

**MICRO CREDIT DEFAULTER PROJECT**

Submitted by:

**ANJANA.P**

**ACKNOWLEDGMENT**

I would like to express my appreciation to team Flyproof for giving such a realistic data for analysis, with a full-length description of the project. My mentor Mr.Harsh Ayush has helped me in many stages of this project where I was stuck with problems. I use this opportunity to thank him for helping me at the right time without any delay.

I also thank DataTrained academy team for their wonderful classes and also their live support team who have been there at any time to help.

Also, this project made me search for a lot of data’s in several webpages and sites, that helped me to rectify my doubts and, I was able to study more about data analysis.

**INTRODUCTION**

* Business Problem Framing

A Microfinance Institution (MFI) is an organization that offers financial services to low-income populations. Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient, efficient, and cost saving. 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 focus 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 from our client database. 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.

* Conceptual Background of the Domain Problem

This project is to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of 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.
* Data is expensive and we cannot lose more than 7-8% of the data while treating the outliers.
* Review of Literature

First of all the data is saved in a csv file. Then its shape, datatypes, column value counts are all checked to get an outline of the data collected. Null values and correlation between the columns are checked using heatmap. Univariate and bivariate analysis are done for more clarification .skewness of the input columns are checked and resolved. Random oversampling is used to balance the dataset. PCA is done to rectify multicollinearity problem and standardscaler is used to resolve the unrealistic values. z score method is used, and outliers are replaced with the median values, so as to lose only very less data. Then train test split is done and checked for the best model and a model with high accuracy score and also relatively high cross validation score is selected as the best model.

* Motivation for the Problem Undertaken

The main objective behind doing this project is to make an understanding of the micro financial services that are widely accepted nowadays as a poverty reduction tool. They also focus primarily on low income families and remote areas. Hope this analysis may help micro financial industries to deliver more offers and help more unbanked poor families.

**ANALYTICAL PROBLEM FRAMING**

* Mathematical/ Analytical Modeling of the Problem

In the describe function we have checked mean, std.deviation, minimum, maximum, 25 percentile, 50 percentile, 75 percentile of each attribute columns.

Mean is the average, median is the central value and mode is the frequency.

Percentile is the value below which the percentage of data falls.

We also use evaluation matrix like confusion matrix, accuracy score classification report, auc-roc. These can be expressed in mathematical formulas as:

Accuracy = TP+TN / TP+FN+FP+TN

Recall (True positive rate TPR) = TP / TP+FN

False negative rate (FNR) = FN / TP+FN

Precision =TP / TP+FP

F1 Score = precision – Recall / Precision +Recall

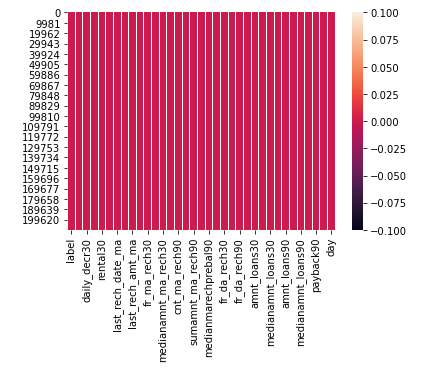
Specificity = TN / FP+TN

* Data Sources and their formats

FlipRobo technologies have provided this dataset for detailed analysis which was collected from a telecom industry in Indonesia. The data collected was in an excel sheet with a very detailed description of each columns with it. The data is converted into csv file and loaded in Jupiter notebook first. There are 37 columns and 209593 rows in this dataset. The data was in integer, float, object and date datatypes. The output column name is ‘label’, which is filled with only 0’s and 1 value. Thus, it is understood that Logistic regression and classification methods should be used for the model prediction.

* Data Preprocessing Done

First the important libraries for preprocessing and also csv file for analysis is imported. Shape and datatypes of columns are checked. There are 209593 rows and 37 columns in our dataset. There are object, integer, float and date datatypes. The object and date datatypes should be converted to integer datatypes for the analysis. Unnecessary columns like index, year, msisdn, pcircle are dropped as they provide no necessary information for our analysis. Null values are checked and should be cleared if there is any.

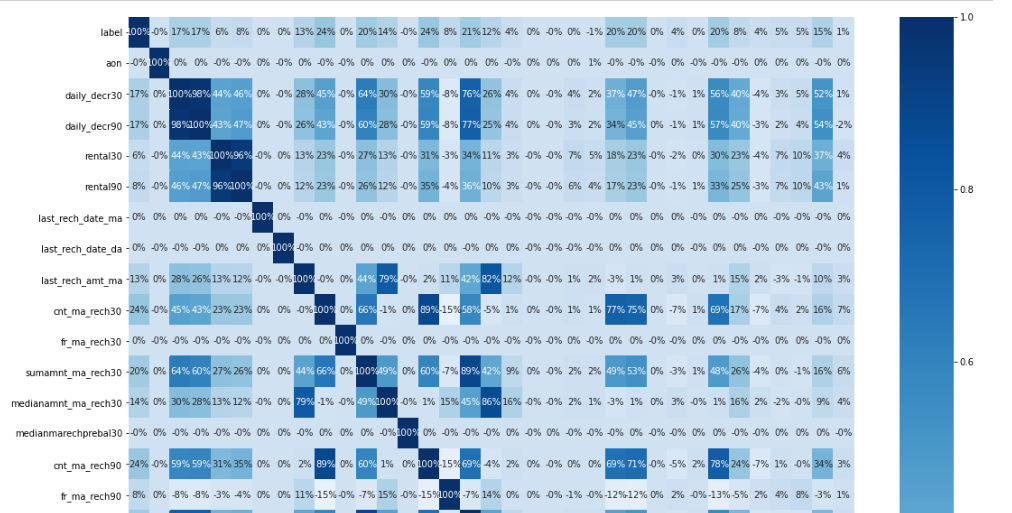


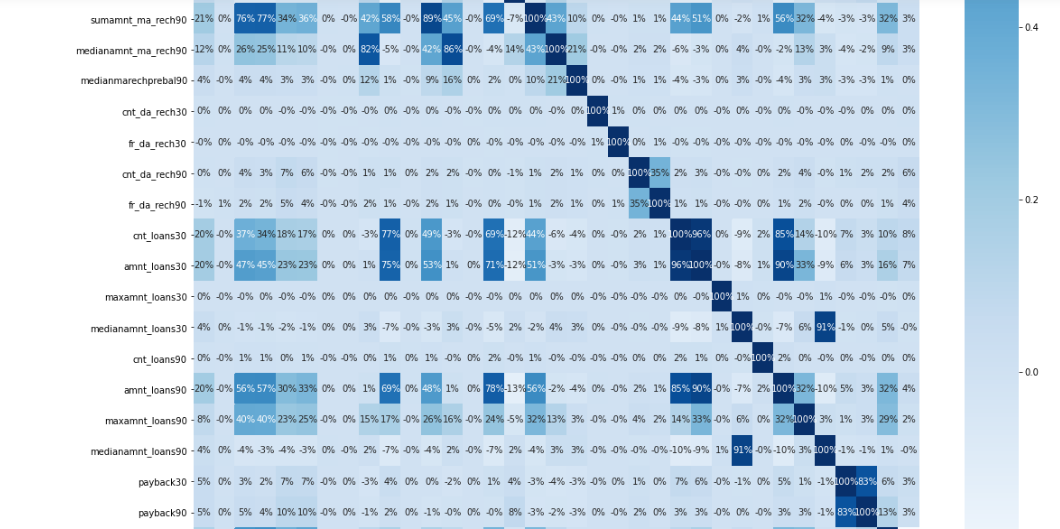
The uninterrupted lines show there are no null values in our dataset.

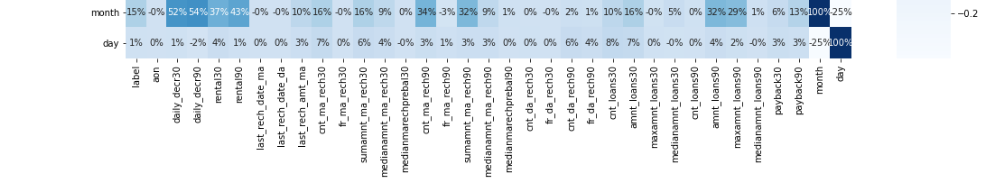
Describe functions provide information with minimum,maximum,mean,std.deviation,25th percentile,50th percentile 75th percentile of each column. We can understand that mean is highest for daily\_decr\_90 and lowest for medianamnt\_loans90. There are a large no. of negative valued columns which they cannot be negative. This shows there are unrealistic datas present. There is large difference between 75% and maximum for many columns, this represents the presents of outliers. So we must treat this unrealistic datas and outliers for a better results and predicting the best model.

* Data Inputs- Logic- Output Relationships

The relation between each column and also relationship of each column with the output column can be represented graphically using a heatmap.







Label is having a strong positive correlation with cnt\_ma\_rech30, cnt\_ma\_rech90, sumamnt\_ma\_rech90, cnt\_loans30, amnt\_loans30, amnt\_loans90.And label is having a negative correlation with fr\_da\_rech90.Also we can see highly correlated columns other than label column. This shows the presence of multicollinearity between feature columns. Multicollinearity can be identified by various methods such as correlation heatmap, pairplot, Variance inflation factor etc. Multicollinearity may lead to low accuracy results and also may affect our model prediction. So multicollinearity must be treated. They can be treated using PCA (principle component analysis) method. Skewness is checked to check whether the data is skewed or not. Highly skewed datas shows the presence of outliers. The threshold value of skewness ranges from -0.55 and +0.55. Skewness can be resolved using log transformation, boxcox method or yeo-johnson method . Here I used yeo-johnson method because it does not strictly needs the input variables to be positive.

Zscore is calculated and instead of removing the outliers I filled outlier values with the median of each column so that we won’t lose much data.

* Assumptions under consideration

Label is having a very positive correlation with no.of times main account recharged, total amount of recharge and total amount of loan taken. That means we can predict whether the customer is defaulter or not by checking his account balance and recharge amount history. Customers who recharge their accounts frequently and who maintains their account balance high may not be defaulters.

Label is having a very poor relation with frequency of data account recharged, median of account balance before recharge and age on network. So, we can assume that a defaulter or non-defaulter doesn’t depends on old or new customer. Also, his data account history and balance before recharge doesn’t play a key role in this prediction.

* Hardware and Software Requirements and Tools Used

I used intel core i3 processor, 4GB RAM and 64 bit operating system as hardware and windows 10, MS excel, MS word and python 3 Jupyter notebook as software for the completion of this project. In jupyter notebook various libraries are also used. They include pandas, numpy, matplotlib , seaborn , imblearn and sklearn.

**MODEL/S DEVELOPMENT AND EVALUATION**

* Identification of possible problem-solving approaches

The major problems we dealt with this dataset. They are.

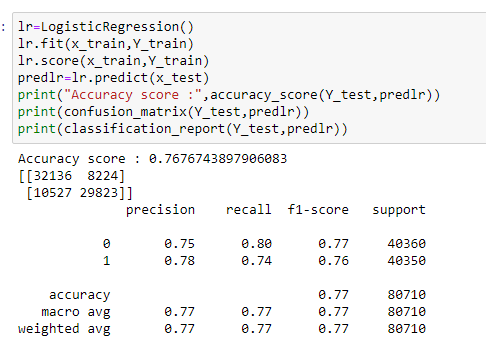
1. imbalanced – An imbalanced data is a common problem in machine learning. It puts accuracy down. It occurs with a disproportionate ratio of observations in each class. They can be dealt with upsampling the minority class or down sampling the majority class. I have used oversampling of the minority class. It randomly selects the minority class and add them to the training dataset. It increases the ratio by replicating.
2. Unrealistic data-We dealt with unrealistic datas in this dataset. They are solved using standard scaler. standard scaler standardises the features by removing the mean and scaling to unit variance
3. Multicollinearity- multicollinearity refers to the collinearity between the features. Multicollinearity occurs when our model includes multiple factors that are correlated with each others other than with label. It makes more difficult to predict the correct model and also affects the accuracy. They are treated using PCA (principle component analysis) method. This algorithm reduces the no. of columns by removing highly correlated feature columns.

