Micro-Credit Defaulter Model

Submitted by:

Harneet Kaur Rehsi

ACKNOWLEDGMENT

I want to thank my SME Khusboo Garg and Flip Robo Techniques to express our sincere thanks to the following people, without whom we would not have been able to complete this project

INTRODUCTION

Business Problem Framing : A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on. 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.

Review of Literature : In this project we need to build a machine learning 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.

Motivation for the Problem Undertaken : The objective behind making this project 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 the hour.

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem : In this project, the dataset contains several rows and columns containing all the necessary information. For unwanted columns based on my analysis, I have drop them by using drop technique, also remove duplicate data present in the dataset and for outliers present In the dataset we have used several statistical methods and exploratory data visualization for better understanding and model building for predictions.

Data Sources and their formats : The datset contains 209593 rows and ,37 columns .

```
Unnamed:0
label
msisdn
aon
daily_decr30
daily_decr90
rental30
rental90
last_rech_date_ma
last_rech_date_da
last_rech_amt_ma
cnt_ma_rech30
fr_ma_rech30
sumamnt_ma_rech30
medianamnt_ma_rech30
medianmarechprebal30
cnt_ma_rech90
fr_ma_rech90
sumamnt_ma_rech90
medianamnt_ma_rech90
medianmarechprebal90
cnt_da_rech30
fr_da_rech30
cnt_da_rech90
fr_da_rech90
cnt_loans30
amnt_loans30
maxamnt_loans30
medianamnt_loans30
cnt_loans90
amnt_loans90
maxamnt_loans90
medianamnt_loans90
payback30
payback90
pcircle
pdate
```

Data Preprocessing Done : Dataset contains some duplicate data so I have removed duplicate data and drop all the unwanted columns based on my analysis of the dataset. The dataset contains outliers in some of the columns so I have used describe method for understanding the data and distplot and boxplot for checking the skewness and outliers in the dataset then I have used Z_score method for removing the outliers from the dataset. As the

numbers of columns are huge. For feature selection I have selected the best 10 columns on the basis of their score. Then I have used the 10 best features for further in the model,

Hardware and Software Requirements and Tools Used :

These are the algorithms, techniques or model etc used in this project

- import pandas as pd
- import numpy as np
- import seaborn as sns
- import matplotlib.pyplot as plt
- from scipy.stats import zscore
- %matplotlib inline
- import warnings
- warnings.filterwarnings('ignore')
- from sklearn.feature_selection import SelectKBest , f_classif
- from sklearn.preprocessing import StandardScaler
- from sklearn.ensemble import RandomForestClassifier
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.linear_model import LogisticRegression
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.metrics import roc_curve, roc_auc_score
- from sklearn.metrics import plot_roc_curve
- from sklearn.model_selection import train_test_split , GridSearchCV
- from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods) :

The various analytical and statistical techniques used in the project are: describe for checking the health of the dataset and isnull().sum() for checking the null values if any are present in the dataset or not and shape for checking the numbers of rows and columns dataset contains. For outliers here I have used Z_score for

removing the outliers from the columns in the dataset. And feature selecting technique SelectKBest , f_classif as the number of columns as huge so I have picked top 10 best features from all the features

Testing of Identified Approaches (Algorithms)

- from sklearn.preprocessing import StandardScaler
- from sklearn.ensemble import RandomForestClassifier
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.linear_model import LogisticRegression
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.metrics import roc_curve, roc_auc_score
- from sklearn.metrics import plot_roc_curve
- from sklearn.model_selection import train_test_split , GridSearchCV
- from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

Run and Evaluate selected models :

all the algorithms used : we have scaled and split the train test data with train_test_split and standard scaler.

- from sklearn.ensemble import RandomForestClassifier

```
scalar = StandardScaler()
X_scalar = scalar.fit_transform(new_X)

X_train,X_test,y_train,y_test =train_test_split(X_scalar,y, test_size=0.2,
random_state=0)

rfc = RandomForestClassifier()
rfc.fit(X_train,y_train)

y_pred =rfc.predict(X_test)

print("confusion matrix","\n",confusion_matrix(y_test,y_pred))

print("\n","accuracy rf normal:",accuracy_score(y_test,y_pred))
```

print("\n","report :",classification_report(y_test,y_pred))

```
confusion matrix
 [[ 2117  2741]
 [  962 30679]]

 accuracy rf normal: 0.898545165620976

 report :               precision    recall  f1-score   support

           0       0.69      0.44      0.53      4858
           1       0.92      0.97      0.94     31641

    accuracy                           0.90     36499
   macro avg       0.80      0.70      0.74     36499
weighted avg       0.89      0.90      0.89     36499
```

