Reveal Estate

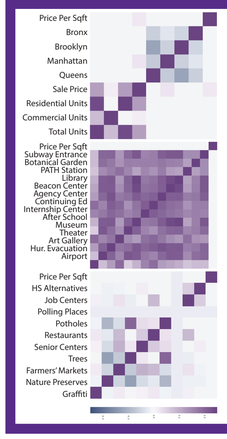
Nora Barry, Laura Buchanan, Jacqueline Gutman , New York University, Center for Data Science

Overview

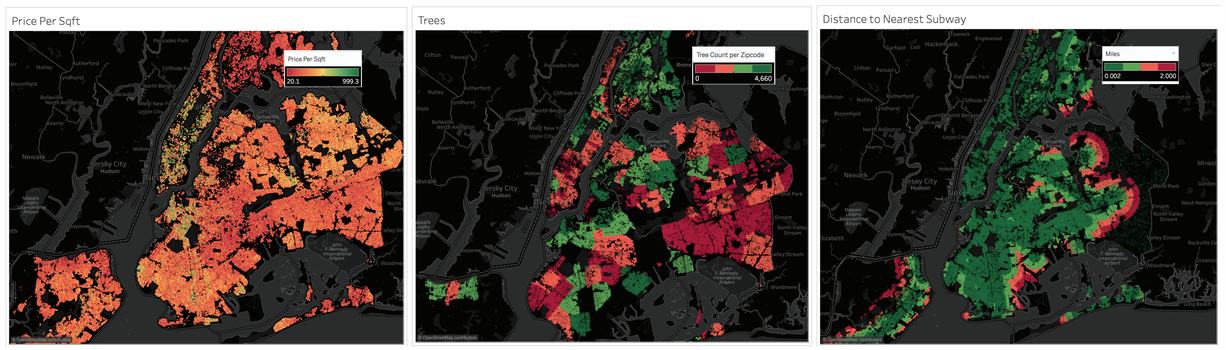
We built a Random Forest model to predict NYC real estate prices based on building features and neighborhood characteristics. With this model, we predict an estimated 200% real estate value increase for the properties within a half mile of the proposed Brooklyn Queens Connector. Furthermore, this property value increase could lead to a $53 million increase in tax revenue for the City of New York.

NYC OpenData NYC Department of Finance

Data



Real estate sale prices were gathered from the NYC Department of Finance (DoF) Annualized and Rolling Sales data from 2003-2016 for all NYC boroughs. We use the 'BBL' (unique ID for each property), 'Sale Price,' and 'Sale Date.' We then merge NYC PLUTO data, which provides property features for every BBL in NYC. With the 'Gross Square Feet' feature, and the previously obtained 'Sale Price,' we derive our model's target variable 'Price Per Square Foot.' We then merged additional datasets from NYC Open Data to incorporate additional spatial features into our model. Left: heatmaps of feature correlation for PLUTO features, distances features, and count features. To the right: spatial distributions of selected features.



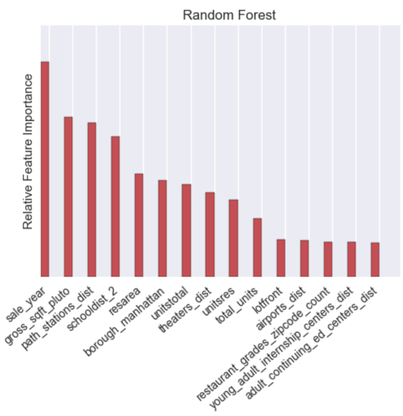
Model

	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	✓	-

After data was curated, cleaned, and merged, we performed a gridsearch to find the best method to model the 'Price Per Square Foot' for each building in NYC that was sold between 2003 and 2016. The gridsearch parameters tested are summarized to the left. Because we assumed our features could have interaction effects, it was important to include non-linear models in our gridsearch. We discovered Random Forest to be the best performing model for our data. The final parameters for our Random Forest included:

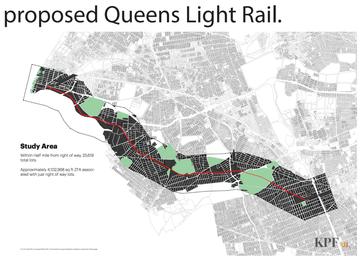
Maximum Percent of Features: 40%
Minimum Samples Split: 20
Minimum Samples Per Leaf: 10
Maximum Depth: 100
Bootstrap: False

Due to the large size of our merged data, we used NYU's High Performance Computing (HPC) services to perform the gridsearch. Median absolute error was 35.6. Gradient Boosting was the second best performing model.

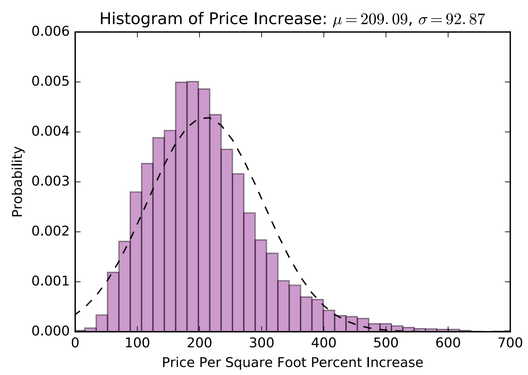


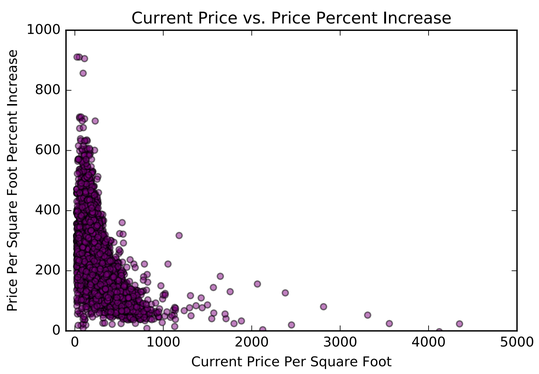
Model Application

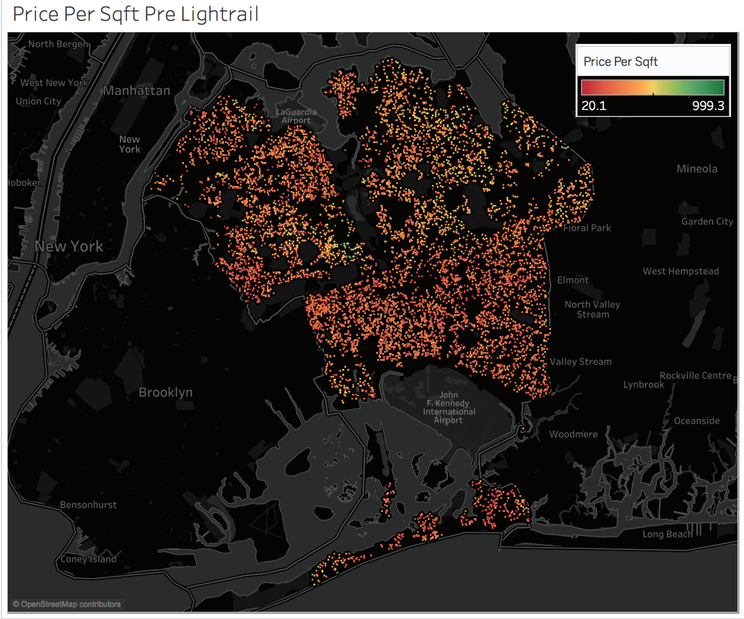
To demonstrate the usefulness of our model, we applied it to the proposed Queens Light Rail.

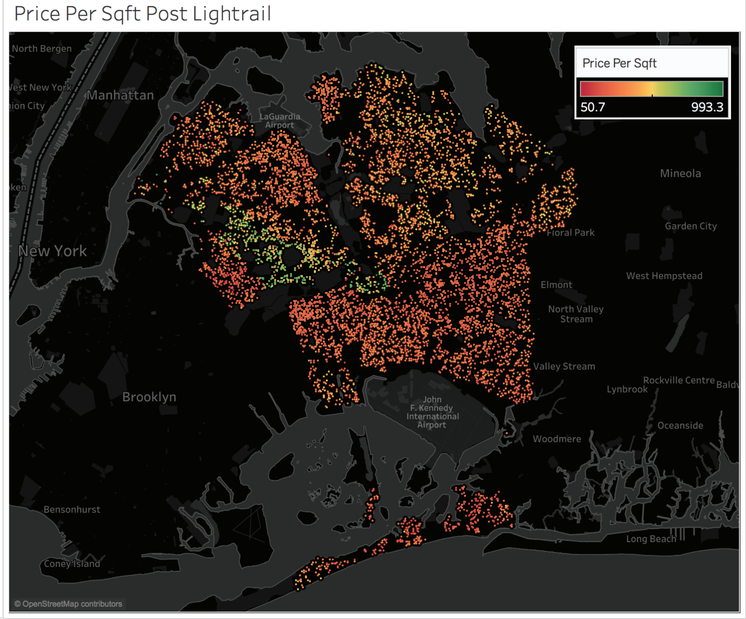


To estimate the effect of the proposed streetcar, we implemented our model to predict the increase of value on the BBL's within a half mile of the Queens Light Rail. Since there are currently no streetcar lines in NYC, we used 'Distance to Subway Entrances' as our proxy to 'Distance to Streetcar.' BBL's within a half mile with greater than 0.5 miles, to the nearest subway were replaced with 0.5 miles. We then predicted 'Price per Square Foot' for these BBL's with out









We found that the average 'Price per Square Foot' of the BBL's within a half mile of the proposed Queens Light Rail increased by 209%. The average effective property tax rate in Queens is 0.796. Given this increase in property values, the expected additional income for the City of New York would be $53 million per year in property taxes, an increase of 120% the current revenue.

Conclusion

- We completed an extensive data preparation project on NYC Department of Finance Building Sales data and spatial NYC Open Data.
- With this data, we built a Random Forest model to predict 'Price per Square Foot' for properties in NYC.
- Our model had a performance of 35.6 in terms of median absolute error.
- We displayed the usefulness of our model by applying it to the predict building value increase given the proposed Queens Light Rail.
- We found that this project could raise $53 million for the City of New York per year in tax revenue and increase property values within a half mile of the Queens Light Rail by an average of 209%.

References

"PLUTO and MapPLUTO Archive." PLUTO and MapPLUTO Archive. N.p., n.d. Web.
"Annualized Sales Update." Annualized Sales Update. N.p., n.d. Web.
"NYC Open Data." NYC Open Data. N.p., n.d. Web.
"Brooklyn Queens Connector (BQX)." NYCEDC. N.p., n.d. Web.
"Build the BQX NOW!" Brooklyn Queens Connector. N.p., n.d. Web.
Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.