King County Housing Prediction

By: Steven Kyle

Project Overview

The purpose of this project is to create a Linear Model that can predict housing prices in King County, Washington.

The Three main questions we are going to ask today are:

1. Can we actually use a linear model to predict housing prices?
2. What features are important when predicting houses?
3. How accurate is this model?

Data and approach to modeling

The data that was used was provided by Kaggle.
After cleaning the data looked at:

- Looks at houses sold in 2014/2015
- Looks at 21,187 houses (after taking out outliers)
- Price of houses range from $124,000 - $1,730,000
- Average price of a house is $513,000

Went through 7 modeling iterations (7th being the final model)

- Multicollinearity
- Stepwise selection looking at p-values
- Feature manipulation (scaling, adding/subtracting)

Modeling Features that were used

Features from the dataset and there coefficients:

- ★ - Zipcodes (categorical, varies greatly -.05 to 1.04 logged price)
- ★ - Sqft_living (0.46 increase in logged price for every 1 log increase in sqft)
- View (0.0648 increase in logged price for every view)
- Grade (0.1126 increase in logged price for 1 increase in grade)
- Condition (0.0477 increase in logged price for 1 increase in condition)
- ★ - Waterfront (0.42 increase in logged price if it is waterfront)
- Renovated (0.0798 increase in logged price if it was renovated)
- Floors (-0.0188 increase in logged price for every floor)
- Bathrooms (0.0104 increase in logged price for every bathroom)

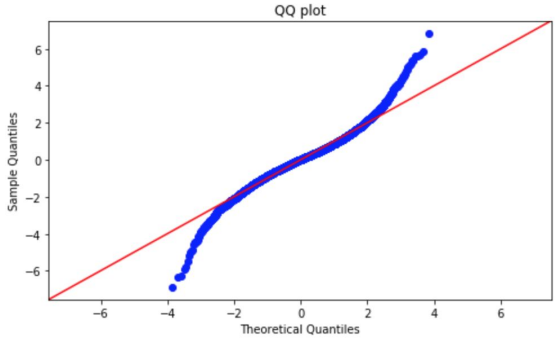
Engineered Features:

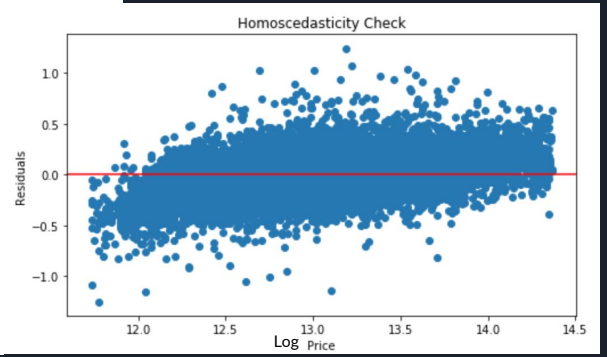
- ★ - Distance from the center of Seattle (-0.1757 increase in logged price for every log increase in miles)

Final Model Summary

Dep. Variable:	price	**R-squared:**	0.857
Model:	OLS	**Adj. R-squared:**	0.856
Method:	Least Squares	**F-statistic:**	1550.
Date:	Sun, 01 Nov 2020	**Prob (F-statistic):**	0.00
Time:	17:12:08	**Log-Likelihood:**	4786.2
No. Observations:	16949	**AIC:**	-9440.
Df Residuals:	16883	**BIC:**	-8930.
Df Model:	65		
Covariance Type:	nonrobust		

- The Adj. R-squared is 0.856. This means that 85.6% of the variance in price can be accounted for by the model.
- The QQ-plot shows that the residuals have a light tail distribution.





- The Homoscedasticity plot shows that the model tends to overpredict for homes in the lower end of the market and underpredicts for the homes at the higher end.

Results of error rate for Final Model

```
Train Mean Squarred Error: 10684571498.746878
Test Mean Squarred Error: 11163046254.656796
Train Mean Error: 103366.20094956996
Test Mean Error: 105655.31815605306
Difference in Mean Error: 2289.1172064831044
```

Error rate is around $100,000 for both train and test set.

There is only a ~2.2% error difference between train and test set, meaning that the model is able to predict on unseen data as accurately as it does on trained data.

The average error is 20% of average housing price.

Conclusion

In conclusion Linear Regression may not be the best modeling technique to use on predicting housing prices.
- Adj. R-squared of 0.856
- Avg error of ~$100,000 (~20%)
- Model overpredicts for lower end houses and underpredicts for higher end houses

In the end the final model used a logged form of the price (dependent variable) meaning that the trend is more likely exponential than linear.

The top four features that had the most impact:
1. Zipcodes
2. Sqft_living
3. Waterfront
4. Distance from center of Seattle

Future Steps

If I am allowed more time to work on this project I would like to look into school districts. I believe that it would be a more accurate representation of housing prices instead of zipcodes.

I would also like to explore other types of modeling to see if it will give more accurate results.

Acknowledgments

Thank you
-Yish
-Everyone in the online-ds-pt-070620

Resources:
-Kaggle
-GIS Open Data King County

Links and contact info:
https://github.com/stevenkyle2013/HousingPrediction
https://www.linkedin.com/in/steven-kyle-b3b771158/
stevenkyle2013@gmail.com