BFS Capstone Project

SUBMISSION

Submitted By:

1. Aakanksha Maurya
2. Snehalakshmi Balasubramanian
3. Udaya Prakash ST

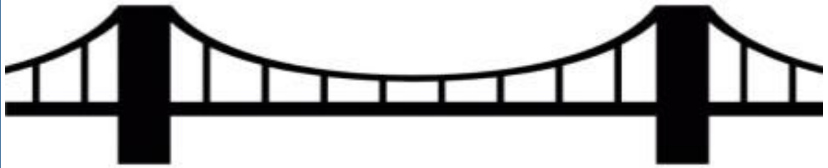
CredX, a leading credit card provider wants to minimize credit loss by acquiring the ‘right’ customers

Questions:

1. What are the key factors that impact credit risk?
2. What are the steps to be taken to mitigate credit risk
3. How demographic attributes of a customer impact credit risk?

Current State

- The acquisition team of CredX is responsible for identifying the right prospects to target and provide suitable product(s)
- In the past few years, CredX has experienced an increase in Credit Loss
- CredX wants to determine the factors affecting credit risk,and create strategies to mitigate the risk and assess the financial benefit of the risk model.



Desired Future State

CredX acquires the right customer base using the risk model, leading to overall decrease in credit loss

Data Available

- **Demographics/Application Data:** provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
- **Credit bureau**: Taken from the 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.

Executive Summary

UpGrad

Actions

1. Identify the top variables impacting the likelihood of default?

2. Build an application scorecard to identify the likely defaulters at applications stage using predictive models

3. Estimate the potential financial benefits of using the models for auto approval of credit cards

Insights & Recommendations

1. Predictive modelling should be used for auto approval of applicants
   - Credit loss can be reduced from 4.4% to 1.65%. (2.75% saved)
   - A revenue loss of 36% might occur if the model is used for auto approval
2. Significant variables from logistic regression model indicate the behavior of defaulting customers is indicated by recent interactions with bank

High level process flow

UpGrad

Identification of significant variables through EDA

IV & WOE analysis

Application Score-Card

Determine threshold score for rejecting an applicant

Understanding data

- *Understanding the credit bureau and demographic attributes*
- *Data wrangling and preparation*

Predictive Modelling

- *Model on demographic data*
- *Model on both demographic and credit bureau attributes*

Assess financial benefits

Assess financial benefit of using the risk model for auto approval

Detailed analytical process flow

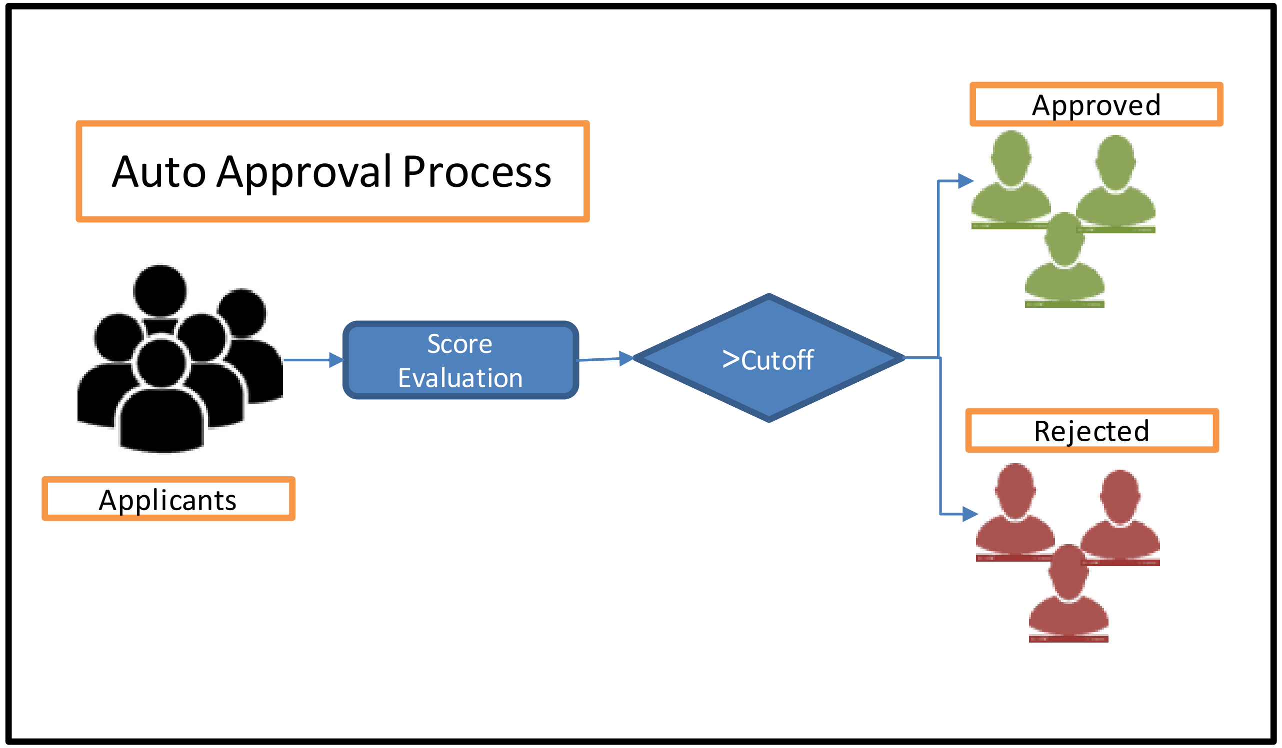
UpGrad

Data understanding and preparation
- Understand different attributes of customer and demographics data
- Remove duplicates
- Perform univariate and multivariate analysis
- Outlier treatment
- Null Value treatment

IV and WOE analysis
- Bin/Bucket continuous variables
- Compute WOE and IV
- Label the predictive power of variables using the IV values

Model Building And Evaluation
- Build simplistic logistic regression model
- Compute model evaluation metrics
- Build Decision trees/ Random forests if logistic regression model evaluation results are not satisfactory
- Evaluate model metrics of Decision trees/Random forests.

Application Score card and Financial Benefits assessment
- Build application score card by comparing the predicted scores of approved and rejected customers
- Find optimum rejection score eliminating maximum rejected applicants
- Assess financial benefit of the model assessing revenue and credit loss



Data understanding and preparation

Data Sources	Short Description	Primary Key
Demographics data	Obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.	ApplicationID
Credit Bureau	Taken from the 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.	ApplicationID

Demographic(Columns)
Application ID
Age
Gender
Marital Status (at the time of application)
No of dependents
Income
Education
Profession
Type of residence
No of months in current residence
No of months in current company
Performance Tag

Credit Bureau(Columns)
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

Data cleaning involved deduplication and removing NA and blank values across the dataset

