence and Information Values)*
Models Tried *(**Ensemble technique**: multiple models tried to get the optimum results)*	▪ Logistic Regression ▪ Random Forest ▪ ADA Boost ▪ Light Gradient Boosting ▪ Gradient Boosting

List of Variables

Customer's Demographic Data

- Application ID
- Age
- Gender
- Marital Status
- No of dependents
- Income
- Education
- Profession
- Type of residence
- No of months in current residence
- No of months in current company
- Performance Tag

Customer's Credit Bureau Data

- Application ID
- No of times 90 DPD or worse in last 6 months
- No of times 60 DPD or worse in last 6 months
- No of times 30 DPD or worse in last 6 months
- No of times 90 DPD or worse in last 12 months
- No of times 60 DPD or worse in last 12 months
- No of times 30 DPD or worse in last 12 months
- Avgas CC Utilization in last 12 months
- No of trades opened in last 6 months
- No of trades opened in last 12 months
- No of PL trades opened in last 6 months
- No of PL trades opened in last 12 months
- No of Inquiries in last 6 months (excluding home & auto loans)
- No of Inquiries in last 12 months (excluding home & auto loans)
- Presence of open home loan
- Outstanding Balance
- Total No of Trades
- Presence of open auto loan
- Performance Tag

Approach

Data Preparation

1. Merge Demographic and Credit Bureau Data on Applicant ID
2. Check for missing values
3. Missing value and Outlier Treatment
4. EDA (Univariate and Multivariate)

WOE and IV

5. Information Value
6. Populate feature with their WOE values
7. Create data frames with and without WOE Values

Model Building

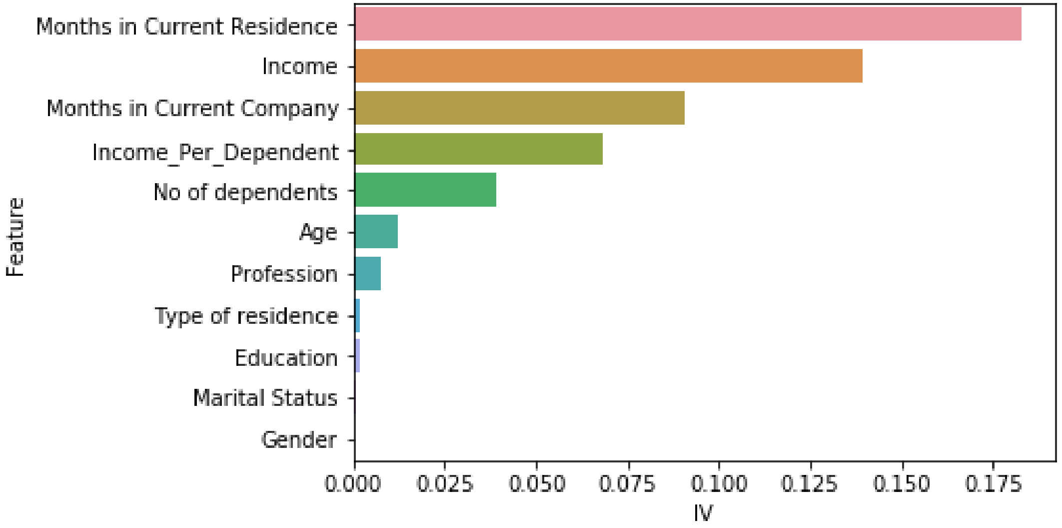
8. Split data to train and test
9. Addressing Class imbalance with SMOTE (Synthetic Minority Over-sampling Technique)
10. Building model on demographic data and combined data separately
11. Rebuilding Model using (Logistic Regression, Random Forest, ADA Boost, Light Gradient Boosting and Gradient Boosting)

Evaluation Metrics

12. Check Model on Test Data
13. Evaluate Model Metric
14. Build Application Score Card
15. Build Financial Strategies using the optimum model

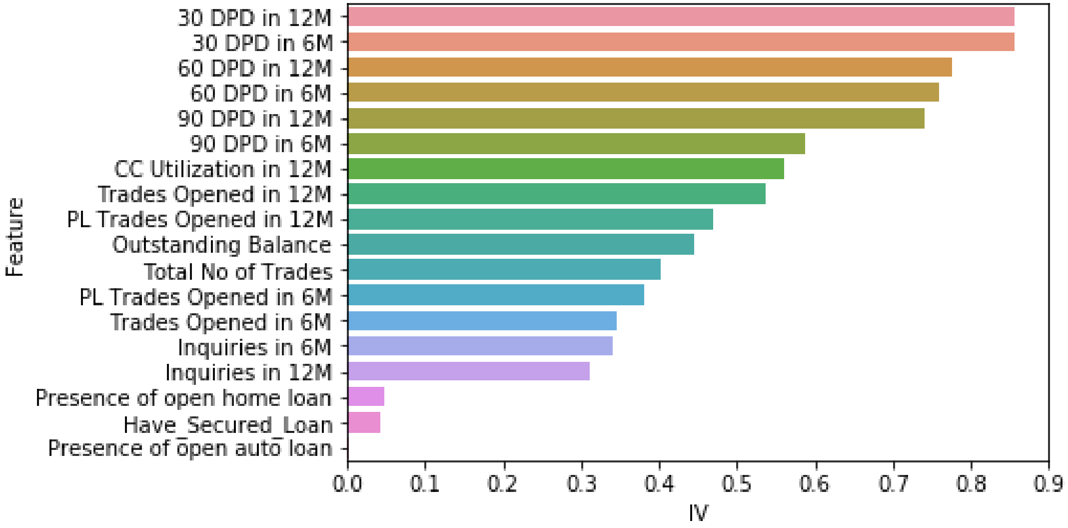
Variables Considered for Analysis Basis WOE and Information Value

Demographic Data



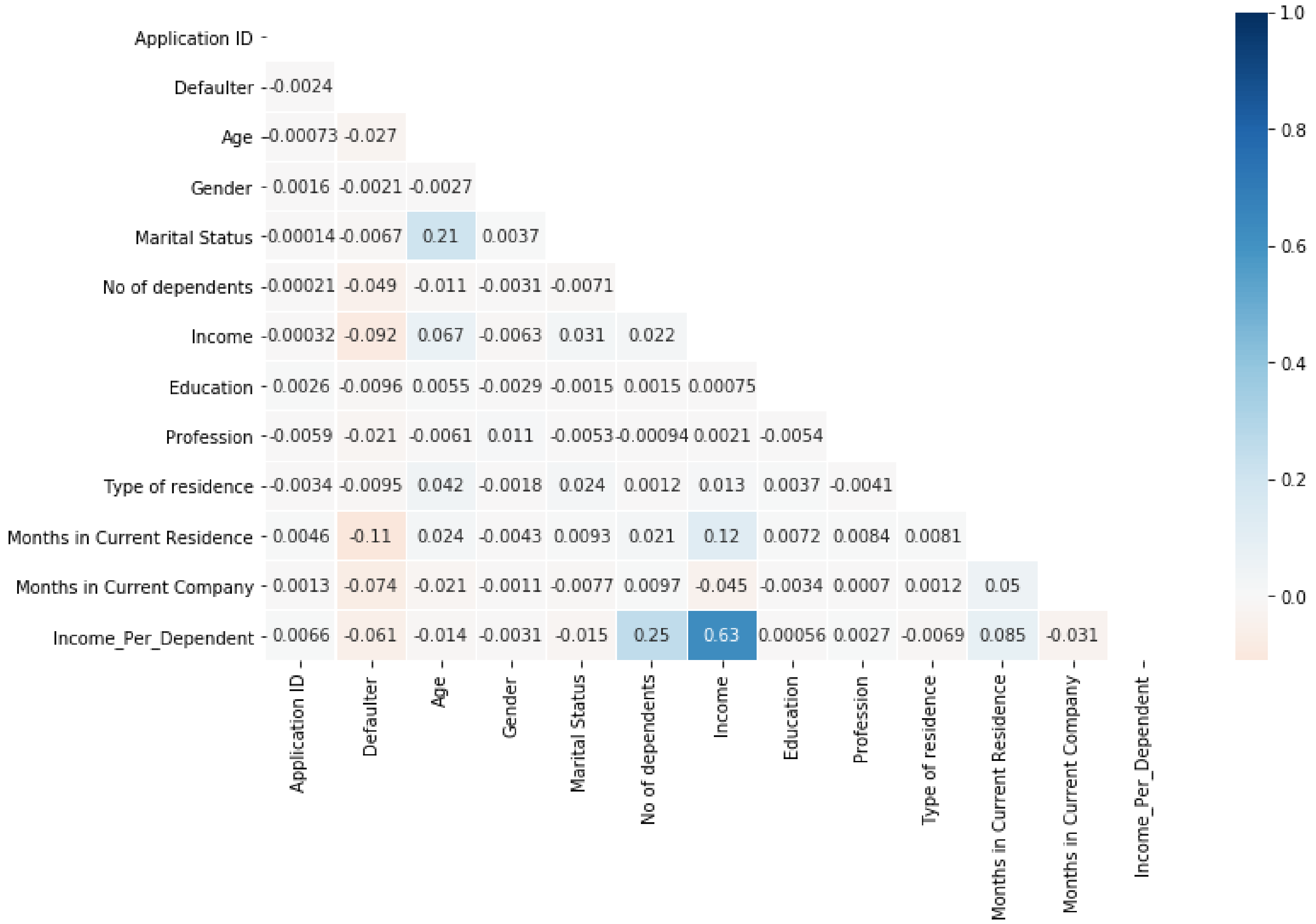
Feature	IV
Months in Current Residence	0.182606
Income	0.139102
Months in Current Company	0.090508
Income_Per_Dependent	0.068326
No of dependents	0.039303
Age	0.012360
Profession	0.007569
Type of residence	0.001889
Education	0.001575
Marital Status	0.000751
Gender	0.000078

Credit Bureau Data



Feature	IV
30 DPD in 12M	0.856121
30 DPD in 6M	0.855173
60 DPD in 12M	0.774534
60 DPD in 6M	0.760158
90 DPD in 12M	0.740845
90 DPD in 6M	0.586412
CC Utilization in 12M	0.561875
Trades Opened in 12M	0.536770
PL Trades Opened in 12M	0.470251
Outstanding Balance	0.444760
Total No of Trades	0.402015
PL Trades Opened in 6M	0.379986
Trades Opened in 6M	0.347077
Inquiries in 6M	0.339339
Inquiries in 12M	0.311907
Presence of open home loan	0.048302
Have_Secured_Loan	0.042299
Presence of open auto loan	0.001875

Correlation using Heatmap for Demographics Data

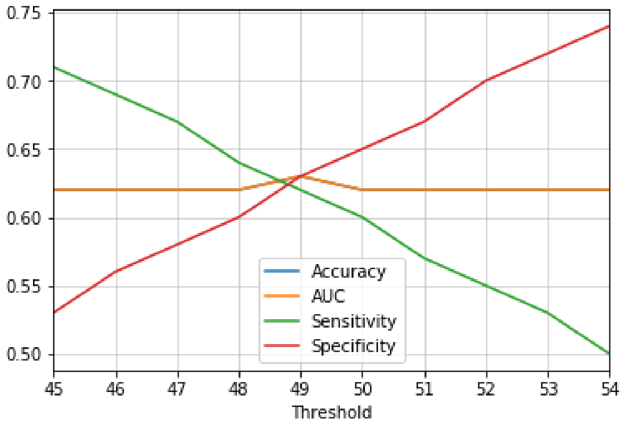


- No high correlation within variables.

