

**MICRO CREDIT DEFAULTER**

submitted by: - Ankit soran

ACKNOWLEDGEMENT

The internship opportunity, I have with FlipRobo Technologies is a great chance for learning and professional development. I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills acknowledge in the best possible way.

I would like to extend my appreciation and thanks for the mentors from Data Trained and professionals from FlipRobo Technologies who had extended their help and support.

References:

https://sklearn.org/supervised\_learning.html#supervised-learning

<https://www.datacamp.com/community>

<https://github.com/mxc19912008/Andrew-Ng-Machine-Learning-Notes>

https://www.analyticsvidhya.com/blog/category/machine-learning/

**MICRO CREDIT DEFAULTER**



**Problem Statement:**

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).

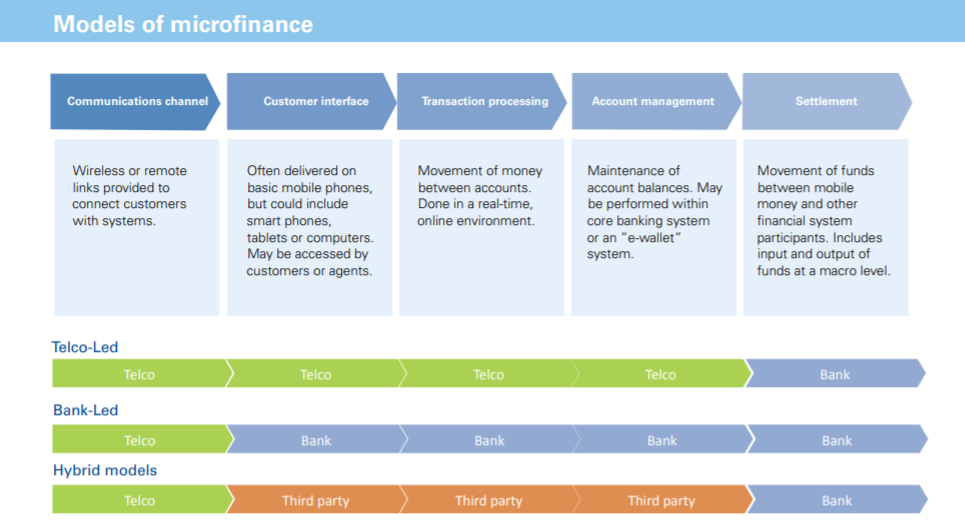
The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

**Exercise:**

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 paid i.e. Non- defaulter, while, Label ‘0’ indicates that the loan has not been paid i.e. defaulter.

**Points to Remember:**

* There are no null values in the dataset.
* There may be some customers with no loan history.
* The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records.
* For some features, there may be values which might not be realistic. You may have to observe them and treat them with a suitable explanation.
* You might come across outliers in some features which you need to handle as per your understanding. Keep in mind that data is expensive and we cannot lose more than 7-8% of the data.



Based on data provided from our client database, customer’s repayment of loan is assessed based on different factors. By building the model, we can assess which customers are highly likely to repay the loan, thereby it will be useful for those needy people who will repay the loan and also prevent the loss to the customer by avoiding loans to the defaulters.

**ANALYTICAL PROBLEM FRAMING**

**MATHEMATICAL MODELLING OF PROBLEM:**

**Mathematical modeling is simply the method of implementing statistical analysis to a dataset where a Statistical Model is a mathematical representation of observed data.**

While analyzing the data, there are an array of statistical models we can choose to utilize.

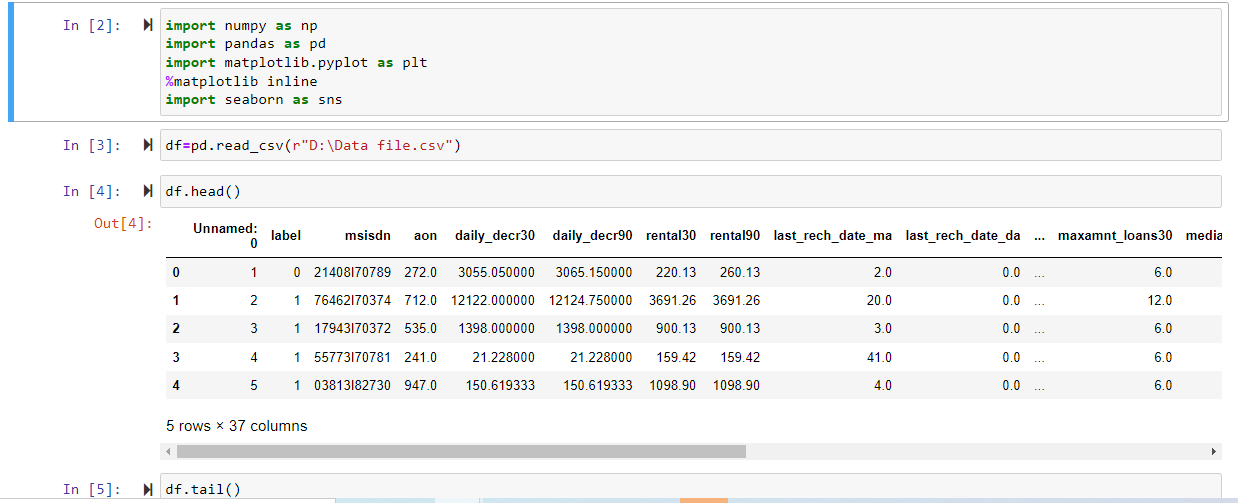
For the given project, we need to predict whether the customer is a defaulter or not.

This is a classification problem. There are wide varieties of classification models like decision trees, random forests, nearest neighbor, Logistic Regression.

**DATA SOURCE AND FORMAT:**

The data has been provided by client in a comma separated Values (.csv) format.

The data will be loaded into Pandas Data frame.

After doing basic exploratory data analysis (EDA) with the above Data Frame like:

* Checking Info
* Checking shape (209593, 36)
* Identifying features.
* Removing unwanted columns.

