CAPSTONE PROJECT - BFSI

SUBMISSION

Group Members:

1. Shivam Kakkar (Facilitator) - Roll Number - DDA1730346
2. Ashwin Suresh
3. Manohar Shanmugasundaram
4. P Sai Prathyusha

Business Understanding

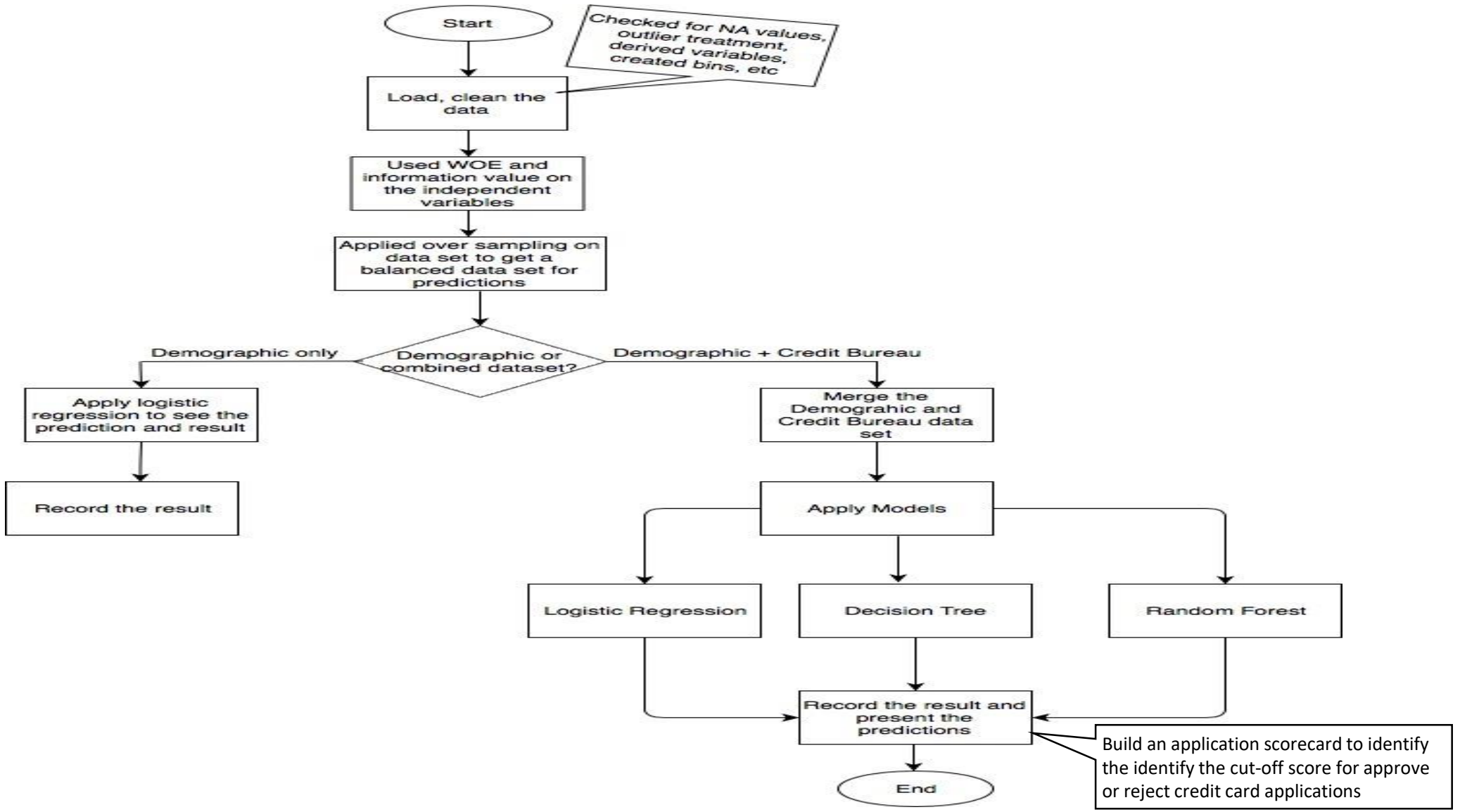
- Credx is a leading credit card provider that gets thousands of credit card applicants every year.
- In the past few years, they have been experiencing an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.
- The objective of this project is to find the right customer to reduce the credit loss.

Data Understanding and Preparation

- The demographic data is obtained from the information provided by the applicants at the time of credit card application, which includes customer level information on age, gender, income, marital status etc.
- The credit bureau data contains variables such as number of time 30 DPD (Days Past Due) or worse in last 3/6/12 months, outstanding balance, number of trades etc.
- The demographic and credit bureau csv files are read and named as below:

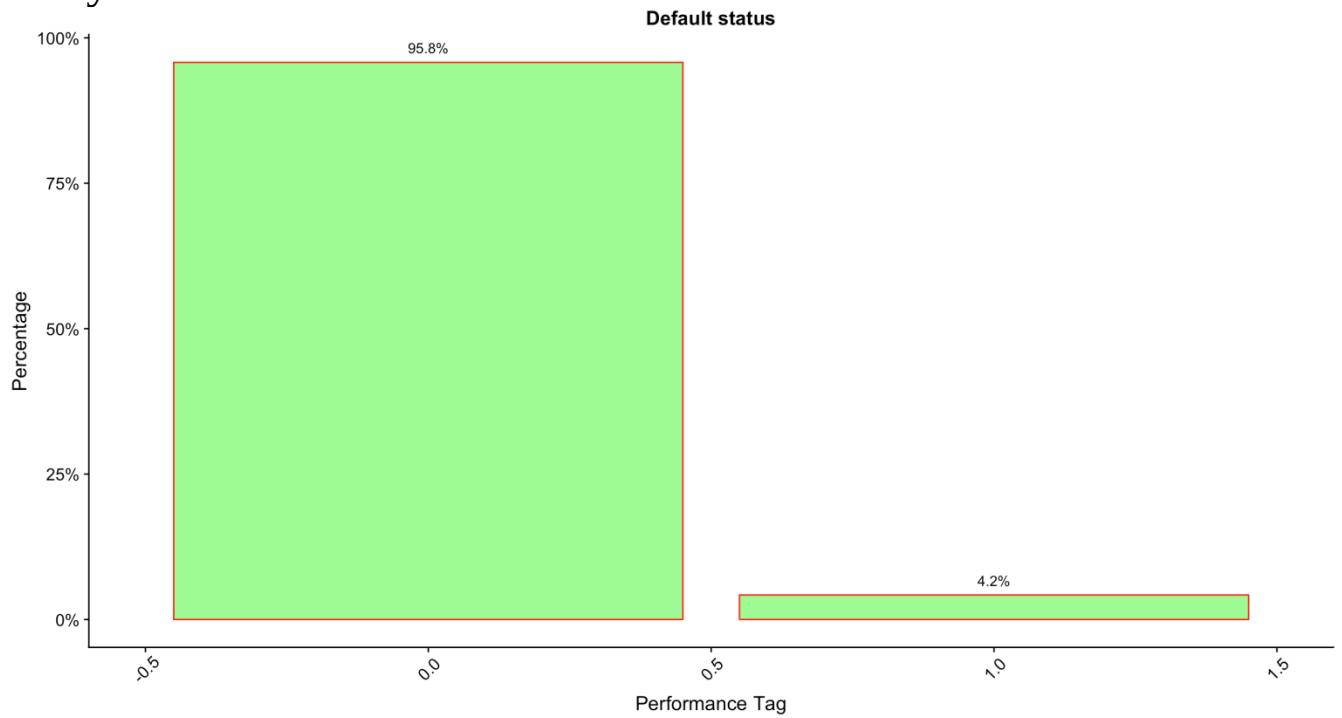
  inspect_demo (Demographic) and inspect_credit (Credit Bureau)
- The demographic data consists of 71295 observations with 12 variables including 1577 NA’s and 3 duplicates application id.
- The credit bureau data consists of 71295 observations with 19 variables including 3028 NA’s and 3 duplicates application id.

Flow chart for the activities

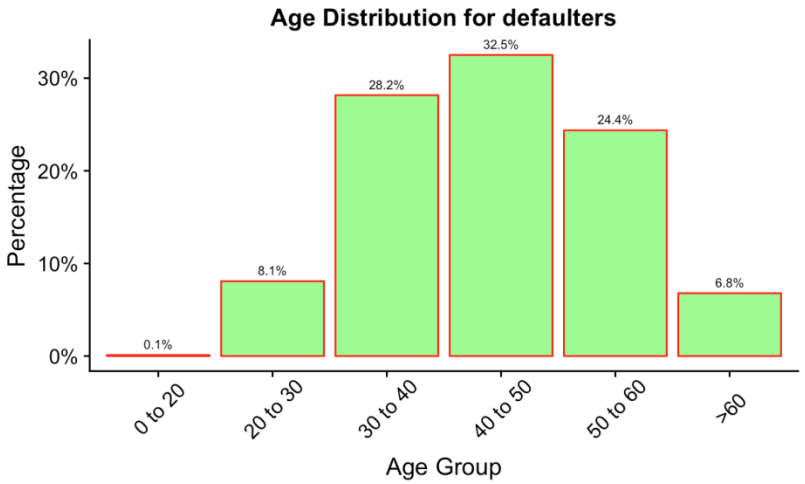


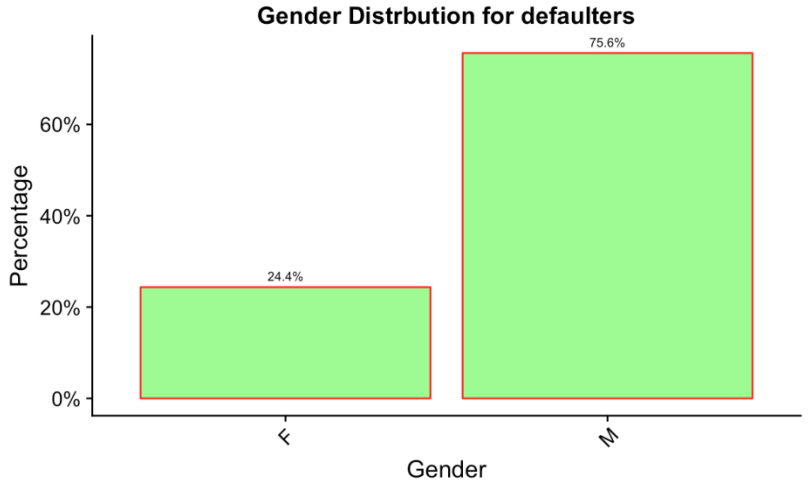
Exploratory Data Analysis (Demographic Data)

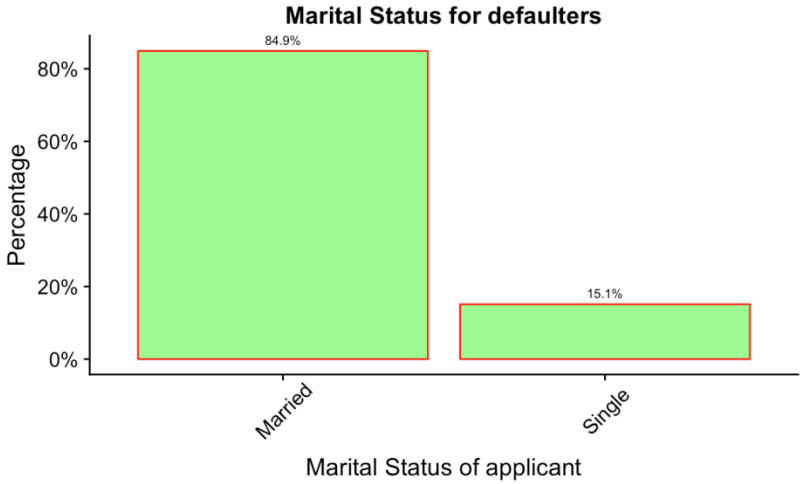
- Exploratory Data Analysis conducted on independent variables on both the datasets i.e. demographic data and merged (demographic data and credit bureau data)
- The analysis is conducted for the folks who are defaulters.
- **There is only 4.2% of bad data and 95.8% of good data available**
- Following are the graphs and the inference from it:
- We have considered only defaulters.

