INFSCI 2160: Data Mining
Python Lab

For the Python Lab, we will use a real, anonymized dataset provided by UPMC regarding surgical procedures between June 2017 and June 2018. The data is already divided into training and test populations for you. A data dictionary is available on Piazza. The data is not permitted for any distribution outside of class.

The goal of the lab is to develop an algorithm that can accurately predict patients that are at risk of a "long" length of stay post-surgery. There are many different definitions of an abnormally long LOS in literature. We will define a long LOS as > 5 days post-surgery.

In a group of 1 – 2 people, your assignment is as follows:

- Develop a model(s) to predict patients that will have an abnormal length of stay.
    - Final models will be evaluated by AUC.
    - Team with highest AUC within 0.03 points between train and test will be the winners and receive 100%.
- Display your results and evaluate your model(s).
- Develop a strategy as to how you would classify patients in terms of risk (Low vs. high; low vs. medium vs. high)
- Explain why or why you would not deploy this model.
    - If you think the model should be deployed, explain how you would flag high-risk patients to physicians

All code should be included in an .html or .pdf document.

You also must turn in a 1 – 2-page essay explaining how you performed your analysis, the evaluation of your model(s), how you stratified patients in terms of risk, and deployment strategy if applicable. If you decide not to deploy, explain why.

Indicate your final AUC on your training and testing data on the top of your essay submission.

The lab will be graded as a weighted homework assignment (50 points). It is due at 5:59 PM EST 5 weeks after it is assigned (week of October 28th)**.**

Grading Rubric:

Python:

The goal of the project is to generalize well on both your training and test data. You will be graded on the following in your code:

- **5 points:** Choosing algorithm sthat solves the type of problem at hand
    - Continuous outcome, binary outcome, or multinomial outcome
- **10 points:** Trying different models to predict the outcome (2 minimum)

- **5 points:** Creativity
    - Creating of new features
    - Implementing cluster analysis, trying new models, combining models, etc.
    - Deployment strategy
- **5 points:** Proof of deployment strategy
    - Example of confusion matrix or matrices, or other deployment strategies
- **5 points:** Appropriate evaluation between training and testing
    - Correct interpretation of results between train and test

Essay

- **5 points:** Description of which algorithms you tried and why
- **5 points:** Description of any data preparation or feature engineering you completed
- **5 points:** Interpretation of raw results (over or underfit, confidence in predictions, etc.)
- **5 points:** Deployment ideas and explanations