NYC PROPERTY SALES PREDICTION REPORT

BYTES MATTER TEAM
Sri Manasa Santhoshi Penmetsa (ssp2187)
Sanchitha Balasubramanian (sb3610)
Tanvir Kahlon (tk2952)
Praveen Sivashangaran (ps3247)
Benarivo Benarivo (bb3052)

Introduction

New York City is home to one of the most expensive, and competitive real estate markets in the world. Creating a robust property price prediction model would be useful to real estate agents and agencies. Therefore, we obtained and studied the New York City property sales publicly available financial record data from nyc.gov. This data represents details regarding properties sold as well as financial details of the sale over the 12-month rolling period between 2016-2017. After preprocessing, we employed data exploration and data visualization strategies to understand the pertinent features of property price prediction in NYC. Subsequently, we created machine learning models including Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, kNN Regression, Decision Tree, Random Forest, Gradient Boosting Regression and Neural Networks to improve price prediction of NYC property sales.

Data Preprocessing

Initial dataset included 84548 sale entries across 22 features. These features included property borough, block, address, Zip Code, gross square feet, total units, year built, as well as sale price, tax class, date of sale. Cursory review of data demonstrated multiple features with missing values or nonsensical values. Additionally, certain ordinal and non-ordinal categorical features were in numerical data formats rendering them inappropriate for informative data analysis.

We proceeded to remove empty or redundant data columns, and duplicate entries were also removed. After the literature search, we determined that property sale prices below $10,000 and above $169 Million in NYC were unlikely to be accurate and therefore removed these row entries. Row entries with either data on year built or year built before the 18th century were also removed. Missing square footage data of properties constituted over 30% of the data. Given the potentially large effect of removing these entries, we planned for missing value imputation. Missing square footage data was filled using the median square footage of the borough that the property was in. The central measure of median was selected due to the heavy positive skew of the sale price data (Figure 1A).

Additionally, a new feature termed “Year Sold” was created by extracting the year from the sale date of each property sale.

Preliminary literature search of property sales in the US highlighted the importance of seasonality and age of buildings. We wanted to understand if these variables played an equally important role in the unique real estate market of New York City. Therefore, we feature-engineered “Sale Season” variable and a “Building Age” variable. Moreover, to study geographic distribution of key features, we mapped NYC zip codes to latitudes and longitudes for choropleth mapping.

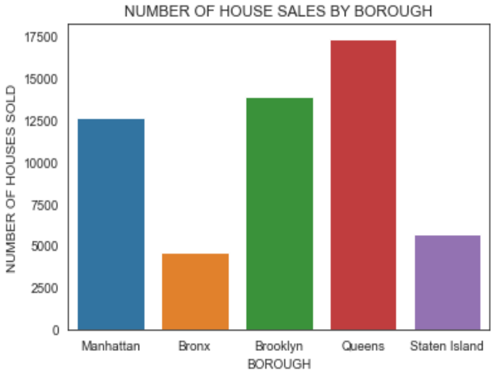
Categorical features underwent One-Hot encoding for purposes of price prediction modeling. The result included multiple features that had unique values in the hundreds. Therefore, we created two separate models based on the number of unique categories per feature to evaluate their respective significance for price prediction. This is further explained in the Modeling section.

Data Exploration and Visualization

Data visualization led to improved understanding of features impacting sale price as well as if novel feature engineering would yield relevant trends in sale price, our target variable.

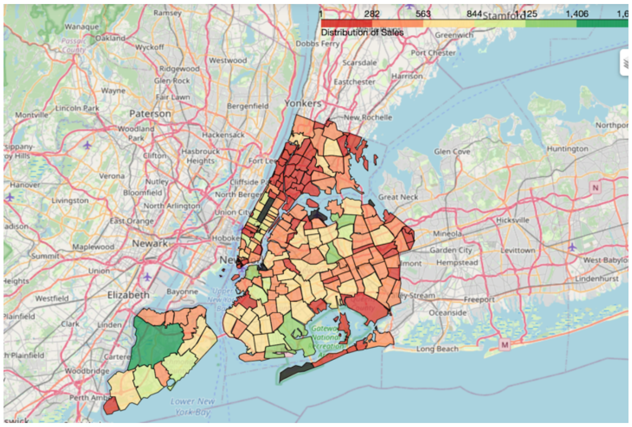
Comparison of number of property sales in each borough revealed that Queens had the highest number of property sale transactions while Bronx had the lowest (Figure 2.1).

Figure 2.1



Unexpectedly, a zip code level drill down of the number of sales demonstrated 10314 zip code in Staten Island to have the largest number of property sales (Figure 3).

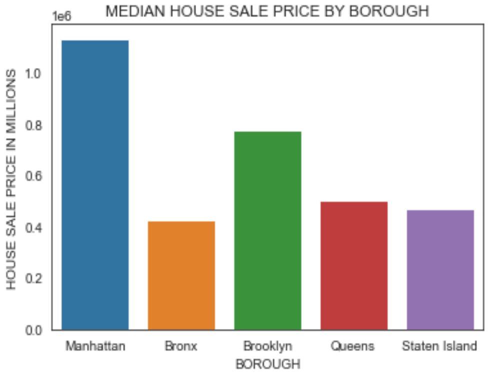
Figure 3: Choropleth map of number of property sales by each zip code



Though boroughs are often considered united by not only their municipal governance but also due to significant cultural and market trend similarities, above findings highlighted the considerable variation within each borough as it relates to property sales.

Intuitively, borough level comparison of median property sale prices yielded Manhattan as the borough with most expensive sale prices. Bronx, on the other hand, had the lowest median property sale prices of all the five boroughs (Figure 2.2).

Figure 2.2



Furthermore, seasonality does not appear to play a significant role in property sales across the five boroughs (Figure 2A) in contrast to rest of US property sales which are notably higher in spring and summer months.

Model Training and Evaluation

To develop robust machine learning models, we created a correlation heat map of numerical features to delineate feature redundancy and linear dependency (Figure 3A). We identified certain features that were highly correlated to each other but retaining them improved model performance, so we chose to keep them for final model development. Moreover, after "One Hot" encoding, we removed an additional three features due to the significant percentage of unique values, such as address, neighborhood, and apartment number, as they would be non-contributory to modeling.

We created two separate data frames of the resulting features for model training. The two separate data frames, termed the Simple Model and the Complex Model from here onwards, were created to reason if the additional complexity would result in an increase in model accuracy. Though intuitively, many unique values for a categorical variable renders it irrelevant, we wanted to identify the ideal cut-off for a unique number of categories per feature. The Simple model, thus, had all categorical features with less than 50 unique categories one hot encoded, whereas the Complex model had all categorical features with less than 200 unique categories one hot encoded. The Simple Model resulted in 70 features, excluding the target variable, while the Complex model resulted in 544 features. We, subsequently, used Principal Component Analysis to reduce the number of features from 544 to 205 for the Complex Model. This was done such that 80% of the total variance was retained, while reducing dimensionality. It should be noted that prior to using PCA, the features were standardized such that the data had a mean of 0 and unit variance, to ensure a fair comparison between the explained variance in the dataset.

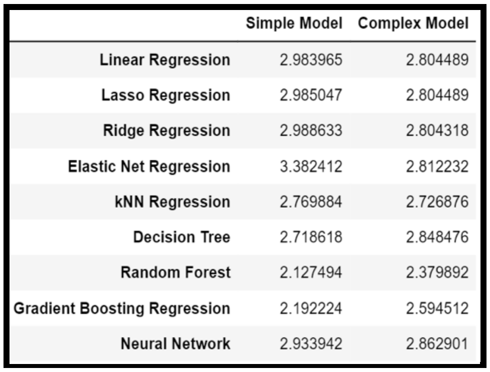
We trained and evaluated a total of nine learning algorithms for both the Simple Model and the Complex Model as follows: Linear Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, kNN Regression, Decision Tree, Random Forest, Gradient Boosting Regression and Neural Networks. Due to the additional complexity and computational time required to train the Complex Model, it was determined that the Simple Model would be preferred unless the Complex Model significantly outperformed the Simple Model.

