Assignment 2: due Nov 09

Your solutions should be gathered together into a single pdf including code, results, plots, comments and other relevant outputs.

Question 1 (10 points)

Data: “Ames Iowa Housing Prices Dataset” (https://www.kaggle.com/datasets/emurphy/ames-iowa-housing-prices-dataset)

a) Use multiple linear regression to fit the ‘SalePrice’ column using data in the file ‘train1.csv’. Select up to 10 variables for your model. The choice is not key at this stage but a reasonable choice of a subset it needed with a brief selection exploration. Report the RMSE upon using the ‘test1.csv’ dataset.
b) Fit a decision tree to the set of independent variables chosen above using ‘train1.csv’. Report the RMSE from using ‘test1.csv’. (library(Metrics); rmseDt = rmse(actual=testpricestarget, predicted=predictedprices)
c) Fit a Random Forest to the set of independent variables chosen above using ‘train1.csv’. Report the RMSE from using ‘test1.csv’.
d) Fit an XGBoost model to the set of independent variables chosen above using ‘train1.csv’. Report the RMSE from using ‘test1.csv’.
e) Which parameters of XGBoost can you tune which change the RMSE from your experience?
f) Comment on the quality of the fits produced in each case? What can you conclude from this predictive task about the nature of the sales price prediction?

Question 2 (10 points)

Data: “Pokemon for Data Mining and Machine Learning”
(https://www.kaggle.com/datasets/alopez247/pokemon)

a) Randomly select 70% of the rows as training data and the remaining rows as testing data.
b) Fit to the training data a decision tree, random forest and XGBoost model where the dependent variable is the ‘Type_1’ column.
c) Report on the accuracy for predicting ‘Type_1’ on the testing rows for each model.
d) Produce a confusion matrix for each model and then comment on the model performances. Discuss in terms of the confusion matrix and accuracy.