

Micro Credit Defaulter Model

Submitted by:

ASHISH KUMAR DEWANGAN

**INTRODUCTION**

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person’s life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

**Objective**

Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label ‘1’ indicates that the loan has been payed i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been payed i.e. defaulter.

**Exploratory Data Analysis**

* Data Pre-processing and Visualizations

The dataset has 209593 entries where each entry has 36 features and 1 label. 21 features have float values, 13 features have integer values and 3 features have object as their datatype. The features which have categorical values are 'msisdn', 'pcircle', 'pdate'. The feature ‘Unnamed: 0’ is serial number,'msisdn' has nominal values, all the entries in the feature 'pcircle' have duplicate values. Hence, it is safe to drop these two features. At the same time the since the feature 'pdate' is date, it does not contain any other significant information hence, this feature is also dropped. Finally in our dataset all the features have numerical values an in total we have 32 features. Let us move ahead to further stage of pre-processing.

**Univariate Analysis**

The following conclusions can be derived from the first look at our dataset:

1. There are no null values in the dataset.
2. The features : 'aon', 'daily\_decr30', 'daily\_decr90', 'rental30', 'rental90','last\_rech\_date\_ma', 'last\_rech\_date\_da', 'medianmarechprebal30' and 'medianmarechprebal90' have values less than 0. From the data description it is clear that the features should not have negative values. Hence, the negative values are replaced with 0.
3. In most of the columns mean is less than standard deviation that means the data is widely spread and there are many outliers in the dataset.
4. 12.5 % of customers are defaulters.
5. The box plot and violin plot of label vs number of days in the network shows that there are a large number of outliers.
6. The distribution plot of all the dataset showed that the data is right skewed.