**Moving forward with Data frame we observed that most of the features value are either integer of float type very few are object type separating them into different data frame.**

**DATA PRE PROCESSING:**

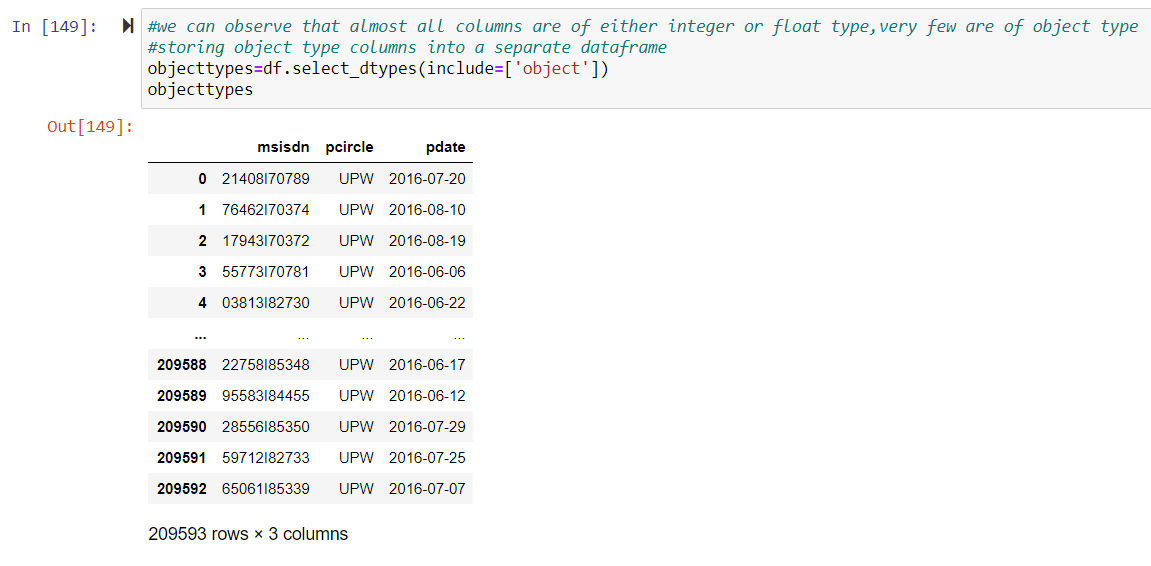
Data preprocessing is a technique of converting raw data into useful format.

Data cleaning is a part of preprocessing technique which involves filling missing values.

For the given dataset it has been mentioned that there are no null values.

Firstly, I dealt with object type columns.

Checking what columns are of object type and what type of data is stored in them.



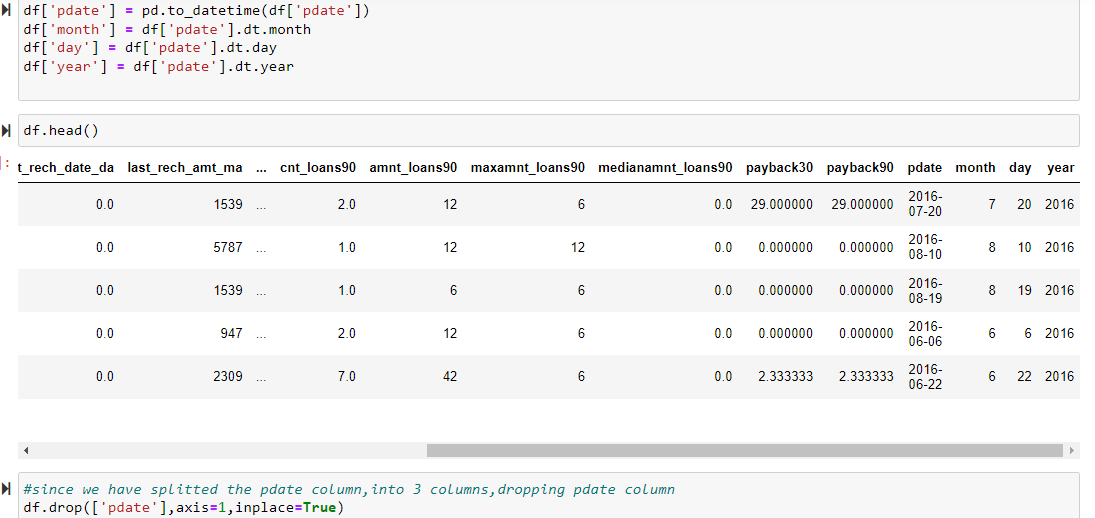
After separating OBSERVATION:

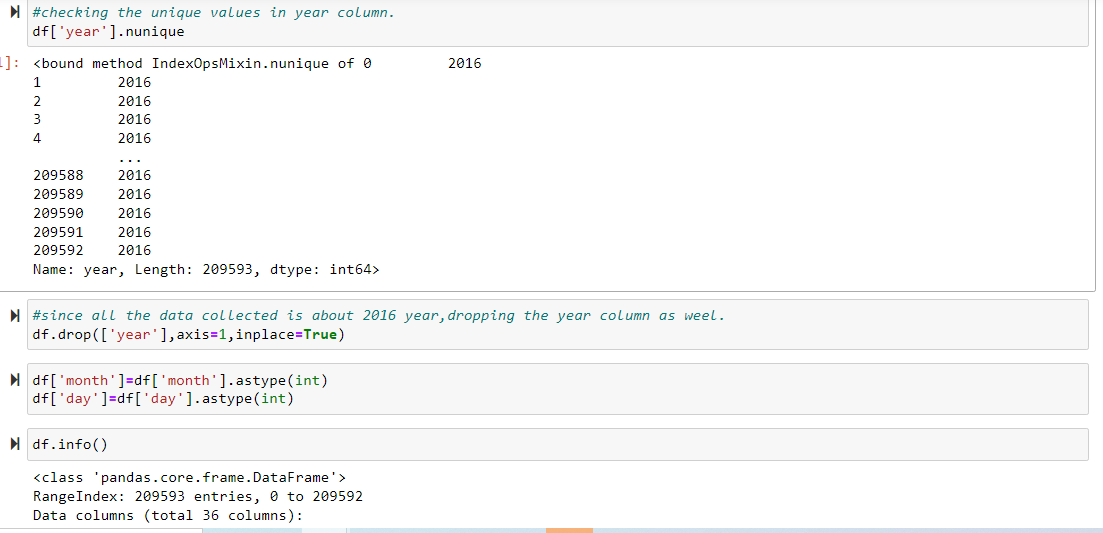
* It has numeric entries, so converting this column into integer type.
* Checking the pcircle msisdn entries, seems all the rows has same telecom provider name.