Demographic Model: Logistic Regression Performance

Iteration 1: Results

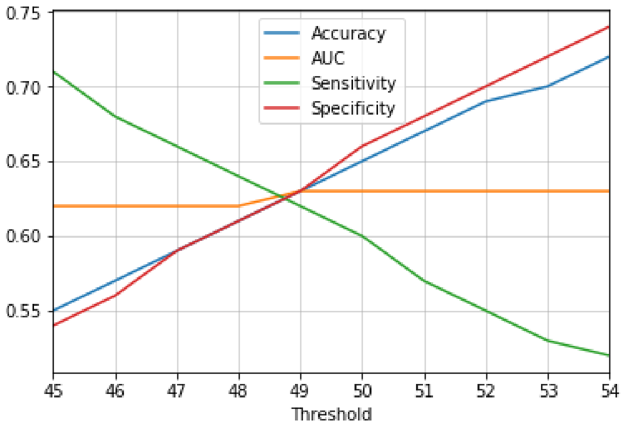
At threshold **49** we have **0.63 & 0.62 as Precision and recall** respectively.



Threshold	Accuracy	Sensitivity	Specificity	AUC
45.0	0.62	0.71	0.53	0.62
46.0	0.62	0.69	0.56	0.62
47.0	0.62	0.67	0.58	0.62
48.0	0.62	0.64	0.60	0.62
49.0	0.63	0.62	0.63	0.63
50.0	0.62	0.60	0.65	0.62
51.0	0.62	0.57	0.67	0.62
52.0	0.62	0.55	0.70	0.62
53.0	0.62	0.53	0.72	0.62
54.0	0.62	0.50	0.74	0.62

Iteration 2 (Final Model): Results

At threshold **49** we have **0.63 & 0.62 as Precision and recall** respectively.



Threshold	Accuracy	Sensitivity	Specificity	AUC
45.0	0.55	0.71	0.54	0.62
46.0	0.57	0.68	0.56	0.62
47.0	0.59	0.66	0.59	0.62
48.0	0.61	0.64	0.61	0.62
49.0	0.63	0.62	0.63	0.63
50.0	0.65	0.60	0.66	0.63
51.0	0.67	0.57	0.68	0.63
52.0	0.69	0.55	0.70	0.63
53.0	0.70	0.53	0.72	0.63
54.0	0.72	0.52	0.74	0.63

