## Santander Customer Transaction Prediction

## Introduction

In fact, for any financial-oriented business to succeed, they must know if the client would repeat the transactions with them in the future. Customer attrition is a huge financial and operational drain for any company. If we could determine the cause for customers' churn and when it occurs, it would significantly assist the business to develop strategies for long-term retention.

In intended to facilitate Kagglers in identifying clients who might potentially make transactions regardless of the number of funds already transacted, Santander bank presents a challenge. The set of data that we've been given is more like the actual data, however, it is all numerical and disguised. No personal customer information has been shared with the competitors in the competition. The training and testing dataset has a total of 200k data points, as stated in the dataset. Training Data has 202 features, with the first 200 of these features all containing var 1 to var 200 values, along with a feature that contains the ID code, and other feature that contains the result of the transactions. For the test data, the entire features exist except for the target variable.

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: is a customer satisfied? Will a customer buy this product? Can a customer pay this loan?

The purpose of this report is to provide a classification model for Santander bank in order to determine which customer will make a transaction using an untagged dataset of 200 variables and 200,000 training observations. This report will provide a summary of the dataset provided to train the model and preliminary data exploration of the potential model features. The report will also outline methods attempted to reduce the dimensionality of the dataset. The report will then go over the best resulting model and other models attempted to classify which customer will make a transaction. Finally, the report will provide details on the best model selected and highlight difficulties encountered in the project and suggestions for improvements going forward.

# 1.1 Background and Data

The purpose of this report is to predict whether if a customer will make a transaction in the Santander bank data. The report will provide an overview of the variables and exploratory data analysis. We were provided a training dataset with 202 variables 200 of which are features we can train our models on. The training dataset had 200,000 observations that we need to use to predict the target variable for the 200,000 observations in the test dataset. The variables were provided to us with no descriptive variable names and no additional explanation in the form of a data dictionary. As a result, we did not have a way to logically infer which are variables of special importance (in case this could be things like the age of the account, the net worth of the customer, and other demographic features). In addition, it was hard for us to add additional features from the existing list of variables as we were unsure what the relationship of the different variables . We observed that the dataset is highly unbalanced where only 10% of the training data is tagged as 1 in the variable target which is an indicator flagging whether if a customer made a transaction. This is an important feature to keep in mind as we may need to oversample the customers that made a transaction to create a better model. A bar chart showing the unbalanced distribution in the training data can be seen below.

# 1.2. Aim and Objectives

**Aim:**

To help Santander bank we will use machine learning techniques to perform predictive analytics on the data to classify the customers who could make the future transactions with the bank and the customers who will not make any further transactions. This will allow the bank to classify the customers and make business decisions to retain the customers and will give an advantage over the competitors.

**Objectives:**

* Train robust machine learning algorithms that could predict the customers transactions with high accuracy.
* Eliminate the overfitting and underfitting in the machine learning algorithms to achieve accurate results on the future unseen data.
* With the help of data visualizations plot each feature to understand the behavioural patterns in the data.

# 1.3. Research Questions

## A research topic is a topic that a study tries to find answers for. This refers to something in the study that's addressed by examining the data and telling the storey it reveals. The research objective is typically formulated in a way that highlights a variety of aspects, such as the study's research population and factors, and also the study's main purpose. Research is commonly centred around scientific research. It's not surprising that researchers frequently revisit and revamp their research questions: Research questions tend to be evolving rather than static. Researchers must reassess and adjust questions as they conduct literatures and build a framework for the study. The research questions that we are trying to address in this project are:

1. What are the Statistical methods that can be used in the data to extract more useful information ?
2. What type of machine learning algorithm will help us achieve the best accuracy among the ones we are using in this project?

# 1.4. Tools and Techniques

The tools that will be used in the following project are the following:

Python – A high-level programming language that is used for general purposes.

Jupyter notebook – An interactive environment that is extensively used by Data Scientists to run Machine Learning and Deep Learning codes.

NumPy – Numerical Python, as the name suggests it is responsible for performing the math operations on the dataset.

Pandas – A tool that is used for manipulating the data.

Matplotlib, Seaborn – Plotting tools that are used for visualizing the data.

SciKit Learn – A library that allows the users to import the Machine Learning algorithms and train these algorithms.

Techniques: Machine Learning, Data Analysis, Data Visualization, Data Manipulation, etc.

# 1.5. Evaluation Metrics

Evaluation of the algorithms is the process of analyzing the performance of the algorithms to know how well they perform on the data. The evaluation metrics that will be used in this project to evaluate the performance of the algorithms is the accuracy metric.

## 2. Legal, Ethical, Professional, and Social Issues

**Legal Issues:**

Using the customer transaction data and sharing the data with others is a big breach of legal laws. Protecting the data is highly important since this data is private to the customers and contains sensitive information related to transactions.

To avoid the legal issues related to this project from the real-world perspective any data of the customers should not be shared with others and should be kept disclosed.

