

MACHINE LEARNING APPROACH FOR DEFAULT ANALYSIS IN A TELECOM DATASET

Submitted by:

PRIYANKA

**ACKNOWLEDGMENT**

Firstly, I would like to express my gratitude to Data Trained for giving me this opportunity to learn and practice through various practical ways. I would like to thank Flip Robo for giving me this opportunity to expand my learning horizons. This project would not have been a reality without the assistance of our mentor Mr.Harsh Ayush who dedicatedly replied to all the queries in no time. I would also like to mention Mr. Mahendra from the technical support team of Data Trained Academy to have helped me during the project. I have been continuously following Krissh Naik’s YouTube channel which helped me clear my basics.

I have gone through certain blogs and research papers to get insight about the problem mentioned below:

1. Does Microfinance Affect Poverty Reduction and Inequality in Indonesia?

<https://www.ijstr.org/final-print/apr2019/Does-Microfinance-Affect-Poverty-Reduction-And-Inequality-In-Indonesia.pdf>

1. Transformative Technology in Microfinance: Delivering Hope Electronically?

<https://www.researchgate.net/publication/305875278_Transformative_Technology_in_Microfinance_Delivering_Hope_Electronically>

1. A Brief Overview of Outlier Detection Techniques.

<https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561>

1. How to remove outliers for machine learning:

<https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/>

1. Data: How to handle Imbalanced Classification Problems

<https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/>

1. Top 10 Data Visualization Techniques, Concepts & Methods In Business (datapine.com)

<https://www.datapine.com/blog/data-visualization-techniques-concepts-and-methods/>

1. Performance Metrics for Classification Machine Learning Problems

<https://towardsdatascience.com/performance-metrics-for-classification-machine-learning-problems-97e7e774a007>

# Accelerating Indonesian microfinance with high tech and high touch

<https://asianbankingandfinance.net/banking-technology/commentary/accelerating-indonesian-microfinance-high-tech-and-high-touch>

**INTRODUCTION**

* **Business Problem Framing**

A product is being offered by a telecom company to their users in need of the hour. Understanding the importance of communication, with the help of an MFI the company has come up with a product through which they extend micro credits on the balance to their users to make sure the users have an uninterrupted communication. They are trying to focus the low-income users and users from the poor family. This credit extension comes with an attached risk. Risk of defaults in the repayment of the micro credits. The company wants to make a fair choice of users when extending the micro credit to minimize the risk of facing a non-repayment. For that we have to come up with a machine learning model that would predict the behaviour of the user and predict the probability of defaults attached with the user.

This problem of loan default is actually a very big risk to all the institutions who lend loans and all these institutions face a huge loss with loan defaults. Much attention is paid on this issue and different institutions use different methods to detect and predict this risk of default.

* **Background of the domain problem:**

Mobile financial services are an effective way of delivering microfinance services to the poor as they are efficient and cost effective as compared to traditional high-touch models which are in use for a long time now. One such micro financial service is offered by a telecom with the help of an MFI to their users. Telecom companies know the importance of communication and how it affects a person’s life. To allow their users to have continued communications they offer a micro credit on balance their users to continue using phones company extends micro credit to the poor families and low-income customers. The company offers micro credit that has to be to be paid back in 5 days. For credit loan amount of, 5 user has to repay 6 and a credit loan of amount 10 one has to repay 12 (in Indonesian rupiah). If a user fails to pay back in desired time he is considered as a defaulter and if a person repays within the mentioned time he is considered as a non-defaulter.

The mass that this product is mainly delivered to are the low-income families which mostly have no formal financial histories which makes it difficult to study their behaviour and to know if they will repay the credit loans on time. But the telecom company has a rich raw user data that will be used to build a machine learning model that would predict the behaviour of the user and would successfully predict the loan defaulters.

* **Review of Literature:**

Mobile revolution is the biggest revolution one has witnessed in the last decade. Mobile phones have changed the traditional method of communication. People even in the remote areas can now communicate at affordable prices. With this revolution in communication the globe has inter connected like never before. Now mobile phones are taking a step closer to revolutionizing financial inclusion of the unbanked masses. Globally, around 38% of the bankable population lacks the basic access to the formal financial institution. It is very important for a country to focus on financial inclusion as it directly affects the growth of economy and the GDP. Two most powerful players are the traditional banks and the telecom industry that can take financial inclusion to the masses. As these telecom companies have access to the millions of users even from the remote areas where formal institutions have no roots, it would one day outperform other players in reaching the goal of financial inclusion. There is no barrier of time and geography for the telecom industry. Beyond basic access to the formal financial system, telecoms are also providing digital payment methods, loans and financial products to their users. Also, the telecoms lend micro credit to their users in need. Telecoms usually extend such credit support for uninterrupted consumption of the network by the users. However, all the micro credits face a problem of estimating whether a potential borrower would repay the credit on time. There are many variables to check the behaviour of defaults. Many countries use credit scoring to predict the defaults. But credit scoring can only be done for those who are inclusive in the financial system. Telecoms have millions of those users too who are not a part of financial system.

Unbanked population usually lacks financial histories which makes it difficult for them to access credits assistance and also for potential lending institutions to extend credit support to them as they don’t have any information regarding them to rely upon based upon which credit support can be extended to them at a lower risk.

In contrast to the unavailability of financial history, these masses have a potential data linked to mobile phone usage which if used properly can come out as rich data that can help study the behaviour of the borrower and the risk attached to the consumer. This raw data can help telecoms understand their user behaviour.

The dataset that we have in this project is also a formal interaction of unique users with the mobile phone itself. The telecoms have the metadata of their clients. We are using the same data to predict the default behaviour. From this raw metadata we are going to build a machine learning model which will combine these features into a model of probabilistic prediction. This model would help the telecoms to disburse micro credits to the consumers at the same time reducing their risk of defaults.

* **Motivation for the Problem Undertaken**

The telecoms are coming up with different strategies and products to retain their users. One such product is micro credit on the balance. But a complexity attached to lending is the repayment. Telecoms to minimize the loss are trying to figure out the methods to predict customers behaviour. Machine learning algorithms can be effectively used to serve this purpose. Here, I will be working upon loan defaulter’s prediction using machine learning models.

**Analytical Problem Framing**

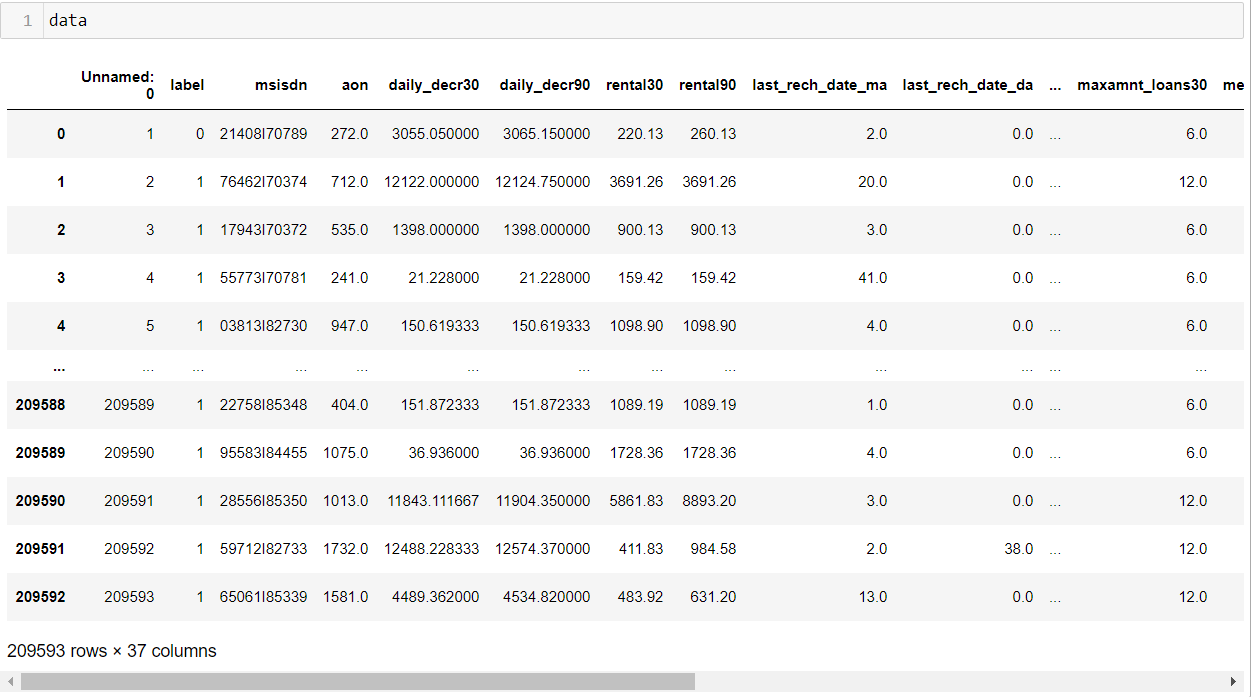
* **Mathematical/ Analytical Modelling of the Problem**

The data that is provided is of a telecom organization which along with an MFI has launched a product of extending micro support to the low-income users in need of hour. This credit facility comes under a basic trap of whether a user would repay the loan on time or not. We are given the meta data of the users using which we have to predict which person would probably be a defaulter.

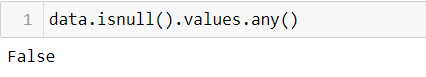
When analysing the data, it was observed that a few features had some negative entries which is not realistic in context of the dataset and the features, these negative values must be dealt with in order to have a better predictive machine learning model. Also, there were some noise entries also known as outliers observed in some of the features that were ranging from the other entries. A few features were in a format that would not give any insights about the data, so we must reframe those features so as to build a model in which the features could positively contribute in the prediction. Also, some features were observed to be redundant. We should drop such features. Our class label is highly imbalanced with majority entries as 1 so model might give us biased results or even might exclude all the zero entries considering them as noise. To start working on the project we must deal with these problems to get an unbiased model.

* **Data Sources and their formats:**

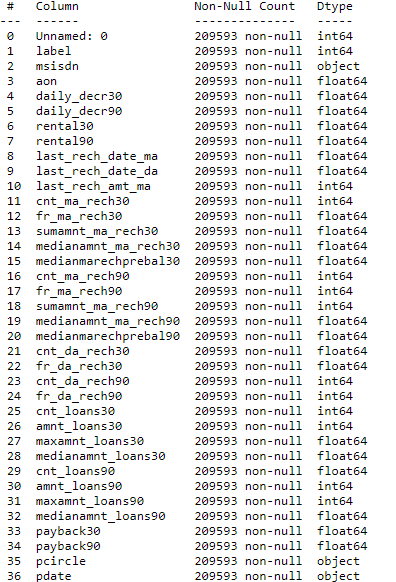
The dataset we are going to work upon covers the information of 2,09,593 users of a telecom company. The dataset has 37 features which includes the information about the mobile phone number of the user, age of the user on the network, daily consumption of the mobile balance by the user, recharges made by the user in the span of 30 and 90 days, credits loans taken by the user in 30 days and 90 days, number of credits taken by the user, amount of the credits taken by the user, maximum amount of the credit taken by the user etc. Along with these features, we have a feature Label which is a response variable which takes the value 1 in case of non-defaulter i.e. the user has paid back the credit on time and 0 otherwise. We have to predict the Feature Label on the basis of other features provided.



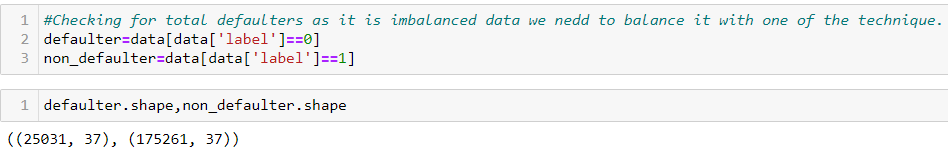
The dataset has no null-values.



The features in the dataset are of different datatypes. We can see integer, float and object datatype in the dataset.

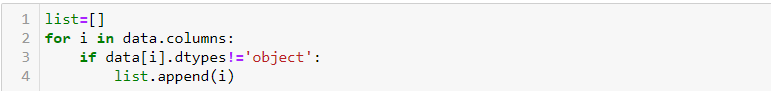


The feature label is of int datatype and takes binary entries either 0: defaulter or 1: non-defaulter. The feature label is an imbalance class comprising only 12% entries as 0 and majority entries as 1.



* **Data Pre-processing:**

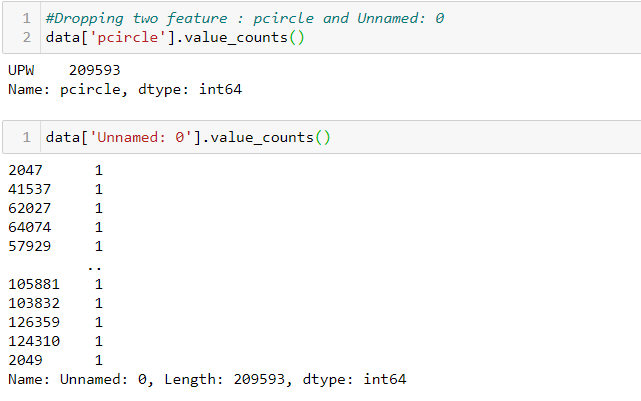
Before passing the data to the models firstly we have to pre-process the data to clean, integrate, reduce and transform the data into machine understandable standards. Observing the features, we came to know that the data has negative entries in some of the features: aon, daily\_decr30, rental30, rnetal90 has negative values to name a few. AON is the age of the user on the network, age can never be negative so having negative values is unrealistic. Also, rental30 is the credit average in 30 days credit loans can never be negative, so the negative data is unrealistic in the data. After observing negative values, the negative values were in context with the positive values just the sign was a problem else values were according to the features. So, it seemed important to keep the values and just remove the negative sign so as to not lose any information.





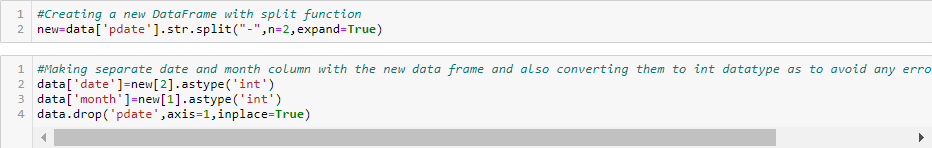
We have dealt with the negative sign using the abs method to make sure there is no loss of information.

It was also observed that features: Unnamed: 0, pcircle were redundant in the context that they were either constant or unique for the rows. Such features never help machine learning model to derive a pattern to predict the dependent feature. So, it is important to drop redundant features.



One can observe from the picture that pcircle has a constant value for every row and Unnamed: 0 has unique index value for every row. Dropping these two features would take our project in direction.

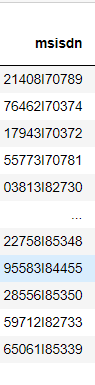
Two of the features: pdate and msisdn were in a format that machine would not draw a pattern from. pdate is in date-month-year format in which year is same for all the rows. So that is of no use to the machine. We split the date into date moth and year and kept the date and month information.



Now we have kept the new date and month feature and dropped the original pdate feature.



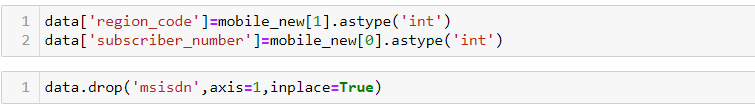
msisdn is a number uniquely identifying a network subscription in a global system for mobile communications. msisdn comprises of the county code, national destination code and subscriber number.



msisdn number is split into two parts with I might be indication of the country Indonesia. So, we will split msisdn upon I as I is common in every row we don’t need a constant variable.

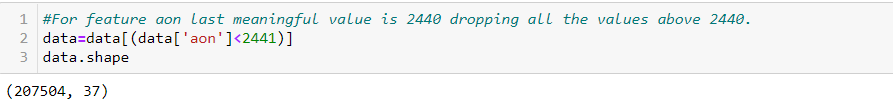


We will keep the number following I as the region code as there are only 887 unique entries in it and number prior to I as the subscriber number as there are 84580 unique values in it.



