BFS CAPSTONE PROJECT

Group Members:

SANDEEP KUMAR DUGYALA

SHARATH CHANDAN REDDY SAMA

SHAYAK ROY

SIBANANDA SARANGI

Business Requirement

- **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.
- The Objective is to reduce the Credit Loss by finding the Right Customers.

Solution

- We ued CRISP-Framework for solving the problem

where we have applied mutiple prediction models and used the best one to predict the defaulters and list down the customers using score card mechanism and Did financial loss and befit analysis.

CRISP Framwork

- Business Understanding and Data Understanding
- Data Cleansing and Preparation
- EDA
- Data Transformation and Model Building
- Model Evaluation

Business Understanding

After digging into the Business Requirement, found there are some other low level objectives also need to accomplised the to resolve the issue and listed them below:

- To mitigate credit risk to ‘acquire the right customers’.
- Identify the right customers using predictive models. Using past data of the bank’s applicants
- To determine the factors affecting credit risk
- Create strategies to mitigate the acquisition risk and assess the financial benefit to CredX

Data Understanding

- The Company has provided the two datasets.
  1. *Demographic/application data*
  2. *Credit bureau data*
- In both the datasets, the performance tag has the values of 0 (non-default) and 1 (default).
- **Demographic/ Application data :**
  a. The Demographic data has the information which was provided by the credit card applicant.
  b. Like age, gender, income, marital status, education, No. of moths in current company, etc.
  c. The demographic data has total of **71295** Observations and among them 3 customers are duplicated.
- **Credit bureau data :**
  a. The Credit data has the information of whether the customer is paying the dues in time or not like number of time 30 DPD or worse in last 3/6/12 months and so on..
  b. The Credit Bureau data has total of **71295** Observations and among them 3 customers are duplicated.

Data Cleaning :

- **Handling missing Values:**
  1. For **EDA**, **KNN imputation** of the dataset has been done.
  2. For **modeling** purpose the **WOE** imputation method was done.
  3. **NA** values in Performance tag were not imputed.
- **Outlier treatment :**
  1. Outlier treatment was done by using Quantile and Box Plots.
  2. The capping of the outliers treated as follows for the categorical variables:
     a. Capped the age which has <15 to 15
     b. Capped the Income which has <0 to 0
     c. Capped the No.of.months.in.current.company which has >74 to 74.
     d. Replaced the Null values of No.of.trades.opened.in.last.6.months with 0.
- **Duplicate Value** handling of the application ID was checked using duplicated command in R and the first entry of duplicate values are kept and the remaining were removed.
- **IV and WOE Analysis:**

  Below is the interpretation of the IV value has been taken in to consideration:

  a. < 0.02 Useless for prediction
  b. 0.02 to 0.1 Weak predictor
  c. 0.1 to 0.3 Medium predictor
  d. 0.3 to 0.5 Strong predictor

Data Sub-setting Approach

- **Creating Datasets**
  - ➢ Before proceeding with EDA approach, 2 subsets of whole dataset was created:
    a. Dataset having Performance.Tag as NA
    b. Dataset not having Performance.Tag as NA
- **Merging Datasets:**
  - ➢ Based on above the two datasets were merged based on common Application.ID variable.

- An attempt was made to check if the Performance.Tag in both datasets match for same Application.ID, and consecutively one of the tag column was removed.

- **Data Sub-Setting before Model creation:**
  - ➢ A Train and Test Dataset are created before Model creation based on Independent Variable, having the ratio of 7:3 respectively.

*The approach can change slightly while model building ahead.

IV Analysis

Below are the Results obtained for the **IV** and mentioned the type of predictor based on the Analysis.

Variable	IV	Type of Predictor Variable
Avgas.CC.Utilization.in.last.12.months	0.3099292	STRONG
No.of.trades.opened.in.last.12.months	0.2979723	STRONG
No.of.PL.trades.opened.in.last.12.months	0.2958971	STRONG
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	0.2954176	STRONG
Outstanding.Balance	0.2462796	MEDIUM
No.of.times.30.DPD.or.worse.in.last.6.months	0.2415512	MEDIUM
Total.No.of.Trades	0.2366296	MEDIUM
No.of.PL.trades.opened.in.last.6.months	0.2197272	MEDIUM
No.of.times.90.DPD.or.worse.in.last.12.months	0.2138633	MEDIUM
No.of.times.60.DPD.or.worse.in.last.6.months	0.2058259	MEDIUM

WOE Analysis

Below are the WOE values obtained for each variable:

No.of.times.90.DPD.or.worse.in.last.6.months	N	Percent	WOE	IV
[0,0]	54664	0.7824	-0.2606781	0.04725916
[1,3]	15203	0.2176	0.622455	0.16010599
No.of.times.60.DPD.or.worse.in.last.6.months	N	Percent	WOE	IV
[0,0]	51870	0.74241	-0.3363664	0.07220016
[1,5]	17997	0.25759	0.6225361	0.20582586
No.of.times.30.DPD.or.worse.in.last.6.months	N	Percent	WOE	IV
[0,0]	50098	0.71705	-0.3867918	0.09018455
[1,1]	9500	0.13597	0.4643187	0.12658538
[2,7]	10269	0.14698	0.7428448	0.24155115
No.of.times.90.DPD.or.worse.in.last.12.months	N	Percent	WOE	IV
[0,0]	50492	0.72269	-0.3566331	0.07830347
[1,1]	11663	0.16693	0.5088234	0.13311253
[2,5]	7712	0.11038	0.7219824	0.21386327
No.of.times.60.DPD.or.worse.in.last.12.months	N	Percent	WOE	IV
[0,0]	45868	0.6565	-0.3519211	0.06940922
[1,1]	12816	0.18343	0.2141538	0.078697
[2,7]	11183	0.16006	0.6940858	0.18548895
No.of.times.30.DPD.or.worse.in.last.12.months	N	Percent	WOE	IV
[0,0]	44857	0.64203	-0.376396	0.07681744
[1,2]	17590	0.25176	0.2805525	0.09938446
[3,9]	7420	0.1062	0.7994935	0.198241
Avgas.CC.Utilization.in.last.12.months	N	Percent	WOE	IV
NA	1023	0.01464	0.11147371	0.000191525
[0,4]	5524	0.07906	-0.80175843	0.035981452
[5,6]	5471	0.07831	-0.80150239	0.071409116
[7,8]	6869	0.09832	-0.79452312	0.115245013
[9,11]	9597	0.13736	-0.67240619	0.161411607
[12,14]	6595	0.09439	-0.46800192	0.178186893
[15,21]	6854	0.0981	-0.07900431	0.178777535
[22,37]	7122	0.10194	0.47504577	0.207487506
[38,51]	6746	0.09655	0.584589	0.250865829
[52,71]	7017	0.10043	0.56373098	0.292410668
[72,113]	7049	0.10089	0.38134147	0.309929165
No.of.trades.opened.in.last.6.months	N	Percent	WOE	IV
[0,0]	12194	1.75E-01	-0.6576285	0.0564612
[1,1]	20121	2.88E-01	-0.4795153	0.1099231
[2,2]	12116	1.73E-01	0.232861	0.1203956
[3,3]	9402	1.35E-01	0.4351239	0.1515994
[4,4]	6297	9.01E-02	0.5242769	0.1832473
[5,12]	9736	1.39E-01	0.1368556	0.1860271

No.of.trades.opened.in.last.12.months	N	Percent	WOE	IV
[0,0]	4956	0.07093	-0.65346215	0.02269765
[1,1]	11377	0.16284	-1.01908605	0.13168755
[2,2]	9323	0.13344	-0.81646884	0.19394762
[3,3]	4678	0.06696	0.003598878	0.19394849
[4,5]	9397	0.1345	0.109294271	0.19563796
[6,7]	8297	0.11875	0.447981607	0.22500374
[8,9]	7175	0.1027	0.571340073	0.26879653
[10,12]	6699	0.09588	0.491781025	0.29796776
[13,28]	7965	0.114	0.006306206	0.2979723
No.of.PL.trades.opened.in.last.6.months	N	Percent	WOE	IV
[0,0]	31080	0.44485	-0.6492118	0.1407488
[1,1]	13546	0.19388	0.1993619	0.1491979
[2,2]	12565	0.17984	0.4384356	0.1916027
[3,6]	12676	0.18143	0.3619618	0.2197272
No.of.PL.trades.opened.in.last.12.months	N	Percent	WOE	IV
[0,0]	25824	0.36962	-0.8938108	0.2002061
[1,1]	6641	0.09505	-0.1310168	0.2017433
[2,2]	6830	0.09776	0.2513399	0.2086806
[3,3]	8130	0.11636	0.4122959	0.2326462
[4,4]	7903	0.11311	0.5000753	0.2683711
[5,5]	6189	0.08858	0.4261494	0.2879895
[6,12]	8350	0.11951	0.2431575	0.2958971
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	N	Percent	WOE	IV
[0,0]	25069	0.35881	-0.71823049	0.134963
[1,1]	13175	0.18857	0.1770721	0.141379
[2,2]	12831	0.18365	0.21609676	0.1508557
[3,4]	11506	0.16468	0.50980053	0.20516
[5,10]	7286	0.10428	0.01241548	0.2051762
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.	N	Percent	WOE	IV
[0,0]	20581	0.29457	-1.06753664	0.2122103
[1,1]	3899	0.05581	-0.06177455	0.2124173
[2,2]	7907	0.11317	0.14214469	0.2148588
[3,3]	8978	0.1285	0.16434931	0.218603
[4,4]	7113	0.10181	0.24810534	0.2256323
[5,5]	4927	0.07052	0.58818059	0.2577593
[6,8]	8951	0.12811	0.48413154	0.2953973
[9,20]	7511	0.1075	0.01370484	0.2954176
Presence.of.open.home.loan	N	Percent	WOE	IV
NA	272	0.00389	-0.37379739	0.000459914
[0,0]	51524	0.73746	0.07370543	0.004604115
[1,1]	18071	0.25865	-0.23665793	0.017619389

