BFS Capstone Presentation

Group Name:

1. Dhir Chandan Roll - DDA1710293
2. Arun Naudiyal
3. SINAN SAHIN
4. Ankur Shrivastava

Problem Statement

- CredX , a leading credit card provider, is experiencing an increase in credit loss.
- To mitigate credit risk, it wants to identify the right customers.

Goals: 1)Identify the right customers for CredX, using predictive models using past customer data.

2) Determine the factors affecting credit risk and assess the financial benefits of the model.

Overall Approach

Load Datasets “Demographic data.csv” and “Credit Bureau Data.csv” in R

Check for duplicate records in both the datasets on Application Id and remove the records which cannot be reconciled.

Merge the 2 datasets into a master dataset based on Application Id.

For records having “NA” in the Performance.tag , remove those records, and keep them in a separate dataset.

Carry out WOE analysis on all the variables to identify the most important ones.

Impute the missing values by WOE values and treat the outliers.

Create an initial logistic model only on demographic data to assess the predictive power of applicant data.

Build various models on the merged dataset

Evaluate the models using various parameters like KS-statistic, ROC curve, etc.

Predict the likelihood of default for the rejected candidates using the model

Build an application scorecard of all the datapoints.

Assess the potential final benefits of the model.

Demographic Data Summary

Variables	n	Mean	Standard Deviation	Median	Mean Absolute Deviation	Min	Max	Range	Skew	Kurtosis	Standard Error	Blanks	NAs
Application.ID	71295	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	None	None
Age	71295	44.94	9.94	45	11.86	-3	65	68	-0.01	-0.69	0.04	None	None
Income	71295	27.2	15.51	27	19.27	-0.5	60	60.5	0.19	-1.03	0.06		
No.of.months.in.current.residence	71295	34.56	36.76	11	7.41	6	126	120	0.99	-0.44	0.14		
No.of.months.in.current.company	71295	33.96	20.41	34	25.2	3	133	130	0.12	-1.07	0.08		
Gender	n	Male	Female	Blanks	NAs								
	71295	54456	16837	2	None								
Marital.Status..at.the.time.of.application	n	Married	Single	Blanks	NAs								
	71295	60730	10559	6	None								
No.of.dependents	n	1	2	3	4	5	Blanks	NAs					
	71295	15387	15289	16279	12222	12115	None	3					
Education	n	Bachelor	Masters	Phd	Professional	Others	Blanks	NAs					
	71295	17697	23970	4549	24839	121	119	None					
Profession	n	SAL	SE	SE_PROF	Blanks	NAs							

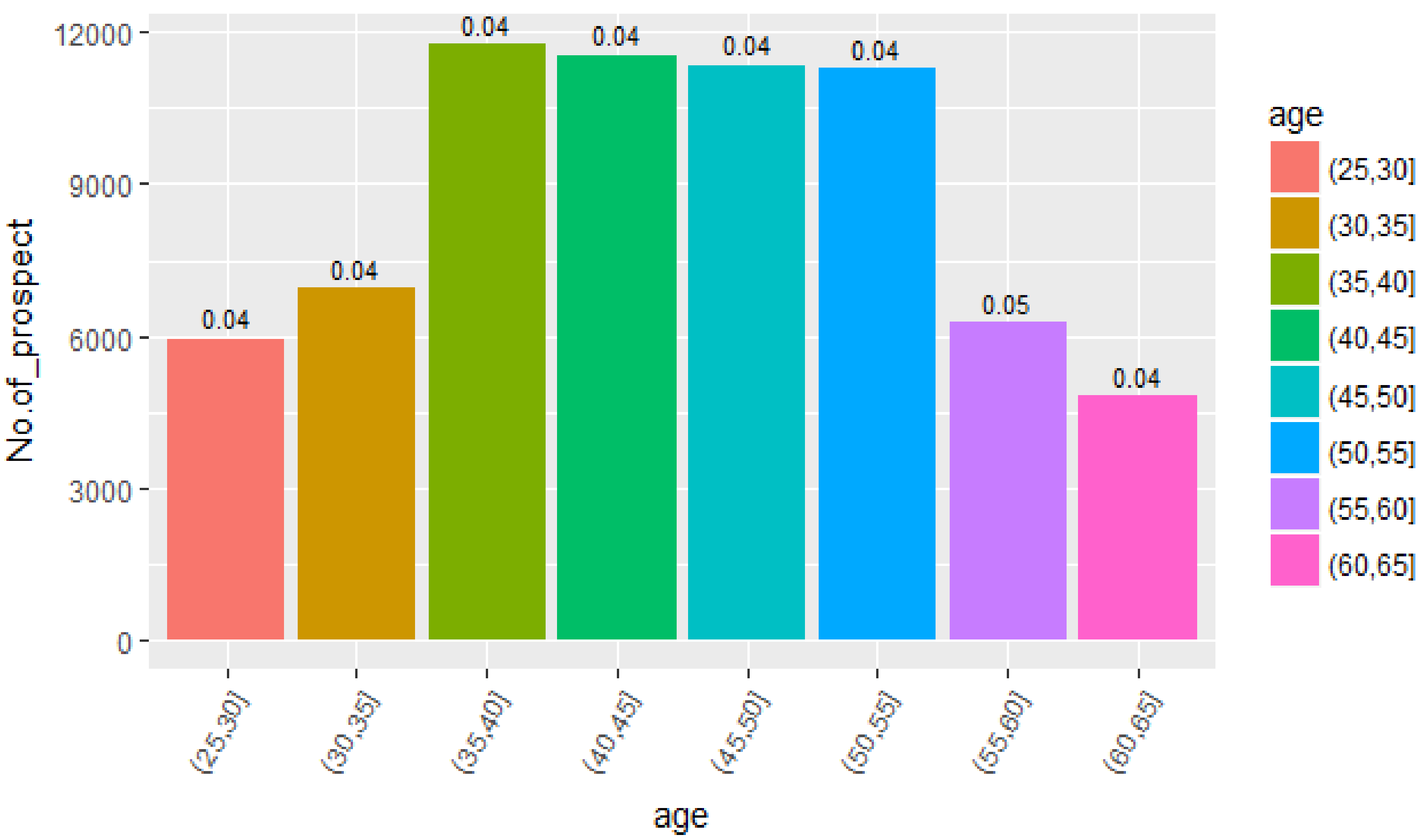
Credit Bureau Data Summary

Variables	n	Mean	Standard Deviation	Median	Mean Absolute Deviation	Min	Max	Range	Skew	Kurtosis	Standard Error	Blanks	NAs
Application.ID	71295	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	None	None
No.of.times.90.DPD.or.worse.in.last.6.months	71295	0.2703134+ C3:N188621 923	0.53	0.00	0.00	0.00	3.00	3.00	2.02	4.02	0.00		
No.of.times.60.DPD.or.worse.in.last.6.months	71295	0.430535	0.83	0.00	0.00	0.00	5.00	5.00	2.16	4.70	0.00		
No.of.times.30.DPD.or.worse.in.last.6.months	71295	0.577207	1.07	0.00	0.00	0.00	7.00	7.00	2.11	4.38	0.00		
No.of.times.90.DPD.or.worse.in.last.12.months	71295	0.45034	0.81	0.00	0.00	0.00	5.00	5.00	1.90	3.37	0.00		
No.of.times.60.DPD.or.worse.in.last.12.months	71295	0.655488	1.09	0.00	0.00	0.00	7.00	7.00	1.91	3.55	0.00		
No.of.times.30.DPD.or.worse.in.last.12.months	71295	0.800912	1.33	0.00	0.00	0.00	9.00	9.00	1.92	3.50	0.00		
Avgas.CC.Utilization.in.last.12.months	70237	29.69693	29.53	15.00	14.83	0.00	113.00	113.00	1.37	1.04	0.11		
No.of.trades.opened.in.last.6.months	71294	2.298048	2.07	2.00	1.48	0.00	12.00	12.00	1.22	1.34	0.01		
No.of.trades.opened.in.last.12.months	71295	5.826888	5.07	5.00	4.45	0.00	28.00	28.00	1.06	0.64	0.02		
No.of.PL.trades.opened.in.last.6.months	71295	1.206901	1.35	1.00	1.48	0.00	6.00	6.00	0.95	0.14	0.01		
No.of.PL.trades.opened.in.last.12.months	71295	2.397447	2.42	2.00	2.97	0.00	12.00	12.00	0.73	-0.31	0.01		
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.	71295	1.763532	1.97	1.00	1.48	0.00	10.00	10.00	1.36	1.70	0.01		

Data Cleaning Steps

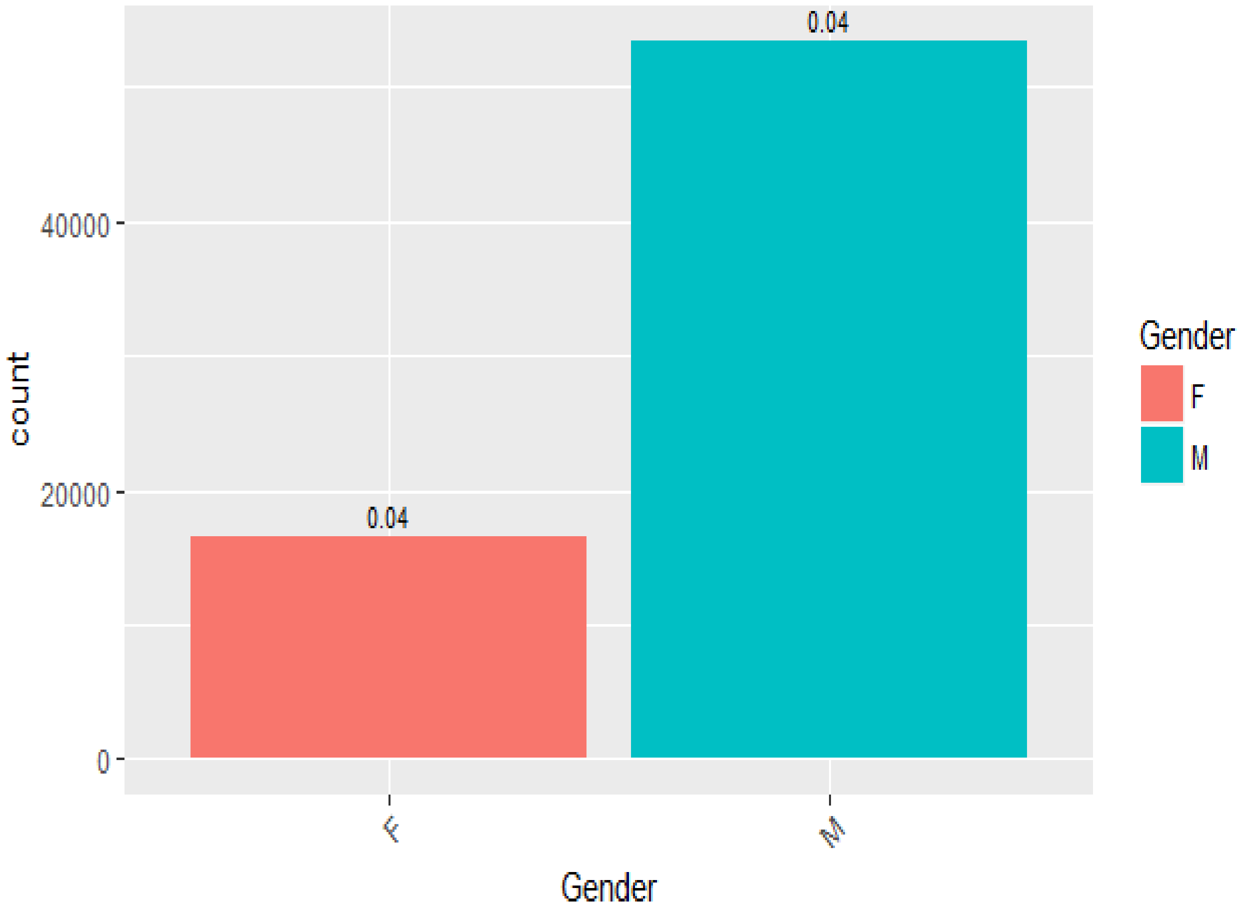
- All rows which have no performance tag haves been removed as they were rejected and saved in a different dataframe.
- Both Datasets had 3 duplicate rows each which were removed.
- The two dataframes were merged with 71295 rows in total with 29 diff variable.
- Missing values were be imputed with WOE & IV data for final model building.

EDA - Age



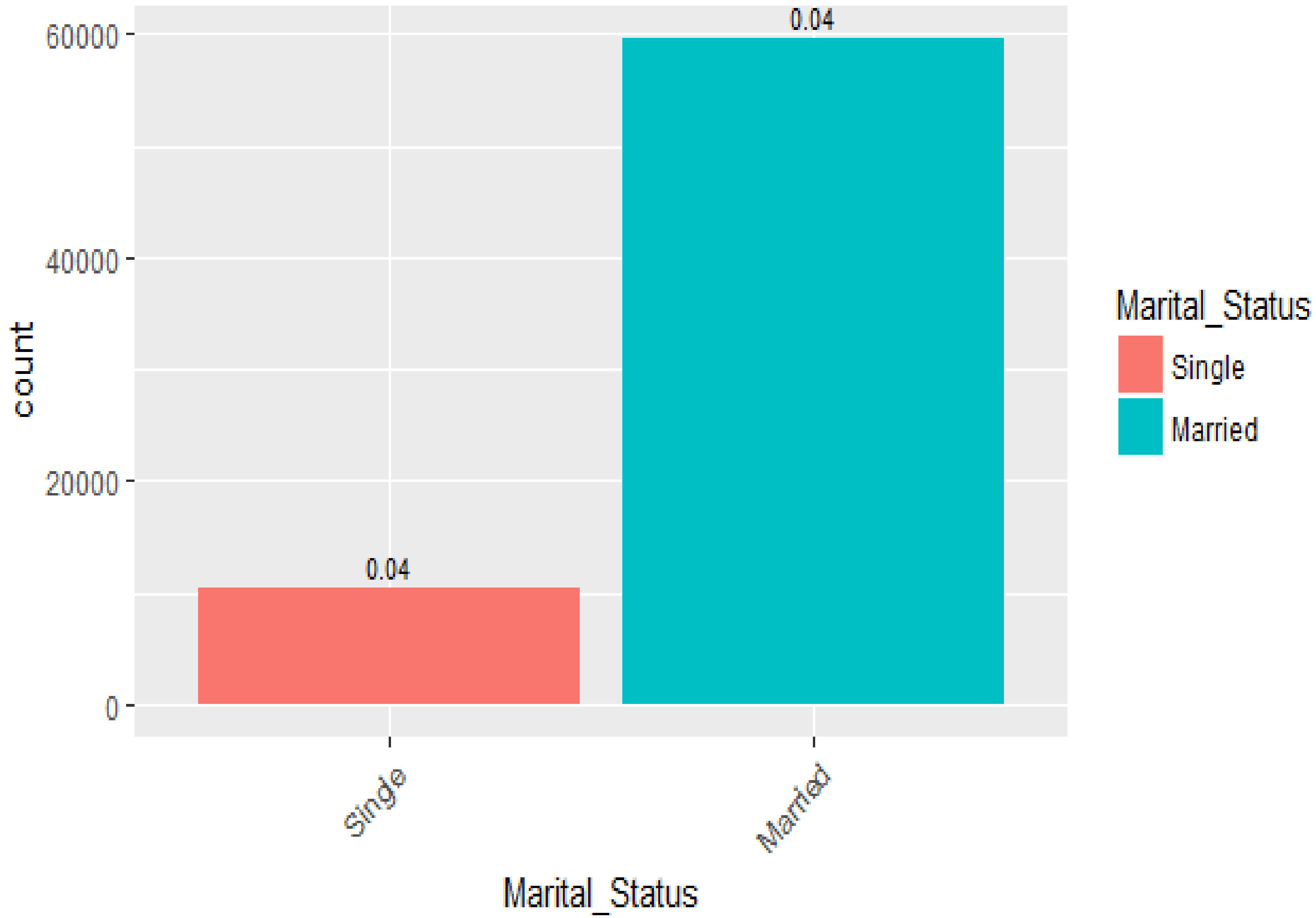
Plotting the Age variable doesn't show much difference in the default rates across various age categories.

EDA - Gender



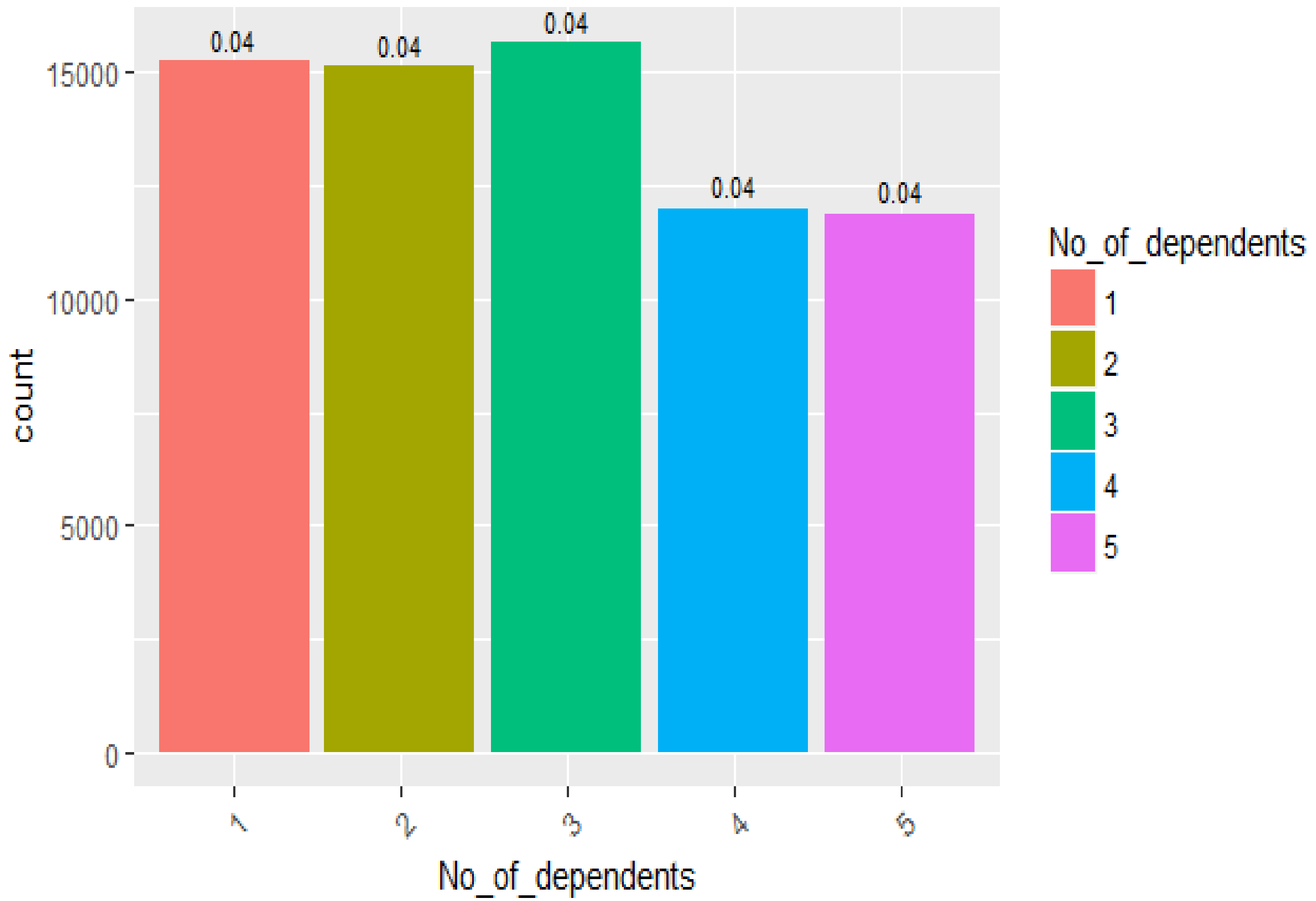
Default Rates don’t show any difference across gender as well.

EDA – Marital Status



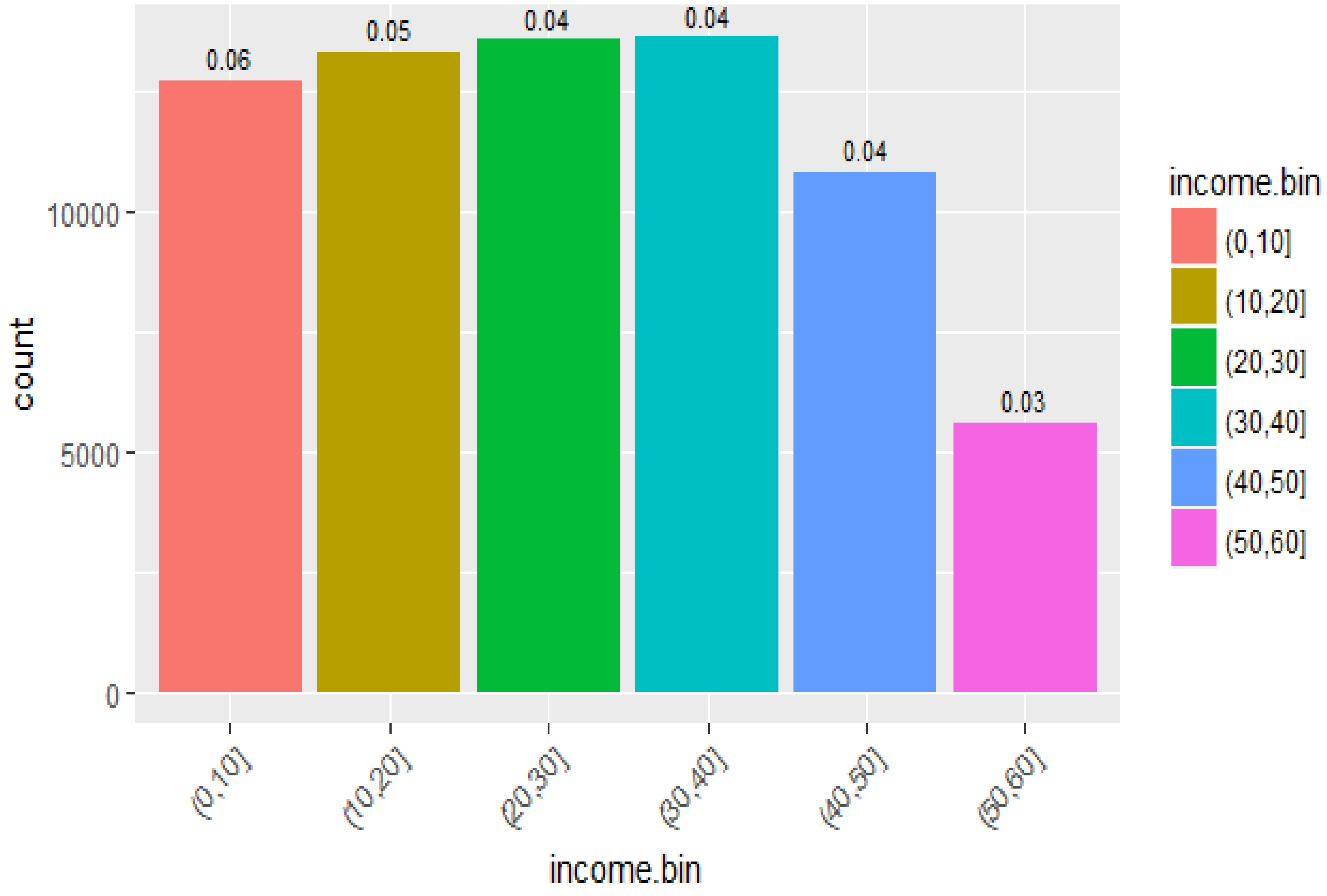
No difference in default rates across the marital states as well.

EDA Number of dependents



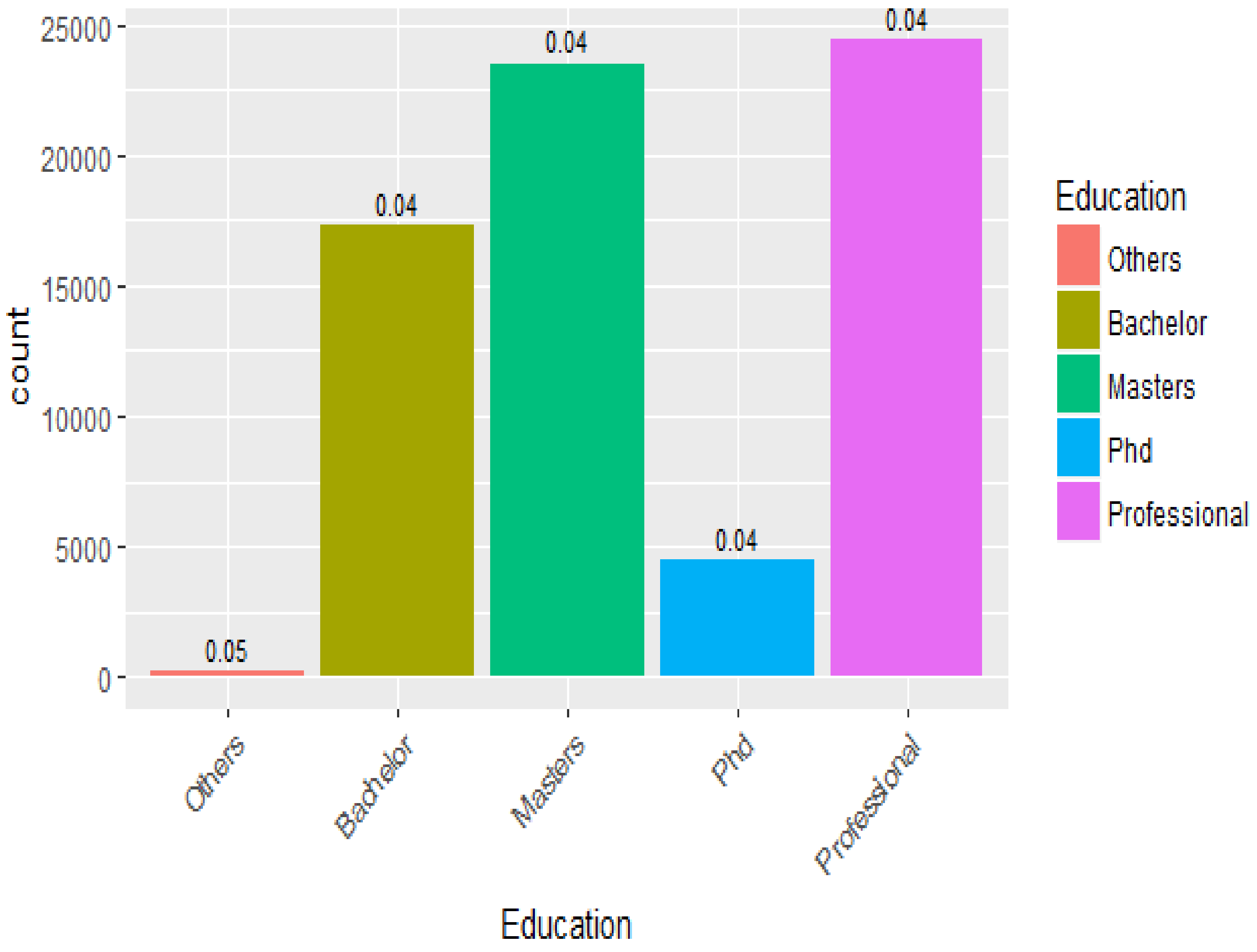
No difference in default rates across different number of dependents

EDA Income



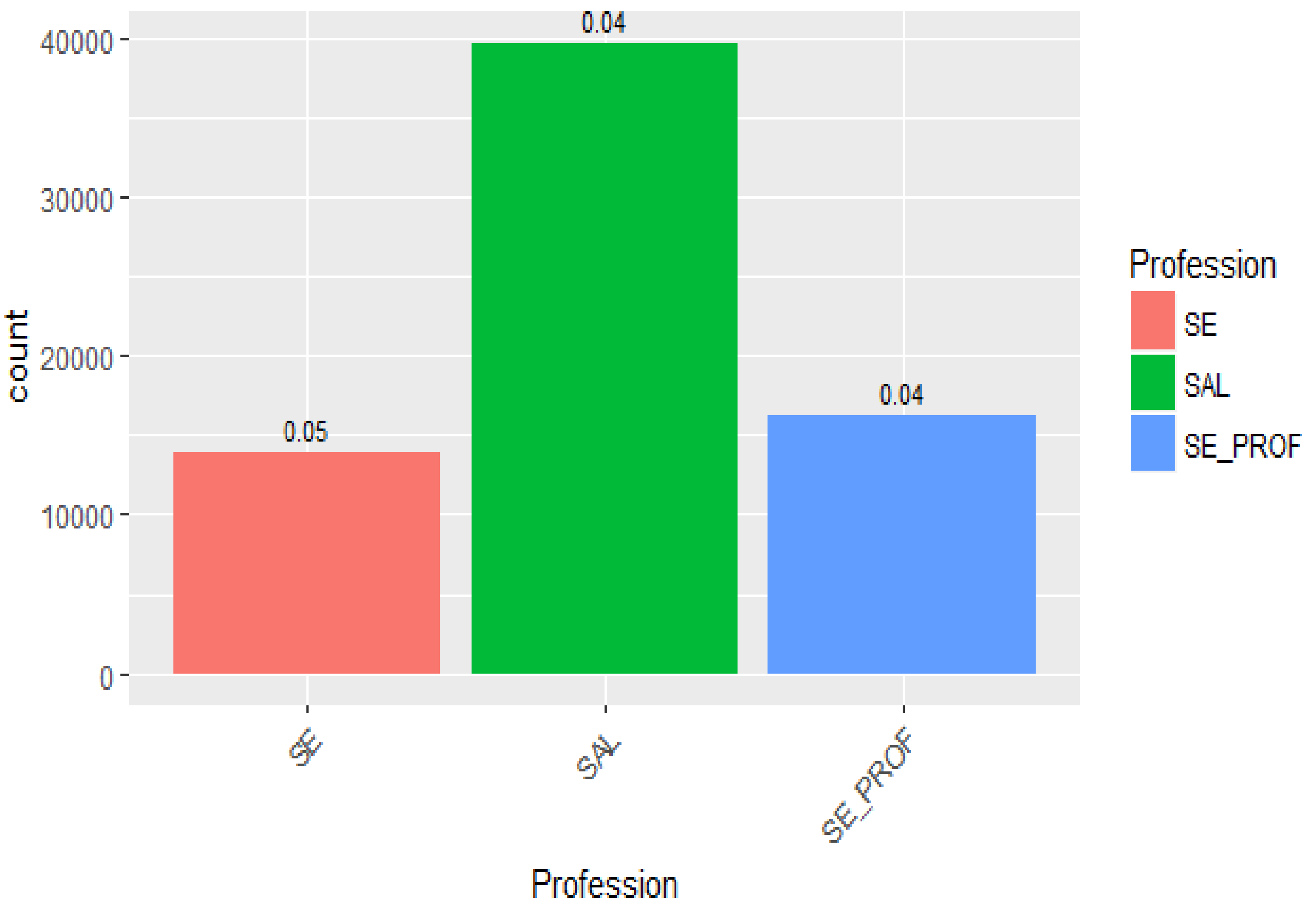
Prospects in the 0,10 income bracket have twice the default rate of prospects in the (50,60) income bracket which would make sense Hence income can be an important predictor of default.

EDA Education



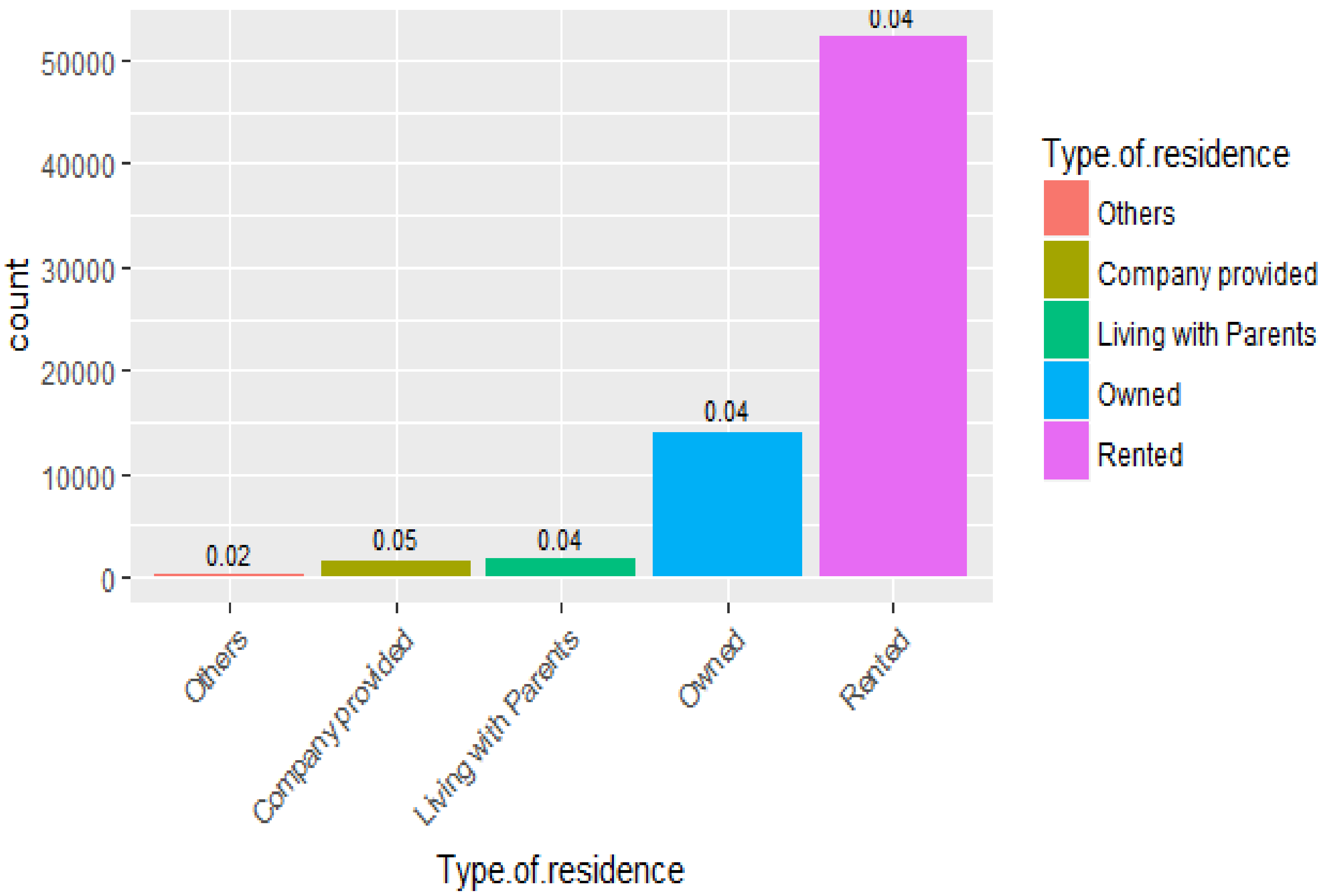
As can be seen from the plots, there is not much difference in the default rates across education levels. Hence education might not be good predictor of default.

EDA Profession



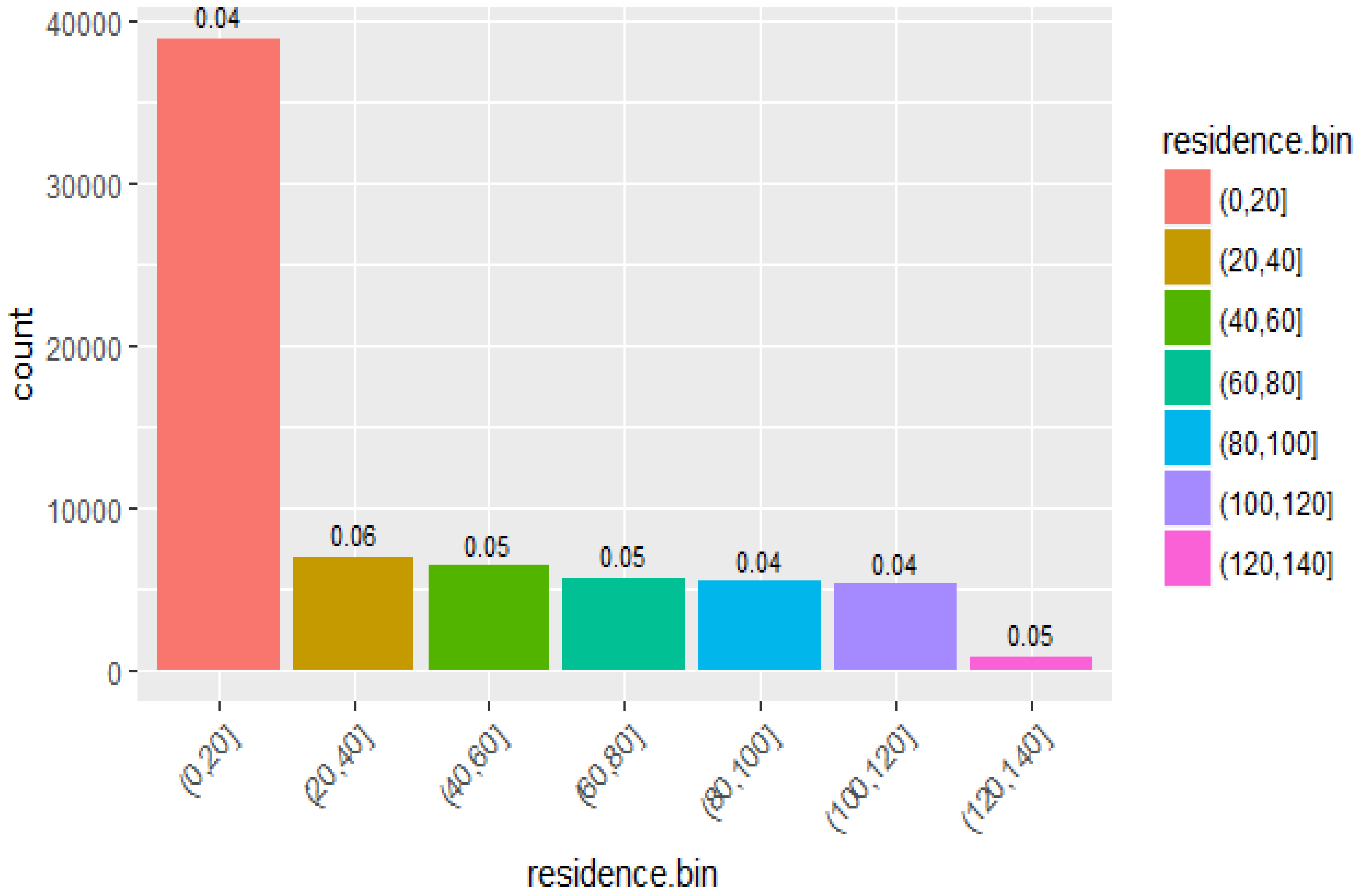
Default in SE level is slightly more than other 2 levels, hence Profession might be weak indicator of default.

EDA Type of residence



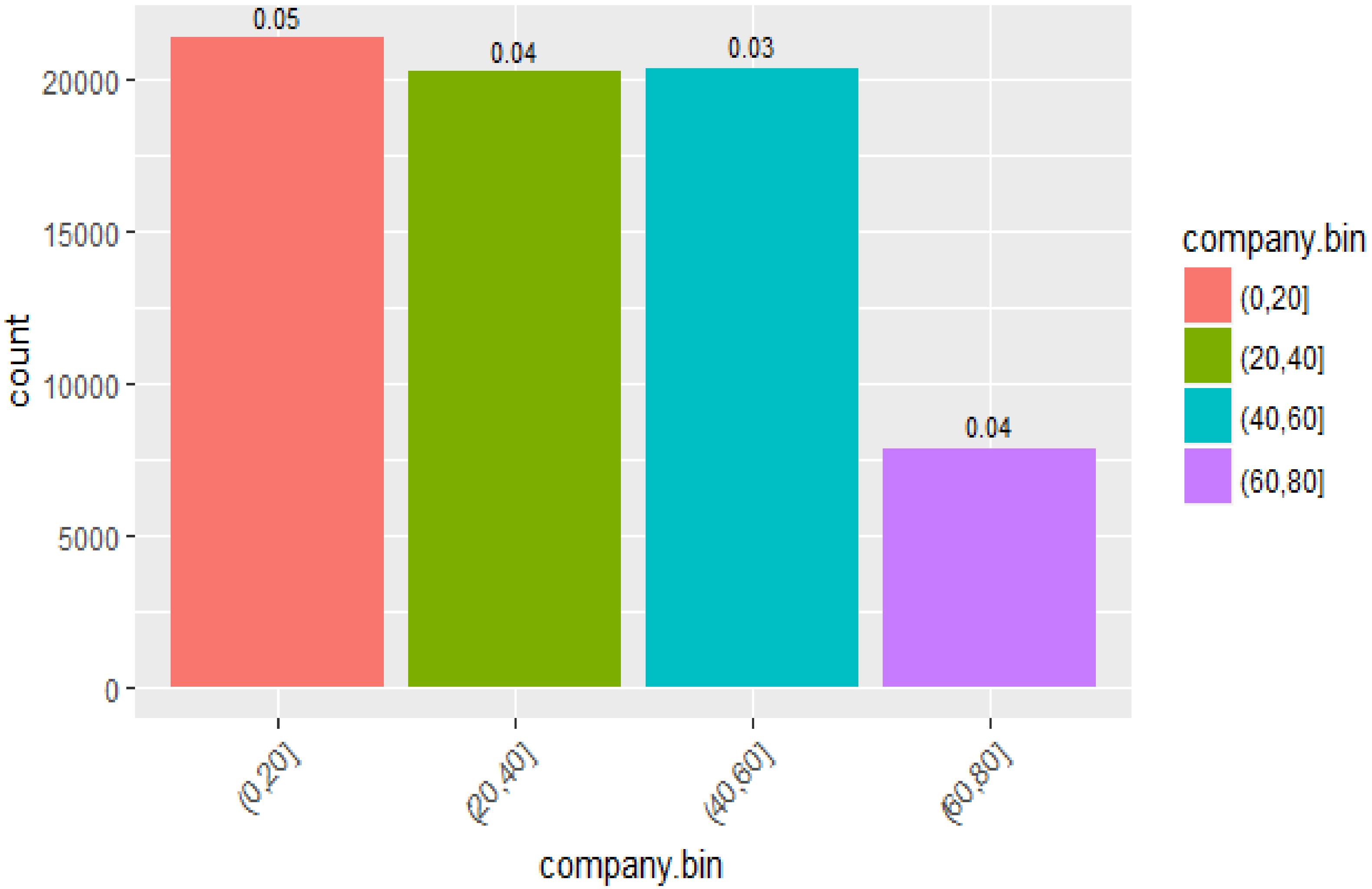
Since the rate of default in Company provided is more than twice of that in Others, type of residence might be an important predictor of default.

EDA No of months in current residence



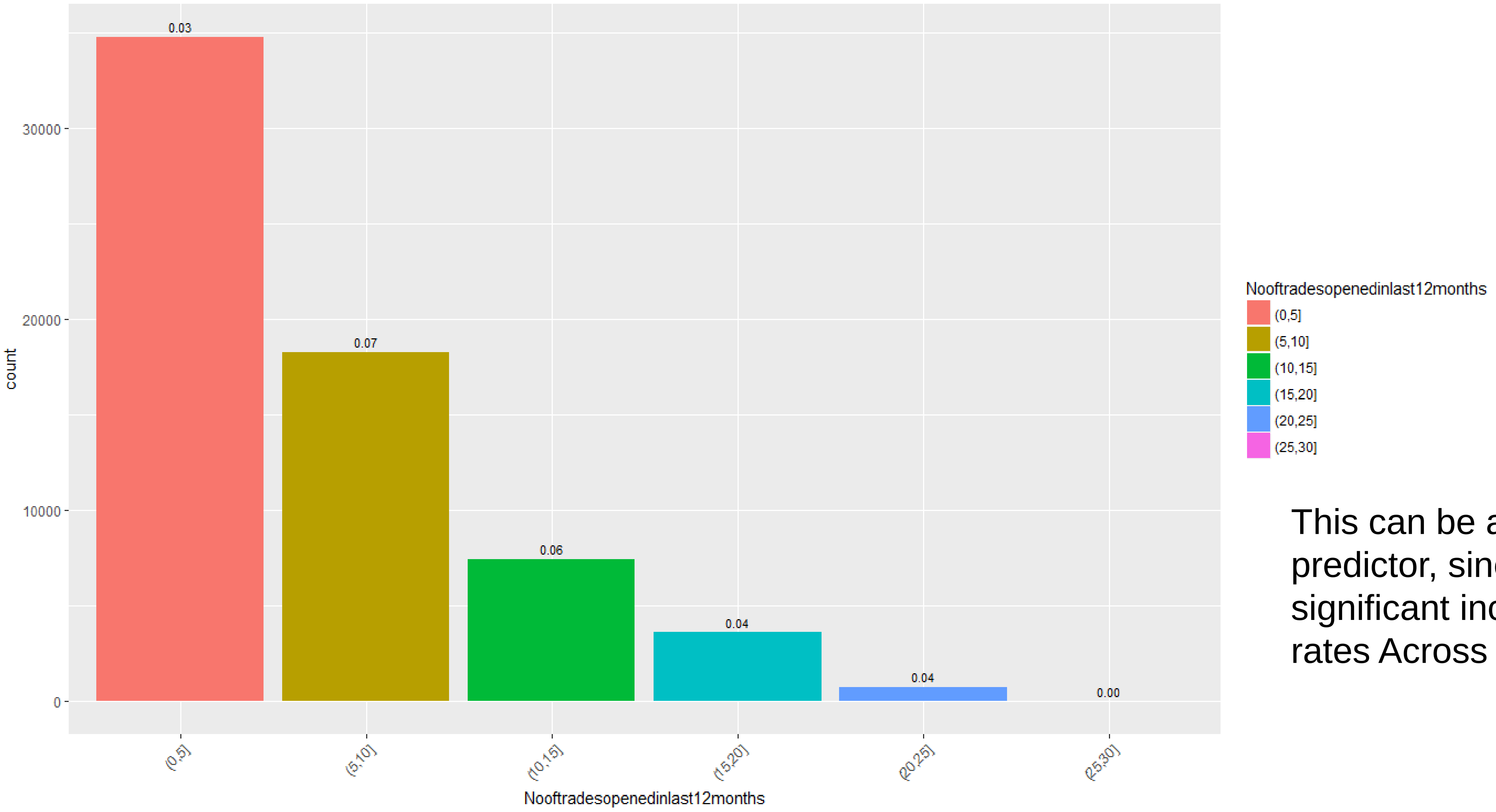
Default rate is significantly higher in 20- 40 months bracket than in other bins . Hence number of months in current residence might be an important predictor of the default rate

EDA No of months in current company



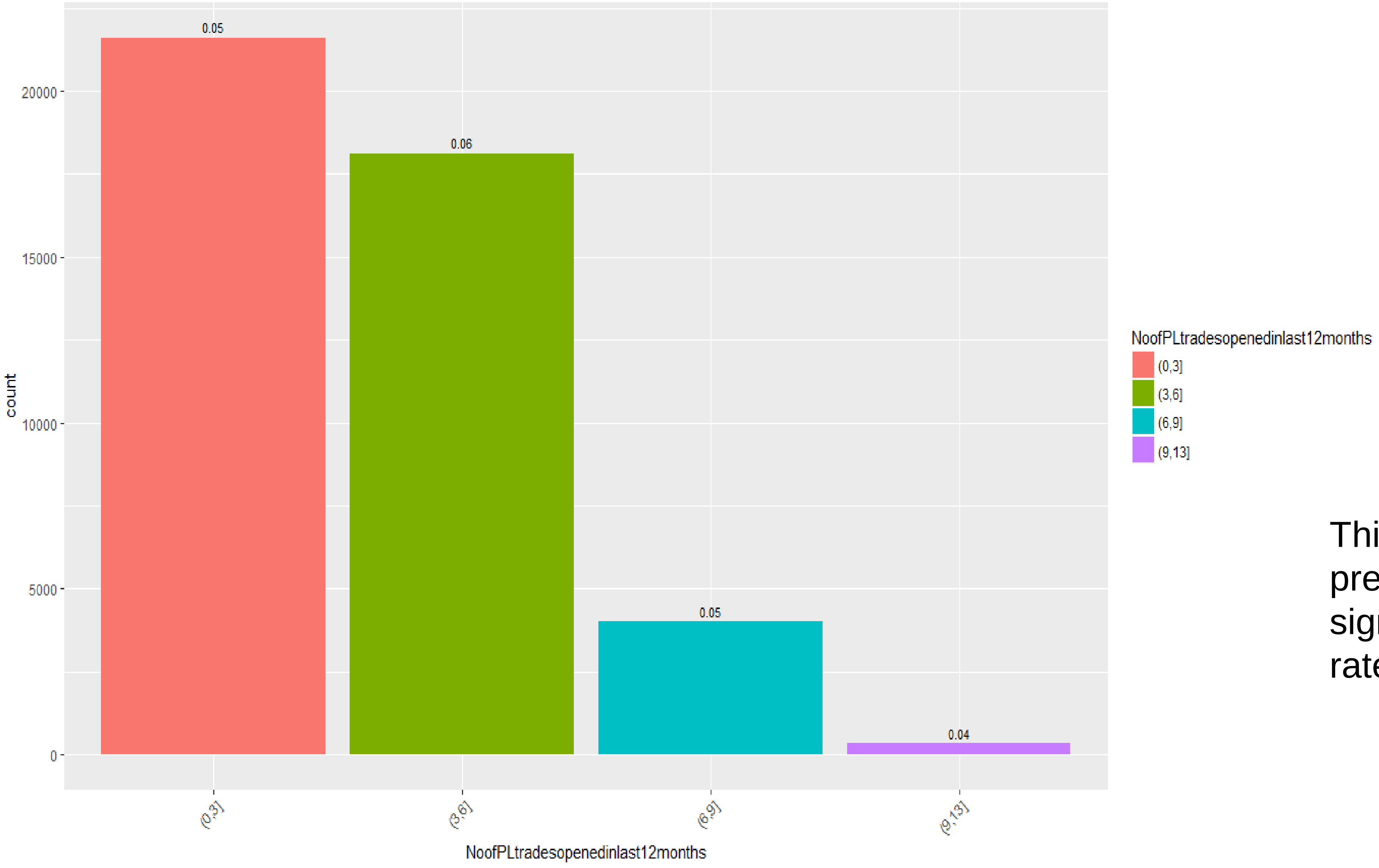
Default rate in the 0-20 months bin is significantly higher than in the 40-60 months bin, hence Number of months in current company might be an important predictor of default rate.

EDA No of trades in last 12 months



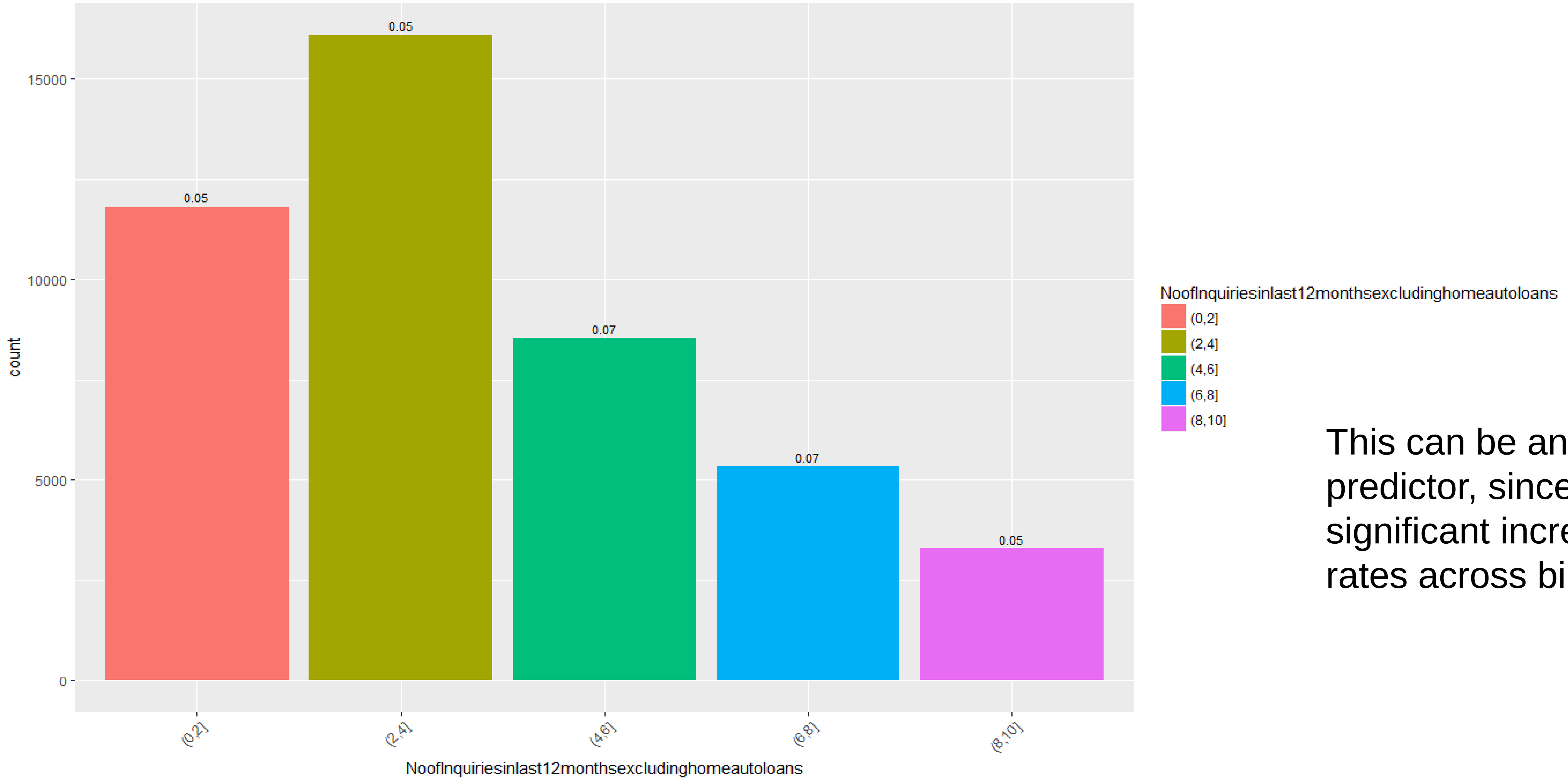
This can be an important predictor, since there is significant increase in default rates Across bins - [0,5] – (5,10]

EDA No of PL trades open in last 12 months



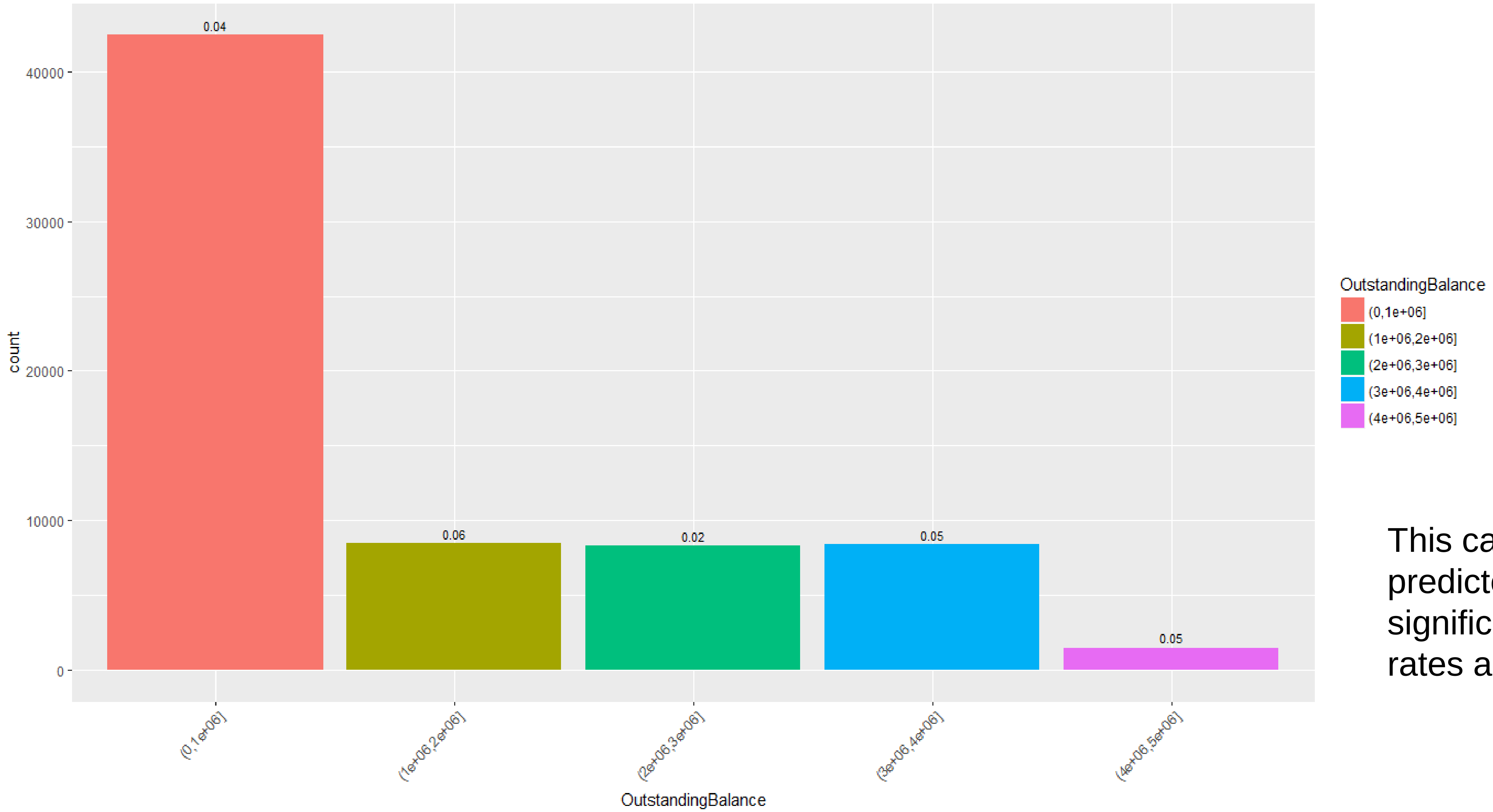
This can be an important predictor, since there is significant increase in default rates across bins.

EDA No of Inquiries in last 12 months excluding home & auto loans



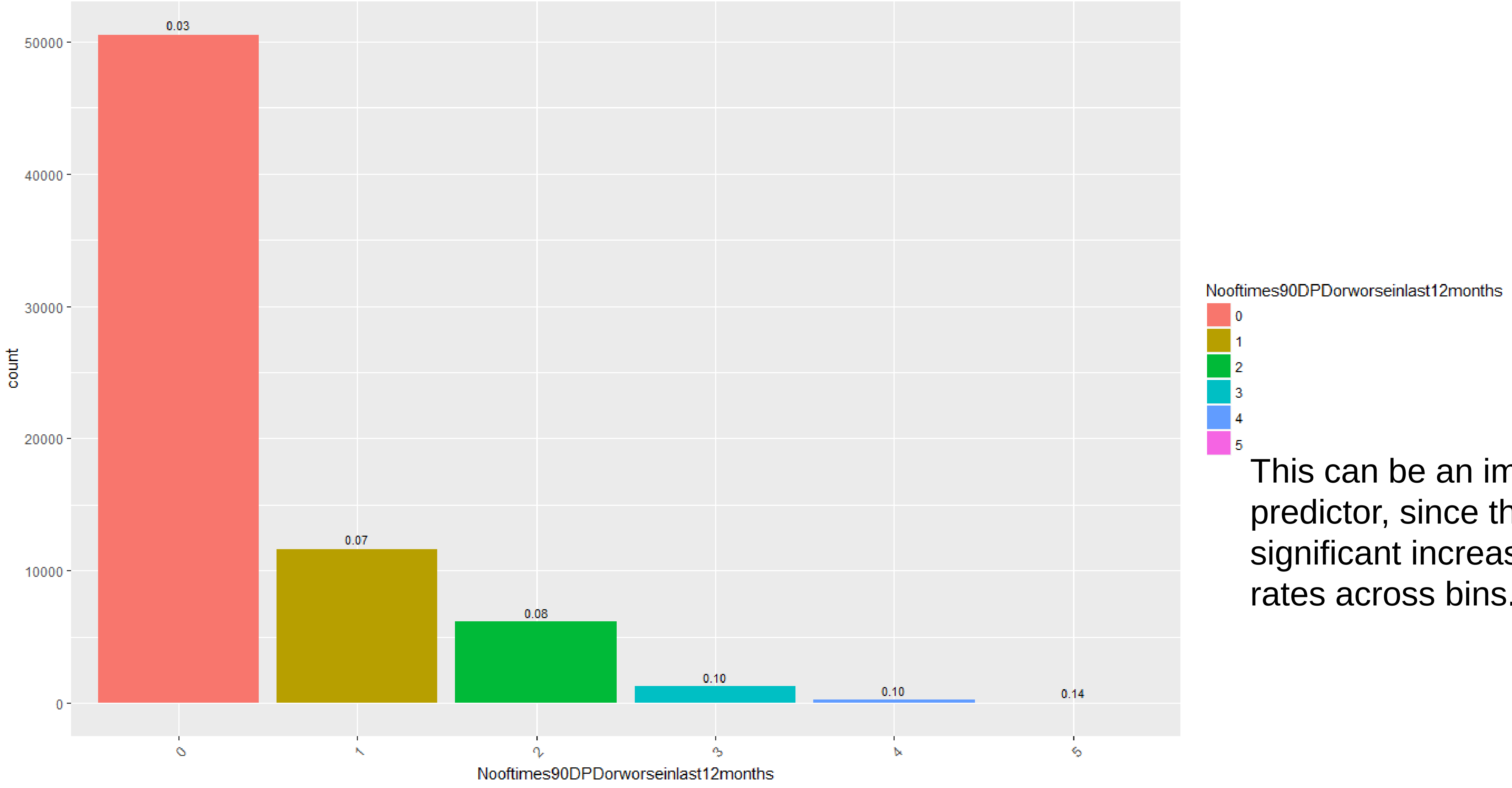
This can be an important predictor, since there is significant increase in default rates across bins.

EDA No of PL trades open in last 12 months



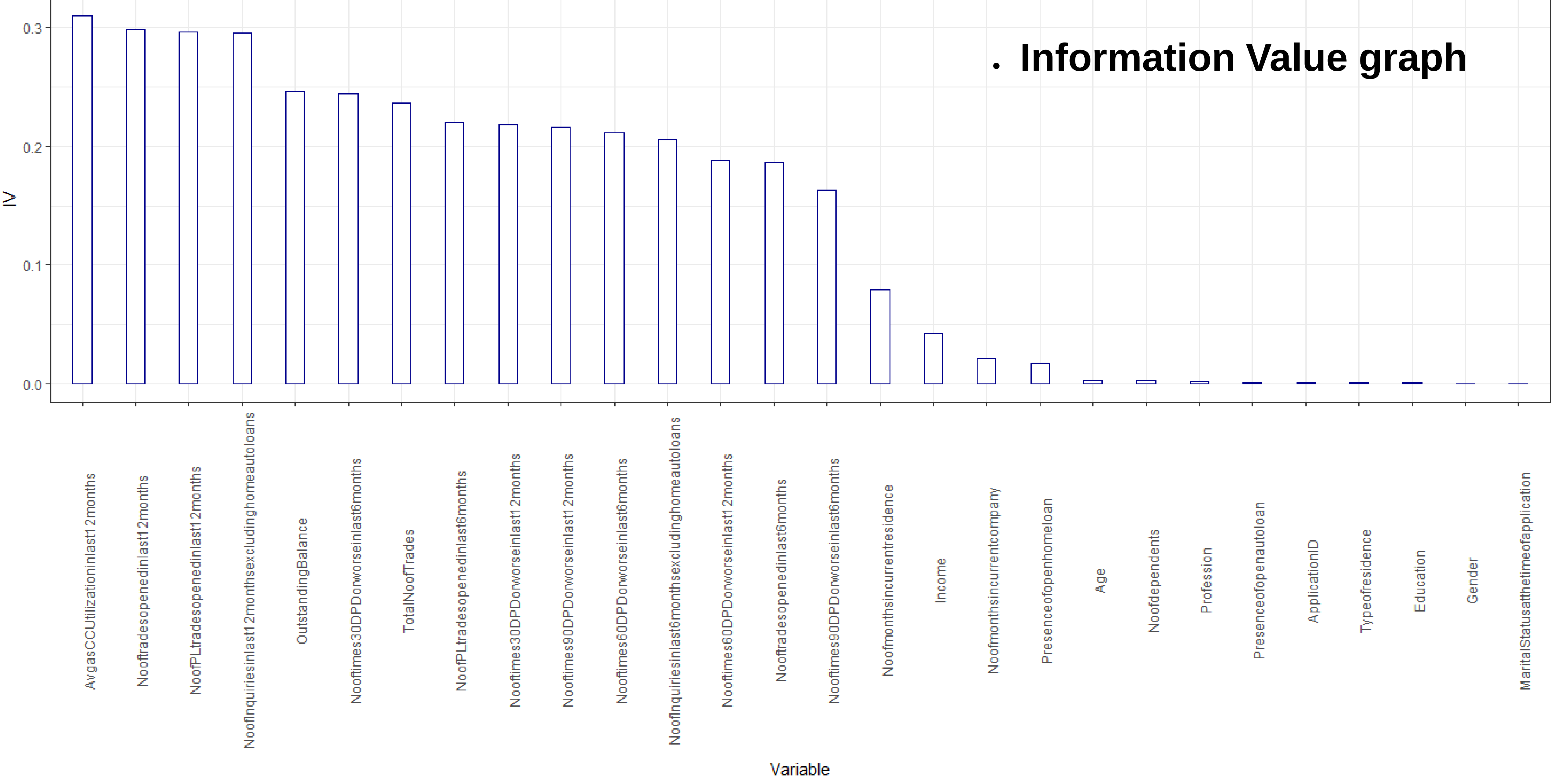
This can be an important predictor, since there is significant increase in defaul rates across bins.

EDA No of times 90 DPD or worst in the last 12 months



This can be an important predictor, since there is significant increase in default rates across bins.

- **Information Value graph**



- --**Information Value graph**:
  - Information Value Graph in the previous slide shows important variables in decreasing order of the Information value to the dependent variable PerformanceTag.
  - Most of the Important variables are from Credit Bureau data.

- --**woe_data**:
  - This data set created contains woe values for all variables, this will also take care of missing values.

- --**Model building**:
  - Based on these important variables a Logistic regression model is built and is evaluated on its accuracy, sensitivity and specificity .
  - Built models using other methods of classification like Decision Trees, Random Forrest and chose best one out of it using Model evaluation techniques like ROC curve and k-fold Cross validation.

- --**Application Scorecard**:
  - We will built the application score card as per the business problem and the final model using the scorecard package.
  - Financial benefit analysis o the model is carried out.

Logistic Regression on complete data - 1

- Results of Logistic regression:

```
summary(final_model)


Call:
glm(formula = Performance.Tag ~ Avgas.CC.Utilization.in.last.12.months +
    No.of.trades.opened.in.last.12.months + No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. +
    Outstanding.Balance + No.of.times.30.DPD.or.worse.in.last.6.months +
    Total.No.of.Trades + No.of.PL.trades.opened.in.last.6.months +
    No.of.times.30.DPD.or.worse.in.last.12.months + No.of.times.90.DPD.or.worse.in.last.12.months +
    No.of.times.60.DPD.or.worse.in.last.6.months + No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.,
    family = "binomial", data = bal_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7691  -1.1075   0.7409   1.0456   1.8908
```

Logistic Regression on complete data - 2

```
Coefficients:
                                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                   0.0006513  0.0095229   0.068  0.94547
Avgas.CC.Utilization.in.last.12.months                        0.3039553  0.0186172  16.327  < 2e-16 ***
No.of.trades.opened.in.last.12.months                         0.1688853  0.0250729   6.736 1.63e-11 ***
No.of.Inquiries.in.last.12.months..excluding.home...auto.loans. 0.2165643  0.0229600   9.432  < 2e-16 ***
Outstanding.Balance                                           0.1074644  0.0230076   4.671 3.00e-06 ***
No.of.times.30.DPD.or.worse.in.last.6.months                  0.1387625  0.0274502   5.055 4.30e-07 ***
Total.No.of.Trades                                            0.0597766  0.0239145   2.500  0.01243 *
No.of.PL.trades.opened.in.last.6.months                       0.0765000  0.0249790   3.063  0.00219 **
No.of.times.30.DPD.or.worse.in.last.12.months                 0.1466754  0.0267269   5.488 4.07e-08 ***
No.of.times.90.DPD.or.worse.in.last.12.months                 0.1052879  0.0250280   4.207 2.59e-05 ***
No.of.times.60.DPD.or.worse.in.last.6.months                  0.1241348  0.0279691   4.438 9.07e-06 ***
No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.  0.1738025  0.0245889   7.068 1.57e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 67799  on 48906  degrees of freedom
Residual deviance: 62966  on 48895  degrees of freedom
AIC: 62990
```

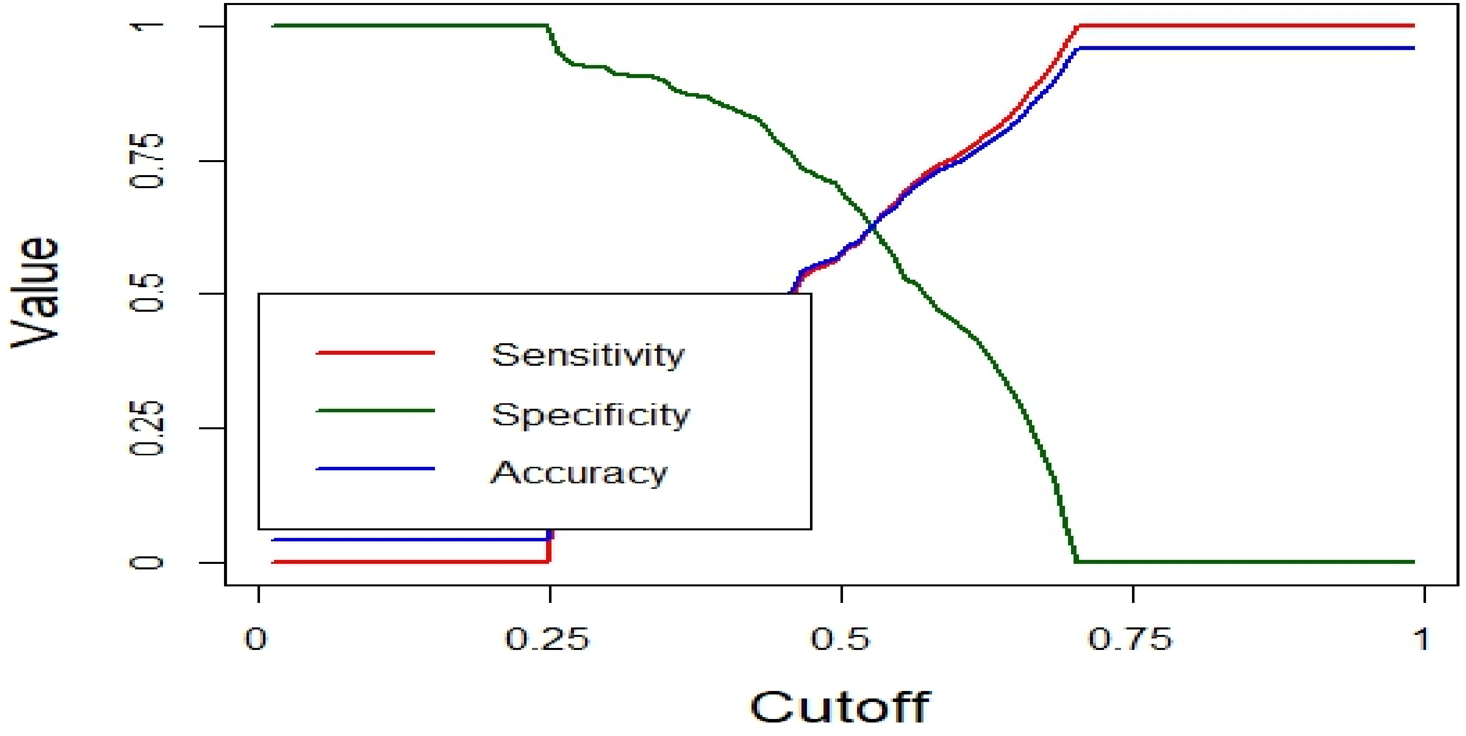
Logistic Regression on complete data - 3

- Thus most significant variables from Logistic Regression on Demographic data are:

1) Avgas.CC.Utilization.in.last.12.months
2) No.of.Inquiries.in.last.12.months..excluding.home...auto.loans.
3) No.of.Inquiries.in.last.6.months..excluding.home...auto.loans.

