



Predicting Home Prices in Vancouver and Seattle

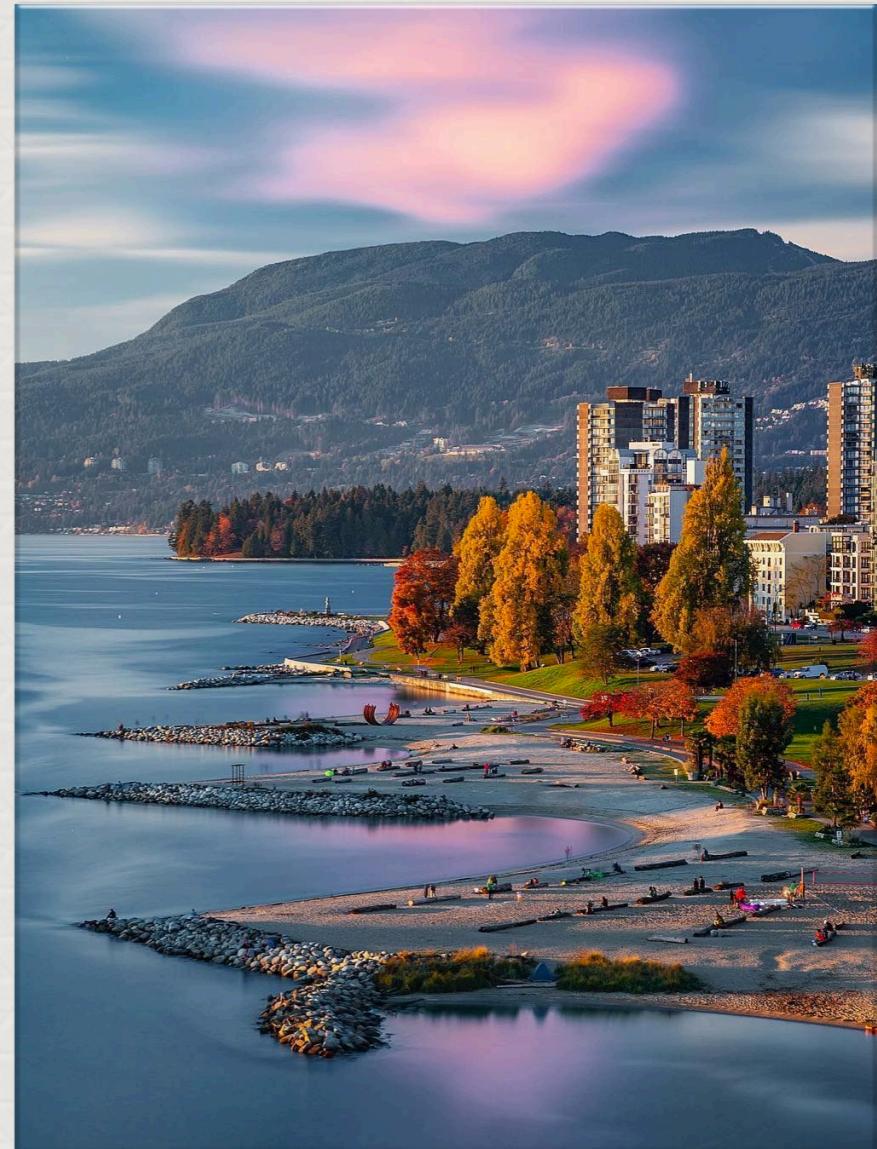
Web Scraping and Regression Project

Xinru (Flora) Cheng



Objectives

- ◆ Web API to compare predictions based on filters
— can expand to more cities
- ◆ How much does a certain categorical variable increase the home price?
- ◆ Predict home price for Vancouver and Seattle using a simple linear regression model



Methodology

- ◆ Data

- ◆ Scraped Vancouver data from redfin.ca
- ◆ Obtained Seattle data from redfin.com

- ◆ Tools

- ◆ Python, Scikit-learn, Pandas, NumPy, Seaborn
- ◆ HTML, Selenium for web scraping
- ◆ JupyterNotebook, Visual Studio Code