Random Forest Classifier give accuracy of 90%

After hyper parameter tuning the accuracy is

```
confusion matrix
 [[ 1937  2921]
 [  504 31137]]

 accuracy rf normal: 0.9061618126524015

 report :               precision    recall  f1-score   support

           0       0.79      0.40      0.53      4858
           1       0.91      0.98      0.95     31641

    accuracy                           0.91     36499
   macro avg       0.85      0.69      0.74     36499
weighted avg       0.90      0.91      0.89     36499
```

Random Forest Classifier give accuracy of 91%

- ☆ from sklearn.neighbors import KNeighborsClassifier
- ☆ from sklearn.linear_model import LogisticRegression
- ☆ from sklearn.tree import DecisionTreeClassifier

lr =LogisticRegression()

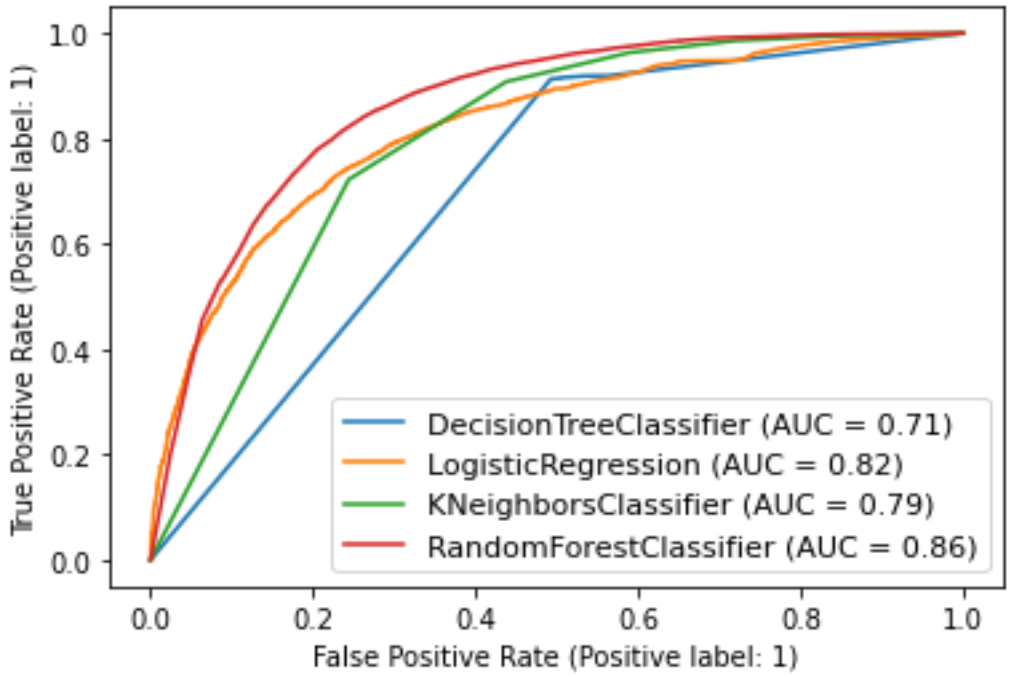
```
dt =DecisionTreeClassifier()
rfc =RandomForestClassifier()
kn =KNeighborsClassifier()
```

```
# training with all classifiers

lr.fit(X_train,y_train)
dt.fit(X_train,y_train)
rfc.fit(X_train,y_train)
kn.fit(X_train,y_train

# all models score captured

lr.score(X_test,y_test)
dt.score(X_test,y_test)
rfc.score(X_test,y_test)
kn.score(X_test,y_test)
```



