Property Price Evaluation for Manhattan

Junjie Cai, Jianwei Li, Wenjie Zheng

Table of Contents

Abstract

New York City is home to 8 million people and makes up a huge and dynamic property sale market. Numerous factors have an impact on the housing and real estate pricing, including location, amenity, and community service, etc. This project predicts the value of a property given various characteristics of a real estate property in Manhattan, New York, which leads to a deeper understanding of the real estate market. The analysis used rolling sales data for the properties, MTA station information for transportation accessibility, school district-based SAT scores for a measure of education quality, census data as well as parks and crime statistics for measurement of amenities and neighborhood desirability. Multiple machine learning models were used including Linear Regression models, Decision Tree and the corresponding ensemble models, and KNN model. Overall, the Gradient Boost model scored the highest performance for both accuracy and robustness.

Keywords: property value, sales, housing, regression

Introduction

The factors affecting the property value of a particular real estate consist of numerous features including not only the layout of the real estate itself but also the extrinsic characters that the real estate possesses. The characteristics of a property often include the number of bedrooms, number of bathrooms, area, floor, elevator accessibility, as well as the neighborhood desirability, transportation accessibility, school district, and general security. Getting an understanding of the structure that makes up the property value leads to valuable insights into multifarious subjects and fields, which inevitably aid in solving other urban problems including noise control, traffic congestion, education quality, and etc. This paper seeks to explore the structure of the property value, to evaluate and predict the property values with the Rolling Sales Data between April 2018 and March 2019 for Manhattan, New York City integrated with a variety of information about a property and the neighborhood where it is located.

Literature

Relevant researches and papers provide crucial contextual knowledge in helping us better approach our objectives and understand the property sales market. To begin with, the Zillow's Zestimate Forecast predicts the change in property value for the next 12 months. It uses mainly two datasets which include the county level Zillow Home Value Index and property characteristics to formulate a time-series. By comparing with the real sale record retrospectively, the Zestimate Forecast achieves an error of 5.84%. (Rao, 2014)

The Crompton paper investigated the impact of the green area and parks on the proximate property values. (Crompton, 2017) It started out with a simple economic approximation and calculation that involve park area, neighborhood area, down payment on land acquirement, lot sizes, property taxes, and the willingness to pay for the premium of enjoying and being close to the park, and other factors, etc. The paper talked about some negative impact of parks on the nearby neighborhood. One of many is induced crimes, noise, parking, traffic congestion, and littering due to the influx of visitors. The earliest recognition of the relationship between parks and nearby lots dates back to the effort of acquiring the land of the nowadays Central Park by the City and State Park Authorities in 1856, which the Central Park is expected to be funded and financed from the incremental tax from nearby appreciated lands resulted from the establishment of the Central Park. (Metropolitan Conference of City and State Park Authorities, 1926, p. 12) In the study of the general impact that parks have on the nearby properties, the researchers analyzed several earlier case studies from the cities across the U.S. including the Clinton Park and the San Antonio Park in the Oakland case, the Pennypack Park in Philadelphia, and the green belt in Boulder, Colorado. Most case studies lead to the conclusion that the nearby property values increase from the adjacent green areas. The paper also discussed a theory called the Proximate Principle, resulting from a number of studies which suggest that properties located closer to the parks and forests receive higher appreciation in land value than those located further away.

Accessibility to the nearby railroad for great transportation convenience is deemed to be a strong determinant of the property values and housing values while numerous studies have conducted relevant researches on the relationship between the two aforementioned subjects. Many have done studies on the cities around the world but few about New York City and New York City Transit. Having convenient access to the public transportation system enjoys a variety of benefits. The paper suggests that the benefits include lower transportation costs, better accessibility to other places, shorter time wasted on the road, and the potential rise in diversity and prosperity of commercial and human activities, offices, and changes in the zoning plan for greater potentials. (Lewis-Workman & Brod, 1997) The study performed a hedonic price estimation by performing multiple regressions aided by the geographic information system (GIS) technique. To be specific, the change in property value is in the function of distance to the railway stations, and other variables. The paper talks about measuring the distance in transit access models and focused on local effects of a metro station of a 1-mile radius from a station for the transit systems in San Francisco, California, New York City, New York, and Portland, Oregon. The study used Zillow property transaction data and GIS data from local governments. It recognized the shortcoming in calculating the distance between properties and transit stations by assuming the residents traveling in a straight line which may actually not. The study used the GIS software to precisely determine the walking distance from each property to the proximate transit station provided by Criterion Inc. and did a case study for the two neighborhoods including Forest Hill and Rego Park in Queens, NY. The result of the model from the case study showed that the properties within the walking distance from nearby subway stations enjoyed a high level of benefits. That is, property values within the walking-distance are 13% higher than those outside of the zone. Having the scale as close as neighborhood level is congruent with our project's scale of analysis.

Data and Methods

Data Sources

The data sources table below contains different categories of data, dataset names, and links that include the datasets we collected from various sources. (Table 1)

Data Examination, Cleaning, and Integration

Given the datasets and sources mentioned in the table above, every dataset is then downloaded individually and cleaned individually before being integrated into a general data frame.

Building Footprints (BBL). The first dataset collected is the Building Footprints also referred to as the Borough Block Lot (BBL) from the NYC Open Data Portal. Every BBL basically has the shape of each individual building. It is one of the fundamental datasets needed for the analysis since it contains some essential information about a building. The dataset has 45,590 individual records of BBL for BBLs in Manhattan, and 17 columns for different types of information including ground elevation, shapefile geometry polygon coordinate, roof height, and DOITT ID, etc. For every BBL, a single point feature, the centroid, is calculated and a buffer zone with the radius of a quarter mile (0.25 mile) is then created for every BBL's centroid. These will be used for future processes, such as calculating the distance to the closest subway entrances, bus stops, parks, and etc.

