BFSI Capstone Project – CredX Credit Risk Analysis

Dinesh Challa

Bhumesh Arkal

23-Dec-2019

Background

- Problem Statement :

- Increasing Credit Loss of CredX(Credit Card Provider Institution). Looking at past data, there are about 4% of customers who defaulted.

- Goal :

- Improve Customer base. And thus offer credit cards to right customers who are less likely to default.

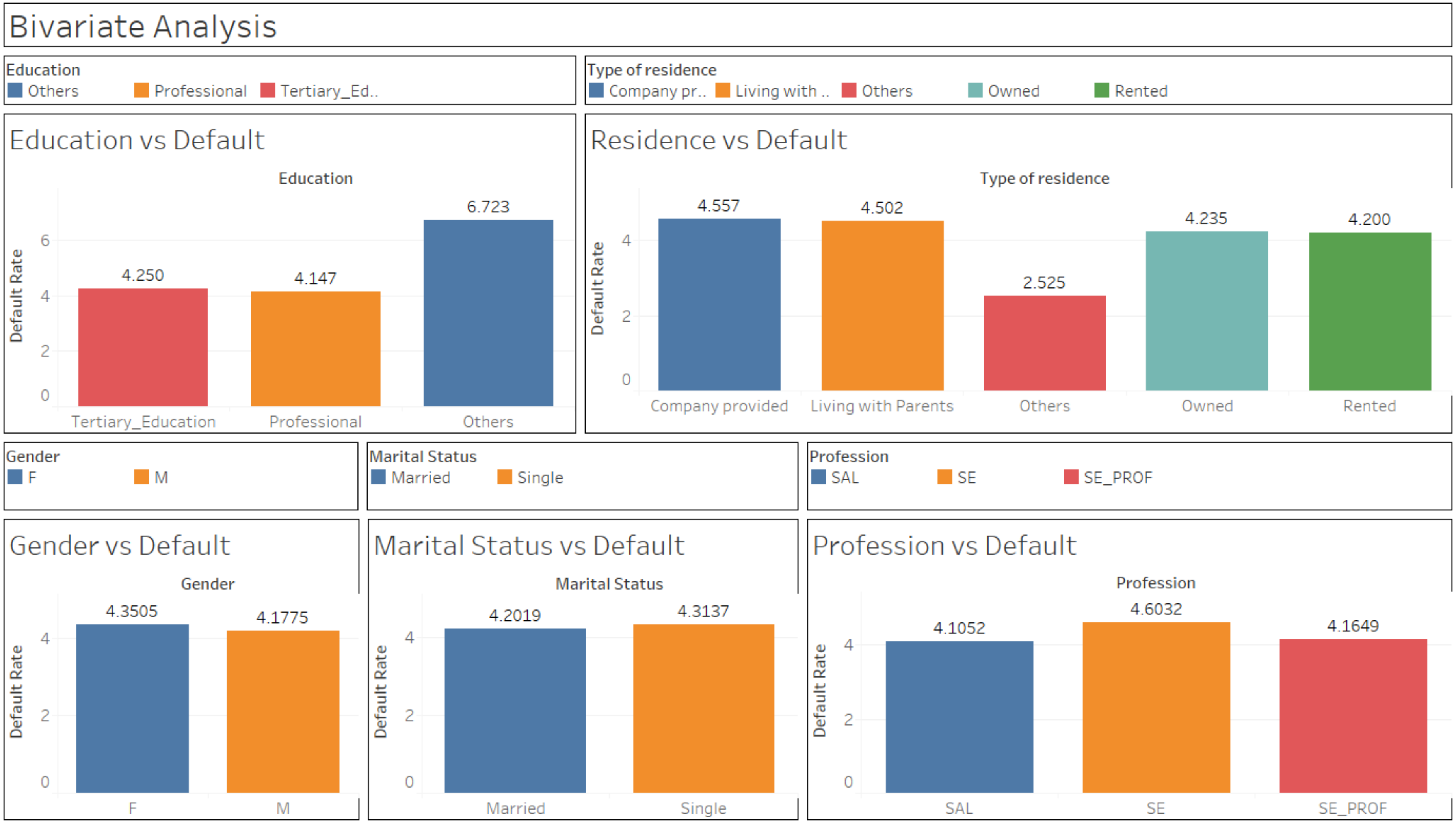
- Expected Outcome of Assessment :

- Identify important attributes using the Weight Of Evidence (WOE).
- Build predictive models and identify the best performing model.
- Build an application scorecard and identify cut-off score to grant credit card to applicants..

Actions done for this Case Study

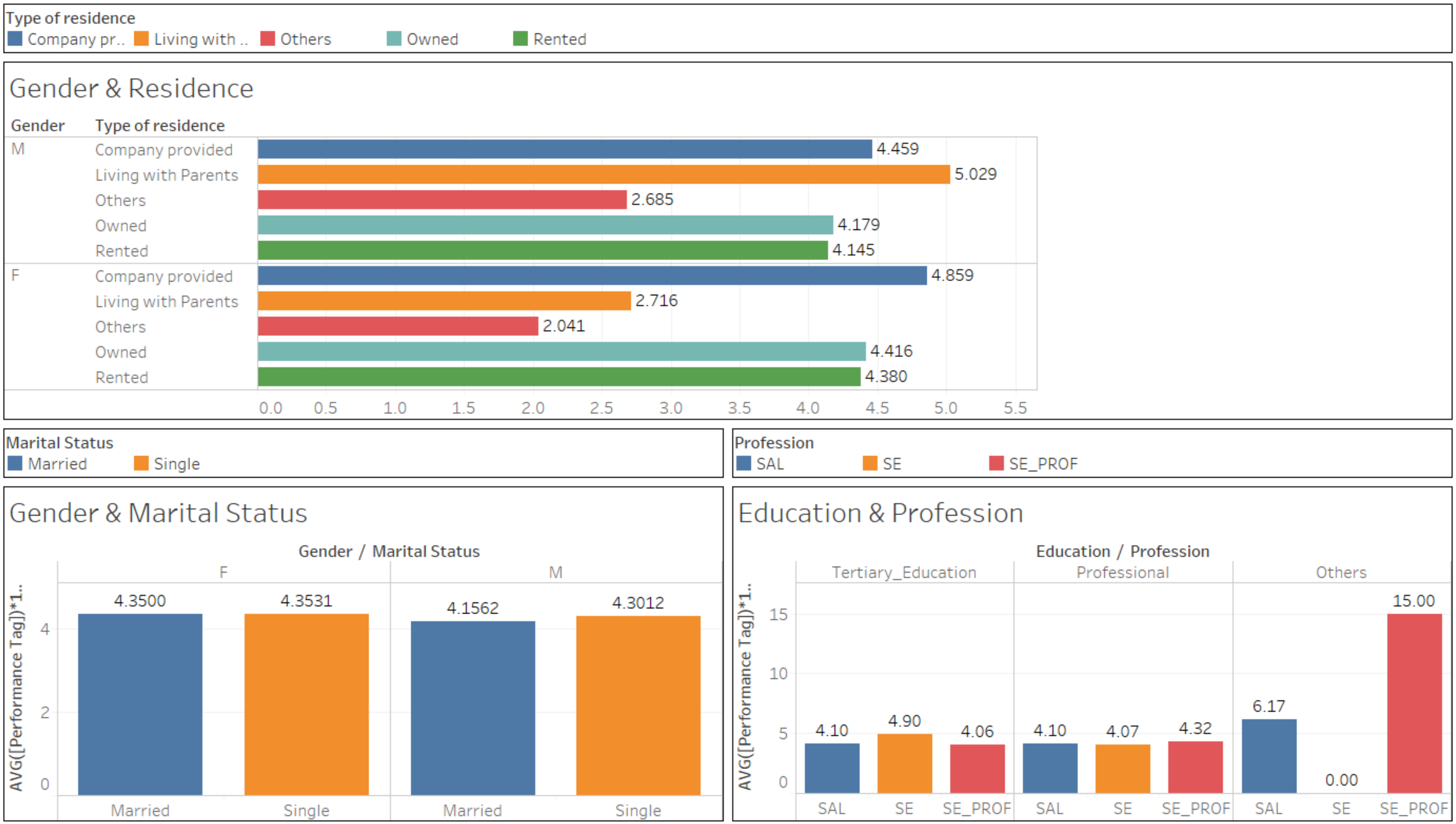
1. Data Understanding, EDA and Cleaning of Demographic and Credit Bureau Data
2. Identify important variables by WOE and IV assessment
3. Model Building and evaluation
4. Creating Application Scorecard and suggesting cut-off for auto-approvals
5. Financial Benefit Analysis

Data Understanding, EDA and Cleaning of Demographic data

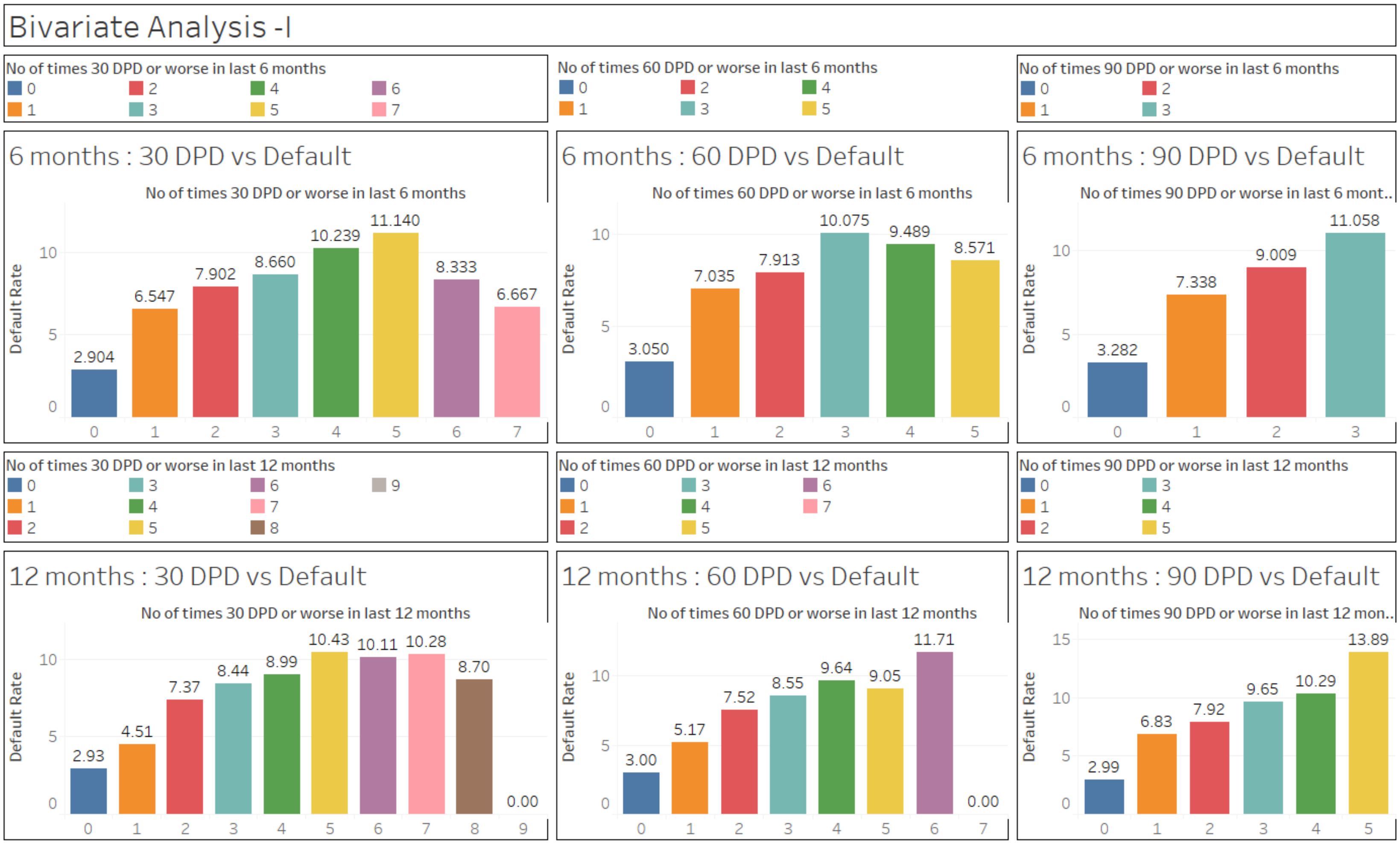


Data Understanding, EDA and Cleaning of Demographic Data

Multivariate Analysis

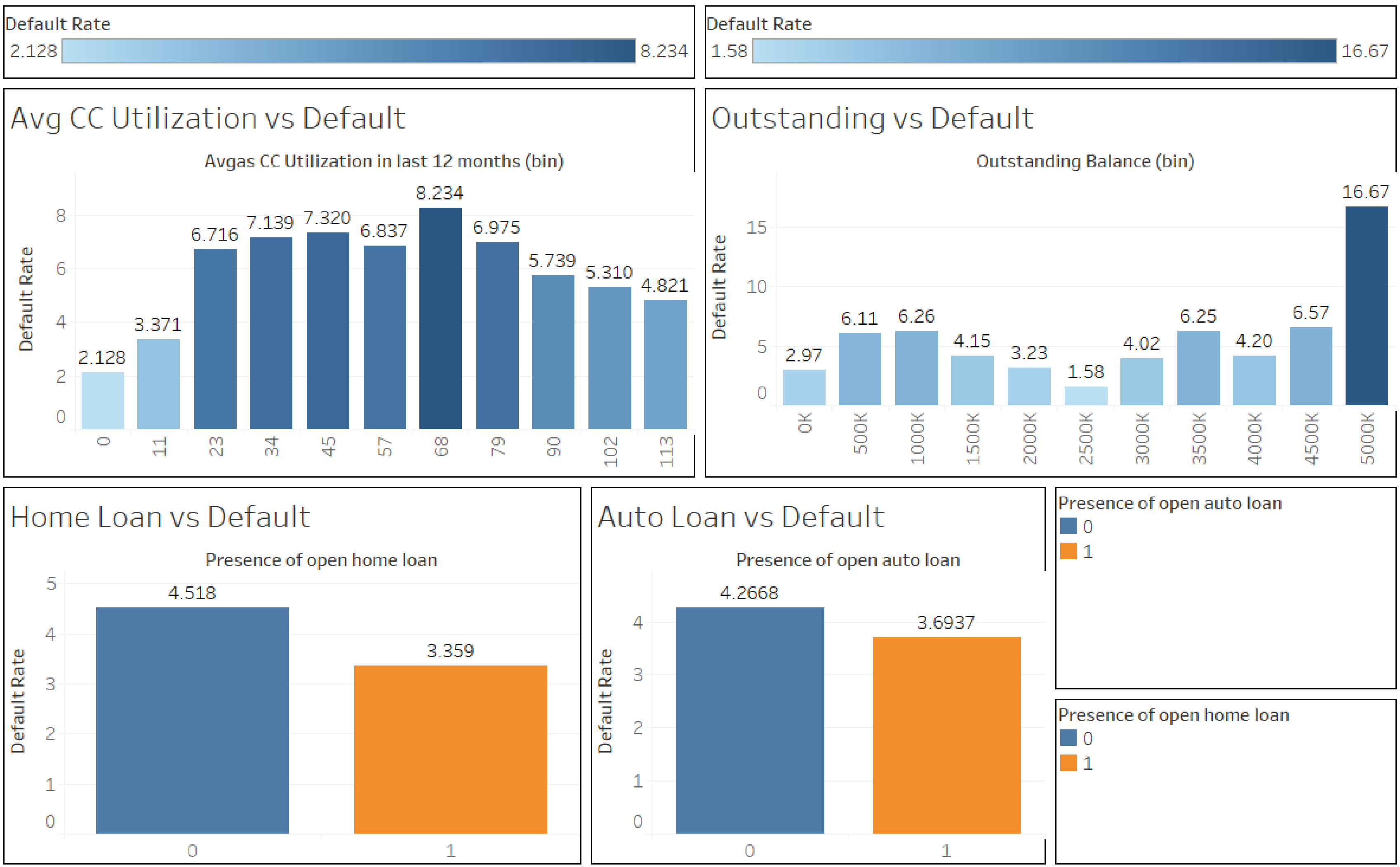


Data Understanding, EDA and Cleaning of Credit Bureau Data



Data Understanding, EDA and Cleaning of Credit Bureau Data

Bivariate Analysis - II



Identify important variables by WOE and IV assessment

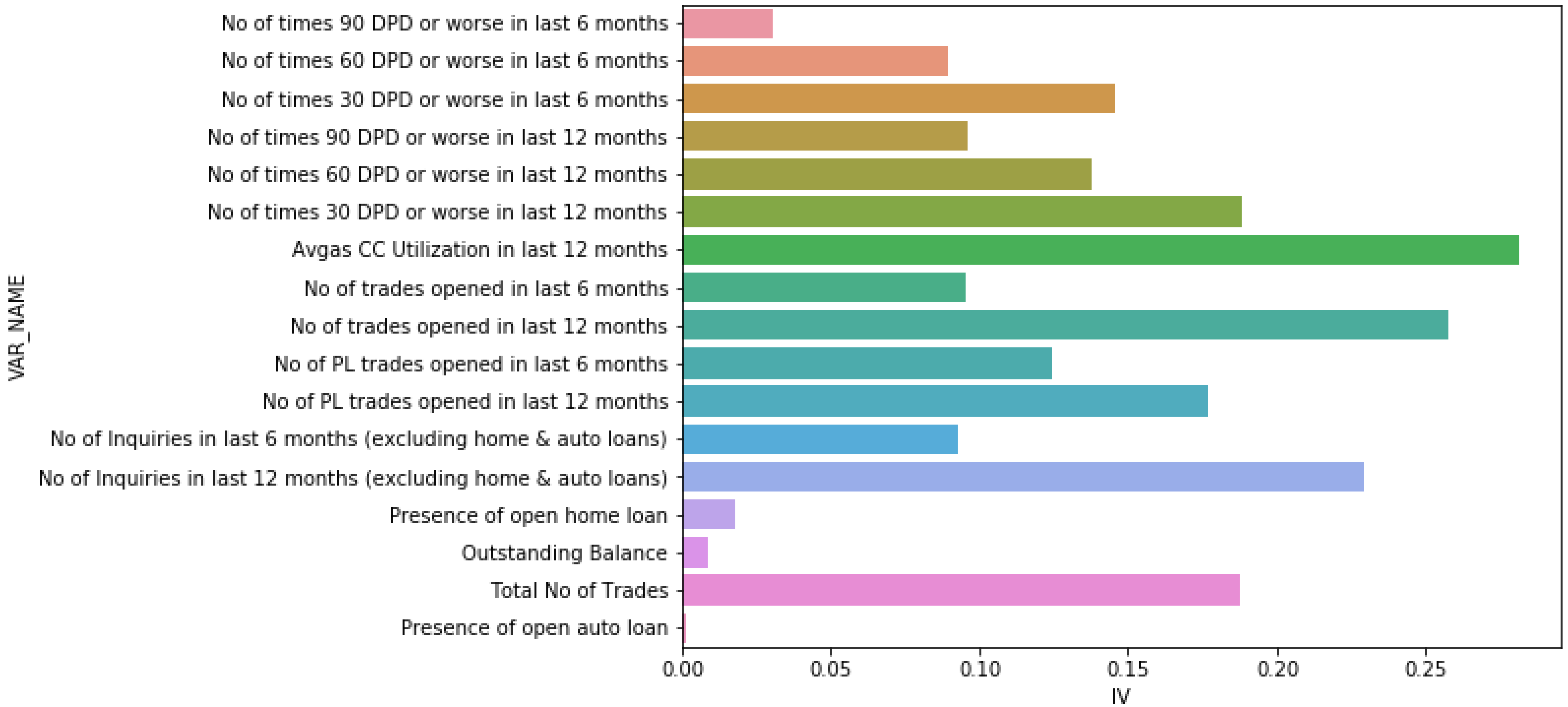
	VAR_NAME	IV
7	No of months in current residence	0.048779
3	Income	0.032527
6	No of months in current company	0.010968
8	Profession	0.002226
9	Type of residence	0.000925
1	Education	0.000672
0	Age	0.000627
2	Gender	0.000327
4	Marital Status	0.000096
5	No of dependents	0.000056

Variables of Demographic data sorted on IV

	VAR_NAME	IV
0	Avgas CC Utilization in last 12 months	0.281539
11	No of trades opened in last 12 months	0.257429
1	No of Inquiries in last 12 months (excluding h...	0.229218
5	No of times 30 DPD or worse in last 12 months	0.188045
16	Total No of Trades	0.187303
3	No of PL trades opened in last 12 months	0.176644
6	No of times 30 DPD or worse in last 6 months	0.145708
7	No of times 60 DPD or worse in last 12 months	0.137676
4	No of PL trades opened in last 6 months	0.124744
9	No of times 90 DPD or worse in last 12 months	0.095714
12	No of trades opened in last 6 months	0.095337
2	No of Inquiries in last 6 months (excluding ho...	0.092939
8	No of times 60 DPD or worse in last 6 months	0.089574
10	No of times 90 DPD or worse in last 6 months	0.030711
15	Presence of open home loan	0.017627
13	Outstanding Balance	0.008569
14	Presence of open auto loan	0.001655

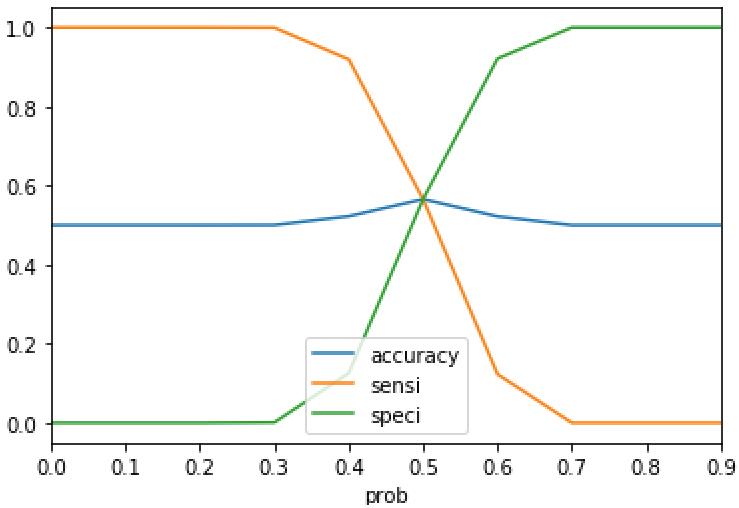
Variables of Credit Bureau data sorted on IV

Identify important variables by WOE and IV assessment



Visualisation of IV for Credit Bureau data

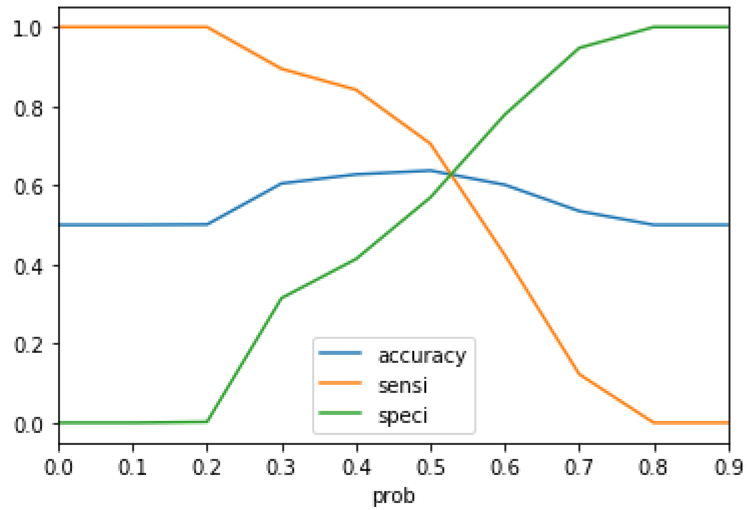
Model Building and evaluation (Demographic Data)



	precision	recall	f1-score	support
0.0	0.96	0.68	0.80	20033
1.0	0.06	0.44	0.10	926
accuracy			0.67	20959
macro avg	0.51	0.56	0.45	20959
weighted avg	0.92	0.67	0.76	20959

With Logistic Regression model the overall accuracy is around 0.57 and sensitivity is around 0.44. Thus, we can say that the demographic data has got decent predictive power but the sensitivity and accuracy needs to be better. Let's find out how the metrics do on the overall dataset.

Model Building and evaluation (Overall Data)



	precision	recall	f1-score	support
0.0	0.97	0.60	0.74	20033
1.0	0.07	0.64	0.13	926
accuracy			0.60	20959
macro avg	0.52	0.62	0.43	20959
weighted avg	0.93	0.60	0.72	20959

