Ithaca Housing

Marc Davila and Emmanuel Lambrakis
(md934 and el668)
ORIE 4741

May 10, 2024

Abstract

The rapid evolution of real estate markets demands robust predictive models to understand and forecast housing prices. In this study, we employed two datasets from Zillow Home Value Index (ZHVI) and Kaggle to analyze the impact of market dynamics, including the influence of large real estate firms, on housing prices in Ithaca. We investigated the effectiveness of linear regression models with the quadratic loss and Huber loss functions. Despite challenges in data collection and quality, our study shows the importance of key factors affecting housing prices in Ithaca and highlights areas for future research and model improvement.

1 Introduction

1.1 Background

The purpose of this project is to understand how big real estate firms like BlackRock and other companies affect the housing market. As time progresses, we see that many of these firms take advantage of the market and begin to buy much of the real estate. For that reason, it is important to predict and understand the prices of today and tomorrow. Therefore, we are here to analyze the impact of these companies and possibly predict the next prices in the Ithaca area. The reason we chose a city like Ithaca is because it is a very small city, and most people would be looking to buy a house. We will use a dataset that focuses on the houses in Ithaca that reflect the typical value for homes in the 35th to 65th percentile range from 2021 to now. Additionally, we will use a dataset containing 2022 Ithaca house prices and data.

1.2 Data Description and Analysis

We modified the original dataset and used two datasets to help with our analysis. The first was the Zillow Home Value Index (ZHVI), a dataset that determines the value of a house in certain areas. In our case, we will be analyzing from 2021 to now. The reason is that before the pandemic, buying a house was a lot harder, so using data from that period may show some sort of bias and incorrect analysis. Moreover, during the pandemic, most houses were very cheap and nothing compared to what we are facing now. The second dataset will be one from Kaggle where someone was able to web scrape numerous houses from all of NYC. This dataset has multiple features, including property-url,

property-id, address, street-name, apartment, city, state, latitude, longitude, postcode, price, bedroom-number, bathroom-number, price-per-unit, living-space, land-space, land-space-unit, broker-id, property-type, property-status, year-built, total-num-units, listing-age, RunDate, agency-name, agent-name, agent-phone, and is-owned-by-zillow.

1.3 Data Cleaning

We decided to first clean up the first dataset that focuses on ZHVI. Since we are only looking for the city of Ithaca, we deleted all other rows and focused on the Ithaca row. Next, we also deleted all the other entries from 2001 until January 2021. We will use this data to determine the growth of real estate over the years and then increase all the values in the other dataset to have something to predict to. Additionally, we decided to look at the increasing value, and the output is shown in Figure 1:

Figure 1: Increase in Cost Post-Pandemic

Additionally, in the second dataset, we are given data on some houses in 2022 in NY. The first step was to clean the data to only include Ithaca houses. Then, we dropped any unnecessary columns like property-url and property-id. Next, for any columns that are missing values, we decided to put the mean of the column as the missing value/the most common category. Then, we also changed any houses that have their land measured in acres into square feet. Moreover, we got rid of any outliers by removing the top 75th percentile and the bottom 25th percentile. After that, we focused on the type of data we had. In our case, we knew we were going to make a feature that included the latitude and longitude of Cornell and Ithaca College. This will help us get the distance between the two colleges. However, this will be a feature more than anything else.

2 Methodology and Models

2.1 Quadratic Model

We decided to use a Linear Regression Model since it is very common when predicting real estate. In this case, we first used a quadratic loss without any regularization:

$$\text{minimize} \quad L(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

Then we tried the feature of longitude and latitude. The reasoning behind this is that people are more inclined to want to live closer to colleges. For that reason, it makes more

sense. This is a very abstract feature as this is tailored more toward the city of Ithaca only. However, most other cities in NY have universities/colleges around them. Based on whichever is better, we kept the best one. Then we added a regularizer. More specifically, an L1 regularizer, since our data is based on a human market. There could still be very interesting outliers that mess with our data a lot. For that reason, we will not be using the ridge regularizer since it may mess up our model due to its insensitivity:

$$\text{minimize} \quad L(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

2.2 Huber Model

Using the same intuition as in the previous section, we wanted to try to fit a linear model using a loss that does not penalize outliers. As we do not want to penalize our data too much, we will focus on using the Huber loss function. This function punishes less than the quadratic function when it comes to outliers, as the Huber loss function controls both the linear and quadratic parts. This is what the model looks like:

$$L(\beta) = \sum_{i=1}^{n} \phi_\delta(y_i - x_i^T \beta)$$

$$\phi_\delta(r) = \begin{cases} \frac{1}{2} r^2 & \text{for } |r| \leq \delta, \\ \delta(|r| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

