Housing Price Prediction

An Nguyen

March 20, 2018

Abstract

This paper explores the question of how house prices in five different counties are affected by housing characteristics (both internally, such as *number of bathrooms*, *bedrooms*, etc. and externally, such as *public schools' scores* or the *walkability score of the neighborhood*). Using data from sold houses listed on Zillow, Trulia and Redfin, three prominent housing websites, this paper utilizes both the hedonic pricing model (Linear Regression) and various machine learning algorithms, such as Random Forest (RF) and Support Vector Regression (SVR), to predict house prices. The models' prediction scores, as well as the ratio of overestimated houses to underestimated houses are compared against Zillow's price estimation scores and ratio. Results show that SVR gives a better price prediction score than the 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, this paper's models reduce the overestimated to underestimated house ratio of 3:2 from Zillow's estimation to a ratio of 1:1. This paper also identifies the four most important attributes in housing price prediction across the counties as *assessment*, *comparable houses' sold price*, *listed price* and *number of bathrooms*.

ACKNOWLEDGEMENTS:

I would like to thank my thesis advisors - Prof. Chris Fernandes, Nick Webb, and Harlan Holt - for their advice and inputs on this project. Many thanks to my friends and family as well, who spent countless hours to listen and provide feedbacks.

Contents

1 Introduction

According to the US Census Bureau, 560,000 houses were sold in the United States in 2016 [11]. In addition, 65% of all American families owned houses in 2016 [12]. For the Americans who sold and bought these houses, a good housing price prediction would better prepare them for what to expect before they make one of the most important financial decisions in their lives. A recent report from the Zillow Group, a popular housing database website, indicates that house sellers and buyers are increasingly turning to online research in order to estimate house price before contacting real estate agents [4]. Researching how much the house you are interested in is worth on your own can be difficult for multiple reasons. One particular reason is that there many factors that influence the potential price of a house, making it more complicated for an individual to decide how much a house is worth on their own without external help. This can lead to people making poorly informed decisions about whether to buy or sell their houses and which prices are reasonable. Because houses are long term investments, it is imperative that people make their decisions with the most accurate information possible. Therefore, housing websites such as Zillow, Trulia and Redfin [1], exist to provide estimations of housing valuations based on the houses' characteristics, at no cost.

However, the estimations provided by these housing websites are not always accurate. For example, Zillow states that their housing price prediction algorithm, called "Zestimate", only estimates 54.4% of houses within the 5% of their actual sale prices [22]. For Trulia, only 48.2% of houses have Trulia-estimated prices to be within the 5% range of their actual sold prices [20]. Therefore, the first question of this project is whether I can outperform Zestimate's prediction score or come close to it. In this project, I define the prediction score as the percentage of houses whose estimated prices fall within the 5% range of their actual sold prices. Using this project's datasets and Zestimates as the pre-

[1] Zillow: https://www.zillow.com; Trulia: https://www.trulia.com; Redfin: https://www.redfin.com.

dictions, I compute Zillow's prediction scores and use them as the baselines to see how well my own models perform. I chose Zillow's estimator as a benchmark instead of its competitors' because Zillow is widely regarded as the most popular housing website due to its large databases of 110 million houses and their 11 years of expertise in pricing estimations. According to Hitwise, a consumer analytics company, Zillow's market share, based on online visits to the site, is 27.2% in 2016, while the numbers for Trulia and Redfin are 9.4% and 3.7%, respectively.

Zillow tends to overestimate their listed properties, meaning the Zestimates are higher than the actual sold prices of the houses. In the dataset of 1,457 sold houses I collected, the ratio of overestimated houses to underestimated houses is 3 to 2. Hollas, Rutherford and Thomson (2010) studies Zillow's estimations of single family houses and finds that 80% of their housing sample gathered from Zillow are overpriced by Zestimate [8]. For a house seller who prices his house based on Zillow's suggestion, he/she is likely to list his/her house for more than what it is worth. According to a Zillow research in 2016, if a house is priced above its true market valuation, it tends to stay on the market five times longer compared to a house that is well-priced, suggesting a string penalty for overpricing houses [19]. Moreover, the same research suggests that houses that have been on the market for two months can lose 5% of its original listed price. Asabere and Huffman (1993) also supports the theory of a reversed correlation between a house's time on the market and its final sold price [1]. Therefore, the second question of this project is whether my models can get rid of this overestimation problem.

The final question of this project is what the most important factors affecting housing prices are. In order to answer the three questions listed above, this project proposes using both the hedonic pricing model and various machine learning algorithms.

2 Literature Review

Sirmans, Macpherson and Zietz (2005) provides a study of 125 papers that use hedonic pricing model to estimate house prices in the past decade [16]. The paper provides a list of 20 attributes that are frequently used to specify hedonic pricing models. This dataset contains 12 attributes on this list. Moreover, Sirmans, Macpherson, and Zietz (2005) also discusses the effects of some variables on housing price. For example, number of bathrooms is usually positively correlated to the final sale price. Out of 40 times appearing in housing price studies, this attribute has a positive effect 34 times and is statistically significant 35 times. On average, keeping other variables unchanged, an increase of 1 bathroom leads to 10% to 12% increase in the property's value. Similarly, my paper shows that, based on the dataset of sold houses in five counties, the number of bathroom has a statistically significant and positive effect on sold price. On average, an increase of 1 bathroom could increase a house's price by $15,787.

Cebula (2009) conducts a study on the housing prices in the City of Savannah, Georgia using the hedonic pricing model [3]. The paper's data contains 2,888 single-family houses for the period between 2000 and 2005. Cebula (2009) shows that the log price of houses is positively and significantly correlated with the number of bathrooms, bedrooms, fireplaces, garage spaces, stories and the total square feet of the house. Additionally, the paper adds three dummy variables, MAY, JUNE, and JULY, to account for seasonable factor with regards to the houses' prices. If the house is sold in May, the variable MAY is set to be equal to 1 and 0 otherwise. The other variables, JUNE and JULY are constructed in a similar fashion. The paper finds that the log sale prices of houses are significantly and positively correlated with MAY and JULY while JUNE is insignificant. This implies that houses that are closed in May or July tends to have a higher price. Similar to Cebula (2009), my paper includes sold month of the house as dummy variables. However, these attributes do not appear to be statistically significant.

Selim (2009) seeks to study the effects of different housing characteristics on housing prices in Turkey using two different methods: hedonic pricing model and artificial neural network [15]. The paper's dataset, which was collected from the 2004 Household Budget Survey Data for Turkey, contains 5,741 observations with 46 housing characteristics. For the hedonic pricing model, the author uses the semi-log form, $ln(P) = \beta x + u$, where P denotes the price of the house, x is the set of independent variables and u is the error term. As for the artificial neural network model, the paper uses 2 hidden layers, with nine and four nodes for the first and second layer, respectively. The results are consistent with other studies on housing price. The author finds that the total number of rooms, the size of the house, the heating systems, appliances such as garbage disposal, garage and pool, etc. have a significant and positive effect on the house price. More importantly, Selim finds that the artificial neural network model has a lower error score than the hedonic model. When the hedonic model's mean squared error is 2.47, the same error measurement by the neural network model is 0.44. Similarly, Tay and Ho (1991/1992) compared the pricing prediction between regression analysis and artificial neural network in predicting apartments' prices in Singapore [18]. They found that the neural network model outperforms regression analysis model with a mean absolute error of 3.9%.

Jirong, Mingcang, and Liuguangyan (2010) uses support vector machine (SVM) regression to forecast the housing prices in China in between 1993 and 2002 and in certain district in Tangshan city in between 2000 to 2002 [9]. The paper utilizes the genetic algorithm to tune the hyper-parameters in the SVM regression model. The error scores for the SVM regression model for both China and a Tangshan City's district are both lower than 4%. This indicates that the SVM regression model perform well in forecasting housing prices in China. In the Singapore's housing market, Fan, Ong and Koh (2006) uses decision tree model study the housing characteristics' effects on prices [6]. The paper concludes that the owners of 2-room to 4-room flats are more concerned with the flats' basic characteris-

tics such as model type and age more than the owners of 5-or-more-room flats. Moreover, owners of executive flats care more about the services characteristics such as the neighborhood location and recreational facilities than basic housing characteristics.

3 Methods

3.1 Hedonic Pricing Model

In Economics, the hedonic pricing model is frequently used to measure a property's price. The model is based on the theory of consumer's demand by Lancaster (1966), which states that utilities of a good is not based on the good itself but on the individual "characteristics" of the good [10]. However, it's not until Rosen (1974) that the idea of pricing is added to the model. Rosen argues that a good can be evaluated based on the individual values of its composite attributes [14]. Since then, this pricing model has been adapted to evaluate properties based on their internal and external characteristics.

Hedonic pricing model combines both a house's internal characteristics (such as *number of bedrooms*, *number of bathrooms*, etc.) and its external characteristic (such as *neighborhood's walkability score*, *public schools' scores*, etc.) to estimate its values. A hedonic model can be written as a linear regression model, as follows:

$$P_i = \sum_{i=1}^{k} w_{i,m} E_{i,m} + \sum_{i=1}^{k} w_{i,n} I_{i,n} + b \tag{1}$$

In equation (1), there are k observations with m number of External housing attributes and n number of Internal attributes. Moreover, *b* represents the constant term. This model explores the linear relationship between various characteristics of a house and its actual sold price. For example, if the coefficient of the variable "bathroom" (w) in the hedonic model is 15000, keeping other variables constant, if a house has one more bathroom, its

sold price could go up by $15,000.

In this project, the hedonic pricing model or Linear Regression is used as the baseline model to compare more complex machine learning algorithms against. This model is chosen for its frequent appearance in Economics papers on housing price prediction and its simplicity in explaining relationships among attributes.

3.2 Machine Learning Algorithms

This project uses WEKA[2], a suite of machine learning algorithms, developed at the University of Waikato. There are various algorithms[3] tested, based on their abilities to handle regression analysis and their appearances in previous literature. The best performing ones are Random Forest and Support Vector Regression, which are explained in details in the next two subsections.

3.2.1 Random Forest

Random Forest is a learning algorithm first created by Tin Kam Ho [7], a computer scientist at IBM, and later extended by Leo Breiman and Adele Cutler [2] [13]. It operates by constructing a multitude of decision trees to fit the observations into groups based on their attributes' values and outputs the mean prediction of the individual trees. As the name suggests, "decision tree" model builds a reversed tree-like structure, where the "root" is at the top, followed by multiple branches, nodes and leaves. The end of each branch is a decision leaf, which is the model's predicted value, given the values of the attributes represented by the path from the root node to the said decision leaf. Figure 1 presents a sample decision tree where the dependent variable or the decision leaf is the sold price of a house, and the dependent variables or the nodes are the number of bath-

[2]https://www.cs.waikato.ac.nz/ml/weka/

[3]These include IBk, Artificial Neural Network (Multilayer Perceptron), and Decision Tree.

rooms, bedrooms, and the size of the house. In Figure 1, the leftmost decision leaf (Price = $50,000) is derived using the path in which the root node "Baths" has a value of less than two and the node "Beds" has a value of less than or equal to two. This tree's maximum depth, which can be defined by the longest distance from a decision leaf to the root node, is therefore three. In building a decision tree, the best attribute of the dataset, in terms of error deduction, is placed at the top of the tree (root node). The next node used in each tree branch produced by the root node is considered the next best attribute if the attribute at the root is removed from the dataset, and so forth. The process of choosing which node to use at each tree branch is described below:

1. For each of the independent variables, fit a regression between the independent and the dependent variable.
2. For each of the independent variables, the observation set is split into several disjoint subsets at certain values of the variable.
3. At each split point, the error between the predicted and actual value is squared to create the sum of squared errors (SSE).
4. The SSE is compared across all independent variables and the split points. The variable/split point with the lowest SSE is chosen to be the root node and the split point for the root node.

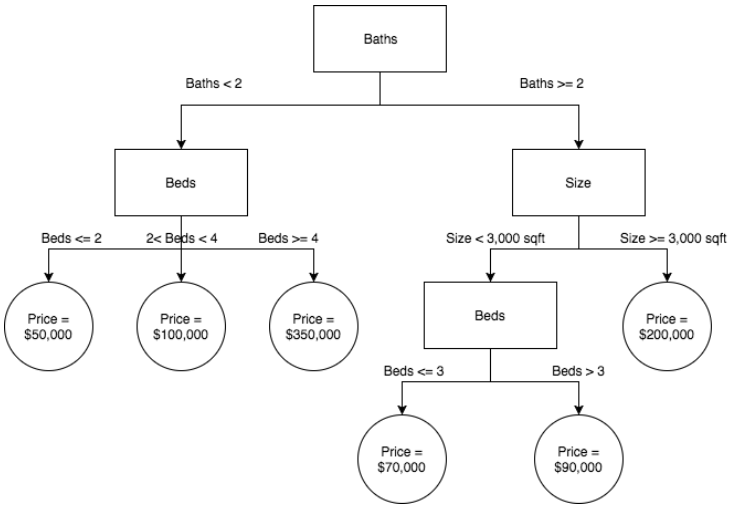


Figure 1: A sample decision tree model

After following the four steps above, the tree in Figure 1 chooses the root node as "Baths" with a split point of 2. This means that this variable along with the split point produces the smallest SSE compared to other variables and split points. After identifying the root node, the algorithm uses steps 1 to 4 again for each branch of the tree (when Bathrooms $<$ 2 and when Bathrooms $\geq$ 2) until all the data is processed and the decision leaves contain house price. Based on the decision tree model from Figure 1, a house that has 1 bathroom with 3 beds is estimated at $100,000 whereas a house with 3 bathrooms, size of 2,000 square feet and 3 bedrooms is estimated at $70,000.

