Real Estate Regression

An Exploration of the Price of Real Estate in King County

Carlos Garza

Business Problem

- Seattle based real estate company wants to automate their initial appraisal process
- Given information on a property, can we predict its selling price?

Data Utilized

- Sales Data for 21,597 property sales across King County
- Dependent Variable: Price

Independent Variables

Bedrooms	Bathrooms
Sqft Living	Sqft Lot
Floors	Waterfront
View	Condition
Grade	Sqft Above
Sqft Basement	Year Built
Year Renovated	Zipcode
Latitude	Longitude
Sqft Living 15	Sqft Lot 15

Baseline Model

- After minimal data cleaning and dealing with missing values, I used the StatsModels library to create a baseline linear regression model.
- Baseline Coefficient of Determination = .694

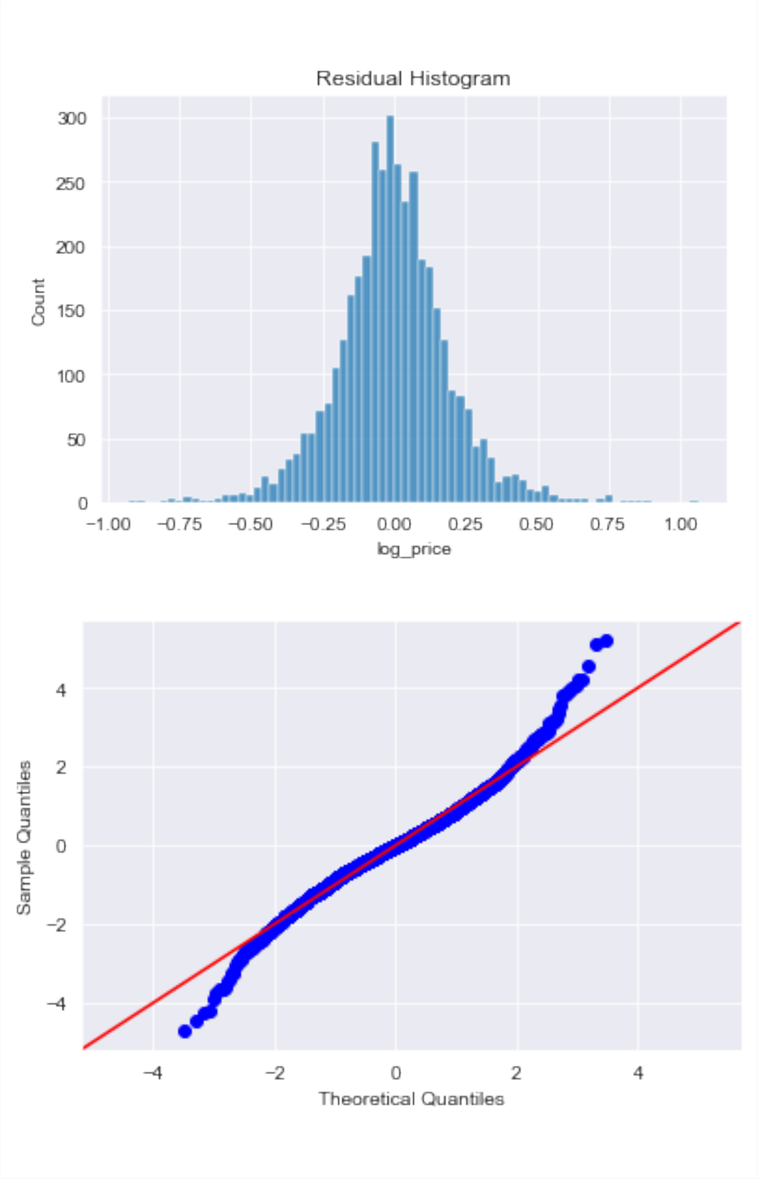
Tuning and Reiteration

- Various statistical techniques were implemented to normalize the data and achieve a better fitting model.
  - Removed statistically insignificant variables (p-value > 0.05)
  - One hot encoded categorical variables
  - Implemented log transformation and removed outliers to normalize variables

Results

Final Model

- Final model captures 86% of the data's variance
- Using a random Train-Test split, the mean squared error of training data vs test data was very similar
- Testing revealed residuals follow a normal shape, but show some heteroscedasticity



Deployment

- Model can be exported and used in appraisal applications

```
deployment — -zsh — 93×17
(learn-env) carlosgarza@x86_64-apple-darwin13 deployment % python3 house_price_prediction.py
-----
bedroom quantity: 3
bathroom quantity: 2.5
waterfront property (y=1, n=0): 0
renovated (y=1, n=0): 1
condition (1-5): 3
grade (3-13): 7
year built: 1951
sq.ft of living area: 2570
zipcode: 98125
latitude: 47.7210
longitude: -122.319
-----
estimated cost of property: $ 573557.55
-----
(learn-env) carlosgarza@x86_64-apple-darwin13 deployment %
```

Future Work

- Create GUI that takes data and delivers a price estimate
- Explore using multiple models or different transformations to achieve higher accuracy
- Modify function to accept an address as an input rather than taking zip code, latitude, and longitude separately

Conclusions

- Via multiple linear regression, A model was developed that captures 86% of our data's variance.
- Training and Testing data have similar MSE values, indicating a properly trained model
- Model has been deemed fit to be used in a new appraisal automation software for the client real estate company

Thank You

carlosgarza.io