Train and Test Scores

Training Data Score

********* 49 ***********

Accuracy : 0.63

Sensitivity / Recall : 0.62

Specificity : 0.63

ROC - AUC : 0.63

Test Data Score

********* 49 ***********

Accuracy : 0.63

Sensitivity / Recall : 0.62

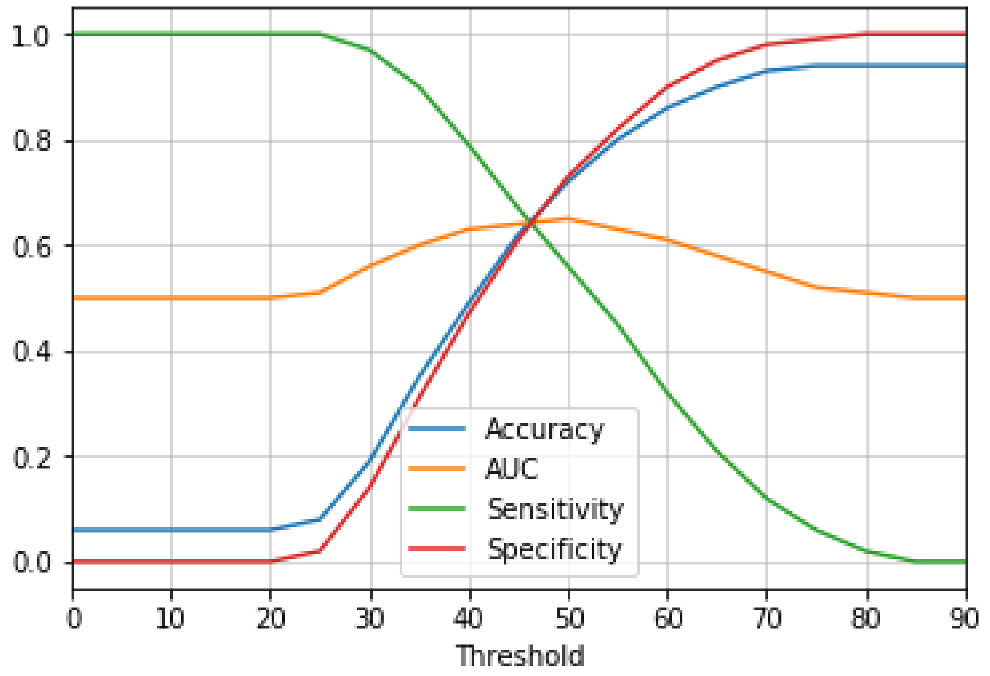
Specificity : 0.63

ROC - AUC : 0.63

Demographic Model: Random Forest Performance

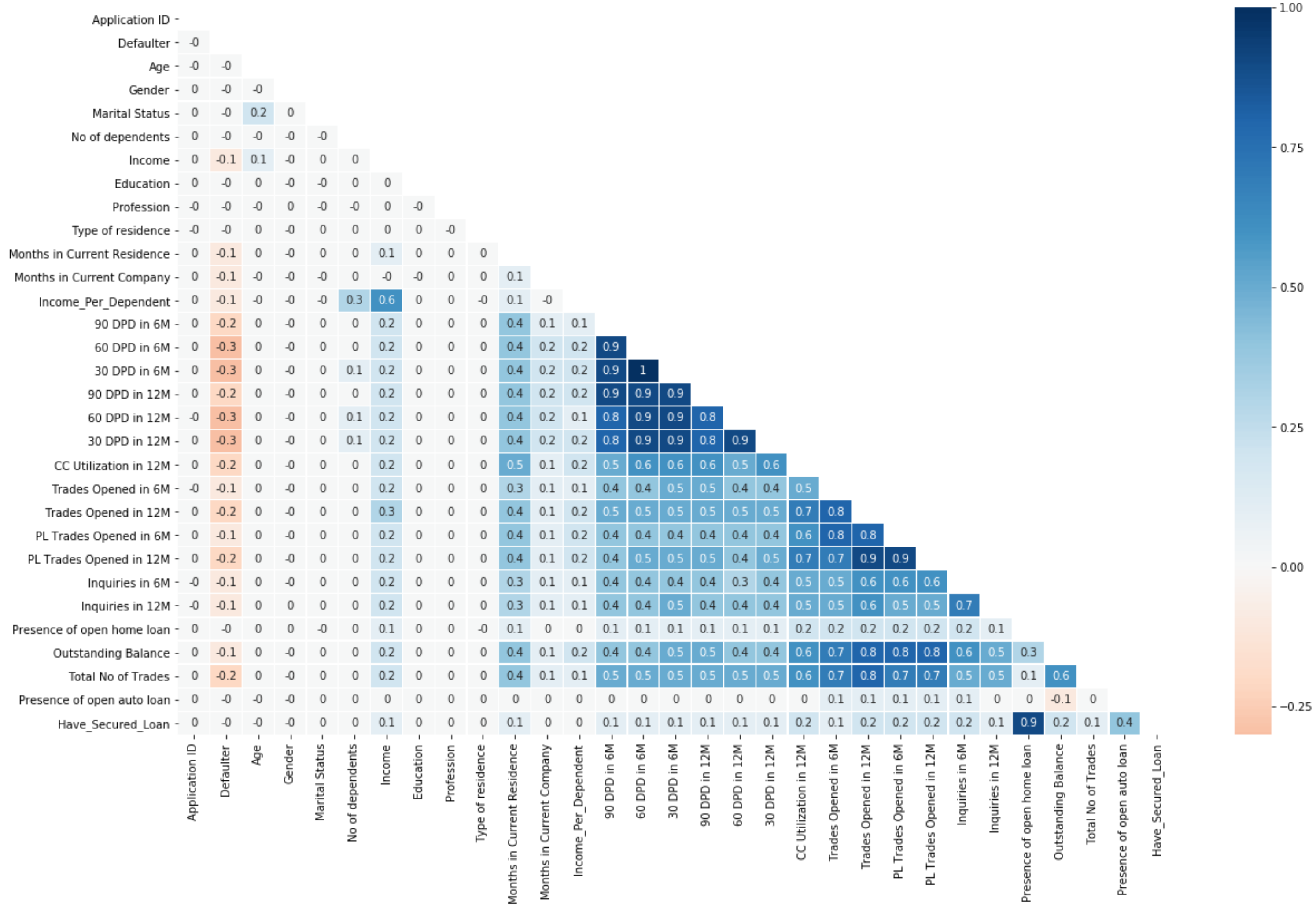
Results

Sensitivity & Specificity of model using demographic data is not good which is 0.63



Threshold	Accuracy	Sensitivity	Specificity	AUC
0.0	0.06	1.00	0.00	0.50
5.0	0.06	1.00	0.00	0.50
10.0	0.06	1.00	0.00	0.50
15.0	0.06	1.00	0.00	0.50
20.0	0.06	1.00	0.00	0.50
25.0	0.08	1.00	0.02	0.51
30.0	0.19	0.97	0.14	0.56
35.0	0.35	0.90	0.31	0.60
40.0	0.49	0.79	0.47	0.63
45.0	0.62	0.67	0.61	0.64
50.0	0.72	0.56	0.73	0.65
55.0	0.80	0.45	0.82	0.63
60.0	0.86	0.32	0.90	0.61
65.0	0.90	0.21	0.95	0.58
70.0	0.93	0.12	0.98	0.55
75.0	0.94	0.06	0.99	0.52
80.0	0.94	0.02	1.00	0.51
85.0	0.94	0.00	1.00	0.50
90.0	0.94	0.00	1.00	0.50

Correlation using Heatmap for Combined Data



- High collinearity between independent variables in data set.
- Will drop variable with more than 80% collinear to each other.

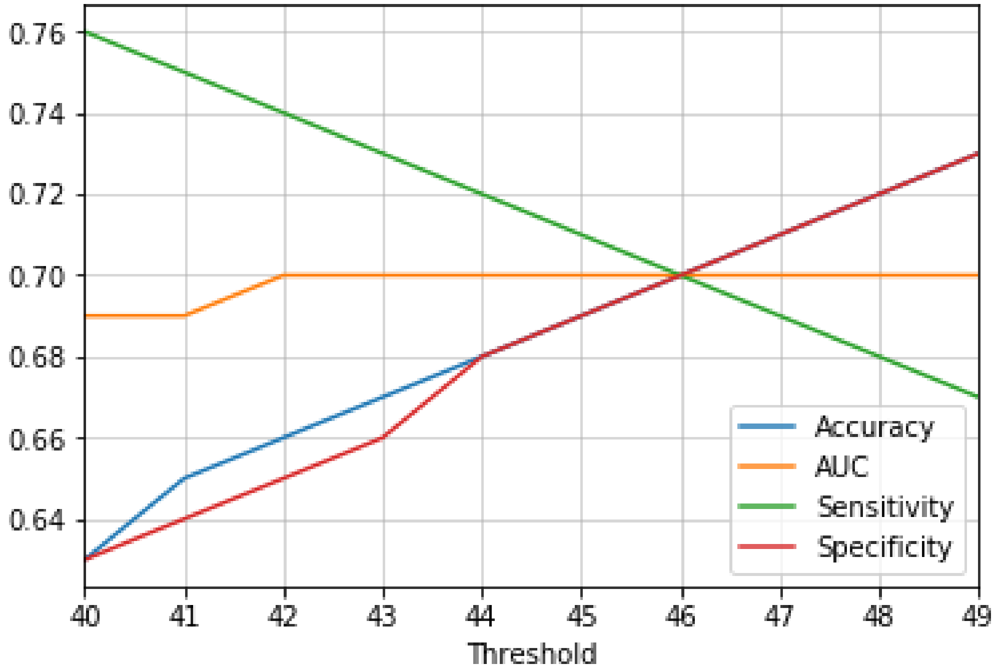
Feature Selection: RFE (List of features selected post RFE)

Features	VIF
Age	1.07
Marital Status	1.05
No of dependents	1.03
Income	1.17
Education	1.00
Profession	1.01
Type of residence	1.00
Months in Current Company	1.09
30 DPD in 12M	1.73
CC Utilization in 12M	2.09
PL Trades Opened in 6M	1.72
Inquiries in 6M	1.56
Presence of open auto loan	1.28
Have_Secured_Loan	1.35

