BFS CAPSTONE PROJECT

CredX CREDIT RISK ANALYSIS

Anupam Majhi
Hari Nyshadam
Lijo Thomas
Rituraj Achuthan

Business Objectives & Strategy

Background & Objective

- *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.
- Mitigate Credit Risk by '*Acquiring the Right Customers*' using Predictive Modelling.
- Use bank's past applicants data, and
  - Create strategies to mitigate the acquisition risk.
  - Assess the financial benefit of the project.

Data Understanding

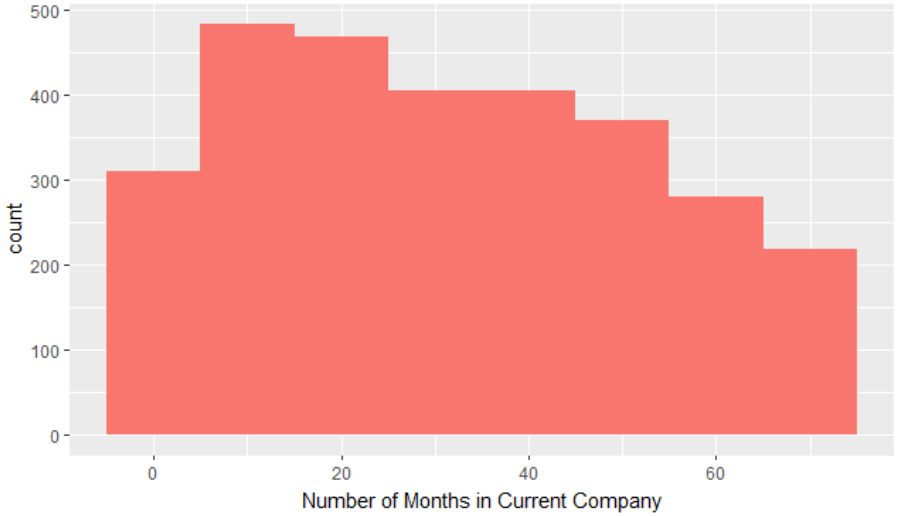
- There are two data sets in this project — demographic and credit bureau data.
  - **Demographic/application data**: This is obtained from the information provided by the applicants at the time of credit card application. It contains customer-level information on age, gender, income, marital status, etc.
  - **Credit bureau**: This is 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.

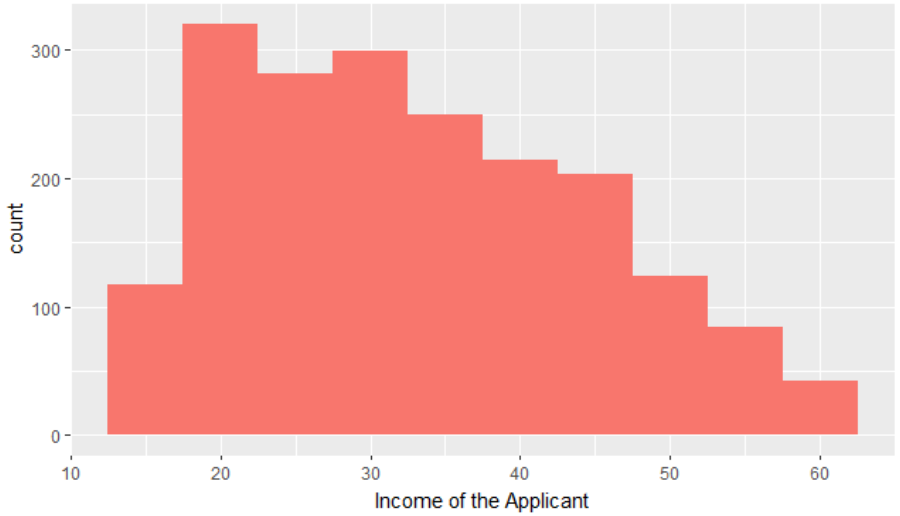
Problem Solving Methodology – Analysis Flow

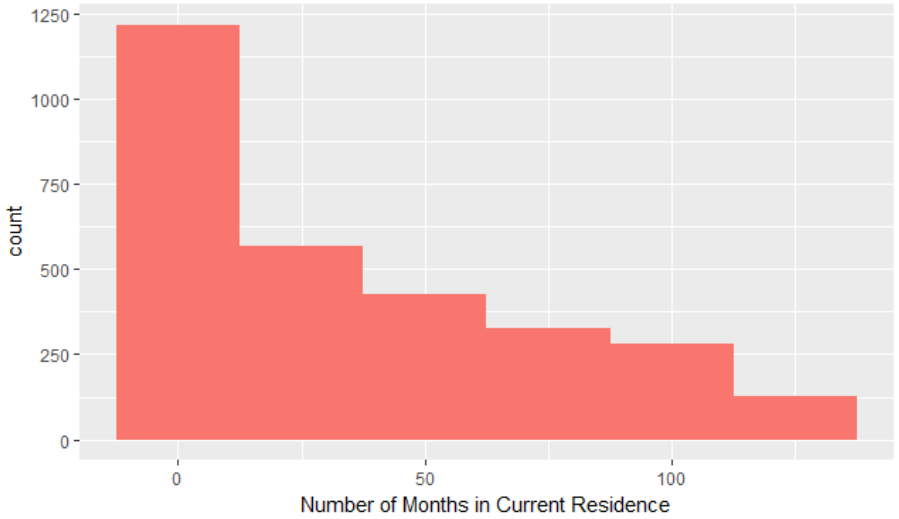
- Data Cleaning and Prepping:
  - Dataset is highly unbalanced with only 4% Rejected Customers information. Building prediction model on top of this data will not be effective.
- For the fluctuations in the data, we will be using '**Weight of Evidence (WOE)**' and '**Information Values (IV)**'. These are perfect frameworks for variable screening and exploratory analysis for predictive modelling. Using IV we will be imputing the missing values.
- To overcome the unbalanced dataset issue, we are using '**SMOTE** from DMwR package' in R. This package artificially generates new examples of the minority class using the nearest neighbors of these cases. Furthermore, the majority class examples are also under-sampled, leading to a more balanced dataset.
- Run Logistic Regression, Decision Trees and Random Forest on the prepped data and choose the best performing model.
- Create a Scorecard based on the final model and estimate the profit to the company

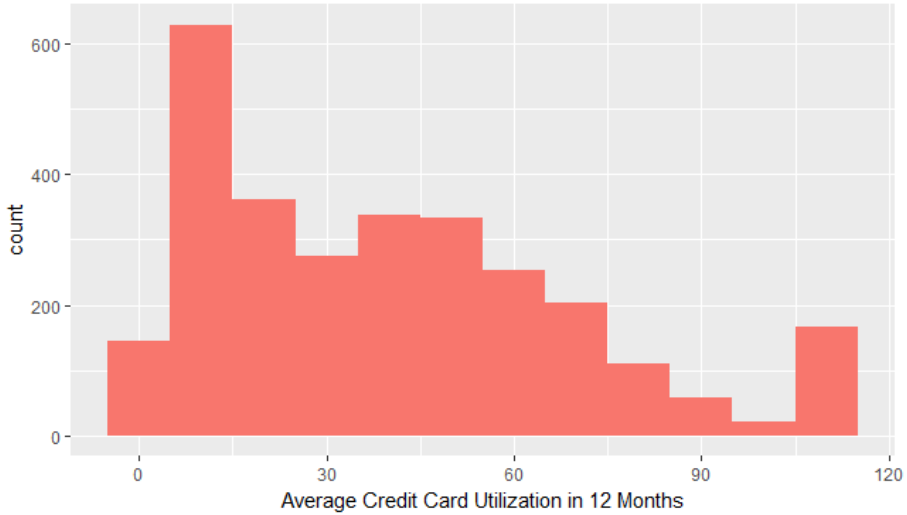
- **ASSUMPTIONS**:
  - Dataset has several records with Performance Tag = NULL. Assuming that the bank didn't process their application and are rejected candidates. Hence, replacing NULLs with 1 to validate the model.

Exploratory Data Analysis









As per our analysis, these were the major factors affecting the defaults.

- A person with higher number of months in current company – lower the chances of defaulting.
- People with higher income tends to have low default rates.
- People who are in same residence for very few months have high chances of defaulting.
- Frequent Credit Card users tends to default lesser.

CORRELATION MATRIX

- In the dataset we see that there is a high correlation between DPD's, which is expected.
- We have similar correlations among PL Trades
- Similarly, Inquiries within last 6 and 12 months have high correlation, which can be expected.

SMOTE SCORECARD

- We have highly unbalanced dataset as the rate of default is just 4%.
- We are using SMOTE to balance this dataset, which undersamples the majority and oversamples the minority classes.
- After several trial and errors, we obtained the best model values for the values where the minority and the majority percentage were at 50%, as well as we have increased the number of observations for the defaulters by 3 times.
- This was obtained by making the '**perc.over = 300**' and '**perc.under = 130**'.

perc.over	perc.under	Minority %	Model	Accuracy	Sensitivity	Specificity
200	350	30%	Logistic	63.13%	91.83%	57.79%
			Tree	84.30%	0.00%	100.00%
			Forest	84.39%	4.82%	99.21%
200	220	40%	Logistic	63.06%	91.59%	57.74%
			Tree	81.74%	80.02%	82.06%
			Forest	82.66%	32.17%	92.06%
200	150	50%	Logistic	63.06%	91.45%	57.72%
			Tree	84.49%	74.20%	86.40%
			Forest	82.82%	32.87%	92.12%
200	100	60%	Logistic	63.01%	91.45%	57.72%
			Tree	71.01%	91.45%	57.72%
			Forest	77.39%	62.22%	80.21%
300	310	30%	Logistic	63.00%	91.40%	57.71%
			Tree	84.30%	0.00%	100.00%
			Forest	84.27%	1.33%	99.70%
300	200	40%	Logistic	63.06%	91.51%	57.73%
			Tree	86.42%	21.21%	98.56%
			Forest	84.33%	5.73%	98.96%
300	**130**	**50%**	**Logistic**	**63.10%**	**91.72%**	**57.77%**
			Tree	**84.49%**	**74.20%**	**86.40%**
			Forest	**83.62%**	**19.90%**	**95.49%**
300	89	60%	Logistic	63.06%	91.59%	57.74%
			Tree	66.21%	87.14%	62.31%
			Forest	80.36%	42.88%	87.34%
300	58	70%	Logistic	62.81%	90.81%	57.60%
			Tree	40.80%	96.46%	30.43%
			Forest	71.79%	68.23%	72.45%

FINAL MODEL

		Logistic Regression	Decision Trees	Random Forest
Demographic Data	Accuracy	68.33%	61.47%	55.64%
	Specificity	69.02%	61.27%	54.14%
	Sensitivity	62.37%	63.20%	68.64%
Demographic + Credit Bureau Data	Accuracy	80.40%	66.67%	69.74%
	Specificity	81.30%	66.15%	69.79%
	Sensitivity	72.47%	71.29%	69.28%

- **Demographic Data Model**: On creating the model based just on the demographic data, we get a Accuracy of – **68.33%** from the Logistic Regression.
  - The predicting power of this model is not that high, but it does help in basic understanding of the trends.
  - The demographically dominant factors are –
    - Number of months in current residence.
    - Current Income of the applicant.
    - Number of months in current company.
    - Number of dependents.

- **Demographic + Credit Bureau Model**: On creating the model based on the combined data, we get a Accuracy of – **80.40%** from the Logistic Regression.
  - This model is efficiently predicting the defaulters.
  - The dominant factors are –
    - Average Credit Card Utilization in last 12 months
    - Number of Inquires in last 12 months
    - Number of Trades Opened in 12 months
    - Number of 30 Days Past Dues or worse in Last 12 months
    - Number of months in Current Company

FINAL MODEL EVALUATION

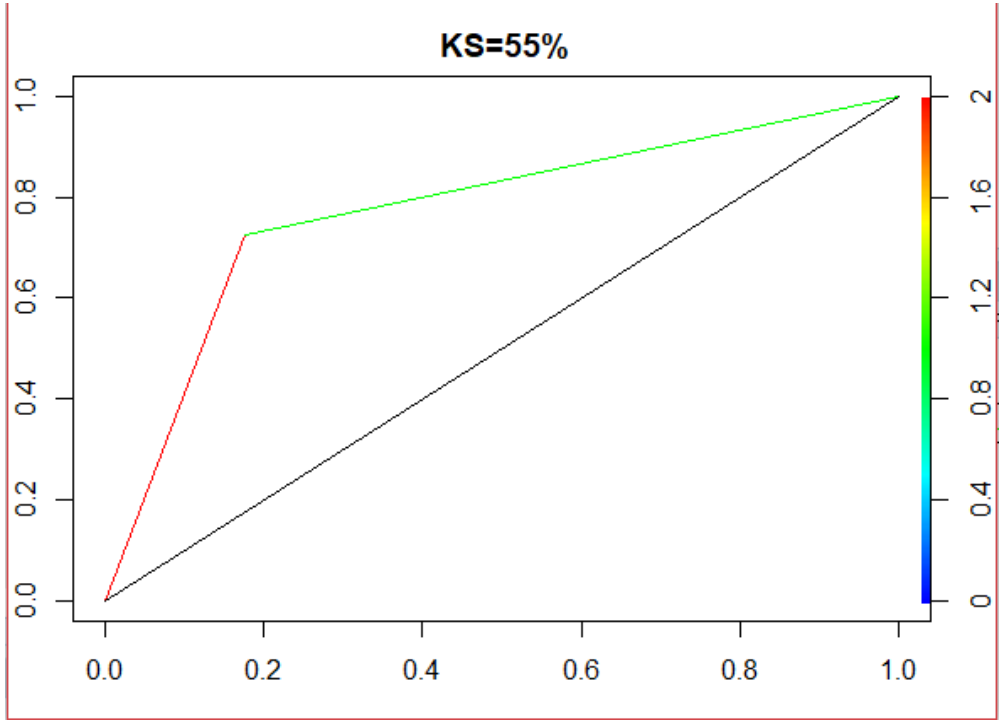
METRICS FOR COMBINED DATA

Gain Table

bucket	total	totalresp	Cumresp	Gain	Cumlift
1	2234	722	722	31.50087	3.150087
2	2234	722	1444	63.00175	3.150087
3	2234	285	1729	75.43630	2.514543
4	2233	78	1807	78.83944	1.970986
5	2234	77	1884	82.19895	1.643979
6	2234	91	1975	86.16928	1.436155
7	2233	84	2059	89.83421	1.283346
8	2234	68	2127	92.80105	1.160013
9	2234	86	2213	96.55323	1.072814
10	2233	79	2292	100.00000	1.000000

By the third decile we are having almost 75% gain and by the 4th decile, we are getting 78%

ROC Curve



Area Under Curve = **77.44%**

KS Statistics = **55%**

SCORECARD

This is the formula we have used for calculating the scorecard for our model:

$$Score = Offset + Factor * \ln(odds)$$
$$Score + pdo = Offset + Factor * \ln(2 * odds)$$

$$pdo = Factor * \ln(2)$$
$$Factor = pdo / \ln(2)$$
$$Offset = Score \{Factor * \ln(Odds)\}$$

- We have created the scorecard on the test data without NA values.
- We have identified a cutoff of **324.3**. This was then evaluated against the rejected population and was found to be 97% accurate.

Probability Good	Odds Good	Natural Logs of Odds	Original Response	Score	Predicted Response
0.447395917	0.809613846	-0.211197879	0	327.47	0
0.506143703	1.024880531	0.024576051	0	334.27	0
0.539007356	1.169232013	0.156347134	0	338.07	0
0.403577918	0.67666495	-0.390579032	1	322.29	1
0.535099239	1.150996696	0.140628259	0	337.62	0
0.500821789	1.003292567	0.003287158	0	333.66	0
0.421992991	0.730082827	-0.31459729	0	324.48	0
0.415088461	0.709660238	-0.342968962	1	323.67	1
0.421992991	0.730082827	-0.31459729	0	324.48	0
0.421992991	0.730082827	-0.31459729	0	324.48	0
0.598295697	1.489393299	0.398368855	0	345.06	0
0.421992991	0.730082827	-0.31459729	0	324.48	0
0.409929252	0.694712037	-0.364257854	1	323.05	1
0.431789731	0.759911875	-0.274552806	0	325.64	0
0.421992991	0.730082827	-0.31459729	0	324.48	0
0.426574328	0.74390518	-0.295841699	0	325.03	0
0.421992991	0.730082827	-0.31459729	0	324.48	0
0.42719411	0.745792105	-0.293308398	0	325.10	0
0.507817855	1.031768136	0.031273968	0	334.46	0
0.498058692	0.992264802	-0.00776527	1	333.34	0
0.423488601	0.734571081	-0.308468513	0	324.66	0

Snapshot of the Scorecard

FINANCIAL BENEFIT

IMPLICATIONS OF USING THIS MODEL

- Auto rejection rate for our scorecard is **~30%**.
- Auto approval rate for our scorecard is **~70%**.
- Using this model, we get some False positives, but true positives are more significant in the model and therefore, the bank always remains in profit.
- The loss of potential revenue gain is covered up by the prevention of credit loss from potential defaulters.
- Applicants with credit score close to our cutoff of 324.3, we have to be cautious while providing larger loans.

	* Values Obtained
Total Revenue Gained	20,593,708,889
Total Credit Loss	6,698,003,758
Financial Benefit	13,895,705,131

** We assume that the outstanding balance present in the dataset represents the total exposure at default. Adding the balance gives us the values in the table.

CONCLUSION

- Demographic Data alone does not have high predictive power. But when combined with Credit Bureau Data gives a pretty good predictive model.

- Few dominant factors affecting the model were :
  - Average Credit Card Utilization in last 12 months
  - Number of Inquires in last 12 months
  - Number of Trades Opened in 12 months
  - Number of 30 Days Past Dues or worse in Last 12 months
  - Number of months in Current Company

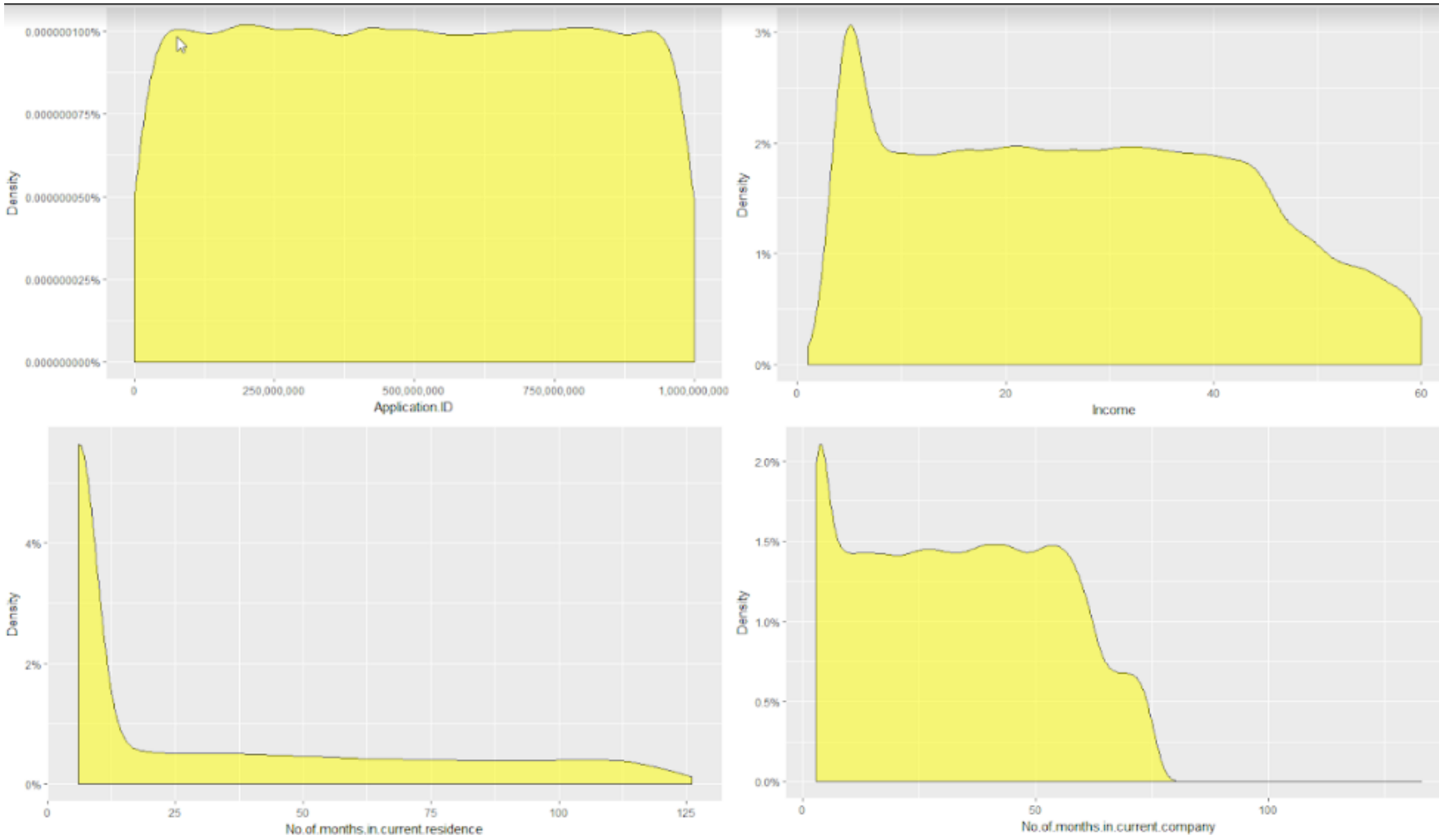
- We have built a scorecard to eliminate the defaulters applying to the bank. Using this scorecard, we have a 30% rejection rate which means that we might get some False positives (potential customers might get rejected). But the candidates predicted correctly are more significant and therefore, the bank always remains in profit.

- For applicants with credit score close to our cutoff of 324.3, we have to be cautious while providing larger loans. These candidates should be carefully monitored and the recovery strategy should be appropriately designed

- The scorecard with the cut off score of 324.3 was evaluated on the rejected applicants and found to be 97% accurate.

