LDSSA Hackathon #1 - Binary Classification

Schedule

Hour	***Activity***
8:00	Arrival, Breakfast, Student setup
8:30	SLU18 - Hyperparameter Tuning
9:30	SLU 19 - Workflow
10:30	Hackathon Prompt, Team Assignment
10:45	**Starting hacking!**
13:00	Lunch Served (no need to stop hacking)
14:00	Goal - make first submission
15:00	Goal - make improved submission
16:00	Start Working on Presentation
17:00	**Stop Hacking!**
17:30	Team Presentations
18:20	Instructor’s Presentation
18:30	Winners Announced
18:45	Closing Remarks by Pedro Martinho
20:00	Building Closed/Move Party Elsewhere

Overview

Today you will get to experience what is like to work at...

On a daily basis at Feedzai, predictive models process millions of transactions for large banks and online merchants. These models are generated with historical data using the latest Machine Learning techniques.

Feedzai works with different financial crime cases: Transaction Monitoring, Money Laundering and Account Opening. Today your team will be dealing with an example of transaction monitoring dataset, with some characteristics similar to the datasets processed by Feedzai.

The main goal is to predict which transactions are fraudulent and which ones are genuine. You will work with training data that spans 8 days, and make predictions for test data, that spans the 2 following days.

Objective

The main objective is to predict if a transaction is fraudulent.
For each transaction in the test dataset, you'll have to predict a probability of it being fraudulent.

Your submission file should be a csv with two columns:

- **id**: the id of the transaction
- **isfraud**: probability of the 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 transactions
- Your file has the same transaction 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. 8 days of transactional data
- test.csv - Test set. 2 days of transactional data for scoring
- sample_submission.csv - Submission file example

Data dictionary

- id - an anonymous id unique to a given transaction
- timestamp - timestamp in unix ms of the transaction
- product_id - product id of the product present in the transaction
- product_department - product department of the product present in the transaction
- product_category - product category of the product present in the transaction
- card_id - card id of the card used in the transaction
- user_id - user id of the user that did the transaction
- {C15, C16, C17, C18, C19, C20, C21} - anonymized **categorical** variables that characterize the transaction
- amount - amount of the transaction
- isfraud - binary variable that marks a transaction as fraud or not

Recommendations

❗ Bear in mind that the dataset is **imbalanced**. The provided dataset to illustrate Feedzai use case, has a rate of positive class cases (fraud) of about 10%. Though this is larger than typical fraud rates at Feedzai (which can be 1% or smaller), it will already allow you to explore some strategies adapted to imbalanced datasets.

❗ Note that the dataset contains time dependencies so you will have to be careful on how to split your dataset for training and validation of the model (hint, hint: sorting on the timestamp sorts on time)

❗ You may have high cardinality categoricals.

❗ There are categorical values that may exist in the test set, but not in the train set. You'll need to be clever in how you deal with this.

❗ If at any moment if looks like your computer is about to fly, you might find useful to work with samples. You may also find that heavier operations may take really long (or even crash your machine) on such a big dataset, so be smart about how your use your resources.

❗ Remember: "*weeks of programming can save hours of planning*", 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 characterize user past behavior.

❗ 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 SLU11 - 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 SLU19):
    - 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 it and edit your copy).
- Charts/tables/great visuals are encouraged in your presentation. We actually have an evaluation criteria 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 having everyone presenting a part

Procedure

- **Marking attendance:** go to the hackathon page on the portal and flag yourself as present in the hackathon. There is also an option if you're joining remotely.
- **Taking presences:** instructors will double check that you are here, so expect to hear your name shouted by an instructor during this phase
- **Team selection:** you'll be assigned to a team **randomly** by the portal
- In 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 in this page that your team will **submit prediction files**.
- Send **@Sofia Jerónimo** your presentation through Slack (shared link or 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!
Also, next week we'll share with you the code for a possible solution for the problem.

Summary of Resources

- Github repo
- Leaderboard
- Template for Presentation

Good luck!

Lisbon Data Science Starters' Academy team