Kings County Housing Data Project February 2019

Ann Marie

FlatIron School

Data Science Course Module 1 Final Project

Our Goal:

- Help buyers in Kings County, Washington determine a fair price for houses in the neighborhood
  - Use rigorous statistical methods
  - Be transparent about our methods
  - Provide some conclusions and interpretations
- Our final product
  - A multivariate regression model with specified coefficients
  - Data needed for the regression should be easily available in a house listing
  - The buyer can use data and the model to come up with predicted price

Kings County Dataset

- 21,597 homes
- Categories of information available for forecasting:
  - Price
  - Date of sale
  - Square footage of living space, basement and overall lot
  - Number of floors, bedrooms and bathrooms
  - Condition and grade of the house
  - Location and proximity to the water
  - Year house was built and/or renovated
  - How often was the house viewed when listed for sale
  - All this information should be available when a house is listed for sale

Our Process

- We used the OSEMiN method to create our model:
  - Obtain: Pull data into our environment
  - Scrub: Pre-process data. Check for inconsistencies, missing data and outliers
  - Explore: Visualize data. Format data to be compatible with models
  - Modelling: Look at relationships between each predictive variable and price. Then look for overall relationships.
  - Interpretation: A few key conclusions

Our result:

- Our model explains 85% of the variation in the price of a house in Kings County
- If you give us the following information:
  - Square feet of living space
  - Number of bedrooms
  - Zipcode
  - Square feet of the lot
  - Square feet in the basement
  - Latitude (on the map)
  - How many times has the house been viewed at listing
- We could give you a predicted price range that will be correct 85% of the time

In more detail:

- No single variable is sufficient to explain prices.
- Square feet of living space has the highest predictive power. Larger living space means higher prices (all else equal)
- Larger lot means higher prices (but the magnitude of change is small)
- Location is very important to price
- If you have two houses with similar characteristics, the house with lower square feet in the basement will generally have a higher price
- If a house was viewed often during a listing, it can have a higher price.
- If you have two houses with similar characteristics, the house with average or less than average bedrooms will have a higher price. Perhaps this points to open layouts being more popular than before (or large master suites)

Caveats

- We have used 77 different variables to create our model. This in itself can cause a few problems:
  - The model could be overfitted. This means that its possible that a lot of the variables we are using are correlated to each other more than we can measure and there is bias in our prediction
- It’s possible that the data set we used is not a good representation of the entire universe of Kings County housing
- Explaining 85% of the variance on average does not equal explaining 85% for a unique data point. We can be generally right but wrong for some points.