Reveal Estate

Real Estate Model Applied to Proposed Queens Light Rail

Nora Barry
Center for Data Science
New York University
New York, USA
neb330@nyu.edu

Laura Buchanan
Center for Data Science
New York University
New York, USA
lcb402@nyu.edu

Jacqueline Gutman
Center for Data Science
New York University
New York, USA
jg3862@nyu.edu

***Abstract*—The real estate market in highly populated cities such as New York is ever-changing, largely due to architectural developments, changes in the environment, and trends in human preference. Models that can predict these changes, and give users insight into the effects that certain developments may have on real estate value, are extremely valuable. In this project, we use publicly available real estate and urban data for New York City to predict the increase in property value for buildings surrounding the proposed Queens Light Rail [6]. We consider multiple regression models, but get the best performance with a random forest, which predicts within 10% of the true price per square foot 36.1% of the time. With this model, we are able to predict, on average, a 143% increase in real estate value for properties within a half mile of the proposed Queens Light Rail. This would amount to an extra $80 million in tax revenue for the New York City area.**

Keywords—real estate; PLUTO; NYC Open Data; machine learning; random forest, BQX

I. Introduction

Real estate in populous cities are a major investment, on the microeconomic level of personal residence or on the macroeconomic level as financial vehicles. The capacity to predict value of real estate would be immensely valuable to a broad range of individuals and organizations. However, predicting future real estate values in a city like New York is very challenging. No one feature of a property strongly correlates with how valuable the property will be per square foot.

We built a model of NYC real estate prices based on building features and neighborhood characteristics, with the intention of evaluating financial opportunities for the City of New York. As an example, we show the predicted financial gain that can be expected from the proposed Queens Light Rail, both in real estate value increase in that neighborhood, and the increased tax revenue for NYC.

II. Data Description

A. *Department of Finance*

The New York City Department of Finance Real Estate Detailed Annual Sales Reports by Borough [1] were gathered for the years 2003 to 2015 for all five New York City boroughs. For 2016, the rolling sales data contains everything through the end of August 2016 [7]. These data contained sale date and sale price for all properties sold in all tax classes over the given time period. The Department of Finance data also contain information such as neighborhood, building class, tax class at time of sale, and per-unit square footage, among other features. The features of interest collected from the Department of Finance data were sales price and sales date, with their corresponding borough, tax block, and unit lot number, the unique property identifiers.

B. *PLUTO*

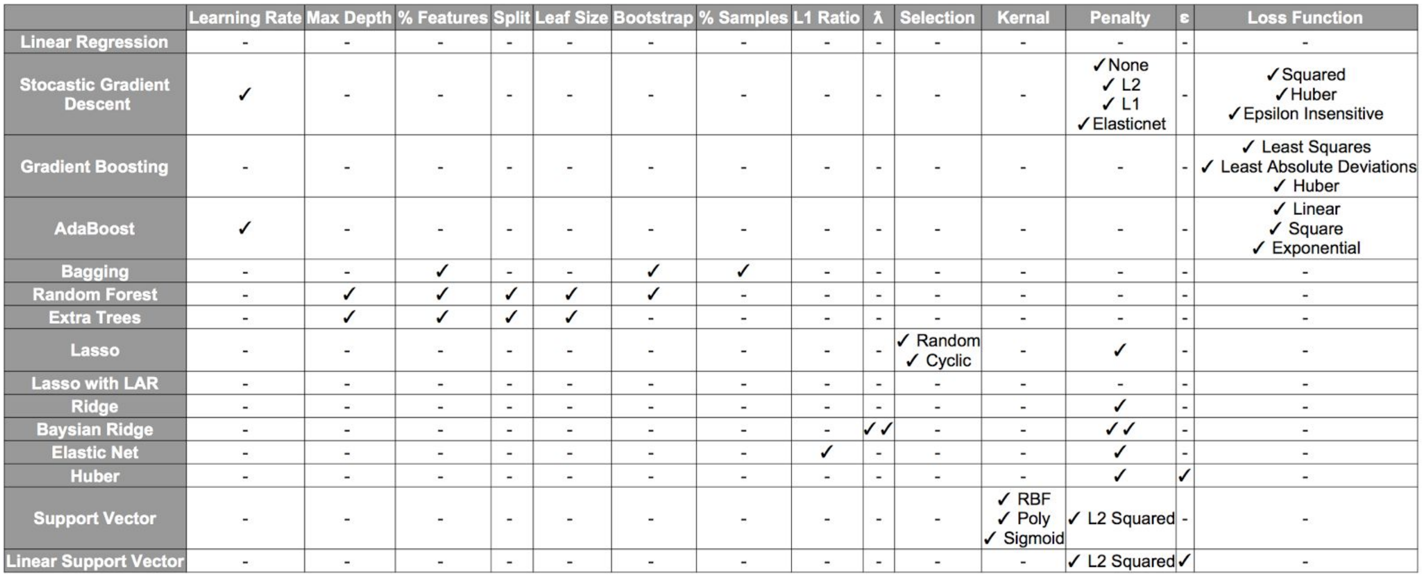
PLUTO data is open dataset provided by the NYC Department of City Planning [5] that contains building characteristics, land use, and geographic data for every real estate property in New York City, with more than seventy features. Archived versions of the PLUTO data are available and the data contains up-to-date building information for every borough as of March 2016. Numerous features from this dataset were used in our model including unique property identifiers, property and building size, and building type.