If all the entries are same in pcircle column, will be deleting the column.

Will be converting the pdate column into like year, month and date.

As the date column is splitted into 3 columns,will be deleting the pdate column.





Also, the data gathered belongs to 2016 year, hence it won’t be impacting the output due to same entry in all the columns. So dropped the year column.

Later converted the month and date columns to integer.

Checking whether all the columns are of integer type. After checking all the columns are integer type except msisdn which is object type.

Describe method is used to view some basic statistical details like percentile, mean, standard deviation etc. of a data frame or a series of numeric values.



We can note that there is a huge difference in 75% value and max value for most of the columns which indicate presence of outliers.

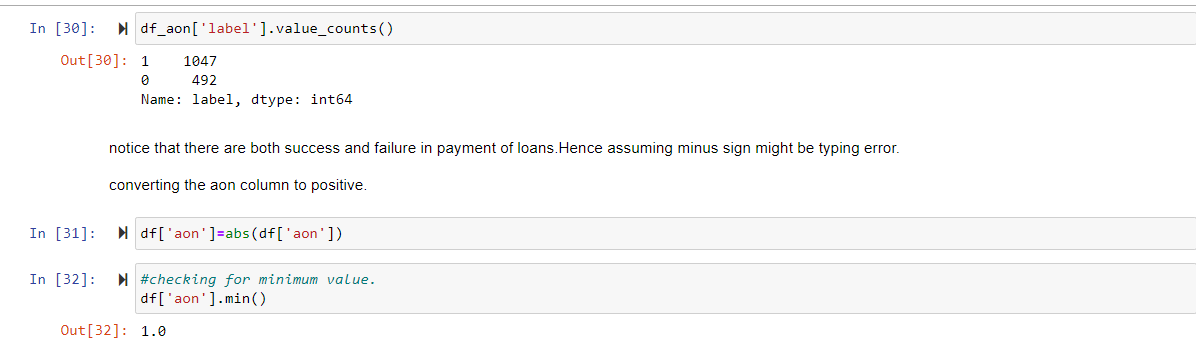
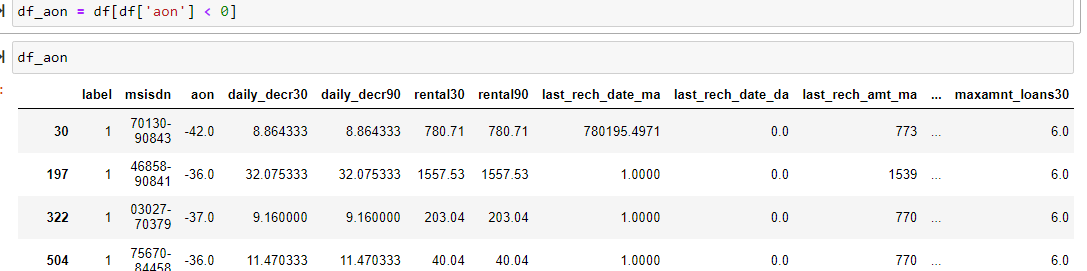
Observations:

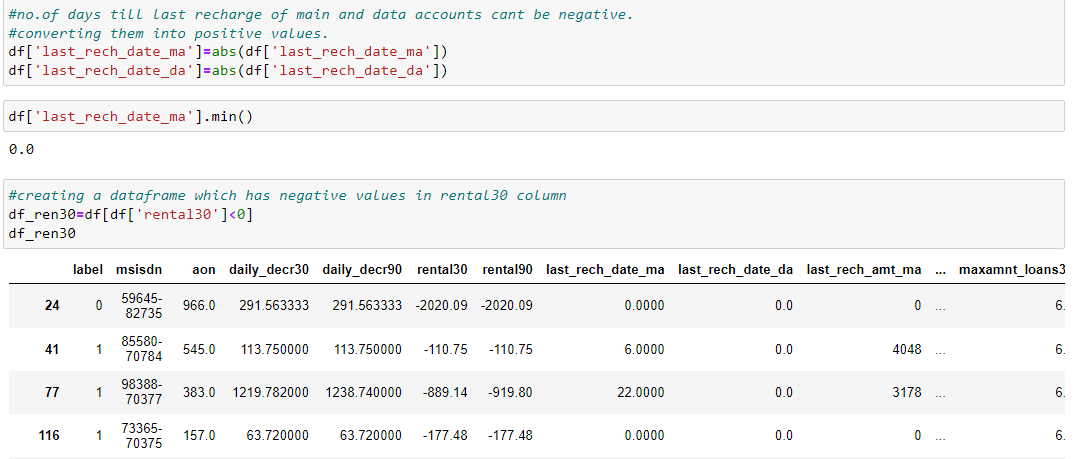
COLUMNS WITH NEGATIVE MINIMUM VALUES:

* Aon
* daily\_decr30=>Daily amount spent from main account, averaged over last 30 days (in Indonesian Rupiah)
* daily\_decr90=>Daily amount spent from main account, averaged over last 90 days (in Indonesian Rupiah)
* rental30=>Average main account balance over last 30 days
* rental90=>Average main account balance over last 90 days
* last\_rech\_date\_ma=>Number of days till last recharge of main account
* last\_rech\_date\_da=>Number of days till last recharge of data account

