Credit EDA Assignment

By Nikhita Varma Polakonda

Index

- Problem statement
- Approach & Methodology
- Data imbalance – ratio (Target variable)
- Univariate/Bivariate Analysis
- Segmented Univariate/Bivariate Analysis
- Correlations, Heatmaps & Top 10 Target variables
- Analysis on Previous application data
- Merged-Data(Current+Previous application) Analysis
- Conclusion

Problem statement

- This case study 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 those who are capable of repaying the loan are not rejected.
- The company can utilise this knowledge for its portfolio and risk assessment

Approach & Methodology

- Start off by importing, reading and understanding the data
- Drop columns and rows with significant missing data(greater than 40%), after ensuring the information being dropped is not important.
- Wherever necessary, impute missing data with the Mode, Median, Mean or with "Unknown" or "Others".
- Ensure that the data is in the right format by standardizing the values.

Approach & Methodology (continued)

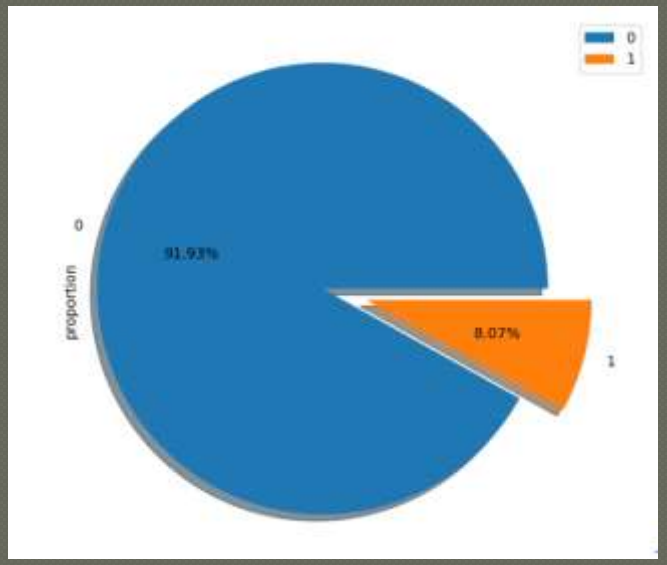
- Bin numerical data wherever necessary.
- Identify outliers and treat them by imputing them with upper and lower whiskers.
- Identify the imbalance of the Target column in the data.
- Split the data based on targets – defaulters, and others (Non-defaulters).
- Perform univariate and sub-univariate analysis on different variables.

Approach & Methodology (continued)

- Combine the Current ID and Target columns from the application data set to the previous application data set.
- Perform univariate and sub-univariate analysis on variables in the second data set.
- Identify the top 10 correlation between variables, for each target.

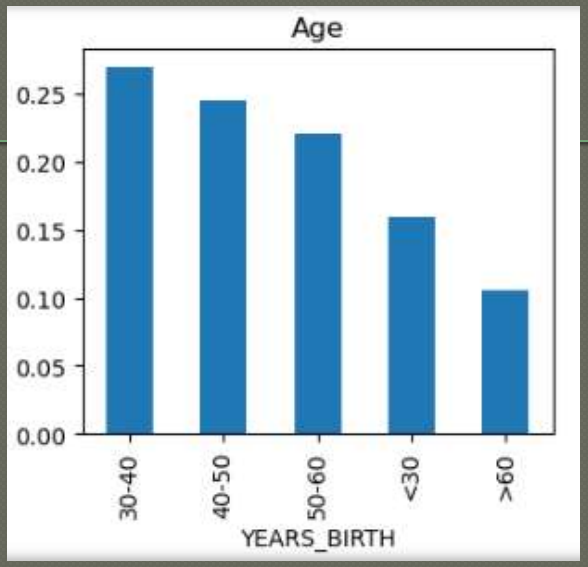
Imbalance Percentage

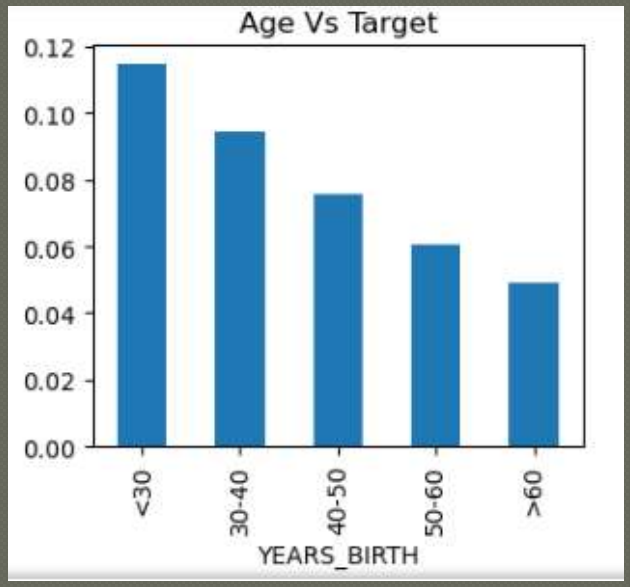
This Pie plot on target variable indicates that out of the total data the number of defaulters are very less (8.07% of the total data) as compared the non defaulters (91.93% of the total data).



Univariate Analysis

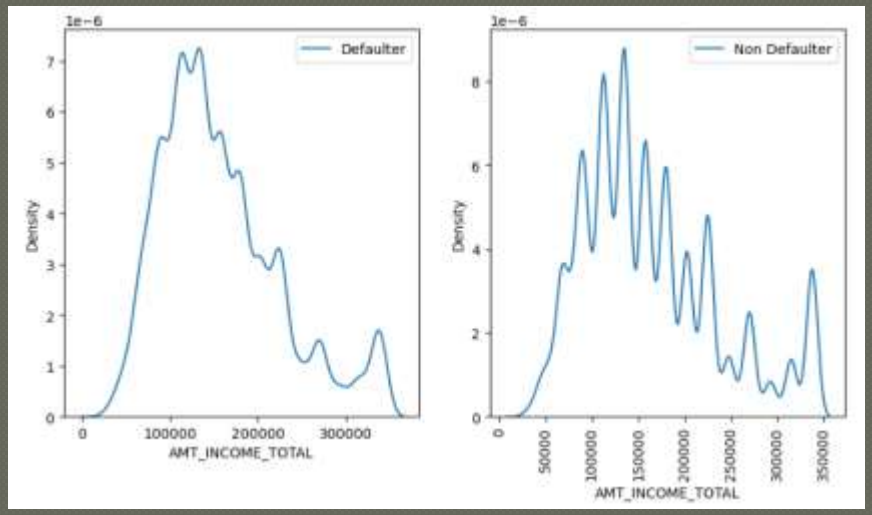
- We can say that the clients whose age is less than 30 are less in number after the senior citizens.
- However they are the ones with highest default rates. While the senior citizens whose age is greater than 60 are less in number and at the same time they have least default rates.
- Hence, banks should focus more on the senior citizens and be more careful with the young clients whose age is less than 30. Here, there is a linear relationship between the age and the default rate. As the age increases the defaulters reduce.





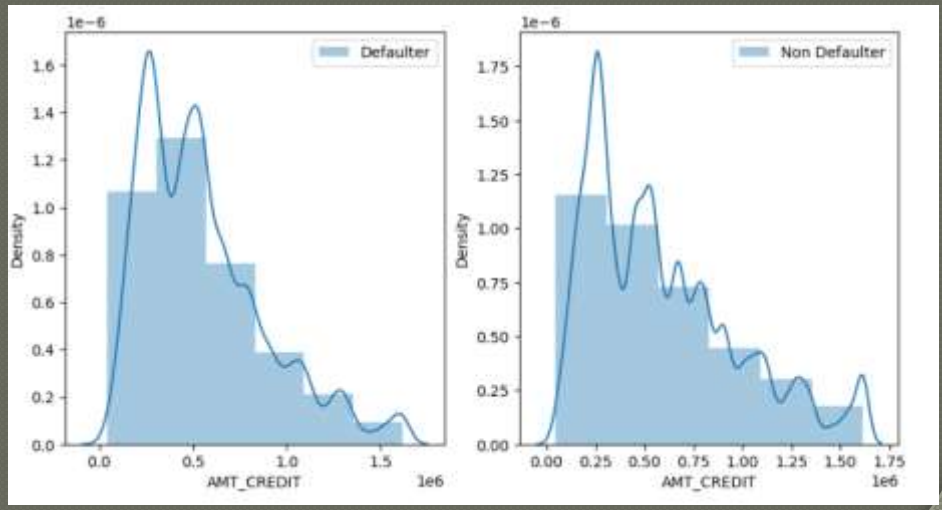
Income

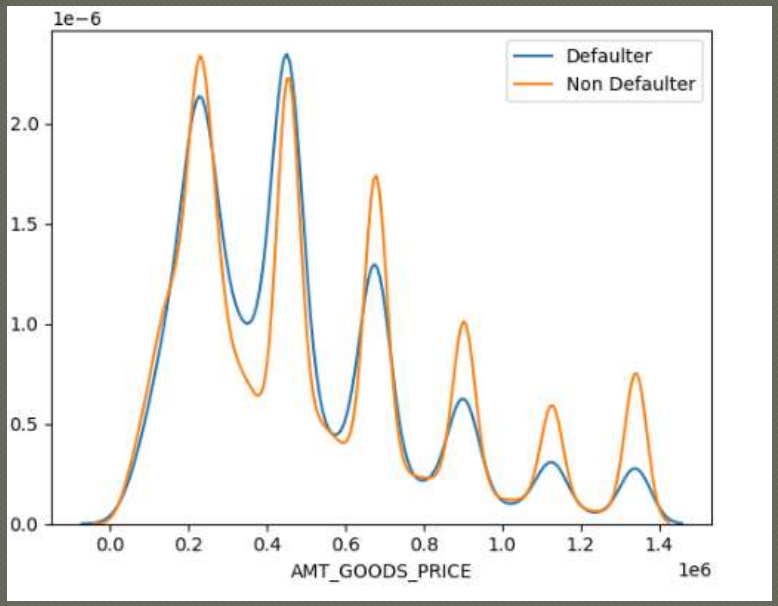
It has been observed that maximum defaulters are those people in whose income is in range 1 to 1.5 lakhs. Also as the income increases to 3 and 3.5 lakhs the count of defaulters decreases.On the contrary in non defaulters there is a mixed trend. People with high and low income group are all non defaulters.



Credit Amount

Under defaulters the credit values have peaks at 250000 and 500000. In non defaulters the credit values have peaks at 250000 then at 500000 and still peaks keep on having their rise and falls with peaks at 750000,1250000 and so on but the density keeps on decreasing post 500000.





Goods Price

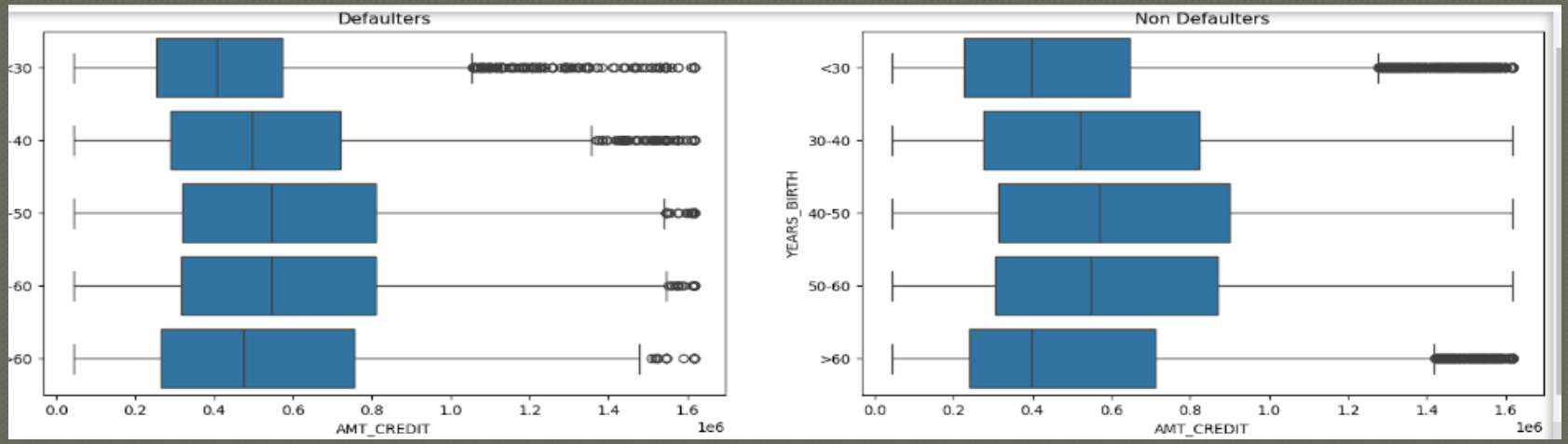
The curve indicates that both defaulters and non defaulters show a similar trend when compared on basis of price of goods.