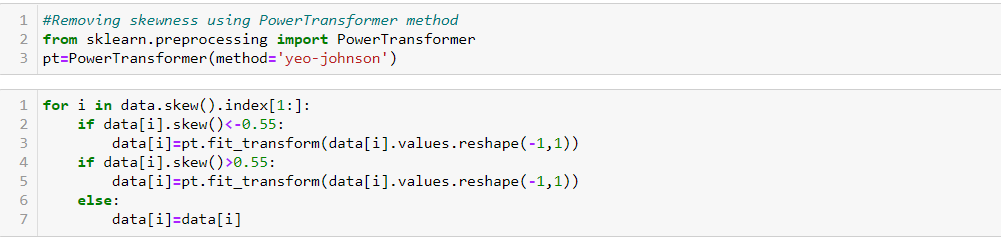
The original dataset has outliers which would create a biased model if not dealt. There are many ways to deal with outliers, when zscore was used to remove the outliers there was a loss of 15% of the data which is not a good amount to loss as we have real time data with us. IQR can also be used to remove the outliers, the IQR uses 25th and 75th percentile to remove the outliers, for many features the 75th percentile was even removing the logically correct data. For e.g. one of the feature maxamnt\_loans30 has either 0,6 and 12 as entries. Entry 6 made almost the data between 25th and 27th percentile, IQR technique removed the columns bearing values 0 and 12 which were actually important to consider. Many other features lost some of the very important information required to draw a pattern for machine learning. So, we were losing more than just outliers, which was not favourable to the dataset as we were losing correct information. At last, I settled with using the last correct value for each feature and dropped all the entries above the last correct value. This way the outliers were removed and the data loss was just of 5% which is acceptable.

According to my analysis of the data, features: aon, last\_rech\_date\_ma, last\_rech\_date\_da, fr\_ma\_rech30, cnt\_da\_rech30, fr\_da\_rech30, maxamnmt\_loans30 and cnt\_loans90 had outliers.

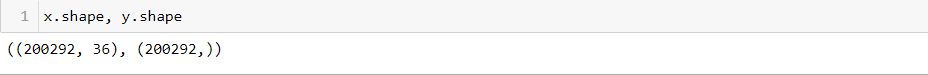


The picture above is how I dealt with outliers for the eight mentioned features. This ensured that the outliers were dropped and no meaningful information was lost also keeping the data loss to be as low as 5%. Initially we had 209593 observations now after the removal of outliers we are left with 200292 observations.

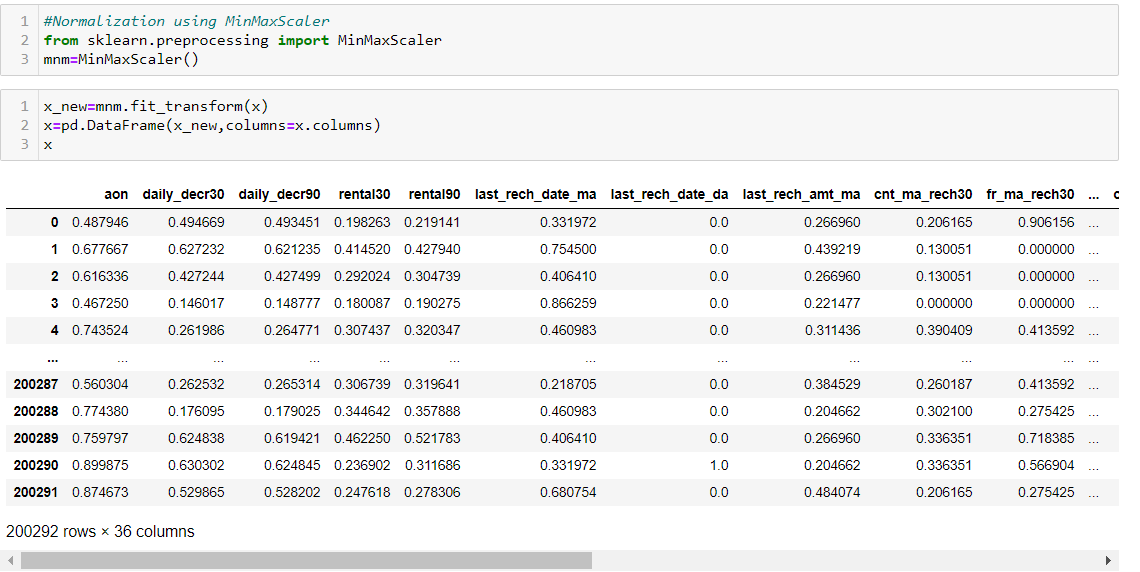
The data is also skewed, skewness must be reduced to have a gaussian like distribution so as to achieve better performance of the model. There are many methods that can help us reduce skewness like log transform, box-cox and the power transformer with method yeo-johnoson. We will use yeo-johnson to reduce the skewness as this method effectively reduces the positive as well as the negative skewness.



After successfully achieving a Gaussian like distribution next we would normalize the data using Min max scaler. We will split the data and separate our label to be predicted.



It is very important to scale the data as we have features with different ranges which would create complexity for machine to understand, so it is required to either normalize or standardize the data in order to have a single scale for the features so as to increase the performance of the model. We are normalizing the data here as it will scale the data between 0-1.



* **Data Inputs- Logic- Output Relationships**

The input features of the dataset are the information about the recharges and rentals taken by user. The output feature is a binary feature having 0 and 1 as the observations. The input features will be used to predict the output feature and tell whether a person would pay back the loan on time.

* **Assumptions:**

Feature msisdn here is mentioned as the mobile number of the user instead it is the unique number which identifies a network subscription in the global system. It is the sim number. Also, many users have the same msisdn number so instead of dropping it considering it to be unique for the users I have kept it and split it into parts to reduce the uniqueness of the feature.