C. *NYC Open Data*

All datasets made available by NYC Open Data [4] were inspected to determine if they included relevant spatial data that could be used as features in our real estate model. The data is collected from various NYC agencies and organizations for the purpose to “improve accessibility, transparency, and accountability of City government.” Website links to all NYC Open Data datasets used in project available in the Appendix.

III. Data Preparation

The features in the PLUTO dataset were carefully re-engineered into a useful format for our model. A number of features, such as whether a property had a basement or was located in a historical district, were binarized. The values of other features, such as building type, were first collapsed into



	Learning Rate	Max Depth	% Features	Split	Leaf Size	Bootstrap	% Samples	L1 Ratio	λ	Selection	Kernal	Penalty	ε	Loss Function
Linear Regression	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Stocastic Gradient Descent	✓	-	-	-	-	-	-	-	-	-	-	✓None ✓ L2 ✓ L1 ✓Elasticnet	-	✓Squared ✓Huber ✓Epsilon Insensitive
Gradient Boosting	-	-	-	-	-	-	-	-	-	-	-	-	-	✓ Least Squares ✓ Least Absolute Deviations ✓ Huber
AdaBoost	✓	-	-	-	-	-	-	-	-	-	-	-	-	✓ Linear ✓ Square ✓ Exponential
Bagging	-	-	✓	-	-	✓	✓	-	-	-	-	-	-	-
Random Forest	-	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-
Extra Trees	-	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-
Lasso	-	-	-	-	-	-	-	-	-	✓ Random ✓ Cyclic	-	✓	-	-
Lasso with LAR	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ridge	-	-	-	-	-	-	-	-	-	-	-	✓	-	-
Baysian Ridge	-	-	-	-	-	-	-	-	✓✓	-	-	✓✓	-	-
Elastic Net	-	-	-	-	-	-	-	✓	-	-	-	✓	-	-
Huber	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-
Support Vector	-	-	-	-	-	-	-	-	-	-	✓ RBF ✓ Poly ✓ Sigmoid	✓ L2 Squared	-	-
Linear Support Vector	-	-	-	-	-	-	-	-	-	-	-	✓ L2 Squared	✓	-

Figure 1: Gridsearch parameters

broader categories. Then, the broad categories were used to create dummy variables.

The re-engineered PLUTO dataset was then merged onto the Department of Finance Sales data on BBL, the unique property identifier created from concatenating the borough code, block number, and tax lot number. BBL does not appear in the Department of Finance data, so this feature was engineered from borough, block number, and unit lot. For sales of condominiums, where the unit lot number provided in the Finance data did not concord with the tax lot, Digital Tax Map [3] data for all five boroughs provided an alternative intermediary mapping between these two data sets. This was done by connecting the borough code and condominium number, provided in the PLUTO data, to the borough code, block number, and unit lot number provided in the Finance data.