- RFE (Recursive Feature Elimination) technique to select most important features.
- 13 features out of 23 features considered post multiple iteration of model evaluation variable selection

Model Results and Insights: Logistic Regression

- **SMOTE (Synthetic Minority Over-sampling Technique)** : SMOTE is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.
- **Model Output**:



Train and Test Scores

Training Data Score

********* 49 ***********

Accuracy : 0.69

Sensitivity / Recall : 0.71

Specificity : 0.69

ROC - AUC : 0.7

Test Data Score

********* 49 ***********

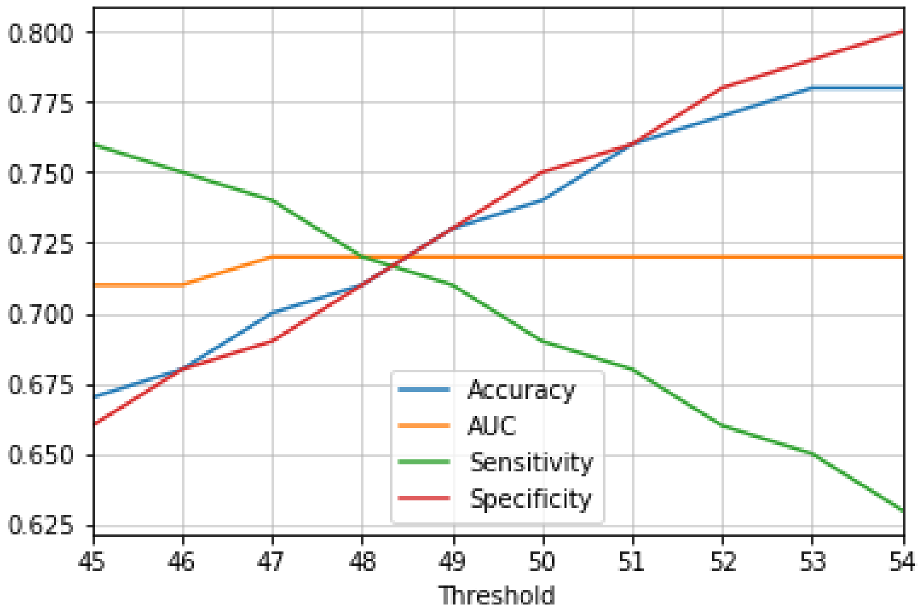
Accuracy : 0.69

Sensitivity / Recall : 0.69

Specificity : 0.69

ROC - AUC : 0.69

Model Results and Insights: Random Forest

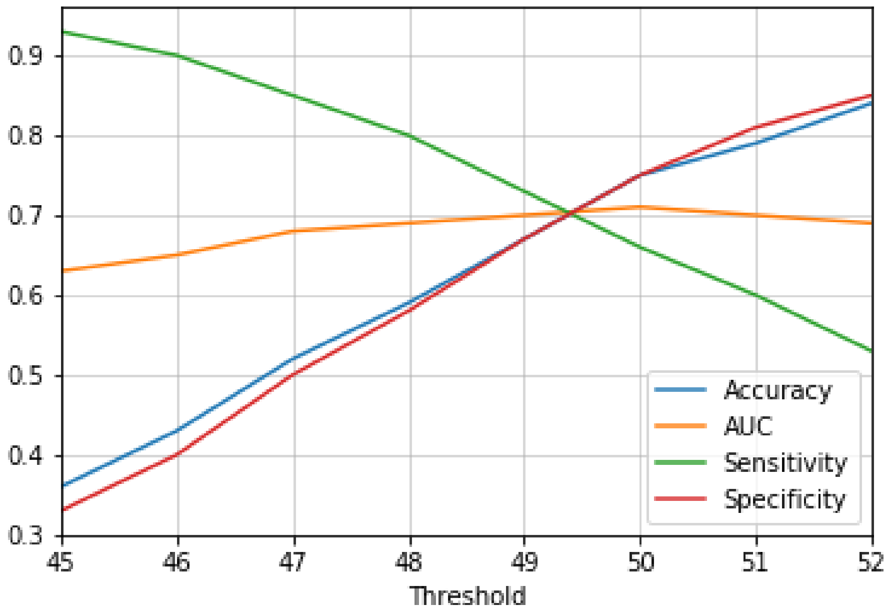


Threshold	Accuracy	Sensitivity	Specificity	AUC
45.0	0.67	0.76	0.66	0.71
46.0	0.68	0.75	0.68	0.71
47.0	0.70	0.74	0.69	0.72
48.0	0.71	0.72	0.71	0.72
49.0	0.73	0.71	0.73	0.72
50.0	0.74	0.69	0.75	0.72
51.0	0.76	0.68	0.76	0.72
52.0	0.77	0.66	0.78	0.72
53.0	0.78	0.65	0.79	0.72
54.0	0.78	0.63	0.80	0.72

Train and Test Scores

Training Data Score	Test Data Score
********* 48 ***********	********* 48 ***********
Accuracy : 0.71	Accuracy : 0.71
Sensitivity / Recall : 0.7	Sensitivity / Recall : 0.67
Specificity : 0.71	Specificity : 0.72
ROC - AUC : 0.71	ROC - AUC : 0.69

Model Results and Insights: AdaBoost Classifier



Threshold	Accuracy	Sensitivity	Specificity	AUC
45.0	0.36	0.93	0.33	0.63
46.0	0.43	0.90	0.40	0.65
47.0	0.52	0.85	0.50	0.68
48.0	0.59	0.80	0.58	0.69
49.0	0.67	0.73	0.67	0.70
50.0	0.75	0.66	0.75	0.71
51.0	0.79	0.60	0.81	0.70
52.0	0.84	0.53	0.85	0.69

Train and Test Scores

Training Data Score

********* 49 ***********

Accuracy : 0.67

Sensitivity / Recall : 0.73

Specificity : 0.67

ROC - AUC : 0.7

Test Data Score

********* 49 ***********

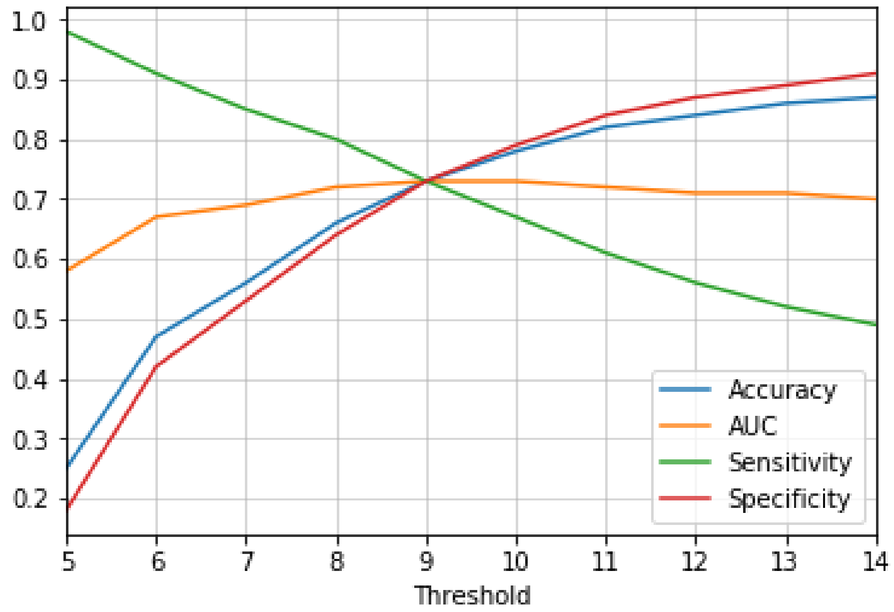
Accuracy : 0.68

Sensitivity / Recall : 0.71

Specificity : 0.67

ROC - AUC : 0.69

Model Results and Insights: Gradient Boosting Classifier



Threshold	Accuracy	Sensitivity	Specificity	AUC
5.0	0.25	0.98	0.18	0.58
6.0	0.47	0.91	0.42	0.67
7.0	0.56	0.85	0.53	0.69
8.0	0.66	0.80	0.64	0.72
9.0	0.73	0.73	0.73	0.73
10.0	0.78	0.67	0.79	0.73
11.0	0.82	0.61	0.84	0.72
12.0	0.84	0.56	0.87	0.71
13.0	0.86	0.52	0.89	0.71
14.0	0.87	0.49	0.91	0.70

