Visualizing New York: Real Estate Prices vs. School Quality

Emily Christian, Tom Deloi, Danny Park, Amirtha Pugazhendhi, Andy Sun, Will Trawick

1 Introduction

For the current prospective homebuyer, information about real estate prices and school performance values is fragmented and difficult to evaluate simultaneously. Local school ratings and reviews are commonly discussed among individuals in city-data forums and other online communities while separate real-estate websites such as zillow.com offer information on real-estate prices. For many parents who are prospective homebuyers, these two sources of information together are very critical in determining which property they should buy for their family.

1.1 Problem Definition

The goal of this project is to use school ranking and real estate data to easily visualize which neighborhoods in New York City offer the best value for school systems in comparison to its real estate prices. We created two models to determine the school rankings and real estate prices dependent upon specific features. These models will inform our interactive choropleth map of New York City to display key performance indicators for both valuable real estate and high-quality schools so that parents and real-estate agents can quickly and easily determine in which geographic areas they should invest.

1.2 Survey

Choropleth maps have been shown as an effective way to communicate geographically distributed data [1]. We will also consider using a cumulative frequency legend as a way to enhance our choropleth visualization [2]. We plan to use a time series (ARIMA) approach for real-estate valuation [3] and take conservative approach to rate schools by using well established indicators that have been shown to correlate with quality [4]. School rating can be nebulous. Harvey and Green identified five approaches to considering school quality: high achievement by its students, the absence of mistakes, preparation for a purpose, cost effectiveness, and personal development [5]. There have been a considerable number of studies done on the relationship between school performance and real estate prices. In general, there does appear to be a premium on house price for good schools in a study in Orange County, California [6]. There are a lot of factors besides school quality that influence house price, however, studies in Chicago, North Carolina, and California found that after controlling for other factors, school quality did appear to have a positive impact on house price [7,8,9]. Another study by Bogarts found that house prices in Shaker Heights, Ohio dropped by 10% after school districts were disrupted [10]. On the other hand, a study of Connecticut schools found mixed results when comparing test scores and house prices [11]. A study by Seo and Simons found that school district ratings and performance index were the most appropriate metrics of school quality for housing price prediction [12]. While a relationship exists, it is not directly causal so there still exists room to find value housing in terms of school quality. The strength of that correlation can differ based on geographic region and experimental design.

There are similar services to our project, but none show both real estate and school quality in a single, easy to use interface. For example, greatschools.com creates a map of high

schools in a selected area with rankings determined by graduation rates, standardized test scores, college enrollment, etc [13]. However, it is hard to find any real estate information without conducting an additional web search. Our method provides a convenient way to overview the school quality and the price of the neighborhood together.

Our project would be useful to both parents and real estate agents. Studies have found that parents cite academics as a significant factor when choosing a home [14, 15]. As a result, another study found that 27% of Florida public school students attended a school other than the one they were assigned to [16]. This desire is justified in a study by Card and Krueger which found that quality of school provided a significant economic benefit for students [17]. Real estate agents could use this map to provide better visualizations to their customers which would most likely result in increased customer satisfaction as has previously occurred with other visual analytic real-estate systems [18].

2 Proposed method

For this project, we decided to work with two different datasets. One focusing on representing the quality of the school and other focusing on the price of the real estate. At present, these datasets are generally analyzed and visualized separately. Hence, we have taken up the challenge to innovatively and meaningfully integrate those approaches to bring better understanding and visualization dataset. The first step we took was to prepare the necessary dataset to cater the needs of the model.

2.1 Assessment of school quality

The New York City High School dataset was obtained from the New York Open data department [22]. The dataset is composed of 427 schools with 473 attributes such as address, number of students, etc. Because the goal of this project was to integrate the high school quality with the housing price, it was important to rate each school based on their attributes. To make the process more efficient, we decided to use a 10-point scale rating from Greatschools.org. After combining the high school dataset with quality score from Greatschools.org, we found 54 out of 427 schools were not rated due to insufficient data from the website [19]. As a result, we used a classification method, Random Forest, to predict the rest of them.

Before we run any machine learning algorithm, we preprocessed the dataset to make it more usable. First, we reduced the number of columns by removing all the unrelated variables such as the school website and number of buses. Then we removed all the unexpected characters, filled the empty values with mean and mode and transformed some variables into more interpretable ones. For example, we combined school start and end time to create a school duration column. Similarly, the AP course column, a collection of advanced courses, changed into a column that shows how many AP courses each school offers. To consolidate the dataset further, we screened out more columns that are less influential than the others using a feature selection technique called Chi-squared feature selection. As shown in the figure below, we selected 15 features out of 25 based on their correlation value with the response variable.

