LDSSA Hackathon #1 - Binary Classification

Schedule

Hour	***Activity***
8:00	Arrival, Student setup
8:30	Hackathon Prompt, Team Assignment
9:00	Get to know your team!
9:15	**Start hacking!**
12:30	Lunch (no need to stop hacking)
14:00	Goal - make the first submission
15:00	Goal - make an improved submission
16:00	Work on the presentation
17:00	**Stop Hacking! (Submissions close)**
17:30	Team Presentations
18:30	Instructor Baseline Presentation
19:00	Winners Announcement and Closing

Overview

In this hackathon, with the support of **Farfetch** we will enter the intriguing and dynamic world of fraud detection.

Over the past few years, the landscape of credit card fraud has evolved dramatically, especially within e-commerce. Every day millions of transactions occur across the world and with the exponential growth of online transactions, the sophistication of fraudulent activities has surged, posing a significant threat to both consumers and businesses alike.

In response to this surge, machine learning has emerged as a crucial tool in the fight against credit card fraud. Its ability to analyse vast amounts of data and detect intricate patterns is pivotal in identifying fraudulent transactions in real-time. Advanced algorithms can discern anomalies, recognize behavioural patterns, and adapt to evolving fraudulent strategies, thereby bolstering the security measures of e-commerce platforms.

Today we will focus on developing a machine learning model to predict whether a set of transactions are considered fraudulent or not. We will be using a sample of a real dataset that has been provided by Farfetch, with various e-commerce purchases with different types of payment methods. Each transaction corresponds to a purchase of a particular customer in a particular moment.This data has been collected from August 2021 to November 2022. In total ~195k of these purchases (or events) were collected, from which ~17k resulted in fraudulent transactions. **Thus, our goal is to predict whether an order will result in a fraudulent transaction or in a genuine purchase, based on the order's characteristics and previously observed fraud patterns.**

To give you some intuition, a fraudulent pattern sometimes can be detected when it differs considerably from what is considered to be the customer's usual behaviour when purchasing products.

In order to save you time, you can use the requirements.txt in the Bootcamp directory which identifies some basic packages that you should install right away when setting up your virtual environment for the hackathon. Please bear in mind that they are not mandatory to use, and if you are more comfortable using other packages feel free to do so.

Disclaimer: This exercise is only meant for educational purposes. Fraud detection is a very hard subject where larger datasets and state of the art machine learning models are required. Don't get frustrated if the results aren't as good as you might have expected.

Objective

The main goal is to predict whether a transaction is considered fraudulent or not.
For each event in the test dataset, you'll have to predict the probability of it being an actual fraud.

Your submission file should be a CSV with two columns:

- **order_id**: the id of the event
- **result**: the probability of a transaction being fraudulent

When you submit your predictions, some validations will be run that will check the following:

- Your file has the two columns with the right name
- Your file has the right number of events
- Your file has the same event ids as the test dataset. The submission is sorted by id, so the order doesn't matter
- Your predictions are probabilities and not just 0s and 1s

Data files

You can find all these files in data/ under the hackathon directory.

- train.csv - Training set (195k events)
- test.csv - Test set (58k events)
- sample_submission.csv - Submission file example

Data dictionary

- **order_id** - id of the order
- **order_date** - date and time when the order was made by the customer (UTC timezone)
- **label** - 0 = legitimate order, 1 = fraudulent order
- **amount_usd** - amount spent by the customer in the order

- **payment_method** - payment method used by the customer
- **billing_address_country_code** - 2-letter country code of the billing address
- **shipping_address_country_code** - 2-letter country code of the shipping address
- **avs_code** - code sent by the payment processor indicating the degree of the customer's address matching(usd)
- **email_amount_spent_avg_180d** - average amount spent by this customer email in the past 180d (usd)
- **email_orders_count_180d** - number of orders made by this customer email in the past 180d (usd)
- **session_visitor_type** - type of visitor
- **shipping_method_type** - shipping method chosen by the customer to receive each of the items bought in the order
- **card_expiration_date** - expiration date of the credit card used in the transaction (MM/YYYY)
- **bin_brand -** card network brand
- **bin_type** - card funding type (Debit or Credit)
- **suspicious_activity_score** - output from a set of rules for card activity

Recommendations

❗ Remember: "*weeks of panning can save hours of programming*", so work with your team to plan and distribute work before diving in!
❗ Focus on feature engineering and data understanding/exploration, which type of features you can build to better detect fraudulent patterns.
❗ Make sure that you get to and submit a baseline ASAP! Then work on improving it.

Evaluation criteria for your model

Evaluation Metric - Area Under Receiver Operating Characteristic Curve (AUROC).
You learned about this metric in SLU10 - Metrics for classification.

Hackathon Rules

- The selection of the teams is **random**.
- Instructors will be available to help at any time. The instructors will **not** help your team solve the challenge but they will help your team to be on track and answer technical questions that your team might have.
- **No more submissions and questions** to the instructors shall be done after the end of the challenge.
- Your team will have to prepare a presentation to share your findings with everyone. See the presentation guidelines below. This presentation will be considered in the

overall evaluation of your team, so don't consider it less important than the ML model!

- You can submit your predictions up to **five** times to evaluate your AUROC score. The best will be chosen for the team's best score.
- The **final rank** is calculated as:

$$FinalRank = 0.5 * AUROC\_rank + 0.5 * Presentation\_rank$$

Where:

- ***AUROC_rank*** is the rank of your team in the leaderboard, considering the score of your **best submission**
- ***Presentation_rank*** is the rank of your team in the presentation evaluation

The teams will be sorted by *FinalRank* ascending, and the first team wins!

Feel free to ask any questions about the scoring function!

Presentation guidelines

- The presentation can take a **maximum of 4 minutes**. This is a hard limit! We'll literally silence you and move on to the next group after the 4 minutes have passed.
- The presentation should approach the following topics (following the data science workflow from SLU16):
    - Problem description
    - Data science workflow
        - Data preparation (data analysis, dealing with data problems, feature selection and engineering)
        - Model selection (which models did you try, how did you evaluate them, which one you ended up choosing and how good it was)
    - Recommendations / Future work if you had more time to work on the problem
    - A funny pun at the end (not mandatory, but everyone loves it)!
- You can use this template if you want (make a copy and edit your copy).
- Charts/tables/great visuals are encouraged in your presentation. We actually have an evaluation criterion for the presentation which is "Used **relevant** visuals" (note the relevant!)
- The team can decide who is presenting. There are no rules here, you can go with one person presenting everything or have everyone presenting a part.

Procedure

- **Team selection:** you'll be assigned to a team **randomly** by the portal
- On the hackathon page, you'll have to select a **name** and **gif** for your team. This is what will be displayed on the leaderboard.
- It's also on this page that your team will **submit the prediction files**.
- Send **@João Gomes** your presentation through Slack (**PDF file**) before "Submission of presentations".

After the hackathon

The leaderboard will be reopened after today so that you can keep trying to improve your score!

Summary of Resources

- Github repo
- Leaderboard
- Template for Presentation

Good luck!

Lisbon Data Science Starters' Academy team