* **Hardware and Software Requirements and Tools Used:**

**Hardware specifications-**

Processor- Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz

RAM-8.00 GB

Edition-Windows10

This project was completed using Python 3.0 which is an interpreted high-level and general-purpose programming language. Jupyter notebook which is an open-source web application provided GUI environment for the python notebook.

Important libraries that were used in the completion of this project are:

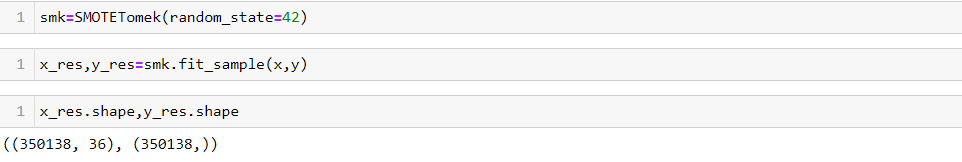
* Pandas- It is a python library which provides tools for data analysis and reading the data in the jupyter notebook.
* Numpy- Also know as Numerical python is a python library used to perform algebraic and statical analysis.
* Seaborn- It is a python data visualization library. It allows us to draw attractive and informative graphs for statistical understanding.
* Matplotlib.pyplot- Matplotlib is a plotting library in python used for creating visualizations in python. Pyplot is the sub-module of matplot library which is the collection of functions that makes matplotlib works like pyplot. Pyplot in simple language helps in making the figure decorative and informative. We can put labels with the help of pyplot, we can insert some lines and points in the figure using pyplot etc.
* Warning- The warning module of python warns the user of certain situation that aren’t fatal. It is distinct from an error and cause no stoppage in the program.
* Sklearn- Scikit-learn is a library in python programming language. It features various classification, regression and clustering algorithms like Support Vector Machines, Random Forests, etc.
* Sklearn.preprocessing- preprocessing is a sub-module of sklearn module which provides various utility and transformer methods to transform raw features into a suitable representation that is easy for machine to understand.
* Power transformer- power transformer is a module from the sklern.preprocessing library which is used to make the data more Gaussian-like by stabilizing the variance and minimizing the skewness.
* MinMaxScaler- is a module in the sklearn.preprocessing which is used to transform the features by scaling them to a given range. MinMaxScaler normalizes the data and the features are scaled in the range 0-1.
* Imblearn- imblearn is a module in python that has various techniques to help us in balancing the ratio of the data by either upscaling or downscaling.
* SMOTETomek- is a technique in the imblearn module which is used to handle the class imbalance. It Is a hybrid method which combines an upsampling technique with a downsampling technique to handle class imbalance.
* Train\_test\_split- It is a technique in the sklearn.model\_selection module which can be used for any supervised machine learning algorithm to randomly split the data for training and testing phase.
* Cross\_val\_score- It is a technique in the sklearn.model\_selection module which divides the data into n-folds and give an estimate of the score.
* GridSearchCV- It is a library function of sklearn.model\_selection package which optimizes the hyperparameters of the models based on the dataset.
* Sklearn.metrics- It is a module of the sklearn package which has several functions like classification report, accuracy score, F1 score, roc-auc-score and many more which helps to measure the performance of the model.

**Model/s Development and Evaluation**

* **Problem-solving approaches**:

Our data is imbalanced and a machine learning model would give a biased result towards the majority class. Balancing the data is crititval. We can either under sampling technique or over sampling technique to handle class imbalance. In under sampling the observations in the majority class are reduced in order to balance the class distribution, which means we are at risk of losing the crucial information. Oversampling means creating dummy entries in the minority class to match the number of observations of the majority class. Various techniques can be used to handle class imbalance like NearMiss, Tomek, SMOTE, SMOTETomek. We will use SMOTETomek technique to handle class imbalance as this technique is a hybrid of under-sampling and over sampling. It uses an under-sampling method Tomek with an over-sampling method Smote. This is present in the imblearn.combine module.

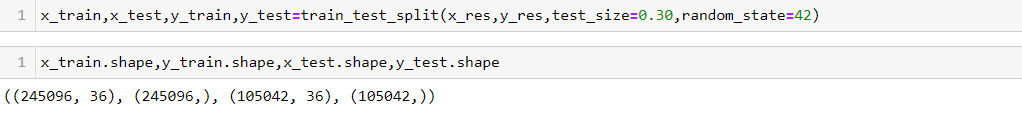




This way we have successfully handled the class imbalance. Now our data is ready to be fit in different models.

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

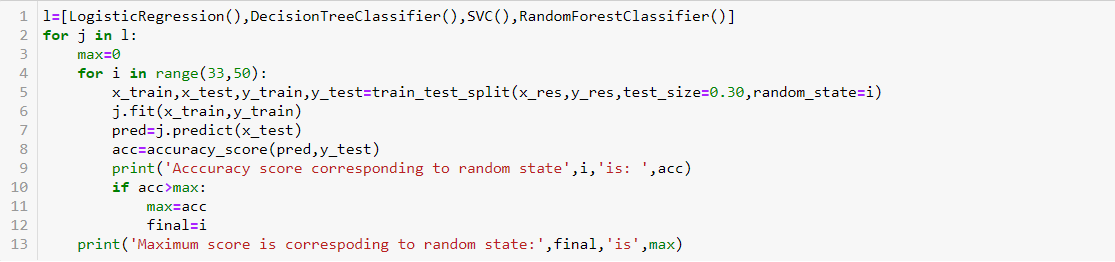
After handling the imbalanced data it’s time to split the data into training and testing for that we have used train\_test\_split algorithm from the sklearn.linear\_model module.



Now we will pass the training and testing data to four models. As it is a classification

problem, we will use classification models to predict the test data. We algorithms that we used here are: Logistic Regression, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier.

After the hyperparameter tuning of the random state, we have proceeded with the random state 42 to run the four models.



We have further optimized the parameters of the algorithms with the help of GridSearchCV.



* **Run and Evaluate selected models:**

Using the optimized parameters, we have run the models to predict the Label class.

The accuracy score obtained for different models is:

Logistic Regression: 0.772

SVC: 0.865

Decision Tree Classifier: 0.917

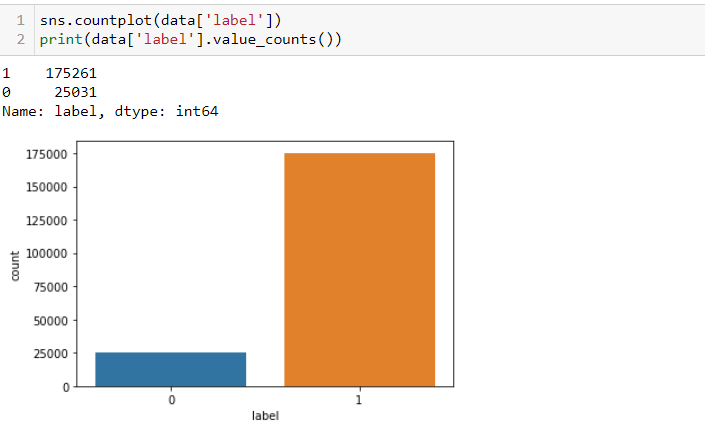
Random Forest Classifier: 0.952

* **Key Metrics used:**

As we have a binary classification problem statistical metrics like means squared error, mean absolute error and root mean squared errors cannot be used, we will use confusion metrics, classification report to check the performance of the models. As we had class imbalance, we will also use ROC curve to evaluate the performance of different classification algorithms.

* **Visualizations:**

To know more of how the independent features might be helpful in the prediction of the dependent variable visualization is required. We have visualized a few independent features with dependent feature label to study their relation.



The distribution of feature label is visualized using count plot which clearly depicted how imbalanced this feature is. With 0 entries at 25031 and 1 at around 175261.

**Label v/s aon-age on number:**



As highest count of user falls in the range of 0 year i.e. they are new to the network, we can observe counts for both label 1 and label 0 at highest. And counts for both the label decreases further as the years on network increases. There might be a pattern and dependency of target variable with aon.

**Label v/s medianamnt\_loans30:**

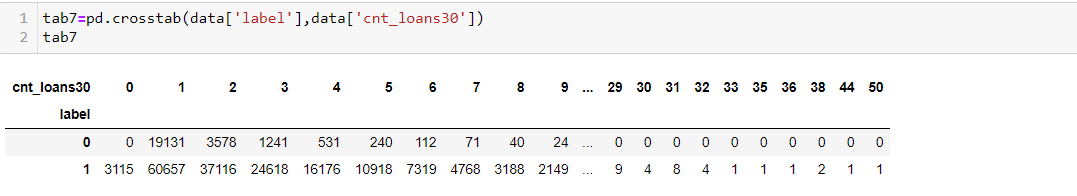
****

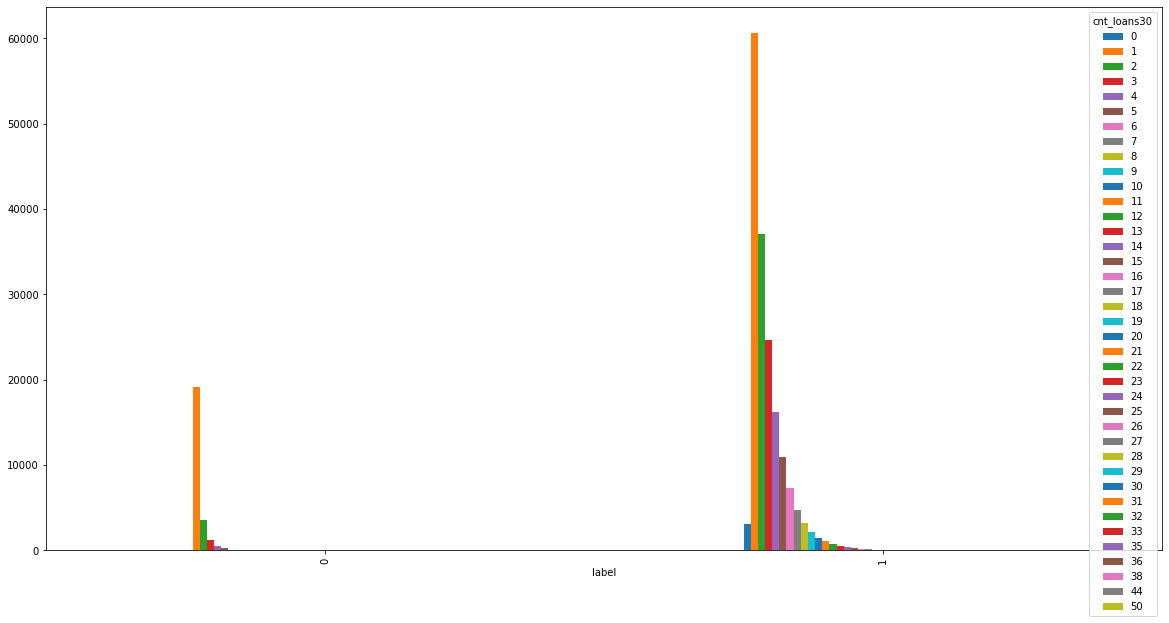
A huge difference between the labels is clear as we know the data is imbalanced. Still, it can be observed that users with a lower median amount are more likely to be defaulters.