Given the objective of effectively predicting sales price of a property in NYC based on historical data is a regression problem, we selected the model evaluation metric of normalized root mean squared error. Due to the large range in the target variable (lowest house price was $10,000 while the highest house price was approximately $169 million), the root mean squared error was normalized by dividing the root mean squared error by the mean of the target variable instead of the range.

$$NRMSE = \sqrt{\frac{\Sigma\,(\hat{y} - y)^2}{N}} \;/\; \bar{y}$$

The data was split into a 70/30 training and test set split for all learning algorithms to ensure consistency. It should also be noted that a random seed of 42 was used for reproducibility purposes. Hyperparameter tuning was performed for some of these models. For example, for the kNN regression model, we iteratively look for the k-parameter (number of nearest neighbors) with the lowest normalized root mean square error (NMRSE). Moreover, for the Gradient Boosting Model (Figure 4A), we chose the maximum depth of the trees that gave the lowest NMRSE (Figure 5A). For the neural network, two hidden layers were used with a dropout rate of 20% as a means of variance reduction, and a "rule of thumb" methodology was adopted to determine the number of neurons in the hidden layers. The number of neurons in the first hidden layer was determined by taking the mean of the number of neurons in the input and output layers, and the number of neurons in the second hidden layer was one half of the number of neurons in the first hidden layer.

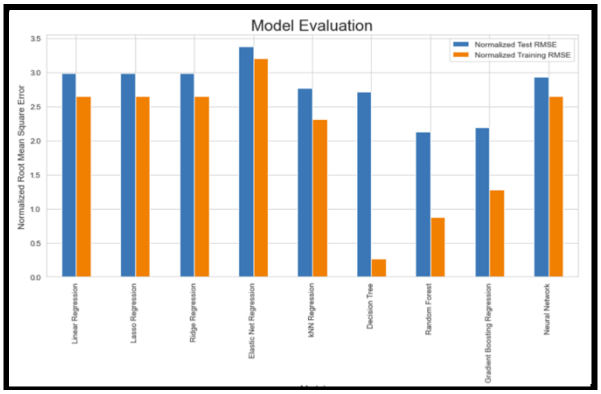
Table 1: Normalized test root mean squared error of the Simple and the Complex Model



	Simple Model	Complex Model
Linear Regression	2.983965	2.804489
Lasso Regression	2.985047	2.804489
Ridge Regression	2.988633	2.804318
Elastic Net Regression	3.382412	2.812232
kNN Regression	2.769884	2.726876
Decision Tree	2.718618	2.848476
Random Forest	2.127494	2.379892
Gradient Boosting Regression	2.192224	2.594512
Neural Network	2.933942	2.862901

After all, nine learning algorithms were evaluated for both the Simple Model and the Complex Model, it was determined that the best learning algorithm in both cases was the Random Forest Model and in fact, the Simple model had the lowest normalized root mean squared error (Figure 4). It was observed that the increase in complexity due to the additional features in the Complex model led to better performance in the variants of the linear regression models (Figure 1A) but the absolute best model was the Random Forest model trained with fewer features (Simple model criteria) (Table 1).

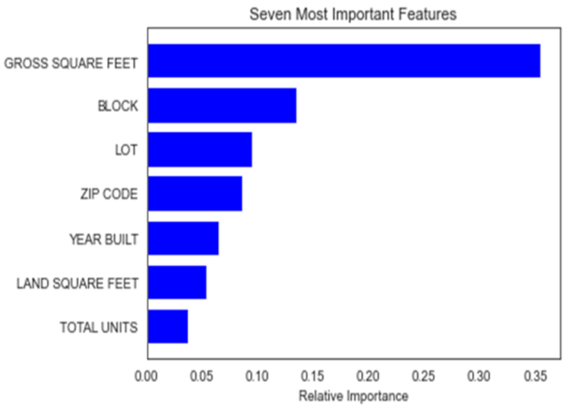
Figure 4: Normalized test and train root mean squared error for the Simple Model



Conclusion, Implications and Future Work

Overall, data analysis of 12-months of property sales in NYC leads to certain key highlights of the market. Our data analysis demonstrates that seasonality plays less of a role in NYC sales than it does in the rest of the country. Moreover, geographical distribution of sales and sale prices creates unique dynamics in property sales in NYC, namely that sales behaviors can be markedly different in a short distance from each other. As such, our absolute best model, Random Forest, highlights the smallest geographical feature of a block to be one of the most important features for model training, second only to gross square feet (Figure 5).

