King County House Sales

Morgan A. Keith

Business Understanding

I will create a model that evaluates a house’s characteristics that affect the price in the King County area.

Data

The data received is a data frame of characteristics about each house. Each characteristic is categorical or continuous data.

Categorical Data

- Date the house was sold
- Number of bedrooms
- Number of bathroom
- Number of floors
- Having a waterfront
- If the house has been viewed
- The house condition
- The overall grade of the house
- Zipcode

Continuous Data

- Price of the house
- Square footage of the house
- Square footage of the lot
- Square footage of the house apart front he basement
- Square footage of the basement
- Square footage living space of the nearest 15 neighbors
- Square footage land of the nearest 15 neighbors

Methods

- Clean data to fix missing values, delete duplicates, and replacing date with the house’s month sold.
- Exploring data for categorical and continuous characteristics. Checking distribution and correlation between variables.
- Baseline model
- Filtering data model
- Categorizing data model
- Geographic Sector model
- Adjust for multicollinearity model

Baseline Model

The baseline model shows our initial accuracy of the model, and compare adjusted models to it. It shows an intercept of 223800 and an r-squared of 0.692.

Filtered Data:Model 1

This model will be adjusted for outliers that can be extraneous info to accurately form a model.

Filtered columns:

- Bedrooms less than or equal to 7
- Bathrooms between 1 and 5
- Square Foot lot less than 25,0000
- Square Foot above basement less than 4,000
- Square Foot basement less than 1,500
- Square foot lot of 15 nearest houses, less than 60,000
- Floors less than or equal to 3
- Grade between 4 and 11

Categorizing: Model 2

For this model, taking into account the categorical variables. Each categorical variable is separated to represent a more detailed model for house attributes that affect the price. The intercept is at -8305000, and r-squared is .637. Though the accuracy for the model has decreased, this maybe due to each category being closely correlated. The model shows how each category effect the price increasing or decreasing the value of the house. An increase in bedrooms lead to a decrease in value while more bathrooms, a higher grade, and additional floors show an increase in value.

Geographic Sectors: Model 3

This model will separate the map into geographic sectors to show a difference in price depending on location. The intercept is 44,550,000, and r-squared .721. Here can see the price effects between each sector. Sector 4 shows the largest increase while sector 2 shows a decrease in price.

Final: Model 4

For the final model, unnecessary and multi-correlated variables were removed. The zipcode, latitude, and longitude were unnecessary. While square footage above basement, square footage of 15 nearest lots, bedrooms between 2 and 4, bathrooms between 2 and 3', 'grade between 5 and 7, and sector 6 of the geographical sectors were removed because of their multicollinearity.

Conclusion

The final model is more limited predicting price than previous models since there are some variables removed, but its predictions are more plausible with removing multicollinearity indicators. With p-values < .01, each variable is independently able to predict the final price.

The End