

**Micro-Credit Defaulter Model Project**

**Submitted by:**

**Aditya Bhushan Pandey**

**ACKNOWLEDGMENT**

would like to thank the management of

MCO EKI for making the data available for this research.

I would like to thank the management of FlipROBO for making the data available for this research.

**INTRODUCTION**

* **Business Problem Framing**

Prediction of success or failure of micro credit loan use payment to improve the selection of customers for the credit. Furthermore, it would help us to know whether the customer will be paying back the loaned amount within 5 days of issuance of loan.

* **Conceptual Background of the Domain Problem**

Machine learning algorithms enable the creation of a new model using existing anonymized historical data that would be used to train the model to to know whether the customer will be paying back the loaned amount within 5 days of issuance of loan. With a good machine learning model, financial institutions could predict the likeliness of a client repaying the loan within the maturity date and then have procedures with predefined preventive actions to be taken, prior to this happening.

* **Review of Literature**

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. Microfinance services (MFS) becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The MFS provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Indonesian Telecom company is collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. 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 which is provided to us 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.

We have used machine learning model to predict the above. Classifier has been used to

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

* **Motivation for the Problem Undertaken**

I have been motivated by challenging imbalance dataset which was provided to us. I have tried my level best to retrieve a proper dataset through data cleansing techniques and further used it in my machine learning model creation to predict the best result.

**Analytical Problem Framing**

* **Mathematical/ Analytical Modeling of the Problem**

Tools which we have used during the project are as follows;

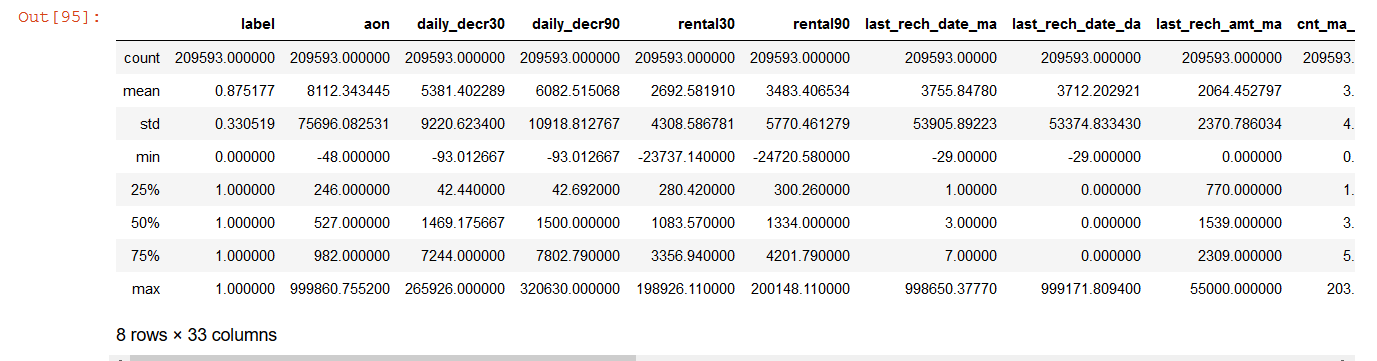
**Median** is used in the aon columned fill the minority data.

**Correlation** is used to find the highly correlated columns to analyse the best featured selection.

**Scikit-learn** is used as analytical tool for data processing & achieving the target. In scikit learn there are many packages which we have applied in the various process.

* **Data Sources and their formats**

Data has been provided by FlipROBO Company. Origin of dataset is Indonesian Telecom Industry. The snapshot of following dataset is provided below;



* **Data Preprocessing Done**

Feature Engineering has been used for cleaning of the data. Some unused columns have been deleted and even some columns have been bifurcated which was used in the prediction. As for example, date column has been bifurcated in days & month to complete the process. Using column aon (age of cellular network), some values have been deleted as outliers; using median to fill the minority data for improvising the outliers.

Even, using column label, we have deleted majority of rows and did the featured engineering for minority dataset.