Then, we will do the same steps as the Quadratic Loss model by only applying the L1 regularization. We emphasize this because simple data is better than very complicated data. As such, trying to combat outliers would be better with very few outliers:

$$\text{minimize} \quad L(\beta) = \sum_{i=1}^{n} \phi_\delta(y_i - x_i^T \beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Finally, for simplicity reasons, we only did an 80-20 split to ensure our data is good. While we should have done an 80-10-10 split for validation, we will explain why later.

3 Results

3.1 Results of Quadratic Model

Whenever we tried to run our data, we were really struggling and figured out very early on that finding data on Ithaca real estate would require web scraping. We will discuss this more later on. However, given the dataset that we used, out of the 28 categories/columns, we decided to use 4 of them. All you really need to see if a house is worth the value is based on how many bedrooms, baths, square feet, and how much land it has. These four are all we really need from our data. This had a mean error of 159,792.963179, which is a very high number, but given the dataset that we are working with high numbers, this makes sense. Next, we decided to look at what exactly has the most correlation. We noticed something very interesting: the number of rooms had a negative correlation to the prices of the house. This was very counterintuitive. As a result, we added parameters

that all connected back to the number of bedrooms. In other words, we added the correlation between the other three parameters to enhance the data's bedroom correlation with price. This caused our mean-squared-error to be 54,626.003741. Here is the graph of this model: Additionally, we included the correlation: Given these two models, we decided

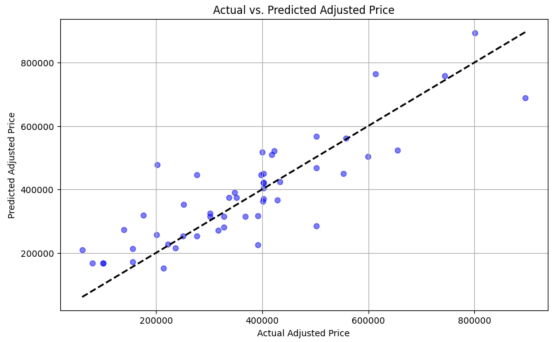

Figure 2: Model with 8 Parameters

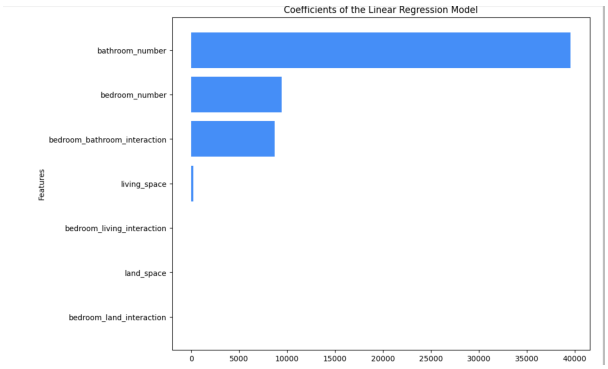

Figure 3: Correlation for Coefficients

to figure out using feature engineering if it mattered whether someone would live closer because of a college. As a result, we added the coordinates of Cornell University, and since our data gave us the coordinates of the houses, we used the Euclidean distance as a feature. It turns out that it did not matter, and in fact, it was worse for our model, giving us a mean-squared-error of 696,751.076354. Next, we decided to use an L1 regularizer, and these were the results for the lasso linear regression model: The mean-squared-error

Figure 4: Lasso Model

for this model was very interesting since, based on looking at the graph, this looks a

little more compact. We used an alpha of 0.01, but the mean-squared-error was higher by 60,537.664017. As a result, the best model to use is the Quadratic Model with regularization.

