

Credit EDA Assignment PPT

LONE RISK ANALYSIS

By :-Jaydeep Manna

Date :- 03-12-2023

Introduction

We will develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

Steps to Follow

1. Import Packages.
2. Reading Dataset.
3. Checking Basic Info.
4. Exploratory Data Analysis.
5. Handling Missing Values.
6. Handling Outliers.
7. Handling the Data imbalances.
7. Plotting Correlation

DATA UNDERSTANDING

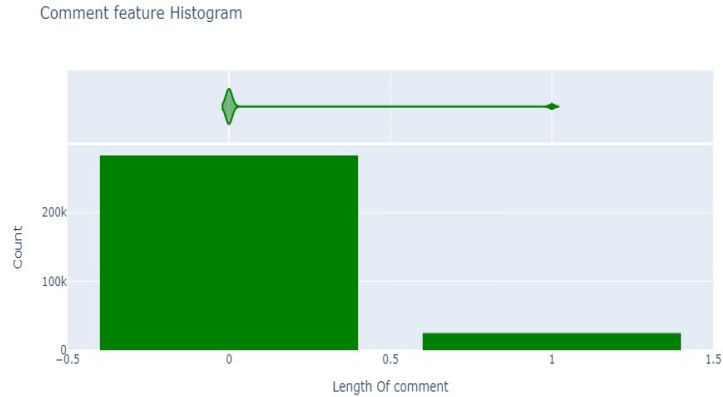
Application.csv-The data is about whether a client has payment difficulties.

Columns_Description.csv - Data dictionary which describes the meaning of the variables.

Previous_Application.csv - It contains the data on whether the previous application had been Approved, Cancelled, Refused or Unused offer.

Exploratory Data Analysis

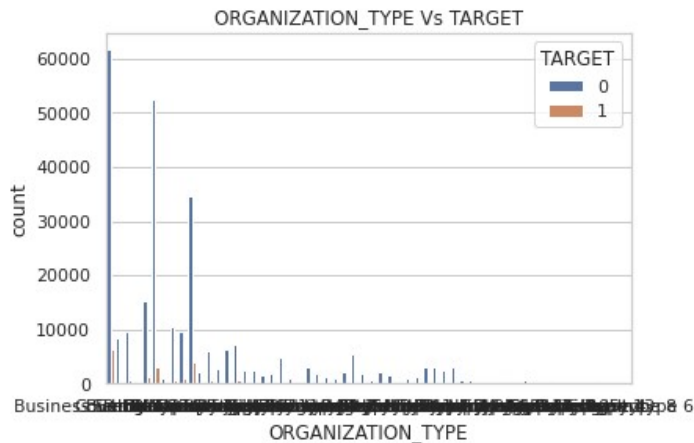
Univariate Analysis



Inference:

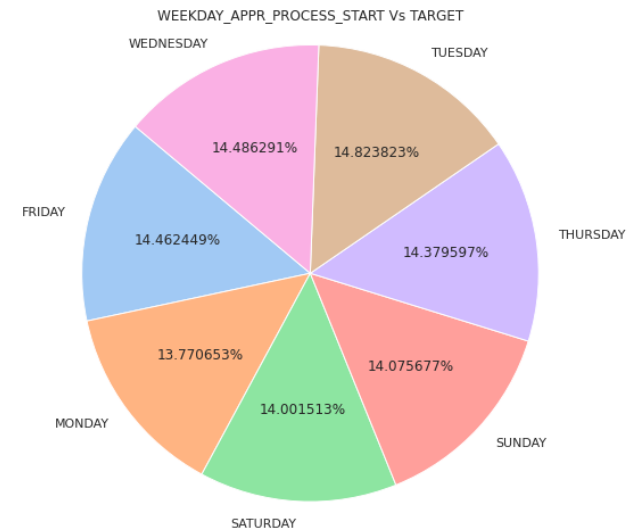
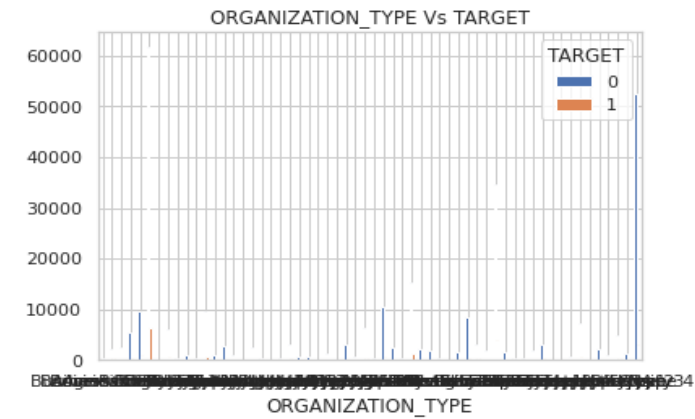
1. Here we can see that there are two class in target col. distribution of data between these classes are not evenly distributed.
2. Because of this data is imbalance.

