Forecasting Sales Prices on Real Estate in Queens, NY

A project for STAT 422 at Wharton Business School

March 3, 2017

By: Carl-Oscar Gustafson

In collaboration with:

Juan Manubens

Jack Soslow

David Baxter

Raul Mendez

William Fry

Liubov Vazhenina

Kwang Jun Lee

Abstract

An online real estate platform called Zillow provides its visitors with so-called "zestimates," which are meant to serve as estimates of future sales prices. Upon discovering the metrics' questionable performance, our team of undergraduate and graduate students set out to create more accurate estimates of future sales prices using a statistical model based on sound predictive modelling principles. This paper will explain the methodologies which went into creating the statistical, discuss the results of the model, and potential ways in which the model could have been improved further.

1. Introduction

A predictive model is a statistical model which is formulated through collected data with the purpose to forecast a particular output, or *response*, based on specific data inputs, or *predictors*, which are hypothesized to be related to the response either through correlation or causation. The data inputs on which the model is formulated are chosen through two processes called featurization and sampling. In the first, the features/attributes/characteristics believed to be related to the response are chosen, so that the data related to already realized responses can be collected. In the latter, the specific observations which are believed to help formulate the most accurate model are chosen. To create a predictive model forecasting real estate sales prices in Queens, NY, we collected data on features such square footage and number of rooms of listings similar to those which we intended to forecast. The response of our model can therefore be said to be "Sold Price" and the units of observation which we used to formulate our model are the listings from which we collected data. We used a combination of open data provided by our professor and data scraped from the Zillow website using an online tool called MTurk. We collected data on a total of n=2300 observations and the data provided by our professor Adam Kapelner made up another 541 observations. After cleaning up the data and imputing missing data using various methodologies described later in the paper, a non-parametric model was made based on an algorithm called Random Forest. Before making any predictions, the model was tested on assigned out-of-sample data and performed an out-of-sample R^2 of approximately 91.5%. Subsequently, the formulated model was used to predict 947 unsold listings.

2. The Data

The two sources used to collect the data were two websites: nycopendata.socrata.com and msli.com. Data on historical real estate listings was collected from the former website and geographically related data such as crime rates, middle- and high schools, as well as demographics related to the listings was collected from the latter website to complement the limited data on each listing. In total, we collected data on n=2841 listings to forecast the sales prices of a total of 947 listings specified in a prediction dataset.

2.1. Sampling

The 2841 historical observations used to formulate the model came from two different data samples: 2300 listings collected by a group of Wharton MBA students, and 541 listings collected by our professor Adam Kapelner. Although we are not familiar with the exact sampling process utilized in collecting information about the listings, the observations were likely collected randomly, since the size of the sample (n=2841) made up a significant proportion of the total historical listings available - approximately 3500. The data includes listings which are representative of those for which we intended to forecast future sales prices in terms of location, square footage, style, etc. However, given that the data includes listings on properties sold between January, 2016 and February 2017, a time frame which does not overlap with the listings in the prediction set, it could be argued that the data cannot be effectively generalized to forecast prices outside of that time frame, since those predictions would be extrapolations and time is arguably an important factor in forecasting real estate prices.

2.2. Featurization

For each listing we collected data from two different sources. First, data was collected from mlsli.com on the following 25 features:

- **Sold Price -** The response variable indicates sales prices and is only included for historical listings. The variable ranged from $55,000 to $1,000,000 and had an average of $315,200 with a standard deviation of $166,740.
- **Full address -** Provides information about the address, city, state, and zip code of each listing and was used in pulling data from open data sources. It had too many levels to be used as a predictor.
- **City -** Refers to the city in Queens, New York where the listed real estate is located. The variable has 52 distinct levels. The most frequent cities were: Bayside (17.8%), Flushing (17.0%), Hills (6.5%), Forest Hills (5.9%), and Rego Park (5.2%). The least frequent cities (less than 1%) were RochDale, Richmond Hill, Richmond Hill N., Laurelton, and Manor.
- **Zip -** Refers to the zip code pertaining to the location of the listed real estate. The variable had 49 distinct levels. The most frequent zip codes were 11375 (11%), 11360 (9%), 11374 (7.3%), 11364 (6.9%), 11354 (6.6%). The least frequent zip codes (less than 1%) were 11416, 11429, 11420, 11434, and 11366.

- **Style -** Provides information on the style of the listing. The variable had four different levels: Co-op (75%), Condo (23%), Colonial (1%), and Cape (1%).
- **Approximate Year Built -** Numeric variable corresponding to the year in which the listed real estate was built. The variable ranged from 1890 to 2017 with an average of 1962 and a standard deviation of 19.9 years.
- **Construction -** Refers to the type of construction of the listed real estate. After consolidation and cleaning, there were 36 different levels. The most common were Brick (92.2%), Concrete (3.1%), Cinder (0.9%), Frame (0.6%), and Stucco (0.5%). The least frequent types were Cement, Brick Siding, Concrete/Cinder, Masonary, and MSN/Frame, which only hold one observation each.
- **Approximate Interior Square Footage -** Continuous variable specifying the interior size of the listed real estate in terms of sq. ft. The variable ranged from 1 to 514 with an average of 316 and a standard deviation of 146.
- **Number of Floors in Building -** The number of floors in the building of the listed real estate ranged from 1 to 33 with an average of 22 and a standard deviation of 9.
- **Fuel -** Represents the type of fuel used to power the listed real estate. Given the variability of inputs, levels were consolidated to the following four levels: GAS (60.9%), OIL (32.0%), OTHER (4.5%), and ELEC (electric, 2.4%) with distribution percentages specified.
- **Total Rooms** – The total number of rooms in listed real estate ranged from 1 to 45 (a potential erroneous value) with an average of 4.1 and a standard deviation of 1.6.
- **Bedrooms –** The number of bedrooms in the listed real estate ranged from 1 to 8 with an average of 2.59 and a standard deviation of 0.73.
- **Baths Full** – The number of full bathrooms in the listed real estate ranged from 1 to 6 with an average of 1.57 and a standard deviation of 1.18.
- **Baths Half** - The number of half bathrooms (no shower or bath) in the listed real estate ranged from 1 to 3 with an average of 1.06 and a standard deviation of 0.23.
- **Kitchen Type** – Refers to the type of kitchen in the listed real estate. After consolidating the levels by changing empty cells and levels of only one observation to Unspecified, the variable had four distinct levels: Eat In (42.9%), Efficiency (39.9%), Combo (15.8%), and Unspecified (1.4%).

- **Garage –** The binary variable indicates the existence of a garage or not in the listed real estate. 88.6% had a garage whereas 21.4% did not.
- **Walk Score –** The variable is an index ranging from 0 to 100 and indicates the walkability of the surrounding area of the listed real estate. The average walk score was 81.81 with a standard deviation of 14.1.
- **Dogs Allowed** – The binary variable indicates whether or not the listed real estate permits dogs with a majority not permitting dogs (65.9%) and a minority permitting them (34%).
- **Cats Allowed –** An identically structured variable, but for cats with 61.3% permitting cats and 38.6% forbidding them.
- **School** – A factor variable representing the school districts of the listed real estate. The variable had 25 distinct levels / community districts. The most frequent were 25 (35.2%), 26 (22.3%), 28 (22.2%), 30 (6.6%) and 24 (5.5%). The least frequent (less than 1%) were 13, 14, 17, 18, and 20.
- **Common Charges** – A continuous variable specifying the cost of common services and amenities associated with the listed real estate. The variable ranges from $0 to $2,499 with an average of $385 and a standard deviation of $223. It is highly right-skewed.
- **Maintenance** – A continuous variable specifying the cost of maintenance for the listed real estate. It is important to keep in mind that since no frequency of payment was specified, maintenance could indicate both monthly and yearly costs, although such information is unknown.
- **Parking Charges -** A continuous variable specifying the cost of parking for the listed real estate. As with maintenance, no specific timeframe is specified. The charges range from $0 to $500 with an average of $86.09 and a standard deviation of $68.5.
- **Total Taxes -** A continuous variable specifying the total cost of taxes associated with the listed real estate. The variable ranges from $0 to $9,300 with an average of $2,249 and a standard deviation of $1,393.
- **% Deductible –** A variable representing the tax deductible portion of mortgage and maintenance based on tax bracket. It ranges from 0 to 100%, with an average of 43.5% and a standard deviation 7.4%.

Second, data on the following 14 additional features related to the location of each listing was collected from nycopenddata.socrata.com:

- **Robbery (1, 3, and 5-year averages) –** Represents robbery rates in the surrounding area with a one year average rate ranging from 44 to 384 cases with a mean of 152 cases and a standard deviation of 87. The three year average rate ranged from 57 to 404 cases with a mean of 163.5 and a standard deviation of 87. The five year average rate ranged from 64 to 417 cases with a mean of 179 and a standard deviation of 91.
- **Burglary (1, 3, and 5-year averages) -** Represents burglary rates in the surrounding area with a one year average rate ranging from 114 to 358 cases with a mean of 229 cases and a standard deviation of 82. The three year average rate ranged from 133 to 419 cases with a mean of 279 and a standard deviation of 110. The five year average rate ranged from 138 to 435 cases with a mean of 289 and a standard deviation of 109.
- **Felony Assault (1, 3, and 5-year averages) -** Represents felony assault rates in the surrounding area with a one year average rate ranging from 58 to 444 cases with a mean of 168 cases and a standard deviation of 103. The three year average rate ranged from 59 to 440 cases with a mean of 177 and a standard deviation of 104. The five year average rate ranged from 60 to 458 cases with a mean of 175 and a standard deviation of 101.
- **Grand Larceny (1, 3, and 5-year averages) -** Represents grand larceny rates in the surrounding area with a one year average rate ranging from 349 to 936 cases with a mean of 602 cases and a standard deviation of 239. The three year average rate ranged from 410 to 908 cases with a mean of 612 and a standard deviation of 207. The five year average rate ranged from 419 to 839 cases with a mean of 587 and a standard deviation of 176.
- **Grand Larceny of a Motor Vehicle (1, 3, and 5-year averages) -** Represents the rate of grand larceny of motor vehicles in the surrounding area with a one year average rate ranging from 72 to 248 cases with a mean of 136 cases and a standard deviation of 49. The three year average rate ranged from 82 to 247 cases with a mean of 147 and a standard deviation of 57. The five year average rate ranged from 86 to 293 cases with a mean of 160 and a standard deviation of 64.
- **Murder and Non-Negligent Manslaughter (1, 3, and 5-year averages) -** Represents murder and non-negligent manslaughter rates in the surrounding area with a one year average rate ranging from 0 to 9 cases with a mean of 2.2 cases and a standard deviation of

1.8. The three year average rate ranged from 0 to 8 cases with a mean of 1.8 and a standard deviation of 1.7. The five year average rate ranged from 0 to 11 cases with a mean of 2.1 and a standard deviation of 2.

- **Rape (1, 3, and 5-year averages) -** Represents rape rates in the surrounding area with a one year average rate ranging from 6 to 37 cases with a mean of 14.6 cases and a standard deviation of 8.8. The three year average rate ranged from 3 to 33 cases with a mean of 15.4 and a standard deviation of 9.5. The five year average rate ranged from 4 to 36 cases with a mean of 14.6 and a standard deviation of 8.5.
- **Total 7 Major Felony Offenses (1, 3, and 5-year averages) -** Represents the total of all seven major felony offense rates in the surrounding area with a one year average rate ranging from 661 to 1948 cases with a mean of 1304 cases and a standard deviation of 509. The three year average rate ranged from 752 to 2019 cases with a mean of 1398 and a standard deviation of 526. The five year average rate ranged from 818 to 2051 cases with a mean of 1411 and a standard deviation of 502.
- **Elementary School** – Specifies to which elementary school district the real estate listing is assigned. The most frequent schools are P.S. 169 Bay Terrace (7.8%), P.S. 196 Grand Central Parkway (5.3%), P.S. 221 The North Hills School (5.0%), P.S. 175 The Lynn Gross, Discovery School (3.8%), and P.S. 205 Alexander Graham Bell (3.6%). The least frequent are P.S. 81Q Jean Paul Richter, New York City Academy for Discovery, P.S. 019 Marino Jeantet, P.S. 056 Harry Eichler, and P.S. 088 Seneca.
- **Middle School** - Specifies to which middle school district the real estate listing is assigned. The most frequent schools are Q157 (12.9%), Q025 (9.5%), Q194 (8.7%), Q067 (7.5%), and Q250 (6.5%). The least frequent (less than 1%) schools are Q192, Q226, Q049, Q008, Q087.
- **High School** - Specifies to which high school district the real estate listing is assigned. The most frequent schools are Bayside High School (17.2%), Hillcrest High School (12.7%), Benjamin N. Cardozo High School (12.1%), Forest Hills High School (11.8%), and John Bowne High School (10.1%). The least frequent schools are Long Island City High School (2.4%), John Adams High School (2.1%), Citywide High School Choice (1.7%), Grover Cleveland High School (0.8%), Richmond Hill High School (0.7%).

- **YTD Enrollment 2011** – The variable specifies the percentage of eligible children who are enrolled in a school in the area of the listed real estate. It ranged from 63.8% to 93.3% with a mean of 89.0% and a standard deviation of 4.9%.
- **YTD Attendance 2011 -** The variable specifies the number of eligible children who are enrolled in a school in the area of the listed real estate. It ranged from 10,200 to 15,370 with a mean of 35,870 and a standard deviation of 5485.
- **NOV Count –** Represents the number of violations the Department of Housing in NYC submitted to a property. It was counted using REGEX to parse out the parts of the address and query the dataset containing all housing violations in 2016. The data ranged from 0 to 333 violations with an average of 2.4 and a standard deviation of 13.1.

2.3. Missingness

In terms of the data collected on the 27 features of the different listings specified in the data sample, missing values were common and sometimes made up significant portions of the total sample size. Before imputation, the data was cleaned by deleting columns which either had no overlap with those in the prediction set or were deemed useless due to the nature of those features. Levels of different predictors were also consolidated by controlling for different spelling or abbreviations of particular words or phrases. In terms of imputing missing values, it was first to be decided for which predictors to impute values. These predictors were determined using the Chi-Squared Test of Independence to check for any predictive power of the predictor with regards to the response, "Sold Price." Only values for those predictors which either demonstrated significant dependence with regards to the response, or which had some significant non-linear relationship with the response were imputed. The imputation process was carried out by a Python script written by William Fry, a fellow student in the class, and followed the following imputation hierarchy: (a) if values exists for other listings with the same address, impute the average of those listings, (b) if values exists for other listings with the same street name and zip code, impute the average of those listings, (c) if values exists for other listings with the same zip code, impute the average of those listings, (d) if listings with missing values do not overlap with other listings in terms of address, street name, or zip code, take the average of all listings. When the predictors were categorical, the missing value was imputed using the mode. With regards to the data collected on the 14 features related to the locations of the listings, no data was missing and thus no imputation was necessary.

Although it seemed much of the missing data was NMAR – “Not Missing At Random” – the data was treated as MAR – “Missing At Random” as imputed so that potential additional predictive power could be extracted and the observations could be included in the formulation of the model. We did not include any dummy variable indicating missingness, although doing so could have been useful given a pattern mixture model where information was assumed missing due to an unwillingness to share.

3. Modeling

In creating the statistical model used to forecast real estate sales prices, the Random Forest algorithm was used to formulate a non-parametric model. A non-parametric model was used since non-linear relationships and interactions were expected between many of the predictors and the response, which could not have been captured effectively by a parametric model. Furthermore, given that prediction accuracy was the main goal and a non-parametric model would minimize the mode misspecification error and approximate the functional form of f in a better way than a parametric model would, losing interpretability as a result of using Random Forest was not a concern. In formulating the model, no custom variables were utilized. Given that Random Forest utilizes technique such as bagging, boosting, and sampling of features, which randomly selects samples of the entire dataset and samples of features in creating each tree, overfit is avoided and is not a concern in using the model for out-of-sample predictions. Because the model is non-parametric, it does not provide any information with regards to which variables were most important in predicting sales prices - such information was lost because of choosing Random Forest.

4. Results

In-sample, the R^2 was 98.2%. At first glance, the value of in-sample R^2 may seem high and raise suspicion of overfitting. However, since it is known that each tree used in Random Forest is a “weak learner” and overfits in-sample, a high in-sample R^2 is to be expected. For that same reason, Random Forest also generated a comparatively low in-sample RMSE of $23,432. Out-of-bag, the model generated an R^2 of 91.51% and an RMSE of $48,580. The R^2 implies that the model successfully could explain 91.51% of the variance through its use of the predictors, a number which is quite high. Unfortunately, however, the out-of-bag RMSE was also quite high compared to the range of sold properties which sold at prices from $55,000 to $999,999, and produced a proportionally wide margin of error. Although an out-of-sample wasn't specified or used in testing our model's future accuracy, out-of-bag provides an equally good estimate since these are effectively the out-of-samples for each tree.

5. Discussion

In building the model, most of our time was spent obtaining data, cleaning it and making decisions with regards to missing data. Since our time and technical abilities were limited, we allocated limited consideration with regards to sampling and featurization and had to employ simplified methodologies in imputing missing values. Instead of thinking deliberately about what listings to gather information from and what information to gather from each listing, we tried to gather as much data as possible – as many features and observations we could capture – and then complement that data with open data we deemed potentially useful. In selecting which features to include and how to deal with missing values, we mostly relied of "common sense" based on cognitive models formed through our life experiences. Although these were helpful as a guideline, we would likely have benefitted from a more structured approach. In considering how the model could have been improved, I recognize several opportunities with regards to sampling, featurization, dealing with missingness, and formulating the model using Random Forest. In terms of reducing the generalization error through improved sampling, a larger number of listings could have been gathered from the website to achieve a wider distribution of various predictors. By considering the features of the prediction set more closely, such as the location and price range of the listings, samples could have been gathered more intelligently to better represent the listings in the prediction set. Featurization also opens up opportunities for improvement. Since many features present in sample dataset were missing from the prediction set, a more accurate prediction could potentially have been obtained by gather those features for the listings in the prediction set as well. The prediction accuracy of our model could also have been improved if we had complemented our sample and prediction sets with additional potentially relevant features, such as the median income of the location, length of the listing description, number of photos, and time since published. Many more features could probably have been thought of in a brainstorming session, but with limited time and knowledge, we likely would not have been to collect those additional features without outside assistance. I believe our biggest opportunity for improvement lies in the most timeconsuming of activities, namely in imputing data based on available information. Instead of using averages by location, chances are more accurate imputations could have been made by the use of Random Forest modelling predicting for missing values of certain features. During the imputation process, we noticed that some columns included unparsed text specifying further details of the listings which could have been helpful in making more accurate imputations. However, we considered that it would likely not be economically efficient to spend time parsing the text without the help of a computer. Additionally, the model could have been improved by tuning the parameters used in Random Forest, such as the number of trees and number of sampled features per tree. Although it likely would not have produced any significant improvements (default settings are generally effective), there is a chance marginal improvements could have been made. Since the model was created with the assumption of a stationary model, I don't believe that the model is production ready just yet. In order to adjust for time dependence, features related to time should first be incorporated before shipping of the model and the previously stated suggestions for improvement should also be implemented. All in all, it has been exciting using the concepts in class to predict something as tangible and practical as real estate sales prices. I am curious to see how the model would have performed with the suggestions implemented, but either way, I believe there is a good chance the model produced will beat the "zestimates" provided by the website given it is to a large extent based on sound statistical principles.

Acknowledgements

I want to thank all my teammates for their help in cleaning and imputing the data as necessary, as well as providing helpful advice with regards to the model building process and every component of it. I also would like to thank the MBAs and professor Adam Kapelner for helping us gather data to use in the modelling process. Lastly, I want to thank William Fry for blessing us with his extensive knowledge of programming.