Credit EDA Case Study

GOAL : THE COMPANY WANTS TO UNDERSTAND THE DRIVING FACTORS (OR DRIVER VARIABLES) BEHIND LOAN DEFAULT, I.E. THE VARIABLES WHICH ARE STRONG INDICATORS OF DEFAULT AND UTILISE THIS KNOWLEDGE FOR ITS PORTFOLIO AND RISK ASSESSMENT.

Submitted By:
1. Vaishali Papneja
2. Sushasree Vasudevan Suseel Kumar

Business Understanding

When the company receives a loan application, the company has to decide for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:

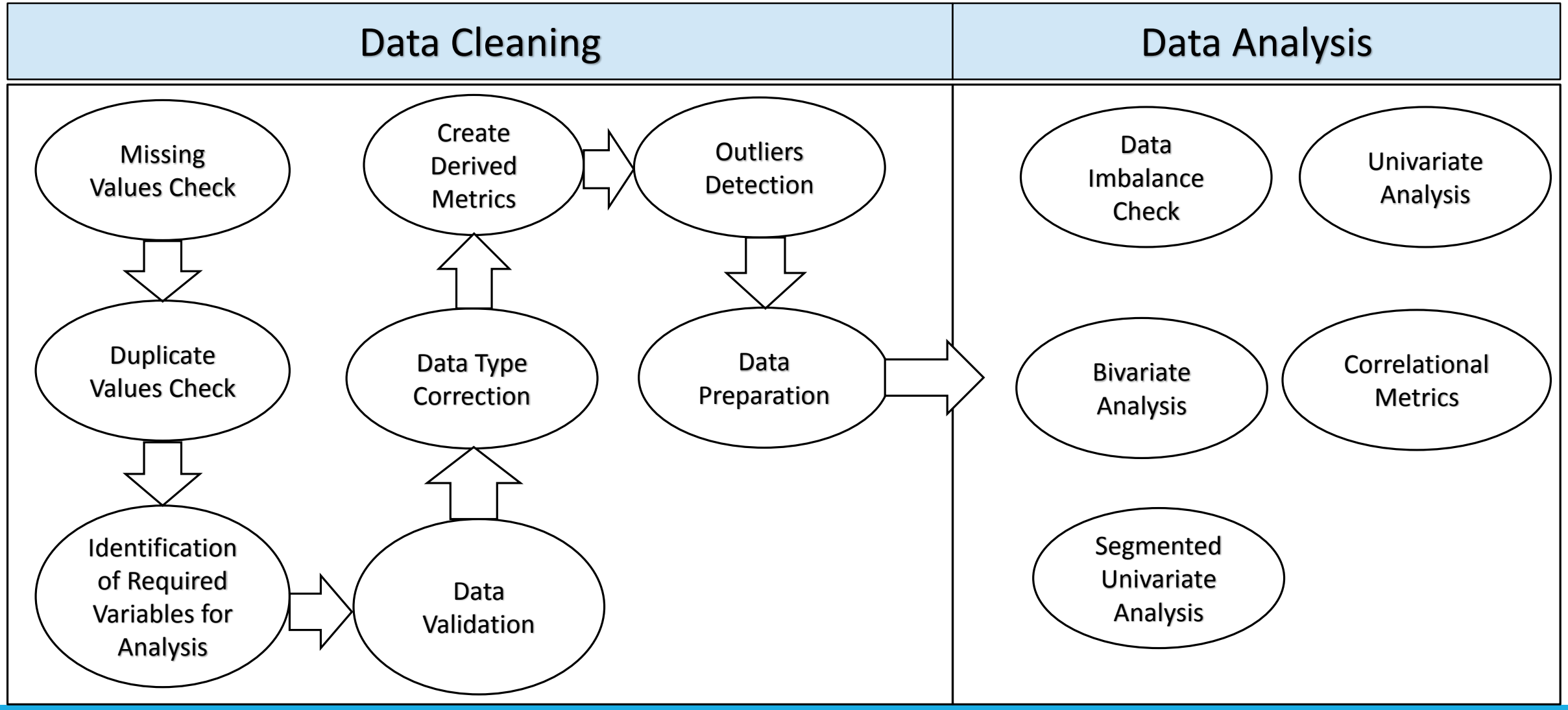
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Steps Involved in Analysis

- Use EDA to analyze the patterns present in the data
- The patterns should indicate if a client has difficulty paying their installments which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- **Approach Used in Analysis**
  - We are given two datasets,
    - The first data is about whether a **client has payment difficulties.**
    - The second data is about whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
  - Approach used in Analysis for few columns : To find the probability of default/rejection
    - In Application dataset, Find the probability of default to identify the defaulters
      - In our case,
        - Probability of defaulter = defaulter/(defaulter + non defaulter)
        - Probability of nondefaulter = nondefaulter/(defaulter + non defaulter)
    - In Previous Application dataset, Find the probability of rejection to identify the rejected applicants
      - We can also calculate the chances of an application getting rejected
      - In our case,
        - Probability of rejected = rejected/(rejected + approved)

  - ****We have done our analysis on both the datasets separately**.**

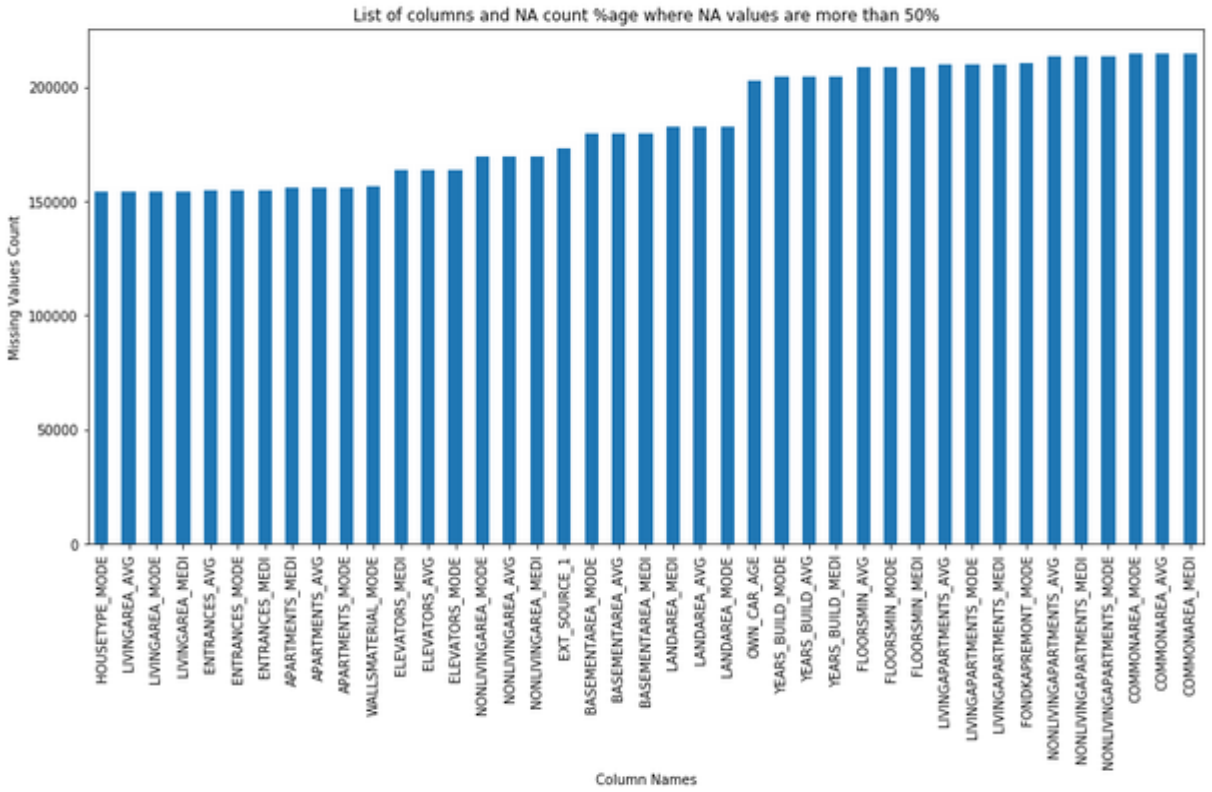
Problem Solving Methodology



Data Cleaning

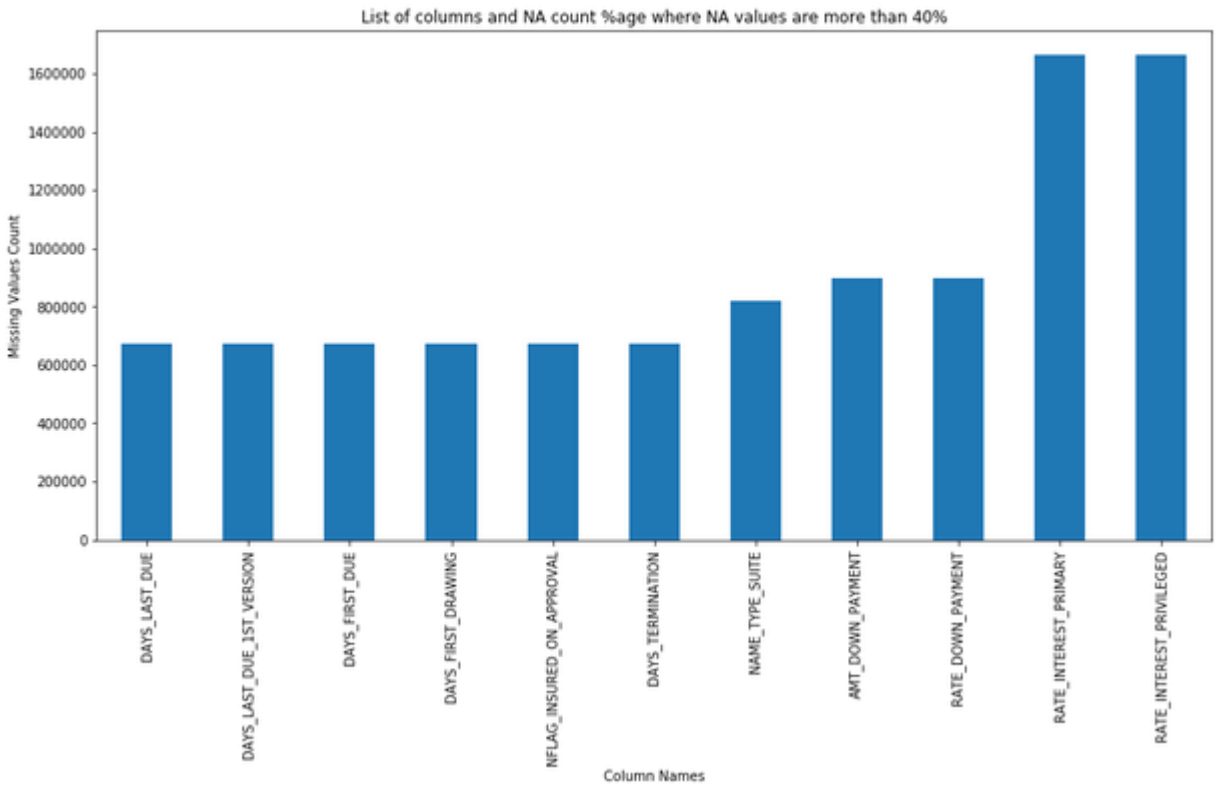
Missing Data Proportion in the Datasets

Application Dataset



There are 41 columns in application dataset which have more than 50% missing values.

Previous Application Dataset



There are 11 columns in previous application dataset which have more than 40% missing values.

Methods to Deal with Missing Data

There are various ways to deal with missing values. Out of which, most common methods are as below:

- Remove those columns if we have higher proportion of missing data
- Replace them with
  - Mean/Median/Mode in case of quantitative variables.
    - Replace them with mean if data in that filed is distributed normally.
    - Replace them with median if there are outliers present in that particular field.
    - Replace with mode if replacing with most repeated value of filed makes sense.
  - Most repeated value in case of categorical variables.
- Replace with a default value
- Leave as it is.

In our datasets, We have replaced missing values of categorical variables with most repeated value. And in case of quantitative variable, missing values are filled with Medians of that column as there are outliers present in the columns..

Outliers Detection and Ways to Deal with them

- An outlier is a data point that differ significantly from other data points.
- It may be due to variability in the measurement or may also indicate an experimental error.
- One should be very cautious while dealing with outliers.

For example –

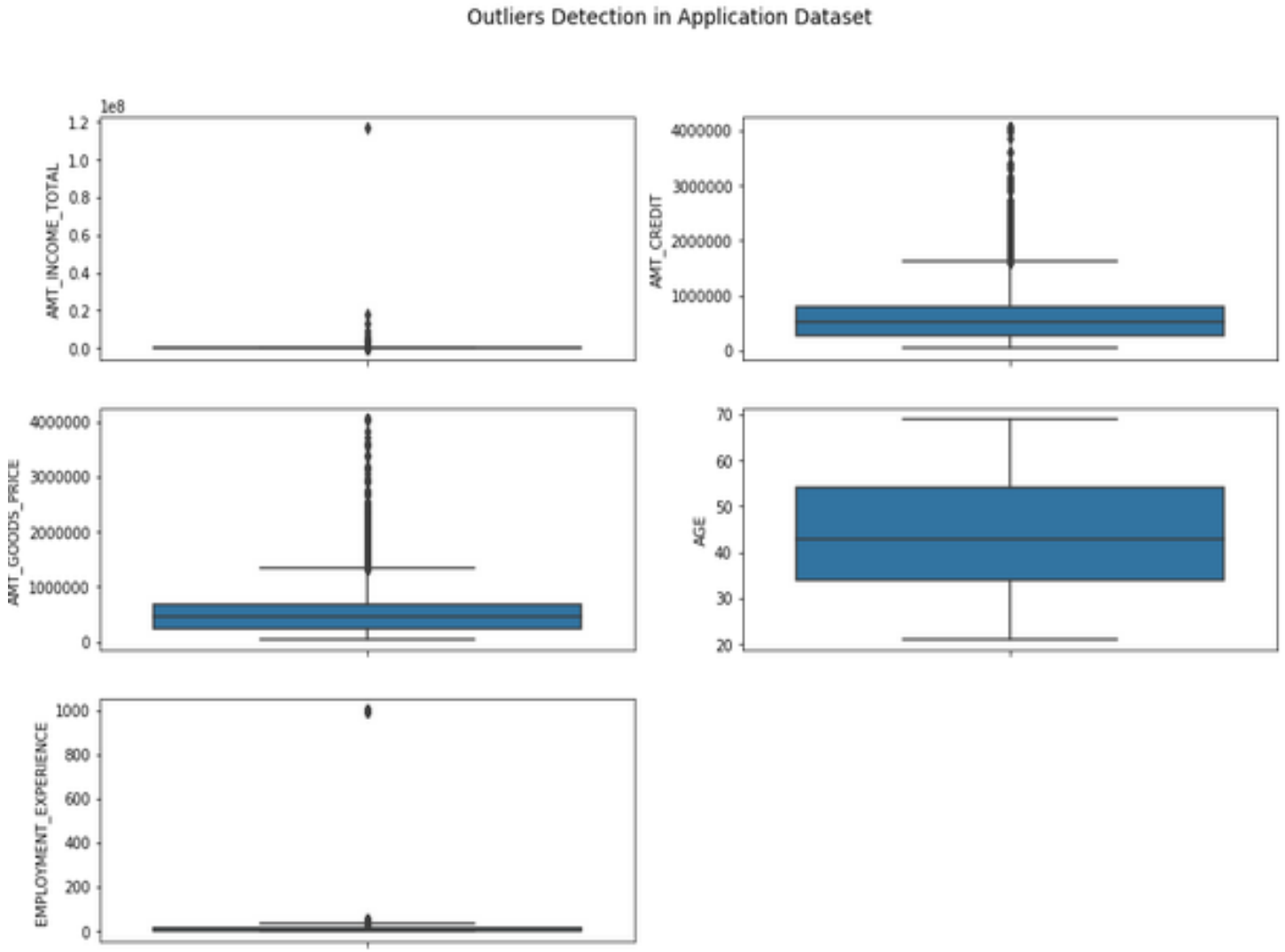
- Outliers introduced due to experimental error can directly be discarded.
- But in the other case, where it has occurred due to variability of measurement, one should take steps to fulfill the purpose.
- Sometimes it is better to remove highly skewed outlier and keep rest of the outliers to get a better insight. While in some other cases, it is better to remove all the outlier data points.

