**Profound Question:**

From frontline support teams to C-suites, customer satisfaction is a key measure of success. Unhappy customers don't stick around. What's more, unhappy customers rarely voice their dissatisfaction before leaving.

A Bank is asking data mining experts to help them classify which customer are dissatisfied early in their relationship. Doing so would allow the bank to take proactive steps to improve a customer's happiness before it's too late.

**Data:**

To abide by customer’s privacy policy and to keep bank’s customer relationship model private the data is provided as hundreds of anonymized features to predict if a customer is satisfied or dissatisfied with their banking experience.

*Total Features –* 373 (Please see the attached data sample with the email)

*Total training data –* 76020 rows (Each row represents a single customer)

*Total test data –* 75818 rows (Each row represents a single customer)

*Feature data type –* Numeric (continuous & classification)

*Target variable –* Binary (1 – Satisfied, 0 – Dissatisfied)

*Missing values –* Yes (9999999999)

**Software:**

*Python 3.4(Spyder framework)* *–* Sklearn, pandas and numpy packages

**Algorithms:**

Since the data has several numerical features all of which are anonymous following algorithms and methods are proposed for model building and evaluation-

*Feature selection –* To identify important features among the several hundred given in order to reduce model bias and variance.

*Cross validation –* This technique will be implemented to eliminate any sampling related bias in the model.

*Extra Tree Classifier –* An ensemble method efficient when working with numeric data, enhances accuracy and reduces variance of the model by generating by generating several decision trees and then combining them for final prediction.

*Boosting (Ada Boost & Gradient Boosting) –* This is another ensemble method which boosts the accuracy of the model by training the features which weak learners and combine them to form a strong learner.

**Evaluation:**

*Area under ROC curve –* Area under ROC curve which plots the True Positive Rate against the False Positive Rate will be generated to determine the accuracy of the model. The final accuracy is obtained by submitting the results on Kaggle website.

**Source:**

The data is taken from Kaggle data science community website and this project is part of an ongoing competition. For more details, please visit the following link -

<https://www.kaggle.com/c/santander-customer-satisfaction/data>