Bank Loan Case Study report

Overall approach of analysis

Problem statement

A consumer finance company which specializes in lending various types of loans to customers must decide if the customer can be lent a loan or not based on the applicant's profile.

Analysis approach

Based on factors and patterns, it must be decided if the application should be accepted or rejected. The primary driving factor should be that the customer should be able to pay for the installments and not default with any payment. There are various other factors and patterns like, previous loan history, interest rate of loans based on user profile, reduction of loan amount, etc.

We first check the customer's application to find out if their application meets the company's metric of approval. We check various metrics like assets and liabilities to arrive on a decision. If the metric is still in the grey, we move on to check if the customer has any previous loans at our company or a different company. If there is no default of payment or any red flags in the checking process, our company can proceed to draw out a loan for this particular customer.

By following this approach, we can filter out loan defaulters easily so as to reject their application and ensure that any potential customer is not rejected by the company which if rejected leads to loss for the company.

Indentify the missing data and use appropriate method to deal with it.

Upon analysis of the given datasets, it was found that were certain columns with missing values. In EDA, since it is not necessary to replace the missing values, specific columns like ***own_car_age (column V)*** have missing values if the applicant has specified ***N*** in ***FLAG_OWN_CAR (column E)*** in ***application_data*** dataset. In cases like these, there is no requirement to replace values or remove column.

In the same dataset, there are few other columns with missing values and no requisite reason was identified to remove or replace them.

In ***previous_application*** dataset, columns related to annuity of loan, rate of interest have missing values that need not be meddled with. But ***NAME_TYPE_SUITE*** column ***(column U)*** has missing values and considering the data column's relevancy to the dataset, it was decided that the column can be removed and thus removed from the dataset.

Outliers in the dataset

An outlier is defined to be as an observation that lies an abnormal distance from other values in a random sample from a population.

In ***application_data*** dataset, columns *AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE can all serve as outliers.*

In ***previous_application*** dataset, columns AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT can serve as outliers.

AMT_INCOME_TOTAL	Outlier	Q1	Q3	IRQ	Upper limit	lower limit		
202500	FALSE	112500	202500	90000	337500	-22500		
270000	FALSE							
67500	FALSE						TRUE	Indicates an outlier
135000	FALSE						FALSE	Not an outlier
121500	FALSE							
99000	FALSE							
171000	FALSE							
360000	TRUE							
112500	FALSE							
135000	FALSE							
112500	FALSE							
38419.155	FALSE							
67500	FALSE							
225000	FALSE							
189000	FALSE							
157500	FALSE							
108000	FALSE							
81000	FALSE							
112500	FALSE							
90000	FALSE							
135000	FALSE							
202500	FALSE							
450000	TRUE							
83250	FALSE							

Using QUARTILE function, we find Quartile 1 and Quartile 3. We use the difference of Q3-Q1 to find IRQ.

Upper limit = Q3+(1.5*IRQ)

Lower limit = Q1-(1.5*IRQ)

And the by using OR function,

=OR(AMT_INCOME_TOTAL>upper limit, AMT_INCOME_TOTAL<lower limit)

All the outliers appear as TRUE.

Similarly, by using these functions we can say that all the columns contain outliers.

Data imbalance in data

Classification of a data with uneven distribution of data is referred to as data imbalance. Data that make up a large proportion of the data set are called majority classes. Those that make up a smaller portion are minority classes. On analysis of the data sets provided, ratio of data imbalance was found out.

The formula used,

Ratio = NUM1/GCD(NUM1,NUM2)&":"&NUM2/GCD(NUM1,NUM2)

	A	B	C	D	E	F	G
1	Results	TARGET		FLAG_OWN_CAR		FLAG_OWN_REALTY	
2	1/Y	1	24825	N	104587	Y	213311
3	0/N	0	282686	N	202924	N	94199
4		0		Y		Y	
5	ratio	0	24825:282686	N	104587:202924	Y	30473:13457
6		0		N		Y	
7	ratio formula	0		N		Y	
8	NUM1/GCD(NUM1,NUM2)&":"&NUM2/GCD(NUM1,NUM2)	0		Y		Y	
9		0		Y		Y	
10		0		N		Y	
11		0		N		Y	
12		0		N		Y	
13		0		N			
14		0		N		Y	
15		0		Y		N	
16		0		N		Y	
17		0		Y		Y	

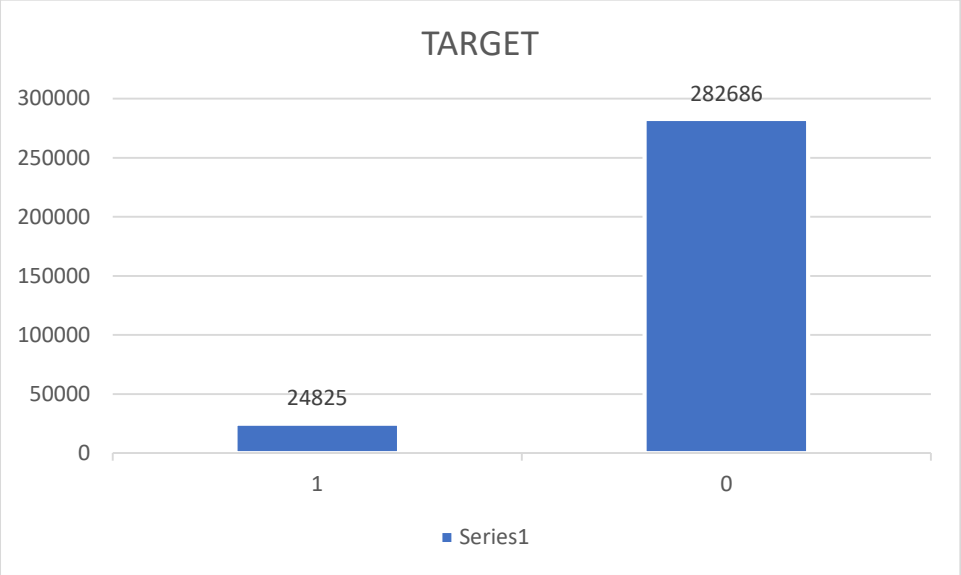
It was found that the Target column in ***application_data*** dataset has a data imbalance ratio of 24825:282686 after the formula was applied.

An excel workbook containing ratios of data imbalance has been attached in the drive link submitted.

Results of variate analysis

Univariate analysis

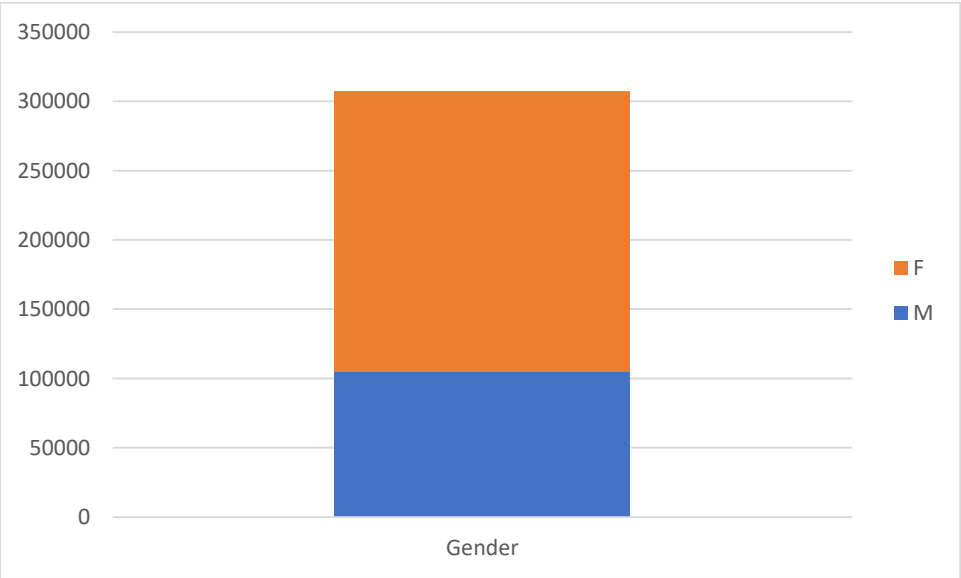
It refers to the analysis carried out on one variable to summarize the variable. On analysis of the Target column in ***application_data*** dataset, a graph was derived for the variable.



Where the Target variable has values 0 and 1 in the column. In this column, 1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases and we can clearly notice that customers falling under the all other cases are more than those with payment difficulties. This helps us understand that the customers with Target value as 0 are potential target customers for the company and this variable can be used as a deciding factor that they are reliable and would not fault on their repayments.

Segmented univariate

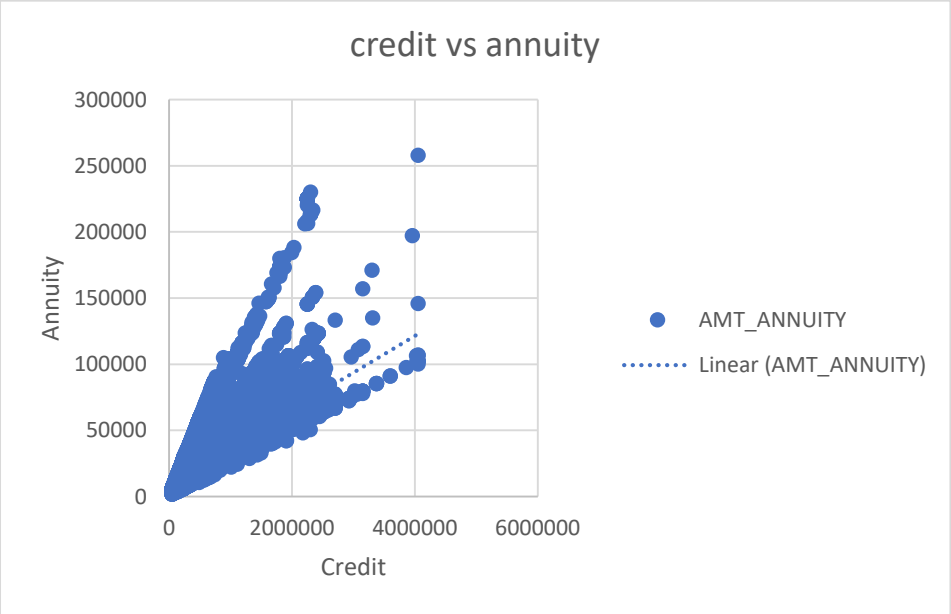
It can be used to find summary of a single data variable in the form of segments. The dataset variable is divided into subsets and patterns. In the ***application_data*** database, we used the ***Code_Gender*** column to implement segmented univariate.



The result of segmented univariate is that the female population contributes to the majority share, share greater than double of the male population. This results indicate that the company has a good name among its female population thus pulling a lot of applications from them. This means that there is a loss of revenue from the male population as their share of applications contributions to less than half of the total applications. The company must strategize on ways to improve the situation and bring in more revenue.

Bivariate analysis

It is the analysis of two variables to determine relationships between them. It is useful to determine whether there is a correlation between the variables and, if so, how strong the connection is. This is incredibly helpful for researchers conducting a study.



In the scatter chart, we use two variables, credit and the annuity to be paid. When comparing the both in a chart we find that the credit is heavily scattered around the low credit region and as the chart increases, the spread decreases. The linear line shows the estimates a straight line that minimizes the distance between itself and where observations fall in the data set.

Correlation

Top 10 correlation was to be determined between the segments of Target column where the division is

1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample

0 - all other cases

Correlation formula

=CORREL(ARRAY1,ARRAY2)

For Target value – 1

- AMT_CREDIT : AMT_GOODS_PRICE 0.983102519 (1)
- AMT_ANNUITY : AMT_GOODS_PRICE 0.752699196 (2)
- AMT_CREDIT : AMT_ANNUITY 0.752194735 (3)
- DAYS_REGISTRATION : DAYS_ID_PUBLISH 0.096832619 (4)
- AMT_GOODS_PRICE : REGION_POPULATION_RELATIVE 0.07604893 (5)
- AMT_ANNUITY : REGION_POPULATION_RELATIVE 0.07169025 (6)
- AMT_CREDIT : REGION_POPULATION_RELATIVE 0.069161087 (7)
- AMT_INCOME_TOTAL : AMT_ANNUITY 0.046421057 (8)
- AMT_INCOME_TOTAL : AMT_CREDIT 0.038131435 (9)
- AMT_INCOME_TOTAL : AMT_GOODS_PRICE 0.037216357 (10)

For Target value – 0

- AMT_CREDIT : AMT_GOODS_PRICE 0.987079085 (1)
- AMT_ANNUITY : AMT_GOODS_PRICE 0.779246915 (2)
- AMT_CREDIT : AMT_ANNUITY 0.773818766 (3)
- AMT_INCOME_TOTAL : AMT_ANNUITY 0.454037761 (4)
- AMT_INCOME_TOTAL : AMT_GOODS_PRICE 0.388458637 (5)
- AMT_INCOME_TOTAL : AMT_CREDIT 0.381544751 (6)
- AMT_INCOME_TOTAL : REGION_POPULATION_RELATIVE 0.182109433 (7)
- AMT_ANNUITY : REGION_POPULATION_RELATIVE 0.118967149 (8)
- DAYS_REGISTRATION : DAYS_ID_PUBLISH 0.104653072 (9)
- AMT_GOODS_PRICE : REGION_POPULATION_RELATIVE 0.097614974 (10)

Visualizations

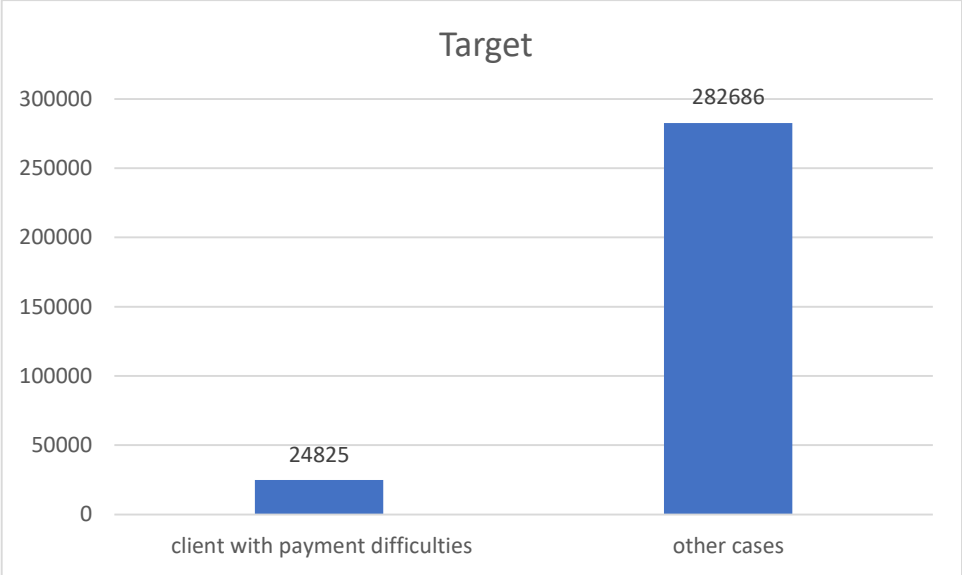


Fig 1 – Target clients

This chart helps the company to make a decision to either approve a loan to the next stage, cancel the application or re-evaluate the loan terms. This chart plays a pivotal role as the repayment ability of the client is depicted here. Continuous non re-payment of annuity leads to a loan defaulting and brings a loss to the company.

