

BFS CAPSTONE

FINAL SUBMISSION

NILANJAN PORIA
MADHURA KELKAR
SAIRAM
ADITYA MENON

OBJECTIVE & APPROACH

Problem Statement:

- CredX is a leading credit card provider that gets thousands of credit card applicants every year. But in the past few years, it has experienced an increase in credit loss due to increase in defaults.
- 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. We need to determine the factors affecting credit risk and create strategies to mitigate the acquisition risk and assess the financial benefit of your project. Build an application scorecard and identify the cut-off score below which one would not grant credit cards to applicants.

Solution Approach:

This is a binary supervised classification problem. We aim at building models such as Logistic regression, Random forest, SVM to identify the customers who are at a risk of defaulting if offered a credit card. We have followed CRISP-DM framework. It involves the following series of steps:

- Business Understanding and Data Understanding
- Data Cleansing and Preparation
- Exploratory Data Analysis (Graphs & plots)
- Data Transformation and Model Building
- Model Evaluation
- Application Score card
- Assessing Financial benefit of the model

DATA UNDERSTANDING

Demographic Data :

- This information is provided by the applicants at the time of credit card application.
- It contains customer-level information on age, gender, income, marital status, education, Profession and number of dependants.

Credit Bureau Data :

- It provides all details of past transaction.
- This information is taken from credit bureau and contains variables such as 'number of times 30 DPD or worse in last 3/6/12 months', 'outstanding balance', 'number of trades', etc.

Nature of Data :

- The demographic data consists of 71295 observations with 12 variables.The credit bureau data consists of 71295 observations with 19 variables.
- Application ID is the common key between the two datasets for merging.
- Performance Tag is the target variable which says if customer is default or not.The values are 0(non-default) and 1(default).

Data Quality Issue :

- The 1425 rows with no performance tag.Thus we can assume that the applicant is not given credit card, hence they are removed.
- Both occurrences of 3 duplicate Application ID records (765011468, 653287861, 671989187) has been excluded from the dataset.
- Since 18 is the minimum age to grant credit card, records with age <18 has been excluded from the dataset.
- The 1425 rejected records have been saved separately and would be used later to predict if they would default if they were given a credit card using a model made from non rejected records for making a model with better performance and application score card calculations.

WOE AND IV ANALYSIS

- Creating Infotables plot to obtain the IV for all variables after merging the datasets. Based on the IV values given in the table below, classified the variables as Strong/Weak predictor

Information Value (IV)	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	weak predictor
0.1 to 0.3	medium predictor
0.3 to 0.5	strong predictor
> 0.5	suspicious or too good to be true

NULL VALUE IMPUTATION

```
> #---Education
> plot_infotables(IV,"Education")
> print(IV$Tables$Education,row.names = FALSE)
```

Education	N	Percent	WOE	IV
<NA>	118	0.001692144	-0.003479673	2.052136e-08
Bachelor	17279	0.247784438	-0.016014162	6.403358e-05
Masters	23440	0.336134454	-0.008493373	8.837594e-05
Others	117	0.001677804	-0.509523239	6.409352e-04
Phd	4456	0.063899963	0.029155704	6.945345e-04
Professional	24324	0.348811197	0.017674128	8.026171e-04

- Consider Education variable as shown above:
- Null value imputation:**
 - Check the WOE value for NULL bin
 - Choose the closest bin having woe close to -0.003479673(NULL bin)
 - Impute Masters bin woe value in the NULL bin

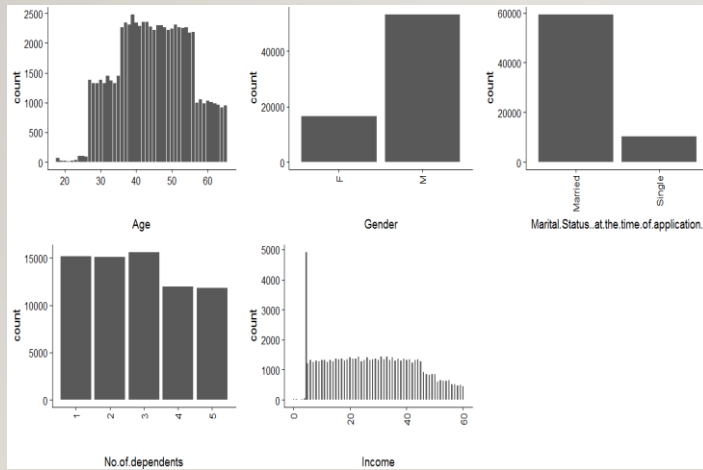
After Classifying based on IV, we observe that most of the Demographic variables are “Not Useful”, while Credit Information are “Strong” or “Medium” predictors