Train and Test Scores

Training Data Score

********* 9 ***********

Accuracy : 0.72

Sensitivity / Recall : 0.69

Specificity : 0.73

ROC - AUC : 0.71

Test Data Score

********* 9 ***********

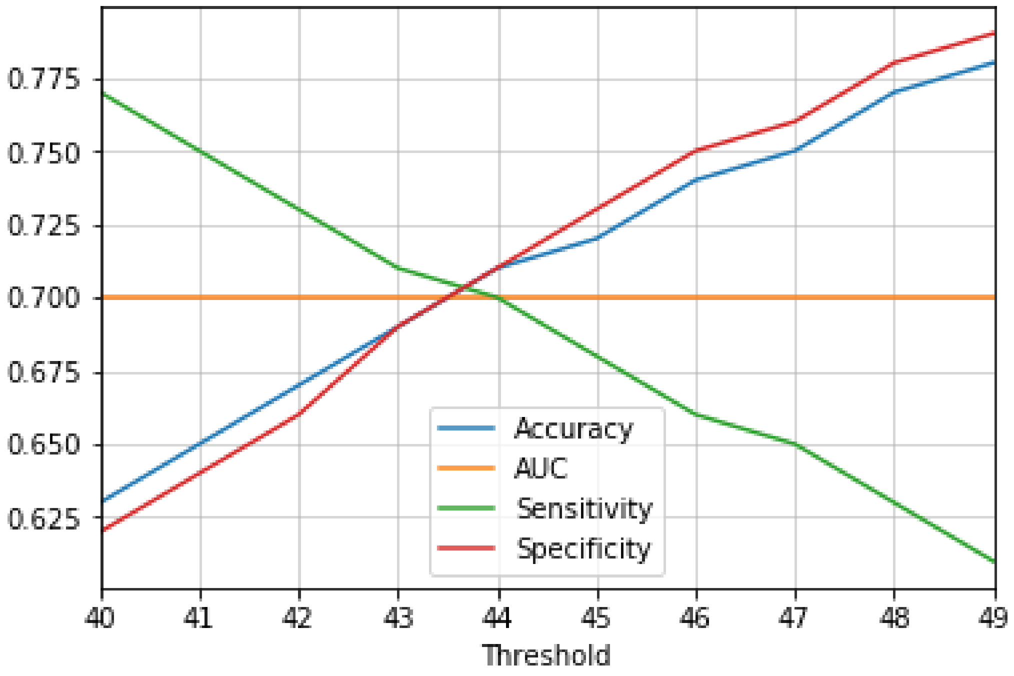
Accuracy : 0.73

Sensitivity / Recall : 0.66

Specificity : 0.73

ROC - AUC : 0.7

Model Results and Insights: Light GBM



Train and Test Scores

Training Data Score

********* 43 ***********

Accuracy : 0.69

Sensitivity / Recall : 0.71

Specificity : 0.69

ROC - AUC : 0.7

Test Data Score

********* 49 ***********

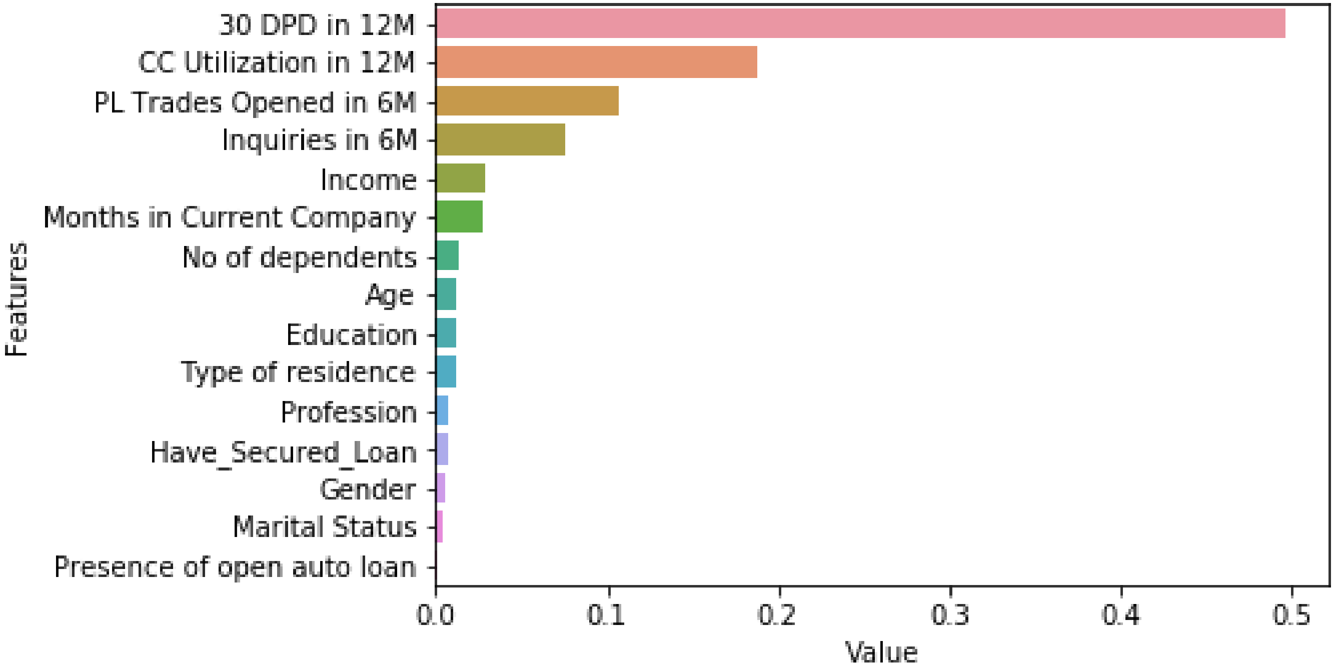
Accuracy : 0.69

Sensitivity / Recall : 0.69

Specificity : 0.69

ROC - AUC : 0.69

Feature Importance as per Random Forest Model

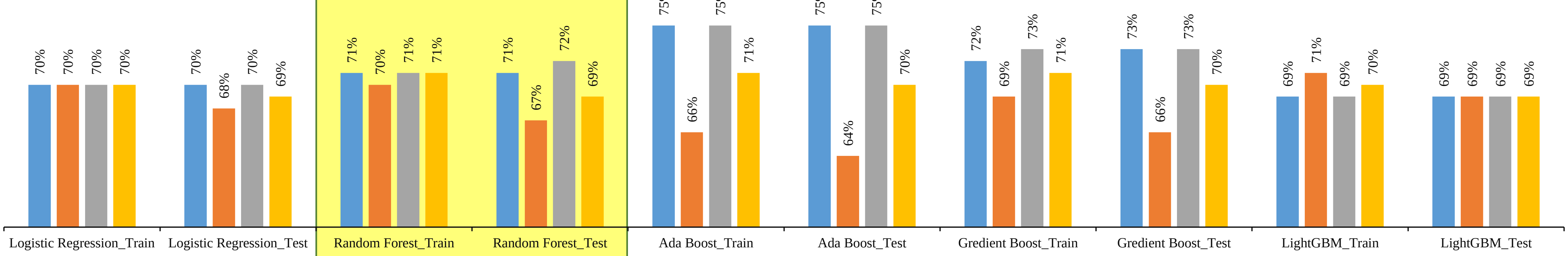


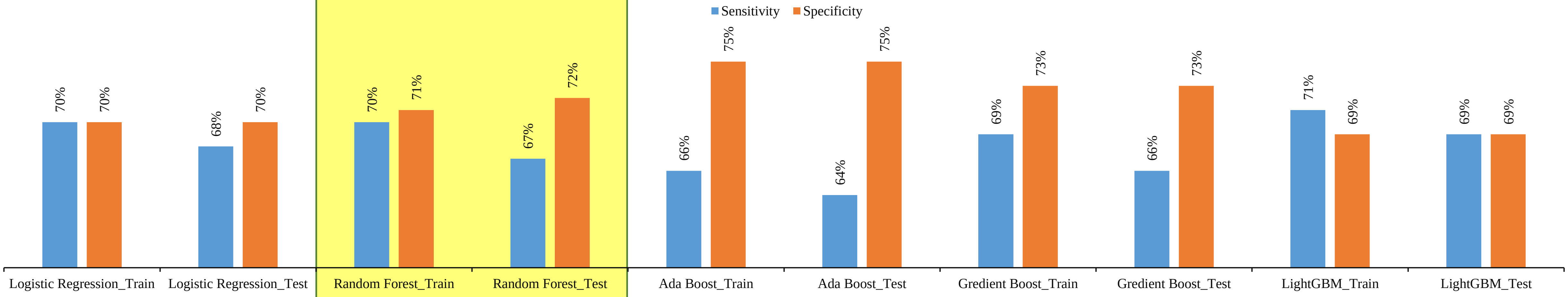
Top 3 features most important used to predict defaulters:

- 30 DPD in 12 Months
- CC utilization in 12 Months
- PL Trades Opened in 6 Months

Best Model Selection

Random Forest is best model for prediction





Application Score Card

- The give Points Double Over is 20, Base Score is 400 and Odds is 10

$$\text{Score} = \sum_{i=1}^{n}\left(-\left(woei * \beta_i + \frac{a}{n}\right) * factor + offset/n\right)$$

$$\text{Factor} = \text{points to double}/Ln(2)$$

$$\text{Offset} = \text{score } - (factor * \ln(odds))$$

- Threshold score is **335** . Score Lower than this would mean beyond this score the customers probability of default would increase to 70%
- Implementing this scorecard means **70%** of the default customer will not be given credit cards below the threshold
- Total customers rejected basis the scorecard: **21,345**
- Correctly defaulter customer identified by the model: **70%**

Odds	Score
10	400
20	420
40	440

Max Score for the Defaulter in 3rd Decile is 335 hence we have considered it as the threshold

