# Banco Santander Customer Transaction Prediction.

## Problem Statement:

Customer preservation is essential in a variety of businesses as acquiring new customers is often costlier than keeping the current ones. Many companies are therefore always trying to answer the question “How can we predict the value of a customer over the course of his or her interactions with the business”. Santander therefore wants to know future customer transaction predictions.

## Company overview

Banco Santander, S.A. is a Spanish multinational commercial bank and financial services company founded and based in Santander, Spain. Their mission is to help promote businesses and people by helping them understand their financial status and how best to achieve their monetary goals. Santander is presently working on how to accelerate its digital transformation and platform strategy to boost growth and increase profitability.

One of these transformations is to identify which customers will make at least one single transaction in the future irrespective of the amount.

This question can best be answered by using predictive analysis of the customer past interactive data with the company and then a machine learning model is used to predict the possibility of these customers being retained or lost.

From these predictions, the company will know how much more efforts they need to keep the old customers which could be through bonuses and discounts, special coupons, better customer services and many more. While also thinking of how to attract more customers into the business through marketing campaigns, offer more discounts and deals, online adverts etc.

# Source of Data and description

The data for this project is taken from Kaggle, an online community of Data Scientists and machine learners owned by Google LLC.

The data set contains two comma separated value files for training and testing on the model with anonymous feature names for security purposes.

The train data set contains:

- Unique ID_code
- 200 numerical features labeled from var_0 to var_199
- Target either a 0 for no future transaction and a 1 for future transaction
- Total of 20000 unique records

# Methodology

This project will undergo a series of processes outlined as:

- Exploratory Data Analysis where the different relationships between the 200 features and the target
- Inferential Statistics applied on the different features and the target.
- After these analyses, stories will then be brought out from these results depending on the visualizations gotten.
- Since this project is a classification problem with two classes (0 and 1), different machine learning models will be tested on the dataset to see which best predicts with highest model accuracy score.
- Applicable extensions of project and it's impact to the business and/or customers too.

# Data wrangling and Storytelling from the Santander data set

The data set provided for Banco Santander Customer Prediction doesn't need cleaning with all records with unique data information.

Doing analysis on both the trained and test data frames, so much information was extracted from it which brings out so much insights from the data.

From the description of train and test data frames, the mean, min, and max are slightly different whereas the standard deviations of both are quite close but large. Standard Deviation being a measure that is used to quantify the amount of variation. This implies a low Standard Deviation value indicates that data points turn to be close to mean while a high Standard Deviation value shows data points are spread out over a wide range of values.

Visualizing the relationship between each feature and the target using density plots of all the 200 features shows sum of all probabilities with the peak of the density plot showing the highest probability of the data points. Most of these features have very low peak values which shows they have little or no influence on a customer making a transaction in the future. Also, their low standard deviation values also show the wide range of variability, so measures could be brought upon these anonymous features to make the values within a shorter range and eliminate lower value data points too. This will therefore decrease the standard deviation and an increased probabilistic value influence to the target variable 1.

Since the aim of this project is to predict customers that will make a transaction in the future, like 5% of the features have the peak value at a probabilistic level of 0.6 which is the highest i.e. they have lower values of Standard Deviation. These few features though unidentified for security purposes could therefore be used as the main target on how to best keep these customers.

To conclude, this analysis shows most of the features instead have high probabilistic values on a customer not doing a transaction in the future. Since this is not the aim of the project, those with lower Standard Deviation values and with higher probability values greater then 0.5 could therefore be used further though the features are unidentified.

# Inferential Statistics

Inferential statistics, is trying to reach decisions that extend beyond the immediate data alone. Inferential statistics is therefore used to make judgments and predictions that an observed variation or correlation between groups of features is a dependable one or not.
Correlation shows the extent to which the variables being compared fluctuate together.

Inferential statistics on the Santander data was concentrated on some of the features which had high probability values when density plots were plotted against the 200 features. Three of these features were considered and each of them, a Pearson Correlation was computed to see their relationship with the target.

Considering a Pearson Correlation between these features and the target.

Also, the Null Hypothesis is set for all the three features as: there is no correlation between each feature and the target variable.

The following observations were made;

- An r value calculated by Pearson Correlation inbuilt function as 0.04 which is less than 0.02 implies there is a positive but weak correlation between the two features. This means the increase in the unidentified feature increases the probability of a customer making a future transaction.
- Also, the p value displayed is greater than 0.05 indicates a weak evidence against the null hypothesis. This therefore implies a fail to reject the null hypothesis. Moreover, we can't say if the correlation is statistically significant or not.
- For var_108 and var_148, the have a weak negative correlation with the target. Since their r values are all negative. This implies a decrease in the value of either of these features increases the probability of a client making a future transaction.

From this analysis, the Pearson correlation therefore makes the var_91 feature statistically important as this feature could be used to increase future customer transactions which is the aim of this project. If for instance this feature was a coupon or some bonuses, increasing these offers if increased could therefore increased the probability of having a customer make a future transaction.