MICRO CREDIT DEFAULTER

Submitted by:

ASHIKA YASMEEN

ACKNOWLEDGMENT

It gives me a great pleasure in presenting the project report on Customer Retention Using Machine Learning.

Sources:

Google.com

Wikipedia

INTRODUCTION

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

Business Problem:

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).

Background:

Machine learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually solve big data would be difficult without the support of machines. Machine learning attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is based on creating algorithms that permit the machine to learn. Machine learning is categorized into two groups, which are supervised and unsupervised.

Supervised machine learning is an approach that is defined by its use of labelled datasets. These datasets are designed to train algorithms into classifying data or predicting outcomes accurately. Supervised learning can be separated into two types, classification and regression.

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabelled datasets. These algorithms discover hidden patterns in data without any human intervention. Unsupervised machine learning models are used for three main tasks, clustering, association and dimensionality reduction.

The performance will be measured upon predicting which of the Indian retailer is most recommended to friends. Since the prediction in many classification algorithms relies not only on a specific feature but on an unknown number of attributes that result in the value to be predicted. E-commerce has different number of features that may result in customer retention.

The data used in this project will be handled by using a combination of pre-processing methods to improve the predictions accuracy.

Literature:

Literature survey is the most important in any kind of research. Most of the literature study is based on articles from websites. This literature study attempts to construct a model based on classification techniques and how it can be applied to predict house price.

The literature study gives an overview of the articles that are related to this study, the feature engineering methods and the evaluation metrics that are used to measure the performance of the algorithms.

This paper presents how to make optimal use of Logistic Regression, Decision Tree Classifier and Random Forest Classifier. The system will give effective result on knowing the causes for customer retention.

Motivation:

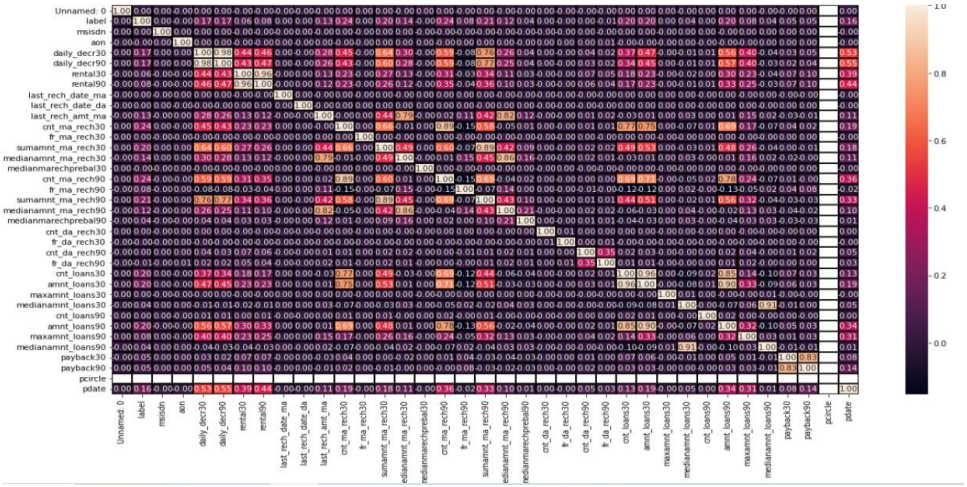
To understand the importance of communication and how it affects a person's life, thus, focusing on providing their services and products to low-income families and poor customers that can help them in the need of hour.

