Exploratory Data analysis

On

Credit application Data

Analysed by:
Swathi Somayaji

Introduction on Analysis report

The report on the “Credit Application data” has been drafted with an objective to present insights on the various aspects which can be a driving factor in deciding whether a credit application can be accepted or is a potential risk for the credit institution if application is passed.

The insights were identified by performing an exploratory data analysis based on the details of the applicants, which comprised of current application details and previous application history.

Objective of Analysis report:

The Exploratory data analysis on the “Credit Application data” is done to identify patterns which could act as driving factors during the process of underwriting for deciding whether a credit application should be accepted or is identified as a risky profile and should be rejected.

Apart from identifying the deciding factors, analysis is to be done on various other aspects to get an insight on the applicants prolife and behavioural patterns

Scope of the Analysis Report:

The scope of the document includes performing Exploratory data analysis on the below data sets.

1. Application_Data: Contains all the information of the client at the time of application.
2. Previous_application: Contains information about the client’s previous loan data.

The analysis was done on various aspects such as Income, Occupation, Gender, Type of loans, Credit rating etc.

This analysis aims to identify patterns which 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.This will ensure that the consumers capable of repaying the loan are not rejected.

Importing and Inspecting the Dataset

This is was the initial basic process which was followed as is followed in every EDA process.

The below data were imported and then inspected:

1. **Application_Data**: Contained details of nearly 307511 applicant's information across 122 data columns.

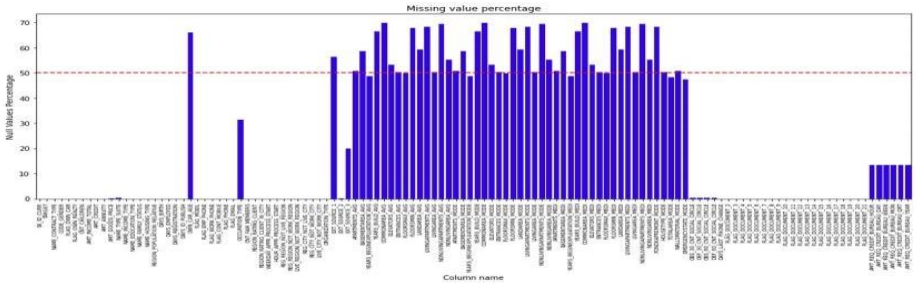
2. **Previous_application**: Contained information about the client's previous loan data. As observed, there were 1670214 rows and the data set also indicated that multiple loan application were applied by some applicants.

Data Cleaning

Missing value and analysis in Application_data:

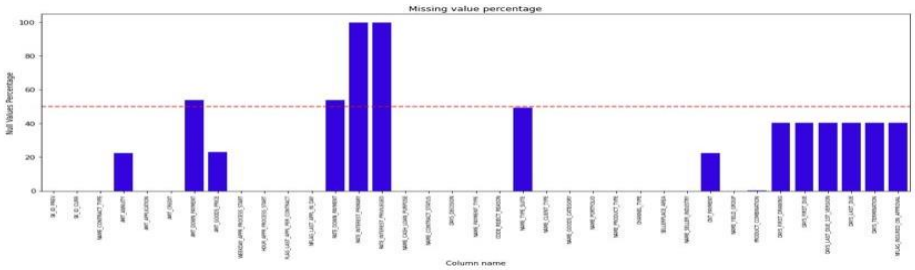
41 columns had more than 50% of missing values

The columns relating to the residential/housing information (APARTMENTS_AVG to EMERGENCYSTATE_MODE) had most of the null values.



From analysing ,Documents columns (FLAG_DOCUMENT_X) it could be inferred that most of the clients who applied had not submitted them.

Missing value analysis value in Previous_application:
Only 4 columns had missing value above 50% .

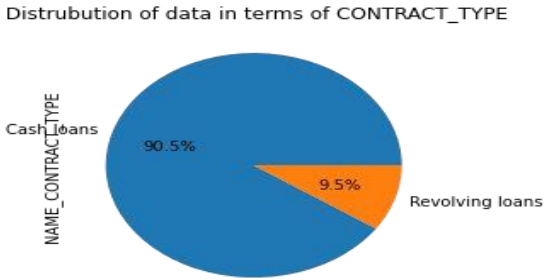


Credit Application Data- An insight

Background – The Analysis is on the total data set of the Applicants who have applied for credit. The data Contains details of nearly 307511 applicant's information across 122 data points ranging from type of loans (Cash/Revolving), Gender, occupation, income etc.