After data cleaning, visualisation is done where population of label has been checked and we found that the defaulters are less than non- defaulters. Further, we have checked the number of defaulters & non- defaulters monthly.

Correlations of columns have been checked through the heatmap plotting and found many correlated columns and deleted the columns.

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

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

After checking the correlation, we have selected 32 important features as input. Except columns – aon, label, msisdn, pcircle & pdate; rest of the columns are chosen as input features. Column label was selected as target.

Depending upon the cellular users data history which we are using as input directly affecting the target.

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

After reading the dataset, we have preassumed that classification ML technique will be used for predicting the data.

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

**Hardware:** minimum 4GB RAM, i5 processor.

**Software:** Python, Jupyter notebook.

Packages: Pandas, Numpy, Scipy, Seaborn, Model\_Selection (Train\_Test\_Split), zscore, matplotlib.pyplot, sklearn.model\_selection( train\_test\_split,cross\_val\_score), sklearn.linear\_model (LogisticRegression), sklearn.svm (SVC), sklearn.tree (DecisionTreeClassifier), sklearn.neighbors (KNeighborsClassifier), sklearn.naive\_bayes (GaussianNB), sklearn.metrics (classification\_report,confusion\_matrix,accuracy\_score,roc\_curve,auc,recall\_score,precision\_score), sklearn.externals (joblib)

**Model/s Development and Evaluation**

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

We have used some statical tools like median, correlation, data skewness, Standard deviation for data analysis.

We have used scikit-learn as analytical tool for achieving the target.

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

In this research, five algorithms are implemented for micro credit loan repayment success predictions.

They are as follows;

1. KNeighborClassifier,
2. Logistic Regression (LR),
3. Decision Tree Classifier (DT)
4. Support Vector Classifier (SVC),
5. GaussianNB

We have run the different algorithms after applying all the previous process of importing, cleaning the dataset to test the dataset. The different algorithms have been described below;

**KNeighborClassifier**

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

**In k-NN classification**, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

**In k-NN regression**, the output is the property value for the object. This value is the average of the values of k nearest neighbors.

In the current case study, we have used KNeighbourClassifier to predict the target.

**Logistic Regression (LR)**

LogisticRegression is utilized to estimate discrete values or Binary values like 0 or 1, yes or no, true or false ) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event. Its output values lie between 0 and 1 as it predicts the probability.

**Decision Tree Classifier (DT)**

It is a type of learning algorithm which is mostly utilized for classification dilemmas. Remarkably, it works for both categorical and continuous dependent variables. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

**Support Vector Classifier (SVC)**

The objective of a Linear SVC (Support Vector Classifier) is to suit to the info you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you'll then feed some features to your classifier to ascertain what the "predicted" class is.

**GaussianNB**

GaussianNB is utilise in classification and it assumes that features follow a normal distribution. According to a normal (or Gaussian) distribution, a typical assumption is that the continuous values associated with each class are distributed while dealing with continuous data. This algorithm is very firmly known for multi class prediction feature. Here the probability of multiple classes of target variable easily get predicted using GaussianNB.

After using all the algorithms, we have found “Decision Tree Classifier” has scored best I terms of test the dataset and is the best suited for training the machine.

Listing down all the algorithms used for the training and testing.

* **Run and Evaluate selected models**

We have used all the classifier algorithms and detailed description along with the snapshots have been included in the mentioned below;

(1) KNeighborClassifier,

(2) Logistic Regression (LR),

(3) Decision Tree Classifier (DT)

