DS 540 TERM PROJECT

- Introduction

In this project, I worked with a large loan data and performed several machine learning models to predict the target value. To increase the performance, I have used cross-validation and hyperparameter tuning methods on the best performed model. In the second part, I created different clusters in train dataset with k means clustering method.

- Data Preparation

First, I started with data cleaning by checking duplicates, NaN values and exclude them from the dataset. I also excluded column “SK_ID_CURR” which might have an unwanted impact on models. Then, I turned categorical variables into dummy variables. Before splitting the dataset into train and test, I checked if there was an imbalance in target feature. I detected a visible difference in numbers of two different target features. Because of that, I used SMOTE to balance this difference. Lastly, I split the dataset into train and test to proceed to model running phase.

- Running Models

I tried 4 different models and measured their performance. The models I have used are decision tree, random forest, gradient boosting model and XGBoost. The performance metrics used were accuracy, AUC and F1 score. Regarding these metrics, XGBoost was the best performer among these models. That’s why I decided to continue with XGBoost for the later phases.

Performance of XGBoost Model:

```
Accuracy: 0.9576137624861265
AUC Score: 0.9576335328919964
F1 Score: 0.9559286827072875
```

- **Cross-Validation and Hyperparameter Tuning**

I used XGBoost in this phase. For cross-validation, I chose cv as 10. After I checked mean values, I have observed that mean accuracy and F1 scores were slightly lower than default XGBoost model. However, there was a visible increase in AUC score.

Performance of Cross-Validation Method:

```
Accuracy Scores: [0.95585461 0.95937847 0.95674251 0.95677026 0.956798   0.95818535
 0.95549266 0.95629734 0.95735176 0.95757374]
AUC Scores: [0.97778714 0.97948724 0.9784794  0.97765413 0.9790814  0.97886357
 0.97837195 0.97835619 0.97843975 0.97869371]
F1 Scores: [0.95399208 0.95779764 0.95497473 0.95495548 0.95504807 0.95655182
 0.95360139 0.95445739 0.95557932 0.95581308]
Mean Accuracy Score is:  0.9570444692439029
Mean AUC Score is:  0.9785214477265647
Mean F1 Score is:  0.9552770990602639
```

After that, I used hyperparameter tuning method. With grid search, I was able to observe performances of several different parameters. Best parameters I could find were:

n_estimators=200, max_depth=7, learning_rate=0.1, subsample=1

In this setting, accuracy (0.9581) and F1 scores (0.9563) were higher than cross-validation method. However, AUC score was lower than cross-validation method although it was still higher than the default model.

- **Clustering**

I used train dataset for clustering process and chose k means clustering method. The plot of elbow method is below: