NYC Proprety Sales data with R: Model for predicting sale price.

Dodema BITENIWE

04 octobre, 2024

Contents

1 Project overview

This project is part of Data Science professional certification pathway offered by HarvardX on edx. It is the second project in the final course of the program. In this project, we propose to analyze data relating to real estate properties in New York City, and to build a powerful model for predicting real estate sales prices in this city, based on the characteristics of these properties.

The data analyzed in this project comes from the (kaggle site, a site on which you can find various data and sizes accessible to the public for data science training. The data from this project is named NYC Property Sales on the site, and contains records on every construction or fraction of a construction (apartment, etc.) in the New York City real estate market over the last twelve months. The data contains information on the location, address, type, sale price and sale date of each building unit.

In this report, we will start with an organization of the data, followed by an exploratory analysis of the data, then the construction of the model and presentation of the results, and finally the conclusion.

2 Data processing and organization

The data was downloaded through this link and then cleaned. The first step was to format the data appropriately and remove any columns or variables that were not relevant to the purpose of the project. In the second stage, we processed the missing data to make it more compact. The final processing step involved removing outliers from the data.

Table 1 shows the first lines of some relevant columns of data. The presence of missing data is easy to spot. We start by deleting rows where the selling price is negative or zero. A zero or extremely low sale price characterizes sales that are in fact transfers of ownership between parties: for example, parents transferring ownership of their home to a child after moving to retire.
Also, for the building_age variable, we require it to be non-zero and non-negative, otherwise it would be counted as missing data.

Table 1: first lines of some relevant columns of data

BOROUGH	BUILDING CLASS CATEGORY	BLOCK	LOT	TOTAL UNITS	LAND SQUARE FEET	GROSS SQUARE FEET	SALE PRICE	building_age
Manhattan	07 RENTALS - WALKUP APARTMENTS	392	6	5	1633	6440	6625000	117
Manhattan	07 RENTALS - WALKUP APARTMENTS	399	26	31	4616	18690	NA	116
Manhattan	07 RENTALS - WALKUP APARTMENTS	399	39	17	2212	7803	NA	116
Manhattan	07 RENTALS - WALKUP APARTMENTS	402	21	10	2272	6794	3936272	103
Manhattan	07 RENTALS - WALKUP APARTMENTS	404	55	6	2369	4615	8000000	116
Manhattan	07 RENTALS - WALKUP APARTMENTS	405	16	20	2581	9730	NA	117
Manhattan	07 RENTALS - WALKUP APARTMENTS	406	32	8	1750	4226	3192840	96
Manhattan	07 RENTALS - WALKUP APARTMENTS	407	18	46	5163	21007	NA	117
Manhattan	08 RENTALS - ELEVATOR APARTMENTS	379	34	15	1534	9198	NA	97
Manhattan	08 RENTALS - ELEVATOR APARTMENTS	387	153	24	4489	18523	16232000	96

After these two operations, Table 2 presents the percentage of missing data for each variable. We note the high percentage of missing data for the LAND SQUARE FEET and GROSS SQUARE FEET variables. We could, if we wished, use GROSS SQUARE FEET to predict some of the missing LAND SQUARE FEET and vice versa, but this would reduce the percentage only slightly. We therefore propose to delete these incomplete data rows and continue the analysis with data without missing elements.

Table 2: Percentage of missing values by variable.

	NA_Percentage
BOROUGH	0.000000
NEIGHBORHOOD	0.000000
BUILDING CLASS CATEGORY	0.000000
TAX CLASS AT PRESENT	1.009241
BLOCK	0.000000
LOT	0.000000
RESIDENTIAL UNITS	0.000000
COMMERCIAL UNITS	0.000000
TOTAL UNITS	0.000000
LAND SQUARE FEET	35.817009
GROSS SQUARE FEET	36.734347
YEAR BUILT	0.000000
TAX CLASS AT TIME OF SALE	0.000000
BUILDING CLASS AT TIME OF SALE	0.000000
SALE PRICE	0.000000
sale_year	0.000000
sale_month	0.000000
building_age	0.000000

The final cleansing step involved removing any outliers from the data, especially those that were more than 3 standard deviations away from the mean of the distribution. This processing involved both the target variable (SALE PRICE) and other variables such as LAND SQUARE FEET and GROSS SQUARE FEET.