For a dataset with many attributes, using decision tree can lead to a large number of splits, which creates a large and complex tree. When a tree is designed so that it can fit all the training data points too well, the over-fitting problem occurs. This leads to inaccuracy when predicting value of data points in the testing sets. By using random subsets of attributes or observations for training on different trees, Random Forest can therefore limit the over-fitting problem. In addition, Random Forest works well with datasets with missing values, using the "surrogate split" method[4].

3.2.2 Support Vector Regression

Support Vector Machine (SVM) is developed by Vladimir N. Vapnik and Alexey Ya. Chervonenkis, two Russian statistician/mathematician in 1963 [21]. In 1996, a version of the algorithm, called Support Vector Regression (SVR), was introduced by Vladimir N. Vapnik, Harris Drucker, Christopher J. C. Burges, Linda Kaufman and Alexander J. Smola [5]. Instead of fitting a best fitted line over the observations like Linear Regression, SVR with a particular kernel (a polynomial kernel with a degree of one), called Linear SVM Regression, fits a flat hyperplane. This Linear SVM is used for the project. Figure 2 shows an example of fitting a hyperplane through a collection of data points in three dimensions.

[4]Subsection 7.1 under the Appendix Section provides an explanation of how surrogate split works.

For any point within the margins of the hyperplane, its error would be 0. The model is described as $y_i = wx_i + b$, in which y_i is the predicted value, x_i is the attribute and w is the weight of the attribute x_i.

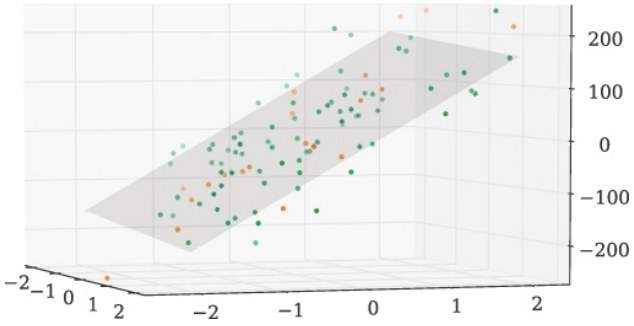


Figure 2: Hyperplane Fitting

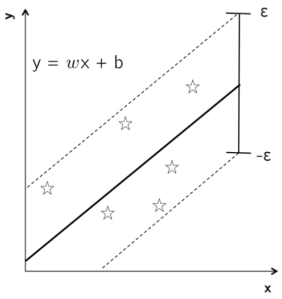


Figure 3: Linear Support Vector Regression

The goal of SVR is to minimize $\frac{1}{2}||w||^2$, which helps reduce over-fitting in training data and therefore reduces test errors [17]. The error is calculated as shown in equation (2).

$$\text{Error} = \sum_{i=1}^{N} [\max(y_i - (wx_i + b) - \varepsilon, 0)] \tag{2}$$

In equation (2), ε is the margin of the hyperplane as shown in Figure 3, and y_i, x_i and w are as explained above.

By minimizing the weights of attributes, Support Vector Regression is less sensitive to noises. Therefore, this model works well in finding patterns in real-life noisy datasets, such as financial data.

4 Data

4.1 Data Collection

This project collects data on 1,457 sold houses from Zillow, Trulia and Redfin using Python and Selenium (a browser automation tool) for data scraping. These houses are selected from five counties in five different regions of the United States. Figure 4 shows the geographical locations of these counties on the United States' map along with their corresponding states, regions and the number of houses scraped from them.

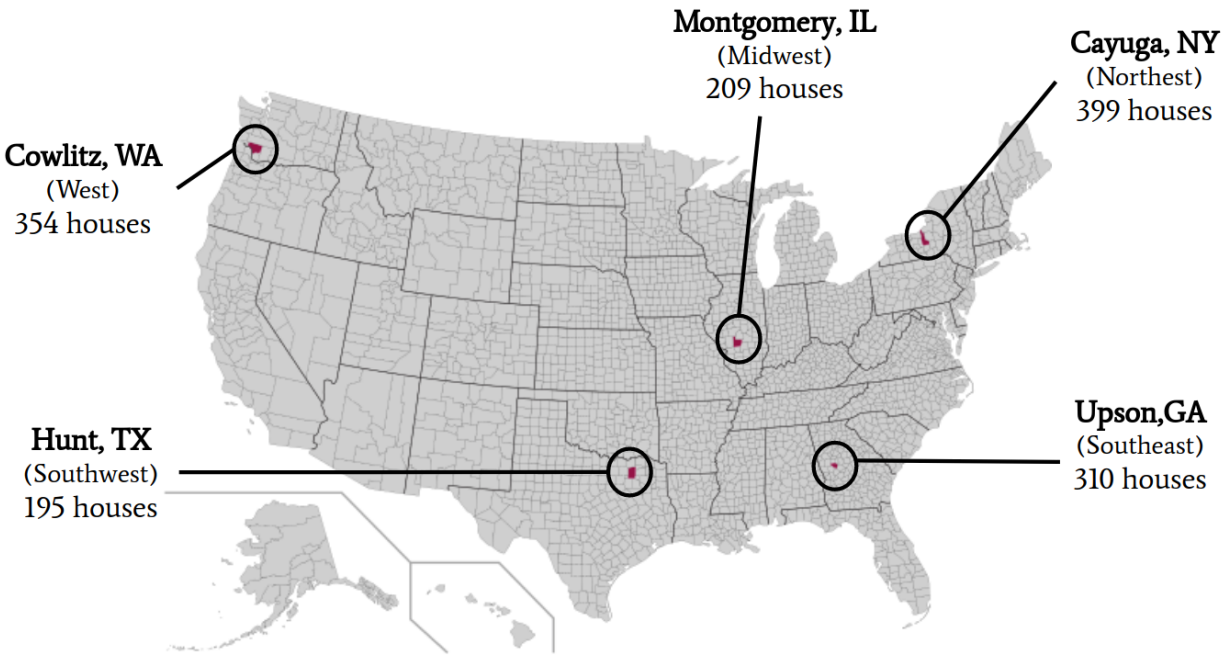


Figure 4: Selected Counties On The US Map

These five counties are chosen based on the following criteria:

1. They are from different regions in the United States.
2. They are among the worst performing counties based on the percentage of houses whose Zestimates fall within the 5% range of their actual sold prices. This suggests that there could be visible improvements in the price prediction algorithms for these particular counties. The list of how well Zestimates perform in different counties

can be found on Zillow's website.

3. These counties' housing information is available on all three housing websites.

Since the result for this project is calculated as the percentage of houses whose predicted prices fall within the 5% range of their actual sold prices, the corresponding evaluation baseline is the percentage of houses whose Zestimates are within the 5% range of their actual sold prices. Therefore, it is important to scrape the Zestimate prediction for every sold house. However, Zillow updates the Zestimate number even after a house is sold, using the Zestimate currently listed on Zillow's page would be inaccurate. Instead, the historic Zestimate prediction right before a house is sold is needed. Since Zillow keeps track of the Zestimate history for every month, with houses' Zillow property ID, I was able to find the Zestimate number for the month right before a particular house is sold, and use this to compute my data's prediction score baseline.

County	**State**	**Region**	**# of Houses**	**Zillow's Countywide Baseline**	**My Data's Baseline**
Cayuga	NY	Northeast	399	16.7	28.6
Montgomery	IL	Midwest	209	8.7	21.1
Upson	GA	Southeast	310	10.7	13.9
Hunt	TX	Southwest	195	19.8	39.5
Cowlitz	WA	West	354	29.3	27.7

Table 1: Selected Counties' Information

Table 1 summarizes the five counties' information, including the *Zillow's Countywide Baseline* for each county, which is the percentage of houses on Zillow whose Zestimates fall within 5% of the houses' actual sold prices. This is also the number Zillow reported on their website as a measurement of their algorithms' price predictability on each county. Since my dataset only contains a fraction of houses that Zillow has, I compute my own data's baseline, which is shown in Table 1 as *My Data's Baseline*. *My Data's Baseline* is

measured in a similar way as *Zillow's Countywide Baseline*, only on the dataset of 1,457 houses.

4.2 Data Processing

This project collects 1,457 houses from three prominent different housing websites (Zillow, Trulia, and Redfin). The reason is to make sure that the scraped housing data is as accurate as possible. Since all three websites get data from listing services or other third-party companies, inconsistency and mistakes in data are unavoidable. Figure 5, which shows how the same house (*215 P G Street Rd, Kelso, WA 98626)* is recorded across Zillow, Trulia and Redfin, will serve as an example of inconsistency in sold price.

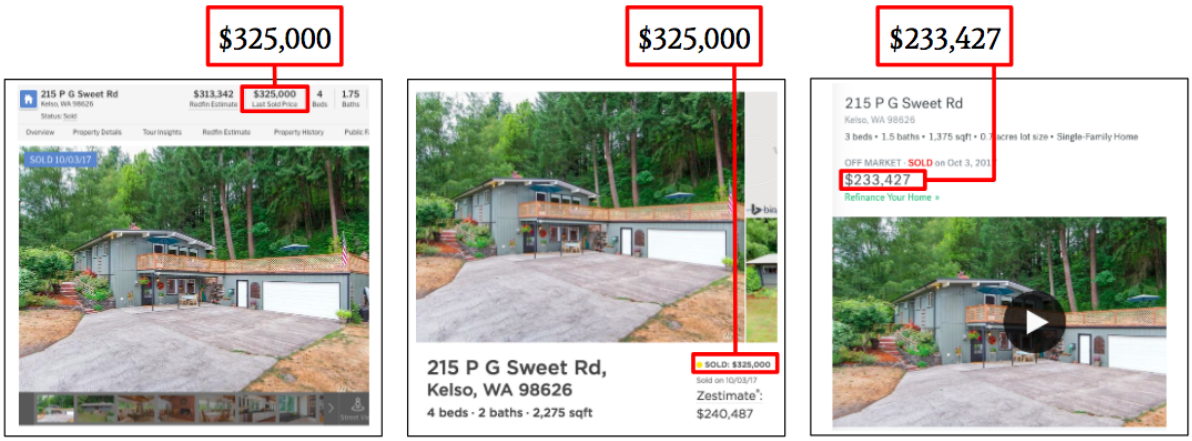


Figure 5: An Example Of Data Inconsistency

All three websites record that the house was last sold in October, 2017. However, the sold price of the houses is inconsistent. On Zillow and Redfin, the house's sold price is recorded as $325,000 while the price is $233,427 on Trulia. In order to handle this consistency, I simply pick the number that appears most often (which, in this case, would be $325,000). If the three websites record three different sold prices, then I take an average

of the three and mark the house for a later check-up. In case that the sold prices vary in a wide range across the three websites (e.g. $1 in Zillow, $16,000 in Trulia and $58,000 in Redfin), the observation will be drop. Besides sold price, other housing characteristics are also subject to the same comparison algorithm.

Besides inconsistency in data values across three websites, there is also inconsistency in data units. For example, size of a house can be recorded in either square feet or acres. Therefore, an extra conversion step has to be taken in order to uniform data units. All data processing steps are done in Excel's Visual Basic for Applications (VBA).

4.3 Data Description

In this dataset, there are 35 housing attributes, including internal attributes and external attributes. Internal housing attributes, such as *number of bedrooms* and *number of bathrooms*, are intrinsic variables to the houses. On the other hand, external housing attributes, such as *the walkability of the neighborhood* and *public schools' scores*, are variables that are not built-in with the houses. Table 6 and 7 in the Appendix Section show a full list of attributes, both numeric and non-numeric, used in this project. If the attributes are numeric, Table 6 displays their summary statistics. If the attributes are non-numeric, Table 7 lists these attributes and the coded dummy/binary variables these attributes are turned into. For example, for the non-numeric attribute *Sold Month*, its dummy variables are the twelve months of the year. If a house is sold in January, then the variable January would take a value of 1, and 0 otherwise.

In this project's dataset, one of the attributes is comparable houses' sold price. In this paper, this attribute is recorded as *Comparables' Sold Price* or *Coms' Sold Price*, for short. Zillow, Trulia and Redfin all provide a list of comparable houses (based on similar features such as location, square footage and beds/baths) to the house currently being looked at,

called "House X". If a comparable house is sold before "House X" and the sold date is within one year of the sold date of "House X", the price of this comparable house is put in a list of comparable houses' sold price. For every house, there are three lists of the comparable houses' sold price from the three housing websites. The attribute *Comparables' Sold Price* is then calculated as the average of these three lists' median values.

5 Results

5.1 Prediction Scores

Figure 6 shows the prediction scores of the three algorithms used in this project (Linear Regression (LR), Support Vector Regression (SVR) and Random Forest (RF)) in comparison to the data's baselines[5] for all five counties. The county dataset with asterisk (*) next to it demonstrates that the best performing algorithm is statistically better than the baseline algorithm, or LR. As mentioned before, the prediction score is measured as the percentage of houses whose estimated prices fall within the 5% range of their actual sold prices.

On the dataset of Hunt (TX), SVR outperforms the baseline by 3.2%. Zestimate predicts 39.5% of houses in Hunt county's data to be within the 5% range of their actual sold prices while SVR estimates 42.7% of houses to be in this 5% range. Moreover, the result produced by SVR is statistically better than LR, the baseline algorithm. For the dataset of Upson (GA), RF gives the same prediction score as the baseline (13.9%). In addition, RF produces prediction scores that are close to the baselines for the datasets of Cowlitz (WA) and Montgomery (IL). For these two counties, the differences between RF's predictions and the baselines are around 3%. However, RF is not statistically different than LR for the datasets of the three counties in Washington, Illinois, and Georgia. Among the five

[5]This data's baselines are the same as the values recorded in the last column of Table 1.

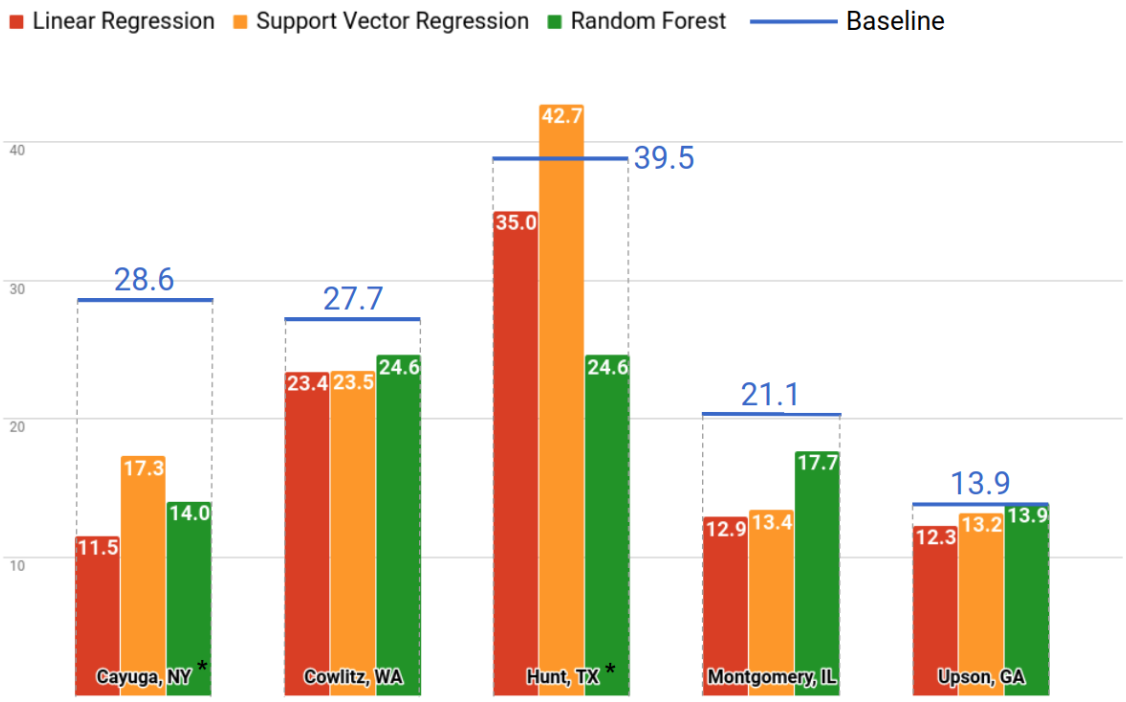


Figure 6: Models' Prediction Scores Compared To Baselines

counties, Cayuga (NY) appears to have the worst performance. For this county's dataset, SVR is the best performing algorithm, followed by RF and LR. For Cayuga (NY), the prediction score gap between the baseline and SVR's performance is 11%. However, SVR is statistically better than LR for this particular county.

In order to produce the results as shown in Figure 6, each county data has a different set of attributes that are considered "most important" in terms of predicting sold prices. These attributes are selected using a combination of WEKA's *Attribute Selected Classifier*, which evaluates the predictability of a subset of attributes by considering the individual predictability of each attribute and the degree of redundancy between them, and through "trial and error" experiments. Table 8 in the Appendix Section shows these most important attributes for each of the five counties. However, it would be beneficial to have the same set of attributes for all counties, so that we can have a uniform frame of reference for comparison of attributes' effects on sold prices. Given the sets of most important at-

