Credit-EDA Assignment

By Manavalla Dilip Kumar

Index

1.Introduction

- This is a case study on Banks Credit Analysis to apply EDA, Explore and Extract insights from the provide data given.
- The analysis is to understand how data is used to reduce the risk factors which are majorly effecting.
- The provided datasets:-
  - Application_data.csv: with all users with difficulties and all other cases.
  - Previous_application.csv: with the approved, cancelled, refused and unused offer data and different cases.
  - Description_data.csv: Is an helping data set with meta data of all columns in both application_data and previous_application data.

2. Read & understand Data

- Here is the fist step we do when starting the EDA.
  - Imports: import all the required packages that is required for the EDA on data sets.
  - Load Data: Load all the required data sets for your analysis and check the shape and sizes so to make an idea of how to work on.
  - Helping Functions: Create functions or import functions required for EDA and make them initialize so that they can be reduce the amount of code and time.

3.Filtering Data

- Based on the requirement we have to filter the columns required for the EDA and make sure those columns set the maximum importance to Analysis.
- Understand the domain to have a good idea on data and filtering.
  - Appication_data.csv : out of 122 columns I find 46 are use full by relating the actual motto of EDA .
  - Previous_data.csv: here also I applied the same where out of 37 columns I find 21 are suitable for analysis.

4.Handling Missing Values $ Data Imputation

- In the datasets there will be many null values and wrongly entered values which in further analysis creates wrong results.
- So to overcome we analyse the data and based on the data and its type we perform different operations like:
  - Removing null values
  - Replacing null values
  - Making changes to data such that suitable for EDA

  We also call it as data cleaning .

5.Handling Outliers

- Here in this case where we draw different graphs on categorical and numerical data for removing the far values or most least occurrence values.
- In the code I used a different functions to make plot the carts and identify the outliers.
  - If there are outliers we have to make remove the outlier content so that the data we have will give meaning full results

6.Data Conversion

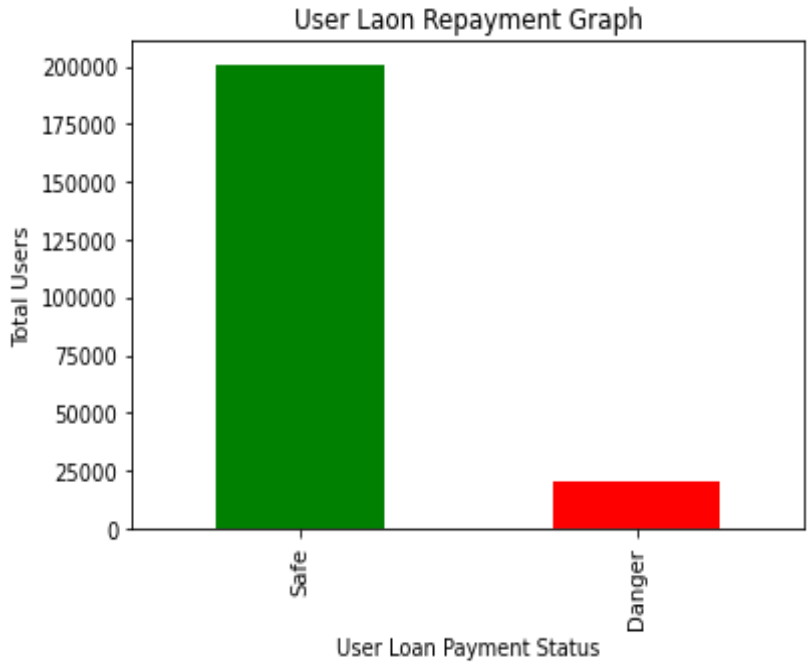
- To make the EDA more flexible and to get insights we need to split the data from ranges of numerical to make them into bins.
- The main motto of data conversion is to make the data neat and understable and easily plot able.
- Here based on the requirement we convert:
  - numerical to categorical data
  - Categorical to numerical data
  - Ordered to unordered
  - Unordered to ordered….etc

7.Analysis

- As per his Case study is on Banks Credit data we majorly focus on the defaulter and payer and there chance of occurrence we proceed with SEGMENTATION analysis.
- It is where we prior on segmenting the data for comparative analysis.
- Here we Perform three types of analysis:
  - Uni-variate analysis
  - Bivariate analysis
  - Multivariate analysis

7.Analysis

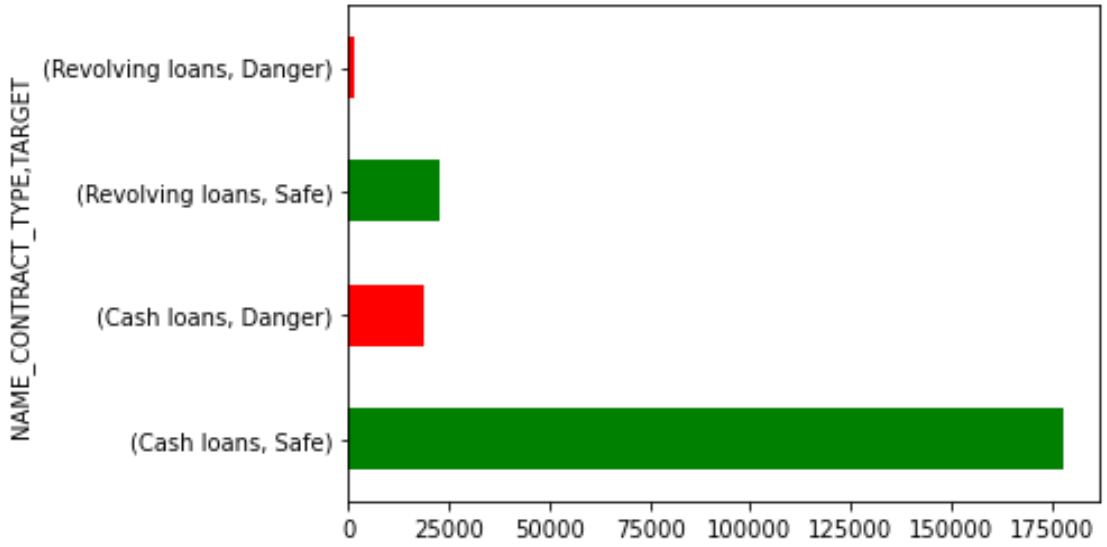
- Uni-variate analysis: The analysis which we perform on single column data is called uni-ariate analysis.
  - We majorly use bar plots, pie charts, boxplots to analyse the single columns.



  - From the graph we can say that the majority of users are repayers.

7.Analysis

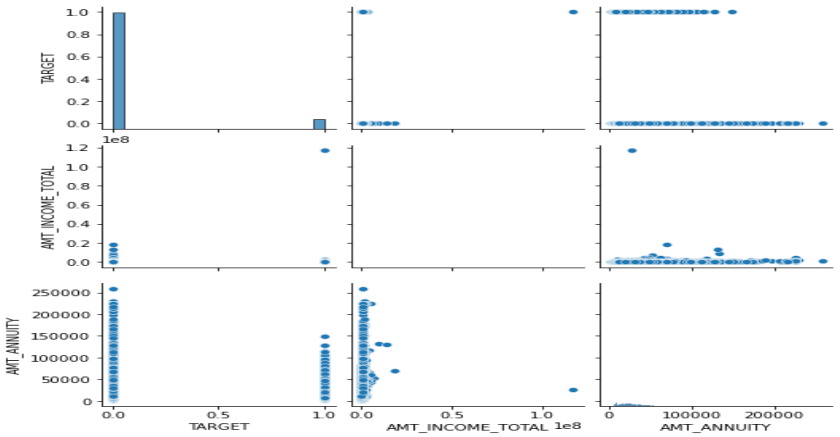
- Bivariate analysis: This where we start relating to other columns and get insights.
- We majorly use scatterplots, bar plots for representing associations.



- Based on the graph we can say that majority users consider cash loans and cash loans also high in default users.

7.Analysis

- Multivariate Analysis: The analysing of data with more that two columns at a time is called multivariate analysis.
- We use heat maps, subplots, pair plots for plotting multivariate analysis.

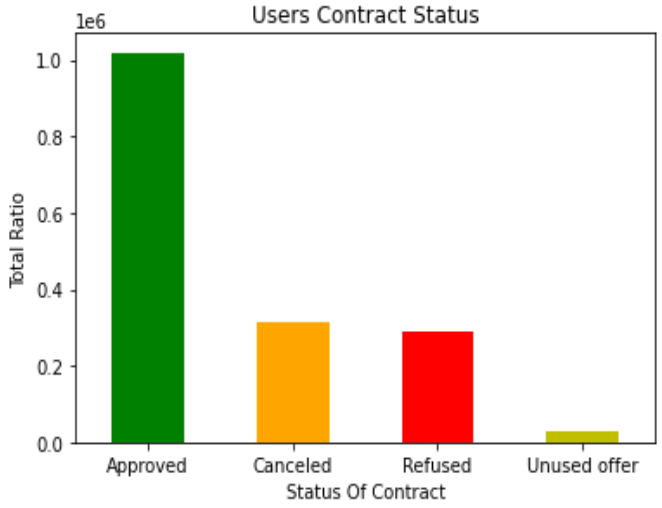


- Based on the subplot we can clearly see that amt_annuity and target are quite far and there is no much relation between amt annuity and amt income.

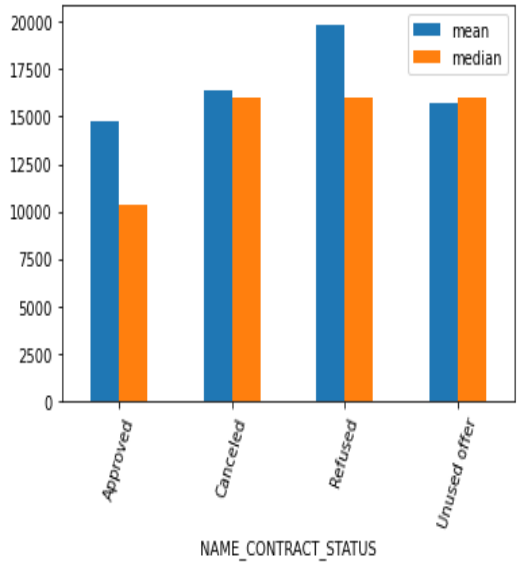
7.Analysis

- Few other plots as example :

Uni-variate



Bivariate



Multivariate

8.Conclusion

- The Majority of users falls under re-payers and few of them are defaulters.
- Users Mostly Prefer Cash Loans than Revolving loan as we can say that the revolving loan will have high interest rates.
- The Female users are mostly taking the loans than male.
- Default users are more in cash flow name_contract_type
- Out of all loan applications 55% of loans are getting refused which is to be considered for banks user experience.
- Applicants with more than 7Lks are very less likely to default.
- The most people who apply for loans are either married or single.
- The people of age 20-40 are chances of defaulting.

Thank You.....☺

By Manavalla Dilip Kumar