Credit EDA Case Study

By

Vigneshwari Chandramohan
Anilkumar Biradar

Business Requirement

- To identify the clients who have payment difficulties.

- To identify the driving factors or pattern behind the loan default, which could assist the financial institution to avoid the credit loss.

Dataset:

To provide a detailed analysis to the bank, we have received two datasets:

1. ***application_data.csv***

   - contains all the information of the client at the time of application. The data is about whether a **client has payment difficulties** (Column: “Target” with value = 1) or not (Target = 0).

2. ***previous_application.csv***

   - contains the information about the previous loan details for the customers given in the 1st CSV file. It contains the status of the previous loan application – Approved, Cancelled, Refused or Unused offer.

With the help of the above datasets, we have to find a pattern and submit a report to the bank which contains the factors that has to be considered before approving the loan.

Implementation:

First we have to cleanse the data in “application_data.csv” file. This involves,

- Analysis on the missing data for each column.
- Find the outliers in the column values and perform the outlier treatment.
- Binning (grouping) of column values
    - Example – Age can be grouped into 20-30,30-40,40-50,50-60 and so on.
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis and Correlation

- After performing all the above steps, we divide the datasets into:
  - Defaulter (customers with payment difficulties Target = 1)
  - Non-Defaulter (Target = 0).
- Below are the columns considered:

SK_ID_CURR	AMT_GOODS_PRICE	DAYS_ID_PUBLISH	AMT_REQ_CREDIT_BUREAU_WEEK
TARGET	NAME_TYPE_SUITE	OCCUPATION_TYPE	AMT_REQ_CREDIT_BUREAU_MON
NAME_CONTRACT_TYPE	NAME_INCOME_TYPE	CNT_FAM_MEMBERS	AMT_REQ_CREDIT_BUREAU_QRT
CODE_GENDER	NAME_EDUCATION_TYPE	LIVE_REGION_NOT_WORK_REGION	AMT_REQ_CREDIT_BUREAU_YEAR
FLAG_OWN_CAR	NAME_FAMILY_STATUS	REG_CITY_NOT_LIVE_CITY	AGE
FLAG_OWN_REALTY	NAME_HOUSING_TYPE	REG_CITY_NOT_WORK_CITY	AGE_GROUP
CNT_CHILDREN	REGION_POPULATION_RELATIVE	LIVE_CITY_NOT_WORK_CITY	INCOME_GROUP
AMT_INCOME_TOTAL	DAYS_BIRTH	ORGANIZATION_TYPE	DEF_30_CNT_SOCIAL_CIRCLE
AMT_CREDIT	DAYS_EMPLOYED	DAYS_LAST_PHONE_CHANGE	DEF_60_CNT_SOCIAL_CIRCLE
AMT_ANNUITY	DAYS_REGISTRATION	AMT_REQ_CREDIT_BUREAU_HOUR	OBS_60_CNT_SOCIAL_CIRCLE
	OBS_30_CNT_SOCIAL_CIRCLE	AMT_REQ_CREDIT_BUREAU_DAY	

For more detailed analysis, we have performed binning in the following two variables:

- Using the column **DAYS_BIRTH**, Age of the client can be calculated. Then, we divide the age into groups as 20-30, 30-40, 40-50, 50-60, and 60-70.
- Using the **AMT_INCOME_TOTAL**, we group the salary into "Very Low", "Low", "Medium", "High", and "Very High“.

Univariate Analysis:

Following variables are considered for the **Univariate Analysis**.

1. NAME_CONTRACT_TYPE
2. CODE_GENDER
3. NAME_INCOME_TYPE
4. NAME_FAMILY_STATUS
5. NAME_HOUSING_TYPE
6. AGE_GROUP
7. NAME_EDUCATION_TYPE
8. INCOME_GROUP
9. OCCUPATION_TYPE

Observation and Inferences from Univariate Analysis:

- More number of defaulters in the Cash loans compared to Revolving loans
- Females are more defaulters compared to that of males.
- People who are married are greater in number in both the defaulter and non-defaulter list.
- People who own a house/apartment are higher in number in defaulters list.
- People in the age group: “30-40” are more defaulters compared to other age group people. These people may have lower income. On analysing, their income level, it is proved that they fall under “Low” and “Very-Low” income category.

Observation and Inferences from Univariate Analysis: (Cont.)