**Label v/s cnt\_loans30:**

cnt\_loans30 is the number of times the user has taken credit in 30 days.

With the help of cross table we will visualize the variables.





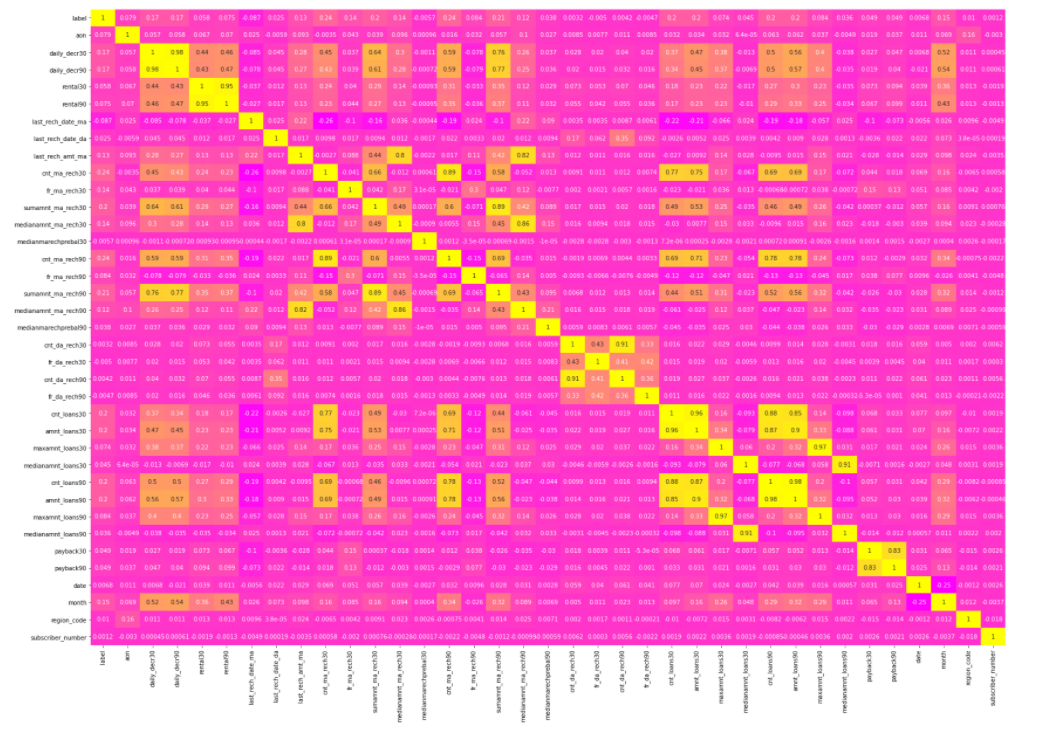
1. For 0 loans taken by a person there is no question of non-repayment.
2. Else, we can see a drastic fall in the defaulters as the count of loans increases i.e people who have taken lesser loan are likely to be defaulters. This might be helpful in predicting the defaulters.

**Label v/s maxamnt\_loans30:**

Maxamnt\_loans30 is the maximum amount that has been taken as credit.

****

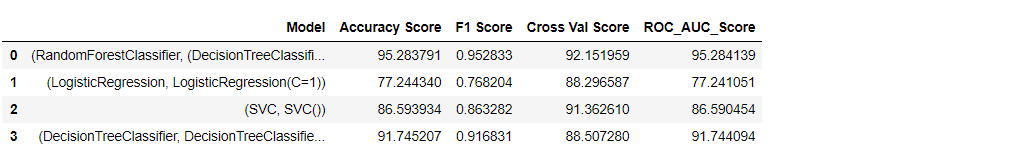
Users who have taken no loans are not defaulters as can be seen in the cross table. The number of users were highest for 6 which is reflected in the highest defaulter and non-defaulter count too. But around 13% of the total users who have taken a maximum loan of 6 are defaulters and only 4% of those who have taken maximum credit of 12 are defaulters. Let us further analyse the relationship among the variable via correlation.