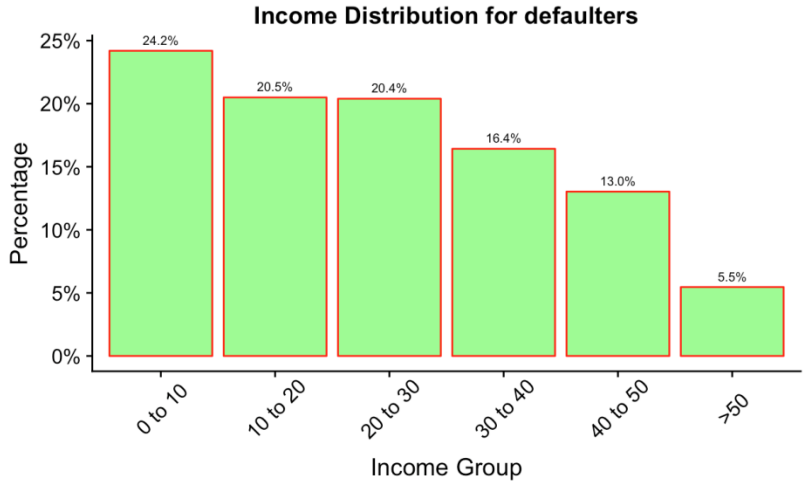


Exploratory Data Analysis (Demographic Data)





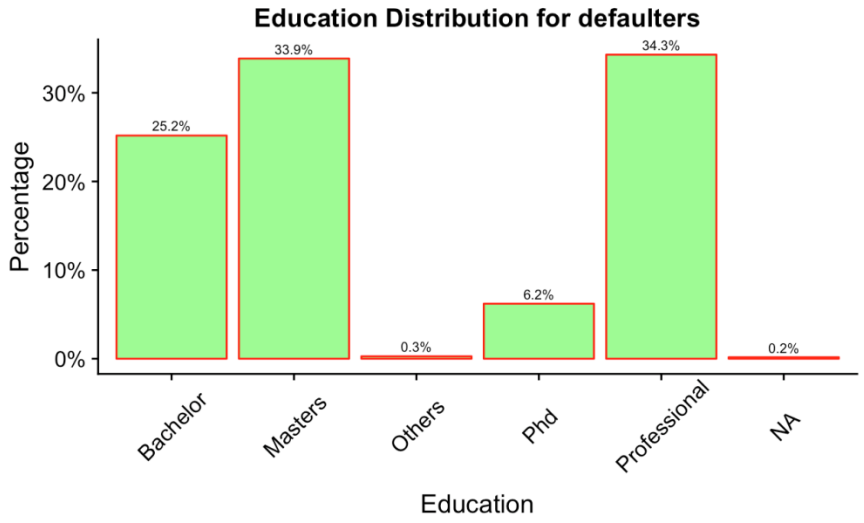


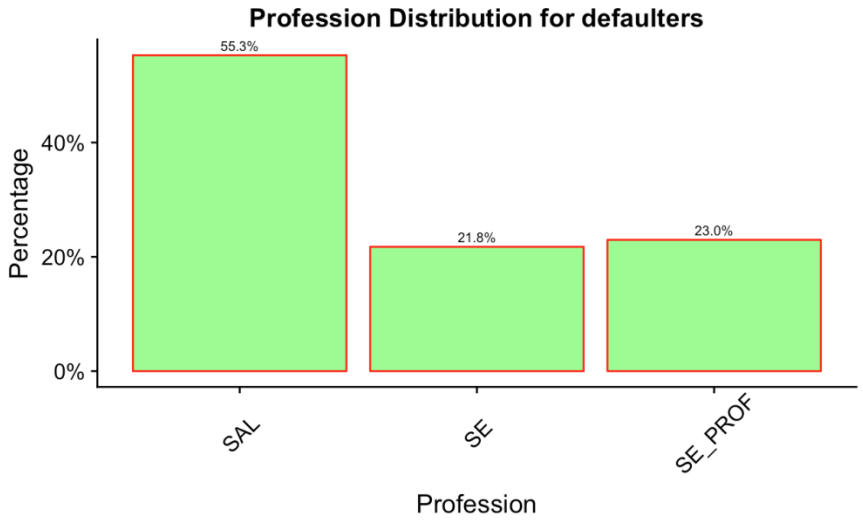


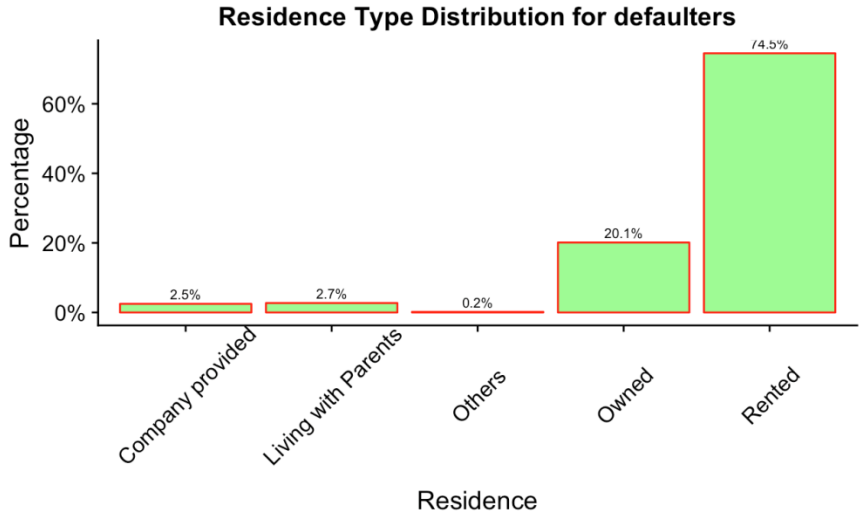
INFERENCE:

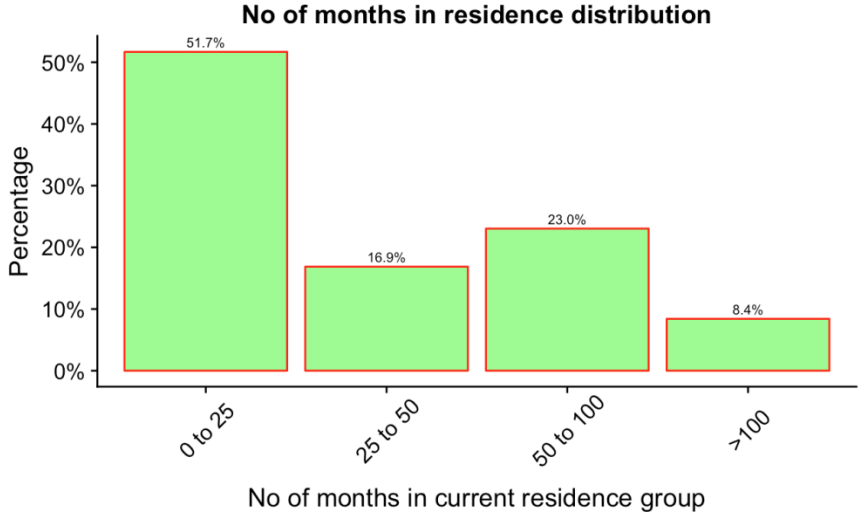
- From the Income distribution (we have binned it) , the highest number of defaulters are in the lowest income range (0-10k) and the lowest number of defaulters are in the highest income range (> 50k)
- From the Age distribution, majority of defaulters are in the age of (30 – 60) and they more or less equally distributed in the brackets of (30-40) ,(40-50) & (50-60). Least number of defaulters are in the age bracket of (> 60)
- No inference can be drawn from “Gender” and “Marital Status”.

Exploratory Data Analysis (Demographic Data)