Bivariate Analysis with categorical columns.



Inference:

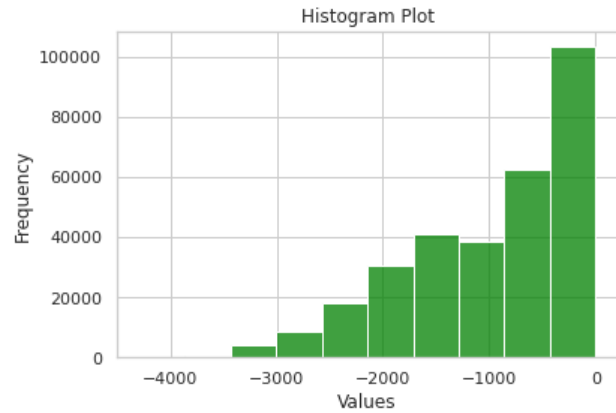
Categorical values are more inclined to the second class which zero. which means that these have higher values in zero



Inference:

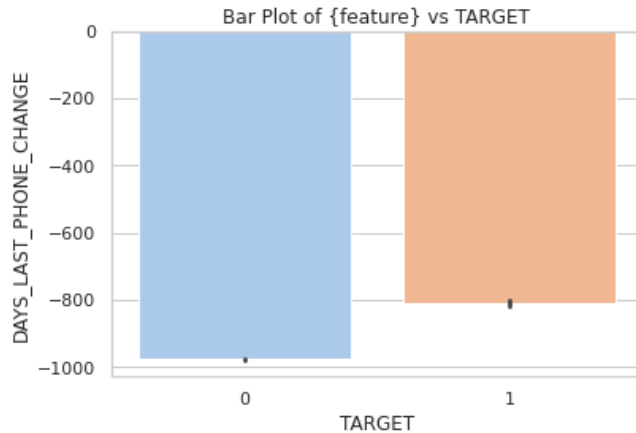
Here we can see that data are balanced between the classes of these category

NUMERICAL VARIABLES BIVARIATE ANALYSIS.



Inferences:

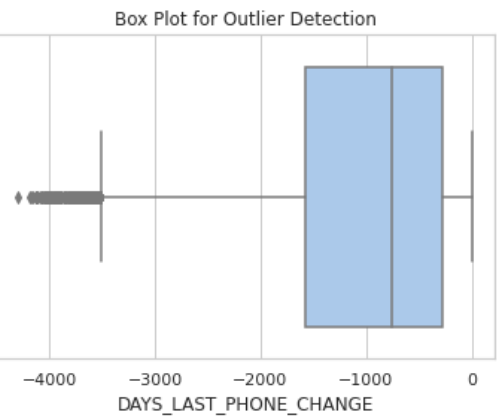
Distribution of these continuous numerical columns are not normalized these are left skewed. Also Some of them have outliers because of these we can see the barr at one side and one single value on other side.