Census Tracts Shapefile Map. The census tracts are created by the Bureau of Census and have the area roughly the same as a neighborhood for demographic information assessment. The census tract dataset is collected from NYC Open Data Portal and it shows that there are 288 census tracts in Manhattan. It has GIS geometry information, borough name, NTA code, and PUMA code, etc. for each census tract which makes this dataset the primary base map source for analysis. The census tract information was mapped for every BBL within a census tract. Some buildings BBL are located in multiple census tract, that is, part of a building is within one census tract while the rest of a building is within the realm of another census tract, which makes deciding the census tract of a BBL ambiguous and biased. As a result, the aforementioned BBL centroid is used in the calculation so that the BBL can be located in an exact census tract.

Census Tracts Demographic Data. The demographic dataset comes at 48 columns of different kinds of information for 288 census tracts and provides information about the population of different racial groups, economic information, employment information for different types of status such as private work, public work, self-employed, family work, and unemployed, as well as commute means including drive, carpool, transit, walk, and etc. The census tract demographic information is then integrated with earlier census tract and BBL dataset.

Transportation Dataset: MTA Subway Station Entrances & MTA Bus Stops. Transportation and accessibility of a building provide great convenience to the amenity of a building. The study done by Lewis-Workman and Brod indicates that transit convenience offers a variety of benefits and affects property values. (Lewis-Workman & Brod, 1997) Thus, a dataset of subway entrance locations is collected from the NYC Open Data Portal. The dataset shows that there are 1928 subway entrances and every single one of them has a unique geographic coordinate. The accessibility of a building to the closest subway station in terms of the distance from the BBL centroid to the nearest subway station entrance is calculated with the *nearest_points* function from the Shapely package. As a result, every BBL has a feature for the closest distance to the nearby subway entrance. The bus stop dataset that has geometry information for 16,231 MTA bus stops is collected from the MTA Google Transit Feed Specification (GTFS). A similar approach is used and a BBL's distance to the closest bus stop is also calculated.

Amenity Dataset: Parks. Accessibility to nearby parks has an impact on the property values as well indicated by Crompton's empirical study that analyzed more than 40 cases from different cities. (Crompton, 2017) The dataset for parks is collected from NYC Parks and there are 384 parks in Manhattan. Each park has 36 features and geometry geographic information. Overall, there are 17 distinct types of parks including community park, garden, playground, neighborhood park, waterfront facility, and etc. They are categorized into 3 categories namely '1,' '2,' and '3' based on the level of significance. The Central Park is categorized as the Flagship Park by NYC Parks and it is assigned a value of 3. Based on the intersection between the BBL buffer zone and parks, the number of parks for each BBL can be assessed along with the information including the total area of the nearby park, categories of nearby parks, and the closest distance to the nearby parks.

NYPD Dataset: Crimes. Low crime rates and a safe atmosphere affects the property value within a neighborhood. The Crimes statistics dataset is obtained from NYPD and has all-time criminal case reports until 2018. The dataset has 464,065 records for 63 kinds of offenses categorized by NYPD such as Assault 3, Criminal Mischief, Robbery, Harassment 2, and etc. These offenses fall into 3 general categories including Violation, Misdemeanor, and Felony. The statistics were then spatial joined to the BBL dataset, so that every BBL has a statistic of a number of violations, misdemeanor cases, and felony cases happened within the buffer of the BBL created earlier.

Education Dataset: SAT Scores. There are 33 school districts in New York City and 6 of them are in Manhattan. Students living in certain school district have the capability of receiving better education and the quality of the education can be indicated by the SAT scores. The 2012 SAT Results dataset was obtained from the NYC Open Data Portal and every school has a number of test takers with the average SAT category scores including reading, math, and writing. This information was calculated and then spatially joined to the school district map such that every school district has average SAT category scores.

Property Sale Dataset: Rolling Sales Data. Last but not least, the Rolling Sales Data dataset was obtained from the NYC Department of Finance that includes the sales record for the transaction and relevant information about the unit being transacted, which include the number of units, land square feet, building area, and etc.

Data Preprocessing

Inspection & Extremum Filter. Before analyzing the data, several steps of data preprocessing are necessary, and it all starts with data inspection. The integrated dataset comes at 8,885 rows and 56 features including property sale price, date of sale, age of the building, number of residential units, crime statistics, park amenities, education quality, and etc. With the *statistics.median()*, the sales record has a median sale price of $825,000. From the Price Decile Map (Figure 1), the top 10% tier average sale price is $9,155,000, which is almost three times higher than the second tier which is from 10% to 20%. The other quartile dropped exponentially from the top 10% to 50%. All features and statistic are on different scales and have various units. The ages and area of the unit in square feet have large values and some discrete categories have small values like ones and zeros. Thus, the features require additional attention, close observations, and potential preprocessing before the analysis goes any further.

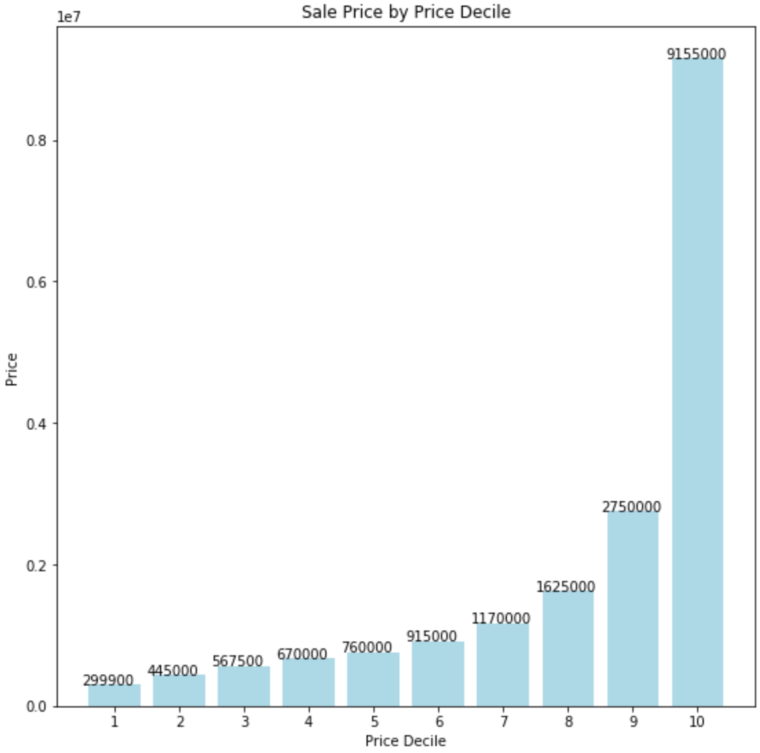


Figure 1 Sale Price by Price Decile

Gaussian Distribution Test. Understanding the distribution of the dataset with normality tests dictates whether to use a parametric statistical model or a nonparametric statistical model. Several normality tests were used including the KS Test, AD Test, and Chi-Square Test.

