2nd Round User Interview Invitation - Machine Learning Engineer INDICO - Problem set 1

Anas Nafis Almustofa

- Fraud Detection
    a. Data prep and setup env
        Download data from the kaggle competition
        https://www.kaggle.com/competitions/ieee-fraud-detection/data
        Activate new python environment using conda or venv
        Install requirements.txt using ```pip install -r requirements.txt```
        Load data into .ipynb
    b. Data Analysis
        Load data into .ipynb
        Gather insight from data. Below are insights that we already gathered:
            There are two datasets: transaction and identity, each of data has been splitted into train and test data
            The transaction contains the transaction history and its attributes such as the transaction's datetime, address, cards, and so on.
            The identity contains the transaction device id and device info
            The training data is imbalance (non-fraud data: 569877, fraud: 20663)
            There are numerical and categorical data in both dataset
    c. Data prerocessing
        Encode categorical data into numerical data using label encoders with train data as base of label encoders
        Sample non-fraud data to balance it same like the number of fraud data
        Split data into train and validation with validation ratio 0.2
    d. Training and evaluation
        Fit the data into Sklearn Random Forest model
        Measure accuracy, sensitivity, specificity of the model. The model performance that we created are stated below:
            accuracy training: 0.9998
            accuracy testing: 0.8487
            sensitifity training: 0.9996
            sensitifity testing: 0.8311
            spesificity training: 0.9999
            spesificity testing: 0.8662
        Predict fraud from test data and submit the result to kaggle competition to get metrics from test data. Here is the result of the model:

Submission and Description	Private Score	Public Score	Selected
submission.csv Complete (after deadline) · 1h ago	**0.802247**	**0.833674**	☐

e. Possible Improvement
    Add identity data into feature value
    Use data standardization
    Research more on suitable model