ANALYTICAL PROBLEM FRAMING

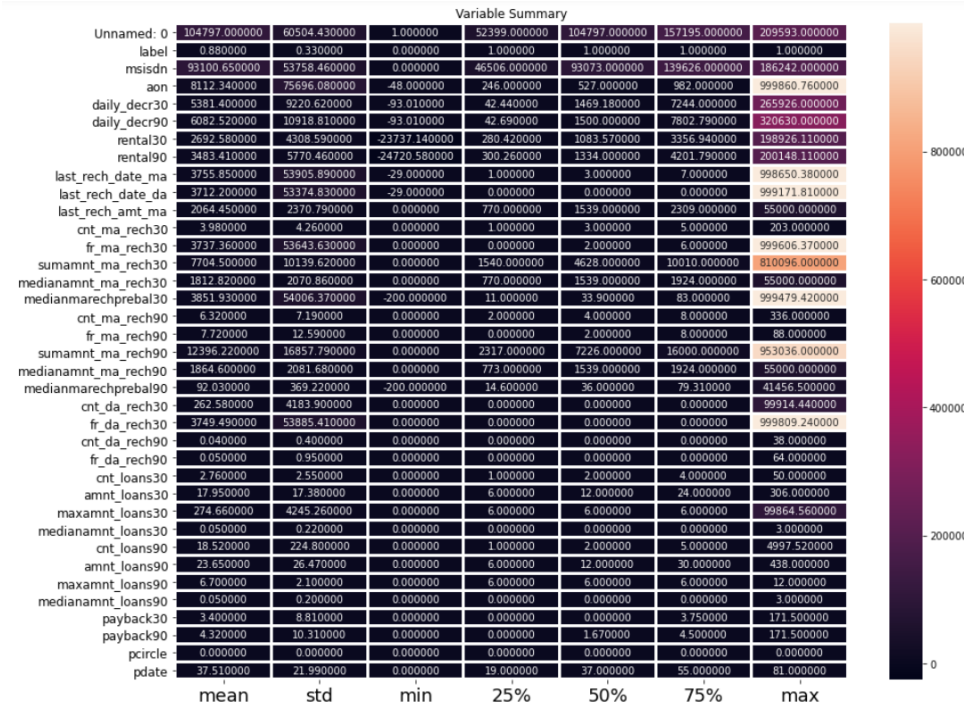
Mathematical/Analytical Modelling:

This experiment is done to pre-process the data and evaluate the prediction accuracy of the models. The experiment has multiple stages that are required to get the prediction results. These stages can be defined as:

- Pre-Processing: The dataset will be checked and pre-processed using certain methods. These methods have various ways of handling data. Thus, the pre-processing is done on multiple iterations where each time the accuracy will be evaluated with the used combination.
- Data splitting: Dividing the dataset into two parts to train the model with one and use the other in the testing. The dataset will be split 75% for training and 25% for testing.
- Evaluation: The accuracy of both datasets will be evaluated by measuring the accuracy score, confusion matrix and classification report when training the model with an evaluation of the on the test dataset with that are being predicted by the model.
- Correlation: Correlation analysis defines the strength of a relationship between two variables, which can be between two independent variables or one independent and one dependent variable. Correlation between the available features and output will be evaluated to identify whether the features have a negative, positive or zero correlation with the output variable.



- Describe of the dataset: Describe of the dataset is plotted using the heatmap which shows us the mean, standard deviation, maximum and minimum value of each column in the given dataset.



- Evaluation Metrics: The prediction accuracy will be evaluated by measuring the accuracy score, confusion matrix and classification report. Accuracy score is the ratio of number of correct predictions to the total number of input samples. It works well only if there are equal number of samples belonging to each class. A confusion matrix is an NxN matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 score and support of our trained classification model.
- Outliers are datapoints that are far away from other similar points. They may be due to variability in the measurement or may be due to experimental errors. Outliers should be excluded from the dataset in order to get the efficient and accurate prediction.

Methods to remove outliers:

1. Z-score: Call scipy.stats.zscore() with the given data-frame as its argument to get a numpy array containing the z-score of each value in a

data-frame. Call numpy.abs with the previous result to convert each element in the data-frame to its absolute value.

2. Interquartile (IQR) range: It can be used to remove outliers present in the dataframe. IQR can be calculated using scipy.stats.iqr module.

- Skewness: Skewness is a measure of the asymmetry of the probability distribution of a real valued random variable about its mean.

