



EDA Case Study



Assignment for Course 3, DS C-32

Siddhee Washimkar and Pratyush Pran



siddheepwashimkar@gmail.com
pratyushpran29more@gmail.com

Agenda

- Approach
- Explanation
- Data Sets
- Reading and understanding the data
- Data Imbalance
- Univariate and Bivariate Analysis of Data
- Data Correlation
- Inferences and Conclusions of Data

Approach

- The following presentation explains how a data scientist can ideally manage risk analysis for the domain of banking and financial services, by analysing data and taking the necessary precautions for the company, to decide on further decisions to be initiated by the employees
- The objective of such an assessment is for the purpose of risk analysis for a bank, which means avoiding the maximum number of losses for the company, and hence to keep the company running on the highest profit possible
- In this case study where the company is a loan providing bank, the past performance & data of the company's clients is used to decide if the bank should approve or refuse loan requests of clients, to be in the highest possible profit
- If the bank approves the loan of clients who do a timely repayment, it will be in profit. If the bank rejects the loan of clients who can do a timely repayment due to lack of data, even then it can miss making a higher profit mark. Hence, EDA becomes important.

Explanation

- We have used the Exploratory Data Analysis (EDA) tool to analyse the patterns present in the available data
- By EDA, some factors and attributes of the data will explain if the customers who are opting to take the loan from the bank are eligible to do the same and if the bank should approve or refuse their loan request
- If the client repays his monthly debt in time, the bank can be profitable. But if the client has a history of not being timely, the bank might go in loss in such cases and might opt to refuse the loan request. Hence we have identified the defaulters on basis of the data and understood this performance.
- Some of the attributes observed from the data for the analysis are - time of repayment of loan in the past, owned assets (car, property), children, income, employment status, occupation, etc.

Data Sets

- We have referred to two data sets in this case study;
 - One is the **applicant's dataset** which contains information of the client at the time of application and has the attributes in his reference. This dataset will tell us whether the client has any payment difficulties.
 - Second is the **client's past performance dataset** which contains information about the client's previous loan data, whether it was been Approved, Cancelled, Refused or Unused and determines the performance



Reading and Understanding the Data

Applicant's Dataset



Data cleaning

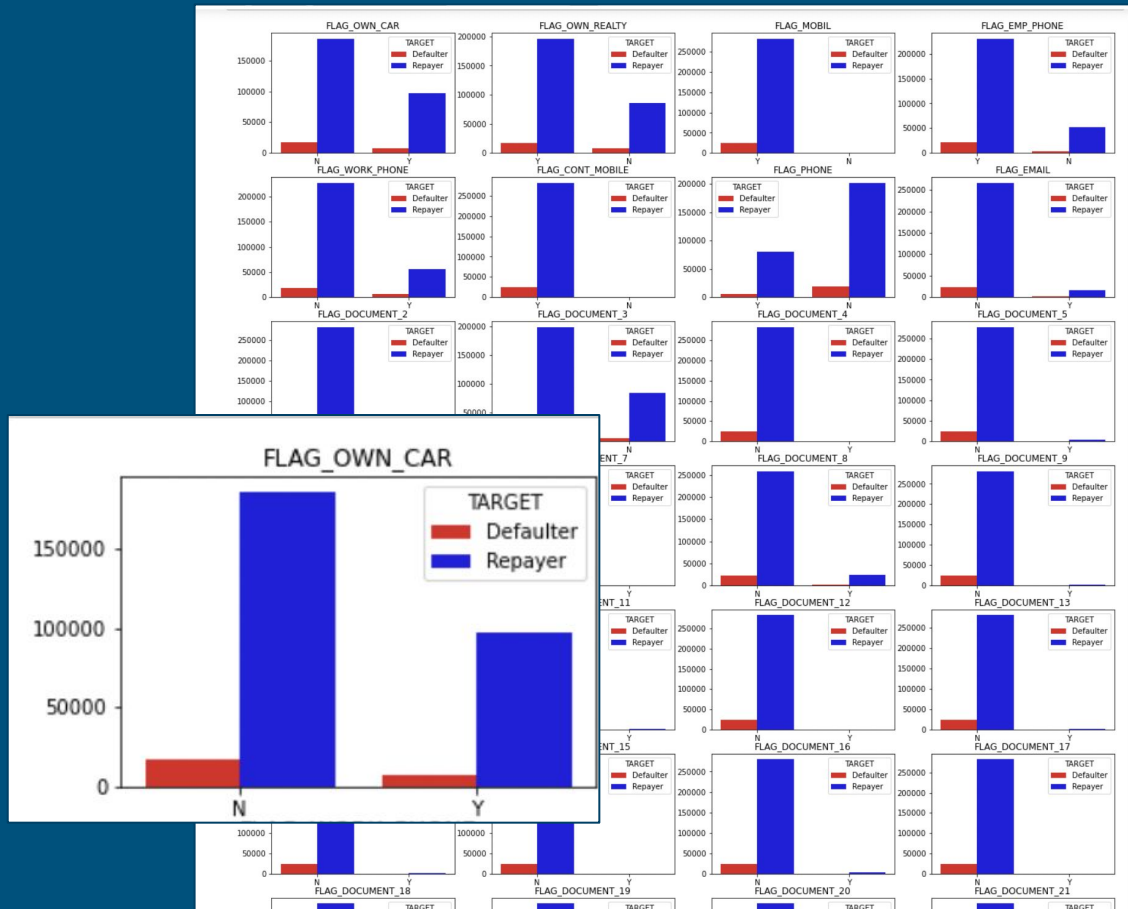
- The applicant's dataset has stored information about **307511** applicants about the attributes mentioned earlier
- To clean the dataset for further analysis, we align it by cleaning (dropping) the columns having missing values and replacing the few necessary columns for analysis which have some missing values with average values, i.e. median values, that can be referred to for our analysis by standardising the dataset.
- We will drop all the columns having missing values of more than 50% as they would not be able to contribute to our analysis due to the high number of missing values and have high chances to hamper our analysis. There are 41 such columns to drop.
- We will also remove columns having missing values of more than 15% that are **not reliable** for our TARGET columns as these are only additional unrequired data that does not make sense for our analysis. There are 8 such columns.
- We will also entirely drop all the columns which are unnecessary for our analysis, to have a focused dataset, and replace the values in the dataset by more understandable and reliable terms to use for a better understanding and reference.

Manipulation

- Furthermore, after dropping and standardisation of columns in one dataset, we will use data manipulation technique to normalise the numerical data into desired and readable integer lengths, in terms of their units. Like for example we have converted the price range columns into lakhs in this case which is easier for everyone's reference and readability.
- Similarly, for the columns related to Days where we have the age of a client in the form of dates, we have converted the column reference into the unit of numerical year.
- After the entire cleaning and manipulation of the data, it leaves us with 53 columns finally.

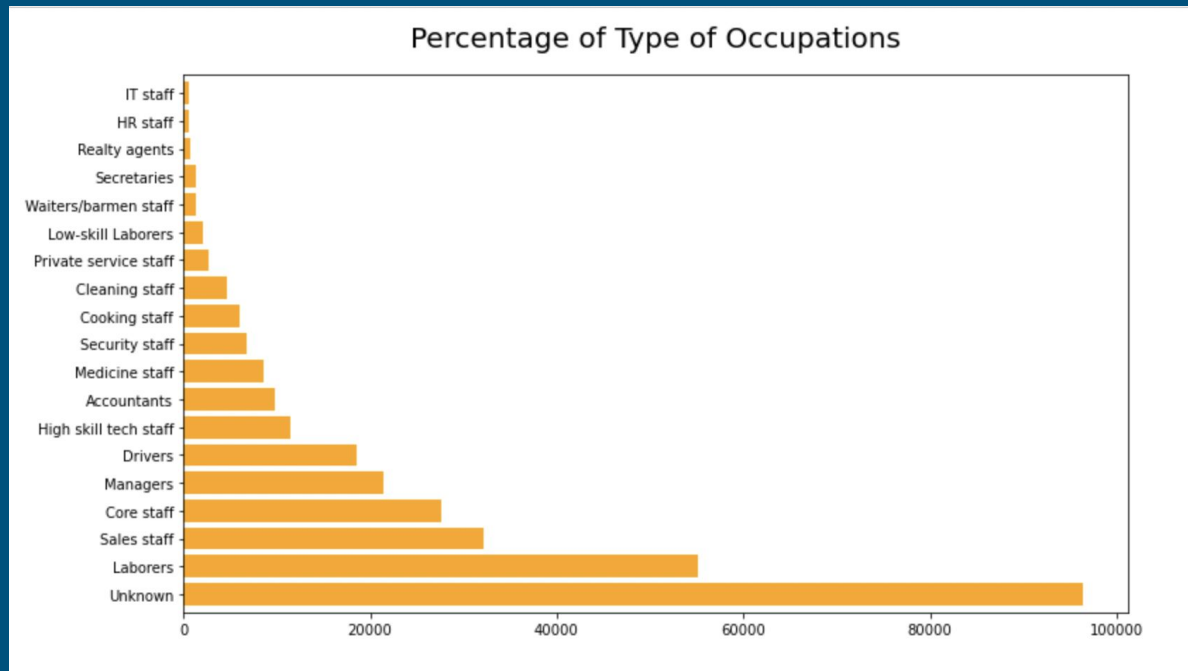
FLAG Target columns

For our further analysis where the applicant's dataset contains columns having all the target columns with the relevant information, let's analyse the information by a bar plot as shown on the right, where N denotes a 'NO' and Y denotes 'YES' for both defaulters and repayers. This gives us an idea of the owned assets the clients have, other attributes and documents submitted, etc. with the bank, explaining the behaviour and present data of both defaulters and repayers at a glance. We can then drop these unnecessary columns which are irrelevant to our dataset.



