Final Report

1. GitHub link of your code: https://github.com/kirito878/final_project.git

1-1.model_link:

https://github.com/kirito878/final_project/blob/main/model/elasticNetCV.pickle

2. Reference:

1. https://www.kaggle.com/code/danielkhromov/tps-aug22-eda-modeling
2. https://www.kaggle.com/code/vishnu123/tps-aug-22-top-2-logistic-regression-cv-fe

3. Brief introduction

在一開始，我使用了 Logistic Regression 作為 model，不過，差了 baseline 一些些，因此，我使用了不同的 linear model 作為 model，最後，總共嘗試了 elastic net、elastic_netcv、AdaBoostClassifier、XGBClassifier、lightgbm。

4. Methodology

1.Data pre-process:經過觀察我發現某些的特徵跟 test data 中的關聯性趨近於零，因此，我將那些 features 給刪除了。

2. Model architecture:在 logistic regression 的部分中，我自己試調了一些 Hyperparameters，但都會在 baseline 附近震盪，因此，我決定使用 elastic net 作為突破 baseline 的 model，然後在此之上我又發現了 elastic_netcv，他是基於 elastic_net 而成的 model，不過它的 Hyperparameters 可以使用一個 range 的值去給，也就是說支援 cross validation 的功能，而效果比起 elastic_net 更好一點。

3. Hyperparameters

```
ElasticNetCV(alphas=[0.0001, 0.001, 0.01, 0.1, 0.3, 0.5, 0.7, 1], cv=5,
             l1_ratio=[0.01, 0.2, 0.4, 0.6, 0.8, 0.9, 0.91, 0.92, 0.93, 0.94,
                       1])
```

5. Summary

這次的 project 我花了蠻多的時間去完成了，這次的 Project 主要的難點在於 data 的預處理，因為需要理解 features 中的意義。而在處理完 data 後，model 的選擇也是一大課題，花了蠻多時間去選擇跟嘗試通過 baseline 的 model。此外，通過這次的作業，我也了解了像是 smote(平衡化資料)、Optuna(搜尋 best Hyperparameters 的工具，比起 grid searich 快很多)等在機器學習方面非常有用的相關知識，不過，可惜的是可能是自己的 data 預處理方式不佳，並沒有太好的結果。

6.some findings

這次的 project 中，我發現到了在 submission 的輸出，需要的是機率，也就是 kaggle 裡會有一個 threshold 去區分 0 或 1(但不是用 0.5)，因此，許多 Scikit-

learn 現成的 Model predict 的部分(預設用 0.5 去切)，因此表現也不佳。

7.best result:

8.Comparisons of different approaches:

Model	Private Score
Logistic regression	**0.58987**
AdaBoost	**0.57345**
Elastic net	**0.59006**
Elastic net cv	**0.59007**
xgboost	**0.58082**
lightgbm	**0.58594**