Santander Customer Transaction Prediction

By Ipsita Sahu

Introduction

Santander bank data set has been given to predict which customers will make a specific transactions in the future , irrespective of the amount of money transacted so that personalized and timely manner service can be provided to that potential customers.

Introduction to data

Both train and Test dataset have same number of samples 20000. Train data set has ID_ code (identifier for the transaction or customer),202 training features and Target which is a label feature for training. Test data set has ID_ code (identifier for the transaction or customer),202 testing features same as training features. All the feature in both the dataset are float. The dependent variable in the train data is integer which is converted into categorical variable. The csv format files are used in both R and Python.

Missing Value Analysis:-

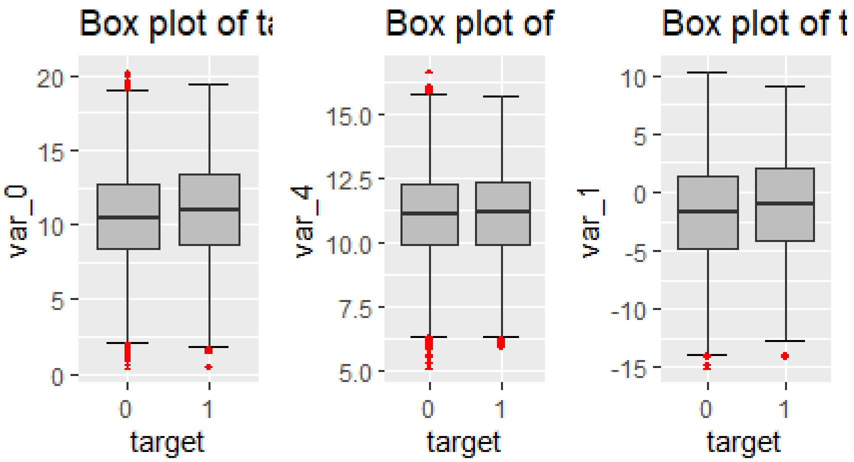
Imputation simply means replacing the missing values with an estimate, then analyzing the full data set as if the imputed values were actual observed values. The best method to impute missing value for a data are imputation using mean, median and KNN method. We can calculate the mean/median of the non-missing values in a column and then replacing the missing values within each column separately. In KNN (distance based method) imputation find the nearest neighbor based on the existing attribute and then find the Euclidean distance. In KNN it will try to find the distance score between missing value and observation and select those observation which are close to the missing value and impute that value.

As per the above analysis, train set do not have any missing value. So I didn't need to perform any imputation for handling the missing values.

Outlier Detection:-

An outlier is a data point that differs significantly from other observations. Box plot diagram is a graphical method typically depicted by quartiles and inter quartiles that helps in defining the upper limit and lower limit beyond which any data lying will be considered as outliers. The purpose of this diagram is to identify outliers and discard it from the data series before making any further observation so that the conclusion made from the study gives more accurate results not influenced by any extremes or abnormal values.

In our case, I have identified the outliers from the series of columns and replaced them with NAs in order to save the information and then calculated the mean of the non missing values in the columns and then replace the NAs with the mean of the respective columns in R and I have removed the outliers in Python environment for some experimental purpose.

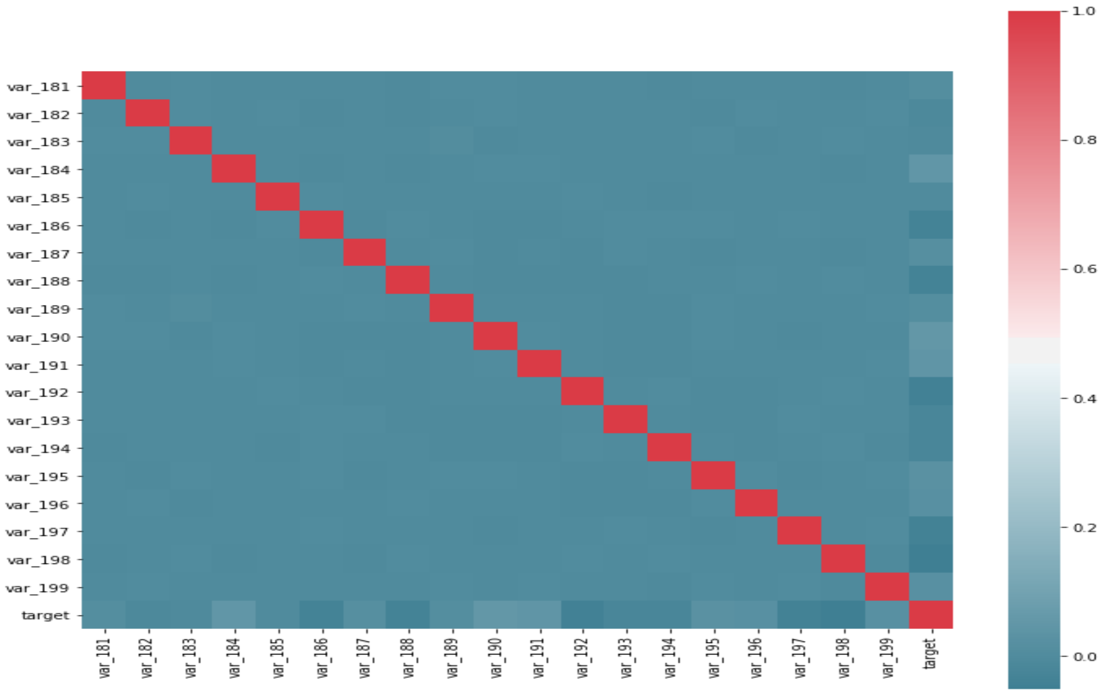


(Fig1.Boxplot diagram of independent variable of train data)

Feature Selection:-

In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant.

Visualizaiton of data relationship, will help us to understand the degree of correlation between features and the dependencies. With the help of heat map we can show the relationship of features.



(Fig2. Heat map showing the correlation between dependent and independent feature)

Correlation –

1. It show whether and how strongly pairs of variables are related to each other.
2. Correlation takes values between -1 to +1, wherein values close to +1 represents strong positive correlation and values close to -1 represents strong negative correlation.
3. Value 0 represents no relation between variables.

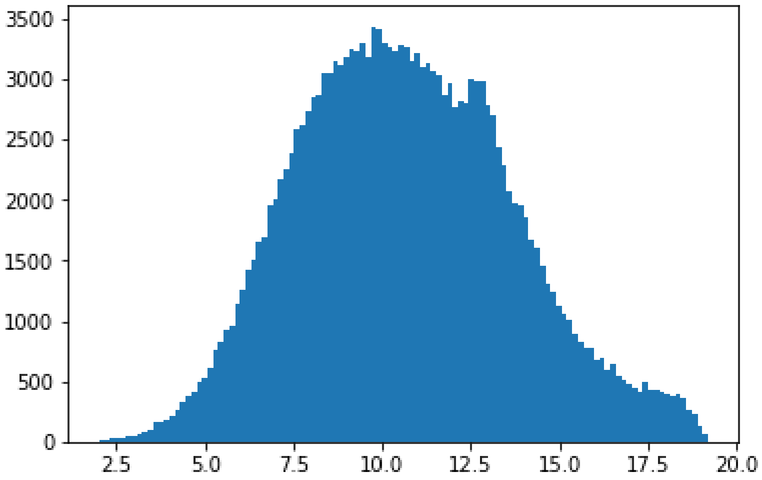
Since all the independent variables are of numeric type no need to perform chi square test as it is only applicable if there is any categorical variable in the dataset.

Feature scaling:-

When you're working with a learning model, it is important to scale the features to a range which is centered around zero. It is done to bring the feature on a same scale. This is done so that the variance of the features are in the same range. If a feature's variance is orders of magnitude more than the variance of other features, that particular feature might dominate other features in the dataset, which is not something we want happening in our model.

Before choosing the data scaling method, we need to check the distribution of data. If the data is uniformly distributed, then Standardization is the suitable method for the scaling purpose. On the other hand, if the data is not normally distributed, we go with Normalization scaling method. Normalization is bringing all the variables into proportion with one another. Normalization is the process of reducing unwanted variation either within or between variables and the range lies between 0 and 1.

From the below visualization it is clearly seen that the data is uniformly distributed or on the other hand you can say it is a normalized dataset. So I went for standardization I the next step.



Model Development

<u>1.Logistic Regression:-</u>

Logistic regression is a statistical model. It uses a logit model to model the probability of a certain class. It is used to explain the relationship between one dependent binary variable and one or more independent variable. If the target variable is categorical variable then go for logistic regression. Logistic regression is used only for the classification model. Input can be continuous or categorical. Output could be class(yes/no) and probabilities (0/1).

Accuracy rate:-91.85

Error rate:-72.96

2.Naïve Bays:-

Naive Bayes classifiers are a family of simple "probabilistic classifier" based on applying bayes theorm with strong independence assumptions between the features. This is only use for the classification purpose. It is the one of the supervised machine learning algorithm which works based on the probability. Basically this allow us to predict a class for a set of features or predictors using the probability. So it is called as probabilistic algorithm for classifier.

Accuracy rate:-92.60

Error rate:-63.95

3.Decision Tree:-

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. Since training dataset has target categorical variable, I opted for decision tress. It works for both categorical and continuous input and output variables. A decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node(terminal node) holds a class label. Entropy is the measures of impurity**,** disorder or uncertainty in a bunch of examples. Entropy controls how a Decision Tree decides to split the data. It actually effects how a Decision Tree draws its boundaries. Information gain (IG) measures how much “information” a feature gives us about the class. is the main key that is used by Decision Tree Algorithms to construct a Decision Tree. Decision Trees algorithm will always tries to maximize Information gain. An attribute with highest Information gain will tested/split first.

Accuracy rate:-83.82

Error rate:-80.93

4.Random Forest:-

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. In the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

As I have multiple no of features, I wanted a way to create a model that only includes the most important features. Random Forests are often used for feature selection in a data science. The reason is because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees (called gini impurity). Nodes with the greatest decrease in impurity happen at the start of the trees, while notes with the least decrease in impurity occur at the end of trees.

Accuracy rate:-90.62

Error rate:-97.5

Since the error rate is quite low in comparison to other model in Naïve Bayes model , I decide to apply this model on the top of the test data and predict the outcomes out of it.

Deployment Of Model:-

The concept of deployment in data science refers to the application of a model for prediction using a new data. Here in this case the naïve bays model is deployed using flask as it is very light web framework.