Inferences:

1. Here we can see that numerical columns are more evenly distributed along the target columns than categorical columns.
2. But here also all little bit skew ness towards the zero category of target column

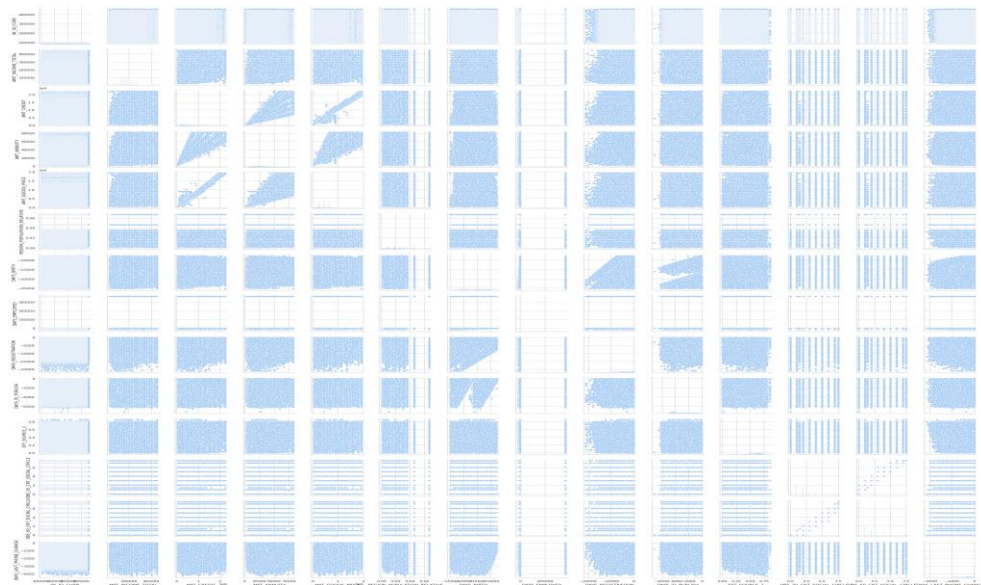
Outlier detection with box plot.



Inference:

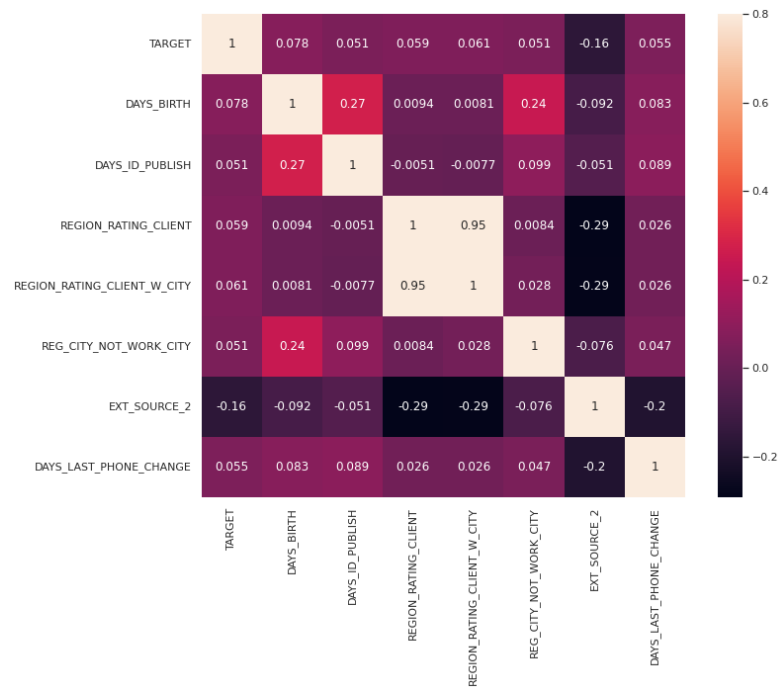
There are features that has outliers for eg: AMT_INCOM_TOTAL,OBS_60_CNT_SOCIAL_CIRCLE

Multivariate Analysis



Inferences:

In the pair plot we can see the scatter plots between multiple independent numerical variables that follows some trend or has some correlation between them.

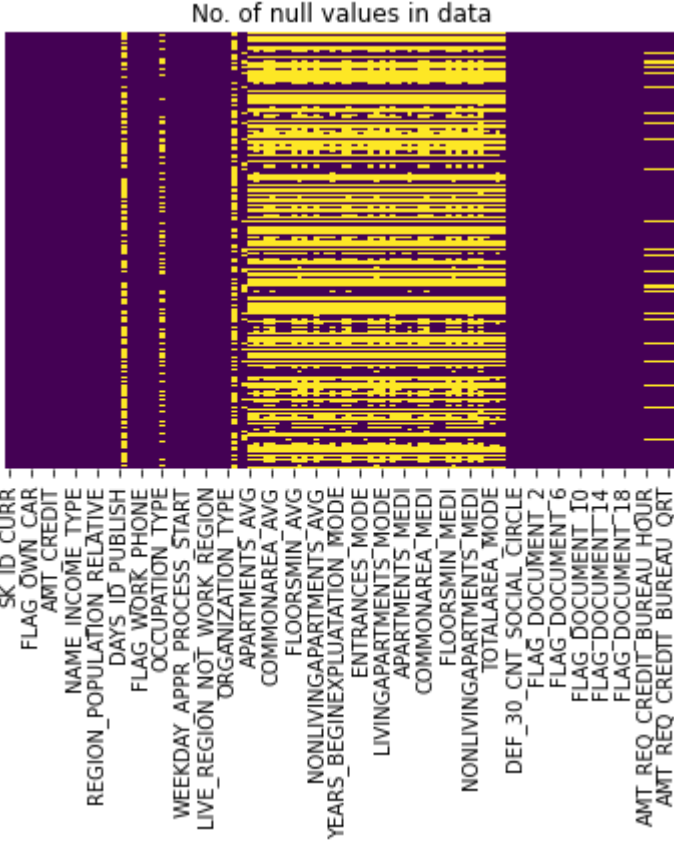


PREVIOUS APPLICATION DATA ANALYSIS.

AMT_ANNUITY : 372235 => 22.29% AMT_CREDIT : 1 => 0.0% AMT_DOWN_PAYMENT : 895844
 => 53.64% AMT_GOODS_PRICE : 385515 => 23.08% RATE_DOWN_PAYMENT : 895844 => 53.64%
 RATE_INTEREST_PRIMARY : 1664263 => 99.64% RATE_INTEREST_PRIVILEGED : 1664263 =>
 99.64% NAME_TYPE_SUITE : 820405 => 49.12% CNT_PAYMENT : 372230 => 22.29%
 PRODUCT_COMBINATION : 346 => 0.02% DAYS_FIRST_DRAWING : 673065 => 40.3%
 DAYS_FIRST_DUE : 673065 => 40.3% DAYS_LAST_DUE_1ST_VERSION : 673065 => 40.3%
 DAYS_LAST_DUE : 673065 => 40.3% DAYS_TERMINATION : 673065 => 40.3%
 NFLAG_INSURED_ON_APPROVAL : 673065 => 40.3% No. Of Columns: 16

Drop Columns

AMT_ANNUITY : 372235 => 22.29% AMT_DOWN_PAYMENT : 895844 => 53.64% AMT_GOODS_PRICE : 385515 => 23.08% RATE_DOWN_PAYMENT : 895844 => 53.64% RATE_INTEREST_PRIMARY : 1664263 => 99.64% RATE_INTEREST_PRIVILEGED : 1664263 => 99.64% NAME_TYPE_SUITE : 820405 => 49.12% CNT_PAYMENT : 372230 => 22.29% DAYS_FIRST_DRAWING : 673065 => 40.3% DAYS_FIRST_DUE : 673065 => 40.3% DAYS_LAST_DUE_1ST_VERSION : 673065 => 40.3% DAYS_LAST_DUE : 673065 => 40.3% DAYS_TERMINATION : 673065 => 40.3% NFLAG_INSURED_ON_APPROVAL : 673065 => 40.3% No. Of Columns: 14



Thank you