Variable	IV	Var_Pred
Avgas.CC.Utilization.in.last.12.months	3.104284e-01	Strong predictive Power
No.of.trades.opened.in.last.12.months	2.987236e-01	Medium predictive Power
No.of.PL.trades.opened.in.last.12.months	2.962713e-01	Medium predictive Power
No.of.Inquiries.in.last.12.months..excluding.home...a...	2.959117e-01	Medium predictive Power
Outstanding.Balance	2.453978e-01	Medium predictive Power
No.of.times.30.DPD.or.worse.in.last.6.months	2.417547e-01	Medium predictive Power
Total.No.of.Trades	2.372198e-01	Medium predictive Power
No.of.PL.trades.opened.in.last.6.months	2.198333e-01	Medium predictive Power
No.of.times.90.DPD.or.worse.in.last.12.months	2.140054e-01	Medium predictive Power
No.of.times.60.DPD.or.worse.in.last.6.months	2.060027e-01	Medium predictive Power
No.of.Inquiries.in.last.6.months..excluding.home...aut...	2.053275e-01	Medium predictive Power
No.of.times.30.DPD.or.worse.in.last.12.months	1.984277e-01	Medium predictive Power
No.of.trades.opened.in.last.6.months	1.862998e-01	Medium predictive Power
No.of.times.60.DPD.or.worse.in.last.12.months	1.856382e-01	Medium predictive Power
No.of.times.90.DPD.or.worse.in.last.6.months	1.602322e-01	Medium predictive Power
No.of.months.in.current.residence	7.915670e-02	Weak predictive Power
Income	4.286994e-02	Weak predictive Power
No.of.months.in.current.company	2.160420e-02	Weak predictive Power
Presence.of.open.home.loan	1.737488e-02	Not useful for prediction
Age	3.404508e-03	Not useful for prediction
No.of.dependents	2.648578e-03	Not useful for prediction
Profession	2.158332e-03	Not useful for prediction
Presence.of.open.auto.loan	1.630187e-03	Not useful for prediction
Application.ID	1.531334e-03	Not useful for prediction
Type.of.residence	9.370664e-04	Not useful for prediction
Education	8.026171e-04	Not useful for prediction
Gender	3.341938e-04	Not useful for prediction
Marital.Status..at.the.time.of.application.	9.833636e-05	Not useful for prediction

EDA - UNIVARIATE ANALYSIS: DEMOGRAPHICS

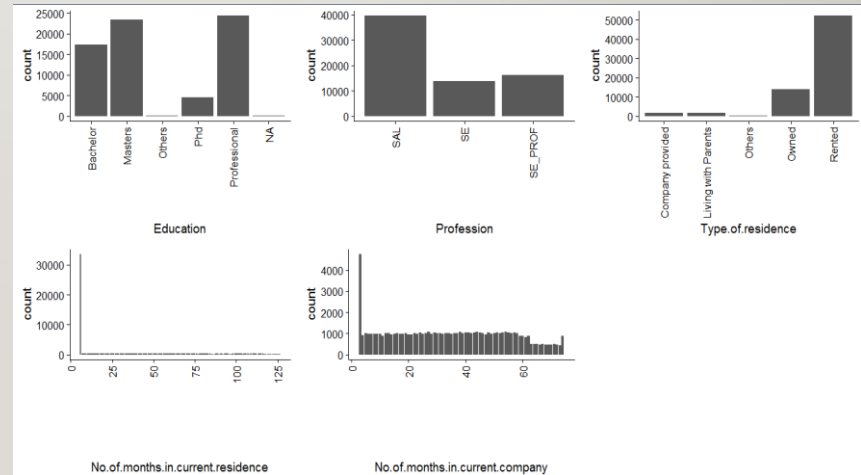
Following important observations can be made from the frequency plots of the demographic variables:

- Gender: Majority of the applicants are males
- Age: Majority of the applicants are in age range 25-55
- Marital Status: Majority of the applicants are Married



Following important observations can be made from the frequency plots of the demographic variables:

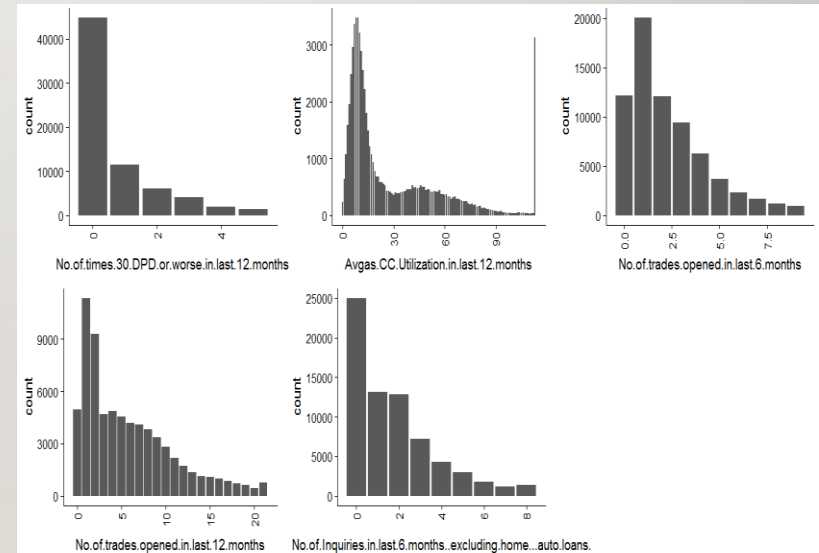
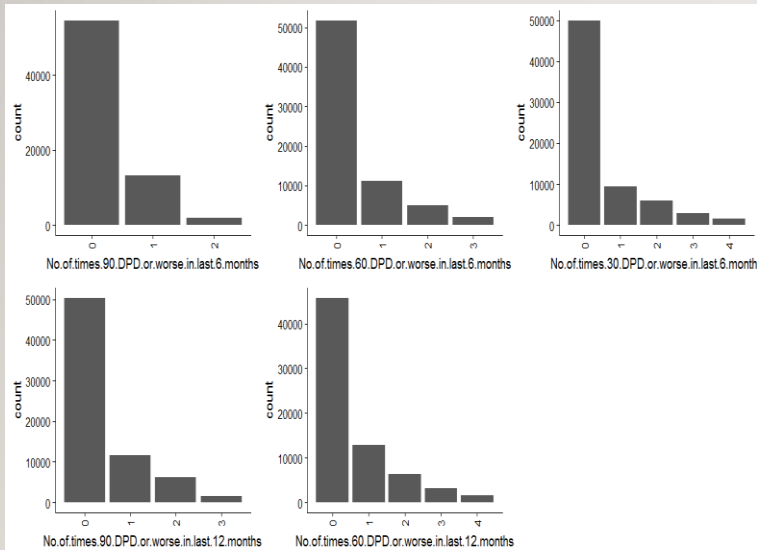
- Education: Majority of the applicants are Professional/Masters
- Profession: Majority are salaried
- Residence: Majority have rented accommodation



EDA - UNIVARIATE ANALYSIS : CREDIT

- For most of the applicants number of times DPD in last 6/12 months is : 0

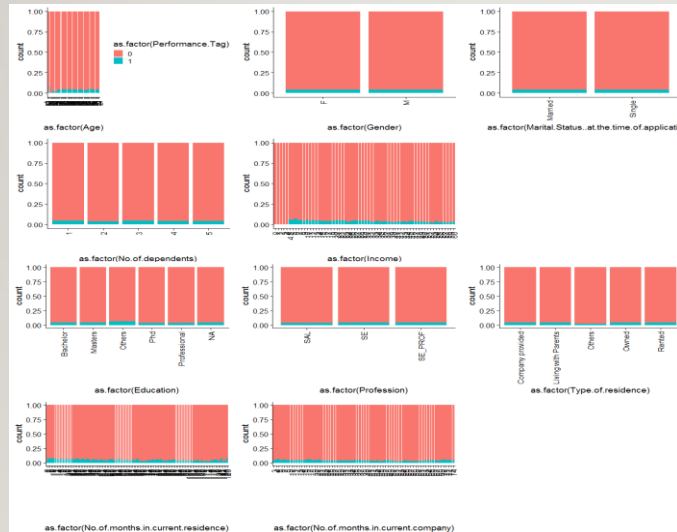
- For most of the applicants, no enquiries were made (value =0)
- For most of the applicants, No of trades opened = 1
- For most of the applicants, avg credit Utilization ranges from 5-20