Data

REDFIN

- ◆ 450 rows, 27 unique features combined
- ◆ Also scraped URLs for each individual listing for Vancouver
- ◆ Vancouver median price ~900k USD
- ◆ Seattle median price ~700k USD

Vancouver Real Estate

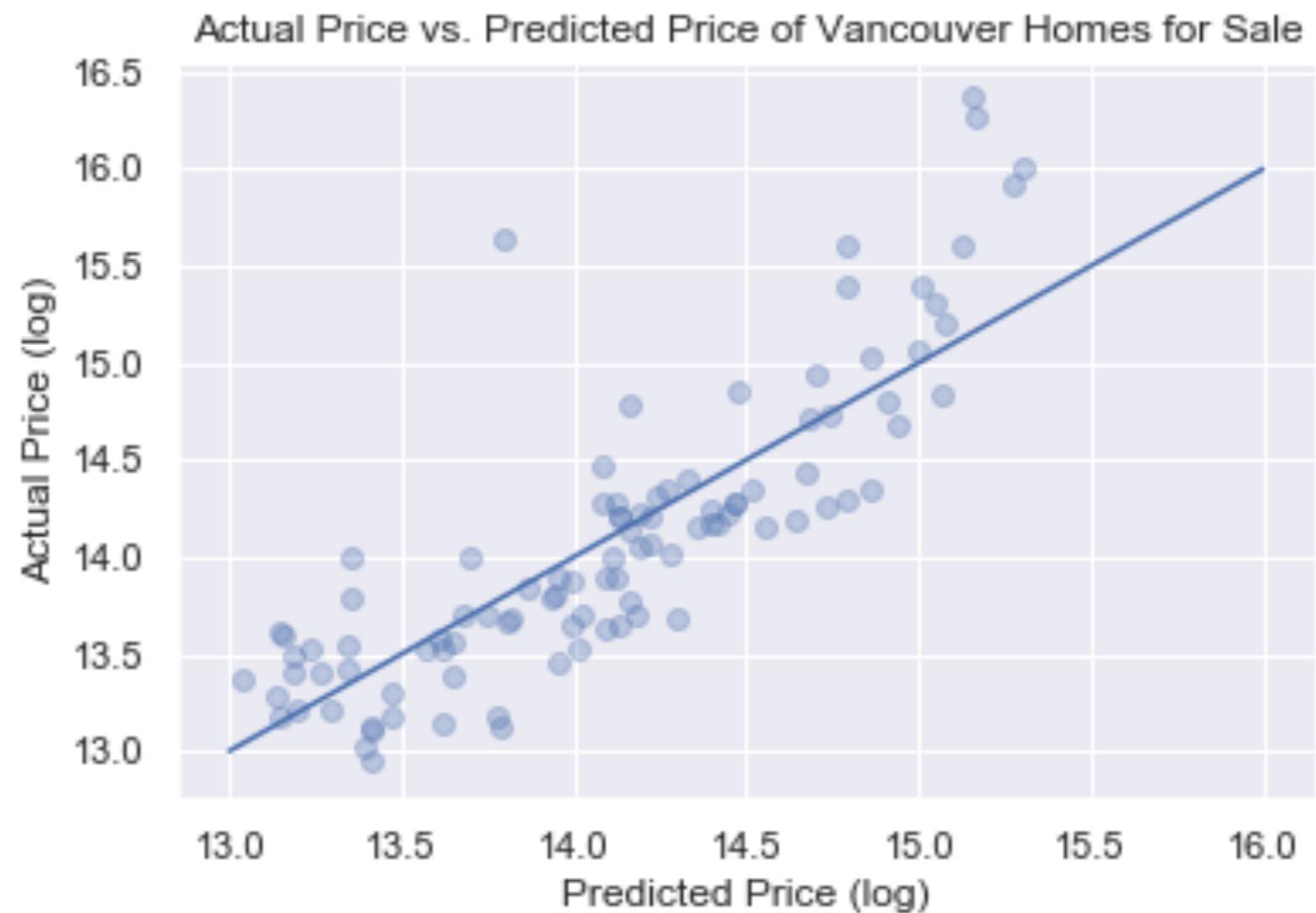
Price No min ▾ to No max ▾ More Filters ▾