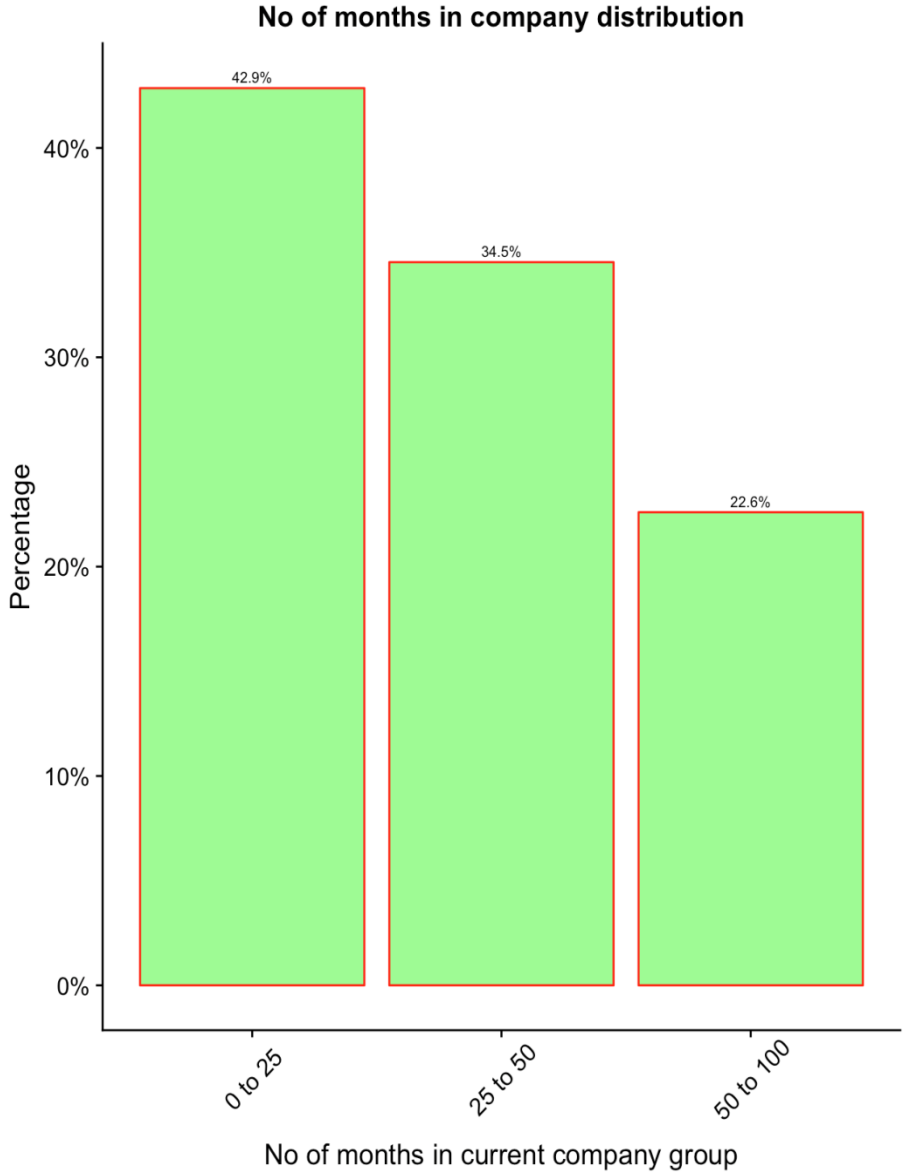


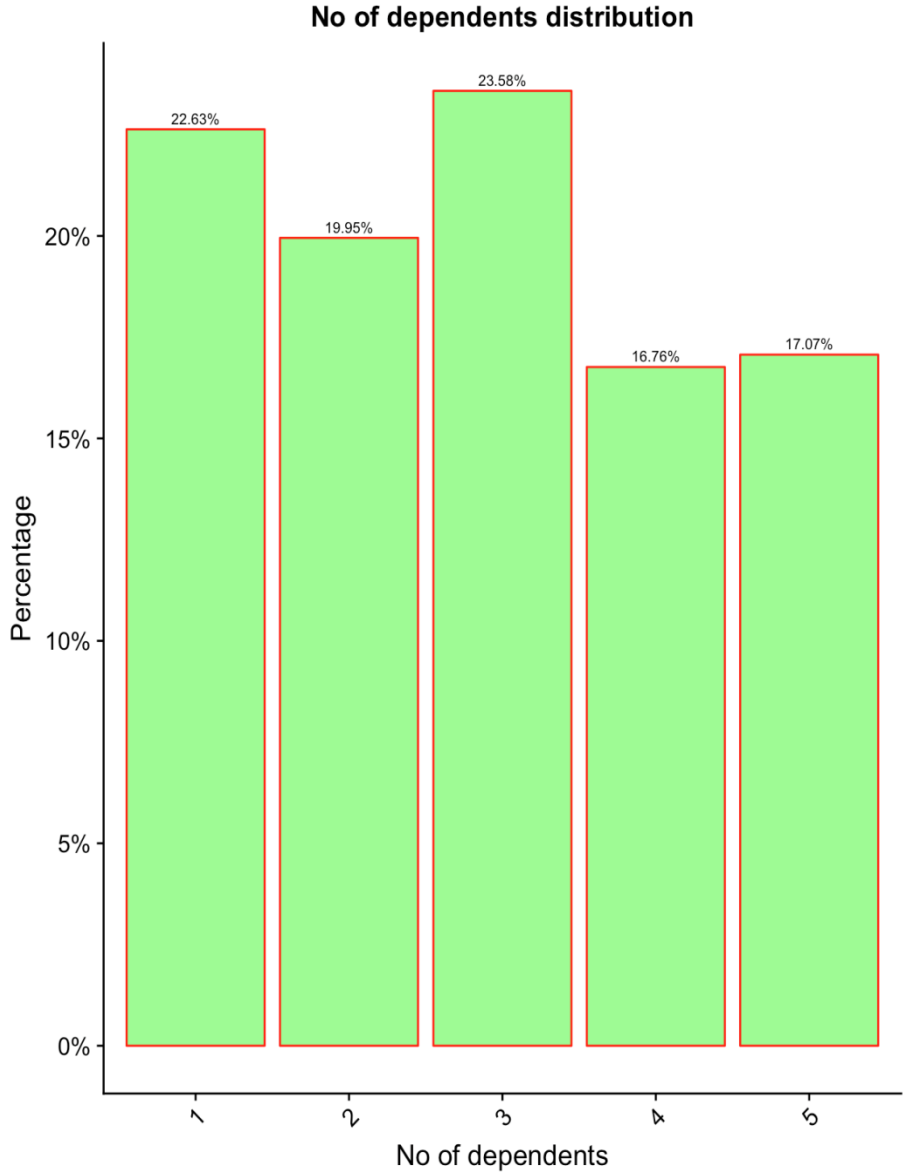


INFERENCE

- From the "no of months in residence" distribution (binned variable) there is a clear declining trend in the default rate, hence lesser duration of stay in current residence, higher the chance of default.
- From the "Residence Type" Distribution , those who stay in Rented default the most.
- From the "Education" Distribution , those who have "Phd" defaulted less but still we cannot see any trend in terms of education.
- From the "Profession" distrbution "SAL" ones who have defaulted the most but still we cannot see any trend.

Exploratory Data Analysis (Demographic Data)





INFERENCE

- From the "no of months in current company distribution (this is binned variable)" there is clear lesser the no of months in current company higher the chance of defaulting , there is clear decreasing trend in default rate as the no of months in current company increases.
- Non inference can be drawn from the "no of dependents"

Information Value of predictors (demographic data)

Variable	info_value
No.of.months.in.current.residence	0.188
Income_grp	0.038
No_mnths_company_grp	0.014
No.of.dependents	0.005
Profession	0.004
Age_grp	0.002
Gender	0.002
Education	0.001
Marital.Status..at.the.time.of.application.	0.001
Type.of.residence	0.001
Application.ID	0.000

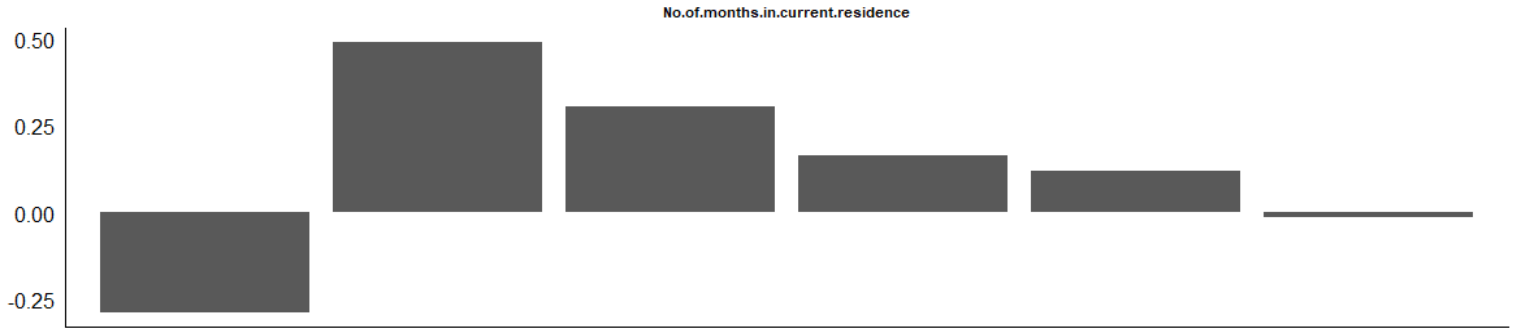
- We have binned the following variables:
  - Income → Income_grp
  - No.of.months.in.current.company → No_mnths_company_grp
  - Age → Age_grp

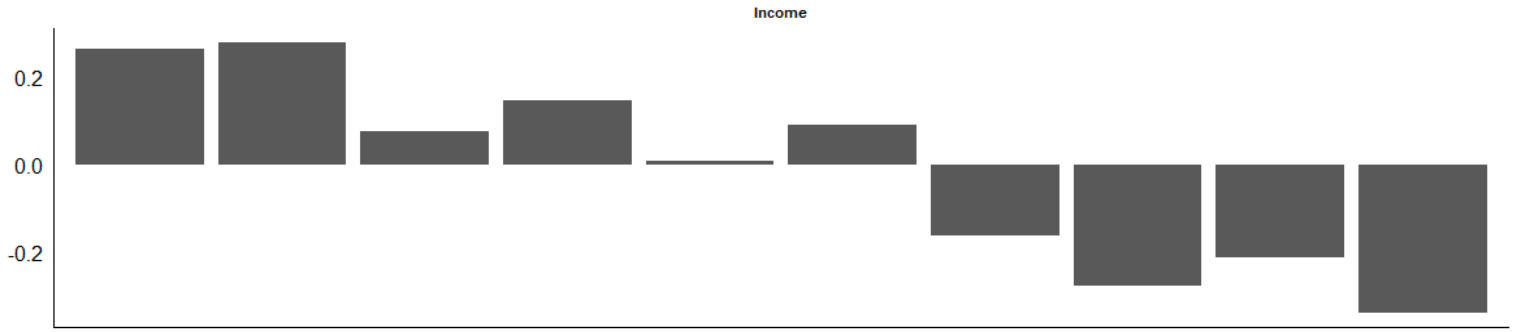
- On the basis of following information:

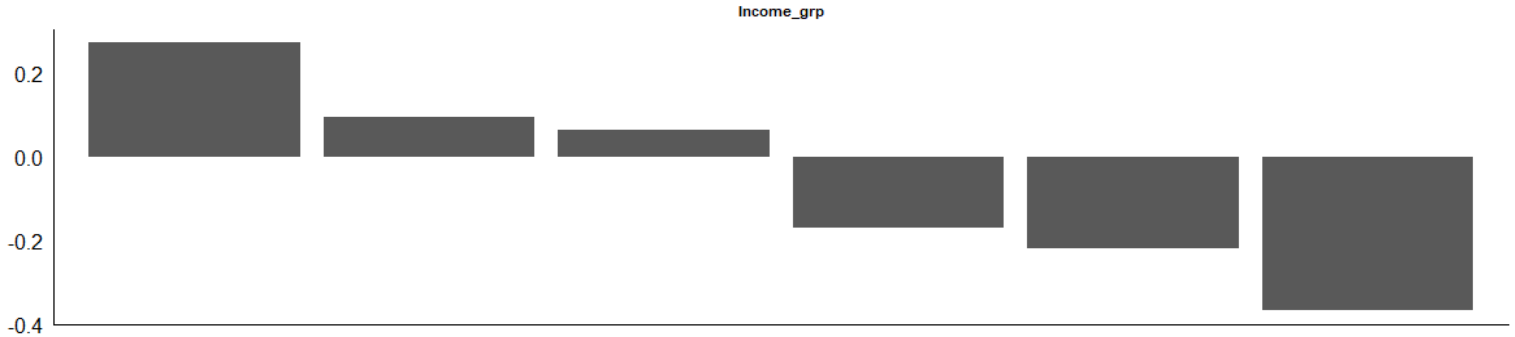
Information Value	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
> 0.3	Strong predictor

Hence we can drop the variables marked in Red.