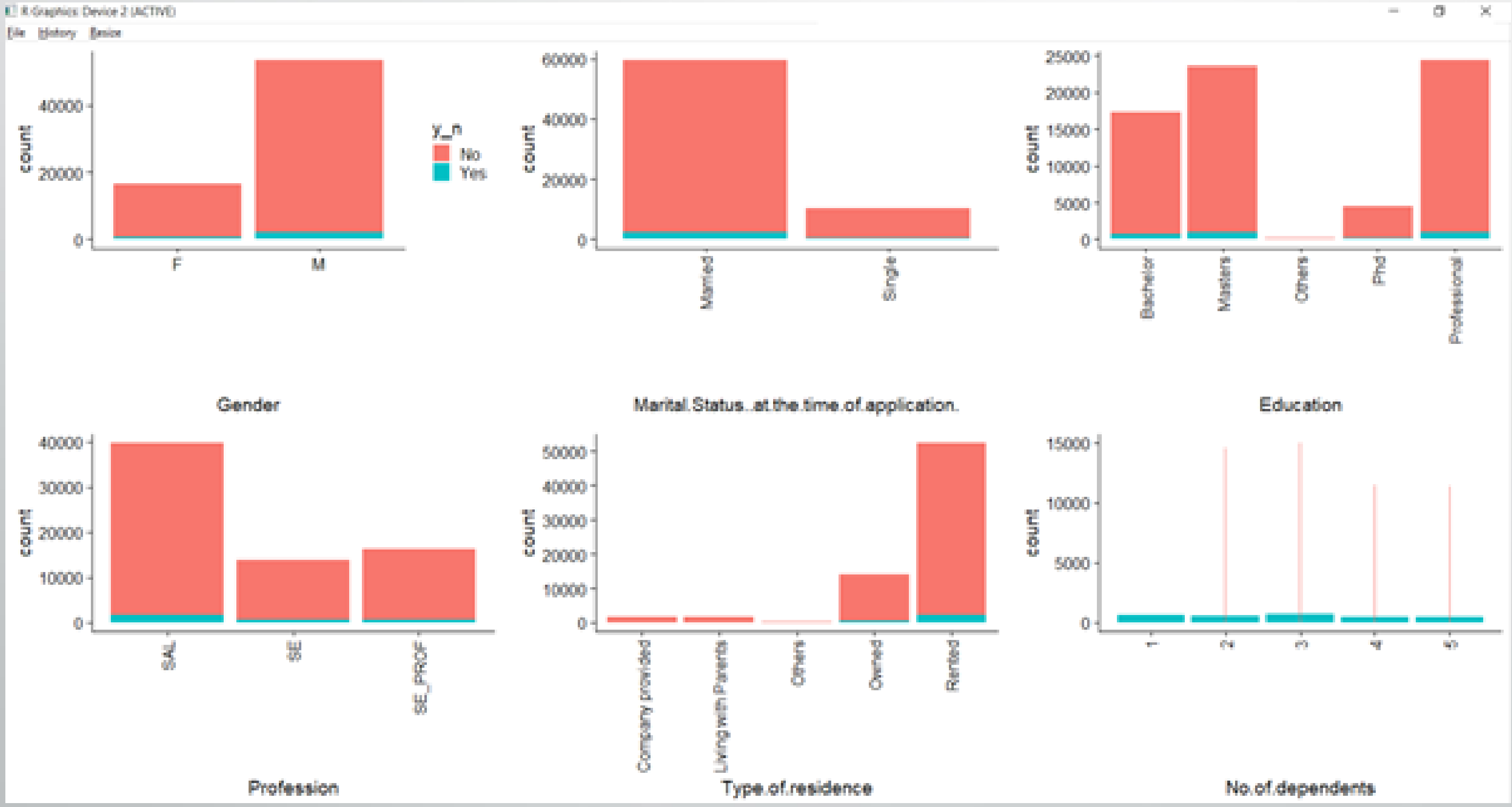
WOE Analysis Contd..

Outstanding.Balance	N	Percent	WOE	IV
NA	272	0.00389	-0.3737974	0.000459914
[0,6843]	6958	0.09959	-0.770284	0.042618466
[6847,25509]	6960	0.09962	-0.9203411	0.099212283
[25522,386813]	6959	0.0996	-0.1343423	0.100903361
[386815,585402]	6960	0.09962	0.2542645	0.108148044
[585423,774228]	6960	0.09962	0.4532364	0.133425761
[774241,972455]	6959	0.0996	0.434264	0.156421066
[972456,1357300]	6959	0.0996	0.4049624	0.176143184
[1357399,2960998]	6960	0.09962	-0.3824181	0.188414102
[2961005,3282314]	6960	0.09962	-0.831026	0.236277242
[3282409,5218801]	6960	0.09962	0.2958682	0.246279591
Total.No.of.Trades	N	Percent	WOE	IV
[0,1]	3914	0.05602	-0.67304028	0.01885887
[2,2]	6766	0.09684	-1.0177255	0.08353841
[3,3]	8615	0.12331	-0.70202474	0.12815199
[4,4]	7490	0.1072	-0.44785257	0.14575218
[5,5]	5714	0.08178	-0.04880056	0.14594265
[6,6]	4966	0.07108	0.12930127	0.1472039
[7,8]	9361	0.13398	0.37936559	0.1702064
[9,10]	7133	0.10209	0.54394026	0.20915717
[11,19]	8476	0.12132	0.42717578	0.23616781
[20,44]	7432	0.10637	-0.06689796	0.23662955
Presence.of.open.auto.loan	**N**	**Percent**	**WOE**	**IV**
[0,0]	63937	0.91512	0.01198467	0.000132165
[1,1]	5930	0.08488	-0.13836752	0.001658061
Age	**N**	**Percent**	**WOE**	**IV**
[15,30]	5948	0.08513	-0.04194167	0.000146916
[31,35]	6927	0.09915	0.034531539	0.000267026
[36,38]	6924	0.0991	0.069071901	0.000755077
[39,41]	7129	0.10204	0.068297625	0.001246199
[42,44]	7007	0.10029	-0.03794166	0.001388093
[45,47]	6830	0.09776	-0.00395867	0.001389622
[48,50]	6743	0.09651	-0.01262931	0.001404927
[51,53]	6841	0.09791	-0.13690543	0.003129372
[54,57]	7619	0.10905	0.043405263	0.003338956
[58,65]	7899	0.11306	-0.01001341	0.003350241
Gender	**N**	**Percent**	**WOE**	**IV**
F	16506	0.23625	0.0321743	0.000248196
M	53361	0.76375	-0.0101473	0.000326473
Marital.Status..at.the.time.of.application.	**N**	**Percent**	**WOE**	**IV**
Married	59550	0.85233	-0.00409243	1.42E-05
Single	10317	0.14767	0.023326708	9.55E-05
No.of.dependents	**N**	**Percent**	**WOE**	**IV**
[1,1]	15218	0.21781	0.04008522	0.000356483
[2,3]	15129	0.21654	-0.0852904	0.001871613
[3,3.4]	15647	0.22395	0.05402167	0.002541594
[4,4]	11998	0.17173	-0.02520439	0.002649435
[5,5]	11875	0.16997	0.00439087	0.002652719

Income	**N**	**Percent**	**WOE**	**IV**
[-0.5,5]	6330	0.0906	0.3024689	0.009536858
[6,10]	6510	0.09318	0.27575091	0.017587277
[11,16]	7923	0.1134	0.06608894	0.018097848
[17,21]	6803	0.09737	0.08080252	0.018757634
[22,26]	6828	0.09773	0.02506399	0.018819737
[27,31]	6817	0.09757	0.07864867	0.019445483
[32,36]	6830	0.09776	-0.15595501	0.021660491
[37,41]	6723	0.09623	-0.26368117	0.027599526
[42,48]	7784	0.11141	-0.17686352	0.030815758
[49,60]	7319	0.10476	-0.36078566	0.042410776
Education	**N**	**Percent**	**WOE**	**IV**
Bachelor	17333	0.24809	0.016850834	7.10E-05
Masters	23525	0.33671	0.007039474	8.77E-05
Others	142	0.00203	0.429585621	5.46E-04
Phd	4483	0.06416	-0.02283908	5.79E-04
Professional	24384	0.34901	-0.0179314	6.90E-04
Profession	**N**	**Percent**	**WOE**	**IV**
SAL	39683	0.56798	-0.0283297	0.000449978
SE	13927	0.19934	0.0912735	0.00218178
SE_PROF	16257	0.23269	-0.01336187	0.00222307
Type.of.residence	**N**	**Percent**	**WOE**	**IV**
Company provided	1603	0.02294	0.080146599	0.000152906
Living with Parents	1784	0.02553	0.0640031	0.000260624
Others	199	0.00285	-0.53571007	0.000904732
Owned	14004	0.20044	0.004074027	0.000908065
Rented	52277	0.74824	-0.004294	0.000921834
No.of.months.in.current.residence	**34694**	**Percent**	**WOE**	**IV**
[6,9]	6922	0.49657	-0.27219153	0.03253516
[10,28]	7210	0.09907	0.4987231	0.06363656
[29,49]	6988	0.1032	0.30118432	0.07440068
[50,72]	6931	0.10002	0.13401754	0.07631146
[73,97]	7122	0.0992	0.13948089	0.07836953
[98,126]		0.10194	-0.07705956	0.07895394
No.of.months.in.current.company	**N**	**Percent**	**WOE**	**IV**
[3,5]	6689	0.09574	0.09851585	0.000972248
[6,12]	6798	0.0973	0.17548049	0.004221095
[13,19]	6933	0.09923	0.20630691	0.008866978
[20,26]	6919	0.09903	0.03919674	0.009021888
[27,33]	7104	0.10168	-0.08567605	0.009739658
[34,40]	7182	0.1028	0.03079397	0.009838521
[41,47]	7217	0.1033	-0.1761485	0.012797367
[48,53]	6169	0.0883	-0.21792183	0.016596495
[54,61]	7824	0.11198	-0.21640008	0.021351021
[62,133]	7032	0.10065	0.06288591	0.021760709

