**Santander Customer Transaction Prediction**

Report By

**Kavin S**

[akkera2@gmail.com](mailto:akkera2@gmail.com)

9944573276

**Problem Statement**

Need to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

**Problem Description:**

At Santander, mission is to help people and businesses prosper. We are always looking for ways to help our customers understand their financial health and identify which products and services might help them achieve their monetary goals. Our data science team is continually challenging our machine learning algorithms, working with the global data science community to make sure we can more accurately identify new ways to solve our most common challenge, binary classification problems such as:

i)is a customer satisfied?

ii) Will a customer buy this product?

iii)Can a customer pay this loan?

**Software Requirements:**

1) conda version 4.8.3

2) Python Version 3.6.10 with required Libraries

3) R version 3.6.3 with required Libraries

**Steps involved to create Machine learning Model:**

**Data Exploration:**

It is the process of analysing thestructure of our data set it is also called as Exploratory data analysis and it is most important part Machine learning because based on the type of data in the dataset we decide the methods to analysis the data and appropriate model used to predict the output.

Steps to check the data;

1)Identify the predictor and target variable

2) Data type of variables and category of variables (Continuous or Categorical).

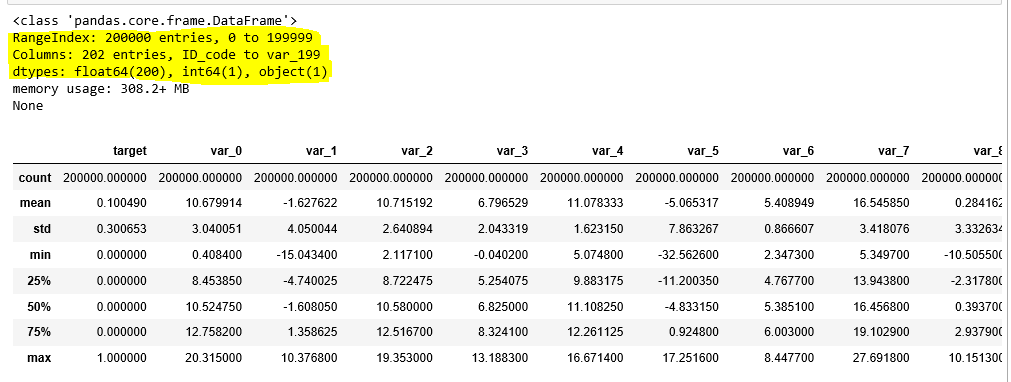
3)Based on the variables category we use different methods to identity relation between variables.

Checking on the training dataset provided, the dataset has target variable that need to be predicted so this project comes under supervised machine learning model. And the target variable has binary categorical value, we can use classification algorithm model to predict the target data.

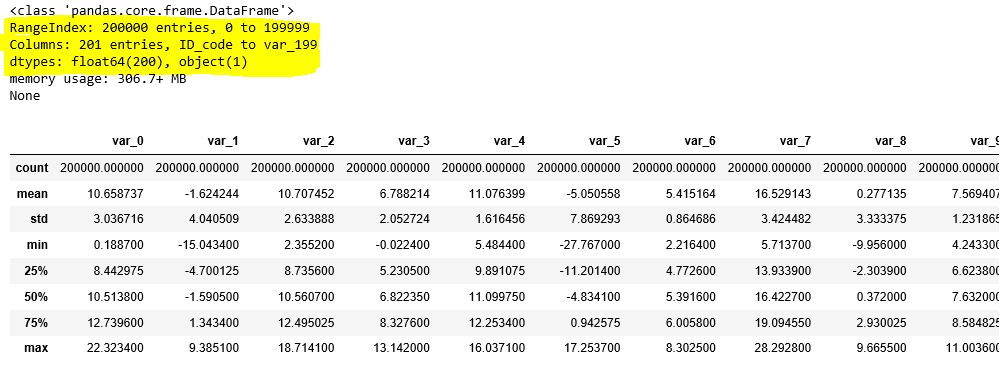
The same kind of attributed we received for test data set too.

Please find the structure of Training and Test data set below.

**Training Dataset:**



**Test Dataset:**



**Missing Value Analysis:**

Missing values in data set are the common scenario which affects the conclusion we get from data. Missing values should be properly handles based on the % and type of the data. We need to decide whether the missing values to be removed or imputed based on % of missing data by variable

i) If a variable has less than 30%, Missing values can be easily treated using various methods like mean, median method, KNN method to impute missing value.

To check suitable impute method (Mean, Mode or KNN) method, first delete the any value from particular index and apply all three methods, check which methods gives nearest value to that deleted value and use the method which generated nearest value

ii)If a variable has more than 30% of its values missing, then those variable columns can be ignored. In our case, none of the columns have a high percentage of missing values.

Both methods have its own pros and cons based on the data set we handle.

Here our dataset doesn’t have any missing values in both training and test .so we have not used any method to solve the same.



**Outlier Analysis:**

Outliers are the data points that were stand out from rest of the data distribution in the data set. The easiest way to detect outliers is to create a graph. Plots such as Box plots, Scatterplots and Histograms can help to detect outliers. In this project I have used boxplot to visualize the outliers.

Box plot is convenient method that displays the distribution of data through their quartile by means of

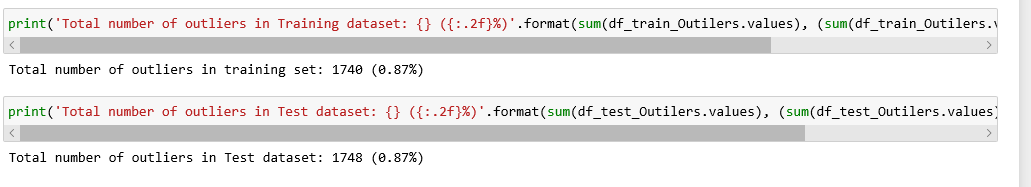
1. minimum value
2. Lower quartile
3. median
4. Higher quartile
5. Highest value

**Advantages of choosing Box plot:**

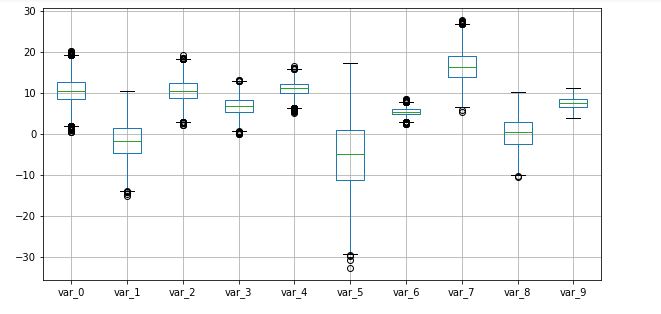
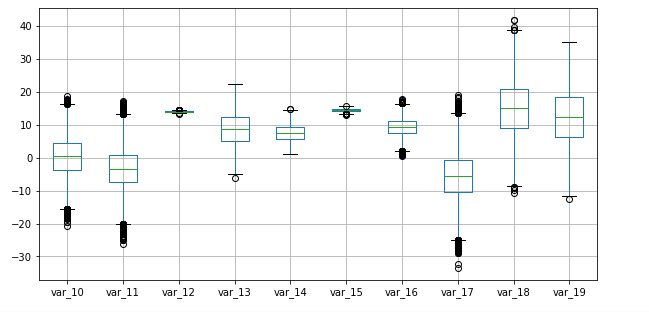
1. It will handle the large dataset by organizing data using above five key concepts which is not possible by other graphs.
2. It shows the simple summary of the distribution of results in graphical manner so that you can quickly view it and compare it with other data.
3. Most important thing, It will identify and display the outliers, which can be easy to determine and to remove those values.

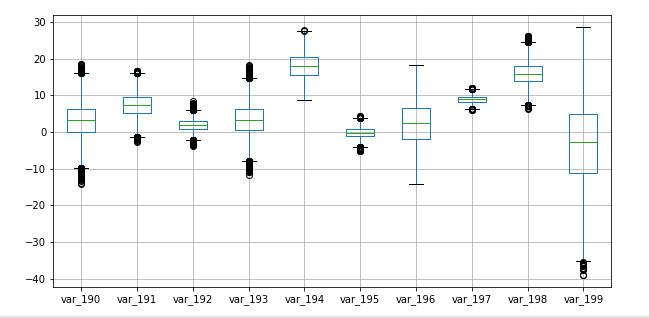
In this project we have outliers in both training and test dataset

**Outliers % in Training and Test Dataset**:



**Box plot created for Training Data set.**

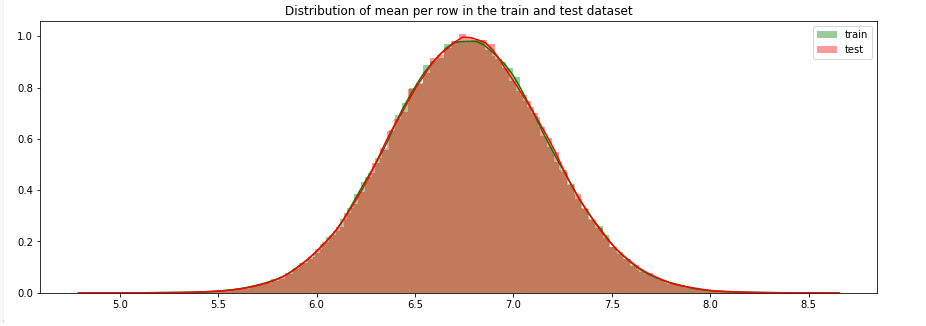
 

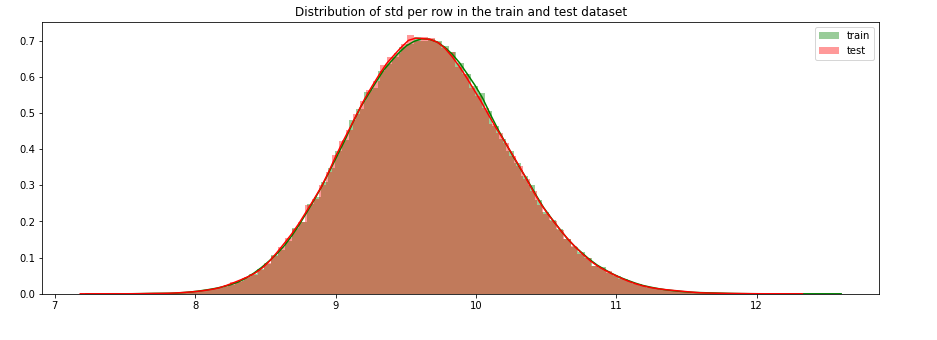


**Data Visualization:**

It is presentation of data in graphical format. Using visualization, we can resemble the data in various visuals like Charts, graphs, maps etc. It’s kind of storytelling, which will visually teach us about the trend’s patterns of the data.

Below distribution plot shows the mean and standard deviation of test and training dataset. And also we can see the data were almost normally distributed.





**Feature Selection:**

Feature selection is the process of reducing the independent variable which carries same information about the dependent variables. If we failed to remove those duplicate data, then it will affect the performance of our model.

Reducing the no of input variables will improve the computation cost and model performance.

The relationship between variable are considered as below.

1. Relation between two independent variables should be low
2. Relation between independent and dependent variables should be low

As our dependent variables are numerical, we can Correlation coefficient method to identify the relation between dependent variables.

**Positive Correlation**: Means that if feature A increases then feature B also increases or if feature A decreases then feature B also decreases. Both features have a linear relationship.

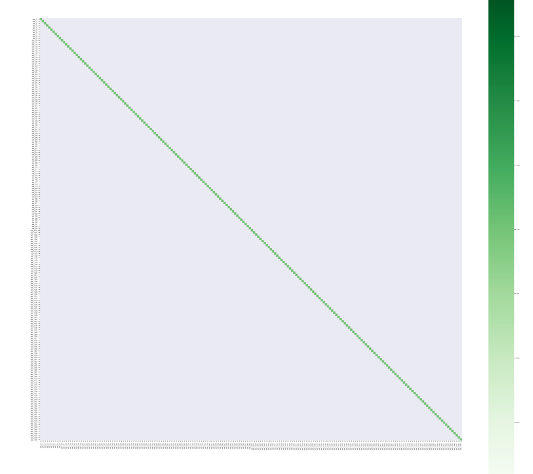
**Negative Correlation**: Means that if feature A increases then feature B decreases and vice versa.

**No Correlation**: No relationship between those two attributes.

If a Correlation coefficient value between two variables is greater than + -0.5 then we can remove any one variable that carries less information of target variable.