Bivariate Analysis

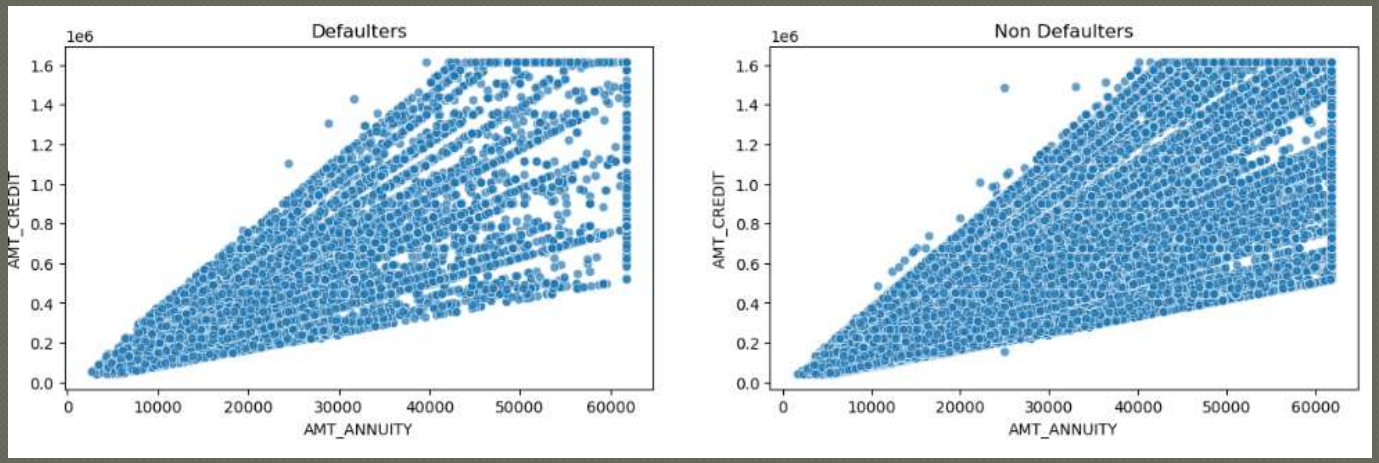
The boxplot indicates the following:

1. In defaulters maximum credit has been taken by people in age group 40-50 and 50-60 followed by people of age more than 60 yrs and then between 30-40 years

2. In defaulters people of age less than 30 years have low credit amount with them.2. In non defaulters category it is evident that the middle aged people i.e in age group of 40-50 years tend to take more loans followed by old people who are 50 and above then followed by young adults in age of 30-40 years.



