Please answer the following questions using github/gitlab/ bitbucket repo (Select the one that is a good fit for you). Finally, please provide a public repo link so the hiring manager can read your answers.

Task 1: General Questions

1. What is your preferred language when building predictive models and why? **Ans.** I prefer using Python more often since it is rich in useful libraries for data pre-processing (e.g., `scikit-learn`, `numpy`, `pandas`), model building and training (e.g., `torch`), and visualization (e.g., `matplotlib`). For model building and prediction, I usually used `torch` to construct a neural network for prediction. One main reason is that it supports the parallel computation of tensors on GPUs, which results in efficient prediction procedures. Another reason is that it also supports the format transformation with `numpy` data. Thus, I can use the preprocessed `numpy` data for prediction easily by transforming it into `torch`.

2. Provide an example of when you used SQL to extract data. **Ans.** According to my working experience in EC data team at LINE, I used SQL to extract the data of customers from the LINE shopping app. In order to find the impacts of decreasing number of customers, we extract the attributes, such as purchase amount, products, time and demographics, to analyze i) which groups of customers that reduce consumption, and ii) the purchase power of the groups evaluated by their historical consumption behaviors.

3. Give an example of a situation where you disagreed upon an idea or solution design with a co-worker. How did you handle the case? **Ans.** According to my experience co-working with my co-advisor to publish our research paper, we often had different ideas. For example, he might suggest using a bar chart to present comparative results, whereas I believed a line graph would better highlight the trends over time. Under the circumstances, I would start by acknowledging his perspective. In the following, I would then explain my reasoning and provide evidence to support my approach, such as referencing examples from similar published papers or showing how the line graph made the trends more visually intuitive. If we still couldn't reach a consensus, I would suggest testing both approaches—either by seeking feedback from peers or reviewers or by incorporating both options in the draft for evaluation. This way, we ensured the final presentation effectively conveyed our findings while maintaining a collaborative and professional working relationship.

4. What are your greatest strengths and weaknesses and how will these affect your performance here? **Ans.** My strong educational background, including a Ph.D. in Electrical Engineering and Computer Science and top-tier academic performance, aligns

well with the job's requirements. With experience at NVIDIA AI, LINE Corp, and Academia Sinica, I have a proven track record in deep learning, computer vision, and generative models, such as Stable Diffusion, which directly relates to the role’s focus. Proficiency in Python, TensorFlow, PyTorch, and familiarity with SQL further strengthen my fit. My publications in top conferences (e.g., CVPR, NeurIPS) highlight my ability to contribute cutting-edge research. Additionally, my leadership and collaboration skills, demonstrated by leading projects and working with international teams, ensure I can excel in this role. However, my experience is more research-oriented, with limited exposure to deploying production systems and tools like Docker or Airflow. While this might require a learning curve, my strong problem-solving abilities and quick adaptability make me well-suited to bridge these gaps effectively.

Task 2: Python model development

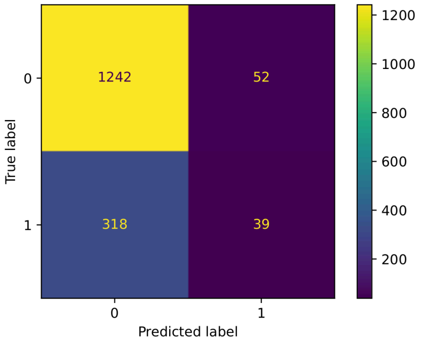
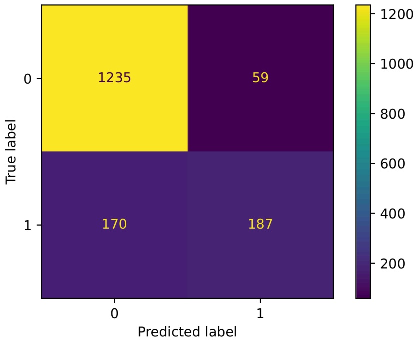
Objective: Given the dataset, **train.csv** - the training dataset; **Exited** is the binary target and **test.csv** - the test dataset; your objective is to predict the probability of Exited, write a python script (main.py) that when run (e.g. python main.py) will output:

- a CSV file containing the following:
  - Predicted Exited from test.csv
  - Evaluation of the predictive ability of the model.(e.g. F1 Score, Confusion Matrix…etc)
- Plots that can help us visualize the classification data and the prediction curves.
- Please submit codes, explanations, and plots when finished. Try to be more **specific**, a README might be helpful.

Preprocessing

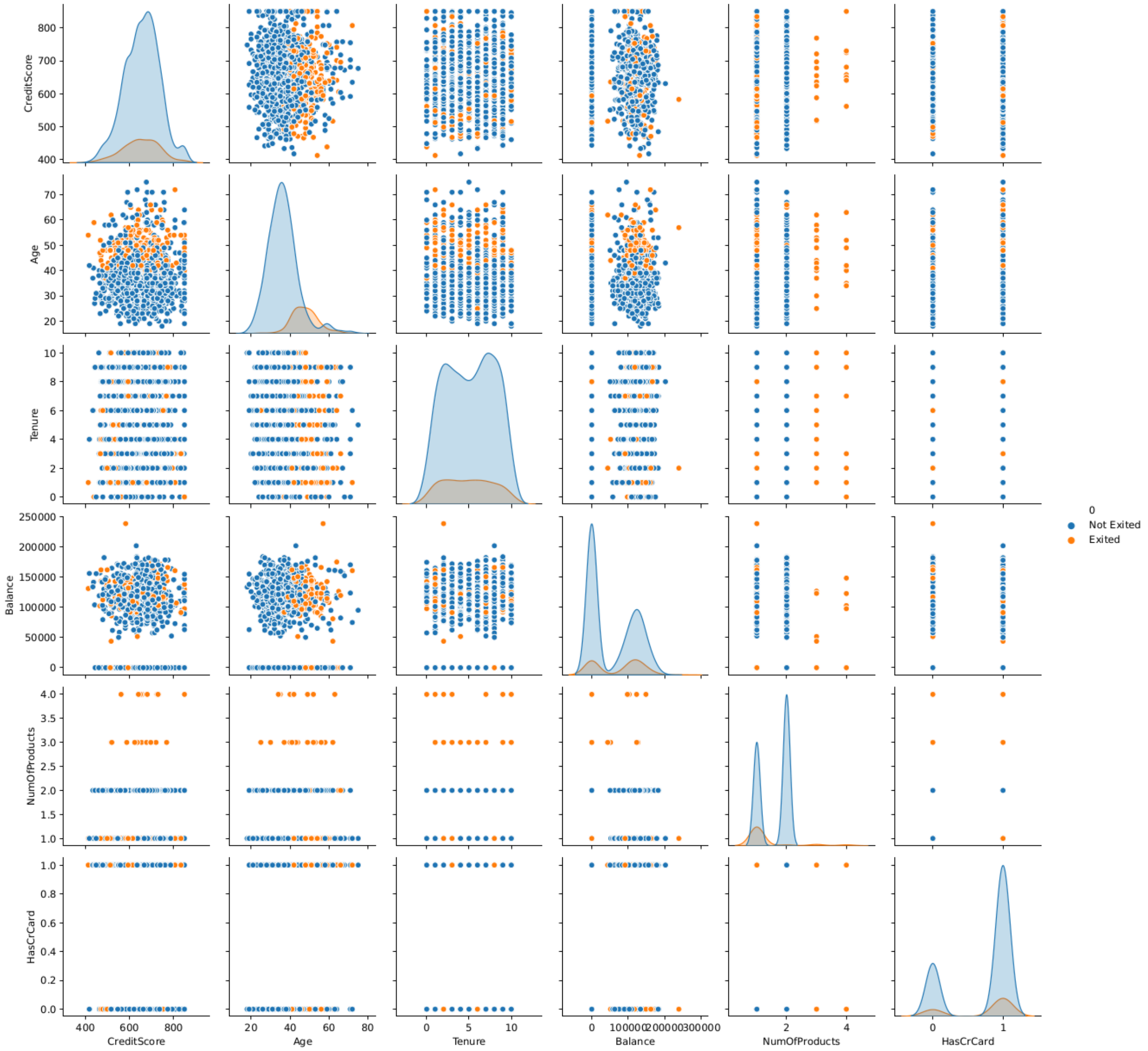
- Remove the columns - 'id', 'CustomerId', 'Surname'
- Turn the categorical variables ('Geography' & 'Gender') into one-hot encodings
- Split a validation set from train data

Methods & Performances

Methods	**Logistic Regression**	**XGBoost**
Accuracy	77.6%	86.1%
F1-Score	17.4%	62.0%
Confusion Matrix	 Num.(predicted as not exited but actually exited) = 318	 Num.(predicted as not exited but actually exited) = 170 (*misclassification rate decreases)

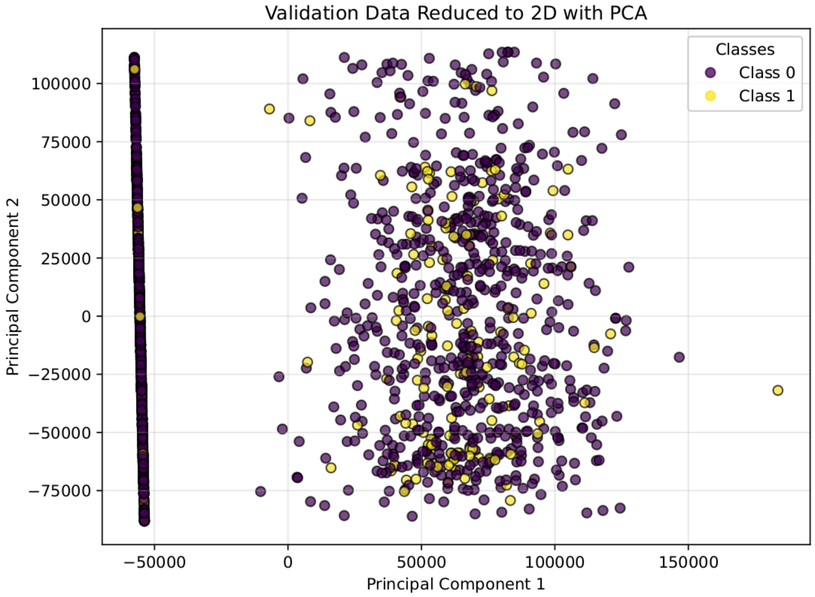
Visualization

- XGBoost
- Pairwise scatter plot (selected variables which are significant for identification)
- Orange: Exited
- Observations
    - The elderly tend to possess higher risks in exit.
    - People with the products 3 to 4 are all exited.



Discussion

Dimension reduction such as PCA has been tried for visualization. However, the data seems not to follow linearity.


Validation Data Reduced to 2D with PCA
Classes
Class 0
Class 1
Principal Component 2
Principal Component 1