

Micro Credit Defaulter Project

Submitted by:

VINAY SAHI

**References**

* **Data trained institute video tutorial**
* [**https://www.geeksforgeeks.org/**](https://www.geeksforgeeks.org/)
* [**https://towardsdatascience.com/machine-learning/home**](https://towardsdatascience.com/machine-learning/home)

**INTRODUCTION**



Microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients. A telecommunications network provider is 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 .The sample data is provided I am going to 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

As the client has been into telecom industry and now venturing into financial market in association with MFI. It will be a great combination as MFI has been into financial market with all that experience put together telecom and finance will do potentially well. The data will help me understanding the pattern of customers

Features like how long the customers has been with the company ,how often does he recharge the pack ,how many days he takes to return the credit.

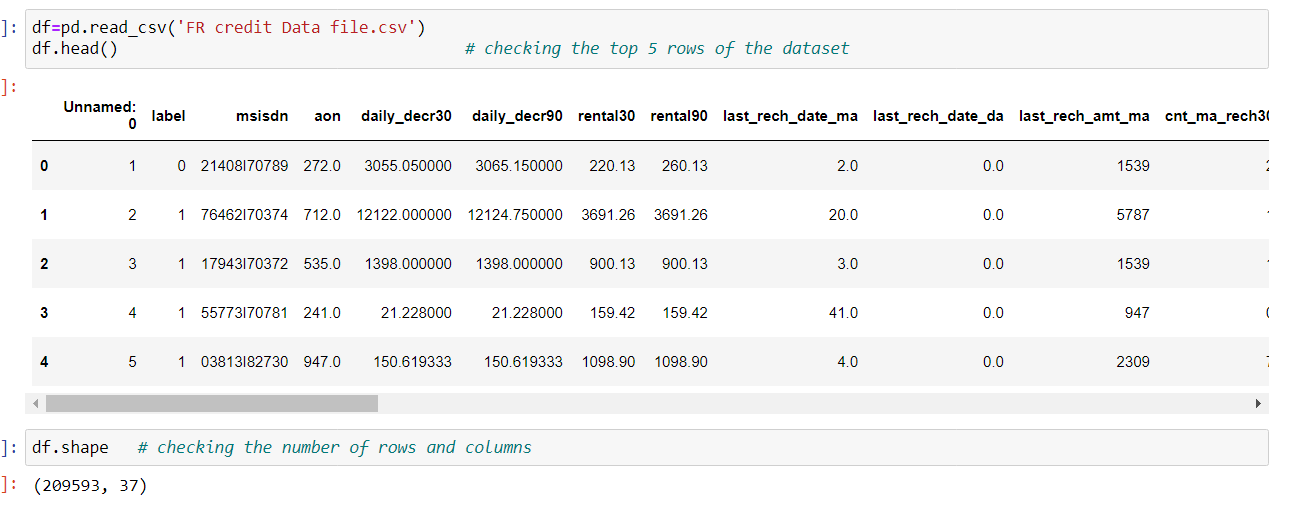
Motivation for the Problem Undertaken

I am a student of data science and machine learning ,I am gaining plenty of hands on experience by working on these projects which will help me to understand the real world scenarios.

I am excited and highly motivated to gain enough hands on experience that will land me a job in reputed organisation.

Data Sources and their formats

The data is been provided by fliprobo ,the dataset is in csv format, it has 209593 rows and 37 columns



The feature includes a target variable which is a binary column consist of 0 or 1 . Independent variable includes last recharge date in days ,last recharge amount of last 30 /90 days ,total loan amount and so on.The data does not have any null values.

Data Pre-processing

Machine learning algorithms don’t work so well with processing raw data. Before we can feed such data to an ML algorithm, we must preprocess it. We must apply some transformations on it. With data preprocessing, we convert raw data into a clean data set

SMOTE- This data has unbalanced data in target variable ,As this is a classification problem ,smote has been used to balance the data so machine can learn best way possible.

DROP- drop method has been used to drop useless column.

DATE TIME- to\_datetime has been used to convert the object type data into datetime.

FEATURE SELECTION- feature selection technique has been used to find out best features with most information gain.