Annuity and credit

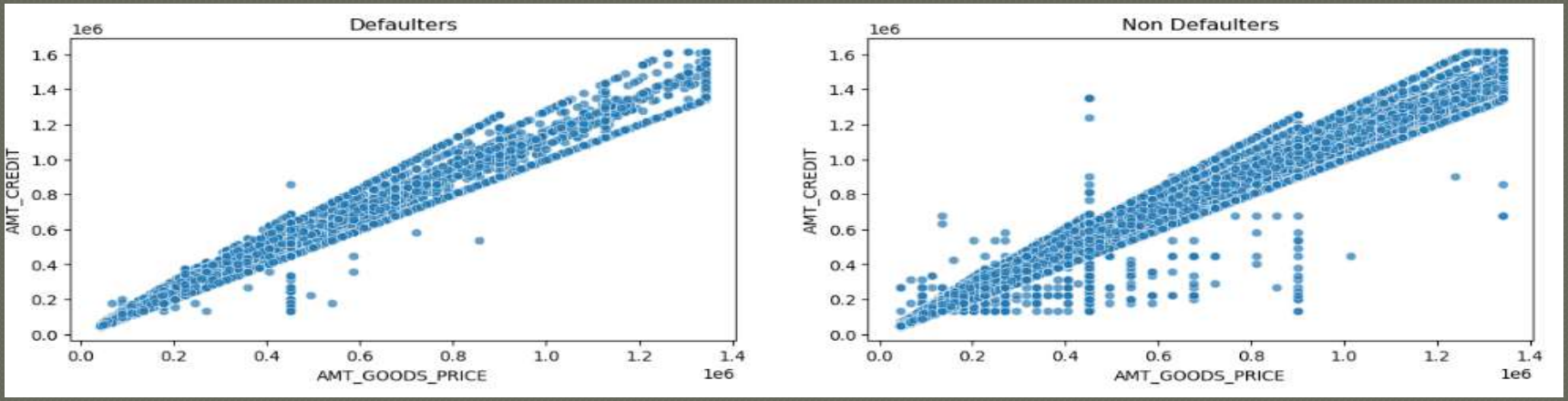


The scatterplots between Annuity and Credit indicate that they have positive correlation and have linear correlation between the two variables. A person with more Annuity assets has better chance to get higher credits and vice versa

Correlation b/w Annuity and Credit for defaulters 0.76

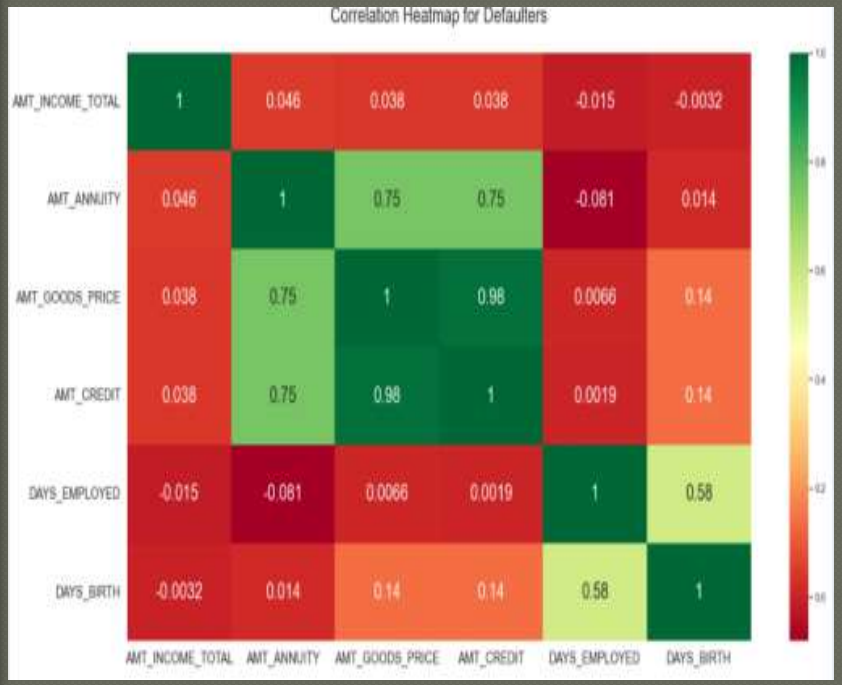
Correlation b/w Annuity and Credit for non defaulters 0.794

Goods Price and Credit



There is very high correlation between Goods Price and Credit for defaulters. In Non defaulters also the correlation is positive and linear but some outliers are also there. The correlation of 98% is quite high for the two variables. So overall we can say that as the goods price increase the credit amount also increases.

Heatmaps



From both the heatmap comparisons we can conclude that there is certain deviation for Goods_Price Vs Income_Total and Credit_Amount vs Income_Total, although there is a high positive correlation between Credit_amount & Goods_Price .

Top 10 Target Variables

The top 10 correlation variables for Defaulters are:

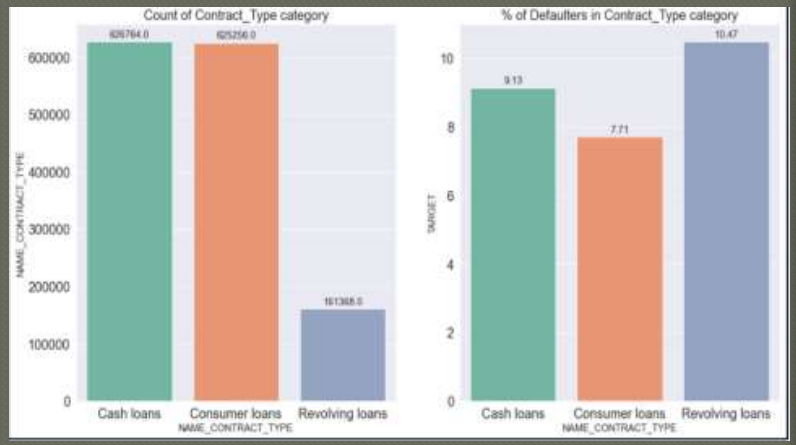
Variable 1	Variable 2	Coefficient
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.99827
AMT_CREDIT	AMT_GOODS_PRICE	0.982566
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.956637
CNT_CHILDREN	CNT_FAM_MEMBERS	0.885484
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.869016
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.847885
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.77854
AMT_CREDIT	AMT_ANNUITY	0.752195
AMT_ANNUITY	AMT_GOODS_PRICE	0.752022
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.497937

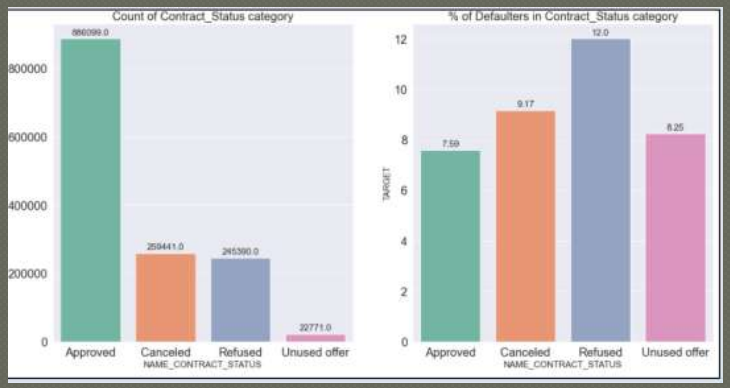
The top 10 correlation variables for Non-Defaulters are:

Variable 1	Variable 2	Coefficient
OBS_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	0.99851
AMT_CREDIT	AMT_GOODS_PRICE	0.98688
REGION_RATING_CLIENT	REGION_RATING_CLIENT_W_CITY	0.950149
CNT_CHILDREN	CNT_FAM_MEMBERS	0.87857
REG_REGION_NOT_WORK_REGION	LIVE_REGION_NOT_WORK_REGION	0.861861
DEF_30_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	0.859371
REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	0.830381
AMT_ANNUITY	AMT_GOODS_PRICE	0.776251
AMT_CREDIT	AMT_ANNUITY	0.771297
REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	0.446101

