

Micro credit defaulter

Submitted by:

Sinchana

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

References:

<https://ideas.repec.org/>

<https://slideshare.com/>

Research papers:

<https://www.researchgate.net/>

Professionals:

**INTRODUCTION**

* **Business Problem:**

MFI is a financial institution that gives financial aid to the economically low class people. As a initiative of this cause, they are tying up with mobile financial services which is telecom provider which provides low cost services to its customers as part of this cause.Partnership is mainly to facilitate low economy background people to get credit on the service availed from their client. So in current situation, client has to predict who are potential customers who are credit defaulters.ie, fail to pay the principal amount within specified period of time.

* **Conceptual Background of the Domain Problem**

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. In order to make sure this underserved population has a positive loan experience, company makes use of a variety of alternative data include transactional information--to predict their clients' repayment abilities.

* **Review of Literature**

This is problem related to Financial services & banking sector.Microfinance institutions play a major role in economic development in many developing countries. However many of these microfinance institutions are faced with the problem of default because of the non-formal nature of the business and individuals they lend money to. This study seeks to find the determinants of credit default in microfinance institutions. With data on 2631 successful loan applicants from a microfinance institution with braches all over the country we proposed a Binary logistic regression model to predict the probability of default. We found the following variables significant in determining default: Age, Gender, Marital Status, Income Level, Residential Status, Number of Dependents, Loan Amount, and Tenure. We also found default to be more among the younger generation and in males. We however found Loan Purpose not to be significant in determining credit default. Microfinance institutions could use this model to screen prospective loan applicants in order to reduce the level of default.

* **Motivation for the Problem Undertaken**

Since the financial services are providing micro credit to the telecom users through telecom partners, they expect the customers to return the offered credit with in certain span of time. If they failed to pay back the same within stipulated time as given as a deadline,they should be labeled as defaulters. So the objective behind giving credit is to provide financial assistance to the financially degraded individuals. The main motivation behind this initiation is to help the poorer section of the society.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

On analysis of the dataset,we got a unbalanced dataset with more than 80% of the target variables as non defaulters and remaining are defaulters. So before balancing the data we have to look into the dataset for any abnormalities in the data. We have most of the features having right or left skewness. So we will first interpret the data for our Analysis by getting some statistical parameters like mean,median & quantiles which decides the factors that will help predict the target variable. So we will apply the formula to identify & delete the outliers for our better model building.

* **Data Sources and their formats**

Data set is of the indonesian financial services which has multiple independent features like Age on cellular network,Daily amount spend, Average main account balance.etc. The dataset file we are using is of the excel format.

* **Data Preprocessing Done**

1. Loading Dataset
2. Shape of our dataset having 209593 rows and 37 columns
3. Looking at first five records
4. Looking at statistical parameters
5. No Null values in our dataset
6. Categorical values are removed
7. Number of Unique value records
8. Number of zeros in our dataset
9. Removing duplicate data
10. Showing correlation with each other features
11. Maximum cutomers pay their loan
12. Dataset is imbalanced.

* **Data Inputs- Logic- Output Relationships**

The input data provided, helps to understand the behaviour of the customer, their various transaction records, their frequency of transaction during a period of time etc, all these helps to predict the customer’s intension toward the repayment of loan.

* **State the set of assumptions (if any) related to the problem under consideration**

No as such assumption been done related to the circumstances.

* **Hardware and Software Requirements and Tools Used**

Data Science task should be done with sophisticated machine with high end machine configuration. But unfortunately, the machine which I’m currently using is powered by intel core i3 processor with 4GB of RAM. With this above-mentioned configuration, I managed to work with the data set in Jupyter Notebook which help us to write Python codes. As I’m using low configuration machine so it took more time then usual to execute codes. The library used for the assignment are Numpy, Pandas, Matplotlib, Seaborn, Scikit learn

**Model/s Development and Evaluation**

* **Identification of possible problem-solving approaches (methods)**

The data set contain more than 2 lakh data with no null values related to the customer. The dataset is imbalanced. Label 1 has 87.5% of data whereas label 0 has approximately 12.5%. As I went through the dataset, I found lot of outliers and skewness are present in the dataset. The outliers were corrected by replacing them with Q3+1.5(IQR) if it is more than Q3+1.5(IQR). The skewness was also reduced using Log transformation and square root transformation wherever applicable. There were certain columns which had least importance with our target variable, hence those were dropped. After data cleaning and data transformation, data visualization was done to represent data graphically. At last, the most important part was to build model for the data set.

* **Testing of Identified Approaches (Algorithms)**

The algorithms used for getting cross validation score are

1. RandomForestClassifier
2. ExtraTreesClassifier
3. LogisticRegression
4. GradientBoostingClassifier
5. DecisionTreesClassifier
6. BoostingClassifier
7. KNeiboursClassifier
8. SupportVectorClassifier
9. BernauliNB

* **Run and Evaluate selected models**

The algorithm used for hyper parameter tuning and fitting train & test dataset are

1. RandomForestClassifier
2. ExtraTreesClassifier
3. Logistic Regression
4. GradientBoostingClassifier

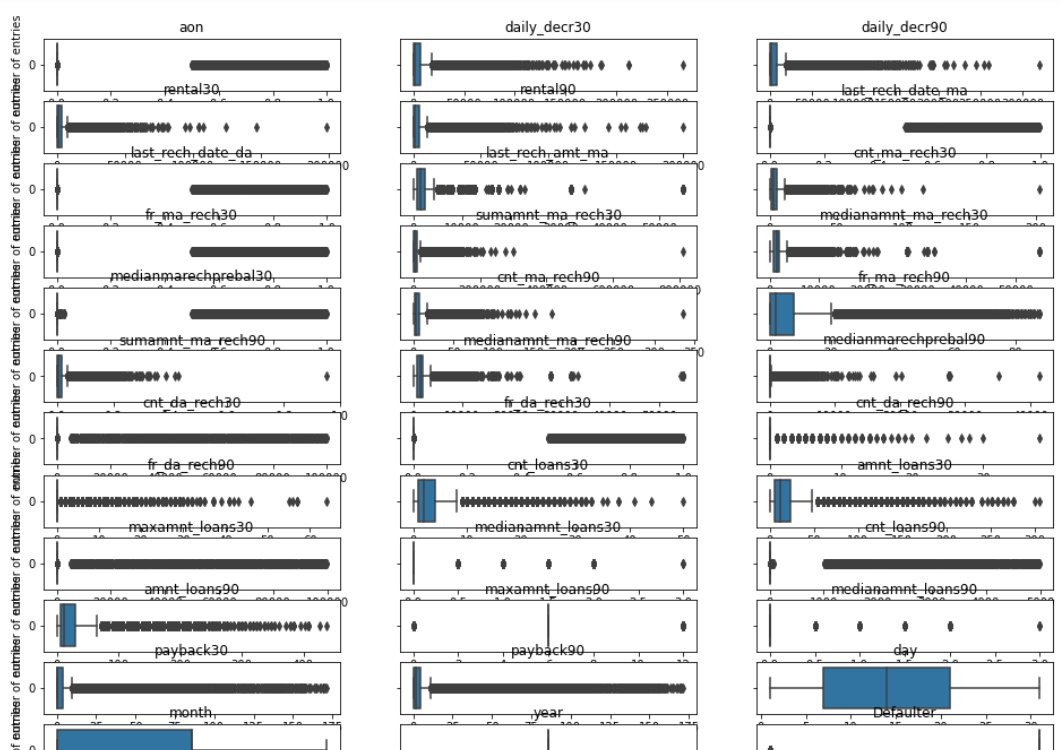
* **Key Metrics for success in solving problem under consideration**

As mentioned earlier, the dataset is unbalanced with 87.5% of label 1 and 12.5% of label 0, which made it clear that, we cannot blindly rely on accuracy score for the prediction as it can lead to biasness. Hence, I have used confusion matrix and AUC ROC curve to determine the accuracy of the model.