Table 2 displays the progression of the correlation with the target variable after each cleaning stage. We notice a very significant evolution in data quality with predictors that are better correlated with the target variable.

Table 3: Correlation with target variable after each cleaning step.

	Variable	Corr_Default	Corr_After_Missing	Corr_After_Outliers
BLOCK	BLOCK	-0.05146	-0.06530	-0.25466
LOT	LOT	-0.01443	-0.01031	-0.04579
RESIDENTIAL UNITS	RESIDENTIAL UNITS	0.12270	0.14637	0.49386
COMMERCIAL UNITS	COMMERCIAL UNITS	0.04722	0.04620	0.31635
TOTAL UNITS	TOTAL UNITS	0.12821	0.14447	0.53572
LAND SQUARE FEET	LAND SQUARE FEET	0.04163	0.04619	0.16333
GROSS SQUARE FEET	GROSS SQUARE FEET	0.45534	0.52919	0.59587
YEAR BUILT	YEAR BUILT	0.00767	0.00144	-0.15627
SALE PRICE	SALE PRICE	1.00000	1.00000	1.00000
sale_year	sale_year	-0.00255	-0.00336	0.00288
building_age	building_age	-0.00767	-0.00149	0.15628

3 Exploratory data analysis (EDA)

In this section, we propose to extend our understanding of the data through an exploratory analysis.We'd like to mention two references (Ermis (2021) and Irizarry (n.d.)) that have inspired us in the following analysis.

3.1 Sale Price Distribution with Density, Mean, and Median

Figure 1 illustrates the distribution of the target variable. We note a distribution that deviates from the normal distribution due to a slightly longer tail on the right, a source of asymmetry as indicated by the mean and median axes.However, for the rest of the analysis, we decided not to transform the data, as several algorithms will be trained and some are robust enough to take this problem into account.