Skewness:

```
cred.skew()
```

```
Unnamed: 0                0.000000
label                    -2.270254
msisdn                    0.000719
aon                      10.392949
daily_decr30              3.946230
daily_decr90              4.252565
rental30                  4.521929
rental90                  4.437681
last_rech_date_ma        14.790974
last_rech_date_da        14.814857
last_rech_amt_ma          3.781149
cnt_ma_rech30             3.283842
fr_ma_rech30             14.772833
sumamnt_ma_rech30         6.386787
medianamnt_ma_rech30      3.512324
medianmarechprebal30     14.779875
cnt_ma_rech90             3.425254
fr_ma_rech90              2.285423
sumamnt_ma_rech90         4.897950
medianamnt_ma_rech90      3.752706
medianmarechprebal90     44.880503
cnt_da_rech30            17.818364
fr_da_rech30             14.776430
cnt_da_rech90            27.267278
fr_da_rech90             28.988083
cnt_loans30               2.713421
amnt_loans30              2.975719
maxamnt_loans30          17.658052
medianamnt_loans30        4.551043
cnt_loans90              16.594408
amnt_loans90              3.150006
maxamnt_loans90           1.678304
medianamnt_loans90        4.895720
payback30                 8.310695
payback90                 6.899951
pcircle                   0.000000
pdate                     0.116409
dtype: float64
```

Data Description:

In Machine Learning, the training data set is the actual dataset used to train the model for predicting the customer retention.

There are 209593 rows and 37 columns.

The column consists of details of users like

label: Flag indicating whether the user paid back the credit amount within 5 days of issuing the loan{1:success,0:failure}

msisdn: mobile number of user

aon: Age on cellular network (in days)

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 30 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

last_rech_amt_ma: Amount of last recharge of main account (in Indonesian Rupiah)

cnt_ma_rech30: Number of times main account got recharged in last 30 days

fr_ma_rech30: Frequency of main account recharged in last 30 days

sumamnt_ma_rech30: Total number of recharges in main account over last 30 days (in Indonesian Rupiah)

medianamnt_ma_rech30: Median of amount of recharges done in main account over last 30 days at user level (in Indonesian Rupiah)

medianmarechprebal30: Median of main account balance just before recharge in last 90 days (in Indonesian Rupiah)

cnt_ma_rech90: Number of times main account got recharged in last 90 days

fr_ma_rech90: Frequency of main account recharged in last 90 days

sumamnt_ma_rech90: Total number of recharges in main account over last 90 days (in Indonesian Rupiah)

medianamnt_ma_rech90: Median of amount of recharges done in main account over last 90 days at user level (in Indonesian Rupiah)

medianmarechprebal90: Median of main account balance just before recharge in last 90 days (in Indonesian Rupiah)

cnt_da_rech30: Number of times data account got recharged in last 30 days

fr_da_rech30: Frequency of data account got recharged in last 30 days

cnt_da_rech90: Number of times data account got recharged in last 90 days

fr_da_rech90: Frequency of data account got recharged in last 90 days

cnt_loans30: Number of loans taken by user in last 30 days

amnt_loans30: Total amount of loans taken by user in last 30 days

maxamnt_loans30: maximum amount of loan taken by the user in last 30 days

medianamnt_loans30: median amounts of loan taken by the user in last 30 days

cnt_loans90: Number of loans taken by user in last 90 days

amnt_loans90: Total amount of loans taken by user in last 90 days

maxamnt_loans90: maximum amount of loan taken by the user in last 90 days

medianamnt_loans90: median amounts of loan taken by the user in last 90 days

payback30: Average payback time in days over last 30 days

payback90: Average payback time in days over last 90 days

pcircle: telecom circle

Data Pre-Processing:

Data pre-processing is an important step in machine learning to get highly accurate and reliable result. Data pre-processing helps in increasing the quality of data by filling in missing data's(NaN values), removing outliers, scaling the data.

There are many steps involved in data pre-processing:

- Data Cleaning helps to impute the missed values and removing outliers from the dataset.
- Data Integration integrates data from multiple sources into single dataset.
- Data Transformation such as normalization helps in improving the accuracy and efficiency of algorithms involved in machine learning.
- Data Reduction reduces the data size by dropping out redundant features using feature selection and feature extraction techniques.

Treating null values

Sometimes there can be certain columns which may contain the null values used to indicate the missing or unknown values. In our dataset there are no null values present.

Converting labels into numeric

In machine learning, we usually deal with datasets which contain multiple labels in one or more column. These labels can be in the form of alphabets or numbers. To make the data understandable or in human readable form, the training data is often labelled in words.

In our dataset three columns have categorical values. These columns are converted using Label Encoder.

Label Encoder refers to converting the labels into numeric form so as to convert it into the machine-readable form. It is an important step in data pre-processing.

