**Santander customer transaction prediction report**

**1. Obejective**

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.

**2. Dataset**

Data contains train and test datasets with 2 lakh of rows and 200 independent variables. Rows contains the information pertaining to the customers of Santander bank. Target variable contains zero and one’s, stating the customer who made the transaction with Santander as one’s.

**3. Analysis of data**

On the 200 variables contains the data for each customer for the transactions has been made however, as the labelling of the variables has not been given it is unlikely to predict what the data is denoting under each variable. All the variables helps to predict the outcome of whether the customer will transact with the Bank or not. Also, except the target variable all are numeric in nature with ID\_code might be denoting the customer id.

**4. Pre-processing**

We can now start to work on the data to reach our goal of working on the train data to predict the outcome on further customer databases if whether the given terms and variable how likely it is a customer to make a transaction.

**5. Missing Value analysis**

Dataset might contain some missing values which should be dropped as part of cleaning the data of any anomalies which will result our code to throw errors. Data should be free of any missing values. If a column contains more than 30% of missing values and the column can be dropped from the analysis and this would again would not be worth to add in our analysis. We the datapoint should not be deleted then we can replace the missing value using mean, median or knn methods. Our data set does not contains any missing values so we can comprehend that all the variables should be available to guess the outcome of the transactions. However, having said this the importance of each variable can vary on different lines considering the object each cell holds.

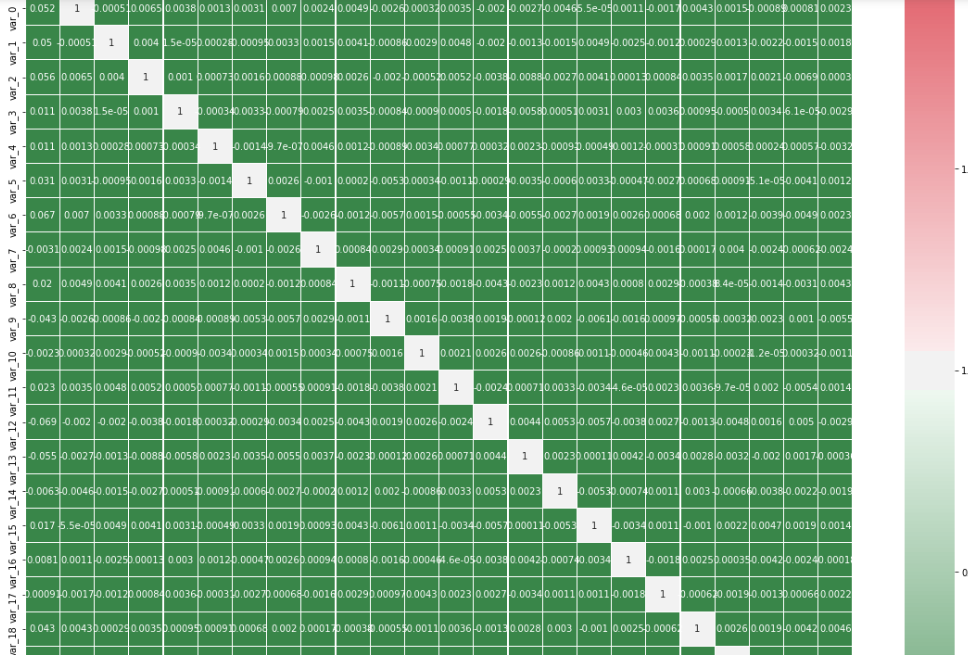
**6. Outlier Analysis**

Outliers are the data points which contains extreme values and should be treated before model deployment as this can deviate the outcome of our model. Using various techniques we can check whether the variable can be dropped or we can replace the outlier using any computation method viz mean, median or knn. Our dataset contains 20 such outliers and can be dropped from the data.

**7. Feature selection**

In a huge database not every variable contribute equally to predict the outcome. Some have sensitive data which might enhance the prediction of the target variable and some variable can contain noise which have mere any impact on the prediction. We can bring to use various plots and heat maps which will show us the importance of the variable and whether the act as a duplicate to other variable in which case only one of them will suffice for our model development. Co-relation analysis determines whether the correlation between two variables is positive, negative or have no correlation at all. Say if correlation between two variable is 1 we can drop one of them as both of the variable will fetch the same output.

Below is the heatmap between variables in python. Observing the variables we can conclude there is not much positive correlation between variables on the given dataset.



**8. Feature Scaling**

We can convert each variable on a single comparison number starting from 0 to 1. This can be applied on the dataset which contains multiple variables and huge size of data like our current dataset. Variables containing different measure scale can cause anomalies which while doing a comparison may not be interpreted in the correct manner. This step will reduce unwanted variation either within or between variables. It also brings all of the variables into proportion with one another. Normalization was used to get all the variables in line to a single range.

**9. Model development**

This starts the part where we pull insight from the data after all the processes have been done on data cleaning. All the above steps are as required and important for the data to make sense after applying various algorithms. As the size of our dataset was large to run all the models, I had to test run considering a lot of sample data.

Models used to test the dataset were:

1. Decision tree (Python)

2. Random forest (Python & R)

3. Logistic Regression (Python and R)

4. Knn imputation (Python and R)

5. Naïve Bayes (Python and R)

6. SGDClassifier (Python)

All of the models faired pretty close compared to each other except Knn did not give expected results in Python and R as well.