Aon

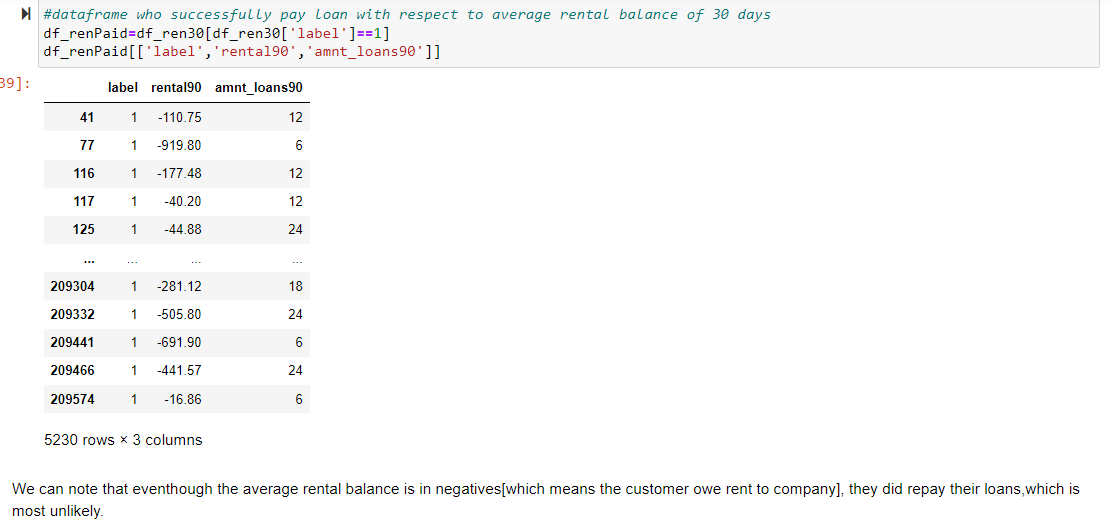
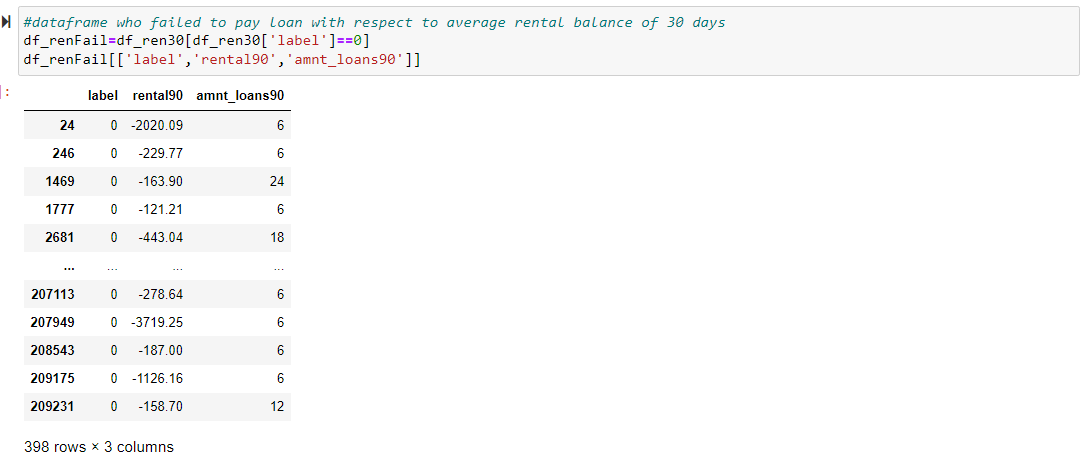
* =>indicates Age on Cellular Network in days.
* =>minimum value is -48, usually minimum value should be zero.
* So checking the other columns values where Aon has negative values.
* Converting the Aon column to positive.
* last\_rech\_date\_ma, last\_rech\_date\_da: these two columns indicate number of days till last recharge of main and data accounts. This count of days also can’t be negative. Converting them to positive.



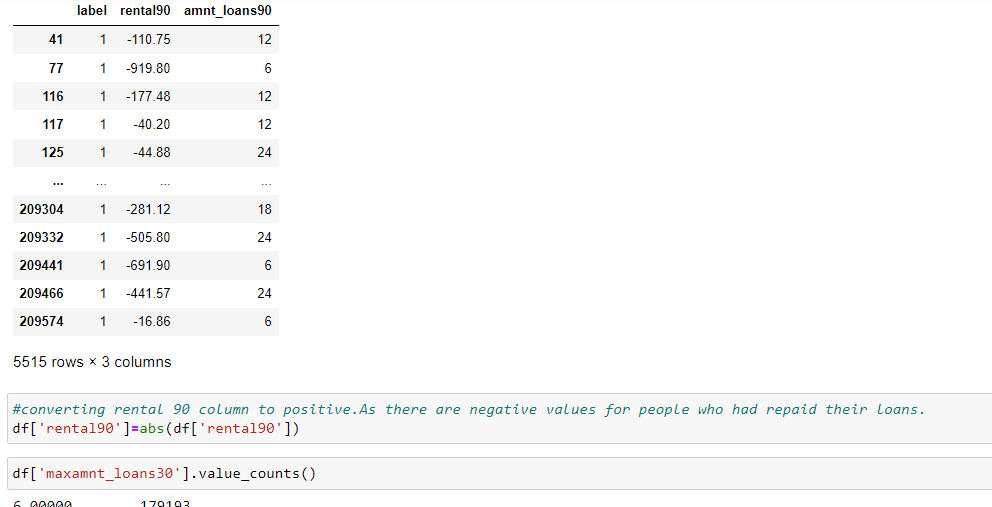
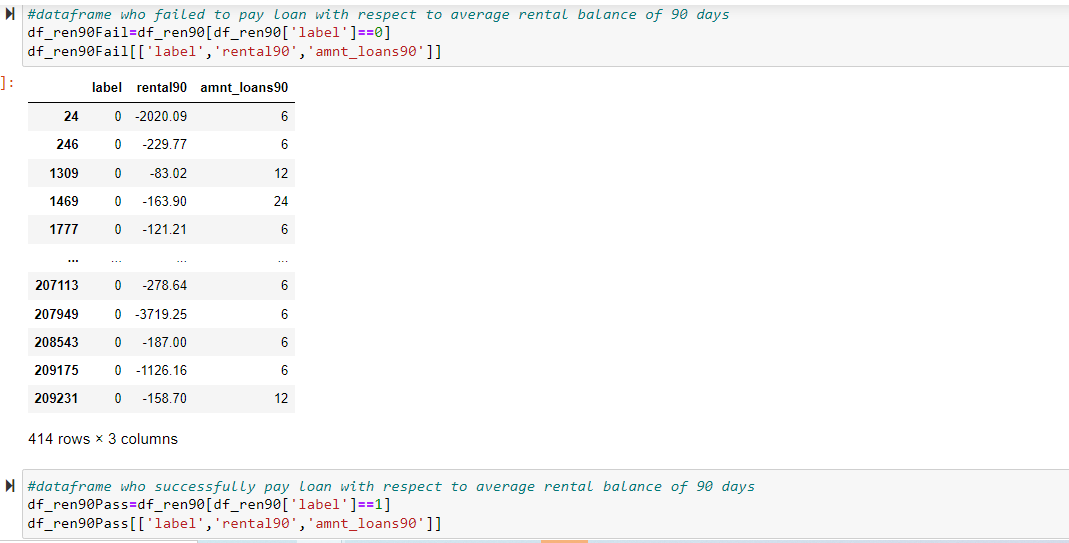


Moving Forward I created two different data frames in respect to negative values in rental 30 column.