Label encoding in python can be imported from Sklearn library. Sklearn provides a very efficient tool for encoding.

Label Encoding:

```
from sklearn.preprocessing import LabelEncoder

le=LabelEncoder()

cred['msisdn']=le.fit_transform(cred['msisdn'])
cred['pcircle']=le.fit_transform(cred['pcircle'])
cred['pdate']=le.fit_transform(cred['pdate'])

cred
```

Input-Output Relationship:

The dependent variable or target or output depends on the features or input variables given in the dataset.

Tools Used:

Hardware: The needed time to train the model depends on the capability of the used system during the experiment. Some libraries use GPU resources over the CPU to take a shorter time to train a model.

Operating System	Windows 10
Processor	CORE i3
RAM	16GB

Language: Python

- Python is widely used in numeric and scientific computing.
- Scipy is a collection of packages for mathematics, science and engineering.

- Pandas is a data analysis and modelling library

Libraries:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scikit Learn

MODEL DEVELOPMENT AND EVALUATION

Model:

Classification model is used. Classification refers to a predictive modeling problem where a class label is predicted for a given set of input data. It is a supervised technique. Regression models a target prediction based on independent variables.

Logistic Regression:

Linear regression is a

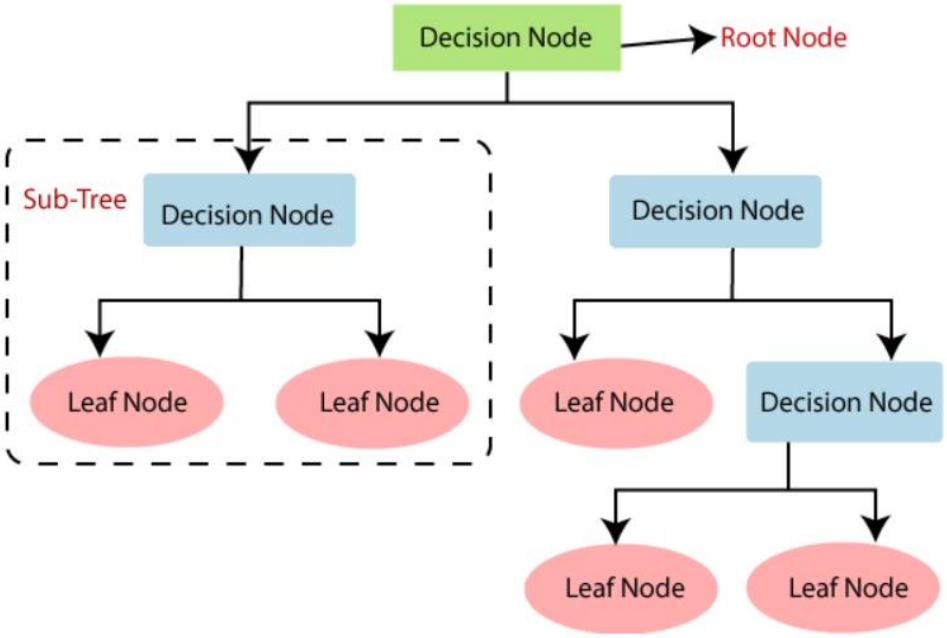
learning algorithm based on supervised learning. Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. It is most commonly used when the data has binary input, so when it belongs to one class or the other, or is either a 0 or 1.

Decision Tree Classifier:

It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

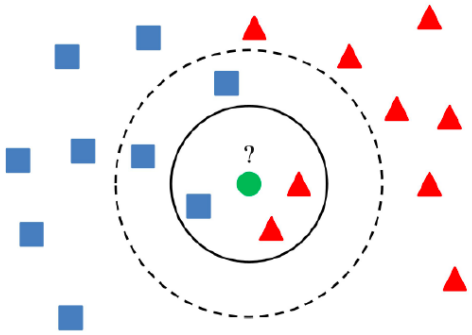
In a decision tree, there are two nodes, which are the decision node and leaf node. Decision nodes are used to make any decision and have multiple branches, whereas leaf nodes are the output of those decisions and do not contain any further branches.

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



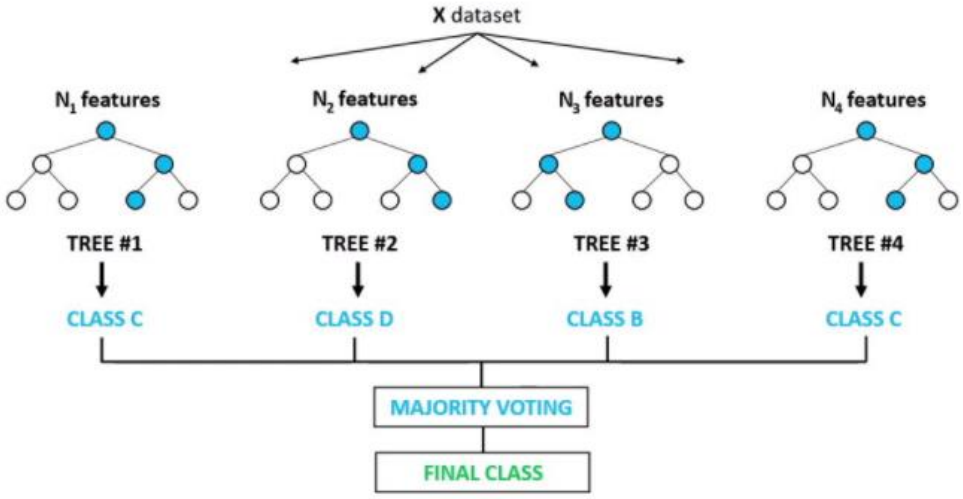
Neighbors Classifier:

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve classification problems. It's easy to implement and understand, but has a major drawback of becoming significally slow as the size of the data in use grows. It stores all available cases and classifies new cases based on a similarity measure. KNN has been used in statistical estimation and pattern recognition as a non-parametric technique.



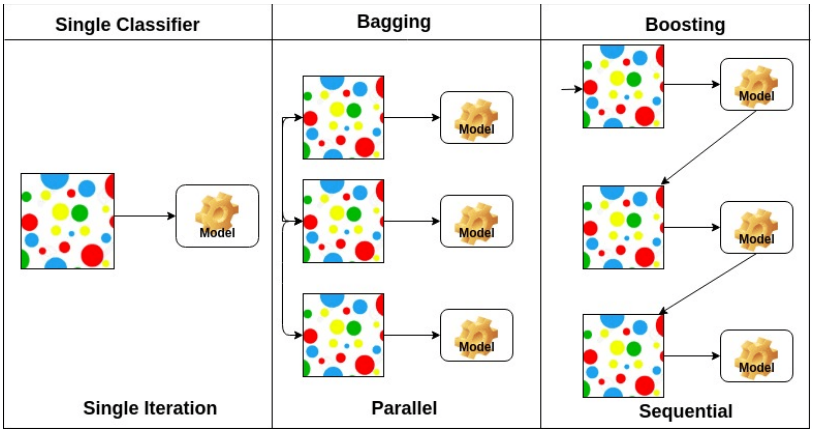
Random Forest Classifier:

It is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each individual tree and try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



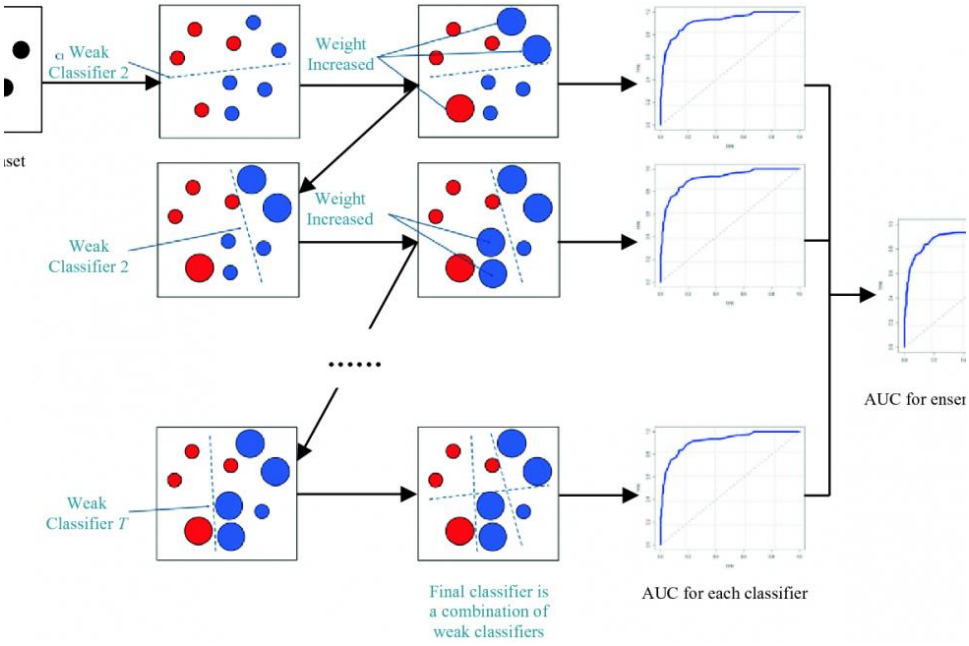
Ada Boost Classifier:

An Ada boost or Adaptive boosting is one of the ensemble boosting classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.



Gradient Boosting Classifier:

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting.



Voting Classifier:

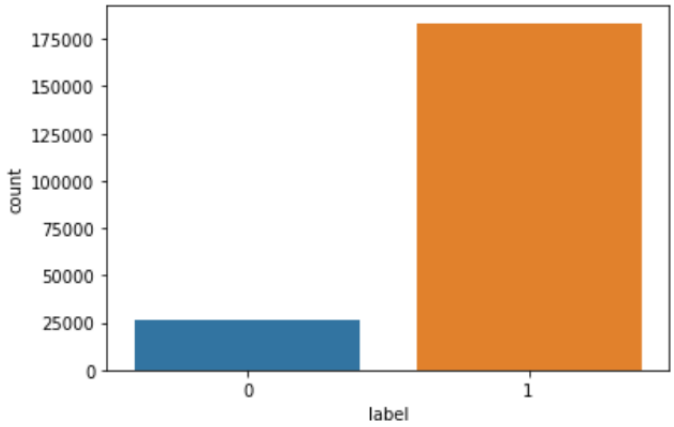
A voting classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output based on their highest probability of chosen class as the output. It simply aggregates the findings of each classifier passed into voting classifier and predicts the output class based on the highest majority of voting. Instead of creating separate models and finding the accuracy for each of them, we create a single model which trains by these models and predicts output based on their combined majority of voting for each output class.

Algorithms Used:

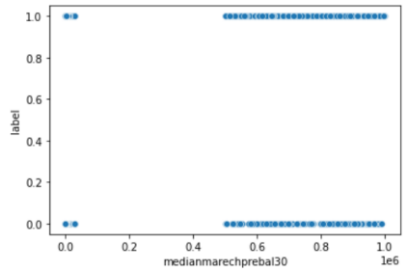
- Logistic Regression
- Decision Tree Classifier
- KNeighbors Classifier
- Random Forest Classifier
- Ada Boost Classifier
- Gradient Boosting Classifier
- Voting Classifier

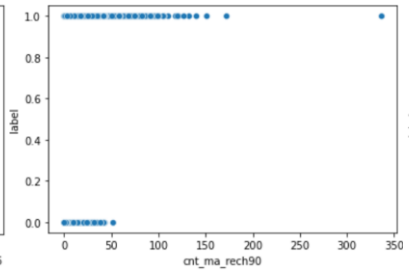
Visualizations:

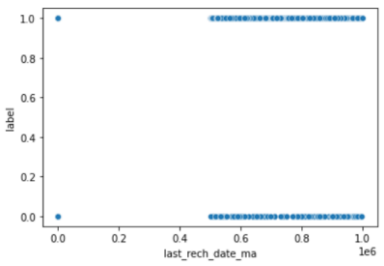
Count Plot:

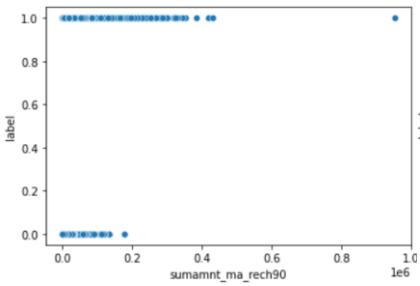


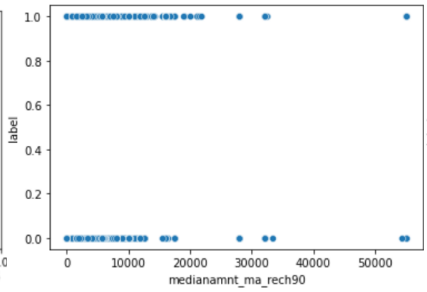
Scatter Plot:

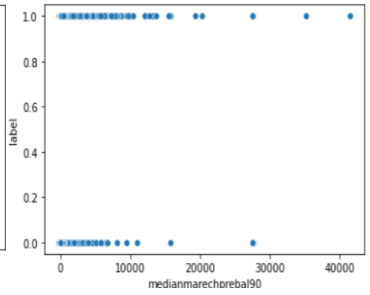

label
medianmarechprebal30
1e6

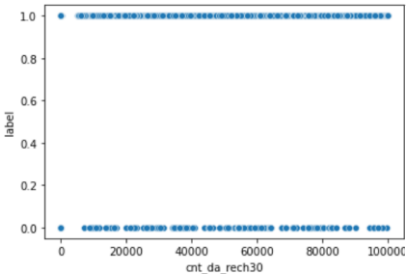

label
cnt_ma_rech90

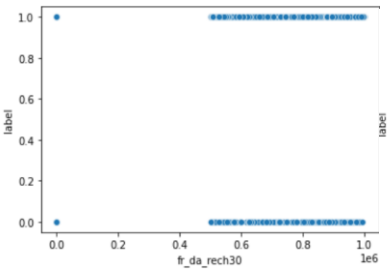

label
last_rech_date_ma
1e6

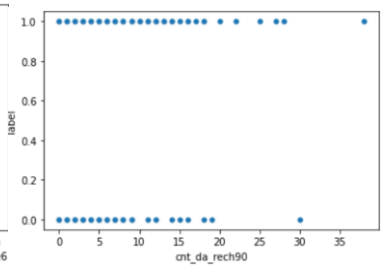

label
sumamnt_ma_rech90
1e6

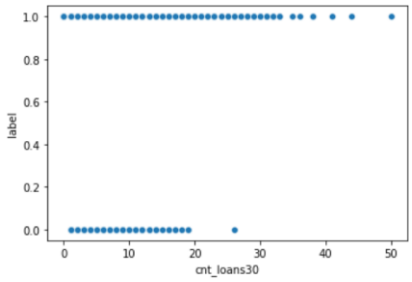

label
medianamnt_ma_rech90

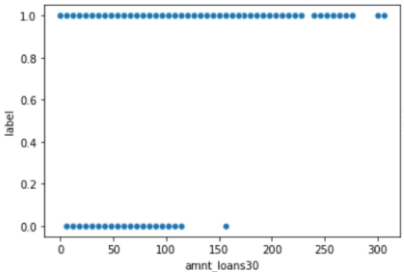

label
medianmarechprebal90

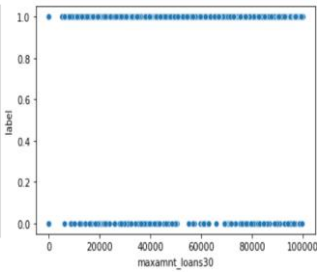

label
cnt_da_rech30

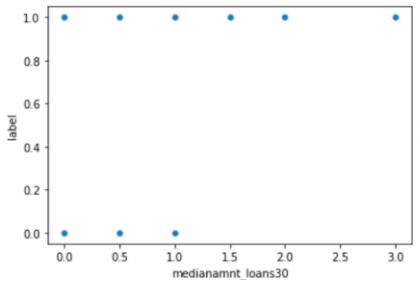

label
fr_da_rech30
1e6

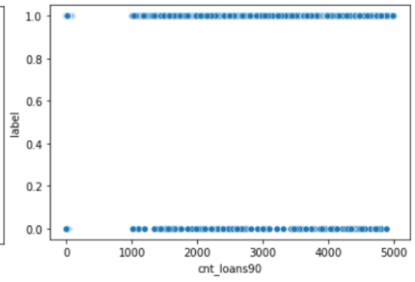

label
cnt_da_rech90

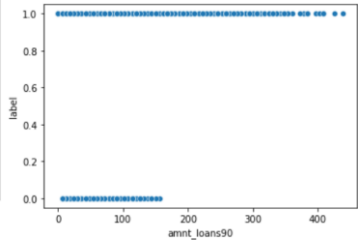

label
cnt_loans30

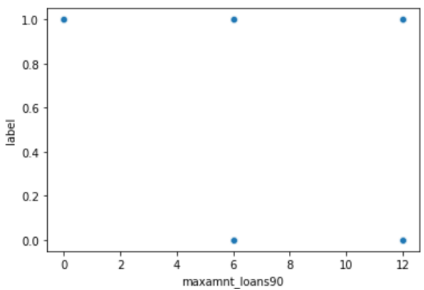

label
amnt_loans30

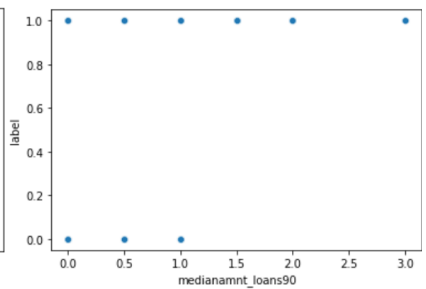

label
maxamnt_loans30

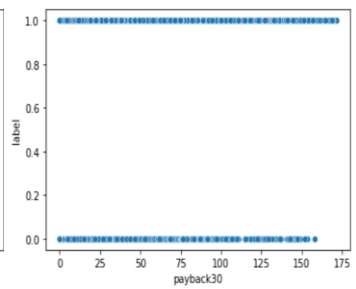

label
medianamnt_loans30

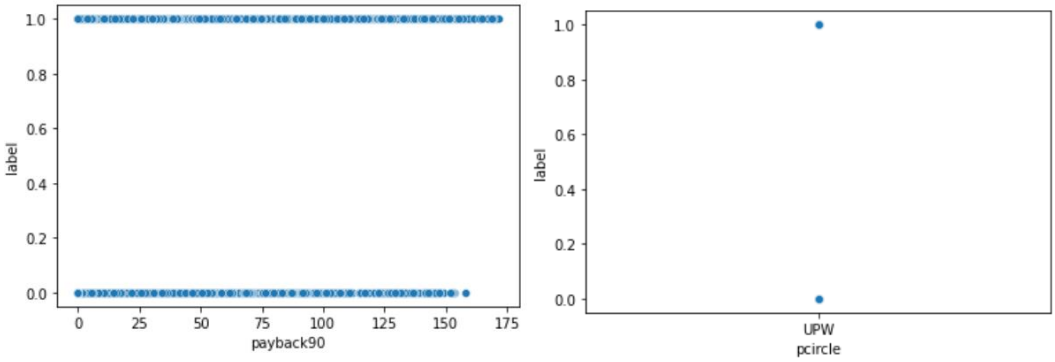

label
cnt_loans90


label
amnt_loans90


label
maxamnt_loans90


label
medianamnt_loans90


label
payback30



Interpretation of Results:

Many machine learning algorithms are used to predict. Using these algorithms is beneficial so that the result can be as near to the claimed results. However, the prediction accuracy of these algorithms depends heavily on the given data when training the model. If the data is in bad shape, the model will be overfitted and inefficient, which means that data pre-processing is an important part of this experiment and will affect the final results. Thus, multiple combinations of pre-processing methods need to be tested before getting the data ready to be used in training.

From Visualizations it is clear that how each feature from the dataset is relatable to the house price. In pre-processing, unnecessary feature is removed, outliers are removed and skewness from the input variables are removed. The data's are scaled and transformed for better training purpose. However, the chosen algorithm is Random Forest Classifier with 94.5% accuracy.

CONCLUSION

The objective of this research was to introduce relational learning method and to determine if it was better in default prediction over traditional prediction method. This research predicted the model with 94.5% accuracy and the chosen algorithm is Random Forest Classifier.

Limitations:

This study did not cover all the classification algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced techniques.