Interim Report

Progress started off a bit slow while I learned how to create a decision tree. I'm a little behind from where I'd like to be. Overall the project has been rewarding (when I get it to work haha). I understand the general format of how a Random Forest should be implemented and don't think it should take too long to code after I have the decision tree properly done.

I've completed all of my milestones up to this point (see below). Although I don't have my random forest implemented yet my decision tree is functioning with a few bugs. Since a random forest is essentially a bunch of decision trees taken into consideration I believe the bulk of the project is finished. My decision tree works on Entropy and Information Gain instead of the Gini index. I've read in a lot of places that information gain is more computationally intensive so I may swap over the Gini index after some testing, but for now it seems to be running at an appropriate speed.

Currently I'm getting an accuracy of ~80%, which isn't too bad considering. The best depth seems to be ~7. I've restricted the minimum number of features to be 20 in each node so it doesn't overfit. I'll definitely modify the restrictions when I move over to a random forest model to see what is better there. I may even create a function to loop over different restrictions to see which is the best if I have time.

Milestones/Dates:

1. ~~**2021-11-12:** Submit the proposal~~
    a. ~~Short gap here to finish the final problem set.~~
2. ~~**2021-11-23:** Finish sanitizing the data for duplicates and errors~~
    a. ~~I could do this manually but I plan on building a function to do it automatically~~

       The tree building function takes into account "NA" values and deletes any duplicates
3. ~~**2021-11-24:** Feed in the data to several graphing functions to check for similarities and get a first look into what I will be working with~~
    a. ~~I'll most likely use some sort of bar graph to check the average prices and I will compare each of the features to see which ones are included in the higher/lower priced homes.~~

       I have created several graphs to view the common factors in the data. This will help when I need to fine-tune anything. For example: removing features that don't contribute to the prediction to cut down on computational time and misclassification
4. ~~**2021-11-25:** Confirm the desire to use Random Forest and begin coding~~
    a. ~~Confirmation is just a step to research more on Random Forest/Ensemble Models~~

       A random forest model should function really well considering the predictions I'm getting with a decision tree already.
5. ~~**2021-11-29:** Have at least the pseudo-code written~~
    a. ~~Creating a skeleton of the project is important to get a scale of things and a larger picture~~

       I have very basic pseudocode that I will take into consideration when finishing up the random forest.
6. ~~**2021-11-30:** Write interim report~~
7. ~~**2021-12-01:** Finish and submit interim report~~

8. **New - 2021-12-01:** Finish the decision tree algorithm and test for predictions
9. **2021-12-02:** Finish implementation of Random Forest
    a. I hope to have this finished before a vacation on December 3
10. **2021-12-06:** Begin writing final report and finishing touches on code
11. **2021-12-08:** Finish writing final report and submit to Canvas