* One being the people who failed to repay the loan.
* Other being the people who did repay the loan.



There might be other possibility that user will not be granted loan if they have negative balance. This might be due to erroneous entry. So converting them to positive.



Further Processing,

Checking the entries of maxamnt\_loans30 column.

It has been mentioned that this columns values has to be either 6 or 12.we can notice that there are huge numbers of entries other than 6, 12. Ignoring 0 because there might be users who hasn’t taken loans. Converting the other numbers to zero because there is no probability of loan repay amount other than 6 ad 12.There are 1047 rows that has values other than 6,12 and0.

There are 1047 records of values that are other than 6, 12 and 0.

Converting these 1047 records to zero because we can’t predict their repayment amount.

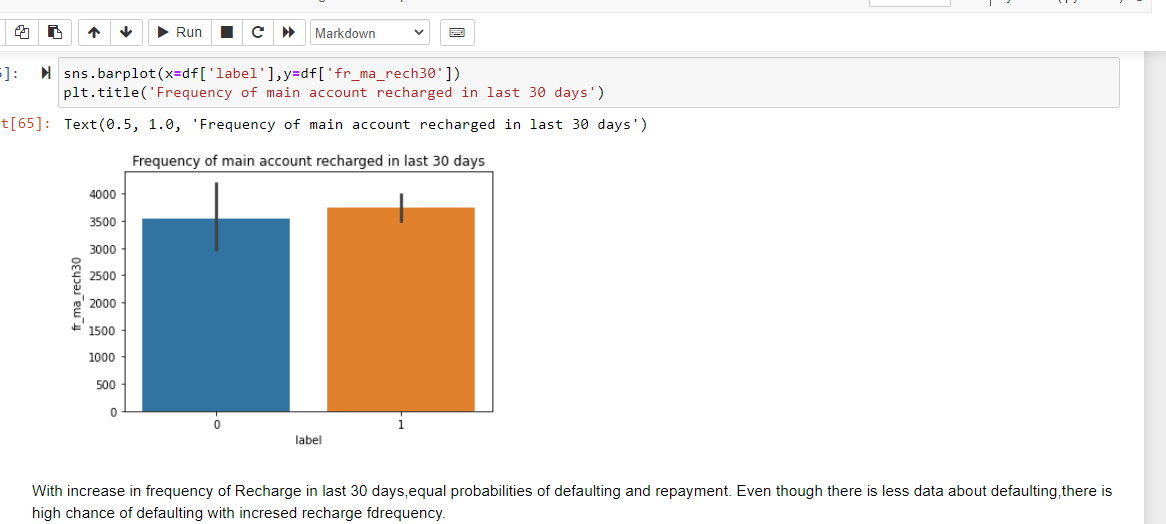
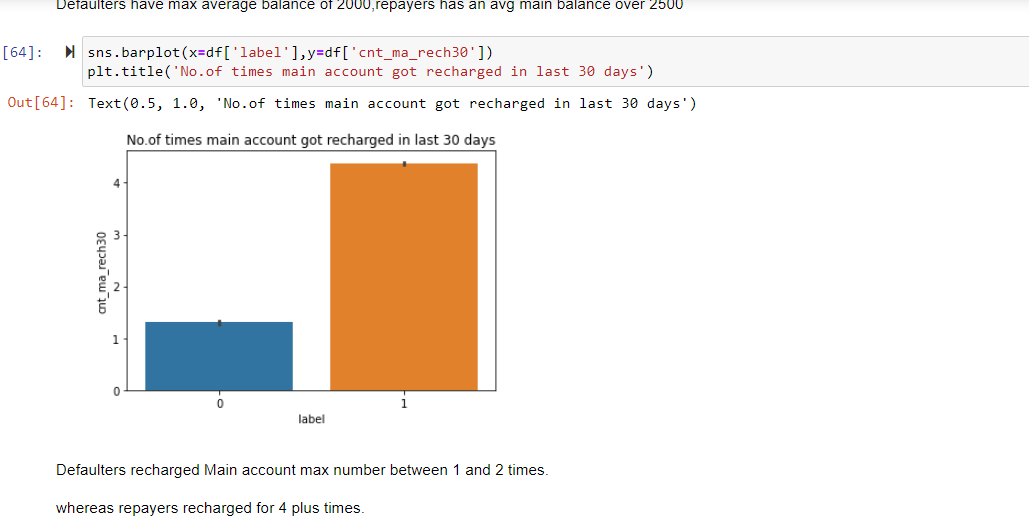
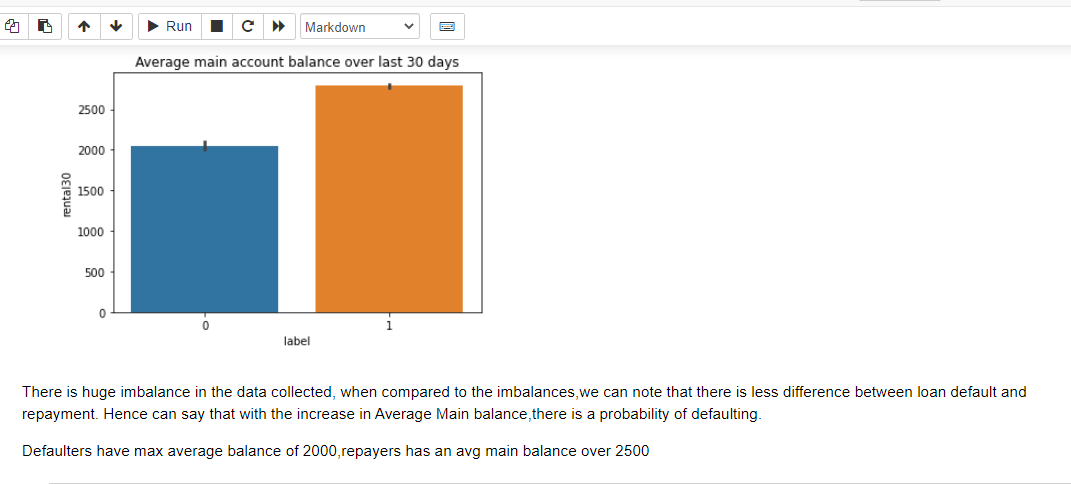
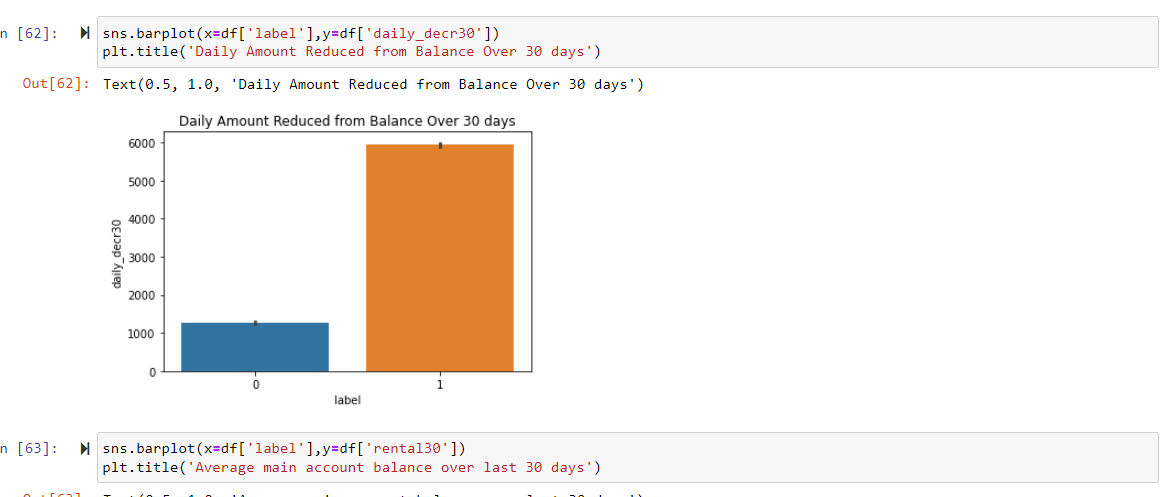
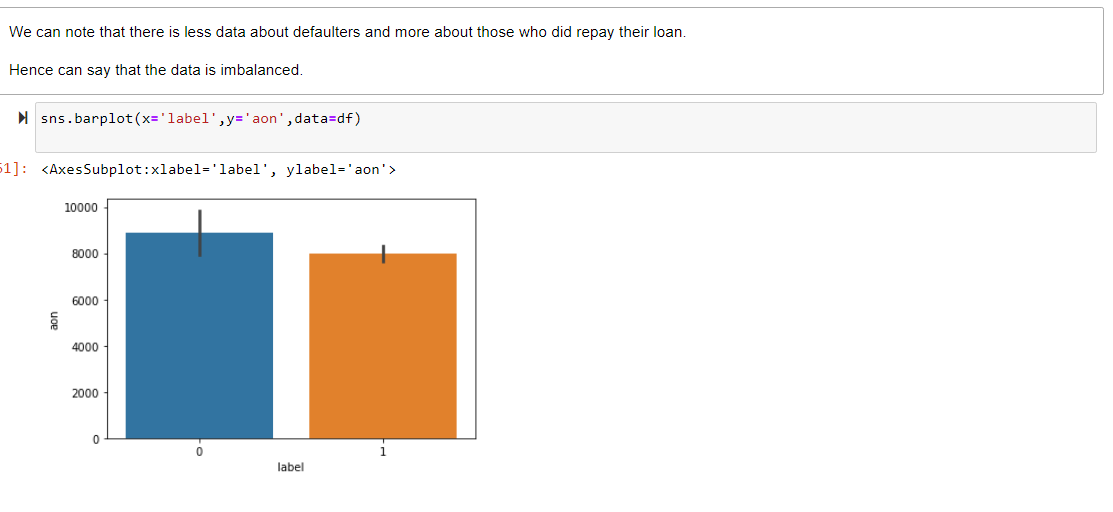
Checking the users who haven’t taken any loan.

Amt\_loans90 column describes the total amount of loans taken by the user in span of 90 days. The presence of zero in this column indicates that the user hasn’t taken any loans

There are 2043 rows in the data frame with zero in amt\_loans90 column. Dropping the rows which has zero in the amt\_loans 90 column because such rows won’t be useful in predicting the loan repayment.

msisdn is nothing but phone number of the user, it has nothing to do with the predictions of loan payment. Hence dropping it.

