Intro to DS 2019 HOMEWORK 2 - Churn Prediction

https://www.kaggle.com/c/introduction-to-data-science-nyu-spring-2019/

Goal:

Given a training set of past churn data, your goal in this homework is to predict whether a person will leave the network (churn) or stay. You may use packages like Numpy, Pandas, SciPy, scikit-learn. You should work in a team of **2 people**.

Steps:

1. Join the competition and form a team
    a. Create a Kaggle account
    b. Join the competition by clicking this link
    c. Search your teammate by his/her 'team name' on the 'Team' Tab.
    d. Request Merge
    e. Rename the team name into ***netid1_netid2*** (eg. ayw255_kw142)
2. Train and build a model based using the training data (**train.csv**): any model studies in class is acceptable, such as a decision tree, logistic regression, or support vector machine.
3. Use your model to predict the outputs for the test data.
4. Upload your prediction as a file to the kaggle competition to be evaluated and ranked. You can use sample_submission.csv as a template for submission. You can at most submit 3 times a day.

Data Description:

Each row represents a customer of the network, with the features for each customer described in the 'Data' tab in Kaggle competition page. The data consists of 20,000 customers, split into 90% (18,000) for training data and the remaining 10% (2,000) as test data (holdout). You can find the labeled training data in 'train.csv' and unlabeled test data in 'test.csv'. You can find the features of each customers in the Dataset tab in Kaggle competition page.

Submission Format (in kaggle competition) [Due March 6]:

For every student in the competition, submission files should contain two columns: 'ID & 'LEAVE'. ID will have values from 0 to 1999 and 'LEAVE' column should have predicted outputs (0 or 1) of 2,000 test samples of test.csv

Note: You can download 'test_submit.csv' to know how should your submission file be.

Submission Format (in NYU Classes) [Due March 6]:

In addition to the test result you submit to Kaggle competition page, you must also submit

either a Jupyter Notebook or a Python Script with the code for your preprocessing steps and modelling steps to NYU Class.

Evaluation

- **Metrics: AUC**
- **Private & Public Leaderboard:** To prevent overfitting the test set, the test data is split to 2 parts, where 30% is used to rank Public Leaderboard, and 70% for Private Leaderboard. You will only see Public Leaderboard ranking, but the final ranking will be based on Private Leaderboard.

Your will be graded based on your rank on the leaderboard in Kaggle Competition as well as your code uploaded in NYU Class.