DS 310 Machine Learning

Project #4: Drug User Prediction

Fall 2020 / Amulya Yadav

NOTE: This is a two-member team project.

Task: This is a multi-class classification on the homeless youth dataset. The task is to predict what kind of drug the homeless youth use, or whether he/she does not use any drug. There is no test set in this dataset, so your target is to maximize the 10-fold cross validation loss.

Just a reminder: you are required to predict the behavior of homeless youth, so the first thing you should do is to remove those who are not homeless.

Dataset: Homeless youth dataset (we have seen it before). You can download it on Canvas.

Note that all datapoints which have missing labels can be discarded. On the other hand, NaN or any other kind of spurious feature values in some datapoints should be replaced (and missing values should be filled) with some reasonable approach; Or you can simply choose to discard the feature containing spurious NaN type values altogether. Please make sure you have enough data for training.

Deliverables: Each team needs to accomplish the following tasks:

1. Measure Score: At the end, we will evaluate the AUC as the team's final submission result.

2. Jupyter notebook: Each team posts its final model in Jupyter notebook, named as `submission.ipynb`.

3. Project Report: Project report in PDF that contains all the details of the major steps of the project.

Grading Rubric (100 points)

- 80 points: performance (65 points) and ranking (15 points) , measured in AUC. 50 points for baseline performance and 15 points for better performance.

- 10 points: submitted Jupyter notebook, `submission.ipynb`.
- 10 points: quality of report.

Grading Exceptions

The goal of this project is to encourage you to emerge yourself in a real-life data science process to learn the materials better in a hands-on fashion. **DO NOT CHEAT IN ANY WAYS.** If you do not attempt to get your hands dirty in this project, you will not learn much.

All team members need to participate in the project and contribute as much as he/she could. Each is expected to bring different skills to the table, i.e, some are good in programming, while others in analytics, etc. Find out what you can contribute within the team, and play the team work. While both members get the same points by default, I will intervene in an exceptional situation (e.g., one member does not participate at all).