EE 510: Machine Learning Theory & Algorithms **Spring 2020**

Mini Project 3

Due: June 5 (official) / June 12, 2020, 11:59PM PT

Student Name: *Instructor Name: John Lipor*

Note: The "official" due data will be June 5, meaning I will not give feedback or answer questions after that date. However, you are welcome to take the extra week to continue working on your project, turning it in on June 12.

Introduction

On this final mini project, instead of predicting who already died or who will die, you will instead predict who is likely to kill somebody by working with the Porto Seguro's Safe Driver Prediction challenge on Kaggle [1]. The task for this data is to predict the probability that an auto insurance policy holder will file a claim.

Notes:

1. DSS rules apply for all mini projects.
2. Be sure to cite any sources you used, including/especially Kaggle notebooks.
3. There are a number of new challenges in this dataset that will require special care. You may wish to consider notebooks for the Credit Card Fraud Detection dataset [2] as well.
4. While our boosting method of choice has been `XGBoost`, you may also wish to consider reading about or using `LightGBM` or `CatBoost`.

Requirements

You must create a report in either LaTeXor a Jupyter notebook that contains the following sections:

1. Problem description
   - Explain what data you have to work with, what algorithms you will use, what your goal is, and why anyone should care.
2. Exploratory data analysis (EDA)
   - There are many excellent tutorials on EDA online. Be sure to cite whichever ones you found helpful and comment on them in Sec. 6.
   - Be sure to make any significant findings stand out and try to keep this concise.
3. Challenges
   - What were the challenges you encountered when applying machine learning to this dataset?
   - Did these challenges mainly result from the data? From results? Installing libraries to perform preprocessing?

4. Approach
    - Provide a thorough but concise description of your approach, including (but not limited to) your approach to wrangling, preprocessing, and improving the performance of your predictor.
    - This should be much clearer than what is typically found in Kaggle notebooks, but make sure it stays concise. A summary paragraph at the beginning of this section could be helpful.
    - Since we have now formally studied model selection, you must use either a validation set or cross validation to tune your model parameters. Be sure to describe how you went about this and why.
5. Evaluation and summary
    - **(NEW)** For this project, you must take special consideration as to which metrics you use to evaluate your approach. In addition to evaluating your performance, you should explain why the competition has chosen the specific metric and which other metrics would or would not be worth considering (and why).
    - **(NEW)** In addition to your evaluation plots, you should make a late submission to the competition and report your score.
    - Consider diving into the data to see if you can determine any trends regarding which points were misclassified.
    - Likewise consider if there were any features that were particularly important or unimportant.
    - Summarize your solution, describing what worked, what didn't, what the main limitations are, and any general conclusions about the dataset.
6. What I learned
    - Describe the main skills/tools that you learned and used for this project and how you learned them.

Grading

The grading breakdown is below. Each section will be graded according to technical correctness, effort, and creativity. Note that clarity of writing is a major component. You should put yourself in the place of writing for a boss or senior in a workplace. If your writing is terrible, they will assume your work is terrible.

Item	**Percentage**
Description	5%
EDA	20%
Challenges	10%
Approach	30%
Evaluation	20%
What I learned	5%
Clarity/conciseness of written communication	10%

References

[1] (2018) Porto seguros safe driver prediction. [Online]. Available: https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data?select=train.csv

[2] (2018) Porto seguros safe driver prediction. [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud