EDA Assignment

Name: Kunal

Problem Statement

The aim is to identify the patterns which can help to understand the clients :-

1. Who have difficulty paying their instalments should not be given loan as it will cause business loss.
2. Who can make the payments on time and should be given loan as it will help to get business profit

Procedure

- Understand the columns in application dataset and the column dictionary will provide the guidance to understand all the parameters involved in the problem statement.

- Followed by that the describe and shape function helps to get the overall information on rows and columns and also mean, minimum, maximum and standard deviation values.

- Then the next part is data cleaning and it is important to understand the % of null values present in individual columns.

- After getting the complete range of null values, I decided to eliminate columns with more than 40% null values for better analysis.

- Initially the dataset had 112 columns and post filtering out columns with 40% or more null values it gets reduced to 73.

- After that with the remaining columns we can check for the outliers and I chose 7 columns for outliers understanding.

- The columns were categorised into continuous and categorical variables

- Continuous Variables = ‘EXT_SOURCE_2’, ‘AMT_GOODS_PRICE’

- Continuous Variables = ‘EXT_SOURCE_2’, ‘AMT_GOODS_PRICE’

- Categorical variables = 'OBS_30_CNT_SOCIAL_CIRCLE','OBS_60_CNT_SOCIAL_CIRCLE','DEF_60_CNT_SOCIAL_CIRCLE','DEF_30_CNT_SOCIAL_CIRCLE','NAME_TYPE_SUITE’

- With the help of boxplot we can understand the outliers for Continuous Variables and we get an understanding that ‘EXT_SOURCE_2’ has no outliers but ‘AMT_GOODS_PRICE’ has outliers which can be imputed by median value.

- Then from application data the columns which are not required they can be filtered out. There are 31 columns which will have no utilization so after removing them we get 42 columns for use which can be determined using shape feature in pandas.

- After that using inner join, we need to merge application data and previous application data

- Then from the combined data we can drop the columns which are not required for any analysis.

- After that we will perform univariate analysis to understand the distribution of contract status with purposes, distribution of purposes with target.

- Then we can check for the anomalies in the columns of gender and it can be determined by mode and as frequency of female values is high so anomalies can replaced by female.

- Then in next step converting all the –ve values into +ve values for columns like DAYS_BIRTH, DAYS_EMPLOYED, DAYS_REGISTRATION, DAYS_ID_PUBLISH, DAYS_LAST_PHONE_CHANGE as –ve values are not possible.

- Then with the bins creation, we can analyse amount income and amount credit range.

- Then segregating defaulters and non-defaulters ana calculating the imbalance ratio.

- Then preparing graphs for Income and loans type to understand and compare the defaulters and non defaulters data

- Then we will perform univariate analysis on Code_Gender and Flag_Own_Car dataset

- Then box plot on analysis on Amt_Annuity, Amt_Goods_Price, Days_Birth, Days_Employed, Days_Last_Phone_Change, Days_ID_Publish and analyze

- Then Bivariate analysis on defaulters data for Credit amount and Education Status

- Then box plot for different families to compare Income amount vs Education status

- Then box plot for different families to compare Credit amount vs Education status.

- Followed by that we need to determine the top 10 correlation variables between defaulters and non-defaulters.

- Then we need to read the previous application dataset to understand the data in different rows and columns.

- Similar to application data we need to find out null values% in previous application and using filter eliminate columns with more than 50% null values

- After that with the help of shape we can determine the final set of rows and columns

- Then box plot for different families to compare Credit amount vs Education status.
- Followed by that we need to determine the top 10 correlation variables between defaulters and non-defaulters.
- Then we need to read the previous application dataset to understand the data in different rows and columns.
- Similar to application data we need to find out null values% in previous application and using filter eliminate columns with more than 50% null values
- After that with the help of shape we can determine the final set of rows and columns

Insights

Data Cleaning Imputation for continuous variables

- While computing boxplot for ‘EXT_SOURCE_2’ no outliers were present which means there is no major difference between mean and median value so median value can be used for imputation.

- For ‘AMT_GOODS_PRICE’, outliers are present and here also median values can be used for imputation

For categorical variables

- For Gender values we can calculate mode to know the frequency, here Female data has more frequency so XNA can be replaced by ‘Female’.

- Box plot for CNT_CHILDREN signifies that majority of the data lies in 1st quartile.

- For AMT_INCOME_TOTAL there is only high value outlier and it can be considered for analysis purpose.

- AMT_CREDIT,AMT_ANNUITY, DAYS_EMPLOYED, DAYS_REGISTRATION all these columns have smaller 1st quartile data compared to 3rd quartile

- Amt Income Range - Customers with no difficulties shows that customers with 100000 – 200000 most no of loans and have more chance of becoming defaulters. The income segment greater than 500000 are less chances of becoming defaulters

- Amt Credit Range - The customers with loan < 100000 are less defaulter and more than 100000 have high level of defaulters

- Name Income Type – Student pensioner and business have higher percentage of loan repayment. The working class, state servant and commercial associates have higher default percentage. Maternity category has high level of problem in loan repayment

- Name Contract Type – ‘Cash Loans’ has higher credit than ‘Revolving Loans’ contract type.

- Revolving loans are small compared to cash loans but non payment of revolving loans are on higher side.

- Categorical Univariate Analysis on ‘Code Gender’ and ‘Flag Own Car’ states that defaulters are more in male and customers owning car are higher percentage defaulters

- Univariate Analysis for continuous variable depicts –
- Customers with higher age have higher chance of repayment.
- Outliers are present in 'AMT_ANNUITY','AMT_GOODS_PRICE','DAYS_EMPLOYED', DAYS_LAST_PHONE_CHANGE in the dataset.
- DAYS_ID_PUBLISH plot suggests that customers who have changed their ID are prone to become defaulters.

- Bivariate analysis for numerical variables for non defaulters shows that in credit amount suggests that credit are higher in number for customers in family status ‘civil marriage’, ‘marriage’ and ‘separated.

- Most credits in 3rd quartile are for civil marriage.

- Bivariate analysis for numerical variables in education type shows that for non defaulters implies that the income amount is mostly similar to family members.

- Bivariate analysis for numerical variables for defaulters
- Credit Amount vs Education status -
- It shows that most of the outliers in Education type are in ‘Higher Education’ and ‘Secondary’.

- Most credits are for civil marriage for Academic degree in 3rd quartile.

- Income amount Vs Education status for defaulters –
- Income level for lower secondary is the least.
- The income amount is mostly similar to family members.

CORRELATION –

- The highest correlation is 1.0 and it is between (OBS_60_CNT_SOCIAL_CIRCLE with OBS_30_CNT_SOCIAL_CIRCLE) and (FLOORSMAX_MEDI with FLOORSMAX_AVG)

Univariate Analysis on combined data

- Most cancelled loan applications happened for ‘Repairs’
- ‘Repairs’ are facing most difficulty in repayment of loans on time.

Conclusion

- Contract type ‘Student’, ‘Pensioner’ and ‘Businessman’ with housing type make payments properly so they should be given loans on time.
- ‘Working’ income type make unsuccessful payments so their loans should not be approved.
- Banks should focus on housing type ‘with parents’, ‘House’ and ‘municipal apartments’ as the loan repayments are received on time.

That’s All Folks!

Thank You