Zillow Regression Analysis to Inform Purchase Decisions

March 19, 2023
By Andrew Levinton

Overview

- Business Problem Discussion
- Data used to conduct study
- Linear Model - Flaws, and plans for improvement
- Interpretation, and Recommendations
- Limitations
- Plan for Future work

Business Problem:

- Reduce cost
- Improve pricing
- Resale for profit

Business Understanding

- Focus: Real estate market of King County in Seattle.

- Determine opportunity cost for resale.

The Data - KC Housing Dataset - Link Below

https://info.kingcounty.gov/assessor/DataDownload/default.aspx.

The full list of columns with descriptions from the data can be located in the readme file of the repository.

- 30,155 Data Points
- After Nulls, outliers, and data cleaning approximately 27,446 data points remain.
- ~8% of the data is removed from data cleaning as a result.

- All house ages are within the years of 1900-2022.
- All house sales in the dataset are in the years of 2021-2022.

Features from Base Model

Location

Square Footage

Age

Condition

Additional Relevant features:

- On/near a waterfront?
- Have a nice view?
- Near a Greenbelt?
- Near a nuisance?

Additional Model Improvements - Engineered Data

- **Water Location by zip code:**
  - **Lake Sammamish**
  - **Lake Washington**
  - **Elliot Bay**
  - **Puget Sound**

- **School Ratings by Zip code**
  (scraped From GreatSchools API)

Great! SCHOOLS

Conclusion and Interpretation of Final Model

Most Positive Impacts on Price

- **Latitude - Houses Further North**
- **Water proximity - near Lake Sammamish**
- **Grade**
- **Square footage of home(no basement)**

Conclusion and Interpretation of Final Model

Negative Impacts on the Price

- **Older houses and lower grade.**
- **Houses near Lake Washington.**
- **Houses near Puget Sound.**
- **Floors**

Recommendations

- **Look near Lake Sammamish or further north.**
- **Buy older homes with higher grade/condition.**
- **Avoid houses near Lake Washington, Puget Sound.**
- **Avoid houses with more floors.**

Limitations of Model

- **Model explains 70.1% of dataset.**
  - **~30% of the data cannot be explained.**

- **Mean average percentage error of 23.2%.**
  - **On average data is 23.2% off of model prediction.**

Future Work

- It is worth revisiting the value of the homes on the remaining waterfronts.
- The views that are highlighted in the column_names.md documentation can be explored and onehotencoded and could be a potential candidate feature.
- A look at school Districts was initiated. Data was scraped from the GreatSchools API and will be further explored.