All datasets available at NYC Open Data were inspected to determine if they contained desirable spatial data. These datasets were downloaded and the relevant features were individually extracted. In many cases, the feature of interest was the distance from a property to a community feature, or the count of a particular feature in an area. The latitude, longitude, borough, and zip code of the properties in the Department of Finance Sales data were used to compute these distance and count metrics. Once computed, these measures were merged onto the Department of Finance and PLUTO merged data.

IV. Data Analysis

After data was curated, cleaned, and merged, we performed a grid search for 15 different model types to find the optimal hyperparameters that provided the best predictive performance for each model type. Optimal performance was defined as minimizing the median absolute error in predicted price per square foot, averaged across five cross-validated held out folds to form the validation set. The validation set performance was then compared across the 15 best performing models selected by the grid search to select the best performing model across all model types. Because we assumed our features could have interaction effects, it was important to include non-linear models in our grid search. Median absolute error was selected as the loss metric because mean squared error produced highly biased and skewed estimates that were overly sensitive to properties with extremely high values for price per square foot, with the regression errors on these properties carrying disproportionate weight in less robust loss metrics such as mean squared error in comparison to median absolute error.

For hyperparameter grids containing over 100 unique model specifications per grid, the grid search randomly selected a maximum of 100 model specifications to validate. Due to the computational load of running up to 100 model specifications for 5 folds each for up to 15 model types for a large dataset, we used NYU's High Performance Clustering (HPC) services to perform the model selection and evaluation in parallel. The grid search hyperparameters for each model type are summarized in Figure 1.

V. Results

We found random forest to be the best performing model for our data. We further tuned the hyperparameters for the minimum samples per leaf, maximum depth of each estimator, the minimum samples required to split a leaf node, the maximum number of features to consider splitting on, and whether bootstrap samples were used to increase model performance. We optimized median absolute error for both

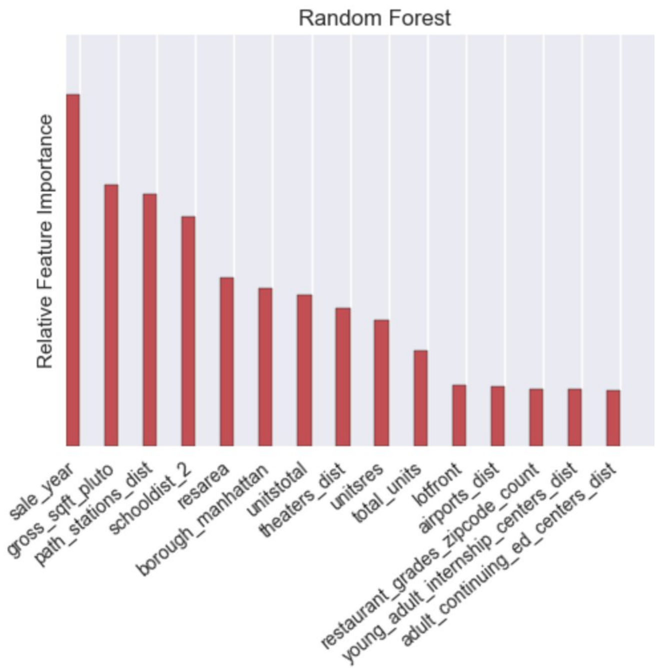


Figure 2: Relative Feature Importance

grid search within each class of models and the model selection between different classes of models. In the end, random forest minimized this loss metric, attaining a final performance of 35.7 on the validation set and 35.6 on the test set. We also reported the accuracy within 10% and 15% of the true price per square-foot, which was 36.1% and 50.4%, respectively for our final model.

We evaluated the features on which the random forest split most often. The top five most "important" features were sale year, gross square feet, distance to the nearest path station, the amount of residential space in a particular building and a dummy variable that represents a particular school district in Manhattan. The fact that our model discovered sale year to be the most important feature was reassuring, as we all know real estate value generally increases over time. In addition, we know that residential property is generally more expensive, so it makes sense that the model used residential area in its prediction of price per square foot. Finally, the school district feature from PLUTO data (schooldist_2) is likely important to the model because it's the school district that covers most of Manhattan, and Manhattan is the most expensive borough. The full feature importance list is shown in Figure 2.

VI. Application

To demonstrate the practicality of our model, we applied it to the proposed Queens Light Rail, a "state-of-the-art streetcar system" part of the Brooklyn Queens Connector [2]. The full connector would run a 16-mile route between Astoria and Sunset Park and "connect up to 10 ferry landings, 30 bus routes, 15 subway lines, 116 Citi Bike stations, and 6 LIRR Lines."

To estimate the effect of this Queens potion of the proposed streetcar, we utilized our model to predict the increase of value on the BBL's within a half mile of the Queens Light Rail. Since there are currently no streetcar lines in NYC, we used 'Distance to Subway Entrance' as our proxy for 'Distance to Streetcar.' BBL's within a half mile with greater than 0.5 miles to the nearest subway entrance were replaced with 0.5 miles. We then predicted the price per square foot for these BBL's with our model. A visualization of the increase in price per square foot for these affected properties is shown in Figures 3 and 4.

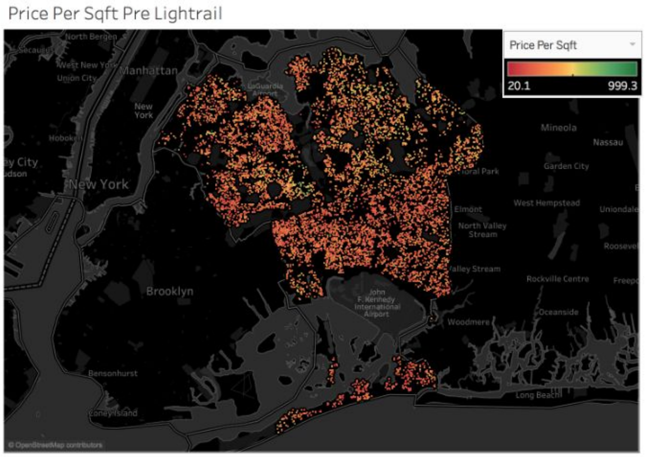


Figure 3: Predicted Queens Real Estate Value

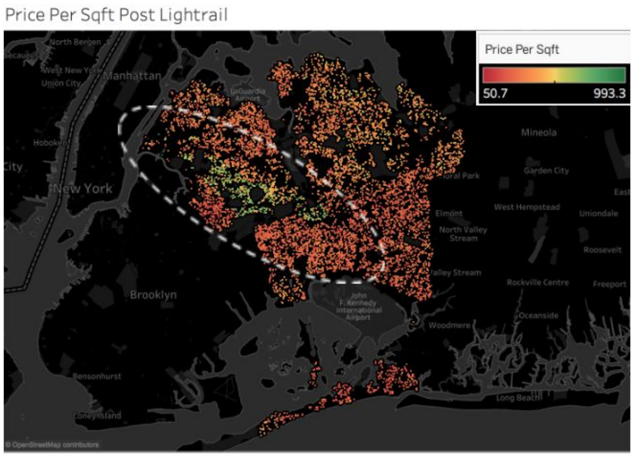


Figure 4: Predicted Queens Real Estate Value with Proposed Queens Light Rail

We found that the average increase in price per square foot for the properties within a half mile of the proposed Queens Light rail was 143%, with a median increase of 129%. This increase in property value would amount to an additional $80 million in property taxes for the City of New York, over a period of sales equivalent to the sales rate observed in these properties between 2003-2016. Furthermore, these predicted increases in property sales value were not uniformly distributed. The largest expected relative increases were estimated for the properties with the lowest and most affordable current sales value, while the expected increases were more modest and predicted prices more stable in properties that had already attained a relatively high sale price.

