CREDIT EDA

Sharhedha Raghavan

Jallepalli Prerna

Introduction

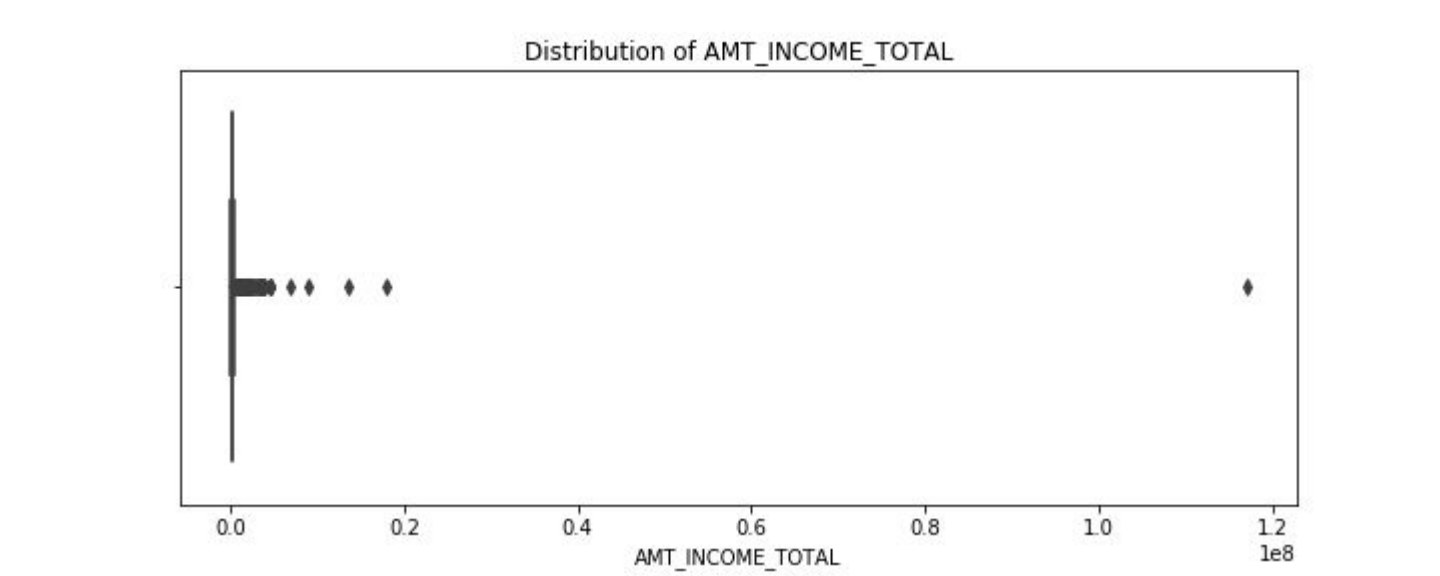
Given two datasets perform Exploratory Data Analysis based and infer how likely would a client default his loan based on some factors.

The two datasets are:

1. Application Data → Loan Application of a particular client and other client data
2. Previous Application Data → Given a loan id if he had any previous loans and analysis related to that

Given these datasets we need to merge, analize etc.

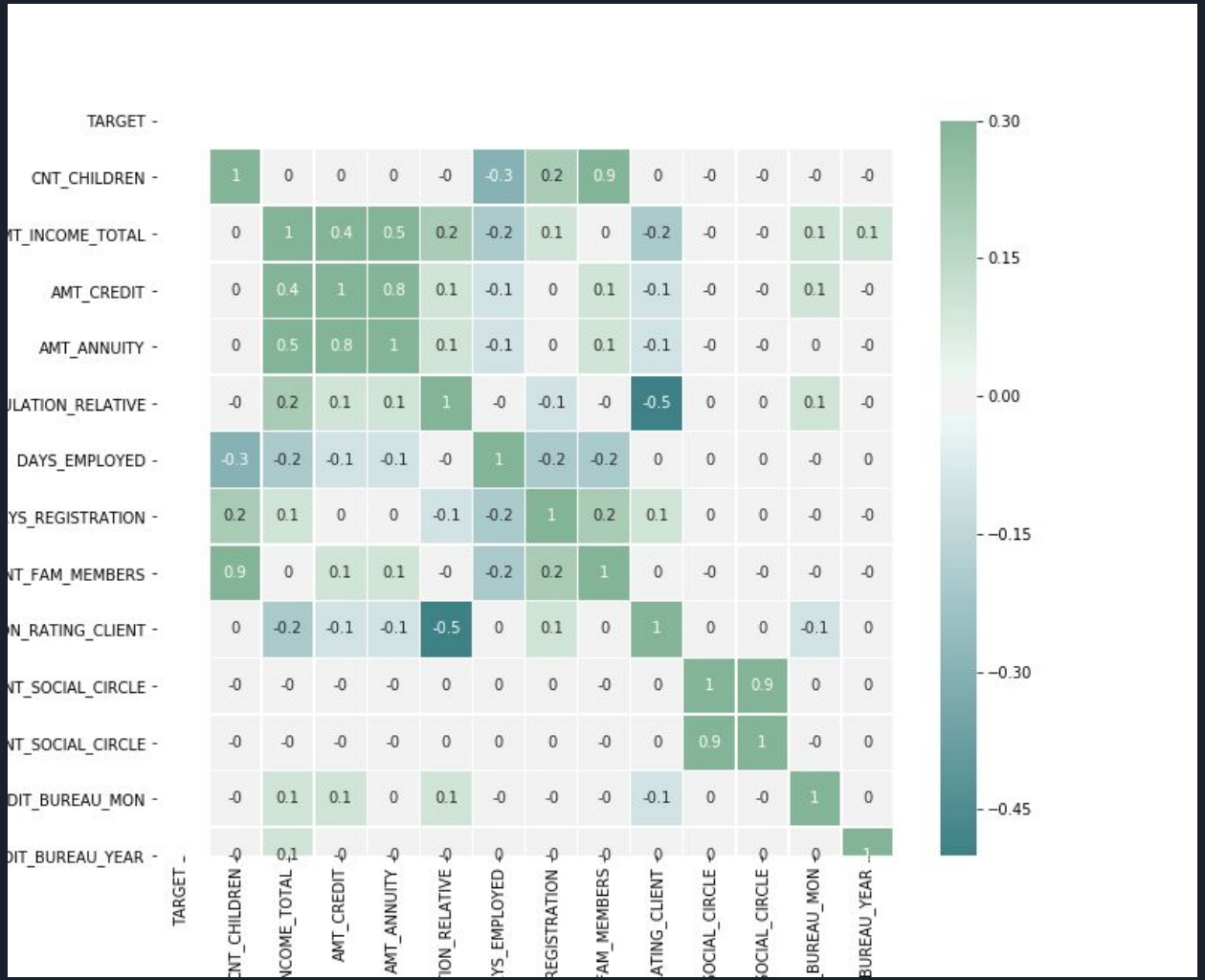
OUTLIER DETECTION



We observe that the data is highly skewed have significant amout of outliers.
Two of the main ways in which we can treat them are either removing the outliers or performing a log transformation. Similarly, Box Plot helps in finding various other outliers.

Finding the Imbalance Ratio

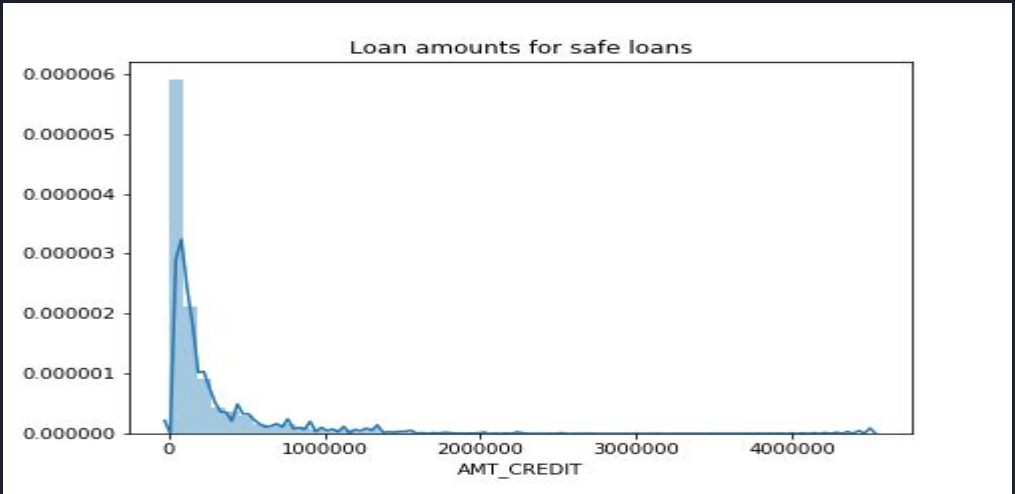
We see that there is a lot of Imbalance in the Target Variable.
Targert-0 has 92% while Target-1 has 8%
Which implies that the people who have defaulted are just 8% while the rest 92% of them have paid without any difficulties

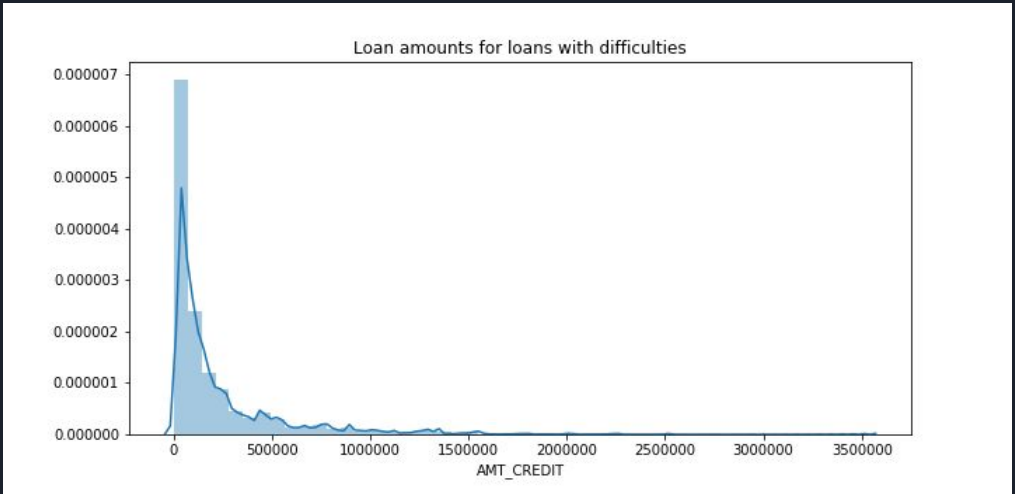


The correlation plot helps in finding the correlated variables between two numerical variable.

This is a correlation plot for Target variable-1

(For clear image refer to the notebook and the attached images)