Showing 20 of 3456 Homes · ⚡ Sort Photos Table

| Address | Location | Price | Beds | Baths | Sq.Ft. | \$/Sq.Ft. | On Redfin | |
|---|----------|-------------|------|-------|--------|-----------|-----------|--|
| R 2770 Burrard S... Fairview VW MLS# R2379412 Listed by: Redfin | | \$849,000 | 2 | 1.5 | 958 | \$886 | 64 days | |
| R 3939 Knight St... Knight MLS# R2395281 Listed by: Redfin | | \$448,000 | 1 | 1 | 592 | \$757 | 64 days | |
| OPEN 838 W Hasting... Downtown V... MLS# R2412257 Listed by: Angell, Hasman & Associates Realty Ltd. | | \$5,350,000 | 3 | 2.5 | 2,368 | \$2,259 | 33 hrs | |
| OPEN 2640 Point Gre... Kitsilano MLS# R2412070 Listed by: RE/MAX Treeland Realty | | \$1,488,000 | 2 | 2 | 1,212 | \$1,228 | 31 hrs | |
| OPEN 939 Expo Blvd ... Yaletown MLS# R2412083 Listed by: Oakwyn Realty Downtown Ltd. | | \$808,800 | 2 | 1 | 715 | \$1,131 | 2 days | |

Results - Vancouver

Small dataset
RidgeCV
 $K_f = 5$
 $R^2 = 0.713$



Results - Seattle



RidgeCV

Kf = 10

R² = 0.501

Reduced RMSE by ~25% after removing outliers

Results - Seattle



RidgeCV

Kf = 10

R² = 0.501

Reduced RMSE by ~25% after removing outliers

Conclusion

- ◆ We can get a decent prediction of home sale price with just a few features and a simple linear model
- ◆ Lots of tools to choose from even when dealing with a small dataset; room for improvement
- ◆ Possible bias in Seattle dataset

Future Work

- ◆ Get more data - from other websites; both numerical and categorical features
- ◆ Transform different features with different methods
- ◆ Compare effects of *individual features* on prediction
- ◆ More feature engineering - dummy variables
- ◆ Evaluate using metrics such as adjusted-R²

*“All models are wrong,
but some are useful.”*

– George E. P. Box (1919 - 2013)

Appendix

- ◆ Standardization of data:
 - ◆ Tried different scalers (Standard, MinMax, Robust, QuantileTransform, PowerTransform)
 - ◆ Also plan to transform different features in X differently
- ◆ Feature engineering:
 - ◆ Adding polynomial terms (degrees 2 and 3)
 - ◆ Adding interaction terms (ratio of number of bathrooms to bedrooms)

Appendix

- ◆ Removing outliers
 - ◆ Seattle Dataset (filtered out prices > 3 million USD)

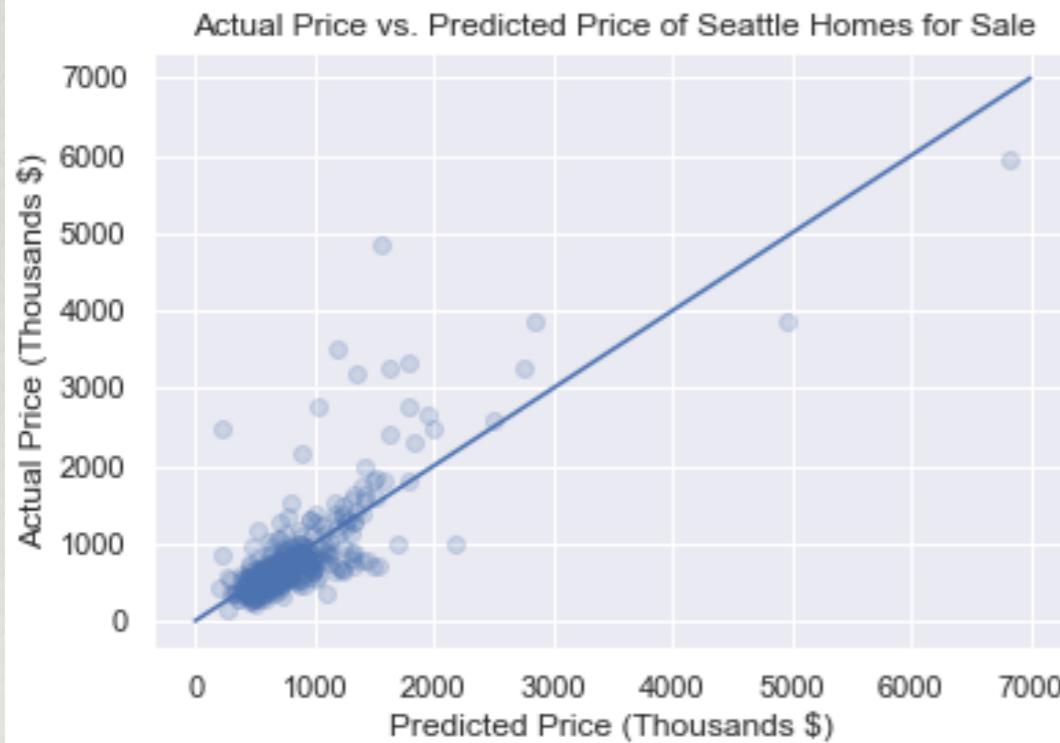


Before: RMSE 411k (~57% of median)

After: RMSE 298k (~40% of median)

Appendix

- ◆ Removing outliers
 - ◆ Seattle Dataset (filtered out prices > 3 million USD)



Before: RMSE 411k (~57% of median)



After: RMSE 298k (~40% of median)

Appendix

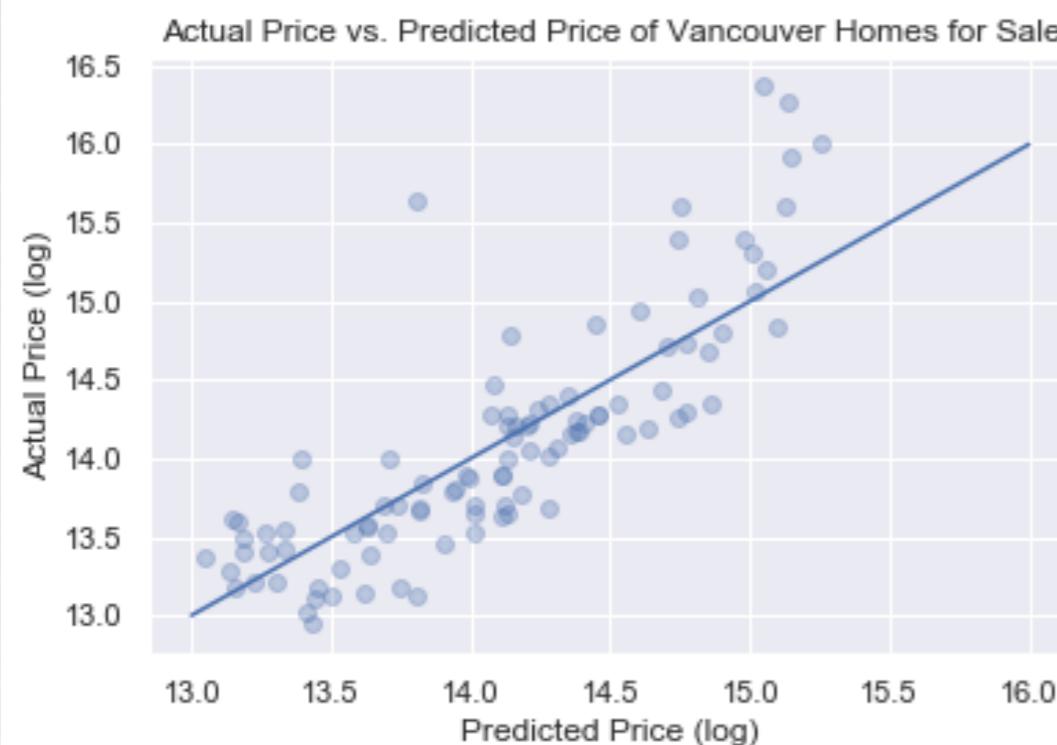
◆ Evaluation metrics

- ◆ Used **R²** (coefficient of determination) and **RMSE** (root mean squared error)
- ◆ Problem: if new features are added to our model, R² only increases or remains constant but it never decreases
- ◆ The **Adjusted R²** is the modified form of R² that has been adjusted for the number of predictors in the model. The adjusted R-Square *only increases if the new term improves the model accuracy*

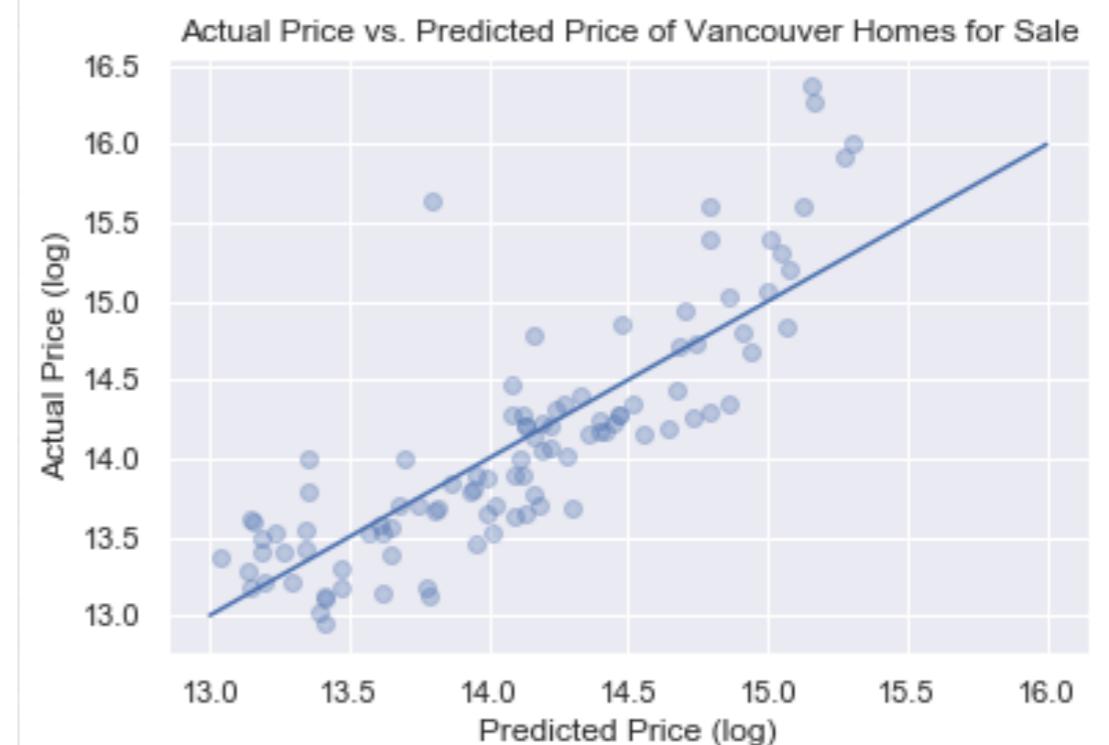
Appendix

◆ Regularization: Ridge vs. LASSO

- ◆ Have a small number of features in our dataset, decided to use Ridge, despite similar R² values for certain alpha values. Lasso would eliminate some features we want to keep.



RidgeCV, alpha = 10, R² = 0.698



LassoCV, alpha = 0.05, R² = 0.712

Appendix

- ◆ Regularization: Ridge vs. LASSO
 - ◆ Ridge regularization:
 - ◆ Minimizes cost function: changing alpha is controlling the penalty term. The higher the values of alpha, the larger the penalty and the magnitude of coefficients are reduced.
 - ◆ It shrinks the parameters, therefore it is mostly used to prevent multicollinearity.
 - ◆ It reduces the model complexity by coefficient shrinkage.
 - ◆ If our model is overfit (variance too high), regularization can often improve generalization error by reducing variance