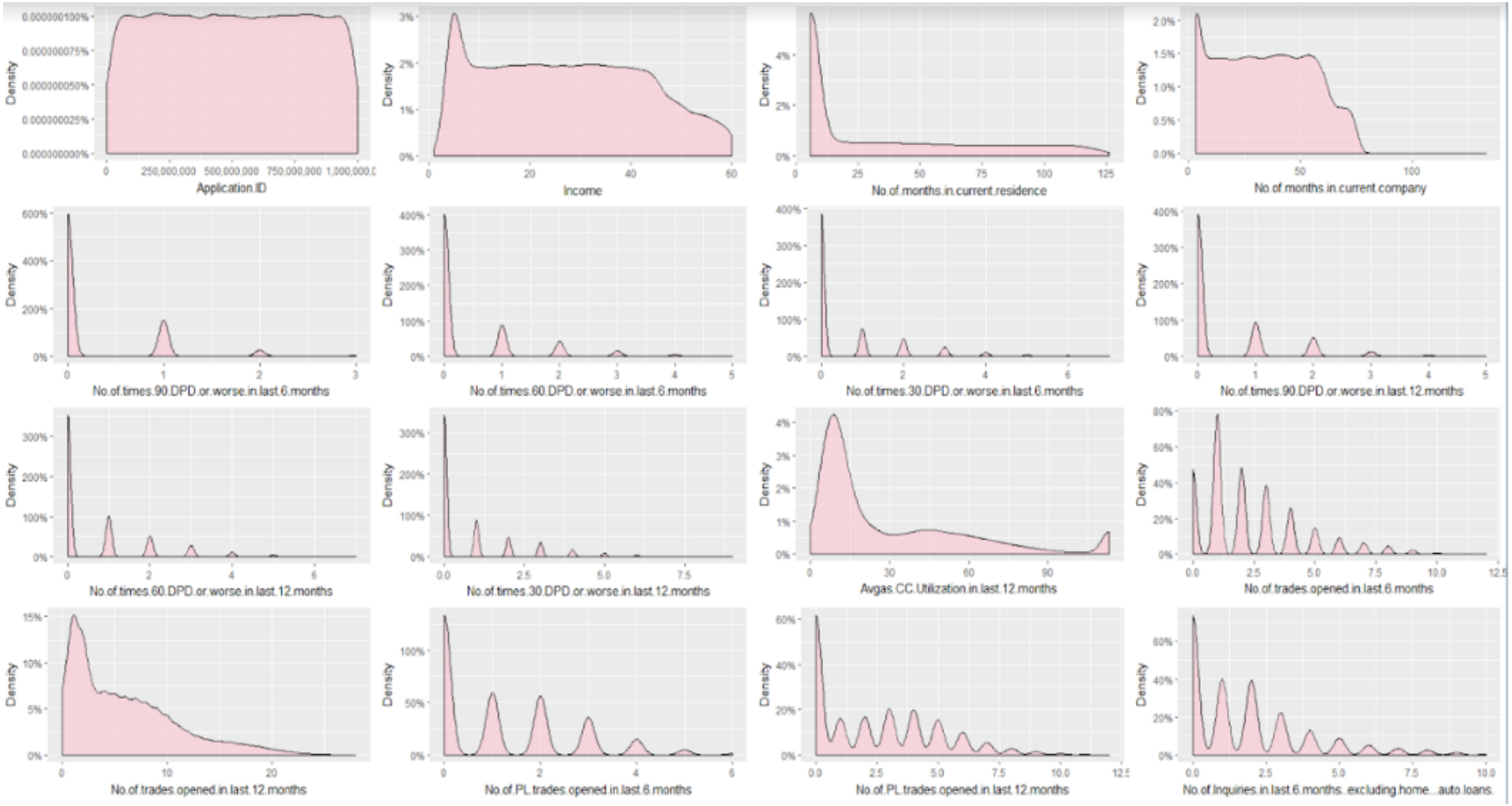
APPENDIX I

Density plot for Demographic Data



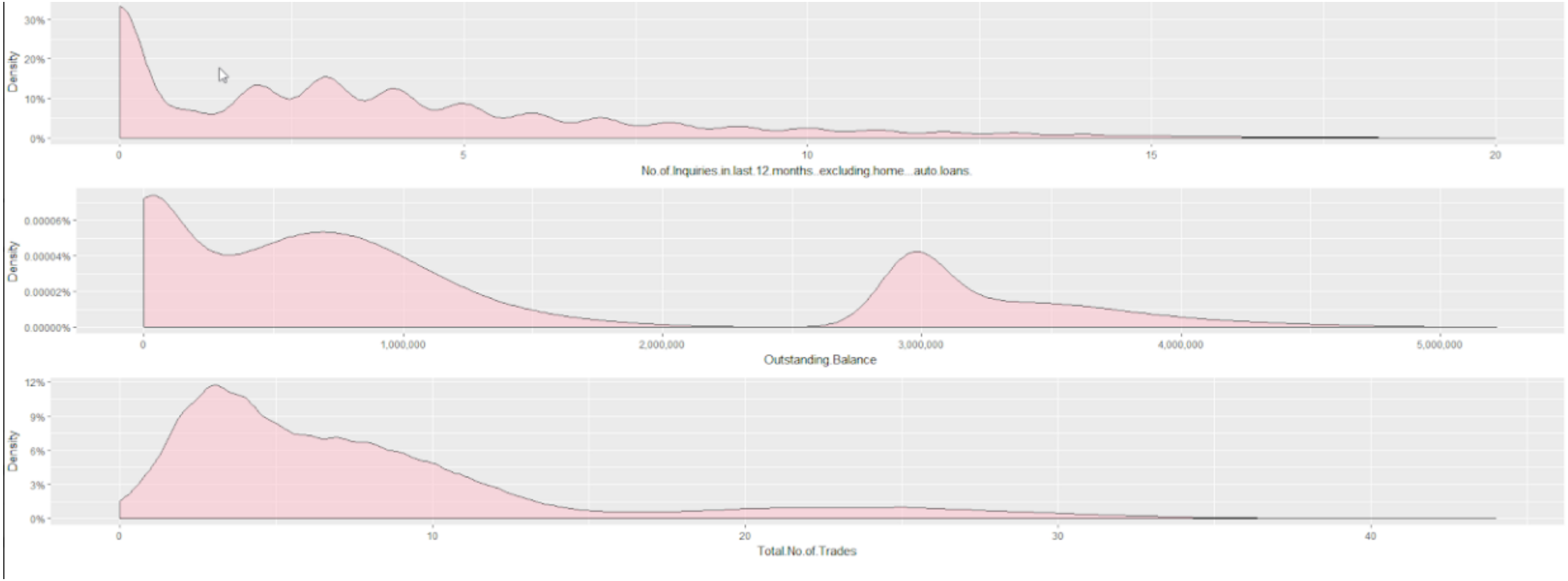
APPENDIX II

Density plot for Combined Data - I



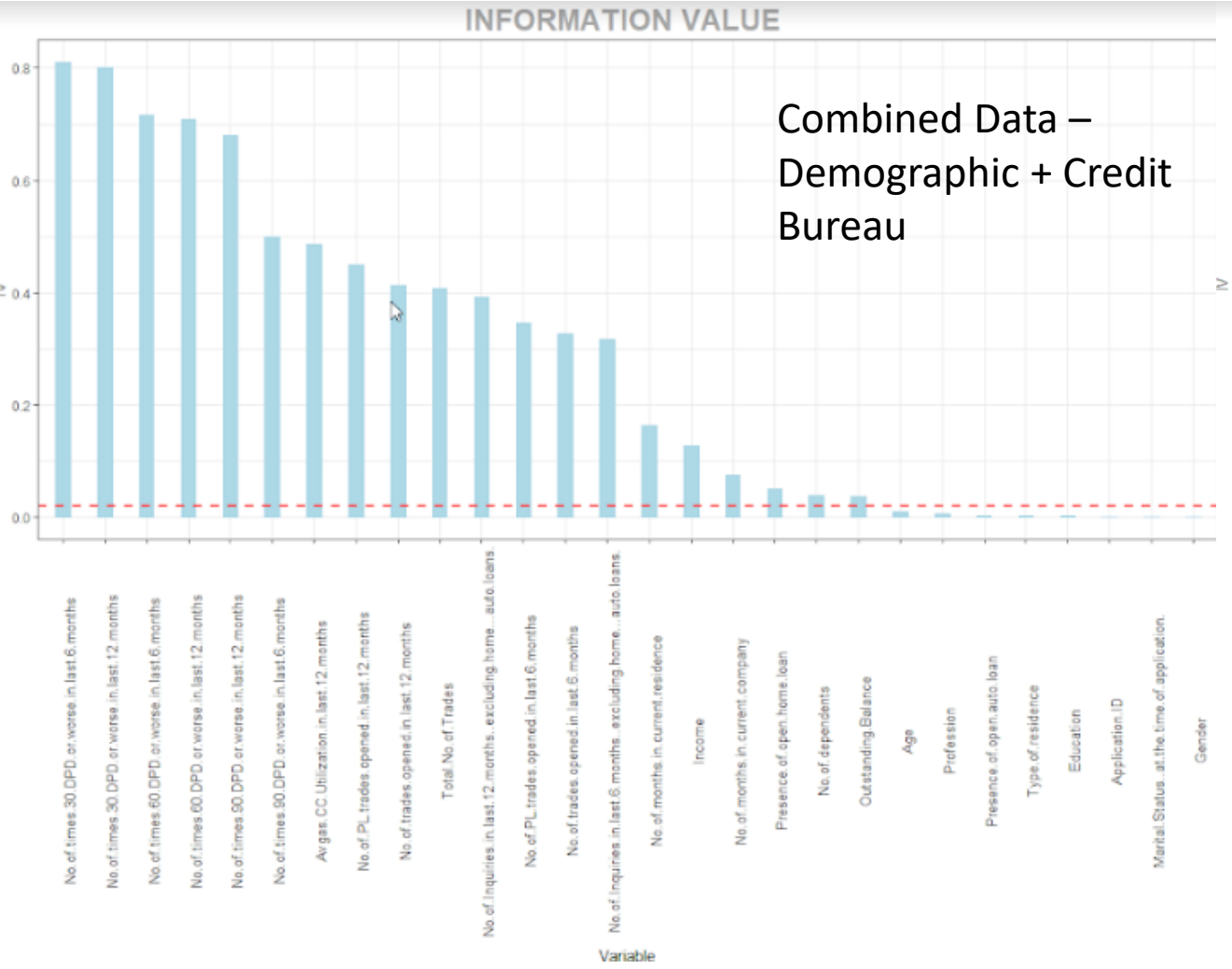
APPENDIX III

Density plot for Combined Data - II