Random forest classifier fit the best model from all the others

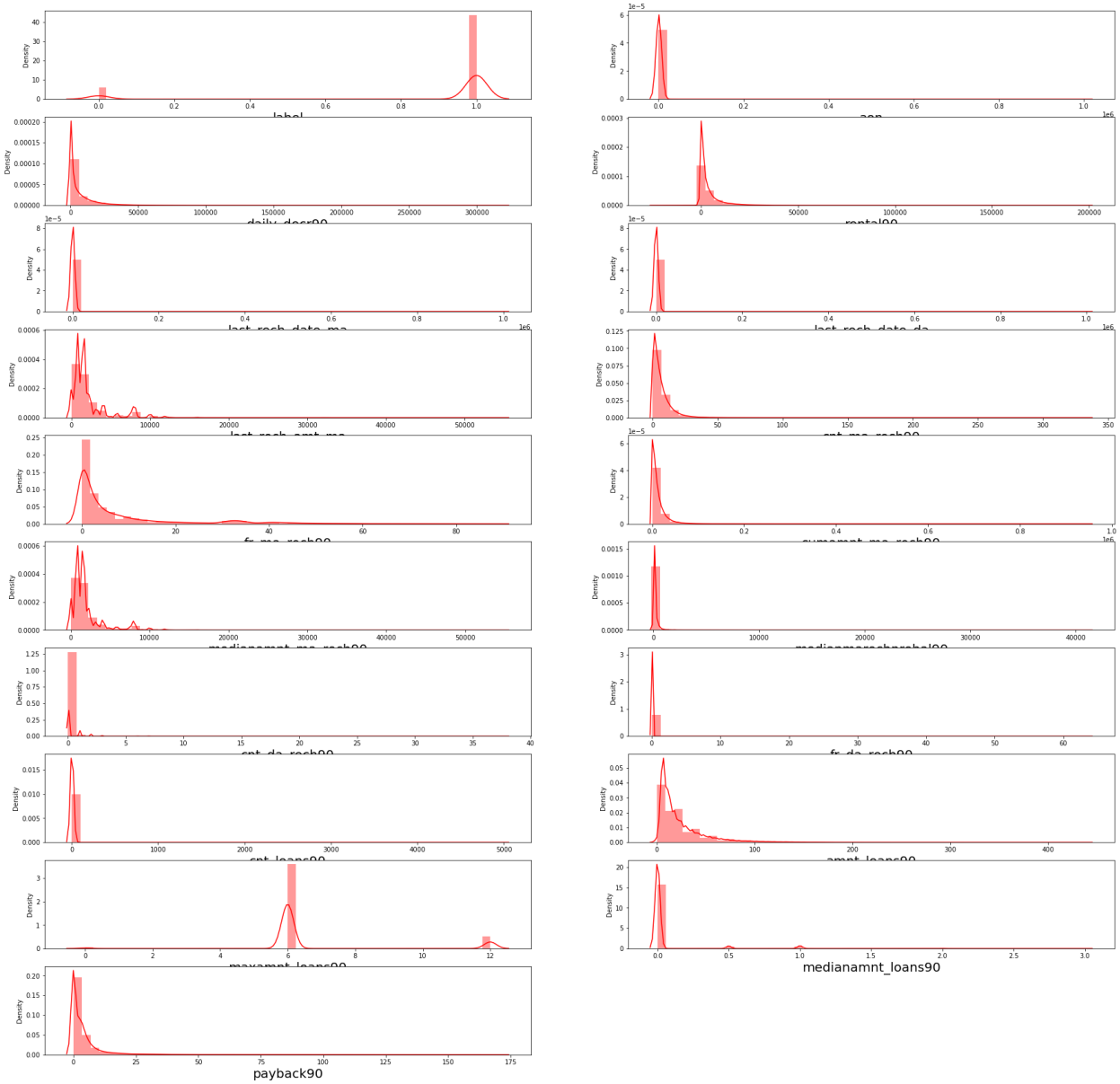
Logistic Regression

KNeighbours classifier

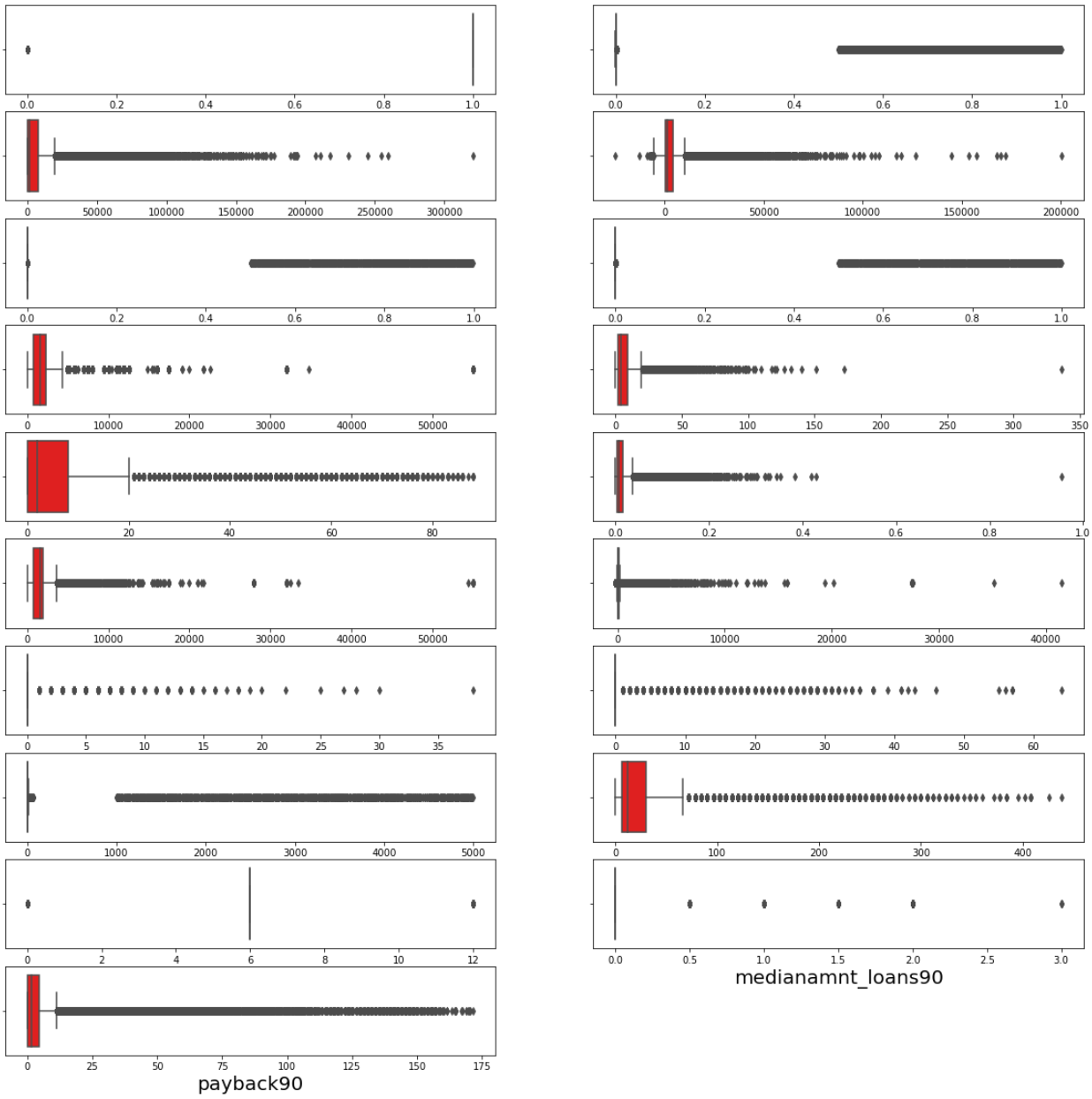
Decision Tree

Key Metrics for success in solving problem under consideration: Z_score and Selecting features For outliers here I have used Z_score for removing the outliers from the columns in the dataset. And feature selecting technique SelectKBest , f_classif as the number of columns as huge so I have picked top 10 best features from all the features

Visualizations

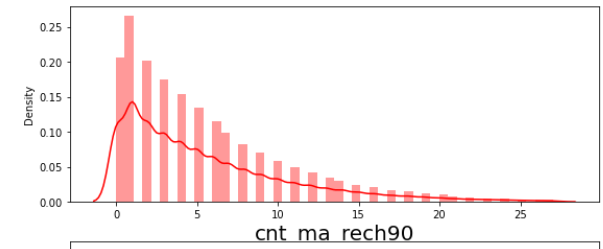


Here I have used distplot for checking the skewness in the data.

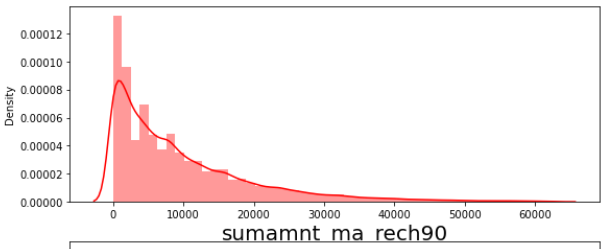


Here I have used boxplot for checking the outliers present in the data

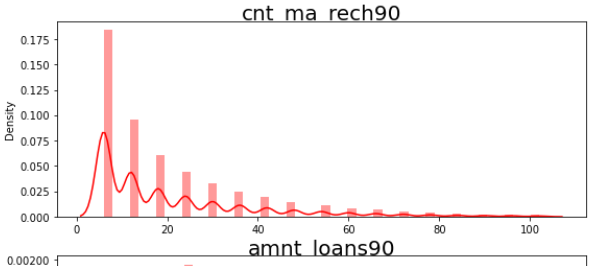
These graphs are after applying the techniques to remove shewness and outliers from the data of top 10 selected features



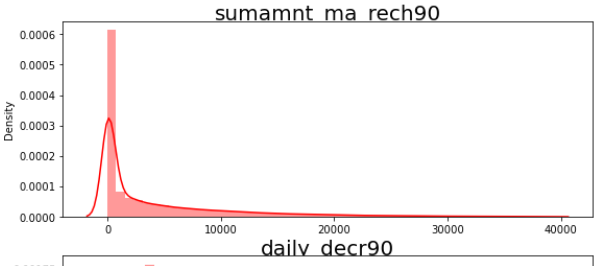
cnt_ma_rech90



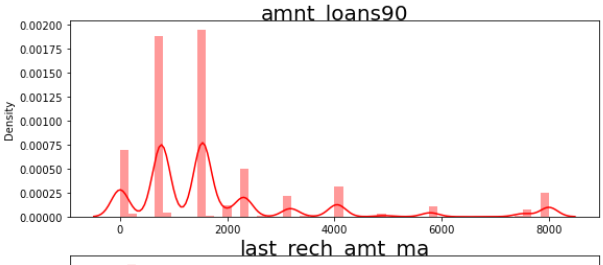
sumamnt_ma_rech90



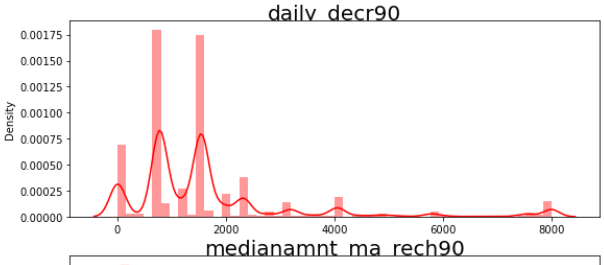
amnt_loans90



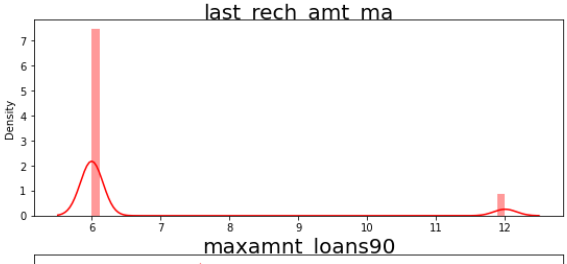
daily_decr90



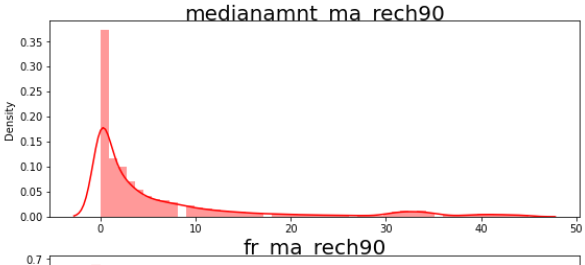
last_rech_amt_ma



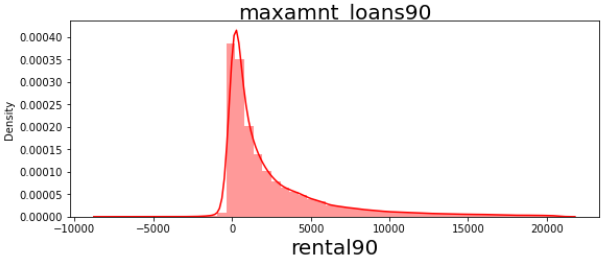
medianamnt_ma_rech90



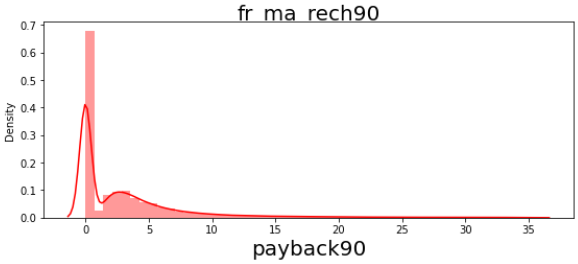
maxamnt_loans90



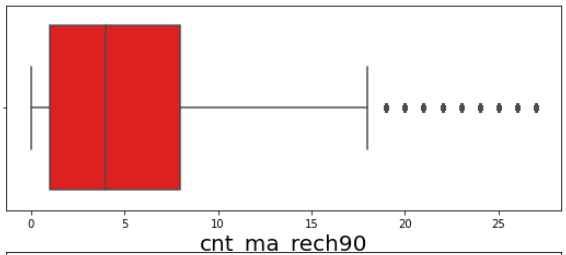
fr_ma_rech90



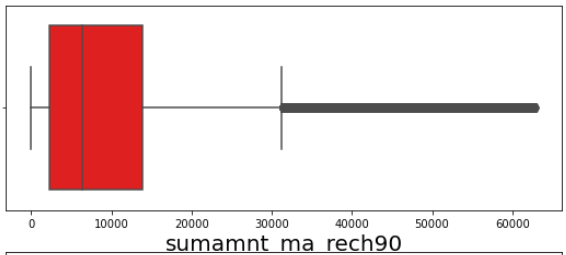
rental90



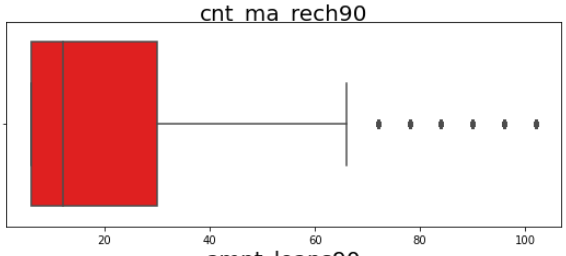
payback90



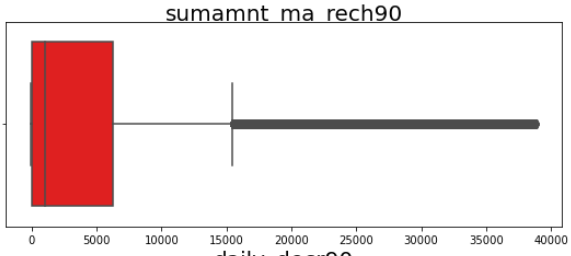
cnt_ma_rech90



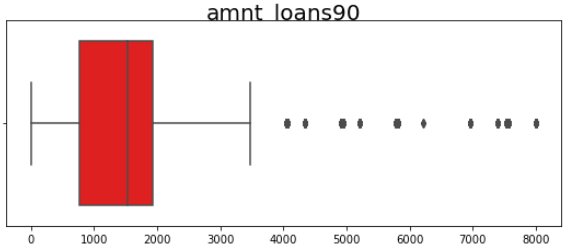
