UNION COLLEGE

Housing Price Prediction

An Nguyen

Advisors: Chris Fernandes, Nick Webb, Harlan Holt

Abstract

This project studies how house prices in 5 different counties in the US are affected by internal and external housing characteristics. Using data on 1,457 sold houses scraped from Zillow, Trulia, and Redfin, three prominent housing websites, we utilize both the hedonic pricing model (Linear Regression - LR), and machine learning algorithms (Random Forest - RF and Support Vector Regression - SVR), to predict house prices. Results show that SVR gives a better price prediction score than Zillow's baseline on the same dataset of Hunt County (TX) and RF gives close or the same prediction scores to the baseline on three other counties. Moreover, our models reduce the overestimated to underestimated house ratio of 3:2 from Zillow's estimation to a ratio of 1:1. We also identify the four most important attributes in housing price prediction across the counties as *assessment*, *comparable houses' sold price*, *listed price* and *number of bathrooms*.

Introduction

Housing websites such as Zillow, Trulia, and Redfin, provide estimations of houses' valuations based on the houses' characteristics, at no cost. However, these evaluations are not always accurate. For example, Zillow and Trulia only estimates about half of the houses within the 5% range of their actual sold prices [2][4]. Moreover, in our dataset, the ratio of overestimated to underestimated houses by Zillow is 3 to 2. Since over-valuating a house can lead to a longer time on the market and reduction of its original listed price [3][1], this project attempts to eliminate this overestimation problem. Lastly, this project seeks to find the most important housing attributes that appear across 5 counties.

Questions

1. Can our models outperform/come close to Zillow's prediction score?
2. Can our models reduce the overestimated to underestimated (OE:UE) house ratio?
3. What are the most important factors affecting houses' sold prices across 5 counties?

Methods

- Hedonic Pricing Model, or **Linear Regression** (LR), serves as the baseline model for this project, because LR is frequently used in previous works on housing price prediction.
- **Random Forest** (RF) is chosen as one of the machine learning algorithms for its ability to handle datasets with missing values.
- **Support Vector Regression** (SVR) is another machine learning algorithm used in this project because of its ability to find patterns in noisy data.

Data

- **Data Collection:** This project scrapes 1,457 sold houses and 35 housing attributes from Zillow, Trulia, and Redfin, using Python and Selenium. The 5 counties chosen for this dataset (as shown in Fig. 1) are among the worst performing ones, based on the percentage of houses whose Zillow's estimations fall within the 5% range of their actual sold prices. Moreover, these counties are in 5 different regions of the US and all the 3 housing websites used in this project have information on them.