Merged - Data Analysis

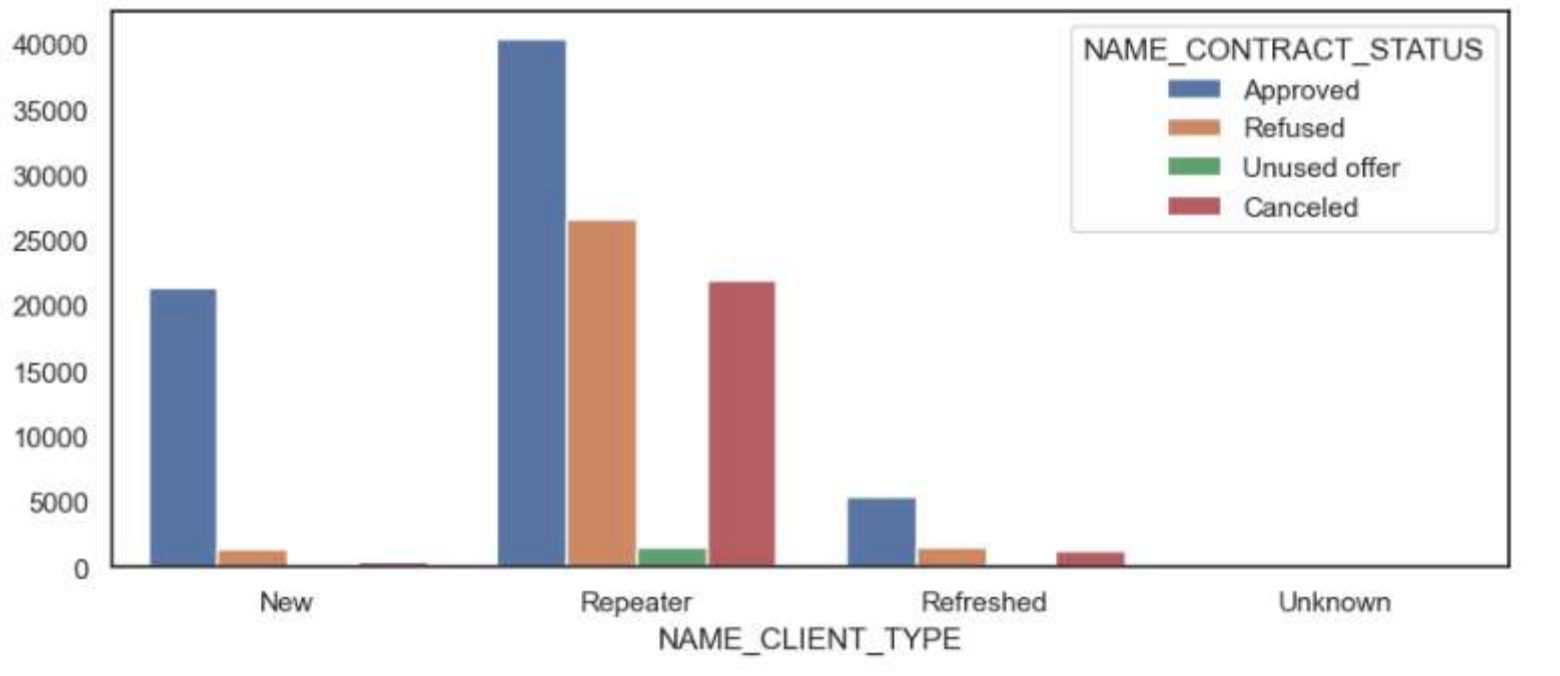
• From the right graph, we can see that the amount of applications for Consumer & Cash loans are pretty close, although we can see that defaulters of current application faced maximum % difficulties for Revolving loans(10.47%) in previous application and minimum % of difficulties in Consumer loans(7.71%).





• From the left graph, we can see that majority of the applications have been approved, although we can also observe that the refused % which is 12% is the maximum % of difficulties faced by current applicants in their previous application & Approved % which is 7.59% is the minimum % of difficulties faced by current applicants in their previous applicatio

Merging the current defaulter dataframe (app_data_1) with the prev_data dataframe



The observation from the above countplot is that there are more repeaters from the previous data are in the current data who have applied for loan previously and are defaulters at present.

Conclusion

Due to data imbalance the outcomes aren't conclusive. But some of the key factors observed are :

Age: • Those between the age of 25-35 have higher chances of defaulting. Considering they may not have a steady a income or a job, it does make them likely to default.

Amount of Annuity: • Clients with middle to higher annuity payments have a higher chance of defaulting. The range of annuity is from 22,000 to 38,000.

ncome: • Customers part of the lower income group, have a higher chance of defaulting. It is likely that lower income groups won't have the capacity to repay loans.