Capstone Project – BFS
CredX

Minerva S N

Muhammad Inamullah V

Prathap Reddy

Problem Statement:

- CredX is a leading credit card provider that gets thousands of credit card applications every year. But in the past few years, it has experienced an increase in credit loss. The CEO believes that the best strategy to mitigate credit risk is to ‘acquire the right customers’.

Objective:

- To identify the right customers for CredX by building a predictive model.
- To identify the variables that factors in affecting the credit risk.
- Assess the Financial benefit of this project.

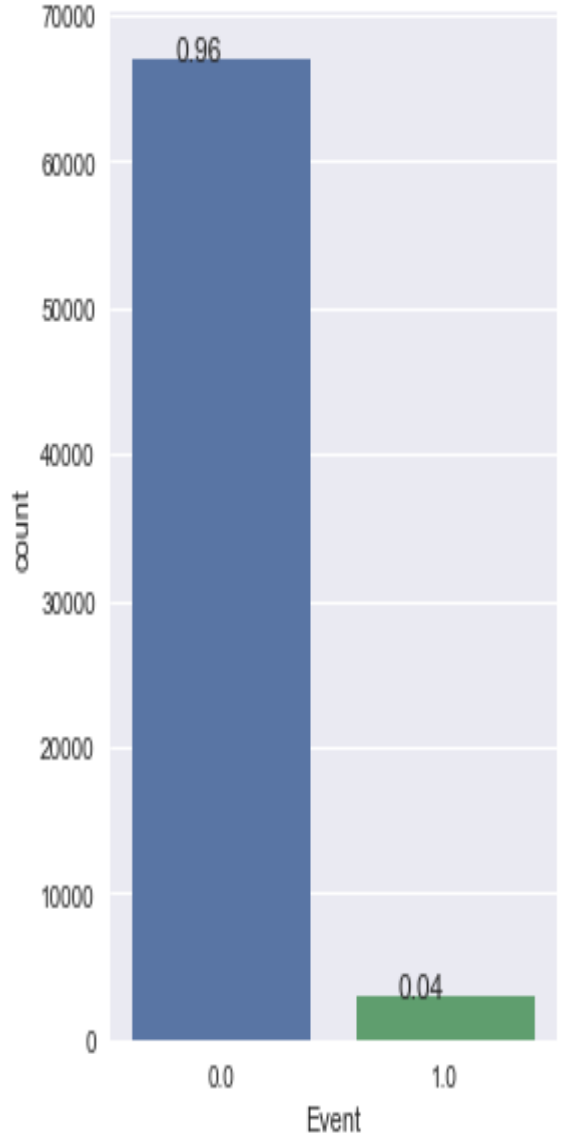
Data Understanding:

- Structure of the data: “Demographic data” has 11 independent variables and “Credit Bureau” has 18 independent variables. Total number of records are about 71295 in both datasets.
- “Credit Bureau” record does not have any duplicates, but the “Demographic data” has 3 duplicates. Only the “Application IDs” are duplicates, the rest of the columns are different in “Demographic data”. Since the count of duplicates were less and only 1 of them has a event, those records are removed from the data sets.
- 2% of the target values i.e., “Performance Tag” had missing values, which means the bank has rejected their application. Those needed to be removed.

Missing Values(in %):

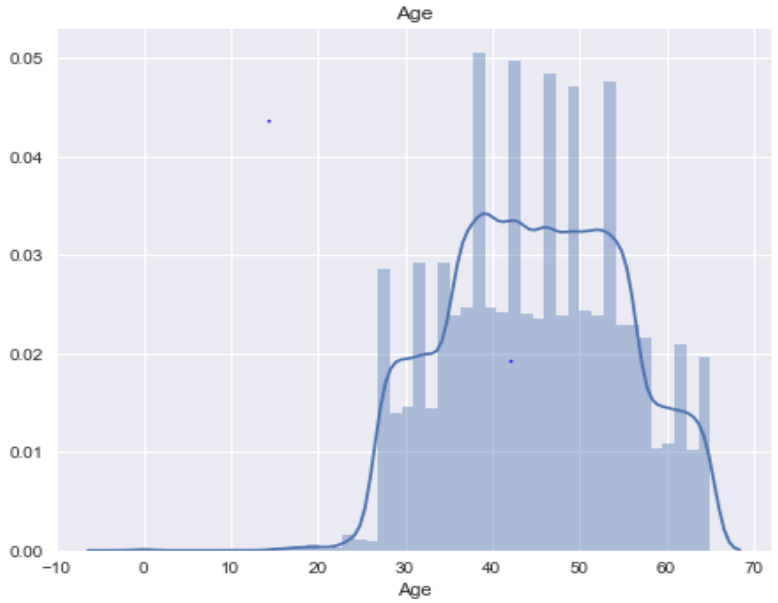
Application ID	0.00
Age	0.00
Gender	0.00
Marital Status (at the time of application)	0.01
No of dependents	0.00
Income	0.00
Education	0.17
Profession	0.02
Type of residence	0.01
No of months in current residence	0.00
No of months in current company	0.00
Performance Tag	2.00