**Ethical Issues:**

Data Control & Monetization – The users should be given access to their data that is being tracked by the financial institutions. It is highly important that the customers should be aware of the data that is being collected by these institutions and the customer should also be given all the rights not to share any data.

Breach of this is a violation of the ethical laws Risk Zone 6.

Implicit Trust & User Understanding – The customers should be informed clearly about the data that is being collected and how the data will be used in the terms of service that are signed by the customers. Also, it is the responsibility of the financial institutions to verify whether the customer is aware of their policies regarding the private data that is being used.

Breach of this issue is a violation of the ethical laws Risk Zone 7

**Social Issues:**

Automation of the prediction could possibly impact the jobs of the employers who manually do these tasks for the organization. The automation of the work may lead to employee termination from the organization. If all the organizations implement automation in every sector thousands of jobs could be terminated, and this might result in the economic downfall of the nation which is a serious social problem.

## 3. Methodology

## 3.1. Installing Setup

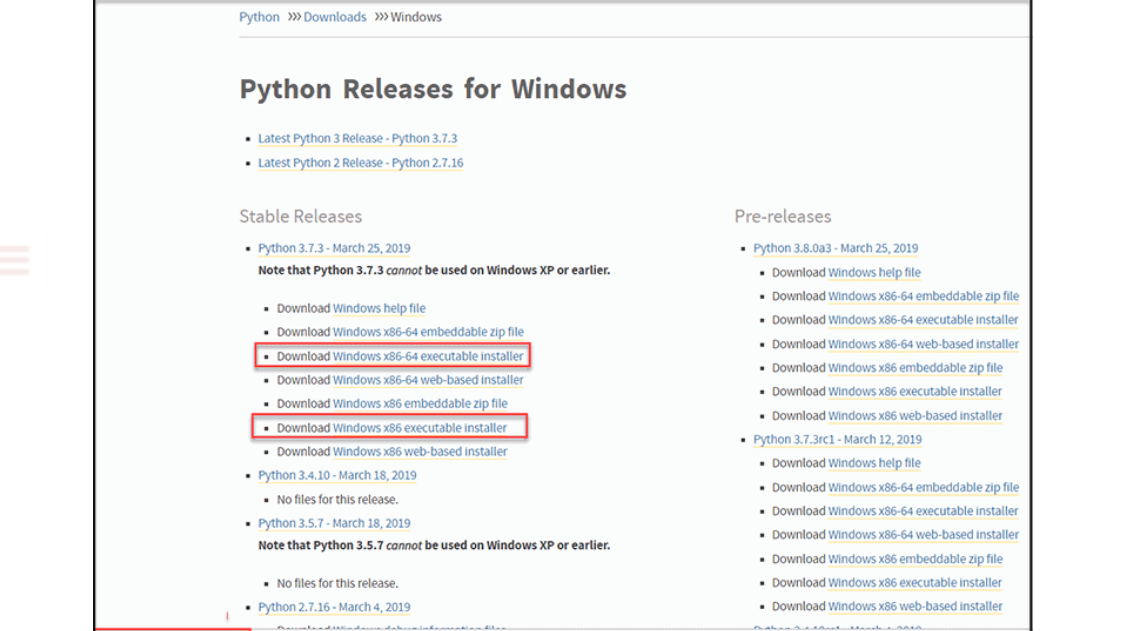
I have used Python 3.7, which I downloaded from the official Python website <https://www.python.org/downloads/> , for my project. I installed Python on my machine by setting it up on my hard drive and adding Python to my path. These are the following steps I have used while installing the setup:

Step 1: Select the suitable version of python to install

According to the need of the project we have to select the suitable version of python. For example if you are working on the project coded in python 2.6. you need to install python 2.6.

Step 2: Download python Executable installer

* Open your web browser and navigate to the Downloads for windows section in the official site of python.
* Search your desired version and select a link to download either the **Windows x86-64 executable installer** or **Windows x86 executable installer.**