sumamnt_ma_rech90



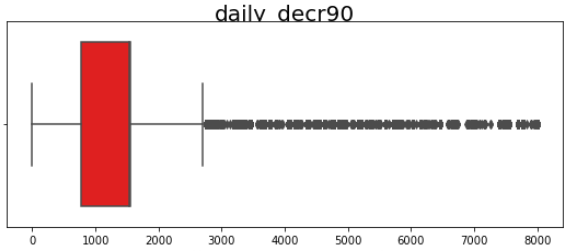
amnt_loans90



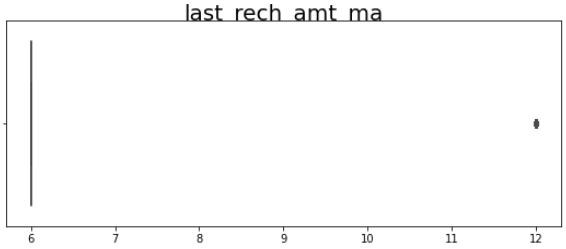
daily_decr90



last_rech_amt_ma

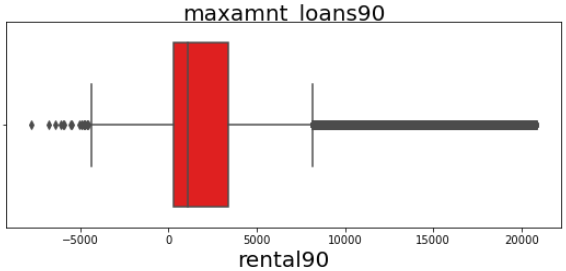


medianamnt_ma_rech90

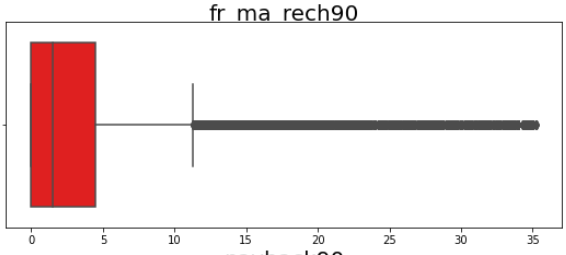


maxamnt_loans90

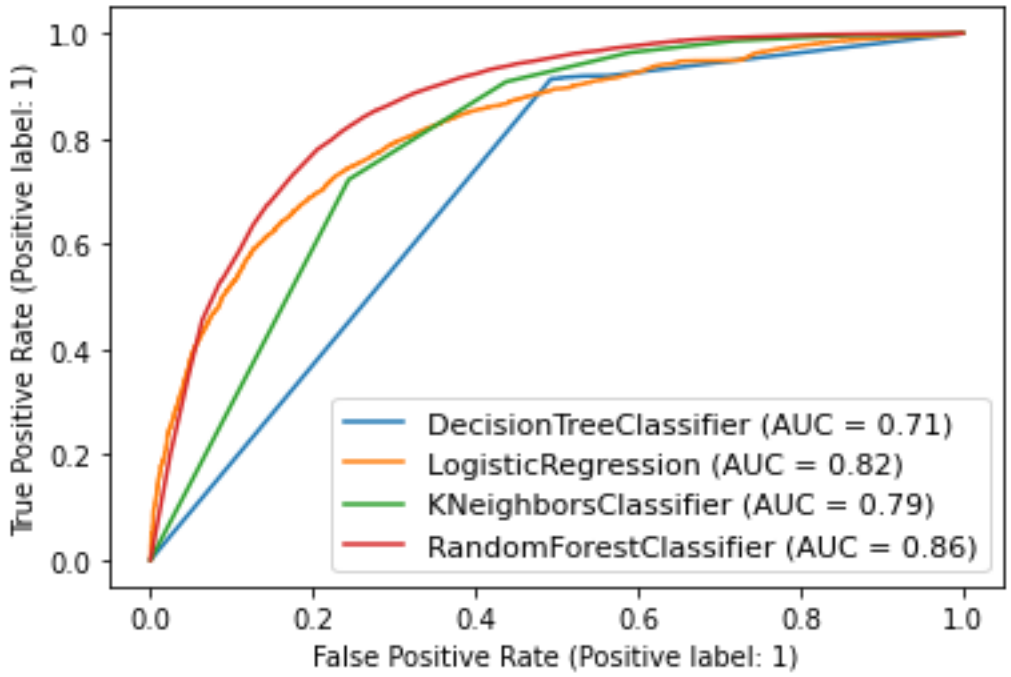
fr_ma_rech90



rental90



payback90



This curve contain the score of all the model I have used and these are the scores

Random Forest Classifier model is fit best

Interpretation of the Results : After visualizing the graph the dataset contain skewness and outliers in the datset with the help of distplot and boxplot we can visualize them easily. And Random Forest Classifier model fits best for the model.

CONCLUSION

Key Findings and Conclusions of the Study: After visualizing the data and graph the dataset contain skewness and outliers in the datset with the help of distplot and boxplot we can visualize them easily. And Random Forest Classifier model fits best for the model. The model that I have built “91” accuracy rate. I have tried different machine learning models the accuracy rate is the different in all the models. Random Forest Classifier gives the highest accuracy rate as compare to all other models.

Learning Outcomes of the Study in respect of Data Science:

Data contains some duplicate values. To clean the data I used the drop. duplicate method and on basis of my analysis I have dropped several columns present in the particular dataset. . For removing outliers I have used z_score technique. For selection of top 10 best columns for the data used SelectKBest , f_classif as the dataset is huge so I have applied for the betterment of the result.

Limitations of this work and Scope for Future Work:

Random Forest Classifier model fits best for future prediction with an accuracy rate of 91% and AUC Score of 86%. so I analyzed that use a random forest classifier for better predictions