Assignment 2

Tree-Models and XGBoost: Range Rover Pricing Prediction

Instructions. This assignment focuses on predicting selling prices for used Range Rovers using tree-based machine learning models. Explore the Week 2 web app and answer the questions below. Stick to 3 sentences maximum per question. Upload your responses to BruinLearn as a PDF by the end of the day (11:59pm PT).

Question 1. Train a single regression tree with max_depth=8, then train a Random Forest with n_estimators=100 and max_depth=8. Use max_features = 0.7. Report and compare the training MAE and cross-validation (CV) MAE for both models. What do the results tell you about the value of “ensembling” (i.e., combining many model predictions)?

Question 2. Train a Random Forest with n_estimators=100 and max_depth=8. Look at the feature importance chart. Which are the three most important features? Are there any features that barely matter for prediction?

Question 3. Set max_depth=8. First train a Random Forest with n_estimators=100 and max_features=0.7. Then train an XGBoost model with n_estimators=100, learning_rate=0.1, and subsample=0.8. Describe how the training MAE curves differ between the two models (e.g., consider when the number of trees is very high or very low). Provide a short explanation for the difference in behavior.

Question 4. Train an XGBoost model with max_depth=6, subsample = 0.8, and the following four parameter combinations:

learning_rate=0.01, n_estimators=50
learning_rate=0.3, n_estimators=50
learning_rate=0.01, n_estimators=200
learning_rate=0.3, n_estimators=200

Report the CV MAE for all four configurations. How does the learning rate affect how many trees the model needs for CV MAE to stabilize?

Question 5. Carvana wants to use your best model to set listing prices for incoming Range Rover inventory. Looking at the data, prices range from $1000 to $135,000, but most cars cluster in the $20K–$50K range. For which vehicles would you trust the model's predictions most, and for which would you flag them for human review? For the segments where the model is weakest, is there an additional feature variable you would you want to collect to improve predictions?