Machine Learning Engineer Nanodegree

Capstone Proposal

Natu Lauchande March 11th, 2019

Proposal

Santander Customer Transaction Prediction (Kaggle Competition)

Domain Background

Financial services are at the cornerstone of the modern society opportunity value chain for indivuals and business. Santander is a bank that thrives to know their customers better to serve them correctly. Part of providing the customers with the right financial choices is to be able to know and predict their desires [1] .

The topic of this project is to predict if bank customers will make a specific transaction in the future based on a set of a features. Predicting customer propensity/suitability for transaction or willigness is paramount on bringing inclusion and low cost financial services for underprivilliged communities [6] .

Predicting if a customer for instance will make a specific transaction might help the financial institution provision resources if it makes business sense . A special important application of predicting transaction is in the context of financial fraud . Having a performant transaction fraud detection system can lower financial risk for institutions [7].

Problem Statement

The goal of this project is to predict which type of transactions belongs each observation of the dataset and create a prototype of open source re-usable pipeline for binary classification of financial transactions given features for a customer and training data.

Datasets and Inputs

The datasets are provided by Santander on Kaggle competition website.

Input Data fields

- *id* - the id of a training/test set
- *var0..var199* - unitentified numeric features presented on the dataset.

Files provided

1. **Training dataset**

Description : Dataset to be used for training.

Filename : train.csv

Number of rows : 200000

Number of collumns : 202 [id + target + var0..var199]

Target Variable : 0 - Customer didn't make transaction 1 - Customer made transaction

2. **Test dataset**

Description : Dataset to be submitted.

Filename : test.csv

Number of rows : 200000

Number of collumns : 201 [id + var0..var199]

Target Variable : N/A

3. **Submission Sample**

Description : Sample dataset to be used during submission.

Filename : sample.csv

Number of rows : 200000

Number of collumns : 2 [id + target]

Target Variable : 0 - Customer didn't make transaction 1 - Customer made transaction

4. **Target class distribution**

   Target distribution (Percentage Wise)

   ```
   0    0.9
   1    0.1
   ```

   Percentage distribution

   ```
   0 +++++++++ (~90%)
   1 + (~1%)
   ```

   A clear class imbalance issue with regards to the target need that needs to be addressed during modelling and data preparation phase.

There is no hint on the datasets on the meaning of the features so exploratory analysis on variable ranking and importance needs to be considered.

Solution Statement

The solution for this project will consistist of the following :

1. The solution will be predictions of each feature of the dataset with the test set data submitted to the kaggle competion.
2. A notebook depicting the solution process as outlined by this process.
3. A software protype of an ML system that will productionize the solution with the following features : - Setup a new transaction prediction pipeline - Setup training - Productionize batch + api serving layer

Benchmark Model

The benchmark model will be a simple logistic regression model trained from the training data .

Evaluation Metrics

A set of metrics will be used to optimize and choose between the different models:

- AUC ROC curve metric (as suggested in the official competition) [2]
- F1 Score [2]
- Matthew correlation coefficient [3]

Project Design

Step 1 : Domain literature review

Look at relevant papers in the literature around techniques and applications on the domain of the problem.

Step 2 : Exploratory Data Analysis

Variables will be explored and analysed using Pandas and visualisations will be constructed using a suitable visualizatio tool like Seaborn.

Step 3 : Data cleaning and preparation

Datasets will be checked for quality and issues including missing values , duplicate data and outliers.

Step 4 : Build baseline/benchmark model

A baseline model will be created using a very simple logistic regression model and the most important features.

Step 5 : Feature analysis and exploration

Feature importance and rankings will be made with the data to improve familiarty of the data.

Step 6 : Iterate on model development :

As noticed before the training dataset suffers from class imbalance that needs to be addressed during model iteration and taken into consideration by mitigating using the most appropriate strategies(example : Smote, sampling, penalties, etc.)[8].

For each method cross validation and hyparameter random search will be used to find the best model in class.

```
a. Evaluate tree based boosting methods (GBM's, XGboost, Catboost).
b. Evaluate distance methods (KNN's and SVMs).
c. Evaluate suitability of deep learning methods.
d. Explore most promising alternative.
e. Submit to Kaggle leader the most promising solution.
```

Step 8 : Explore software tooling to aid on productionization of ML pipelines

Tools like MLFlow and Kubeflow present themselves as possible aids to the delivery of Machine Learning solution and aiding significatively the Machine Learning Engineering process. Part of this project will involve evaluating this tools as tools to be used for the productionization of the current project. [4,5]

A generic prototype standalone pipeline system will be created that will accept unidenfied variables and allow for a Machine Learning API for transaction type prediction : - Trained - Predicted with batch inference - Served with an API

Step 9 : Write final report and conclusions

Analyse and compare performance of different methods chosen on step 7. Final writeup of the project.

Reference

- [1] - Kaggle - (https://www.kaggle.com/c/santander-customer-transaction-prediction)
- [2] - Wikipedia Matthews correlation coefficient - (https://en.wikipedia.org/wiki/Matthews_correlation_co
- [3] - Steven S. Skiena - The Data Science Design Manual, Springer, 2014
- [4] - Databricks MLFlow - https://mlflow.org/, 2019
- [5] - Kubeflow - https://www.kubeflow.org/, 2019

- [6] - Zhang. E. et al - Financial Forecasting and Analysis for Low-Wage Workers - arXiv:1806.05362v3 , 2018
- [7] - Zheng. X et al - FinBrain: When Finance Meets AI 2.0* -Front Inform Technol Electron Eng , 2010 https://arxiv.org/pdf/1808.08497.pdf ,
- [8] - Brownlee J - https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ - 2019