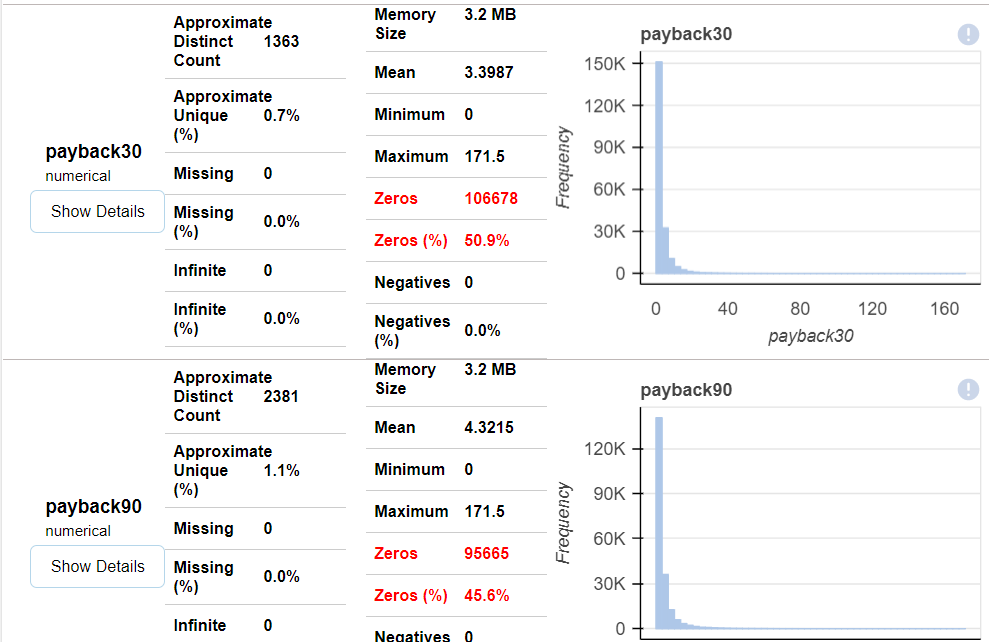
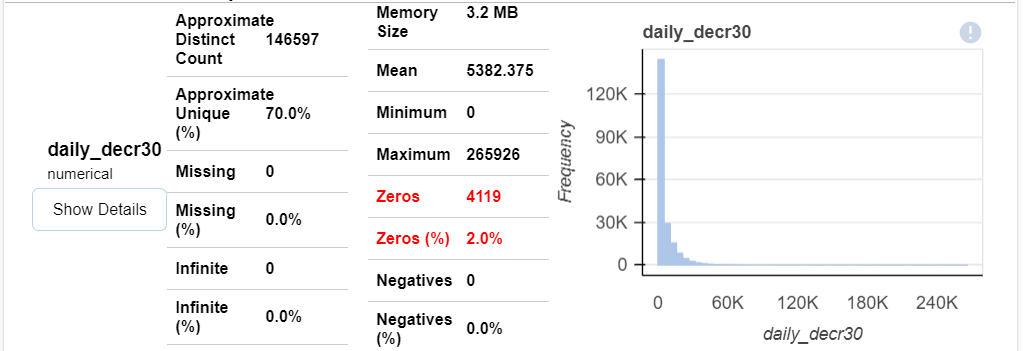
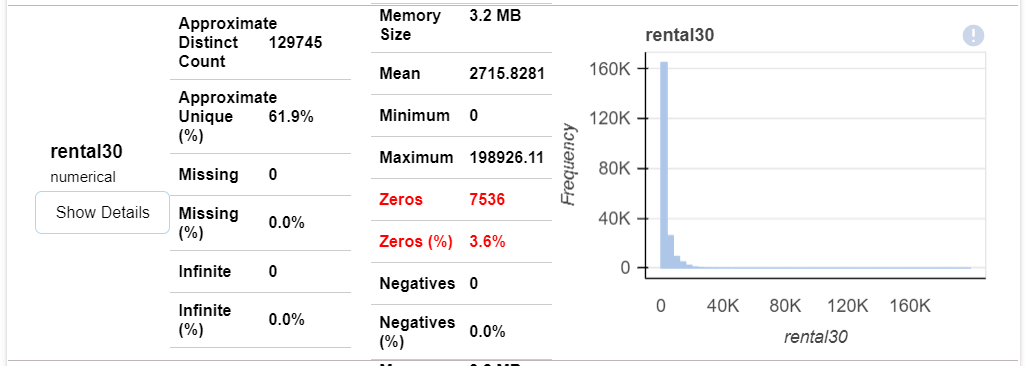
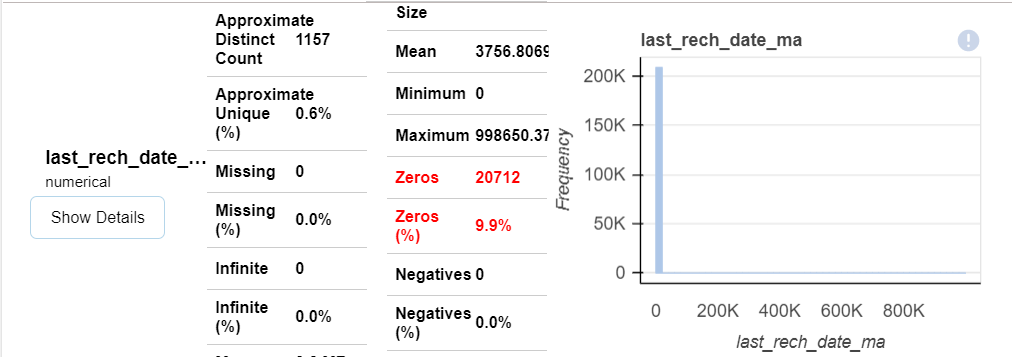
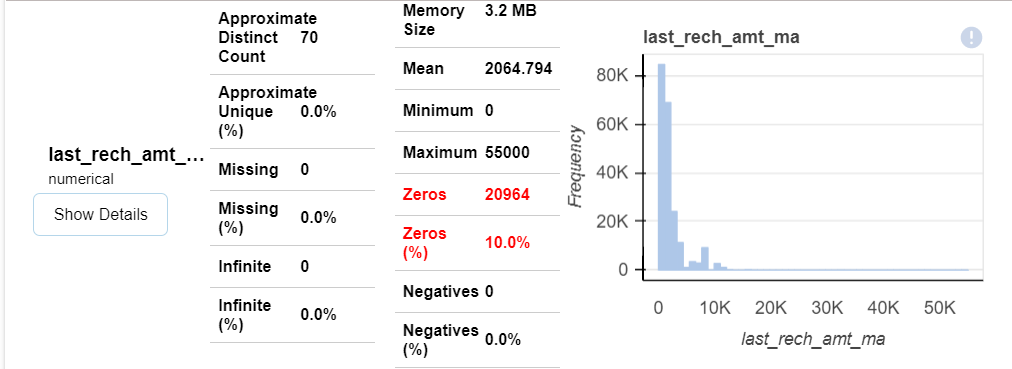
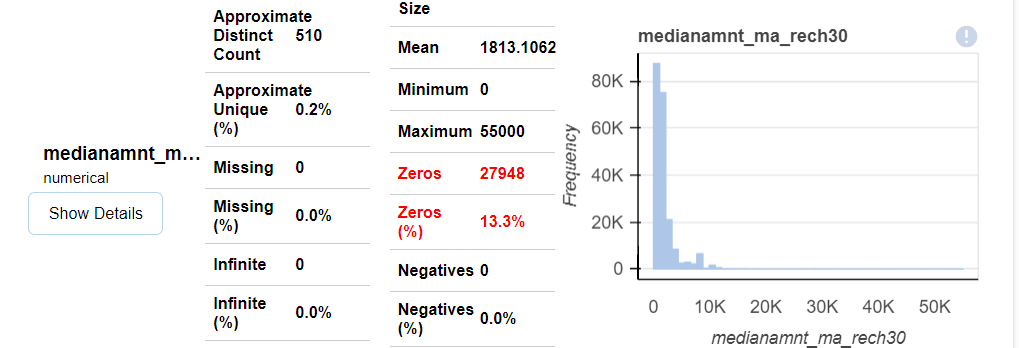
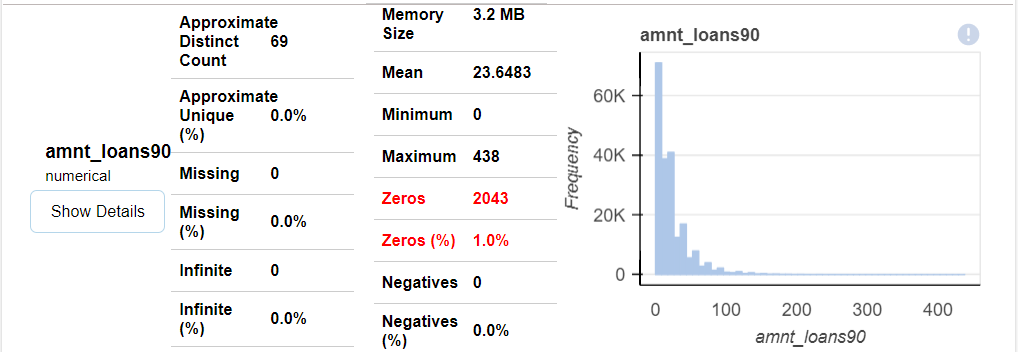
* **Visualizations**

Plots that were plotted to do Univariate, Bivariate,Multi-variate Analysis

Boxplot to know Outliers

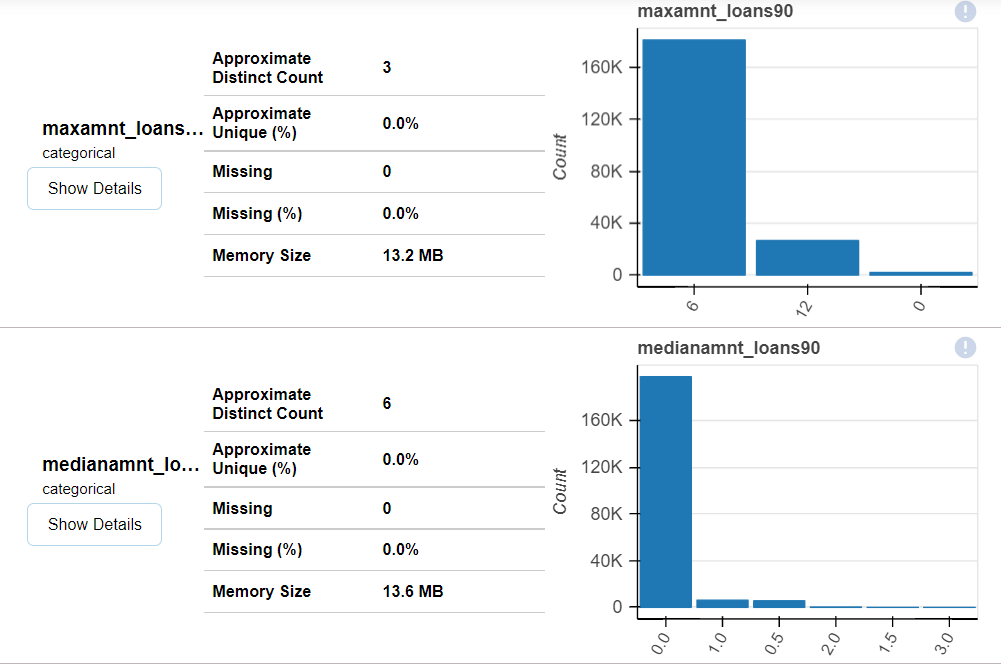


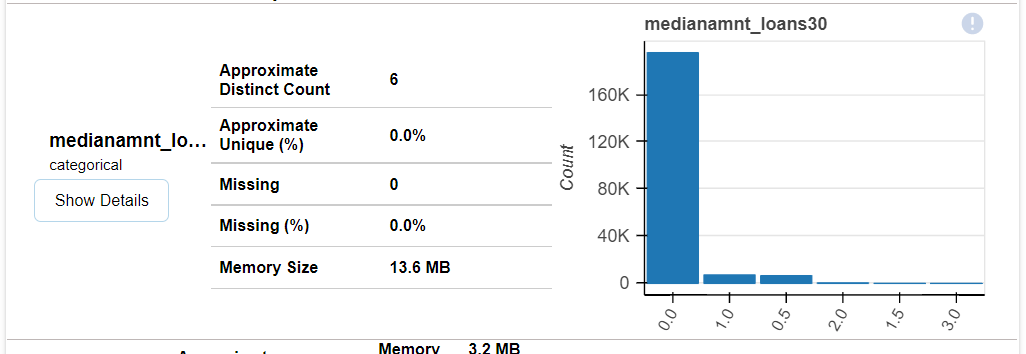
Histogram to know the frequency

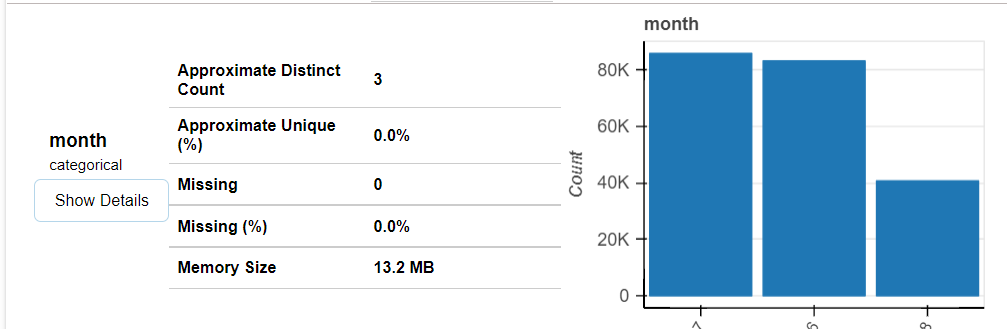


These histograms show that most of our data are left skewed and having high percentage of zeros