WOE plots of predictors (demographic data)



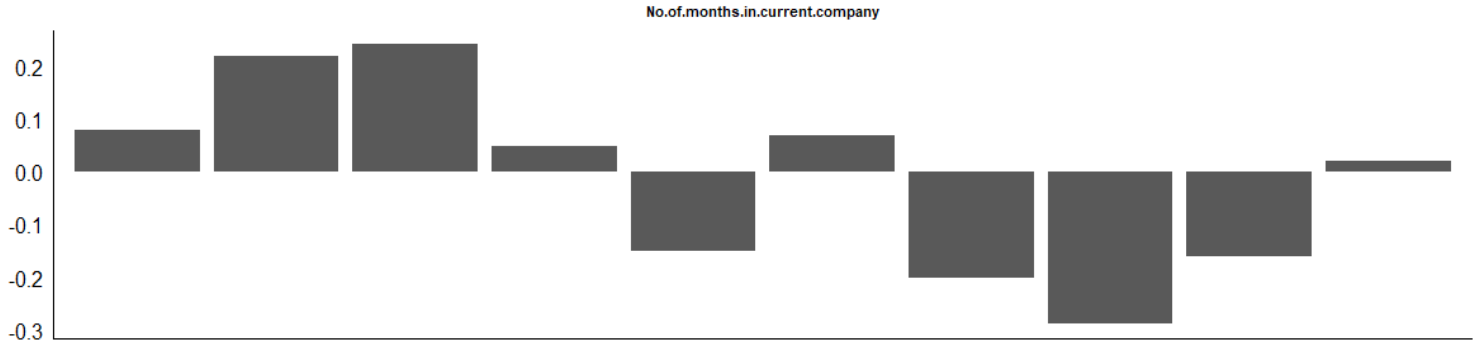


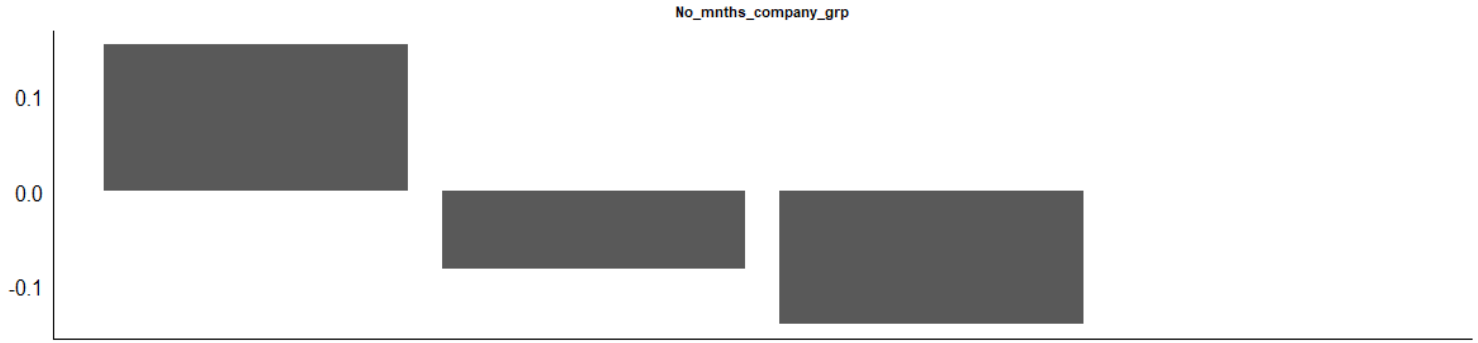


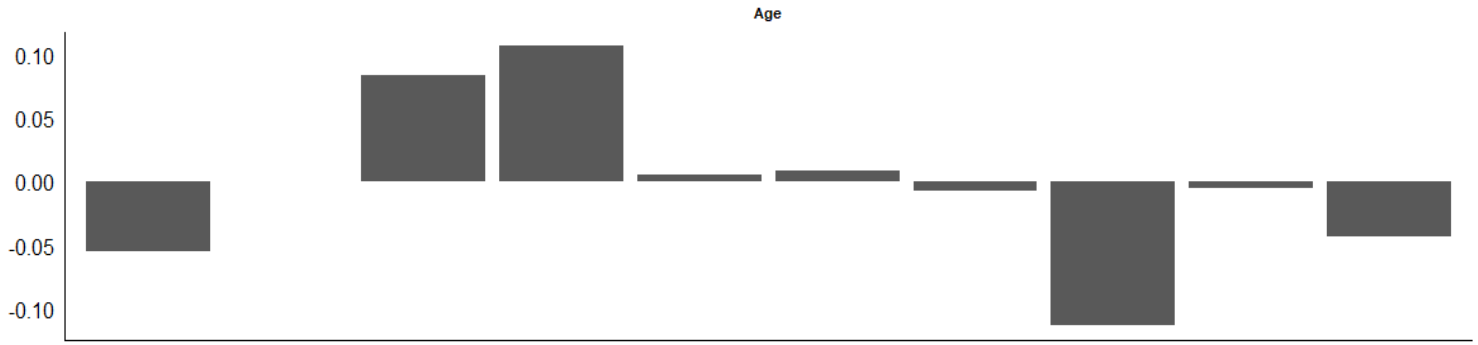
INFERENCE

- From the income_grp we can see it is following a monotonic trend so we can say that our binning of income to income group is correct
- From the “no of months in current residence” we see a negative woe for the months in 0-25 but then post that we see a downward monotonic trend, hence it shows that for defaulters “no of months in current residence” (0-25) play a significant role.

WOE plots of predictors (demographic data)



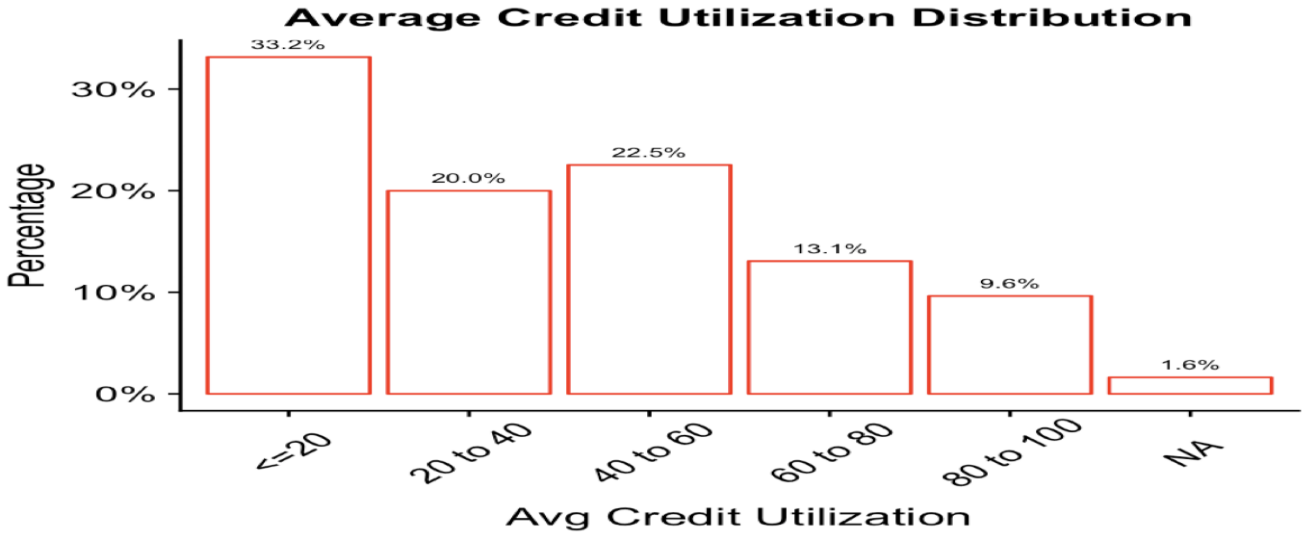


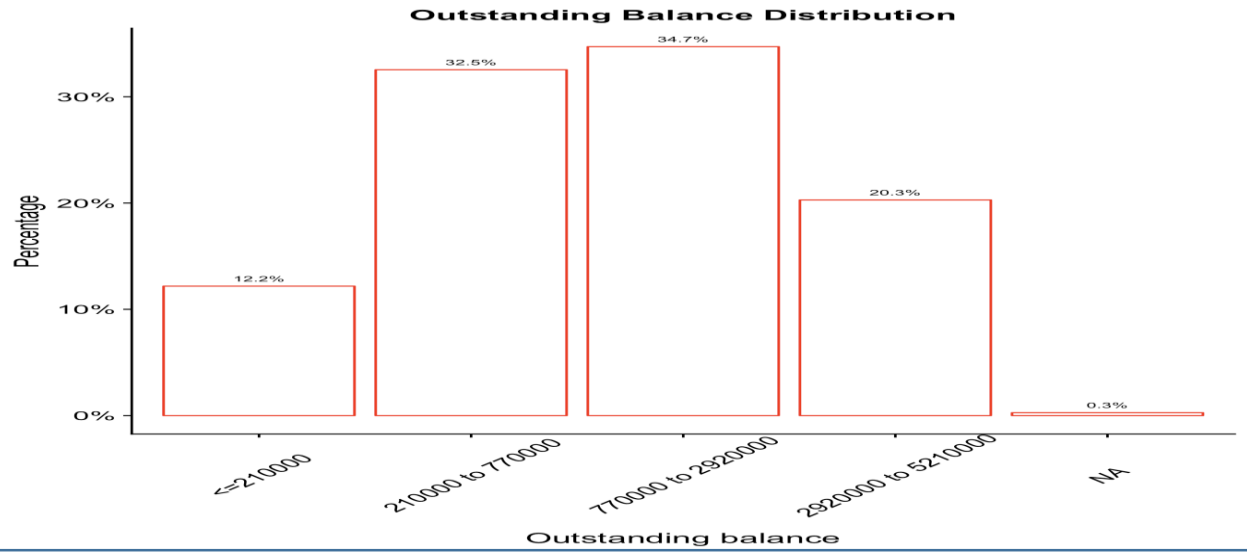


INFERENCE

- Binned the “no of months in current company” to “no_mnths_company_grp” to make the woe trend monotonic.
- For Age we saw that that defaulters are distributed quite evenly in the bracket of 30-40, 40-50 and 50-60 and then defaulters decreases post age of 60
- Later binning has been done to make it monotonic.

Exploratory Data Analysis (Merged Data)

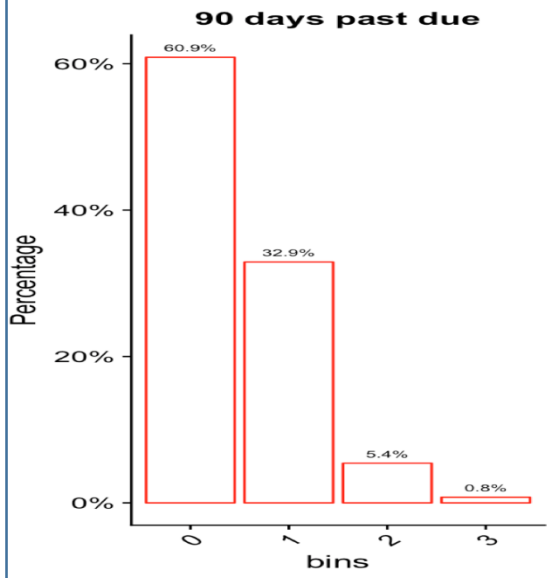


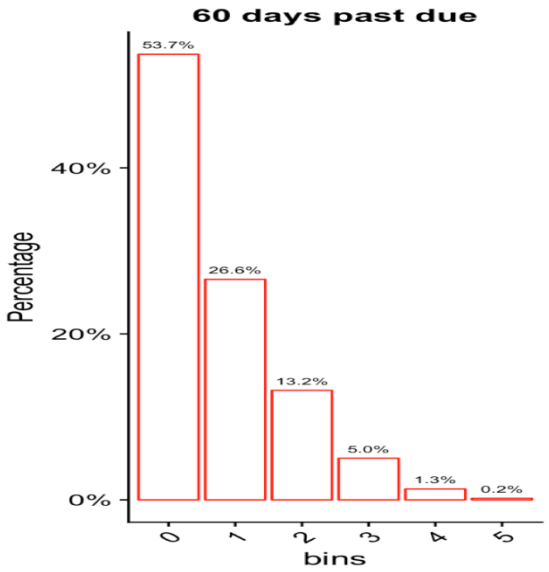


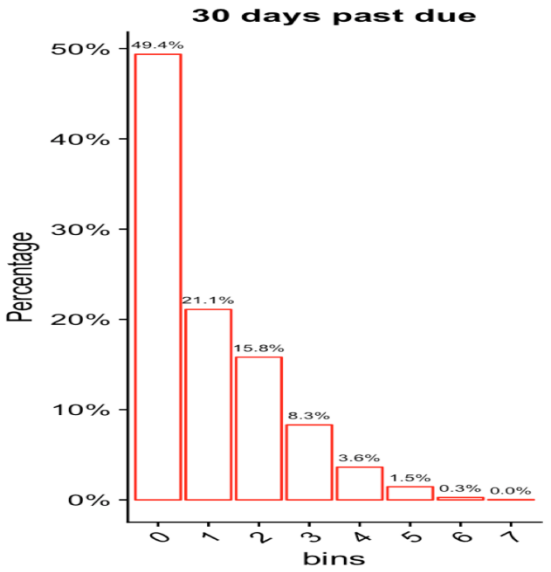
INFERENCE

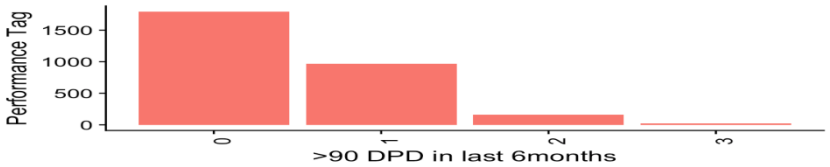
- The default rate of people who use their avg credit less than 20 is slightly high, but doesn't show strong indication since the default rate increased in case of utilization from 40-60 than 20-40.
- From the above graph there is an increase in the default rate where the outstanding balance is from 210000 to 2920000

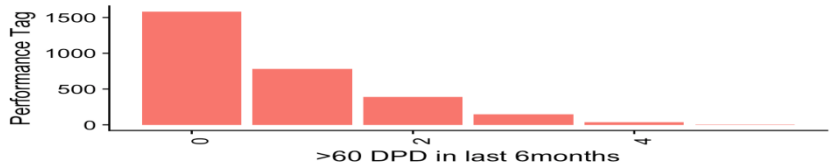
Exploratory Data Analysis (Merged Data)

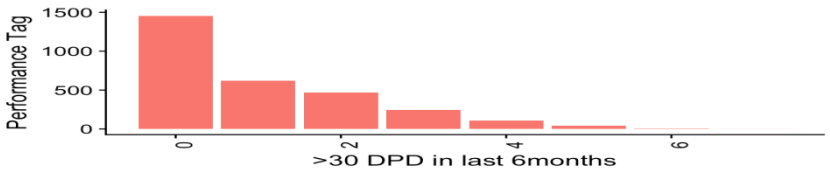


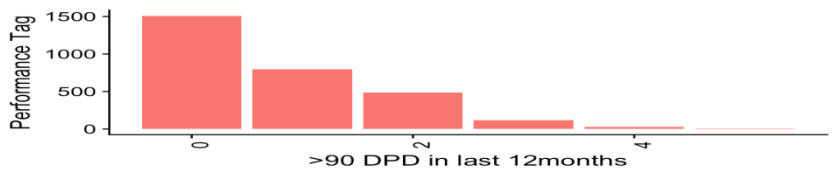


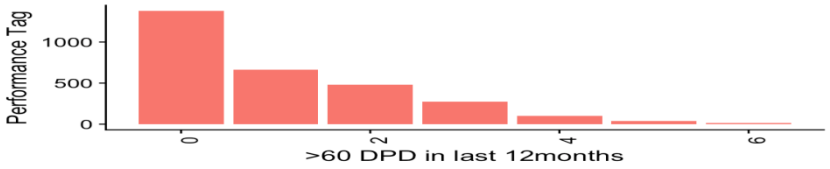


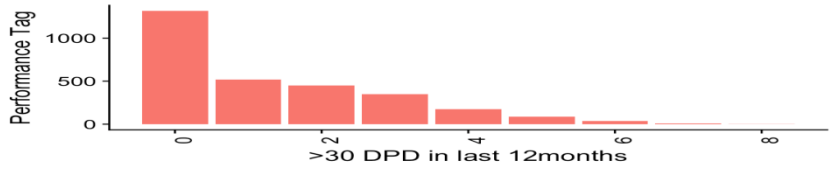












INFERENCE

The above graphs indicate that the frequency customers who have not paid their dues 0 times have defaulted more that who have more number of dues

Information Value of predictors (merged data)

Variable	Info value	Variable	Info value
No.of.trades.opened.in.last.12.months	0.313	No.of.months.in.current.residence	0.171
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.303	No.of.times.90.DPD.or.worse.in.last.6.months	0.167
No.of.PL.trades.opened.in.last.12.months	0.302	Income_grp	0.048
Avgas_CC_bin	0.281	Presence.of.open.home.loan	0.015
Total.No.of.Trades	0.275	No_mnths_company_grp	0.014
No.of.times.30.DPD.or.worse.in.last.6.months	0.234	Profession	0.004
No.of.PL.trades.opened.in.last.6.months	0.234	No.of.dependents	0.003
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	0.219	Gender	0.002
No.of.times.90.DPD.or.worse.in.last.12.months	0.214	Presence.of.open.auto.loan	0.002
No.of.times.30.DPD.or.worse.in.last.12.months	0.213	Age_grp	0.002
No.of.times.60.DPD.or.worse.in.last.6.months	0.205	Type.of.residence	0.002
No.of.trades.opened.in.last.6.months	0.196	Education	0.001
No.of.times.60.DPD.or.worse.in.last.12.months	0.186	Marital.Status..at.the.time.of.application.	0.001
Outstanding_Balance_bin	0.171		

We have binned the following variables:

- Income
- No.of.months.in.current.company
- Age

On the basis of following information:

Information Value	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
> 0.3	Strong predictor

Hence we dropped all the variables marked in red in the table

MODEL BUILDING APPROACH

1) Approach

1. Weight of evidence (WOE), Information Value (IV) and Score values are common terms that we encounter in credit risk modelling in the financial industry.
2. The models using them can be constructed using various algorithms, like logistic regression, decision trees, Random forest, boosting, neural networks.
3. Among these the logistic regression is a time-tested popular model, which is widely used in industries, even now.
4. We are first going to build this as our base model.

2) Key Issues to consider before modelling:

1. How to treat NA's effectively without deleting observations?
2. How to overcome, the severe imbalance of the categories, so that the model is not biased towards the majority category?

3) Woe approach has a two-fold objective:

1. Na's themselves are treated as a separate category in both continuous and categorical variables. These will be replaced by their respective woe, along with other data observations.
2. The second one, is extremely important. The conventional way to treat a categorical variable is to use dummy coding. In dummy coding, the different levels of a categorical variable have same importance or weights. On the contrary, in woe replacement, the values are replaced by weight of evidence, which brings in weights to the levels.
3. As such, woe technique is one of the most widely used tools in credit scoring problems.

4) Tackling Class Imbalance

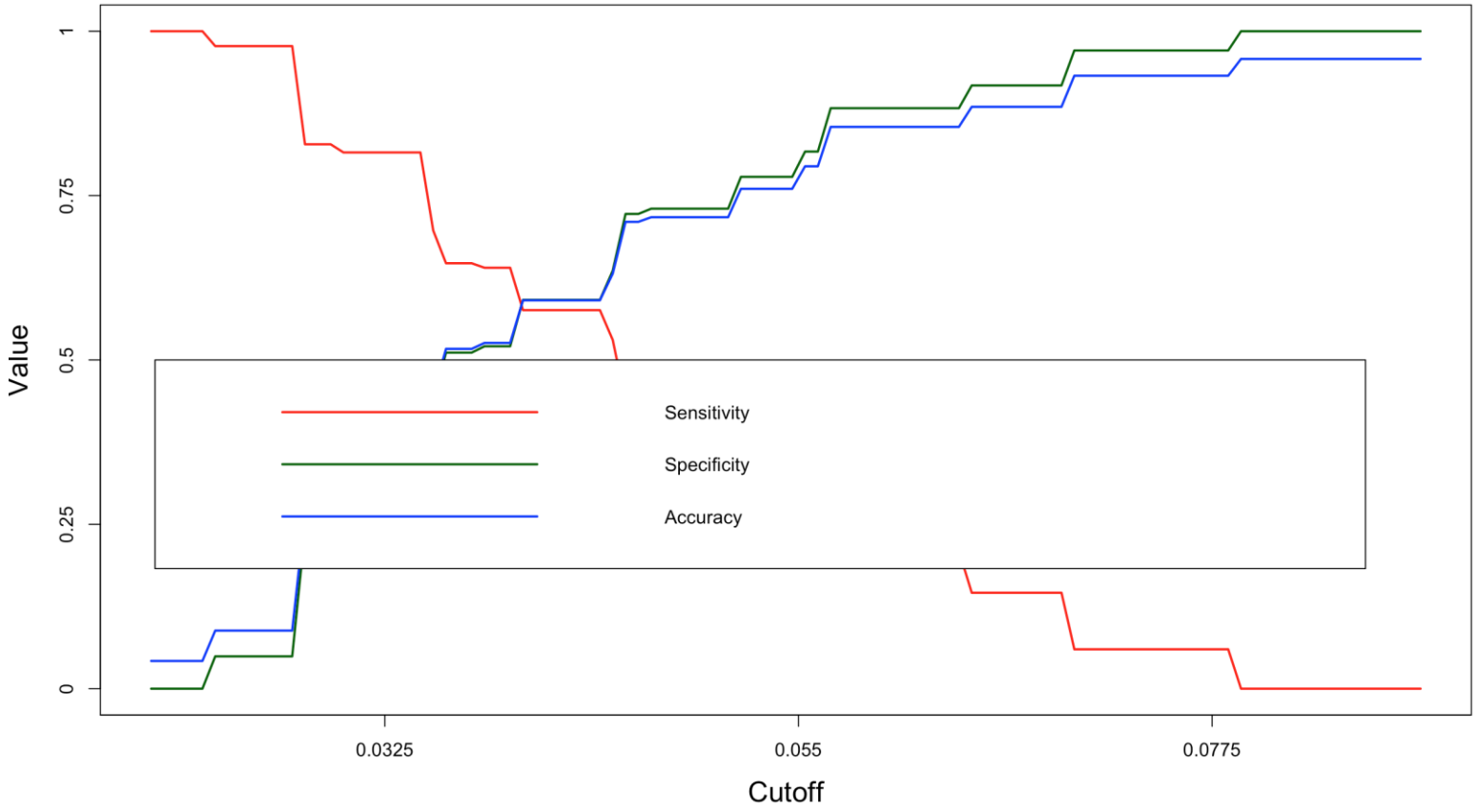
1. The class imbalance has been tackled using various techniques viz., **under, over, both and SMOTE**. **The important thing is , that it is done only on the train and not on the test**. This is done to maintain the sanctity of "unseen" test data.
2. **To support the above procedure, one may look into "ROSE" documentation, where the same approach is used in the data set "Hacide".** This way, we have ensured the generalizability of our model.

MODEL BUILDING STEPS

1. Data cleaning steps of removing outliers was done as logistic regression is very sensitive to outliers.
2. Duplicate records removed.
3. Some of the variables took only discrete values. So, they were factored. This helped in getting good results in RF algorithm.
4. Creating binned variables helped which finally appeared in the final model for various algorithms.
5. One such example is “Avgas.CC.utilization.in.last.12.months”. This variable turned out to be the most important variable in Random Forest.
6. In case Age column, initially binning was not done. Also the corresponding woe was not monotonic. We revisited Feature Engineering ,and binning Age it automatically , became monotonic.
7. We have ensured that the variables (except 2) that enter into algorithms are monotonic. This adds to the stability of the model.
8. Among various options for feature selection like KNN , SVM, LDA , we chose , IV value. The reason for this is that IV and WOE are intimately connected. The models so built have produced good accuracy.
9. The range of values for the continuous were examined. Many of them did not vary much. So, transformations like log was not suitable. However, we tried the transformations and model performance dropped. So, we also dropped transformations.
10. Another reason for not experimenting with transformations, was the values of WOE. The values for different variables did not vary greatly. Hence transformations were not done in the end.
11. As the prime objective is to make the company reduce losses, in models that we have built, the specificity(positive class “0”) has been increased, without sacrificing accuracy much.

Model Building on Demographic Data (Only)

Demographic data only has a partial influence on the defaulters. As such, we prepared only a basic model using it, without any optimization. However we used the various sampling techniques on the combined (demographic & credit) data set for analysis and scorecard preparation

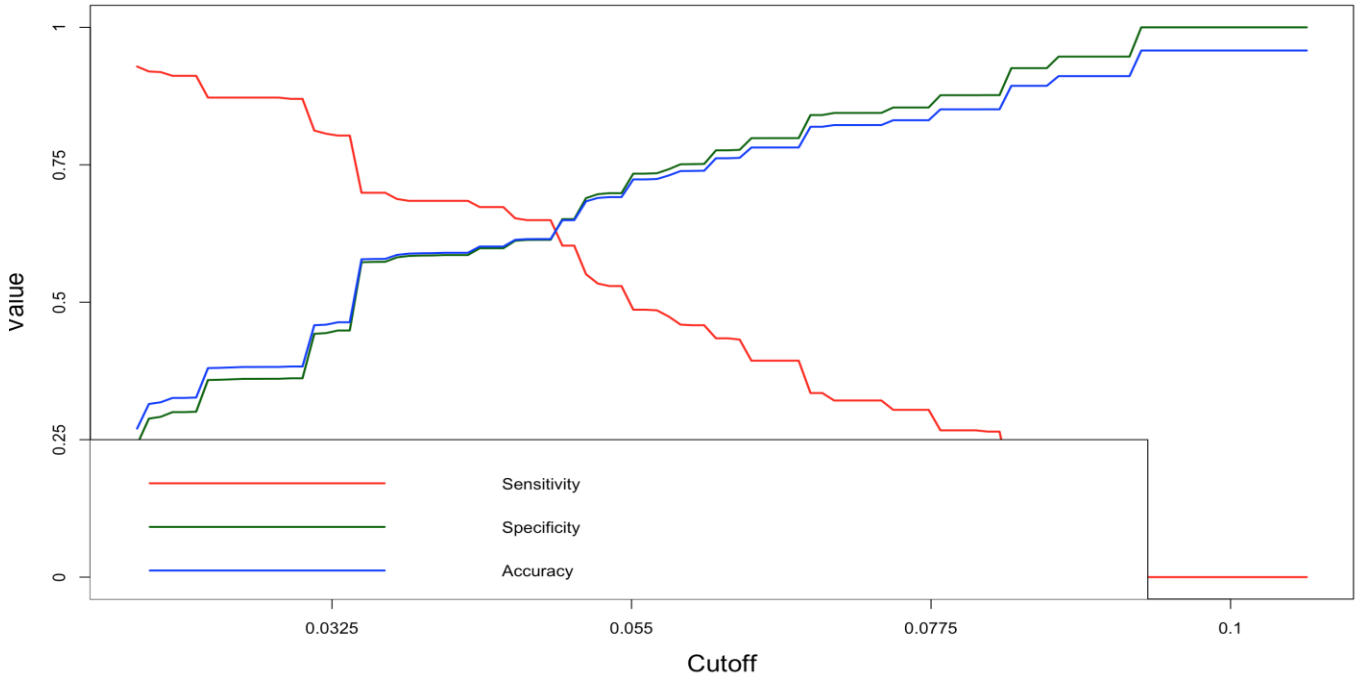


Cut-off : **0.0435**
Accuracy : **60.64%**
Sensitivity : **57.69%**
Specificity : **60.77%**

Model Building on Merged set (Logistic Regression)

- We have built 4 different models on demographic data, different sampled datasets
  - Original
  - Sampled (method = over)
  - Sampled (Using SMOTE method)
  - Under sampled

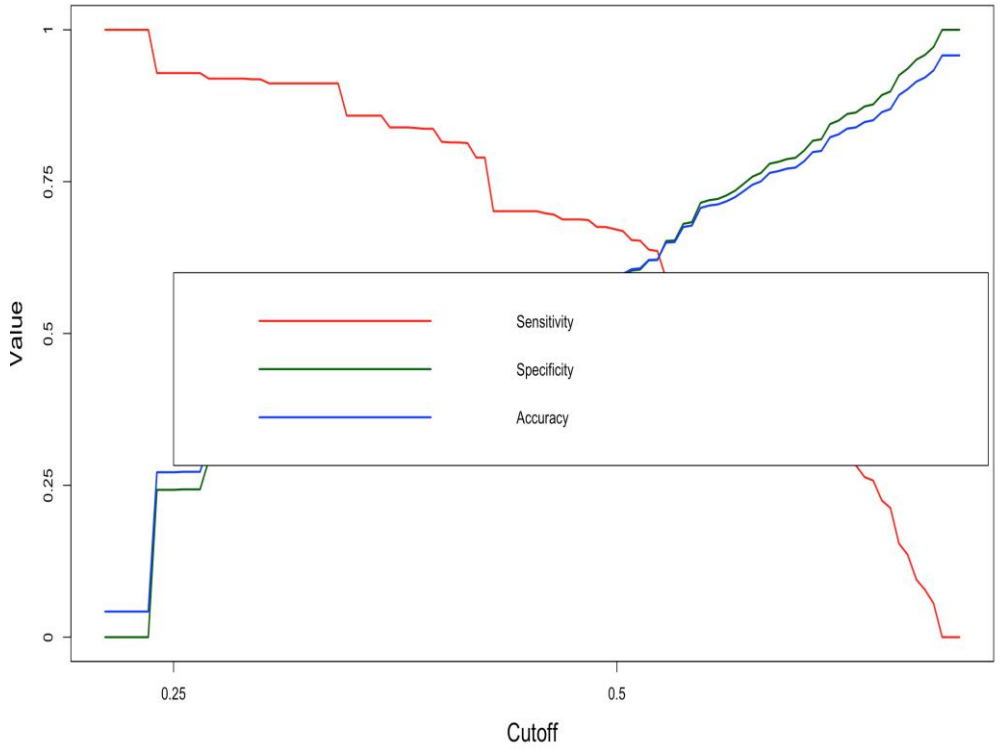
Original:



Cut-off : **0.04714**
Accuracy : **61.48%**
Sensitivity : **64.93%**
Specificity : **61.33%**

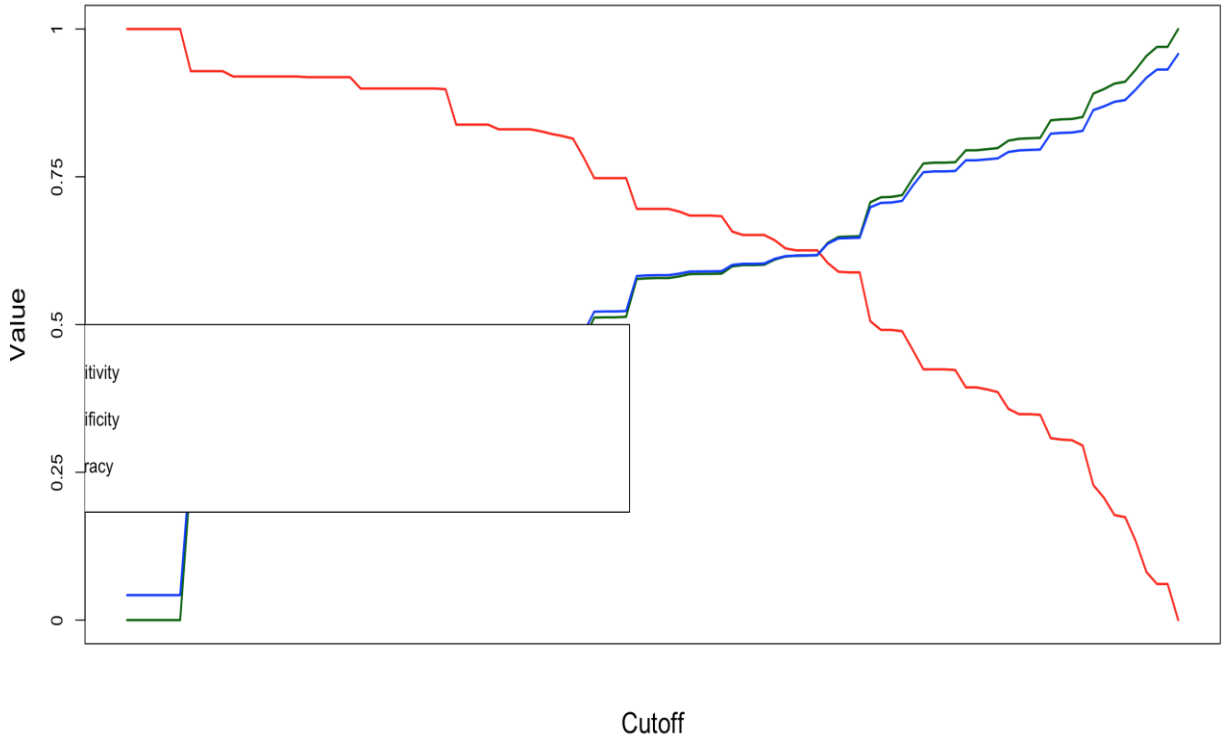
Model Evaluation (Logistic Regression)

Sampled (method = over)



Cut-off : **0.518**
Accuracy : **62.10%**
Sensitivity : **63.80%**
Specificity : **62.03%**

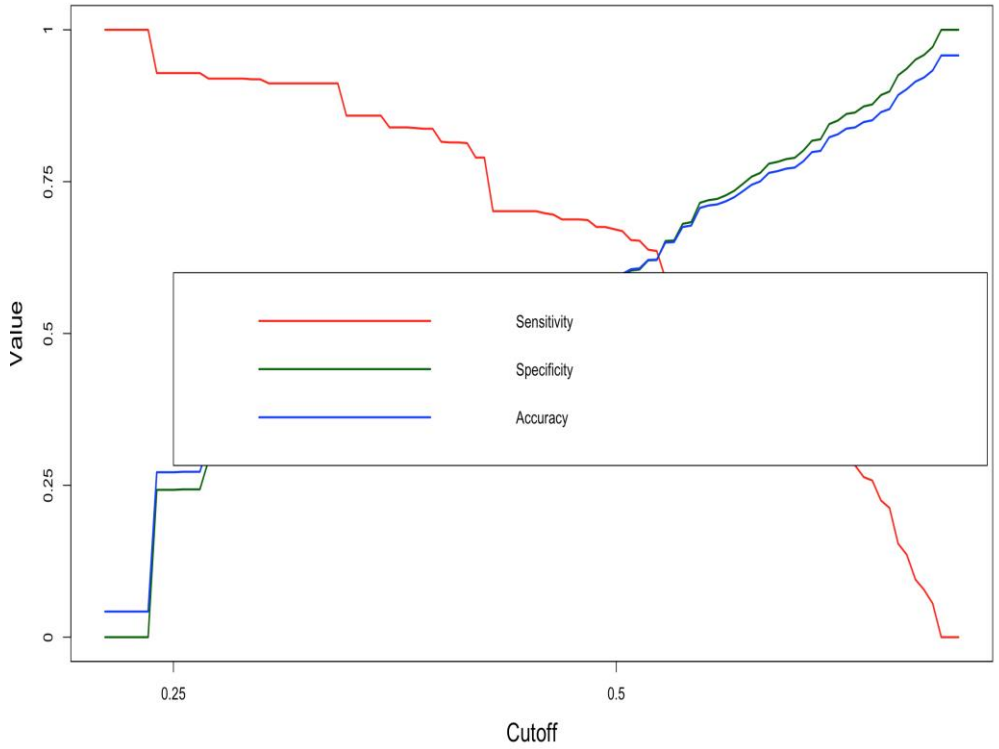
SMOTE Sampled



Cut-off : **0.542**
Accuracy : **61.12%**
Sensitivity : **64.25%**
Specificity : **60.98%**

Model Evaluation (Logistic Regression)

Under Sampled



Cut-off : **0.532**
Accuracy : **61.52 %**
Sensitivity : **63.91 %**
Specificity : **61.41 %**

From the above 4 sampled datasets for logistic regression, we could see over sampling data (method = under) giving a better result.

Model Building and Evaluation on Merged set (Decision Tree)

Over Sampled

Confusion Matrix		
	0	1
0	11784	297
1	8291	587

Performance	Percentage
Accuracy	59.02%
Sensitivity	58.69%
Specificity	66.40%

Over Sampled (SMOTE)

Confusion Matrix		
	0	1
0	13947	449
1	6128	435

Performance	Percentage
Accuracy	68.61%
Sensitivity	69.47%
Specificity	49.20%

Under Sampled

Confusion Matrix		
	0	1
0	11784	297
1	8291	587

Performance	Percentage
Accuracy	59.02%
Sensitivity	58.69%
Specificity	66.40%

From the above 3 sampled datasets for decision tree, we could see the under sampling data is giving a accuracy (59.02%) and specificity (66.40%).

Model Building on Merged set (Random Forest)

RANDOM FOREST

Random Forest , by far, is a very efficient algorithm .The algorithm produces highly accurate results.

Also, RF is known to work very well for data sets that have a large number of categorical data, like in CredX.

The RF method gave very good accuracy and was implemented with following details of hyper parameters

1) The mtry =c(4,5)
2) ntree =500
3) Node_size =seq(2,10,by=2)
4) Sampe_size =c(.55, .632, .70,.75, .80)

The parameters were selected after a few initial runs with specific values. Then the grid with the above was constructed.

RF with 500 trees creates 500 bagging samples with their boot samples for each different set of other hyper parameters . This way the model is highly superior to mere cross validation. Finally the ensemble of trees cuts down over fitting drastically.

Model Evaluation (Random Forest)

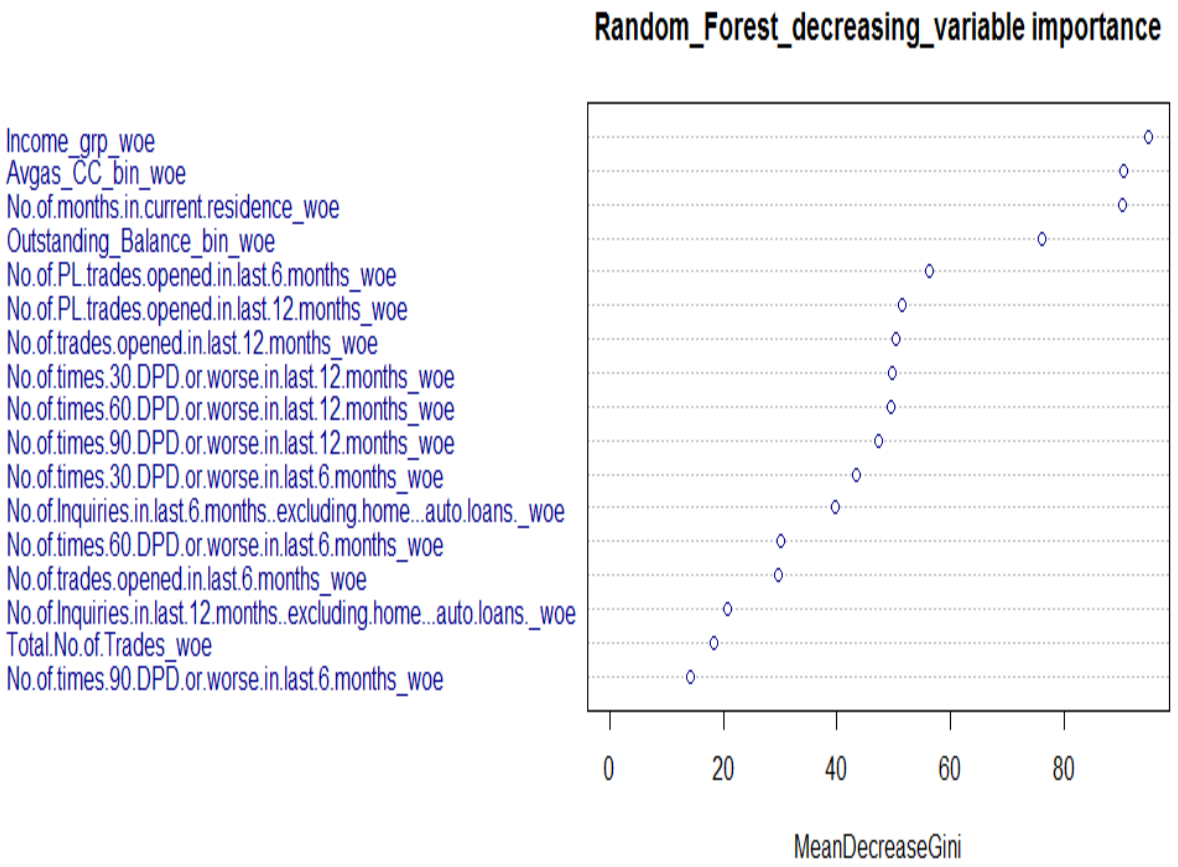
```
# Confusion Matrix and Statistics
#
#                Reference
#    Prediction     0     1
#             0 12025   313
#             1  8050   571
#
# Accuracy : 0.601
# 95% CI : (0.5943, 0.6076)
# No Information Rate : 0.9578
# P-Value [Acc > NIR] : 1
#
# Kappa : 0.0473
# Mcnemar's Test P-Value : <2e-16
#
# Sensitivity : 0.59900
# Specificity : 0.64593
# Pos Pred Value : 0.97463
# Neg Pred Value : 0.06623
# Prevalence : 0.95782
# Detection Rate : 0.57374
# Detection Prevalence : 0.58867
# Balanced Accuracy : 0.62247
#
# 'Positive' Class : 0
```

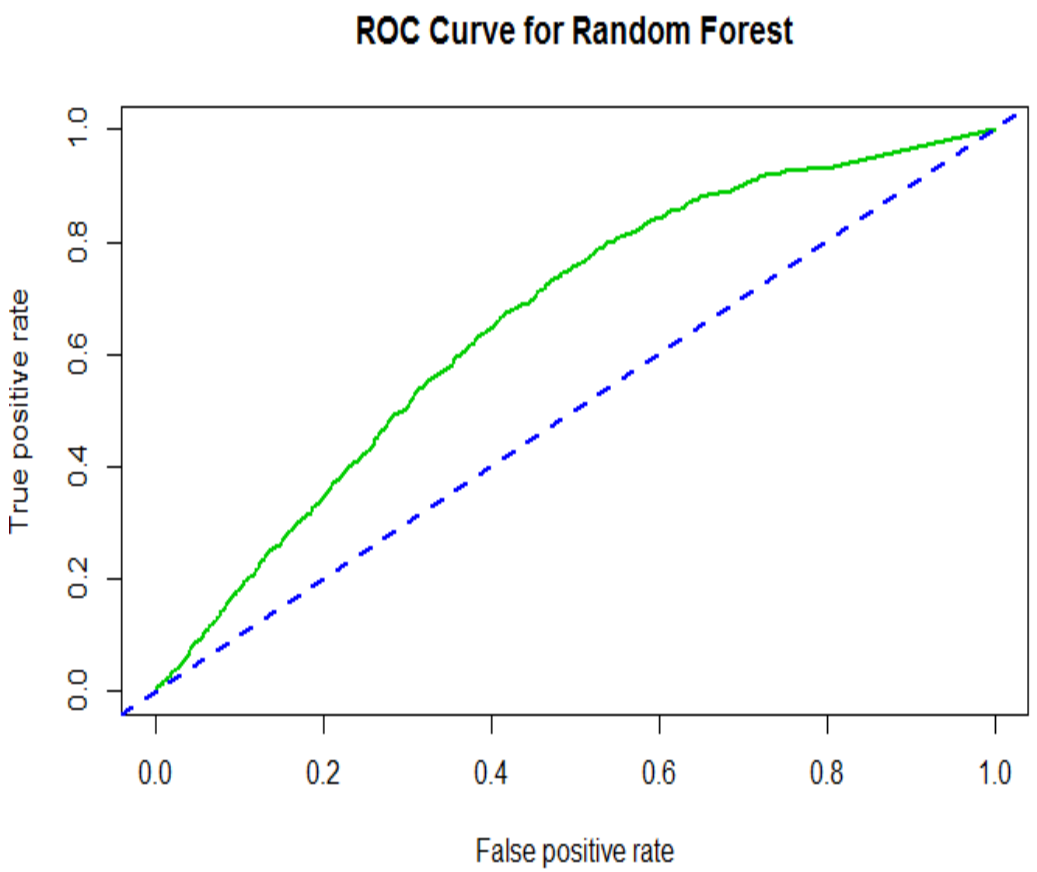
COMMENTS:MERITS OF THE MODEL

The positive class is 0(good). So, for the company to reduce the losses, because of defaulters , we have to increase (1,1) entry as much as possible, without sacrificing too much of accuracy. It can be seen that 571 out of 884 defaulters are predicted correctly , giving the specificity =64.59%.

Further, the sensitivity and accuracy are very close to each other. Since the RF algorithm is an ensemble technique, overfitting is greatly reduced.

Model Evaluation (Random Forest)





Application Scorecard & Financial benefits

- We have gone with the following approach to calculate the scores:

- “good to bad odds of 10 to 1 at a score of 400 doubling every 20 points”

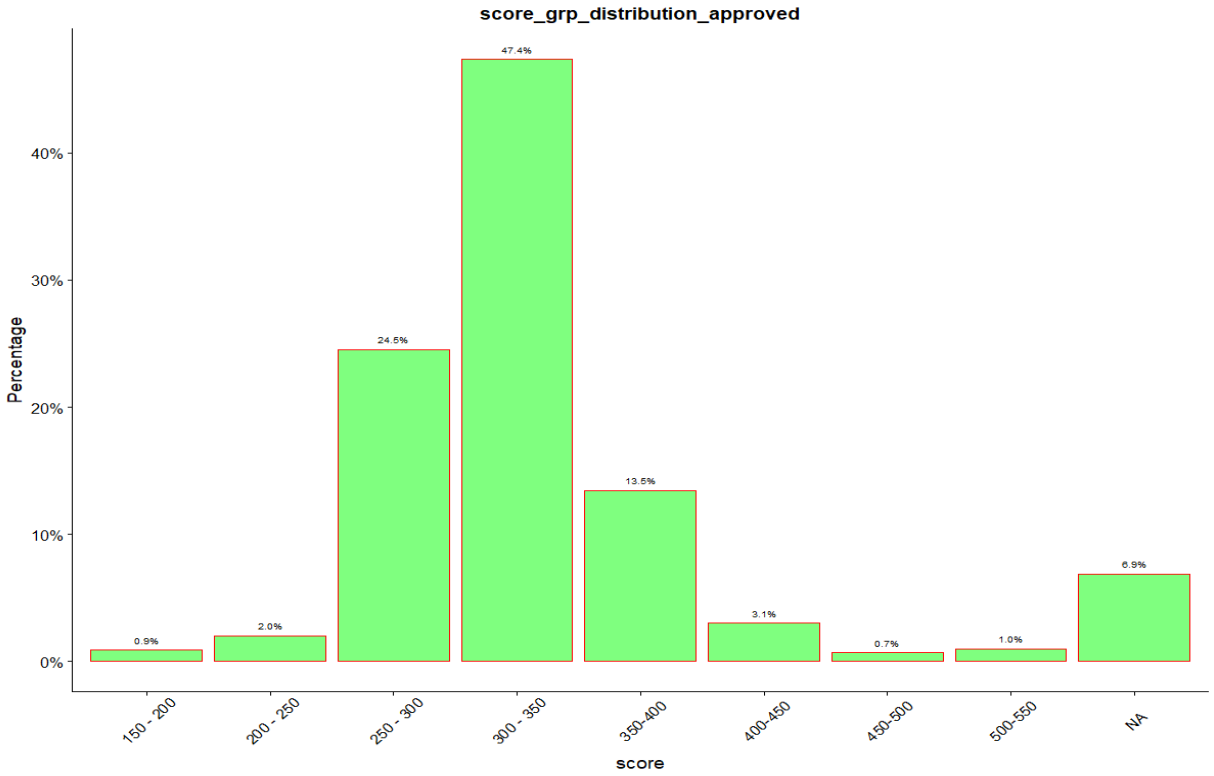
- So going with the approach of "y = mx + c"
  Score = m(ln(odds)) + c

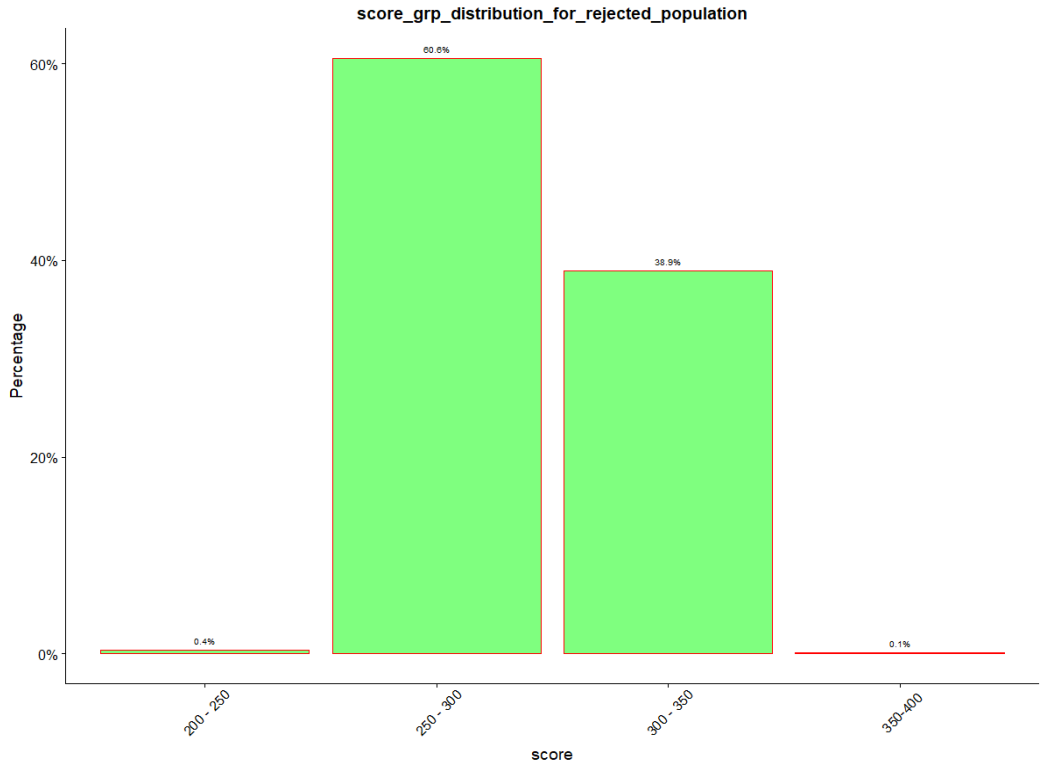
- 400 = m * (ln(10)) + c
- 420 = m * (ln(20)) + c

- 420 - 400 = m(ln(20) - ln(10))
- 20 = m (ln(10*2) - ln(10))
- 20 = m(ln(2) + ln(10) - ln(10))
- 20 = m(ln(2))
- m = 20/ln(2)
- So base score = 400 - 20/(ln(2)) * ln(10)
  = 400 - 20/log(2) * log(10)
- = 333.56

- **Hence score = (20/log(2))*(odds_good_to_bad) + (333.56)**

Comparison of application scores of approved vs rejected





- From the graph of “approved” it is clearly visible that the majority of the scores of the approved population > 300
- # Only 27% is less than 300
- # From the graph it is clearly visible that for rejected population , majority of the score is less than 300
- # Nearly 61% is less than 300

COST ANALYSIS FOR THE RANDOM FOREST MODEL

- A correct decision means that the “CredX” predicts an application to be good or credit-worthy and it actually turns out to be credit worthy. When the opposite is true, i.e. CredX predicts the application to be good but it turns out to be bad credit, then the loss is 100%..
- The main objective of the model is to reduce the losses suffered due to defaulters. It can be noted that the CredX has been very generous in approving the loans, since it resulted in losses.
- Let us assume one defaulter costs $1000 to the CredX
- Without the model , the loss is 884*1000= $ 884000
- With the model , the loss reduces by 571*1000= $ 571000
- If the CredX predicts an application to be non-creditworthy, then loan facility is not extended to that applicant and bank does not incur any loss. However it suffers opportunity loss .By applying the Random Forest model, to applicants where the Performance Tag is missing, we are able to capture 71 good customers.
- If need be, a score that is slightly more than the cut off score of 333.56 can be set. This will again further bring down the number of defaulters, thereby reducing the loss further. The risk appetite depends on the financial position or profit of the CredX.