CSCI 4146/6409 - Process of Data Science (Summer 2023)

Assignment 3: Ensemble Methods

This assignment is designed to deepen your understanding of ensemble methods and the hyperparameter optimization process. This time, we moved from Airbnb to Credit Card Fraud Detection dataset due to its applications/relevance for ensemble methods. By using this dataset, you will be required to develop models, evaluate their performance, and interpret the outcomes. Successful completion of this task will enhance your competency in the deployment of advanced machine learning techniques in real-world scenarios.

Due Friday Jun 16, by 11:59 pm - see submission details below.

Dataset: Credit Card Fraud Detection- https://www.kaggle.com/mlg-ulb/creditcardfraud

1. [1.0] Data Preparation
    a. Display and review the Train, Test, and Original Datasets. Make note of any peculiarities or unique aspects in the data that may require attention. [0.25]
    b. Merge the Train and Original datasets. Provide a justification for this action. [0.25]
    c. Identify and remove any irrelevant features. Substantiate your choices. [0.25]
    d. Divide the dataset into features and targets. Detail your method and reasoning behind it. [0.25]

2. [2.0] Modeling and Understanding Ensemble Methods
    a. Train a baseline model using a single algorithm (e.g., a linear model such as logistic regression) and assess its performance. [0.5]
    b. Implement Bagging using a set of homogeneous models (e.g., Decision Trees). Incrementally increase the number of models (from 1 to 10) and analyze the impact on performance. [0.75]
    c. Implement Boosting, also using a set of homogeneous models. Similar to Bagging, increase the number of models incrementally and observe the impact on performance. [0.75]

3. [1.0] Hyperparameter Tuning and Model Quality Evaluation
    a. From the models you have developed in question/step 2, select one and proceed with hyperparameter optimization. Discuss the selection of parameters you've decided to tune and the range of values you are considering. [0.5]
    b. Evaluate your optimized model using relevant metrics and graphical representations (e.g., performance graphs). The selected metrics should

effectively demonstrate the performance of the model as a function of the hyperparameters. [0.5]

4. [1.0] Predictions and Submission
    a. Make predictions from the model that performed best in question/step 3. Also, provide an analysis of the prediction results. [0.5]
    b. Prepare your final submission. This should include a .csv file of your results. In addition, reflect on your results, their implications, and any potential improvements that could be implemented. [0.5]

Submission Details

Submissions should be made through Brightspace, adhering to the due date and time specified under the Assignments section. To prepare your assignment solution, make use of the provided assignment template notebook on Brightspace. The detailed requirements for writing and coding can be found within the evaluation rubric document available on the platform. Keep in mind that questions will be graded individually using letter grades, with their respective weights indicated in parentheses.

You may complete the assignment individually or with another individual. In the case of a pair submission, only one student should submit the assignment on Brightspace. Be aware that plagiarism detection tools will be employed to identify any instances of cheating or copying in both your code and the accompanying PDF.

Your submission should consist of a single Jupyter notebook as well as a generated PDF that contains the compiled results generated by the notebook. This PDF should include both the code and the results as part of the final printout. Name your files as follows:

- A3-<your_name1>-<your_name2>.ipynb and
- A3-<your_name1>-<your_name2>.pdf.

Failing to submit both files will result in a zero mark for both students.