EDA – BIVARIATE ANALYSIS

Following inferences can be derived from the Bivariate analysis:

- Gender: Approx. 10% of defaulters are seen per gender
- Education: Defaulters in Others category is more than rest of the educational buckets
- No. of dependents: With increase in dependents the defaulter ratio increases

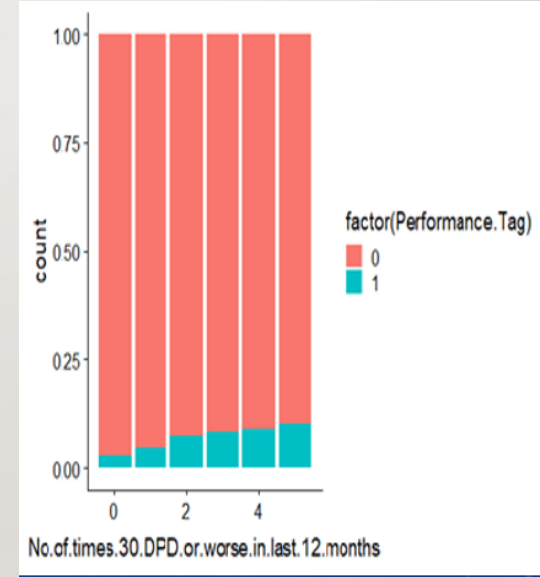
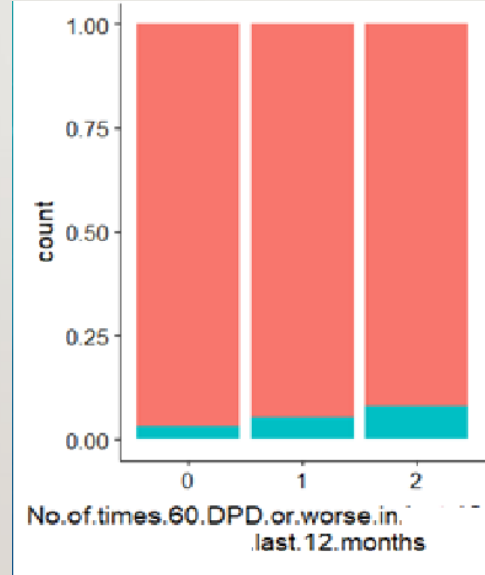
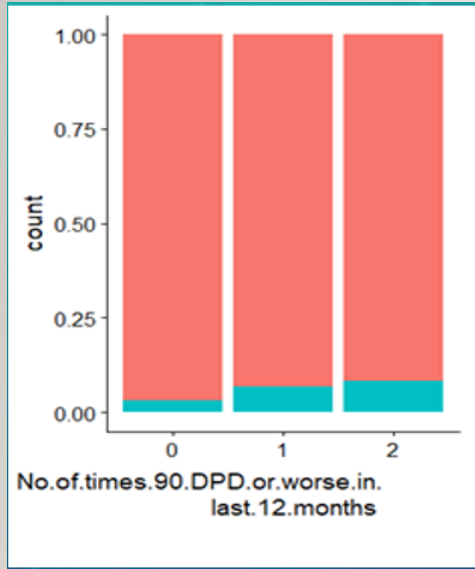


Following inferences can be derived from the Bivariate analysis:

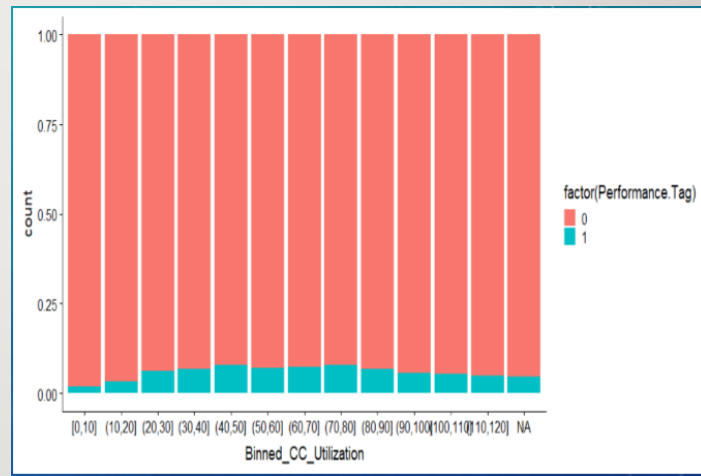
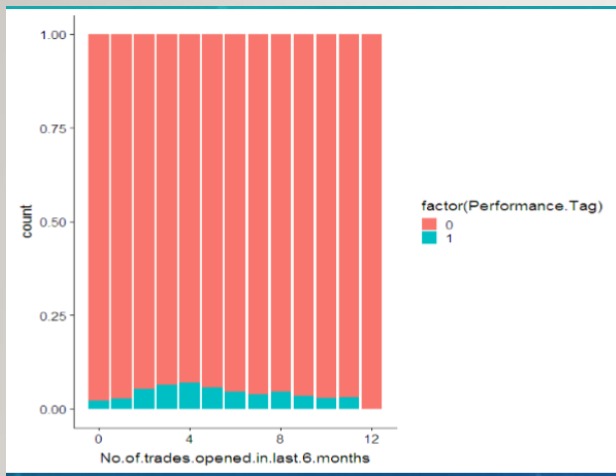
- No of trades opened in last 12 months: Maximum defaulters fall under 5-15 bucket
- No of times 90 DPD or worse in last 6 months: With increase in delinquency the defaulter ratio increases



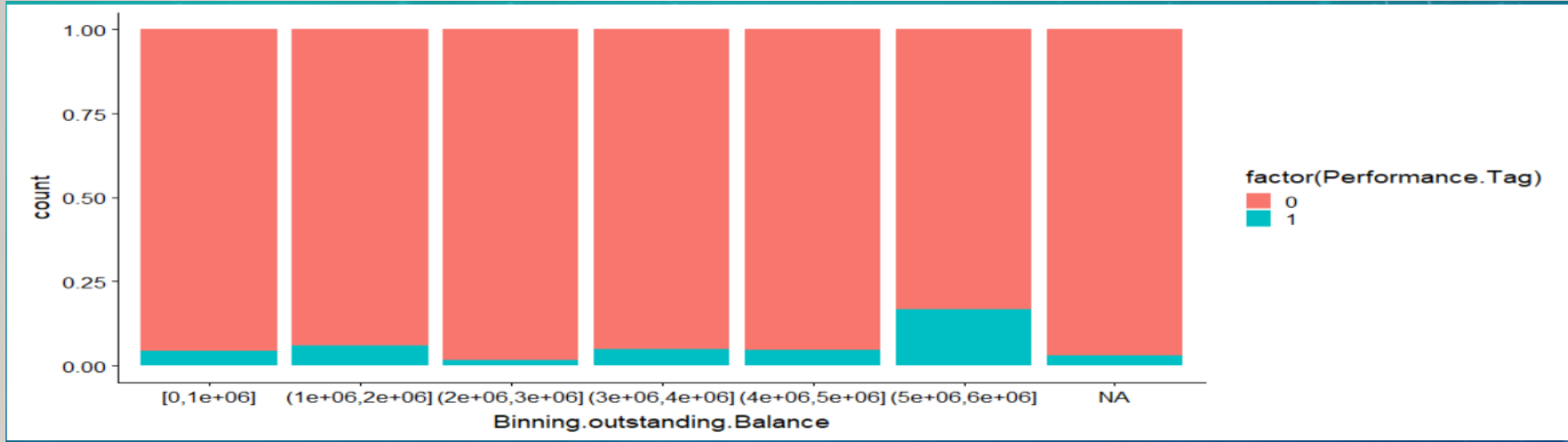
Insights derived



Percentage of defaulters is increasing with increase in Number of 30/60/90 DPD or worse in last 12 months variable values. Hence these variables can be important predictors.



- Percentage of defaulters is lower among customers with Average Credit card utilization between 0 to 20.
- Applicants who opened trades four times in the last six months tend to default more.



Plot for bivariate analysis of Outstanding balance binned in segments of Rs.10 Lac
Outstanding balance field shows higher percentage of defaulters in 50L-60L bin compared to lower outstanding balance bins. This can be an important predictor of default.

CORRELATION MATRIX

Following we could infer from Correlation Matrix:

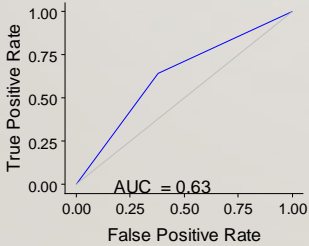
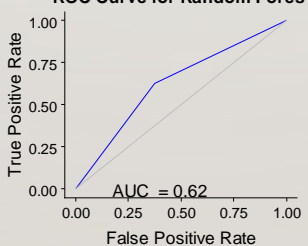
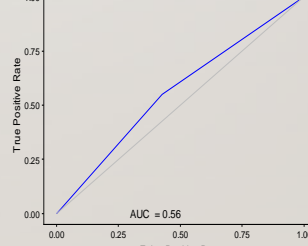
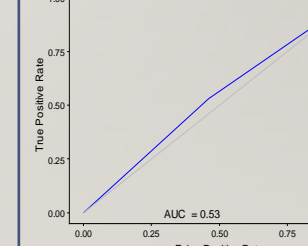
- Two highly correlated “groups of variables” observed :
- No of delinquency in last 6 or 12 months: This shows customer with 60 DPD is likely to move to 90 DPD
- No of trades opened in last 6 or 12 months and Total no of trades

	Age	No of dependents	Income	No of months in current residence	No of months in current company	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	Outstanding Balance	Total No of Trades
Age		0.17	0.06	-0.07	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	0.02	0.02	0.01	0.01	0.01	0.02	-0.01	0.02
No of dependents	0.17		0.03	-0.01	-0.01	0	0	0	0	0	0	-0.01	0	0	0	-0.01	-0.01	-0.01	0	0
Income	0.06	0.03		-0.09	-0.1	-0.19	-0.21	-0.22	-0.21	-0.2	-0.21	-0.18	-0.14	-0.16	-0.18	-0.2	-0.15	-0.16	0.01	-0.11
No of months in current residence	-0.07	-0.01	-0.09		-0.08	0.15	0.15	0.15	0.17	0.19	0.19	0.27	-0.05	-0.04	0.04	0.04	-0.07	-0.08	-0.01	-0.1
No of months in current company	-0.02	-0.01	-0.1	-0.08		-0.1	-0.12	-0.12	-0.12	-0.12	-0.13	-0.04	0.02	0.01	-0.01	-0.01	0.02	0.03	0	0.05
No of times 90 DPD or worse in last 6 months	-0.02	0	-0.19	0.15	-0.1		0.89	0.84	0.89	0.81	0.79	0.36	0.15	0.18	0.25	0.28	0.15	0.16	-0.03	0.04
No of times 60 DPD or worse in last 6 months	-0.02	0	-0.21	0.15	-0.12	0.89		0.95	0.84	0.92	0.9	0.36	0.16	0.2	0.28	0.31	0.17	0.18	-0.04	0.05
No of times 30 DPD or worse in last 6 months	-0.02	0	-0.22	0.15	-0.12	0.84	0.95		0.83	0.9	0.95	0.36	0.17	0.21	0.29	0.32	0.18	0.19	-0.04	0.05
No of times 90 DPD or worse in last 12 months	-0.02	0	-0.21	0.17	-0.12	0.89	0.84	0.83		0.81	0.8	0.39	0.16	0.21	0.28	0.32	0.17	0.19	-0.04	0.05
No of times 60 DPD or worse in last 12 months	-0.02	0	-0.2	0.19	-0.12	0.81	0.92	0.9	0.81		0.9	0.35	0.14	0.17	0.26	0.28	0.14	0.15	-0.04	0.02
No of times 30 DPD or worse in last 12 months	-0.02	0	-0.21	0.19	-0.13	0.79	0.9	0.95	0.8	0.9		0.35	0.15	0.19	0.27	0.3	0.15	0.17	-0.04	0.02
Avgas CC Utilization in last 12 months	-0.01	-0.01	-0.18	0.27	-0.04	0.36	0.36	0.36	0.39	0.35	0.35		0.1	0.14	0.24	0.27	0.09	0.1	-0.03	0
No of trades opened in last 6 months	0.02	0	-0.14	-0.05	0.02	0.15	0.16	0.17	0.16	0.14	0.15	0.1		0.94	0.88	0.84	0.68	0.74	0.08	0.89
No of trades opened in last 12 months	0.02	0	-0.16	-0.04	0.01	0.18	0.2	0.21	0.21	0.17	0.19	0.14	0.94		0.89	0.93	0.73	0.79	0.09	0.94
No of PL trades opened in last 6 months	0.01	0	-0.18	0.04	-0.01	0.25	0.28	0.29	0.28	0.26	0.27	0.24	0.88	0.89		0.9	0.6	0.66	0.09	0.79
No of PL trades opened in last 12 months	0.01	-0.01	-0.2	0.04	-0.01	0.28	0.31	0.32	0.32	0.28	0.3	0.27	0.84	0.93	0.9		0.67	0.73	0.1	0.84
No of inquiries in last 6 months excluding home auto loans	0.01	-0.01	-0.15	-0.07	0.02	0.15	0.17	0.18	0.17	0.14	0.15	0.09	0.68	0.73	0.6	0.67		0.91	0.05	0.72
No of inquiries in last 12 months excluding home auto loans	0.02	-0.01	-0.16	-0.08	0.03	0.16	0.18	0.19	0.19	0.15	0.17	0.1	0.74	0.79	0.66	0.73	0.91		0.06	0.79
Outstanding Balance	-0.01	0	0.01	-0.01	0	-0.03	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	0.08	0.09	0.09	0.1	0.05	0.06		0.09
Total No of Trades	0.02	0	-0.11	-0.1	0.05	0.04	0.05	0.05	0.05	0.02	0.02	0	0.89	0.94	0.79	0.84	0.72	0.79	0.09	

MODEL BUILDING APPROACH

- **OUTLIER TREATMENT:** Outlier detection is done using boxplot on continuous variables and quantiles function and the variables with outliers has been corrected by capping the outliers to the nearest non-outlier values.
- **DATA SCALING:** Scaling is performed for all variables except Application ID and performance tag to standardize the data into common scale.
- **DATA SPLIT:** The final dataset is split into Train and Test in 70:30 ratio for model building. • All models are trained on training datasets and regularization was done by tuning of hyper parameters with cross validation on validation datasets. • All the models are tested on test datasets that were kept separate from training and validation datasets.
- The cut off value for the probability of default was chosen such that model evaluation metrics like accuracy ,sensitivity and specificity were almost equal to each other.

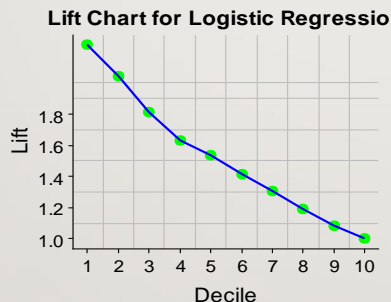
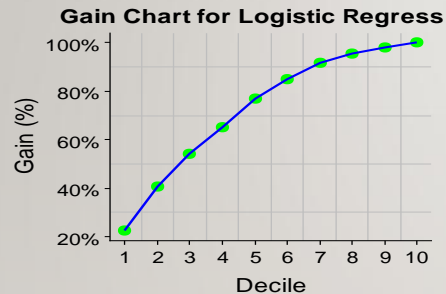
MODEL TYPES - COMPARISON OF PARAMETERS

	Logistic Regression on Merged Data	Random Forest on Merged Data	Logistic Regression on Demographic Data	Random Forest on Demographic Data
Accuracy	62.34%	62.47%	57.26%	52.99%
Sensitivity	64.32 %	62.48%	54.89 %	53.97 %
Specificity	62.25 %	62.46%	57.36 %	52.94 %
KS-statistic	26.58 %	24.95%	12.25%	6.9%
ROC Curve	<p>63 %</p> <p>ROC Curve for Logistic Regression</p>  <p>AUC = 0.63</p>	<p>62.47%</p> <p>ROC Curve for Random Forest</p>  <p>AUC = 0.62</p>	<p>56.12%</p> <p>ROC Curve for Logistic Regression</p>  <p>AUC = 0.56</p>	<p>53.45%</p> <p>ROC Curve for Random Forest</p>  <p>AUC = 0.53</p>

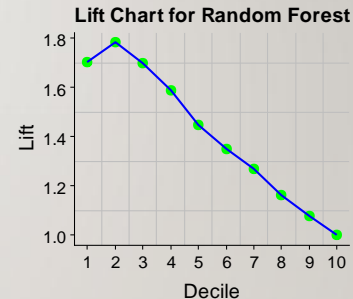
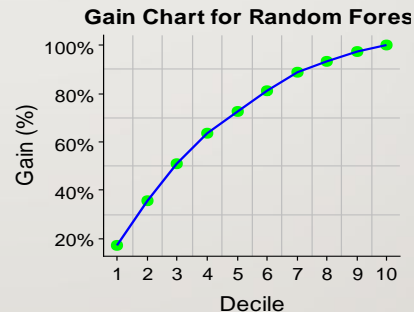
Based on the parameters “Logistic Regression model on Merged Data” is chosen as final model.