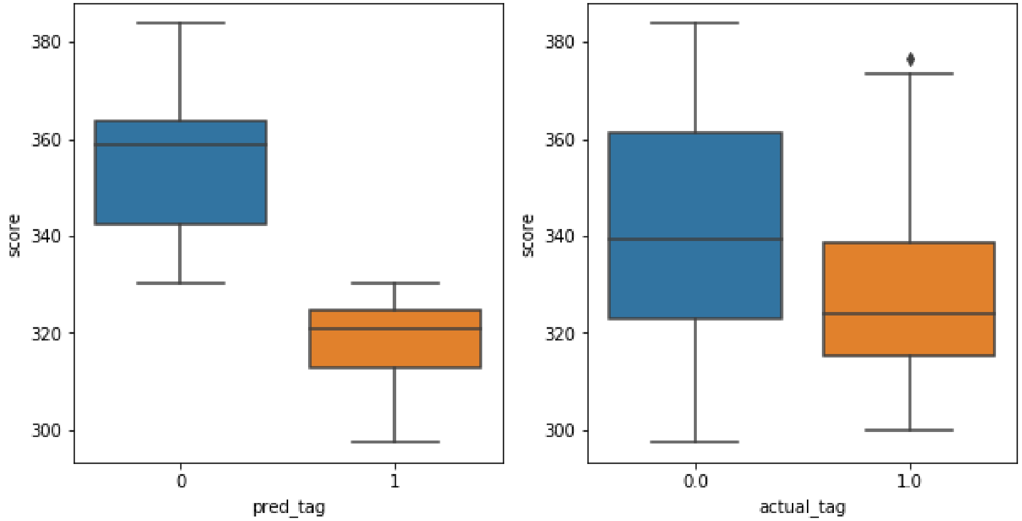
With Logistic Regression, we can observe that the metrics (sensitivity and specificity) from the model evaluation are consistent with that of the training model. Both sensitivity and specificity are above 60%.

Model Building and evaluation (Lift Gain Chart)

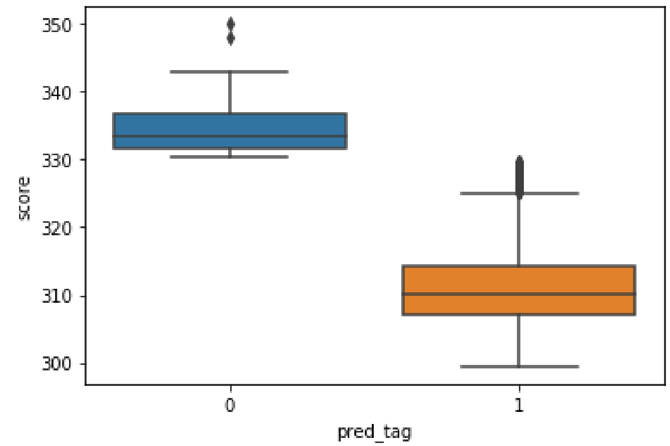
	decile	total	actual_tag	cumresp	gain	cumlift
9	1	6986	593	593	20.122158	2.012216
8	2	6986	529	1122	38.072616	1.903631
7	3	6986	401	1523	51.679674	1.722656
6	4	6986	372	1895	64.302681	1.607567
5	5	6986	319	2214	75.127248	1.502545
4	6	6986	241	2455	83.305056	1.388418
3	7	6986	184	2639	89.548694	1.279267
2	8	6985	119	2758	93.586698	1.169834
1	9	6982	88	2846	96.572786	1.073031
0	10	6992	101	2947	100.000000	1.000000

It can be seen from Lift Gain Chart that the likelihood of default is about double in the first 3 deciles. The increased risk in these groups is almost double.

Creating Application Scorecard and suggesting cut-off for auto-approvals



Overall Population Score



Rejected Population Score

In Overall Population Score plot, we can see the scores compared to predicted tags and actual tags. For scores against predicted tags, we can see distinct boundary between default-cases and non-default cases.

And for scores against actual tags, we can see that there is no clear boundary. There will be some mis-classifications.

In Rejected Population Score, the boundary is more clear and is around 330.

The score of 330 can be considered as cut-off for auto-approvals.

Financial Benefit Analysis

Confusion Matrix based on scorecard cut-off of 330.

		Predicted	
		Good	Bad
Actual	Good	40514	26400
	Bad	1034	1913

Assumptions:

- Let the average credit loss from a defaulter be 100.0 units
- Let the average profit from a good customer be 50.0 units

Default rate without mode = 2947/69861 = 4.2%

Default rate with model = 1034/40514 = 2.5%

Net profit = (Profit from good customers predicted as good + Profit from bad customers predicted as bad) –
(Loss from bad customers predicted as good + Loss from good customers predicted as bad)

Net profit = (40514*5 + 1913*100) – (26400*5 + 1034*100) = 158470 units