Outliers Detection in Application Dataset

In application dataset, we looked for datasets in 5 numeric columns:

- AMT_INCOME_TOTAL
- AMT_CREDIT
- AMT_GOODS_PRICE
- AGE (which is derived metric based on DAYS_BIRTH)
- EMPLOYMENT_EXPERIENCE (which is again a derived metric calculated from DAYS_EMPLOYED)

- By looking at the box plot, it is clear that there are outliers in the dataset for these columns(other than AGE) as some of the data points are located outside the fences.
- Also, by looking closely at the dataset, we found that there are some infinite values present in the dataset for EXPERIENCE field(which is an experimental error), so we can directly discard them.
- For the other columns (AMT_INCOME_TOTAL, AMT_CREDIT and AMT_GOODS_PRICE), we removed top 1% of the data points as they were highly different from other data points.

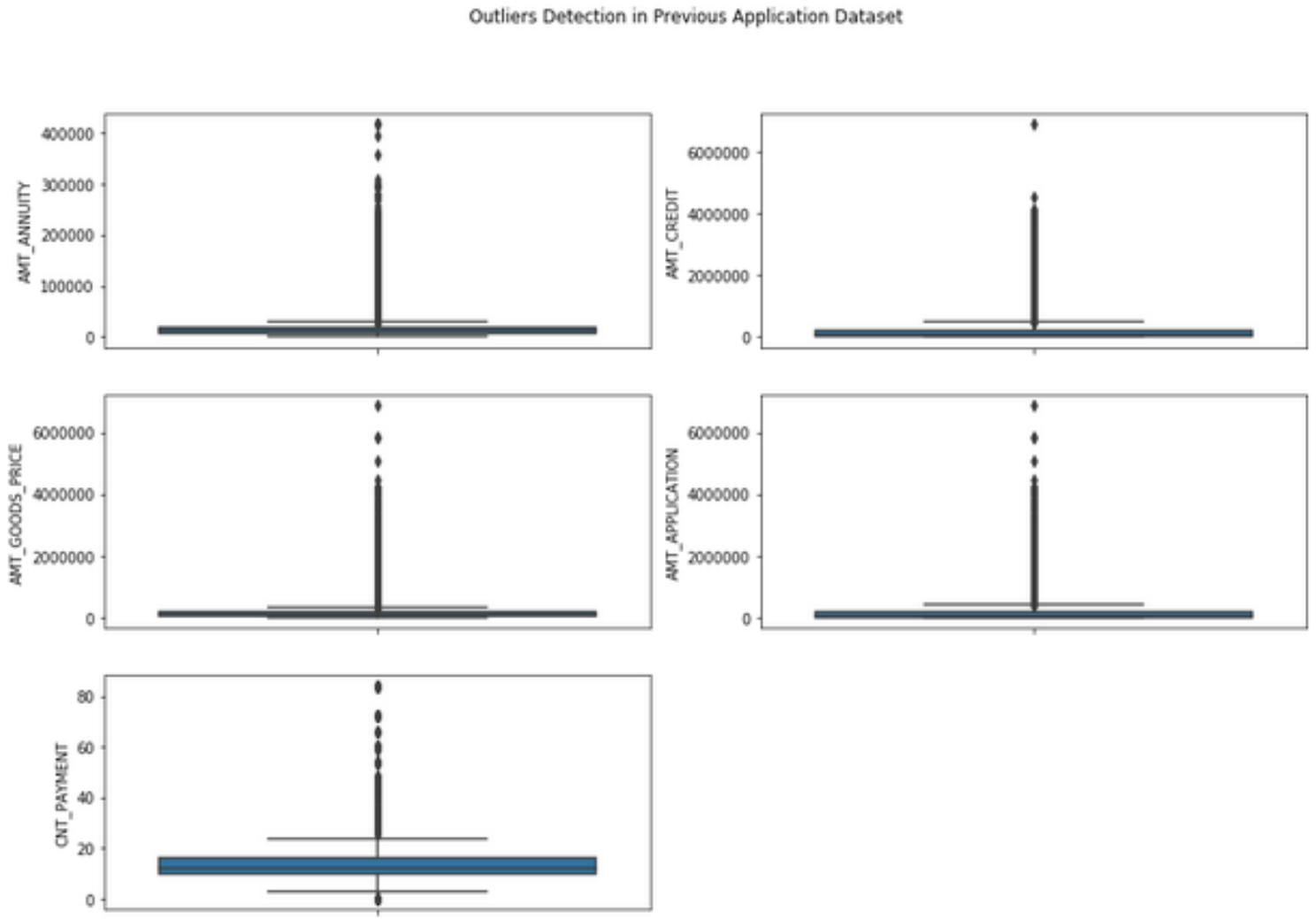


Outliers Detection in Previous Application Dataset

In previous dataset, we looked for datasets in 5 numeric columns:

- AMT_ANNUITY
- AMT_CREDIT
- AMT_GOODS_PRICE
- AMT_APPLICATION
- CNT_PAYMENT

- By looking at the box plot, it is clear that there are outliers in the dataset for these columns as some of the data points are located outside the fences.
- Here also, we removed top 1% of the data points as they were highly different from other data points.



Assumptions used in both Datasets

- Few of these columns have the XNA and XAP values
  - XNA = not available
  - XAP = not applicable (X as a logical not !?)
  - Removed XNA and XAP for better visualization in the plots in few columns

Data Analysis

BASED ON CONSUMER ATTRIBUTES

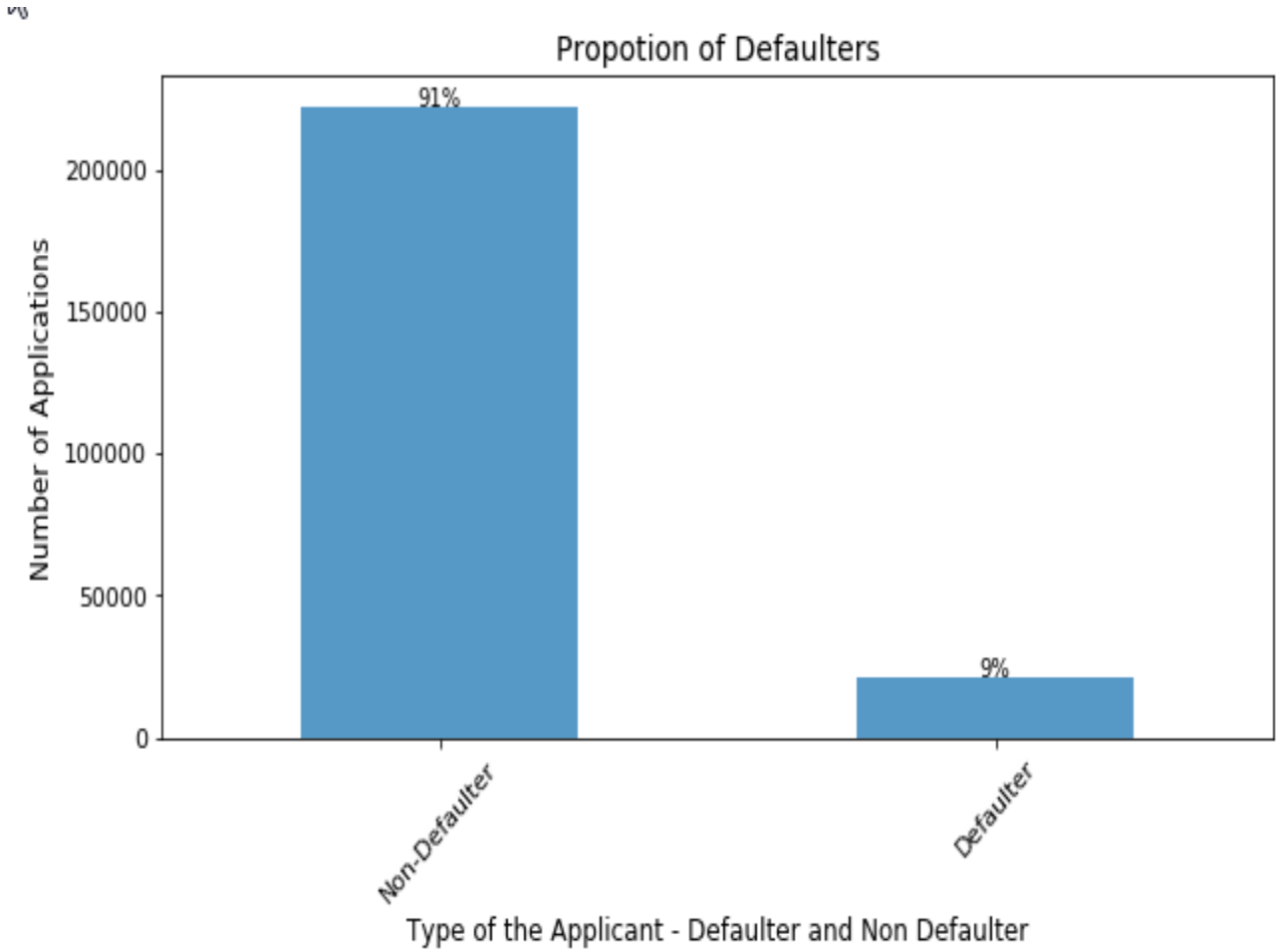
Data Imbalance Calculation

This graph implies that we have less number of applicants with payment difficulties.

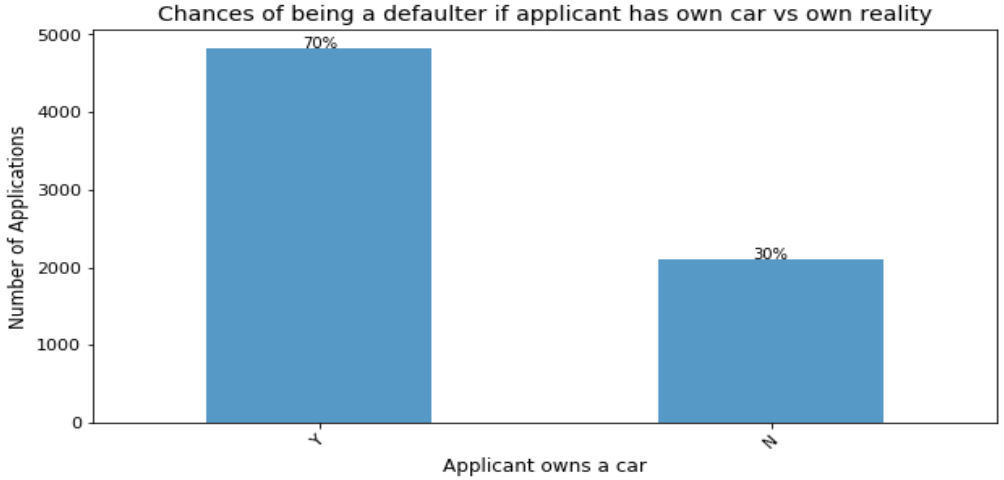
Also,

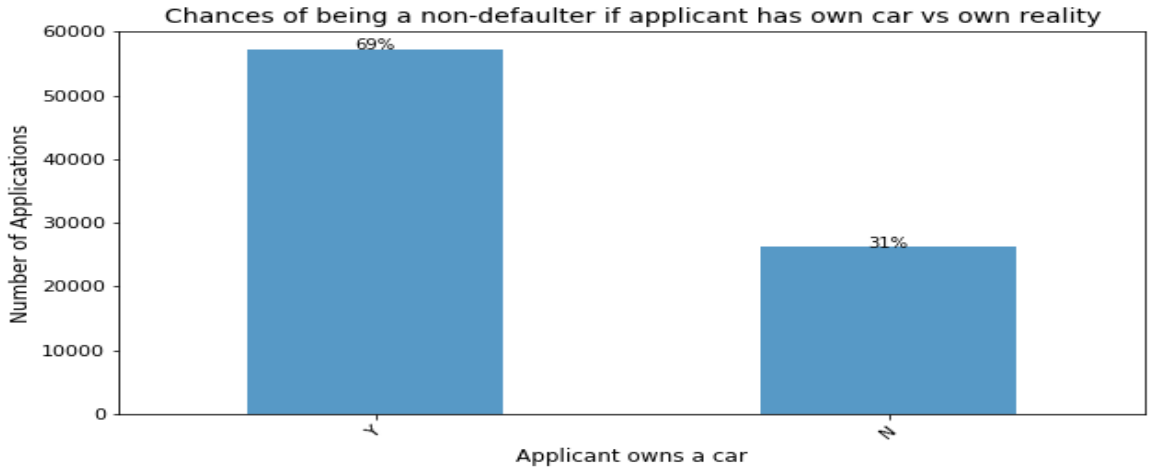
Percentage of Defaulter applicants in our dataset: 8.81 Percentage of Non-Defaulter applicants in our dataset: 91.19

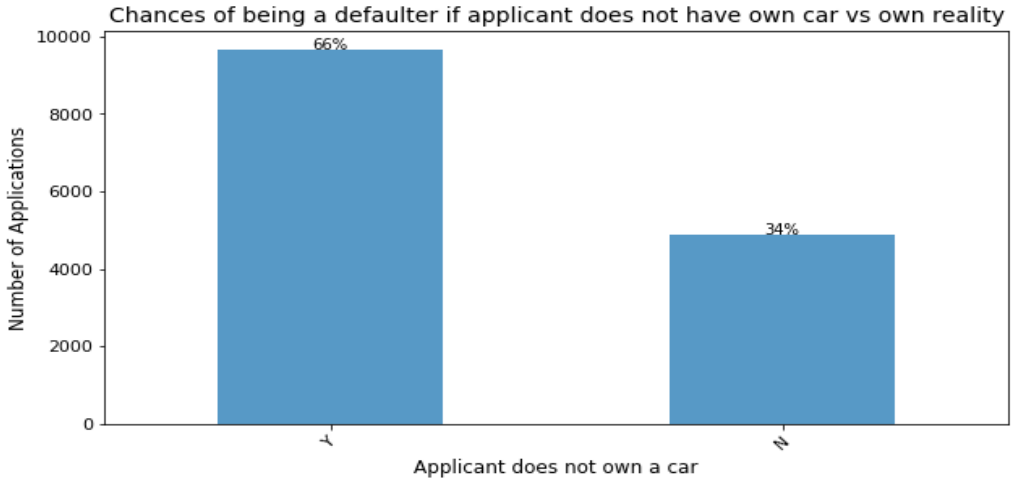
Data Imbalance Percentage in our dataset is: 9.66

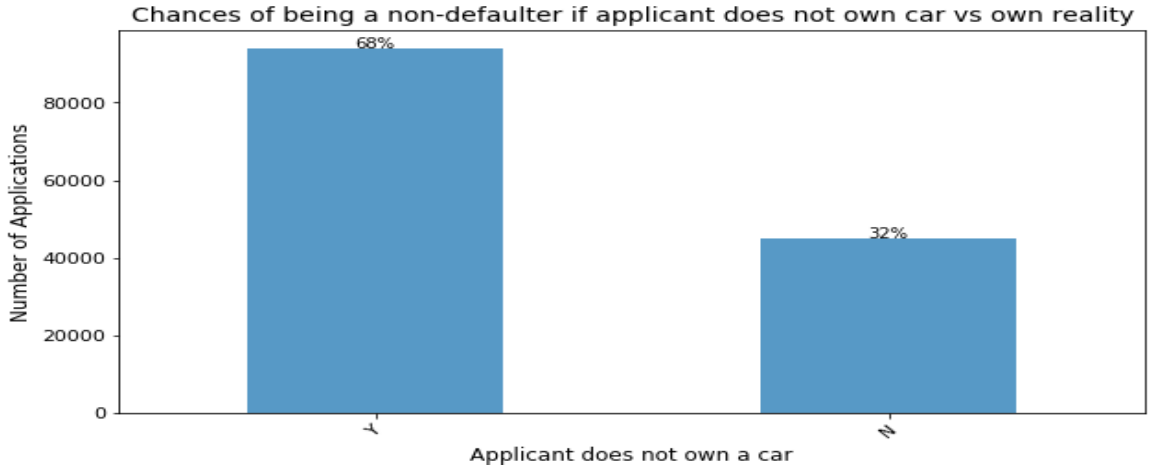


Analysis – Based on Own Car and Own Realty





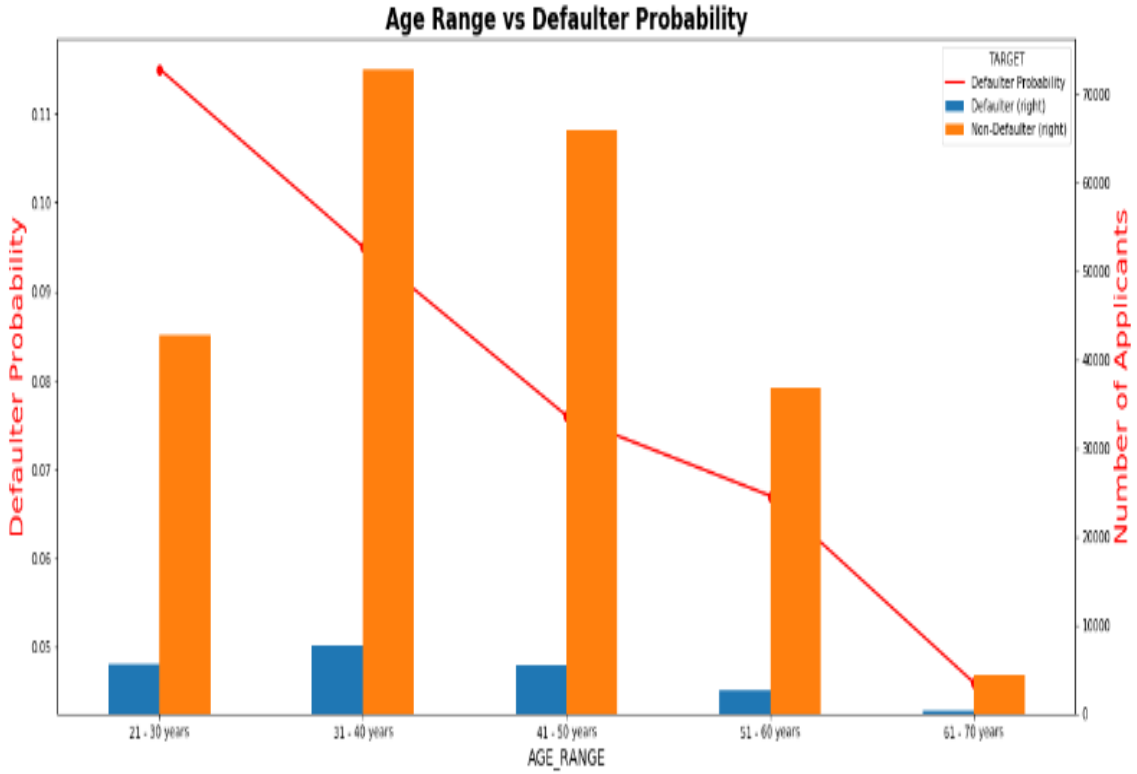


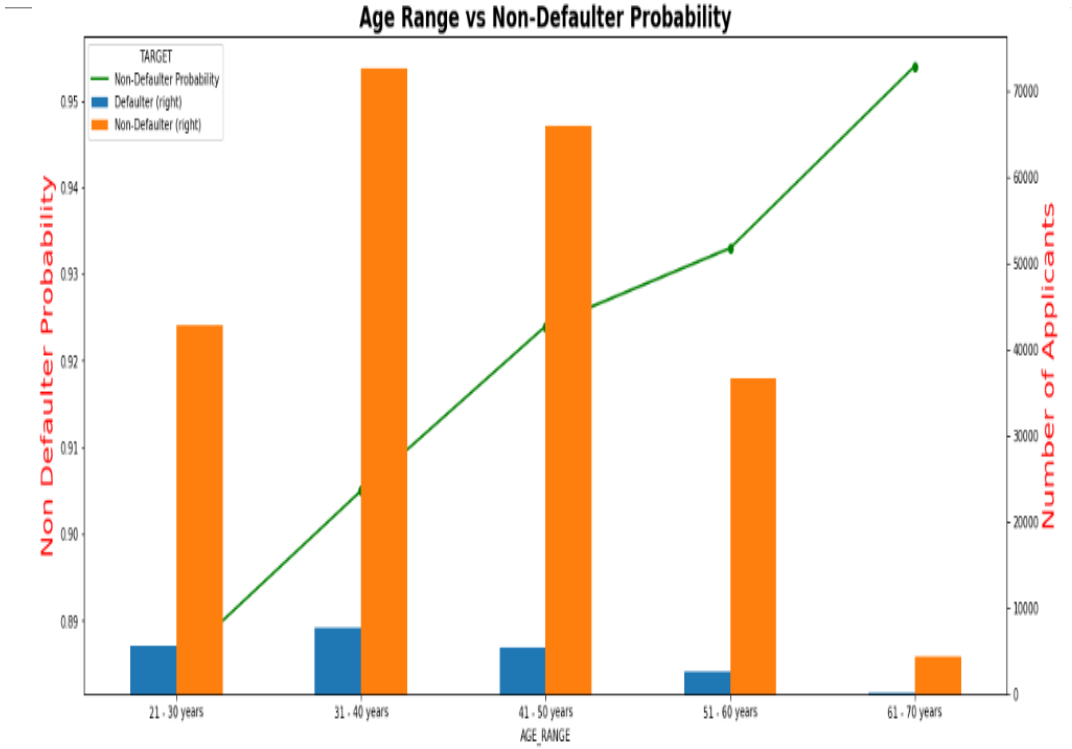


Insight:

- Minimum chances of being a defaulter - if applicant has own car but no realty
- Maximum chances of being a defaulter - if applicant has no own car but he has own realty
- Proportion of being a non-defaulter is more if applicant has no own car but he has own realty

Analysis - Based on Age Range

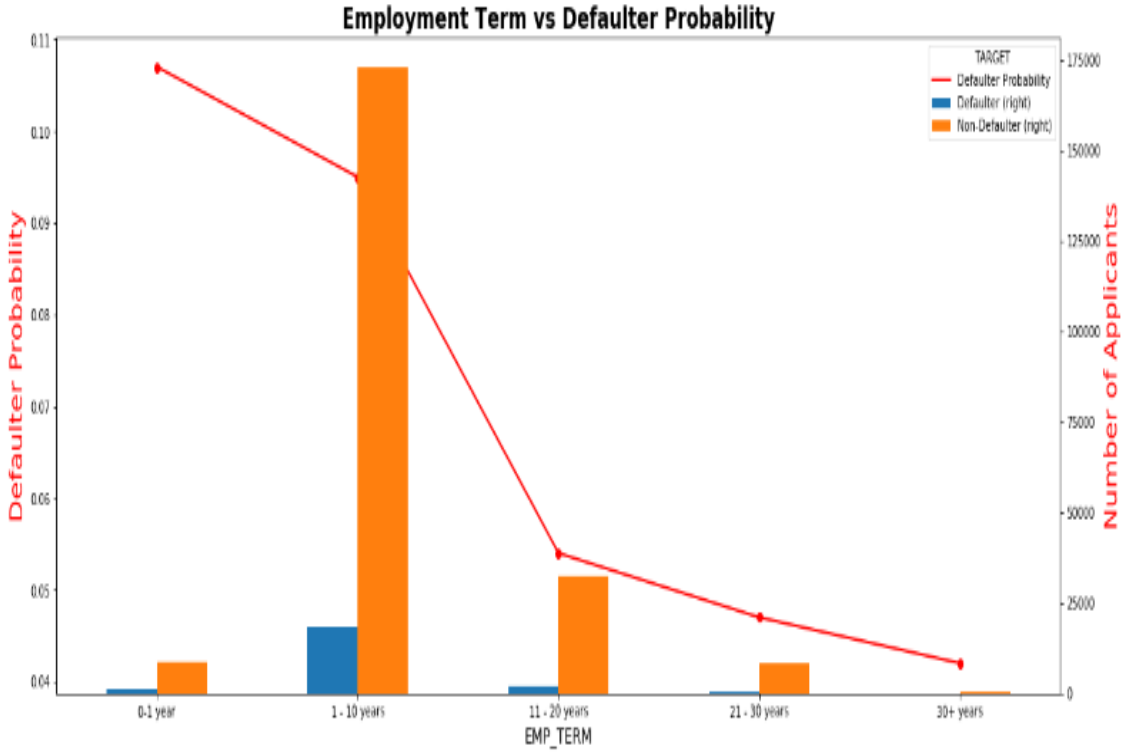


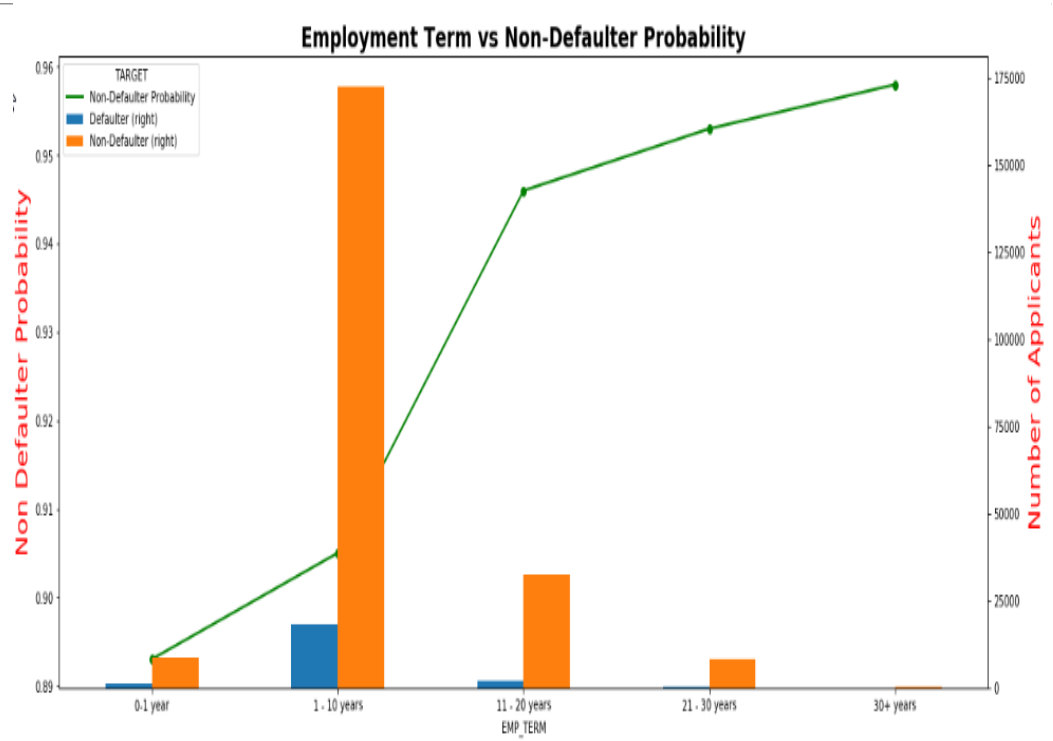


From the above plots, we can infer that,

- We have highest default probability in the age group of 21 to 30 years
- As the age increases, the chances of an applicant being an defaulter reduces.

Analysis – Based on Employment Term

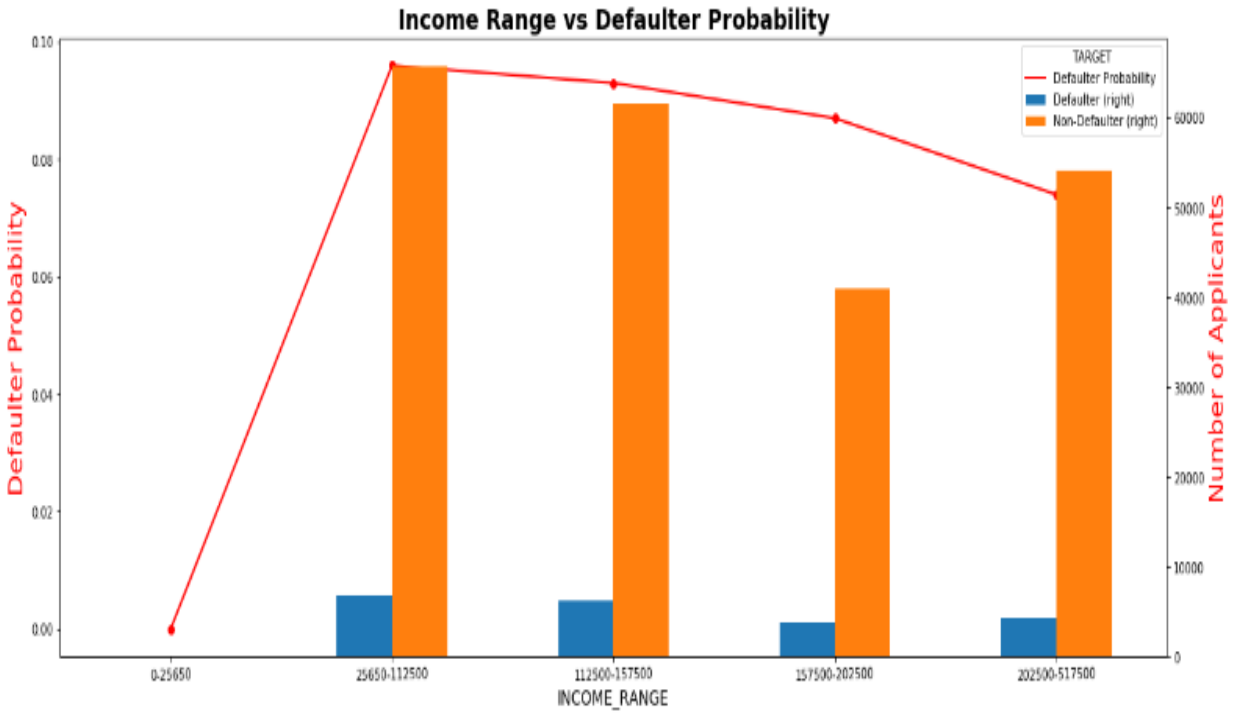


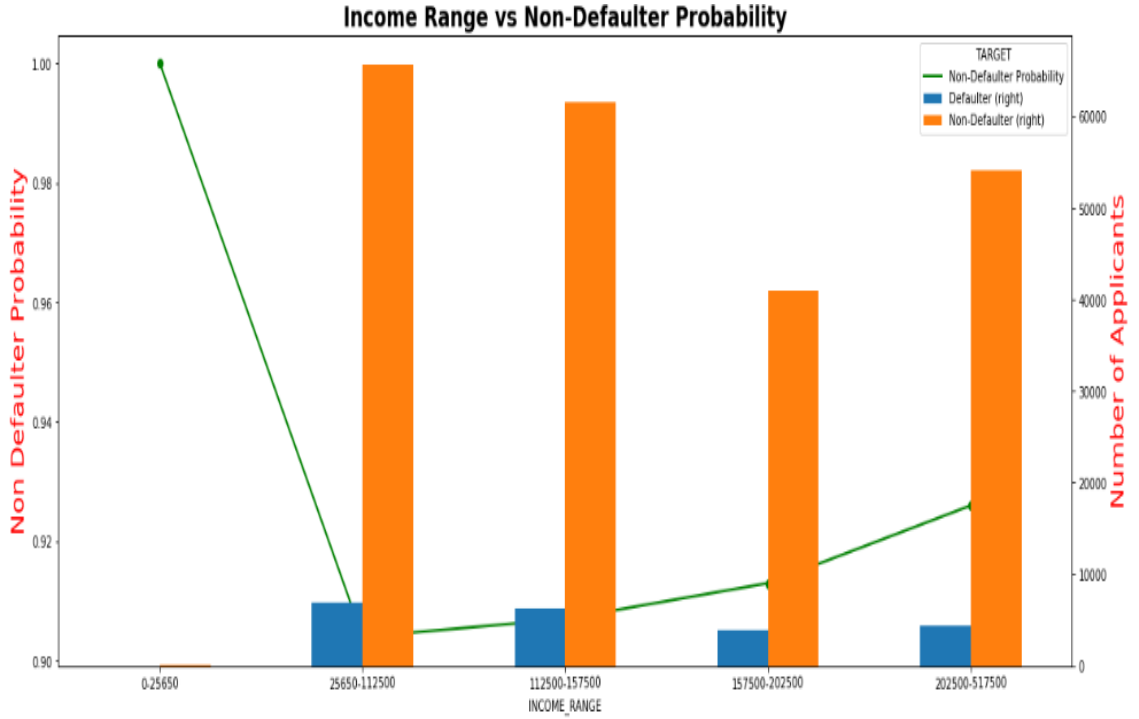


From this plot, we can infer that,

- We have more defaulters when the employment term is less than 10 years.
- As the employment term increases, the chances of an applicant being a defaulter reduces.

Analysis – Based on Income Range

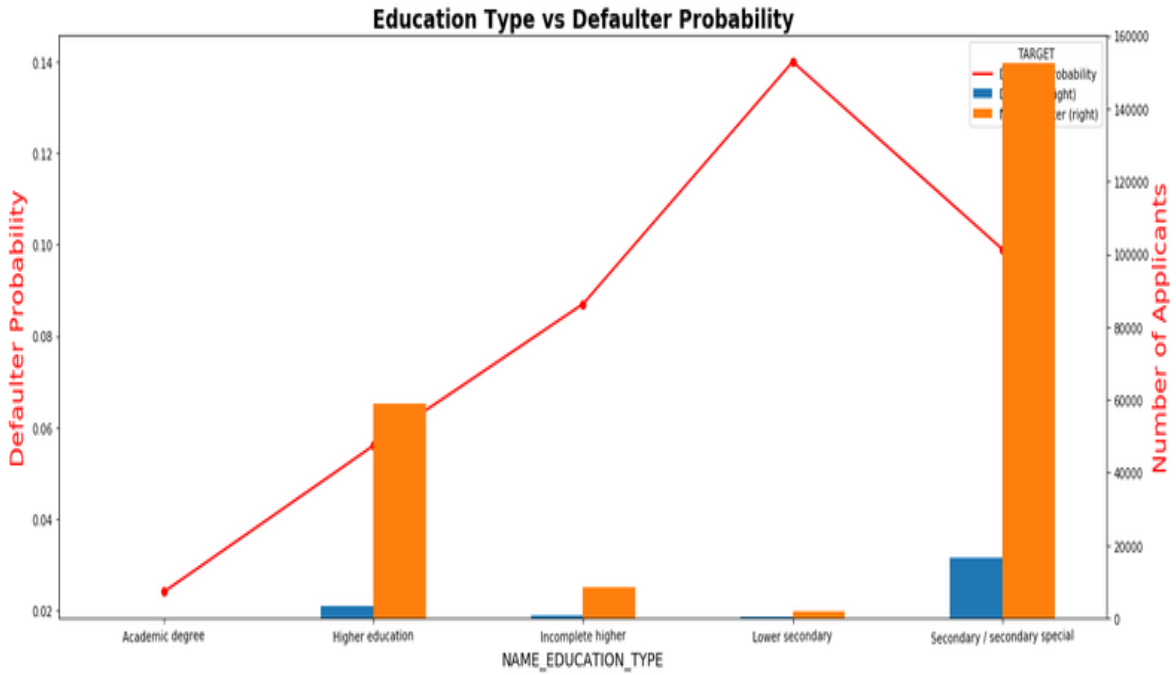


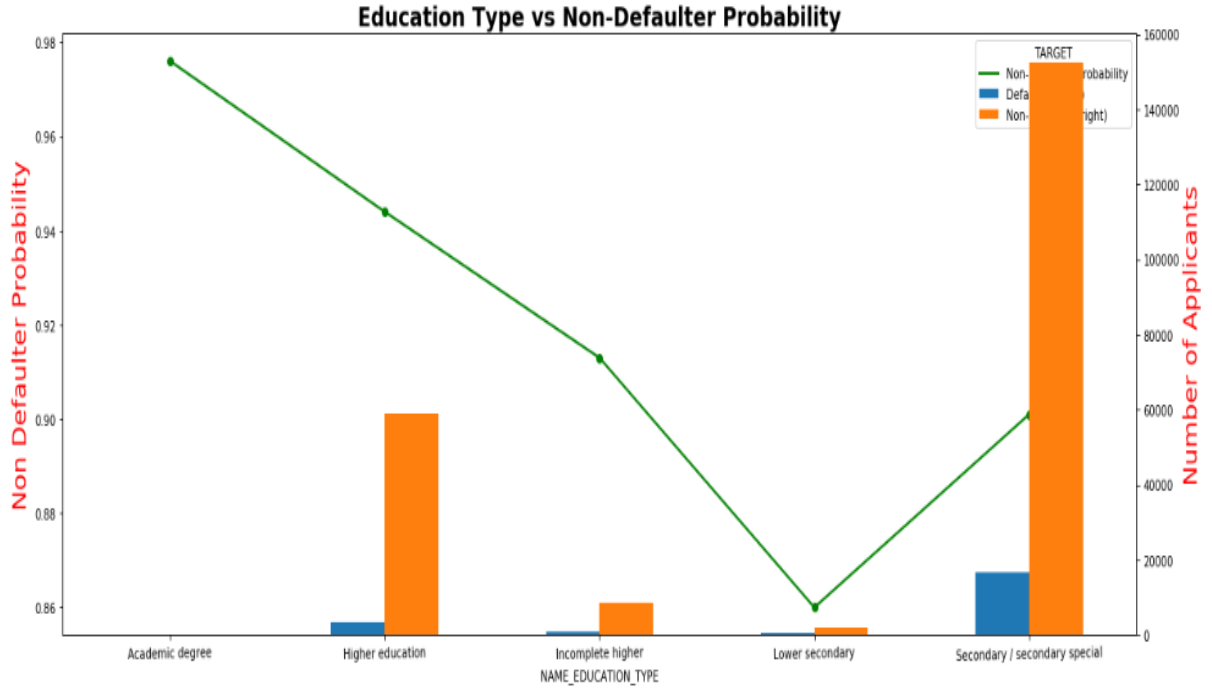


From the above plot, we can infer that,

1. we have highest default probability in the income range 25K to 1.125K.
2. As the Salary increases, the chances of an applicant being an defaulter also reduces.

Analysis – Based on Education Type

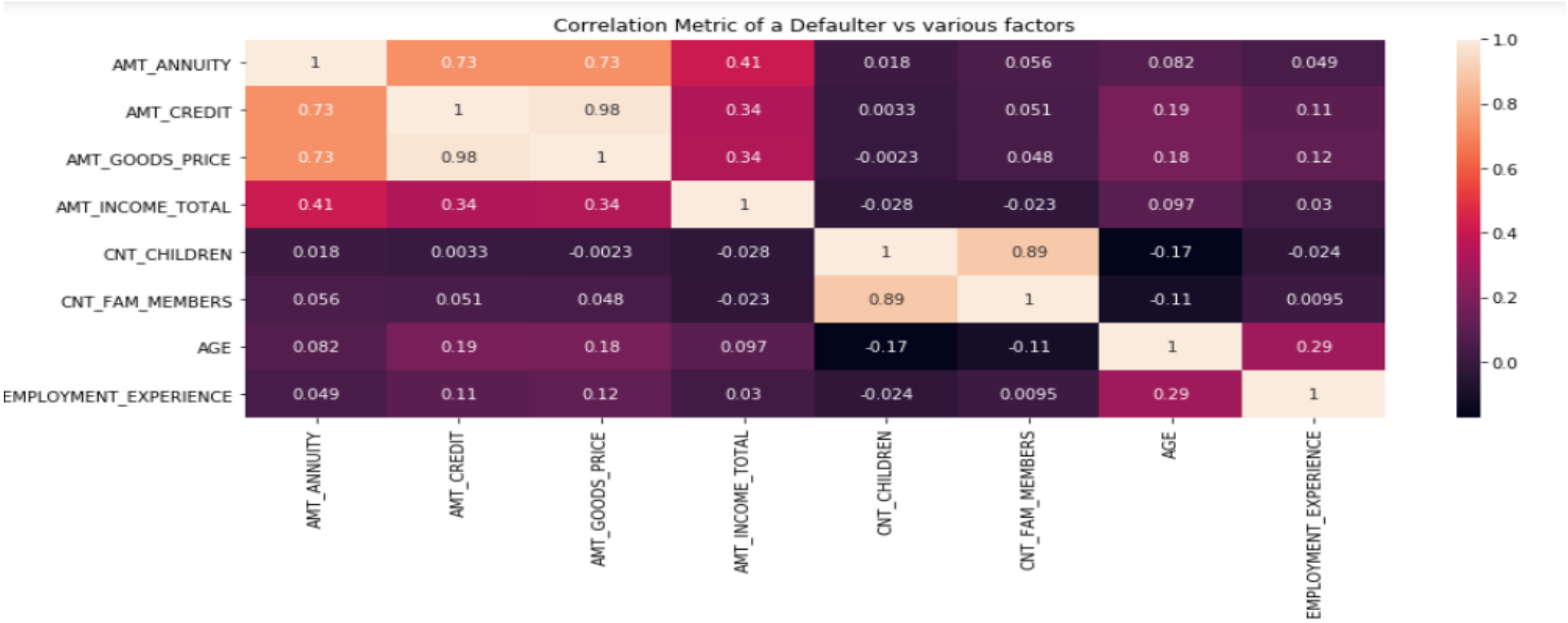




From the above plot, we can infer that,

1. The probability of being a defaulter is high when the Education type is Lower secondary or Secondary/Secondary Special.
2. The default probability is less when the level of education is more.

Correlation Metric: Factors Influencing a Defaulter's Behavior



Observations:

Following Factors influence an applicant behavior the most:

1. AMT_ANNUITY
2. AMT_CREDIT
3. AMT_GOODS_PRICE
4. CNT_CHILDREN
5. CNT_FAMILY MEMBERS

Below is the relation between above factors in case of a defaulter applicant:
Greater than 50% correlation: ['AMT_ANNUITY', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'CNT_CHILDREN','CNT_FAM_MEMBERS']
Greater than 75% correlation ['AMT_CREDIT', 'AMT_GOODS_PRICE', 'CNT_CHILDREN', 'CNT_FAM_MEMBERS']
Greater than 95% correlation ['AMT_CREDIT', 'AMT_GOODS_PRICE']

It is clear from the HEAT map that

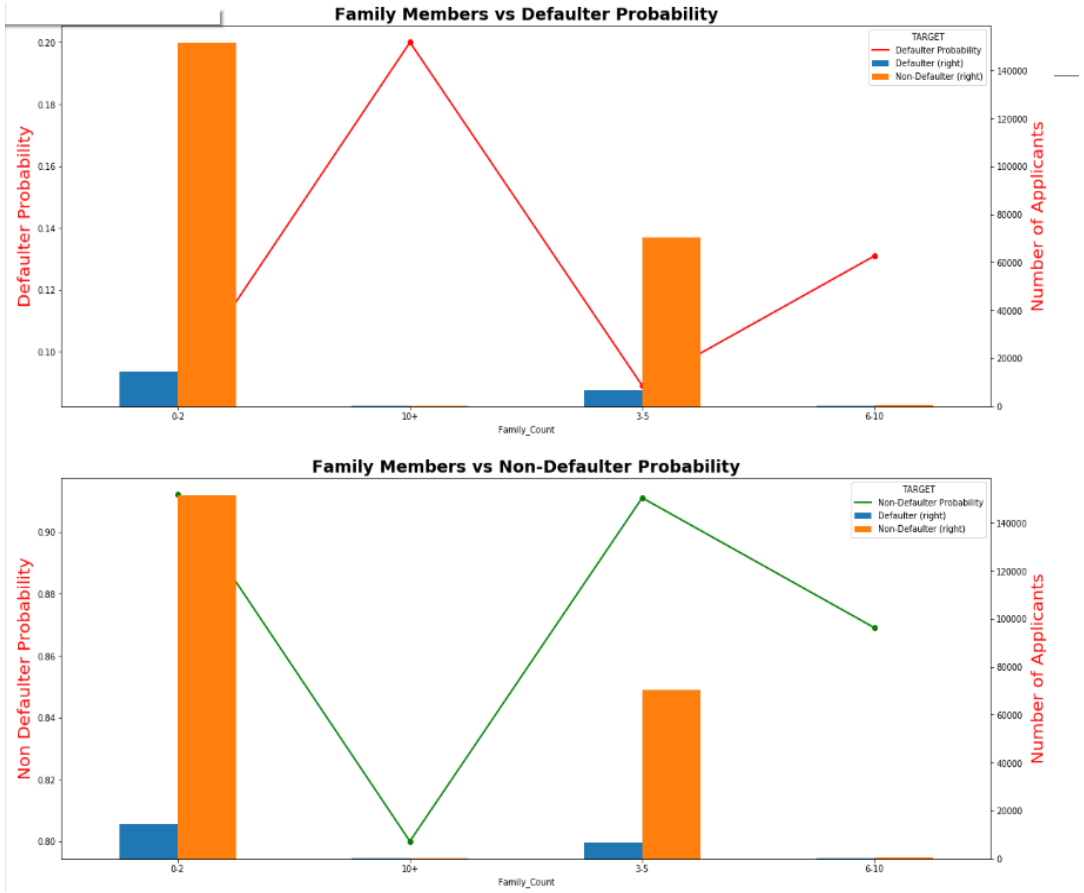
- The loan amount credited, annuity amount is highly correlated with the goods price.
- The count of children is highly correlated with count of family members
- The days of birth is also correlated with days employed.

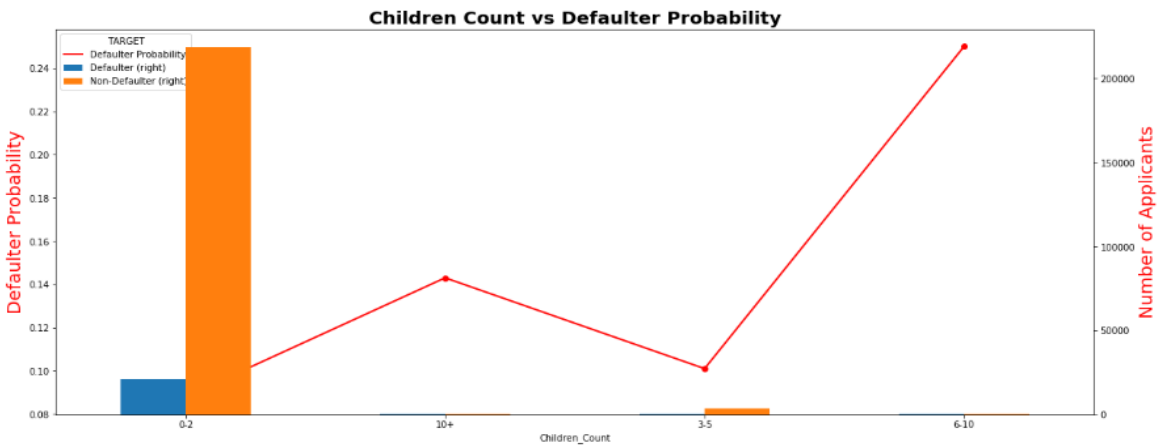
****As a result, we can use one of these columns out of each for analyzing our data and providing the expected result****

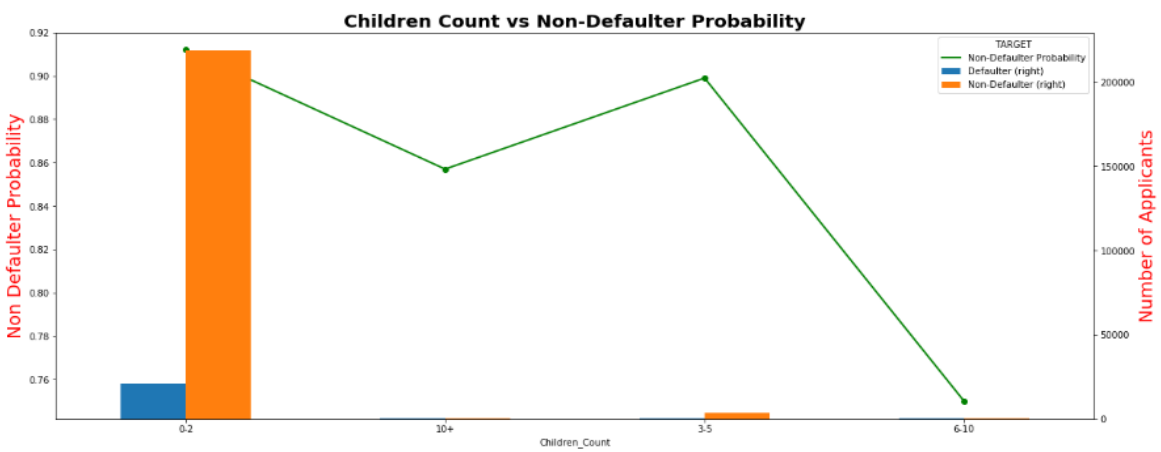
Top 10 Correlational variables w.r.t Target

- OBS_60_CNT_SOCIAL_CIRCLE
- OBS_30_CNT_SOCIAL_CIRCLE
- FLOORSMAX_MEDI
- FLOORSMAX_AVG
- YEARS_BEGINEXPLUATATION_AVG
- YEARS_BEGINEXPLUATATION_MEDI
- FLOORSMAX_MODE
- FLOORSMAX_MEDI
- AMT_CREDIT
- AMT_GOODS_PRICE

Analysis – Based on Family Members/Children Count



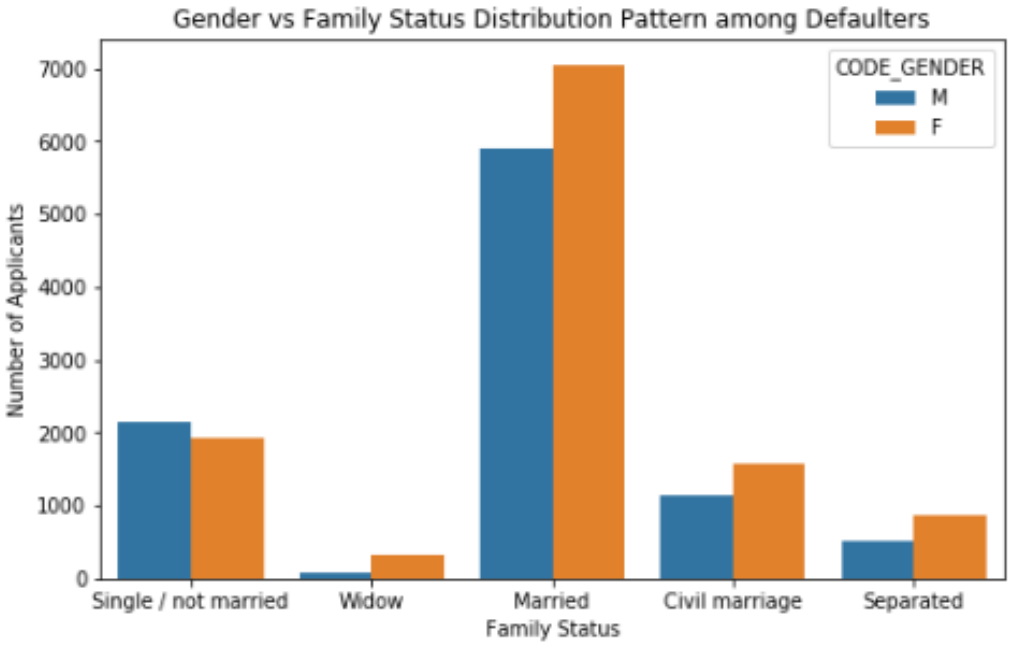


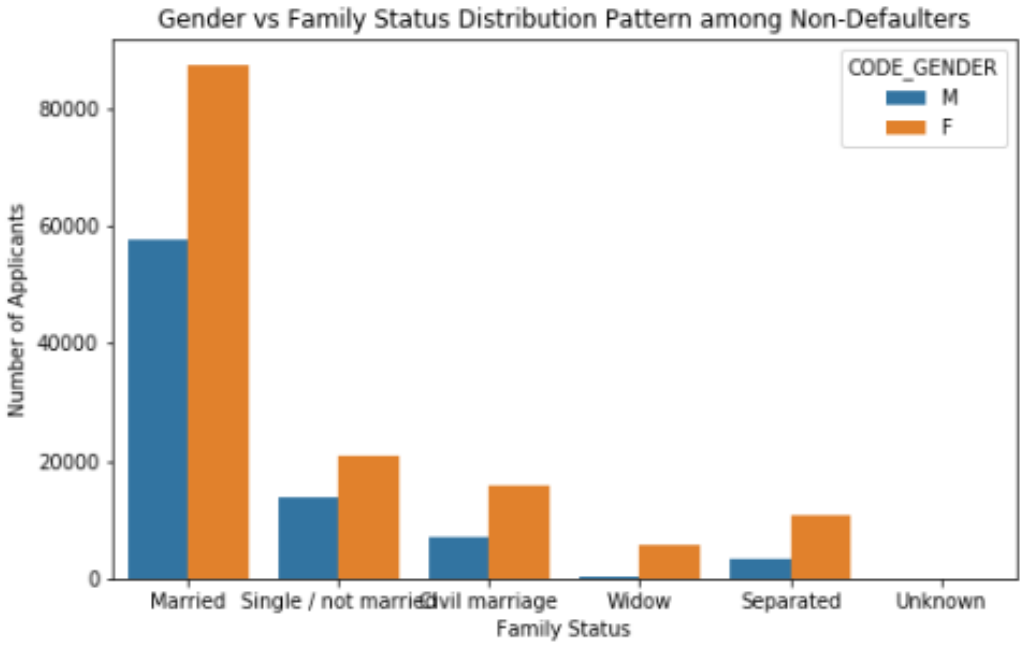


Insight

- Probability of default increases as the number of family members increases.
- Probability of non defaulters increases as the number of family members decreases.
- **We can also infer that Since, Family member count and Children count are highly correlated, we will get the same result after analysis**

Analysis – Based on Gender vs. Family Status

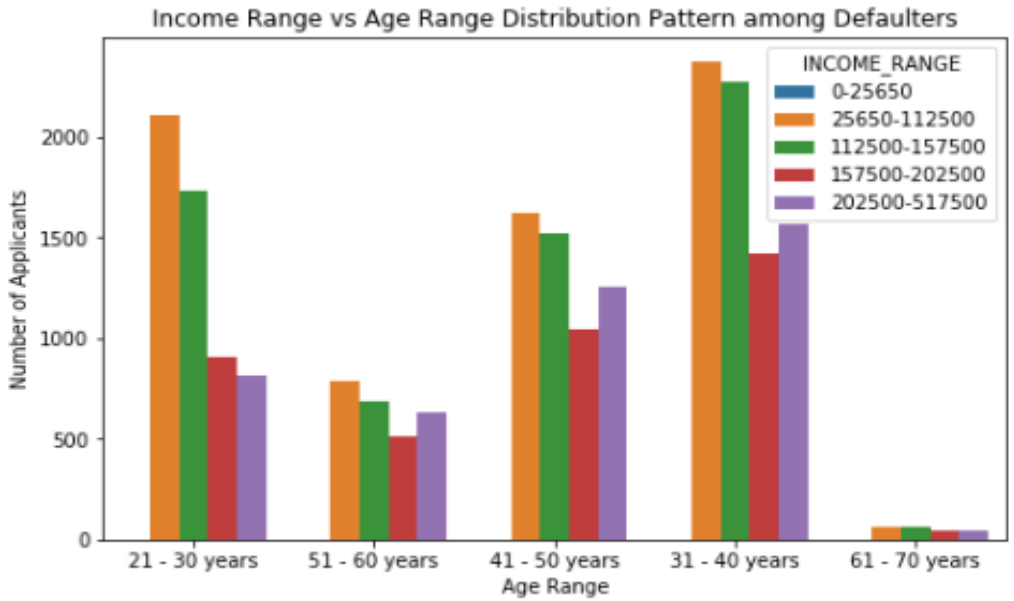


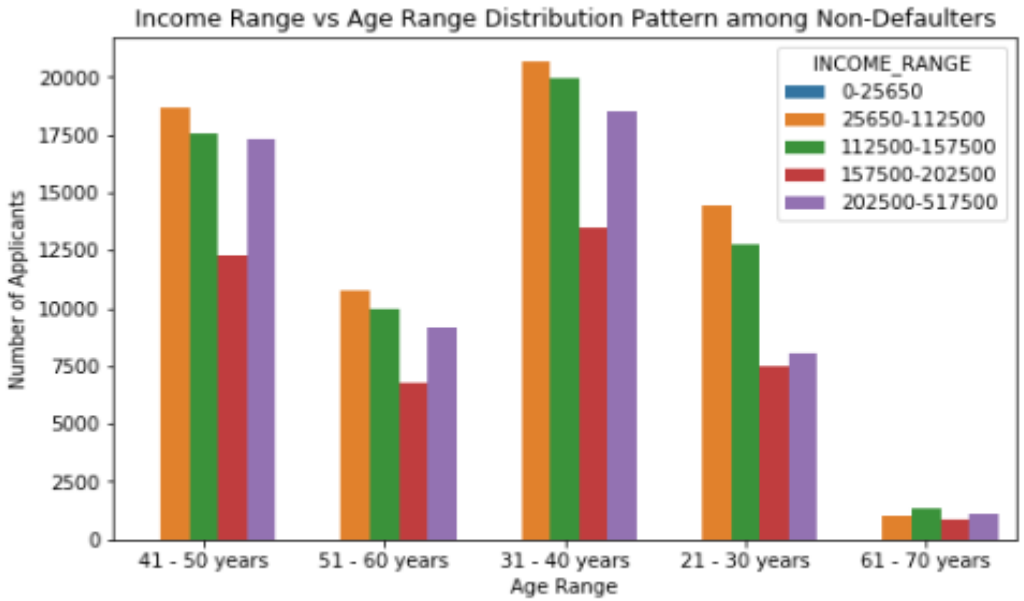


From the above plot, we can infer that,

- Single men tend to have payment difficulties in repaying the loan
- Also, The probability of married/Separated/Civil marriage women being a defaulter is more than married men.

Analysis – Income Range vs. Age Range

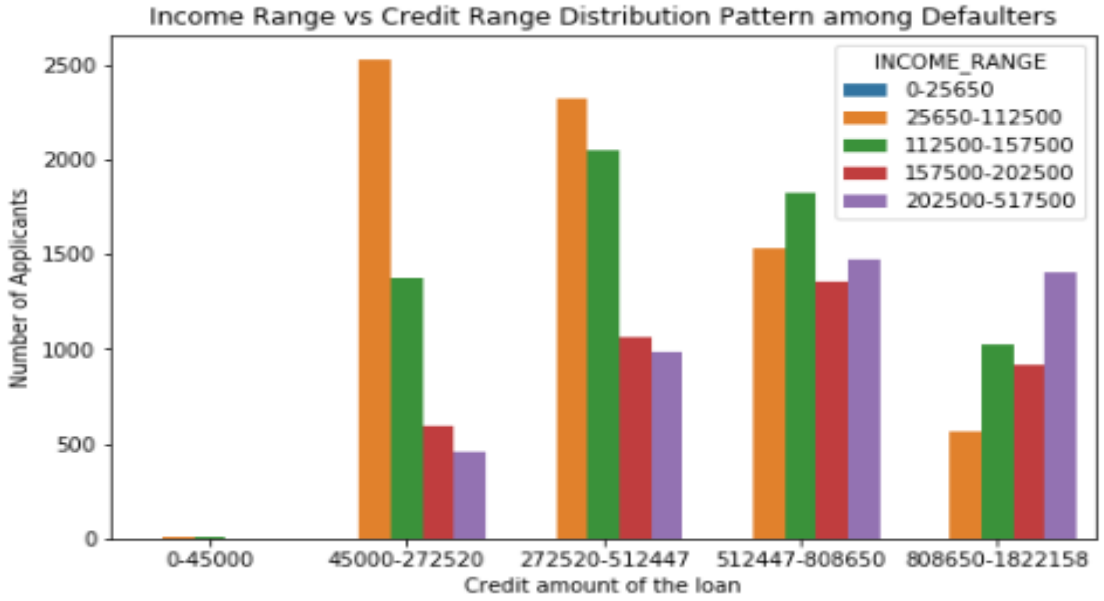


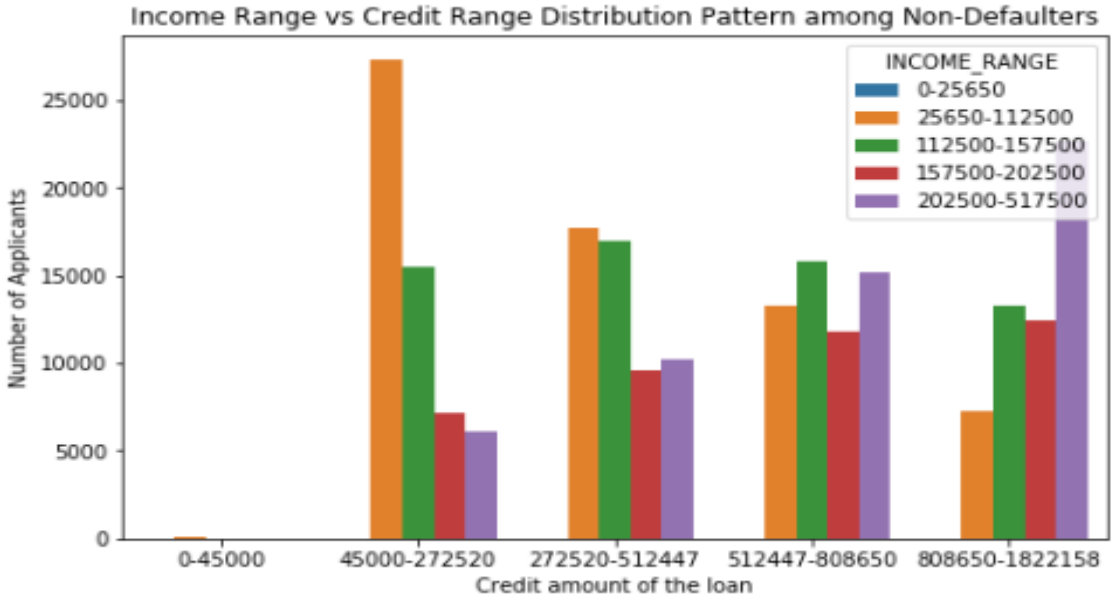


From the above plot, we can infer that,

- People earning between 2-5 lakhs and in the age group of 21-40 age have chances of being a defaulter

Analysis – Income Range vs. Credit Range

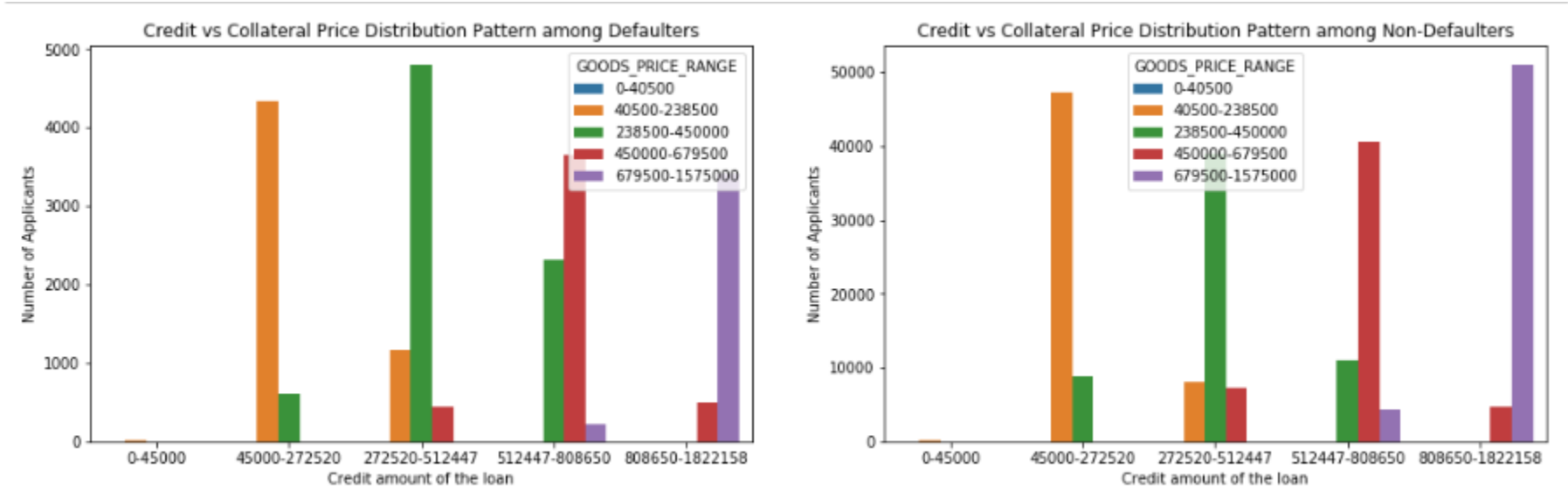




From the above plot, we can infer that,

- There are more chances of being a defaulter if a person earning less than 25000 applies for a loan.
- A person earning more than 25k but less than 1.6lakh, if goes for higher loan - more chances of paying his debt on time than the cases if he goes for less loan amount

Analysis – Credit Range vs. Goods Price Range



From the above plot, we can infer that,

- There are highest chances of being a defaulter if collateral amount is same or less than loan amount.

Recommendations:

Here are the suggestions company can utilise for its portfolio and risk assessment. Company should take care in providing loans to customers with following behaviours :

- Applicant's employment experience is less than 10 years.
- Applicant falls in age group of 21-30.
- Applicant earns in income bracket of 25K to 1.125K
- Education type is Lower secondary or Secondary/Secondary Special.
- Applicant is a Low Skilled Labourer.
- Collaterals with value in between 40K to 4.5 Lakhs.

Data Analysis

BASED ON LOAN ATTRIBUTES

ANALYSIS PERFORMED ON ONLY APPROVAL AND REFUSAL LOAN STATUS

Data Imbalance Calculation

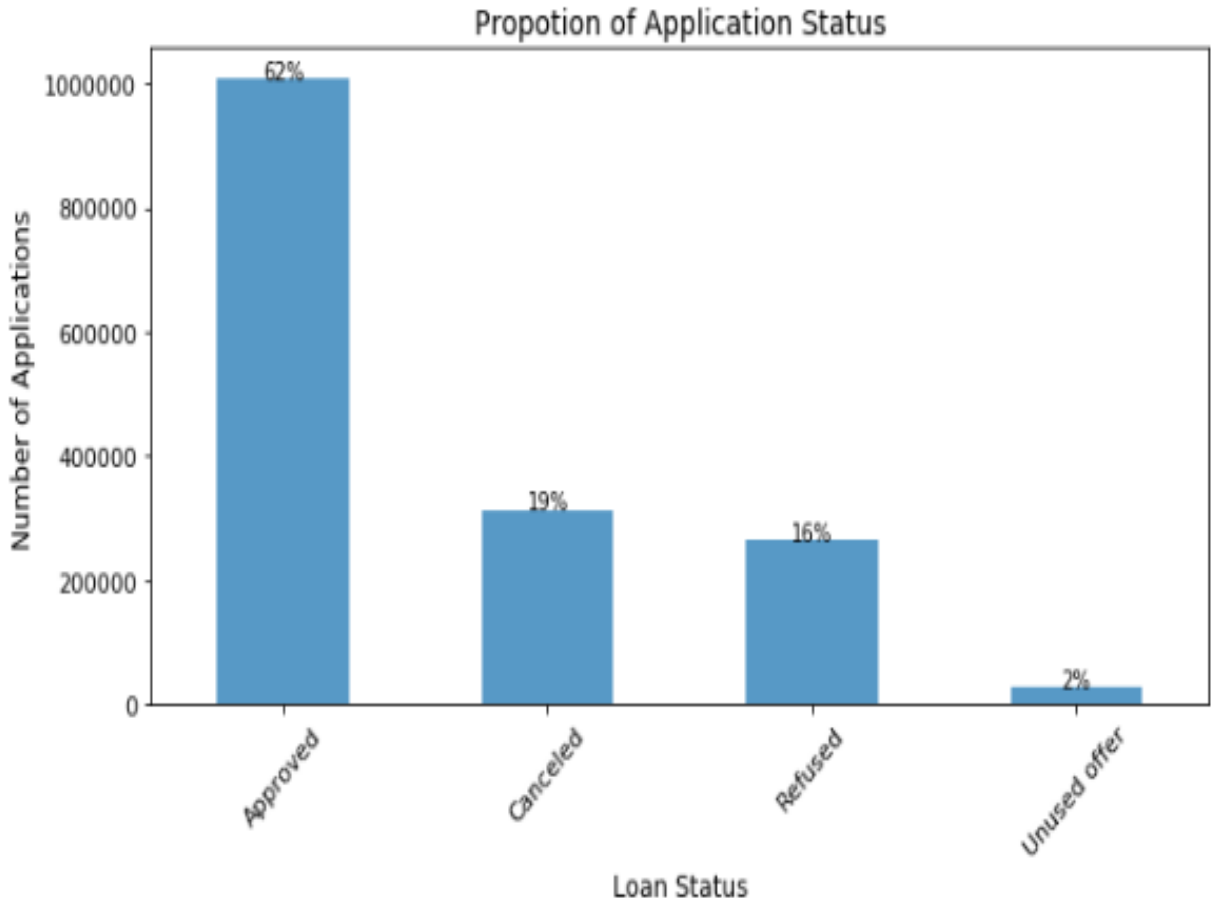
This graph implies that we have more applicants whose loan was approved

Also,

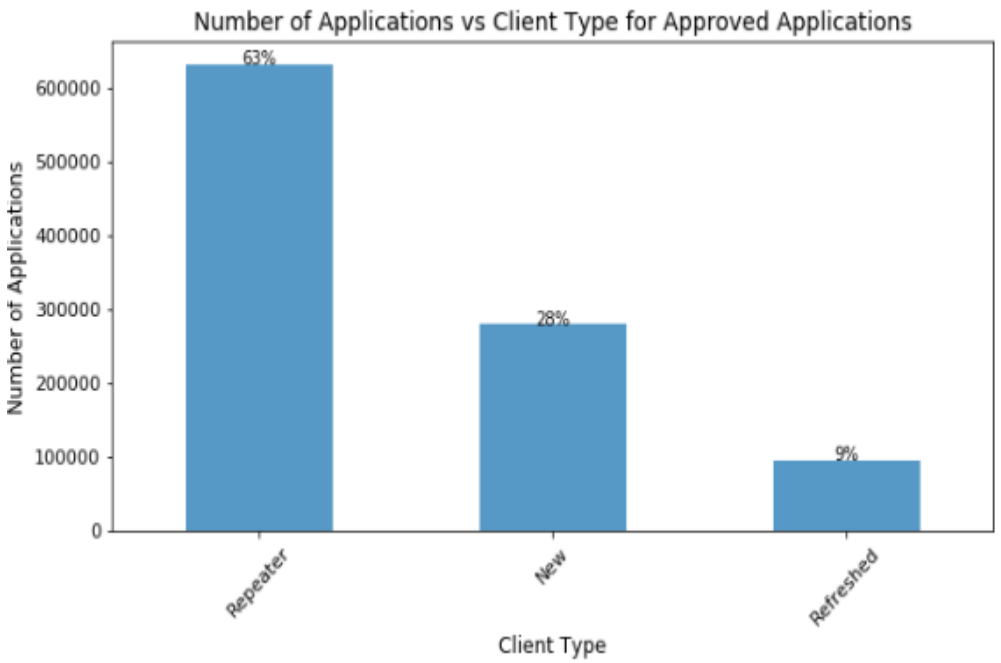
Percentage of Approved applicants in our dataset: 62.48

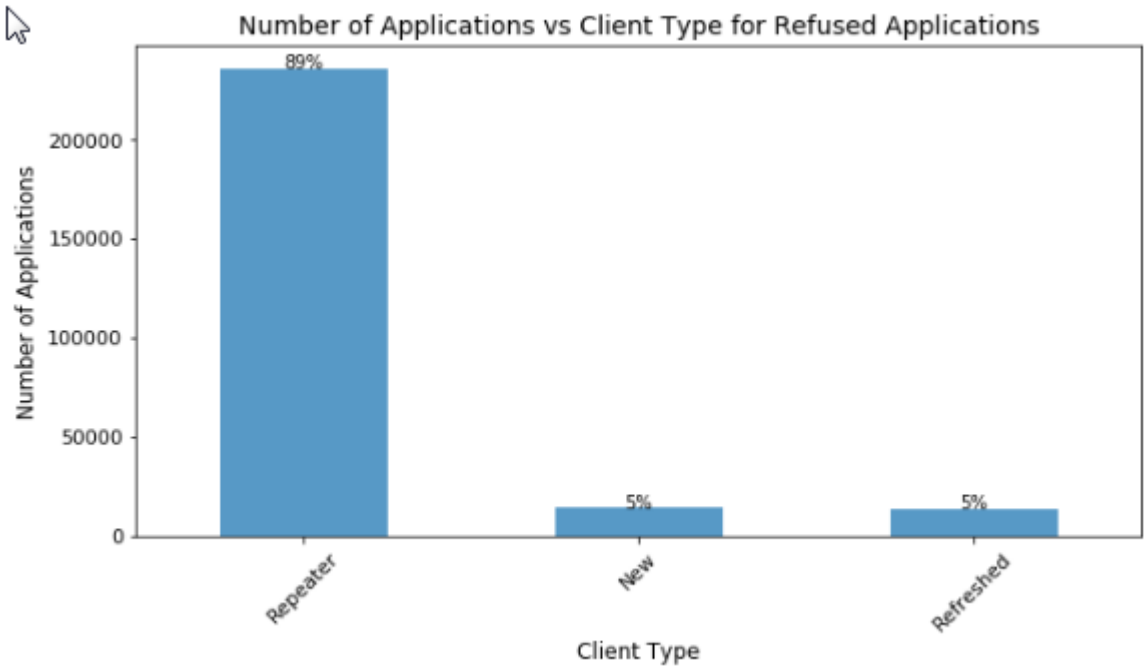
Percentage of Refused applicants in our dataset: 16.4

Data Imbalance Percentage in our dataset is: 26.25



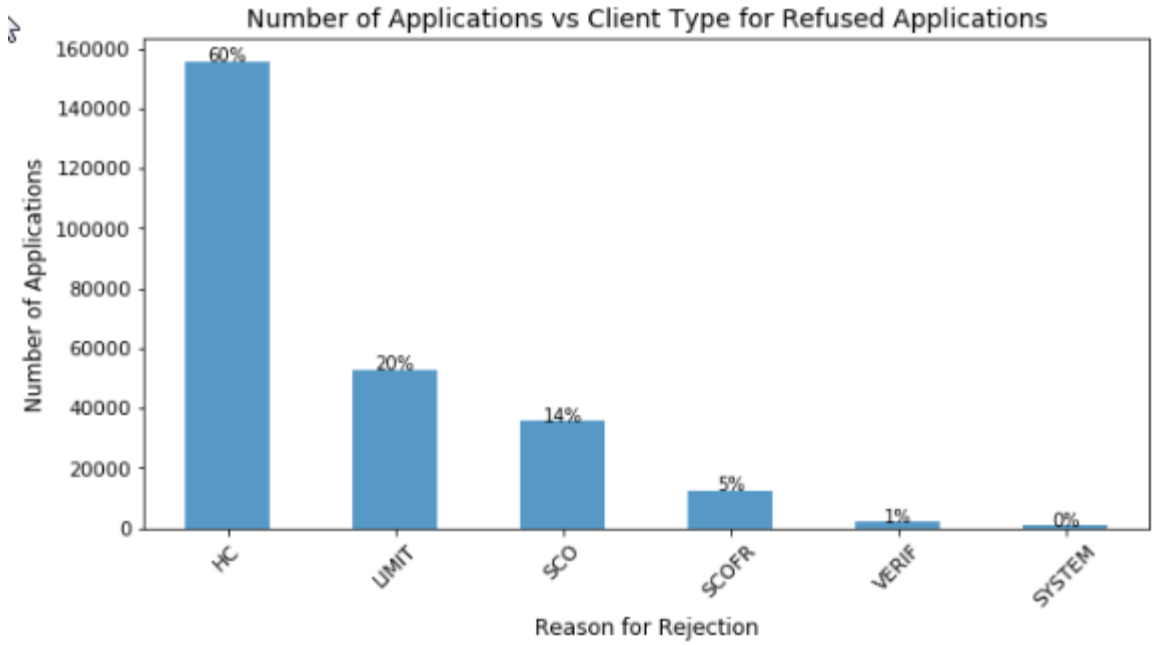
Analysis – Type of Client





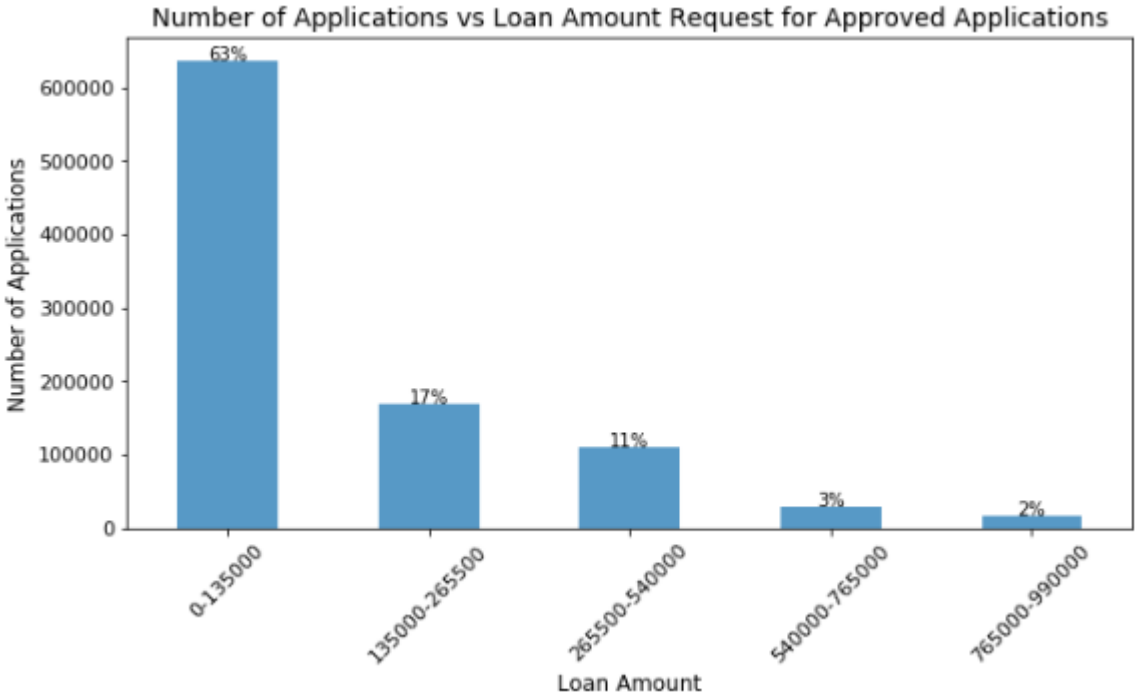
Insights -
There are highest probability of Repeater client's application getting approved followed by new applications.

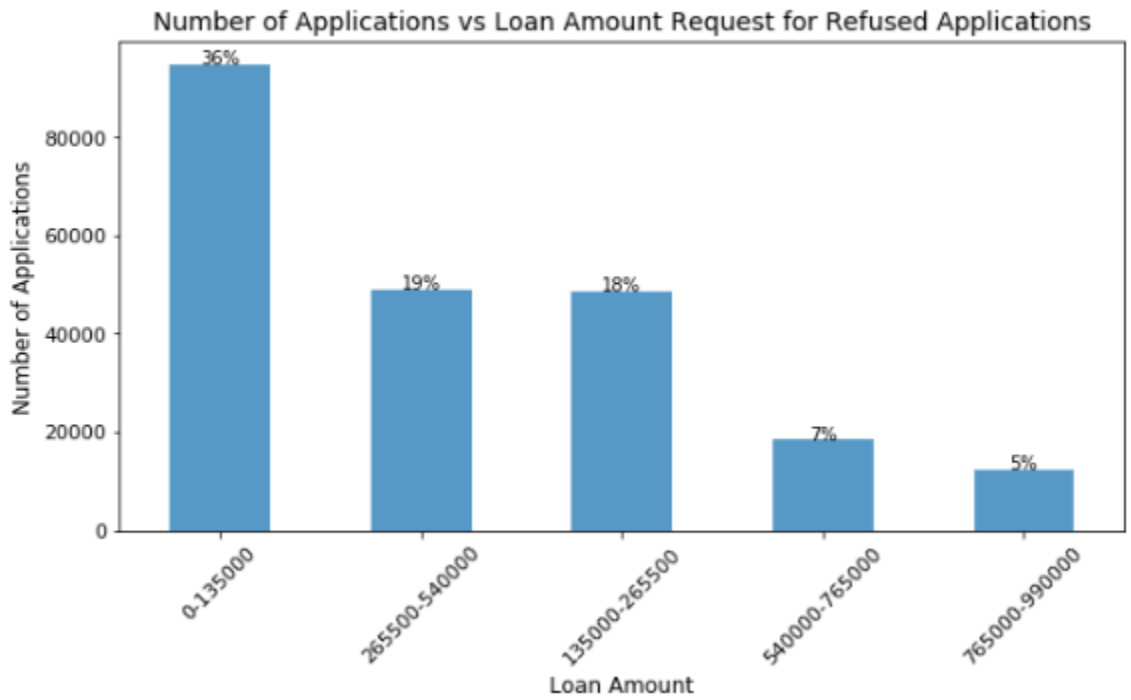
Analysis – Reason for Rejection



Insights - Most number of applications are rejected because of HC.

Analysis – Loan Amount



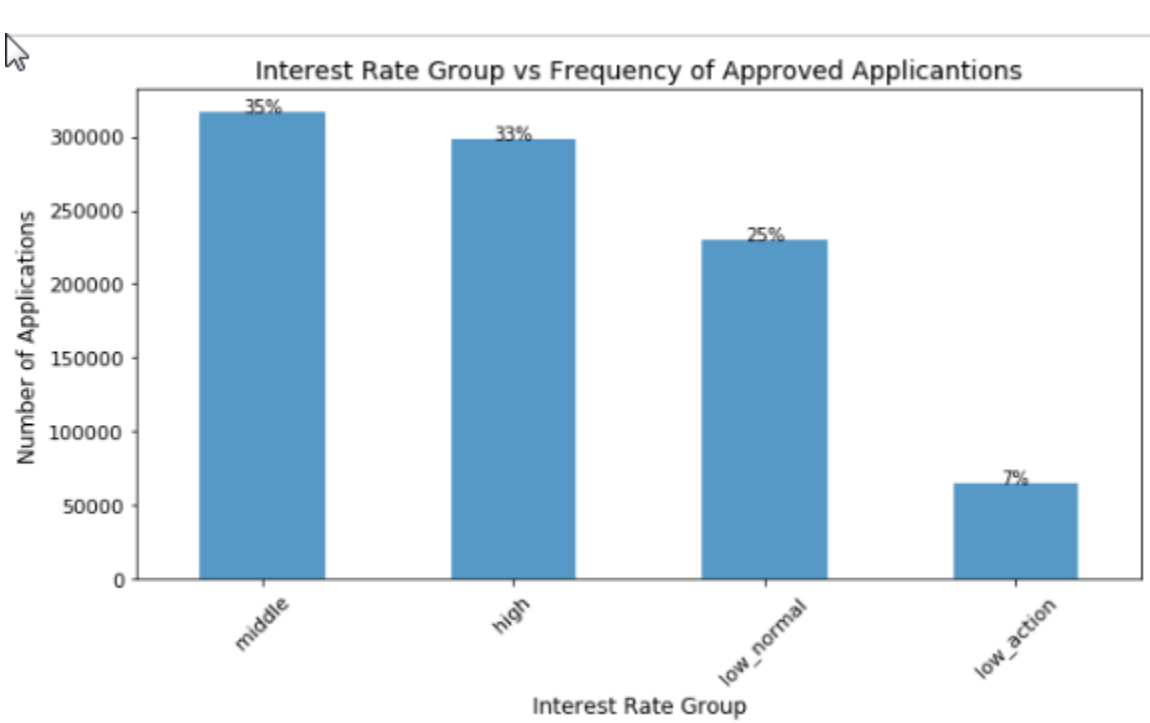


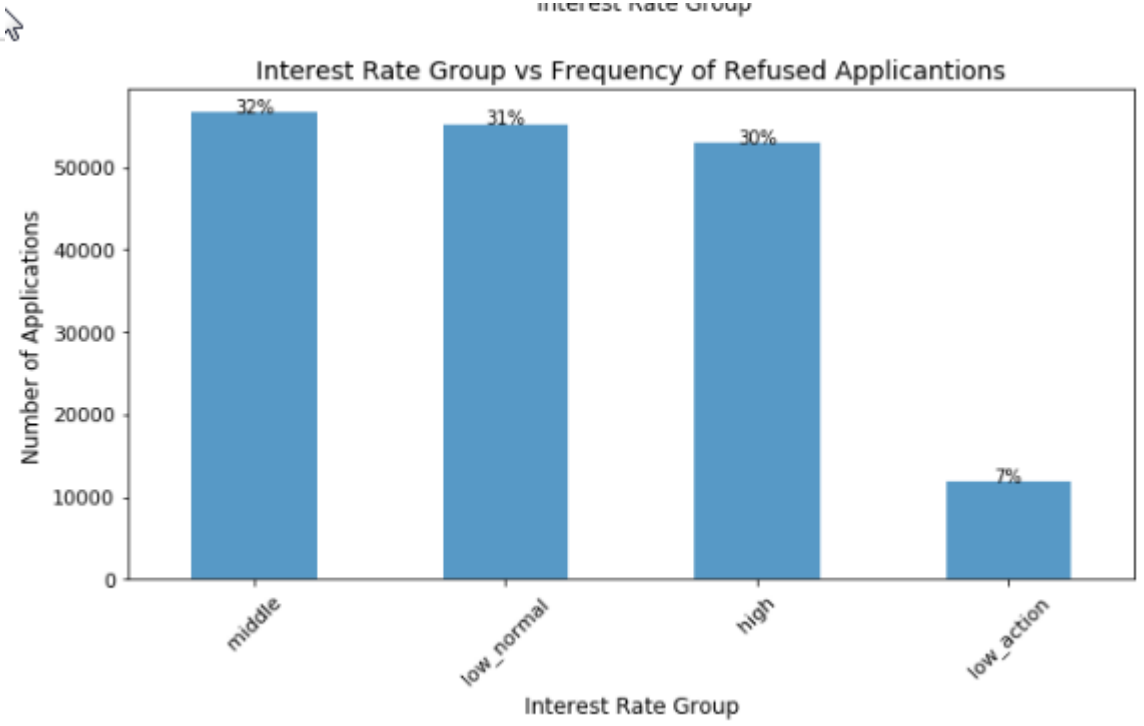
Insight-

More number of applications are approved if loan is <135000.

But with increase in loan amount request, chances of it getting rejected gets increased.

