

Machine Learning Engineer Nanodegree

Capstone Proposal

Santander Customer Transaction Prediction

Megha Patil
March 16, 2019

Proposal

Domain Background

This project is based on Kaggle competition described at <https://www.kaggle.com/c/santander-customer-transaction-prediction>

The problem is in finance domain. Santander is a commercial bank and financial services company. They are looking for ways to help their customers understand their financial health and identify which products and services might help them achieve their monetary goals.

The challenge is to identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted. The data provided for this competition has the same structure as the real data Santander has available to solve this problem.

The main reason for choosing this challenge is its goal. It is a real world problem and provided data structure is same as real world data. This will allow me to explore various machine learning algorithms.

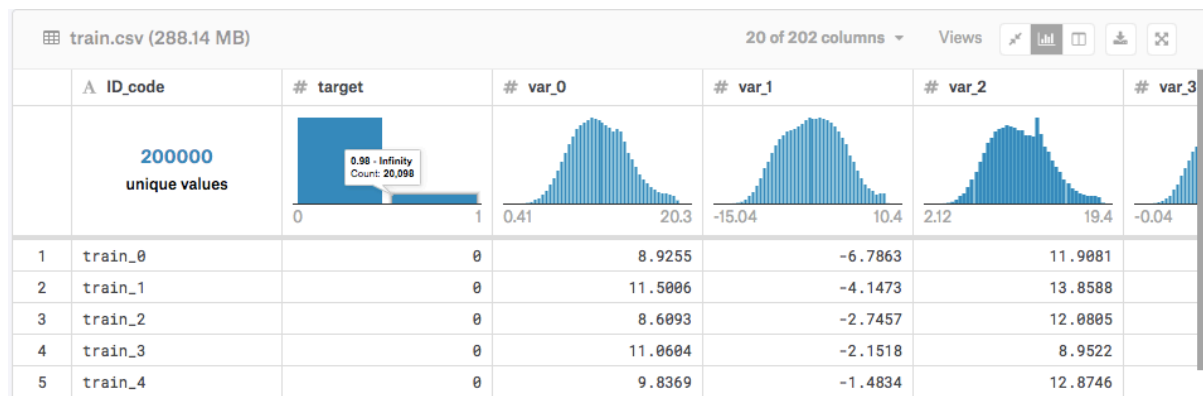
Problem Statement

As described in challenge, we have to use various machine learning techniques on given dataset to predict which customer will make specific transaction in future.

We are provided with an anonymized dataset containing numeric feature variables, the binary target column, and a string ID_code column.

The task is to predict the value of target column in the test set.

Datasets and Inputs



Santander has provided a single data file – train.csv for this problem. See above a sample of the data provided. The file is in CSV format. It contains 202 columns and 200K rows. Each row has an ID and target, followed by 200 attributes. “target” column is a binary value. Value of 1 represents that the customer will make a transaction. This is the column that needs to be predicted.

Solution Statement

The solution is expected to predict which customer will make a transaction. The prediction will be based on the 200 attributes provided in Santander data. The prediction will be specified by a binary variable against the ID specified in the input. Value of 1 indicates a prediction that the customer will make a transaction in future.

Benchmark Model

The training data provided by Santander contains the “target” column. This data will be used as benchmark to validate the solution.

Evaluation Metrics

Solution will be evaluated on area under the ROC curve between the predicted probability and the observed target.

Project Design

The project will be done in following phases:

Data Exploration

Study the training data set. Understand various attributes, their type, their characteristics (like statistical distribution etc). We will use graphs to visualize the data and their correlations.

Features Selection

We will use the data characteristics and correlations to decide the feature to use in our model.

Choosing a Model

We will try various Machine Learning models and techniques – logistic regression, decision trees, random forest, gradient boosting etc. We will choose the model that fits the problem. We will take into consideration factors like, accuracy, recall, performance etc.

Training

We will split the dataset into test and train data. We will train the model using the training data set provided.

Evaluation

We will run the model with test data set and evaluate it.

Parameter Tuning

We will tune the parameters of the model to get the best outcome on evaluation metrics.

Prediction

We will use the tuned model to do the prediction on test dataset provided by Santander. Once we submit the predictions, we will get an evaluation from Kaggle. We will use this evaluation to further improve our model.