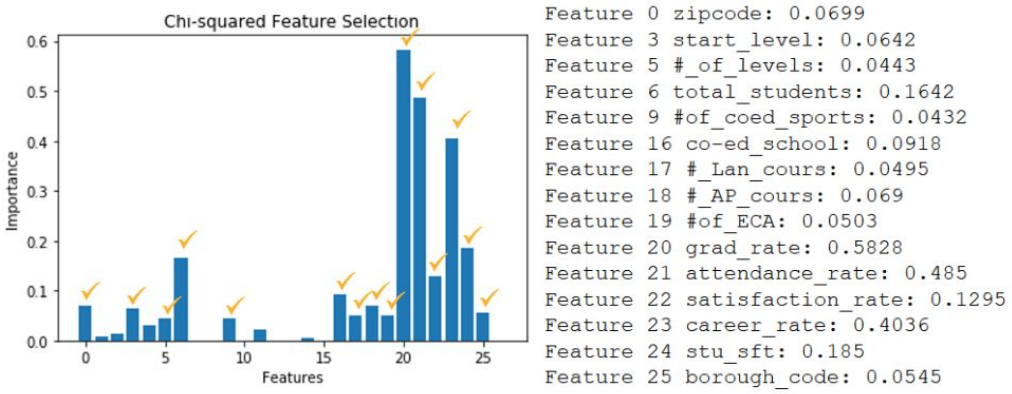


Figure 1: Chi-Squared Feature Selection

To build a model prediction, we used a Random forest method which is one of the most popular machine learning techniques. It consists of multiple decision trees and classifies the data based on a majority vote. For tuning parameters, we used a Random Hyperparameter grid to find the optimal solution for the model. With the optimal parameters, we trained the model, and tested. The performance of the model was evaluated by creating a confusion matrix and more details are illustrated in the evaluation section later. Achieving over 0.8 accuracy, the model predicted unlabeled school qualities which were then re-combined with the previously labeled school data and saved as a CSV file which would be integrated with housing price data for the visualization part of this project.

2.2 Real Estate Valuation

In order to provide value to consumers of our product we decided to forecast real estate prices at a town level. Instead of simply providing consumers with an average real estate price per town, we forecast 1-5 year appreciations/depreciations. Our team utilized data from [20] to create our forecasts. Real Estate valuation at a grouped level (town) lends itself to time series analysis. In order to properly take into account the multiple factors that predict

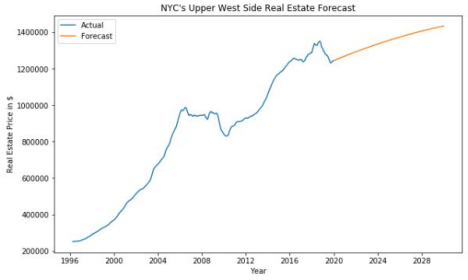


Figure 2: Real estate forecast for Upper West Side neighborhood

price at any moment a Seasonal Auto Regressive Integrated Moving Average (SARIMA) model was used. SARIMA models have 7 hyperparameters that need tuning in order to find the best model. Multiple iterations of data cleaning/feature engineering and hyperparameter tuning were done. Ultimately our final model utilized monthly data from 1996 - 2020 for our input and a robust stepwise grid search was used to determine the optimal hyperparameters of our model, with the optimal hyperparameters resulting from the lowest Akaike Information Criterion Score.

2.3 Visualization

We use an interactive choropleth map to display the results of our school quality assessment and real estate valuation. It displays data broken down by neighborhood [21] and school district[22]. In order to align the neighborhood border data with the real estate data we had to manually clean the neighborhood names. For example, the neighborhood Chelsea-Travis was originally called Chelsea, Staten Island in the border data. We then created several interactive customization features to the visualization which allow users to filter the data by price range and set the weight of school performance in the overall neighborhood score. Results are color-coordinated from red to green and a list of the top five neighborhood

