**Machine Learning Engineer Nanodegree**

## Capstone Project

# Definition

## Project Overview

For every business, customer relationship is most critical. Customer satisfaction is a key factor for success of any company. Even though company tries to collect customer satisfaction through surveys and polls. Most of the users don’t respond to these. So, there is a need to come with new ways to implicitly measure customer satisfaction. Due to recent advancements in Big Data and machine learning, companies can store large amount of data about the customer interactions. This data can be analyzed using machine learning techniques to get an estimate of customer satisfaction. Santander bank is trying to identify unhappy customers using machine learning techniques. They posted this problem as a completion in Kaggle. They have provided both training and testing data which can be downloaded and analyzed.

## Problem Statement

Main goal of this project is to predict how unsatisfied the customer with Santander bank is. Tasks of the projects are as follows:

1. Download the training and test data from Kaggle site.
2. Pre-process the data and make it more ready for the model.
3. Build a model and fit it on the training data.
4. Using the trained model get the predictions of the test data.
5. Submit these predictions and get score of the model.
6. Based on the results tune model to perform better.

Final python script should be able to take new customer data and give a prediction of how unsatisfied the customer is.

## Metrics

In any business most of customers will be happy with the company. So, percentage of unsatisfied customers will be pretty less. Accuracy will not be a correct measure for these kinds of problems. Area under Receiver Operating Characteristic (AUROC) curves will be more appropriate measure in this situation. ROC curves are constructed by taking ratio of Sensitivity (Identification of truly unsatisfied customers as unsatisfied) to Specificity (Identification of Satisfied customers as satisfied) at different threshold levels. A ratio of one means all the unsatisfied and satisfied customers are identified correctly.

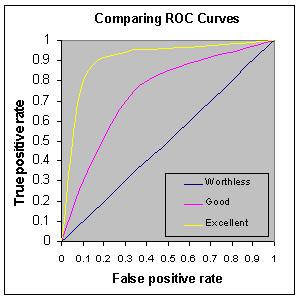


Figure 1: ROC CURVES

From the above figure we can see that, Yellow curve which is near top left performs the best and Red curve performs average. So, an Ideal model will be on the top left most, covering an area of 1. So, AUROC value of 1 is highest score that can be achieved by the model and value of 0 indicates worst model. 0.5 is the value of the benchmark model.

Resources:

1. https://www.kaggle.com/c/santander-customer-satisfaction
2. Fig1 : <http://gim.unmc.edu/dxtests/roccomp.jpg>

# Analysis

## Data Exploration

* Both training and testing data sets are provided separately. Training set has 76020 rows and 371 features including Target variable. Test set has 75818 rows and 370 features. Training data has to be separated into training features and training labels for supervised learning. There is also ID variable which is just a numerical representation of the customer that has to be extracted from both training and testing data.
* There are no missing values in both training and testing data sets.
* 4807 Duplicates were found in training data after extracting the customer Id.
* All the features of the training and testing set are numerical. Of these 258 features are discrete (‘int64’) and 111 are continuous (‘float64’) variables.
* Number of features is very large for such a small dataset. So, Feature Selection has to be done for removing unwanted features.