

**Micro-Credit Defaulter Model**

Submitted by:

Ankita Chaudhari

**ACKNOWLEDGMENT**

1. Abd-Allah, M. N., Salah, A., and El-Beltagy, S. R. (2014). Enhanced Customer Churn Prediction using Social Network Analysis. Proceedings of the 3rd Workshop on DataDriven User Behavioral Modeling and Mining from Social Media, pages 11–12.
2. Backiel, A., Baesens, B., and Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. Journal of the Operational Research Society, 67(9):1135–1145.

**INTRODUCTION**

* Business Problem Framing

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.

This is a classic Business problem which helps Micro Financing Institutions and other Lending companies reduce Credit risks by recognizing potential Defaulters.

* Control Background of the Domain Problem

No matter what the problem is, but data science has the same framework to follow for every problem. These are the steps to follow for solving the problem from the higher viewpoint:

1. EDA(Exploratory Data Analysis) and Data Cleaning
2. Feature Engineering
3. Feature Selection
4. Modeling
5. Model Evaluation
6. Conclusion of the problem

* **Review of Literature**

This is example of classifications of the incorporated machine learning models against traditional methods, such as logistic regression, Random forest classification, Decision Tree classification, KNN classification, Gradient Boost classification. Their results identified that machine learning models had demonstrated superior performance and forecasting accuracy through the financial credit rating cycle.

* **Motivation for the Problem Undertaken**

Motivation is a dynamic process of internal psychological factors encompassing the needs, wants, and goals of an individual and very powerful motive in entrepreneurship that direct to human behavior for reaching to aim and tendencies. There are several ways in which motivation can be described, but motivation is broadly categorized as an intrinsic or extrinsic. Intrinsic motivation is the one in which task performance is for the sake of task performance, while extrinsic motivation involves an element of external reward. In other words, in the extrinsic motivation goals are of interim type at the service of a much more important achievement Most compelling, emphasize reputation-based measures, known as dynamic incentives, which appeal to the intrinsic borrowers. In this Project Data Normalization Since the data was not normal, I normalized all the features except the target variable which was dichotomous(Values '1' and '0').

**Analytical Problem Framing**

**1.What is Statistical Modeling and How is it Used?**

Statistical modelingis the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.

When [data analysts](https://www.northeastern.edu/graduate/blog/what-does-a-data-analyst-do/) apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically. Rather than sifting through the raw data, this practice allows them to identify relationships between variables, [make predictions](https://www.northeastern.edu/graduate/blog/predictive-analytics/) about future sets of data, and visualize that data so that non-analysts and stakeholders can consume and leverage it.

“When you analyze data, you are looking for patterns,” says Mello. “You are using a sample to make an inference about the whole.”

## Important Statistical Techniques in Data Analysis

Before any statistical model can be created, an analyst needs to collect or fetch the data micro credit on a database, clouds, social media, or within a plain excel file. To do this, analysts must also have a solid grasp of data structure and management, including how and where data is stored, fetched, and maintained. Those working in this field should thus share a passion for facts and data, and understand the basics of data manipulation, as well.

Once it comes time to analyze the data, there are an array of statistical models analysts may choose to utilize. According to Mello, most common techniques will fall into the following two groups:

* Supervised learning, including regression and classification models.
* Unsupervised learning, including clustering algorithms and association rules.

### Regression Models

Data analysts use **regression models** to examine relationships between variables. Regression models are often used by organizations to determine which independent variables hold the most influence over dependent variables—information that can be leveraged to make essential [business decisions](https://www.northeastern.edu/graduate/blog/data-driven-decision-making/).

“The most traditional regression models that have been used for a long time are logistic regression, linear regression, and polynomial regression,” Mello says. “These are the most common.”

Other examples of regression models can include stepwise regression, ridge regression, lasso regression, and elastic net regression.

### Classification Models

**Classification** is a process in which an algorithm is used to analyze an existing data set of known points. The understanding achieved through that analysis is then leveraged as a means of appropriately classifying the data. Classification is a form of machine learning that can be particularly helpful in analyzing very large, complex sets of data to help make more accurate predictions.

“Classification models are a form of supervised machine learning which is often used when the analyst needs to understand how they got to a certain point,” Mello says. “They give you more than just an output; [they give you] more information that you can use to explain the results of the prediction to your boss or stakeholder.”

Some of the most common classification models include decision trees, random forests, nearest neighbor, and  Naive Bayes.

There are also the neural networking models that are more used in AI. “These are very powerful models, and they can make accurate predictions very well,” Mello says, “but you typically cannot explain what is happening behind the scenes.”

**Types of the Model:**

Here it is the classification modeling technique use because in this dataset **label** is target variable and in this column there are 2 values are present like 0 and 1 so there was classification problem so use classification models.

**Problem Definition:**

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

**Data Analysis:**

The first and foremost step involves importing necessary libraries and packages and loading the dataset as a pandas dataframe.

**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.

**Data Loading and Visualisations:**

The first and foremost step involves importing necessary libraries and packages and loading the dataset as a pandas dataframe. Data visualization is the graphical representation of information and data. By using [visual elements like charts, graphs, and maps](https://www.tableau.com/learn/articles/data-visualization/glossary), data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Our eyes are [drawn to colors and patterns](https://www.tableau.com/learn/whitepapers/tableau-visual-guidebook). We can quickly identify red from blue, square from circle. Our culture is visual, including everything from art and advertisements to TV and movies. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we [quickly see trends and outliers](https://www.tableau.com/reports/business-intelligence-trends). If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be.

**Importing libraries**

We will start by importing the libraries we will require for performing EDA. These include NumPy, Pandas, Matplotlib, and Seaborn.

### Reading data:

We will now read the data from a CSV file into a Pandas DataFrame in this there are micro credit dataset.

Install This project requires anaconda python, because below libraries already available.

• Numpy

• Matplotlib

• Seaborn

• Skit-learn

• Pandas

Also need to have software.

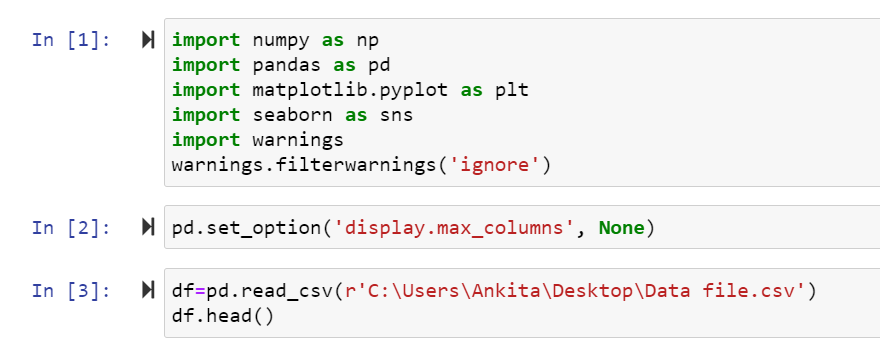
Install, run and execute a Jupyter notebook.

**EDA**:

There are no shortcuts in a machine learning project lifecycle. We can’t simply skip to the model building stage after gathering the data. We need to plan our approach in a structured manner and the exploratory data analytics (EDA) stage plays a huge part in that. I can say this with the benefit of hindsight having personally gone through this situation plenty of times. In my early days in this field, I couldn’t wait to dive into machine learning algorithms but that often left my end result hanging in the balance. I discovered, through personal experience and the advice of my mentors, the importance of spending time exploring and understanding my data.

## The Importance of Exploratory Data Analysis (EDA): There are no shortcuts in a machine learning project lifecycle. We can’t simply skip to the model building stage after gathering the data. We need to plan our approach in a structured manner and the exploratory data analytics (EDA) stage plays a huge part in that. I can say this with the benefit of hindsight having personally gone through this situation plenty of times. In my early days in this field, I couldn’t wait to dive into machine learning algorithms but that often left my end result hanging in the balance. I discovered, through personal experience and the advice of my mentors, the importance of spending time exploring and understanding my data.

**Loading dataset :**



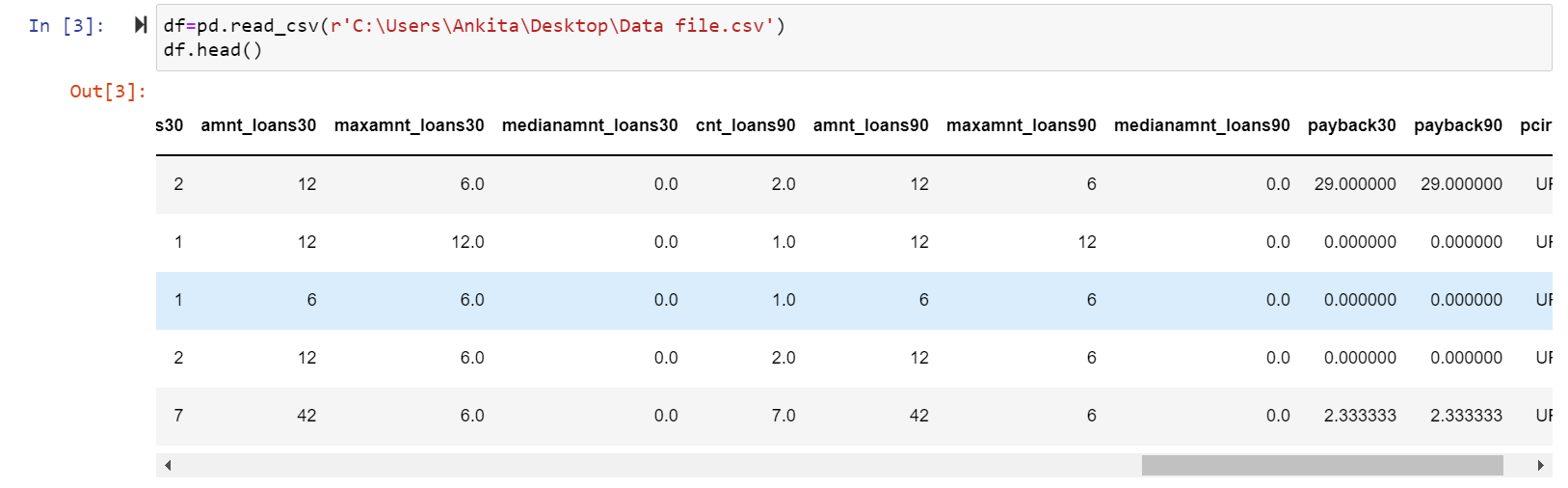
Here importing all necessary library and also load dataset.

**Data Sources and their formats**

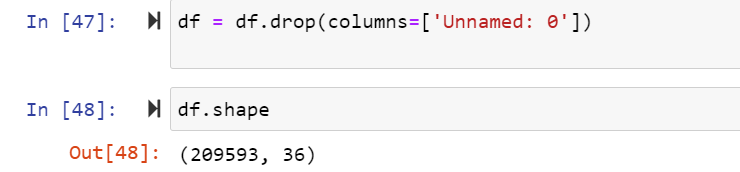
Then need to read and load dataset. And Then this below code isusefor

display all the columns in dataset. Let us have a look at how our dataset

The structure and details of the data are given below in this there are dataset in which no of row and columns are present:



Here in this dataset label are target variable and there was numeric values are present and there are only 1 and 0 are only two values was present so this is **Classification Problem** so Need use Classification Algorithm.



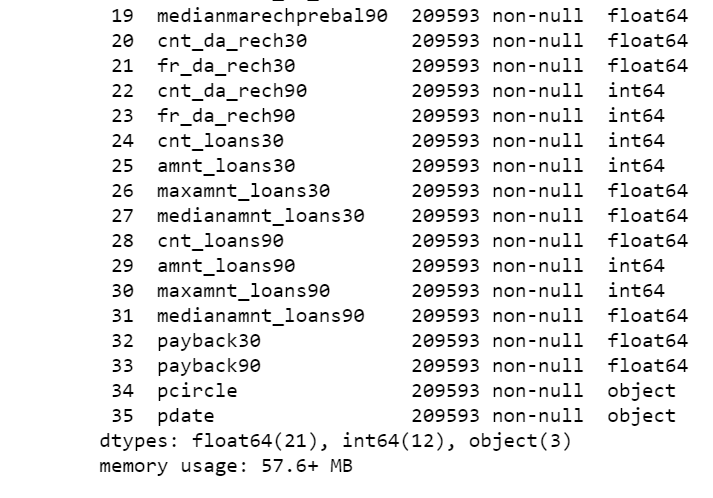
Check the how much Columns and rows present in dataset using df.shape() is use to check the rows and columns count in this above dataset there are 209593 rows and 36 columns are present in this dataset And drop the Unnamed: 0 columns from dataset because it is not much important.

## Classification Algorithms:

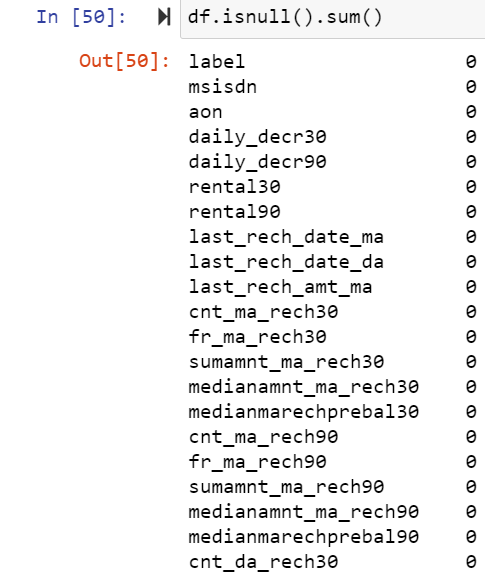
### The study of classification in statistics is vast, and there are several types of classification algorithms you can use depending on the dataset you’re working with. Below are five of the most common algorithms in machine learning. The algorithms we are going to cover are:

* Logistic Regression.
* Naive Bayes.
* K-Nearest Neighbors.
* Decision Tree.
* Support Vector Machines.



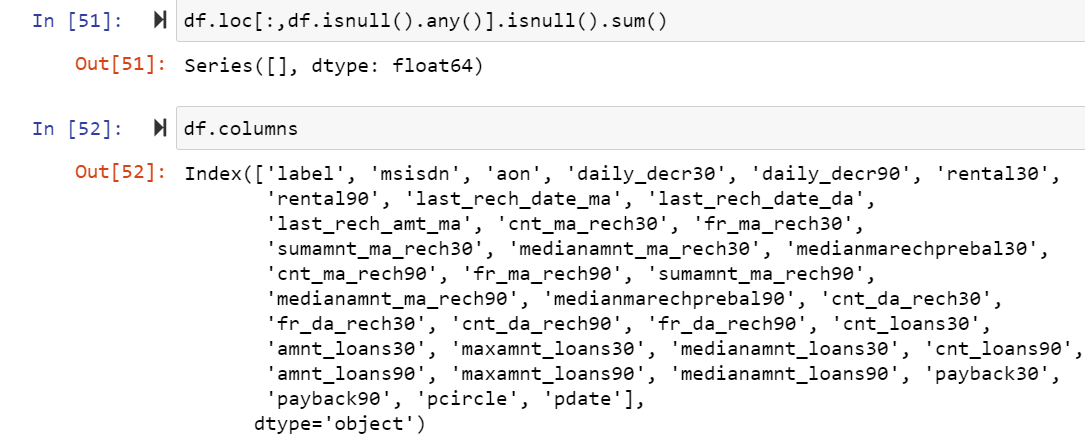


In above using df.info() is used to get the information all the columns in dataset and its data types. In above dataset there are 35 columns in this dataset. And 3 columns are object data types.

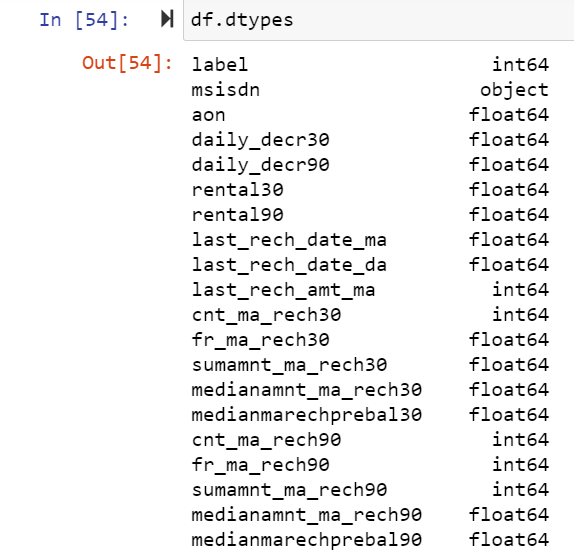


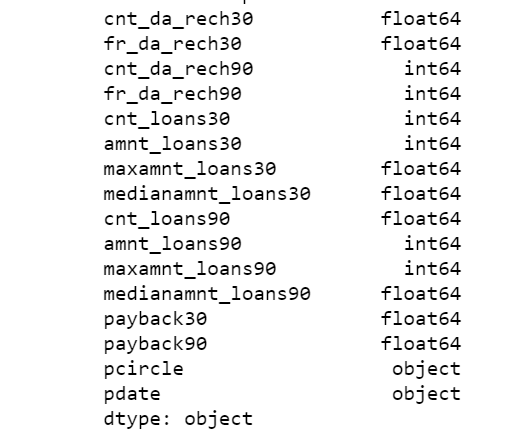


Check the how much missing values present in dataset using df.isnull() is use to check the missing values in dataset in this dataset there are no of missing values are present in train dataset.