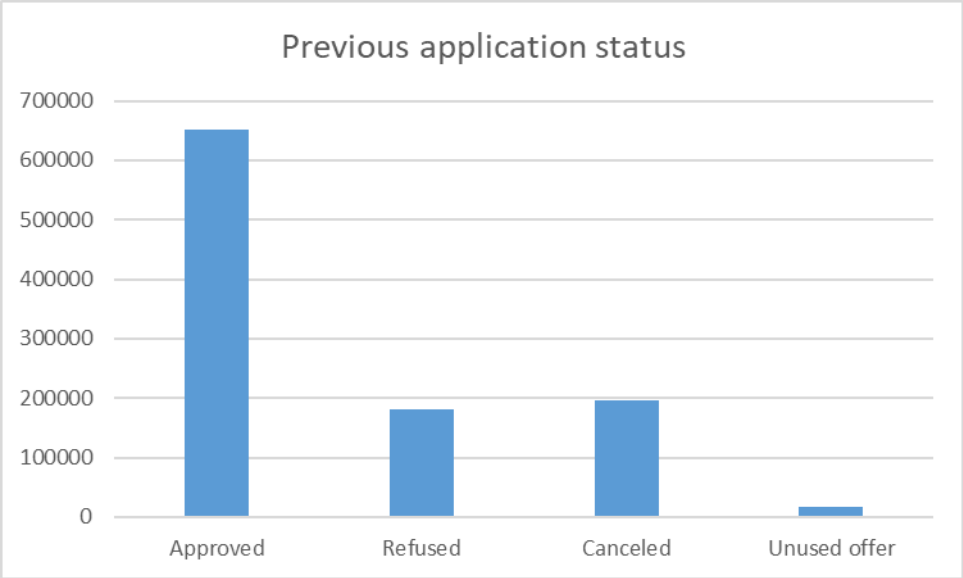


Fig 2 – Application status of previous application

When a company is looking into an application of the applicant, they initially check if the client is new to the company or has already had some business with the company. Based on this data, they check if the client has any previous applications with/without the same company and check the status of the previous application. Out of the results acquired via this test, the company digs into the

reason for the status when the approval condition is not “Approved”. If and when the status is not approved, the company reviews the application of the applicant.

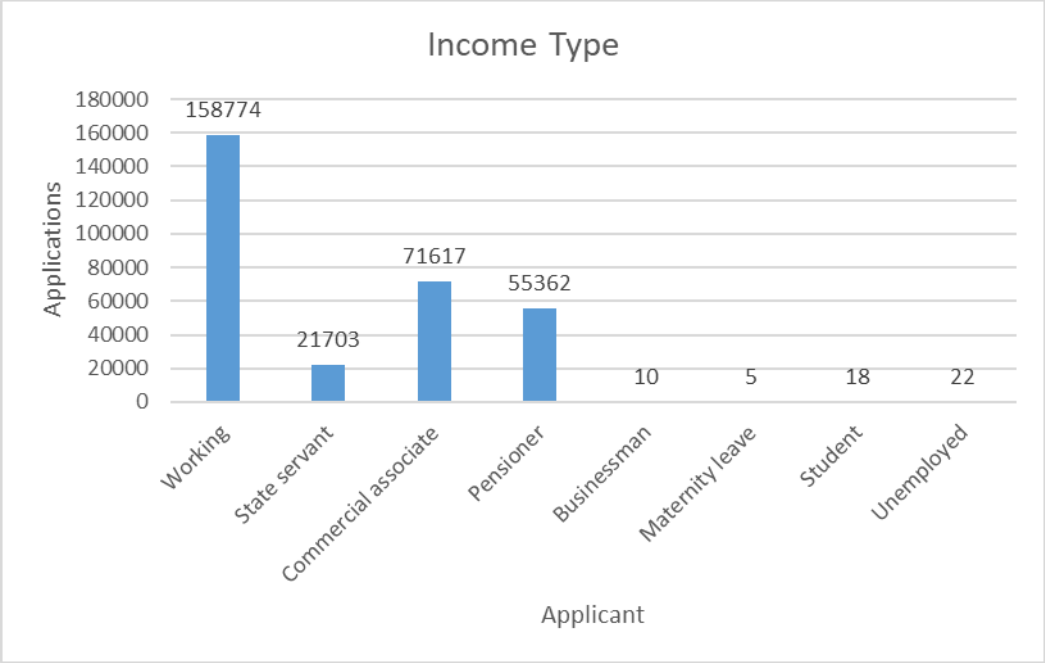


Fig 3 – Income type of client

Figure 3 displays the class of the income to which the application belongs. This data helps the company understand their customer database and helps the company analyse on their moves ahead on how to turn the tide to their advantage. This will help then widen their customer database which in turn leads to better customer satisfaction, more loan disbursing power and a good market share among competitors.