Logistic Regression on complete data - 4

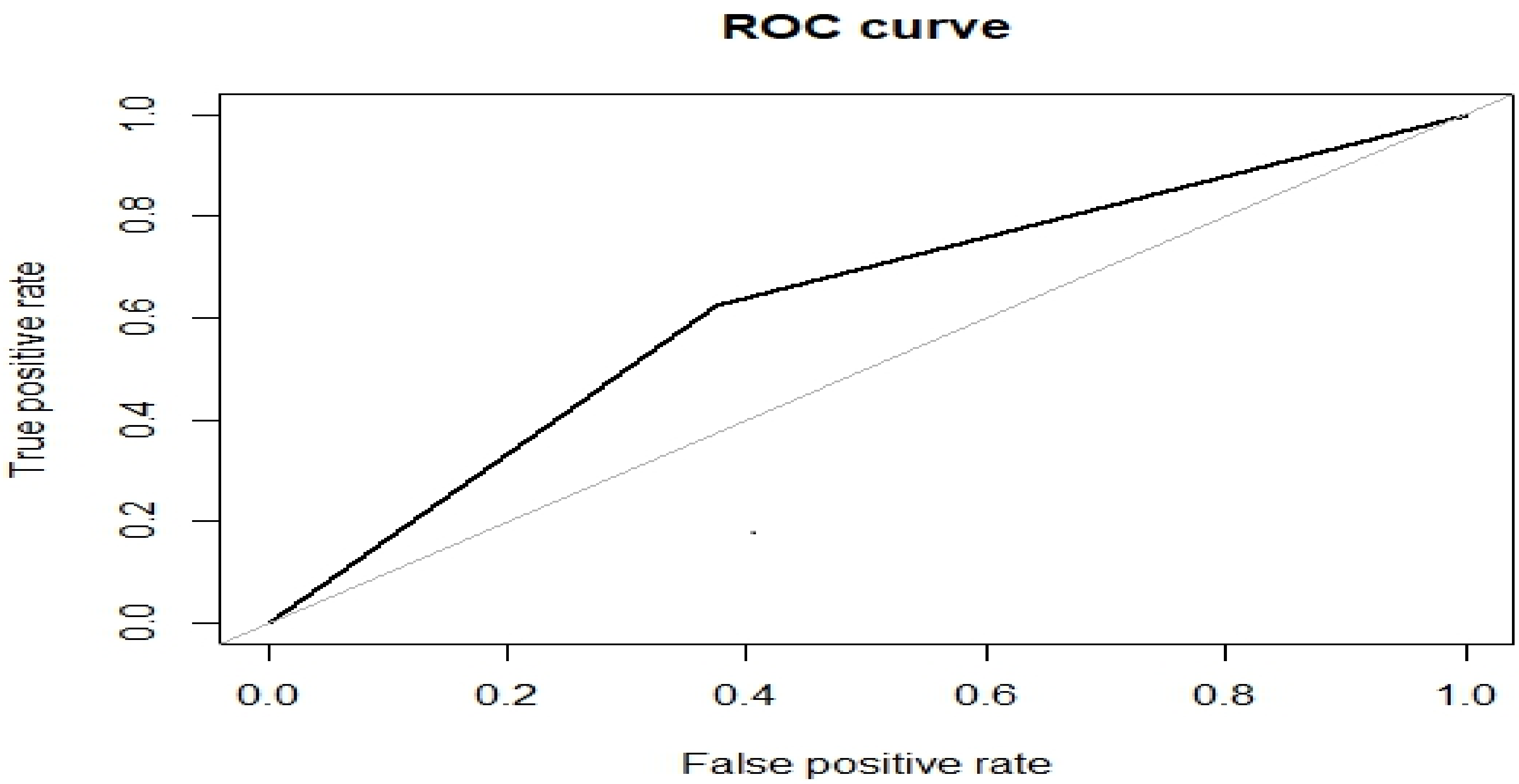
- Optimum value of cut-off



Logistic Regression on demographic data - 5

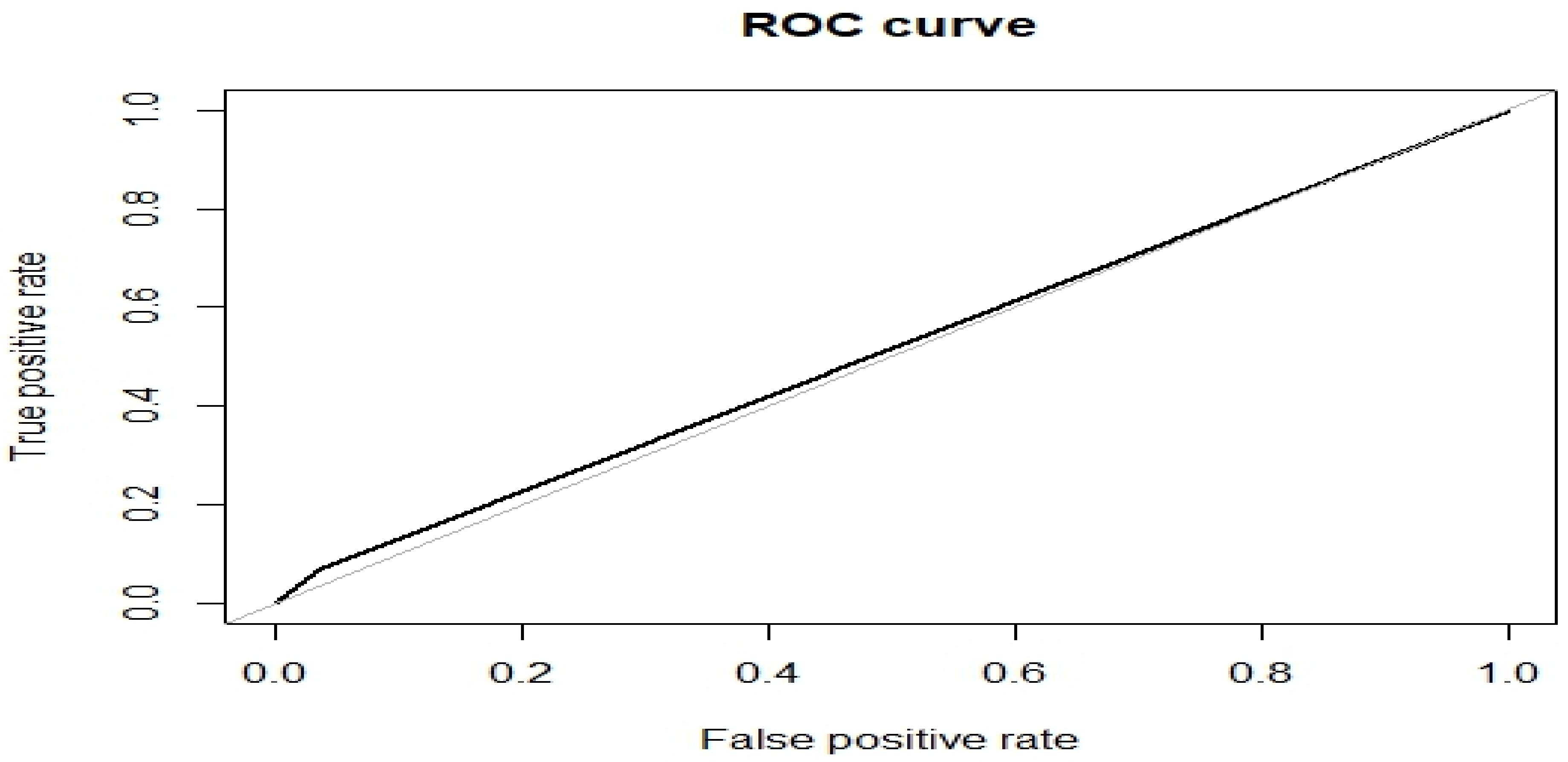
- Cut-off value = 0.5247475
- Accuracy, Sensitivity and Specificity at this Cut-off value
- Accuracy = 0.626
- Sensitivity = 0.625
- Specificity = 0.625

ROC Curve

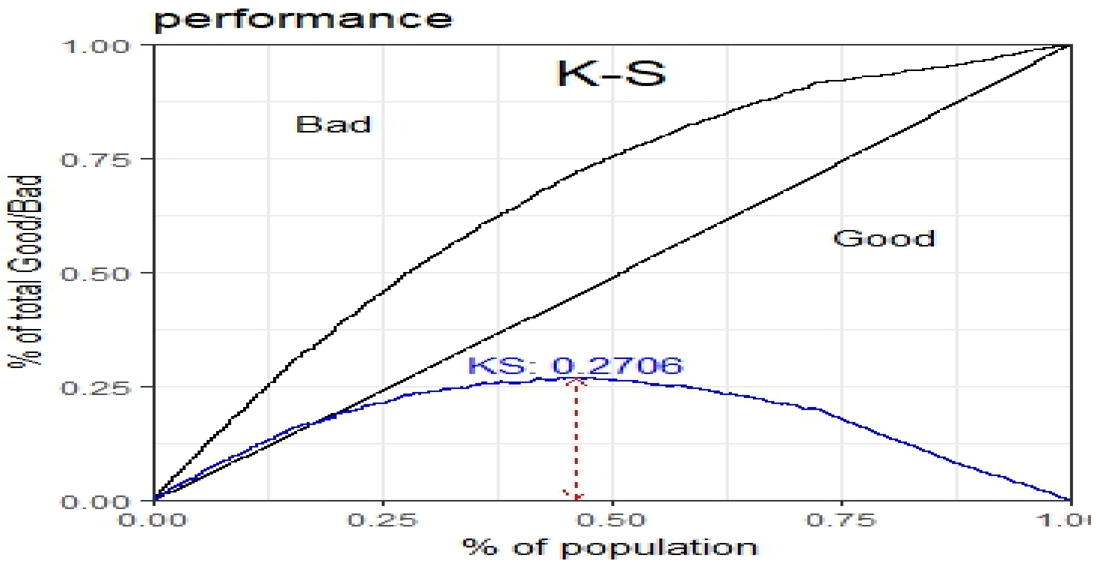


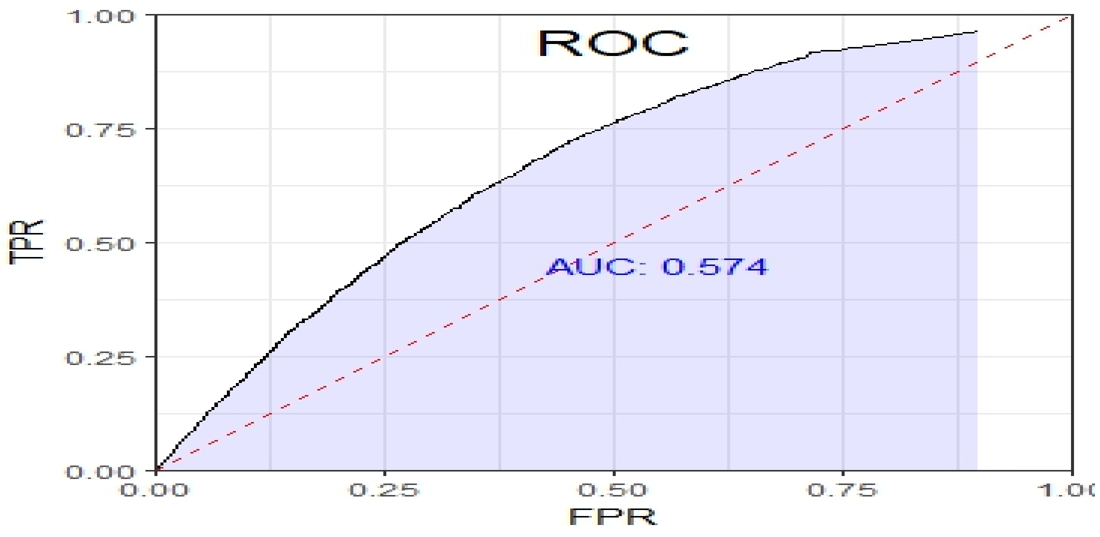
ROC curve
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

ROC Curve –Random Forest



Model Performance





Financial Benefit Analysis - 1

Financial analysis----

assumption -
everyone is give a credit limit of 1 lakh
good customer - 30% profit (rs. 30,000)
bad customer - 100% loss (rs 1,00,000)

total - 69867
without model - total credit - (66920+2947)*100000 = 6986700000
66920 - good - profit - 66920*30000 = 2,00,76,00,000
2947 - bad - loss - 2947*100000 = 29,47,00,000
4.2% defaulters

Financial Benefit Analysis - 2

total -
with model -
score_model

actual	0	1
0	41572	25348
1	1067	1880

only 41572+1067 people will receive the credit card. out of which 1067 will default as per the score cut off.
total credit - (41572+1067)*100000 = 4,26,39,00,000
profit - 41572*30000 = 1,24,71,60,000
loss - 1067*100000 = 10,67,00,000
2.5% defaulters

A credit loss of 294700000-106700000 = 18.80 Crore is saved by using the model