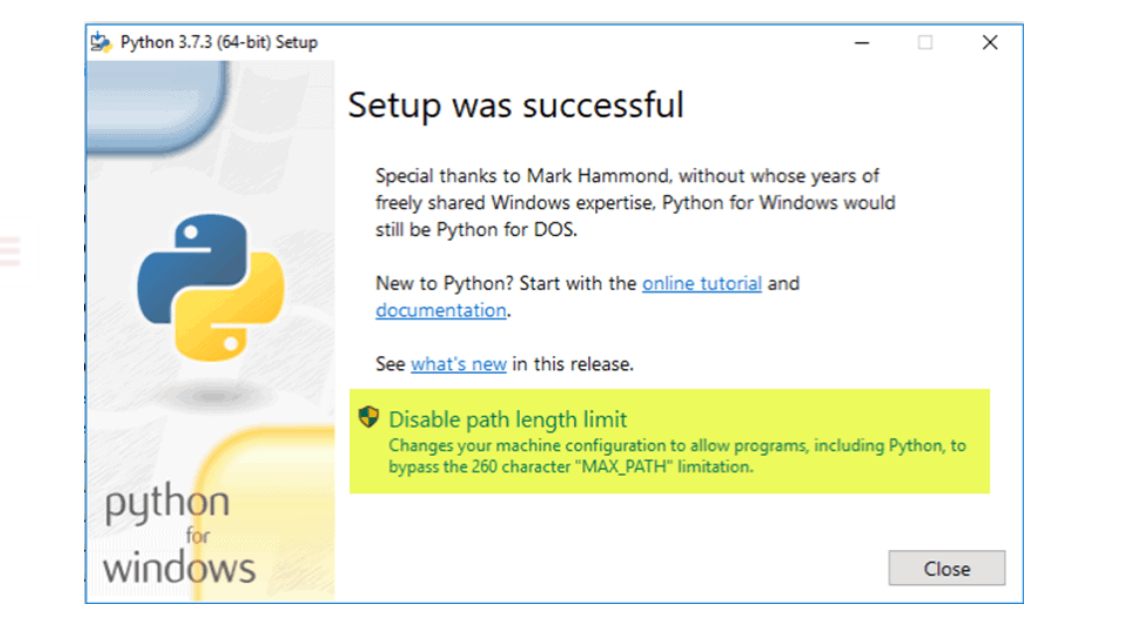


Step 3: Run Executable Installer

* In this step we have to run the executable installer. Double-check that you've selected the Install launchers for all users and have also selected the "Add Python 3.7 to PATH" options. When the interpreter is placed in the execution path, this placement gives the interpreter the ability to introspect.
* Select Install now button.



Step 4: Then setup will successfully installed.



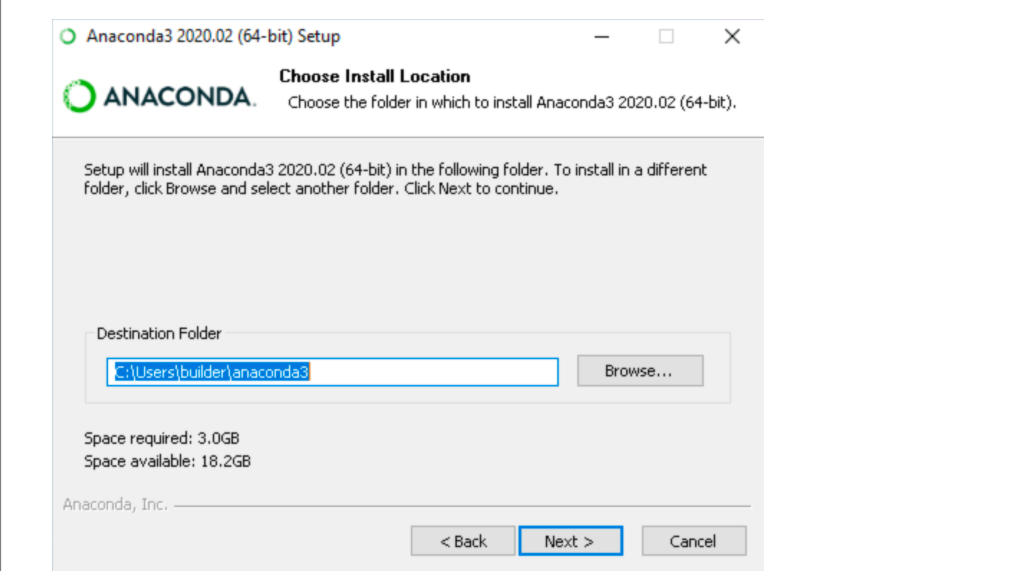
I have installed Anaconda for this project. And Additionally, I've installed Anaconda on my computer to get everything ready for running Python via Jupyter. After that, I was able to execute this project by using command prompt to install a few libraries. These are the following steps I have used while installing the setup:

Step 1: Download the anaconda installer from the official site <https://www.anaconda.com/products/individual#windows>