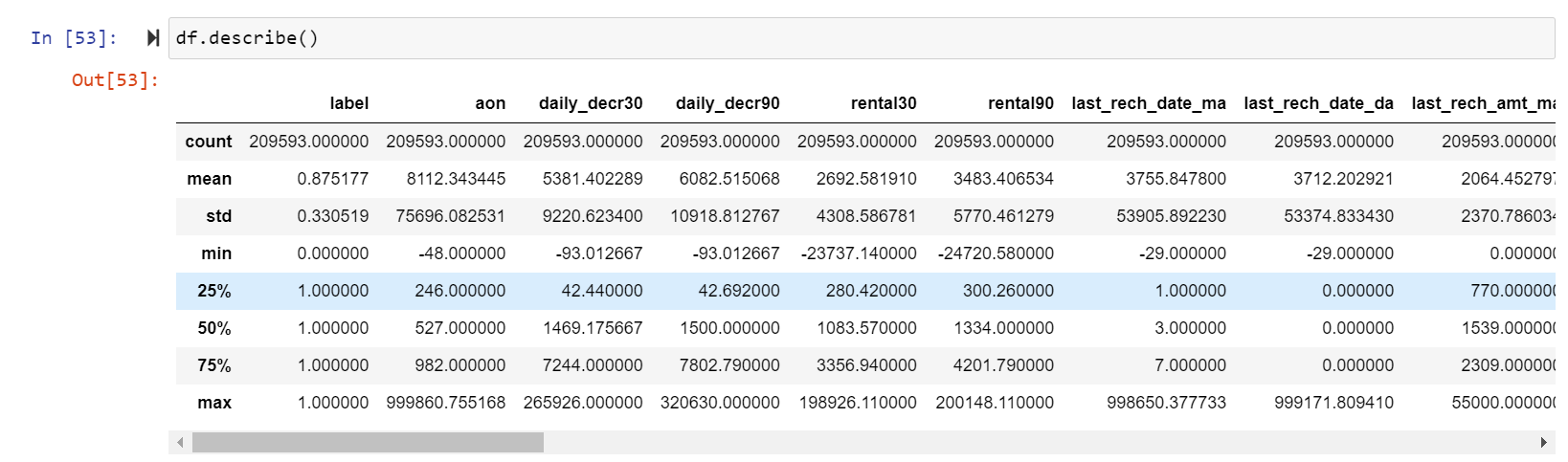


In above df.colums is use to get all the column names in this dataset.

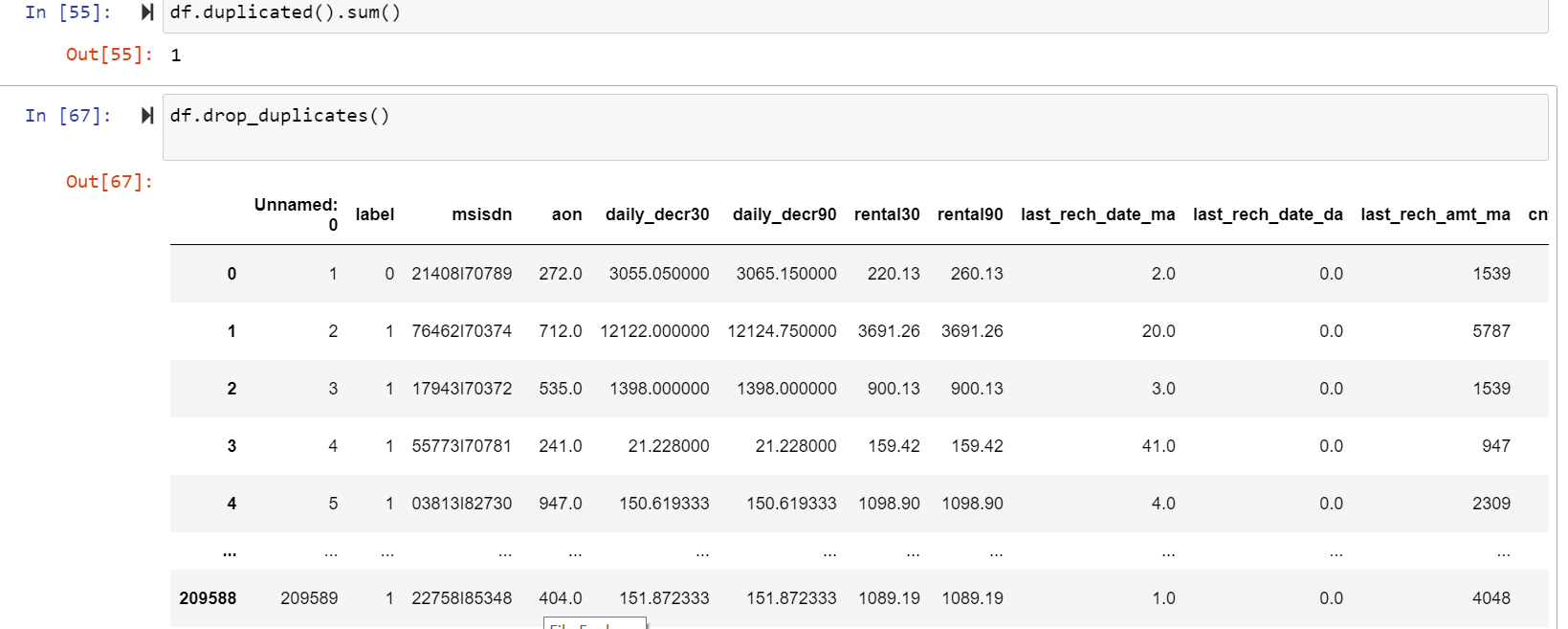
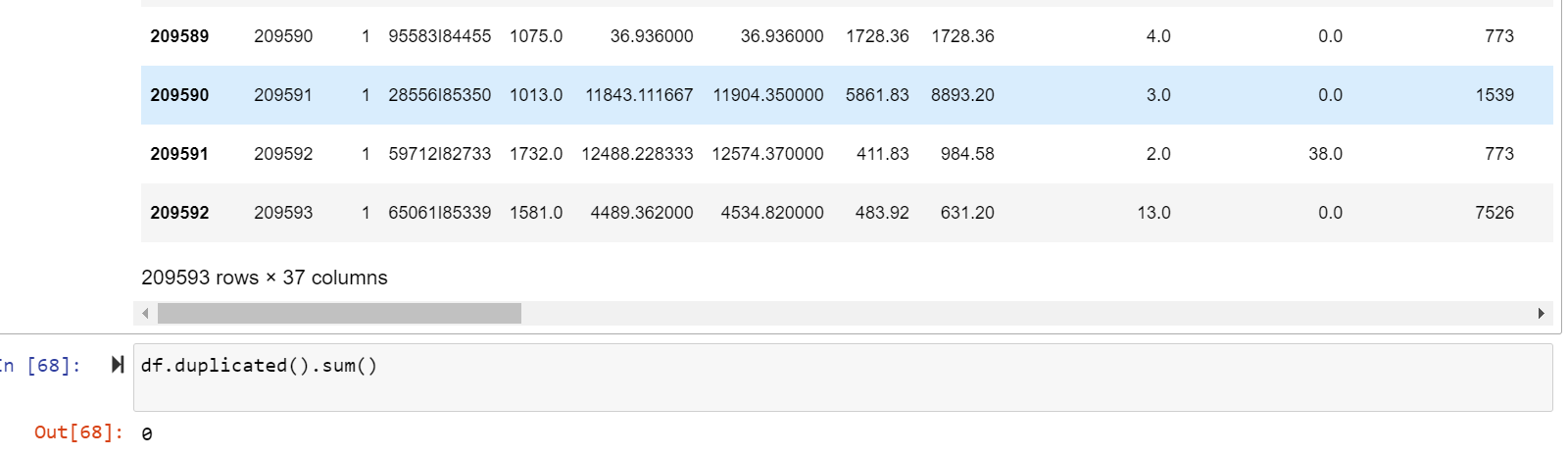




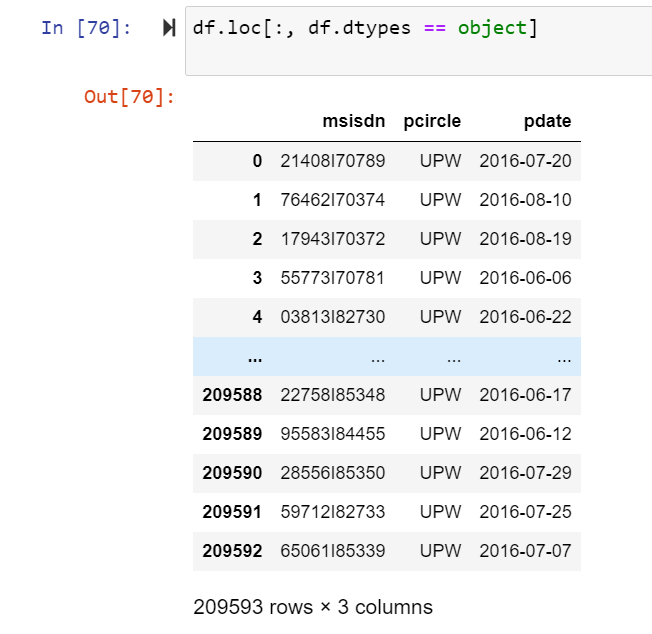
In above code df.dtypes is used to check the data types of all the column In above dataset In this 3 columns data types was object.



Above describe the information of above dataset using df.describe().

Above check the duplicates values of above dataset there was 1 duplicates values are present in dataset and drop that duplicates values from dataset there was no duplicate values was present in dataset.

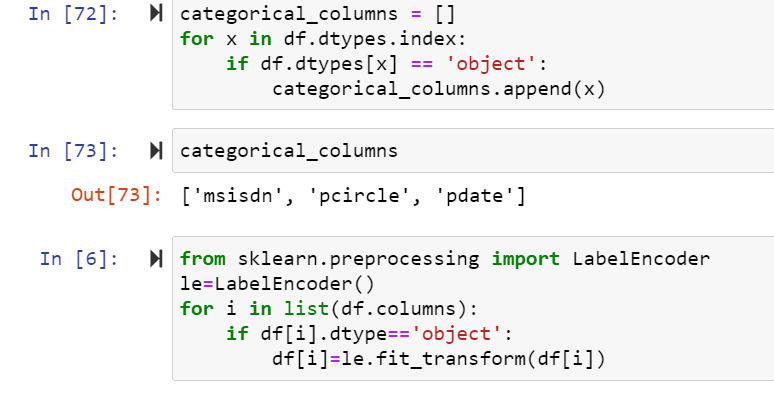


Here above are list of all columns in dataset in which data types are object.

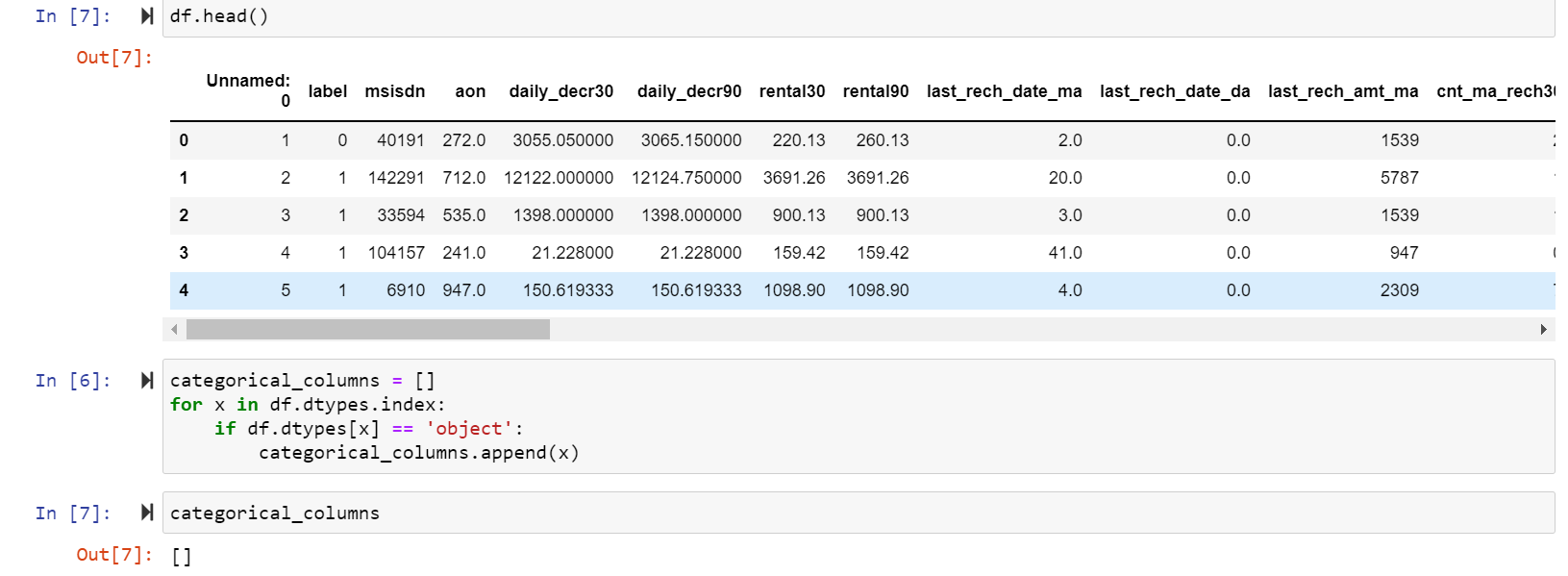
**Data Preprocessing Done:**

**Data Inputs- Logic- Output Relationships**

Here we converted categorical data into numeric in dataset because without converting categorical data into numeric we cannot find out correlation of the dataset.

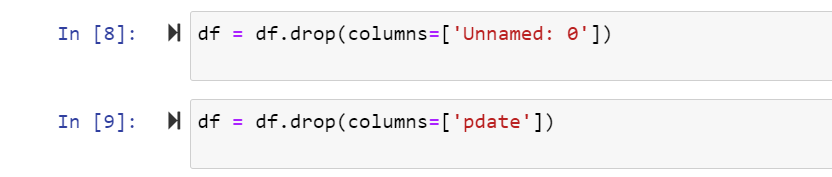


In this above code display the categorical columns in and use lable encoder technique to convert categorical columns into numeric dataset.



Here use label encoder to replace the categorical values into numeric in this

dataset.



Here drop Unnamed :0 and pdate column because it is not much important.

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

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.

**Hardware and Software Requirements and Tools Used**

1. Software Requirements:

* 1. Coding Language: Python3, Python
  2. Coding software : Anaconda, Jupyter Notebook

1. Microsoft Office Word.
2. Snipping Tools (For Screenshots).
3. Microsoft Excel

**Non Functional Requirements:**

1: Platform Independent: The application would be platform independent if all the requirements are installed in the device.

2: Performance: The application should have better accuracy and should provide the information in less time.

3: Capacity: The capacity of the storage should be high so that large amount of data can be stored in order to train the model.

**Hardware Requirements:**

1 GB RAM.

200 GB HDD.

Intel 1.66 GHz Processor Pentium 4

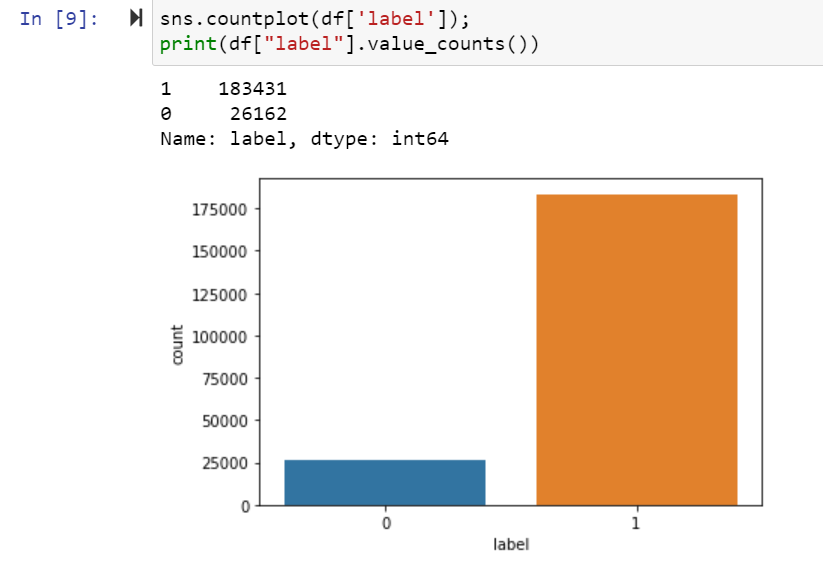
**Graphical representation**

### We will start with Univariate Analysis. We will be using a bar graph for this purpose.

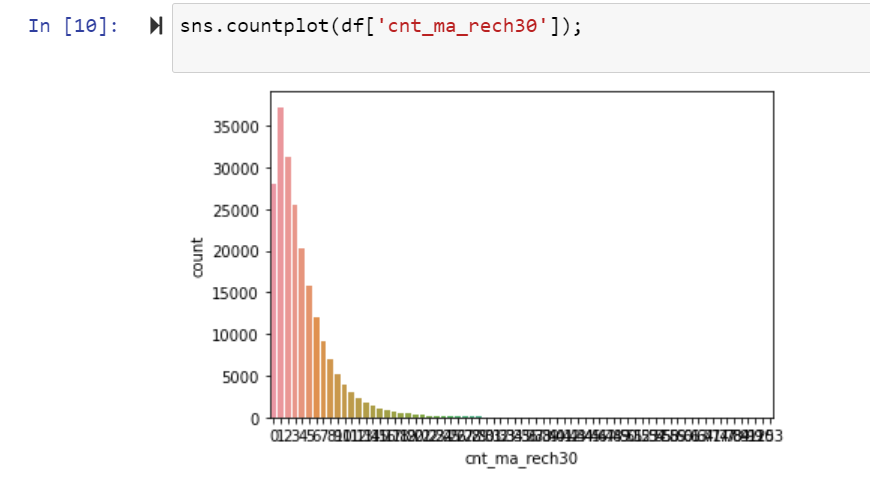
1. Univarient Analysis.
2. Bi Varient Analysis
3. Multivarient Analysis

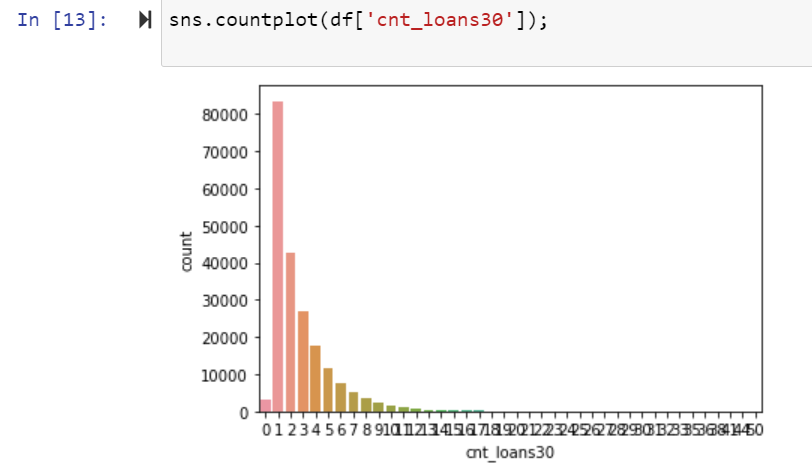
**Univarient Analysis of dataset**:

Univariate analysis **explores each variable in a data set**, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.

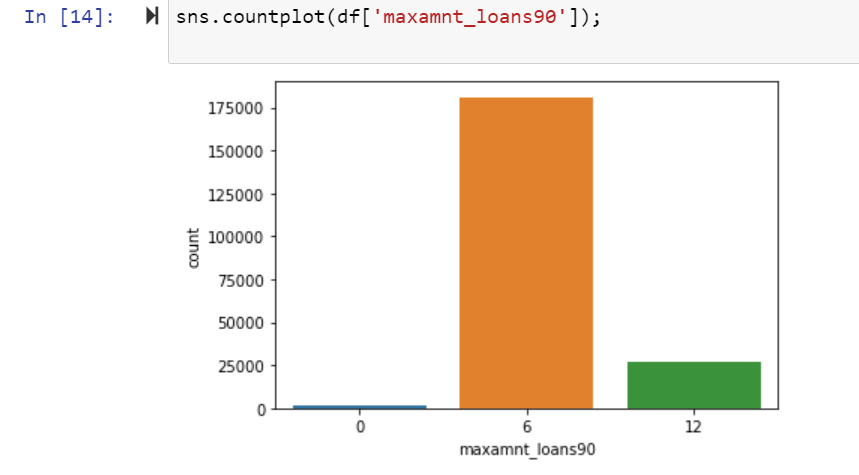


Univarient analysis of ‘label’ column in this dataset in this label was the target column and there was class imbalance problem is occur.

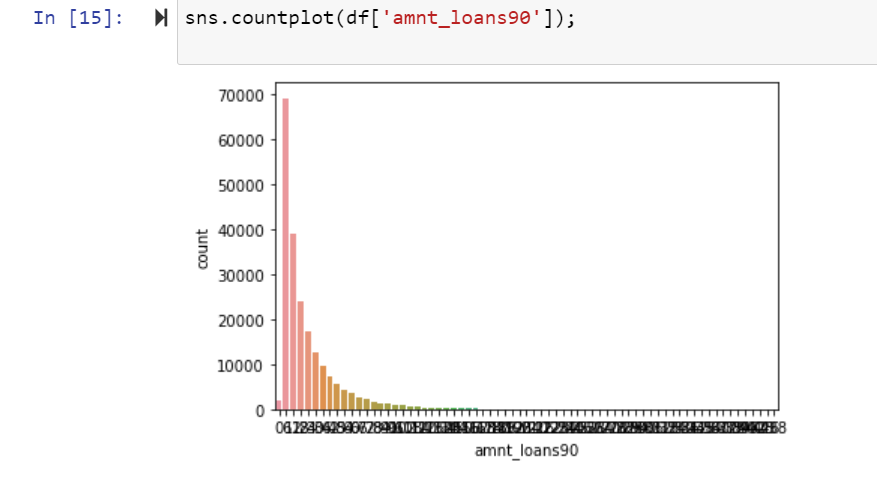
Univarient analysis of cnt\_ma-rech30 column in this dataset.



Univarient analysis of cnt\_loans30 column in this dataset.



Univarient analysis of max\_loans90 column in this dataset.

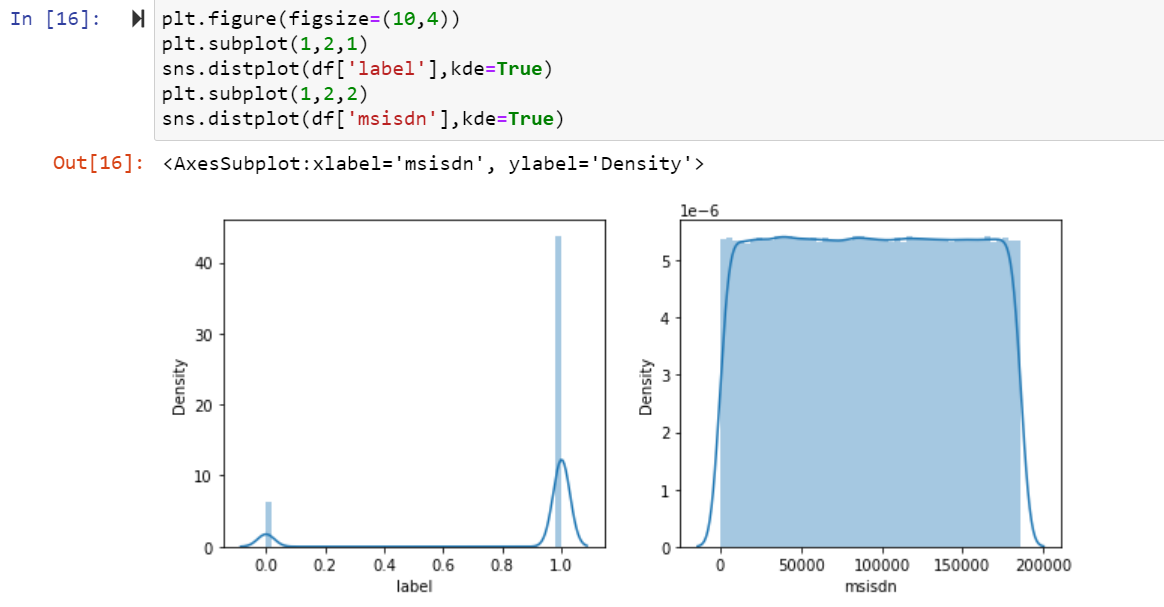


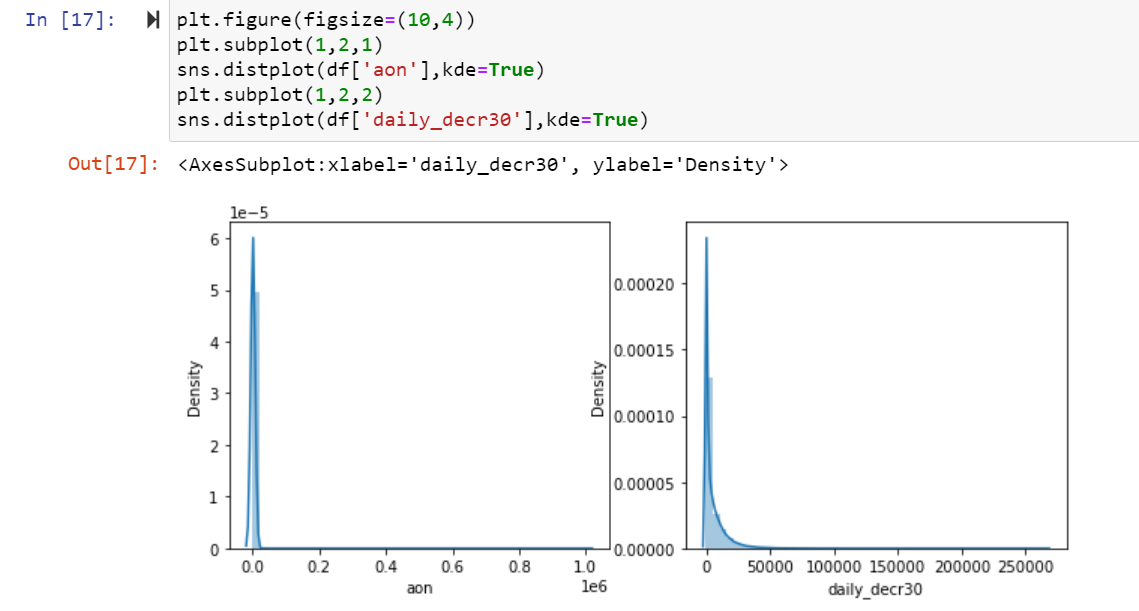
Univarient analysis of amnt\_loans90 column in this dataset.

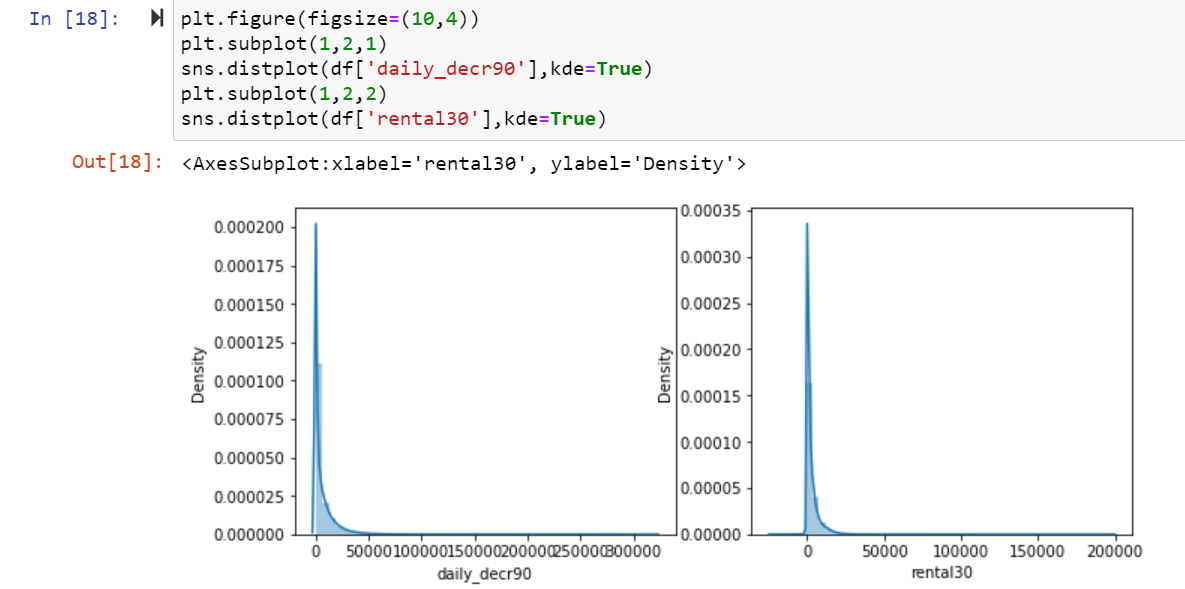
**Checking Distribution of columns in this dataset they are normally distributed or not:**

Now Check The all the columns data are normally Distributed or not:

Normally distributed data, there is a constant proportion of data points lying under the curve between the mean and a specific number of standard deviations from the mean. Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.







In this above distribution there was sum columns are normally is distributed and some columns are not normally distributed.

**Bi varient Analysis:**

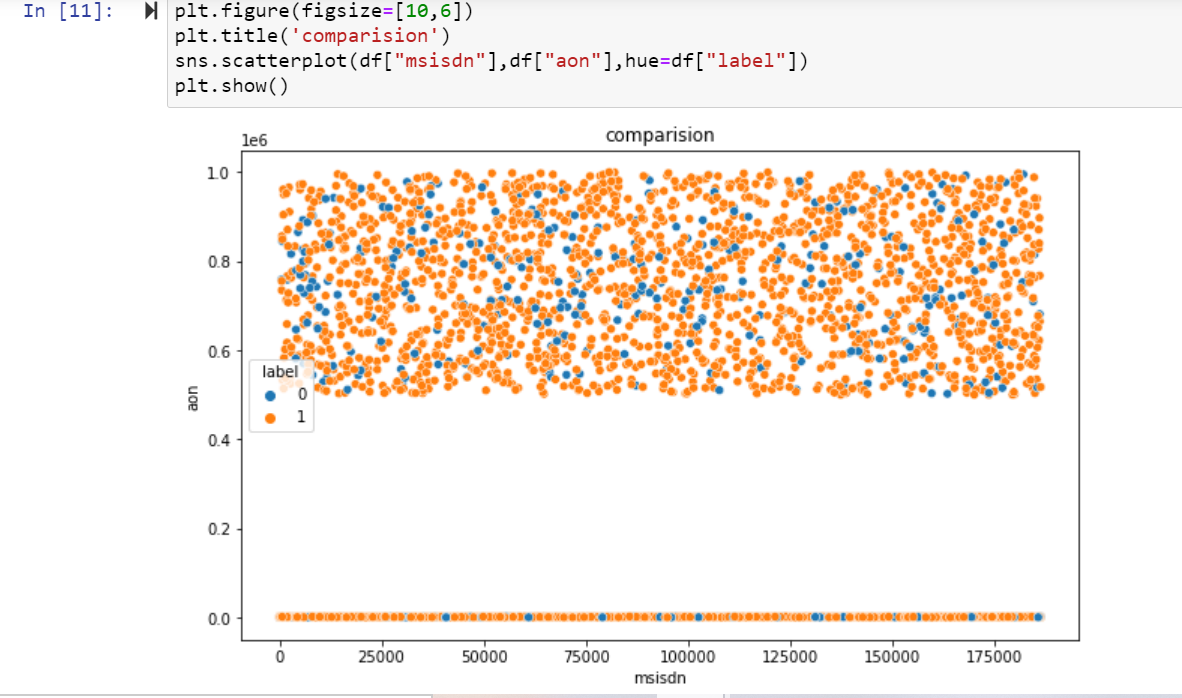
A scatter plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.

#### Scatter Plot of Dataset:

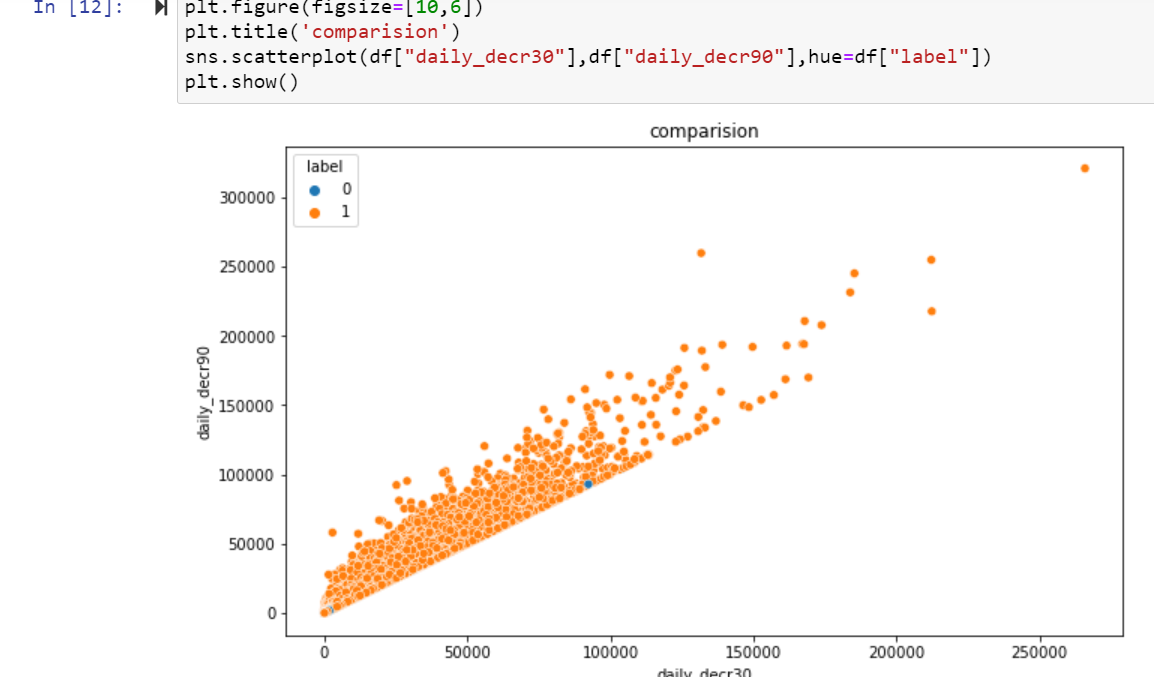
Scatter plot is a graph of two sets of data along the two axes. It is used to visualize the relationship between the two variables.

If the value along the Y axis seem to increase as X axis increases(or decreases), it could indicate a positive (or negative) linear relationship. Whereas, if the points are randomly distributed with no obvious pattern, it could possibly indicate a lack of dependent relationship.

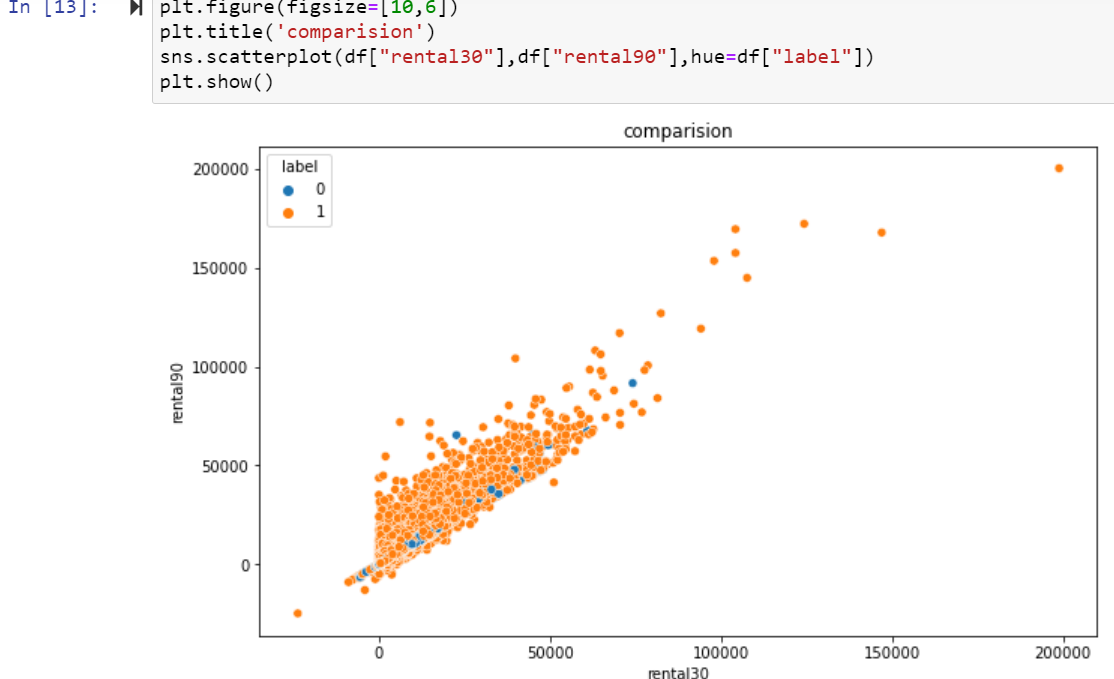
In python matplotlib, the scatterplot can be created using the pyplot.plot() or the pyplot.scatter(). Using these functions, you can add more feature to your scatter plot, like changing the size, color or shape of the points.



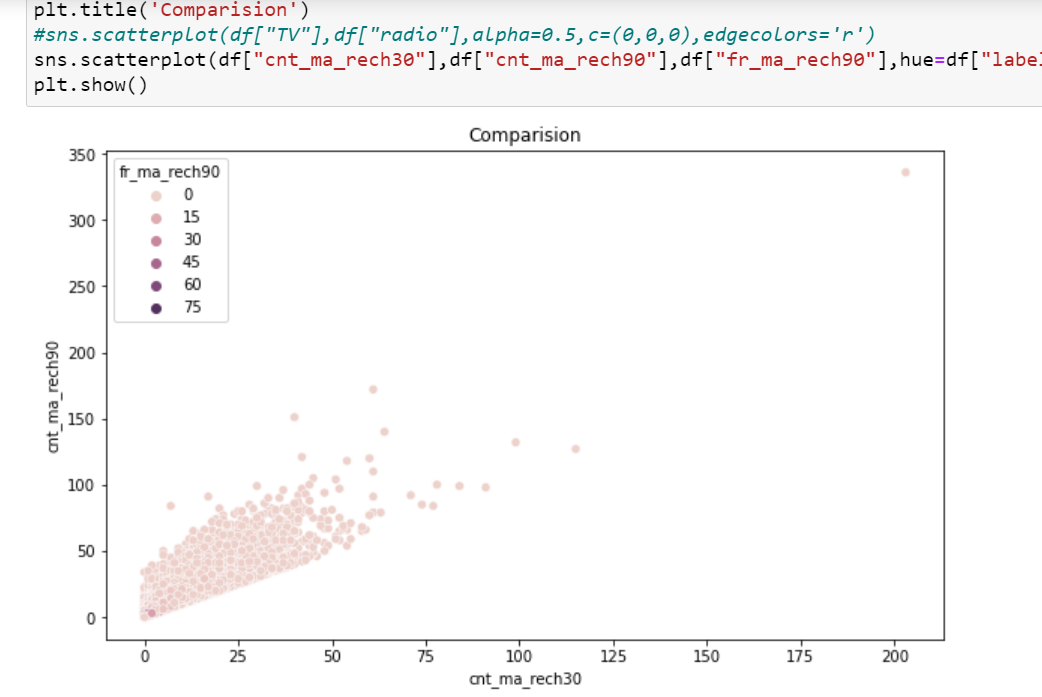
Here Above Scatterplot shows the comparison between two columns.



Here Above Scatterplot shows the comparison between two columns.



Here Above Scatterplot shows the comparison between two columns.

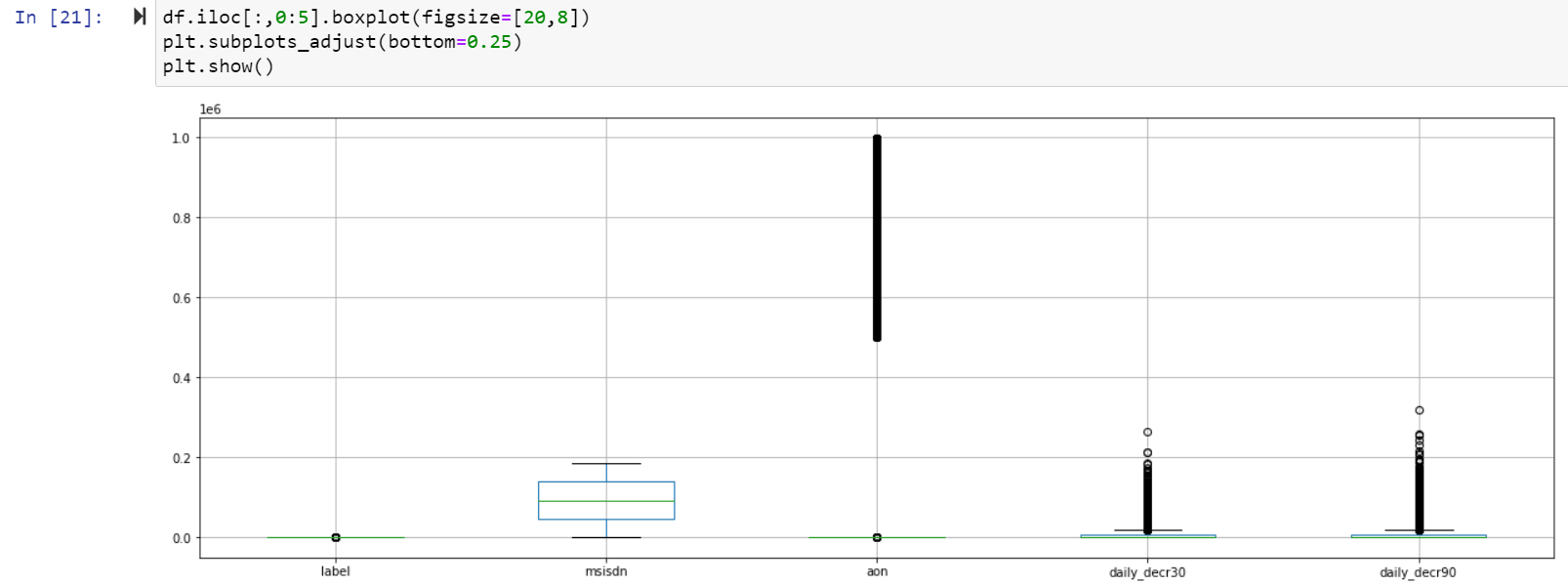


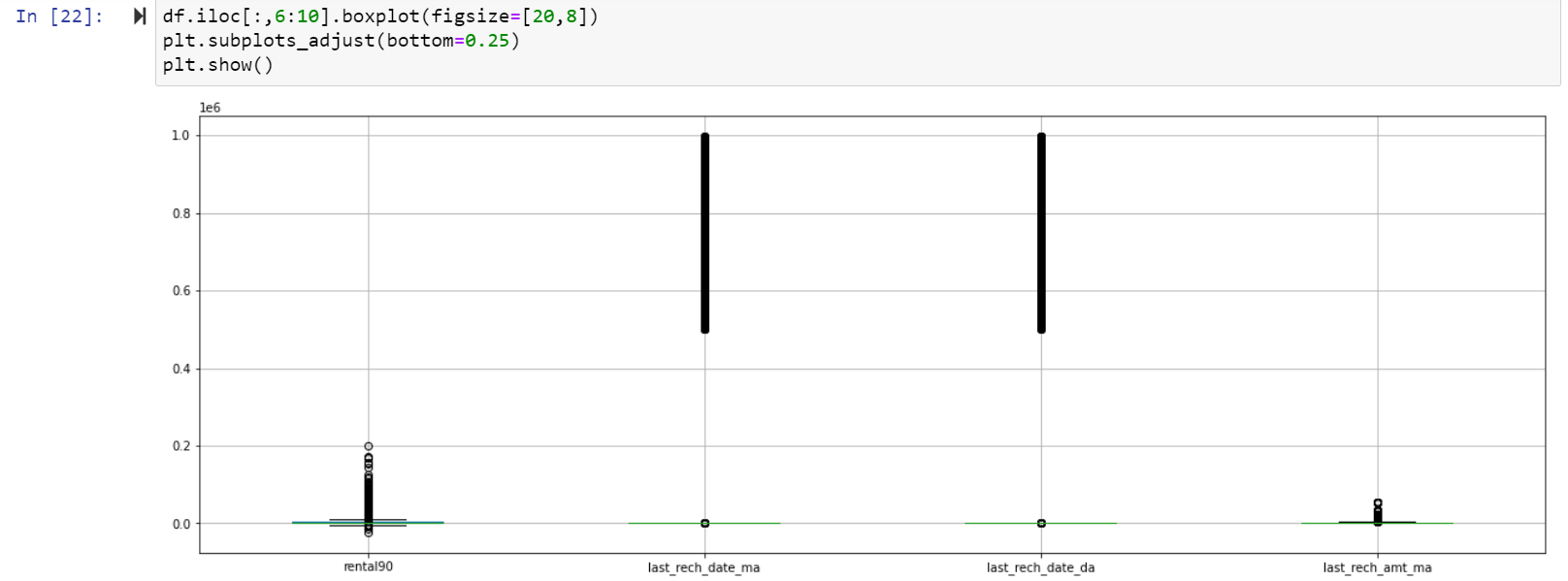
Here Above Scatterplot shows the comparison between three columns.

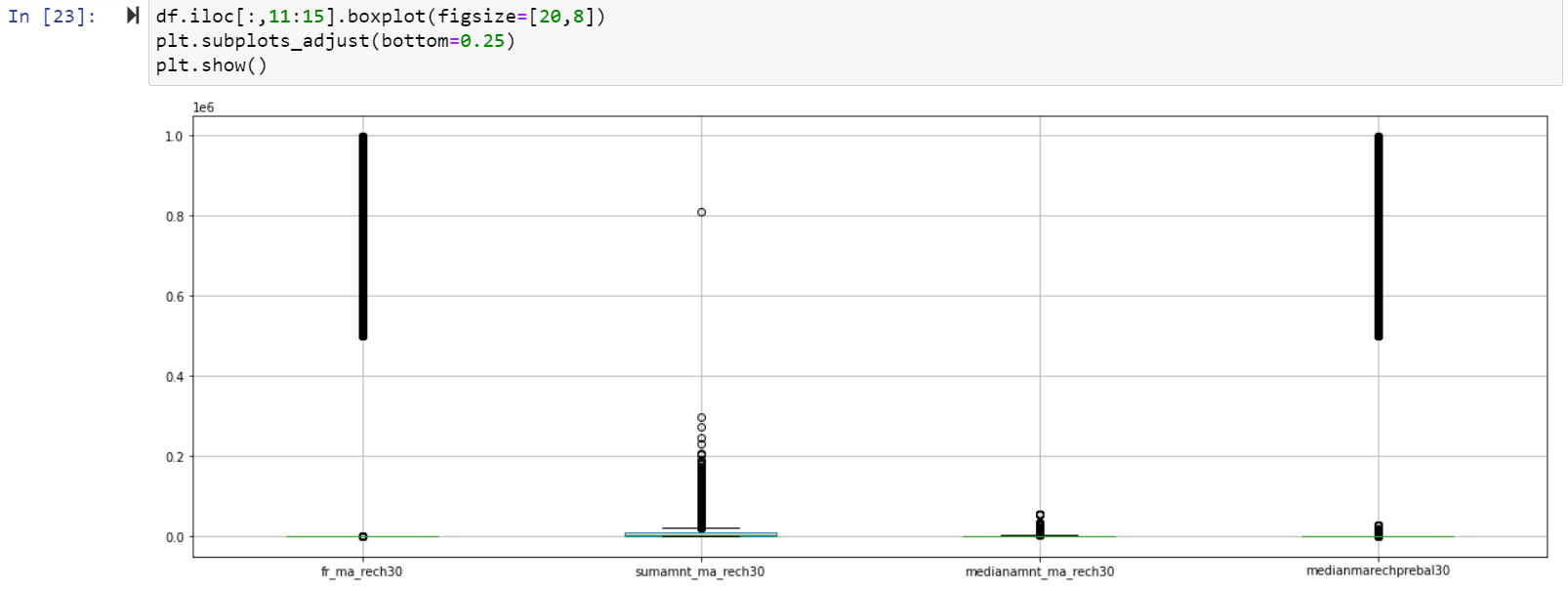
.

**Checking Outliers of Dataset:**

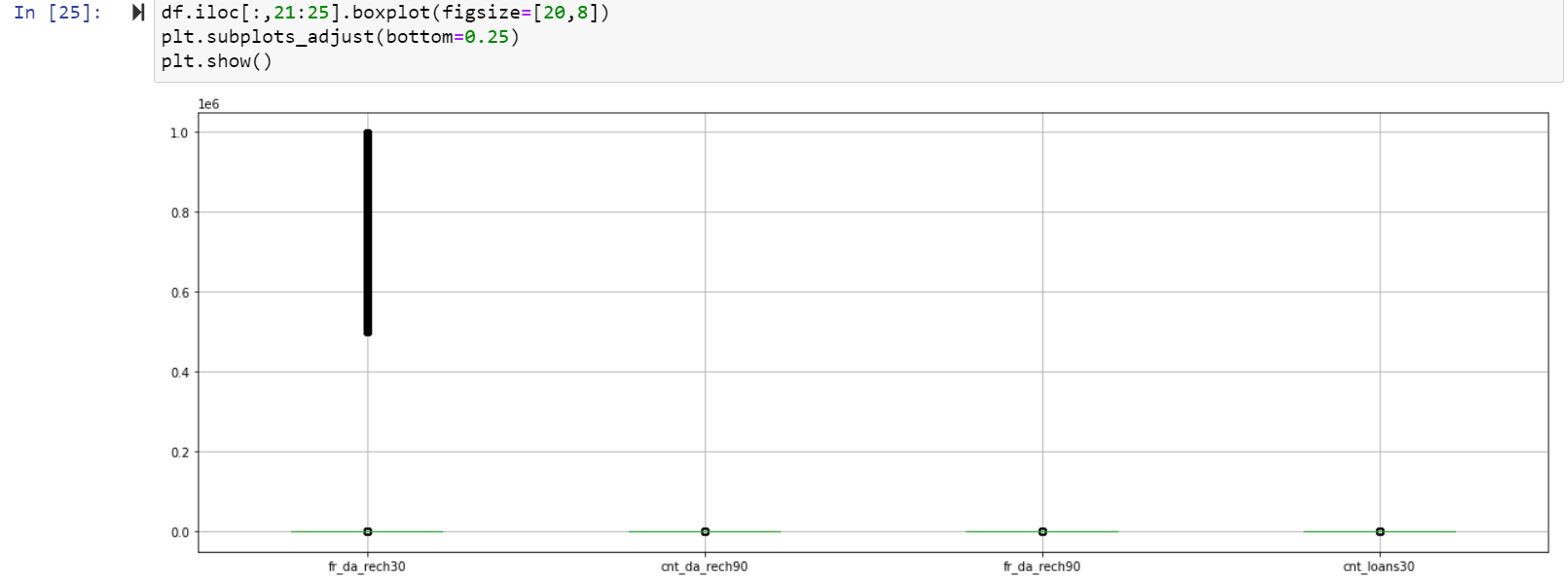
An outlier is an **object(s) that deviates significantly from the rest of the object collection**. It is an abnormal observation during the Data Analysis stage, that data point lies far away from other values. An outlier is an observation that diverges from well-structured data.

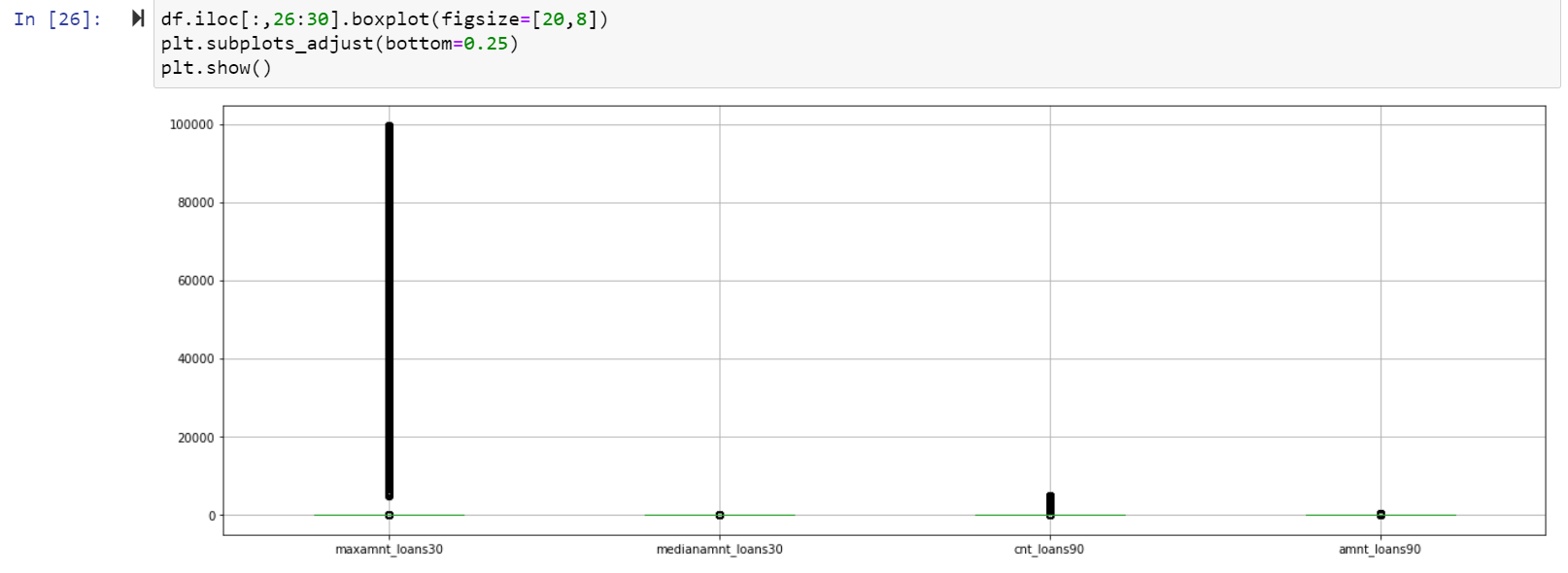


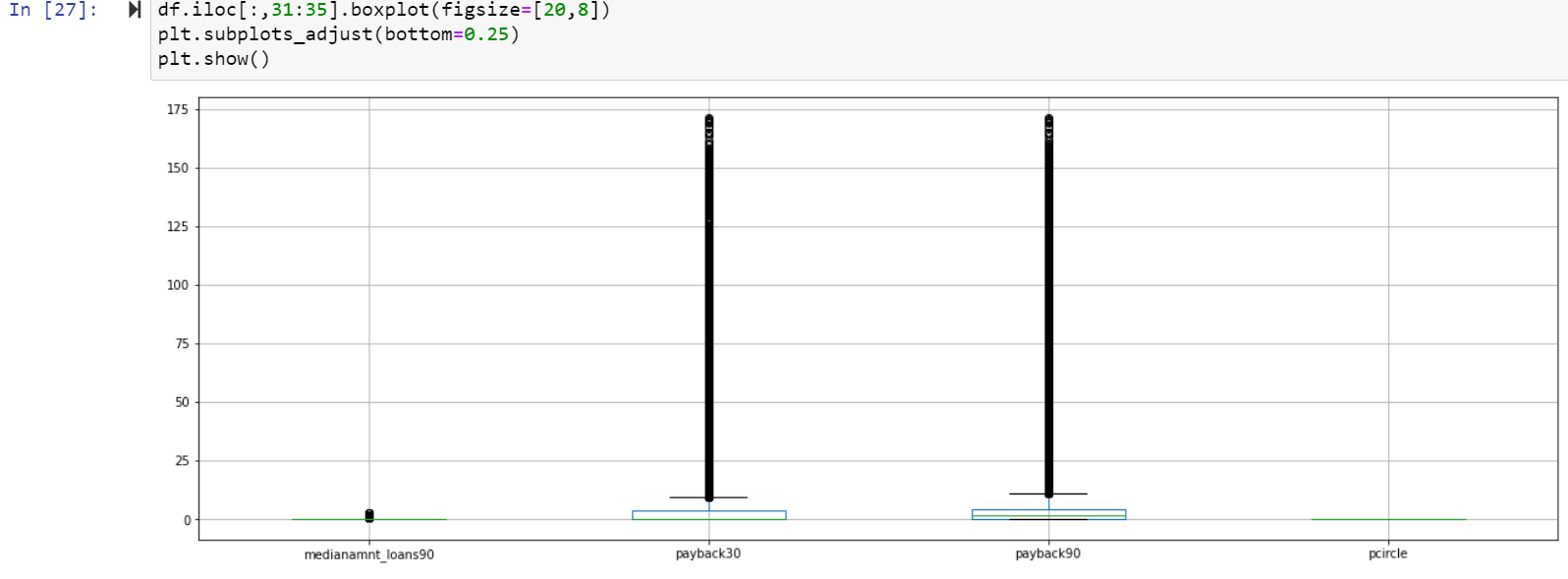




### C:\Users\Ankita\Desktop\pp\Untitled37.png







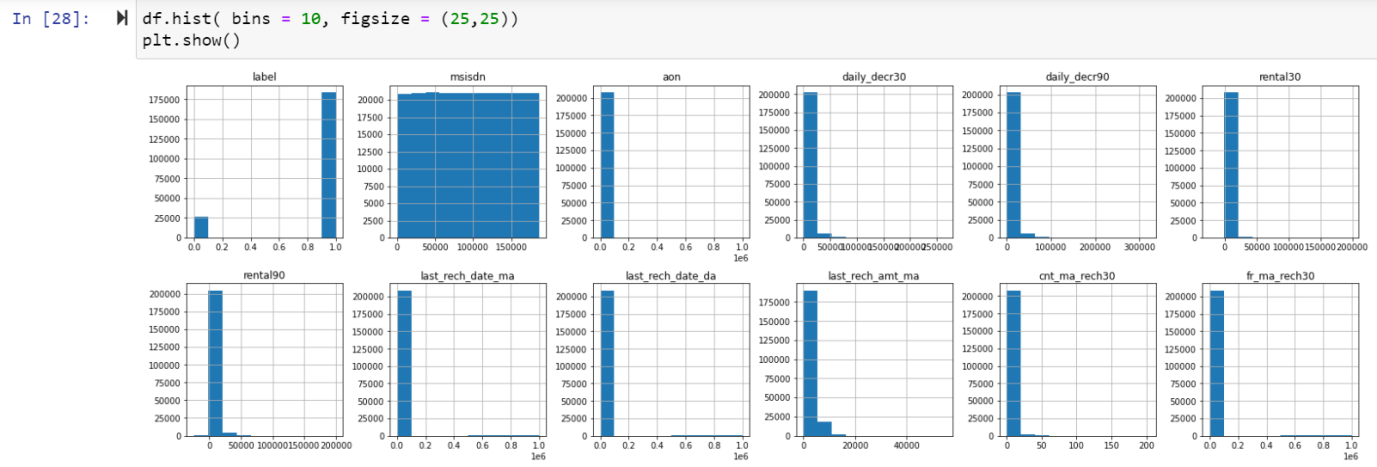
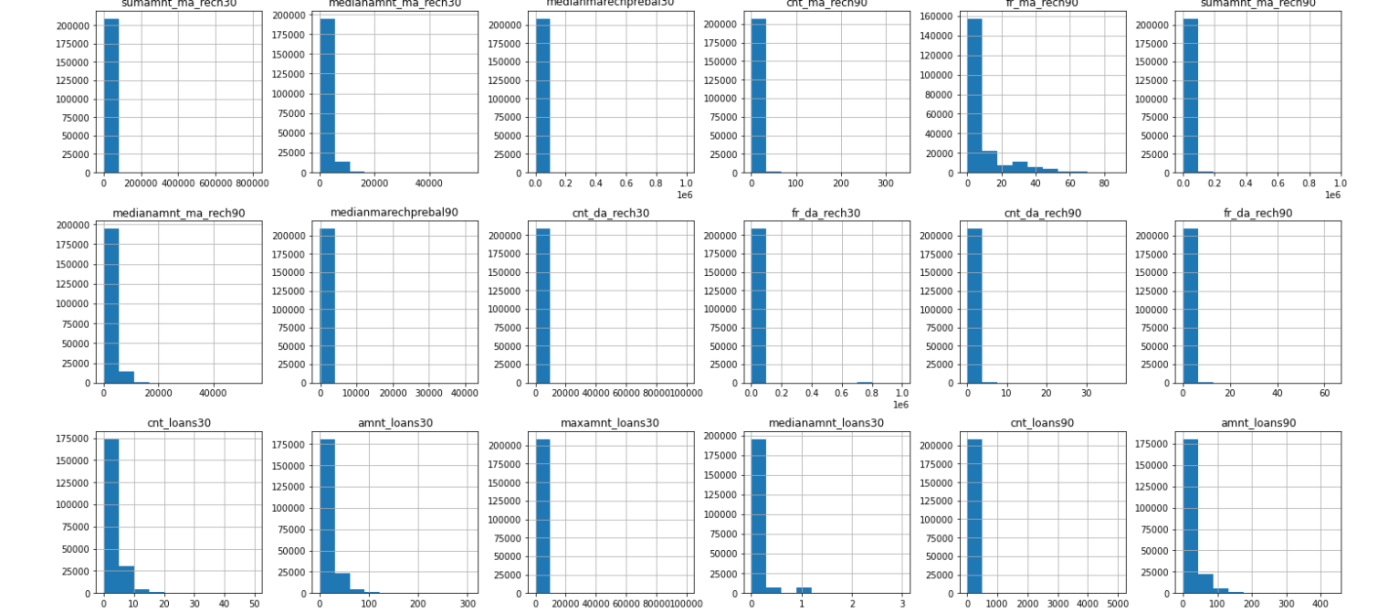
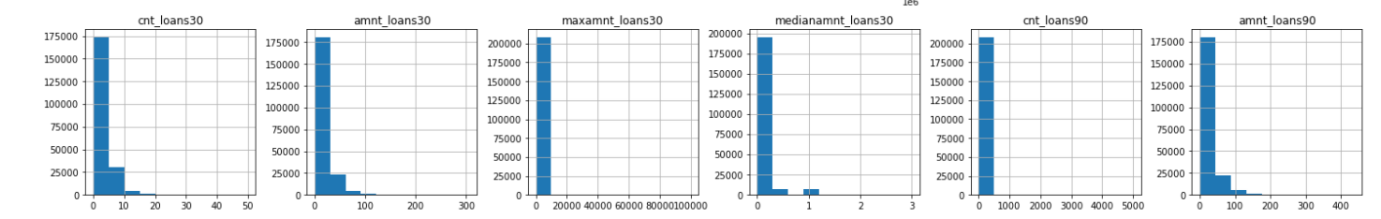
Here many columns Outliers was present.

**Multi Varient Analysis of Dataset:**

Multivariate analysis is used **to study more complex sets of data than what univariate analysis methods can handle**. ... Multivariate analysis can reduce the likelihood of Type errors. Sometimes, univariate analysis is preferred as multivariate techniques can result in difficulty interpreting the results of the test.

**Histplot of Dataset**:

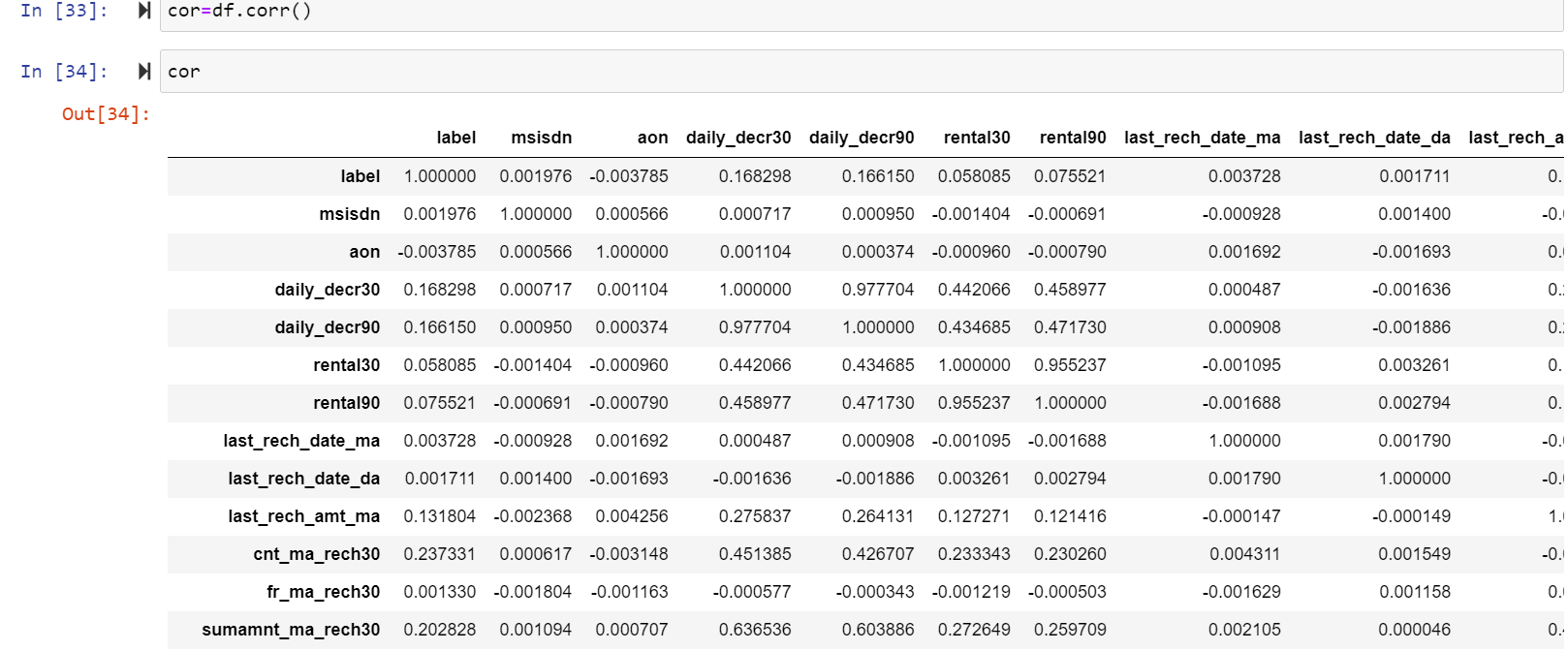
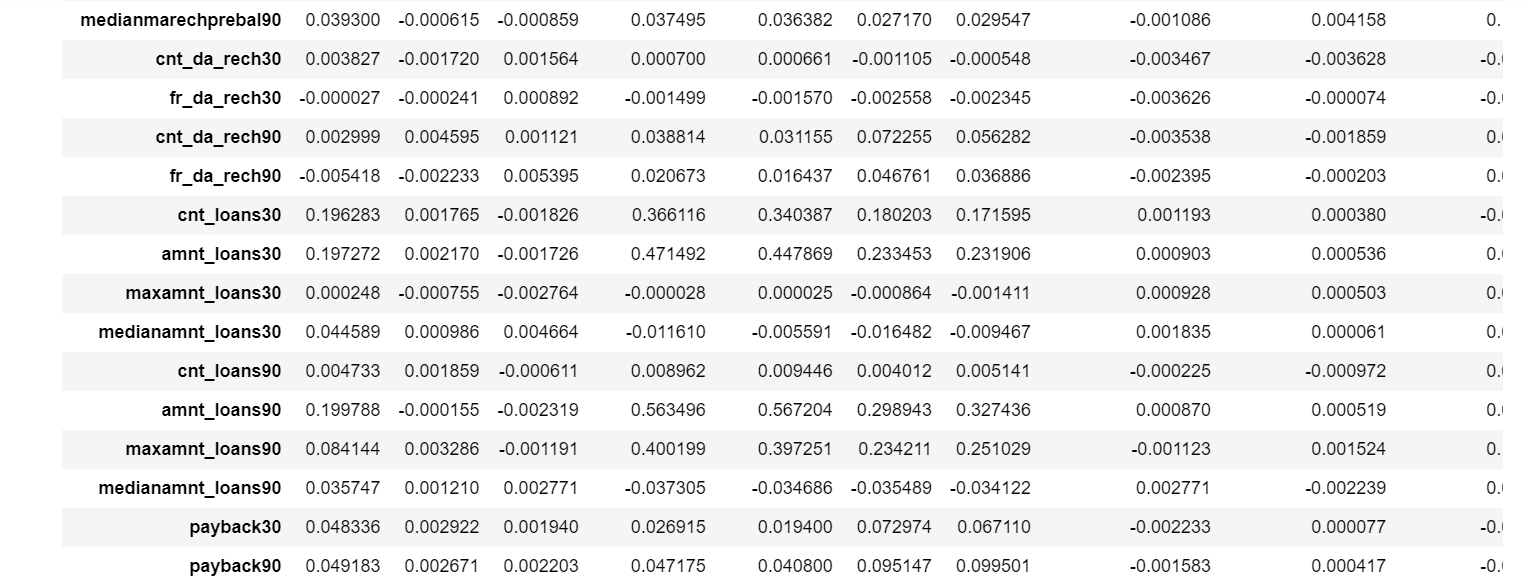
A histogram is a classic visualization tool that represents the. A histogram provides **a visual representation of the distribution of a dataset**: location, spread and skewness of the data; it also helps to visualize whether the distribution is symmetric or skewed left or right. ... In brief, a histogram summarizes the distribution properties of a continuous numerical variable.

**Visualizations:**

**Heatmap:**

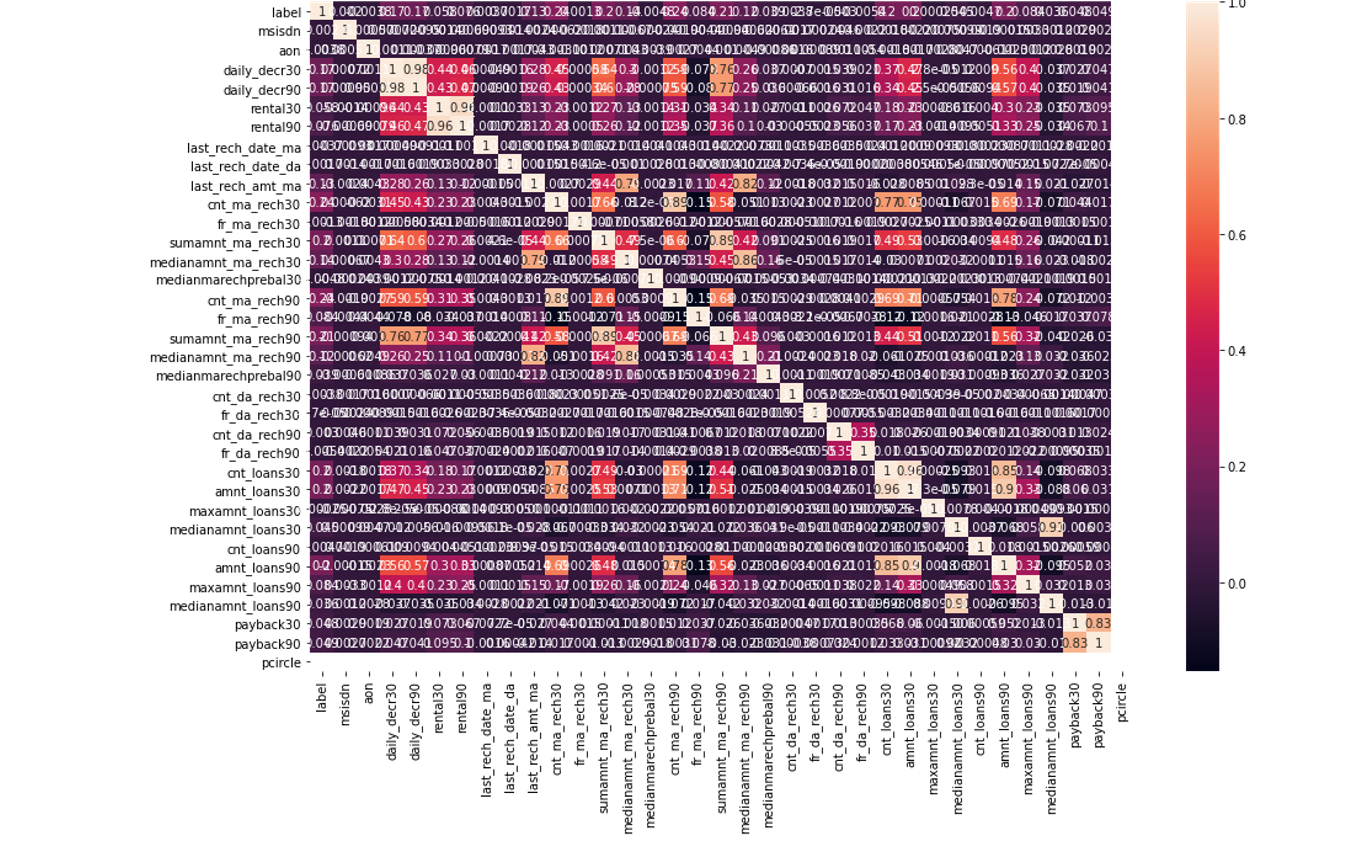
Correlation heatmap is graphical representation of correlation matrix representing correlation between different variables. The value of correlation can take any values from -1 to 1. ... Correlation between two variables can also be determined using scatter plot between these two variables.

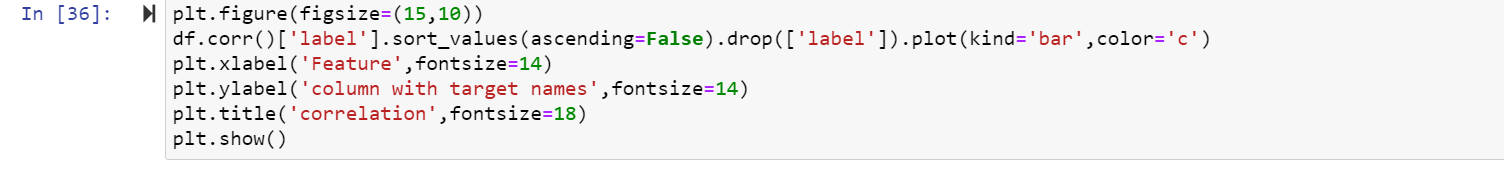
After that Checking Correlation of all independent columns with Target column.

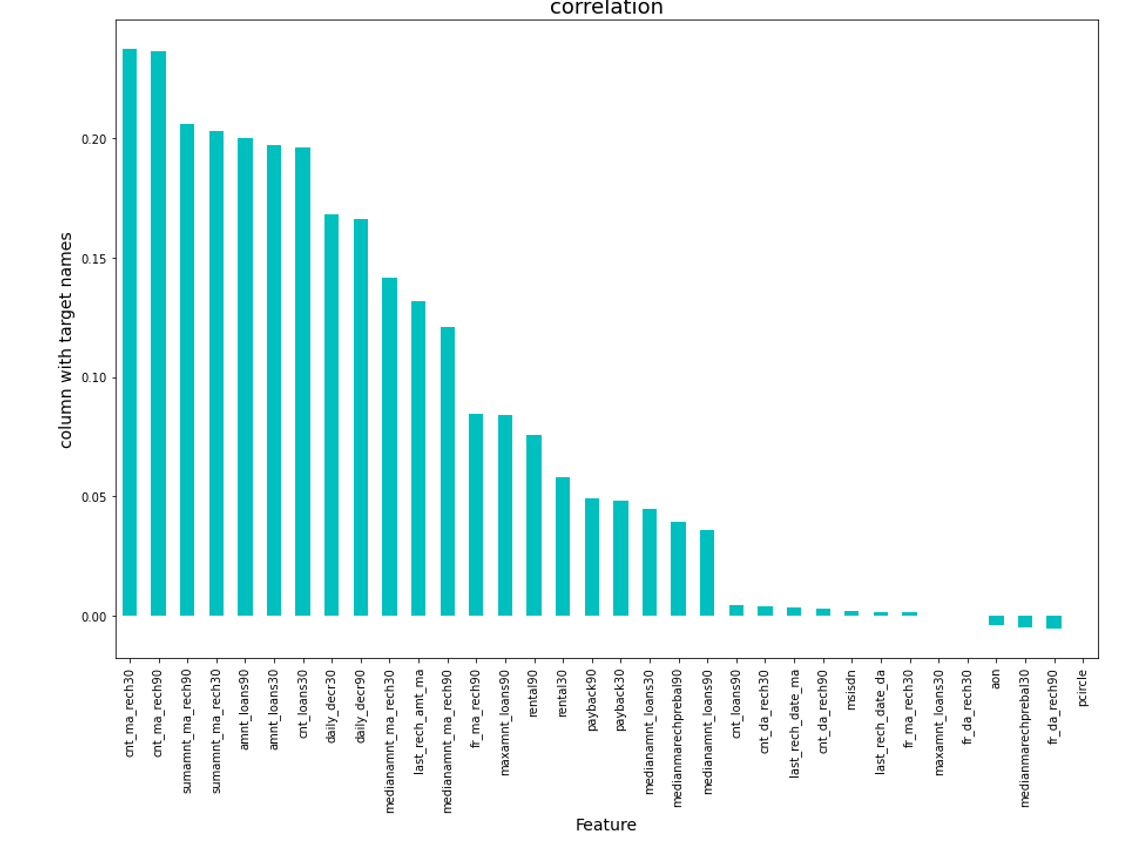
plt.figure(figsize=(15,10))

sns.heatmap(df.corr(),annot=True)



Heat map shows the correlation of every independent variable in dataset with target variable. Here above heatmap the every independent variable check correlation with label target variable.





Here we shows the correlation in another way in this above we show some columns are positively and some columns are negatively correlated with target variable.

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

In this we analyse the accuracy of predicting loans when using Multiple classification algorithms. Thus, the purpose of this is to depend the knowledge in classification methods in machine learning. In addition, the given datasets should be processed to enhance performance, which is accomplished by identifying the necessary features by applying one of the selection methods 2 to eliminate the unwanted variables.

**Z-score:**

Take your data point, subtract the mean from the data point, and then divide by your standard deviation. That gives you your Z-score. You can use Z-Score to determine outliers.One of the most commonly used tools in determining outliers is the Z-score. Z-score is just the number of standard deviations away from the mean that a certain data point is.In your future data science life, Z-scores are gonna be a really useful way to think about how usual or how unusual a certain data point is. And that’s going to be really valuable once we start making inferences based on our data. In this story, we will take a deep dive into our notebooks and learn how to detect outliers using Z-Score.



After that Checking Data loss of the dataset after using preprocessing steps here using zscore method checking data loss here after zscore there 22.96 is dataloss.

**IQR:**

The interquartile range rule is useful in detecting the presence of outliers. [Outliers](https://www.thoughtco.com/what-is-an-outlier-3126227) are individual values that fall outside of the overall pattern of a data set. This definition is somewhat vague and subjective, so it is helpful to have a rule to apply when determining whether a data point is truly an outlier—this is where the interquartile range rule comes in.

## What Is the Interquartile Range?

Any set of data can be described by its [five-number summary](https://www.thoughtco.com/what-is-the-five-number-summary-3126237). These five numbers, which give you the information you need to find patterns and outliers, consist of (in ascending order):

* The minimum or lowest value of the dataset
* The first quartile Q1, which represents a quarter of the way through the list of all data
* The [median](https://www.thoughtco.com/what-is-the-median-3126370) of the data set, which represents the midpoint of the whole list of data
* The third quartile Q3, which represents three-quarters of the way through the list of all data
* The maximum or highest value of the data set.

These five numbers tell a person more about their data than looking at the numbers all at once could, or at least make this much easier. For example, the [range](https://www.thoughtco.com/what-is-the-range-in-statistics-3126248), which is the minimum subtracted from the maximum, is one indicator of how spread out the data is in a set (note: the range is highly sensitive to outliers—if an outlier is also a minimum or maximum, the range will not be an accurate representation of the breadth of a data set).

Range would be difficult to extrapolate otherwise. Similar to the range but less sensitive to outliers is the interquartile range. The [interquartile range](https://www.thoughtco.com/what-is-the-interquartile-range-3126245) is calculated in much the same way as the range. All you do to find it is subtract the first quartile from the third quartile:

IQR = Q3 – Q1.

The interquartile range shows how the data is spread about the median. It is less susceptible than the range to outliers and can, therefore, be more helpful.

## Using the Interquartile Rule to Find Outliers

Though it's not often affected much by them, the interquartile range can be used to detect outliers. This is done using these steps:

1. Calculate the interquartile range for the data.
2. Multiply the interquartile range (IQR) by 1.5 (a constant used to discern outliers).
3. Add 1.5 x (IQR) to the third quartile. Any number greater than this is a suspected outlier.
4. Subtract 1.5 x (IQR) from the first quartile. Any number less than this is a suspected outlier.

Remember that the interquartile rule is only a rule of thumb that generally holds but does not apply to every case. In general, you should always follow up your outlier analysis by studying the resulting outliers to see if they make sense. Any potential outlier obtained by the interquartile

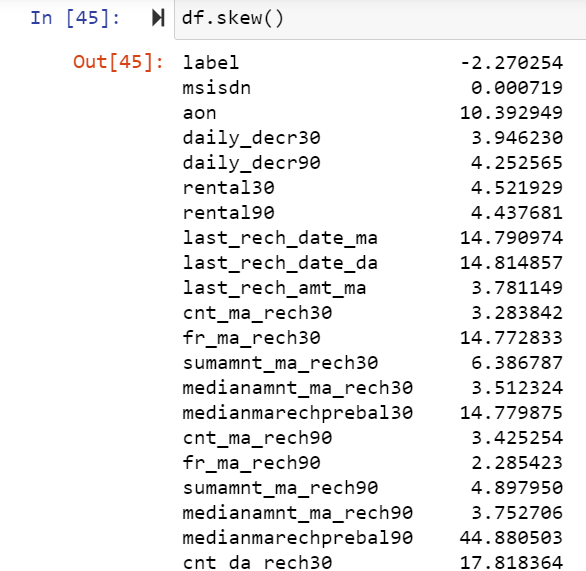


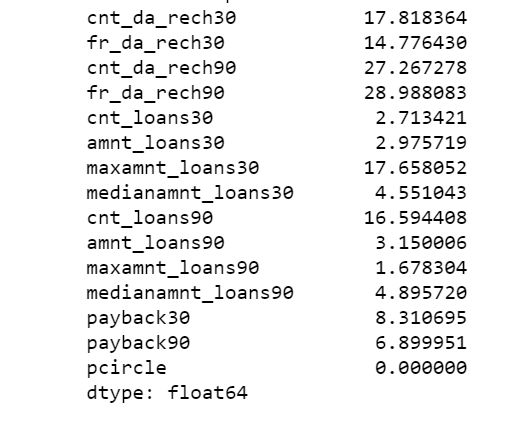
Here After Applying IQR method checking dataloss here using IQR method more dataloss.

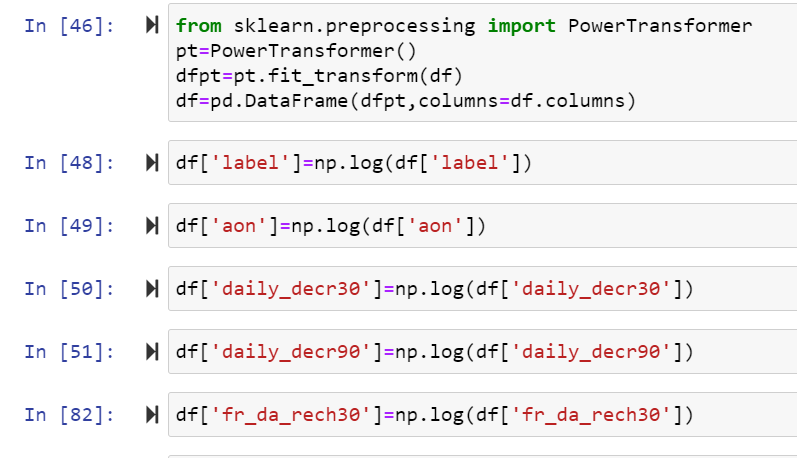
**Skewness:**

The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

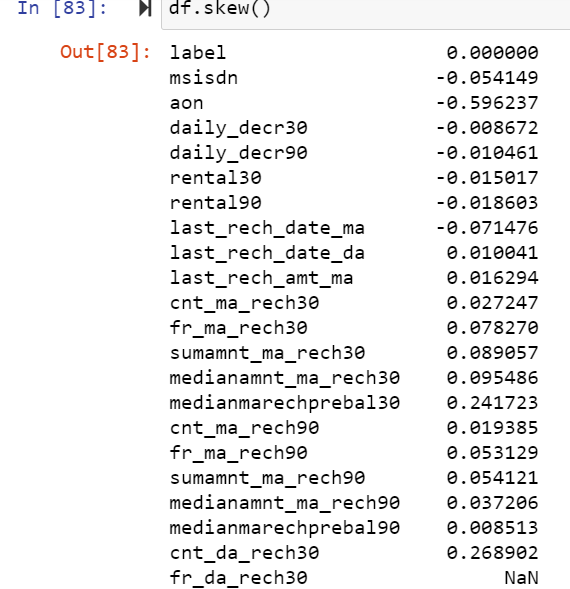
**Checking skewness in dataset:**







To remove skewness use PowerTransformer and log transformer technique in dataset.