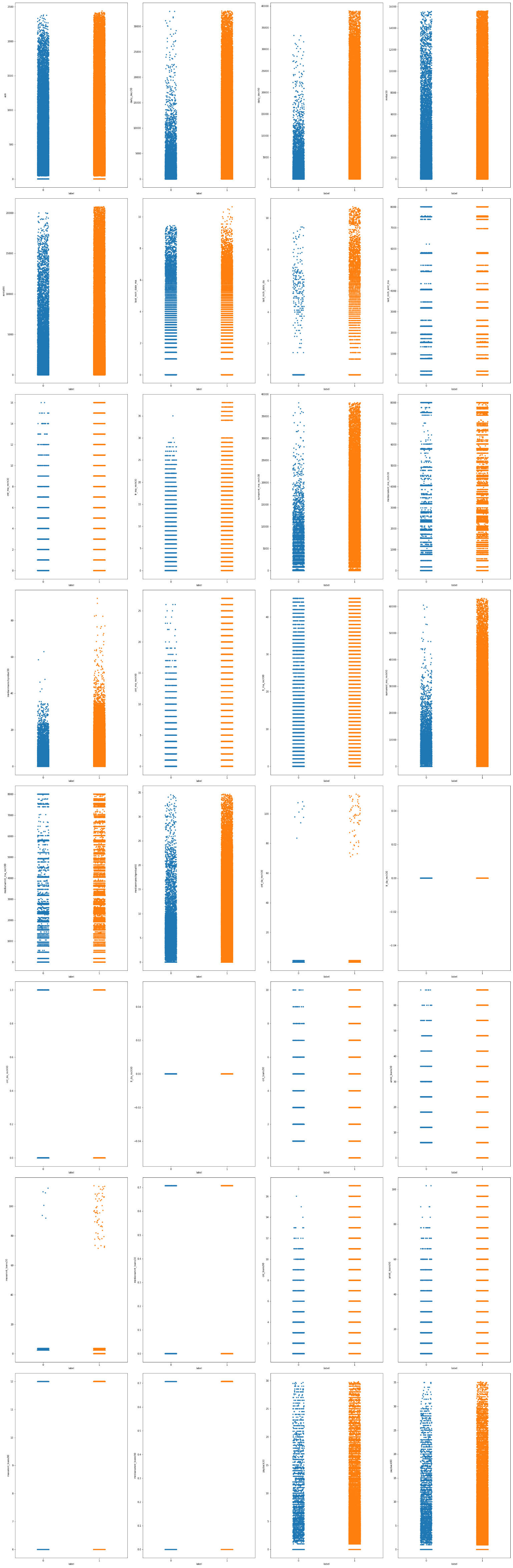
**Data Pre-Processing**

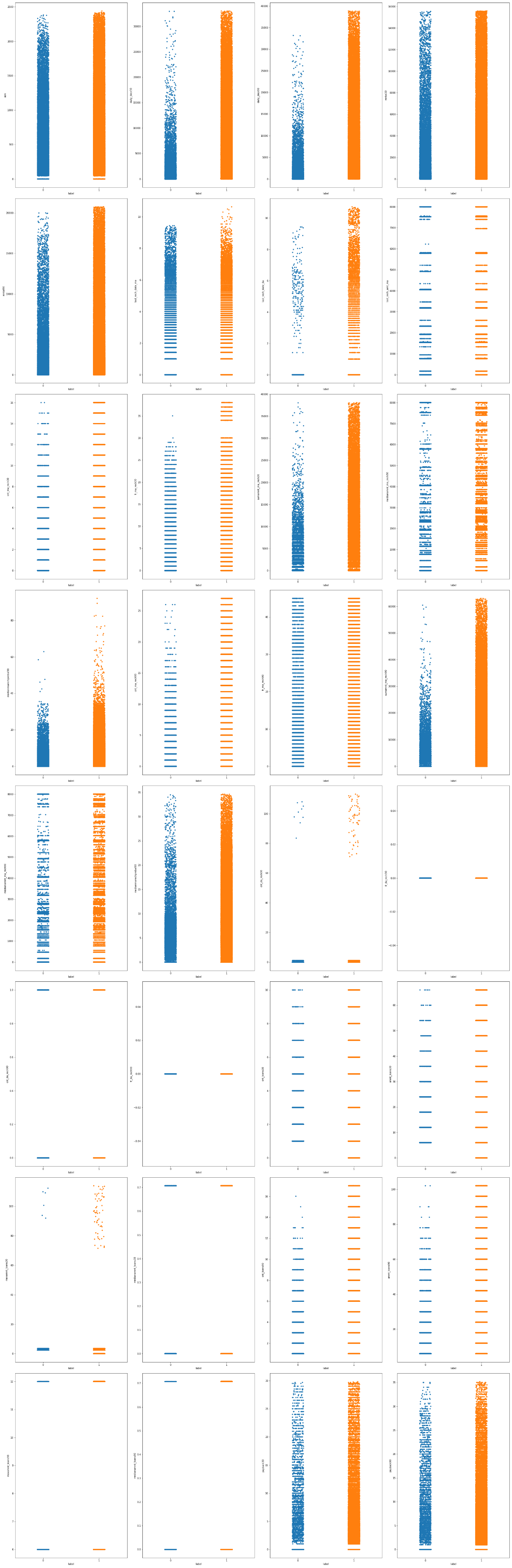
The boxplot of all the features showed that there are lot of outliers in the dataset. To eradicate such a huge amount of outliers the method of Z score was used. After outlier removal we were left with 161457 entries in the dataset. Even after getting rid of most of the outliers there were still few outliers left in the data. But, a decision was taken to move ahead with a fraction of outliers as complete removal of extreme data points may affect the predictability.