Here as far we checked the variables has less correlation in both train and test data set .so I didn’t remove any variables by this method.

Please find the correlation heat map generated for training data set. Dark blue represents the high correlation. Here we can see the dark blue colour only for same variables, other than our variables didn’t have correlation.



**Mutual Information Classifier:**

Mutual information is a measure between two (possibly multi-dimensional) random variables XX and YY, that defines the amount of information obtained about one random variable, through the other random variable.

it helps us to choose the right set of independent variables that carries the useful information about the target variables.so that we can neglect all the unwanted variable/noise from our dataset which will reduce the model performance.

It’s a kind of feature reduction process of reducing the number of variables from the dataset.

The Mutual Information value will not be negative and zero if the input and output variable are independent.

As our target variable is classification type I have used mutual information classifier technique to drop the independent variables that doesn’t have correlation with target class.

**In our training data set I have reduced the no of input variables from 199 to 152variables.**

**Handling Imbalanced Dataset:**

Imbalance data set is the special scenario we face in classification problem where the distribution of target class is not uniform. It frequently found in problem like fraudulent transactions in banks, identification of rare diseases, etc.

If we didn’t handle the imbalanced dataset which will create noise in training data set and in turn the model will have performance impact on new data.

We can apply resampling techniques to handle this imbalanced dataset.

#### Random Under-Sampling:

#### Random Under sampling aims to balance class distribution by randomly eliminating majority class examples. It improve run time and storage problems by reducing the number of training data samples if our dataset has millions of records.

#### But the disadvantage of using this method it can neglect the most useful data that will be used as one of the main rile in algorithm and the sample we choose may leads to biased output.

#### Random over-Sampling:

#### Over-Sampling increases the number of instances in the minority class by randomly replicating them in order to match the level of majority class. Unlike under sampling this method will not leads to information loss and it will outperform under sampling.

#### The main disadvantage of over sampling is that it will leads to overfitting as we replicate the minority class samples.

#### Synthetic Minority Over-sampling Technique:

#### In this method, a subset of data is pulled from minority class and new similar data were created like the minority class and make the minority count equal to majority class.so the main advantages of using this SMOTE technique will avoid overfitting of the model. Even this method has it own disadvantages as it will not perform well for high dimensional data.

#### We also have other advanced Ensemble techniques like bagging, boosting kind of methods to handle imbalanced datasets. We need to try it out different methods and based on accuracy parameters will finalize the technique.

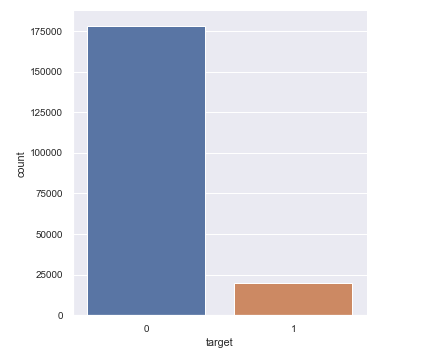
. Model accuracy is not of great importance in imbalanced datasets. In those cases, we look at other performance metrics like sensitivity, specificity , f1-score etc.

**False Positive(FP) or False Negative(FN) is important under which scenarios:**

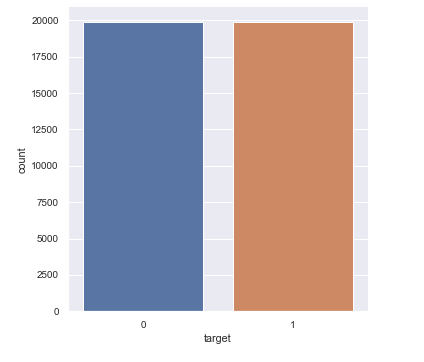
1) Spam detection, where false positives are important(FP to be reduced ). Reason being, its more problematic when model says your email is a spam but, it’s not. So, you will miss out on an imp one. Also, stock market crash, where the model predicted that the market is not going to crash, investor invests in loads of money, but it crashes and then the investor loses all of the money invested. False negative in stock market crash is a lost opportunity.