* Testing of Identified Approaches (Algorithms)

After the train test split of the data many models like logistic regression, Gaussian naïve bayes ,knearest neighbour, svc, decision tree classifier ,auc\_roc curve and decision tree curve are checked. Decision tree classifier gives the best accuracy score, so the cross-validation score of dtc is checked.

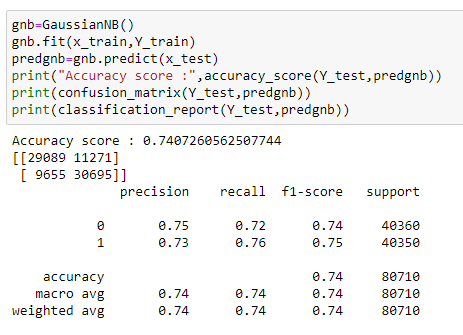
* Evaluation of selected models

1. Logistic regression-It is a classification algorithm used to predict the probability of categorical dependent variable.



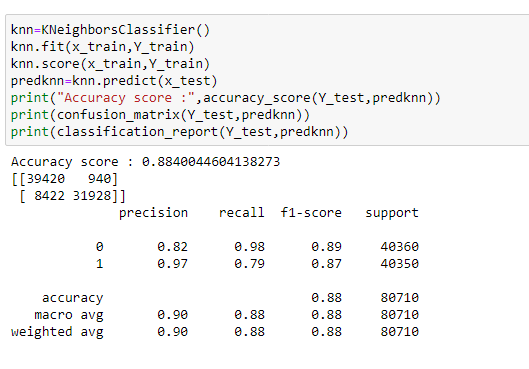
Logistic regression is giving an accuracy score of 0.76.

1. Gaussian Naïve Bayes-It is a variant Naïve Bayes that follows Gaussian normal distribution and supports continuous data.



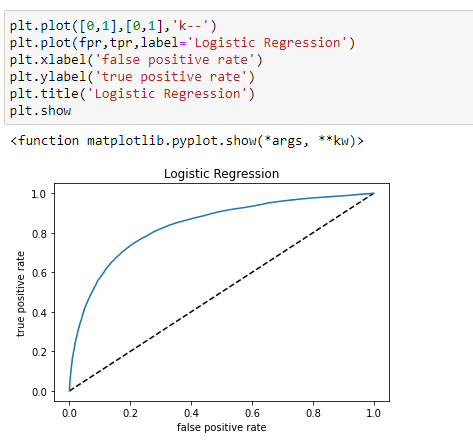
Gaussian NB is giving an accuracy score of 0.74

1. K Nearest Neighbor-It is found that two neighbors who have identical distance but different labels,the result will depend on the ordering of the training data.



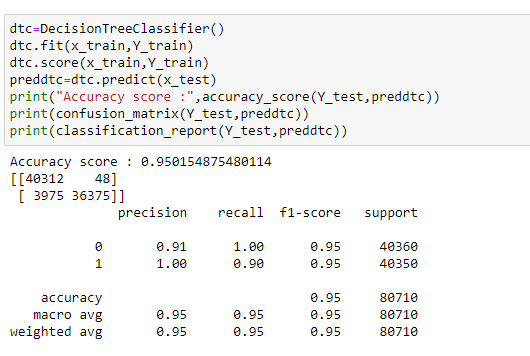
KNN is giving an accuracy score of 0.88

1. AUC\_ROC –It is an important evaluation metrics to check any models performance.auc\_roc graph is drawn with different threshold valued outputs based on true positive rate and false positive rate. The more the area is under the curve, it shows that the model performs well.



We get an auc\_score of 0.76.

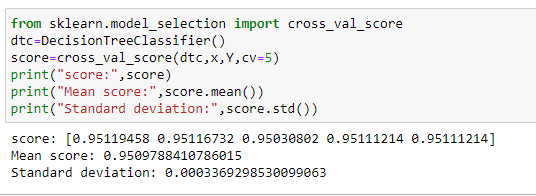
1. 5)Decision tree classifier – A tree structure is constructed that breaks the dataset into smaller subsets eventually resulting in prediction.The root node partitions the data based on most influential feature partitioning.There are two measures for this. They are gini impurity and Entropy.



Decision tree classifier is giving an accuracy score of 0.95

* Key Metrics for success in solving problem under consideration.

Using logistic regression,GaussianNB, KNN classifier,Decision tree classifier etc, Decision tree classifier gives maximum accuracy score of 95%.So its cross validation score is checked.

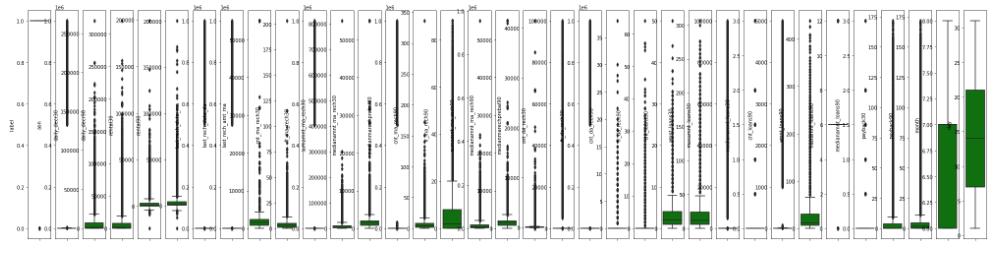


Cross\_val\_score of DTC also gives a score of 95%. That means our data is not overfit.We can also see that std.deviation is very much low.

After checking cross\_val\_score,I used ensemble technique to boost the score.For that RandomForestClassifier is used.And it gives a score of 0.99 .Cross\_val\_score of Randomforest classifier is checked and it gives a score of 0.97 . Since it is the best score, I saved this as the best model in picklefile format.

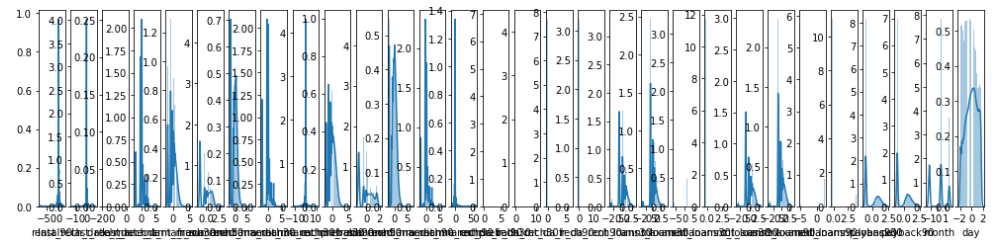
* Visualizations

1. Boxplot- Boxplots are the best methods to check for the presence of outliers.



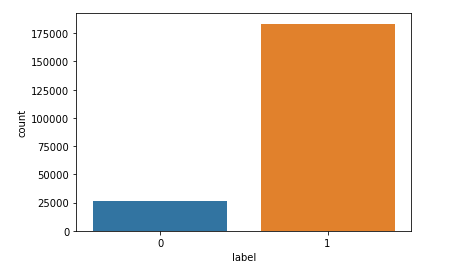
Here the black dots above and below the green coloured boxes represents outliers.

1. Skewness

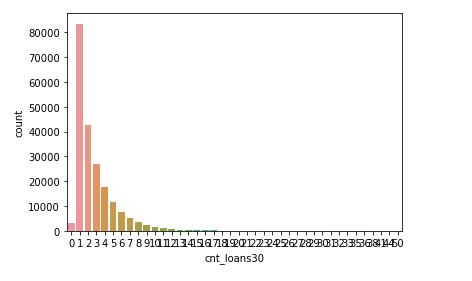


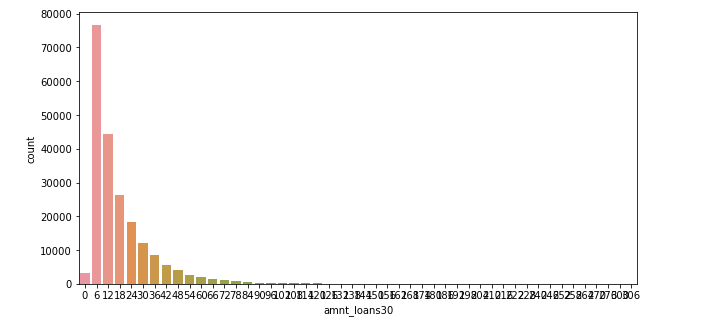
This is a skewness graph.We can see that almost all columns are skewed .This shows the presence of outliers.It can be resolved using log tyransformation method,boxcox method or yeo Johnson method