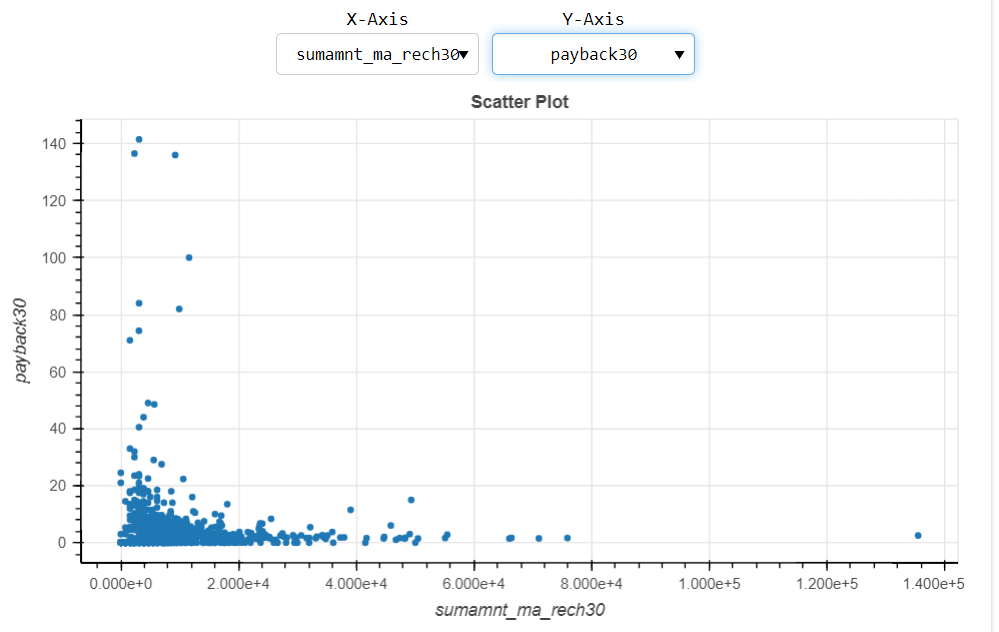
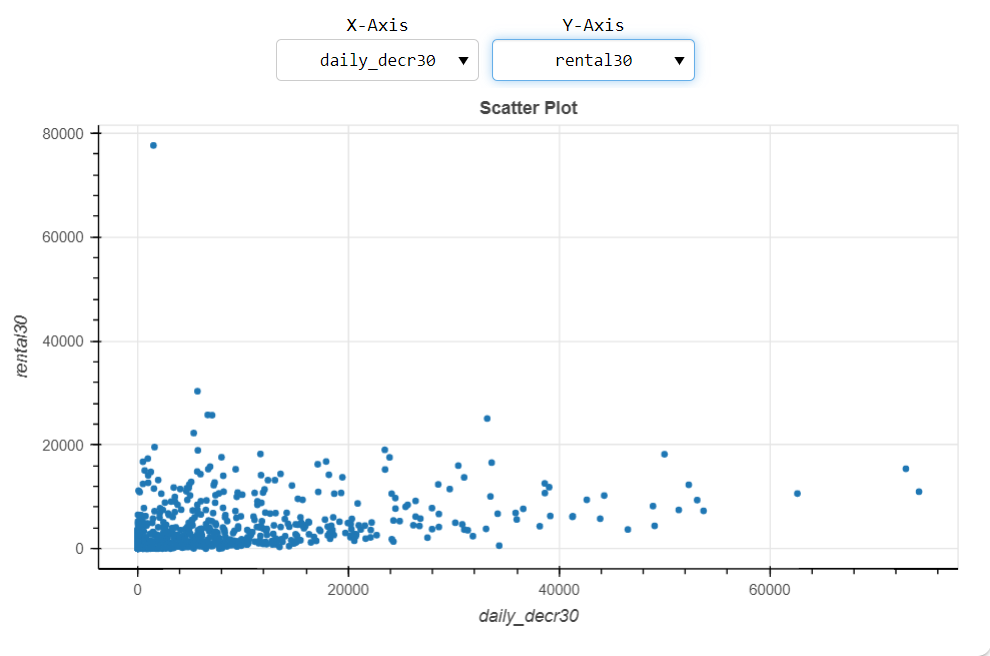
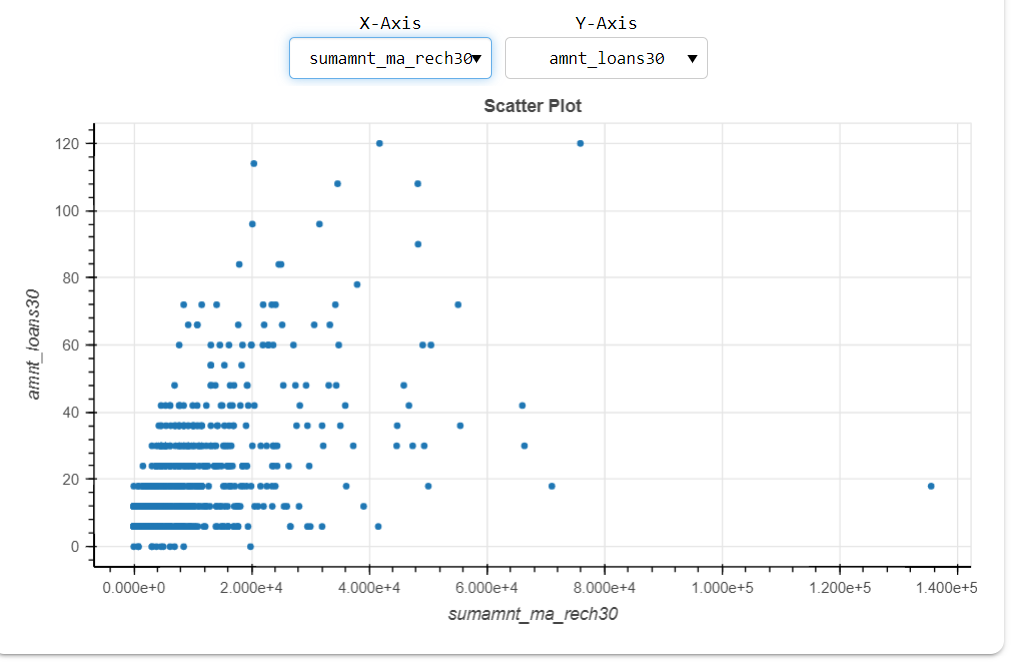
