FEATURE SPACE

OUTSMART RISK

Interview Task

1 Introduction

For the next stage we will be focusing on data science code and project reviewing skills, as our Lead role involves reviewing the work done by more junior members of the team. In the “Task” section below, we outline a task that you could imagine you have assigned to a junior member of the team. There will then be 1 hour interview where we will show you a response to this task. We will ask you to review the response, discuss areas that have been done well and suggest areas in which there could be improvements to the code, the data science, and the presentation. The code will be written in Python using a few commonly used libraries (e.g. `pandas`) and the script will not include any bugs that prevent the code running.

2 Task

Imagine you are a Lead Data Scientist at Featurespace. You are approached by a prospective customer who are looking for Featurespace to demonstrate the value of their product. You are provided historical data by the prospective customer and delegate the following task to a junior member of the Data Science team. During the interview, we will provide you with a sample response to both the coding and presentation parts of this task, which will be used for discussion.

Task Summary

Featurespace has been approached by prospective customer in the banking sector. The customer is an issuing bank seeking a machine learning solution to solve a growing fraud problem. The bank’s customers are having their cards defrauded and their money is being spent by fraudsters. This is causing customer dissatisfaction, so the bank is looking into introducing better transactional monitoring on activity on their customers’ cards.

They plan to have a small team of fraud analysts who review risky-looking purchases and decide whether to allow or block the transaction. This team will have the capacity to review 400 transactions a month. The scores from your model will be used to decide which transactions the fraud analysts should review. The bank is requesting that after working these alerts, as much fraud value as possible has been prevented.

The bank has provided 1 year of historical transactional data and fraud flags and asked you build a model which predicts the likelihood that a transaction is later marked as fraud.

Format of Report

Please return the code used to generate the model and a 2-slide presentation pitched at the bank’s executive board, which describes the performance your model achieved on an appropriately chosen test set, and the uplift this would provide the bank as a business.

3 Example Data

The prospective customer provided historical data in the form of two files:

- `transactions_obf.csv` - a file containing 13 months of historical transaction data
- `labels_obf.csv` - a file containing the IDs of transactions reported as fraud

3.1 Transaction Data

The following fields are included in the transaction data:

- transactionTime - The time the transaction was requested
- eventId - A unique identifying string for this transaction
- accountNumber - The account number which makes the transaction
- merchantId - A unique identifying string for this merchant
- mcc - The merchant category code of the merchant
- transactionAmount - The value of the transaction in GBP
- posEntryMode - The Point Of Sale entry mode
- availableCash - The (rounded) amount available to spend prior to the transaction
- merchantCountry - A unique identifying string for the merchant's country
- merchantZip - A truncated zip code for the merchant's postal region

3.2 Label Data

The following fields are included in the label data:

- reportedTime - The time at which the fraudulent activity was reported
- eventId - A unique identifying string matching the eventId of a record in the transaction file

4 Assessment

You will be assessed on your ability to review and suggest improvements to:

- The quality of the presentation
- The quality of the Data Science used to solve the task
- The quality and structure of the code written to solve the task

You will **not** be assessed on the following skills:

- Finding bugs which prevent the code running (for example you wouldn't need to check that column names referenced in the code actually exist in the data)
- Your ability to provide feedback in a constructive way and coach a Junior to be able to spot their mistakes. Although this is an important skill, we are not testing it in this task, so please be explicit with your review of the code/artefacts
- Suggesting slightly more computationally efficient feature engineering strategies - we're interested in the values the features will take, rather than which functions were used to compute them
- Your ability to read code for specific python packages (we expect you to have coding experience, but the code is clearly commented and we wouldn't expect you to have detailed knowledge about specific packages like `pandas`)