Capstone Proposal

Machine Learning Engineer Nanodegree

FEBRUARY 18

Domain background

This project is mainly focused on Banking and Financial industry Customer satisfaction. As like other industries, Customer satisfaction is very important for any bank to success with given technology revolutions and change in ways of banking.
Unhappy customers rarely voice their dissatisfaction before leaving. So, it's important to understand the customer satisfaction levels in bank and update the products/services accordingly to maintain good market share.

This project will be mainly focused on Santander bank, please refer the following link to know more about the bank **Banco Santander - Wikipedia**

Problem Statement

The objective of this project is to identify dissatisfied customers early in their relationship with use of machine learning algorithms. Doing so would allow Santander to take proactive steps to improve a customer's happiness before it's too late. Solutions from this project will be feed into Kaggle Competition hosted by Santander

Datasets and Inputs

Santander customer data points would be used to predict customer satisfaction as highlighted in the competition, details as follows

Two datasets are provided with Target Variable as 0/1 - It equals one for unsatisfied customers and 0 for satisfied customers.

- train.csv - the training set including the target
- test.csv - the test set without the target

Data will be sourced from Kaggle - **Santander Customer Satisfaction | Kaggle**

Solution Statement

The sophisticated Machine model will be built to predict the customer satisfaction and also the drivers of unsatisfaction through EDA on the data points provided. This will help the bank to take right proactive steps

Benchmark Model

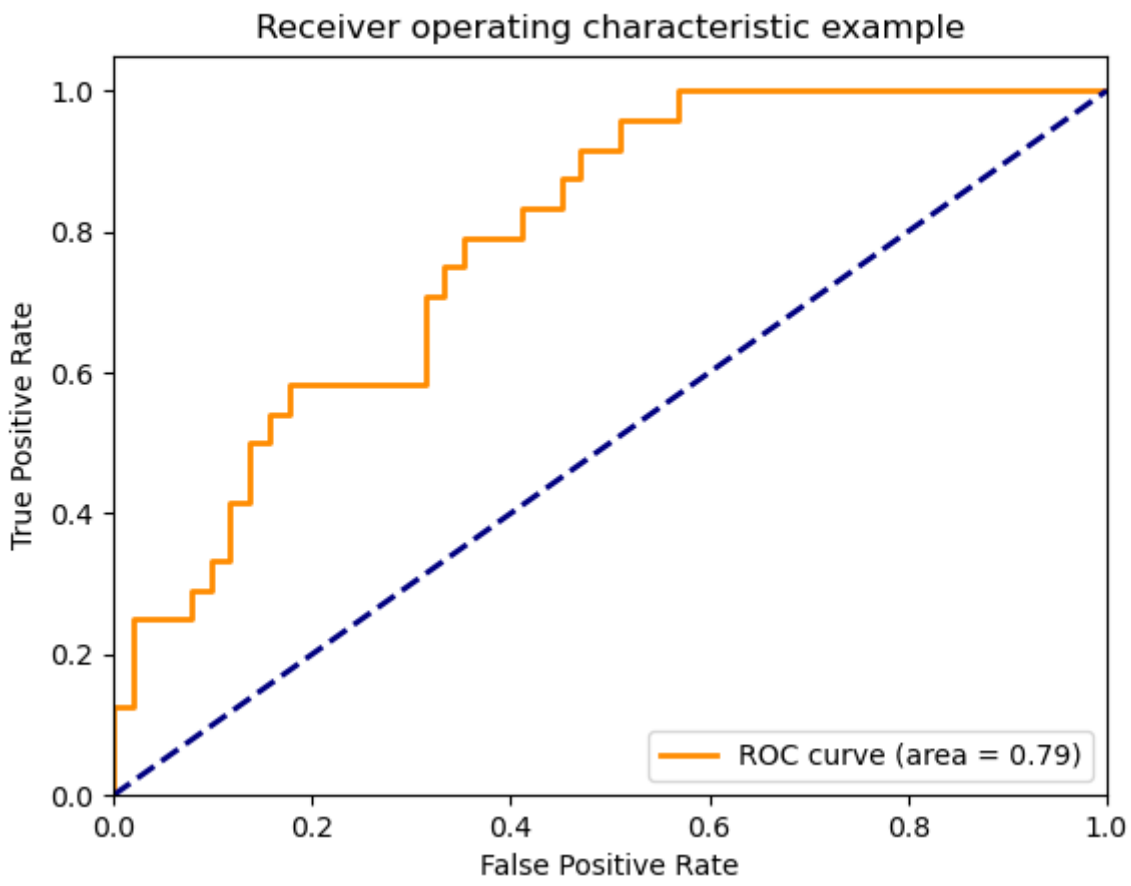
We would be using logistic regression model as a bench mark model with AUC of at-least 50%.

Evaluation Metrics

Evaluation metrics for the outcome on this project is area under the ROC curve (AUC).

AUC is calculated from a graphical plot curves typically feature true positive rate on the Y axis, and false positive rate on the X axis. This means that the top left corner of the plot is the “ideal” point - a false positive rate of zero, and a true positive rate of one.

Example –



Project Design

Step 1 – Data processing

AWS Sage maker notebook will be initiated with data from Santander competition

Step 2 – EDA

Exploratory data analysis on this data will help us in understanding the data patterns and insights

Step 3 – Feature Engineering

Based on EDA, new features will be created for our model if required

Step 4 – Model Selection

Logistic regression, XGBoost and Many other classifications will be tested and final model would be picked based on objective metric

Step 5 – Model training
Hyperparameters tuning & Testing

Model will be trained with using AWS Sagemaker capabilities and right hyperparameters tuning as required by model training. Will test the model with using batch processing and final model will be chosen based on Highest AUC

Step 6 – Deployment

Final model will be deployed with using Sagmaker