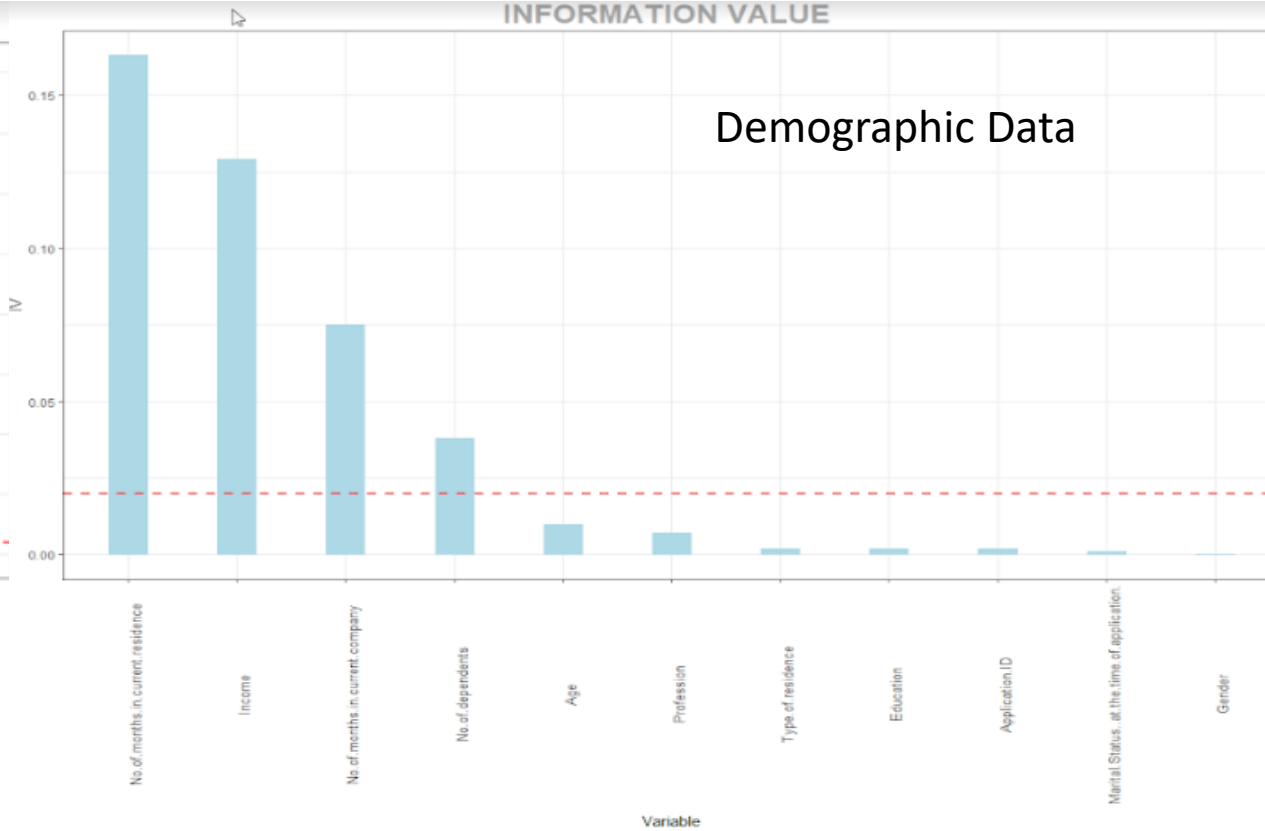


APPENDIX IV

Information Value Charts



Combined Data – Demographic + Credit Bureau



Demographic Data