UpGrad

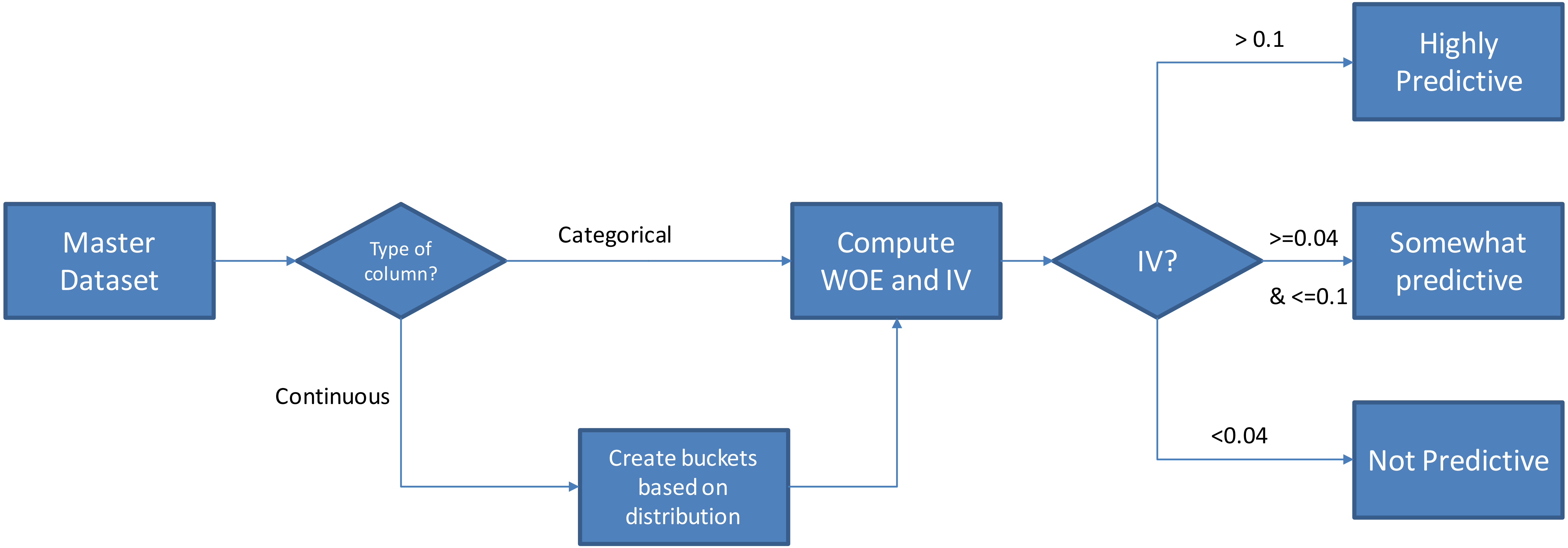
1. There are 3 identical duplicate application IDs in both the datasets(765011468,653287861,671989187). Since they are less than 10% of records, they are discarded. Post that the datasets are merged.

2. There are 3031 Missing values in the dataset.

3. Following are the columns having NA values:
   - No of dependents – 3
   - Performance tag - 1425(Indicates rejection)
   - Average credit card utilization in last 6 Months – 1058
   - No of trades opened in last 6 months – 1
   - Presence of open home loan – 272
   - Outstanding balance - 272
4. Rule Defined(If missing NA values less than 10% of observations, discard the data).
5. Both files have a Synonym field "Performance Tag". Hence we can use only one column

Weight Of Evidence and IV analysis - Process Flow

UpGrad

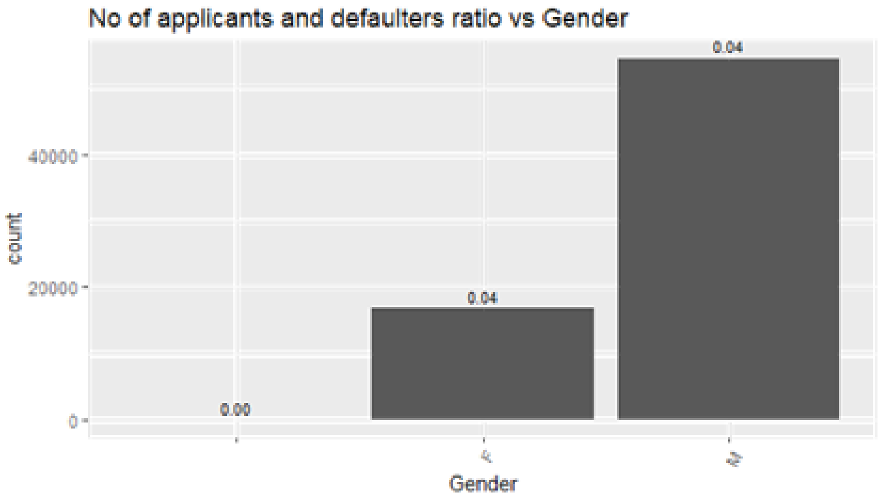


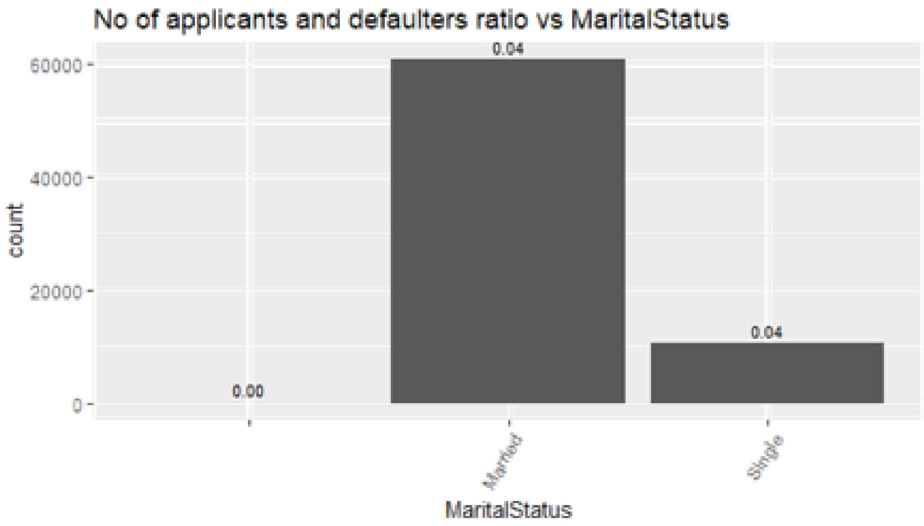
Weight Of Evidence and IV analysis suggest that *demographic variables hold very little predictive power*

