Proposal

June 19, 2019

1 Machine Learning Engineer Nanodegreee

1.1 Capstone proposal

Fredrik Skatland, June 19th 2017

1.1.1 Domain Background

Mechanisms for detecting fraudulent activity is getting more and more important. There are strict laws that require financial institutions providing payment solutions to have appropriate measure in place to abate and avoid criminal activity such as money laundering, terror financing and fraud. The customers also expect the payment solution provider to have mechanisms and systems for protecting them against fraudulent transactions. It is important from a compliance point of view and a customer centric point of view.

Financial payment solution providers wants to stop and report frauduelent tranactions as they occure. To miss a fraudulent transaction can be expensive, so why not double check every single tranaction? There are far too many and that would be far too costly. Hence, we need a machine to do it for us. A machine can do this task quickly and for free without the customer or fraudster even noticing. But to teach this machine what is fraud and what is legitimate we need data. There are many many payment solution providers and they have many customers, generating billions of transactions, data should be abundant. However, payment transactions are very private in nature so there are no good public datasets available. For this project a synthetic dataset will be used.

1.1.2 Problem Statement

The objective of this project is to test multiple machine learning algorithms and compare their performance on identifying fraudulent transactions in the dataset. The primary goal is to create a single best model to classify transactions, and the secondary goal is create two models with a different beta coefficients for the F-beta metric. The low beta represents high the "high precision model", the high beta represents the "high recall model".

- One model with high precicion
- One model with high recall

1.1.3 Datasets and Inputs

The dataset is the synthetic datasets generated by the PaySim mobile money simulator, published on Kaggle.com in April 2017. The dataset consists of simulated mobile payment transactions

generated by the simulator called PaySim. The original logs were provided by a multinational company, who is the provider of the mobile financial service which is currently running in more than 14 countries all around the world.

This synthetic dataset is scaled down 1/4 of the original dataset and it is created just for Kaggle. Variables:

- step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
- type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
- amount - amount of the transaction in local currency.
- nameOrig - customer who started the transaction
- oldbalanceOrg - initial balance before the transaction
- newbalanceOrig - new balance after the transaction
- nameDest - customer who is the recipient of the transaction
- oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).
- newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
- isFraud - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.
- isFlaggedFraud - The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

1.1.4 Solution Statement

The solution to this project will be to test out implementations the following algorithms (with hyperparameter tuning and different architectures where relevant).

- Logistic regression (benchmark model)
- Decision Tree Classifier (C4.5)
- Random Forest
- Gradient booting (xgboost or lightGBM)
- Neural network (keras)

In this project I will ignore possible real life considerations such as training and prediction time and transparancy. However, efforts will be made to explain which features contribute to the model and in what way.

1.1.5 Benchmark Model

The benchmark model in this case a logistic regression without any feature engineering, with a logit link function. I also considered classifying every case either class (Fraud/Non-fraud), but since my secondary objective is to generate high recall and high precision models this would not make that much sense. Cohens Kappa will be evaluated vs. the "modal class model".

1.1.6 Evaluation Metrics

The model will be evaluated based on the following criteria. The final model chosen based on Cohens Kappa, which is a kind of balanced accuracy metric that accounts for the severe class imbalance and is fairly standard metric.

https://en.wikipedia.org/wiki/Cohen%27s_kappa

For the two models in the secondary goals the F-beta statistic will be used with the following beta:

- High precision model - beta = 0.5
- High recall model - beta 1.5

1.1.7 Project Design

The project consists of the following steps.

1. Loading the data and exploratory data analysis
    - Visualizing the different variables
    - Identify correlations
    - Identify outliers
    - Identify missing values
2. Data pre-processing, transformation and cleaning, potential operations depending on the algorithm
    - Centering
    - Scaling
    - Log transformation
    - Removing outiers
    - Removing or imputing missing values
3. Modelling
    - Implement algorithms listed in solution statement
    - Tune hyperparameters with cross validation, where relevant
4. Evaluation
    - Calculate relevant metrics
    - Inspect overfitting
    - Plotting results

This will be an iterative approach where I will cycle through these four steps multiple times as I get to know the data and the challenges with the approach.