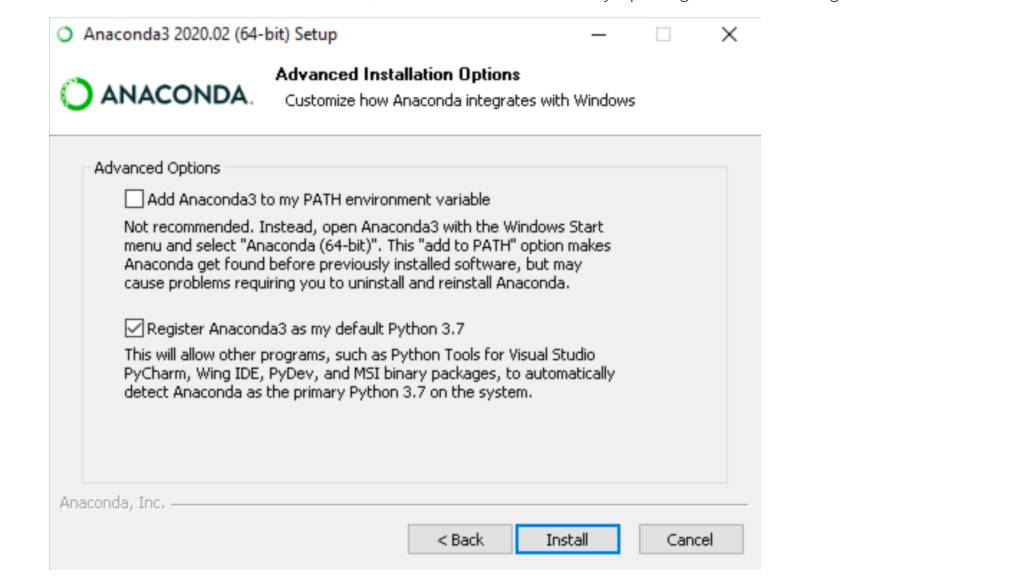
Step 2 : Double click on the installer to launch and click next.

Step 3 : Then read the terms and conditions and click “I Agree “.

Step 4: Then we have to select the destination folder to install the Anaconda prompt and click the next button.



Step 5 : To determine whether to install Anaconda to the PATH environment variable, pick one of the following options: To avoid the risk of other applications, we advocate against including Anaconda in the PATH environment variable. Open Anaconda Navigator or perhaps the Anaconda Prompt from the Start Menu, rather than using Anaconda programme.



Step 6 : Then click the install button. It takes time to install all the libraries.

## 3.2 Data Exploration

In Data Exploration part we will explore the dataset. The dataset for this project is available into Kaggle at – <https://www.kaggle.com/lakshmi25npathi/santander-customer-transaction-prediction-dataset>

We were provided a training dataset with 202 variables 200 of which are features we can train our models on. The training dataset had 200,000 observations that we need to use to predict the target variable for the 200,000 observations in the test dataset. The variables were provided to us with no descriptive variable names and no additional explanation in the form of a data dictionary.

We observed that the dataset is highly unbalanced where only 10% of the training data is tagged as 1 in the variable target which is an indicator flagging whether if a customer made a transaction.

## 3.3 Exploratory Visualization

Exploratory Data Analysis is a vital method that involves conducting initial investigations on data in order to identify trends, identify discrepancies, evaluate assumptions, and verify conclusions using summary statistics and data visualizations. Exploratory Data Analysis (EDA) is a computational data analysis technique focused on John Tukey's pioneering work. EDA offers a basis for a wide variety of data analytic activities and addresses the diverse types of data and architecture encountered by applied researchers. EDA's fundamental conceptual and computational tools include the use of graphics and interactive data visualization, a focus on model creation, diagnosis, and evaluation, addressing fundamental measurement issues associated with various distributions.

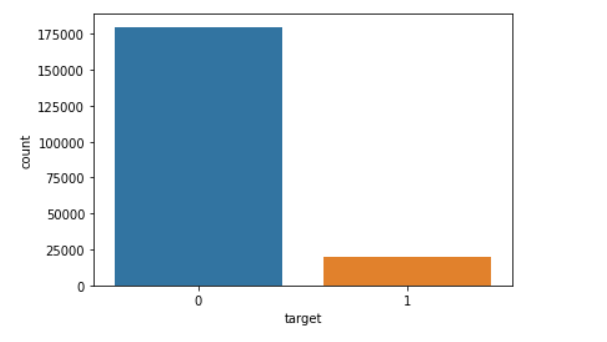
Although these methods serve as a foundation for all research, EDA places a high value on data - based learning from data to enhance standard hypothesis testing procedures that may neglect critical unanticipated aspects of data and their effect on modelling and estimation. The EDA, it is claimed, is critical both in the early stages of science, where hypotheses and model development must be well-informed.

## 3.3.1. Univariate Analysis

Univariate analysis is the most basic form of data analysis. Since "Uni" means "single," the data contains just one component. It is not concerned with triggers or relationships (unlike regression) and its primary objective is to describe; it takes evidence, summarizes it, and looks for correlations in it. In univariate analysis, a vector is simply a state or subset into which the data falls. Consider it a "variable." For instance, the study can focus on a variable such as "gender," "height," or "weight." However, only one variable is examined each time.

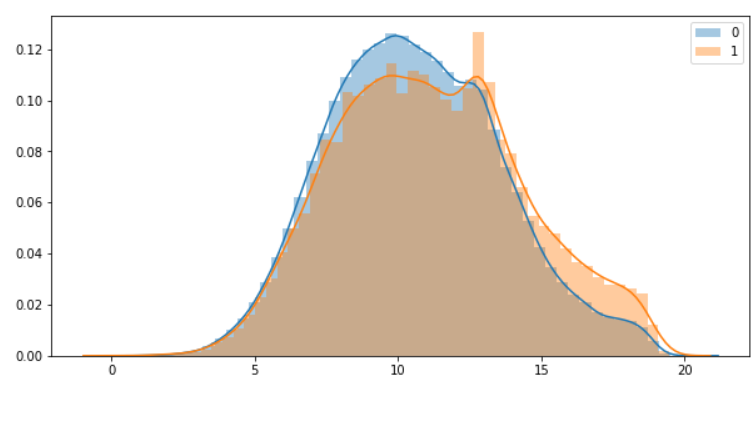
## 3.3.1.1 Target Variable

Target Variable is dependent variable in our dataset which is categorical feature. There are two categories which is 0 and 1.