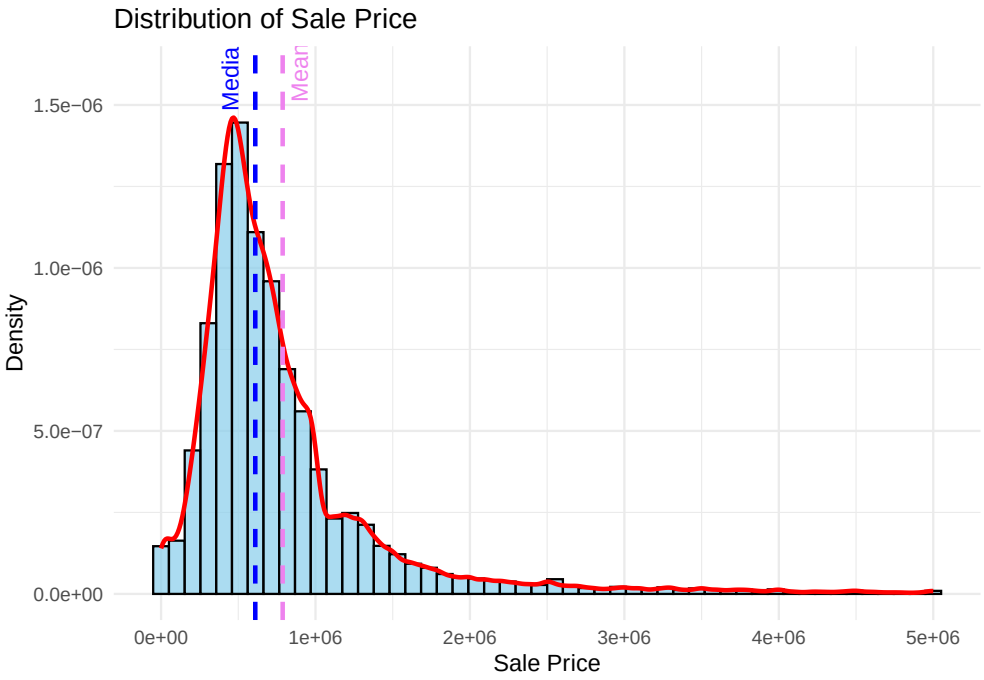

Figure 1: Sale Price Distribution with Density

3.2 Box Plot of Sale Price by Borough

Figure 2 shows the box plot of sales prices by borough. It can be seen that there is a clear difference in property prices between the different boroughs. Manhattan has the most expensive real estate, with heterogeneous sales prices.The Bronx and Staten Island have the lowest and most homogeneous prices.

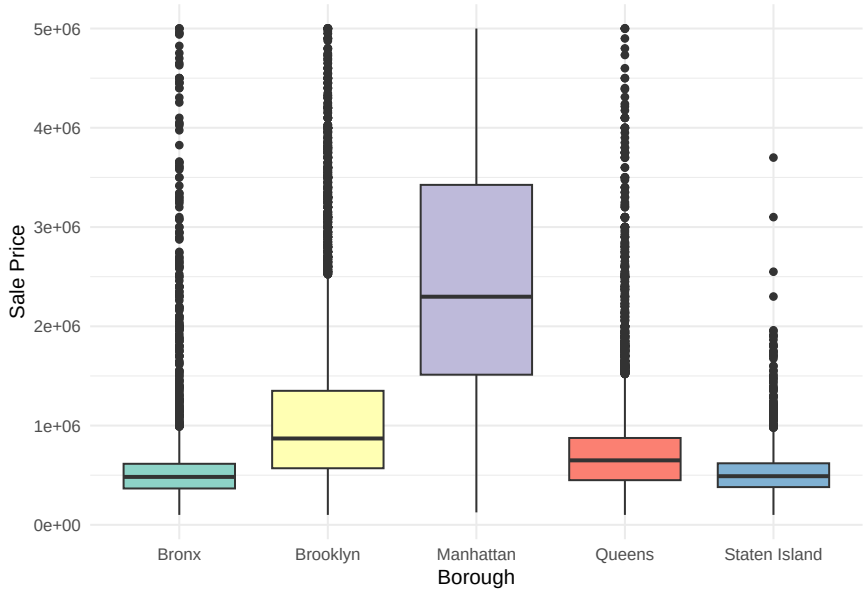

Figure 2: Box Plot of Sale Price by Borough

3.3 Box Plot of Sale Price by Month

Figure 3 shows the box plot of sales prices by month of the year. It can be seen that there is no great difference in property prices compared with the different months of the year. Prices are relatively homogeneous for each month.

