STAT 422 - Project Final Writeup

By: Juan Manubens

Collaborators: William Fry, Raul Mendez, David Baxter, Jack Soslow

Abstract - ***While it would be naively convenient to believe that Real Estate Sale Prices can be modeled as a function of features of both the listed property and external attributes related to it such as economic factors or ZIP-code characteristics, an honest Data Scientist would caution us to be skeptical. Despite massive amounts of available data points related to these factors, many predictive models found on various real estate platforms such as Zillow are remarkably inaccurrate. This paper attempts to build a model on a first-principles basis, supported by previous Statistics Knowledge and further supported by this class - a densely-packed half-semester course on Predictive Modeling and Statistical Theory, and to determine whether the current benchmarks can be outperformed.***

I. Introduction

Our prediction targets are the Future Sale Price of Real Estate listings in Queens, New York. Thus, our response variable in this context is the sold price in USD,

$$\hat{Y}^*_{ik,t>0} = \hat{P}(x_{ik,t>0})$$

while a given listing, $x_k i$, is the unit of observation, and k represents the features of each observation. These features include relevant data points about each listing, its surrounding neighborhood, and other factors we considered could be relevant. Once a set of features was determined, a combination of Amazon's MTurk and Open Data from NYC was used for collection. The sample of historical data collected a total of $n_{h_1} = 2300$ listings in Queens, sold over the past 3 months. We added a second group of observations consisting of $n_{h_2} = 541$ observations, given to use by our professor Adam Kapelner, PhD. After extensive cleaning, the predictions were made based on a non-parametric model generated using the Random Forest algorithm. We did not assume nor expect relationships to be exactly linear and were more concerned about predictive accuracy rather than simplicity or interpretability. The resulting model has an OOS R^2 of 91.49% and was used to predict the sale prices of $n_p = 947$ listings currently for sale.

II. The Data

The data used for this project comes from two different sources: *MLSLI.com* and *NYCOpenData.socrata.com*. The observations all came from the former and were extracted from the website using Mechanical Turk. After the observations were sampled, we used three main datasets from NYC Open Data to add a clearer picture to a given listing's geo-spatial context, given that school, crime and demographic data were unavailable to be extracted from *MLSLI* because of the MTurk task constraints. The prediction set consisted of $n_p = 947$ observations currently for sale. These observations were processed through the same scripts used for the $n_{h_{1+2}} = 2841$ historical observations.

- **Sampling**

The $n_{h_{1+2}} = 2841$ historical observations were sampled via Mechanical Turk. This historical set was composed of two samples, $x_{ij \in n_{h_1}}$ and $x_{ij \in n_{h_2}}$. One was received from a group of Wharton MBAs who created a batch of 2300 hits for listings in Queens on *MLSLI*. The second was received from our professor Adam Kapelner, PhD, who created a batch of 541 hits for us, originating from the same source. While the exact sampling procedure that the MTurks followed is unknown, we have assumed that the listings were sampled randomly.

The sampled data are representative of listings on *MLSLI* in Queens from January 1, 2016 to roughly February 15, 2017. Since time, t, is a non-negligible factor in Real Estate pricing, while often a proxy for other indicators, it can be argued that predicting a price for a property being sold outside of the time period in which this data was collected, $\hat{Y}^*_{ik,t_{data} \notin t^*} = \hat{P}(x_{ik,t_{data} \notin t^*})$, is an extrapolation. A model with only a year's worth of data will not be able to effectively generalize beyond the data collection timeframe, t_{data}, by much. Fortunately, since we are only considering listings used as units of analysis for prediction set to be sold in the next month, we expect the majority of our informa allowing us to comfortably apply our model to that data without great risk.

- **Featurization**

Featurization is the process of manually designing and selecting inputs which will link raw data to a response via specifying features. This study split raw, naturally represented data into two feature categories: (a) features belonging to a listing and (b) features of the surrounding geo-spatial area.