After removing skewness in dataset.

**Model/s Development and Evaluation**

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

**Approach**

Importing the required libraries and reading the dataset.

Merging of the two datasets

* Understanding the dataset

1. Exploratory Data Analysis (EDA) –

* Data Visualization

1. Feature Engineering

* Duplicate value removal
* Missing value imputation
* Encoding of categorical variables
* Dropping of redundant feature columns
* Check for the outliners and removal of outliers.

1. Model Building

Performing train test split

* Feature Scaling
* DecisionTree Algorithm
* RandomForest Classification
* SVC
* Gradient Boosting Classification
* AdaBoost Classifier
* Logistic Regression
* GaussianNB Classifier
* k-nearest neighbors

1. Model Validation
2. Hypermeter Tuning (GridSearchCV)

For Random Forest Classification

1. Checking for Feature Importance
2. Creating the final model and making predictions

**Testing of Identified Approaches (Algorithms)**

* Feature Scaling
* DecisionaTree Algorithm
* RandomForest Classification
* SVC
* Gradient Boosting Classification
* AdaBoost Classifier
* Logistic Regression
* GaussianNB Classifier
* k-nearest neighbors

**Oversampling:**

**SMOTE:**

Imbalanced classification involves developing predictive models on classification datasets that have a severe class imbalance.

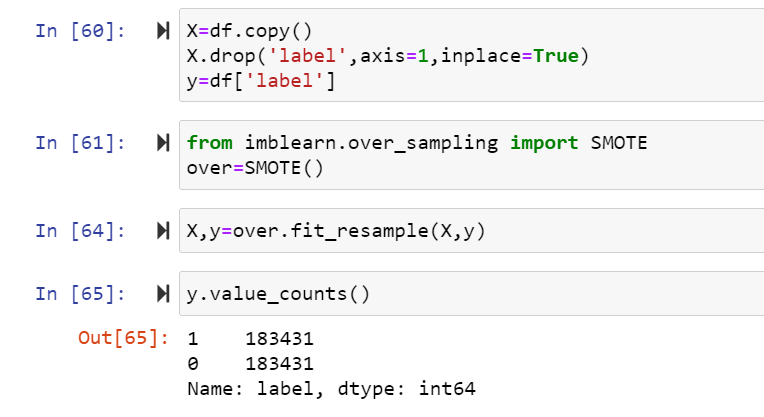
The challenge of working with imbalanced datasets is that most machine learning techniques will ignore, and in turn have poor performance on, the minority class, although typically it is performance on the minority class that is most important.

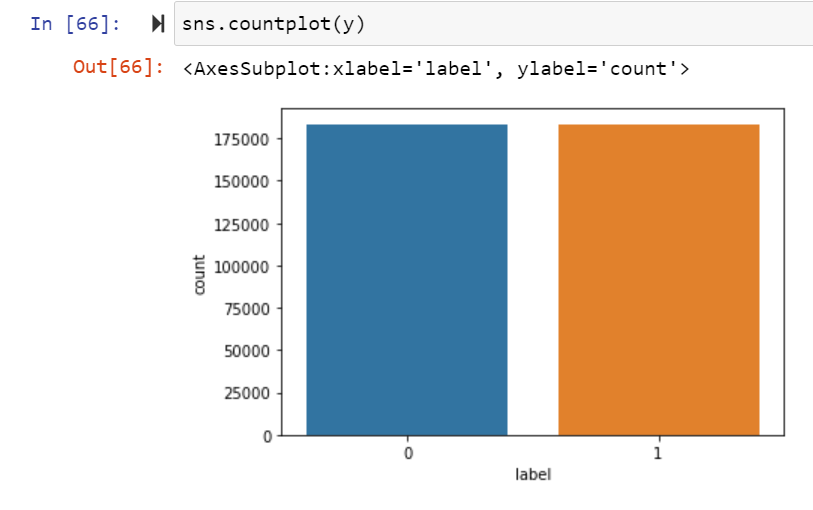
One approach to addressing imbalanced datasets is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don’t add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of [data augmentation](https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/) for the minority class and is referred to as the **Synthetic Minority Oversampling Technique**, or **SMOTE** for short.

In this tutorial, you will discover the SMOTE for oversampling imbalanced classification datasets.

After completing this tutorial, you will know:

* How the SMOTE synthesizes new examples for the minority class.
* How to correctly fit and evaluate machine learning models on SMOTE-transformed training datasets.
* How to use extensions of the SMOTE that generate synthetic examples along the class decision boundary.





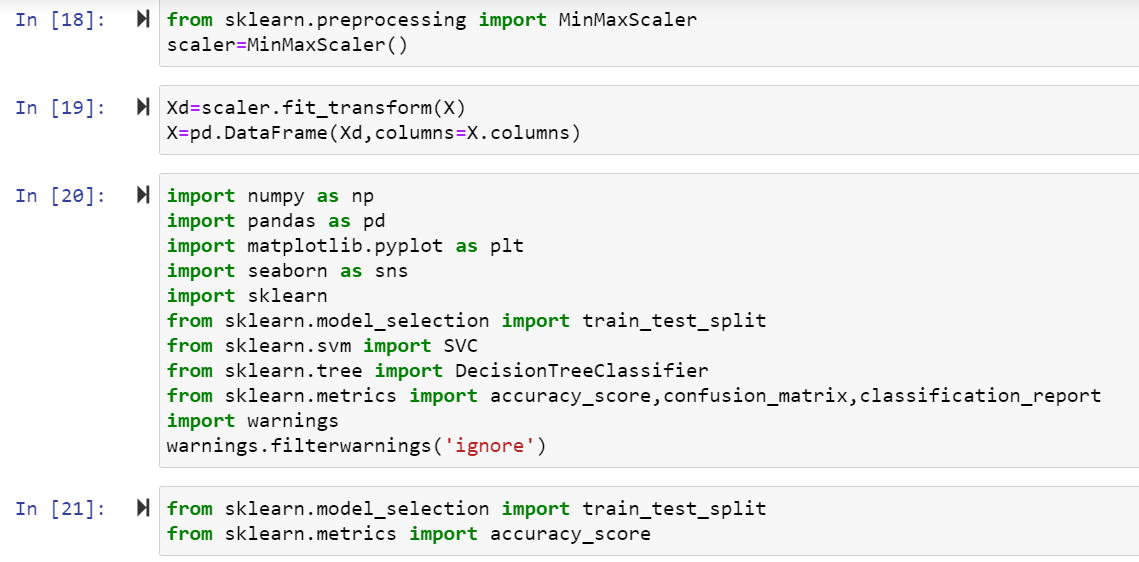
After using Oversampling SMOTE now class are balance.

**MinMax Scaler Technique:**

Here use MinMaxScaler Techique. The MinMax scaler is one of the simplest scalers to understand.

**Building Machine Learning Models:**

Run and Evaluate selected models:



# Confusion Matrix:

# Confusion Matrix There is a need to look at the confusion matrix to assess a classification model’s quality of classification. For an ideal confusion matrix, we expect to get values only on the leading/principal diagonal, since they represent correct classification; values off-diagonal are those that were misclassified. Hence, confusion matrix for each of our ensemble classifiers. values lie along the principal diagonal for all the ensemble classifiers, and the more values we record on the principal diagonal, the more evidence we have of correct classification. However, one thing that is easily noticeable is that most of the misclassifications are recorded between the average and good risk classes for all the classifiers. This is most likely because the decision boundary between the two classes is not so visible; hence, the classifiers cannot easily identify it, leading to some misclassifications between the two classes. However, for each classifier, the hyperparameters were tuned to obtain the best possible model that could quickly and easily identify the decision boundaries for a better classification experience.

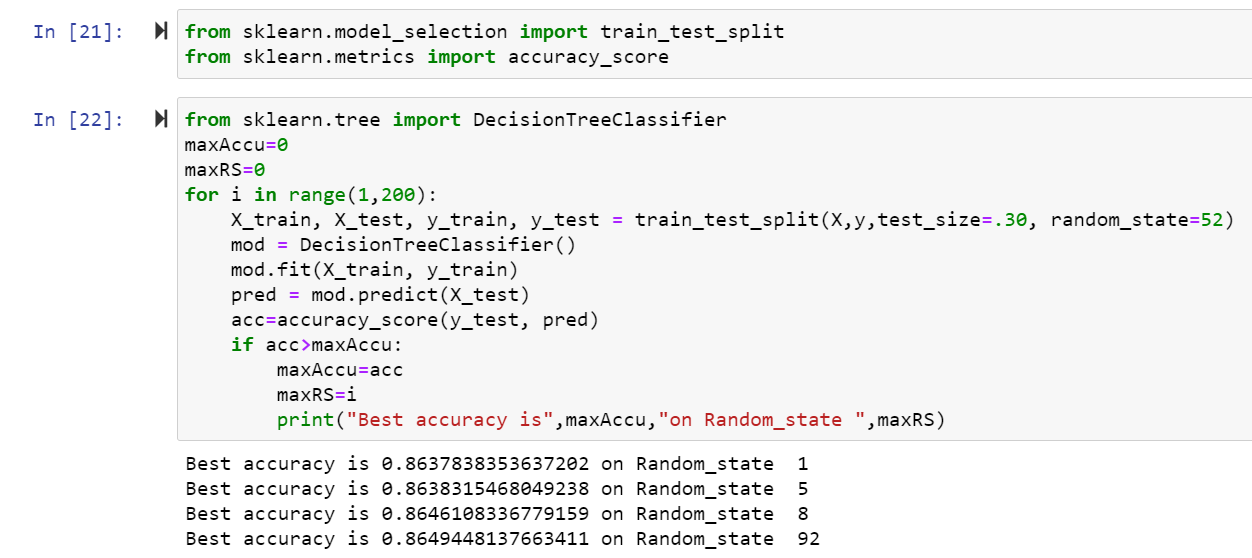
# Decision Tree Classifier:

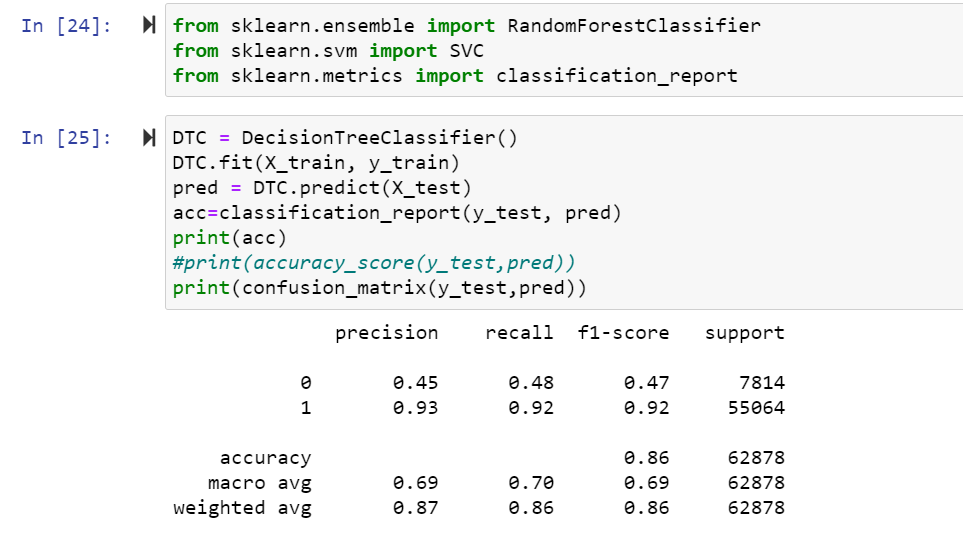
Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcomes** In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset.

**It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.**

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

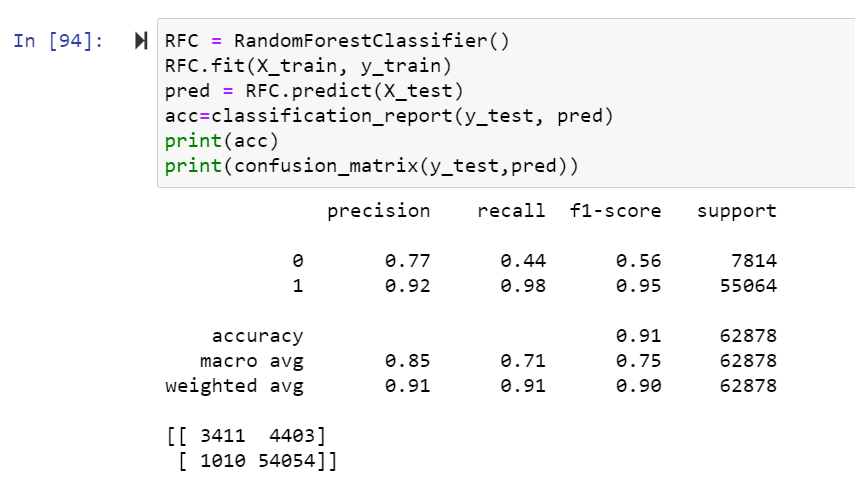




Here we use decision tree classifier model check accuracy score and confusion matrix of above accuracy score is 0.86

**Random Forest Classifier:**

What is a Random Forest Classifier? The term “Random Forest Classifier” refers to **the classification algorithm made up of several decision trees**. The algorithm uses randomness to build each individual tree to promote uncorrelated forests, which then uses the forest's predictive powers to make accurate decisions.

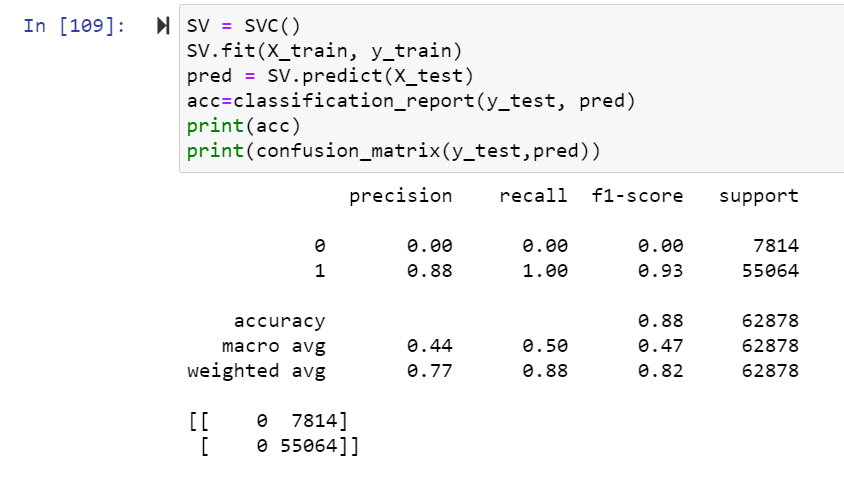


Here we use Random Forest classifier model check accuracy score and confusion matrix of above accuracy score is 0.91.

**Support Vector Classifier:**

The most applicable machine learning algorithm for our problem is [Linear SVC](http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html). Before hopping into Linear SVC with our data, we're going to show a very simple example that should help solidify your understanding of working with Linear SVC.

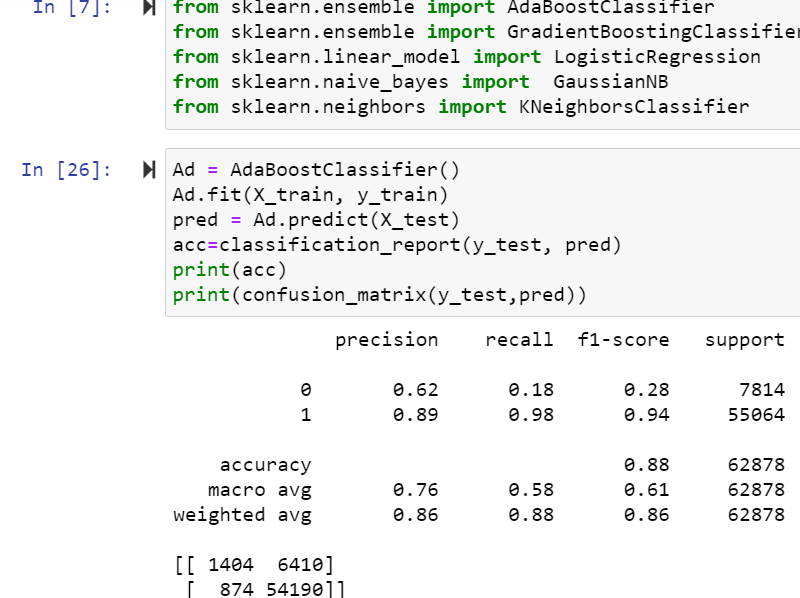
The objective of a Linear SVC (Support Vector Classifier) is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the "predicted" class is. This makes this specific algorithm rather suitable for our uses, though you can use this for many situations.



Here we use Support Vector Classifier model check accuracy score and confusion matrix of above accuracy score is 0.88.

**AdaBoost Classifier:**

AdaBoost (Adaptive Boosting) is a very popular boosting technique that aims at combining **multiple weak classifiers to build one strong classifier**. ... The classifier mentioned here could be any of your basic classifiers, from Decision Trees (often the default) to Logistic Regression, etc.



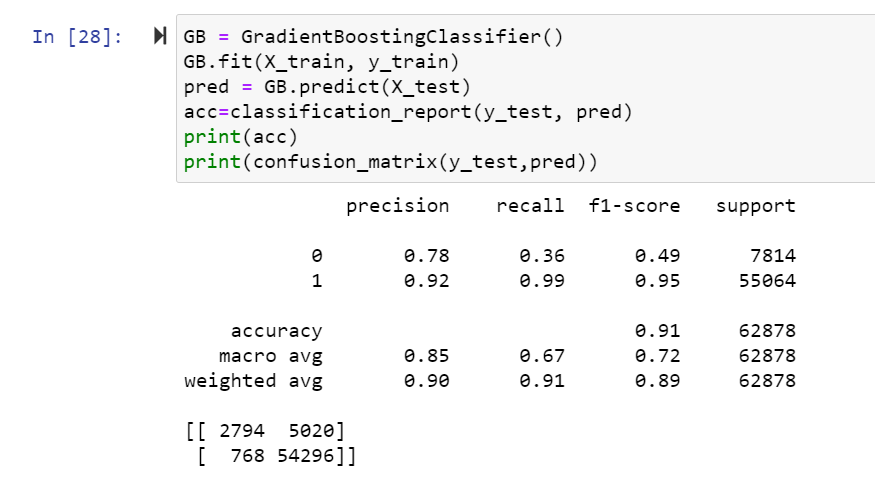
Here we use AdaBoost Classifier model check accuracy score and confusion matrix of above accuracy score is 0.88.

[**Gradient boosting classifiers**](https://en.wikipedia.org/wiki/Gradient_boosting)**:**

[Gradient boosting classifiers](https://en.wikipedia.org/wiki/Gradient_boosting) are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets, and have recently been used to win many [Kaggle](https://www.kaggle.com/" \t "_blank) data science competitions.

The Python machine learning library, [Scikit-Learn](https://scikit-learn.org/stable/" \t "_blank), supports different implementations of gradient boosting classifiers, including [XGBoost](https://xgboost.readthedocs.io/en/latest/" \t "_blank).

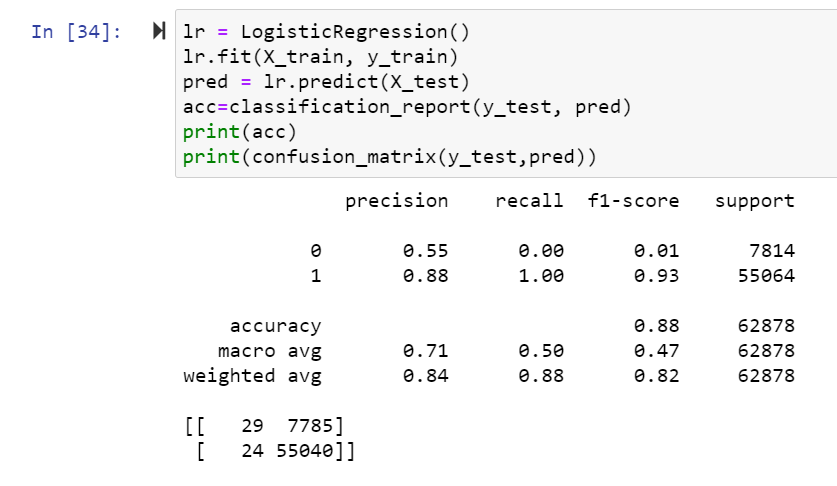
In this article we'll go over the theory behind gradient boosting models/classifiers, and look at two different ways of carrying out classification with gradient boosting classifiers in Scikit-Learn.



Here we use [Gradient boosting classifiers](https://en.wikipedia.org/wiki/Gradient_boosting) model check accuracy score and confusion matrix of above accuracy score is 0.91.

**Logistic regression:**

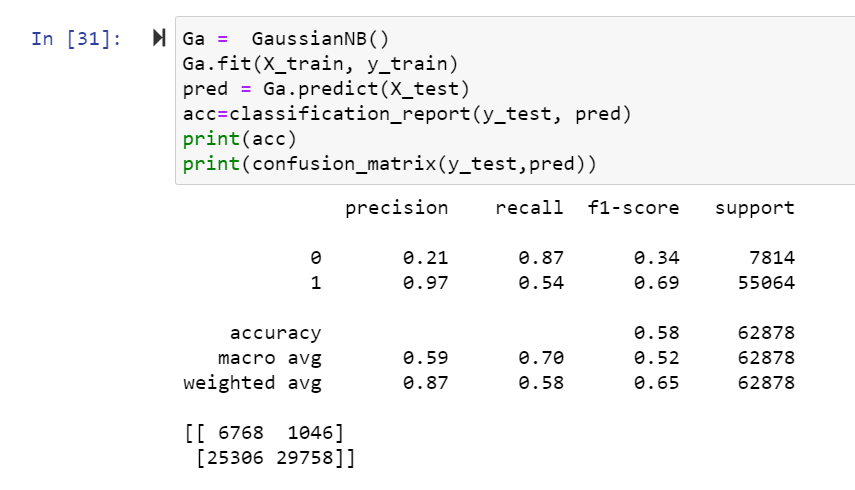
Logistic regression is a statistical analysis method used to predict a data value based on prior observations of a data set. ... Based on historical data about earlier outcomes involving the same input criteria, it then scores new cases on their probability of falling into a particular outcome category.



Here we use Logistic regression model check accuracy score and confusion matrix of above accuracy score is 0.88.

**GaussianNB Classifier:**

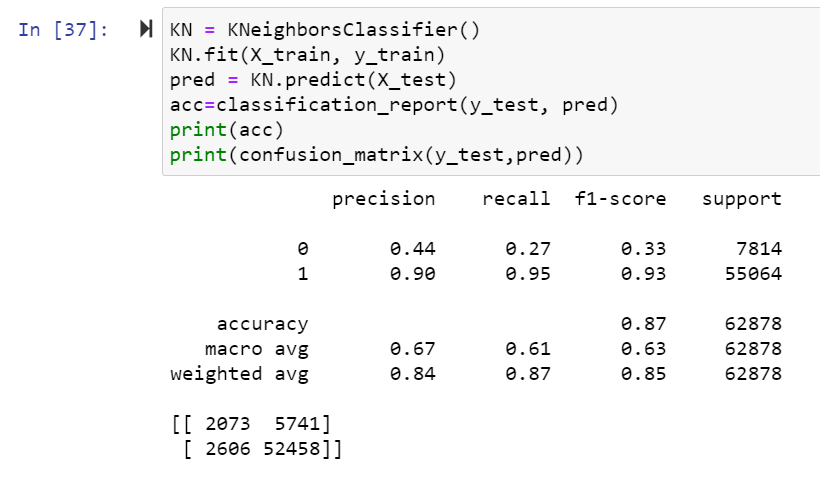
One extremely fast way to create a simple model is to assume that the data is described by a Gaussian distribution with no covariance between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all you need to define such a distribution.



Here we use GaussianNB model check accuracy score and confusion matrix of above accuracy score is 0.58

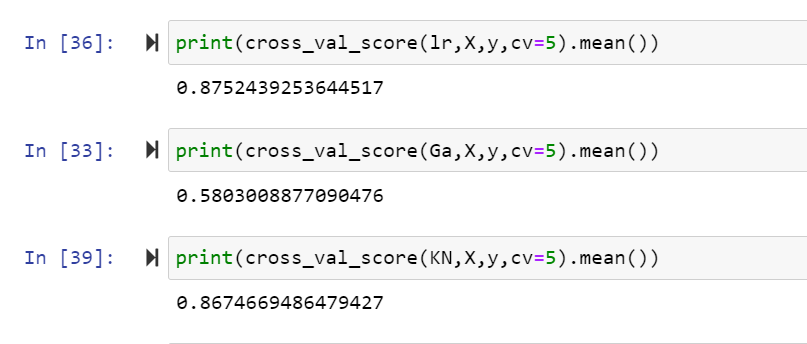
**k-nearest neighbors:**

The k-nearest neighbors (KNN) algorithm is a **simple, supervised machine learning algorithm** that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.



Here we k-nearest neighbors (KNN) algorithm model check accuracy score and confusion matrix of above accuracy score is 0.87.





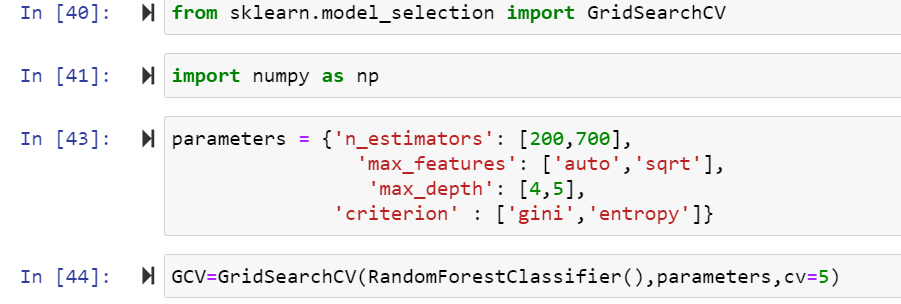
Here we check accuracy score and after that check cross validation score of all the model the Random forest classifier is the best model because accuracy score and cross validation score is 1.0 So apply hyper parameter tuning on it.

**GridSearchCv**:

GridSearchCV is **a library function that** is a member of sklearn's model\_selection package. It helps to loop through predefined hyperparameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyperparameters.

**Parameterlist:**

There is a list of different machine learning models. They all are different in some way or the other, but what makes them different is nothing but input parameters for the model. These input parameters are named as **Hyperparameters.**These hyperparameters will define the architecture of the model, and the best part about these is that you get a choice to select these for your model. Of course, you must select from a specific list of hyperparameters for a given model as it varies from model to model.



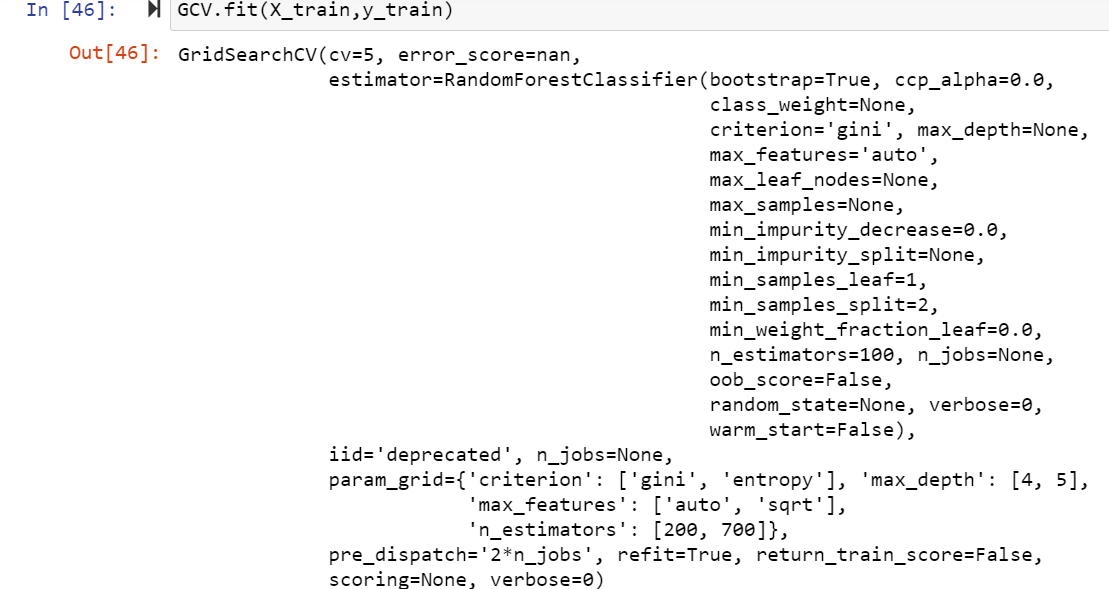
Here above are the parameter list of RandomForestClassifier model .

The above code block we have the following parameters:.  
max\_features: In this maximum features there are two values auto and auto, sqrt.

Criterion: In criterion there are two parameter values are present gini and entropy

n\_estimator is 200 ,700

and in max\_depth 4,5.





Here after use different parameter list the best parameter list is select.

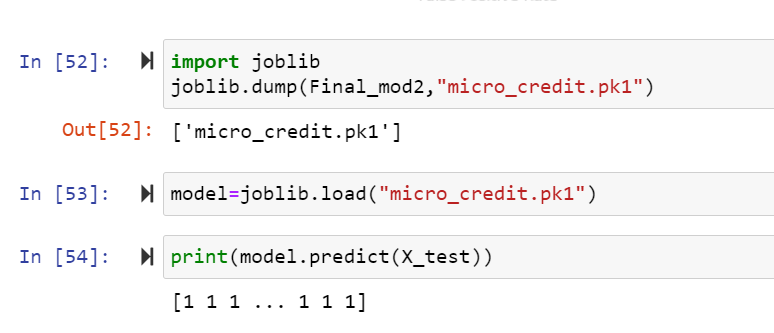
Above are the best parameter list RandomForestClassifier. Put this parameters into the model so output is finally best score is RandomForestClassifier is 90 so it is the best score.

**AUC and ROC Curve:**

In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (**Area Under The Curve**) ROC (**Receiver Operating Characteristics**) curve. It is one of the most important evaluation metrics for checking any classification model’s performance. It is also written as AUROC (**Area Under the** **Receiver Operating Characteristics**)



Finally Load the model and predict the values.

****

Finally Load the model and predict the values.

# Methodology

1. Data Exploration and Cleaning On data exploration, I found that the dataset was imbalanced for the target feature(87.5% for Non-defaulters and 12.5% for Defaulters). Also, I found that the data had some very unrealitic values such as 999860 days which is not possible. Also, there were negative values for variables which must not have one (example:frequency,amount of recharge etc). All these unrealistic values were dropped which caused a data loss of 8% only.
2. Feature Selection Since there were 36 features, many of which I suspected were redundant because of the data duplication. It was imperative to select only most significant of them to make ML models more efficient and cost effective.
3. Data Visualization On visualizing data, there were two important insights I gathered. a. Imbalance of data b. Distribution was not normal
4. Data Normalization Since the data was not normal, I normalized all the features except the target variable which was dichotomous(Values '1' and '0').
5. Oversampling of Minority class Since the data was expensive, I did not want to lose out on data by undersampling the majority class. Instead, I decided to oversample the minority class using SMOTE.
6. Build Models Since it was a supervised classification problem, I built 5 models to evaluate performance of each of them: a. Logistic Regression b. Linear SVM c. Decision Tree d. Random forest e. Gradient Boost Classifier Since the data was imbalanced, accuracy was not the correct performance metric. Instead I focused on other metrics like precision, recall and ROC-AUC curve.

**Visualizations:**

For visualization purpose use heatmap for dataset: data visualization is a method of graphically representing numerical data where the value of each data point is indicated using colors. ... More importantly, heatmaps help to classify the sections that are performing sub-par and need optimization

**Interpretation of the Results:**

House price prediction project there are two dataset Train and Test dataset in Train dataset the SalePrice are the target variable was present and there was continues values are present so it regression problem need to use regression algorithm.

After that check the shapes of the dataset and check null values are present that that dataset and remove the null values .

Check the data types of the dataset and need to convert categorical columns into numeric.

And also the use zscore and IQR to check the data loss.

And check the skewness and remove the skewness using power transform and log transform method.

Use scaling technique.

And use different Regression models and depend on r2 score and cross validation score select Random forest repressor is the best model and apply hyper parameter tuning on it.

After that load that model and predict the values.

**CONCLUSION**

**Conclusion:**

According to the performance metrics, Random Forrest scores highest in accuracy. Also, the curve is tending towards the ideal shape. Hence, Random Forrest looks like the best fit for this data.

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

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

In this project drop some columns because not much important.

There are no null values was present in dataset.

Use label encoder to convert categorical values into numeric.

In this using zscore and IQR technique more data loss.

Remove the skewness is difficult because dataset size is big and so much skewness is present so time is require to remove the skew ness.

Use different Classification modelling technique and using accuracy score and cross validation score select random forest is the best model and apply hyper parameter tuning on it and load the model predict the values.

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

**Limitation**:

Low- and middle-income countries shows that giving small loans in the form of microcredit did not lead to transformative impacts on income or long-term consumption on average, but it did help households better manage financial choices.

**Future work:**

## We suggest that working on ****How can the microcredit system be reformed to have greater benefits for borrowers?****

And for better result We have taken only classification applying different classification algorithm to improve the performance of model.