Kolmogorov-Smirnov Test (KS Test). The KS Test assumes that the data follows a specified distribution as its null hypothesis, which the specified distribution is Gaussian distribution in this case. (National Institute of Standards and Technology, 2012) The KS Test with *SciPy kstest()* against normal distribution returns a value of 0, which rejects the null hypothesis that the data is drawn from a Gaussian distribution.

Anderson-Darling Test (AD Test). The AD Test is slightly different from the KS Test which it assigns more weight towards the tails of the distribution. (National Institute of Standards and Technology, 2012) The AD Test with *SciPy Anderson()* against normal distribution rejects the null hypothesis that the data is drawn from a Gaussian distribution.

Chi-Square Test. The Chi-Square Test with *sp.stats.chisquare()* compares the dataset distribution with an expected chi-square distribution. The chi-square test on the original dataset returns a p-value of 0, which suggests that the dataset distribution is not a chi-square distribution while result from the chi-square test on the dataset after taken the log returns a p-value of 1 which suggests that the log-processed dataset is a chi-square distribution.

Feature Selection with Correlation & P-Value. Correlation Table was done for all the features in the dataset with pandas.corr() function. Figure 2 shows the correlation plot for all features from the original dataset. (Figure 2) Only the 44 features were selected after filtering out features with correlation value higher than 0.9. (Figure 3) The resulted feature table was then analyzed on how they affect the p-value if any one of these features is removed from the dataset. (R, 2018) 16 features turned out to be not as effective in influencing p-values and only 28 features were left. The later modeling with 28 features experienced drastic drop in model accuracy and performance. In comparison, the modeling with 44 features resulted in better performance. Thus, the p-value selection approach was not being taken into further considerations.

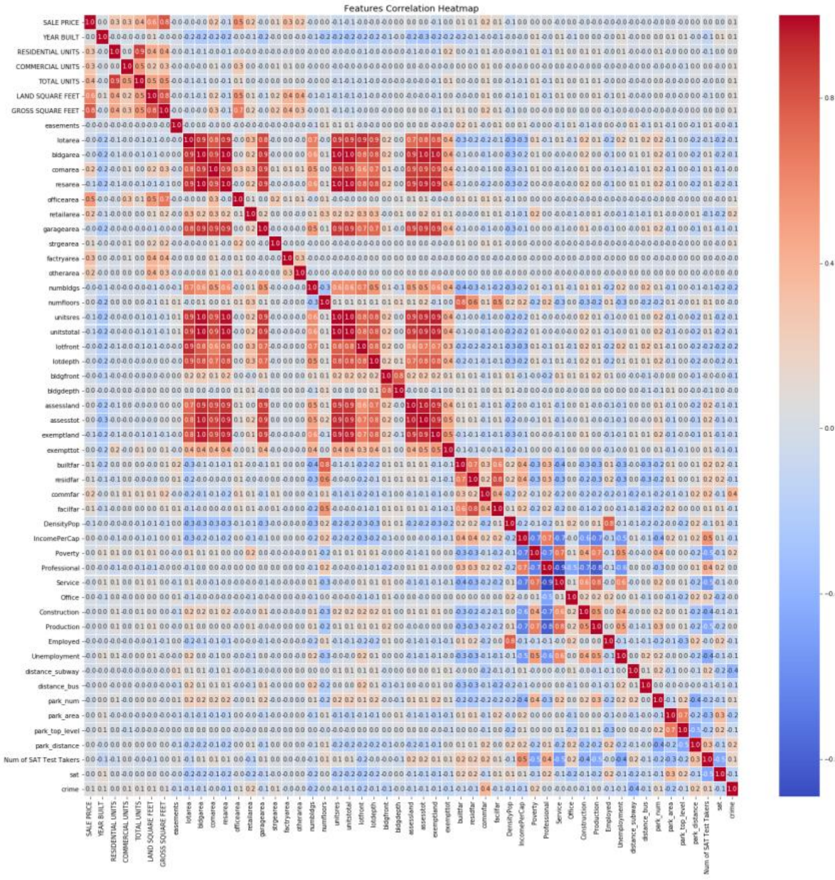


Figure 2 Correlation Heatmap for the Original Dataset

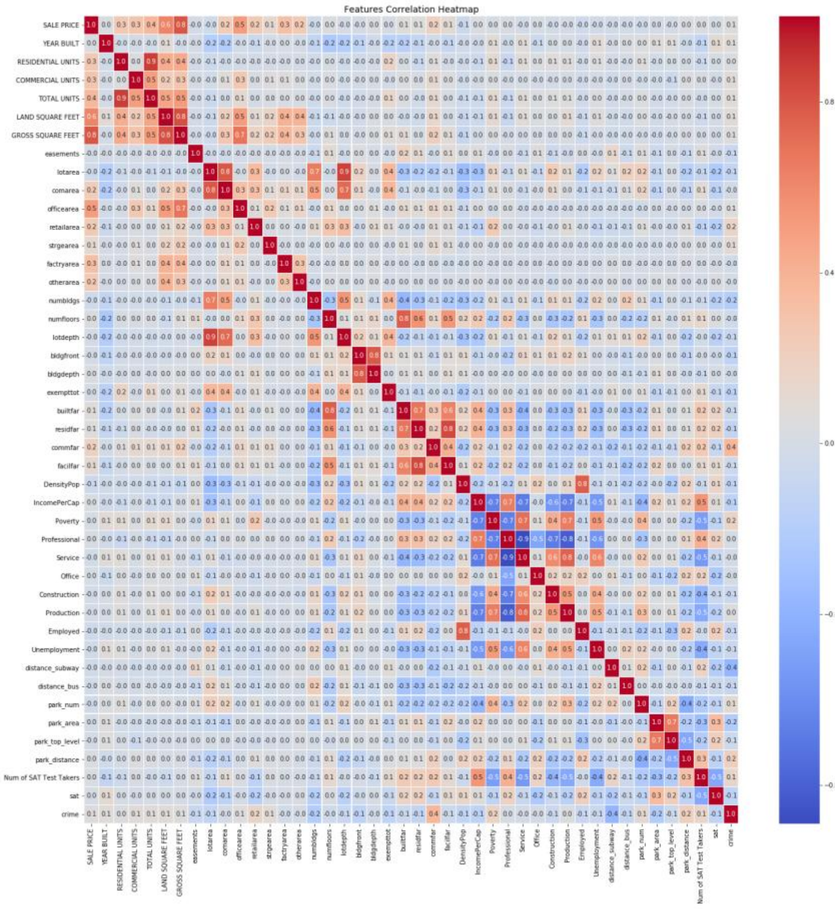


Figure 3 Correlation Heatmap after Feature Selection

Anomaly Detection. The anomaly detection is the identification of rare observation which is significantly different from the majority of the data. Due to the result of the rejection of KS Test and AD Test, it is necessary to delve into the distribution of the original data and perform the anomaly detection analysis. Firstly, we look into the general distribution with a total of 8,885 sales records in Manhattan Borough with 56 features. The minimum and maximum of the sales records are 1 USD and 9,800 Million USD, the median price at 825,000 USD and the average price at 4 Million USD, with a high standard deviation of 28,732,320. The distribution of the sale price shows a long right -tail with 55 high sale prices from 100 Million USD to 1 Billion USD (Figure 4), while the log-process distribution of the sale price shows a long left-tail with 61 sale prices from 1 USD to 10,000 USD. (Figure 5) To achieve a better model, it is possible that the price below a certain threshold could be the unreality sale price. It is possible for the transaction between family members or companies to achieve a lower tax rate.

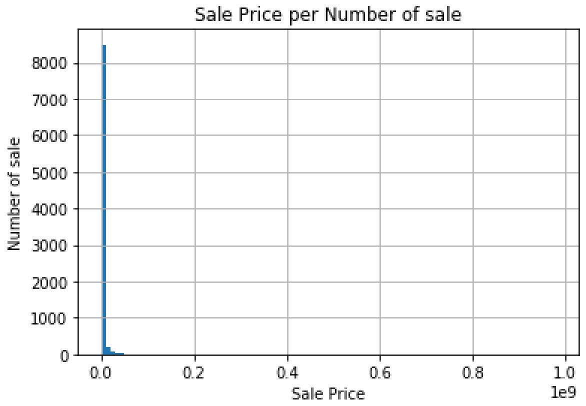


Figure 4 Sale Price per Number of Sale

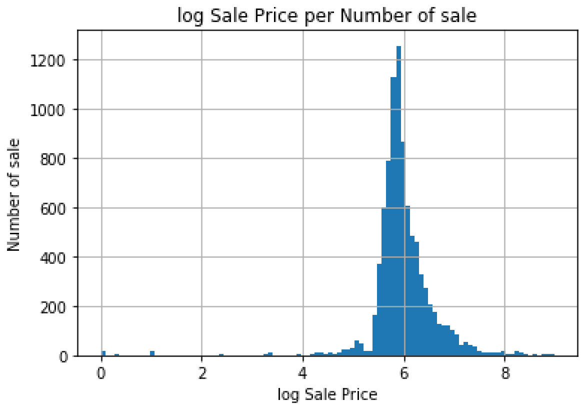


Figure 5 Log Sale Price per Number of Sale

Using the Linear Regression Model is a bench march. We use the Z-score to remove the left-tail with two standard deviations. The in-sample R squared increase from 0.64 to 0.65, and the out of sample rise from 0.58 to 0.78. The options to remove the right-tail or to use IQR (interquartile range between Q1 - 1.5IQR to Q3+1.5IQR) method to remove both tails had undermined the model's in-sample and out sample performance. After filtered by Z-score, we applied the OneClassSVM (Figure 6) and Isolation Forest (Figure 7), which are suitable for novelty detection in the large unlabeled dataset. Apart from the Linear Regression Model benchmark, we use Cross-validation to evaluate the robustness with different input data. The benchmark is 0.34 for the data filtered using Z-score. However, OneClassSVM got the cross-validation score of 0.25, which is lower than 0.34. The reason might be its features for multimodal data, which only match the two uneven small peaks at around $10,000 to 20,000K and another high peak at $1000,000. The Isolation Forest, instead, achieved the 0.66 for the ten folds of cross-validation score. In this case. The isolation forest is an ideal anomaly detection method to extract the outliers and increase the robustness for the model performance.

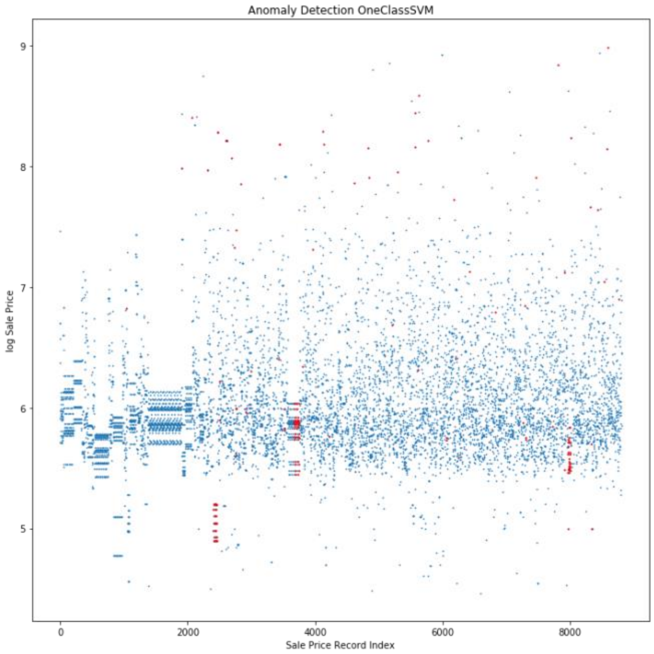


Figure 6 Anomaly Detection OneClassSVM

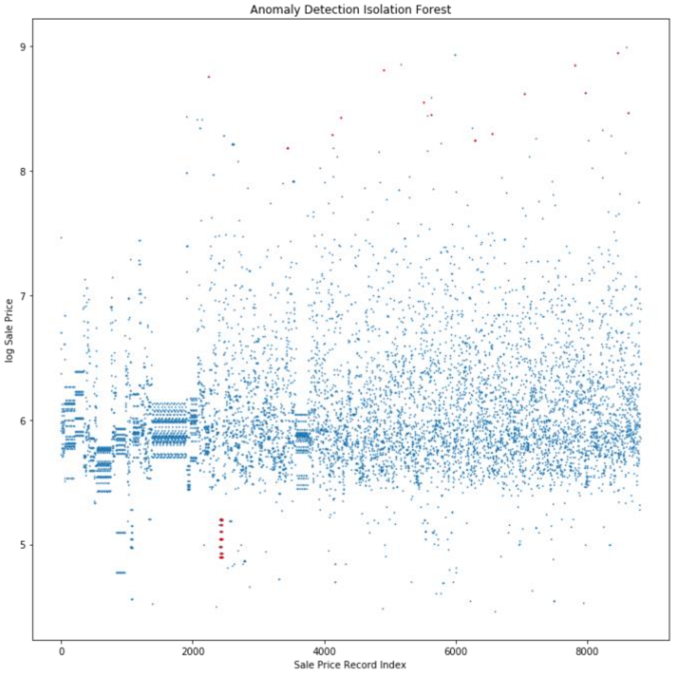


Figure 7 Anomaly Detection Isolation Forest

Standardize. Two methods were used for data standardization. The first one used *StandardScaler()* and *fit_transform()* from the *sklearn.preprocessing* so that the dataset has the mean of 0 and standard deviation of 1 instead of various scales and extremely large or extremely small numbers. Another method of standardization is similar to the previous one but implemented manually, which is to transform the data by subtracting every number by the mean and divided by the standard deviation.

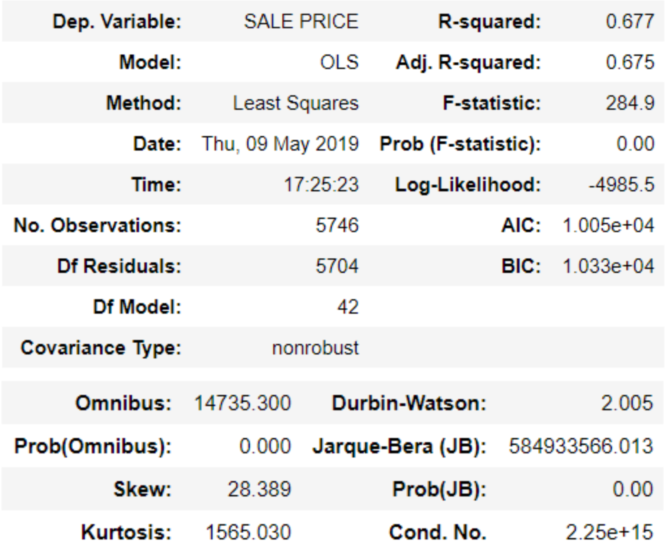
Principal Component Analysis (PCA) & Whitening. Due to the sheer number of features and strong correlations among features from earlier feature correlation table, PCA and feature reduction was considered necessary. With *PCA()* from *sklearn.decomposition*, 15 principal components were able to explain 80% of variance, while 22 components and 28 components were able to explain 90% and 95% of variance respectively.

Train & Test Split. A train test split was done to the dataset so that the accuracy can be tested against the test portion of the dataset with the *train_test_split()* function. The split ratio was set at 0.33 and the random state was set at 42.

Model Selection

Linear Models: OLS, Lasso, and Ridge.

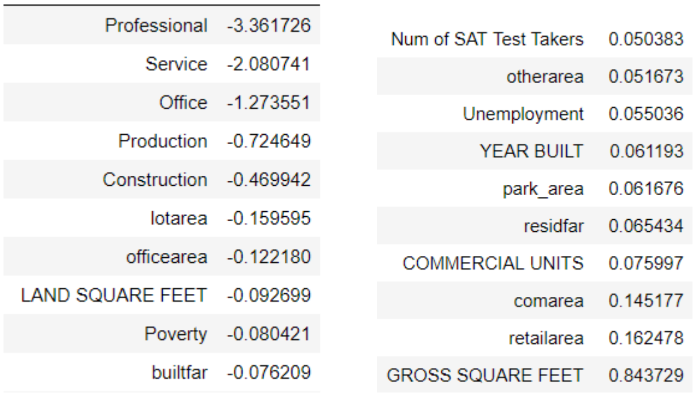
First of all, we built *multivariate linear regression models* as a benchmark. They achieve a best out-of-sample r-squared value of 0.73. (Figure 8) The gross square feet, number of units, year built, easements, and the percentage of people employed in management & business & science & arts have the highest coefficients, which represent the highest importance in the normalized data. (Figure 9) However, the simple multivariate linear regression models show overfitting problems. This is because the smallest eigenvalue is 1.12e-15, indicating the strong multicollinearity. At the same time, *K-Fold cross validations* are applied to measure the overfitting, selection bias, and generalization of the models. The multivariate linear regression models show significant selection bias, which means that the models are not robust enough and their parameters are sensitive to the data. In conclusion, multivariate linear regression models show serious flaws. More models are required to reduce the overfitting problem, multicollinearity problems, and selection bias, and increase the accuracy as well as the robustness.