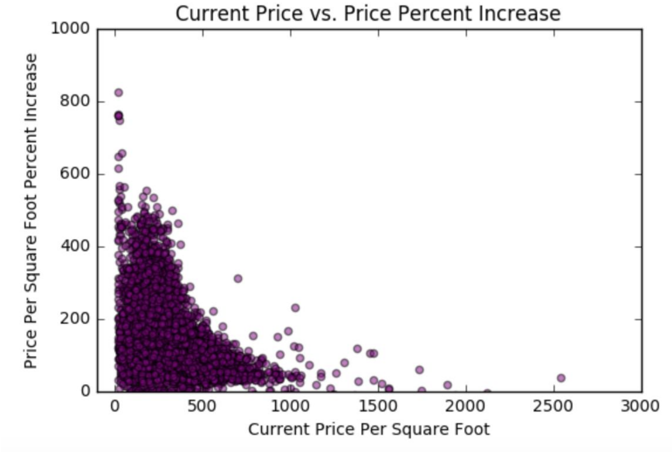


Figure 5: Distribution of Predicted Price Increase

VII. Discussion

A. Challenges with Dataset

As with any extensive data analysis, we ran into a number of setbacks with our chosen datasets. For instance, the Department of Finance data did not have complete records of price per square foot, so we had to compute it ourselves given the gross square feet column in PLUTO data. During this process, we found that many properties had sales prices of $0; in other words, the sale was actually an inheritance. We excluded these “sales” from our model. Unfortunately, this resulted in a great reduction of the number of sales transactions that we had hoped our model would be able to train on.

Another issue we faced was the variance in values for price per square foot. Many times, either the gross square feet reading or sales price reading was extremely off, leading to an unreliable price per square foot calculation. Asking our model to learn from these outliers resulted in huge mean square errors and poor performance overall. After finding that the average commercial real estate price for Manhattan is about $75 [http://spacesny.com] and the average residential real estate for Manhattan is about $1,450 [8], we decided to drop any instances in which the price per square foot was below $20 or above $5,000.

NYC Open Data does not generally have a standard format. Therefore, each dataset had to be curated manually and individually cleaned and merged with the DoF/PLUTO dataset.

B. Limitations

In our Queens Light Rail application, we had to make some simplifying assumptions. Because there are no streetcars currently in NYC, we had to use distance to subway entrances as a proxy for distance to streetcar entrance. Furthermore, we had the BBL’s within a half mile of the proposed streetcar. First, this is problematic in that compared to distance to subway entrances, a property might be within a half mile of the streetcar line but further from a streetcar line entrance. However, each stop on the proposed streetcar line is only a half mile apart. This means that generally, a property would not be further than 0.7 miles from a streetcar entrance. We felt this potential 0.2 mile discrepancy would not be overwhelmingly influential in our model.

Secondly, we had the properties that would fall within a half mile of the proposed streetcar, but not the explicit individual distances. We did not estimate distances possibility shorter than a half mile. Rather, we left the estimate at the higher bound of 0.5 miles. While this gives less specificity to the properties that fall within the half mile of the proposed light rail, we felt that this kept a reasonable, conservative control on our prediction, and would prevent us from overvaluing the effect of the streetcar.

C. Future Directions

A number of datasets from NYC Open data were left out due to complexity of the source dataset format. It would be possible to incorporate more neighborhood features into this model.

VIII. Conclusion

This project included extensive data preparation of the NYC Department of Finance Building Sales data and spatial NYC Open Data. With this data, we were able to built a model to predict ‘Price per Square Foot’ for properties in NYC, with the intention of predicting the impact of future architectural developments such as the Queens Light Rail. After performing model selection and hyperparameter tuning, our final model predicted within 10% of the true price per square foot 36.1% of the time, and within 15% of the true price per square foot 50.4% of the time. We then employed our model to predict property value increase given the proposed Queens Light Rail, a statistic that could be particularly useful in weighing the costs and benefits of a new subway line. In the end, we found that this development could potentially raise $80 million additional income for the City of New York through increased tax revenue over a period of 10 years of property sales, increasing the sale price per square foot of each affected property within the light rail’s vicinity by 143% on average.

