Bank Loan Case Study

PROJECT DESCRIPTION

Here we are trying to build a system through which we will be able to track the risk of providing or declining a loan prospect. Through this Bank loan analytics we will measure certain parameters which in turn will help us grow as a company and also most certainly help in understanding our flaws. By this way the banks get to know about the defaulter, the high risk customer and those who pay in time. These analytics are the foundation pillar of the success of any banking organisation. Situation such as- Approved, Cancelled, Refused or Unused offer are those situation which can occur when a client seeks for loan from a bank so in order to arrive at any conclusion we would first be able to analyse the clients previous loan background.

I have been given a dataset of a bank various columns of different parameters is given. Knowledge in statistics and different formulas in excel are used to draw necessary conclusions about the defaulters.

APPROACH

- My approach is to first find out whether the client is facing difficulty in paying the interest then accordingly I can take actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- The variables which are strong indicators of default.
- Presenting the overall approach of the data analysis, cleaning the dataset, finding outliers, data imbalance, univariate, segmented univariate, bivariate analysis etc.
- The top 10 correlation for the Client with payment difficulties and all other cases .

TECH-STACK

I had used MS-Excel provide by Microsoft. I have used the office home and student version of 2019.I had used Google Colab as it is used extensively in the machine learning community with applications

The reason for using it is that it has very user friendly interface and it is also hassle free with all the provided services such as creating visual illustrations, administering it, modifying it etc. I have particulary used it to create several required charts and graphs to perfectly understand the data then I have used multiple pivot tables to derive the outcome I required out of the given dataset.

RESULT

- **Presenting the overall approach of the analysis and also mentioning the problem statement and the analysis approach briefly**

- **Problem Statement:**

This case study aims to identify patterns which indicate if a client has difficulty paying their instalments 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. Identification of such applicant's using EDA is the aim of this case study.

- **Analysis Approach:**

1. Imported the NumPy, pandas, matplotlib and seaborn python libraries.
2. Imported the datasets (Application_Data & Previous_Application)
3. Identification: We have identified how we will approach the data, finding missing dataset and working on it accordingly to gain the required results.
4. Outliers: Identified outliers and showed how they play if any role in our dataset.
5. Imbalance: Understanding the ratio of imbalance in our data.
6. Correlation Analysis: Finding the correlation between the variables with respect to the target variables and find the top three correlation.
7. Visualisation: Visualized the data with the help of charts and graphs.

- **Identify the missing data and use appropriate method to deal with it. (Remove columns/or replace it with an appropriate value)**

1. Total of 49 columns in Application_Data and 11 columns in Previous_Application which have missing values greater than 40%.
2. Moreover I found that "EXT_SOURCE_2","EXT_SOURCE_3" has no correlation with the "TARGET" column.

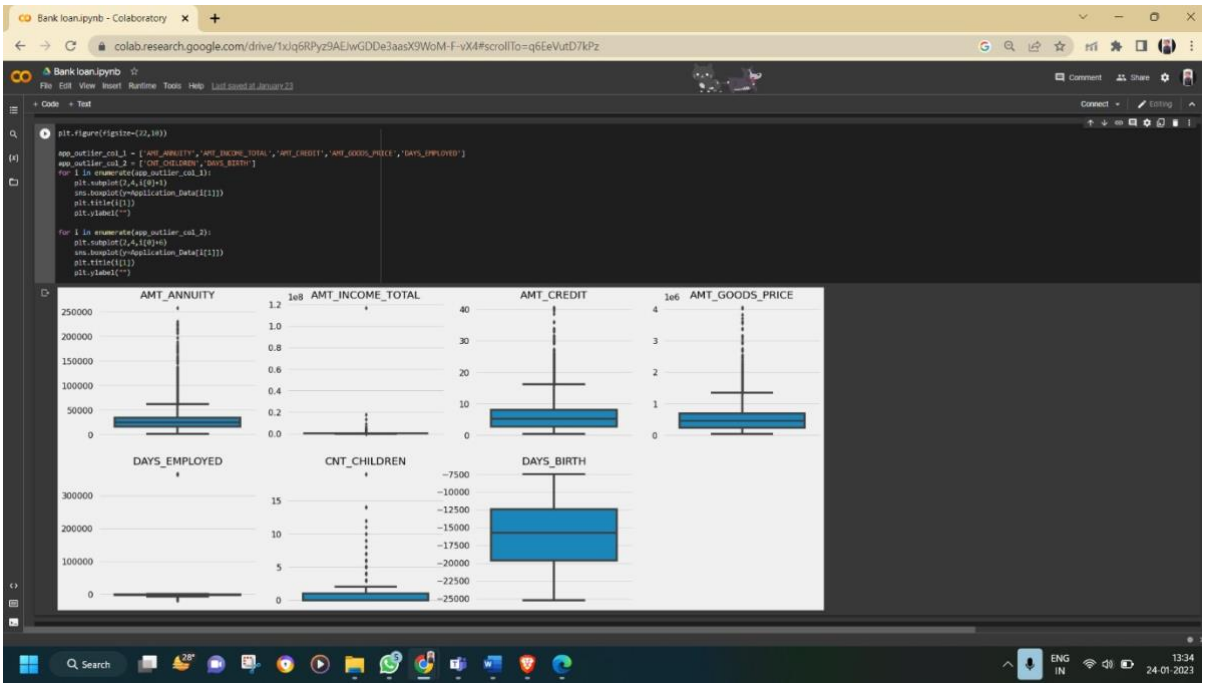
3. On checking the relation of ‘FLAG_DOCUMENT_X’ with loan repayment status, we found that the clients applying for loans only submitted the ‘FLAG_DOCUMENT_3’.
4. There is almost no correlation of ‘FLAG_MOBIL’,‘FLAG_EMP_PHONE’, ‘FLAG_WORK_PHONE’, ‘FLAG_CONT_MOBILE’, ‘FLAG_PHONE’, ‘FLAG_EMAIL’ with the “TARGET” column.
5. ‘WEEKDAY_APPR_PROCESS_START’, ‘HOUR_APPR_PROCESS_START’, ‘FLAG_LAST_APPL_PER_CONTRACT’, ‘NFLAG_LAST_APPL_IN_DAY’ are not needed for the analysis.
6. Dropping total 76 columns in Application_Data and 15 in Previous_Application.
7. Converting the negative days column into positive days.
8. Imputed categorical variable 'NAME_TYPE_SUITE' using mode, 'OCCUPATION_TYPE' by adding an ‘Unknown’ category, numerical variables
   'AMT_REQ_CREDIT_BUREAU_HOUR',
   'AMT_REQ_CREDIT_BUREAU_DAY',
   'AMT_REQ_CREDIT_BUREAU_WEEK',
   'AMT_REQ_CREDIT_BUREAU_MON',
   'AMT_REQ_CREDIT_BUREAU_QRT',
   'AMT_REQ_CREDIT_BUREAU_YEAR' with median.
9. Imputed AMT_ANNUITY with median, AMT_GOODS_PRICE with mode, CNT_PAYMENT with 0 as the NAME_CONTRACT_STATUS for these indicate that most of these loans were not started.

- **Identifying if there are outliers in the dataset. Also, mention why do you think it is an outlier.**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. An outlier can be identified from a box-plot graph. If the value lies above maximum and below minimum, they are considered as outliers.

Application_Data:

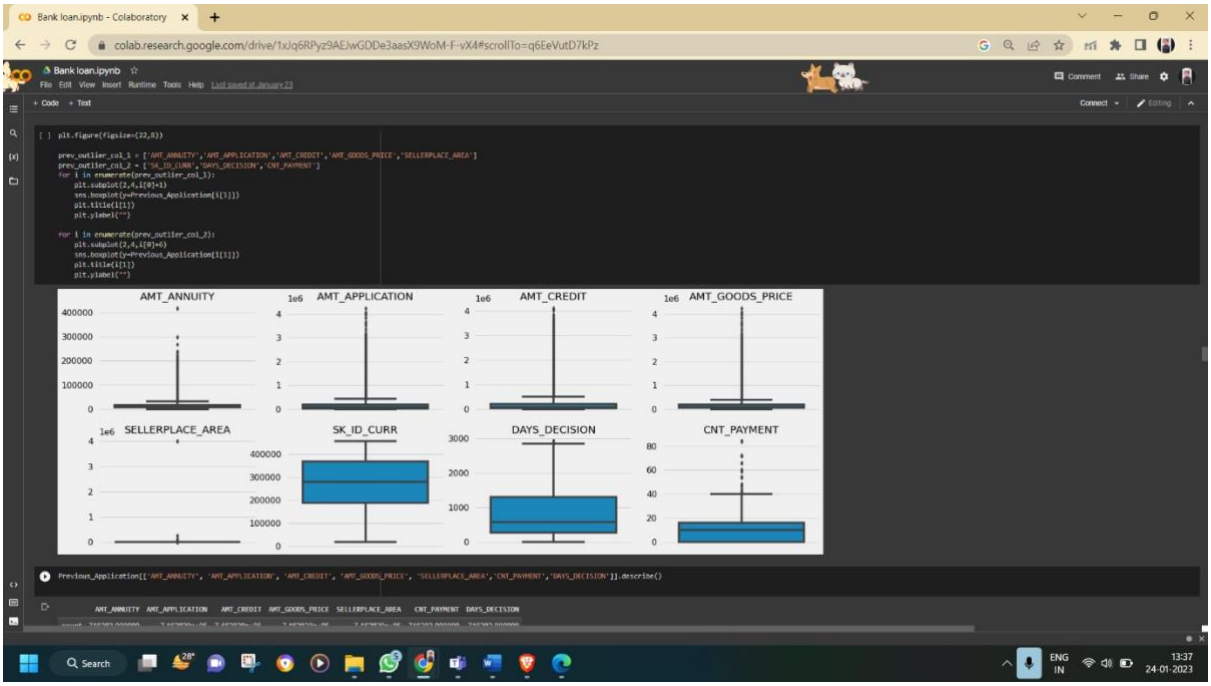
1. AMT_ANNUITY, AMT_CREDIT, AMT_GOODS_PRICE,CNT_CHILDREN have outliers.
2. AMT_INCOME_TOTAL has many outliers which shows that few of the loan applicants have high income compared to the others.
3. DAYS_BIRTH has no outliers which means the data available is reliable.
4. DAYS_EMPLOYED has outlier values around 350000(days) which is around 958 years which is impossible, so it is an incorrect entry.



Previous_Application:

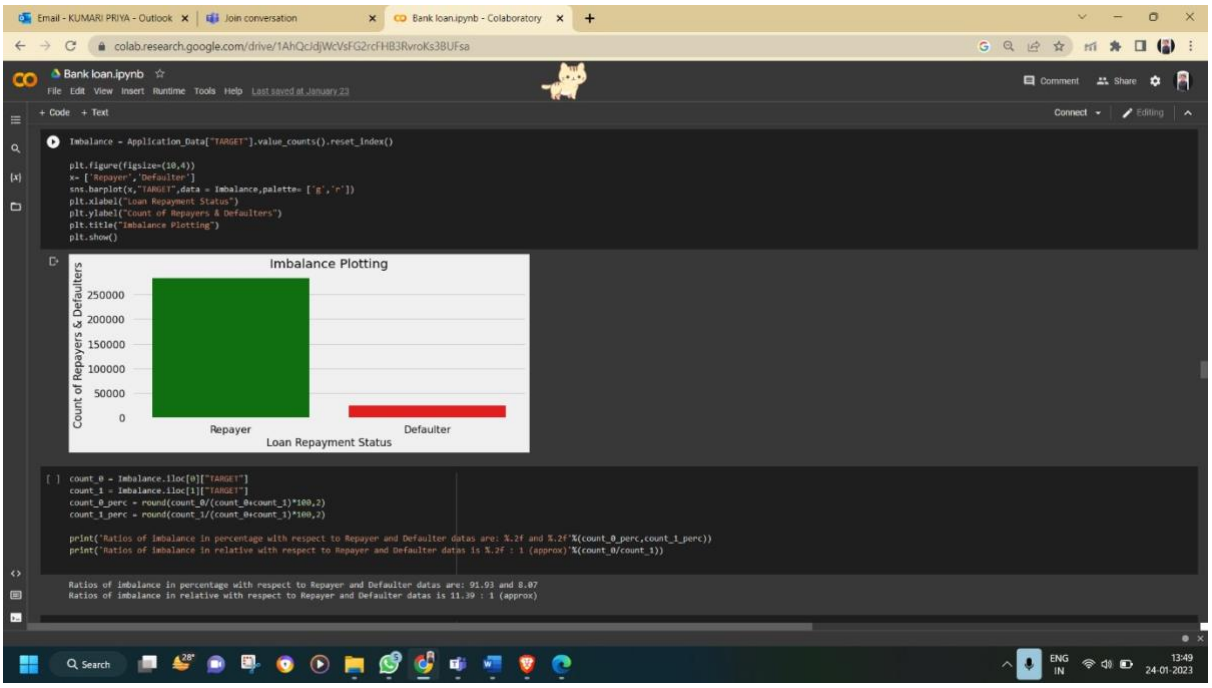
1. AMT_ANNUITY, AMT_APPLICATION, AMT_CREDIT, AMT_GOODS_PRICE, SELLERPLACE_AREA have large number of outliers.

2. CNT_PAYMENT has some outlier values.
3. SK_ID_CURR is an ID column so no outliers.
4. DAYS_DECISION has little number of outliers showing these previous applications decisions were taken long back.



- **Identify if there is data imbalance in the data. Find the ratio of data imbalance.**

This data is highly imbalanced as number of defaulter is very less in total population. Data Imbalance Ratio is: 11.39 : 1 (approx.)



- **Explain the results of univariate, segmented univariate, bivariate analysis, etc. in business terms.**

1. The number of female clients is almost double the number of male clients. Based on the percentage of defaulted credits, males have a higher chance of not returning their loans (~10%), comparing with women (7%).
2. Clients who own a car are half in number of the clients who don't own a car. But based on the percentage of default, there is no correlation between owning a car and loan repayment as in both cases the default percentage is almost same.

3. The clients who own real estate are more than double of the ones that don't own. But the defaulting rate of both categories are around the same (~8%). Thus there is no correlation between owning a reality and defaulting the loan.
4. Majority of people live in House/apartment
5. People living in office apartments have lowest default rate
6. People living with parents (~11.5%) and living in rented apartments(>12%) have higher probability of defaulting
7. Most of the people who have taken loan are married, followed by Single/not married and civil marriage
8. In terms of percentage of not repayment of loan, Civil marriage has the highest percent of not repayment (10%), with Widow the lowest (exception being Unknown).
9. Majority of the clients have Secondary / secondary special education, followed by clients with Higher education. Only a very small number having an academic degree
10. The Lower secondary category, although rare, have the largest rate of not returning the loan (11%). The people with Academic degree have less than 2% defaulting rate.
11. Most of applicants for loans have income type as Working, followed by Commercial associate, Pensioner and State servant.
12. The applicants with the type of income Maternity leave have almost 40% ratio of not returning loans, followed by Unemployed (37%). The rest of types of incomes are under the average of 10% for not returning loans.
13. Student and Businessmen, though less in numbers do not have any default record. Thus these two category are safest for providing loan.
14. Most of the applicants are living in Region_Rating 2 place.
15. Region Rating 3 has the highest default rate (11%).
16. Applicant living in Region_Rating 1 has the lowest probability of defaulting, thus safer for approving loans.
17. Most of the loans are taken by Laborers, followed by Sales staff. IT staff take the lowest amount of loans.
18. The category with highest percent of not repaid loans are Low-skill Laborers (above 17%), followed by Drivers and Waiters/barmen staff, Security staff, Laborers and Cooking staff.
19. 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 to be approved for loan or provide loan with higher interest rate to mitigate the risk of defaulting.

20. Most of the people application for loan are from Business Entity Type 3.
21. For a very high number of applications, Organization type information is unavailable(XNA).
22. Category of organization type has lesser defaulters thus safer for providing loans:

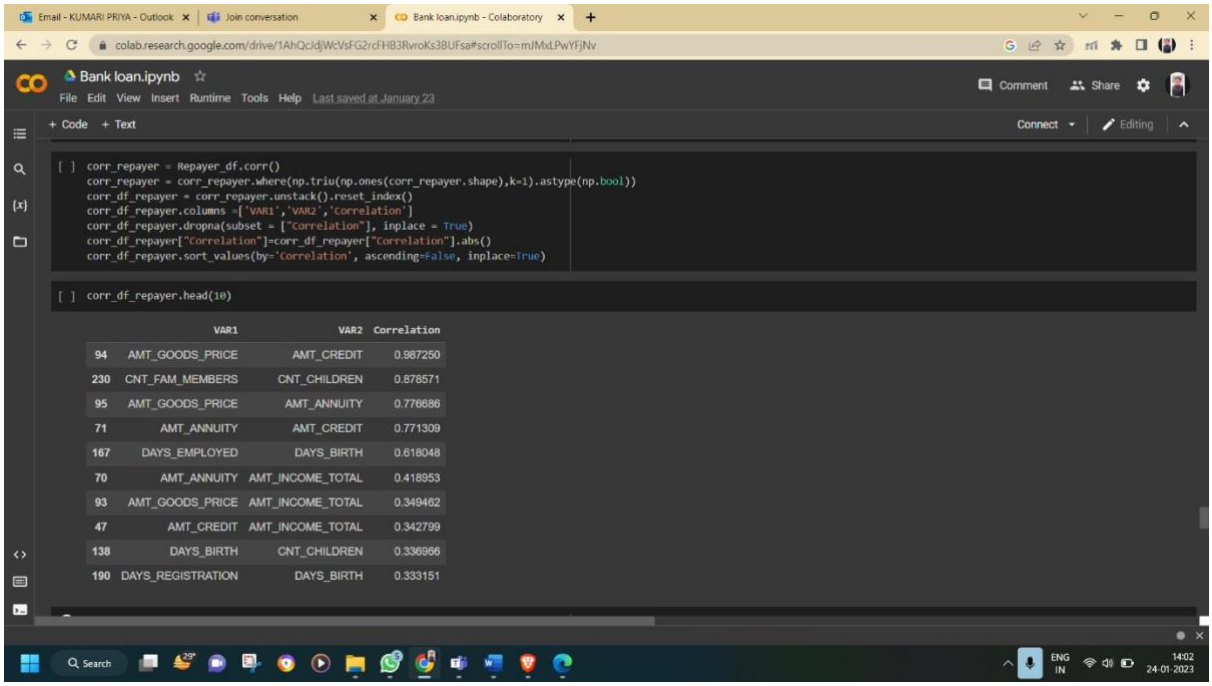
    Trade Type 4 and 5

    Industry type 8
23. There is no significant correlation between non-defaulters and defaulters in terms of submitting document 3 as we see even if applicants have submitted the document, they have defaulted a slightly more (~9%) than who have not submitted the document (6%).
24. People in the age group range 20-30 have higher probability of defaulting. And people above age 50 have low probability of defaulting.
25. Majority of the applicants have been employed in between 0-5 years. The defaulting rating of this group is also the highest which is 10%
26. With increase of employment year, defaulting rate is gradually decreasing with people having 40+ year experiences having less than 1% default rate
27. More than 80% of the loan provided are for amount less than 900,000. People who get loan for 300-600k tend to default more than others.
28. 90% of the applications have Annual Income less than 300,000. Application with Income less than 300,000 has high probability of defaulting. Applicant with Income more than 700,000 are less likely to default.
29. Most of the applicants do not have children. Very few clients have more than 3 children. Clients who have more than 4 children has a very high default rate with child count 9 and 11 showing 100% default rate. Family members follow the same trend as Children, where, having more family members increases the risk of defaulting.
30. Business man's income is the highest and the estimated range with default 95% confidence level seem to indicate that the income of a business man could be in the range of slightly close to 4 lakhs and slightly above 10 lakhs.

- **Find the top 10 correlation for the Client with payment difficulties and all other cases (Target variable).**

  - **The top 10 correlation for the Client with repayment:**

1. Credit amount is highly correlated with amount of goods price, loan annuity, total income
2. We can also see that repayment have high correlation in number of days employed.