3.2 Results of Huber Model

We immediately noticed that whenever we finished the Huber model using an epsilon of 1.5, the model was really good at predicting around the median but horrible otherwise. As a result, this was once again worse with a mean-squared-error of 108,348.053839. Look at the data: It is very interesting that the model is predicting 0; it may mean that

Figure 5: Huber Model

our data really does not correlate well with Huber. Regardless, we will still try it with the L1 regularizer. As expected, this increased to 109,562.389294, and the model looks basically identical to Figure 5. Therefore, we can strongly assume that our dataset does not prefer Huber whatsoever.

4 Discussion

4.1 Limitations

1. **Data Collection and Web Scraping Challenges:**
    - Web scraping the Kaggle dataset and other real estate sources presented difficulties due to frequent changes in website structures, anti-scraping measures, and limited availability of comprehensive data.
    - As a result, our dataset was incomplete and required extensive cleaning, which may have impacted model accuracy and reliability.
    - Additionally, certain features like crime rates and neighborhood quality, which could significantly influence housing prices, were not included due to data unavailability.
2. **Data Quality and Outliers:**
    - The Kaggle dataset contained missing data and outliers that reduced overall quality.

- Missing values were imputed using the mean or most common category, potentially introducing bias.
- Outliers were removed using interquartile range (IQR) filtering, which may have led to the exclusion of legitimate yet extreme values.

3. **Model Simplification and Validation:**
   - We used an 80-20 train-test split rather than cross-validation or a dedicated validation set, which may have affected model robustness and generalizability.
   - Regularization helped improve model performance, but further feature selection and engineering were needed.
4. **Feature Selection and Dimensionality Reduction:**
   - While we utilized basic feature engineering (e.g., proximity to educational institutions), more advanced techniques like Principal Component Analysis (PCA) could have been employed to reduce noise and highlight the most impactful features.
   - Some relevant features (e.g., neighborhood crime rates, transportation access) were missing, which may have affected predictive accuracy.

4.2 Improvements and Future Work

1. **Advanced Feature Engineering and Selection:**
   - Incorporate additional features like neighborhood crime rates, school quality, and transportation access.
   - Apply PCA or other dimensionality reduction techniques to improve feature selection and reduce dimensionality.
2. **Improved Data Collection Techniques:**
   - Utilize advanced web scraping tools like Selenium to gather more comprehensive data while mitigating anti-scraping measures.
   - Explore open datasets from government sources, real estate agencies, or academic institutions to supplement existing data.
3. **Advanced Modeling Approaches:**
   - Experiment with more sophisticated models such as Gradient Boosting Machines (GBM), Random Forest, or Neural Networks.
   - Use cross-validation and hyperparameter tuning to optimize model performance.
4. **Temporal Analysis:**
   - Conduct a time series analysis to forecast housing price trends using ARIMA, Prophet, or other forecasting models.
5. **Comparative Regional Analysis:**
   - Expand the analysis to other cities to understand regional differences and develop more generalized predictive models.

5 Conclusion

In conclusion, although our data was not adequate for the task, we were able to make the following assumptions about choosing a fair dataset. We did not have a great model because our dataset was not as good as we needed it to be. This shows us how important fairness is because, in this example, our lack of data gave our model very high values, causing it to be very volatile. Additionally, we saw that even though we expected Huber to be better, the bias caused it to be worse, which really emphasizes how fair your dataset should be to avoid bias. We should not use this model in our company, but one thing is clear based on Figure 1: the houses are increasing at an alarming rate every month. Soon enough, a house that costs 300 thousand today will be double that. I would like to say that this issue is one that we need to acknowledge and deal with as soon as possible before it is too late and we are forced to buy houses from corrupt people.

6 References

Kaggle dataset: https://www.kaggle.com/datasets/polartech/new-york-state-real-estate-dataset

Zillow dataset: https://www.zillow.com/research/data/

Inspiration and guide on how we constructed this paper: https://github.com/woshibobo/ORIE-4741-Project/blob/master/final_report.pdf

Link to GitHub: Ithaca Housing