WHARTON FINANCIAL ANALYTICS

Lab Questions

Diamonds

1. Read, verify, and clean, as necessary, the diamond data from Blue Nile (20-diamonds.pq). Each row corresponds to a unique diamond offered for sale by Blue Nile and the columns contain information on the diamond:

    a. Shape,

    b. Cut,

    c. Color,

    d. Clarity,

    e. Carat, and

    f. Price.

2. Split the sample into training and test datasets. Test your sample split to ensure the training and test samples are "similar."

3. Perform exploratory data analysis keeping in mind that one of the goals of the lab is to construct a model of diamond prices.

4. Estimate a linear regression of diamond prices on the diamond features and interpret your results. You should estimate it on the training data sample.

5. Explore the predictive ability of the following regression models:

    a. Linear regression,

    b. Lasso regression,

c. Decision Tree,

d. Random Forest, and

e. Gradient Boosting.

Make sure to employ cross-validation on your training sample in assessing the different models. For each model, you may also want to optimize over different hyperparameters. This process should lead to the selection of a preferred model (on the training data).

7. Using your preferred model from the previous problem, predict the diamond prices in the test data. How does the model perform? How does this performance compare to the performance on the training data? Is the model performance good enough to warrant its use for Kerry's purposes?

8. Use your preferred model to predict the prices of *all* diamonds – in both training and test datasets. Compute the pricing errors in absolute and relative (in logs) terms. Why might Kerry find this information useful?