1. Univariate analysis

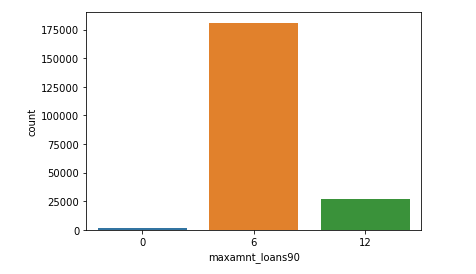


This is a boxplot showing count of label column.From this we can understand that there are almost 175000 non default customers and nearly 25000 default customers.

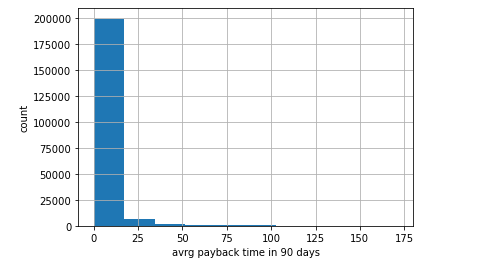
This is a countplot showing no. of loans taken by the customers. From this graph we can understand that there are customers with no loan history in last 30 days. Most of the customers are under one loan in a month category. The count of customers decrease with no.of loans. Also we can see that there is no customers who have taken more than 9 or 10 loans.



This is the countplot showing amount of loans and no.of customers who have taken them. Here maximum customers falls under Rs.6 loan. Nearly 80000 people are there under this category. Second highest loan amount is Rs.12, nearly 50000 customers comes under this.AS amount increases, no. of customers taken also decreases. No. of customers almost reaches zero when the loan amount is 80 and more.

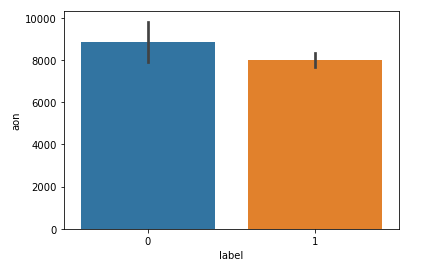


This box plot shows count of maximum amount of loans taken in last 90 days. It is clear from the graph that around 180000 people have taken Rs.6 loan. Around 25000 customers have taken Rs.12 loan and very few have not taken any loan.

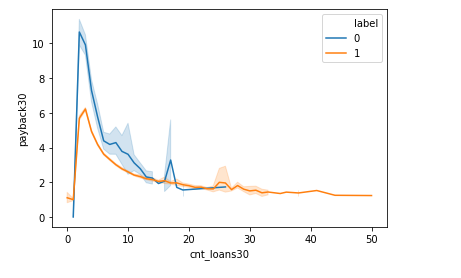


This is a histogram showing average payback time taken .Majority class ranges between 0 and 15.Around 200000 customers have paid the loan back within 0 to 15 days.Very less have taken 20 to 30 days and very few people have taken 50 to 100 days.

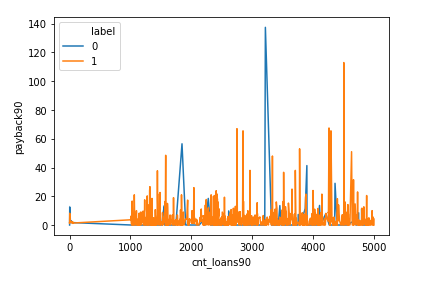
1. Bivariate Analysis



This is a bar plot of label vs age on network. This shows that more no.of defaulters are old customers than non-defaulters.



This is a line plot of no.of loans taken in 30 days vs payback in 30 days.We can see that defaulters are more than non-defaulters. Most defaulters have taken no. of loans between 0 and 7.Also they have taken maximum days to return it.Non defaulters have paid back their loans within 6 days.



This is a line plot of no.of loans taken in 90 days vs payback 90.Here maximum defaulters are in the range 3200 and 3500.They have taken a maximum of 140.

* Interpretation of the Results

From this analysis we can interpret that non default customers are more than default customers for this particular cellular network under this circle. only few default customers are there when compared to total no.of users in this network. Above average people have taken one or two loans. And also above 50% of loans have gone for Rs.6 category, which is the minimum amount. So, we can interpret that there is no fraud cases. Maximum payback time of customers are below 15 days. And number of loans taken by the defaulters are below 8. That means we understand that loans may be taken at crucial situations like, no money to recharge, forgot to recharge, no recharge shops nearby and so on.

**CONCLUSION**

* Key Findings and Conclusions of the Study

From this data analysis it is understood that the age on network doesn’t play a main role predicting the defaulters and non-defaulters. Factors which affect mainly are no.of recharges done ,amount of recharge done and amount of loans taken. And factors which affects the prediction least are frequency of recharge and median of account balance. We can conclude from this analysis that default customers are only minority class when compared to total network users. Also, major customers have taken only 1 or 2 loans in a month and also, they have paid back within 10 days. Since the due date to pay back the loan is before 5 days, they become default customers. Maximum number of loans taken by defaulters are below 8. So, we can identify a default customer by checking his recharge history, loan payback history, and also his main account balance before recharge.

Since major the customers have paid back the amount within 10 days and most of them have taken only one loan, There may be no fraud cases. Loan may be taken at crucial situations like no recharge shops nearby, no money to recharge, or due to travelling.So if the MFC’s increase the due date to payback loan amount, we can get more customers and also will have a large decrease in number of defaulters.

* Learning Outcomes of the Study in respect of Data Science

I have tried various models for this dataset. They are logistic regression,gaussianNb ,KNN and decision tree classifer. From this decision tree classifier gives a maximum score of 0.95. cross\_val\_score of DTC is checked and it gives a result of 0.95. Then to boost the result, Random forest classifier of ensemble technique is used. And my score has raised to 0.99. again cross\_val\_score for this is checked and gives a score of 0.97. Since it is a best score, I suggest this model for my dataset and thus saved random forest classifier as my model.

* Limitations of this work and Scope for Future Work
* Execution of programme was not easy as it took a lot of time for execution.
* Data visualisation also took a lot of time and so was unable to visualise each column and check relation between the columns as planned.
* There were also a lot of problems during each phase of the project, they were all resolved by searching webpages and also with the help of mentor and datatrained support team.
* Some of the url’s which made possible for me to do this project are listed below:

**REFERENCE**

1. <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>
2. <https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/>
3. <https://cxl.com/blog/outliers/>
4. <https://elitedatascience.com/imbalanced-classes>
5. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html>
6. <https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/>
7. <https://www.google.com/search?q=variance+inflation+factor&oq=variance+inflation+factor&aqs=chrome..69i57j0l7.13678j0j7&sourceid=chrome&ie=UTF-8>
8. <https://www.google.com/search?q=multicollinearity+in+regression&oq=muticollinearity&aqs=chrome.2.69i57j0i10i433j0i10l6.7595j0j7&sourceid=chrome&ie=UTF-8>
9. https://imbalanced learn.org/stable/generated/imblearn.over\_sampling.RandomOverSampler.html
10. <https://www.google.com/search?sxsrf=ALeKk02HmEy5i9YItlkcqpojW8oOavLcIg%3A1613537096613&ei=SJ8sYJD3JKyc4-EPkJS4oAI&q=yeo+johnson+transformation&oq=YEO&gs_lcp=Cgdnd3Mtd2l6EAEYATIECCMQJzIFCAAQkQIyBwgAELEDEEMyCgguELEDEIMBEEMyBwguELEDEEMyBAgAEEMyBwguELEDEEMyBwguELEDEEMyBQguELEDMgUIABCxAzoICAAQsQMQgwE6AggAULabT1ibpE9gsMFPaABwAngAgAGXAogB-AWSAQUwLjMuMZgBAKABAaoBB2d3cy13aXrAAQE&sclient=gws-wiz>