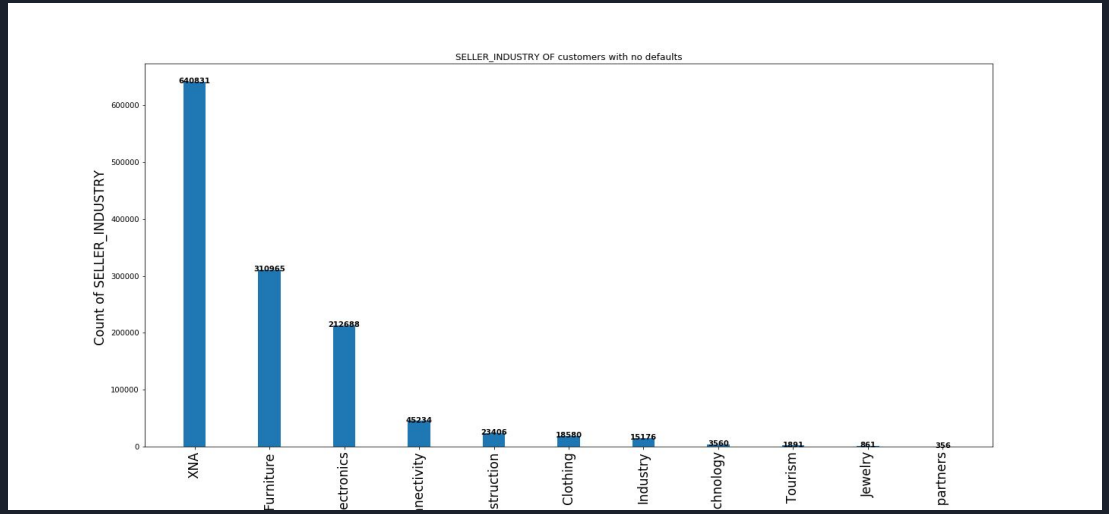


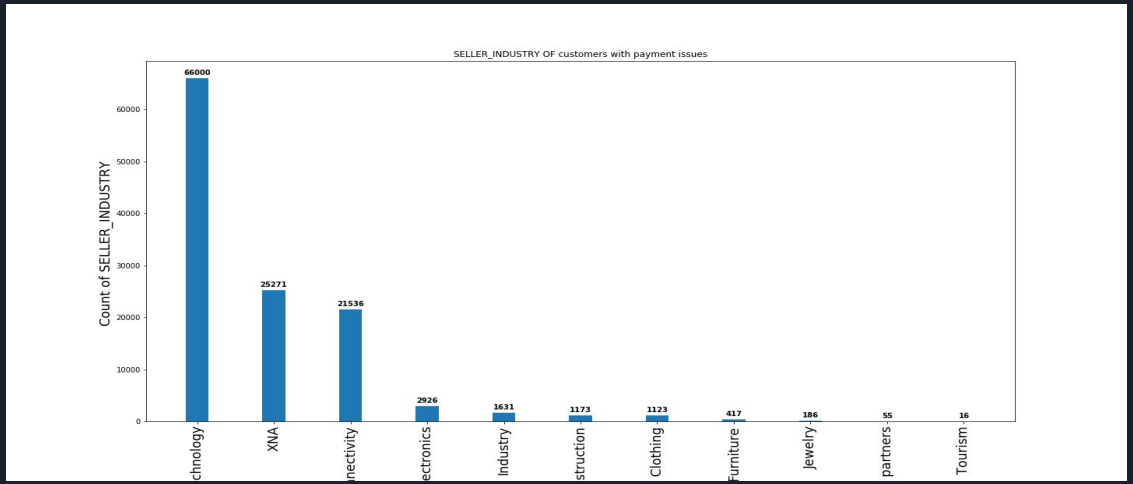


The distribution of data is same for both types of customers.No particular loan amount carries additional risks.

The data has very high outliers compared to the median value.We noticed that all loan amounts have the same risk in previous analysis.

(For clear image refer to the notebook and the attached images)





Auto loans and Connectivity are two high risk areas. Auto loans are mostly hard to recover Connectivity sector has a 50% failure rate

(For clear image refer to the notebook and the attached images)

RECOMMENDATIONS AND INFERRENCES

1). If the loan credit is higher the payment period is also longer

2). If the loan amount is high the annuity is high

3). If there is high income of the client then higher is the loan credited

4). Revolving type loans have a significantly higher risk of non payments.It is higher than even the paid loans of revolving type

5). Loan given to purchase vehicles have a high risk of not being repaid next risky segment is for photo equipment and mobiles with nearly 25 % unreturned loans . Furniture, XNA loans are low risk

6). All yield groups have nearly 10% risk which is consistent with the data distribution. As expected low risk groups have a much better percentage of returns.

7). New clients are riskier. They have a nearly 50 % chance of not repaying. Repeated clients are safer and have around 3% non returned loans

8). Auto loans and Connectivity are two high risk areas. Auto loans are mostly hard to recover Connectivity sector has a 50% failure rate

9). Card loans have the highest risk at around 30% recovery difficulties are detected. Cash loans are safer and POS Loans have highest number of recovery difficulties

10). The distribution of data is same for both types of customers.No particular loan amount carries additional risks

The data has very high outliers compared to the median value.We noticed that all loan amounts have the same risk in previous analysis.

11). For having better results and better inferences, it’s good to deal with Imbalance data.

THANK YOU