Garrett Yamane

Capstone Project 2 Proposal

Problem Statement

Moving can be a major hurdle for one's life. Whether it is the first time moving out on your own after graduating from college or if you are beginning to start a family and looking to settle down, buying one's own place can be quite expensive. Knowing the housing market and what to expect if you are moving to an unfamiliar place can be a challenging task. But what if there was a way to model housing trends and predict what you may have to pay? One model cannot accurately summarize every housing market, but being able to take two geographically separate locations and compare how house features are weighted differently when building housing price models can prove beneficial to both real estate agents and their clients when deciding where to move and what to buy.

Target Client

Real estate agents need to be very educated and aware of the housing market when helping clients buy a new place. Not only is it important to know how much a house is going to cost, but it is also helpful to know what factors will increase the price and what features carry the highest value for different locations. People moving into the suburbs may value a place with more bedrooms and living space for raising a family, whereas those moving into a larger city may want a view with newer renovations. Real estate agents for major cities and their clients can benefit from a study that targets two different geographical locations through predictive modeling: King County in Washington State and D.C.

Data

The data will come from two different datasets from Kaggle:

- King County, Washington Dataset
- Washington D.C. Dataset

There are 21.6k rows in the King County dataset, and 159k rows in the Washington D.C. dataset. However, the King Country dataset only contains house sales between May 2014 and May 2015, whereas the Washington D.C. dataset contains dating back to 1995. For consistency, I will only sample sale dates that both datasets have in common which will decrease the number of rows of the D.C. dataset down to around 14k. I will also look at the following features that each dataset has in common:

- Price
- Sale Date
- # bathrooms
- # bedrooms
- Living SQFT
- Lot SQFT
- Stories
- Condition
- Grade
- Year Built
- Year Remodeled

Approach

Initially, I would like to explore different parts of the data pertaining to King County and compare them to D.C. For example, the average living square feet or the range in selling prices for houses in each location, it is important to understand the similarities and differences between the two joined data sets. The features from each location will have a certain impact on my models, and I want to make sure that all parts of the data are explored in the storytelling.

Next, I will apply inferential statistics through tests such as a t-test for difference of means for selling prices between D.C. and King County. I can also calculate confidence intervals from the sample proportion to determine a range for where the true average selling prices may lie for each data set.

Lastly, I will build individual regression models for King County and D.C. to predict the selling price for houses using their shared features. My baseline model will begin with a linear regression model. Following, in my extended analysis, I will explore the use of ensemble models such as Random Forest regressors to use for prediction and compare the accuracy of these models to their baseline and among themselves. I will analyze the differences between the models built for D.C. predictions and King Country predictions to see how the shared features have different effects based on location.

Deliverables

As required, I plan on making a comprehensive Jupyter notebook that is annotated to make reading the code and visualizations as clear as possible. I also plan to prepare a final report and a presentation slide deck that addresses my goals for the project, the steps I took to solve the business problem I presented, and the results that I found from analyzing the data.