Hardware and Software Requirements

Hardware used-

Dell laptop with 500 GB SSD, intel icore processer i5 11th generation with 8 GM RAM.

Software requirements along with the tools, libraries and packages used-

Python - Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Pandas- pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Numpy- is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots

Seaborn is one of the world’s most regarded Python libraries that is **purpose-built to create beautiful looking visualizations**. It can be considered as an extension of another library called Matplotlib as it is built on top of that. Data visualization is easily performed in Seaborn

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them.

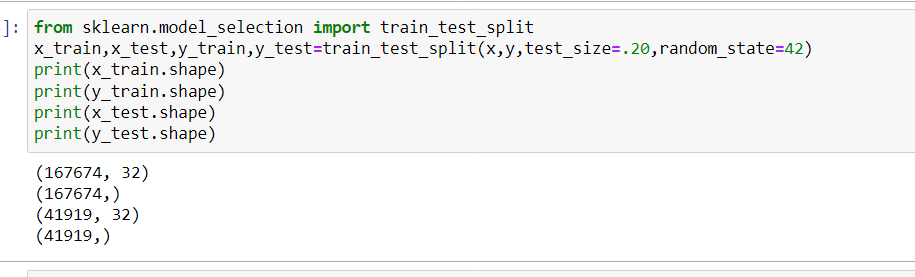
Extra tree classifier - In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**

Identification of possible problem-solving approaches

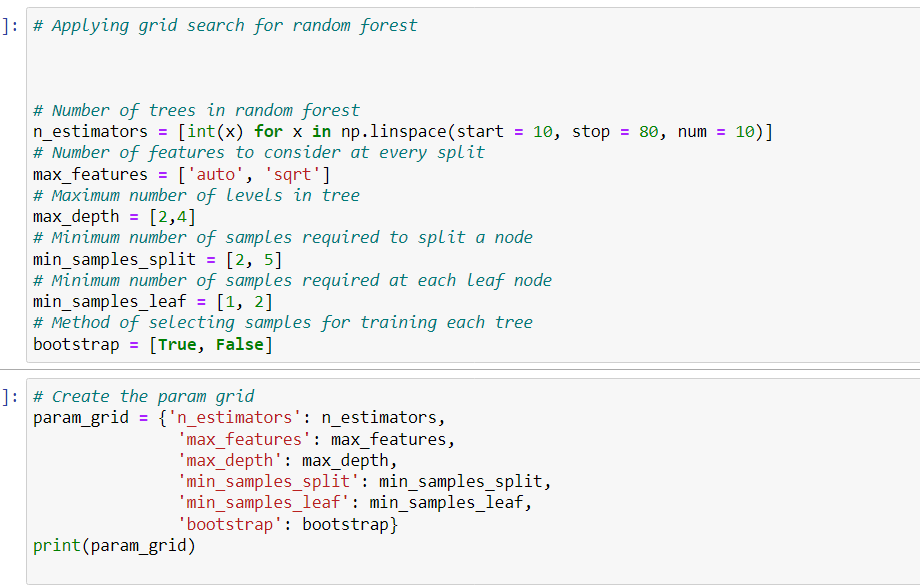
EDA has been performed to understand the data well, Visualisation has been done to understand data behaviour, correlation with heatmap has given understanding of corelation between variables and duplicity.feature selection has been performed to get the idea which variable is providing the best information gain, target variable has been balanced with smote technique,classification algorithms has been applied to build a predictive model to predict whether the customer will be paying back the loaned amount within 5 days .

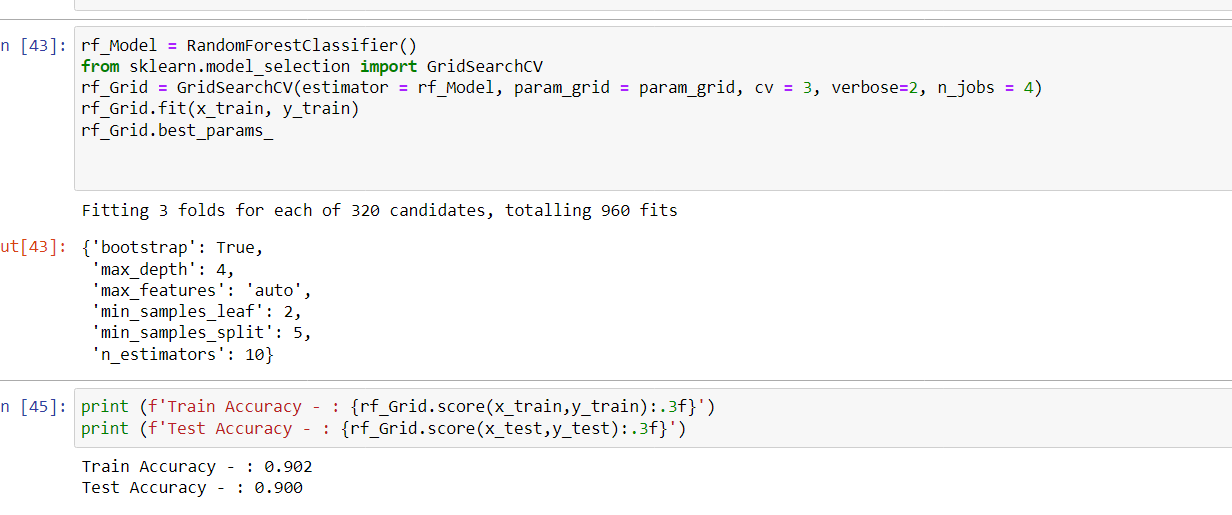
Train test split



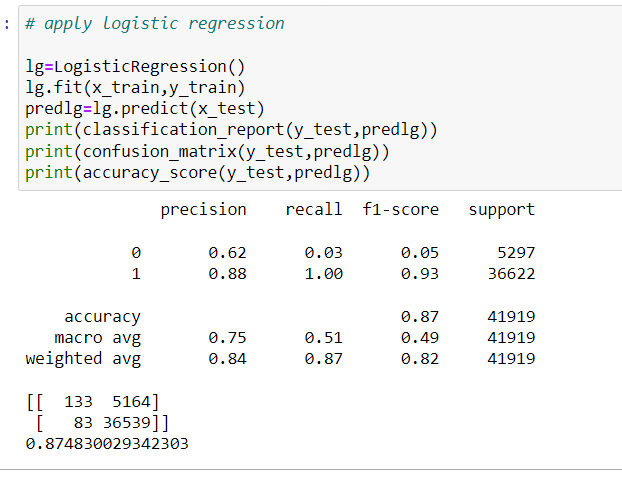
GRID SEARCH- hyperparameter tuning has been performed to get the best parameters