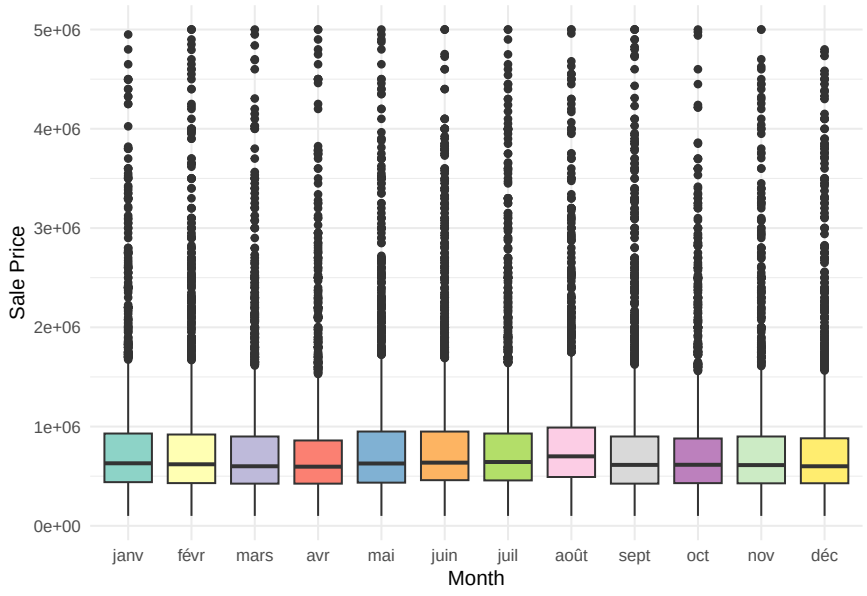

Figure 3: Box Plot of Sale Price by Month

3.4 Sale Count by Month

Figure 4 displays the number of sales by month of the year. It can be seen that there is no great difference in the number of properties sold compared to the different months of the year. Sales are fairly homogeneous over the year. Thus, there is no pronounced seasonality in the target variable, and the sales_month variable provides us with very little information. In the rest of the analysis, it will be removed.

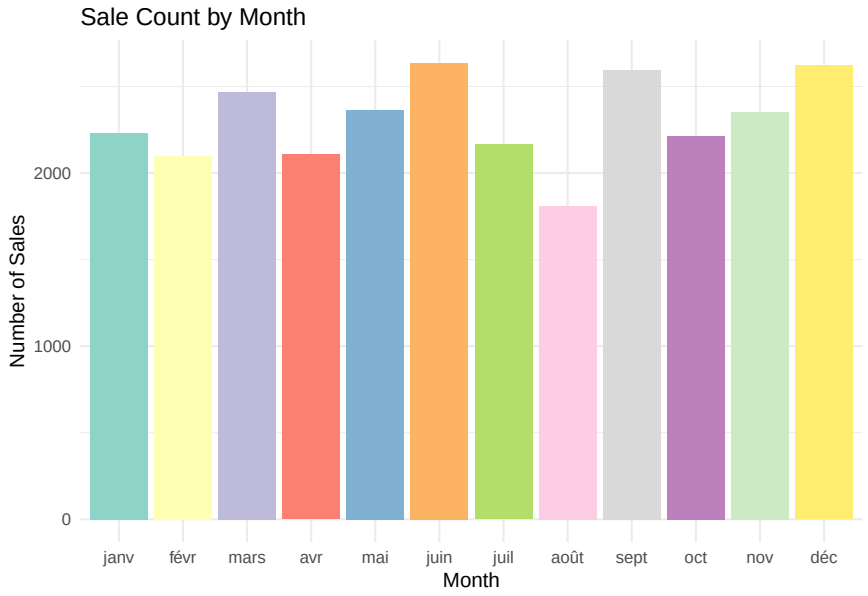

Figure 4: Box Plot of Sale Price by Month

3.5 Average Land Square Feet by Borough

Regarding the Land Square Feet of the buildings sold, figure 5 gives us the distribution according to the borough. We can note that the buildings sold on Staten Island, Queens and Bronx have on average the largest Land Square Feet, ranging from 3000 to 4000.