27 features were selected that specify features of a given listing:

- **Full Address** - Corresponds to the address of the listing, including city, state and zipcode. This was used to pull more features from Open Data, and had too many levels to be a feature itself.
- **Style** - A factor variable corresponding to the style of the listing. Levels included "Cape" (1%), "Co-op" (75%), "Colonial" (1%), "Condo" (23%).
- **Zip** - A factor variable corresponding to the zipcode of the listing. In all there were 49 zipcodes in Queens. The 5 most frequent were 11375 (11%), 11360 (9%), 11374 (7.3%), 11364 (6.9%), 11354 (6.6%). The 5 least frequent were 11416, 11420, 11429, 11366, and 11434 - all of which were less than 1%.
- **City** - A factor variable referring to the city within Queens in which the listing is located. There were 52 distinct cities in the sample that we processed. The most frequent cities were Bayside (17.8%), Flushing (17.0%), Hills (6.5%), (5.9%) Forest Hills, and Rego Park (5.2%). The least frequent cities were Manor, RochDale, Richmond Hill N., Laurelton, Richmond Hill - all of which were less than 1%.
- **Baths Full** - This represented the number of full baths in a listing. As a numeric number it's summary statistics convey a mean of 1.572, a range of 1 to 6 full baths, and a standard deviation of 1.18.
- **Baths Half** - This represented the number of half baths in a listing, meaning there was no shower or bath. As a continuous feature, it ranged from 1 to 3, with a mean of 1.055 and standard deviation of 0.23.
- **Bedrooms** - This was the number of bedrooms in a listing, which we treated as a continuous variable. As such it ranged from 1 to 8, with a mean of 2.59 and a standard deviation of 0.733.
- **Total Rooms** - The total rooms includes dining rooms, kitchens, etc in addition to bedrooms. The

number of total rooms ranged from 1 to 45 (which is definitely an outlier at best and a miscoded value at worst). The average was 4.127, clearly skewed right, with a standard deviation of 1.55.

- **Fuel** - Fuel represents the type of fuel that powers the kitchen and is a factor. As much of the data on MLSLI is unstructured - meaning input forms from agents are likely free-form (not a dropdown), we decided to bin factors into 4 levels: ELEC (electric), OTHER, OIL and GAS. These levels represented 2.4%, 4.5%, 32.0%, and 60.9% of the total count, respectively.
- **Kitchen Type** - Kitchen type described the form of the kitchen in a listing. There were three main forms: Efficiency, Combo and Eat In. The rest were either unspecified or only represented 1 observation, so they were grouped as Unspecified. Eat In accounted for 42.9% of kitchens, Efficiency accounted for 39.9% of kitchens, Combo accounted for 15.8% of kitchens, and the rest were Unspecified.
- **Construction** - Construction refers to the type of material of the house. After cleaning, there are 36 distinct levels. The 5 most frequent are Brick (92.2%), Concrete (3.1%), Cinder (0.9%), Frame (0.6%) and Stucco (0.5%). The 5 least frequent are Brick Siding, cement, Concrete/Cinder, Masonary, and MSN/Frame - all of which only have 1 observation and should be binned together.
- **Approx. Year Built** - This is a numeric corresponding to the year the property was built. We have no reason to believe it represents renovations. That being said, the oldest is 1890 and the most recent is 2017. The average year is 1962, with a standard deviation of 19.9 - meaning that there is a wide spread of property age.
- **Approx. Interior Sqft** - The approximate ft^2 of the interior of properties ranged from 1 (which is erroneous - hopefully) to 514. The average was $315.7 ft^2$, with a standard deviation of 146.
- **Number of Floors In Building** - The number of floors in the building ranged from 1 to 33, with a mean of 22 and standard deviation of 8.9. This is consistent with an initial overview of Queens that lacks the skyscrapers and high rises of Manhattan.
- **Garage** - This is a binary factor, corresponding to whether a listing has a garage or not. 88.6% had a garage, while 21.4% did not. This seems relatively high for an urban area.
- **Walk Score** - Walk Score is a number from 0 to 100 that measures the walkability of a property - i.e. whether daily errands require a car or not. The mean was 81.81, with a standard deviation of 14.1.
- **Dogs Allowed** - A self-explanatory variable, we assumed that properties that allowed dogs would explicitly say so, while those that did not would not have this field or it would say no. With that in mind, the majority of listings did not allow dogs (65.9%), with a minority allowing dogs (34%).
- **Cats Allowed** - The same process was applied to cats. Here the opposite result appeared. 61.3% of listings allowed cats, while 38.6% did not.
- **School** - This was a factor representing the broader community district of a listing. There were 25 unique community districts. The 5 most frequent were districts 25 (35.2%), 26 (22.3%), 28 (22.2%), 30 (6.6%) and 24 (5.5%). The 5 least frequent were 14, 17, 18, 13, and 20 - all of which represented less than 1% of observations.
- **Parking Charges** - The parking charges represents the cost of parking for the property. It is fair to assume most either have street parking (that is free or costs hourly) or parking as part of a complex (that is

free or monthly). It's unclear whether these charges are totals for a month or a year, but assuming that the same process holds for all properties this does not matter. The parking charges thus range from 0 to $500, with an average of $86.09 and a standard deviation of $68.5.

- **Maintenance** - Maintenance is a similar to parking charges in the ambiguity of time period. Again, assuming the time period is the same across properties, then it's a valid feature to use. Maintenance costs range from $120 to $4,659, with an average of $798.6 and a standard deviation of $314.8. This shows a heavy skew to the right.
- **Common Charges** - The third cost feature is common charges. Here it includes charges for "common services and amenities", which covers management and operating expenses. Common charges range from $0 to $2,499, with a mean of $385 and a standard deviation of $223. This shows a heavy skew to the right.
- **Deductible %** - This represents the tax deductible portion of mortgage and maintenance based on your tax bracket. It ranges from 0 to 100%, although both seem relatively unreasonable. The average however is 43.5% with a standard deviation of 7.4%.
- **Total Taxes** - This is a continuous variable representing the amount of taxes for the property. It ranges from $0 to $9,300, with a mean of $2,249 and a standard deviation of $1,393.
- **Sold Price** - This is our response variable and is only present for properties already sold. It ranges from $55,000 to $1,000,00, indicating that *MLSLI* either does not sell homes over $1M or the sold price is capped online at that number. The average is $315,200 with a standard deviation of $166,740.

14 features were selected that specify features of a listing's location and came from open data sources:

- **High School** - This is a factor that corresponds to the school which the property is districted to. The data came from http://schools.nyc.gov/schoolsearch/. As a factor, there are 16 levels representing the different schools available. The most frequent schools to which a property was assigned 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 to which a property was assigned are Richmond Hill High School (0.7%), Grover Cleveland High School (0.8%), Citywide High School Choice (1.7%), John Adams High School (2.1%), Long Island City High School (2.4%).
- **Middle School** - This is a similar factor to high schools, which reflects the middle school to which a property is districted. There are 43 different middle schools in Queens to which properties were assigned, many more than the number of high schools. For the sake of brevity, we will use the middle school codes for reference. The 5 most frequent were Q157 (12.9%), Q025 (9.5%), Q194 (8.7%), Q067 (7.5%), and Q250 (6.5%). The 5 least frequent were Q192, Q226, Q049, Q008, Q087 - which accounted for less than 1% of observations.
- **Elementary School** - Lastly, the same process was applied to elementary schools. There were over 120 elementary schools, reflecting the (unrelated) importance of a high teacher-student ratio at early stages of education. One unique element of this was that many properties had more than one option for elementary schools (this was not the case for later education). The most common

options were 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 5 least common were 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 - all of which represented one observation.

- **Grand Larceny: 1, 3, & 5 Yr Averages** - These three features were pulled from NYC Open Data statistics released by NYPD on a per precinct basis. For these three features and the following crime statistics, we first assigned precinct by property location and then pulled the 1 yr, 3 yr, and 5 yr averages for the given crime. For grand larceny, the minimum 1 year was 349 incidents with a max of 936, a mean of 602, and a standard deviation of 239. For three year averages, the minimum number of incidents was 410 with a maximum number of 908. The mean was 612 with a standard deviation of 207. For five year averages, the minimum number of incidents was 419 with a maximum number of 839. The mean was 587 with a standard deviation of 176. These statistics imply that A) certain precincts have notably higher rates of grand larceny than others, and B) crime has been increasing over time.
- **Murder & Non-Negligent Manslaughter: 1, 3, & 5 Yr Averages** - For 1 year statistics, the minimum was 0 murders with a max of 9, a mean of 2.2, and a standard deviation of 1.8. For three year averages, the minimum number of murders was 0 with a maximum number of 8. The mean was 1.8 with a standard deviation of 1.7. For five year averages, the minimum number of murders was 0 with a maximum number of 11. The mean was 2.1 with a standard deviation of 2.
- **Rape: 1, 3, & 5 Yr Averages** - For 1 year statistics, the minimum was 6 rapes with a max of 37, a mean of 14.6, and a standard deviation of 8.8. For three year averages, the minimum number of rapes was 3 with a maximum number of 33. The mean was 15.4 with a standard deviation of 9.5. For five year averages, the minimum number of rapes was 4 with a maximum number of 36. The mean was 14.6 with a standard deviation of 8.5.
- **Grand Larceny of a Motor Vehicle: 1, 3, & 5 Yr Averages** - For grand larceny of a motor vehicle, the minimum 1 year count was 72 incidents with a max of 248, a mean of 135.5, and a standard deviation of 49.1. For three year averages, the minimum number of incidents was 82 with a maximum number of 247. The mean was 147.2 with a standard deviation of 56.6. For five year averages, the minimum number of incidents was 86 with a maximum number of 293. The mean was 160 with a standard deviation of 63.6.
- **Felony Assault: 1, 3, & 5 Yr Averages** - For felony assault, the minimum 1 year count was 58 cases with a max of 444, a mean of 167.6, and a standard deviation of 102.7. For three year averages, the minimum number of cases was 59 with a maximum number of 440. The mean was 176.7 with a standard deviation of 103.8. For five year averages, the minimum number of incidents was 60 with a maximum number of 458. The mean was 174.7 with a standard deviation of 100.9.
- **Robbery: 1, 3, & 5 Yr Averages** - For robbery, the minimum 1 year count was 44 cases with a max of 384, a mean of 152.3, and a standard deviation of 87.2. For three year averages, the minimum number of cases was 57 with a maximum number of 404.

The mean was 163.5 with a standard deviation of 88.6. For five year averages, the minimum number of incidents was 64 with a maximum number of 417. The mean was 179.1 with a standard deviation of 90.5.

- **Burglary: 1, 3, & 5 Yr Averages** - For burglary, the minimum 1 year count was 114 incidents with a max of 358, a mean of 229.2, and a standard deviation of 82.2. For three year averages, the minimum number of cases was 133 with a maximum number of 419. The mean was 278.7 with a standard deviation of 110.3. For five year averages, the minimum number of incidents was 138 with a maximum number of 435. The mean was 288.8 with a standard deviation of 109.1.
- **Total 7 Major Felony Offenses: 1, 3, & 5 Yr Averages** - The total of all majory felony offenses reflects crime as a whole, not weighted for severity. The minimum 1 year count was 661 offenses with a max of 1948, a mean of 1304, and a standard deviation of 509.4. For three year averages, the minimum number of cases was 752 with a maximum number of 2019. The mean was 1398 with a standard deviation of 526.1. For five year averages, the minimum number of incidents was 818 with a maximum number of 2051. The mean was 1411 with a standard deviation of 502.4.
- **Ytd Enrollment 2011** - YTD enrollment reflects the % of of-age children who are enrolled in school in a community district. It ranges from 63.8% to 93.3%, with a mean of 89.0% and a standard deviation of 4.9%.
- **Ytd Attendance 2011** - YTD attendance reflects the number of children enrolled in school in a community district. It ranges from 10,200 to 15,370 with a mean of 35,870 and a standard deviation of 5,484.8. Both school datasets were from 2011.
- **NOV Count** - NOV count is the number of violations the Department of Housing in NYC submitted to a property. It was counted by using a simple REGEX package to parse out the parts of an address and query a 2GB CSV of all housing violations in 2016, summing the number of violations. The data ranged from 0 to 333 violations, with a mean of 2.4 and a standard deviation of 13.1.
- **Missingness**

We did not encounter any missing data for features sourced from open data. Data within this portion of the feature space was pulled using zipcode and property address, which the Python PIP package was able to process via REGEX. To some degree, this was self-implicated, as we picked relevant datasets that included information on Queens. Luckily most data put out by NYC only relates to property address at the most detailed level, which we had for every property in the sample.

For the remaining portion of the feature space, containing features extracted from the *MLSLI* listings, missing data points were fairly common. For these features, we initially cleaned values and dropped columns that were nonsensical or useless. Examples of the former was consolidating the levels of construction (”BRIKE” to ”Brick”, ”BRIKC” to ”Brick”, etc) or levels of cities (flushing to Flushing, etc). For values that had high proportions of missing data points, we adopted a multi-step process: *(1) First, we tested to see if there was any predictive power when running a Chi-Sq test vs Sold.Price. (2) If there was significant independence between the feature and sold price, we could assume that there was no predictive power and throw it out. (3) If there the variables weren't significant and plotting pairs or a scatterplot showed a useful picture, we then imputed the values.*

We did this via Python scripts that my teammate William Fry wrote, which tried to meet the convenience of imputation by average with the enhanced accuracy of regression. The scripts had basic logic that went as follows: *(1) if values are presents for other listings at same address, take the average and stop. (2) If values are present for other listings on the same street with the same zipcode, take the average and stop. (3) If values are present for other listings with the same ZIP-code, take the average and stop. (4) If none are found, take the average of the value for all listings where value is present.* We repeated this process for factor variables by using the mode instead of the average.

Although the trend of the missing data exhibited NMAR - namely, that rows with missing values had a lot of missing values - we treated is as MAR and imputed. We did not include any missingness dummy variables at the time, although we defaulted to a pattern mixture model where missingness would have an affect on response. Our take - albeit elementary - was that missingness often represented a decision to not share the information. It is in the realtor's interest to put as much positive information as possible, so lack thereof means the information (the level of the factor) is not ideal. We then decided whether to code as "Unspecified" in the case of Kitchen Type or convert all to missing to "No" in case of whether cats are allowed.

III. Modeling

We chose a non-parametric Random Forest prediction model. In view of the problem at hand, we thought it would be very unwise to make assumptions on the functional form of f, and given that our main goal was prediction accuracy, we want $f \approx \hat{f}$ to be as close as possible. A parametric model would not suffice for this problem, so we took the non-parametric route. In this sense, we gained predictive performance at the expense of interpretability.

During our model building process, we never assumed that the relationships between the predictors and our prediction output was linear. From previous experiences in Data Science competitions, a good adjustment we considered was predicting Sale Price per Square Feet. This adjustment tends to make a good price-performance indicator, and linearizes many relationships. Nevertheless, we lacked of information on the space of each listing, and considerations such as whether to include non-interior square-feet, garage space, and other factors made it difficult to consider this adjustment without facing significant uncertainty. In consequence, since we only used Sale Price as our prediction goal, we had to accept that there would likely be non-linear relationships and significant interactions between some of the predictors and Sales Price. Finally, considering that in this particular case, $p < n$, all of these factors made Random Forest the ideal algorithm to suit our needs.

After cleaning and imputing the data, we used the Random Forest Algorithm on a subset of the cleaned predictors. We did not use any explicit derived predictors or custom variables - we relied on the Random Forest algorithm to find interactions as well as curvilinear relationships. Nevertheless, we cannot rank predictors in terms of importance given our choice of algorithm. While we can say with some confidence that variables

such as "Approximate Interior Square Feet" likely have a causal effect on price, we are barred from making inferences of this type within the context of non-parametric modeling.

As we learned in class, the Random Forests Model does not overfit by design, since all the trees are taken together as a unit, given that we sample $p^* < p$ of the predictors and sample rows with replacement.

IV. Results

While we sacrificed model interpretability in favor of predictive accuracy and thus cannot make confidence intervals or prediction intervals, we can still analyze the R^2 as well as the RMSE's size in accordance with values. Starting with in-sample metrics, the R^2 was 98.19%. Although this is high in an absolute sense, it is not to be trusted as each tree is a "weak learner" and overfits on the observations used. The in-sample RMSE is $23,432.07, which is (of course) much better than what we find when looking at out-of-bag goodness of fit.

As for out-of-sample metrics, the model seems to have a pretty high $RMSE$. Although the prices of properties sold ranged from $55,000 to $999,999, a RMSE of $48,580.18 means there is a very wide margin of error. On the other hand, the OOS R^2 was 91.51%, a figure that we're quite proud of. This means that with our data extraction efforts combined with the process of Random Forest, we were able to explain 91.51$ of variance.

As the out-of-sample metrics are the only valid goodness-of-fit indicators at hand, the R^2 shows that the model will be able to explain a good amount of variance. The $RMSE$ is higher than we would prefer, but given the wide range of sold prices and the level of missing data amongst *MLSLI* properties, we believe that this model should be accurately predict real estate prices in Queens for the next couple of months. The previous qualification is important and one that will be addressed in following section, as prices are not stationary despite being assumed so for the purpose of the problem at hand.

V. Discussion

In conclusion, the underlying exercise in this project was to take a very fundamental and basic approach to statistics to build a model which would predict real estate prices for properties in Queens. Drawing upon the key concepts in model building, we began with a reflection upon what features seem relevant to us given the raw, naturally represented data that we've seen over our lifetime. We then used mapped out how we could be connect the disparate data sources on the internet to extract those features for a given observation on *MLSLI*. After collating the data, we encountered the issue of missingness, which required a level of work and effort that we previously did not appreciate. Once we cleaned and imputed the data, we plugged in the dataset into a Random Forest model using a subset of the already cleaned features. After submission of predictors using that model, we then improved it using additional features which were ready after the 5pm deadline on Sunday. The resulting two models both had high levels of OOS R^2 but also relatively high $RMSE$'s.

When reflecting upon what could have been improved, I think that in the featurization aspect, there were a lot of features that we had to neglect in the interest of time. Here I would like to add more demographic data, and more such as the average reviews and prices of restaurants within 500m from FourSquare as well as an index of social media usage (Instagram photos tagged around the block) over time. This would paint a fuller picture of the average income and project the direction of real estate prices as well (an increase of photos per week or prices per week would point to a

future increase in prices). Outside of those features, we could have done a lot more with the text on the pages of the listings. The presence of certain words in describing a property definitely deserves a closer look. The presence of photos as well as basic color analysis of them (there are a few open source python packages for this) could be beneficial - especially for sites that serve as platforms for connecting buyers to sellers.

Another idea that could be explored in future models is cotenant and retail data: there is precedent in this area that the presence of certain retailers, such as Trader Joe's and Whole Foods, has a positive impact on Real Estate prices.

Outside of additional features, we could have also benefited from a more thorough data cleaning process. As only a few of us were technical, much of the scripting / cleaning was rushed through and missingness variables were dropped in favor of forcing missing values into means, modes or pre-existing categories that "intuitively" made sense. We learned in class that this "intuition" is wrong (see PhD study). With high noise and many features, we should have adopted a more systematic way of doing this as well as researched the area more to understand patterns.

While the prediction set missed many of our features entirely and required pure imputation, I believe if we ran our prediction data process with the same hit forms as our initial historical process we would have a better RMSE. This is also more representative of how *Zillow* would predict prices (as they would have access to the underlying dataset). Luckily, many of the open data sources only require location, so leveraging a lot of the geo-spatial data was possible (and many sources were not even touched due to time). What I think we did well here was trying to think outside of the box for alternative but relevant data sources and leveraging programming / data extraction experience. This is something I have to give full credit to William, who pro-actively took the lead in this domain. This is an area where we could drastically improve in breadth and depth of features.

We suspect that our biggest potential area of improvement (and one that William and I plan to investigate after Spring Break) is how to break the stationary requirement of this project and allow time-series data to improve our model. Currently, the trained model works well when training on recent data and predicting values in the near future. This isn't ideal. We missed accounting for changes over time and the effect of velocity of features on listing prices. One really interesting source that I found is called Vigilante. This service transcribes 911 calls to text and open sources them. Understanding the topic, frequency, and geo-location of the calls would be a great improvement upon the crime statistics. Moreover tracking general pricing trends and providing secondary and tertiary forms of them (averages / velocity) over previous time spans would be tremendously interesting.

Finally, as mentioned in the modeling section, I would be very interesting in predicting Sales price adjusted by size. Having complete data on the size of each listing is essential for this step, and it could lead to the possibility of mixing non-parametric and parametric methods.

In conclusion, we're excited with what we were able to build over the course of a few days but we have necessary improvements to do on foundational aspects of the process we followed. We also have a number of exciting ideas for innovative features and data sources. We believe we're ready to compare ourselves against *Zillow* but that in order to be production-ready, a break out of a stationary model is required.

- **Acknowledgments**

First acknowledgment goes to my teammate William. His Computer Science expertise proved essential for

open-data extraction and for several stages of the data-cleaning process. We would like to also thank William's roommate, who is also a data engineer and incoming engineer to Kushner's Cadre fund, which is a platform for real estate investments and secondary sales driven by data analysis. He helped with some of the ideation around data sources. Additionally, we would like to acknowledge some MBAs who bankrolled the dataset extracted via MTurk. It was very nice to meet MBA's with similar interests who where very willing to collaborate with us with data collection. Finally, we would like to acknowledge Professor Kapelner. I will be refering