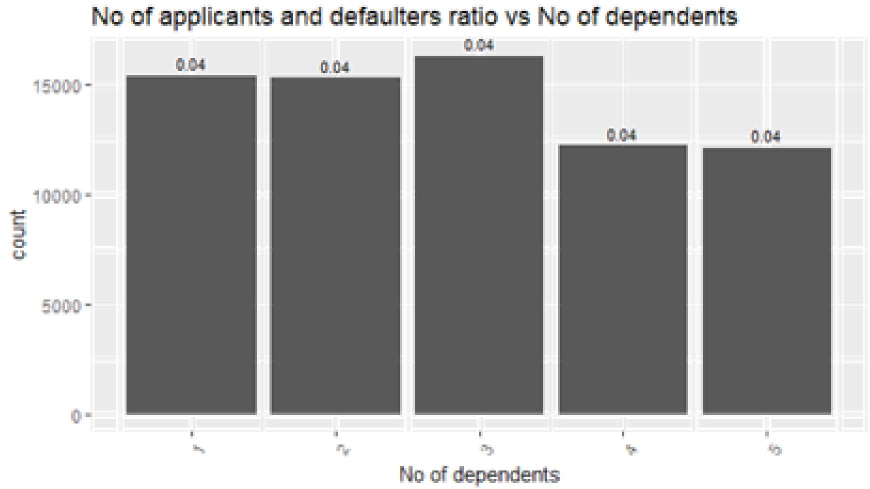
UpGrad

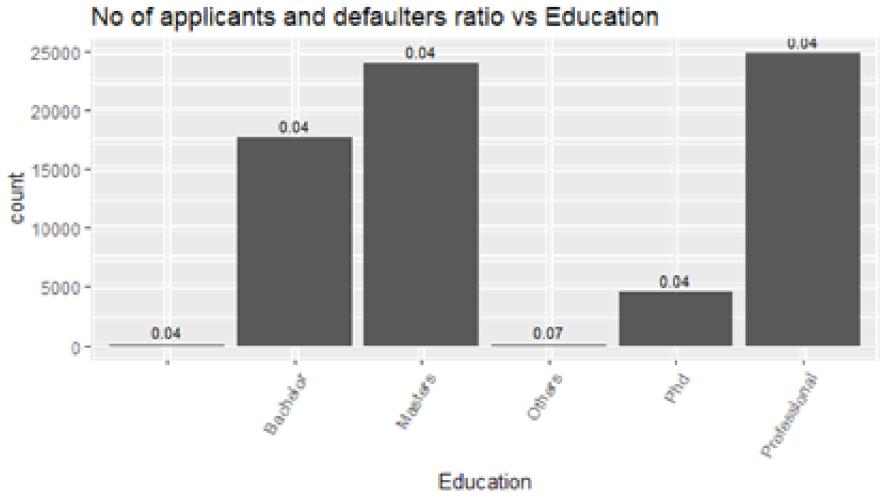
S.No	Variables	Demographics/CreditBureau?	Information Value	Predictive Power
1	No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	Credit Bureau	0.35	Highly Predictive
2	No.of.trades.opened.in.last.12.months	Credit Bureau	0.33	Highly Predictive
3	Avgas.CC.Utilization.in.last.12.months_bucket	Credit Bureau	0.32	Highly Predictive
4	No.of.PL.trades.opened.in.last.12.months	Credit Bureau	0.30	Highly Predictive
5	Outstanding.Balance_bucket	Credit Bureau	0.26	Highly Predictive
6	No.of.times.30.DPD.or.worse.in.last.6.months	Credit Bureau	0.24	Highly Predictive
7	Total.No.of.Trades_bucket	Credit Bureau	0.23	Highly Predictive
8	No.of.PL.trades.opened.in.last.6.months	Credit Bureau	0.22	Highly Predictive
9	No.of.times.30.DPD.or.worse.in.last.12.months	Credit Bureau	0.22	Highly Predictive
10	No.of.times.90.DPD.or.worse.in.last.12.months	Credit Bureau	0.22	Highly Predictive
11	No.of.times.60.DPD.or.worse.in.last.6.months	Credit Bureau	0.21	Highly Predictive
12	No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	Credit Bureau	0.21	Highly Predictive
13	No.of.trades.opened.in.last.6.months	Credit Bureau	0.20	Highly Predictive
14	No.of.times.60.DPD.or.worse.in.last.12.months	Credit Bureau	0.20	Highly Predictive
15	No.of.times.90.DPD.or.worse.in.last.6.months	Credit Bureau	0.16	Highly Predictive
16	Income_bucket	Demographics	0.04	Somewhat Predictive
17	No.of.months.in.current.company_bucket	Demographics	0.02	Not Predictive
18	Presence.of.open.home.loan	Credit Bureau	0.02	Not Predictive
19	Profession	Demographics	0.02	Not Predictive
20	No.of.months.in.current.residence_bucket	Demographics	0.01	Not Predictive
21	Type.of.residence	Demographics	0.01	Not Predictive
22	Marital.Status..at.the.time.of.application.	Demographics	0.01	Not Predictive
23	age_bucket	Demographics	0.00	Not Predictive
24	No.of.dependents	Demographics	0.00	Not Predictive
25	Gender	Demographics	0.00	Not Predictive
26	Presence.of.open.auto.loan	Credit Bureau	0.00	Not Predictive
27	Education	Demographics	0.00	Not Predictive

EDA plots – I Demographic attributes- Default rate do not vary much across Gender , Marital Status, No of dependents and Education

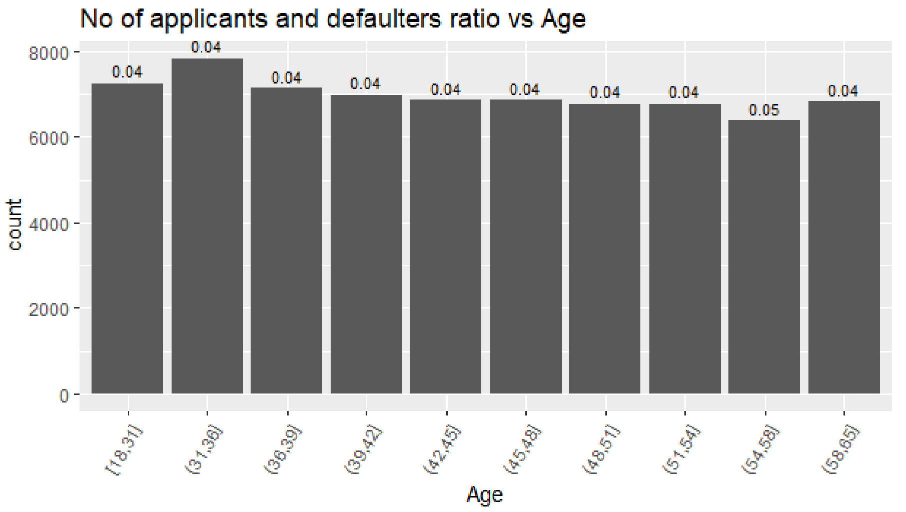


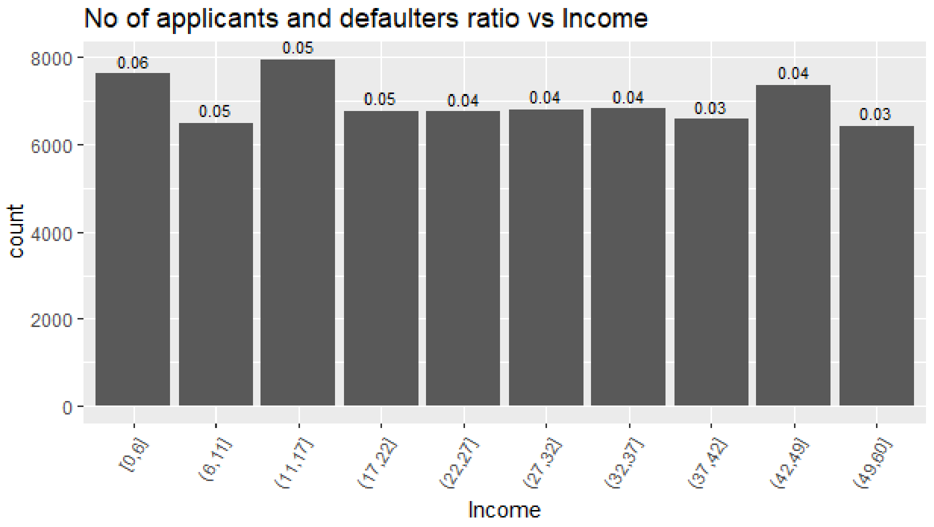


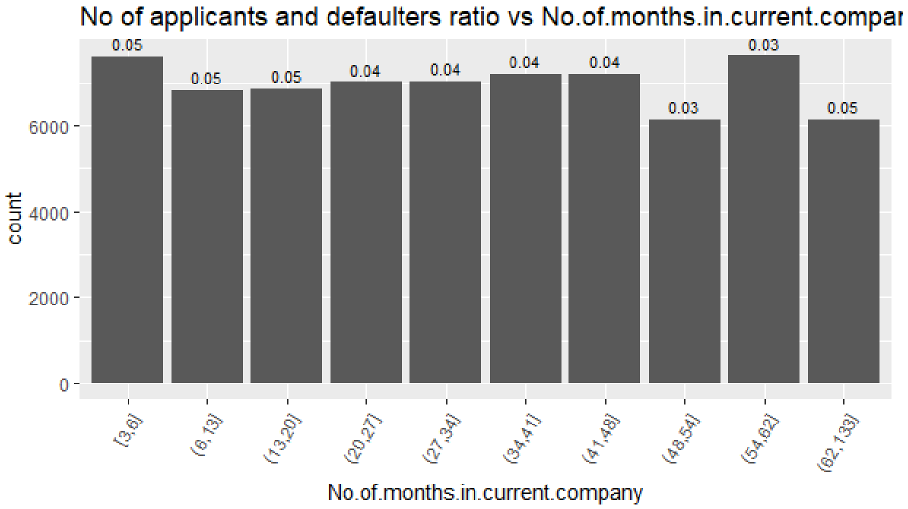


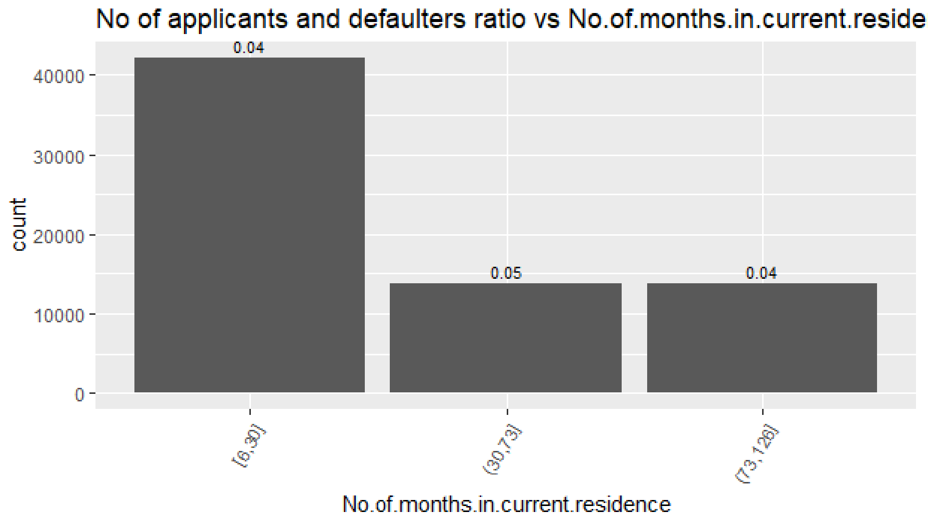


EDA plots – I Demographic attributes: Default rate do not vary much across Age income and number of months in current company and residence