- People with Occupation_Type: “Laborers”, Income type: “Working” and Education type: “Secondary” are more defaulters compared to other categories.
- Performing the same analysis on the Non_Defaulter list, we could find that the people in the below categories are High in “Non-Defaulter” list also
  - Females
  - Type of Loans - Cash Loans
  - Family Status - Married
  - Age-Group - 30-40
  - Occupation - Labourers
  - Education – Secondary

Bivariate Analysis:

Following variables are considered for the Bivariate Analysis:

1. CODE_GENDER and NAME_CONTRACT_TYPE
2. CODE_GENDER and AGE_GROUP
3. CODE_GENDER and AMT_INCOME_TOTAL
4. FLAG_OWN_REALTY and INCOME
5. FLAG_OWN_CAR and INCOME
6. AGE_GROUP and INCOME
7. AMT_ANNUITY and AMT_CREDIT
8. AMT_ANNUITY and DEF_60_CNT_SOCIAL_CIRCLE

Observation and Inferences from Bivariate Analysis:

- In both the Defaulter and Non-Defaulter, Female customers borrowed more Cash Loans than Male customers and also, Female customers with Cash Loans are more in the Defaulter list when compared to males.
- Mean Salary of males is higher than the females.
  - So females face difficulties in paying their loan than males.
- People who have highest mean salary, owns a house/apartment in defaulters list. But in the non-defaulters list, both the categories (i.e.) owning a house/apartment or not owning a house/apartment have same mean salary.
- The mean salary of the people who own a car is high in both defaulter and non defaulter list than the people who don't have a car.
- With the visualization of scatter and pair plots between various numerical variables we were not able to derive any useful insights. We can find the correlation to some more useful information.

Comparison of Correlation HeatMap between Defaulter and Non-Defaulter:

For the purpose of correlation we can consider the following variables,

AGE,AMT_CREDIT,AMT_INCOME_TOTAL,AMT_ANNUITY,REGION_POPULATION_RELATIVE,DAYS_LAST_PHONE_CHANGE,DAYS_EMPLOYED,DEF_30_CNT_SOCIAL_CIRCLE,DEF_60_CNT_SOCIAL_CIRCLE

Observation and Inferences from Correlation:

- Comparatively, **zero correlation** between Age and Income for 'Defaulter' group than 'Non-Defaulter' group.
- In 'Non-Defaulter' group, there is a **positive correlation** among the income, loan amount and the annuity repayment.
  - Noticeably, **zero correlation** among the income, loan amount and the annuity repayment in 'Defaulter' group
- For 'Non-Defaulter' group, there is some **positive correlation** between the income and region population.
  - For 'Defaulter' group, there is **zero correlation** between the income and region population.
- For 'Defaulter' group, there is **no correlation** among the income, loan amount, employment duration & phone number change.

Observations and Inferences – Previous-Application dataset:

- Equal number of people have applied for ‘Cash loans’ and ‘Consumer loans’.
- More number of people falls under ‘Repeater’ category than ‘New’ and ‘Refreshed’.
- Bank have acquired most of the clients through ‘Credit and cash offices’.
- ‘**Approved**’ loans are more in number compared to that of ‘Cancelled’, ‘Refused’ and ‘Unused’.
- There is positive correlation among the application amount, loan amount and the down-payment amount.

Merging of Defaulter dataset and Previous application dataset:

- On merging, we can find whether the client in the defaulter list has any previous loans and we will get to know the status of the previous loan.

Observations and Inferences – Merging:

- On merging, we came to know that there are 23,845 clients who have their ID’s in previous application. So, for the clients who become defaulters, we can check their status of previous application in the merged dataset.

Recommendations for the Bank:

- Bank needs to take additional verifications when approving loans for **Females**.
- Whenever they receive an application for ‘Cash Loans’ they have to check all the following factors:
    - People in the age group of **30-40** are high in defaulters list.
    - If the Occupation type of the client is **Laborers** need to verify before approving their loans.
    - People under **Working** category have payment difficulties.
    - If the education type is “**Secondary**” approving their loans will lead to risk as they are high in number in the defaulters list.
- If a client applies for a loan, Bank has to first look into the Defaulters list and Merged dataset to find if the client already has any loan and status of the previous loan.
- To increase the profitability of the Bank, they have to concentrate more on people whose education is ‘**Academic Higher**” and if the occupation type is “**IT Staff**”, “**HR Staff**” and “**Realty agents**” as these people are less in number in Defaulter list.
- They can increase their business by concentrating more on the clients through “**Credit and cash offices**” as there are more number of clients through the channel.

Key Factors to be considered

Following key factors are to be considered when analyzing the loan applications:

- Gender: Female
- Type of Loan: Cash Loan
- Age group (30-40 years)
- Occupation Type: Laborers
  - Occupation Type: “IT Staff”, “HR Staff” and “Realty agents” – **Profitable Clients**
- Working: Income Type
- Education: Secondary qualification
- Channel through which they apply for loan: Credit & Cash Offices