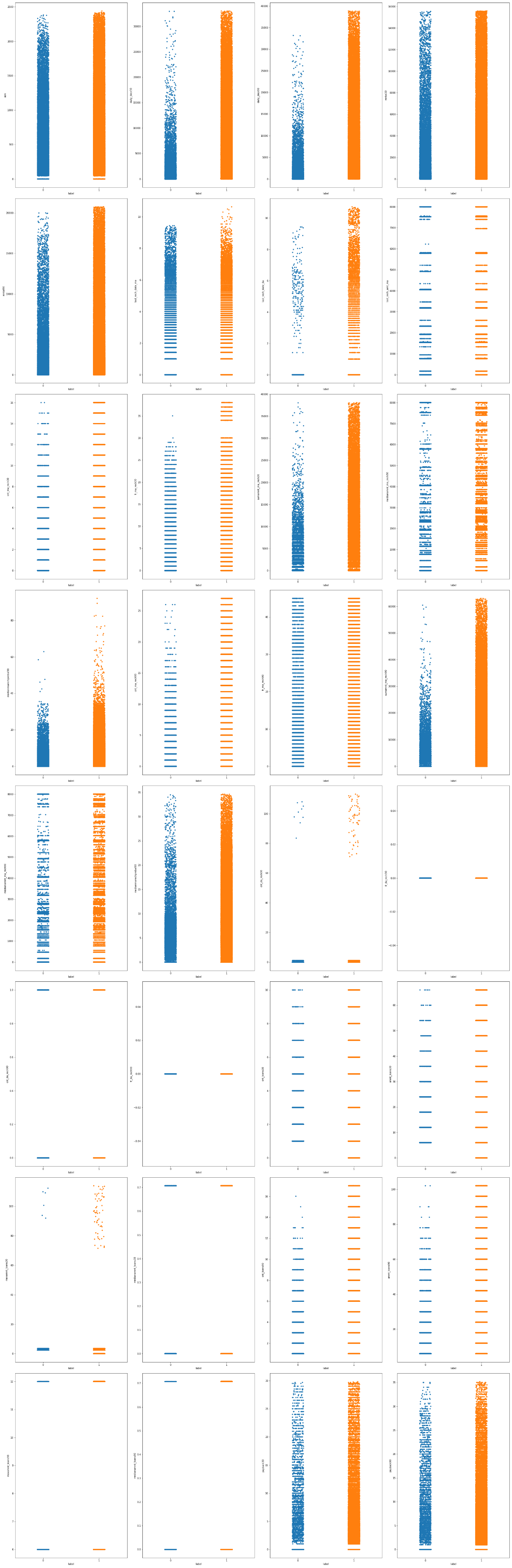
The outlier treatment handled the skewness to a great extent which was verified by visualization the distribution of cleaned dataset.

In order to further address the skewness the features which have skewness greater than 3 were identified and square root transformation was done. This transformation technique could reduce the skewness in some of the features. But the features ‘cnt\_da\_rech30‘ and ‘maxamnt\_loans30’ still have skewness greater than 48.

The next step is to separate the features and labels and visualize the influence of all the features on the outcome. The following conclusions can be noted from the strip plot of the features.

