Problem Set 8

Data Science 602: Data Analysis and Machine Learning

Spring 2022

If you completed the Extra Credit assignment for Week 4, you prepared a dataset to predict the hourly temperature given other weather observations.

For this problem set, use the dataset you prepared for this assignment, or refer to the solution as needed. **From this dataset, remove the prior temperature observation.**. Split the data into test and training datasets.

1. **Regressors**. Build at least 3 regressors using different algorithms to predict the temperature. At least one regressor should implement a tree-based algorithm (random forest, or gradient boosted tree/xgboost.)

2. **Cross-validation** Use cross-validation to test each algorithm, and select the estimator with the highest accuracy score.

3. **Feature importance** Use one of the tree-based models to evaluate feature importance. Which features are the most important?

4. **Residuals plot** For the best model selected above, show a residuals plot ($\hat{y}$ vs. $\hat{y} - y$). Does the residuals plot show evidence of uncaptured explanatory information?

5. **Evaluation** Train the model with the highest accuracy score with the full training dataset. Evaluate the R^2 score for the test data against. Does the model demonstrate predictive validity?