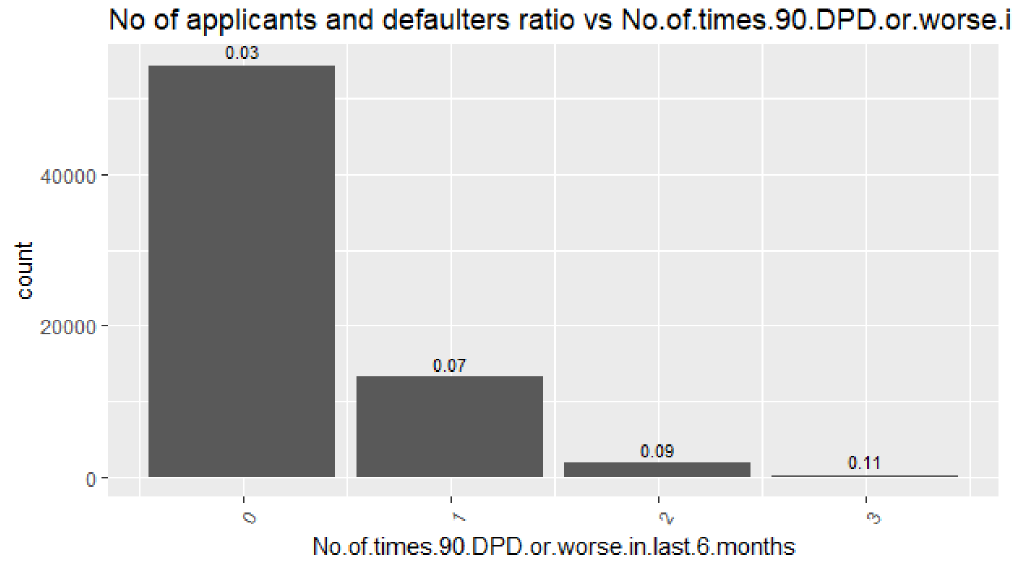


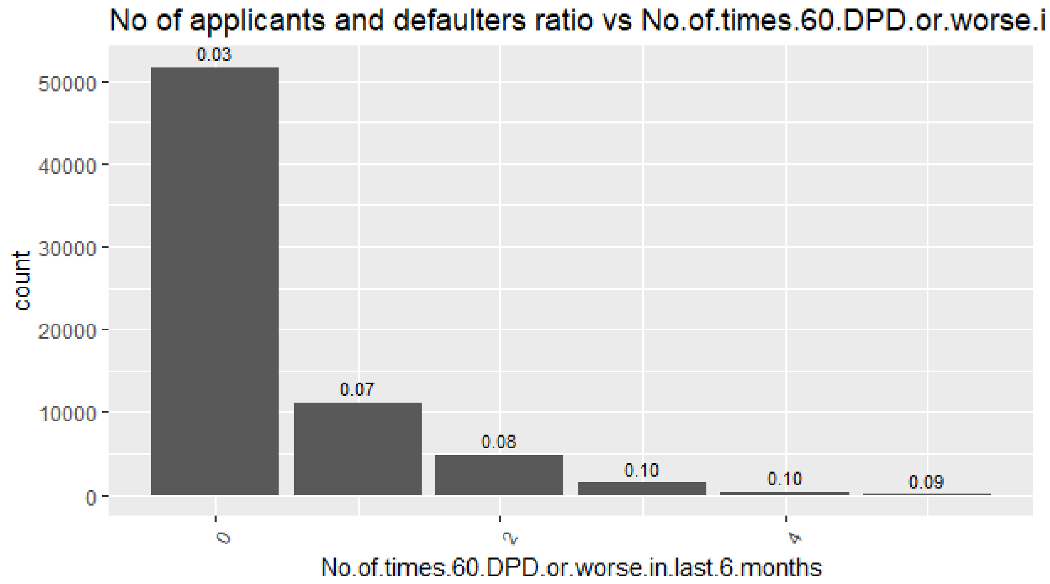


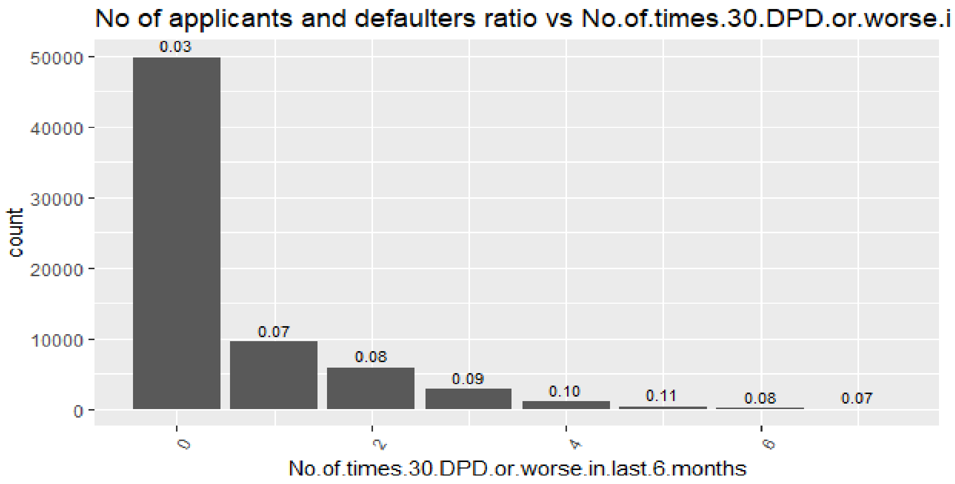


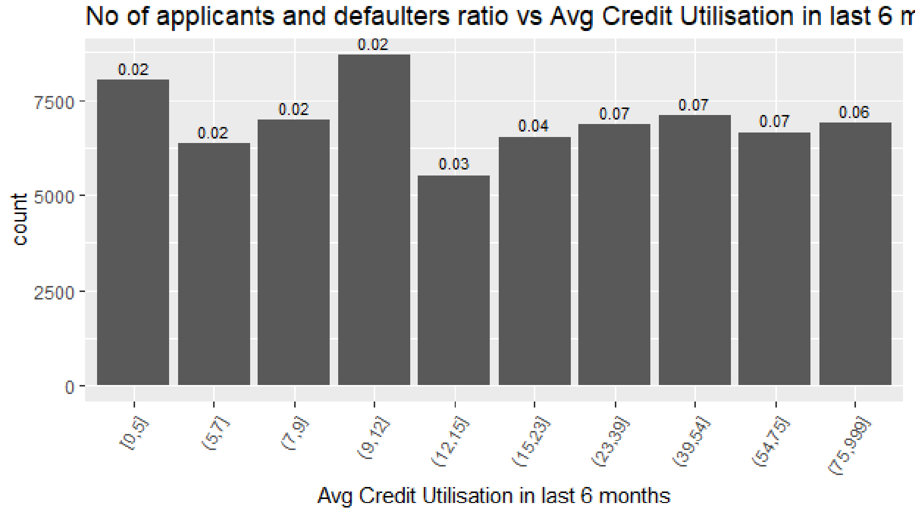


EDA plots – II Credit Bureau : Higher the number of times 30,60,90 DPD in last 6 months, higher the risk of defaults