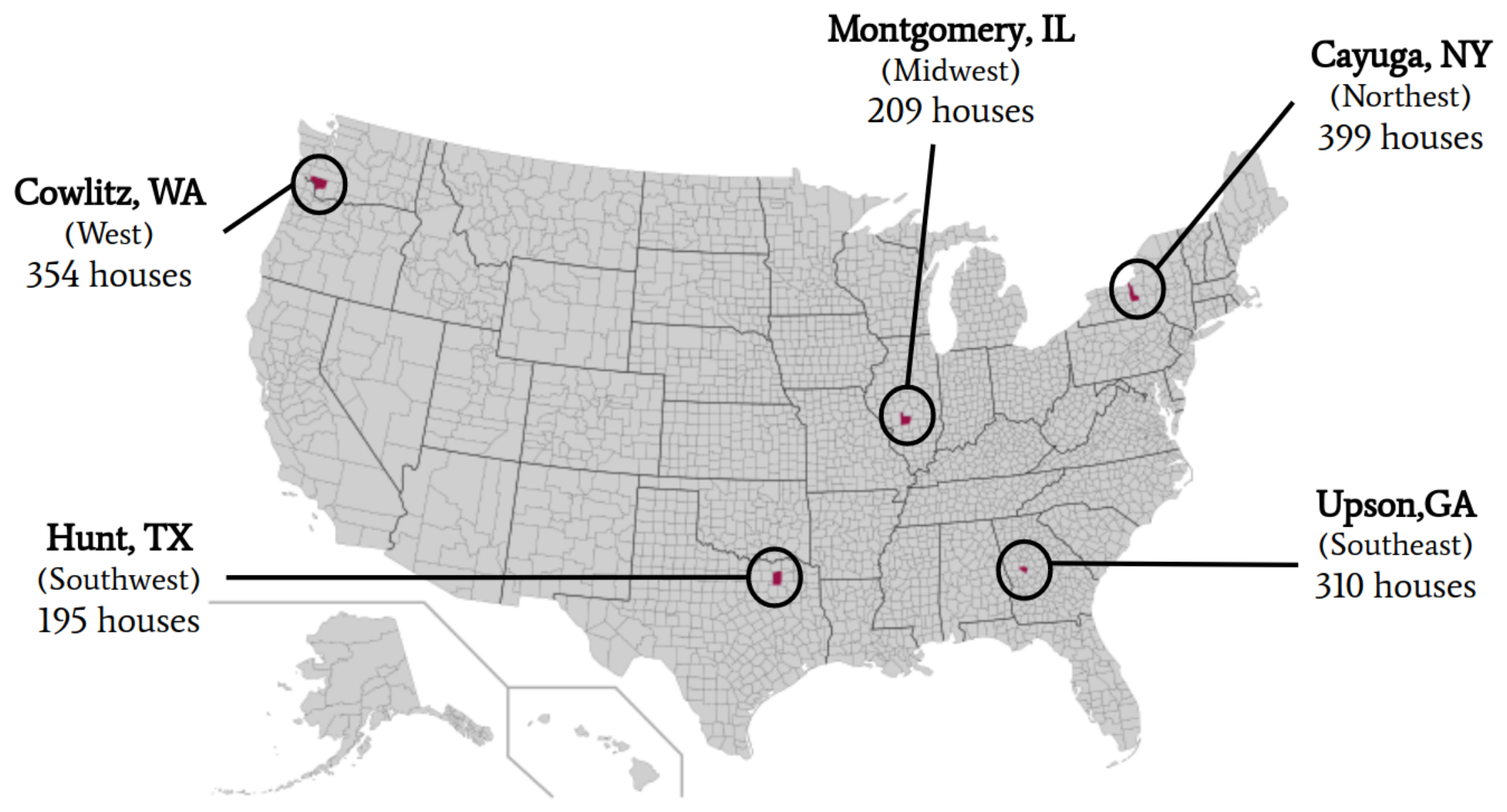

Figure 1: Five Counties On The US Map

- **Data Processing:** Inconsistency in data across 3 websites are unavoidable. Therefore, housing information scraped from these sites are compared against one another, in order to find the most consistent values. In addition, different units (ex: *sqft*, *acres*) could be used to measure the same house attributes (ex: *size*). Therefore, an extra conversion step has to be taken in order to uniform data units. All data processing steps are done in VBA (Excel).

Evaluation Baseline

We measure the prediction scores in this project as the percentages of houses whose predicted prices are within the 5% range of their actual sold prices. Therefore, the baseline prediction score for each county's dataset (as shown in Table 1) is computed as the percentage of houses whose Zestimate (Zillow's predicted prices) are within the 5% range of their actual sold prices. The Zestimate used in this project is the estimated value in the month right before a particular house's sold month.

County	State	# of Houses	Baseline (%)
Cayuga	NY	399	28.6
Montgomery	IL	209	21.1
Upson	GA	310	13.9
Hunt	TX	195	39.5
Cowlitz	WA	254	27.7

Table 1: Prediction Baselines

Results [1]