Figure 5: Random Forest Feature Ranking



Due to the unexpected underperforming of the Neural Network Model, additional time invested in hyperparameter tuning should be justifiable. Due to training time being the main bottleneck in our current setup, the access to additional computing resources should enable further experimentation with a grid search-style approach comparing different parameters including the number of hidden layers, activation function and neurons per layer.

Upon further literature review involving Zillow's property price forecasting methodology [1], a key improvement to this model would be the access to higher quality features that are intuitively strongly correlated with the property sales price such as Median Income, Crime Rate, Number of Public schools, Hospitals and Hospital ratings. If the cost of collecting this additional data for each property sale is reasonable, it is a venue for improvement as model accuracy is strongly anticipated to improve.

For further refinement of our prediction model, given the impact of geographic trend variations over small areas on property prices, we would like to consider sentiment analysis of neighborhood/ block businesses to identify the "cool and trendy" blocks and the respective effect on property prices. We believe such an analysis may portend the block where the next shifts in property prices is to occur to identify intelligent real estate investments in New York City.

References

[1] "Zestimate Forecast Methodology." Zillow Research, 11 Apr. 2016, https://www.zillow.com/research/zestimate-forecast-methodology/.

Appendix:

Figure 1A: Heavy positive skew of NYC property sale prices

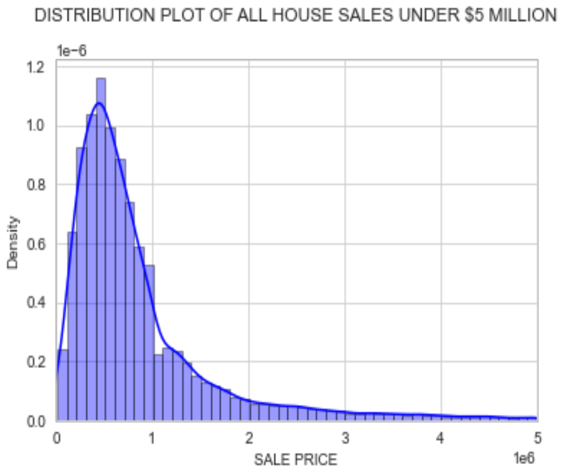


Figure 2A: Number of sales per season in each borough

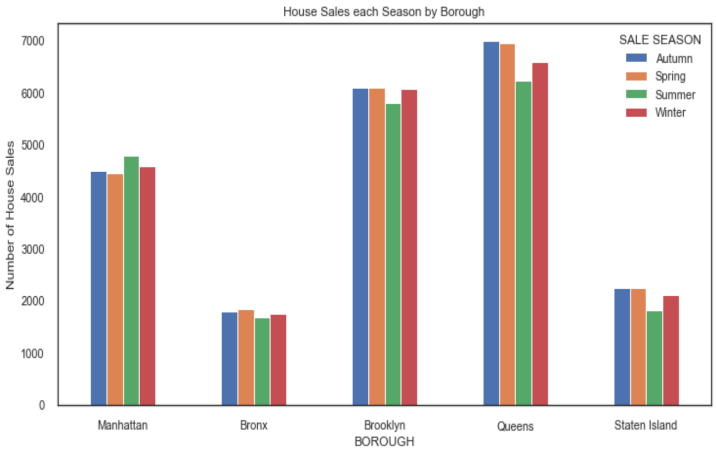
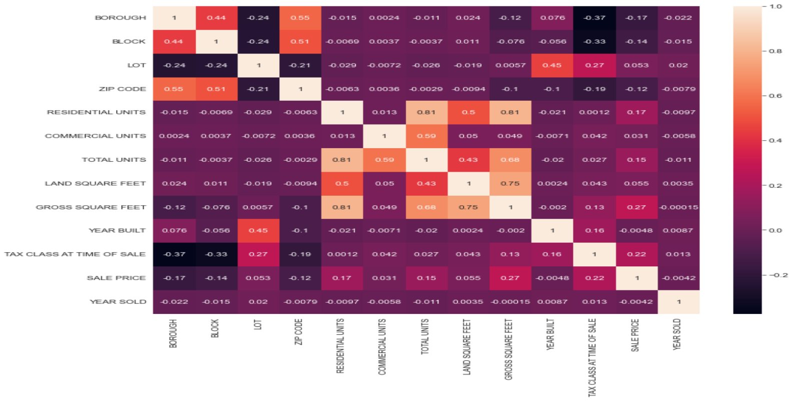
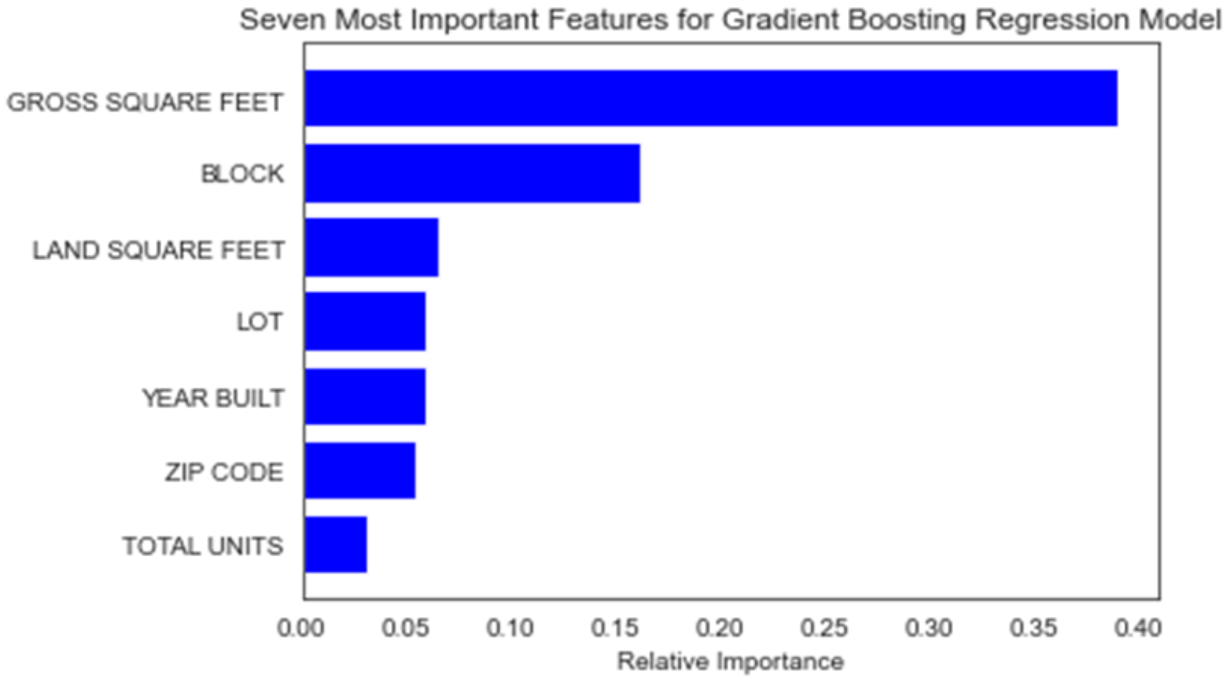