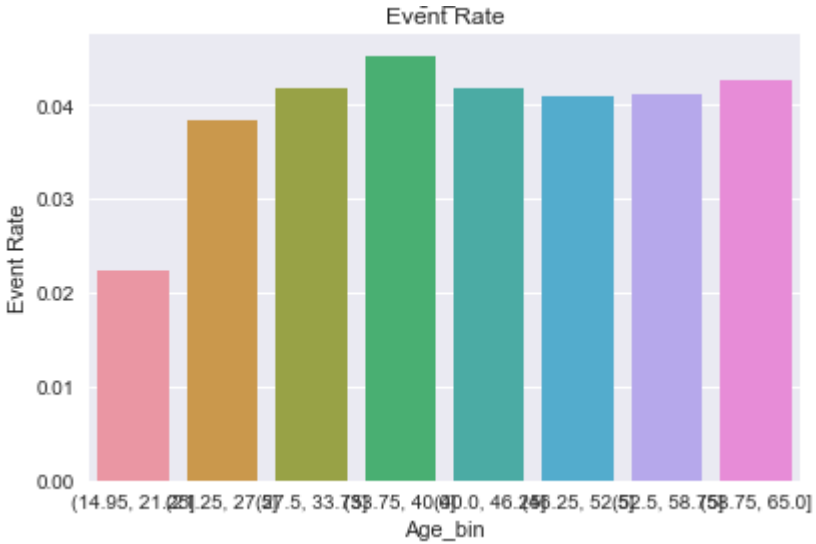
- Apart from Target Variables. “Profession”, “Type of residence” and “Marital Status” have missing values. In these cases, missing values are assigned to a group in that category with higher weightage as the missing values are less than 0.5%.
- “Education” also has missing values but more than one variable have higher weightage. So, the missing values are assigned to a group which have similar default rate as that of missing values.
- “Gender” and “No. of Dependents” have a total of five missing values which are removed as these are negligible.
- **Finally, plotted the “Event Rate” which is 4%.**

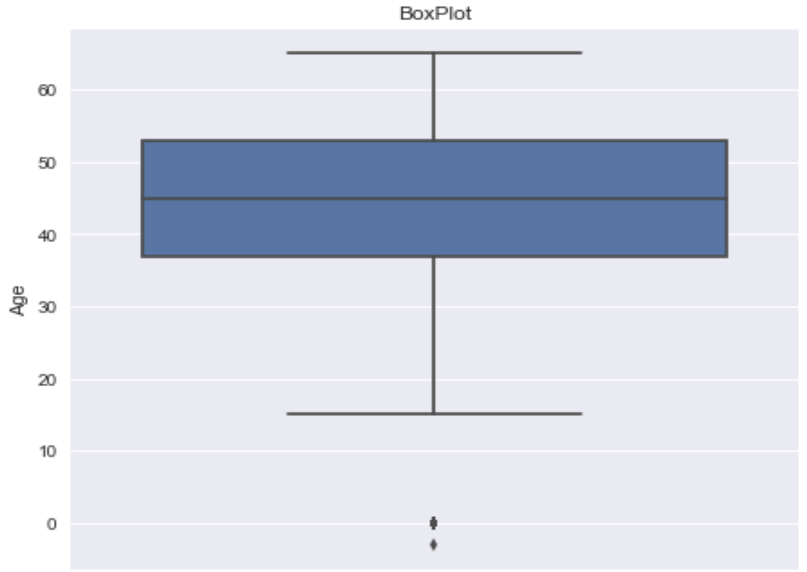


Exploratory Data Analysis:

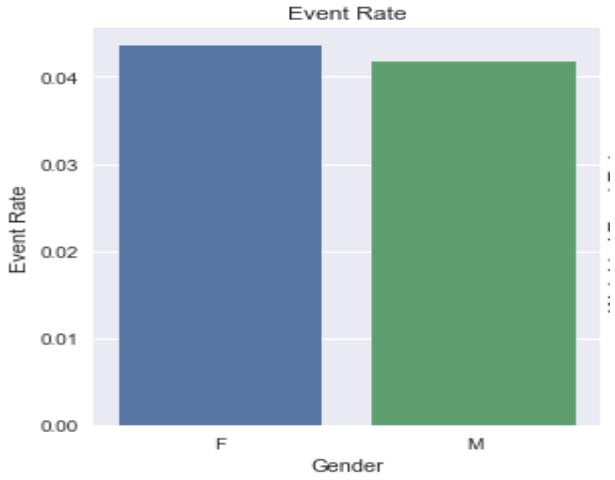
- Variable “Age” is center skewed but has some outlier such as 0 and -3.
- Those outliers are replaced with 5th percentile value.
- “Event Rate” across each “Age” bin are similar, so this variable does not have a good discriminative power.

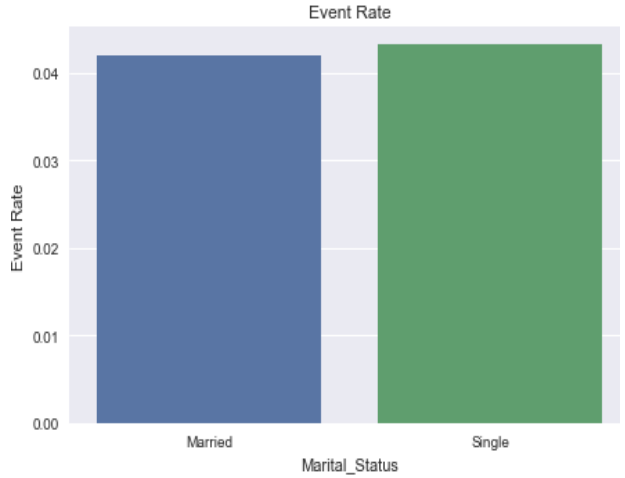




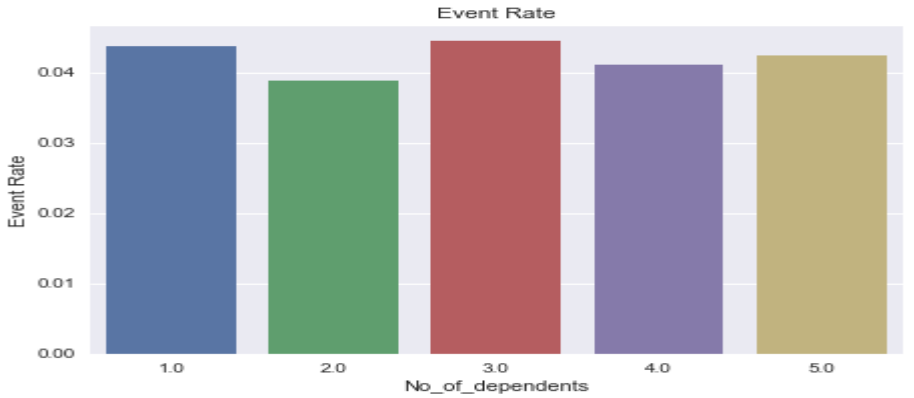


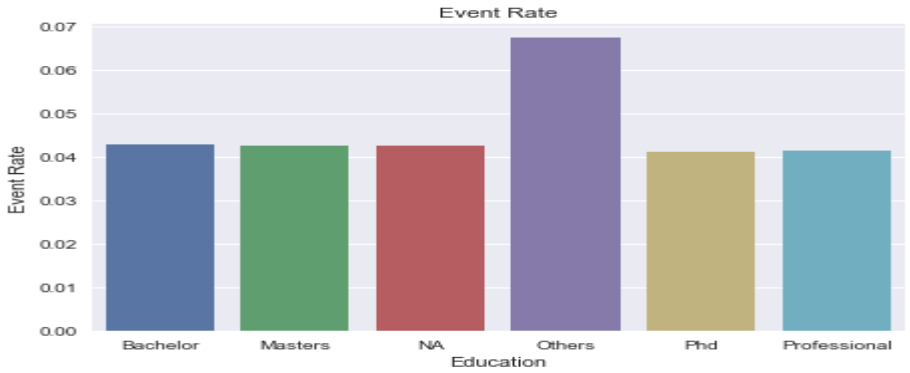
- “Gender” and “Marital Status” also do not have good discriminative power.





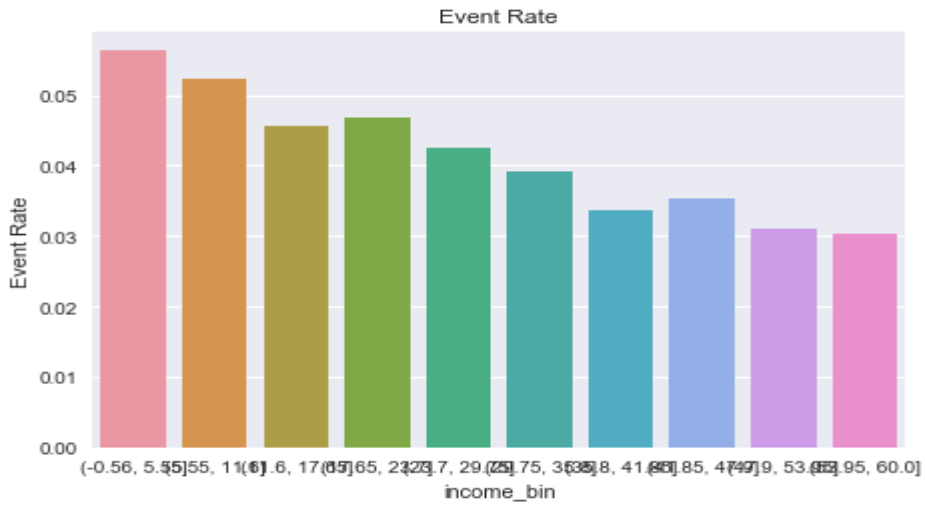
- “No_of_Dependents” and “Education” also are insignificant. “Others” in “Education” had higher default rate but its weightage on the data is very low.