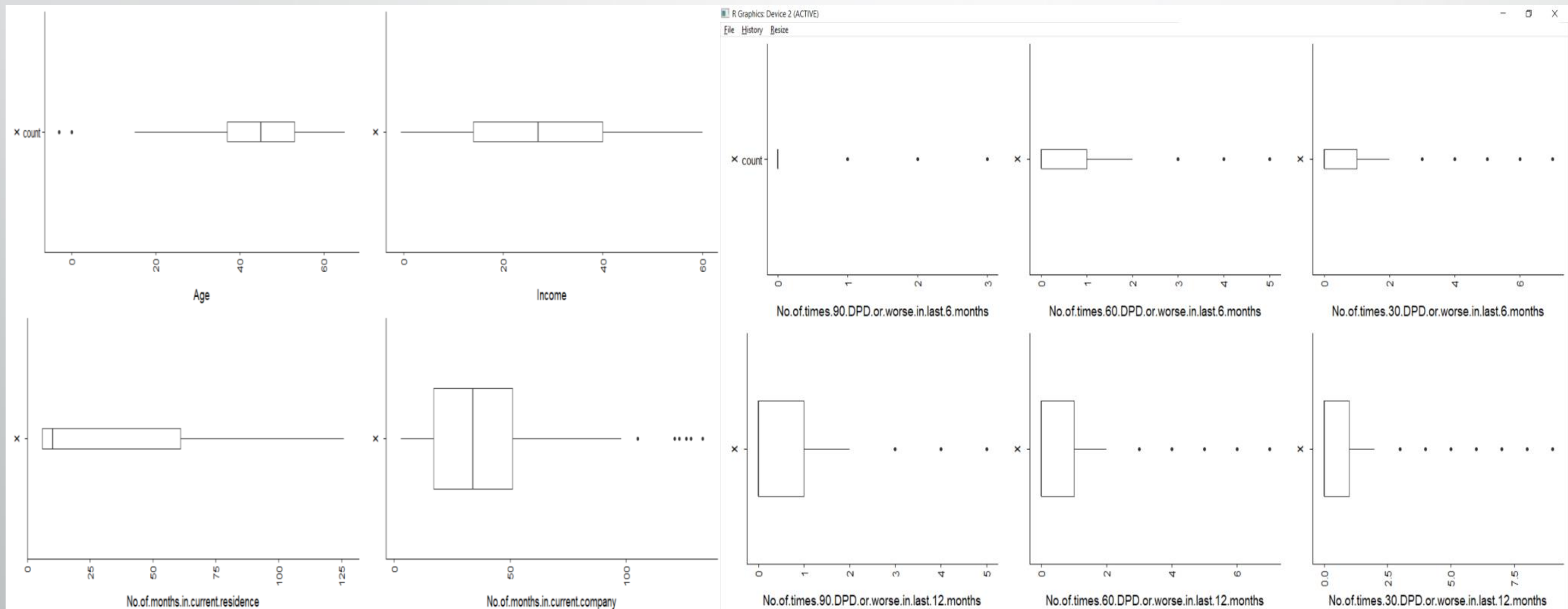
Analysis of Categorical variables

Below Plot showing for the categorical Variables of Demographic data:



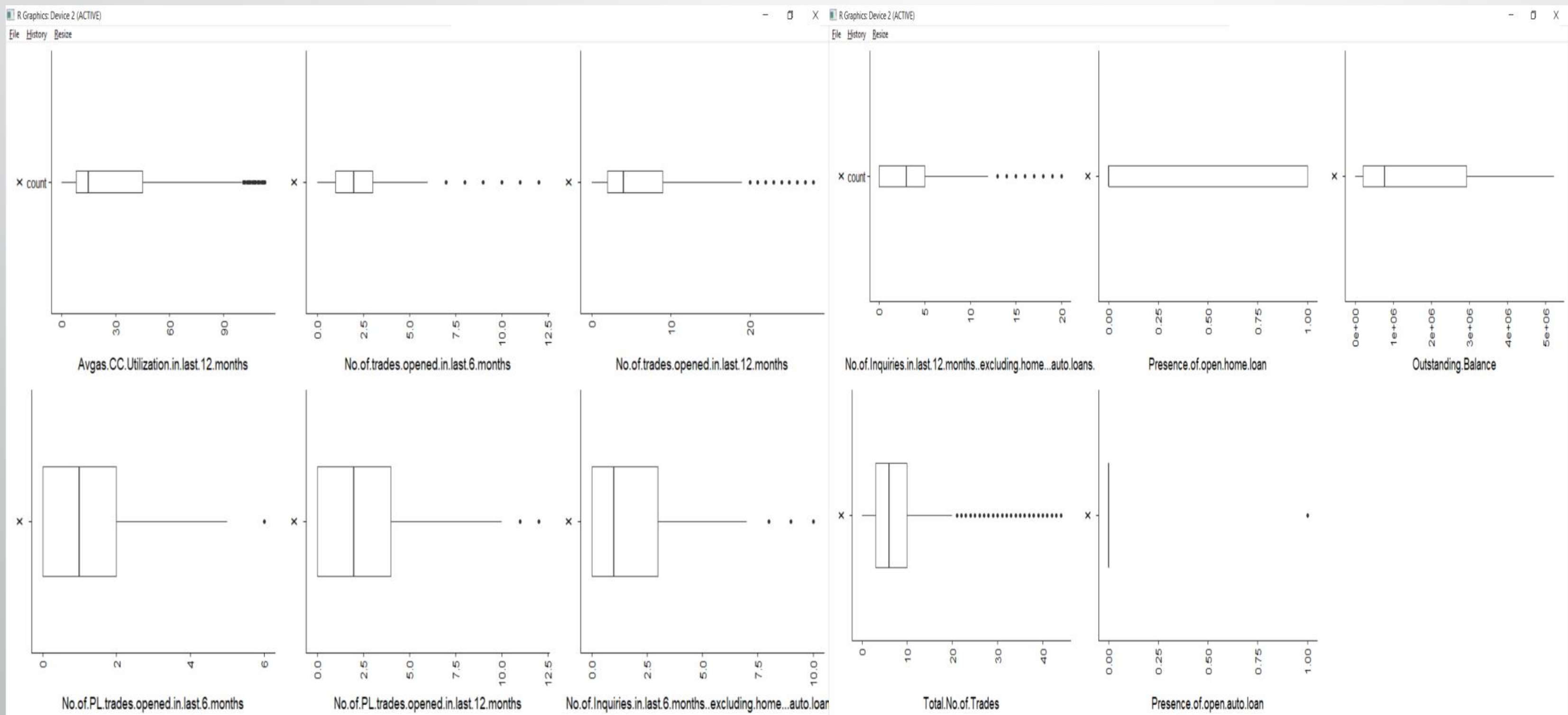
Box plot for Outlier Treatment

Below Plot showing the data for the Outliers:



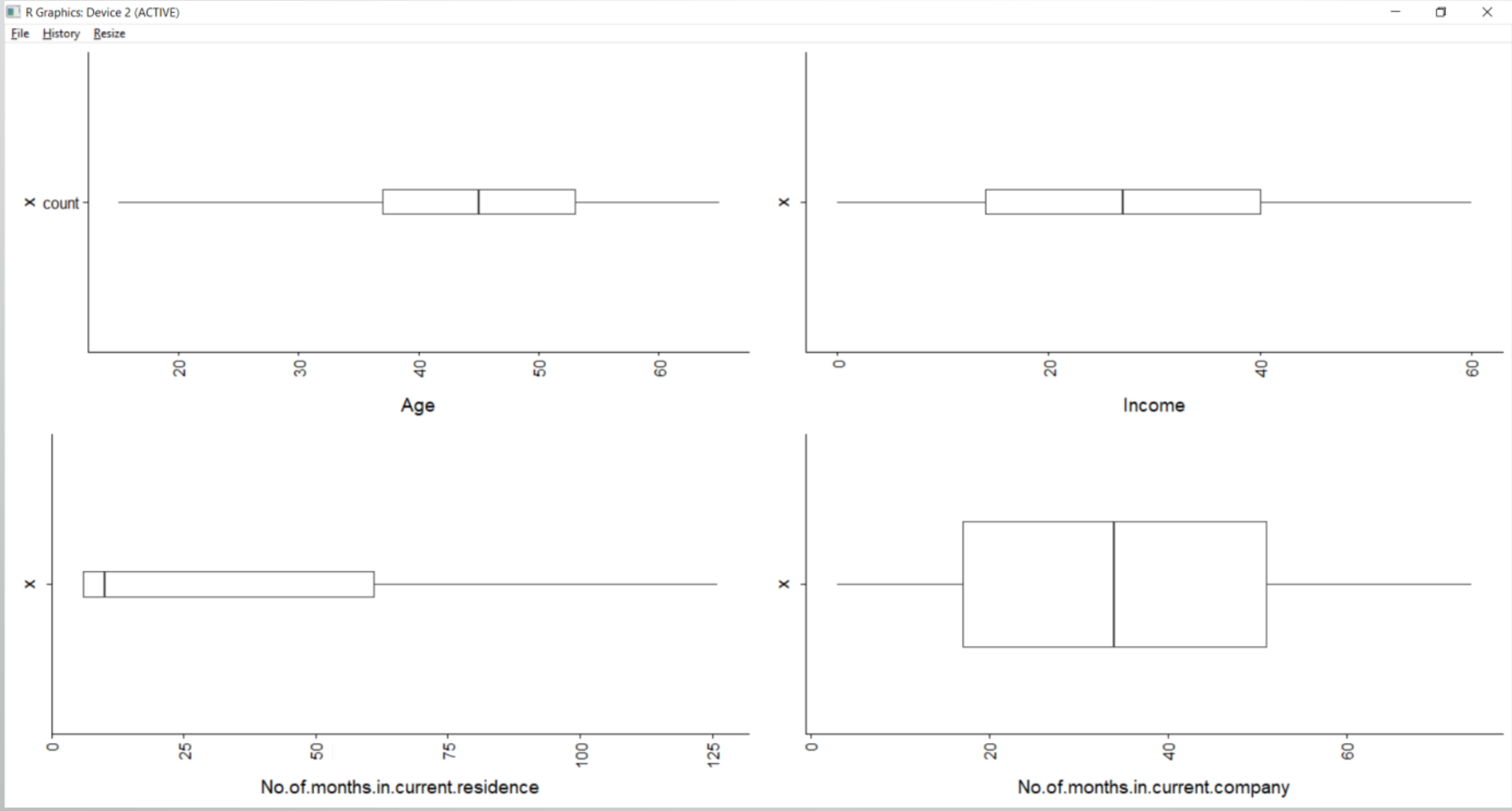
Box plot for Outlier Treatment contd…

Below Plot showing the data for the Outliers:



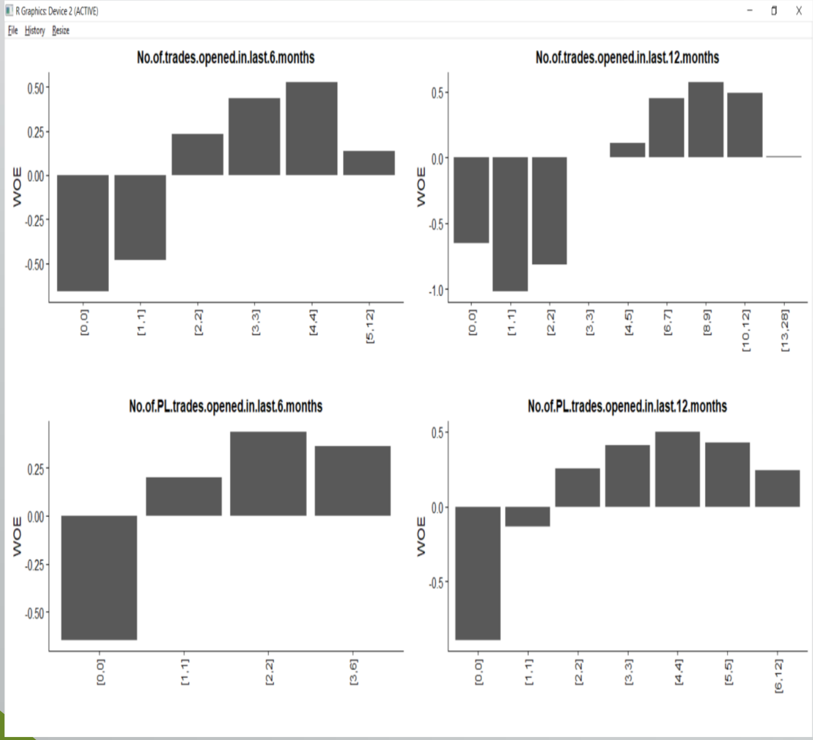
Box plot for Outlier Treatment contd…

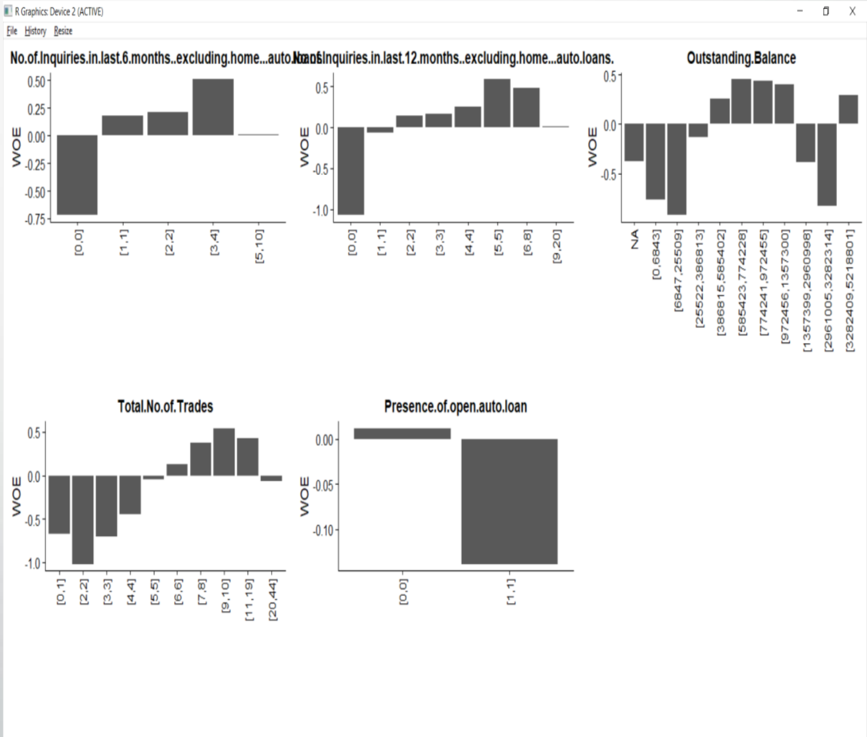
- After removing the Outliers manually for some(Categorical Variables), below are the charts obtained.



Outlier Treatment with WOE

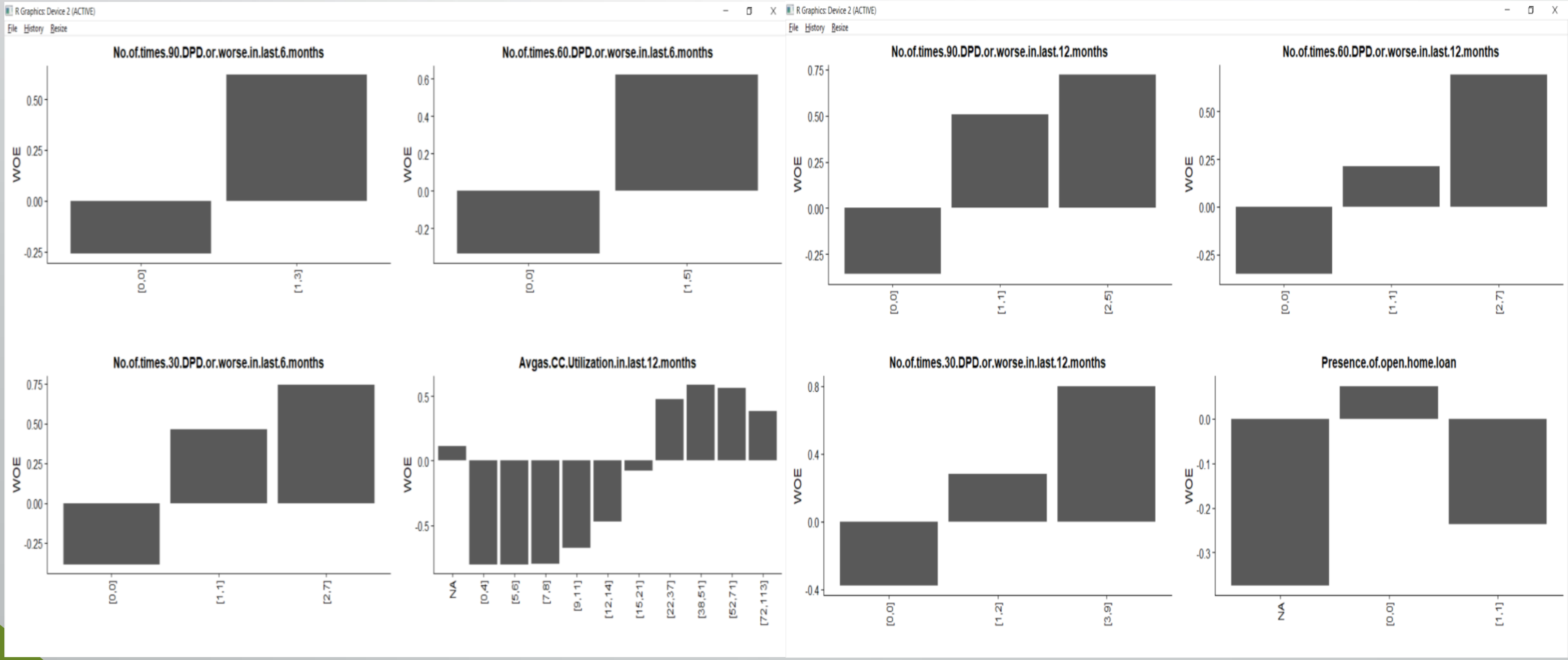
The nature of the graphs for continuous variables are monotonous, which implies the binning of WOE values are proper.