(4) Support Vector Classifier (SVC),

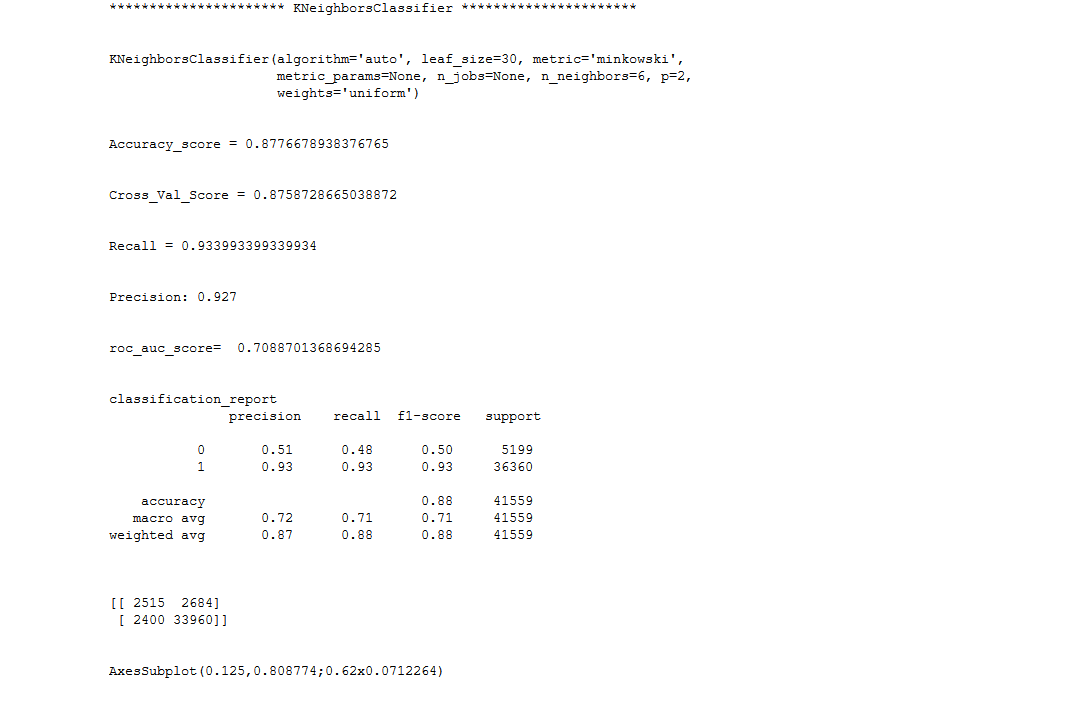
(5) GaussianNB

The following is the snapshot of code written for all the algorithms;

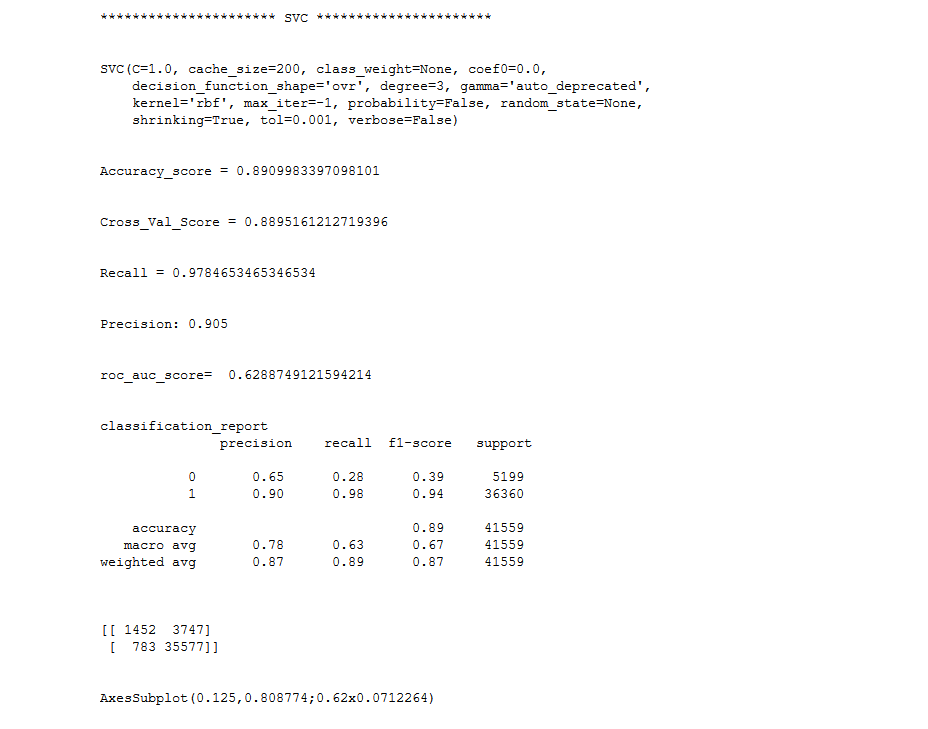


From the above codes, we have observed the following observations;

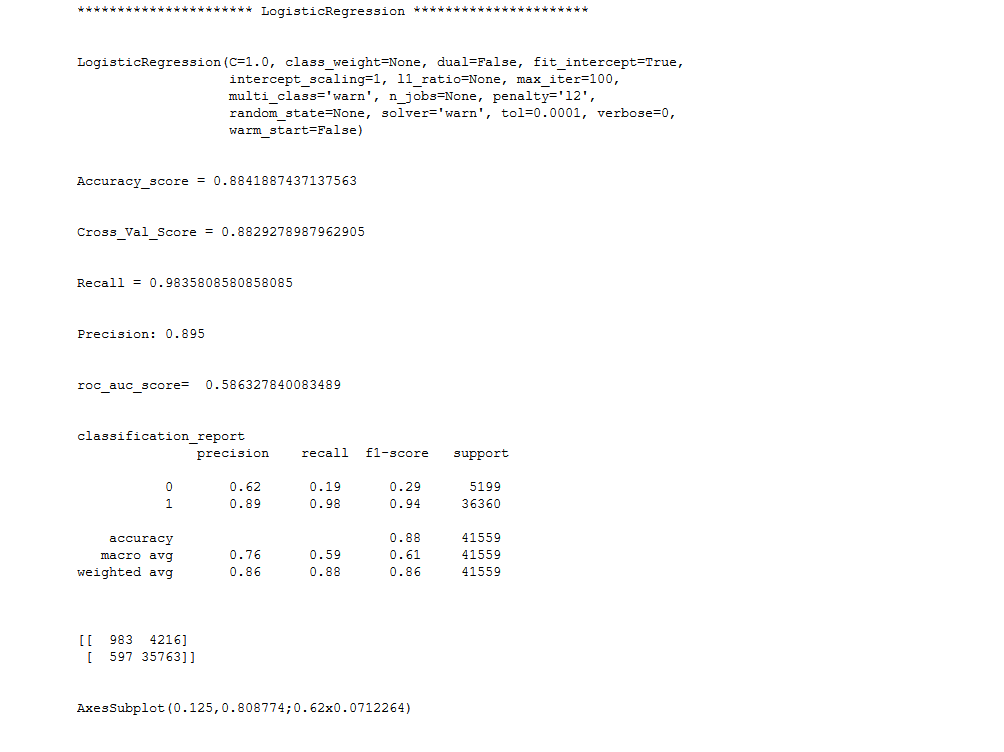
**Snapshot of output of KNeighborClassifier Algorithm**