The above plots show that most of the data points shows that most of the data points (more than 175000) belongs to “0” category and around (25000) data points belongs to “1” category. hence we can say that our dataset is imbalanced and we have to balance it first.

## 3.3.1.2 Var 0

Var 0 is a numerical feature. All features of this dataset is numerical features except target variable.

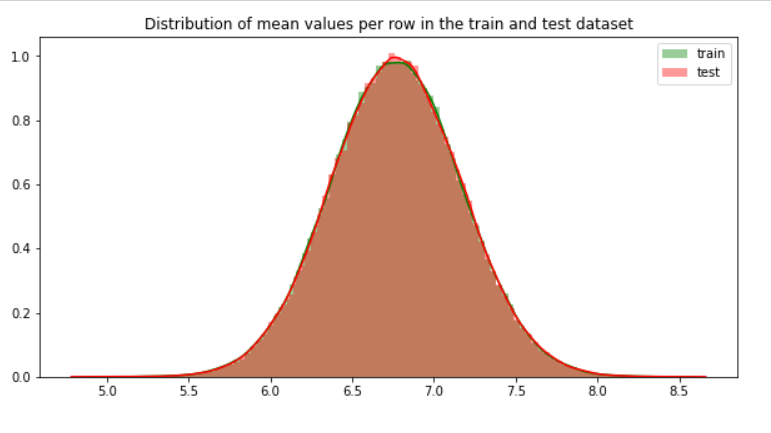
We can say that the graph is normally distributed for both the classes. And In Var 0 column there is no outlier.

## 3.3.2 Statistical Analysis

The process of collecting and interpreting data in order to discover patterns and trends is referred to as statistical analysis. Data analytics, often known as big data, is a constituent of this. Analysis of this sort takes place when a group of elements from which the sample can be drawn is examined, typically with the intent of generating statistical data.

## 3.3.2.1 Distribution of Mean

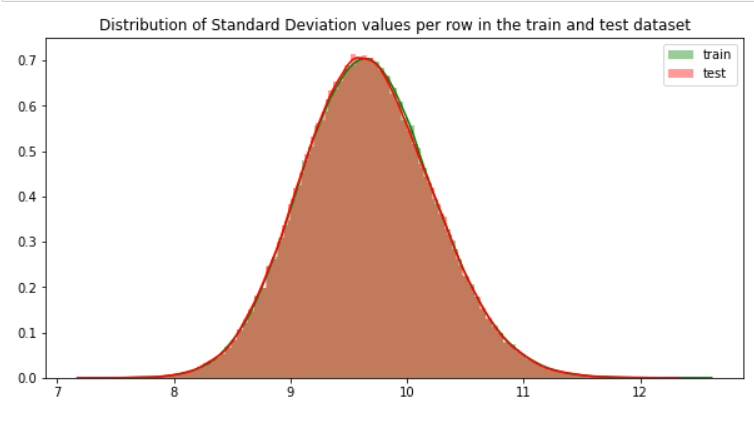
A statistical data set's distribution shows all the possible outcomes and frequencies for the data. As symmetrical distributions have a mean and a median at the same position, the mean and the median coincide, located in the centre of the normal distribution.



Thus, we can see that distribution of mean values per row is of Standard Normal Distribution.

## 3.3.2.2 Distribution of Standard deviation

The standard deviation gauges the disparity in the overall distribution of data. When data is distributed more widely, its standard deviation is larger. Although standard deviation cannot be negative, surprisingly, variance is. Using the normal distribution, standard deviation near zero means that the data points likely to be located close to the mean.

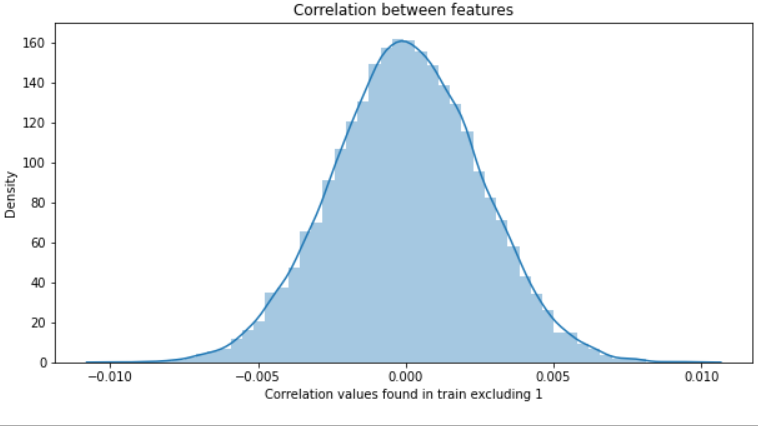


Thus, we can see that distribution of Standard Deviation values per row is of Standard Normal Distribution.