Occupations in Dataset

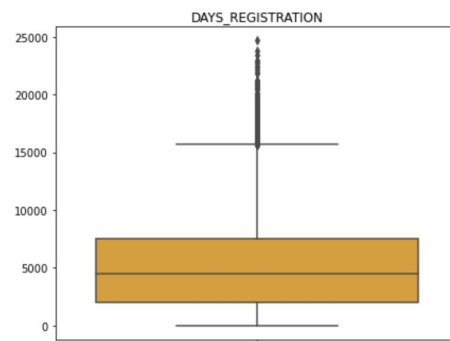
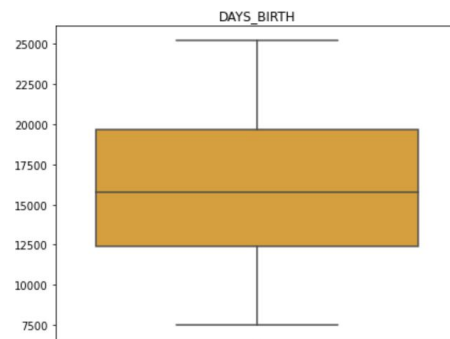
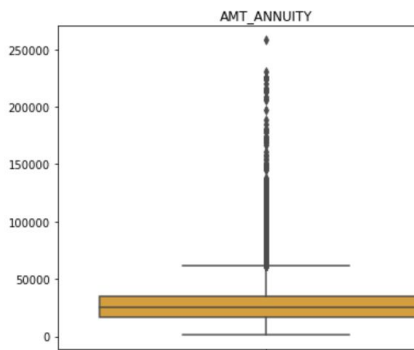
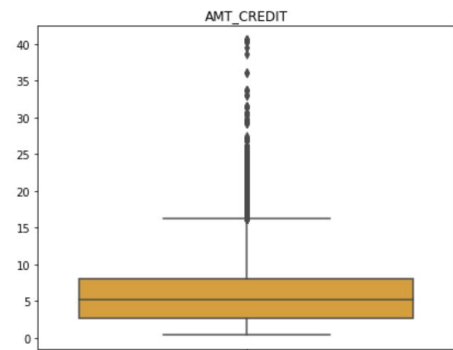
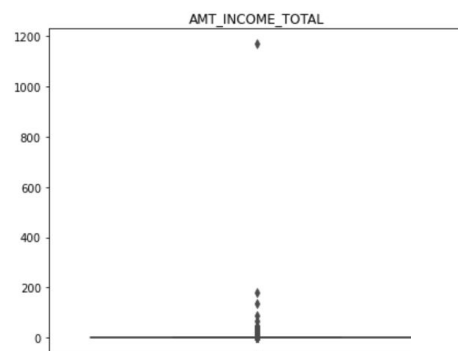
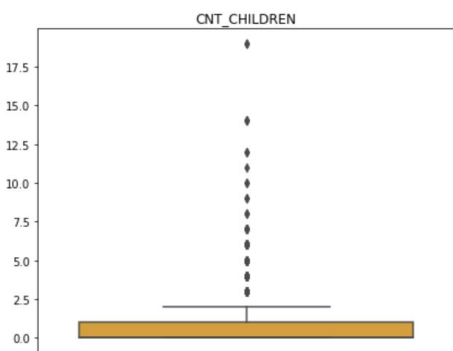
Studying the occupation of the clients becomes very important as it gives us an idea of our clients' industries and while we know how profitable those industries are, it will also give us a rough idea about the overall income types they can fall into. As we can see in the bar plot on the right, most of the client's occupation is unknown and missing in our dataset, but the rest come from the following other occupations starting from lowest being IT staff, to the second highest being Laborers after the unknown category.



Identifying Outliers

- Outliers are values that are much beyond or far from the next nearest data points, that is the data values which are far from the rest of the dataset.
- Outliers can affect the average numerical analysis of any data, as that one value of the outlier being away from the others can affect the range of the dataset and that is how data can be misunderstood.
- Let us identify outliers in this bank dataset as seen in the next slide in the form of boxplots;

As you see in the boxplots here, these are the columns from the dataset which have outliers in them. If we take the income plot on the left for instance, we see that one of the person's income is near the 1200 mark, which is way higher than the entire income range of the other clients that range between 0-200.



See the various univariate and bivariate outliers in these graphs which give us an idea of how the data is scattered.



Reading and Understanding the Data

Client's Past Performance Dataset
(Previous Application)

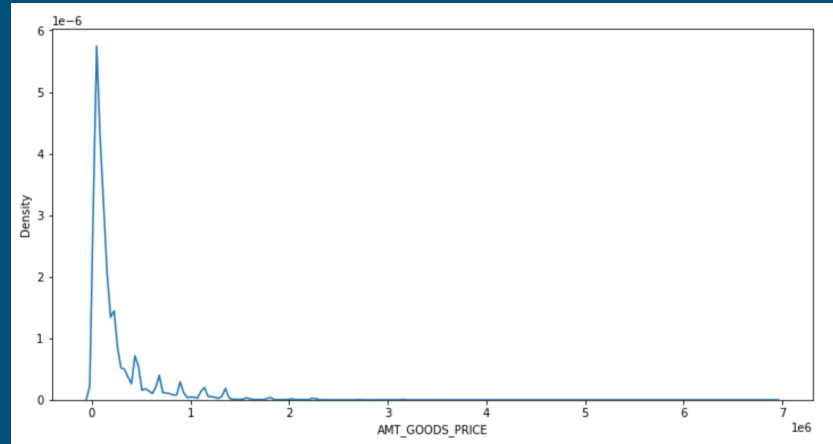
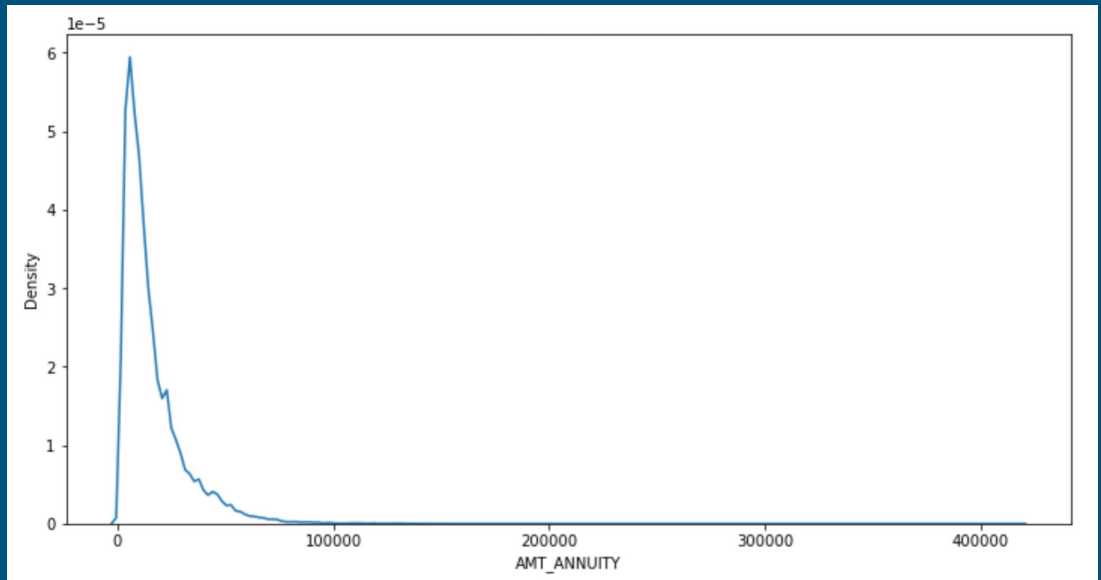


Repeated Steps for Data Cleaning and Manipulation

- Like for the previous applicant's dataset, we will perform the same steps for cleaning this dataset and standardize the values, for us to then read the entire dataset in a better way converting the raw data to a readable data solely for the purpose of readability
- We will similarly drop the null and missing values of more than 50% and the ones that are not relevant for our study.

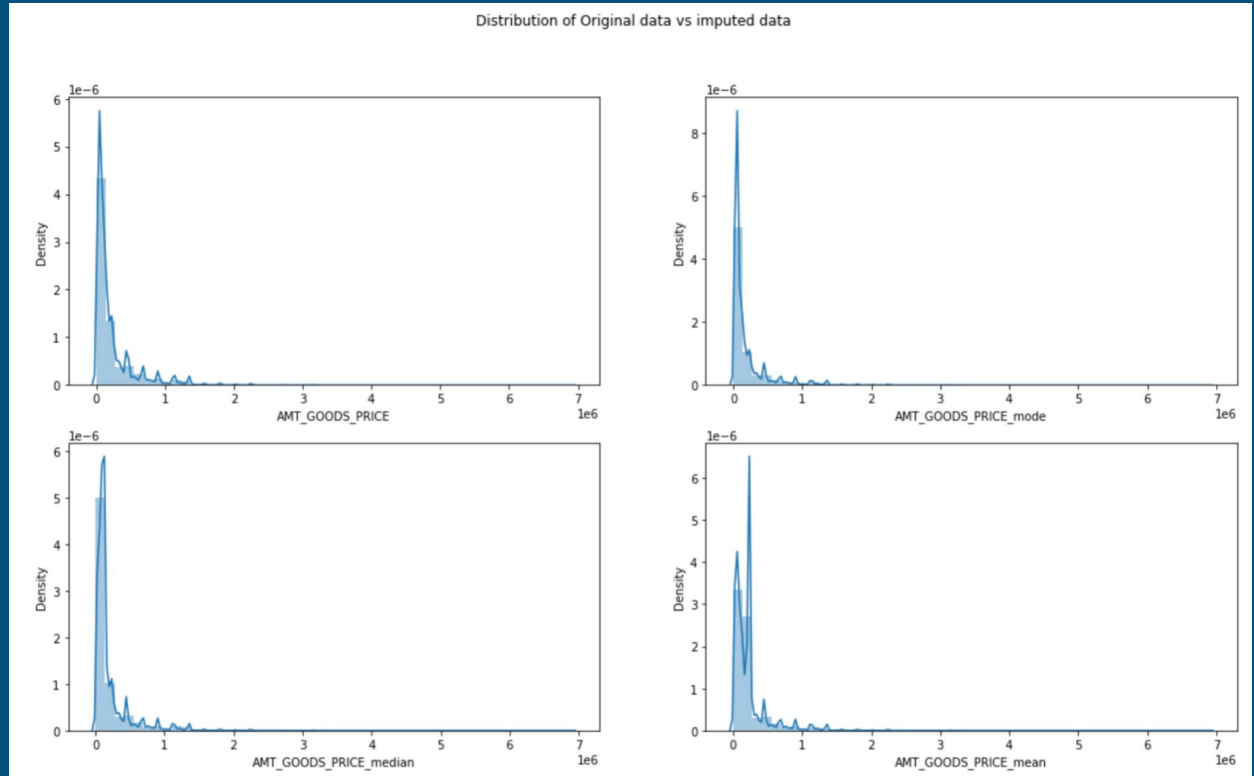
Continuous Variables

In the previous application dataset, we have few continuous variables. To impute null values in continuous variables, we have plotted the distribution of the columns in a Kernel Density Distribution plot and used median values for skewed distribution, and mode for preserved pattern distribution as you can see on the right. Here we have done this in the case of the variable 'Amount for Annuity' and 'Amount for goods price'.

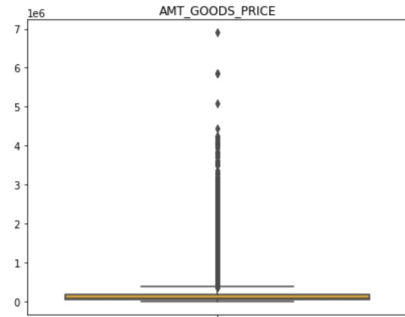
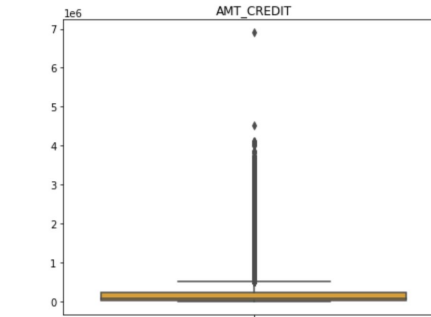
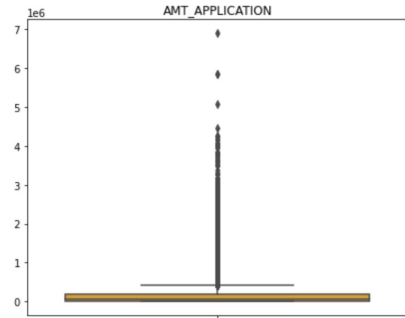
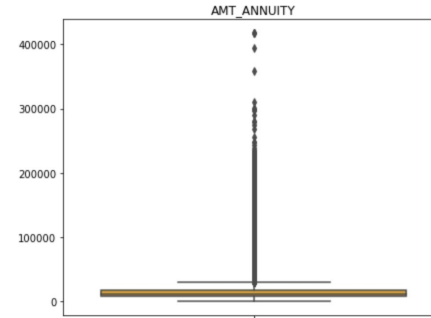
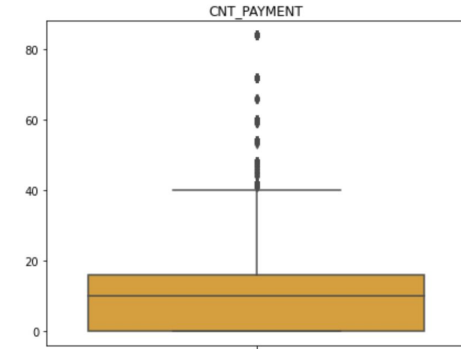
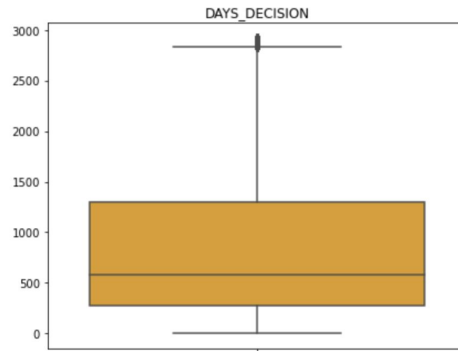
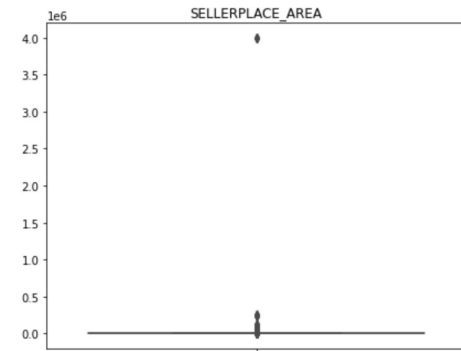


Distribution of Original Data vs Imputed Data

The graph shows the difference between the original and imputed data values through the Kernel Density Distribution plot. As you can see that we have plotted the same graph taking its mode, median and mean values. The original distribution is closer with the distribution of data imputed with mode in this case. Thus we will impute mode for missing values.



Identifying Outliers



As we can see in the graphs, every graph has outliers as there are some data points which lie far away from the entire data set. For example, in the graph of Sellerplace_area above, only one data point lies near 4.0 while the other data range lie between 0 - 0.5. This is a very big difference and hence these points are called outliers for a dataset, which may hamper our analysis.



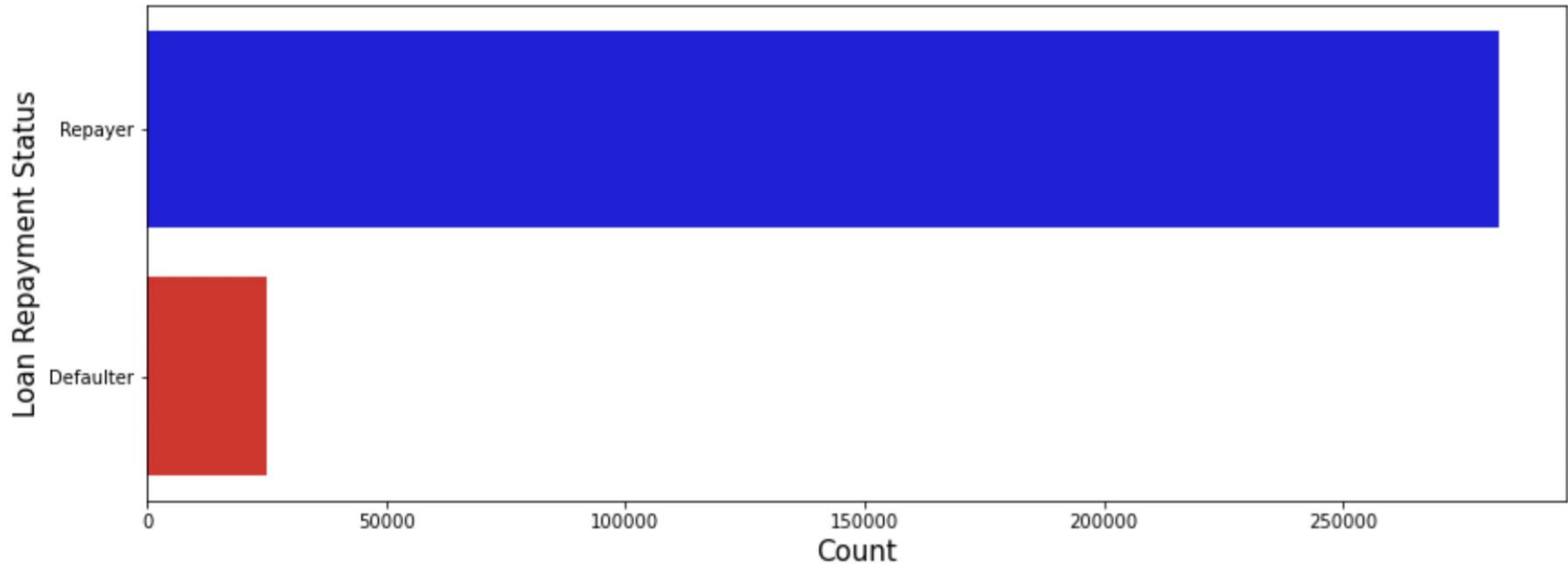
Data Imbalance



Data Imbalance

- An imbalanced dataset is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance. Imbalanced classifications pose a challenge for us to make a predictive analysis of the data and results in hypothesis with poor predictive performance. This is a problem because our analysis and prediction of the data has high chances of being faulty.
- The ratio of data imbalance for our current dataset is approximately - **11.39 : 1**
- Further on, in the next slides we will study this imbalance of data with the help of different types of graph plots, for a visual representation of data imbalance.

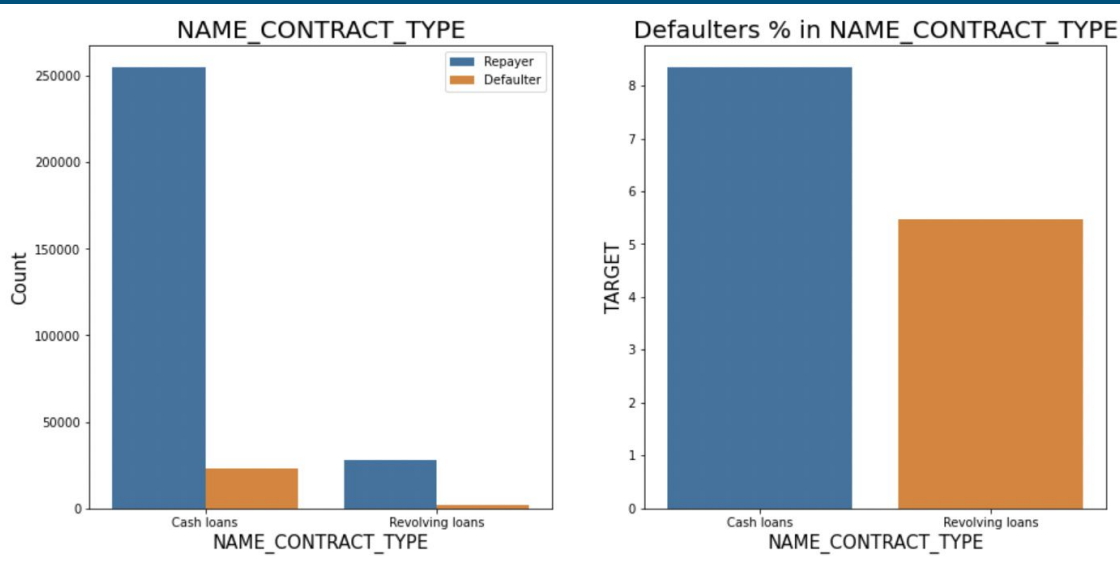
Imbalance Plotting (Repayer Vs Defaulter)



- According to our graph above, we can come to a conclusion that the percentage of repayers in this case is **91.93%** whereas the defaulter percentage is **8.07%**.
- Hence the imbalance ratio with respect to Repayer and Defaulter will be approximately given as **11.39** :

Univariate and Bivariate analysis

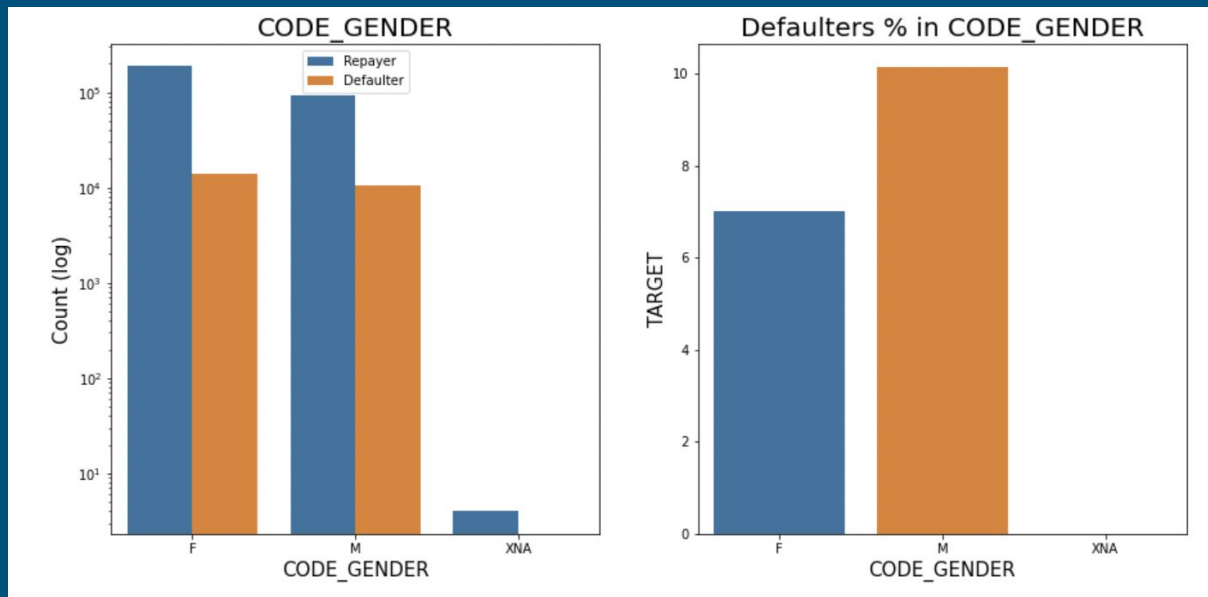
- Univariate analysis deals with analysing a single column and variable at a time. The concept of univariate analysis is divided into ordered and unordered category of variables which we shall see in the further graphs.
- On the other hand, Bivariate or Multivariate analysis of a dataset involves the analysis of two variables or columns in the data set at a time.



In this second graph on the right, by doing a segmented univariate analysis, we figure that the defaulters percent for cash loans is 8% while that of the defaulters with revolving loans is around 5%. Similarly, in the first graph we also see the difference between defaulters and repayers.

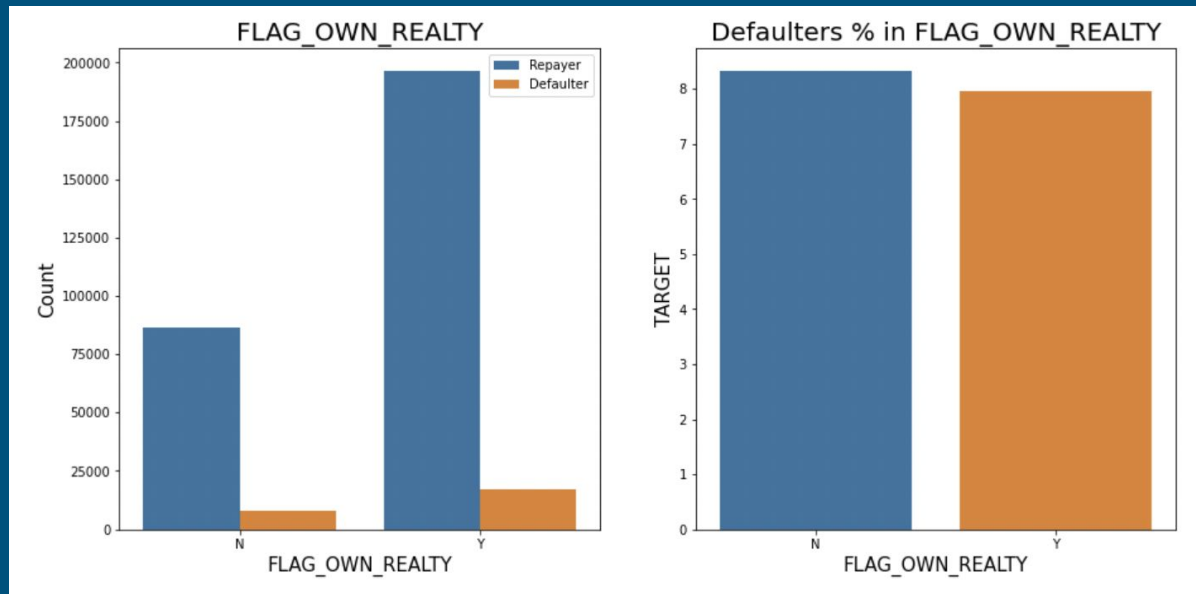
Checking type of Gender on loan repayment status

- In the case of repayers, the number of females is higher than males.
- Even in the case of defaulters, number of females is higher than males.
- This is because the entire number of females for the loan applications is more than males.
- When we look at the percentage, the overall male percentage for the defaulter is higher than females.



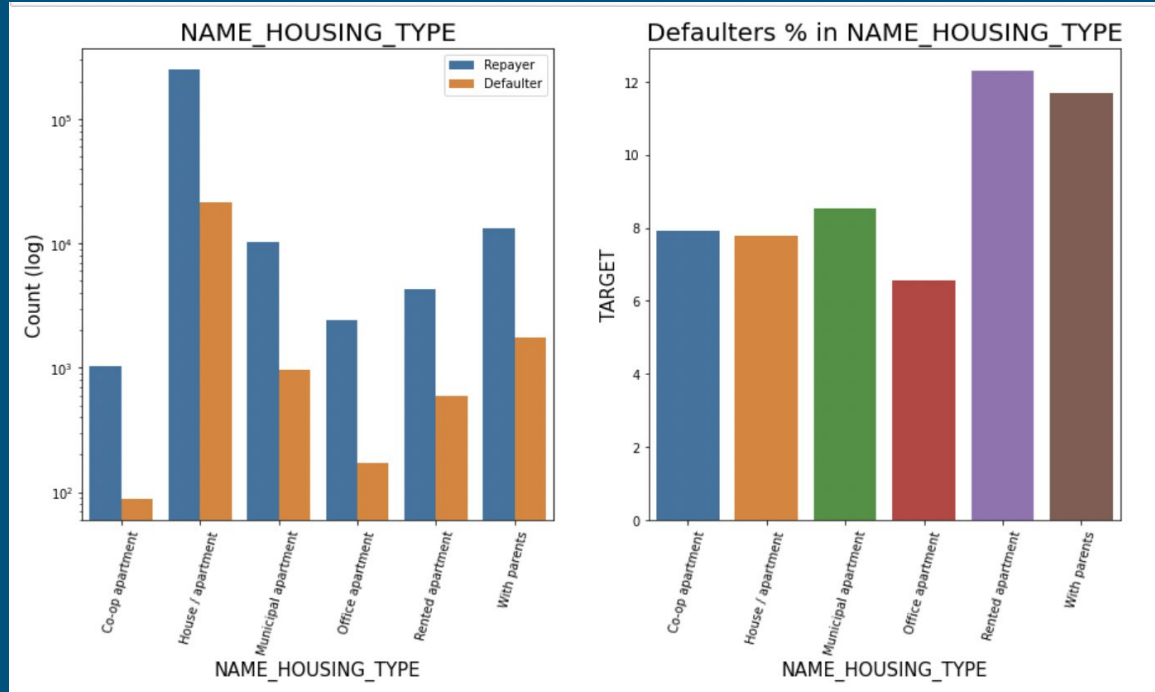
Checking if owning real estate is related to loan repayment status

- The number of defaulters who own a real estate property is actually higher than those who do not own a real estate property.
- But also, the number of repayers is very high for the applicants who own a real estate property.
- The overall picture shows us that owning a real estate property does not affect the repayment status of loan.



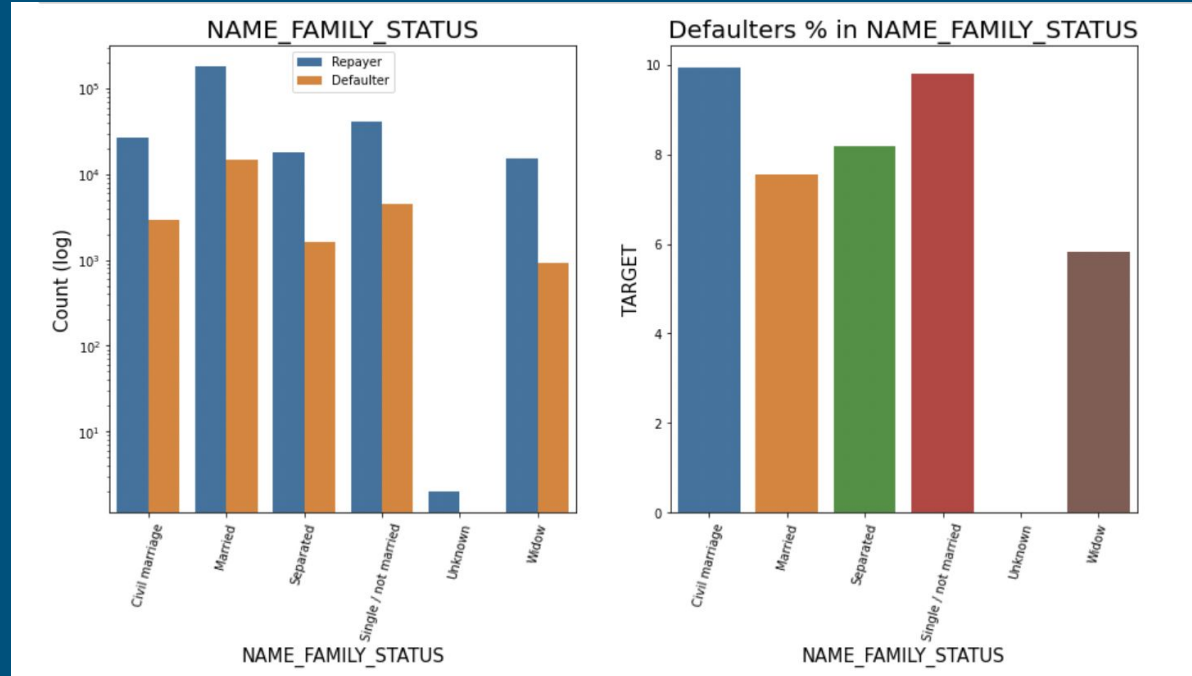
Analyzing Housing Type based on loan repayment status

- Clients who own a office apartment of their own have the lowest defaulter rate.
- Clients having a rented apartment have highest default rate and it is risky for the bank to give them a loan.



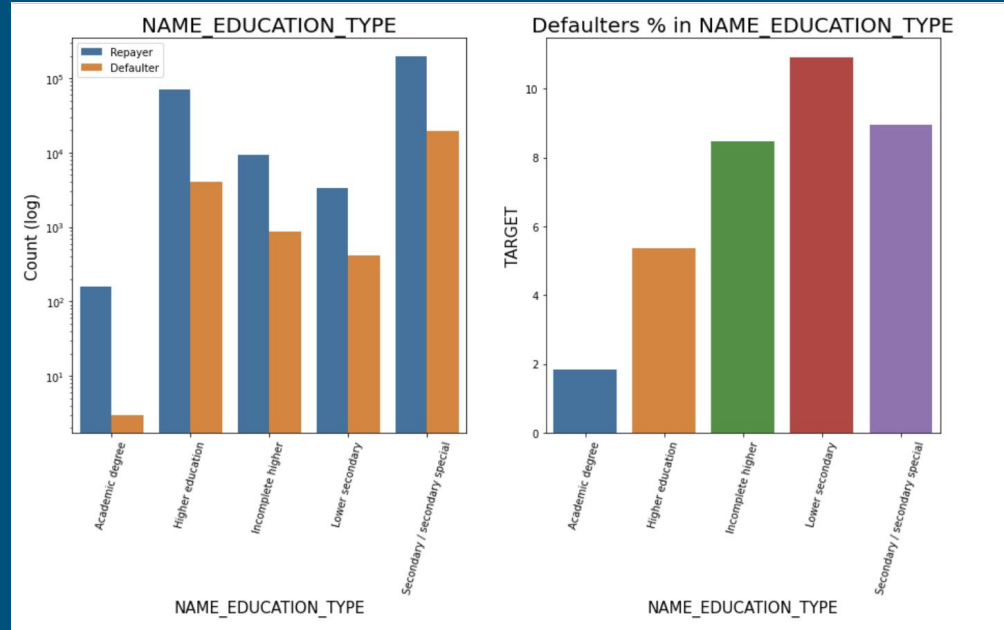
Analyzing Family status based on loan repayment status

Clients who have done a civil court marriage and who are not married have the highest default rate as compared to the other clients who are divorced, married or are single.



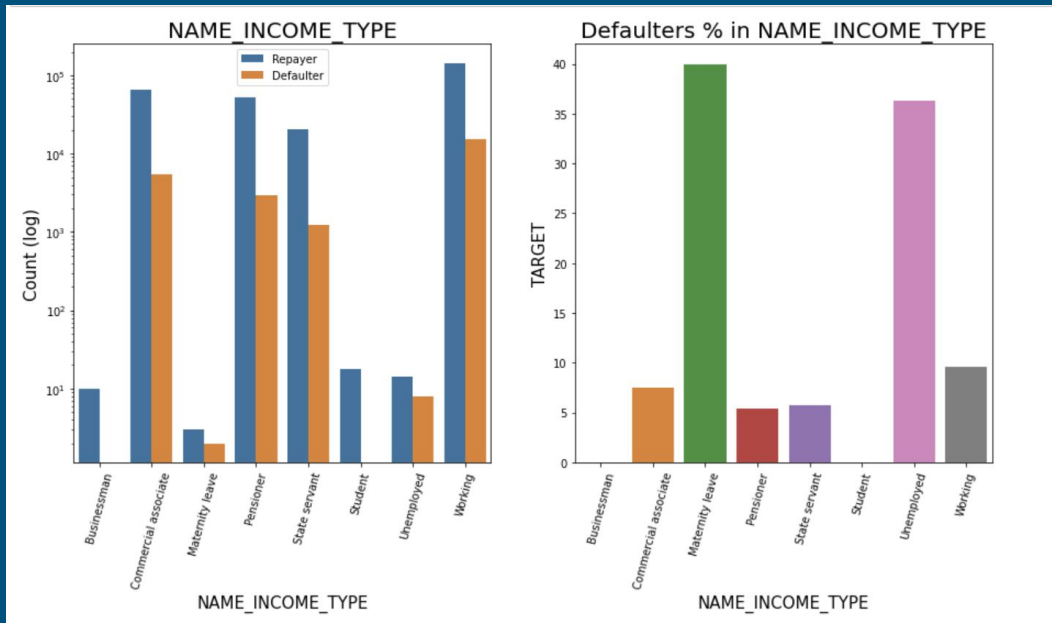
Analyzing Education Type

It is safest for the bank to provide loans to the clients who have a high education with an academic degree, followed by clients who have taken higher education. Clients who have not taken their education seriously and only hold a lower secondary education default the highest.



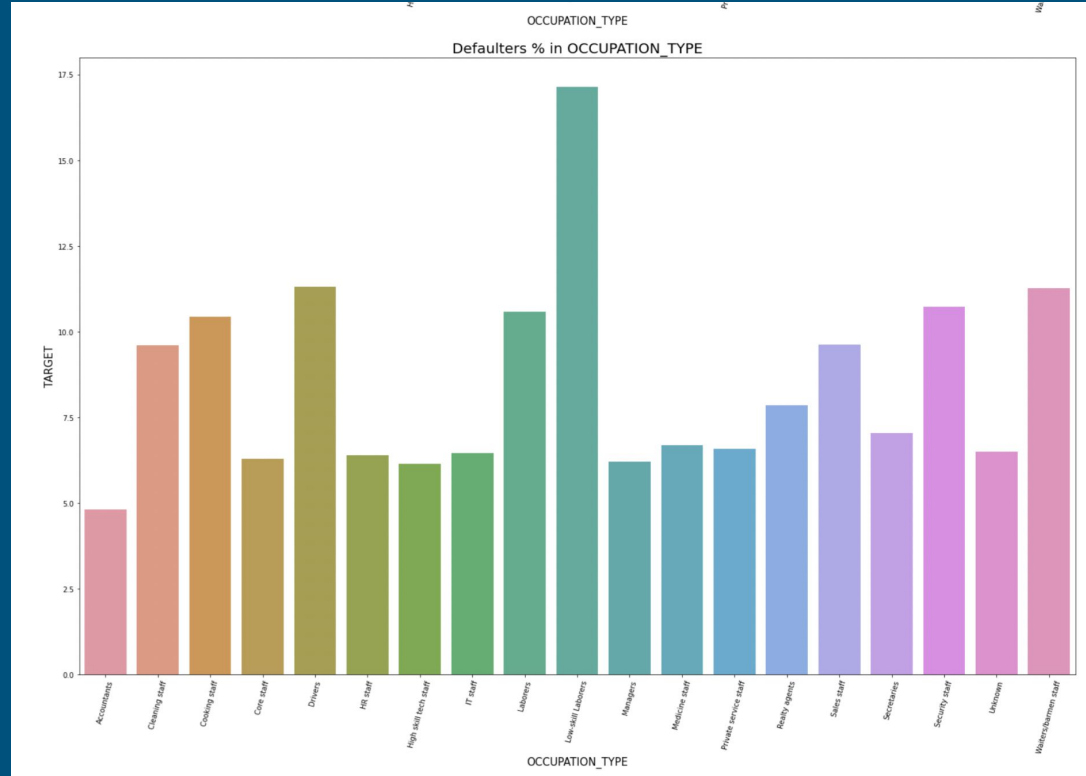
Analyzing Income Type

Clients who have a government job and who would be receiving a pension are safe for the bank to provide loans followed by clients who serve their state. The most risky clients are the unemployed and clients on a maternity leave as they do not know if they will be continuing their job.



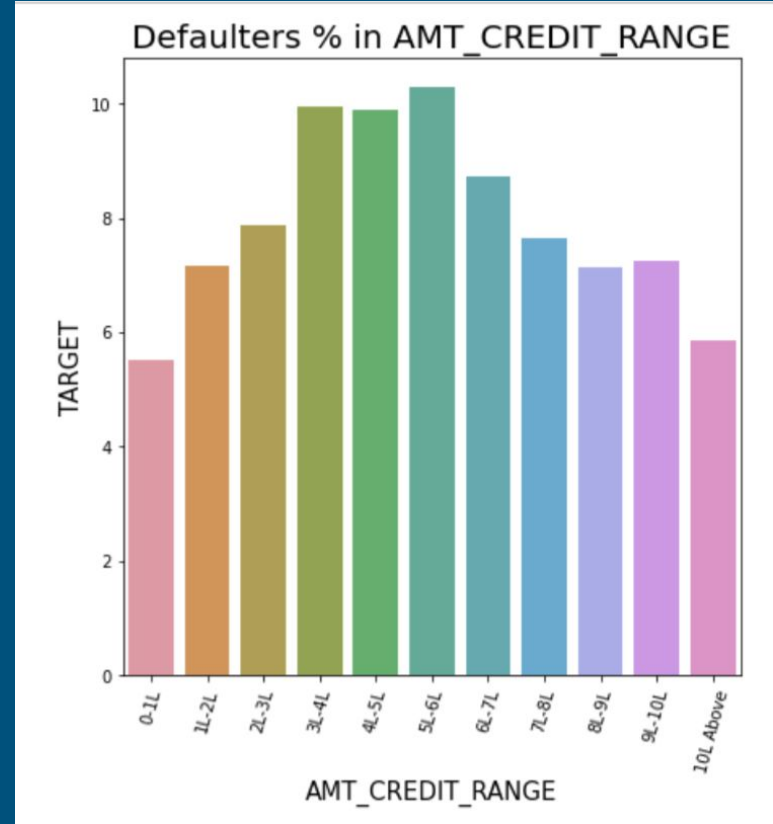
Analyzing Occupation Type

Again, low skilled labourers who do not have an education are at a highest risk of loan repayment to the bank. The safest clients are accountants followed by a highly skilled technology staff, Managers, HR staff and IT staff. Drivers also tend to default a lot.



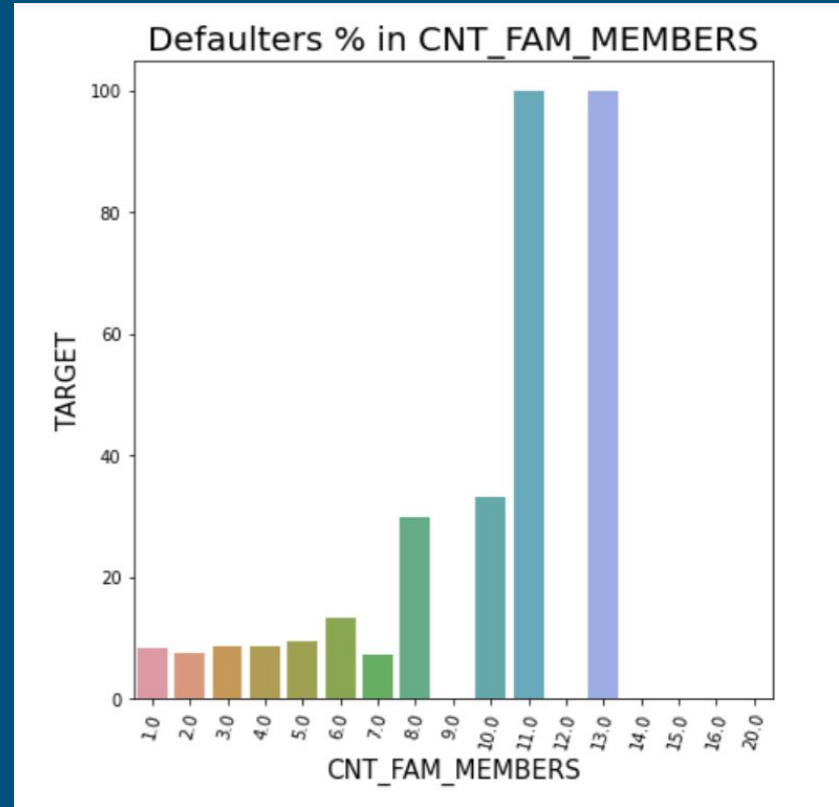
Analyzing Amount Credit Range

Clients which apply for a low amount of loan from 3-6L range have a high chance of defaulting compared to the ones who have a long term loan with a high principal amount of 10L and above. If the loan is also very less below 3L then the chances of default are low.



Analyzing Number of Family Members

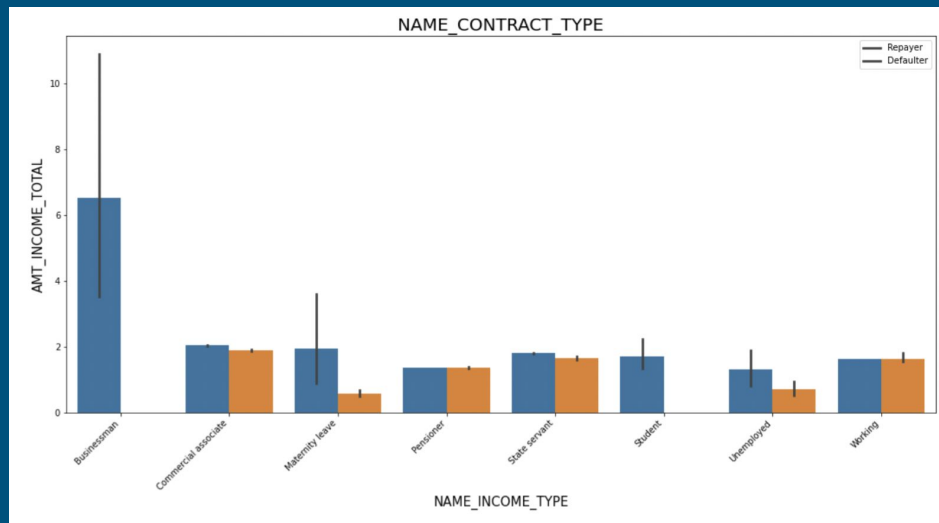
Clients who have a high responsibility of large number of family members at home have a high default rate, compared to the ones who have a responsibility of lower family members at home. Here the clients with 11-13 family members are highest defaulters. Clients with 7 family members are lowest defaulters who are in the mid-range.



Categorical Bivariate or Multivariate Analysis

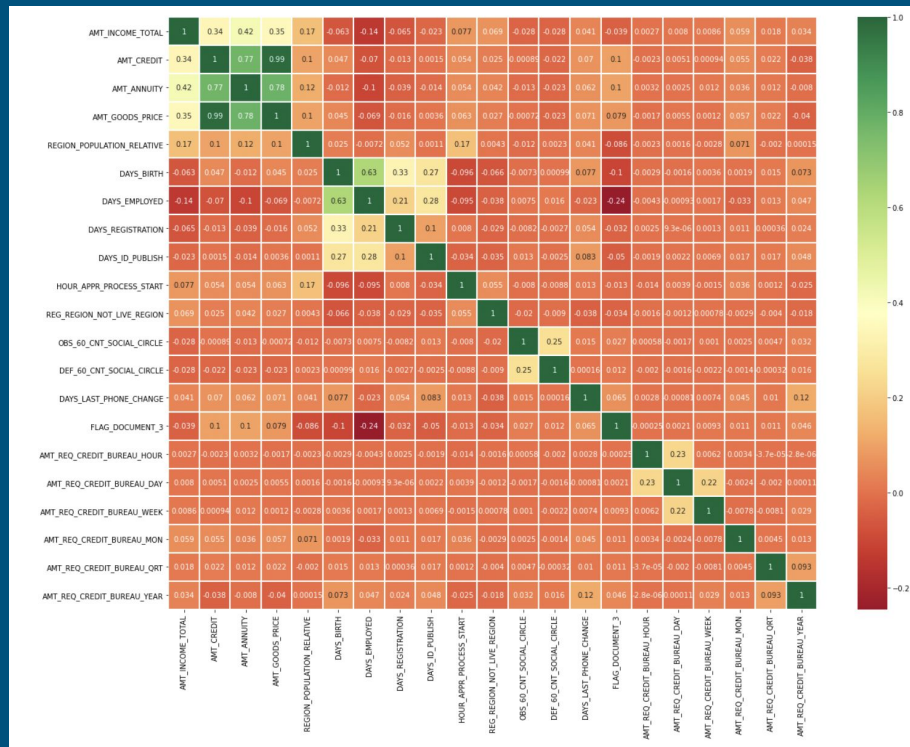
Income type vs Income Amount Range

Businessman who have a high income amount tend to default the least due to their high income range of 6.5 lakh and above. There are also the least number of defaulters in the businessman category. After businessman, students who opt for educational loans also default the least.



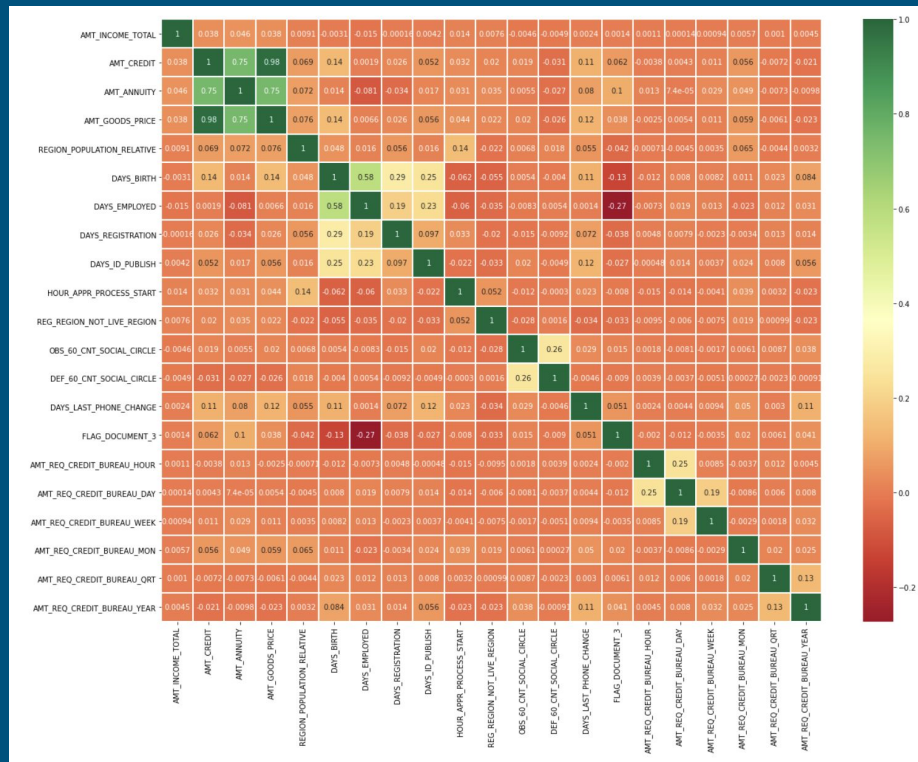
Linear correlation among Repayers

The graph on the right shows us the linear correlation of repayers across various variables where the red colour indicates the lowest values and the green leads to the highest values. The entire graph highlights values ranging from 0-1. The client who has been employed for the least days has a very high fault rate of not providing documents. The correlation is highest in the variables of the amount of goods price and the credit amount which is at 0.99 compared to the lowest between Birth days and published ID which is at 0.27. We can also see that repayers have a high correlation with the number of days employed.



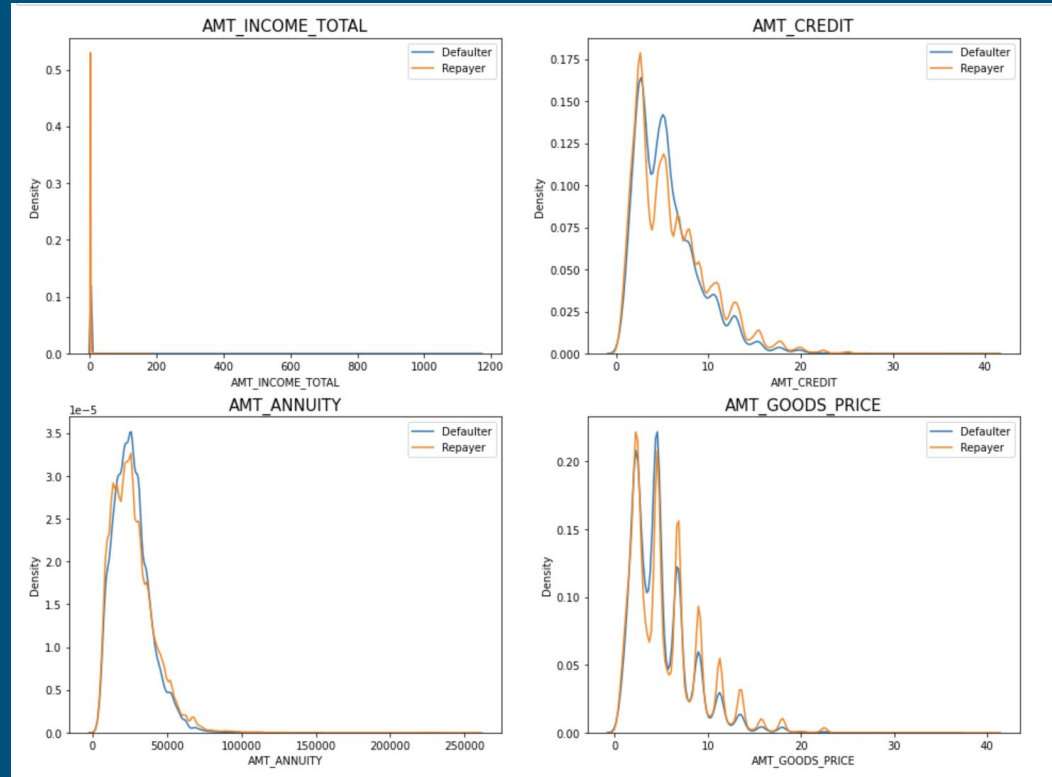
Getting the top 10 correlation for the Defaulter data

In this graph, we see high correlation again between the credit amount and amount of the goods price at 0.98. The credit amount is highly correlated with the goods price amount which is same as the repayers. Loan annuity correlation with credit amount as slightly reduced in defaulters (0.75) when compared to repayers (0.77).



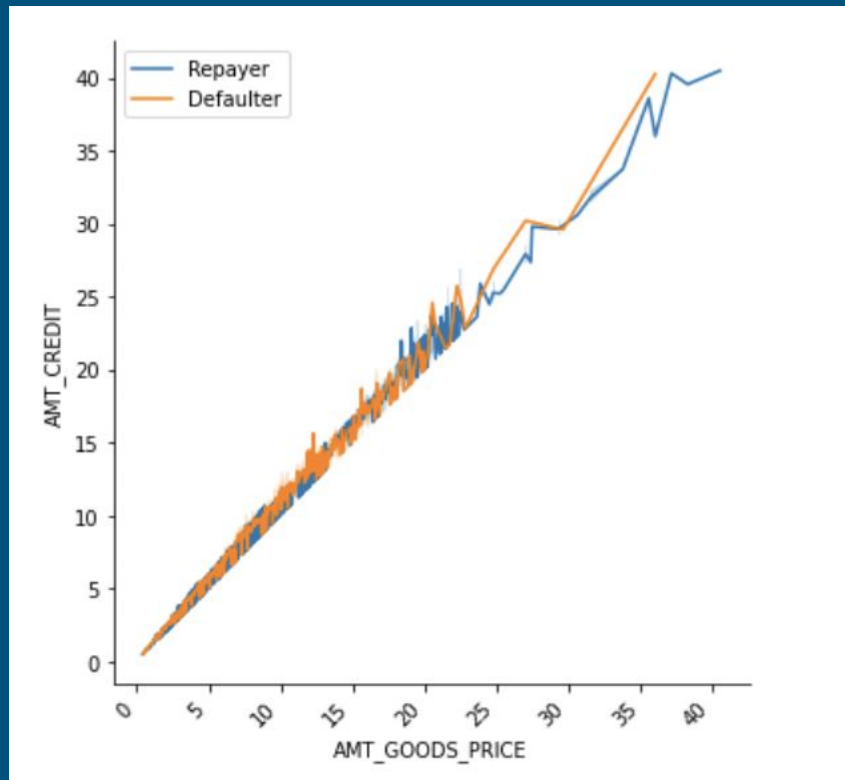
Numerical Univariate Analysis - Plotting numerical columns related to amount to see density

These graphs give us an idea of the numerical distribution of columns by two colours and hence gives us a comparison for the correlations between the defaulters and repayers side by side. The repayers and defaulters distribution overlap in all the plots. Hence we cannot use any of these variables to make a decision.



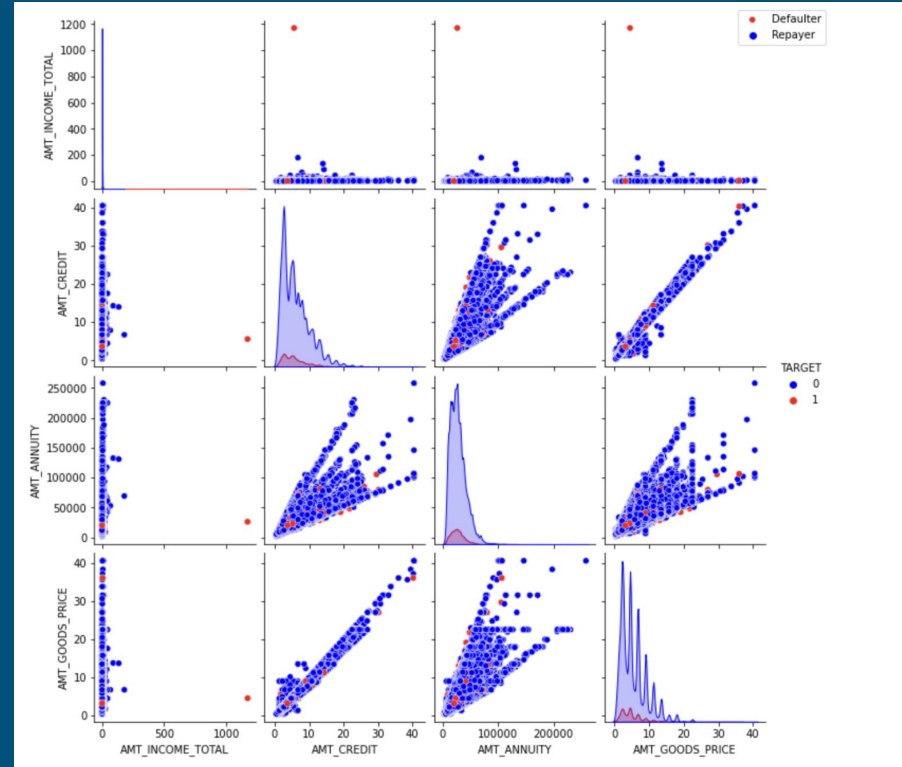
Numerical Bivariate Analysis - Relationship between Goods price and credit

The relationship between the goods price and credit amount is linear for both the defaulters and repayers. When the credit amount goes beyond 30 lakhs there is a steep increase in defaulters.



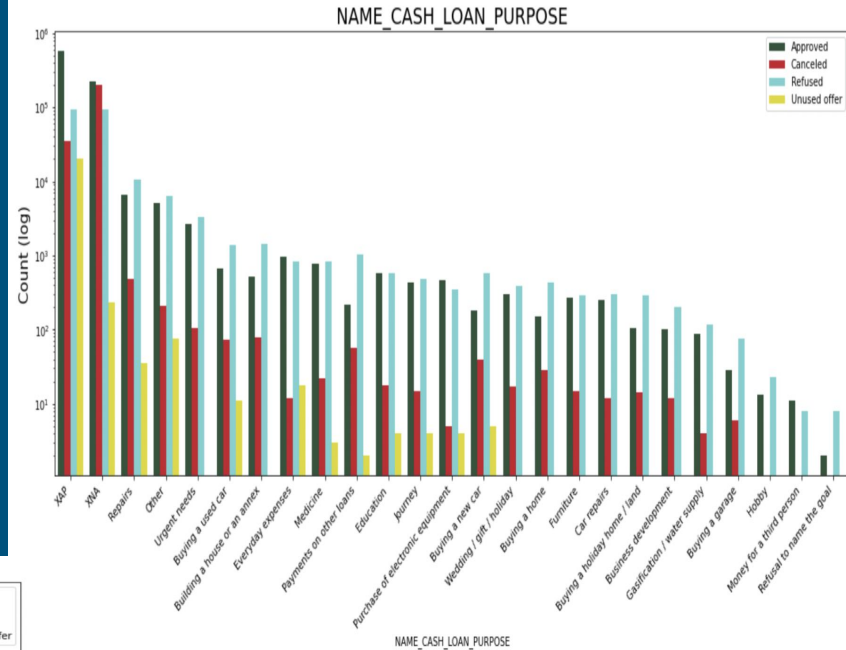
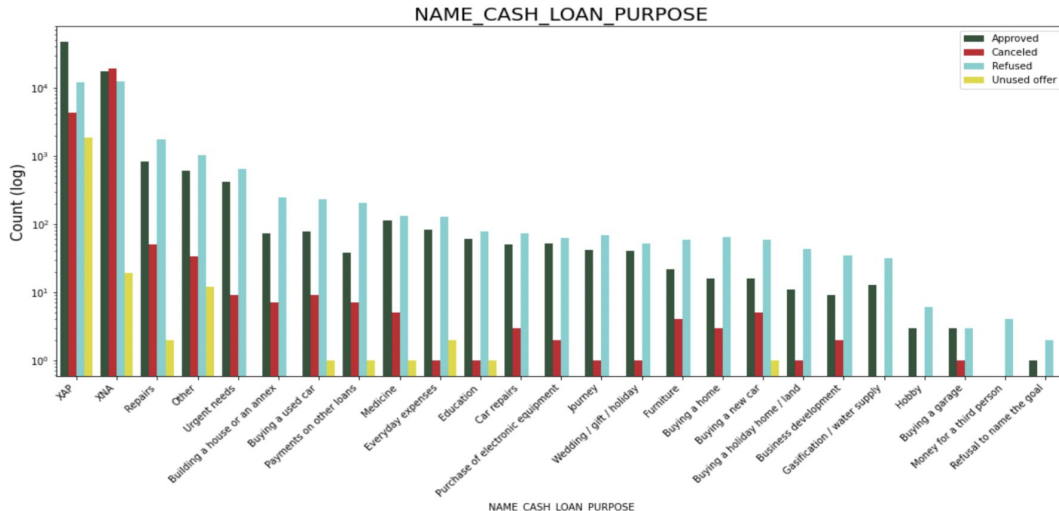
Pairplot between amount variable to draw reference against loan repayment status

- When Annuity amount is less than 15,000, and good price amount is less than 20 lakhs, there is a lesser chance of defaulters.
- Loan amount (AMT_CREDIT) and goods price (AMT_GOODS_PRICE) are highly correlated as based on the scatter plot when most of the data are consolidated in the form of a line.
- There are very less defaulters for AMT_CREDIT which is less than 20 lakhs.



Merged Dataframe analysis

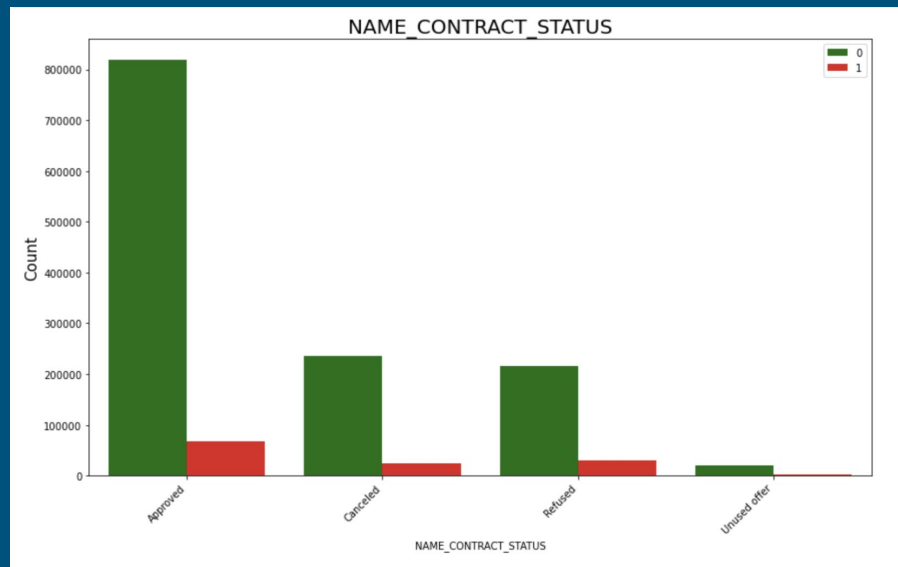
- The loan purpose has high number of unknown values (XAP, XNA)
- Loan taken for the purpose of Repayers looks to have the highest default rate



A huge number applications have been rejected by the bank or refused by client which are applied for a repayer or Other clients. From this we can infer that a repayer is considered as a high risk by the bank. Also, either they are rejected or the bank offers them loan on a high interest rate which is not feasible by the clients and they refuse the loan.

Contract Status - business loss or financial loss

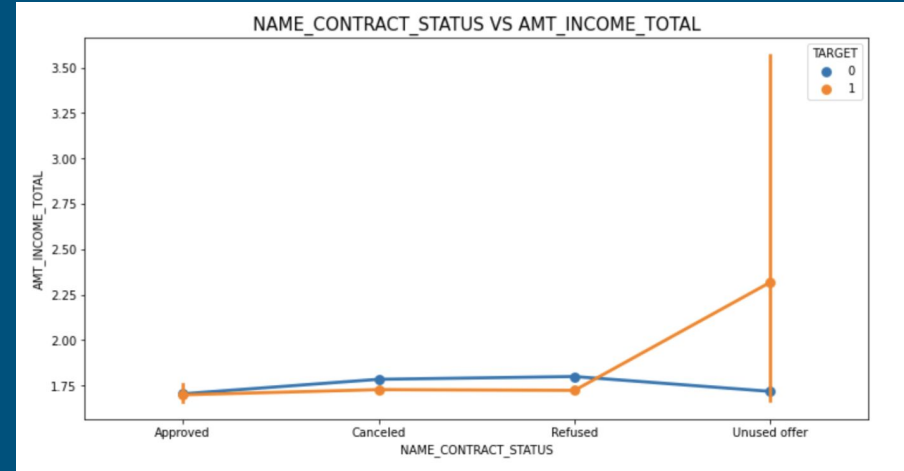
- 90% of the previously cancelled clients have actually repayed the loan. Revising the interest rates would increase business opportunity for these clients
- 88% of the clients who have been previously refused a loan have paid back the loan in the current case
- Refusal reason should be recorded for further analysis as these clients could turn into potential repaying customer



Relationship between income total and contact status

Inferences:

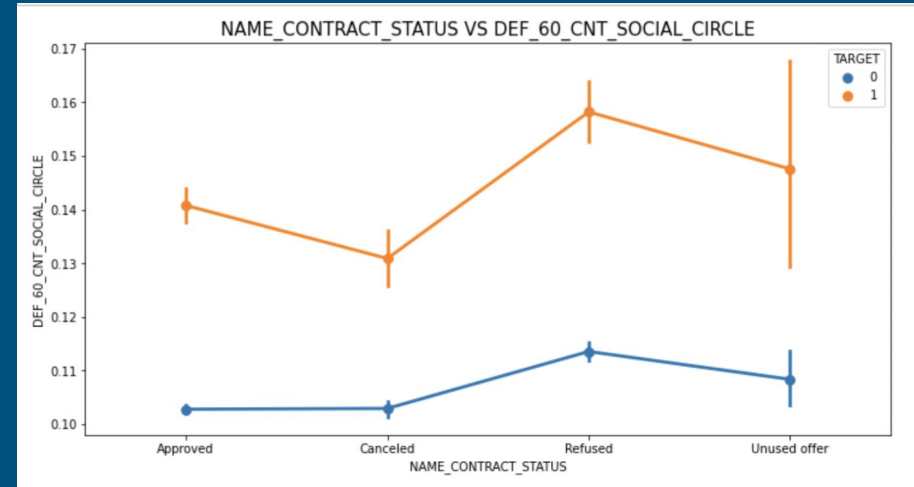
- The point plot shows that the people who have not used the loan offer earlier have defaulted even when their average income was higher than others



Relationship between people who defaulted in last 60 days

Inferences:

- Clients who have an average social circle of 0.13 or higher their DEF_60_CNT_SOCIAL_CIRCLE score tends to default more and thus analysing a client's social circle could help in disbursement of the loan.

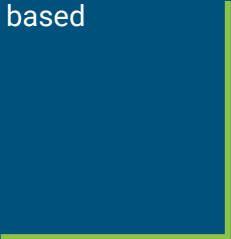




Conclusions



After analysing the datasets, there are few attributes of the client which the bank would be able to identify to denote if they will repay the loan or not. The analysis is based upon the contributing factors and categorization



Decisive Factor whether an applicant will be Repayer

- An applicant should have less defaults in his Academic degree
- Looking at the income range, students and businessmen tend to make the least defaults and have no defaults
- The client which lives in region 1 makes least number of defaults and is safest for the bank to go ahead with
- Clients based from the organisation types of Trade type 4 and 5 and Industry type 8 have defaulted less than 3%. Hence this is industry is a safe one for the bank.
- People above age of 50 have low probability of defaulting
- Clients with 40+ years of work experience have less than 1% of default rate. High experienced person default less.
- An applicant with Income of more than 700,000 is less likely to default
- Loans applied for a Hobby, buying a garage are being repayed the highest
- People with 0 - 2 children tend to repay the loans compared to the ones who have more than 2 children

Decisive Factor whether an applicant will be Defaulter

- Men are at a relatively higher default rate
- People who do a civil marriage or who are single and unmarried default the highest
- People with Lower Secondary & Secondary education are at the highest default rate
- Clients who are either on a maternity leave or unemployed, default the highest
- People who live in the region of rating 3 have highest defaults
- Bank should avoid approving loan of low-skilled laborers, drivers and waiters/barmen staff, security staff, laborers and cooking staff as their default rate is huge
- Organizations with highest percent of loans not repaid are Transport: type 3 (16%), Industry: type 13 (13.5%), Industry: type 8 (12.5%) and Restaurant (less than 12%). Self-employed people have relative high defaulting rate, and thus should be avoided for loan approval or to provide loan with higher interest rate to mitigate the risk of defaulting
- Bank should avoid young people who are in age group of 20-40 as they have higher probability of defaulting
- People who have less than 5 years of employment have high default rate. Loans of people with higher work experience are preferred
- Client who have children equal to or more than 9 default 100% and hence their applications are to be rejected
- When the credit amount goes beyond above 3 lakhs for the goods price, there is an increase in defaulters.

Factors that determine if the Loan can be approved on the Condition of a High Interest rate to mitigate any default risk leading to business loss

- High number of loan applications are from the category of people who live in Rented apartments & who are living with parents and hence offering the loan to them would mitigate the loss if any of them default
- People who get a loan for 3-6 Lakhs tend to default more than others and hence having higher interest specifically for this credit range would be ideal
- Since 90% of the applications have an income range of less than 3 lakhs per annum and have a high probability of defaulting, they could be offered loan with higher interest compared to other income category
- Clients who have 4 to 8 children have a very high default rate and hence higher interest should be imposed on their loans
- Loan taken for the purpose of Repairs seems to have highest default rate. A very high number applications have been rejected by the bank or refused by client in previous applications as well which has purpose of repair and others. This shows that the purpose of repair is taken as a high risk by bank, and either they are rejected, or bank offers very high loan interest rate which is not feasible by the clients, thus they refuse the loan. The same approach could be followed in future as well.

Suggestions for the Bank

- 90% of the clients who previously cancelled their loans actually repaid the loan. The bank should keep a record of the reason for these cancellations as they might help the bank to determine and on negotiate terms with these repaying customers in future for increase business opportunity.
- 88% of the clients who were refused for loan from the earlier have now turned into a paying client. Hence documenting the reason for rejection could mitigate the business loss and these clients could be contacted for further loans.



Thank You



Pratyush Pran
Siddhee Washimkar