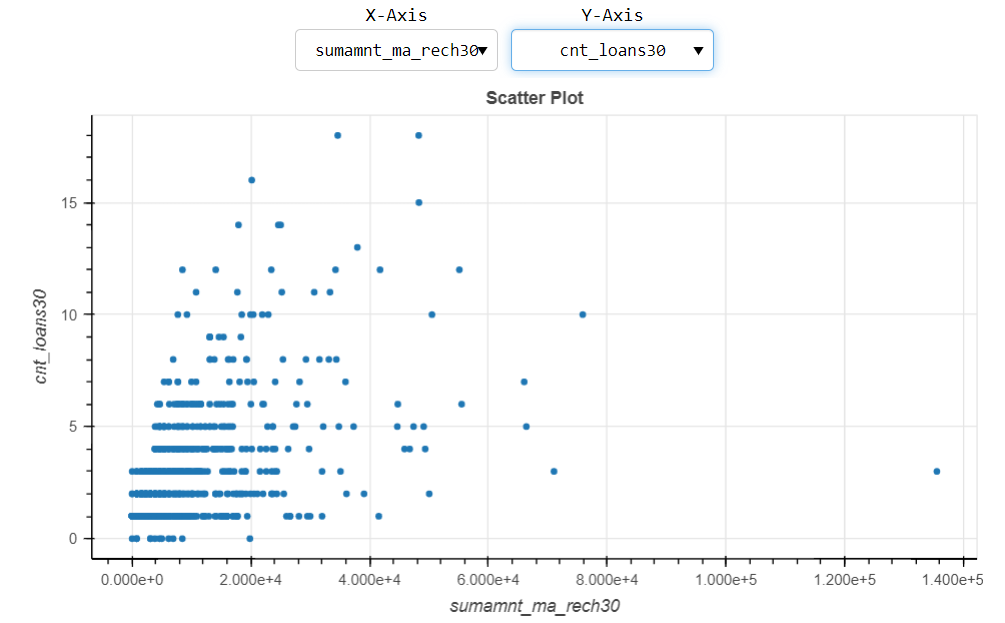
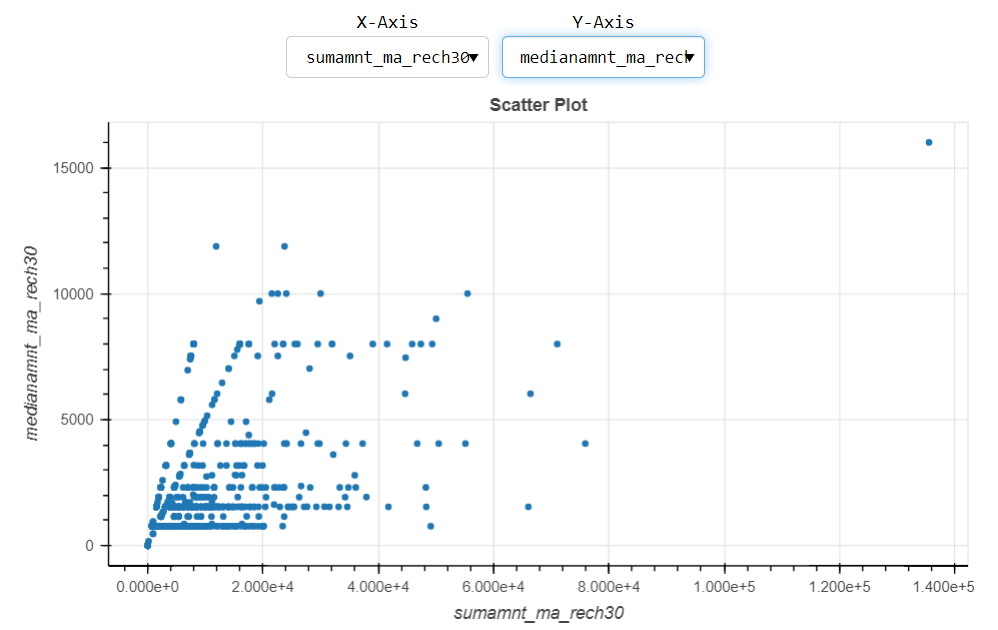
Countplot to know count of the target variable