## 3.3.3 .Correlation

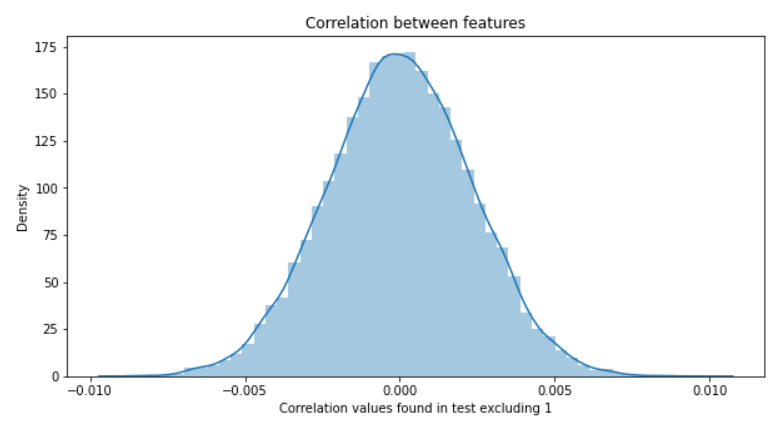
We will check the correlation between features of training data and testing data.

1. Correlation between features of training data



The graph is normally distributed. We can say that there is no correlation between the features of training data.

2. Correlation between the features of testing data



The graph is normally distributed. We can say that there is no correlation between the features of testing data.

## 3.4 Feature Engineering

Feature Engineering is a process of extracting features from raw data using domain knowledge. A feature is anything shared by multiple entities where prediction or analysis is done. Feature usage influences results. feature engineering makes data easier to examine. The world we live in is imperfect, therefore the data we use in it can be messy and unpredictable. Regardless of the type of data source (e.g., linear SQL database, Excel file, etc.), The data, which is normally formed as a table with a new row representing each sample and each column showing a trait, may be difficult to comprehend and handle.

To better understand our machine learning models' data and hence help them perform better, we need execute feature engineering. The task of converting data into a simpler form of comprehension belongs to feature engineering in learning algorithms. In this case, we are working to make the information clearer for a trained model, but feature can be generated such that data visualizations more approachable for non-data professionals can be created. But understanding the concept of clarity in ml algorithms is complex since the technique varies according on the type of data used.

Steps involved in Feature Engineering :

## 3.4.1 . Null Value Imputation

This is the main step of data cleaning .In this project there are no missing values in our dataset(training and testing data ). So there is no need to apply this step.

## 3.4.2. Encoding of Categorical Features

All features of the dataset are in Numerical form. So , there is no need to apply this step.

## 3.4.3 Outlier Treatment

A rogue observation is one that stands far away from the other data points. When we talk about outliers, we are saying that an outlier is much larger or much smaller than the other values in the data set. A statistical outlier may develop due to the erratic data, or due to anthropogenic mistake in the experiment.

These indicators may point to an experimental error or a large degree of skewness in the data (heavy-tailed distribution). Mode, Mean, and Medium are the three measures in statistics. The data are described using these. Meaning is an accurate depiction of the data when outliers are absent. If one data point differs greatly from the rest, the median is utilized. Mode is employed if an outlier is found as well as ½ most or all of the values are identical.

But some outliers are, in fact, good. Some outliers imply data is clearly different from other measurements. It could indicate, for example, something as odd as banking fraud or maybe even a rare condition

Example : Assume the data 1, 2, 1, 6, 4, 3, 100. If these values represent the number of breads eaten in lunch, then 50 is clearly an outlier.

Significance of Outliers:

* Standard deviation and mean are both heavily influenced by outliers in the dataset. Erroneous findings are a possibility with these.
* When outlier is present, many machine learning techniques fail. To avoid getting rid of outliers, they must be discovered and eliminated.

Interquartile Range :

To quantify the inconsistency of a data collection, divide it into quarters and measure it with an IQR. The information has been divided into equal portions with ascending order. Q1, Q2, and Q3 are the names of the three numbers that define the boundaries between four equal divisions. The data is divided into quarters (four equal parts), with the 25th percentile marked as Q1.

The middle 50% of the data is found in Q2.

The 75% of the data is found in Q3.

If a dataset contains an even number of data points and The middle of the data is in Q1.The median of the smallest points is Q2.

The range between first and third quartiles is known as the Interquartile range (IQR), and it can be calculated as Q3 – Q1. The outliers are those values found to be beyond the bounds of Q1.5 IQR below Q1 or Q3 + 1.5 IQR above Q3.

In this dataset All the outliers include our class1 target, so we cannot remove them.

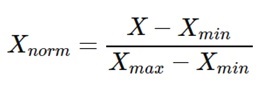
## 3.4.4 Feature Scaling

Feature Scaling is a method to normalize the various features in the data, ensuring that they are located in a specified range. The technique is used in the preliminary processing of data in order to deal with drastically changing values or magnitudes. Feature scaling is important because when this is not done, machine learning algorithms tend to favour larger values and disregard smaller values, regardless of their measurement unit.

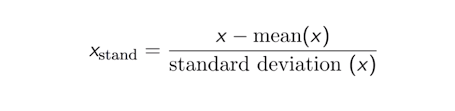
Example : Since the algorithm isn't employing feature scaling, the value 3000 metres can be considered greater than 5 kilometres, even though that's not true. Feature Scaling allows us to make all values equal, solving this problem.

Techniques of Feature Scaling :

1. Min – Max Normalization : Re-scaling features or observations is achieved by combining them with a distribution value that ranges from 0 to 1.



2 . Standardization : This approach, which scales an attributes so that it has zero mean and 1 standard deviation, is particularly effective.



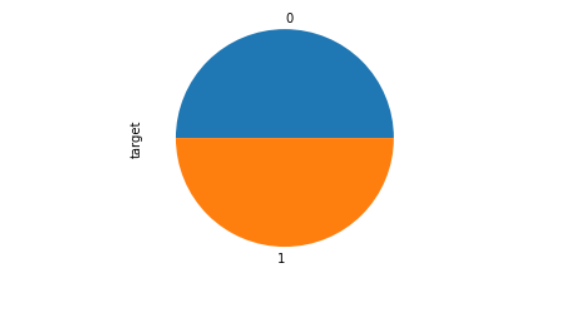
In this project I have used Standardization for scaling the data into a particular range. I have standardize all the data except target variable.

## 3.5 Balancing the Dataset

To gain an understanding about how to handle with an unbalanced dataset, let's first model the dataset. When classifying datasets with an uneven composition, existing Machine Learning techniques tend to generate subpar classifiers. A set of data that is imbalanced is said to have a rare event when the projected event belong to the minority class and the event rate is less than 10%. When dealing with imbalanced datasets, the usual model evaluation approaches will not provide appropriate measures of model performance. Bias against classes that have a high number of instances is commonly found in standard classification methods like Logistic Regression and Support Vector machine ( svm). Because of this, they are good at predicting the majority class values only. In other words, people generally neglect the positive attributes of the minority class. With this in mind, there is a greater likelihood of under classification of the minority group than the majority group.

Synthetic Minority Over-sampling Technique : To prevent overfitting, this method is followed. An exception to the general rule is chosen from the minority class, and then fresh equivalent synthetic examples are constructed. The dataset is augmented with these synthetic occurrences. This dataset is used as a training sample for the classification models.

I have applied SMOTE for balancing the dataset. After applying Smote on the dataset our dataset will be balanced.



## 4. Machine Learning

In this project, I will be training a set of Machine Learning models with optimal hyper parameters to extract the high accuracy with each model. I will be experimenting with different models such as Linear Models – Ensemble Models – Random Forest, Gradient Boosting Decision Trees, and Base Models – K-Nearest Neighbours.

Since this is a two-class classification model we are training, I will be using these models that are popularly used in the real world to experiment in this project.

## 4.1 Random Forest Algorithm

One of the most prominent algorithms for supervised learning is Random Forest, which uses an approach known as regression. ML applications can make use of it to solve classification and regression difficulties. Ensemble learning refers to the methodology of amalgamating several classifiers to deal with an elaborate problem and to better the model performance.

Its name is self-explanatory, as it uses a number of different decision trees on various subsets of the given dataset to take an average and improve the dataset's predictive accuracy. The random forest doesn't only rely on one decision tree. It instead bases its forecasts on majority voting amongst all the predictions and then predicts the ultimate result.

The below diagram explains the working of the Random Forest algorithm:



## Hyper parameter Tuning in Random Forest Classifier

After applying hyper parameter tuning in Random forest Algorithm we can see that for 50 number of estimators and for depth of 50 the algorithm gives higher accuracy The model gives around 97.44% accuracy on training data and 92.24% accuracy for testing data. So there is no overfitting in the model.

## 4.2 Decision Tree

Decisions trees are the most popular and efficient classifiers and predictors available. Decisions are created by dividing a group of information into "decision branches," which separate results into the appropriate categories based on attribute tests.

These are the following benefits of Decision Tree:

* The final outcome of a decision tree is usually well defined.
* A decision tree does not require significant processing for classification.
* For both continuous and categorical, decision trees can be used.
* Decision trees show the which variables to use when predicting or classifying.

## Hyperparameter Tuning in Decision Tree

I have applied Hyperparameter tuning on Decision Tree. After Appling Hyperparameter tuning we can see that

Criterian = gini, max\_depth =3 , splitter= best, max\_depth=3

These are the best parameters. For Training data the algorithm gives 59.000% accuracy and with testing data the algorithm gives 52.85% accuracy.