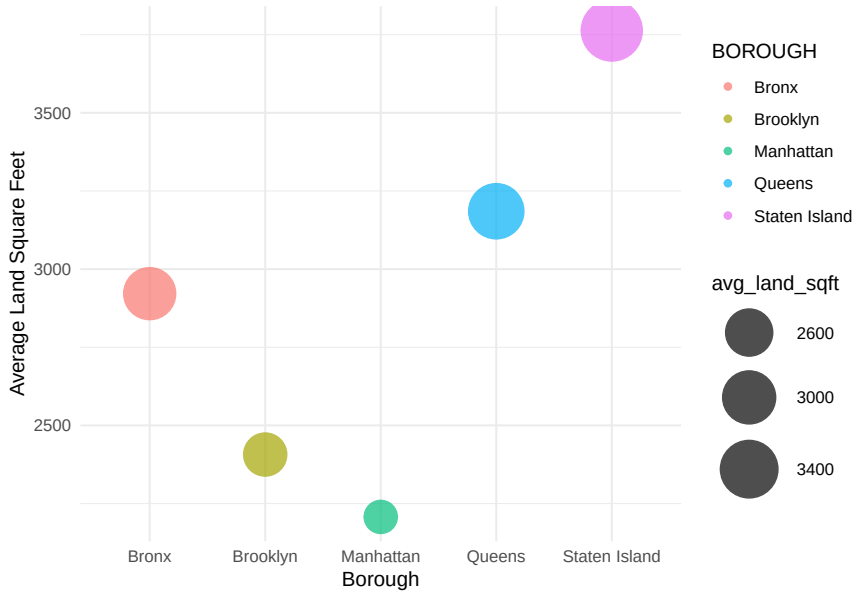

Figure 5: Average Land Square Feet by Borough

3.6 Bar Plot for House Price by Top 10 Neighborhoods

Considering the 10 most dynamic neighborhoods in terms of real estate sales, figure 6 shows us how the average sale price varies in these neighborhoods. We see that prices are high in the neighborhoods of BEDFORD STUYVESANT,FLUSHING-NORTH and BAYSIDE.

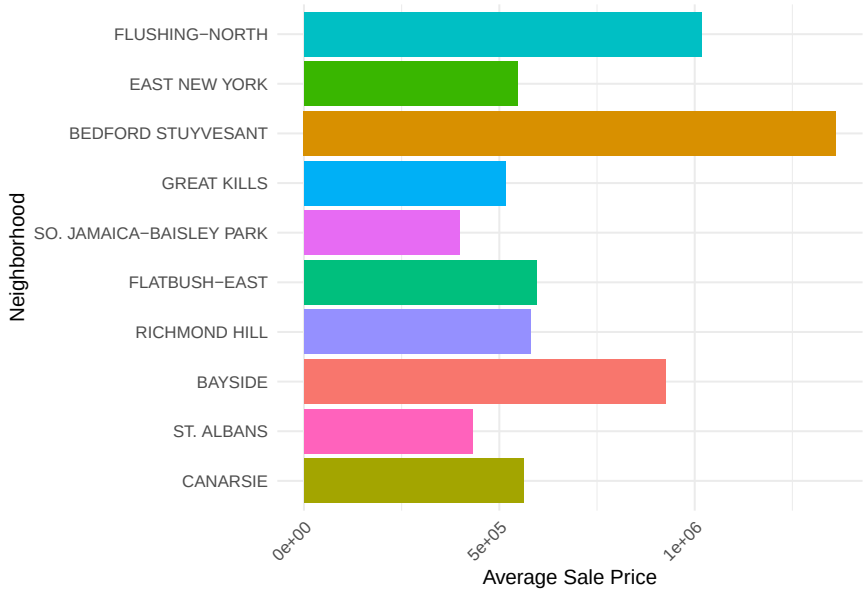

Figure 6: Top 10 Neighborhoods by Number of Sales and Average Sale Price

3.7 Correlation Matrix Heatmap

To address the problem of multicollinearity among the predictors we have used the figure 7. It is evident from inspection of the figure that the variables YEAR BUILT,TOTAL UNITS and RESIDENTIAL UNITS are highly correlated with other variables in the projected model. We therefore decide to remove these variables. Also, the variables LOT, sale_year, building_age, tax_class_at_present, neighborhood, building_class_at_time_of_sale, sale_month do not seem relevant to us for the rest of the analysis because they are weakly correlated with the target variable for some or duplicate information contained in other variables for others. These variables will therefore be removed.