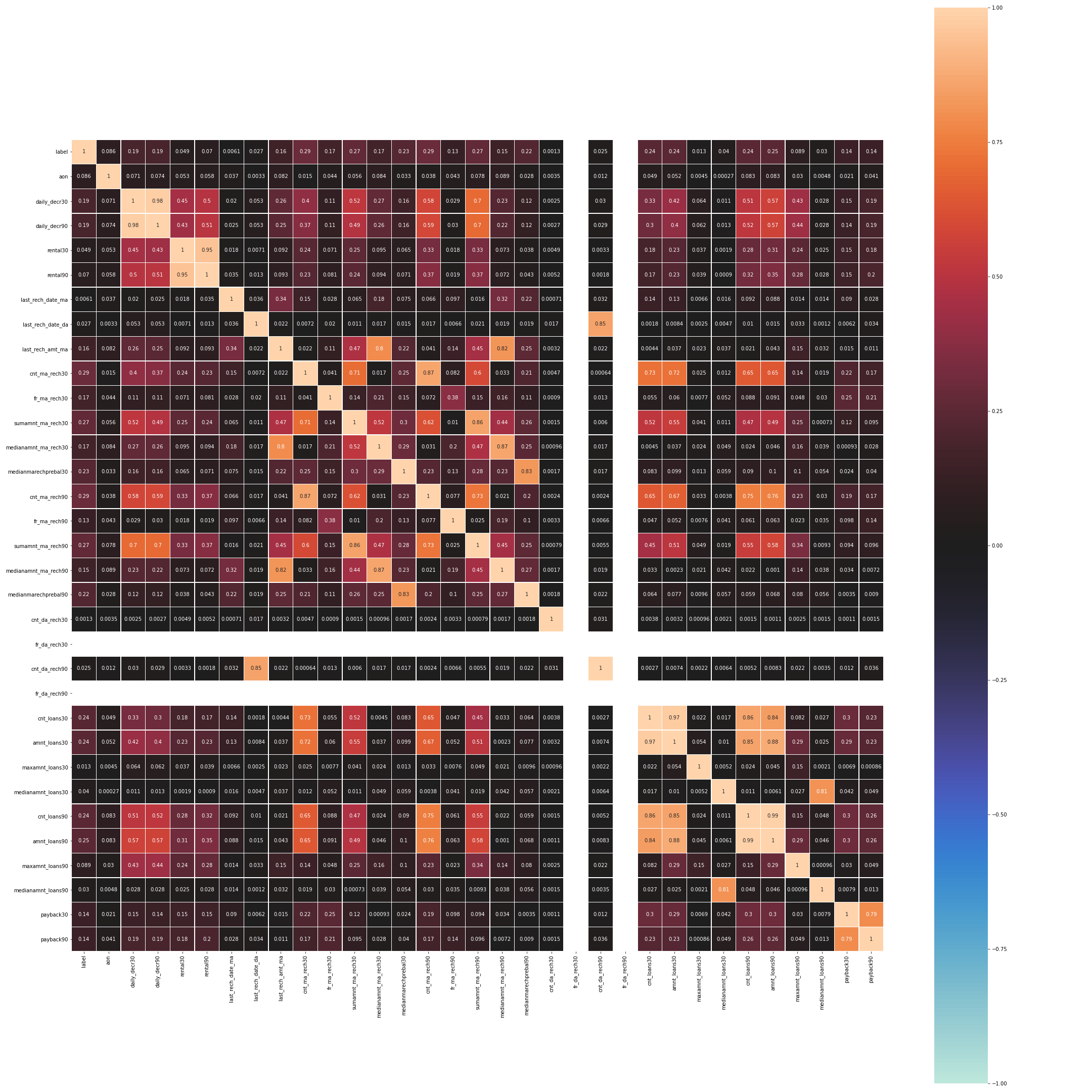






1. If the age on network is more than 1800 days, the customers are more likely to repay the loan and the probability of being a defaulter is less.
2. If the average of daily amount spent from main account over last 30 days is more than 5000 Indonesian Rupiah the probability of repayment increases.
3. If the average of daily amount spent from main account over last 30 days is more than 5000 Indonesian Rupiah the probability of repayment increases and likelihood of repayment is more than the previous case.
4. There are less number of defaulters if the customer’s average main account balance over last 30 days is more than 8000 Indonesian Rupiah.
5. There are less number of defaulters if the customer’s average main account balance over last 90 days is more than 10000 Indonesian Rupiah.
6. If the number of days till last recharge of main account is more than 9, the frequency of loan is less but, they repay the loan for sure without being a defaulter.
7. The customers who are recharging their data account in less than 4 days have very high probability of repaying the loan.
8. The customers who are recharging more than 12 times in the last 30 days are more likely to repay.
9. There is very high chance that the customer will repay the loan if the frequency of main account recharge in the last 30 days is more than 30.
10. If the total amount of recharge in main account over last 30 days is more than 13000 Indonesian Rupiah the chances of being a defaulter is less.
11. If the median of main account balance just before recharge in last 30 days at user level is more than 35 Indonesian Rupiah the repayment probability is significantly high.
12. The repayment probability is more if the customer recharges the main account more than 20 times in the last 90 days.
13. If the total amount of recharge in main account over last 90 days is more than 20000 Indonesian Rupiah the chances of being a defaulter is less.
14. If the median of main account balance just before recharge in last 90 days at user level is more than 20 Indonesian Rupiah the repayment probability is high.
15. The customers who have recharged data account in last 30 days more than 68 times are less likely to become defaulter.
16. If the maximum amount of loan taken by the user in last 30 days is more than 68 they have higher tendency to repay.
17. If the number of loans taken by user in last 90 days is more than 10 their chances of being a defaulter is less.
18. If the total amount of loans taken by user in last 90 days is more than 60 their chances of being a defaulter is less.
19. Higher the average payback time in days over last 30 days or over last 90 days higher are the chances of repayment.

There are 32 features for every entry hence, there are higher chances that the feature will be correlated to each other. Hence, feature selection can prove to be a deciding step to develop an accurate model. We can visualize the correlation matrix with the help of heat map to identify the dependent features.