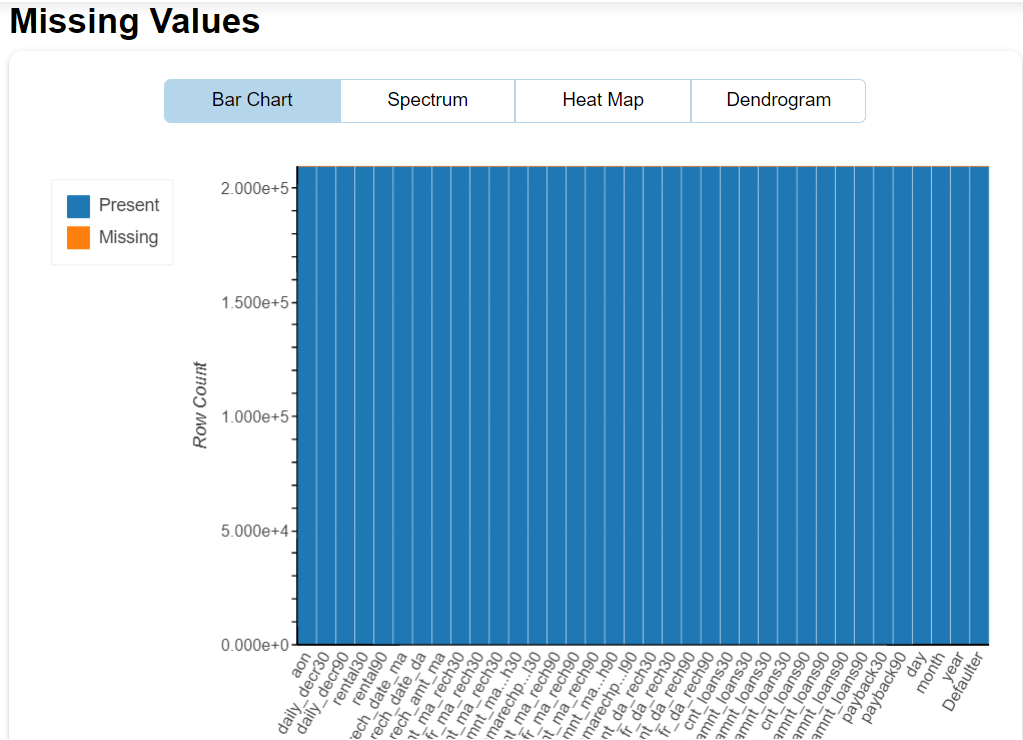




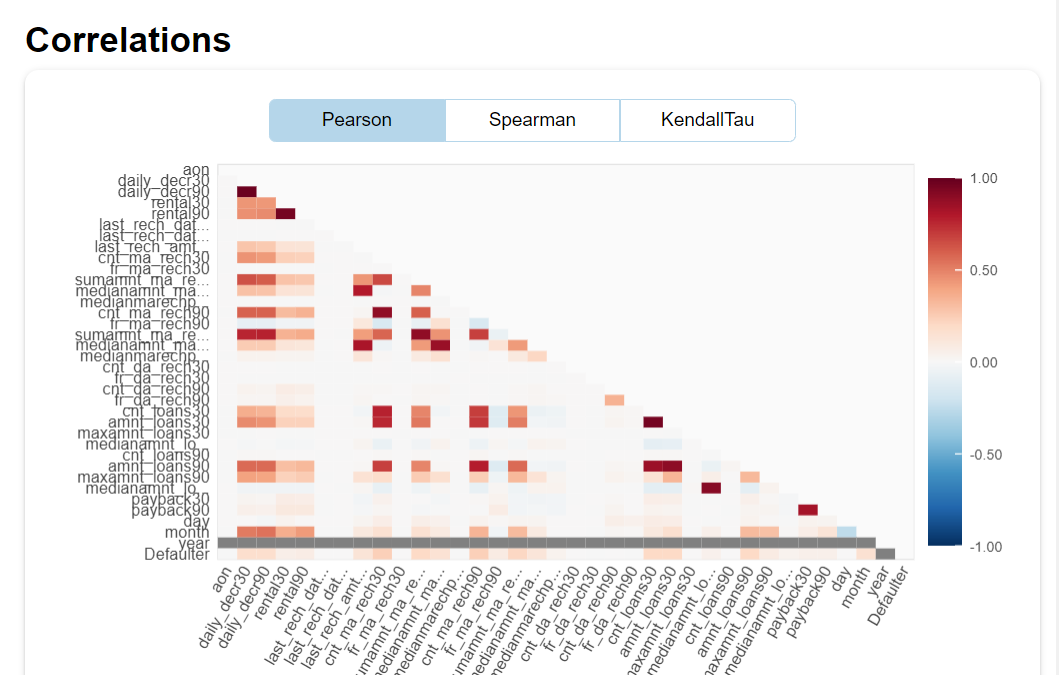
Scatterplot to know the relationship between one variable with other