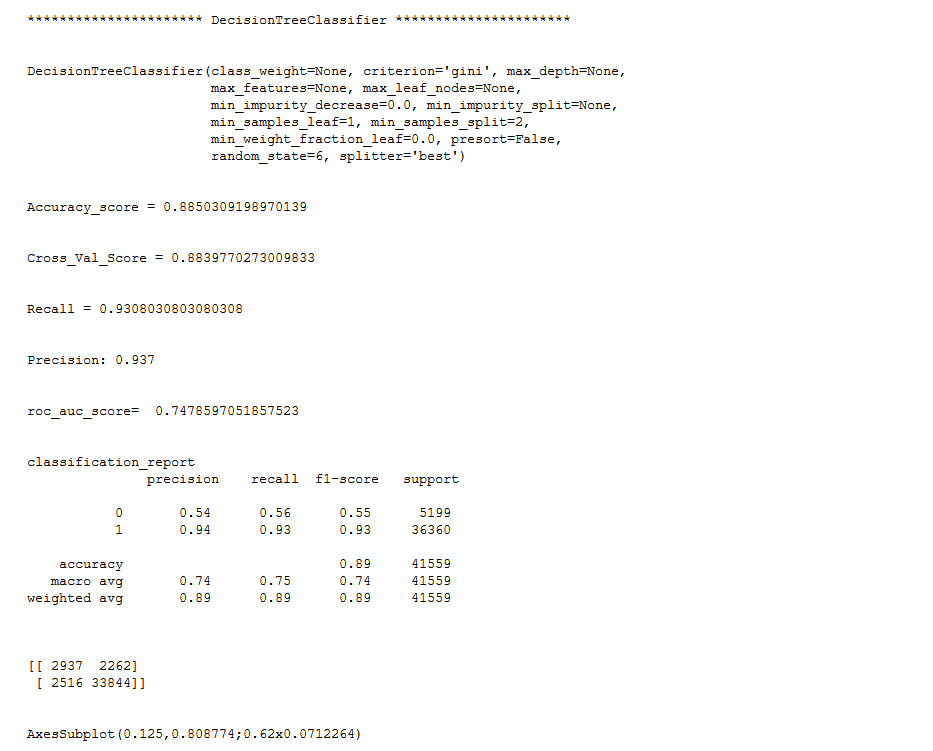
**Snapshot of output of SVC (Support Vector Classifier) Algorithm**



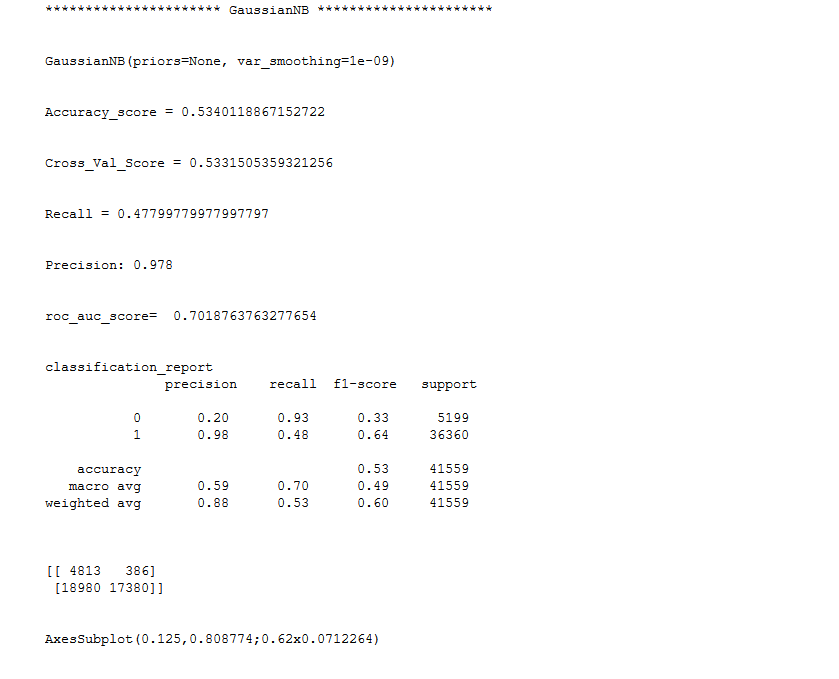
**Snapshot of output of LogisticRegressor Algorithm**



**Snapshot of output of DecisionTree Algorithm**

****

**Snapshot of output of DecisionTree Algorithm**



From the above output of different algorithm, we have seen that DecisionTree Classifier has the highest accuracy rate of different evaluation metrics used within it.

* **Key Metrics for success in solving problem under consideration**

Choosing the right metric is crucial while evaluating machine learning (ML) models. Various metrics are proposed to evaluate ML models in different applications, and I thought it may be helpful to provide a summary of the metrics which I have used in the following case.

Classification is one of the most widely used metric in machine learning. As per my knowledge, classification is best suited to solve the given case.