**Data Visualizations:**

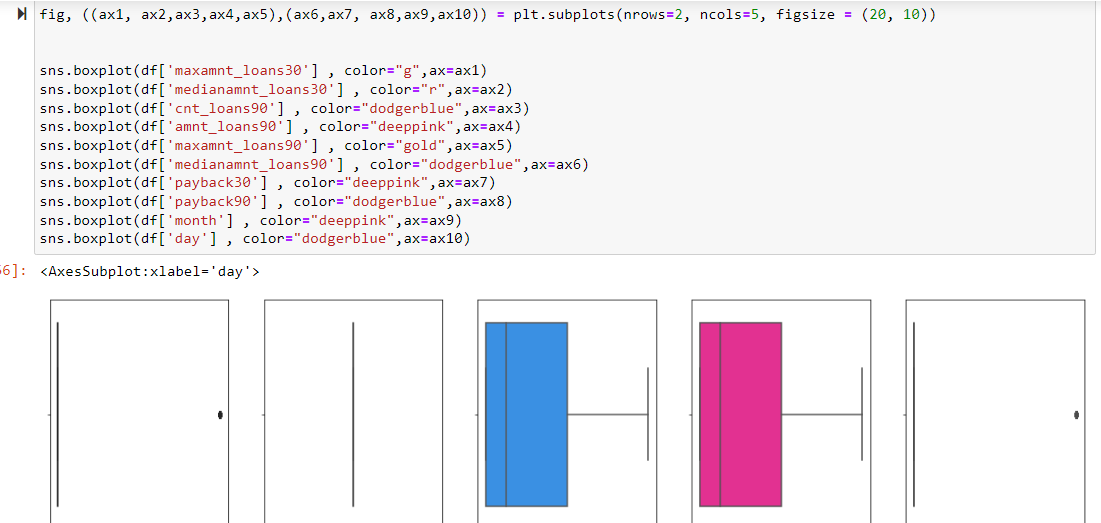
****

After visualizing data with the help of different plots we came across few observations.

* We can note that there is less data about defaulters and more about those who did repay their loan.
* Hence can say that the data is imbalanced.
* With increase in Age on Network, defaulting rate is higher
* There is huge imbalance in the data collected,
* When compared to the imbalances, we can note that there is less difference between loan default and repayment. Hence can say that with the increase in Average Main balance, there is a probability of defaulting.
* Defaulters have max average balance of 2000,repayers has an average main balance over 2500
* Defaulters recharged Main account max number between 1 and 2 times.
* Whereas re-payers recharged for 4 plus times.
* With increase in frequency of Recharge in last 30 days, equal probabilities of defaulting and repayment. Even though there is less data about defaulting, there is high chance of defaulting with increased recharge frequency.
* The defaulters has max limit ranging between 2000 and 3000 of Total recharge amount.
* On an average the defaulters has recharged for a max of 1000 Indonesian rupiah.
* Defaulters has a median Main account recharge amount ranging between 4000 and 5000.
* With increase in Median of Main balance recharge, probability of defaulting is very high.
* Defaulters has recharged the data account for a maximum of 200 to 250 times.
* With increase in No. of times data accounts recharge, probability of defaulting is high.
* With increase in frequency of recharging data account, higher chances of defaulting.
* Defaulters has taken between 1 to 1.5 no. of loans.
* As practically there will be no 1.5 loan, considering only 1 loan.
* Those who repaid had taken max of 3 loans.
* Total Amount of loans took by Defaulters varies between (7.5 to 10).
* Re-payers has took 20 loans which tends to be the max limit.
* A user can take Maximum of 7 loans in 30 days.
* Defaulters took 6 loans whereas re-payers took 7 loans.
* Can say that there not a much difference in loans took by both defaulters and Re-payers.
* Median of loan Amount by defaulters is 0.03 whereas Re-payers took max of median amount which is 0.06
* A potential defaulter might repay in 2 days.
* Re-payers took average of 3.5 days.
* The defaulters has spent a little above 1000
* Re-payers has spent 7000.
* Defaulters average=2000 to 2500.
* Re-payers average= 3500.
* Defaulters recharged for 2 times.
* Re-payers recharged for 7 times.
* Defaulter’s frequency is 5.
* Re-payers frequency is 8.
* Defaulters recharge amount varies from 2000 to 4000.
* Re-payers recharged for 14000.
* Defaulter’s median-1250.
* Re-payers median -2000.
* In terms of No. of times data account recharged, both defaulters and re-payers has approximately equal no. of recharges.
* With increase in frequency of Data account recharge in 90 days, defaulting rate is high.
* Defaulters took max of 15 loans
* Re-payers took max of 20 loans
* Defaulters took max of 10 loans.
* Re-payers took 25 max of loans.
* In terms of maximum loan amount, both defaulters and re-payers max limit is approximately same.
* Defaulters took max. median of 0.03 loan amount
* Re-payers took max. median of 0.05 loan amount
* Average pay back time of potential defaulter in 90 days is 3 days.
* Average pay back time of re-payer in 90 days is greater than 4 days.

CHECKING FOR OUTLIERS:





**Models Development:**

The data set is imbalanced since it has large no. of records which contains data about those repaid the loan and less no. of records of those who defaulted loan.

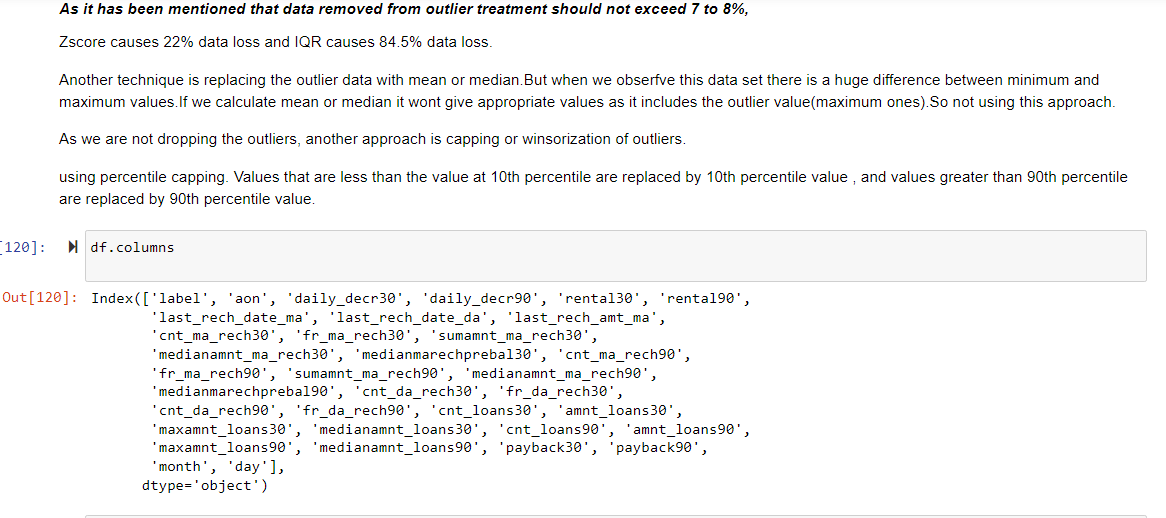
This might result in biased predictions. So, used imblearn library to reduce the imbalances. The imblearn library provides different approaches one is Random under sampling. In context of this problem, Random under Sampling reduces the number of records of those who paid the loan. To be precise, random under sampling deletes data from the majority class such that there will be equal number .of samples of both the classes. Hence reduces the bias.