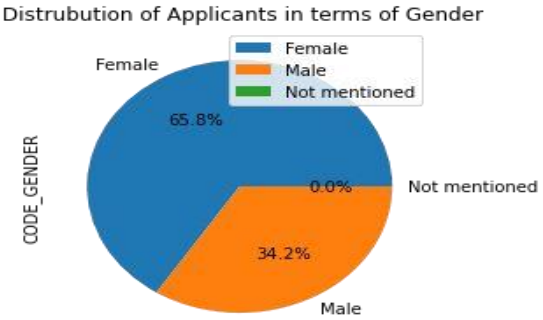
Out of the total no of applicants who applied for the loan, 90% of application was for Cash loans.

Type of loan	Total Applicants
Cash loans	278229
Revolving loans	29279



On analysing the percentage age distribution of gender, majority of the applicants are females at 65.8% and rest 34.2% are males. A very marginal population of 4 applicants have not declared their gender.

Gender	Total Applicants
Females	202447
Males	105057
NA	4



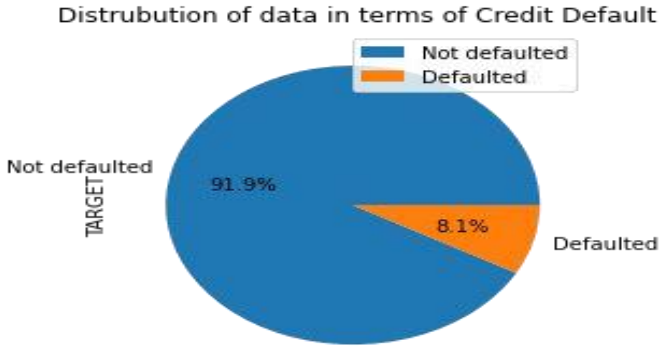
Credit Application Data- An insight

Speaking about the Defaulters, it comprised of 8.1% of the Applicants who defaulted on the credit out of total applicants.

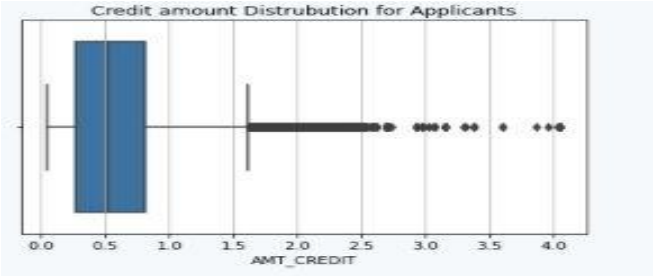
Default Flag	Total Applicants
Not Defaulted	282684
Defaulted	24824

Ratio of defaulted:not defaulted is 11.38:1
This indicates that for every 11.38 clients , one is defaulder.

This is the strong indicator that TARGET is a highly imbalanced data set



- Applicant credit amount mainly ranged between 0.27 to 0.8 million with a median of 0.51.
- The lowest credit amount which was applied was 0.04 millions and the highest Credit amount applied was 4.05 million.

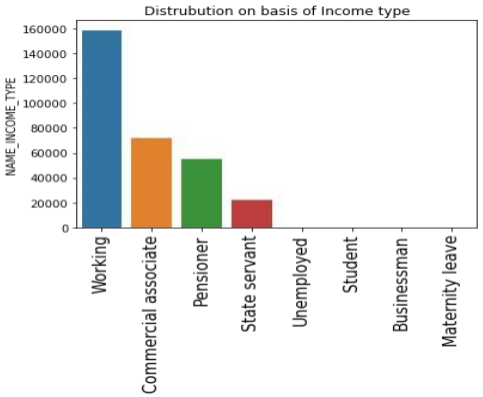


Analysing data on basis of the income type, applicants mainly comprised of

- “Working” group
- “Commercial associate”
- “Pensioner”
- “State servant”

Unemployed, Student, businessman and Maternity leave had a marginal share.

Income Type	No of applicants	% Share
Working	158773	51.632153
Commercial associate	71615	23.288825
Pensioner	55362	18.003434
State servant	21703	7.057703
Unemployed	22	0.007154
Student	18	0.005854
Businessman	10	0.003252
Maternity leave	5	0.001626



Credit Application Data- An insight

An insight to the profile of pensioner & Unemployed

Background - The analysis on the profile of “Pensioner” and “Unemployed” applicants were done by extracting a separate data set named “YEARS_EMPLOYED_MISSING” from the “Credit Application Data” of all applicants.

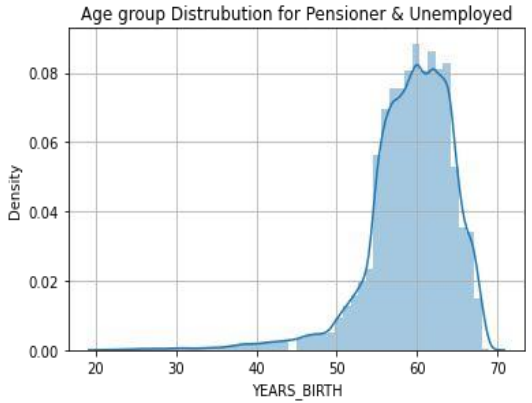
The data set was extracted on basis of the column “YEARS_EMPLOYED” wherein it was observed that a substantial data has been updated with "-1001" value, which was identified during the data cleaning/Standardization process.

The data mainly comprised of applicants with Income type of “Pensioners” and “Unemployed”.

- Pensioners are the third largest profile of applicants at 18% in terms of income source applying for credit at 55394 applicants.
- The Dataset also comprised of a marginal no of applicants at 22 who were unemployed.

- The age group distribution of the Applicants in the data set were primarily concentrated in the range of 57 to 63 years, with an average age of applicants being at 67 years.

Parameters	Age in years
Lowest Age	21
Lower Quartile (25%)	57
Average (mean)	59
Upper Quartile (75%)	63
Highest Age	67



An insight to the profile of pensioner & Unemployed

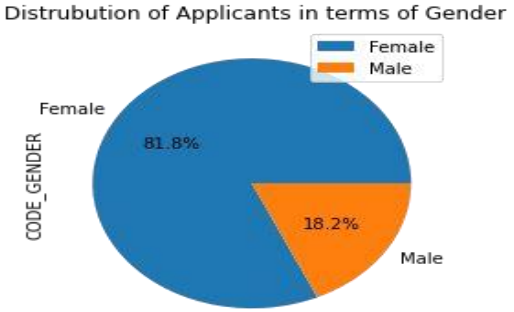
- The percentage of female applicants in the dataset distribution were found to be on the majority at 82%.

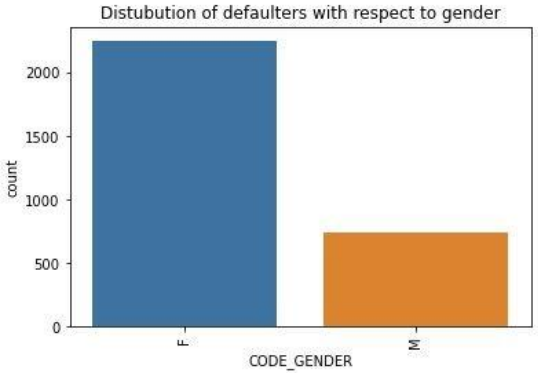
Gender	Total Applicants
Females	45278
Males	10106

- Females' applicants defaulted the most compared to Male applicants at 75% out of a total defaulter applicant of 2990.

Gender	Total Applicants (Defaulter)
Females	2249
Males	741

- On analysing the data further of only the unemployed which comprised of 22 applicants, it was observed that the defaulter applicants have applied for "Cash Loans" and comprised of females who aged between 57-63.





An insight on defaulter Applicants

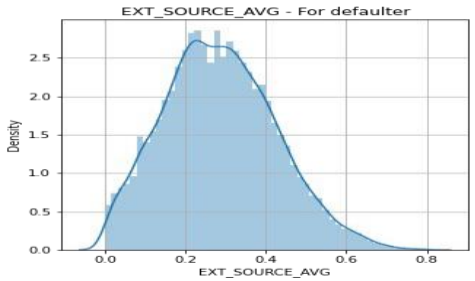
Analysis on basis Normalized Score Data of Applicants – The Analysis was done by calculating the average of the EXT_SOURCE columns in the Applicants Data set. (“EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3).

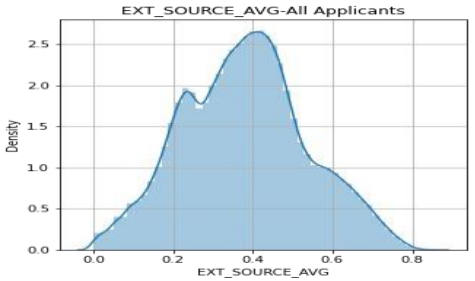
- The normalised score of the defaulter applicants were in the range of 0.2-0.4.

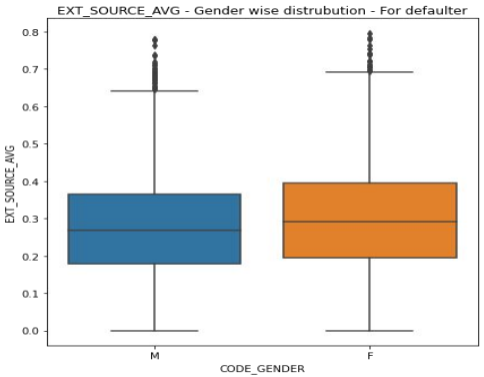
- In case of all applicant data, the maximum no of applicant had normalised score of 0.4 with peaks in 0.2 and some applicants in 0.6 range.

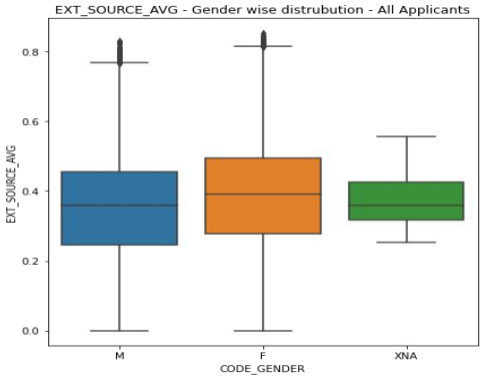
Plotting the Normalized score gender wise:

- The normalised score of male and female (defaulter) is similar within the range 0.2 to 0.4.
- For all applicants, the score was in range of above 0.2 and close to 0.5. For a major no of applicants gender was unspecified, and was in the range of nearly 0.3 to a slightly above 0.4





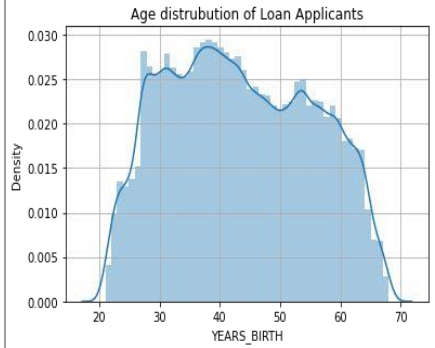


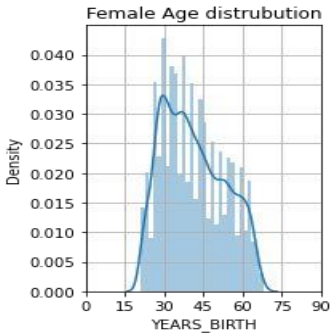


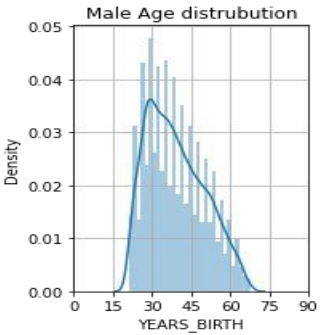
An insight on defaulter Applicants

Analysis on basis of Age:

- The Applicants defaulting the most are in the age group of 25 – 30 and witnessed across all gender.
- No of defaulters decreased as the age of applicants increased.
- Lowest age and highest of both male and female were found to be same, ie 21 and 69 respectively.

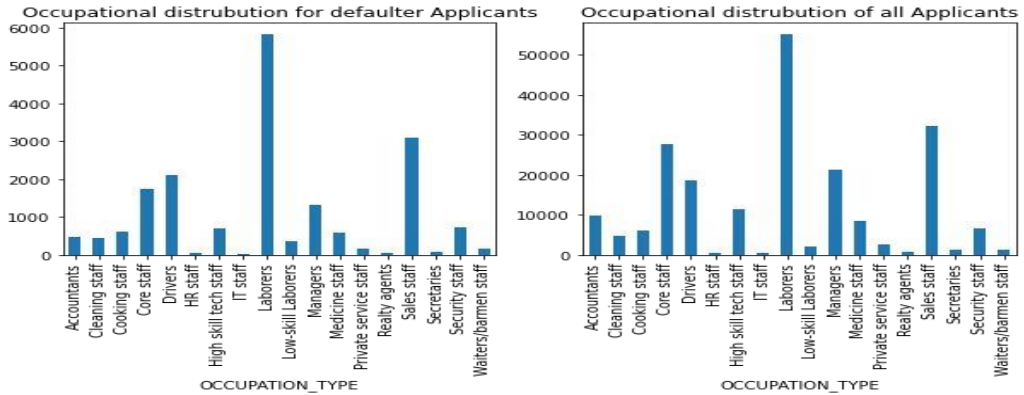






- Below are the top 5 occupation wherein the 56% of the applicants who have defaulted were engaged out of a total defaulter population of 24824.