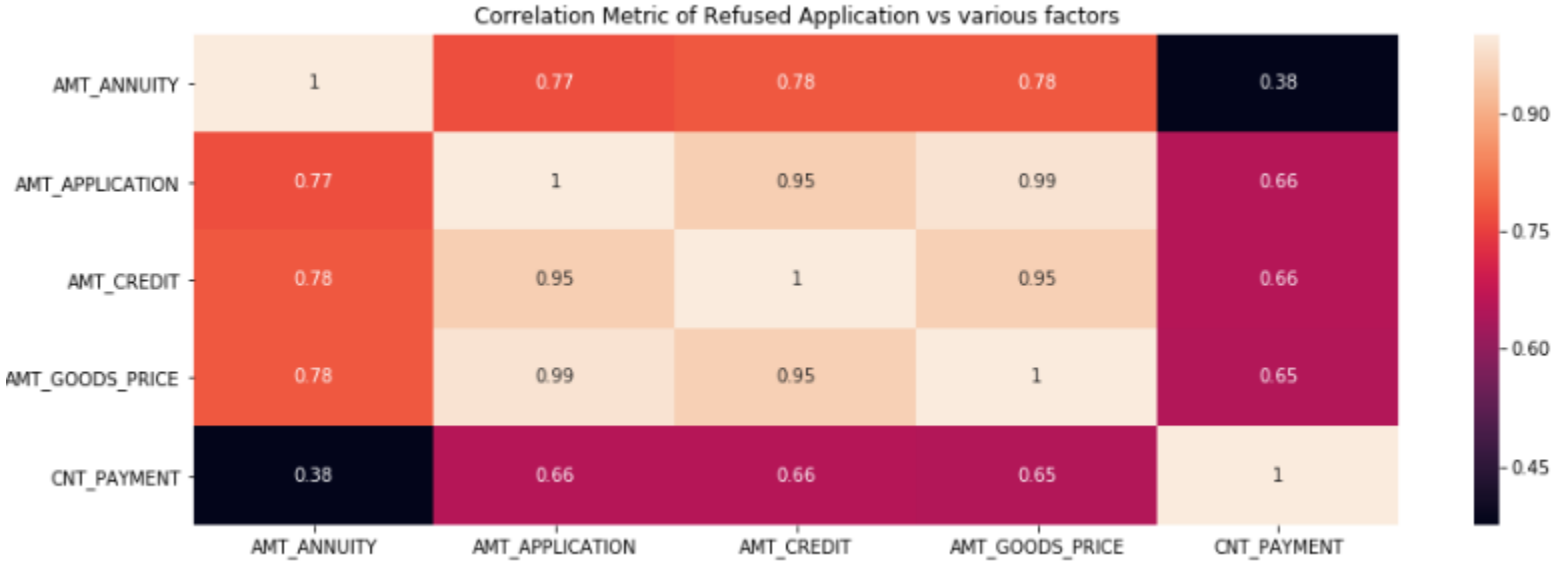
Analysis – Interest Rate





Insight -
There are higher chances of applications being rejected if interest group is low_normal.
There are higher chances of applications getting approved if interest group is medium or high.

Correlation Metric: Factors Influencing a Loan Refusal Behavior



Observations:

Following Factors influence an applicant behavior the most:

1. **AMT_ANNUITY**
2. **AMT_APPLICATION**
3. **AMT_CREDIT**
4. **AMT_GOODS_PRICE**
5. **CNT_PAYMENT**

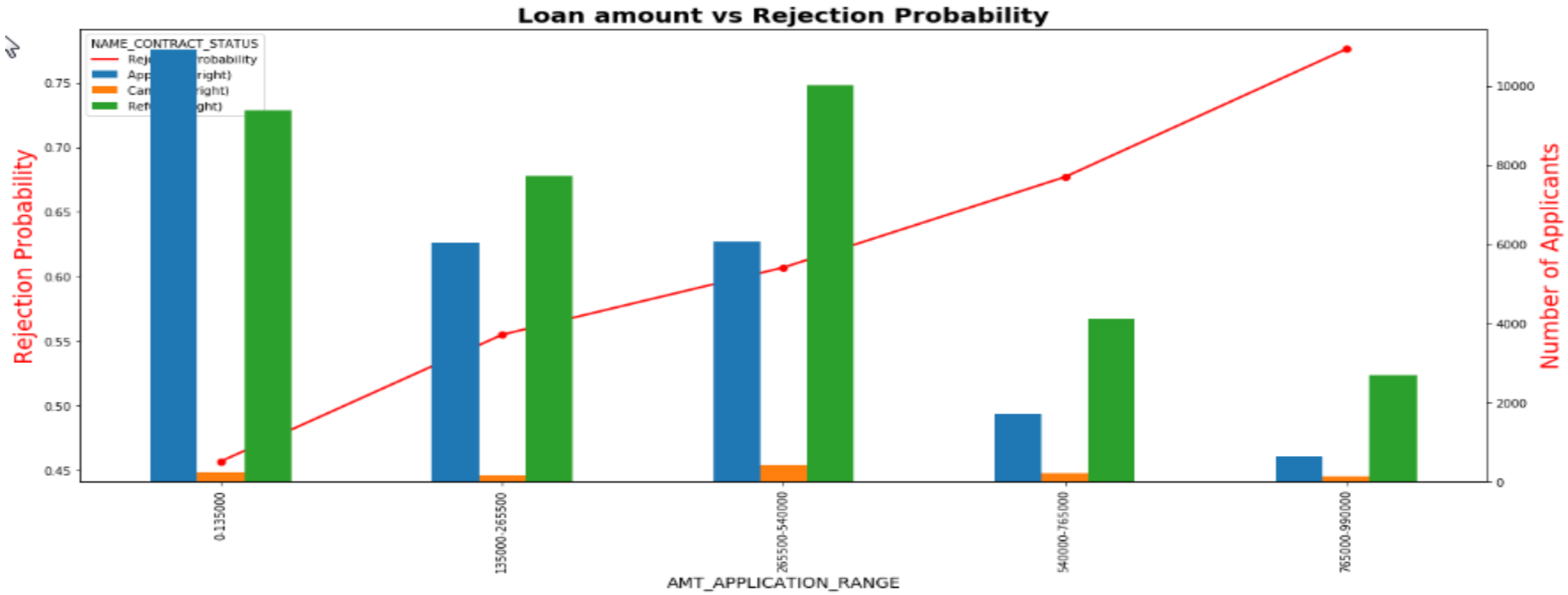
Below is the relation between above factors in case of a defaulter applicant:
Greater than 50% correlation ['AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE', 'CNT_PAYMENT']
Greater than 75% correlation ['AMT_APPLICATION', 'AMT_CREDIT', 'AMT_GOODS_PRICE']
Greater than 95% correlation ['AMT_CREDIT', 'AMT_GOODS_PRICE']

It is clear from the HEAT map that

- The loan amount is highly correlated with the goods price
- The loan amount credited is correlated with the annuity amount
- The count of payment is correlated with loan amount, loan amount credited and annuity amount

As a result, we can use one of these columns out of each for analyzing our data and providing the expected result

Analysis – Loan Amount vs Rejection probability

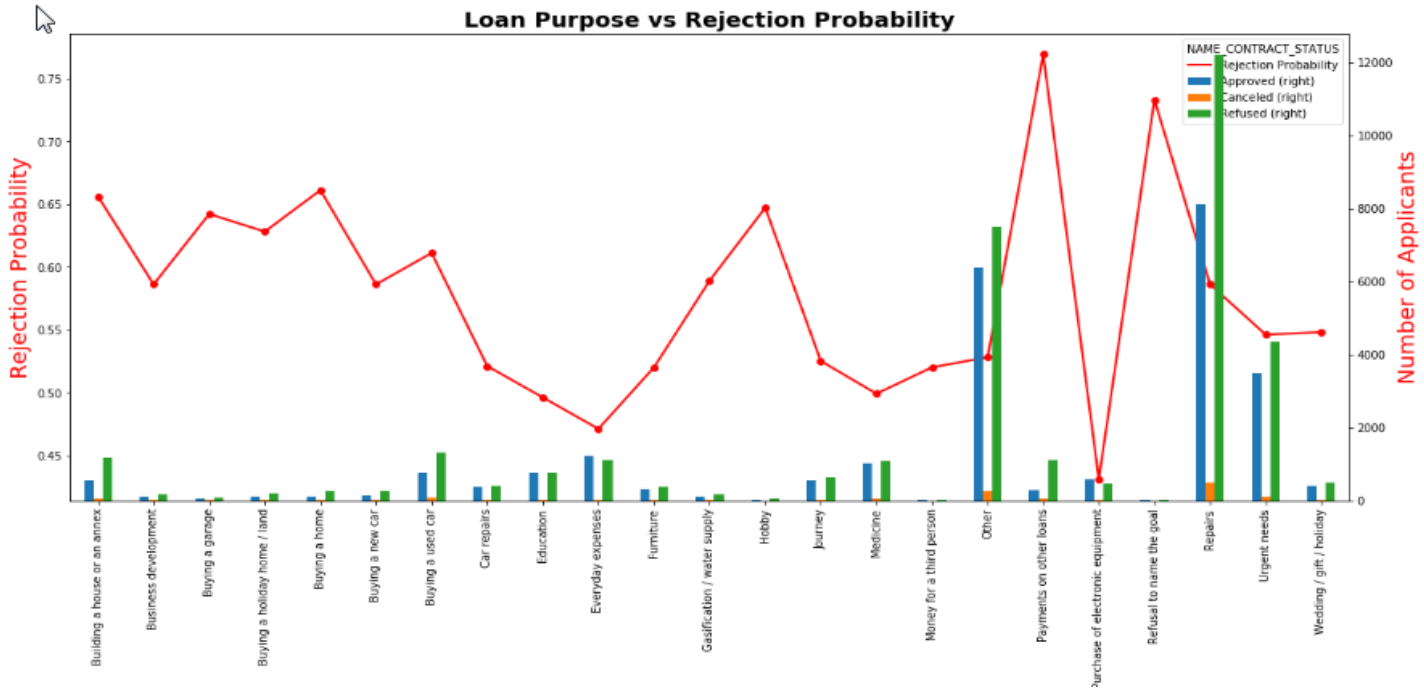


Rejection probability is high when the loan amount is more

Analysis – Loan Purpose vs Rejection probability

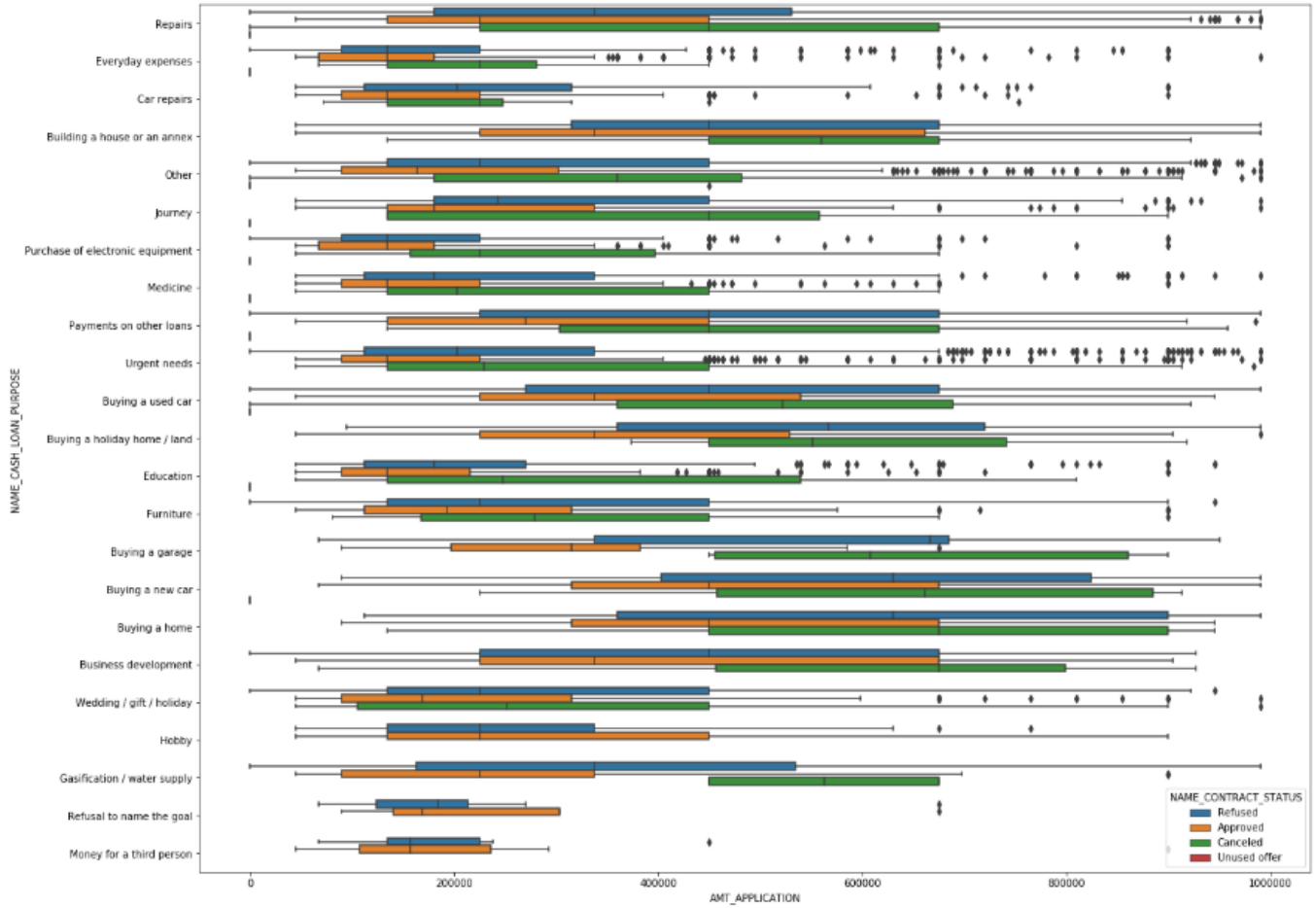
Insight-
Rejection probability is high on the below items

- Buying a Home
- Building a house
- Hobby
- Payment on other loans
- Refusal to name the goal



Relation between loan purpose, loan amount and status

- By looking at above plot we can see Everyday expenses and purchase of electronic equipment have a lot of outliers in Approved Status and which will contribute to more losses for the bank.
- Medicines, Car Repairs, urgent needs and Education also needs to be closely monitored.
- Hence while giving loan for above categories bank should have more proactive checks.



Key-Insights

Here are the suggestions company can utilise while approving or rejecting a loan.

While approving the loan, company should have a close watch on the below loan purposes

- Everyday expenses
- Purchase of electronic equipment
- Medicines, Car Repairs, urgent needs and Education

Rejection probability is high on the below items

- Buying a Home
- Building a house
- Hobby
- Payment on other loans
- Refusal to name the goal