Dep. Variable:	SALE PRICE	**R-squared:**	0.677
Model:	OLS	**Adj. R-squared:**	0.675
Method:	Least Squares	**F-statistic:**	284.9
Date:	Thu, 09 May 2019	**Prob (F-statistic):**	0.00
Time:	17:25:23	**Log-Likelihood:**	-4985.5
No. Observations:	5746	**AIC:**	1.005e+04
Df Residuals:	5704	**BIC:**	1.033e+04
Df Model:	42		
Covariance Type:	nonrobust		

Omnibus:	14735.300	**Durbin-Watson:**	2.005
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	584933566.013
Skew:	28.389	**Prob(JB):**	0.00
Kurtosis:	1565.030	**Cond. No.**	2.25e+15

Figure 8 the Multivariate Linear Regression Model

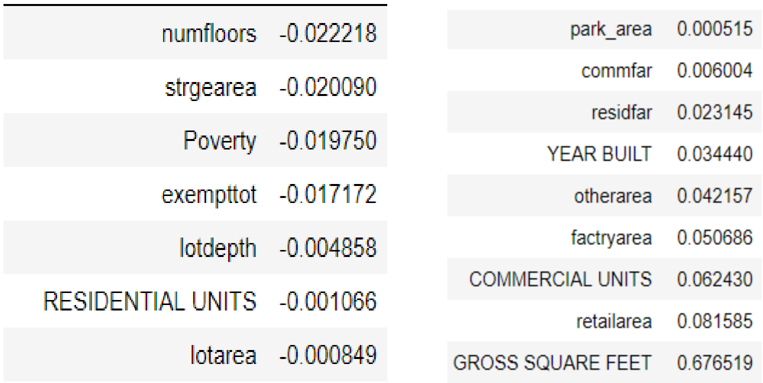


Professional	-3.361726
Service	-2.080741
Office	-1.273551
Production	-0.724649
Construction	-0.469942
lotarea	-0.159595
officearea	-0.122180
LAND SQUARE FEET	-0.092699
Poverty	-0.080421
builtfar	-0.076209

Num of SAT Test Takers	0.050383
otherarea	0.051673
Unemployment	0.055036
YEAR BUILT	0.061193
park_area	0.061676
residfar	0.065434
COMMERCIAL UNITS	0.075997
comarea	0.145177
retailarea	0.162478
GROSS SQUARE FEET	0.843729

Figure 9 Feature Importance of the Multivariate Linear Regression Model

Secondly, *Lasso Regression Models* were built to alleviate the multicollinearity problems**.** In this process, the 10-fold cross validations were applied to tune the optimal alpha parameter. As a result, Lasso regression models eliminate the multicollinearity problems and overfitting problems. 27 features have a coefficient of 0, and only 16 features are left. The gross square feet, area, number of units, year built, number of floors, and the poverty rate have the highest coefficients, which represent the highest importance in the normalized data. (Figure 10) However, the best out-of-sample r-squared value is 0.67. This accuracy is not strong enough for the property price evaluation. In conclusion, Lasso Regression Models can be an alternative model in this case.

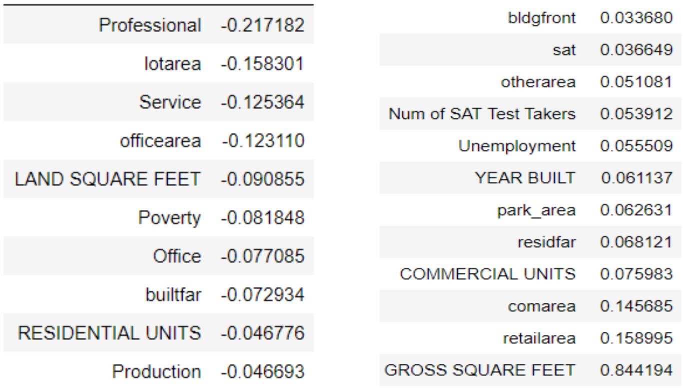


numfloors	-0.022218
strgearea	-0.020090
Poverty	-0.019750
exempttot	-0.017172
lotdepth	-0.004858
RESIDENTIAL UNITS	-0.001066
lotarea	-0.000849

park_area	0.000515
commfar	0.006004
residfar	0.023145
YEAR BUILT	0.034440
otherarea	0.042157
factryarea	0.050686
COMMERCIAL UNITS	0.062430
retailarea	0.081585
GROSS SQUARE FEET	0.676519

Figure 10 Feature Importance of the Lasso Regression Model

Meanwhile, *Ridge Regression Models* are built to alleviate the multicollinearity problems and improve the accuracy**.** Grid Search with 10-fold cross validations were applied to tune the optimal alpha parameter. Ridge regression models eliminate the multicollinearity problems and overfitting problems as well. The gross square feet, area, number of units, year built, and the percentage of people employed in management & business & science & arts have the highest coefficients, which represent the highest importance in the normalized data. (Figure 11) In addition, they achieve a best out of sample r-squared value of 0.73, which is 6% better than the result obtained from the Lasso Regression Models. In conclusion, Ridge Regression Models can be a reliable model in this case.



Professional	-0.217182
lotarea	-0.158301
Service	-0.125364
officearea	-0.123110
LAND SQUARE FEET	-0.090855
Poverty	-0.081848
Office	-0.077085
builtfar	-0.072934
RESIDENTIAL UNITS	-0.046776
Production	-0.046693

bldgfront	0.033680
sat	0.036649
otherarea	0.051081
Num of SAT Test Takers	0.053912
Unemployment	0.055509
YEAR BUILT	0.061137
park_area	0.062631
residfar	0.068121
COMMERCIAL UNITS	0.075983
comarea	0.145685
retailarea	0.158995
GROSS SQUARE FEET	0.844194

Figure 11 Feature Importance of Ridge Regression Model

Extended Models: Decision Tree, Random Forest, AdaBoost, and KNN Models.

Decision Tree, Random forest, AdaBoost, and KNN models were tried but they have poor performance in this case. *Decision Tree models* have the optimal max depth of 5 after tuning hyper-parameters with GridSearchCV. (Figure 12 - 14) But the out of sample accuracy is under 0.7 with serious overfitting problems. *Random forest models* also have an accuracy under 0.7 after tuning hyper-parameters. Additionally, the *AdaBoost models* and the *KNN (K Nearest Neighbors Regression)* models have the optimal out of sample r-squared value under 0.2. (Figure 15 – 18) In conclusion, Decision Tree, Random forest, AdaBoost, and KNN models can be abandoned for further study in this case.

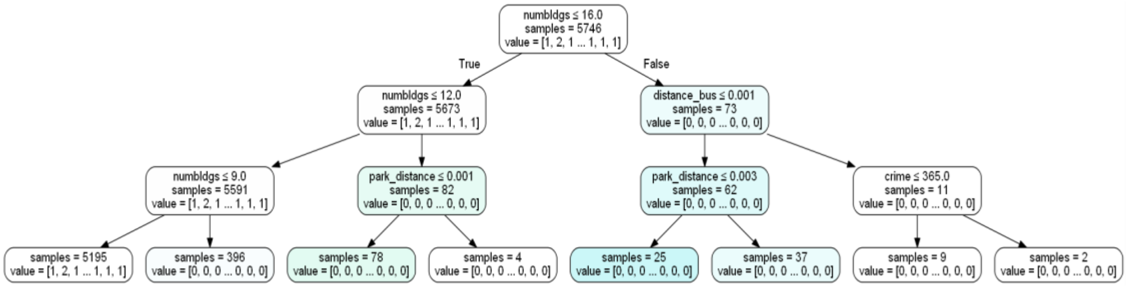
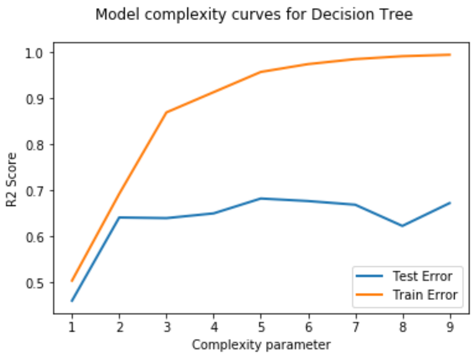

numbldgs ≤ 16.0
samples = 5746
value = [1, 2, 1 ... 1, 1, 1]
True
False
numbldgs ≤ 12.0
samples = 5673
value = [1, 2, 1 ... 1, 1, 1]
distance_bus ≤ 0.001
samples = 73
value = [0, 0, 0 ... 0, 0, 0]
numbldgs ≤ 9.0
samples = 5591
value = [1, 2, 1 ... 1, 1, 1]
park_distance ≤ 0.001
samples = 82
value = [0, 0, 0 ... 0, 0, 0]
park_distance ≤ 0.003
samples = 62
value = [0, 0, 0 ... 0, 0, 0]
crime ≤ 365.0
samples = 11
value = [0, 0, 0 ... 0, 0, 0]
samples = 5195
value = [1, 2, 1 ... 1, 1, 1]
samples = 396
value = [0, 0, 0 ... 0, 0, 0]
samples = 78
value = [0, 0, 0 ... 0, 0, 0]
samples = 4
value = [0, 0, 0 ... 0, 0, 0]
samples = 25
value = [0, 0, 0 ... 0, 0, 0]
samples = 37
value = [0, 0, 0 ... 0, 0, 0]
samples = 9
value = [0, 0, 0 ... 0, 0, 0]
samples = 2
value = [0, 0, 0 ... 0, 0, 0]

Figure 12 Decision tree nodes(3-depth)



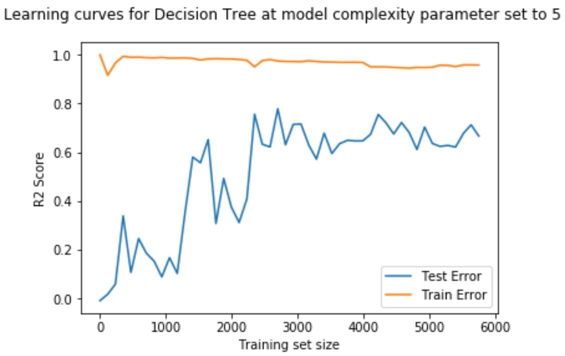
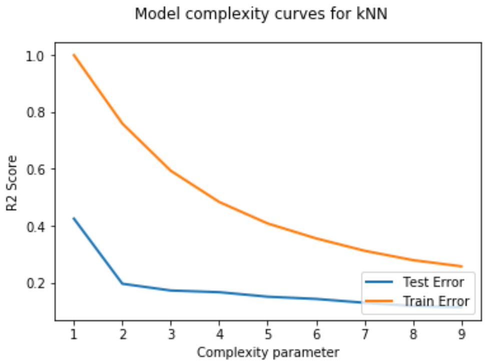


Figure 13 Model Complexity for Decision Tree

Figure 14 Learning Curves for Decision Tree



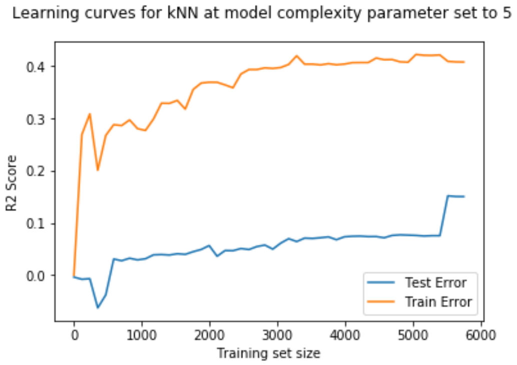
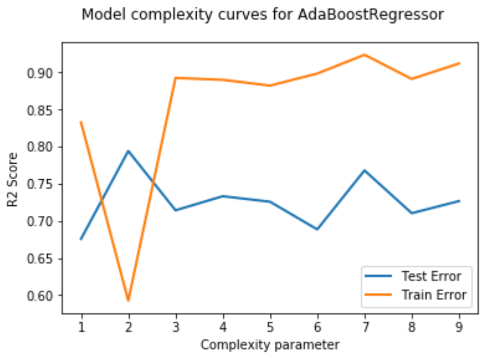


Figure 15 Model Complexity Curves for Decision Tree

Figure 16 Learning Curves for KNN model



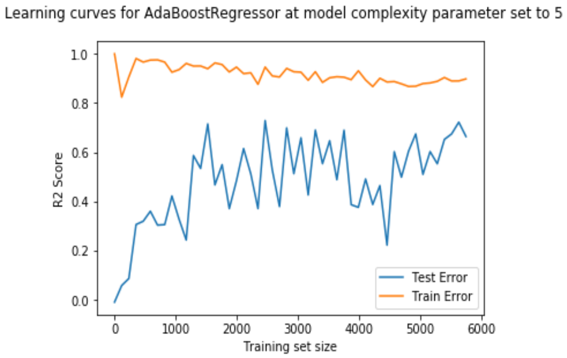


Figure 17 Model Complexity Curves for AdaBoost.