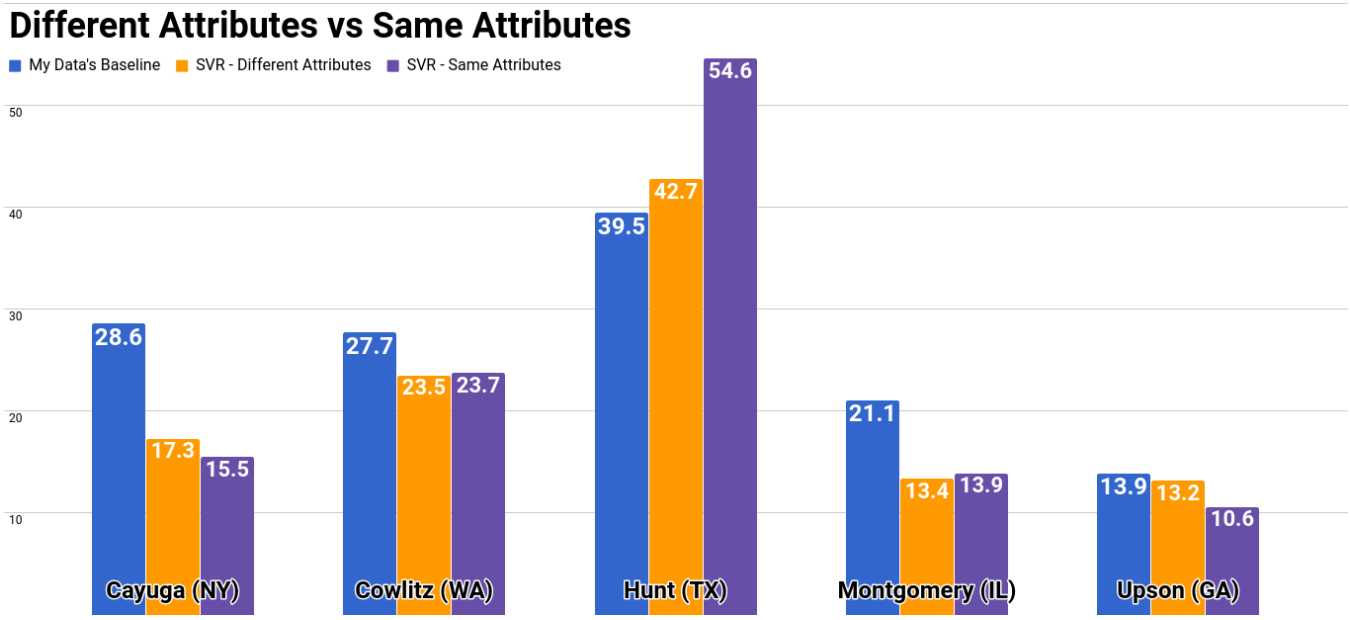


Figure 7: SVR - Different Attributes vs. Similar Attributes

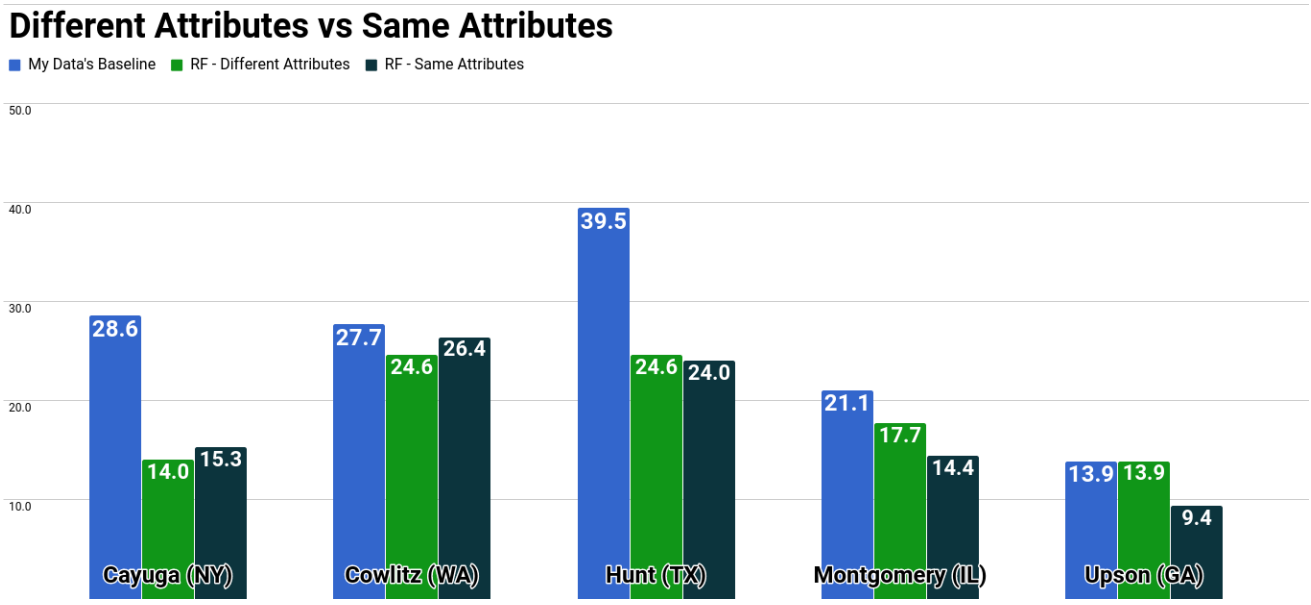


Figure 8: RF - Different Attributes vs. Similar Attributes

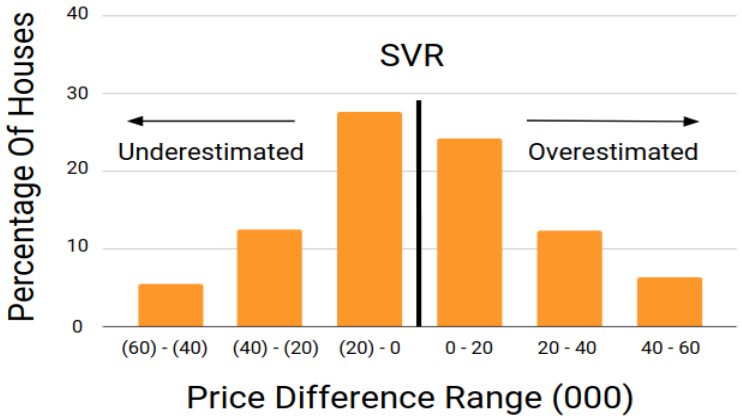
tributes across counties, I picked out 10 attributes that are important to the majority of the counties. In another words, these 10 attributes appear at least 3 times within the 5 counties' datasets. For example, *assessment* appears as the one of the most important attributes in all five counties' datasets and *number of bedrooms* appears in three datasets. Therefore, these two attributes are included in the list of 10 most important attributes across the five counties. Table 9 in the Appendix Section shows a list of these attributes and which counties' datasets they appear on.

Figure 7 and 8 display the prediction score comparison between having one common set of attributes for 5 counties and having different sets of attributes for different counties. Figure 7 uses SVR as the algorithm whereas Figure 8 uses RF. These figures suggest that switching from different set of attributes to a single set don't change the prediction scores by a lot. In some cases, such as the dataset of Hunt (TX) with SVR (Figure 7), using the same set of attributes yields a better result.

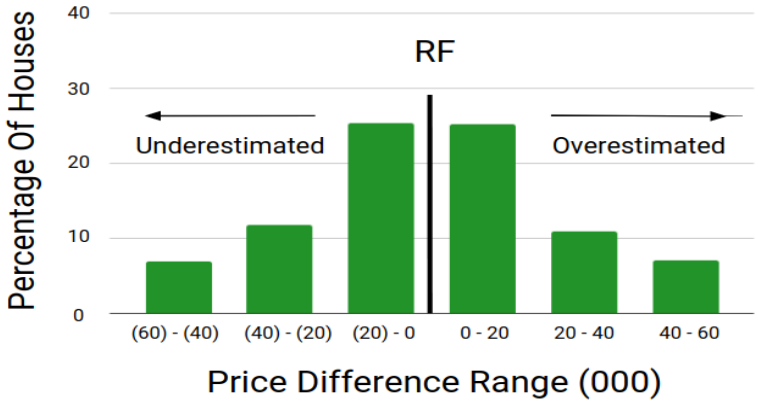
5.2 Overestimation Problem

As mentioned before, Zillow tends to overestimate their properties. The Introduction Section states the reason why over-pricing houses can have a negative effect on the final sale prices. In this dataset of 1,457 houses, the ratio of overestimated to underestimated houses by Zillow is 3 to 2 (Figure 9a). However, my models successfully reduce this ratio to 1 to 1. Figure 9b shows the ratio of overestimated to underestimated houses when using SVR while Figure 9c shows the ratio when using RF as the algorithm. The horizontal axis represents price difference range (in thousand dollars) while the vertical axis shows the percentage of houses that fall within a certain price difference range. Negative price difference, which means the Zestimate/predicted price is lower than the actual sold price, is put in brackets, as shown on the horizontal axis of the graph.

(a) Zestimate Gives a Ratio of 3:2



(b) Support Vector Regression Gives a Ratio of 1:1



(c) Random Forest Gives a Ratio of 1:1

Figure 9: Overestimated To Underestimated House Ratio Comparison

5.3 Attributes' Effects On Sold Price

Table 2 shows the coefficients of the most important attributes across 5 counties that are also statistically significant at 95% confidence level when it comes to predicting house prices. The coefficients shown in the table are calculated by taking average of the attributes' coefficients among the counties. Results show that, on average, $100 increase in the house's assessment could increase its sold price by $54. The same amount increase in the sold price of comparable houses and the house's listed price would increase the sold price by $34 and $38, respectively. Moreover, Table 2 suggests that if a house has one more bathroom, its sold price could go up by $15,787.

ID	Attribute	Coefficient
1	Assessment	0.54
2	Comparables' Sold Price	0.34
3	Listed Price	0.38
4	Baths	15,787

Table 2: Most Important and Statistical Significant Attributes

The effects of these four attributes can vary across counties. For example, the coefficient of *Baths* is higher in Upson (GA) than in Montgomery (IL), suggesting that the number of bathroom has a bigger effect on sold price in Upson county. Table 3 shows the effects of these four attributes on houses' sold prices for each of the five counties. A blank entry in the table means that the attribute is not considered an "important" attribute for that particular county.