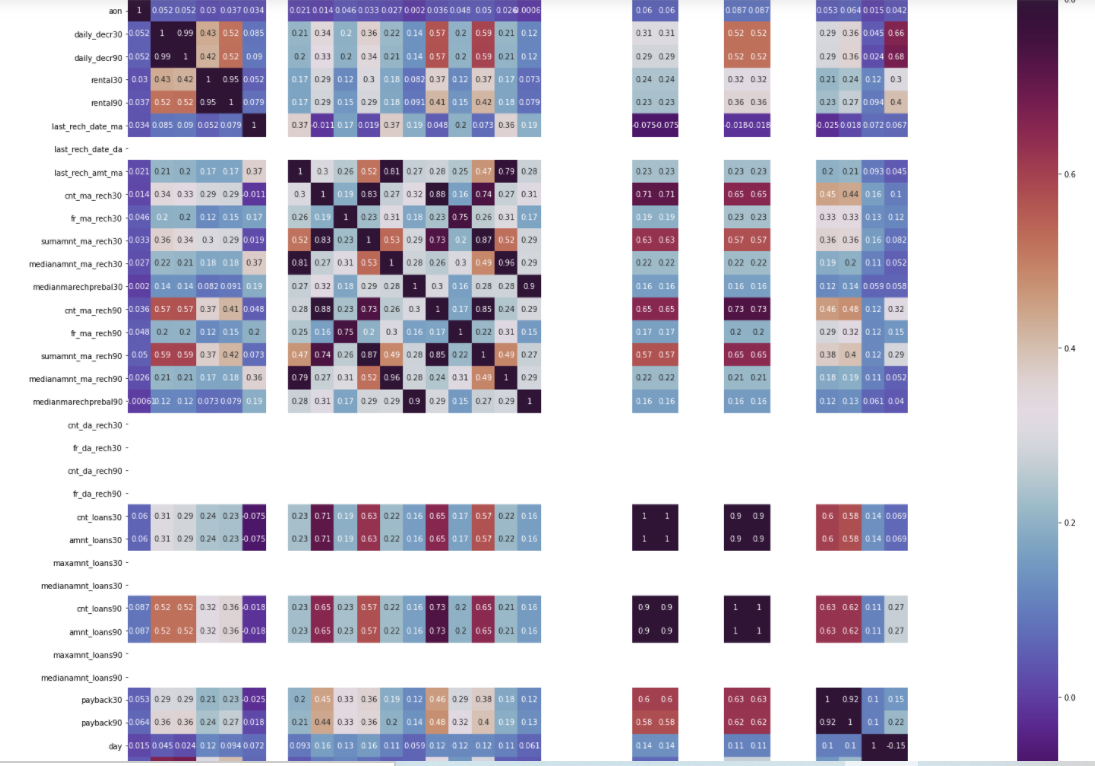
Countplot of Missing Values



Different Heatmaps kendal,pearson etc to know the correlation with target variable



Heatmap to know the correlation of each feature with all other feature



* **Interpretation of the Results**

From the dataset, it clear that most of the customers are inclined to pay the loan as 87.5% of the customer repaid it and only 12.5% of the customers are defaulter.

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

Mostly, the customers have the intension of repaying. There are certain cases, when the customers have no intension of repayment but the number of such customers are few. With the model built, we can certainly determine customers having intension of repayment or not.

* **Learning Outcomes of the Study in respect of Data Science**

The dataset was full of outliers, skewness and unbalanced data which was the biggest challenge to overcome. Hence data cleaning was very important to get proper prediction. I have used Logistic Regression, Gaussian NB and Random Forest Classifier. Among the three algorithms Random Forest Classifier gave the best outcome. As the dataset was unbalanced, the other algorithm may overfit and can come out with wrong prediction whereas Random forest can control overfitting and give best prediction.

* **Limitations of this work and Scope for Future Work**

The solution can be applied to the customer having a transaction history but the model may not perform well with customer having new profile and no transaction history. Nevertheless, the model will perform well with customer having transaction history and can predict whether a person will be a defaulter or non-defaulter. Hence, we can say that this statistical model will be helpful in future for the prediction of micro credit defaulter and non-defaulter customer.