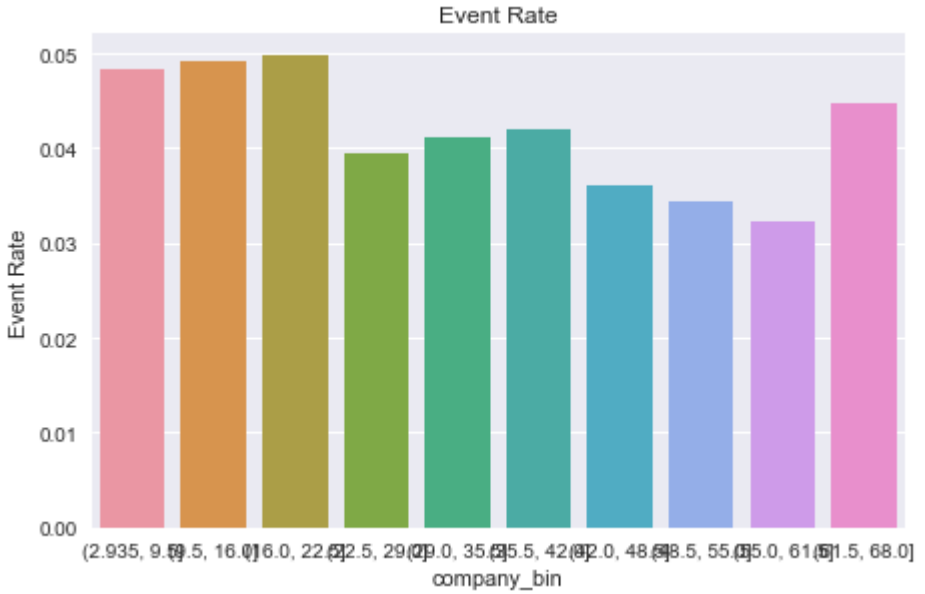




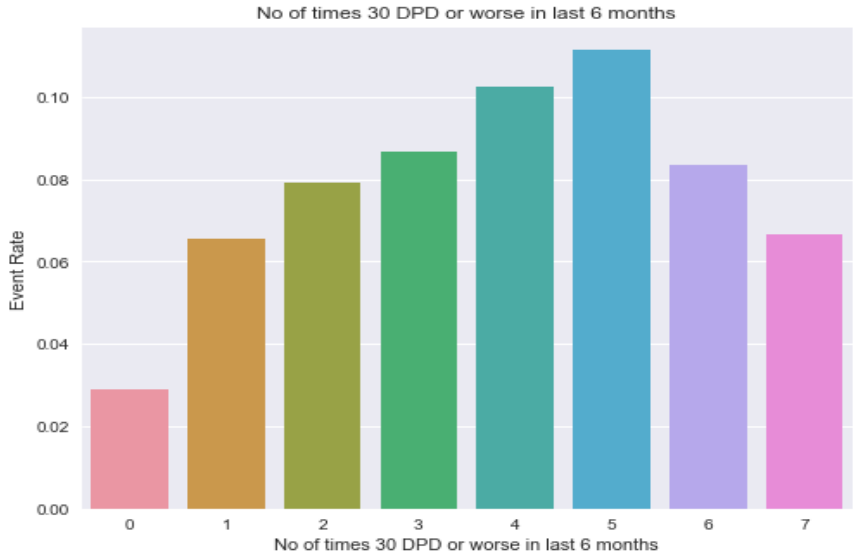
- “Income” follows a downward slope indicating that higher the income lower is the default rate.

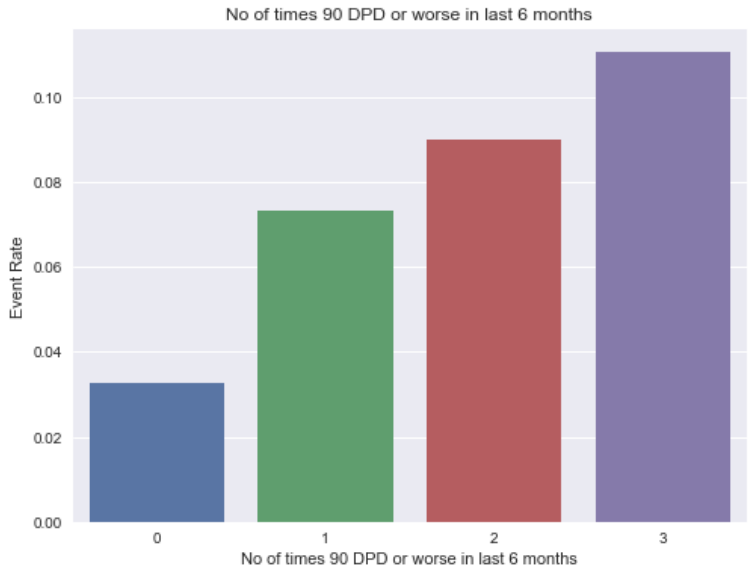
- “No_of_months” in current company: Applicants with less than 20 months on current company has higher default rate than others.
- These two variables are significant from the demographic data.

