**Report**

**Credit Card Approval Model**

The entire dataset consists of two csv files- application records and credit records. The first one contains the demographic data of the customers and credits record contains the details of credit card holders.

There are 438557 rows in the applications dataset with 18 columns.

The 18 columns include:

|-- ID: integer (nullable = true)

|-- CODE\_GENDER: string (nullable = true)

|-- FLAG\_OWN\_CAR: string (nullable = true)

|-- FLAG\_OWN\_REALTY: string (nullable = true)

|-- CNT\_CHILDREN: integer (nullable = true)

|-- AMT\_INCOME\_TOTAL: double (nullable = true)

|-- NAME\_INCOME\_TYPE: string (nullable = true)

|-- NAME\_EDUCATION\_TYPE: string (nullable = true)

|-- NAME\_FAMILY\_STATUS: string (nullable = true)

|-- NAME\_HOUSING\_TYPE: string (nullable = true)

|-- DAYS\_BIRTH: integer (nullable = true)

|-- DAYS\_EMPLOYED: integer (nullable = true)

|-- FLAG\_MOBIL: integer (nullable = true)

|-- FLAG\_WORK\_PHONE: integer (nullable = true)

|-- FLAG\_PHONE: integer (nullable = true)

|-- FLAG\_EMAIL: integer (nullable = true)

|-- OCCUPATION\_TYPE: string (nullable = true)

|-- CNT\_FAM\_MEMBERS: double (nullable = true)

There are no missing and duplicate values in the dataset, but there is 134203 null vales in OCCUPATION\_TYPE column. This is imputed with “NOT MENTIONED” category. There is significant amount of outliers in the dataset.

Two variables, DAYS\_BIRTH & DAYS\_EMPLOYED were transformed into new variables, since they were given in negative values. They were dropped afterwards,

The following are the important insights drawn from the univariate analysis of the applicant dataset:

1. Around 64% of the credit card applicants are females.
2. 60% of applicants doesn’t own a car and 68% of applicants have house ownership
3. Majority of the applicants have no children or have no minor dependents. This is implied by the fact that 50% of the applicants have only 2 family members.
4. Most of the applicants work for their source of income and 68% have Secondary /special secondary education.
5. 69% of the applicants are married and 90% of the applicants have own house.
6. All the applicants possess mobile number
7. About 75% of application have no email, work phone or land line facility.
8. 23% of the applicants are laborers
9. Most applicants are between the ages of 27 to 44. About 11% of the applicants have only 1 year of work experience.

There are 1048575 rows in the credits records dataset and 3 columns. There are no null, duplicate and missing values in this dataset.

The columns in this dataset are:

|-- ID: integer (nullable = true)

|-- MONTHS\_BALANCE: integer (nullable = true)

|-- STATUS: string (nullable = true)

The following are the important insights drawn from the univariate analysis of the applicant dataset:

1. About 43% of the total credit card owners have cleared their dues.

According to the problem statement, the customers who have delayed payment for more than 60 days are considered delinquent or bad customers. This classification or the target variable of this dataset is achieved by this applying this condition on the STATUS variable. It is found that the target variable is extremely imbalanced. This can be seen by the biased insights of the bivariate analysis of the credits dataset.

Bivariate Analysis gave the following insights:

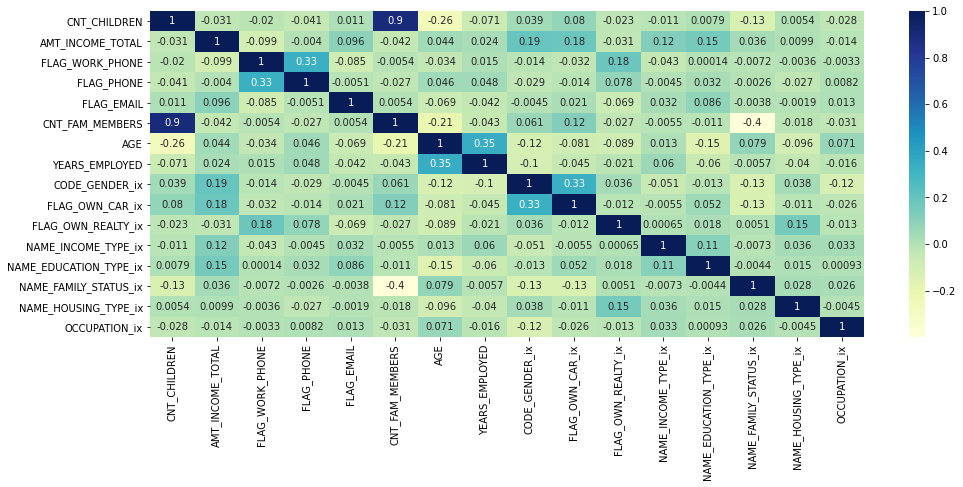
1. Around 65% of female applicants have house ownership compared to the mere 35% in the case of males.
2. As expected salary increases with better educational levels. Correlation value between Educational level and Income is 0.9922547141689341. This is indicated by the heavy positive correlation between both.
3. Average Salary of applicant base = 14855.269960126381

Median of salary of applicant base = 13875.0

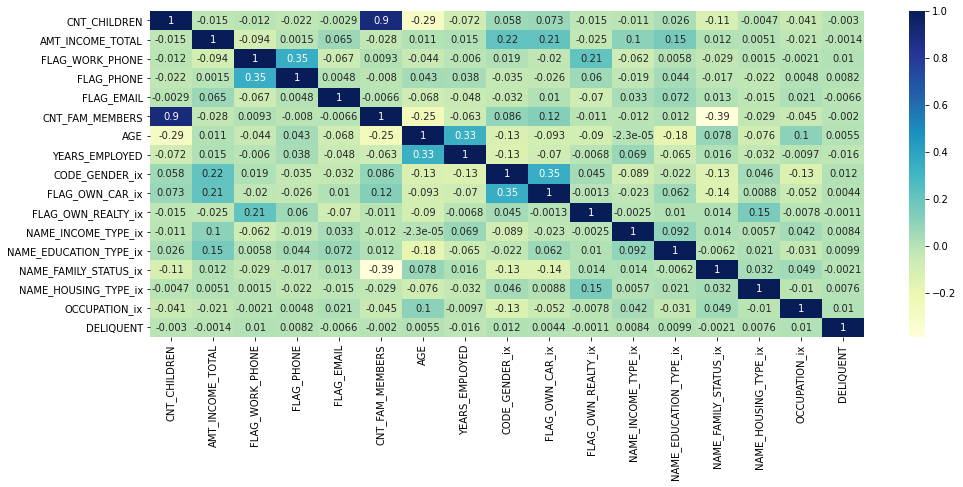
1. Most of the applicants who are married have no children OR are older couples (children who have a family of their own), which in turn implies the insight from univariate analysis of CNT\_FAM\_MEMBERS variable, that majority of applicants have only 2 family members.
2. Managers are paid highest and cleaning staff the least
3. The proportion of bad customers who owns a car is only 0.37%. This is very low compared to the proportion of good customers viz; 99%
4. The proportion of bad customers who stays in a rented apartment is only 0.46%. This is very low compared to the proportion of good customers in staying in rented apartments viz; 99%
5. The proportion of bad customers who are single or unmarried is only 0.44%. This is very low compared to the proportion of good customers who are single or unmarried viz; 99%

Multivariate Analysis:

Correlation matrix of application dataset



Merged data:



There is little to less correlation between variables. i.e. there is no collinerality among them.

Model Development:

Model is developed on the inner joined data of applications and credit card owners. Since most of the variables are categorical, the numerical variables were also binned. They were binned to transform into WoE (weight of Evidence) categorization- a variable transformation method. This will take of any missing, null, duplicate values and also will take of the outliers. No extra one hot encoding is required. The IV (Information Value) values will determine the strong and weak predictors, thus helping in better feature selection for an optimized model.

A custom generate class was used to determine woe and IV values. The woe fitted data frame is then split into train and test sets in the ration 70:30 with a seed value of 2018.

Since this is a binary classification problem- Logistic Regression algorithm is considered for training the model. After initial training the model is evaluated on the test set and the metrics were calculated.

Recall : [1.0, 0.0]

Precision : [0.9842950571695511, 0.0]

Accuracy : 0.9842950571695511

Area undre ROC : 0.6373329464675793

The metrics clearly indicate the bias in the classification. Since he number of good customers is almost 99% the model is biased towards the customers not being delinquent.

The model is retrained with features selected via the IV values. According to the problem statement a threshold of 0.002 is set for IV values and those below this value is considered to be weak predictors and are dropped.

Again the metrics were not that appealing and is as follows.

Recall : [1.0, 0.0]

Precision : [0.9842950571695511, 0.0]

Accuracy : 0.9842950571695511

Area undre ROC : 0.6234276749692563

The dataset is under sampled and is again trained and tested. This time the metrics were promising and it is as follows.

Recall : [0.5393258426966292, 0.5661971830985916]

Precision : [0.5549132947976878, 0.5506849315068493]

Accuracy : 0.5527426160337553

Area undre ROC : 0.5970367146700427