Housing Rent Prediction

Kh.M Fuad Harun
Md.Tarikul Islam
August 5,2019

Abstract

This paper explores the question of how house prices in different area's 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 DSCC&DNCC, 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. 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

Introduction

According to the DSCC&DNCC Census Bureau, 60,000 houses were sold in the Dhaka City in 2016 . In addition, 55% of all families owned houses in 2016 who live in Dhaka Metropolition City. For the people who stay in Dhaka City they 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 housing sell and share 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.

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 and it is. 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.

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 statis- tically 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 fac- tor 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 measure- ment 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 pre- dicting 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) regres- sion to forecast the housing prices in China in between 1993 and 2002 and in certain dis- trict 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 characteristics such as model type and age more than the owners of 5-or-more-room flats.

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" .

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:

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

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. 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 . 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.

rooms, bedrooms, and the size of the house. 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 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.

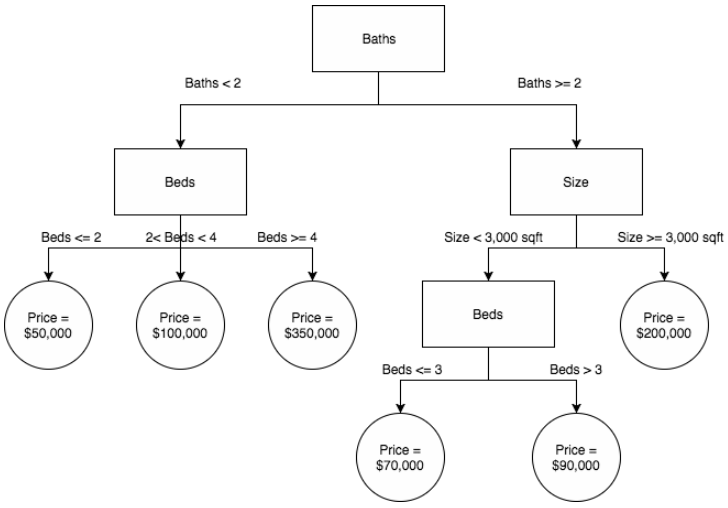


Figure 1: A sample decision tree model

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 ≥ 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. 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 . Instead of fitting a best fitted line over the observations like Linear Regression, SVR with a particular kernel , called Linear SVM Regres- sion, 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.

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.

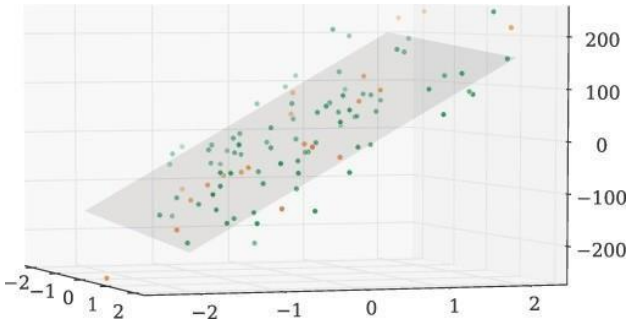


Figure 2: Hyperplane Fitting

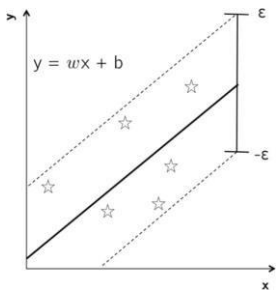


Figure 3: Linear Support Vector Regression

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).

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

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.this model works well in finding patterns in real-life noisy datasets, such as financial data.

DATA

4.1 Data Collection

This project collects data on 1,457 sold houses from DSCC&DNCC, Trulia and Redfin using Python and Selenium (a browser automation tool) for data scraping. These houses are selected from five area in five different regions of the Dhaka City.

These five counties are chosen based on the following criteria:

1. They are from different area in the Dhaka City.
2. They are among the very good performing area 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 particulararea's.

3. These area's 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.

Area	**No.word**	**No.Thana**	**# of Houses**	**Citywide Baseline**	**My Data's Baseline**
Aftabnagar	7	2	399	16.7	28.6
Basabo	3	1	209	8.7	21.1
Banasree	5	3	310	10.7	13.9
Bashundhara	3	2	195	19.8	39.5
Gulshan	11	3	354	29.3	27.7

Table 1: Selected Area's' Information

Table 1 summarizes the five area's' information, including the *Citywide Base- line* for each area, which is the percentage of houses on Zillow whose Zestimates fall within 5% of the houses' actual sold prices.

4.2 Data Processing

This project collects 1,457 houses from different housing websites . The reason is to make sure that the scraped housing data is as accurate as possible. Since all these websites get data from listing services or other third- party companies, inconsistency and mistakes in data are unavoidable.

All three websites record that the house was last sold in October, 2018. However, the sold price of the houses is inconsistent. Besides sold price, other housing characteristics are also subject to the same comparison algorithm.

Besides inconsistency in data values across these 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. 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. .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.

. Among the five

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

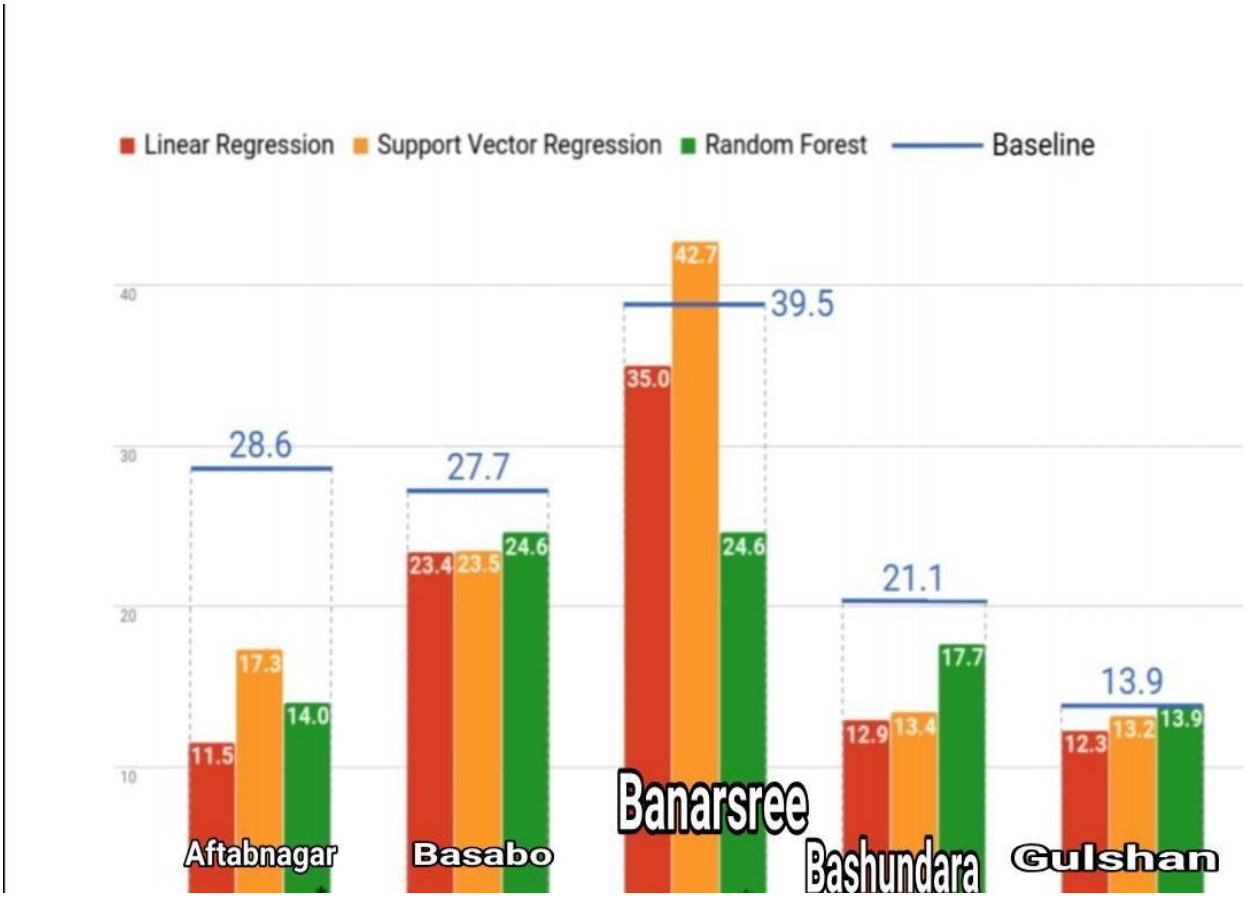


Figure 6: Models' Prediction Scores Compared To Baselines

Area's, Aftabnagar appears to have the worst performance. For this area's dataset, SVR is the best performing algorithm, followed by RF and LR. For Aftabnagar, the pre- diction score gap between the baseline and SVR's performance is 11%. However, SVR is statistically better than LR for this particular area.

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 impor-

tant 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-

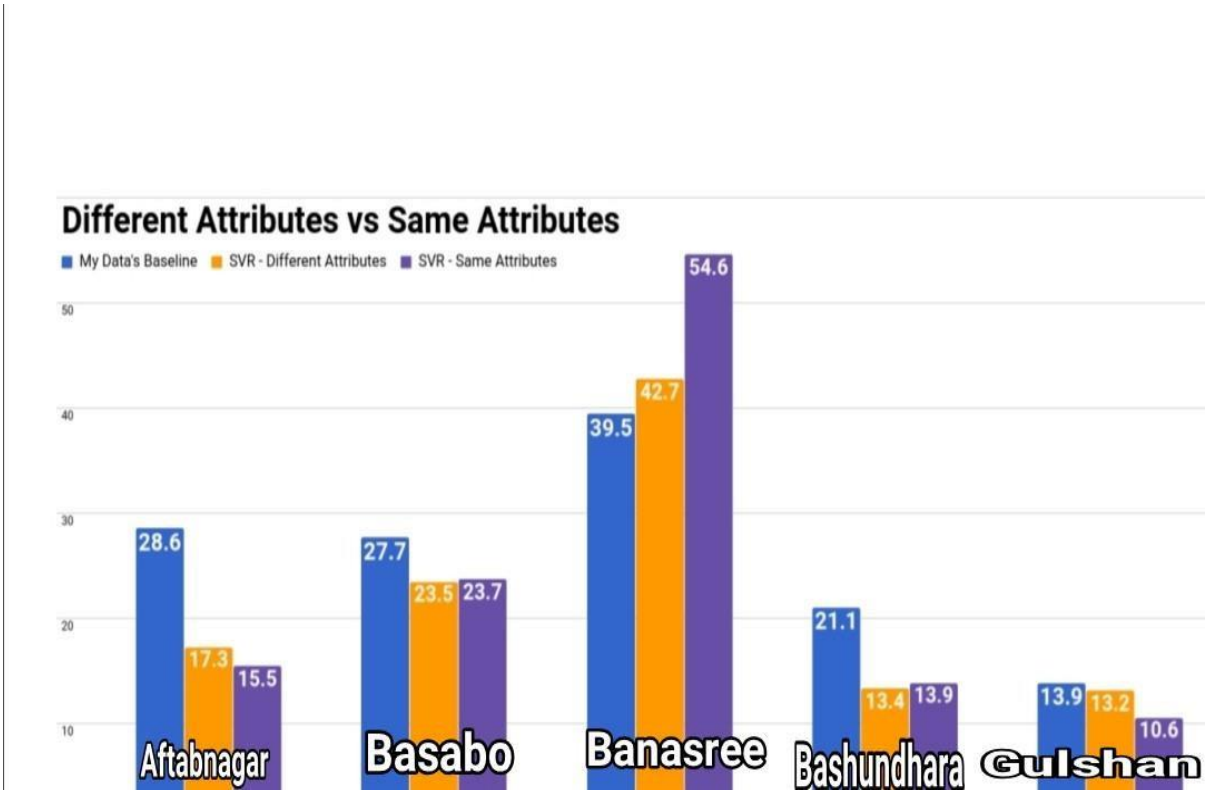


Figure 7: SVR - Different Attributes vs. Similar Attributes

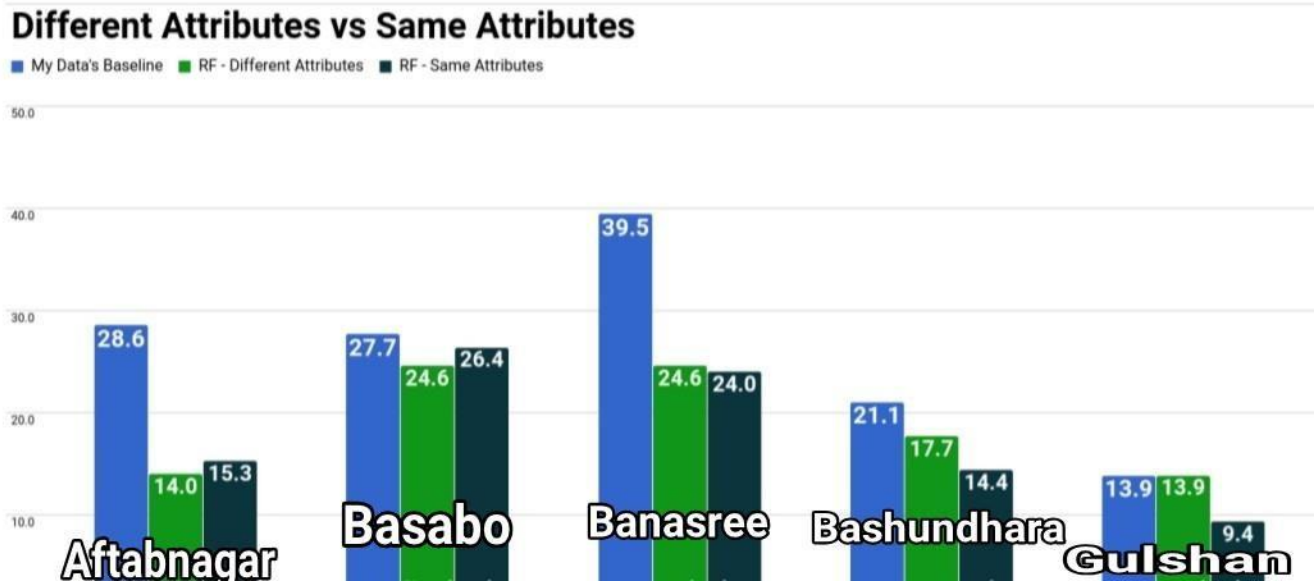


Figure 8: RF - Different Attributes vs. Similar Attributes

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 area's datasets they appear on.

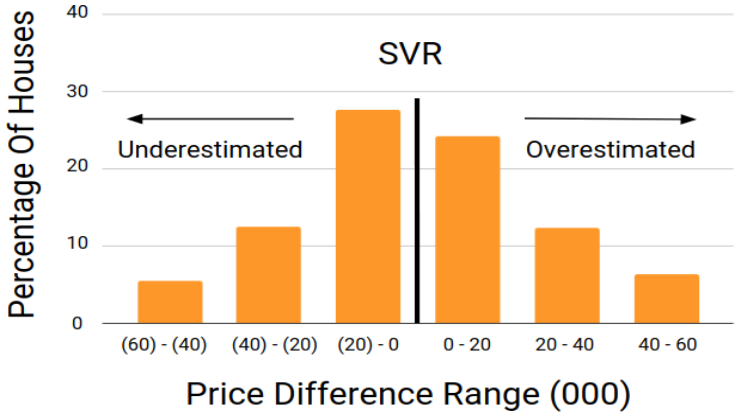
Figure 7 and 8 display the prediction score comparison between having one common set of attributes for 5 area's and having different sets of attributes for different counties. Figure 7 uses SVR as the algorithm whereas Figure 8 uses RF. These figures suggest that switching from different set of attributes 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

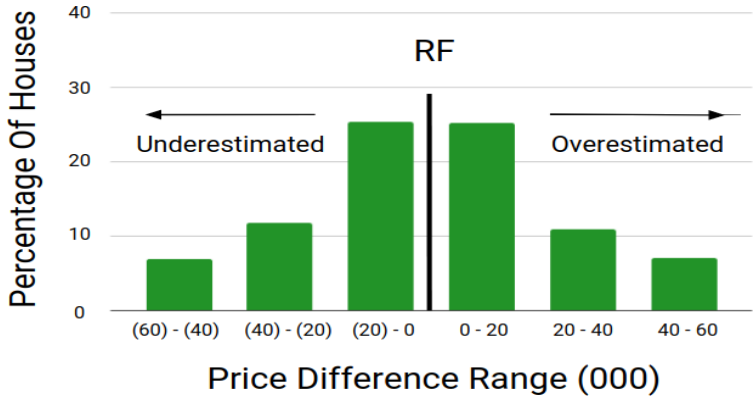
The horizontal axis represents price difference range (in thousand dollars) while the vertical axis shows the percentage of houses that fall within a certain price difference range. Negative price difference, which means the predicted price is lower than the actual sold price, is put in brackets, as shown on the horizontal axis of the graph.

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

6 Conclusion

Using a dataset of 1,457 houses from 5 different area's scraped from website, 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 area's 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 area's. The overesti- mated 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 performingones

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.

.