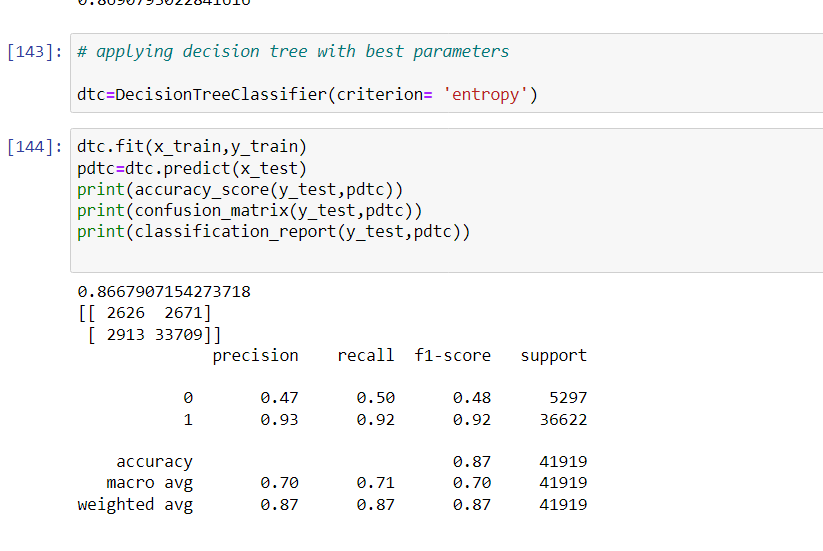


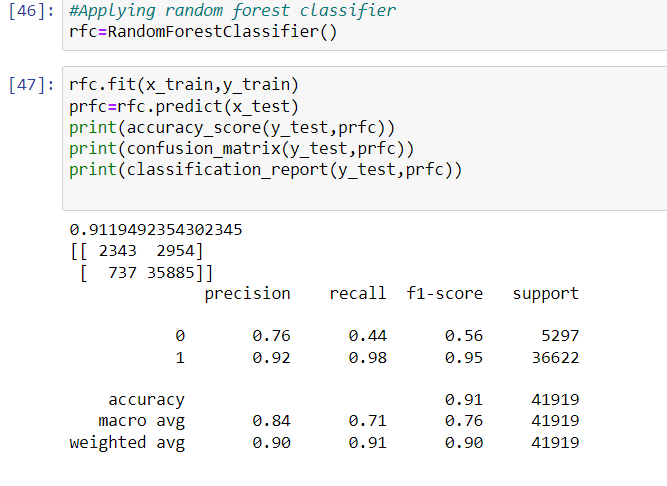




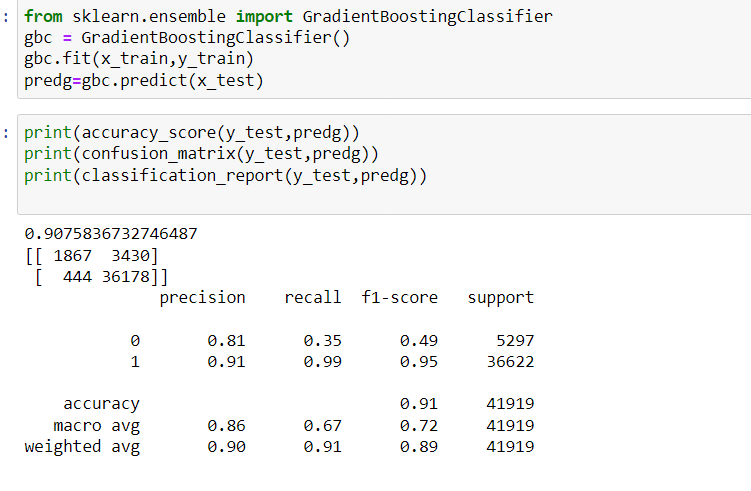
ALGORITHMNS



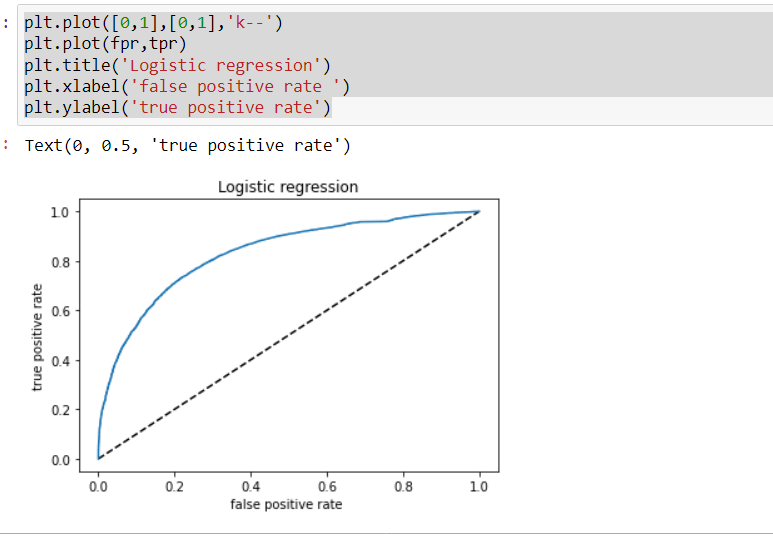




Applying gradient boosting classifier -

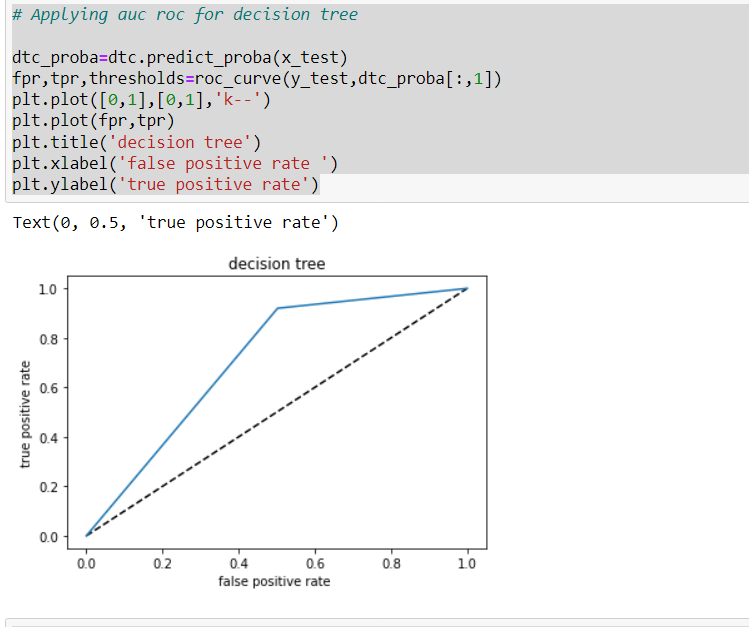
 AUC ROC CURVE –