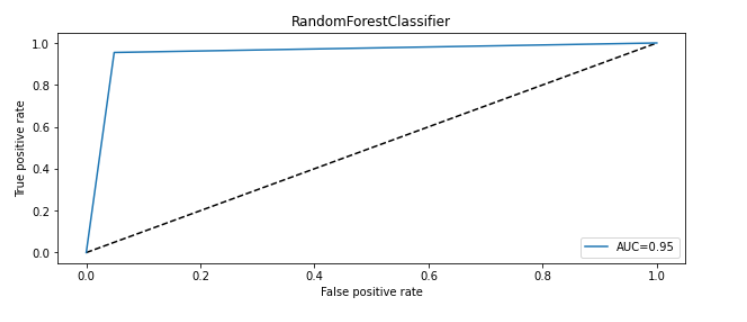
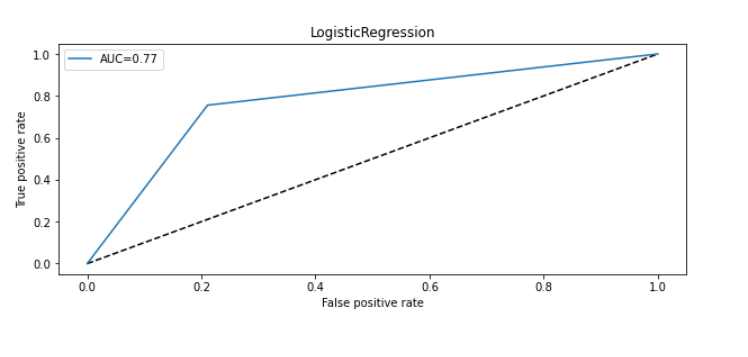
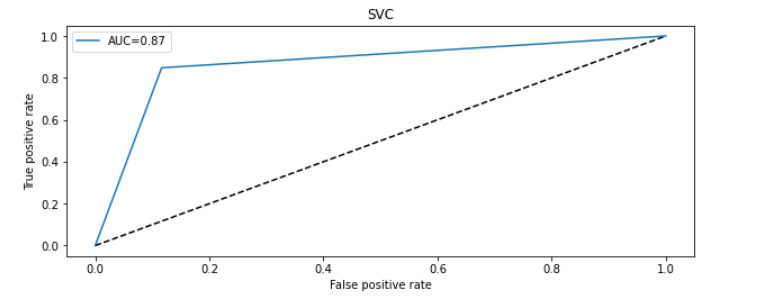
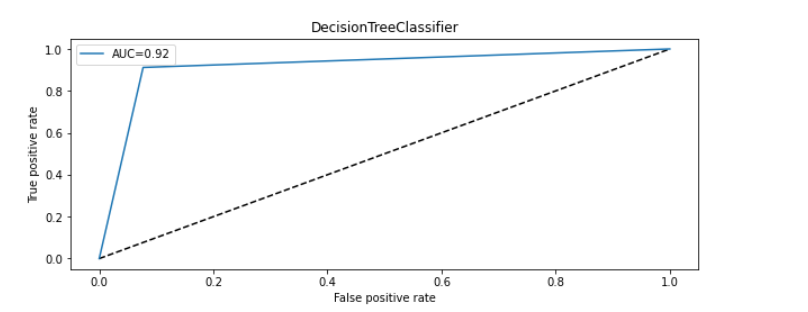
Label holds a positive corelation with daily\_decr30, daily\_decr90, cnt\_ma\_rech30, sumamnt\_ma\_rech30, cnt\_ma\_rech90, sumamnt\_ma\_rech90, cont\_loans30, amnt\_loans30, cnt\_loans90 and amnt\_loans90 these features would help the machine to positively draw a pattern in predicting the label.

* **Interpretation of the Results:**

After pre-processing, visualization, running the model and checking the performance via different metrics. We have results of four models:



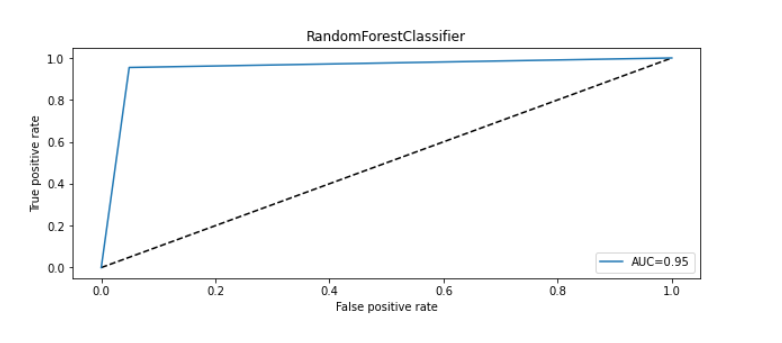
Also, the ROC curve is also used to check the performance of the metrics.

**CONCLUSION**

* **Key Findings and Conclusions:**

It was deduced that Random Forest predicted the class variable ‘Label’ with accuracy 95% and a cross validation score of 92.15% which confirms that our data is not under fit or over fit. F1 score corresponding to the random forest classifier is 95.28%. F1 score tell us that we have low false positives and low false negatives and we are correctly identifying the real defaulters and model is not much disturbed by false alarms. The ROC curve corresponding to the random forest classifier is:

 Random forest classifier is a bagging technique which uses decision tree as the base model and randomly selects subsets of the dataset for training purpose. Random forest as an ensemble technique reduces the risks of over fitting. So, it came out to be the best algorithm in predicting the defaulters.

* **Learning Outcomes of the Study:**

This data held information of the users of a telecom network which was used to predict the credit defaulters among the users. The data was in numeric format for most of the features. A few features were complex but we were successful in bringing out the best information out of those features. The main challenge in this problem was not to lose the data in removing outliers as we have real time data with us and much loss of information is not favourable. Traditional methods which were used to treat outliers were leading to data loss. So, I came up with a solution in which I thoroughly inspected the features with outliers and came out with the last acceptable value for every feature and dropped all the values above the last accepted value. This way the data didn’t suffer much loss of information. We were successful in bringing out a solution with a bagging classifier Random Forest which predicted the class with an accuracy of 95%. Our data was skewed and imbalanced there were chances of overfitting. Random forest performed well and also the model didn’t overfit.

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

It was observed in the feature msisdn some of the users had same mobile number which I believed were unique users as we had completely different information about them. So, there were chances that the same number was used by multiple users. Machine would treat them as different observations. There must have been some score linkage to each user to handle this same number situation. Along with this information if we had access to credit scores which a financial institution creates for the users this model would have performed even better.

Scope of future work:

Telecoms can further expand their business with the financial institutions and link the credit scores with their users data which can further elaborate this study and give precise results based on credit scores as well. This problem can also be extended by categorizing the users as banked and unbanked. Banked users have financial histories which would enable us to take this study further in deep. Unbanked people must be evaluated on certain other parameters apart from this mobile phone information.