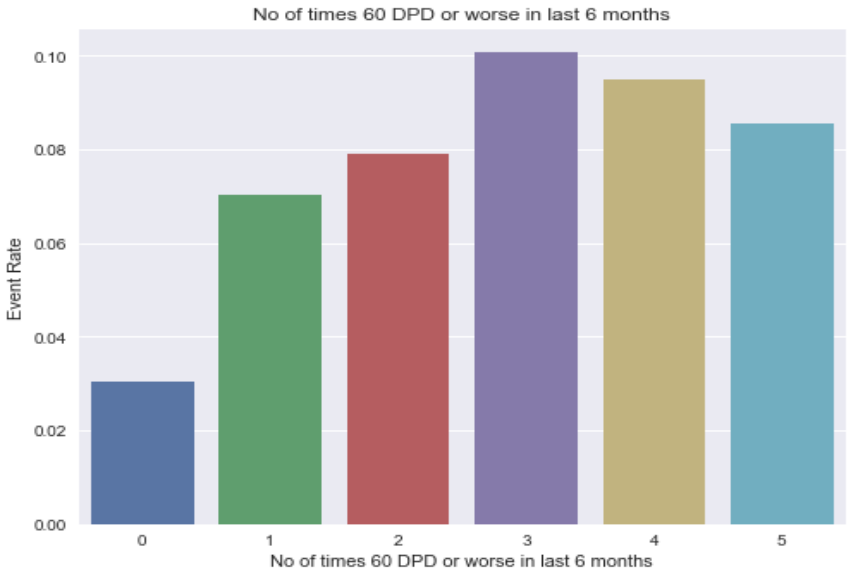




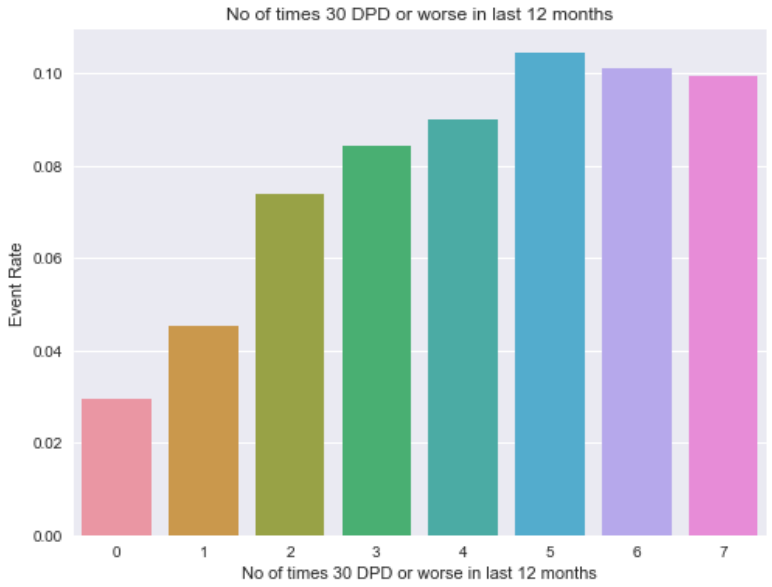
- DPDs in last 6 months: 30,60 and 90 DPD follow an upward slope indicating higher default rates at higher no. of DPDs.
- These variables are highly correlated. So, selecting 30 DPD which has higher IV value.

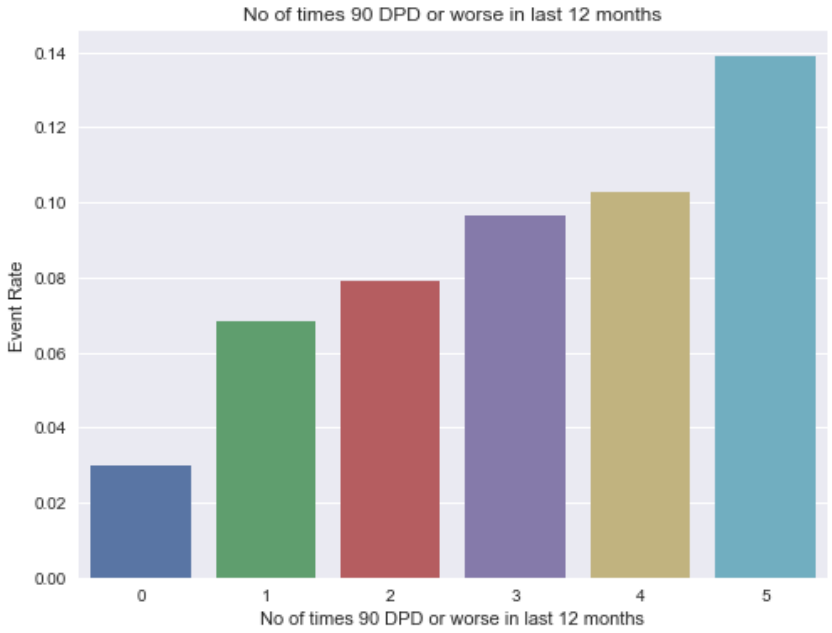


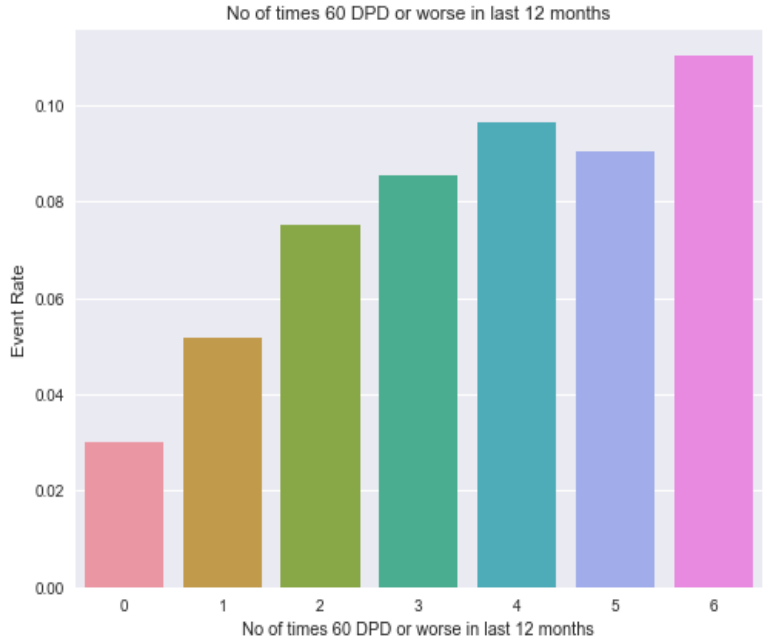




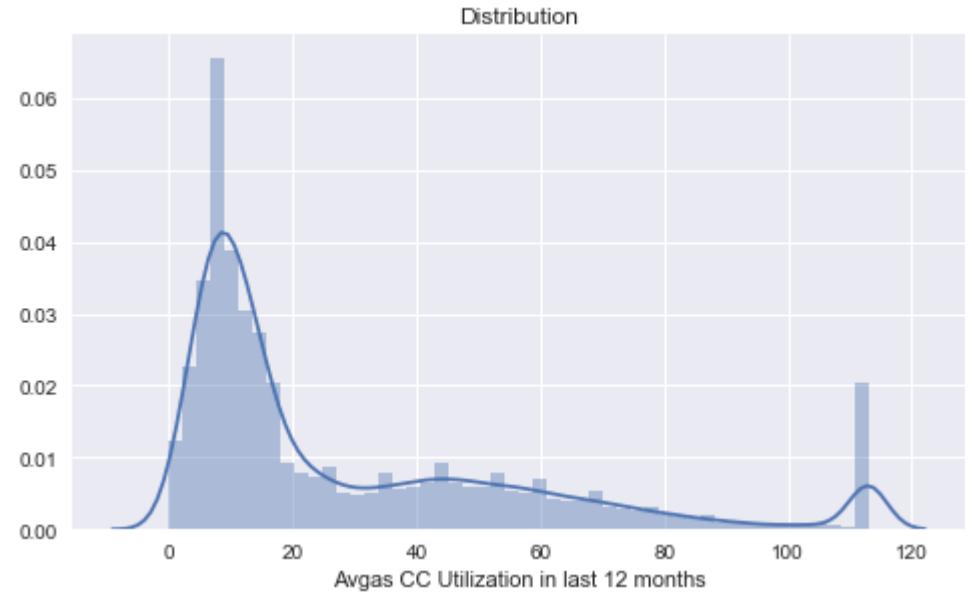
- DPDs in last 12 months: 30,60 and 90 DPD follow an upward slope indicating higher default rates at higher no. of DPDs.
- These variables are highly correlated. So, selecting 30 DPD which has higher IV value.