## 4.3 K-Nearest Neighbors

The Supervised Learning algorithm K-Nearest Neighbor is one of the easiest algorithms, with no prior assumptions on the properties of data. The K-NN method groups the new data/case based on resemblance to other, existing categories, assigning the new case to the category most similar to the current accessible cases. KNN method uses data from all the previous data points to compare similarities and classify new data points. The data can be categorized quickly with the help of K-NN, since the new information is easy to understand.

K-NN is best for Classification problems but may also be useful for Regression. Non-parametric algorithms like K-NN assume nothing about their data. It is sometimes referred to as a lazy learner algorithm since it takes a long time to learn from the training set, instead of doing so instantly. This method conducts an action on the dataset at the time of classification, rather than learning from of the training set immediately. When getting fresh data, KNN algorithms only classify it into a category with a similarity to the new data.

## Hyperparameter Tuning on KNN Algorithm

I have applied hyper parameter tuning in K Nearest Neighbors Algorithm. Basically hyper parameter tuning is used to find the hyper parameters which gives higher accuracy and also reduce the chances of overfitting.

For n\_neighbors=3 the algorithm gives higher accuracy with training and testing data. The accuracy with the training data is 80.25% and the accuracy with the testing data is 78.59%.

## 4.4 Naïve Bayes Algorithm

A naïve Bayes model is an extremely simple model to implement and particularly beneficial for datasets with a lot of data. Simplicity, and Naive Bayes' apparent ability to outperform more advanced algorithms, are further reasons for selecting Naive Bayes. It performs better than average in situations where categorical input variables are present variable in terms of numbers (s). It is believed that a normal distribution is appropriate for numerical variables

Assumption (a strong one) Another feature of numerical variables is that they are very loosely connected. Bayesian naïve Bayes is also referred to as a poor estimator, which makes it prone to various problems.

## Hyperparameter Tuning in Naïve Bayes

I have applied hyper parameter tuning in Naïve Bayes Algorithm. Basically hyper parameter tuning is used to find the hyper parameters which gives higher accuracy and also reduce the chances of overfitting.

The best alpha parameter is alpha =10 and the accuracy for the training and testing data is 80.64 and 78.67 respectively.

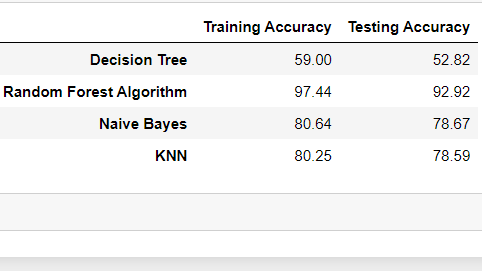
## 5. Performance Evaluation Metrics

While it is important to validate an evaluation of a model's skill, using metrics is necessary to evaluate the performance of a model. Metrics are employed to determine the most suitable problem-solving method. These metrics are what matter most.

Accurately identifying the ratio of correctly predicted instances to the total amount of instances evaluated is a metric for measuring accuracy. While the accuracy might be adequate as a measurement for model performance, it may not be good enough because it fails to include incorrect predictions. If someone treats a fake post as a real one, it could cause a serious issue. False positives and false negative issues that accommodate for misclassification should be considered because of this.



The Accuracy of each model is :



Accuracy for Random Forest Algorithm is highest. That’s why Random Forest Algorithm gives accurate results as compared to other models.

## 5. Conclusion

In this paper, I have worked with different models of Machine Learning to predict the Customer transaction from the given data to identify the better performing models. It was hoped to get multiple perspectives by doing various models, and also compare the performances. This classification challenge was present on a dataset with an uneven set of values, and no missing values. Anonymity and numericness are characteristics of predictors, whereas a categorical target variable is another feature of the prediction. There is no correlation between factors. I determined that there was an imbalance, therefore I tried to fix it A dataset was produced and several models were developed by using the original data with Random Forest as the final result. following which the ultimate result is 92.24% . After that, I proposed a few Several types of modifications, as well as the use of stratified folding, were implemented in order to facilitate separating training and testing data .

Based on the analysis my findings from answering the two research questions:

These are some statistical analysis methods that can be used in the data to extract more useful information:

* Distribution of mean
* Distribution of Standard Deviation

After performing statistical analysis we can see that the distribution on mean is Normally distributed. Random Forest gives better accuracy with testing data and also with the training data and the model is also not overfitted. This project can help company in following ways: Rather than making assumptions about what makes consumers like to each other, and rather than merely looking at predetermined categories, customers are classified according to their actual activities. hiding key details about individual clients through aggregated data. finding and mapping consumers' movement through various sectors over time, including transaction data. Instead of merely looking to the rear-view mirror of past data, it is useful to employ predictive customer behaviour modelling tools to forecast future consumer activity (e.g., transaction prediction).