CASE STUDY 6

Hollie Gardner

PROJECT GOAL

Using the Higgs-Boson dataset, build a model to maximize accuracy of predictions to detect the presence of a particle.

OBJECTIVES FOR ACHIEVING GOAL

- Use a neural network model
- Evaluate highest accuracy score in predictions
- Evaluate precision and recall
- Recommendations for next steps

DATASET OVERVIEW

- 7 million observations with 29 features
- First feature:
  - "# label" column is the dependent variable
  - Binary: Signal = 1 vs. Noise = 0
- Balance between Classes: Almost 50/50 split between observations that are signal and noise

DATASET OVERVIEW

- The remaining 28 features:
  - Columns 2-22 are kinematic properties measured by the particle detectors in the accelerator. (UCI website)
  - Columns 23-29 are functions of the first 21 features, which are high-level features derived by physicists to help discriminate between signal and noise

DATA PREPARATIONS

- Converted target column, '# labels', from float to integer
- No need to worry about class imbalance with 50/50 split
- 80% training; 20% test data split

MODEL OVERVIEW

- Utilizing a Neural Network Model:
  - 1 input layer
  - 2 dense layers
  - 1 output layer
- Parameters:
  - Optimizer: Adam
  - Loss: Binary Cross Entropy
  - Epochs: 5
  - Batch Size 10,000

MODEL EVALUATION

- **Training Set:** 80.99% - 85.42% Accuracy
- **Test Set:** 86% Accuracy with 50% threshold

WHERE IS THE MODEL STRONGEST?

89 out of every 100 particles are correctly labeled

WHERE IS THE MODEL WEAKEST?

15.8 out of every 100 particle predictions are actually just noise.

NEXT STEPS

- Utilize high performance computing power to tune parameters to potentially achieve greater accuracy
- Work towards reducing the number of Noise datapoints that were classed as Signal.
  - Is this more important than some particles being classed as Noise?
- Closer look at which features are more likely to be correlated with particle status than noise