Models such as support vector machine (SVM), logistic regression, decision trees, random forest, XGboost, convolutional neural network¹, recurrent neural network are some of the most popular classification models.

There are various ways to evaluate a classification model, and I am covering the one which I have applied in solving the current case.

**Confusion Metrix**:

One of the key concept in classification performance is 36confusion matrix, (not metrics) which is a tabular visualization of the model predictions versus the ground-truth labels. Each row of confusion matrix represents the instances in a predicted class and each column represents the instances in an actual class.

Other metrics are;

**Classification Accuracy**

Classification accuracy is perhaps the simplest metrics one can imagine, and is defined as the**number of correct predictions divided by the total number of predictions,**multiplied by 100**.**

**Precision**

There are many cases in which classification accuracy is not a good indicator of your model performance. One of these scenarios is when your class distribution is imbalanced (one class is more frequent than others). In this case, even if you predict all samples as the most frequent class you would get a high accuracy rate, which does not make sense at all (because your model is not learning anything, and is just predicting everything as the top class).

**Recall** is another important metric, which is used to retrieve the fraction of samples from a class which are correctly predicted by the model.

**ROC Curve**

The receiver operating characteristic curve is plot which shows the performance of a binary classifier as function of its cut-off threshold. It essentially shows the true positive rate (TPR) against the false positive rate (FPR) for various threshold values. Let’s explain more.

**AUC**

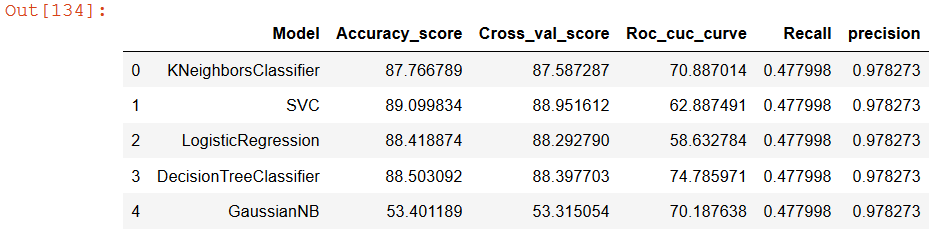
The area under the curve (AUC), is an aggregated measure of performance of a binary classifier on all possible threshold values (and therefore it is threshold invariant).

**Cross Val Score**

Cross-validation starts by shuffling the data (to prevent any unintentional ordering errors) and splitting it into k folds. he results from each evaluation are averaged together for a final score, then the final model is fit on the entire dataset for operationalization.

All the metrics have been used for analysing the behaviour of data along the models after the fitting the algorithms regarding the machine learning.

Following table provide the value of different metrics used in the different algorithm;

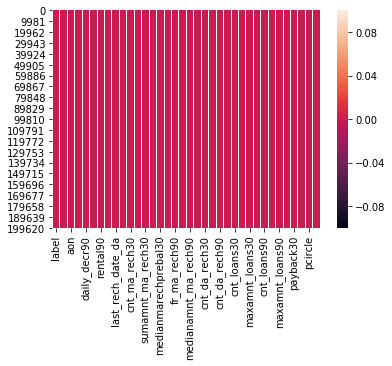


* **Visualizations**

We have used the following plots to interpret the data set to reach the target;

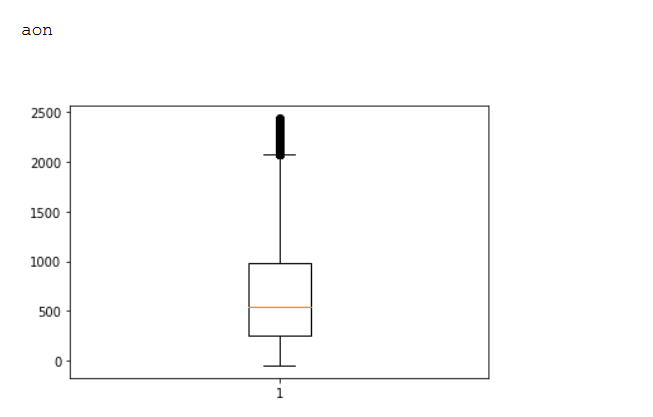
**Heatmap**

After importing the dataset, it has been read and null value has been analysed. Heatmap has been used to find all the null values in the dataset. Even null values are found in the dataframe. Cleaning of data has been done as for example;



Interpretation: Heatmap has been plotted to analyse the null values & no null value exist in the dataset.

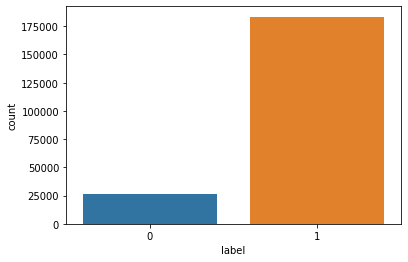
**BoxPlot with aon column**

We have plotted outliers in boxplot for each & every columns, below we have represented the boxplot of aon column.

Interpretation: After plotting the boxplot of different column, we have found the outliers existing in every columns in a dataset.

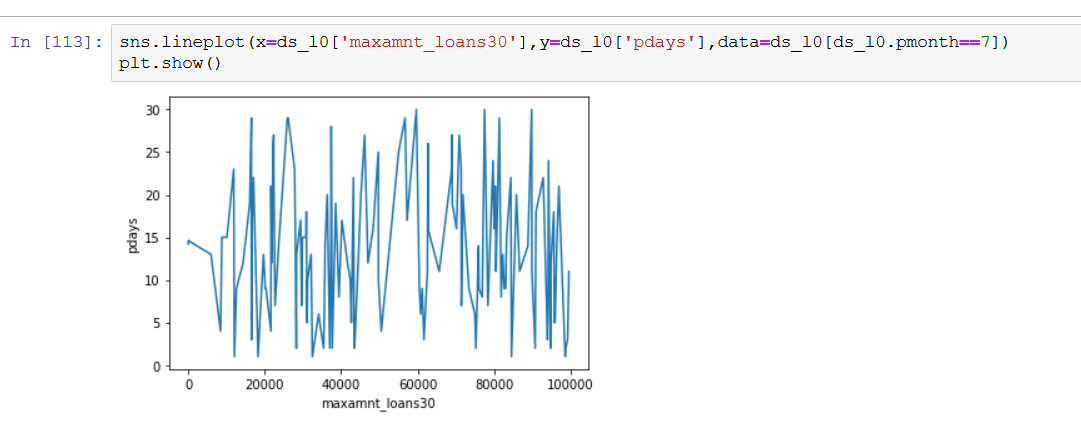
**CountPlot/ Matplot**

We have used countplot to analyse the “0 & 1” i.e.; number of success and failure in the given dataset.



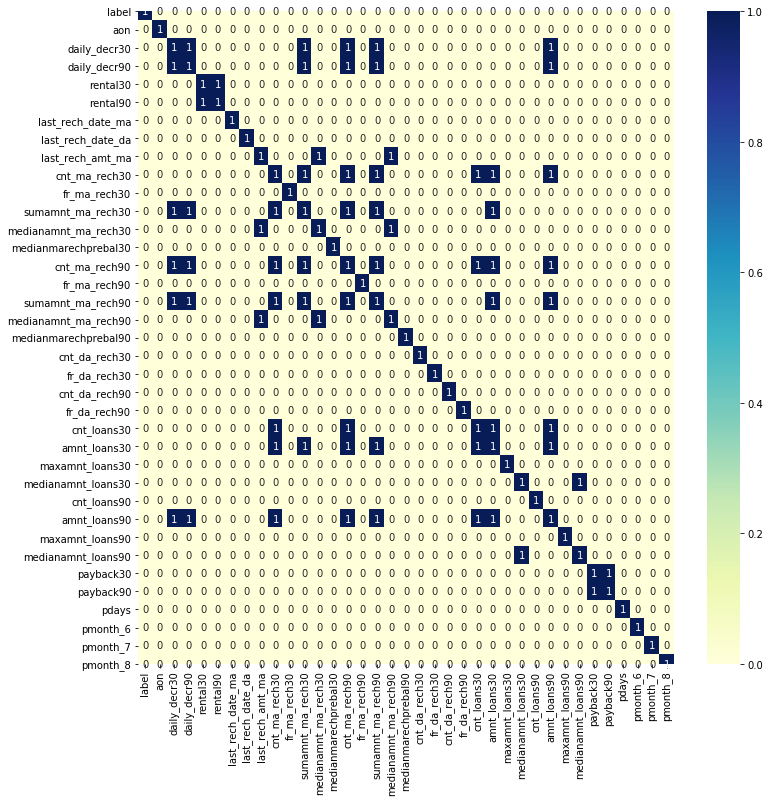
Interpretation: We have found the number of success & failure rate which further tell us the given dataset is imbalanced.

**Lineplot**



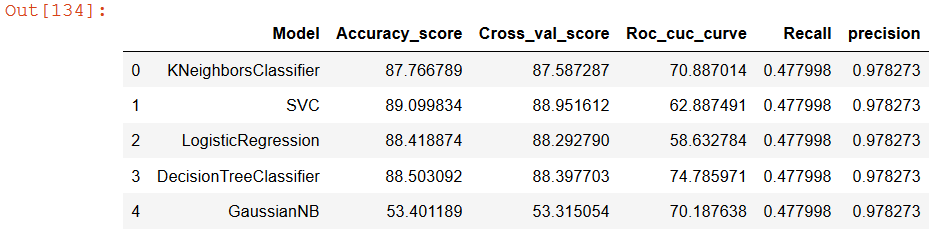
Interpretation: Use of lineplot to analyse the frequency of maximum loan amount on daily basis in a month.

**Heatmap**



**Interpretation**: Use of heatmap to check correlated columns. We have found some of the columns are highly correlated. Some of the columns are daily\_decr90','rental90','cnt\_loans30','payback90'

* **Interpretation of the Results**



From the above table, one can see the different algorithms have different values in their metrics. From a comparative analysis, we have found the best algorithm used for training the machine is Decision Tree Classifier as it has the highest value in all the different metrics used for testing the dataset.

**CONCLUSION**

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

There are many columns which are not be use for modelling and we have neglected them. Some of them are msisdn, pcircle, pdate etc.

We have created two new columns i.e.; day & month from date column and deleted some of the row in aon column as there were many rows which covering the data beyond the specified date. to easily apply in the model.

Number of success & failure rate of loan payers has been found which interpreted that the dataset is imbalanced.

Some of the columns are highly correlated, they are as follows; daily\_decr90','rental90','cnt\_loans30','payback90' which were deleted to apply in model creation.

Different classifier models have been used to test the dataset and we have found DecisionTree Classifier was best suited to train the machine for micro credit loan case study.

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

From the above problem, we have learned that through power of visualization, we can directly check the outliers through analysing the history of the dataset.

The given dataset was too large. According to data cleaning,we have learned further that the dataset was imbalanced. Even, we have found high correlated data which were deleted. Many outliers were seen in the dataset. Within the given time limit, we have tried to delete the outliers as possible as we can.

From the whole case study, we have seen the best algorithm used to train the machine according to the dataset is DecisionTree Classifier as all the values along the metrics were highest.

We have found the challenge while fitting the models due to large dataset. Our laptop was keep on going in sleep mode simultaneously due to large dataset.

The only overcome of the problem in the project was patience as we need time to run the algorithms and laptop was keep on sleep mode simultaneously.

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

The only limitation was time boundation.

The future scope of project is that we can train the machine identify & restrict the fraudulent through micro credit loan business.

This could further aids the telecom business income generation by restricting the fraudulent.

The further steps to improve the study could be;

We can use techniques like featured engineering, PCA and different boosting algorithms further to retrieve better results.