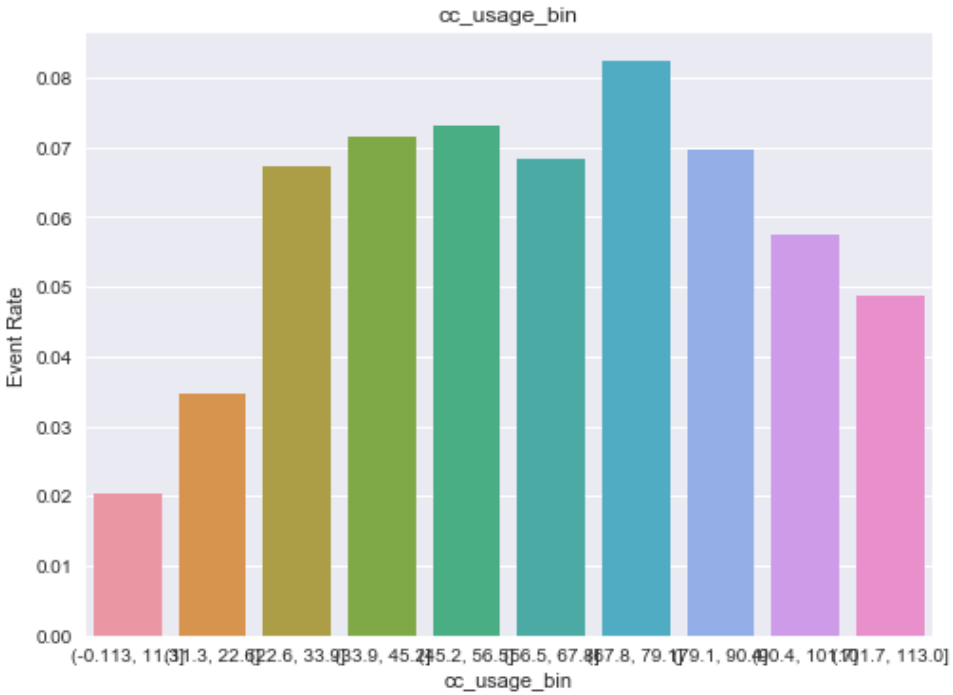




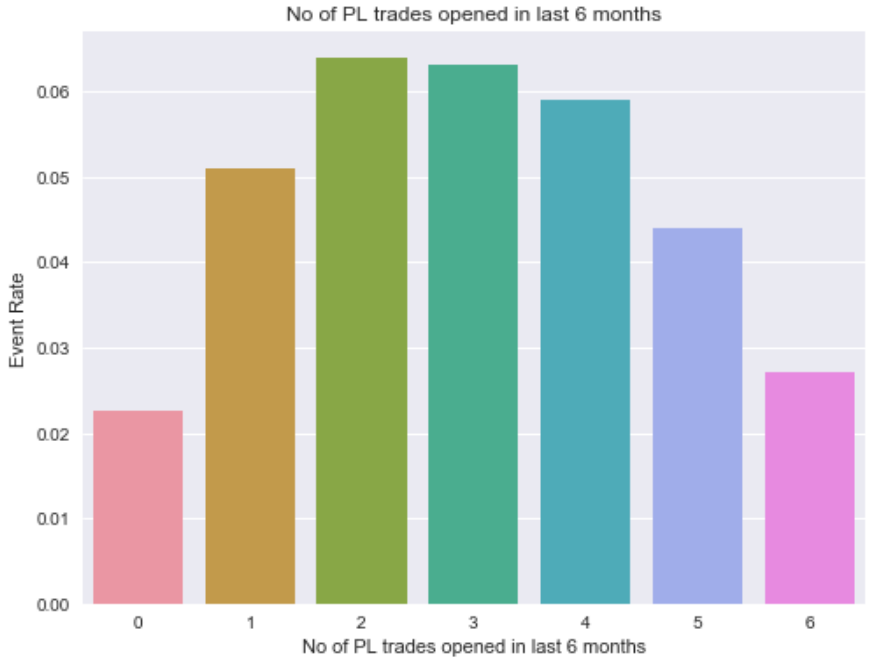


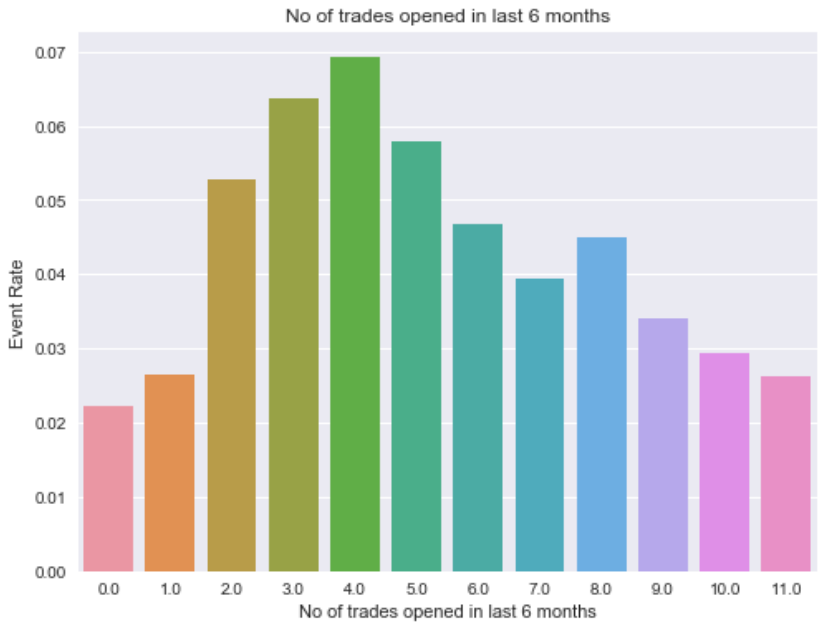
- “Avgas CC Utilization” : It is left queued. It has 1.46% missing values which is imputed using the median to retain the distribution of this variable.
- Default rate are higher with CC utilization between 33 and 80.
- Average CC utilization where applicants have defaulted is approximately 40
- This variable has good discriminative power.



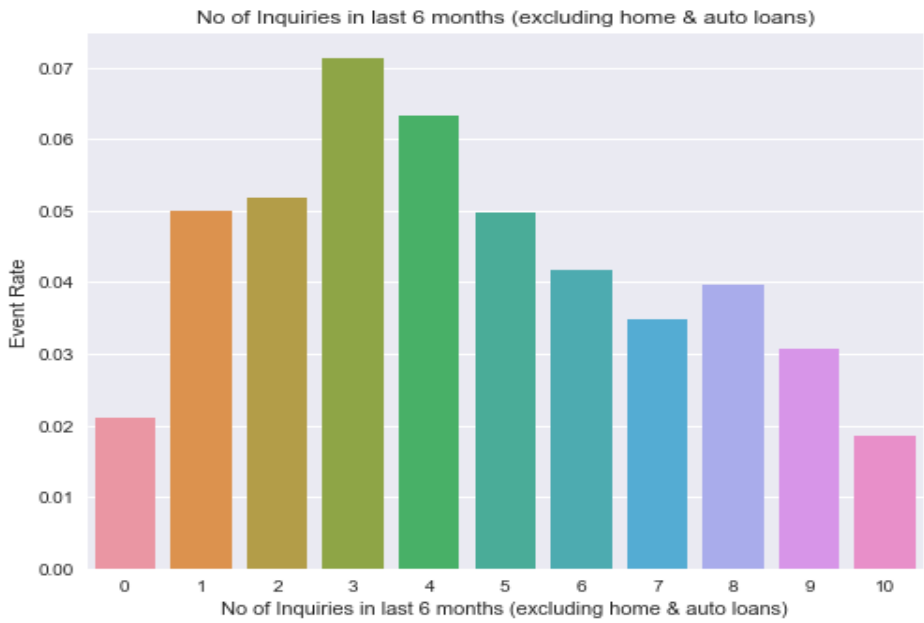


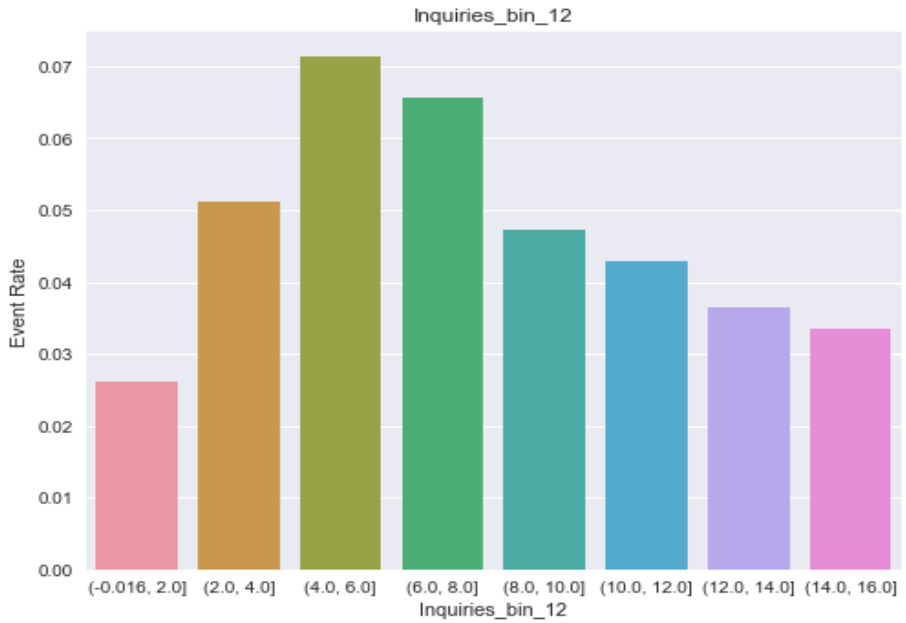
- Trades and PL Trades: Default rate for both of these trades follow a similar pattern. It increases with increasing values of trades up to a point, then it starts to decrease.
- IV values shows that Trades adds more information, so selecting this for the model. But need to validate the model performance with the other variables.





- Default rate for “No. Of Inquiries” also follow the same pattern as “Trades”.
- So, we have identified the discriminative power of the variables from Demographic and Credit Bureau data. Lets now look at IV for all of these variables and decide which variables to choose for the model.





IV

Variable	IV	Variable	IV
Age	0.004143267	No of times 90 DPD or worse in last 12 months	0.215702631
Gender	0.000326378	No of times 60 DPD or worse in last 12 months	0.188267915
Marital_Status	9.53E-05	No of times 30 DPD or worse in last 12 months	0.218655266
No_of_dependents	0.002654867	Avgas CC Utilization in last 12 months	0.321889247
Income	0.0428251	No of trades opened in last 6 months	0.18730502
Education	0.000783693	No of trades opened in last 12 months	0.293595262
Profession	0.00222918	No of PL trades opened in last 6 months	0.2242227
Type_of_residence	0.000925048	No of PL trades opened in last 12 months	0.258575202
Months_in_current_residence	0.0707418	No of Inquiries in last 6 months (excluding home & auto loans)	0.113378867
Months_in_current_company	0.022821479	No of Inquiries in last 12 months (excluding home & auto loans)	0.245269417
Application ID	0.00150062	Presence of open home loan	0
No of times 90 DPD or worse in last 6 months	0.162701693	Outstanding Balance	0.246658269
No of times 60 DPD or worse in last 6 months	0.211320271	Total No of Trades	0.232279858
No of times 30 DPD or worse in last 6 months	0.244295829	Presence of open auto loan	0.001656772

Highlighted variables will be used in building the model because these variables have a good discrimination power and have decent IV values. Some of the variables are correlated those variables are left out because their correlation variables have higher IV.

Test Train Data Split

- The Combined data Set has to be separated into “Test Data” and “Train Data” on 30:70 ratio.
- These Data Sets will be used on build different models for predicting the Target Variable – “Event rate”.

Standardization Technique

- After Test Train split we standardize the features.
- For standardization we transform all the selected variables to WOE values. Also we skip one hot encoding by using WOE on categorical values.

Approach to Model Building

- Building a “Logistic Regression” model with “Demographic Data” alone initially and identify the Driver Fields using Logistic Regression model and RFE(Recursive Feature Elimination) Techniques.
- After getting the Significant drivers from the “Demographic Data” Set, combine these Fields of” Demographic Data” Set with “Credit Bureau” Data set and Perform Logistic Regression. Use IV based elimination Techniques to identify the driver Fields with greater predictive ability for predicting the Target Variable.
- From some 15 driver fields will be plotted on heat map to get the correlated fields. Thus helping us to identify the top features and eliminating one of the correlated features which are redundant.

Handling Class Imbalance

- There are multiple ways to handle class Imbalance, for this project we are selecting Adaptive Boosting since most of the variable's IV say that they are medium to weak classifier.

- To overcome the disadvantage of Adaptive Boosting sensitivity to outliers, we use WOE / appropriate percentile values to impute the outliers

Model Building and evaluation

- Initially the Train Data Set will be used for finding the driver variables and performance parameters of the “Logistic Regression” model.
- Then the same datasets will be used to build a “Decision Trees” and the model performance will be compared against the performance of the model built using a “ Logistic Regression” model.
- If performance of “Decision Tree” model is better, it will be used else “Logistic Regression” model will be used.
- If , the model built using Decision tree is over fitting, then random forest technique will be used to build the model and its performance will be compared and evaluated.

Feature Ranking sorting and selecting to build final model

- The features will be ranked and sorted to select some 10 features and model will be built using these highly predictive variables .
- Then Intercepts and Co-efficients will be extracted from Logistic Model to derive Business Insights necessary to derive recommendations.

Applicant Risk or Likelihood of Defaulting

- Each and every applicant will be identified with the likelihood to default using Application Scorecard.
- The score for each applicant will be calculated and threshold identified for rejecting the Application for Credit Card.

Business decisions and consequent benefits

- By Identifying the prospective Defaulter based on past behavioural attributes which have more predictive ability on the “Event Rate” of defaulting help CredX to automatically identify the Eligible Applicants and thus helps in stemming the foreseeable losses which cannot be otherwise identified using some set of objective criterion.