Decile	Total	Defaulter	Min_Score	Max_Score
1	7115	1744.0	270.0	309.0
2	7115	738.0	309.0	325.0
3	7115	500.0	325.0	335.0
4	7115	338.0	335.0	342.0
5	7115	320.0	342.0	349.0
6	7115	231.0	349.0	357.0
7	7115	200.0	357.0	364.0
8	7115	124.0	364.0	372.0
9	7115	93.0	372.0	377.0
10	7115	79.0	377.0	391.0

Credit Score Card:

- **Score <= 335 : High Risk Customer (1 - 3 Decile)**
- **Score Between 336 - 363 : Medium Risk Customer (4 - 7 Decile)**
- **Score >= 364 : Low Risk Customer (8 - 10 Decile)**

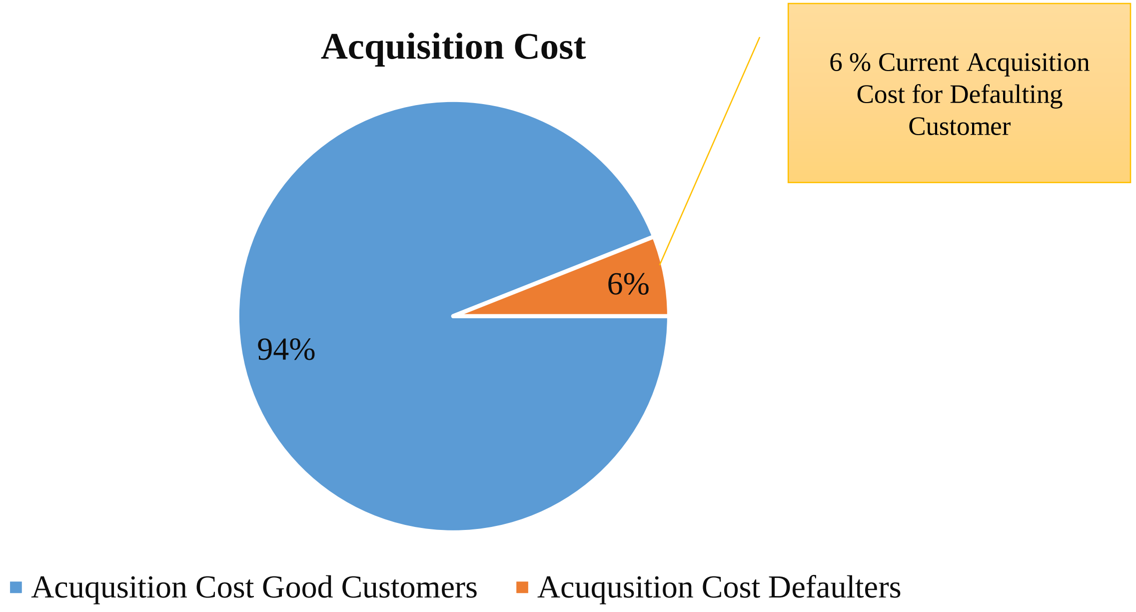
Credit Loss: Impact Analysis of the Score Card

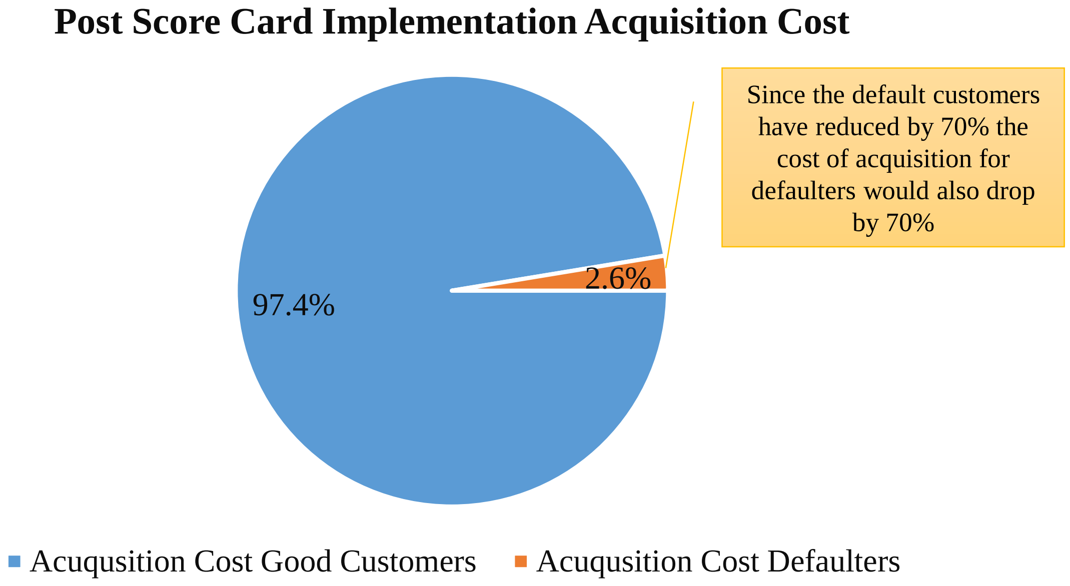
There could be incremental saving in the acquisition cost by targeting the right set of customer

Assuming Current Acquisition Cost is: **100%**

Post Implementing Scorecard Acquisition Cost would Reduce to: **70%**

Net Saving of: 30%





Credit Loss: Impact Analysis of the Model

- Current rate of default customer is at **6%**
- Post model implementation of the Model the default rate is suppose to drop to **2.6%**

Current Credit Loss at Risk: $ 5,189 Mn.

(6% of total Outstanding)

(total outstanding balance of the customer who are marked as defaulter)

- Post model implementation 70% default customer would not be granted credit card.
- Since there is a decrease of 30% in the over all population the amount of credit loss would be $ 1,625 Mn. (2.6% of $61,945 Mn.)
- If similar profile of another 30% customers are acquired then the Credit Loss would be $2,594 Mn.

Net Savings in Credit Loss: $ 2,594 Mn.

(2.6% of total Outstanding)

(total outstanding balance of the customer who are marked as defaulter)