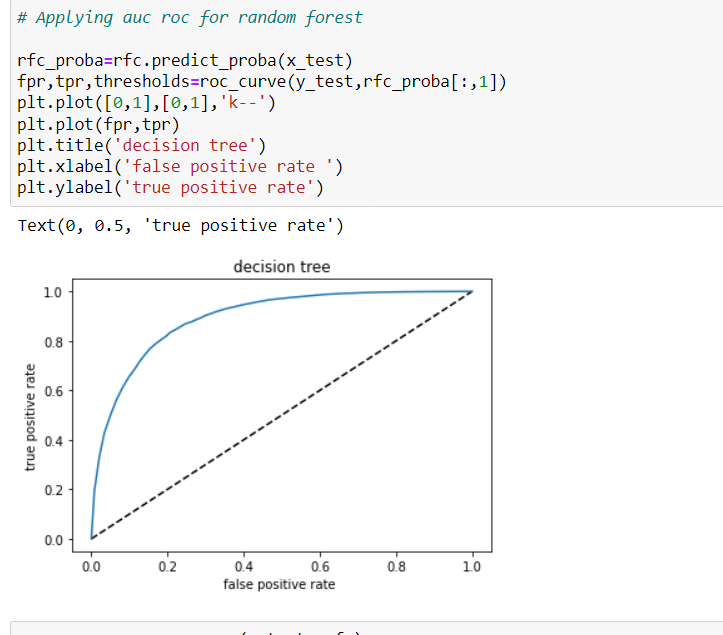
AUC roc for logistic regression



The roc auc is one of the important KPI’S ,the area under the graph shows how well the sensitivity or true positive rate being predicted.The more the area under the graph is better the model will predict.