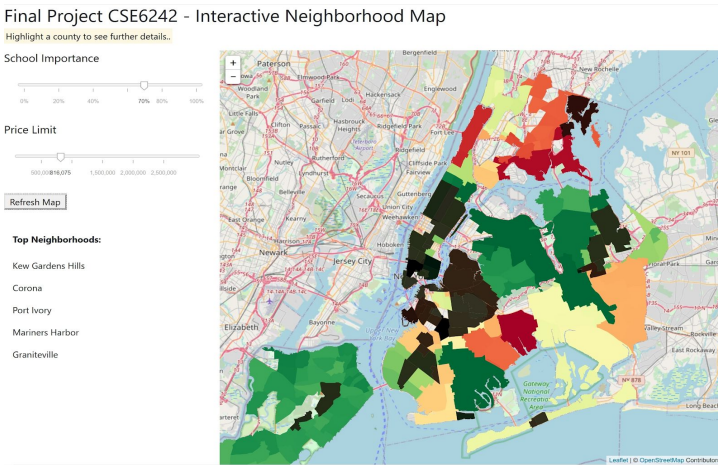
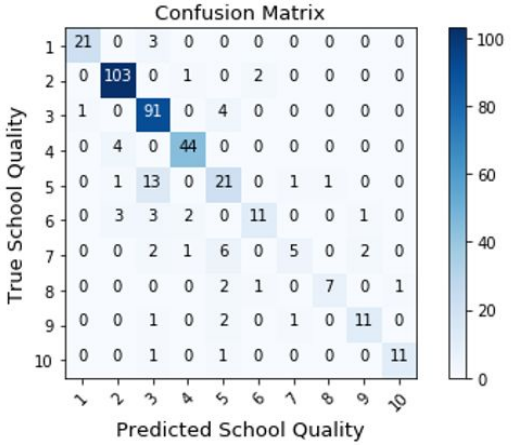


Figure 3: Interactive Choropleth Map

matches are displayed to the left of the map. The price limit filter will discolor some regions of the map according to the extent by which the median real estate price in that area exceeds their budget. Very unaffordable regions will be dark black and largely unaffordable regions will be shades of brown and burgundy.

3. Experiments/ Evaluations

As mentioned previously, the performance of the prediction model for the school quality is illustrated with the confusion matrix as shown in the figure below. To make the outcome more robust, we used 5-fold cross validation.



Result	
Train Accuracy	0.9831
Testing Accuracy	0.8423
Macro Precision	0.8258
Macro Recall	0.7357

Figure 4: Confusion Matrix and Cross Validation Result Table

As shown in the table next to the confusion matrix, we reached the testing accuracy over 0.8. Also, macro precision and recall came out to be 0.8258 and 0.7357 respectively, indicating that the most of the predicted school qualities are accurately classified.

4 Innovations

4.1 Value of the School

Currently, the aspect of Quality of schools and Price of the Real estate were dealt separately till now. In order to improve the decision making easier for stakeholders like parents and real estate owners, we combined the aspects of Quality of Schools and Price of Real Estate by introducing the compound dimension called Value of the Schools. The concept of value is directly related to the concept of consumer Surplus (ie) that how much benefit does a consumer get for the price paid for it. Value can be calculated by using the taking ratio between an overall standardized school quality rating and Standardized real estate price rating.

$$Value\ of\ the\ School\ =\ \frac{Overall\ Standardised\ Quality\ rating}{Overall\ standardized\ real\ estate\ price}$$

From here we give users the option to calibrate the weight of school quality in the overall neighborhood score metric. If the baseline formula is .5(school) + .5(real estate) the slider bar permits users to change those factors from a 0 to 1 scale and recalibrates the color-scales based on the weights of those new scores.

The high valued schools will have the characteristic of High quality and Low Price of Real Estate and conversely low valued schools will have the characteristic of Low quality and Low Price of Real Estate. There can also be a combination of High Quality and High Price, which is labeled as Luxury Schools and There can also be a combination of Low Quality and Low Price, which is labeled as Budget Schools. This helps the parent to take decisions efficiently and effectively. Finally, We have planned to demarcate the Zone to the Value of the school they possess.

4.2 Combined borders

The boundaries of school districts usually do not line up nicely with the boundaries of neighborhoods. In order to interact with these subregions in D3, we needed to create a new dataset with these new subregion boundaries. We used the shapely library in python to automate most of the process, however a little bit of manual intervention was still needed.