Attribute[1]	Cayuga (NY)	Cowlitz (WA)	Hunt (TX)	Montgomery (IL)	Upson (GA)
Assessment	0.39	0.39	0.15	0.57	1.22
Coms' Sold Price	0.30	0.51	0.12	0.47	0.29
Listed Price	0.27	0.23	0.62	0.40	
Baths		16,697		12,650	18,014

[1] These attributes are statistically significant at the 95% confidence level.

Table 3: Attributes' Effects On Sold Price Across Selected Counties

6 Conclusion

Using a dataset of 1,457 houses from 5 different counties scraped from Zillow, Trulia and Redfin, this paper addresses the following questions:

1. Can the models proposed in this paper outperform or get close to Zillow's prediction score baseline?
2. Can the overestimated to underestimated house ratio be reduced?
3. What are the most important attributes that affect sold price?

For Hunt (TX), SVR outperforms the baseline by 3.2%. RF outputs close predictions scores to the baseline with the dataset from Cowlitz (WA) and Montgomery (IL). The differences between RF's predictability and Zestimate for these two counties is around 3%. RF gives a similar score as the baseline for Upson (GA). Moreover, results suggest that using one single set of 10 attributes for all counties will not change the models' accuracy scores by a lot in comparison to using different sets of attributes for different counties. The overestimated to underestimated house ratio is also reduced from 3:2 to 1:1. In addition, the four most important and statistical significant attributes are identified as *number of bathrooms*, *assessment*, *listed price* and *comparable houses' sold price*.

Finally, for future work, it would be interesting to see what results could be yielded from applying the same models on counties that Zillow reports to be the best performing ones.

7 Appendix

7.1 Surrogate Split

Consider a dataset of 10 observations, 3 independent variables (*Baths*, *Beds*, and *Rooms*) and the dependent variable as *Sold Price*, as shown in Table 4 below. Assume that when we apply a tree algorithm to this dataset, *Baths* is picked as the tree's root node with a split point of 1. This is called the "primary split". However, the algorithm still have to determine whether to put x_5 and x_8 in the group of observations with fewer than or equal to 1 bath (x_2, x_6) or in the group with more than 1 bath ($x_1, x_3, x_4, x_7, x_9, x_{10}$), because *Baths* is missing for both of these observations.

Obs.	Baths	Beds	Rooms
x_1	2	4	1
x_2	1	?	3
x_3	3	3	4
x_4	2	3	2
x_5	?	3	4
x_6	1	2	?
x_7	3	3	5
x_8	?	3	4
x_9	2	3	3
x_{10}	2	1	4

Table 4: Sample Data

The surrogate splitting method is then utilized. It uses the attributes *Beds* and *Rooms* to create "alternative splits". If the best split point using the attribute *Beds* is 3, we will have two groups of observations: (x_6, x_{10}) and ($x_1, x_3, x_5, x_7, x_8, x_9, x_4$). If the best split point using the attribute *Rooms* is 4, we have two different sets: (x_1, x_2, x_4, x_9) and ($x_3, x_5, x_7, x_8, x_{10}$). Comparing between the sets produced by using the attributes *Rooms* and *Beds*, we see that the two sets produced by *Beds* bear more resemblance to the two sets produced by the attribute *Baths* in terms of the number of observations that are grouped

together. Therefore, we say that *Beds* produces the best "surrogate split". Using this "surrogate split", the two observations x_5 and x_8 are grouped with observations x_1, x_3, x_7, and x_9, which, in the "primary split", are pooled together in the group that has more than 1 bathroom. Therefore, after using the surrogate splitting method, the tree algorithm knows which tree branch to put observations with missing values into.

7.2 Tables

- **Table 5** shows a summary of the literature, specifying the previous works' datasets, methods, accuracy and error scores.
- **Table 6** shows a list of the numeric attributes.
- **Table 7** shows a list of the non-numeric attributes.
- **Table 8** shows the most important attributes for each county.
- **Table 9** shows the ten most important attributes across all five counties.

Table 5: Literature Summary

Paper	Data	Method	Accuracy	Error
Cebula (2009)	2,888 single family houses in Georgia	Hedonic Pricing Model	R-squared = 0.86	
Fan, Ong, and Koh (2006)	5,589 flats in Singapore	+ Decision Tree. + Training-Test percentage is 75% - 25%. + Uses Tree Node in SAS, which integrates the CHAID, CART, and C4.5/C5.0 algorithms.	R-squared = 0.88	
Jirong, Mingcang, and Linguangyan (2011)	+ National annual selling price in China between 1993 and 2000. + Quarterly selling price of Tangshan city between 2000 and 2002	+ Support Vector Machine. + Uses Genetic Algorithms to optimize SVM.		+ Mean Absolute Percentage Error: 2.33% for China data + Mean Absolute Percentage Error: 1.94% for Tangshan City data
Tay and Ho (1991/1992)	1,055 observations	+ ANN (1 hidden layer with 5 nodes) + Regression		+ Mean Average Error for ANN = 3.9% + Mean Average Error for Regression = 7.5%
Selim (2009)	5,741 houses in Turkey	+ ANN (2 hidden layers, with 9 and 4 nodes each) + Hedonic Pricing Method	Hedonic Pricing Model's R-squared = 0.65	+ Mean Absolute Error for ANN = 0.51 + Mean Absolute Error for Hedonic Pricing Model = 1.23

ID	Attribute	Percentage Missing	Min	Max	Mean	Std. Dev.
1	Sold Price	0	1,127	695,000	141,380	105,372
2	Sold Year	0	2015	2018	1026.9	0.4
3	Beds	0	1	7	3.0	0.8
4	Baths	0	1	6	1.8	0.7
5	Size (sqft)	0	458.5	7481	1715.6	751.9
6	Lot (sqft)	9	1742	958,320	65,696	125,975
7	Date Built	6	1790	2017	1957	40.5
8	Last Remodel Year	65	1900	2017	1977.5	19.1
9	Tax Amount	9	65	12,246	1,996	1,580
10	Assessment	7	850	545,420	104,722	88,491
11	Elementary School Score	1	0	9	4.4	1.7
12	Middle School Score	0	0	9	4.9	1.6
13	High School Score	0	0	9	4.7	1.4
14	Previous Sold Price [1]	66	1	510,000	118,452	84,599
15	Listed Price	41	10,400	850,000	179,605	119,886
16	Changed Price	76	10,400	724,999	174,386	129,463
17	Walkability	11	0	82	20.7	21.7
18	Rooms	71	1	17	6.7	2.6
19	Garage Space	62	1	17	2.1	1.3
20	Floor	51	1	5	1.3	0.5
21	Coms' Sold Price	21	2,000	513,276	142,294	90,583
22	Restaurants	18	0	76	11	14
23	Grocery Stores	18	0	13	1	2
24	Nightlife	18	0	21	2.5	4.6

[1] There are two houses in Cayuga, NY whose previous sold prices are $1. These are considered anomalies but the two houses are still kept in the dataset because their current sold prices and other attributes are within normal ranges.

Table 6: Numeric Attributes

ID	Attribute	Representatives
1	Sold Month	Jan, Feb, etc.
2	House Type	Single Family, Condo, Townhouse
3	Street Parking	Yes, No
4	Floor Type	Hardwood, Carpet, Tile
5	Heating System	Gas, Electric, Center
6	Cooling System	Center, Electric
7	Appliances	Dishwasher, Garbage Disposal, Oven, Fridge
8	External Material	Stone, Wood, Brick
9	House Style	Colonial, Contemporary, Ranch, Bungalow
10	Roof Type	Asphalt, Composition, Metal, Shake Single
11	Garage Type	Detached, Attached

Table 7: Non-Numeric Attributes

Table 8: Most Important Attributes For Selected Counties

ID	Cayuga (NY)	Cowlitz (WA)	Hunt (TX)	Montgomery (IL)	Upson (GA)
1	Size	Size	Size	Beds	Beds
2	Date Built	Tax Amount	Lot	Baths	Baths
3	Last Remodel Year	Assessment	Date Built	Size	Lot
4	Tax Amount	Listed Price	Tax Amount	Date Built	Date Built
5	Assessment	Electric Heating	Assessment	Tax Amount	Last Remodel Year
6	Listed Price	Center Cooling	Listed Price	Assessment	Assessment
7	Walkability	Asphalt Roof	Walkability	Previous Sold Price	Composition Roof
8	Dishwasher	Coms' Sold Price	Hardwood Floor	Listed Price	Coms' Sold Price
9	Garbage Disposal	Beds	Tile Floor	Walkability	Changed Price
10	Contemporary Style	Baths	Floor	Dishwasher	Walkability
11	Coms' Sold Price	Elementary School Score	Coms' Sold Price	Restaurants	Dishwasher
12	Previous Sold Price	Middle School Score		Bungalow Style	Oven
13		High School Score		Coms' Sold Price	Restaurants
14					Grocery Stores
15					Nightlife

ID	Attribute	Cayuga (NY)	Cowlitz (WA)	Hunt (TX)	Montgomery (IL)	Upson (GA)
1	Assessment	x	x	x	x	x
2	Coms' Sold Price	x	x	x	x	x
3	Date Built	x		x	x	x
4	Listed Price	x	x	x	x	
5	Size	x	x	x	x	
6	Tax Amount	x	x	x	x	
7	Walkability	x		x	x	x
8	Baths		x		x	x
9	Beds		x		x	x
10	Dishwasher	x			x	x

Table 9: Set Of Most Important Attributes

References

[1] Paul K. Asabere and Forrest E. Huffman. "Price Concessions, Time of the Market, and the Actual Sale Price of Homes". In: *Journal of Real Estate Finance and Economics* 6 (1993), pp. 167–174. URL: https://link.springer.com/article/10.1007/BF01097024.

[2] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[3] Rochard J. Cebula. "The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and Its Savannah Historic Landmark District". In: *The Review of Regional Studies* 39.1 (2009), pp. 9–22. URL: journal.srsa.org/ojs/index.php/rrs/article/download/182/137.

[4] *Consumer Housing Trends Report 2016*. Zillow Group. Accessed: 11/10/2017. 2016. URL: https://www.zillow.com/research/zillow-group-report-2016-13279/.

[5] Harris Drucker et al. "Support vector regression machines". In: *Advances in neural information processing systems*. 1997, pp. 155–161.

[6] Gang-Zhi Fan, Seow Eng Ong, and Hian Chye Koh. "Determinants of House Price: A Decision Tree Approach". In: *Urban Studies* 43.12 (2006), pp. 2301–2315. URL: journals.sagepub.com/doi/pdf/10.1080/00420980600990928.

[7] Tin Kam Ho. "Random decision forests". In: *Document analysis and recognition, 1995., proceedings of the third international conference on*. Vol. 1. IEEE. 1995, pp. 278–282.

[8] Daniel R. Hollas, Ronald C. Rutherford, and Thomas A. Thomson. "Zillow's estimates of single-family housing values." In: *Expert Systems with Applications* 78.1 (2010). URL: http://www.freepatentsonline.com/article/Appraisal-Journal/220765044.html.

[9] Gu Jirong, Zhu Mingcang, and Jiang Liuguangyan. "Housing price based on genetic algorithm and support vector machine". In: *Expert Systems with Applications* 38 (2011), pp. 3383–3386. URL: http://www.sciencedirect.com/science/article/pii/S0957417410009310.

[10] Kelvin J. Lancaster. "A New Approach to Consumer Theory". In: *The Journal of Political Economy* 74.2 (1966), pp. 132–157. ISSN: 0303-2647. DOI: 10.1.1.456.4367&rep=rep1&type=pdf. URL: http://www.jstor.org/stable/1828835.

[11] *Number of houses sold in the United States from 1995 to 2016*. www.statista.com. Accessed: 11/10/2017. URL: https://www.statista.com/statistics/219963/number-of-us-house-sales/.

[12] *Quick Facts: Resident Demographics*. National Multifamily Housing Council. Accessed: 11/11/2017. 2017. URL: http://www.nmhc.org/Content.aspx?id=4708.

[13] *Random Forests by Leo Breiman and Adele Cutler*. URL: https://www.stat.berkeley.edu/~breiman/RandomForests/.

[14] Sherwin Rosen. "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition". In: *The Journal of Political Economy* 82.1 (1974), pp. 34–55. URL: http://people.tamu.edu/~ganli/publicecon/rosen74.pdf.

[15] Hasan Selim. "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network". In: *Expert Systems with Applications* 36 (2009), pp. 2843–2852. URL: www.sciencedirect.com/science/article/pii/S0957417408000596.

[16] G. Stacy Sirmans, David A. Macpherson, and Emily N. Zietz. "The Composition of Hedonic Pricing Models". In: *Journal of Real Estate Literature* 13.1 (2005), pp. 3–43. URL: http://www.jstor.org/stable/44103506?seq=1#page_scan_tab_contents.

[17] Alex J Smola and Bernhard Schölkopf. "A tutorial on support vector regression". In: *Statistics and computing* 14.3 (2004), pp. 199–222.

[18] Danny P. H. Tay and David K. H. Ho. "Artificial Intelligence and the Mass Appraisal of Residential Apartments". In: *Journal of Property Valuation and Investment* 10.2 (1992), pp. 525–540. URL: http://www.emeraldinsight.com/doi/abs/10.1108/14635789210031181.

[19] *The Price of Overpricing: How Listing Price Impacts Time on Market*. Zillow. Accessed: 03/06/2018. 2016. URL: https://www.zillow.com/research/overpricing-impacts-time-market-12476/.

[20] *Trulia Estimate*. Trulia. Accessed: 11/11/2017. 2017. URL: https://www.trulia.com/trulia_estimates/.

[21] Vladimir N. Vapnik and Alexey Ya Chervonenkis. "On a class of algorithms of learning pattern recognition." In: *Automation and Remote Control* 25.6 (1964).

[22] *Zestimate*. Zillow Group. Accessed: 11/11/2017. 2017. URL: https://www.zillow.com/zestimate/.