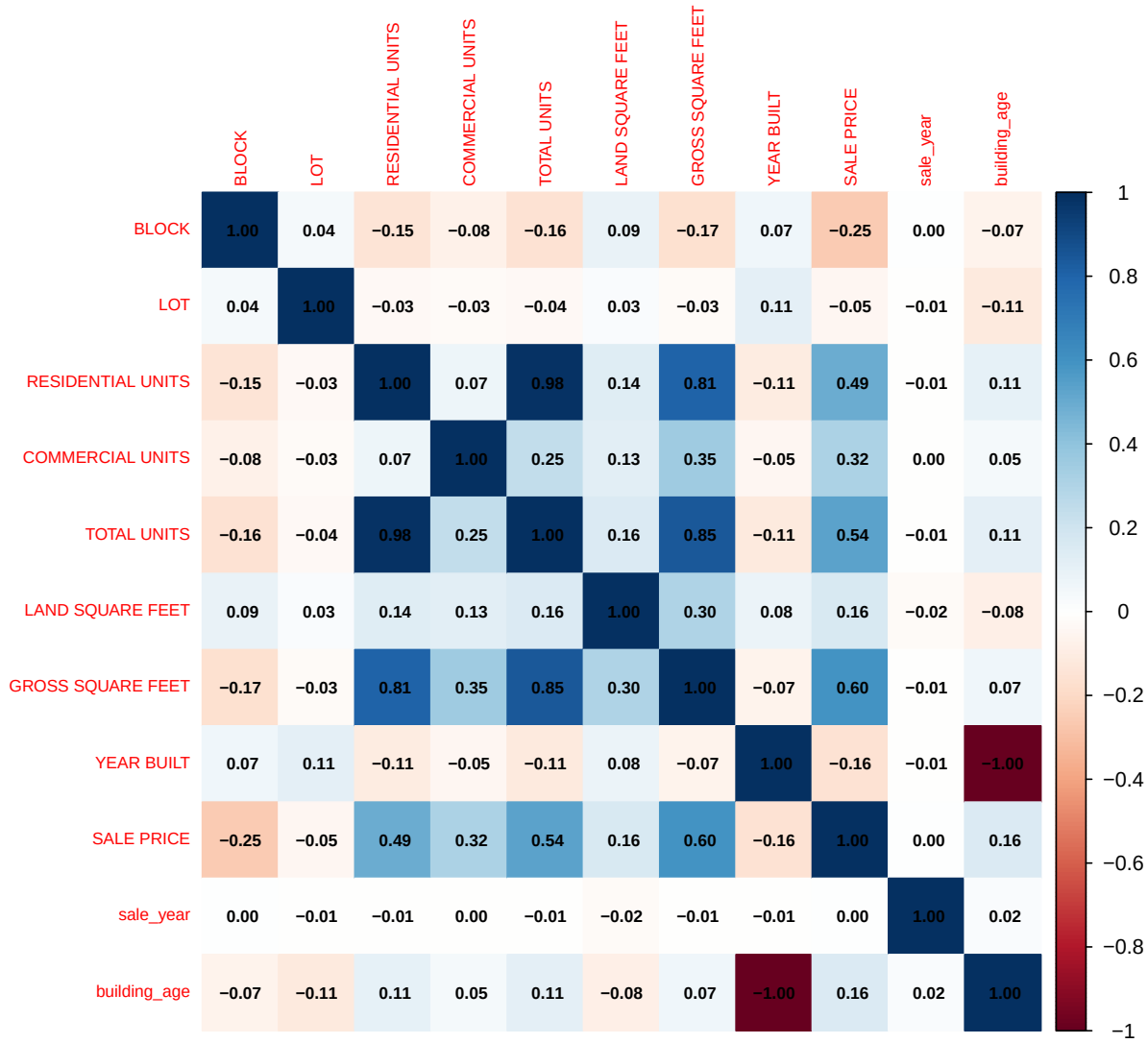

Figure 7: Correlation Matrix Heatmap

4 Machine Learning Model Development

A statistical summary of the variables included in the development of the machine learning model is given below.

```
##           borough                         building_class_category
##  Bronx        : 3333   01 ONE FAMILY DWELLINGS       :12534
##  Brooklyn     : 8238   02 TWO FAMILY DWELLINGS       : 9798
##  Manhattan    :  610   03 THREE FAMILY DWELLINGS     : 2299
##  Queens       :10642   07 RENTALS - WALKUP APARTMENTS: 1639
##  Staten Island: 4843   22 STORE BUILDINGS            :  411
##                        14 RENTALS - 4-10 UNIT        :  318
##                        (Other)                       :  667
##      block       commercial_units  land_square_feet gross_square_feet
##  Min.   :    5   Min.   : 0.0000   Min.   :  200    Min.   :  120
##  1st Qu.: 2834   1st Qu.: 0.0000   1st Qu.: 2000    1st Qu.: 1350
##  Median : 4985   Median : 0.0000   Median : 2500    Median : 1836
##  Mean   : 5643   Mean   : 0.1198   Mean   : 3038    Mean   : 2437
##  3rd Qu.: 7898   3rd Qu.: 0.0000   3rd Qu.: 3800    3rd Qu.: 2560
##  Max.   :16319   Max.   :32.0000   Max.   :14608    Max.   :72781
##
##  tax_class_at_time_of_sale   sale_price
##  1:24651                   Min.   :     200
##  2: 2026                   1st Qu.:  430000
##  3:    0                   Median :  625000
##  4:  989                   Mean   :  918915
##                            3rd Qu.:  940000
##                            Max.   :11435000
##
```

We then proceed to format the data for modeling. First, we convert categorical variables into dichotomous variables using a process known as One Hot Encoding.We then partition the data into two parts (80/20 split), one for training (Train_data) and the other for testing (Test_data).

4.1 Model Selection