The features which have high degree of correlation are - 'daily\_decr30', 'rental30', 'cnt\_ma\_rech30', 'last\_rech\_date\_da', 'sumamnt\_ma\_rech30', 'medianamnt\_ma\_rech30', 'medianamnt\_ma\_rech90', 'medianmarechprebal30', 'cnt\_loans30', 'cnt\_loans90', 'maxamnt\_loans30', 'medianamnt\_loans30', 'maxamnt\_loans90' and 'medianamnt\_loans90'. Hence, we can drop these features. After this step we have 19 features which play a significant role in identifying the defaulters.

When there are a large number of features in the dataset we may face multicollinearity problems. Variance Inflation Factor (VIF) can help us to identify and eliminate this issue. None of the feature have high vif.

We have checked the feature scores using Select K Best Feature technique and try to further reduce the number of feature for improved model performance. Based on the this feature selection technique the feature which we will use to predict the outcome are - cnt\_ma\_rech90, sumamnt\_ma\_rech90, amnt\_loans90, amnt\_loans30, medianmarechprebal90, daily\_decr90, fr\_ma\_rech30, last\_rech\_amt\_ma, payback30, payback90, fr\_ma\_rech90 and aon

After data pre-processing labels and features were separated, the features were scaled and split into train and test dataset in the ratio of 3:1 keeping in mind that the data was imbalanced. Hence, while splitting into train and test stratify was set to yes to remove the chances of overfitting.

The model which were tested in this data are-

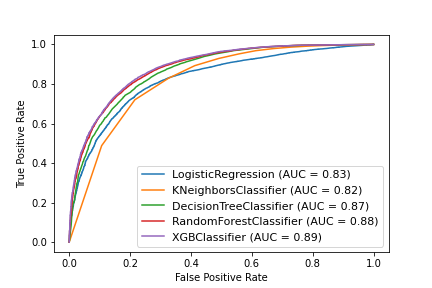
1. Logistic Regression
2. KNN Classifier
3. Decision Tree Classifier
4. Random Forest Classifier
5. XG Boost Classifier

The parameters of the model were tune were ever applicable to improve model accuracy and precision. The parameters precision, recall and F1 score were also compared.

The model performance is tabulated below for reference

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S. No.** | **Model Name** | **Accuracy** | **Precision** | **F1 Score** | **Accuracy after tuning** | **Precision after tuning** | **F1 Score after tuning** |
| 1. | Logistic Regression | 0.86 | 0.43   |  |  | | --- | --- | | 0 | 0 | | 1 | 0.86 | | 0.46 | - | - | - |
| 2. | KNN | 0.88 | 0.76   |  |  | | --- | --- | | 0 | 0.61 | | 1 | 0.91 | | 0.71 | 0.89 | 0.79   |  |  | | --- | --- | | 0 | 0.66 | | 1 | 0.91 | | 0.71 |
| 3. | Decision Tree | 0.85 | 0.69   |  |  | | --- | --- | | 0 | 0.45 | | 1 | 0.92 | | 0.69 | 0.90 | 0.73   |  |  | | --- | --- | | 0 | 0.76 | | 1 | 0.91 | | 0.77 |
| 4. | Random Forest | - | - | - | 0.90 | 0.85   |  |  | | --- | --- | | 0 | 0.79 | | 1 | 0.91 | | 0.73 |
| 5. | XG Boost | - | - | - | 0.83 | 0.85   |  |  | | --- | --- | | 0 |  | | 1 |  | | 0.77 |

**ROC AUC CURVE for Model Evaluation**

****

The area under curve for XG Boost Classifier is maximum hence it is selected as the final model for prediction of micro credit defaulter.

**Conclusion**

We cleaned the data, performed EDA and successfully trained a model and tuned the hyper parameters to predict the response of a loan application based on the features and dataset available with us. The XG Boost Classifier proved to be the best model in prediction due to its high precision and accuracy. I hope this has proven to be an informative topic.

* **Accuracy – 0.90**
* **Precision (Defaulter) – 0.77**
* **Overall Precision – 0.84**
* **Hyper Parameters – {max\_depth=10, learning\_rate=0.01, n\_estimators=1000, colsample\_bytree=0.7, scoring='neg\_mean\_squared\_error', verbose=1}**
* **AUC – 0.89**