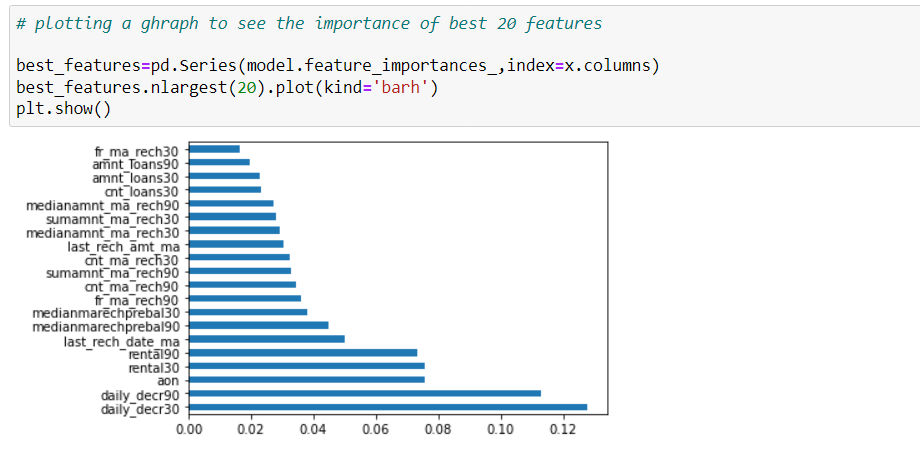
AUC ROC FOR DECISION TREE

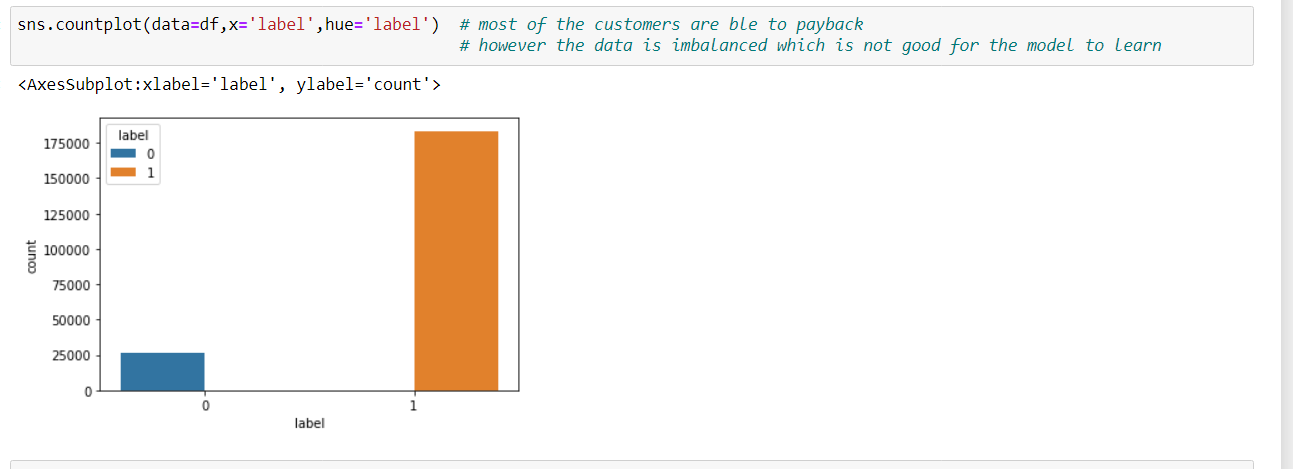




Visualizations

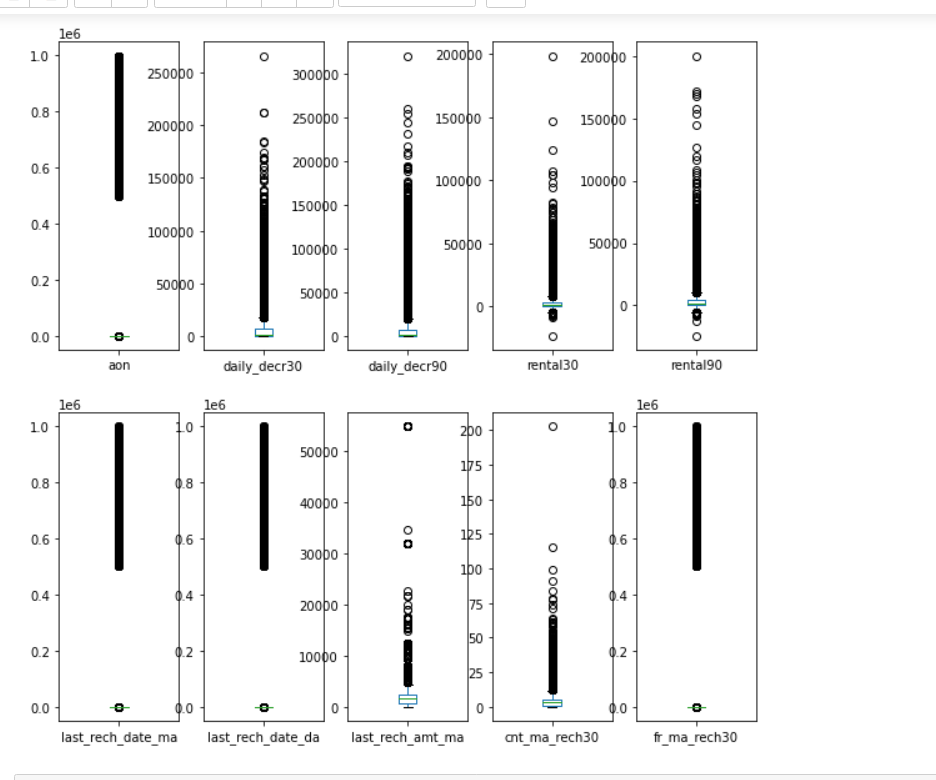
Showing the best features on graph

Countplot showing number of customers as defaulter or not

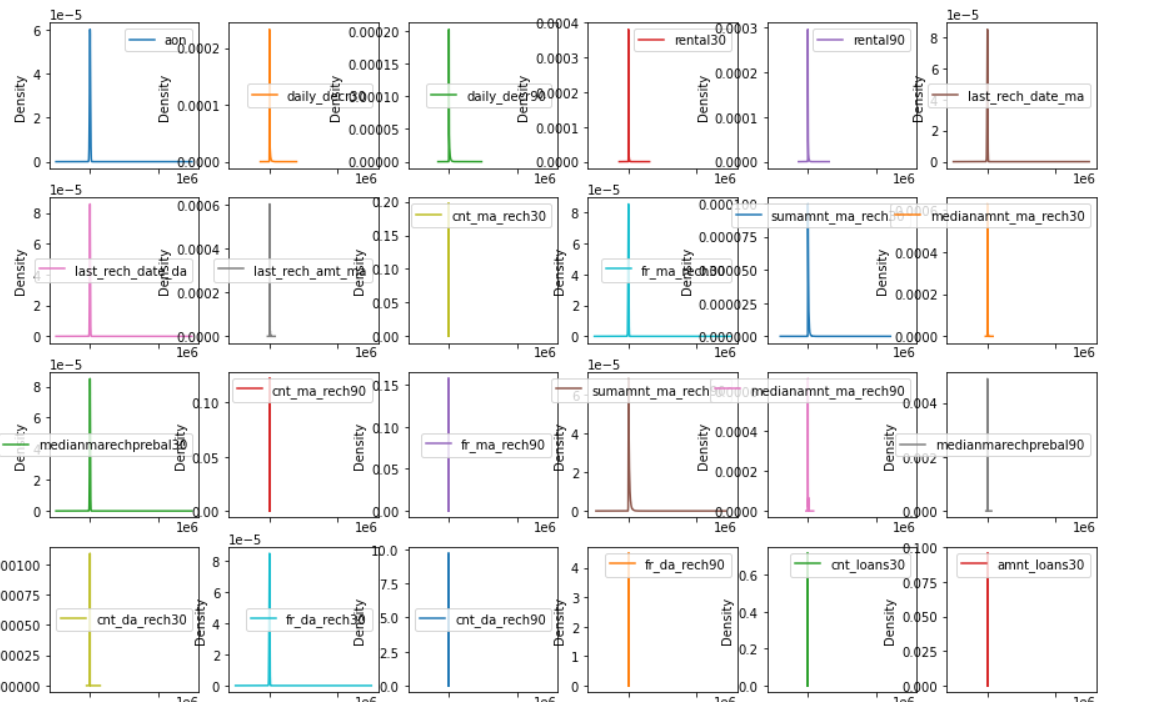


Box plot of different columns to check the outliers –

Box plot is a good way of visualising the outliers in the data ,there are outliers present in this data ,however in attempt of removing the outliers we would loose more than 25% of data so have to let it stay as it is .we can not afford to loose so much of data .



Kernal density plots – To check the data distribution we can use kde plots ,I have applied kde to all the columns to check the spread of data



Interpretation of the Results

The graphs shows that data is normally distributed ,data contains outliers. The majority of consumers are able to pay back the credit on time .

With the implementation of feature selection, we were able to get the top feature that are giving high information gain to the model to predict well.

We can see that random forest classifier is performing the best with 91% accuracy.

The auc roc graphs shows the models are performing average as the area under the graph is not covering the entire area.

Learning Outcomes of the Study in respect of Data Science

After completing this project, I have learned more about the different ways of visualisation techniques, I have used various algorithms and random forest has performed the best.

I had faced a lot of challenges as i was not well versed with hypermeter tuning as I am still in learning phase however while working on this project I have learned and implemented various techniques.