Capstone Proposal:

Adrian Lievano

April 1st, 2019

Domain Background:

Financial institutions, like Santander, help people and businesses prosper by providing tools and services to assess their personal financial health and to identify additional ways to help customers reach their monetary goals. In the United States, it is estimated that 40% of Americans cannot cover a $400 emergency expense[1]. As a result, it is imperative that financial institutions learn consumer habits to adopt new technologies to better serve their financial needs.

Problem Statement:

Santander, a financial institution, is trying to predict the next transaction a given customer is trying to complete based on historical banking information. This is a binary classification problem where the input data contains 17 unnamed normally-distributed feature variables. The solution to this problem will be evaluated on a provided test data set by Santander.

Datasets & Inputs:

Santander, via a posted Kaggle data science competition, provided a train.csv file which contains 288MB of anonymous customer data with 17 unnamed feature variables corresponding to a target variable, also unnamed, that is either a 1 or 0 depending on the classification[2]. This dataset is appropriate given that the Kaggle competition is restricted to using Santander's dataset.

Solution Statement:

The provided train.csv file contains 200,000 unique rows corresponding to customer data. Given the large dataset, and the need to complete binary classification, there are many solutions to this problem: mine will involve using a deep neural network, after preprocessing the inputs by normalizing and scaling features, to classify the two target variables. After the model is trained and validated on a subset of the data from the train.csv file, I will run my trained model on the provided test set from Santander and measure the accuracy of each prediction.

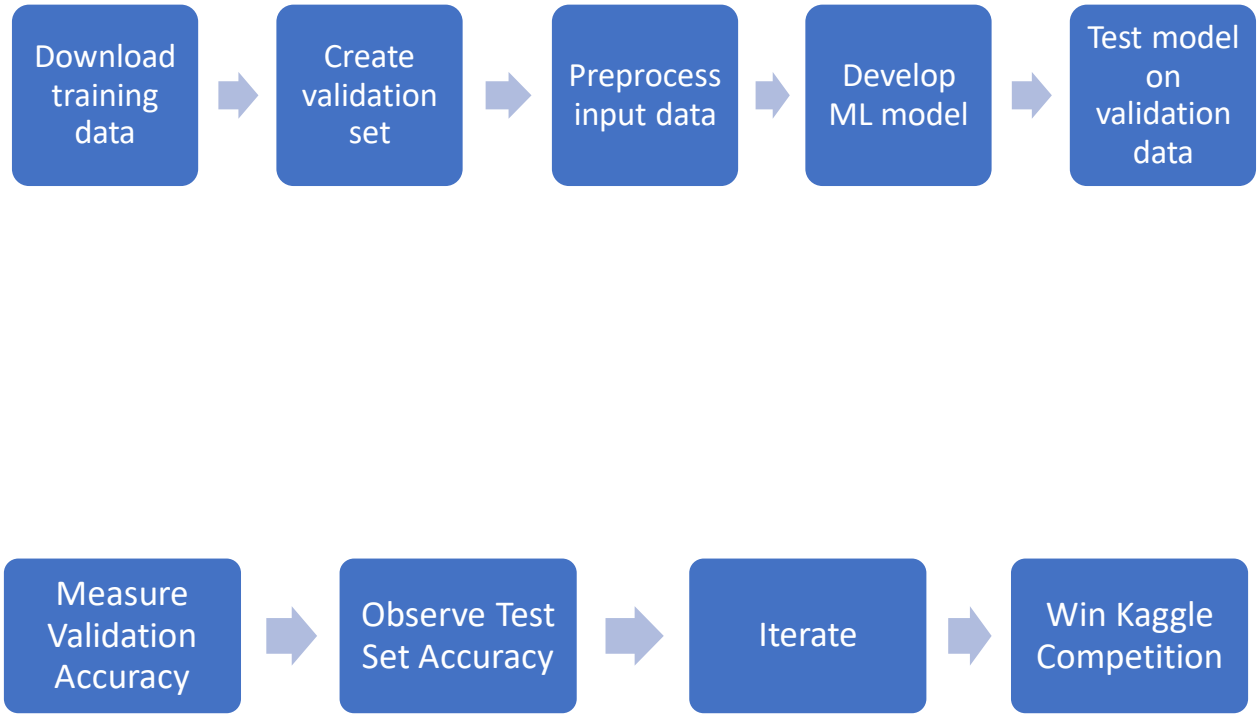
Benchmark Model:

The top five Kaggle models correctly predict the classification of the customers with greater than 92% accuracy. Their models, however, are not disclosed. Some of the public kernels listed on the discussion page use logistic regression as one potential solution. Given that the leaderboard is public with nearly 8,000 entries, we will compare our model to all the results on the Kaggle page.

Evaluation Metrics:

Test set accuracy will determine the performance of both the benchmark and the solution model. This dataset is provided by the Kaggle competition host, Santander.

Project Design/Workflow:



Literature Cited:

1. Bahney, Anna. “40% Of Americans Can't Cover a $400 Emergency Expense.” *CNNMoney*, Cable News Network, 2018, money.cnn.com/2018/05/22/pf/emergency-expenses-household-finances/index.html.
2. Santander, Bank. “Santander Customer Transaction Prediction.” *Kaggle*, www.kaggle.com/c/santander-customer-transaction-prediction/leaderboard.