- **Prediction Scores:**

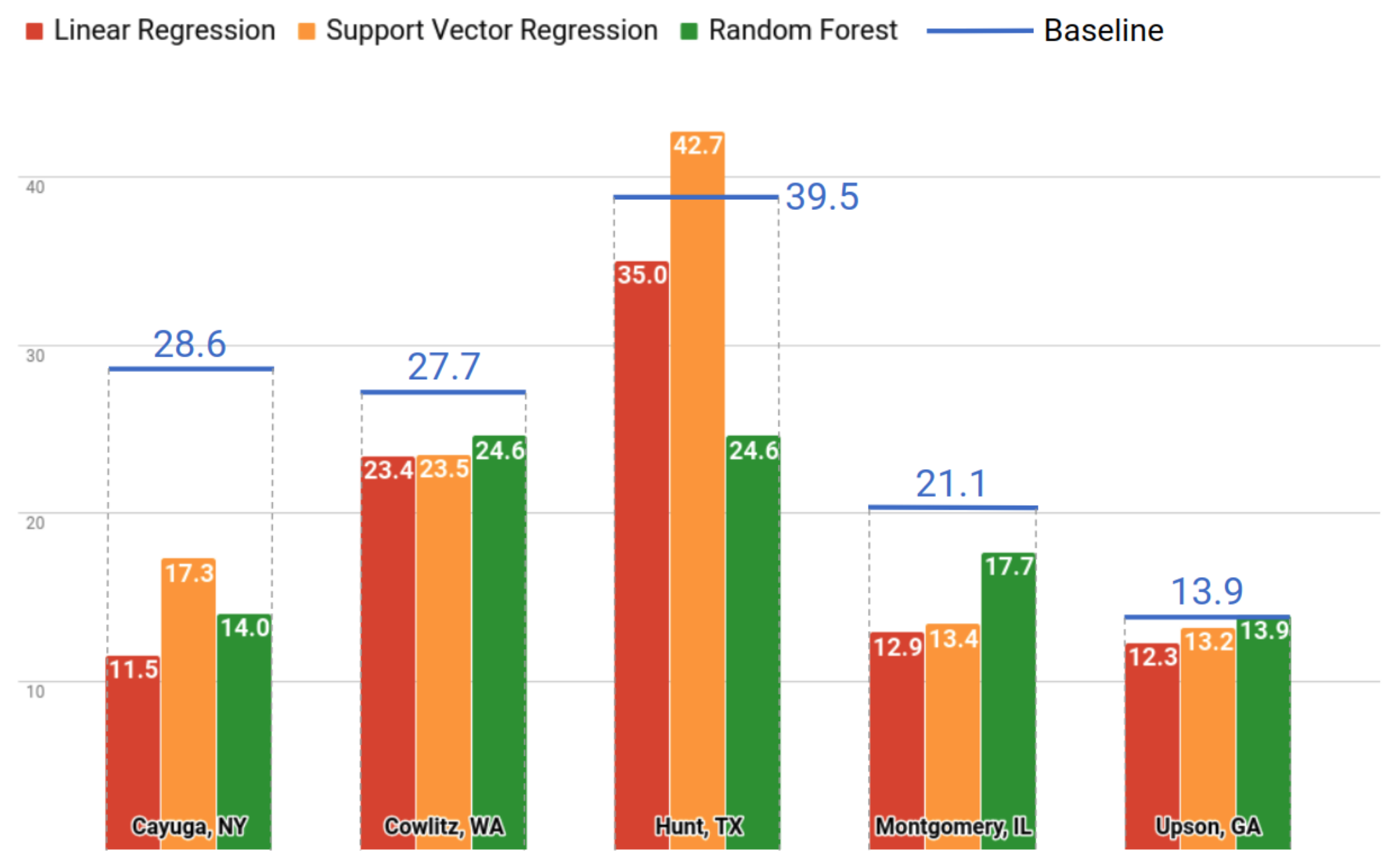

Figure 2: Models' Prediction Scores Compared To Baselines

- **Attributes' Effects:**

Attribute	Unit increased in attribute	$ increased in sold price
Assessment	$100	$54
Comparable Houses' Sold Price	$100	$34
Listed Price	$100	$38
Baths	1 bathroom	$15,787

Table 2: Effects of Most Important and Statistical Significant Attributes Across Five Counties

Results [2]

- **Overestimation Problem:** Our models successfully reduce the overestimated to underestimated house (OE:UE) ratio from 3 to 2 (Fig. 3) to 1 to 1 (Fig. 4).

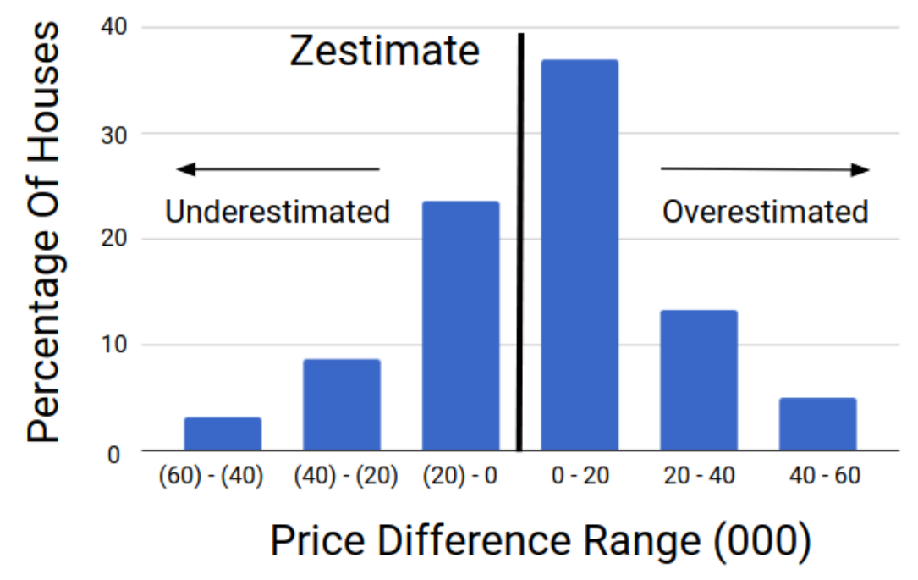

Figure 3: Zestimate Gives a OE:UE Ratio of 3:2

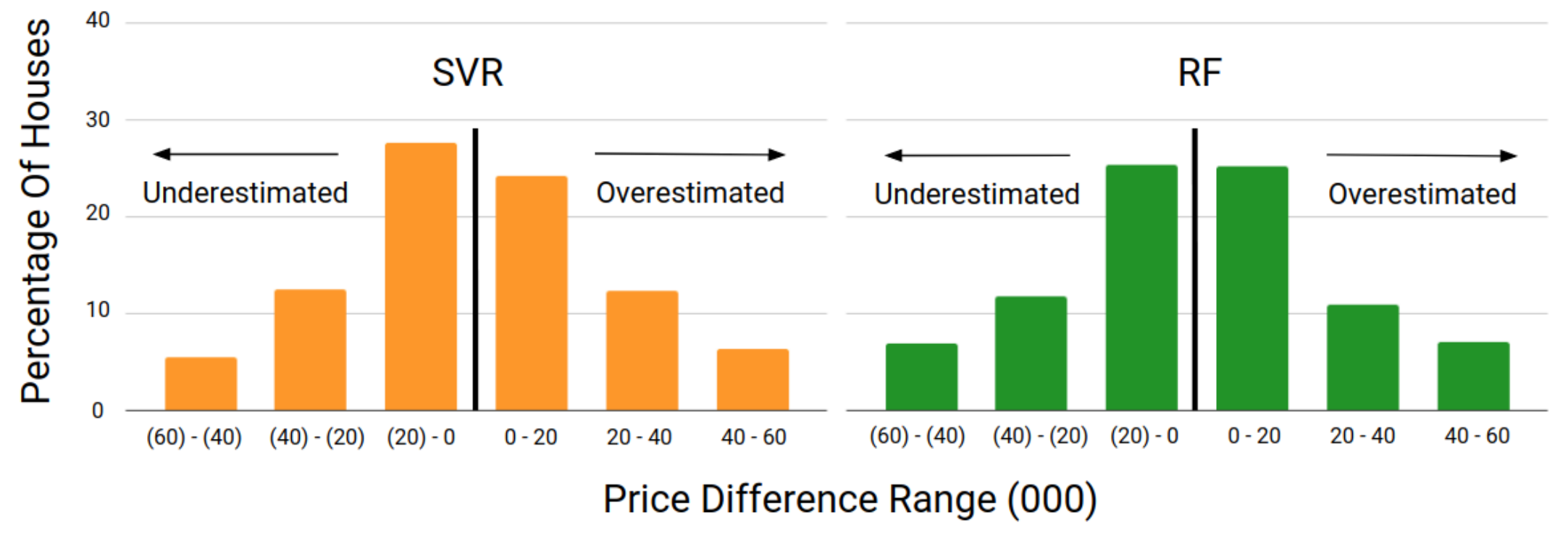

Figure 4: SVR and RF Give a OE:UE Ratio of 1:1

Conclusions

1. For Hunt (TX), SVR outperforms the baseline, and is statistically better than LR. For Cowlitz (WA), Montogomery (IL), and Upson (GA), RF produces results similar or close to the baselines. However, RF is not statistically different from LR.
2. This project produces an even distribution of overestimated to underestimated houses.
3. The 4 most important and statistical significant housing attributes across 5 counties are *assessment*, *comparable houses' sold price*, *listed price*, and *number of bathrooms*.

References

[1] Asabere, P. K., and Huffman, F. E.
Price concessions, time of the market, and the actual sale price of homes.
Journal of Real Estate Finance and Economics 6 (1993), 167 – 174.

[2] Trulia.
Trulia estimate, 2017.
Accessed: 11/11/2017.

[3] Zillow.
The price of overpricing: How listing price impacts time on market, 2016.
Accessed: 03/06/2018.

[4] Zillow.
Zestimate, 2017.
Accessed: 11/11/2017.