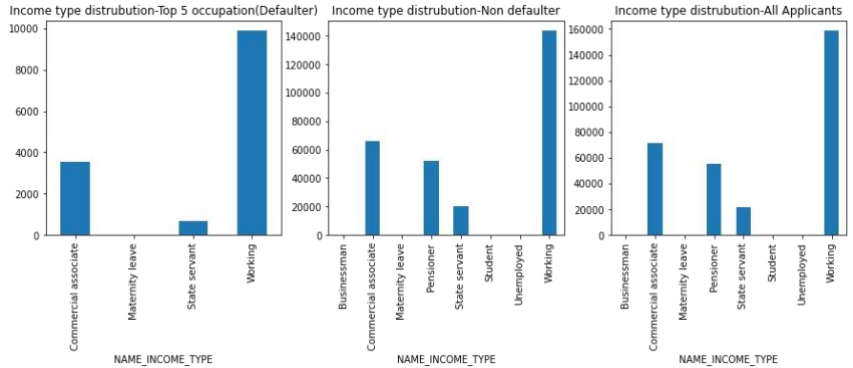
Parameters	No of Defaulter
Labourers	5837
Sales Staff,	3092
Drivers	2107
Core Staff	1738
Managers	1328



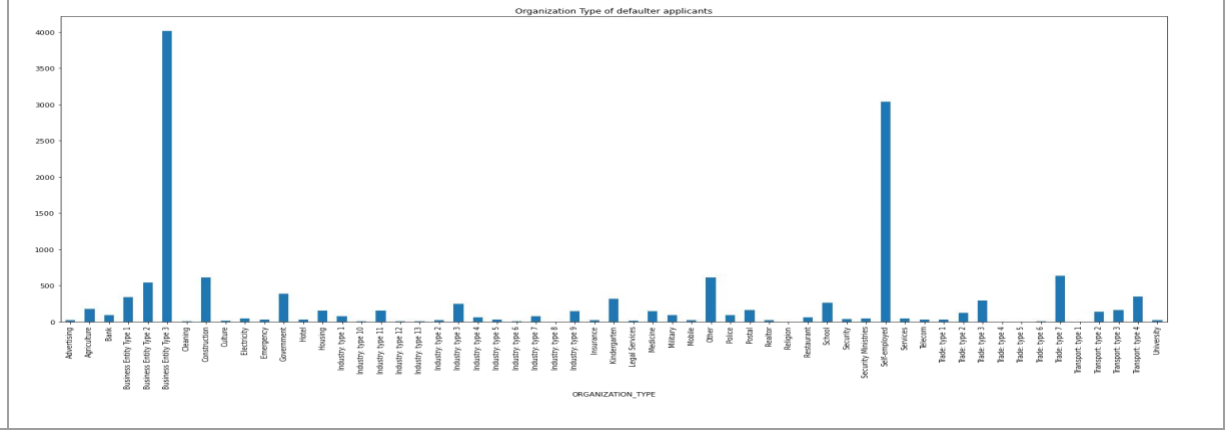
Analysis on basis of Occupation:

The top defaulters are the "working" followed by "Commercial Associate" and "State servant".

“Pensioners” are not among the top 5 defaulters on basis of occupation despite 18% of the total applicants on basis of "Income Type".



On analysis of the applicant's background on basis of their organisation of engagement, it is observed that the maximum no of Defaulters under the top 5 Occupation type belong to "Business Entity Type 3" and "Self Employed".



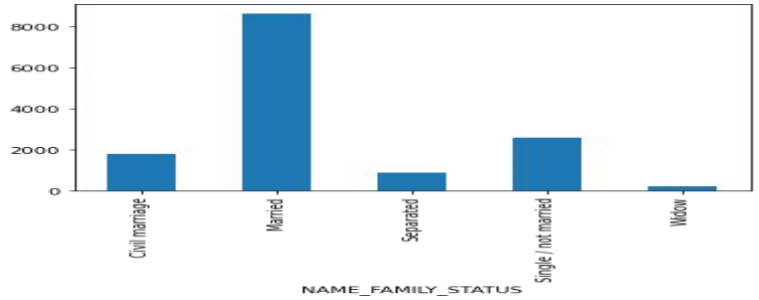
An insight on defaulter Applicants

An insight on defaulter Applicants

Based on the analysis of the Top 5 occupational defaulters

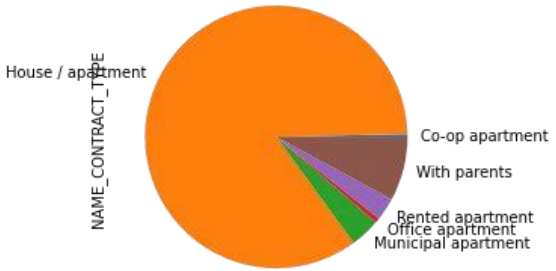
Marital Status of top 5 occupational defaulters:

- Married and single/Not Married are most likely to default compared to "Widow" or "Separated”



Housing type of top 5 occupational defaulters

- Majority of the defaulter applicants resided in “House / apartment”.

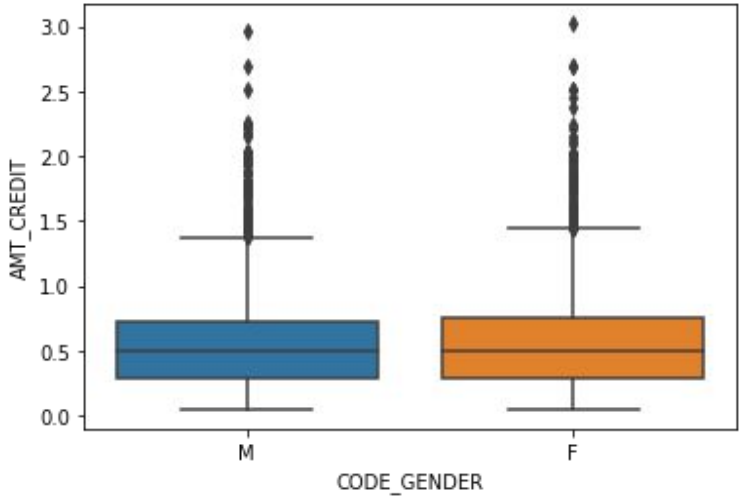


An insight on defaulter Applicants

Based on the analysis of the Top 5 occupational defaulters

Verification of data set in terms of defaulter applicants Credit Value:

- The credit value range observed for both the genders are similar.



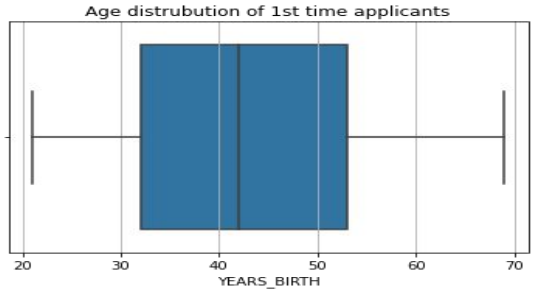
An analysis on the data of 1st time Applicants

Background – In the analysis, data of applicants who did not have any previous history of credit was extracted. The extraction was done by merging the application data with the previous application data set, and then extracting the data of the applicants for whom there was no history of applicant

Out of total applicants of 307511, there were 16454 applicants who did not have any previous application history, and are thus considered as 1st time applicant

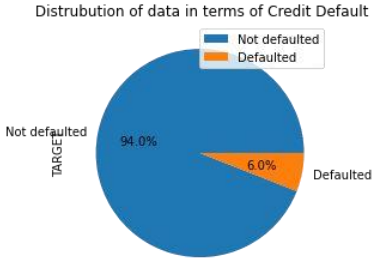
- The age distribution is mainly in range of 32 to 53 Years, with a median of 42 Years.

Parameters	Age in years
Lowest Age	21
Lower Quartile (25%)	32
Median	42
Upper Quartile (75%)	53
Highest Age	69



- 6% of first time applicants were found to be defaulters out of the Total Applicant Data set.

Default Flag	Total Applicants
Not Defaulted	15474
Defaulted	980



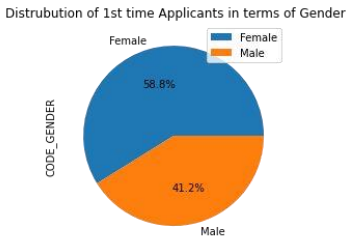
An analysis on the data of 1st time Applicants

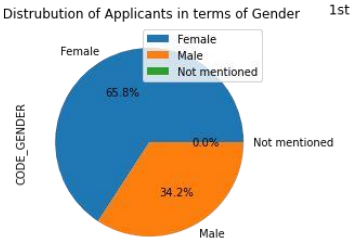
- In the gender wise distribution of 1st time applicants, 58.8% applicants were females and remaining 41% were males. If this is compared with the gender wise distribution of all the applicants, then 65.8% were males and 34.2% were females.

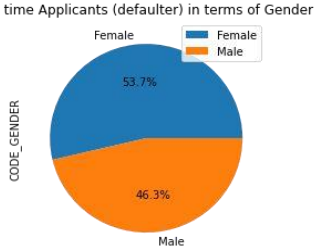
- The gender wise distribution of defaulter among 1st time applicants, is nearly 53% are females and 46% are males.

Gender	Total 1st time Applicants (defaulter)
Females	526
Males	454

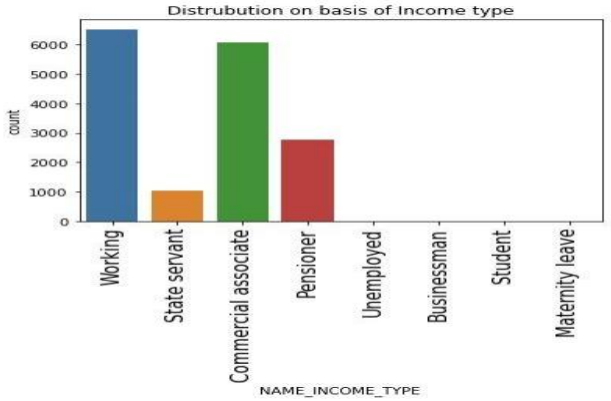
- The income type for 1st time applicants, “Working” group followed by Commercial associate and Pensioner.





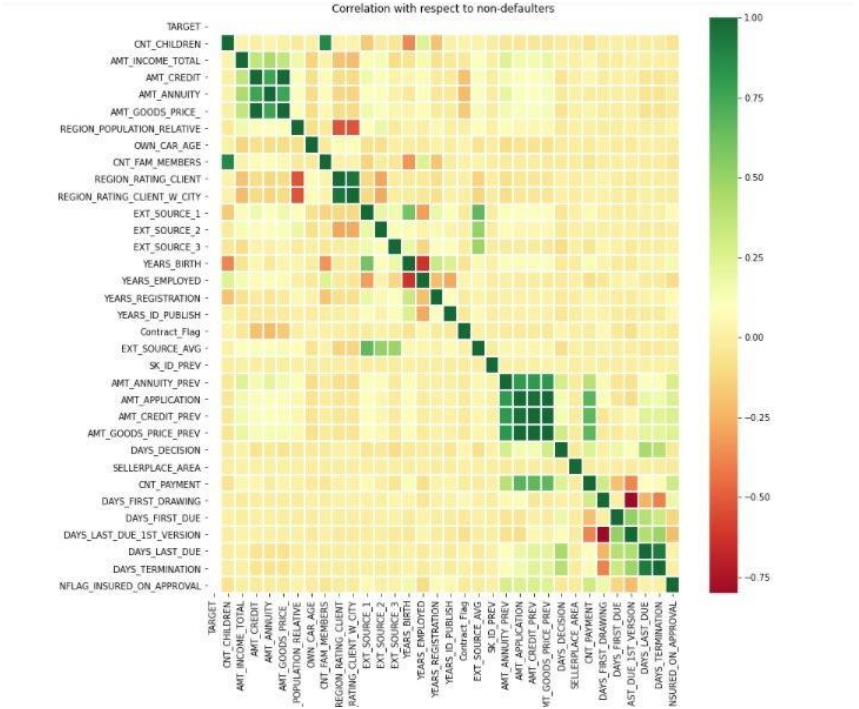


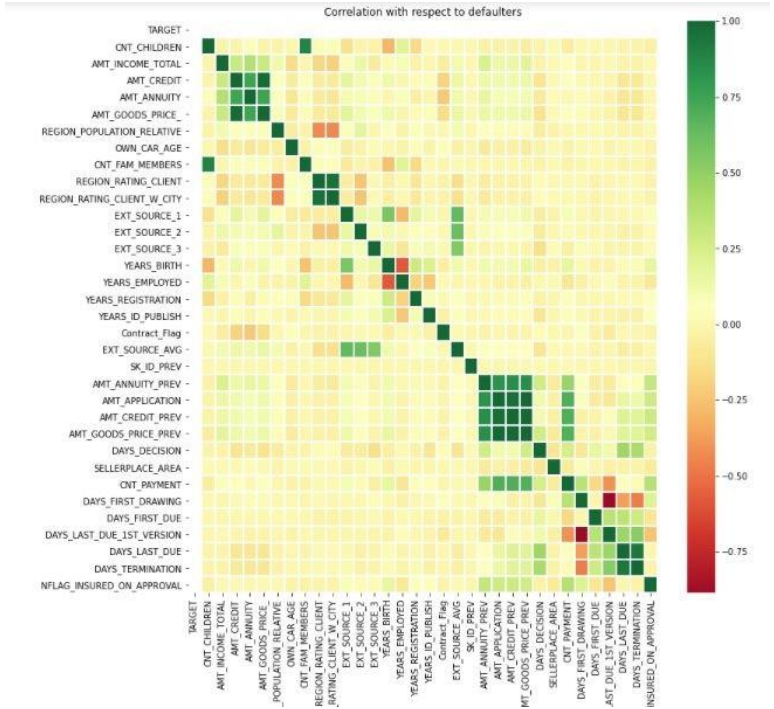
- State servant, unemployed, student, Businessman and Maternity Leave comprised of a very marginal set of applicants.



Analysis of Historical Data set

Background - The Analysis was done by merging post data cleaning and removing unwanted columns of previous_application data to application data, the below is the correlation analysis that was done for defaulters and non-defaulters.


Correlation with respect to non-defaulters
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE_
REGION_POPULATION_RELATIVE
OWN_CAR_AGE
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
YEARS_BIRTH
YEARS_EMPLOYED
YEARS_REGISTRATION
YEARS_ID_PUBLISH
Contract_Flag
EXT_SOURCE_AVG
SK_ID_PREV
AMT_ANNUITY_PREV
AMT_APPLICATION
AMT_CREDIT_PREV
AMT_GOODS_PRICE_PREV
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Correlation with respect to defaulters
TARGET
CNT_CHILDREN
AMT_INCOME_TOTAL
AMT_CREDIT
AMT_ANNUITY
AMT_GOODS_PRICE_
REGION_POPULATION_RELATIVE
OWN_CAR_AGE
CNT_FAM_MEMBERS
REGION_RATING_CLIENT
REGION_RATING_CLIENT_W_CITY
EXT_SOURCE_1
EXT_SOURCE_2
EXT_SOURCE_3
YEARS_BIRTH
YEARS_EMPLOYED
YEARS_REGISTRATION
YEARS_ID_PUBLISH
Contract_Flag
EXT_SOURCE_AVG
SK_ID_PREV
AMT_ANNUITY_PREV
AMT_APPLICATION
AMT_CREDIT_PREV
AMT_GOODS_PRICE_PREV
DAYS_DECISION
SELLERPLACE_AREA
CNT_PAYMENT
DAYS_FIRST_DRAWING
DAYS_FIRST_DUE
DAYS_LAST_DUE_1ST_VERSION
DAYS_LAST_DUE
DAYS_TERMINATION
NFLAG_INSURED_ON_APPROVAL
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Analysis of Historical Data set

Correlating factors amongst Non-Defaulters:

Credit amount in application details as will as historically is highly correlated with

1.amount of goods price

2.loan annuity

3.total income

DAYS_LAST_DUE_1ST_VERSION is highly negatively correlated with DAYS_FIRST_DRAWING(i.e, first due and first disbursement is negatively correlated)

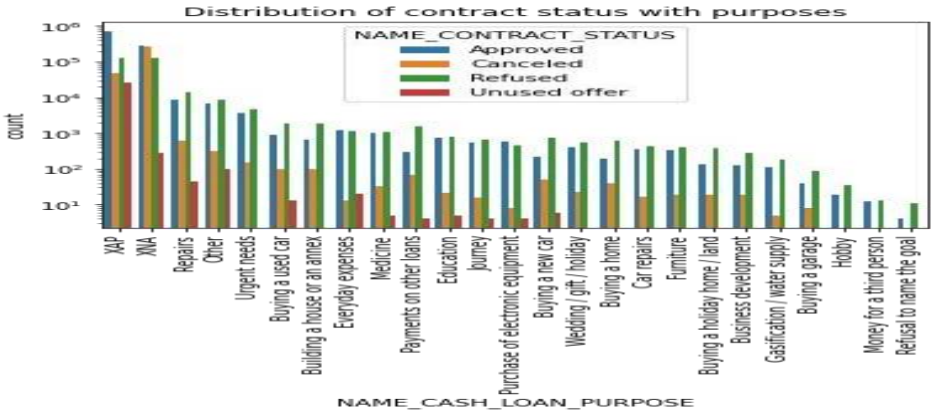
Correlating factors amongst Defaulters:

1.Credit amount is highly correlated with amount of goods price which is same as repayers

2.DAYS_FIRST_DUE and DAYS_TERMINATION correlation has increased in defaulters compared with repayers

Distribution of contract status with purpose:

High number of loan purpose is unknown (XAP,XNA). from the known purposes, Repairs has highest rejection and it is considered risky by the bank.



Conclusions:

- Female applicants are more compared to male Applicants across all data subset (pensioner, defaulter, 1st time applicants).
- The Applicants defaulting the most are in the age group of 25 – 30 and witnessed across all gender.
- The No of defaulter decreased as the Age of Applicants increased.
- Pensioners" are not among the top 5 defaulters on basis of occupation.
- Female applicant in age of 57-63 in with unemployed income type and applying for cash loans was considered to be a risky profile
- The income type applicants mostly comprised of "Working" group followed by Commercial associate and Pensioner.
- Repairs had highest rejection rate and it is considered risky by the bank.