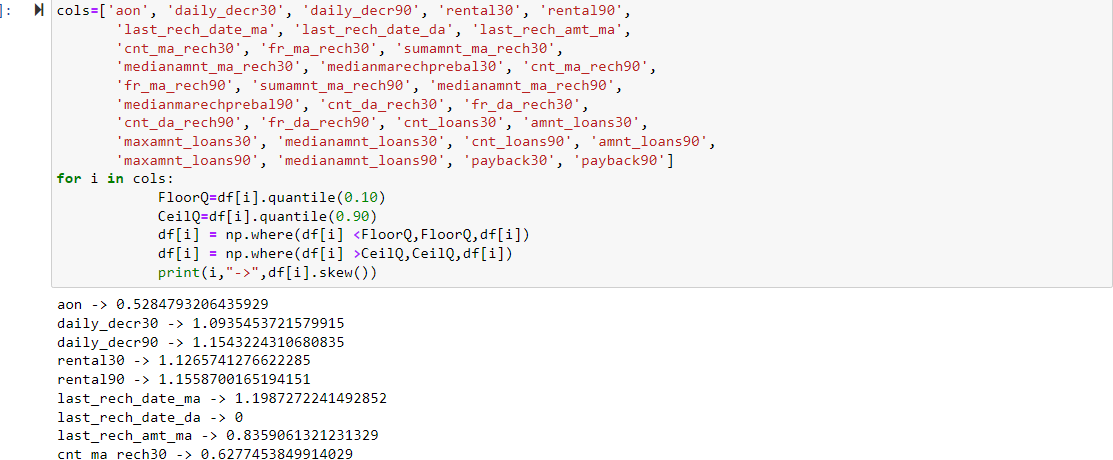
**Statistical methods used:**

Outlier removal: Mostly outliers are removed by either Z score or IQR (Inter Quartile Range).Tried both these approaches first but, the data loss is high in both these approaches. It has been mentioned in guidelines that the data loss should not exceed 7%.So applied capping technique which is also called as Winsorizing.









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

List of algorithms used:

* Logistic Regression
* Decision Tree Classifier
* KNeighborsClassifier
* RandomForestClassifier
* AdaboostClassifier
* BaggingC;assifier
* GradientBoostingClassifier

**Run and Evaluate selected models:**

Cross-validation is used to test the model's ability to predict new data that was not used in estimating it. Cross validation used in scenarios where we need to avoid over fitting.



**After using all the ML algorithms we figured. That Adaboost Classifier has given best results**.

Choosing Adaboost classifier because there both train and test accuracies are same. Rest of the two models there is huge difference between train and test accuracies so not considering them



Key Metrics for success in solving problem under consideration:

A confusion matrix helps us gain an insight into how correct our predictions were and how they hold up against the actual values.

The following metrics are used:

1) Accuracy:  Accuracy is the ratio of the total number of correct predictions and the total number of predictions.

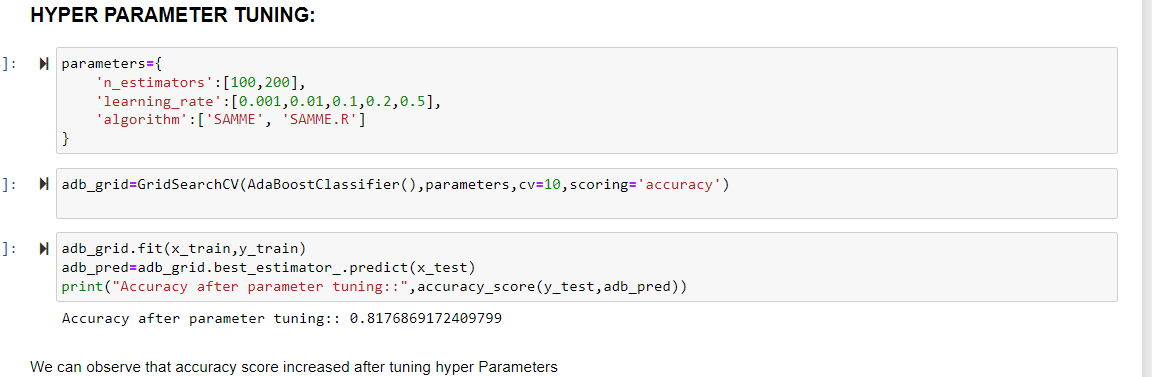
2) Precision: Precision is the ratio between the True Positives and all the Positives

3) Recall: The recall is the measure of our model correctly identifying True Positives

4) F1 score: F1 Score is needed when you want to seek a balance between Precision and Recall.

HYPER PARAMETER TUNING:

Hyper parameter tuning is used to increase the performance of the algorithm.



**CONCLUSION**

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

The defaulting rate is higher in old customers. Defaulters recharge for the main account less Number of times but does recharge for data account more no. of times.

Re-payers recharge the main account more Number of times when compared to defaulters.

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

One of the challenge I faced while data cleaning is outlier removal, in most of the scenarios Z-score will be used as outlier removal technique since it performs quite well with less data loss. In our data set, Z-score has caused 22% data loss. Then I tried another famous technique called Interquartile Range it caused around 80% data loss.

Another technique is replacing the outlier data with mean or median. But when we observe this data set there is a huge difference between minimum and maximum values. If we calculate mean or median it won’t give appropriate values as it includes the outlier value (maximum ones).So not using this approach.

As we are not dropping the outliers, another approach is capping or Winsorization of outliers .Using percentile capping. Values that are less than the value at 10th percentile are replaced by 10th percentile value, and the value greater than 90th percentile are replaced by 90th percentile value.

The other challenge is when I used the imbalanced data, the accuracy was very high but there was bias in predictions. So I used imblearn to reduce the imbalances in the target variable.

Limitations of this work and Scope for Future Work:

If we get data of other years along with other telecom companies we can predict on varied scenarios.

**Submitted by: Ankit soran**