Outlier Treatment with WOE contd…

- As mentioned earlier , some(Categorical Variables) of the Outliers are done manually and while some are automatically treated by replacing with WOE Values.
- The values of WOE having negative are less likely to default and those having positive values are more likely to default.



Model Selection:

1. We choose Logistic regression as our first approach as it serves two purpose:

    a. It gives idea about driving variables

    b. It will act as baseline for other models.

2. Then we will opt for decision tree and compare the results with the logistic regression model.

3. Then among SVM and Random forest models, random tree model is preferred because of constraint in time and performance and also random forests have many advantages over decision trees like it is hard to over fit and there is no need of pruning trees in random forest.

4. Found Logistic and Random forest are perferable models.

Model Evaluation

Model evaluation Metrics to be used:

1. KS Statistics

2. Plotting ROC curve

3. Lift and Gain chart

4. R-Square

5. Confusion Metrics

.

Model Evaluation

Models	Sensitivity	Specificity	Accuracy
LogisticRegression(Demographic) using scaling and KNN imputation	38.43%	69.93%	68.57%
LogisticRegression on Merged dataset (WOE replacement)	64.77%	63.40%	63.46%
LogisticRegression on Merged dataset (Smote)	63.96%	64.11%	64.11%
RandomForest	60.70%	61.46%	61.43%
RandomForest (Smote)	60.71%	63.20%	63.11%

Score card- Accepted Data

Application Scorecard:

Application scorecard will be calculated and the result will be assessed on the rejection candidates. Application scorecard cutoff will be calculated for granting the credit.

Formula used for the application scorecard is mentioned below:

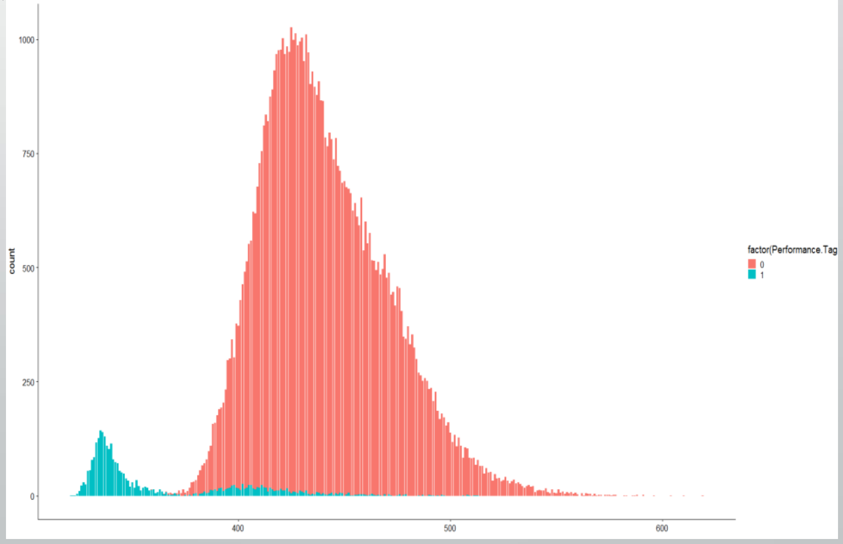
Application score with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 point

Score = floor(400 + ((20/(log(2))) * (odds-(log(10)))))

Min and Max Score: 319 and 592 Cut off : 409

Below Cutoff: 2514

Above Cutoff: 433



Score card- Rejected Data

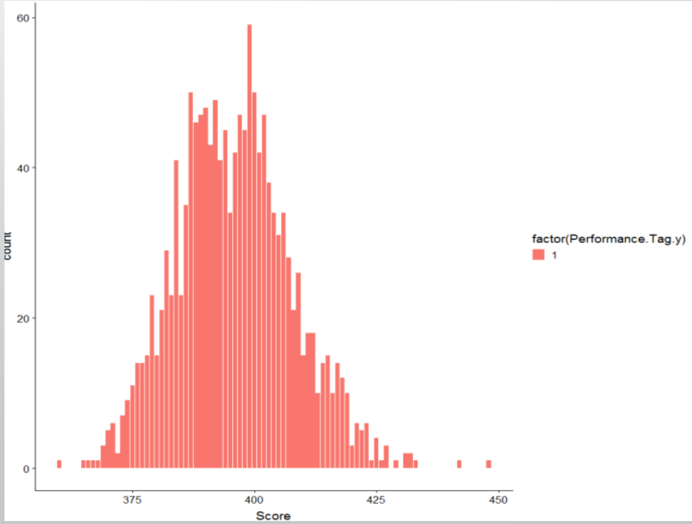
Application score with the good to bad odds of 10 to 1 at a score of 400 doubling every 20 point

Score = floor(400 + ((20/(log(2))) * (odds-(log(10)))))

Min and Max Score: 364 and 463 Cut off : 409

Below Cutoff: 1196

Above Cutoff: 229



Assessing the financial benefits

The potential benefit from the outcomes of the model will be assessed in terms of Profit and loss by the optimization of metrics will be shared with the bank.

Extra_benfit = 3170133356

loss_per_customer(outstanding_balance)*no_of_predicted defaultcustomer

- profit_pef_customer* no_of_predicted_defaultcustomer (Accepted and Rejected data set)

Summary

- Random forest model is chosen as the final Model as the values found in score card seem most consistent with the result.
- Optimal score card cut-off value of **409** is derived to approve and reject the applications and it's between maximum of rejected (404) and min of accepted (417)
- Score in Accecepted Data (1st Quantile and 3rd Quantile is):417 and 457
- Score in Rejected Data (1st Quantile and 3rd Quantile is): 388 and 403
- Net profit % using the model is: 86%.