Five of the most common algorithms are trained and compared in performance on the basis of RSME. These are

- "lm": Linear Regression
- "ranger": Random Forest, a faster version
- "svmRadial: Support Vector Machines with Radial Basis Function Kernel
- "gbm": Stochastic Gradient Boosting
- "xgbTree": eXtreme Gradient Boosting

Mathematically, mean absolute error (MAE) is defined by :

$$MAE = \frac{1}{N}\sum_{i} \|y_i - \hat{y}_i\| \tag{1}$$

and root mean square error (RMSE) by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i}(y_i - \hat{y}_i)^2} \tag{2}$$

We define y_i as the price of property i sold. and denote our prediction with $\hat{y}_i$.

The ranger (Random Forest) model is the best-performing of all the trained models, based on the RMSEs shown in table 4. Based on the test data, we reach an RMSE of **582403**. We then select this model and optimize it using the model parameters.

Table 4: Performance of each algorithm based on different metrics.

Model	Train_Score	Test_Score	Train_RMSE	Test_RMSE	Train_MAE	Test_MAE	Execution_Time_Secs
lm	0.57400	0.58615	758781.3	723469.7	368659.6	363440.6	3.08 secs
ranger	0.92696	0.73386	325023.5	582403.0	149837.9	264432.7	32.60 secs
svmRadial	0.69385	0.60808	654247.0	707522.8	296685.7	317904.4	143.13 secs
gbm	0.65966	0.61826	680503.6	695217.8	322631.6	331347.9	9.70 secs
xgbTree	0.80031	0.70993	522391.3	605993.2	259271.3	284719.3	63.35 secs

5 Results

Our analysis shows that the Ranger Random Forest algorithm is the best performing of the 5 algorithms trained on NYC Property Sales data. Performance in terms of RMSE on the test data is **582031**.

Figure 8 shows the importance or contribution of each variable in the model in explaining the target variable (sale price). We note that the variables gross_square_feet, block, tax_class_at_time_of_sale2 and land_square_feet alone explain 78% of the variability in the selling price of real estate properties.

Table 5: Performance of the final algorithm.

Model	Test_Score	Test_RMSE	Test_MAE
Ranger Random Forest model	0.7341	582031	264572

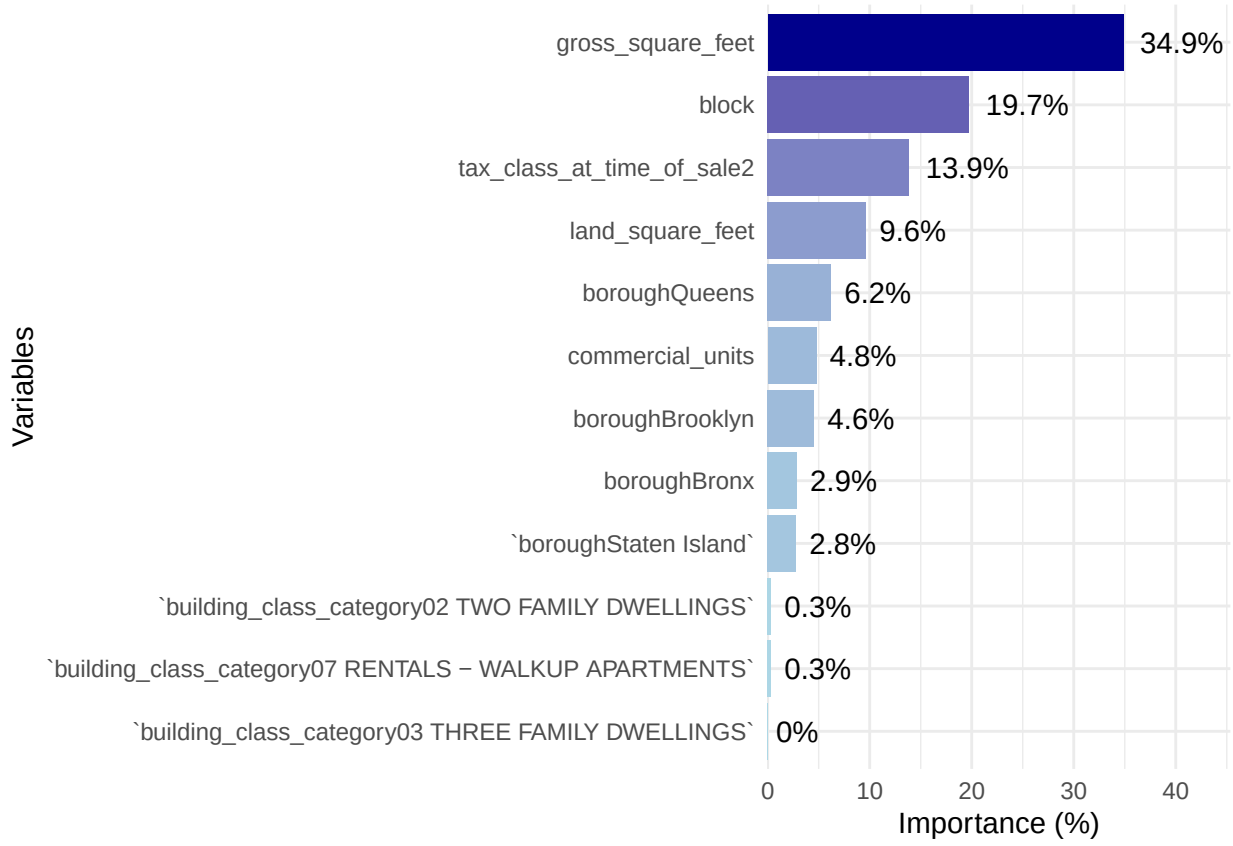

Figure 8: Variable Importance (Random Forest)

6 Conclusion

In this study, we explored the NYC Property Sales data. First, we used various visualizations to understand the data. Then 5 algorithms were chosen and trained on part of the data. The analysis revealed that the Ranger Random Forest model is the best model, with an RMSE performance of **582031**. This model will therefore be ideal for predicting sale price in the New York City property market.

References

Ermis, Mustafa Batuhan. 2021. "End-to-End Machine Learning Regression Project." In. https://www.kaggle.com/code/ermismbatuhan/end-to-end-machine-learning-regression-project.

Irizarry, Rafael A. n.d. *Introduction to Data Science*. https://rafalab.github.io/dsbook/: HarvardX.