5 Conclusion and discussion

Our project lies in a unique space relative to tools and resources that are readily available. We aim to increase access to information for involved parents and homebuyers who may otherwise be left in the dark. So far, we have broken down the task into the three action areas of (1) Real Estate Valuation (2) School Quality Evaluation and (3) Visualization Tools. The initial challenges involved choosing appropriate metrics to use in each of these categories, customizing what has been done to fit the needs of our end users, and keeping in mind the flexibility required to combine each component with the others. From here we move onto the stage of integrating the three into a single tool that is both deeply meaningful

and immediately intuitive to use. We are excited about everything we're doing and look forward to the chance to demonstrate the final product.

6 Work Distribution

All team members have contributed a similar amount of effort.

References

[1] Norman, Kent, et al. "Dynamic query choropleth maps for information seeking and decision making." *Proc. Human-Computer Interaction International*. 2003.

[2] Cromley, R.G., Cromley, E.K. Choropleth map legend design for visualizing community health disparities. *Int J Health Geogr* 8, 52 (2009). https://doi.org/10.1186/1476-072X-8-52

[3] Tse, R. (1997), "An application of the ARIMA model to real-estate prices in Hong Kong", *Journal of Property Finance*, Vol. 8 No. 2, pp. 152-163. https://doi.org/10.1108/09588689710167843

[4] Mayer, Daniel P. *Monitoring school quality: An indicators report*. Diane Publishing, 2000

[5] Harvey, Lee, and Diana Green. "Defining quality." Assessment & evaluation in higher education 18.1 (1993): 9-34.

[6] S. Y. He, "A hierarchical estimation of school quality capitalization in house prices in Orange County, California," *Urban Studies,* 2017.

[7] T. A. D. e. al, "The impact of school characteristics on house prices: Chicago 1987–1991," *Journal of Urban Economics,* 2002.

[8] Clark, David E., and William E. Herrin. "The impact of public school attributes on home sale prices in California." *Growth and change* 31.3 (2000): 385-407.

[9] Kane, Thomas J., Stephanie K. Riegg, and Douglas O. Staiger. "School quality, neighborhoods, and housing prices." *American law and economics review* 8.2 (2006): 183-212.

[10] Bogart, W. T., & Cromwell, B. A. (2000). How much is a neighborhood school worth?. *Journal of Urban Economics*, *47*(2), 280-305.

[11] Clapp, J. M., Nanda, A., & Ross, S. L. (2008). Which school attributes matter? The influence of school district performance and demographic composition on property values. *Journal of Urban Economics*, *63*(2), 451-466.

[12] Seo, Y. and Simons, R. . (2009). *Summary of Previous Literature*, The Effect of School Quality on Residential Sales Price, journal of real estate research vol 31

[13] Lautz, Jessica, et al. "Home Buyer And Seller Generational Trends Report 2017." National Association Of Realtors Research Department, 2017

[14] P. J. Wolf, "Introduction to the Special Issue—School choice: Separating fact from fiction," *Journal of School Choice,* 2017.

[15] S. A. e. al, "School Choice Decision Making Among Suburban, High-Income Parents," *AERA Open,* 2016.

[16] L. M. P. e. al, "Parental preferences in the choice for a specialty school," *Journal of School Choice,* 2018.

[17] Card, David, and Alan B. Krueger. "Does school quality matter? Returns to education and the characteristics of public schools in the United States." Journal of political Economy100.1 (1992): 1-40.

[18] Li, Mingzhao, et al. “Visualization-Aided Exploration of the Real Estate Data.” Lecture Notes in Computer Science, Springer International Publishing, 2016, pp. 435–39. Crossref, doi:10.1007/978-3-319-46922-5_34.
[19] Greatschools, School Quality Rating,Oakland, CA, United States : Greatschools,2020.Accessed on Mar. 30, 2020 [Online]. Available:https://www.greatschools.org/new-york/
[20] Zillow, *Zillow Home Value Index (ZHVI)*, Seattle, USA: Zillow, 2020. Accessed on Mar. 30, 2020. [Online]. Available:https://www.zillow.com/research/data/
[21] datHere*,NYC Neighborhoods*, New York City, USA: BetaNYC, 2014. Accessed on: Mar. 30, 2020. [Online]. Available:https://data.beta.nyc/dataset/pediacities-nyc-neighborhoods
[22] datHere*, NYC School Districts*, New York City, USA: BetaNYC, 2014. Accessed on: Mar. 30, 2020. [Online]. Available:https://data.beta.nyc/dataset/nyc-school-districts