CREDIT EDA ASSIGNMENT

EDA FOR BANK LOAN DEFAULTERS

INTRODUCTION

This assignment aims to give you an idea

applying EDA in a real business scenario. In this assignment, apart from applying the techniques that you have learnt in the EDA module, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.

BUSINESS UNDERSTANDING

- The loan -providing companies find it hard to give loans to people due to their insufficient or non-existent credit history. Because of that, some consumers use it to their advantage by becoming a defaulter. Suppose you work for a consumer finance company specialising in lending various types of loans to urban customers. You have to use EDA to analyse the patterns present in the data. This will ensure that the applicants capable of repaying the loan are not rejected.

- When the company receives a loan application, the company has to decide on loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision:
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business for the company.
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

Data sets

- *1. 'application_data.csv'* contains all the information of the client at the time of application.
  The data is about whether a **applicant has payment difficulties.**
- *2. 'previous_application.csv'* contains information about the applicant's previous loan data. It contains the data whether the previous application had been **Approved, Cancelled, Refused or Unused offer.**
- *3. 'columns_description.csv'* is data dictionary which elaborates the meaning of the variables.

Analysis on application_data

- Approach of Application Data Analysis

→Importing Module

→Reading the dataset

→Divided the columns into small segments and analyzed segment-wise using a smaller Dataframe.

→Data Cleaning,Missing Data Handling,Segment-wise Type casting done.

INSIGHTS

- Divided into segments:

Segment:1:Social Circle Info:

→ The features show similar trend for defaulters and non-defaulters, can be dropped.

segment:2:Contact Info :

→Considered 'FLAG_MOBIL','FLAG_EMP_PHONE' etc. for this segment. No impact on Target, features can be dropped.

Segment:3:Asset Details:

→ i. Most of the applicants own realty ii. Most of the applicants do not own cars iii. People not owning reality and car and have a slightly higher default rate than the people who own reality and car.

Segment:4:Family Related data:

→ This data is highly imbalanced as number of defaulter is very less in total population.

'CNT_FAM_MEMBERS', 'CNT_CHILDREN','NAME_INCOME_TYPE', 'OCCUPATION_TYPE',CODE_GENDER, 'EXT_SOURCE_1' and 'EXT_SOURCE_3' are some of the important driving factors.

Segment:5:Region Related data:

→ Regional Info: Defaulter rate is highest when REG_REGION_NOT_WORK_REGION=0 i.e. permanent address and working address is same

Segment:6:Housing Info:

→ All of the features considered have very high (47-70%) missing data percentage. Hence all these features can be dropped. Plot of 'NAME_HOUSING_TYPE' vs 'TARGET' shows that

→i. Most of the applicants live in House/Apartment ii. Applicants living with their parents or in rented apartment have higher rate of default.

Segment:7:Education and occupation info:

Most of the applicants are working.

→ Applicants on Maternity Leave and Unemployed has highest percentage of Defaulter.

→ Businessman have lowest (0) percentage of Defaulter.

However applicants of income type('Unemployed', 'Student', 'Businessman','Maternity leave') are very few in the dataset to contribute in the analysis.

Segment:8: Document submitted by Applicants:

→FLAG_DOCUMENT_2','FLAG_DOCUMENT_3',...,'FLAG_DOCUMENT_21' for this segment. Majority of the applicants did not submit any documents apart from DOCUMENT_3. FLAG_DOCUMENT_3 has similar impact on defaulters and non-defaulters. Hence these columns can be dropped.

- **Top 10 Correlations for Non-Defaulters**
- (YEARS_BUILD_AVG, YEARS_BUILD_MEDI)
- (OBS_30_CNT_SOCIAL_CIRCLE, OBS_60_CNT_SOCIAL_CIRCLE)
- (FLOORSMIN_AVG, FLOORSMIN_MEDI)
- (FLOORSMAX_AVG, FLOORSMAX_MEDI)
- (ENTRANCES_AVG, ENTRANCES_MEDI)
- (ELEVATORS_AVG, ELEVATORS_MEDI)
- (COMMONAREA_MEDI, COMMONAREA_AVG)
- (LIVINGAREA_AVG, LIVINGAREA_MEDI)
- (APARTMENTS_MEDI, APARTMENTS_AVG)
- (BASEMENTAREA_AVG, BASEMENTAREA_MEDI)

- **Top 10 Correlations for Defaulters**
- (OBS_60_CNT_SOCIAL_CIRCLE, OBS_30_CNT_SOCIAL_CIRCLE)
- (BASEMENTAREA_AVG, BASEMENTAREA_MEDI)
- (YEARS_BUILD_AVG, YEARS_BUILD_MEDI)
- (COMMONAREA_MEDI, COMMONAREA_AVG)
- (FLOORSMIN_AVG, FLOORSMIN_MEDI)
- (NONLIVINGAPARTMENTS_MEDI, NONLIVINGAPARTMENTS_AVG)
- (LIVINGAPARTMENTS_MEDI, LIVINGAPARTMENTS_AVG)
- (NONLIVINGAPARTMENTS_MEDI, NONLIVINGAPARTMENTS_MODE)
- (FLOORSMAX_AVG, FLOORSMAX_MEDI)
- (ENTRANCES_AVG, ENTRANCES_MEDI)

Top 5 important columns:

- Family info: ('CNT_FAM_MEMBERS', 'CNT_FAM_MEMBERS‘)
- Education and occupation info(‘NAME_INCOME_TYPE,”OCCUPATION_TYPE’)
- CODE_GENDER
- DAYS_BIRTH
- 'EXT_SOURCE_1' and 'EXT_SOURCE_3'

Data Imbalance:

- Data imbalance ratio:
- This data is highly imbalanced as number of defaulter is very less in total population. Data Imbalance Ratio
- Defaulter : Non-Defaulter = 8 : 92 = 2 : 23

Analysis on previous_application

- For the Exploratory data analysis, mentioned steps have been followed.

--> **Import Modules**

--> **Read the dataset**

--> **Data Cleaning**

- Missing value handling
- Type Casting
- Fixing Rows and Columns - removing unncessary rows/columns (through missing value handling and correlation)
- Handling Outliers

→ **Univariate Analysis,Bivariate and Multivariate analysis**

INSIGHTS

- columns with 50% or more missing data can be dropped.
- There are feature columns in the dataset that are highly correlated to each other. Which means both will have similar impact on the target value. Those features can be removed before feeding this data to a model to avoid collinearity.
- Following columns should be converted to integer. DAYS_FIRST_DRAWING float64 DAYS_FIRST_DUE float64 DAYS_LAST_DUE_1ST_VERSION float64 DAYS_LAST_DUE float64 DAYS_TERMINATION float64

- This categorical column has only 0 and 1 and hence can be converted into integer column. NFLAG_INSURED_ON_APPROVAL float64.
- 7% of the previously approved loan applicants that defaulted in current loan
- 90 % of the previously refused loan applicants that were able to pay current loan
- This dataset is highly imbalanced
- The applicants whose previous loans were approved are more likely to pay current loan in time, than the applicants whose previous loans were rejected. NAME_CONTRACT_STATUS is an important feature.

- ‘SCO’, ‘LIMIT’ and ‘HC’ are the most common reason of rejection.
- Most of the people did not request insurance during previous loan application.
- For “Cards” defaulter percentage is highest (17%). ‘**NAME_PORTFOLIO**’ is an important feature for analyzing ‘TARGET’ variable.
- 15% loan applicatiant defaulted for AP+ (Cash Loan). ‘**CHANNEL_TYPE**’ is an important feature for analyzing ‘TARGET’ variable.
- Highest percentage (17%) of default cases is for ‘Card Street’. ‘**PRODUCT_COMBINATION**’ is an important driving factor.