[1] "Annualized Sales Update." *Annualized Sales Update*. N.p., n.d. Web.
[2] "Brooklyn Queens Connector (BQX)." *Brooklyn Queens Connector (BQX) | NYCEDC*. N.p., n.d. Web.
[3] "New York City Dept. of Finance Digital Tax Map." *NYC.gov*. N.p., n.d. Web.
[4] "NYC Open Data." *NYC Open Data*. N.p., n.d. Web.
[5] "PLUTO and MapPLUTO." *PLUTO and MapPLUTO*. N.p., n.d. Web.
[6] Rivoli, Dan. "Officials Push for Study to Run Light Rail in Queens." NY Daily News. NY Daily News, 07 June 2016. Web.
[7] "Rolling Sales Data." *Rolling Sales Data*. N.p., n.d. Web.
[8] "Trulia. Your Home for Real Estate." Trulia: Real Estate Listings, Homes For Sale, Housing Data. N.p., n.d. Web. 17 Dec. 2016.

Appendix

2012 NYC Farmers Market List
https://data.cityofnewyork.us/Business/2012-NYC-Farmers-Market-List/b7kx-qikm

2015 Street Tree Census
https://data.cityofnewyork.us/Environment/2015-Street-Tree-Census-Tree-Data/uvpi-gqnh

After School Programs
https://data.cityofnewyork.us/Social-Services/After-School-Programs/6ej9-7qyi

Agency Service Centers
https://data.cityofnewyork.us/Social-Services/Agency-Service-Center/nn5y-wmuj

Airport Points
https://data.cityofnewyork.us/City-Government/Airport-Point/f6st-pb23

Colleges and Universities
https://data.cityofnewyork.us/Education/Colleges-and-Universities/4kym-4xw5

Day Care Centers
https://data.cityofnewyork.us/City-Government/Day-Care-Center/3nxf-gbay

DFTA Senior Centers
https://data.cityofnewyork.us/Social-Services/DFTA-Senior-Center-Map/gtwb-v72z

Directory of Job Centers
https://data.cityofnewyork.us/Business/Directory-Of-Job-Centers/9d9t-bmk7

DSNY Graffiti Information
https://data.cityofnewyork.us/City-Government/DSNY-Graffiti-Information/gpwd-npar

DYCD Beacon Locations
https://data.cityofnewyork.us/Social-Services/DYCD-Beacon/n3et-mfjw

DYCD Out of School Youth Services
https://data.cityofnewyork.us/Social-Services/DYCD-OSY-Out-of-School-Youth-/2n64-63dq

DYCD Summer Youth Employment
https://data.cityofnewyork.us/Social-Services/DYCD-SYEP-Summer-Youth-Employment-/yqz9-aduk

DYCD Runaway and Homeless Youth Services
https://data.cityofnewyork.us/Social-Services/DYCD-RHY-Runaway-and-Homeless-Youth-Services/h682-ywyg

DYCD Young Adult Internship Programs
https://data.cityofnewyork.us/Social-Services/DYCD-YAIP-Young-Adult-Internship-Programs-/s2d8-h5fg

Hurricane Evacuation Centers
https://data.cityofnewyork.us/Public-Safety/Hurricane-Evacuation-Centers/ayer-cga7

Individual Landmarks
https://data.cityofnewyork.us/Recreation/Individual-Landmarks/ch5p-r223

Library Locations
https://data.cityofnewyork.us/Business/Library/p4pf-fyc4

LinkNYC Locations
https://data.cityofnewyork.us/Social-Services/LinkNYC-Map/tgrn-h24f

New York City Art Galleries
https://data.cityofnewyork.us/Recreation/New-York-City-Art-Galleries/tgyc-r5jh

NYC Wi-Fi Hotspot Locations
https://data.cityofnewyork.us/Social-Services/NYC-Wi-Fi-Hotspot-Locations/a9we-mtpn

NYCHA Community Facilities
https://data.cityofnewyork.us/Social-Services/Map-of-NYCHA-Community-Facilities/pv8j-5ywy

Path Station Locations
https://data.cityofnewyork.us/City-Government/Path-Station-Locations/acxp-7ep7

Public Pay Telephone Locations
https://data.cityofnewyork.us/Social-Services/Public-Pay-Telephone-Locations/t7sx-id53

Public Recycling Bins
https://data.cityofnewyork.us/Environment/Public-Recycling-Bins/sxx4-xhzg

Sidewalk Café Licenses and Applications
https://data.cityofnewyork.us/Business/Sidewalk-Caf-Licenses-and-Applications/qcdj-rwhu

Subway Entrances
https://data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56