Lloyds Banking Group

Data Analyst Incubation

Sprint 3 : Transaction Fraud Prediction ‘Hackathon’

Contents

Background

Quay Bank's Fraud Team have traditionally labelled transactions as fraudulent either after customers have reported them as such or when transactions have been placed after cards are stolen/lost. The Bank has so far relied on third party software to detect risk of fraud, with unsatisfactory results.

The Markets and Finance team believe that the development of an in-house approach to discern and potentially block fraudulent transactions is imperative for the bank's operations, especially now that there has been an increase in customers reporting fraud.

Eventually the Markets and Finance Team want to deploy this model against real-time transactions streamed directly into our cloud data store, enabling a proactive approach to Fraud Risk.

As a starting point, it is necessary to prove to the Fraud team that flagging fraudulent transactions can be achieved by the data analysts at Quay Bank, using common algorithms. Your supervisor has proposed a "hackathon" style machine learning competition to prove that fraud can be predicted accurately.

Deliverables

Working alone or in pairs, you will train a Classification model which produces an array of predictions (1 for fraud, 0 for not fraud) against provided validation rows. You will submit your predicted labels to the sprint supervisor, who will compare your submissions with the actual data and tell you how many fraudulent transactions your model failed to flag.

You have up to 5 chances to submit your predictions to obtain your best score. The best scores will be assembled on a leaderboard.

In addition to participating in the competition, your supervisor asks you to compile a simple one-page model summary (template: best_model_summary_template.pptx) describing your best model and the evaluation metrics you observed in training/testing and validation.

Assignment Detail

Data Set provided for Training

The dataset for you to train your model is:

sprint3_transactions.csv

These 236k transactions (including just 2,800 fraudulent rows) do not include any identifiable information about the account holder and all accompanying features, apart from amount, have been pre-anonymised.

Training Steps

- Examine the data and perform a comparative statistical analysis of the features by Target class (0 and 1). Use this to gain insight into the range of values for both legitimate and fraudulent transactions.
- Utilising visualisation tools, examine each feature by Target class, to identify which features show the most significant differences between the two classes.
- Conduct a correlation analysis to determine which features are most strongly associated with the Target variable. This will involve creating correlation matrices and possibly using advanced techniques like feature importance from tree-based models. You can drop features, apply sampling techniques at this stage, but remember the evaluation data will not be oversampled.
- Split the data into an appropriate train and test proportion.
- Select and train classification models. Experiment with model types, feature selection, parameter tuning and methods for validation.

Data Set provided for Validation

The dataset to validate your model is:

sprint3_predictions.csv

These 50k transactions are unlabelled so you cannot confirm your predictions directly. Imagine these are live transactions arriving at the database. We don't yet know if they are fraudulent. Your job is to use what you have seen in other transactions to make your best prediction for each transaction.

Prediction Steps

- Once you have your model trained, retaining all the training parameters, run your model against the 50K transactions, to create your predictions, retaining the same features as in your model training.
- Append your predictions to the prediction dataset. 1 is fraudulent, 0 is non-fraudulent.
- Export and submit your predictions dataset (with labels) as a csv to your sprint supervisor. They will validate your predictions against labelled data and advise your results so that you can try again or complete your model evaluation.