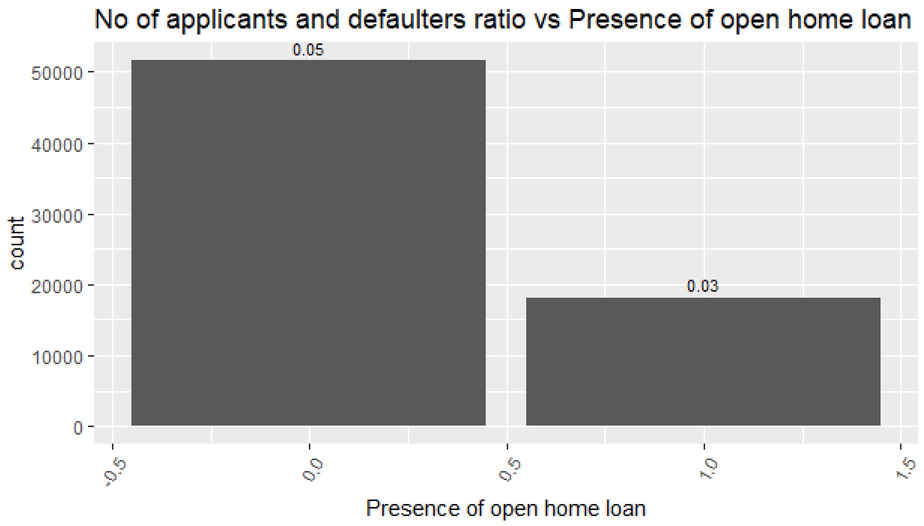


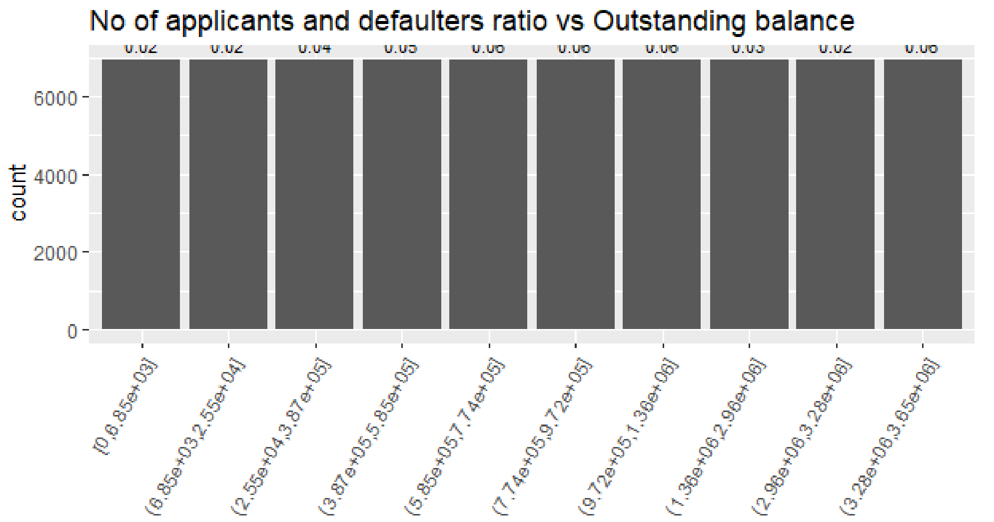


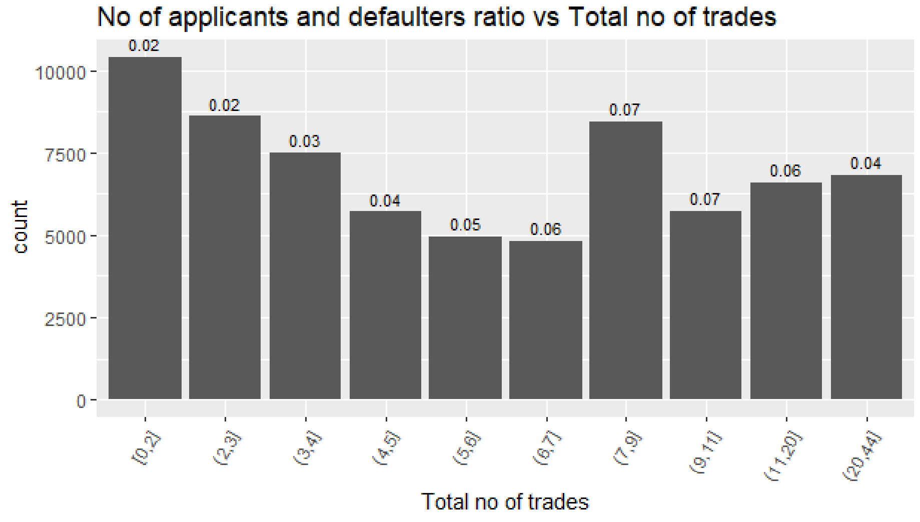


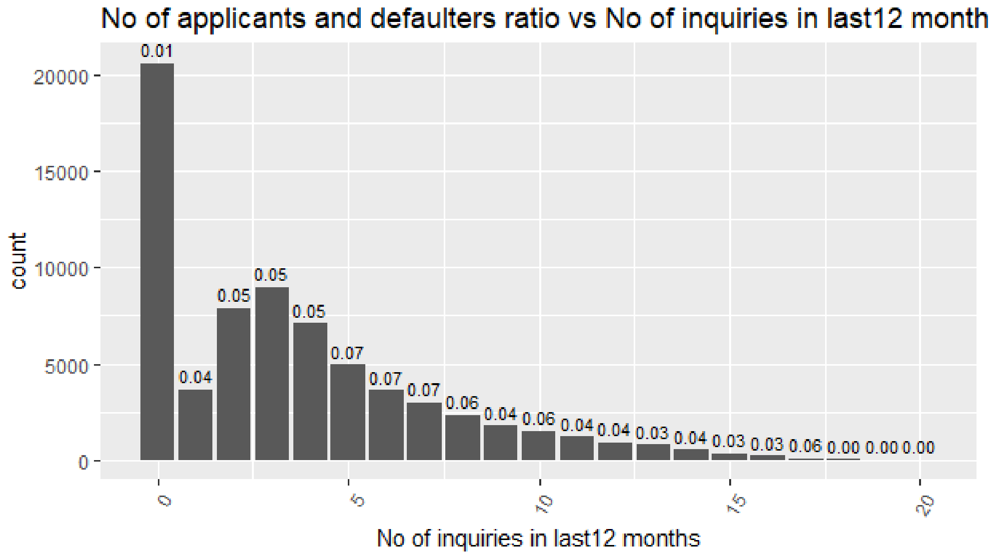


EDA plots – II Credit Bureau: Higher the outstanding balance higher the risk of default









Model developed using only the demographic attributes have very low accuracy, sensitivity and specificity.

UpGrad

I. Sample imbalance problem was observed as number of defaulters in the dataset is low.

II. Model was built using both unbalanced and smote sampled data.

III. Maximum accuracy achieved is 51%

Sample Selected	*Accuracy*	*Sensitivity*	*Specificity*
Unbalanced Sample	51.0%	61.8%	50.5%
Smote Sample	50.0%	49.1%	50.1%

Final model was developed using both the demographics and credit bureau data

UpGrad

I. Sample imbalance problem was observed as number of defaulters in the dataset is low.

II. Model was built using both unbalanced and smote sampled data.

III. Both logistic and random forest models were created and optimized for maximum accuracy

IV. Even though accuracy and sensitivity are the best for Random forest, *specificity is less than 1 %*. Therefore ***Logistic Regression Model is chosen to be the final model***

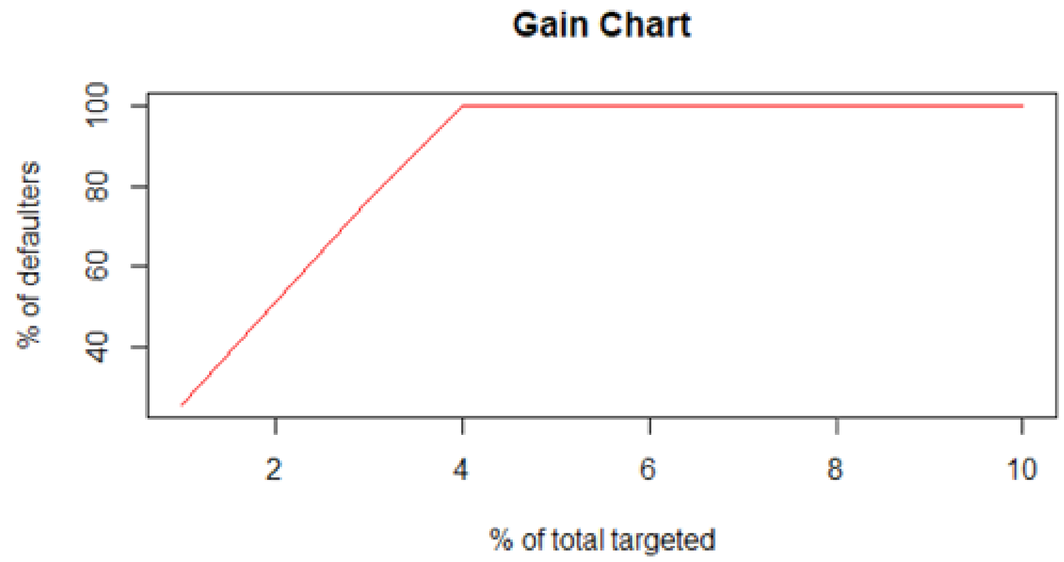
Model Type	*Sample Selected*	*Accuracy*	*Sensitivity*	*Specificity*
Logistic Regression	Unbalanced Sample	63%	61%	63%
Random Forest	Unbalanced Sample	95.8%	100.0%	0.1%
Random Forest	Smote Sample	95.30%	99.48%	0.45%

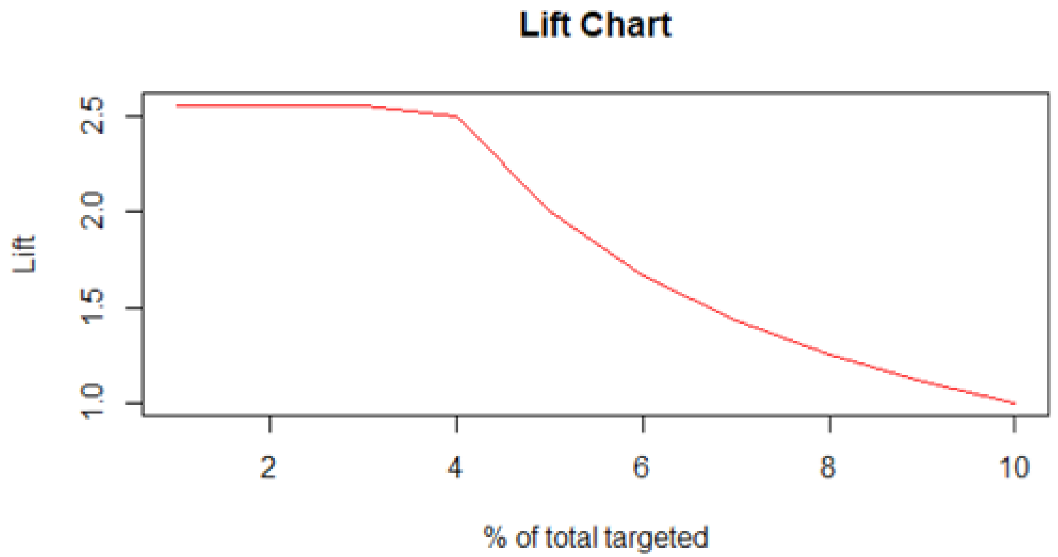
Significant variables from logistic regression model indicate the behavior of defaulting customers is indicated by recent interactions with bank

Significant variables
No.of.times.30.DPD.or.worse.in.last.6.months
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
Total.No.of.Trades
Avgas.CC.Utilization.in.last.12.months
Outstanding Balance
Income
No.of.months.in.current.residence_bucket
age
Education(Others)

Logistic Regression model identifies 76% of the defaulters in top 3 deciles of the population

bucket	total	Defaulters	Cumulative Defaulters	Gain	Cumlift
1	7,102	7,102	7,102	25.59	2.56
2	7,101	7,101	14,203	51.18	2.56
3	7,101	7,101	21,304	76.77	2.56
4	7,102	6,447	27,751	100.00	2.50
5	7,101	0	27,751	100.00	2.00
6	7,101	0	27,751	100.00	1.67
7	7,102	0	27,751	100.00	1.43
8	7,101	0	27,751	100.00	1.25
9	7,101	0	27,751	100.00	1.11
10	7,101	0	27,751	100.00	1.00





Score card is developed based on probability scores of all the observations(Approved and Rejected)

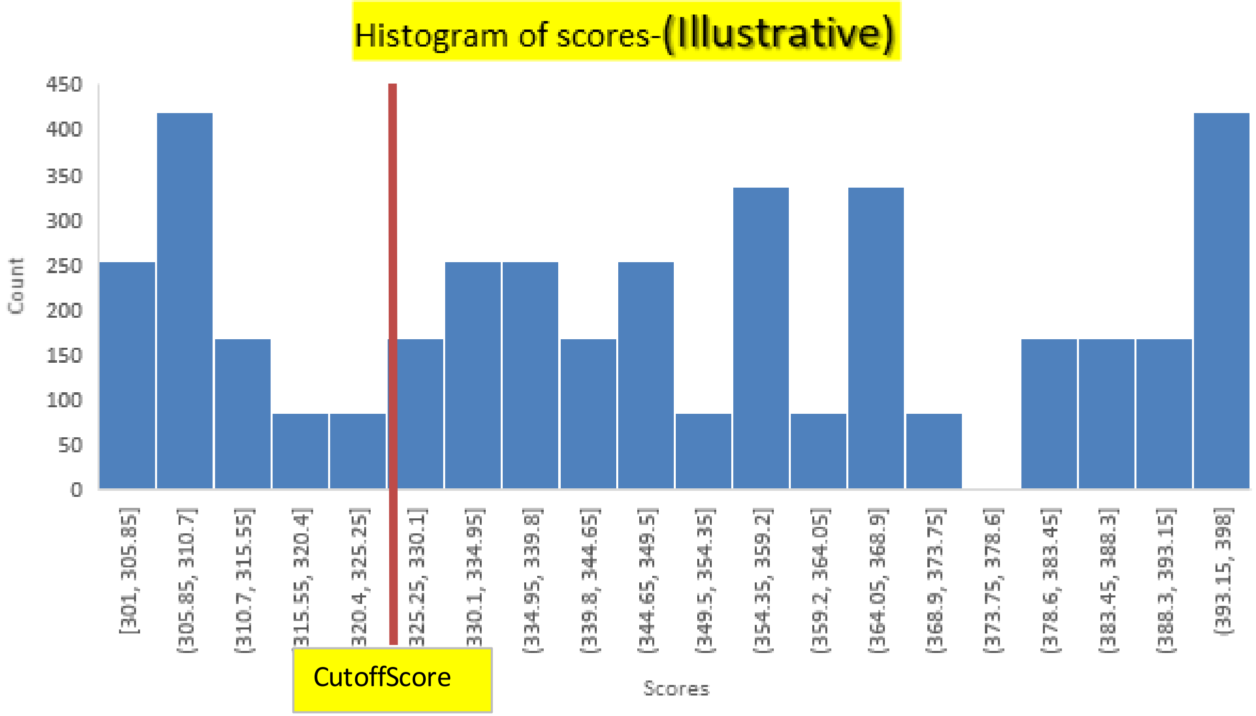
Score card calculation:

1. Points to double the odds = 20, Base Score=400 & odds = 10
2. Score = Offset + { Factor* log(Odds) }

where

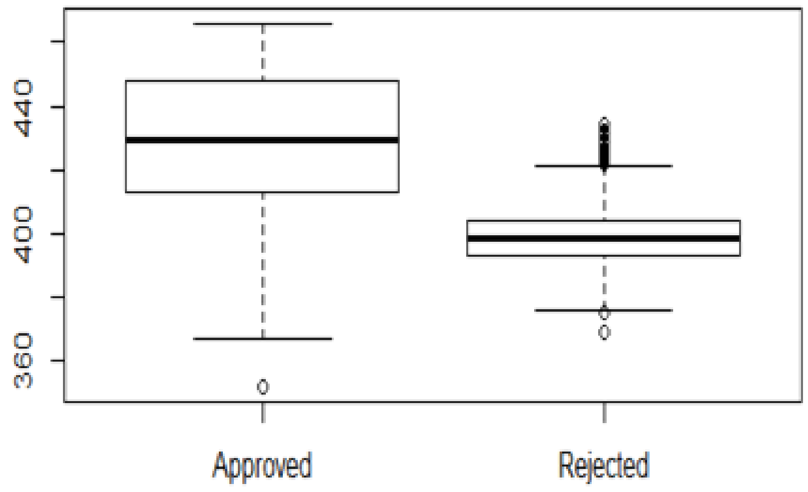
Factor = = 20/log(2) = 28.8539
Offset = 400 - (28.8539*log(10)) = 333.5614
log(odds) = log(odds(good)) = log(probability(0)/probability(1))

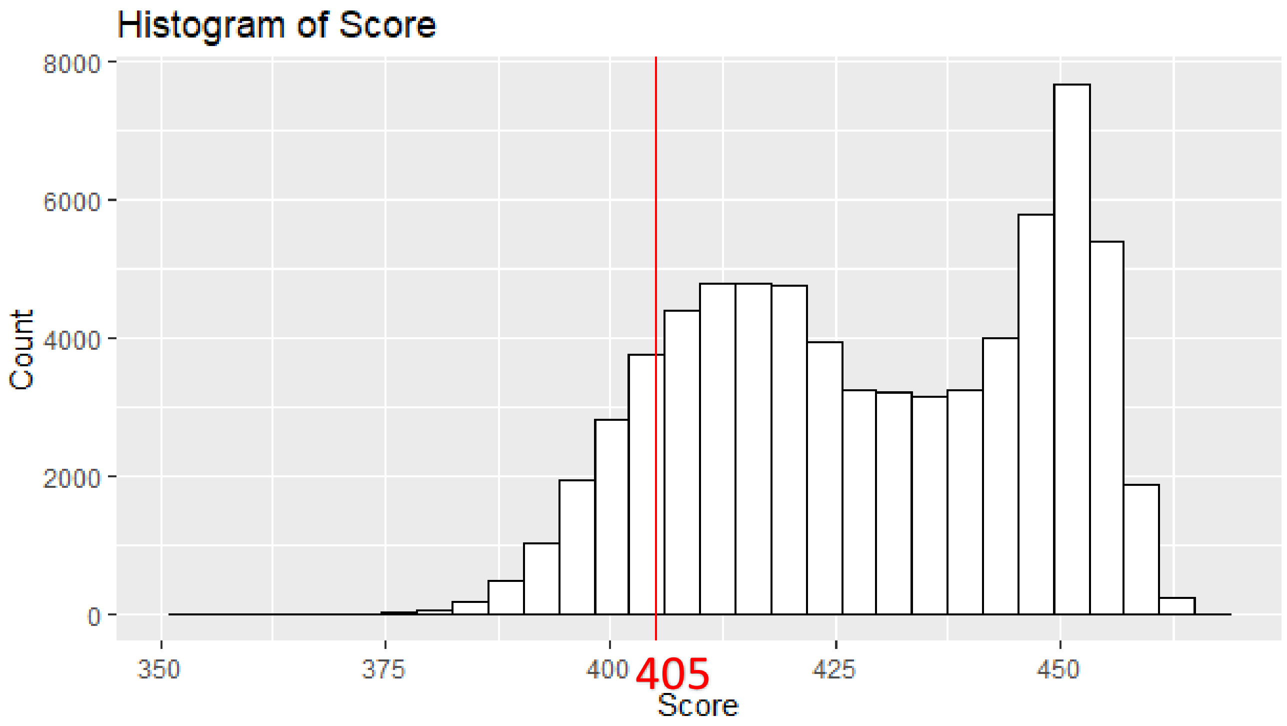
A cut off score will be defined above below which the applicants would be rejected.



Score varies between 352 and 465 and average score of 'rejected' population is very low as compared to average score of approved population

UpGrad





I. Cut off is defined at 405, ***Model predicts 77.12% of the rejected population correctly based on cutoff score***

II. *Rejected by Bank:* ***1425****, Identified correctly by Model at cutoff -* ***1099***

A credit loss of 2.75% can be saved if the model is used for auto approval of applications

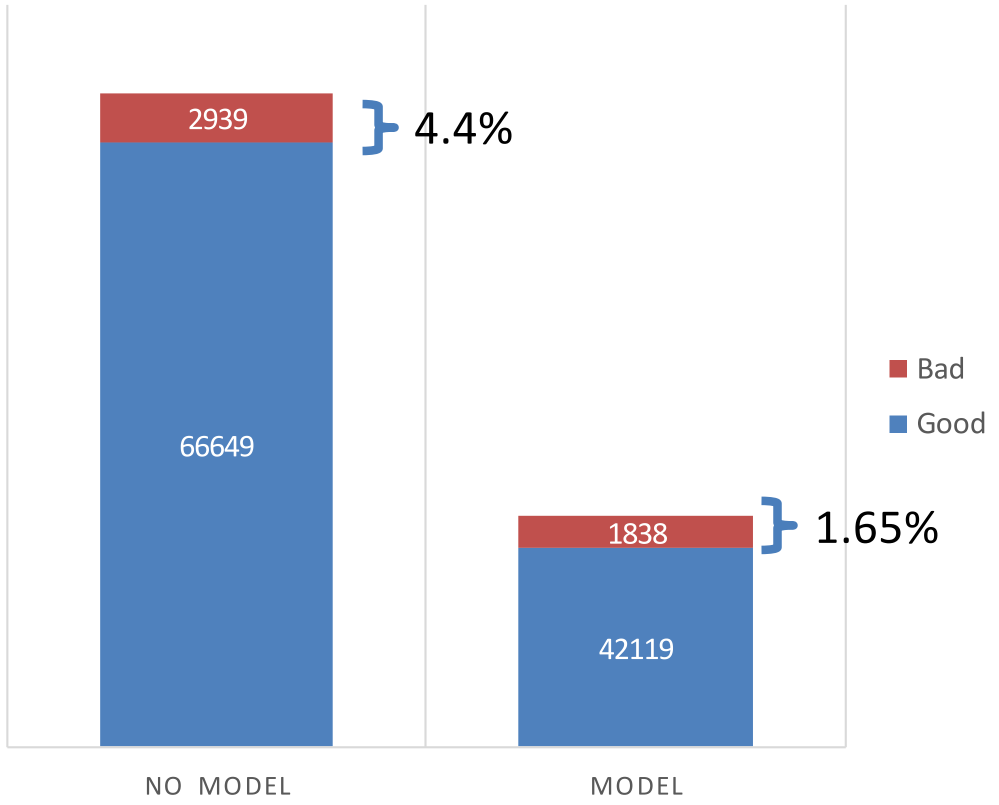
UpGrad

Credit Loss is the loss occurred from bad customers.

Credit Loss with no model in place: ***4.4%*** (2939/66649)

Credit Loss with model in place: ***1.65%***

Credit Loss Saved: 2.75%



A revenue loss of 36% might occur if the model is used for auto approval

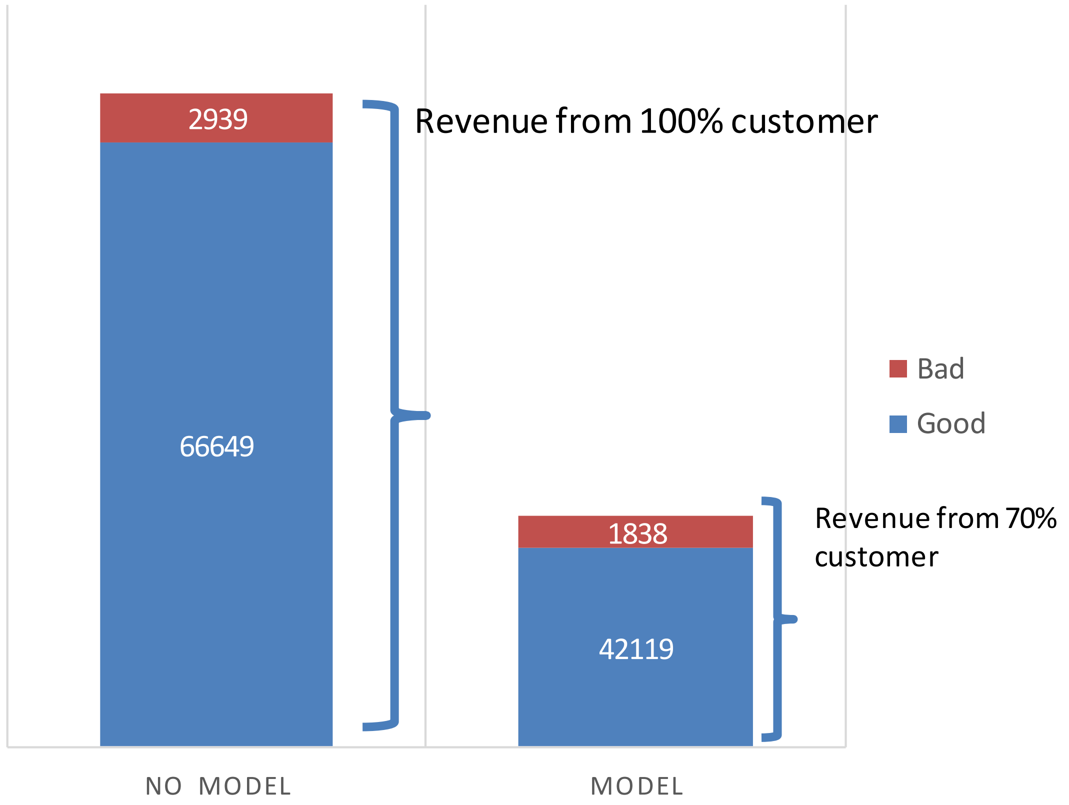
UpGrad

Revenue loss is the loss incurred by wrongly identified ***"bad"*** to the good customers.

Revenue with no model : 100%

Revenue Loss: 36%

Revenue with model: 64%




Thank
You