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.

Similar problems in finance have been solved using machine learning methods. I have referred to the following case studies:

1. The following case study predicts fraudulent credit card transactions using supervised machine learning methods.

https://colab.research.google.com/gist/gjlr2000/b1231aa441e4d5cb2265edde2ba73118/creditcardfraud.ipynb

2. This case study is about sales prospecting (identifying new customers for a product or service)

https://becominghuman.ai/predicting-buying-behavior-using-machine-learning-a-case-study-on-sales-prospecting-part-i-3bf455486e5d

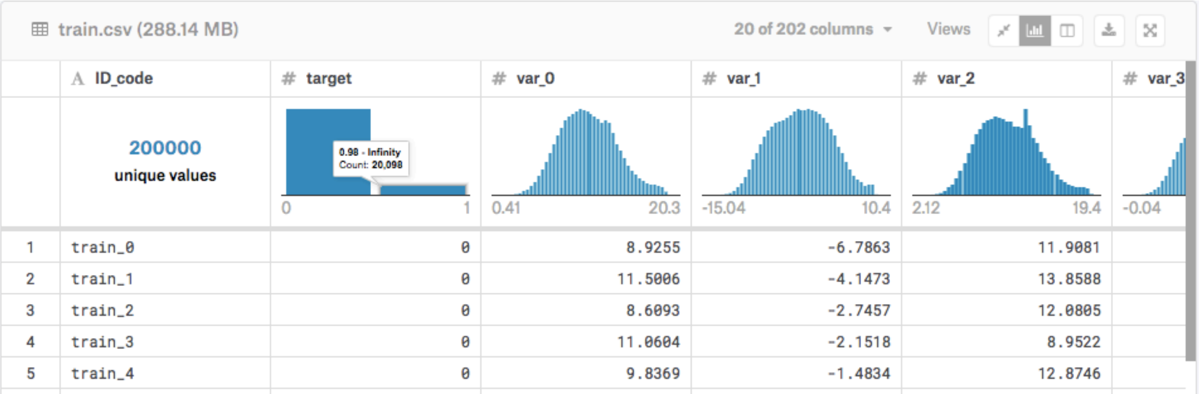
Problem Statement

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



	ID_code	target	var_0	var_1	var_2	var_3
1	train_0	0	8.9255	-6.7863	11.9081	
2	train_1	0	11.5006	-4.1473	13.8588	
3	train_2	0	8.6093	-2.7457	12.0805	
4	train_3	0	11.0604	-2.1518	8.9522	
5	train_4	0	9.8369	-1.4834	12.8746	

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.

The distribution of the “target” column (column to be predicted) in the training dataset is as follows:

VALUE	% OF RECORDS
0	99.02
1	00.98

The value of 1 is found in **less than 1%** of cases. This is an unbalanced dataset.

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

We can implement a naïve solution as benchmark model. The training data distribution shows that "target" has value 1 in less than 1% of records. In other words, the transaction is expected to happen in very rare cases. We will implement a naïve model which always predicts "target" value of 0. This will be our model to use as benchmark. We will use the evaluation metric - "area under the ROC curve between the predicted probability and the observed target" - to measure the performance of benchmark model. The solution we develop must do better than the benchmark model in predicting transactions.

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.