LIFT AND GAIN CHARTS

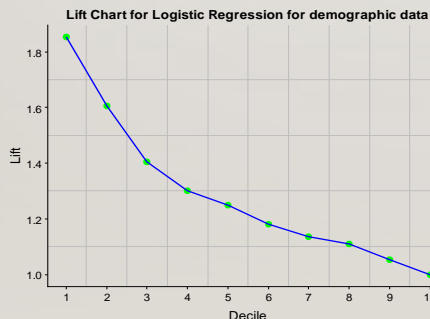
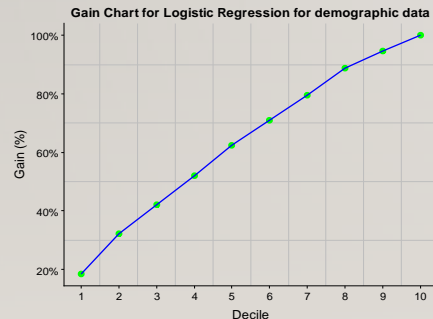
Logistic Regression on Merged Data



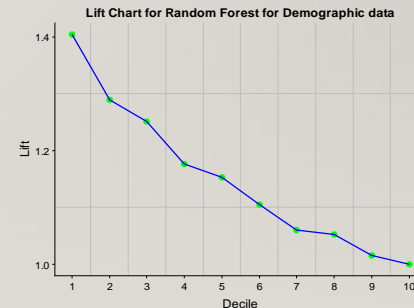
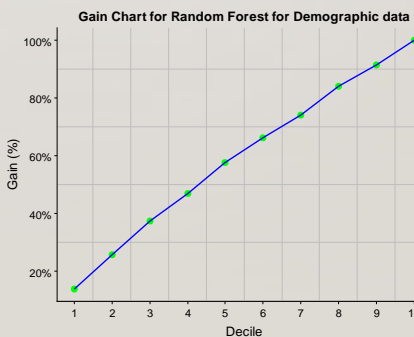
Random Forest on Merged Data



Logistic Regression on Demographic Data

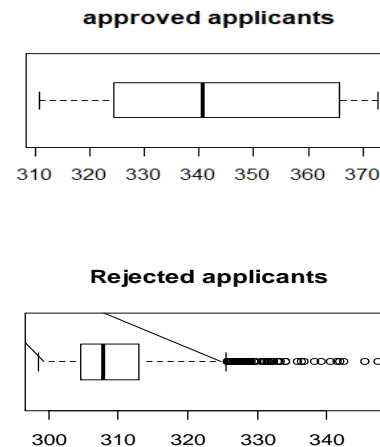


Random Forest on Demographic Data



APPLICATION SCORECARD AND CIBIL SCORES

- Application scorecard is build with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 points.
- CUTOFF SCORE is equal to 334.8953
- Maximum of approved credit-card holders have more than the cutoff score.
- 68.14 % of total defaulters are have less CIBIL scores than cutoff
- Except few outliers all the rejected applicants doesn't cross the cutoff score. Only 11 rejected applicants have CIBIL scores more than cutoff.



REVENUE LOSS AND POTENTIAL CREDIT LOSS

Revenue Loss :

- Occurs when good customers are identified as bad and credit card application is rejected.
- No of candidates rejected by the model who didn't default – 27015.
- About 41% of the non defaulting customers are rejected which resulted in revenue loss.

Credit Loss Saved :

- The candidates who have been selected by the bank and have defaulted are responsible for the credit loss to the bank.
- 68.14 % of defaulters identified by using this model
- Only 923 out of 2897 defaulters cross cut off score card
- 99.2 % of rejected applicants are rejected by this model also.

FINANCIAL BENEFITS OF THE MODEL

- The Confusion Matrix for calculating the Financial gain using our model was made on the dataset without missing Performance tag records, since we need to evaluate how much gain was achieved using our model for applicants who were provided with credit card compared to when no model was used.

Profit calculations – with model Vs without model

- We have considered an average profit of Rs.5000 from each non defaulters
- an average loss of Rs.1,00,000 when each accepted applicant defaults
- Net Profit without model = Rs 3,94,45,000
- Profit using model will be total profit due to each true positive and each true negative minus loss from each false positive and each false negative prediction
- Profit with model = Rs 10,17,10,000
- Net financial gain with using our model = Rs. 6,22,65,000