Figure 3A: Feature correlation heatmap



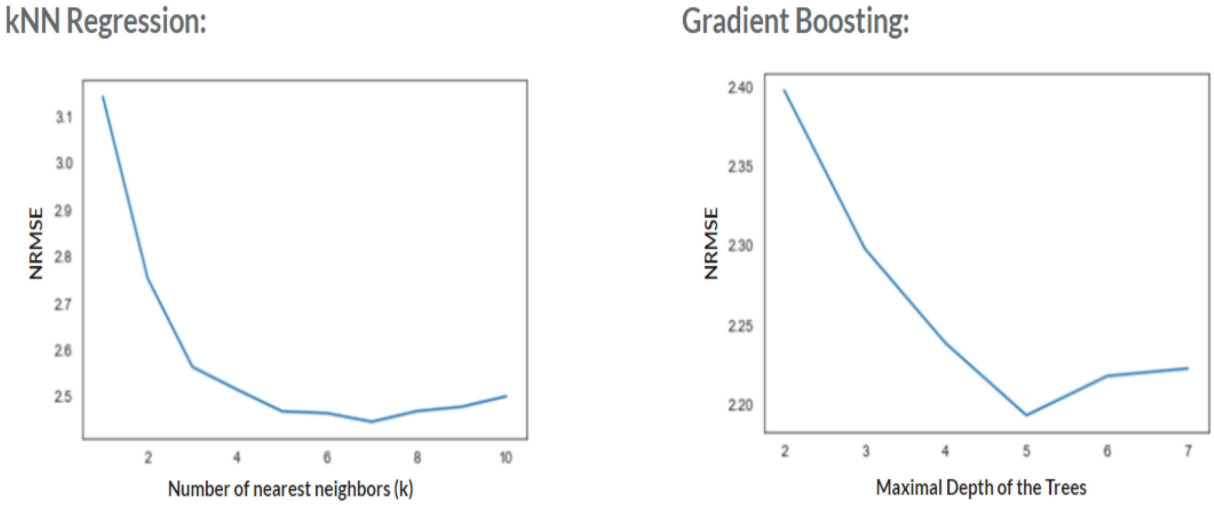
	BOROUGH	BLOCK	LOT	ZIP CODE	RESIDENTIAL UNITS	COMMERCIAL UNITS	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	YEAR BUILT	TAX CLASS AT TIME OF SALE	SALE PRICE	YEAR SOLD
BOROUGH	1	0.44	-0.24	0.55	-0.015	0.0024	-0.011	0.024	-0.12	0.076	-0.37	-0.17	-0.022
BLOCK	0.44	1	-0.24	0.51	-0.0069	0.0037	-0.0037	0.011	-0.076	-0.056	-0.33	-0.14	-0.015
LOT	-0.24	-0.24	1	-0.21	-0.029	-0.0072	-0.026	-0.019	0.0057	0.45	0.27	0.053	0.02
ZIP CODE	0.55	0.51	-0.21	1	-0.0063	0.0036	-0.0029	-0.0094	-0.1	-0.1	-0.19	-0.12	-0.0079
RESIDENTIAL UNITS	-0.015	-0.0069	-0.029	-0.0063	1	0.013	0.81	0.5	0.81	-0.021	0.0012	0.17	-0.0097
COMMERCIAL UNITS	0.0024	0.0037	-0.0072	0.0036	0.013	1	0.59	0.05	0.049	-0.0071	0.042	0.031	-0.0058
TOTAL UNITS	-0.011	-0.0037	-0.026	-0.0029	0.81	0.59	1	0.43	0.68	-0.02	0.027	0.15	-0.011
LAND SQUARE FEET	0.024	0.011	-0.019	-0.0094	0.5	0.05	0.43	1	0.75	0.0024	0.043	0.055	0.0035
GROSS SQUARE FEET	-0.12	-0.076	0.0057	-0.1	0.81	0.049	0.68	0.75	1	-0.002	0.13	0.27	-0.00015
YEAR BUILT	0.076	-0.056	0.45	-0.1	-0.021	-0.0071	-0.02	0.0024	-0.002	1	0.16	-0.0048	0.0087
TAX CLASS AT TIME OF SALE	-0.37	-0.33	0.27	-0.19	0.0012	0.042	0.027	0.043	0.13	0.16	1	0.22	0.013
SALE PRICE	-0.17	-0.14	0.053	-0.12	0.17	0.031	0.15	0.055	0.27	-0.0048	0.22	1	-0.0042
YEAR SOLD	-0.022	-0.015	0.02	-0.0079	-0.0097	-0.0058	-0.011	0.0035	-0.00015	0.0087	0.013	-0.0042	1

Figure 4A: Gradient Boosting Feature Ranking



Seven Most Important Features for Gradient Boosting Regression Model

Figure 5A: Parameter tuning



kNN Regression:

Gradient Boosting: