Credit EDA Assignment

Pooja Rathod

Business Objective

- 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.

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

READ THE DATA

- Understanding the columns in application dataset and the previous application dataset.
- Import the packages ,warnings and files
- Data Inspection

Data Cleaning

- Handling Null Values
- Check Data types and converting them to appropriate types
- Handling Outliers
- Check Imbalance of data between the targets

Handling Null Values

- Drop the columns with more than 35% of null values
- Impute the columns which has <15% of null values
- Impute Categorical columns with their mode value
- Impute Numerical columns with their mode value

Changing datatypes for more convinient imputation

Change variables into categorical columns which contains ‘digits+alphabets’ values.

For eg. *‘SK_ID_CURR’* is int data type which holds ids of customers and this variable cannot be manipulated hence convert it to ‘object’ datatype

Correcting Inappropriate data

'DAYS_BIRTH' ,'DAYS_EMPLOYED' ,'DAYS_REGISTRATION' ,'DAYS_ID_PUBLISH'

- Firstly change them to positve values using abs() Function.
- Secondly Convert days into years.

By making days to years, we can figure out data easily

Impute Continuous Numerical Variables

- Create new columns by dividing them into bins and ranges for easy understanding

- For Eg.
  - Binning `AMT_INCOME_TOTAL` to `AMT_INCOME_TYPE`
  - Binning 'AMT_CREDIT' to 'AMT_CREDIT_RANGE'

OUTLIER ANALYSIS

- Boxplots distribution for each `numerical_col` using `"for"` loop
- INSIGHTS FOR OUTLIERS:

- * IQR for `AMT_INCOME_TOTAL` is very slim and it has a large number of outliers.
- * Third quartile of `AMT_CREDIT` is larger as compared to the First quartile which means that most of the Credit amount of the loan of customers are present in the third quartile. And there are a large number of outliers present in AMT_CREDIT.
- * The third quartile `AMT_ANNUITY` is slightly larger than the First quartile and there is a large number of outliers.
- * Third quartile of `AMT_GOODS_PRICE`,`DAYS_REGISTRATION` AND `DAYS_LAST_PHONE_CHANGE` is larger as compared to the First quartile and all have a large number of outliers.
- * IQR for `YEARS EMPLOYED `is very slim. Most of the outliers are present below 25000. And an outlier is present 375000.
- * `YEARS_BIRTH ,YEARS_ID_PUBLISH and EXT_SOURCE_2,EXT_SOURCE_3` don't have any outliers.

Check for imbalance of data between the targets

- Divide the application dataset into two different dataframes based on target variable's value, df0 and df1
- We found ratio of 0:1 is `11.387:1` indicates that for every target 1 there are almost 11 number of target 0's.
- This is a highly imbalanced data set.
- Defaulted population is 8.1 % and non- defaulted population is 91.9% .

Univariate Numerical Analysis Insights

- From box plot we can note that customer without payment difficulties having year in between 34 to 54 years And customer with payment difficulties having in between 31 to 50 years; While Senior Citizens(60-100) and Very young(19-25) age group facing paying difficulties less as compared to other age groups.

- Customer without payment difficulties lies in between 5 to 11 and Here we can see that the customer with payment difficulties lies in between 3 to 11 when a box plot of days to publish plotted.

Univariate Categorical Analysis Insights

- We found that the customer without payment and customer with payment difficulties both are taking cash loans.

- Female clients applied higher than male clients for loan.
  - 66.6% Female clients are non-defaulters while 33.4% male clients are non-defaulters.
  - 57% Female clients are defaulters while 42% male clients are defaulters.

- Clients having education Secondary or Secondary Special are more likely to apply for the loan. Clients having education Secondary or Secondary Special have higher risk to default. Other education types have minimal risk.

- (Defaulters as well as Non-defaulters) Clients with ORGANIZATION_TYPE Business Entity Type 3, Self-employed, Other ,Medicine, Government,Business Entity Type 2 applied the most for the loan as compared to others

- * (Defaulters as well as Non-defaulters) Clients having ORGANIZATION_TYPE Industry: type 13, Trade: type 4, Trade: type 5, Industry: type 8 applied lower for the loan as compared to others.

- We have found that the payment difficulties in home/ apartment in both the cases where customers take loan for house/ apartment as compared to others.
- We can see that labourers are having more difficulties in repaying the loan and also the core staff and the sales staff.
- But in the case of labourers those who have payment difficulties is way more then with having the payment done

BIVARIATE ANALYSIS

- People without payment difficulties take more credit for the annuity that they have
- We can see that goods price is positively correlated with credit amount.
- There more customers without paying difficulties who have been employed for 0 to 20 years range.

Categorical - Categorical bivariate analysis

- By Boxplot
  - Clients with different Education types except Academic degrees have a large number of outliers
  - Most of the population i.e. clients' credit amounts lie below 25%.
  - Clients with an Academic degree and who is a widow tend to take higher credit loan.
  - Some of the clients with Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special Education are more likely to take a high amount of credit loans

- By Boxplot
  - Clients having Higher Education, Incomplete Higher Education, Lower Secondary Education and Secondary/Secondary Special have a higher number of outliers.
  - some of the clients having Higher Education tend to have the highest income compared to others.
  - Though some of the clients who haven't completed their Higher Education tend to have higher incomes. Some of the clients having Secondary/Secondary Special Education tend to have higher incomes

Merge df and pre dataframes

- check the percentages of each type of contract status in new dataframe.

Approved 62.679378
Cancelled 18.351900
Refused 17.357984
Unused offer 1.610737

- Dividing the new dataframe into 4 parts based on the contract status

  i.e: Approved, refused, cancelled, unused

Examine the variables based on different status from Countplots

- Revolving loan is much more acceptable as compare to the cash and consumer loans by plotting the graphs of newly created status dfs
- The Repeater is getting more Refused but also we can see that the it also getting more approved and even that it is getting more cancelled and more unused.

Contd..

- Female is getting more Refused more approved more cancelled more unused but in case of male it is having average in every category(from countplot)
- Client with Secondary/ Secondary of education category is highest in all the cases of statuses.
- Working type people are applying more loans as compared to others and also Commercial associates are taking more loans.
- Married people are applying and taking loans more than the others

Contd..

- People are taking more loan in format of cash through the bank.
- Most approved loan were through POS and Most refused loans were in cash.
- Labourers are getting most refused and most approved loans; Sales staff is also getting the second most refused and approved loans.
- Most Refused loan is of Mobile and most approved loan is Mobile
- The most accepting loan is Cash X-sell: low And most cancelled loan is Cash and Most Unused loan is POS mobile with interest

Conclusion

- Banks should focus more on contract type ‘Student’ ,’pensioner’ and ‘Businessman’ with housing ‘type other than ‘Co-op apartment’ for successful payments.
- Banks should focus less on income type ‘Working’ as they are having most number of unsuccessful payments.
- Also with loan purpose ‘Repair’ is having higher number of unsuccessful payments on time.
- Get as much as clients from housing type ‘With parents’ as they are having least number of unsuccessful payments.