TRADE REPUBLIC

RECRUITING

FINCRIME
DATA SCIENCE CHALLENGE

To proceed with your interview process at Trade Republic, we have prepared a take home data science challenge that reflects the type of problems we solve on the FinCrime team.

You've been given a synthetic dataset of customer card transactions, split by training and test sets (tr_fincrime_train.csv and tr_fincrime_test.csv). The goal of this challenge is to generate meaningful business insights and create a fraud detection model. Please use the training set for analysis and training your model and the test set to evaluate your model's performance.

Note: tr_fincrime_train.csv has ~1.3million records and is 350MB in size. If you're unable to work with this volume of data due to computation limitations, feel free to reduce the dataset to a manageable size to solve the challenge and please make your logic for selecting the subset of data clear.

Description of columns:

1. index - Unique Identifier for each row
2. trans_date_trans_time - Transaction DateTime
3. cc_num - Credit Card Number of Customer
4. merchant - Merchant Name
5. category - Category of Merchant
6. amt - Amount of Transaction
7. first - First Name of Credit Card Holder
8. last - Last Name of Credit Card Holder
9. gender - Gender of Credit Card Holder
10. street - Street Address of Credit Card Holder
11. city - City of Credit Card Holder
12. state - State of Credit Card Holder
13. zip - Zip of Credit Card Holder
14. lat - Latitude Location of Credit Card Holder
15. long - Longitude Location of Credit Card Holder
16. city_pop - Credit Card Holder's City Population
17. job - Job of Credit Card Holder
18. dob - Date of Birth of Credit Card Holder
19. trans_num - Transaction Number
20. unix_time - UNIX Time of transaction
21. merch_lat - Latitude Location of Merchant
22. merch_long - Longitude Location of Merchant
23. is_fraud - Fraud Flag <--- Target Class

Part 1: Analyse customer transactions

In the first part of the challenge, we would like you to familiarise yourself with the tr_fincrime_train.csv dataset and generate **3-5 actionable business insights in the form of visualisations** to discuss with your product lead.

For example: Maybe you find a correlation between a user's location and the incidence of fraudulent transactions. With this insight, we could look into our KYC (know your customer) process in that location, and recommend potential product improvements to our onboarding process.

Part 2: Fraud Detection

In the second part of the challenge, we would like you to **create a machine learning model that predicts whether a transaction is fraudulent** or not. Please use the training set for training your model and the test set to evaluate your model's performance.

Along with your code and model performance metrics, we would also like to know if you had more time to work on this challenge (as you would at work), what would you do differently? Would you look at other kinds of data or try different models? If yes, please explain what kind and why.

Finalization

Please submit your solution with all required files via email.

We are looking forward to the further discussions!