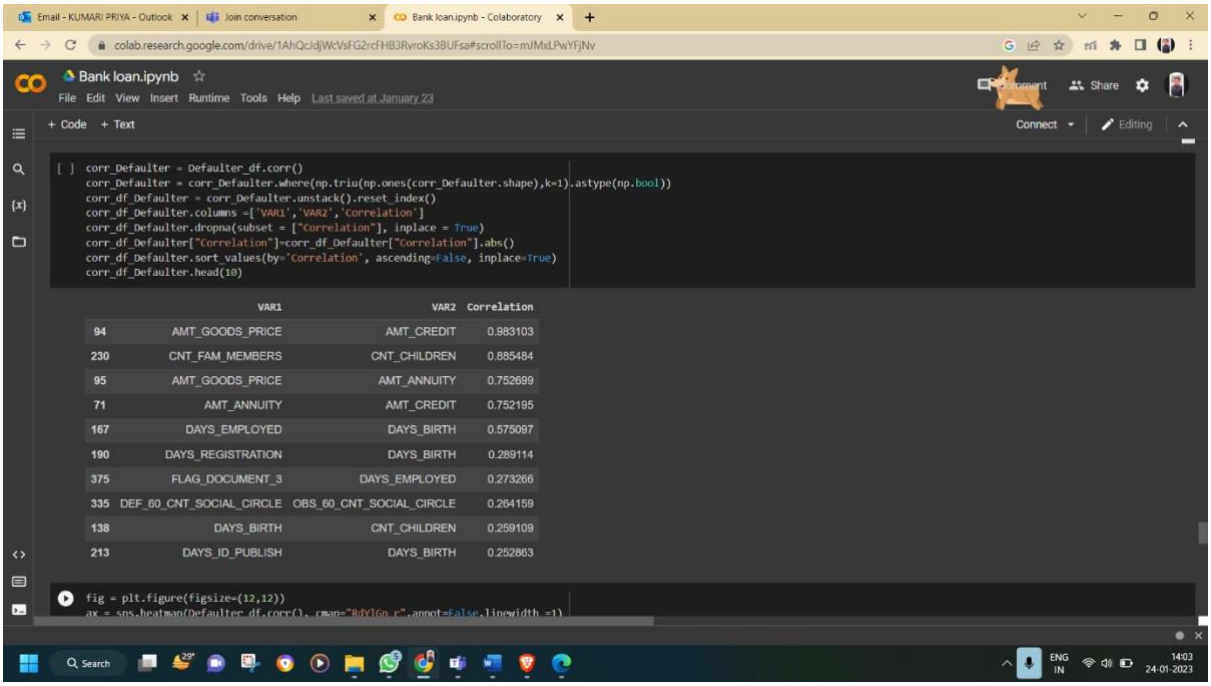


	VAR1	VAR2	Correlation
94	AMT_GOODS_PRICE	AMT_CREDIT	0.987250
230	CNT_FAM_MEMBERS	CNT_CHILDREN	0.878571
95	AMT_GOODS_PRICE	AMT_ANNUITY	0.776686
71	AMT_ANNUITY	AMT_CREDIT	0.771309
167	DAYS_EMPLOYED	DAYS_BIRTH	0.618048
70	AMT_ANNUITY	AMT_INCOME_TOTAL	0.418953
93	AMT_GOODS_PRICE	AMT_INCOME_TOTAL	0.349462
47	AMT_CREDIT	AMT_INCOME_TOTAL	0.342799
138	DAYS_BIRTH	CNT_CHILDREN	0.336966
190	DAYS_REGISTRATION	DAYS_BIRTH	0.333151

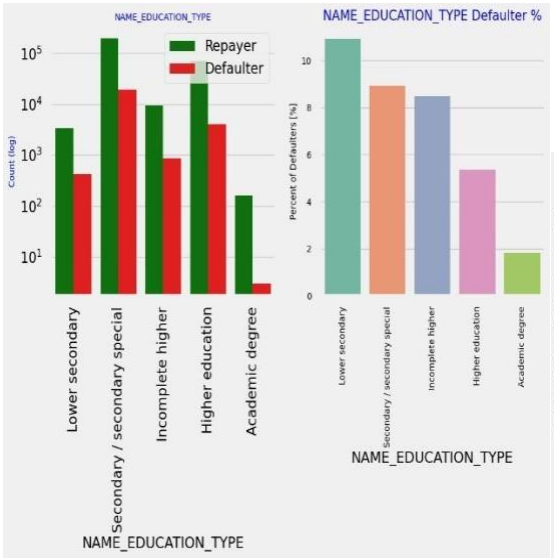
  - **The top 10 correlation for the Client with default:**

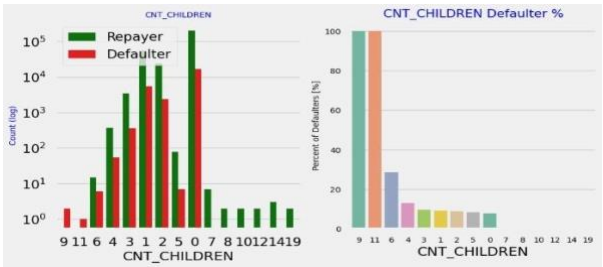
1. Credit amount is highly correlated with amount of goods price which is same as repayments.
2. But the loan annuity correlation with credit amount has slightly reduced in defaulters(0.75) when compared to repayment(0.77).
3. We can also see that repayment have high correlation in number of days employed(0.62) when compared to defaulters(0.58).

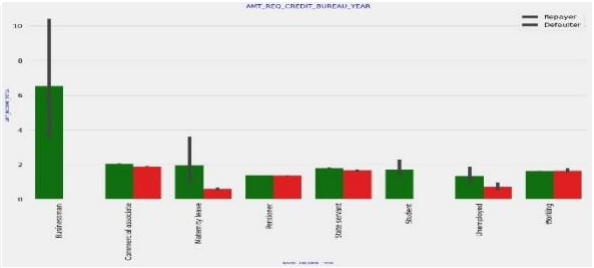
4. There is a severe drop in the correlation between total income of the client and the credit amount(0.038) amongst defaulters whereas it is 0.342 among repayment.
5. Days_birth and number of children correlation has reduced to 0.259 in defaulters when compared to 0.337 in repayment.
6. There is a slight increase in defaulted to observed count in social circle among defaulters(0.264) when compared to repayment (0.254).

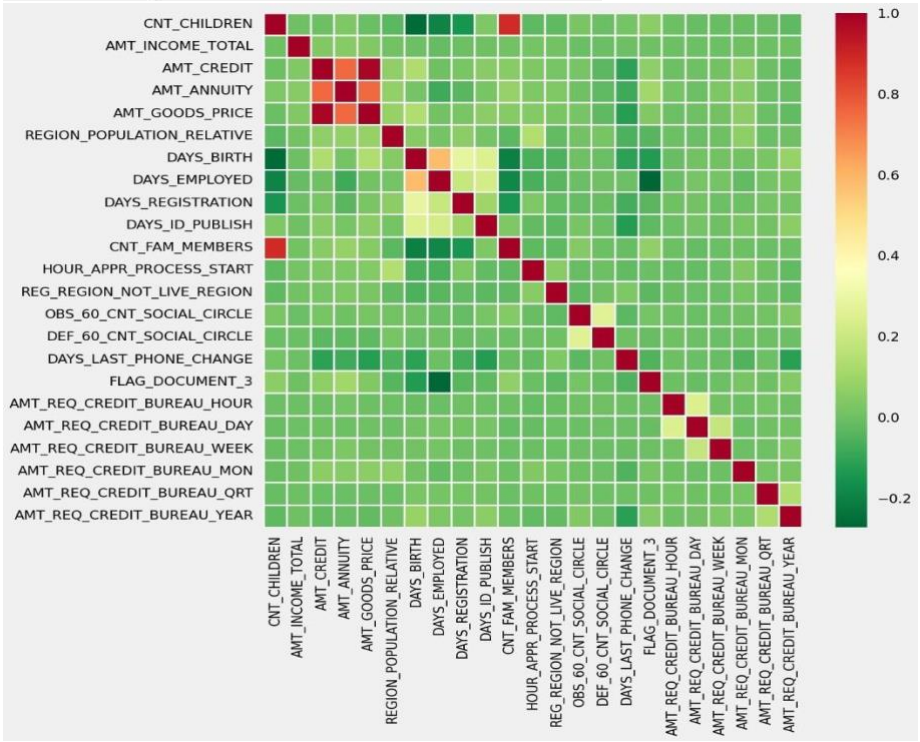


- **Include visualizations and summarize the most important results in the presentation**.









Visualizations and their summarization link:

https://colab.research.google.com/drive/1AhQcJdjWcVsFG2rcFHB3RvroKs3BUFsa?usp=sharing

INSIGHT:-

- In this case study, I applied the EDA in the real business case scenario.
- I learned basic of risk analytics in banking and financial services and understood how data is used to minimize the risk of losing money while lending to customers.
- This case study helped me in learning how to summarize a huge dataset to gain the valuable insights.
- This project was very challenging. I implemented the study of correlation between different variables to extract the necessary insights for the clients.
- I learned about data imbalance, outliers, driving factors for the datasets.
- It helped me in visualizing the huge dataset and summarizing the most important results helpful to the client.

Submitted by :- **Bishwomkar Panigrahi**