2). Examples where False negative is important(FN to be reduced ) are healthcare and risk compliance/sanction models. When model says the person doesn't have cancer, but, the person has cancer. So, the person will miss out on imp medications. Also, risk models like customer risk ranking model, where the model predicts the customer to be non-risky, but the customer is risky. And the consequence being, the bank allows such customers to do business with them and end up in a mess where they perform suspicious activity like money laundering.

Please find the class label count of our imbalanced training dataset



#### In this project I have used under sampling method to handle imbalance dataset and the count of majority class has been reduced to level of minority class.



**Feature Scaling:**

Feature scaling is a method used to normalize the range of independent variables or features of data.it is used to normalize the data in particular range irrespective of data in cm, inch, weight, ounces etc. We have two different scaling methods such as normalization and standardization.

 Normalization usually means to scale a variable to have a value between 0 and 1 , and it can be applied if the data is not in gaussian distribution .

**Normalization formula**=(Values-Min value)/(Max value-Min Value)

While the Data is Normally/Uniformly distributed then we can standardization that transforms data to have a mean of zero and a standard deviation of 1 ,i.e) process of grouping the maximum data near to Mean values and in the range of Std values of 1.

**z=(Value-Mean)/Standard Deviation**

**Classification model used:**

1. **Logistic regression:**

Logistic regression is a classification algorithm that used when the dependent variable is binary in nature. Independent variables can be numerical or categorical, but target variable should be Categorical.

Logistic regression performs logistic sigmoid function to return a probability value that works well if the target variable is either

* 1. binomial—Target variable with two classes i.e ) Pass/Fail
  2. Ordinal -- Target variable in order. i.e ) low, medium, high
  3. Multinomial-- Target variable has more than two classes  i.e ) Cats, Dogs, Sheep)

**Outputs of logistic Regression:**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **0** | **1** |
| Accuracy | 76.94% | |
| ROC\_AUC Score | 76.94% | |
| Precision | 77 | 77 |
| Recall | 77 | 77 |
| F1 Score | 77 | 77 |

1. **Random Forest algorithm:**

Random forest is a supervised learning algorithm and it’s an ensemble technique that consists of many decision trees.

It will make the combination of weak learners to make the strong model. The main hypothesis is that when weak models are correctly combined, we can obtain more accurate and/or robust models. So I have tried out this algorithm expecting higher accuracy than Logistic regression.

**Advantage:**

It can be used for both classification and regression problem

It reduces misclassification error rate and improve the accuracy; we combine multiple decision tree to create a classifier, so it avoids overfitting of the model

**Disadvantage**:

As it will create large number of trees the real time prediction is too slow.

**Outputs of Random Forest:**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **0** | **1** |
| Accuracy | 76.73% | |
| ROC\_AUC Score | 76.76% | |
| Precision | 79 | 75 |
| Recall | 73 | 80 |
| F1 Score | 76 | 77 |

**3)XGBoost Algorithm:**

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance .  Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

It comes under ensemble techniques . Gradient boosting is an approach where new models are created that predict the residuals /errors of prior models and then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models.

**Outputs of XGBoost:**

|  |  |  |
| --- | --- | --- |
| **Metrics** | **0** | **1** |
| Accuracy | 76.53% | |
| ROC\_AUC Score | 76.92% | |
| Precision | 76 | 77 |
| Recall | 77 | 76 |
| F1 Score | 77 | 76 |

**Performance metrics for Classification model:**

Performance of a model is most important part in machine learning and the choice of metric completely depends on the type of model and the implementation plan of the model.

**1)Confusion Matrix:**

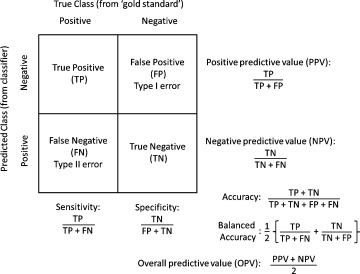
**Four important terms:**

TP--True Positive--Both actual and predicted values are YES

FN--False Negative--actual=YES and predicted=NO

FP--False Positive--actual=NO and predicted=YES--Predicted positive as falsely

TN--True Negative--actual=NO and predicted=NO



* **Accuracy :** Proportion of the total number of predictions that were correct.



* **Positive Predictive Value or Precision** : Proportion of positive cases that were correctly identified.



* **Negative Predictive Value** : Proportion of negative cases that were correctly identified.



* **Sensitivity or Recall** : Proportion of actual positive cases which are correctly identified.



* **Specificity :** Proportion of actual negative cases which are correctly identified.



**ROC – AUC Curve:**

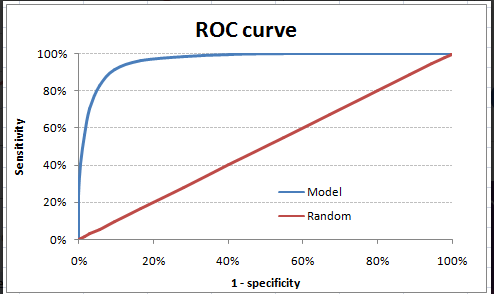
The **Receiver operating characteristic (ROC) curve** is plotted with true positive rate (TPR) versus the false positive rate (FPR) where TPR is on y-axis and FPR is on the x-axis Which graphically displays cut -off value between TPR and FPR that separates the data from noise.

**Area under the curve (AUC):** AUC provides an aggregate measure of performance across all possible classification thresholdswhich perfectly distinguish between all the Positive and the Negative class points correctly.

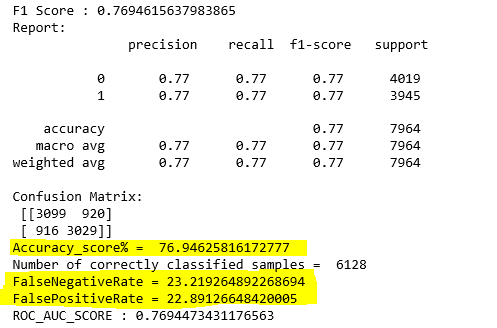
• **True Positive Rate (Sensitivity):** True Positive Rate is defined as TP/ (FN+TP). Proportion of positive cases that were correctly identified.

• **False Positive Rate (Specificity):** False Positive Rate is defined as FP / (FP+TN). Proportion of actual negative cases which are correctly identified

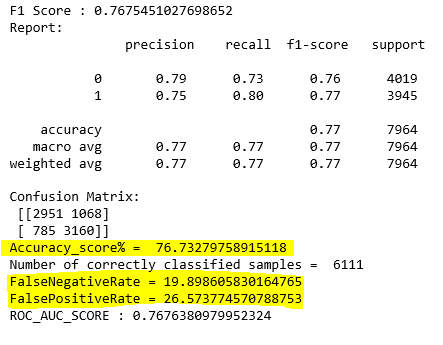
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



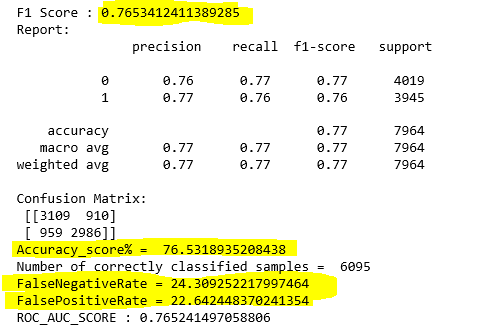
**Logistic Regression Classifier Output:**



**Random Forest Classifier Output:**



**XGboost Classifier output:**



On checking the accuracy and other parameter of training data set we got **77.9 %** of **accuracy** using Logistic regression model and more over **77.7% accuracy** while using Random forest algorithm.

**Conclusion:**

Predicting the Santander customer prediction dataset is very challenging as it got imbalanced target class.

As dataset has biased target data, I gave equal importance to both Accuracy and Precision (Positive Predicted value) as its values should be higher and less FPR( False positive Rate).

In this customer prediction, **FPR** should be lower as possible, because if our model falsely predicted the output as 1 (Yes he will make transaction) instead of 0(No).

Based on our prediction that our customer will transact in future, client will fail to do necessary steps to retain customer to make future transactions and eventually there is chance of losing this customer.

As both logistic regression and Random Forest model boost gave as same accuracy percentage, we can choose any model to feed the test data to predict the target. In my code I feed the test data in logistic regression model.

**Understanding of this Project:**

Based on our model prediction, client will have confident that the target variable with 1 will do transaction in future and for the value 0 ,client will take further steps/follow-ups on the customers to make them transact in future by giving additional offers on products.

So based ML model predictions ,client will take necessary actions and strategies to improve the business and increase their revenue.