Figure 18 Learning Curves for AdaBoost

Best Model: Gradient Boosting.

The Gradient Boosting performs the best, with an out of sample r-squared value of 0.88 after tuning hyper-parameters with GridSearchCV. We chose a set of optimal parameters including n_estimators of 500, max_depth of 3, and min_samples_split of 10. The Gradient Boosting model makes a prediction with an ensemble of weak prediction models which are usually the decision trees. It builds the model in a stage-by-station fashion, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

However, the in-sample r-squared value is over 99%, indicating an overfitting problem. At the same time, the 10-fold cross validation result of 0.55 shows a selection bias and the weak generalization, which means that this model is not robust enough and its parameters are sensitive to the data. To strengthen the final Gradient Boosting model, we did anomaly detection to filter the outliers, such as the fake transaction. After the *IsoForest Anomaly Detection* and outlier filter, the model sacrifices 5.6% accuracy of the out of sample r-squared, but improve the robustness by 11.0%. Finally, the Gradient Boosting model achieves an out of sample r-squared value of 0.83 with a 10-fold cross validation r-squared value of 0.66.

In conclusion, the Gradient Boosting model is the most reliable model in terms of accuracy and robustness in this case.

Results

First of all, the gross square feet, area, number of units, and the age of the building are the most critical features for the property price evaluation. That is, the attributes of the property itself are more crucial than its location factors. Secondly, as for the preprocessing, the extremum filter, anomaly detection, feature selection, and standardization are necessary and they significantly improve the model performance. Thirdly, the Gradient Boost Model is the best model in the balance of accuracy and robustness, achieving an out of sample r-squared 88%.

Discussion

Bias and Limitation

Despite relatively good performance of the model, the model suffers several bias and limitations. First of all, the model analyzed on limited Data Size. The relatively small dataset contains only the sales of properties between April 2018 and March 2019 and has only 8,885 records. The lack of sufficient data records is assumed to be severely hindering the performance of models namely the Random Forest model. Secondly, the model did not produce an error margin. Finally, Gradient Boost model has an overfitting problem which cannot be solved by tuning hyperparameters.

Future Work

Several improvements can be considered and implemented to the model in the future in order to further improve the performance. For example, fine-tuning the hyperparameters and refining the anomaly detections will be able improve the accuracy and robustness of the model. By taking more sales records with a longer time span and different types of models including XGBoost, LightGBM, and Bayesian Networks into consideration and analysis, the models stand a better chance at increasing accuracy and improving stability, as well as conducting temporal prediction. Furthermore, the existing model has the potential for a website application. That is, an interactive website of inquiring the predicted property sale price at a particular location given the property and neighborhood characteristics can be constructed. The aforementioned future improvements would potentially improve the model performance and strongly improve the model.

Bibliography

Crompton, J. L. (2017). The Impact of Parks on Property Values: A Review of the Empirical Evidence. *Journal of Leisure Research* , 1-31.

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

Lewis-Workman, S., & Brod, D. (1997). Measuring the Neighborhood Benefits of Rail Transit Accessibility. *TRANSPORTATION RESEARCH RECORD* , 147-153.

National Institute of Standards and Technology. (2012, April). *1.3.5.14. Anderson-Darling Test* . Retrieved from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

National Institute of Standards and Technology. (2012, April). *1.3.5.16. Kolmogorov-Smirnov Goodness-of-Fit Test*. Retrieved from Engineering Statistics Handbook: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35g.htm

R, V. (2018, September 11). *Feature selection—Correlation and P-value*. Retrieved from Towards Data Science: https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

Rao, K. (2014, March 11). *Zestimate Forecast Methodology*. Retrieved from Zillow Research: https://www.zillow.com/research/zestimate-forecast-methodology/

Tables

Table 1

Data Sources

Category	Dataset Names	Source URL
Building	- Building Footprint	- https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh
Transportation	- MTA Subway Entrances - MTA Bus Stops	- https://data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56 - https://transitfeeds.com/p/mta
Census Data	- Employment - Population/ Population density - Household Income (Poverty Rate)	- https://popfactfinder.planning.nyc.gov/#12.25/40.724/-73.9868 - https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml
Security	- NYPD Crime Data	- https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243/data
Education	- SAT score per School District	- https://data.cityofnewyork.us/Education/2012-SAT-Results/f9bf-2cp4
Commercial & Amenity	- Park	- https://data.cityofnewyork.us/City-Government/Parks-Properties/k2ya-ucmv
Unit Sale Data	- Rolling Sales Data	- https://www1.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

Group Member Contacts and Contribution

Junjie Cai, jc9033@nyu.edu

- Data Collection: Census, BBL, Parks, Crimes, Education
- Preprocessing: Gaussian Distribution Test, PCA,
- Modeling: Linear Models, AdaBoost, Gradient Boost
- Report Writing: Modeling

Jianwei Li, jl9200@nyu.edu

- Data Collection: Sales
- Preprocessing: Feature Selection, Anomaly Detection
- Modeling: Lasso & Ridge, KNN
- Report Writing: Preprocessing

Wenjie Zheng, wz1405@nyu.edu

- Data Collection: Transportation
- Preprocessing: Standardization, Train Test Split
- Modeling: Decision Tree, Random Forest
- Report Writing: Literature Review

Github Link

https://github.com/JunjieTsai/MLC2019_Project