Using Interest Rates To Predict Commercial Real Estate Prices with Random Forest, Linear Regression and Extreme Gradient Boosting Machine Learning Models.

Abstract:

Commercial real estate is a crucial component of the economy, and it can be significantly impacted by interest rates. As interest rates increase, the cost of borrowing money for owners and investors rises, thus it lowers the capacities and number of investors to invest in commercial real estate. In this paper we are trying to investigate the relationship between commercial real estate market prices and interest rates within Manhattan. In addition to that we are trying to predict the prices using such features as year built, neighborhood, etc. but most importantly interest rates. The research was built on 27432 commercial units in Manhattan that were sold in 2017 to 2022. In this paper we are using several regression models to predict the price per sqft and data science techniques to visualize, clean and understand the data we collected. To conduct this paper we are using NYC Open Data, Interest rates from FRED economic data as well as other resembling most updated research papers conducted within 3-5 years.
The source and code can be found in my github:
https://github.com/aidskiy/Research2023IR

1. Introduction:

Background and context for the research. - Covid, Interest Rates, Concerns
Up until 2023, the Manhattan commercial real estate market had experienced various trends and shifts in pricing since 2016. Early in that period, there were indications of a

softening market, with slower price growth and increased availability of office spaces. This was partly due to concerns about oversupply and changing trends in how businesses utilize office spaces, as well as economic factors. Then, in 2020, the COVID-19 pandemic had a significant impact on the commercial real estate market globally, including Manhattan. Lockdowns, remote work arrangements, and economic uncertainties resulted in a decline in demand for office spaces, leading to increased vacancy rates and decreased rental prices in some areas. In 2022 Federal Reserve raised the benchmark interest rate by half a percentage point and elevated the targeted rate to 4.25% and 4.5%, marking the highest level seen in 15 years. Higher interest rates mean higher borrowing costs. Consequently, this makes it more expensive to borrow funds for commercial real estate development. Higher borrowing costs drive down demand from businesses and investors, leading to fewer new projects and a slower transaction volume.

Problem statement or research question. - Can the interest rates affect the prediction
In this paper, we are trying to find out how interest rates are influencing commercial real estate market prices and identify the difference of accuracy in predicting the price with the interest rate and without by running regression models with multiple features. Specifically we are analyzing 27432 sales that were made from 2017 to 2022 in Manhattan, NY. The paper will analyze NYC dataset and will predict real estate market price using various regression models such as linear regression, gradient boosting and random forest.

2. Related Work:

To answer our question and find the evidence of interest rates influencing the market we are using NYC open data “NYC Citywide Annualized Calendar Sales Update” that was updated July 4th 2023 and the Federal Reserve Bank of St. Louis' “FRED” Database for interest rates dataset. In addition to that, we used “Manhattan Commercial Market Trends” to illustrate most recent data from 2023 that is still not available on NYC Open Data. To expand our knowledge and bring new ideas we used few other research papers such as “PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction” by Yaping Zhao; Ramgopal Ravi; Shuhui Shi; Zhongrui Wang; Edmund Y. Lam; Jichang Zhao. Their experiments shows that the characteristics from different aspects i.e. amenities, traffic, emotions, have an economic impact on real estate price. Another interesting paper where they used Machine learning in predicting price by using Neural Networks “Applied research on real estate price prediction by the neural network” by Hu Xiaolong; Zhong Ming. Another inspirational paper is called “Evaluating machine learning algorithms for predicting house prices in Saudi Arabia” where Talal Alshammari used random forest, decision tree and linear regression to predict the price.

3. Methodology:

Description of research design and methodology.

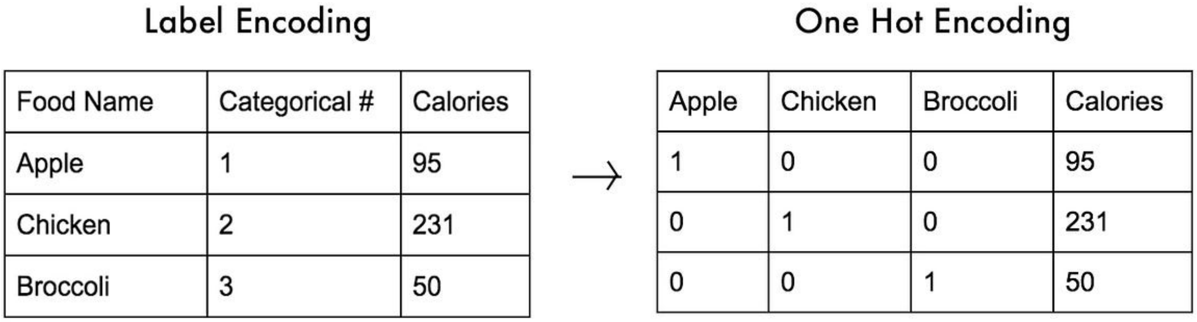
This section outlines the steps and procedures undertaken in this study to achieve the research objectives. The research focuses on investigating the impact of interest rates on commercial real estate market prices in Manhattan and developing a reliable model for real estate prices. In section 3.A we will cover how we collected the data, where and how we chose to preprocess it. In section 3.B we will show some important visualizations through which you can understand the data and see evidence of the objectives of this work. In section 3.C we will go through feature engineering process, this analysis helps to identify which variables, including interest rates, play a significant role in predicting property prices. In addition to that we will go through model development and what models we used and why we chose them.

A. Data Collection - Data Preprocessing

To achieve our goals we chose to collect data from open source provided by NY state - “NYC Open Data”. The dataset includes variables such as Neighborhood, borough, zip code, address, land sqft, year the building was built, sale data and price and other plenty of unnecessary variables. And for additional most important variables, interest rates, we got data from the Federal Reserve Bank of St. Louis’ “FRED” Database. We combined data from 2017-2022 and started deleting columns we will not be using as our features.

Removed instances with NAN values and zeros as it can ruin the results of prediction if we find the mean and replace missing values with them. We removed the outliers by cutting the highest percentile and lowest. Percentiles are statistical measures that divide a dataset into 100 equal parts. The 1st percentile represents the value below which 1% of the data falls, and the 99th percentile represents the value below which 99% of the data falls. This technique helps mitigate the influence of outliers and extreme values, making the data more suitable for analysis and modeling.

Also, most of the models can’t understand strings and prefer floats as variables therefore we had to do one-hot-encoding and replace strings with unique float values. One-hot encoding is used to convert categorical variables into binary vectors. In many machine learning algorithms, categorical data needs to be converted into numerical form.



Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

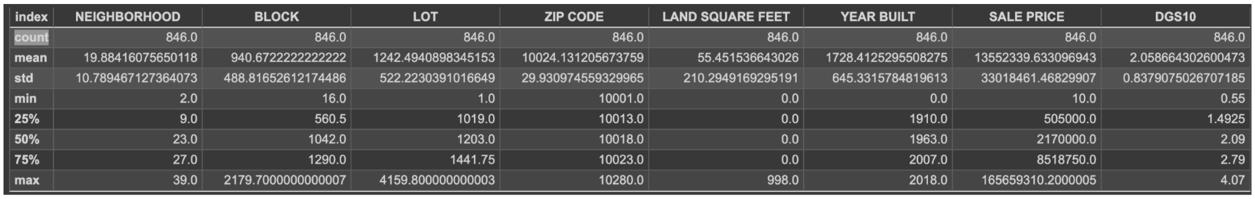
Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

(One-Hot Encoding example)

When mapping is applied, the categorical values are replaced with corresponding numerical values. This is particularly useful when the categorical data has an ordinal relationship, where the categories have a meaningful order.

B. Data visualization

To understand the data we have preprocessed we can start with .describe() function to understand the general information. Due to large amount of randomly missing data that and the importance of some features we can’t simpy delete we decided to delete the number of items and from it 27432 units sold we were left with 846.

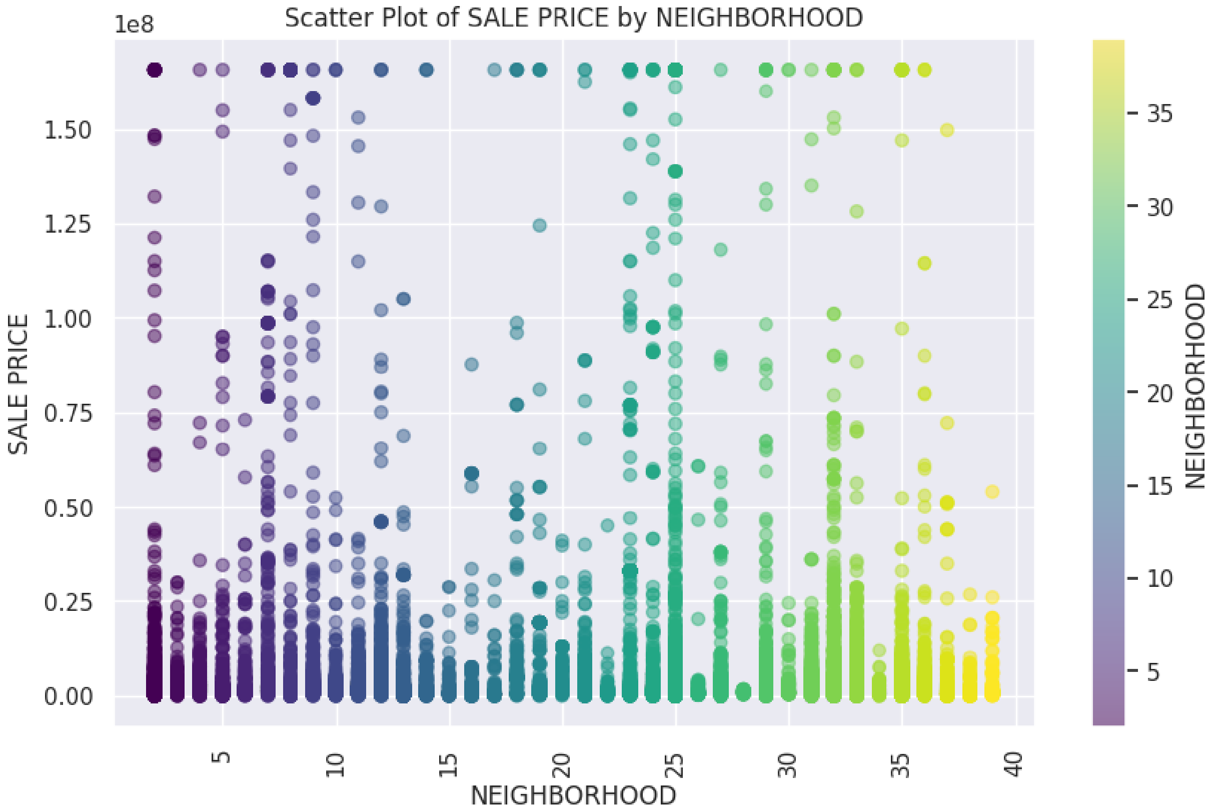


index	NEIGHBORHOOD	BLOCK	LOT	ZIP CODE	LAND SQUARE FEET	YEAR BUILT	SALE PRICE	DGS10
count	846.0	846.0	846.0	846.0	846.0	846.0	846.0	846.0
mean	19.88416075650118	940.6722222222222	1242.4940898345153	10024.131205673759	55.451536643026	1728.4125295508275	13552339.633096943	2.058664302600473
std	10.789467127364073	488.81652612174486	522.2230391016649	29.930974559329965	210.2949169295191	645.3315784819613	33018461.46829907	0.8379075026707185
min	2.0	16.0	1.0	10001.0	0.0	0.0	10.0	0.55
25%	9.0	560.5	1019.0	10013.0	0.0	1910.0	505000.0	1.4925
50%	23.0	1042.0	1203.0	10018.0	0.0	1963.0	2170000.0	2.09
75%	27.0	1290.0	1441.75	10023.0	0.0	2007.0	8518750.0	2.79
max	39.0	2179.7000000000007	4159.800000000003	10280.0	998.0	2018.0	165659310.2000005	4.07

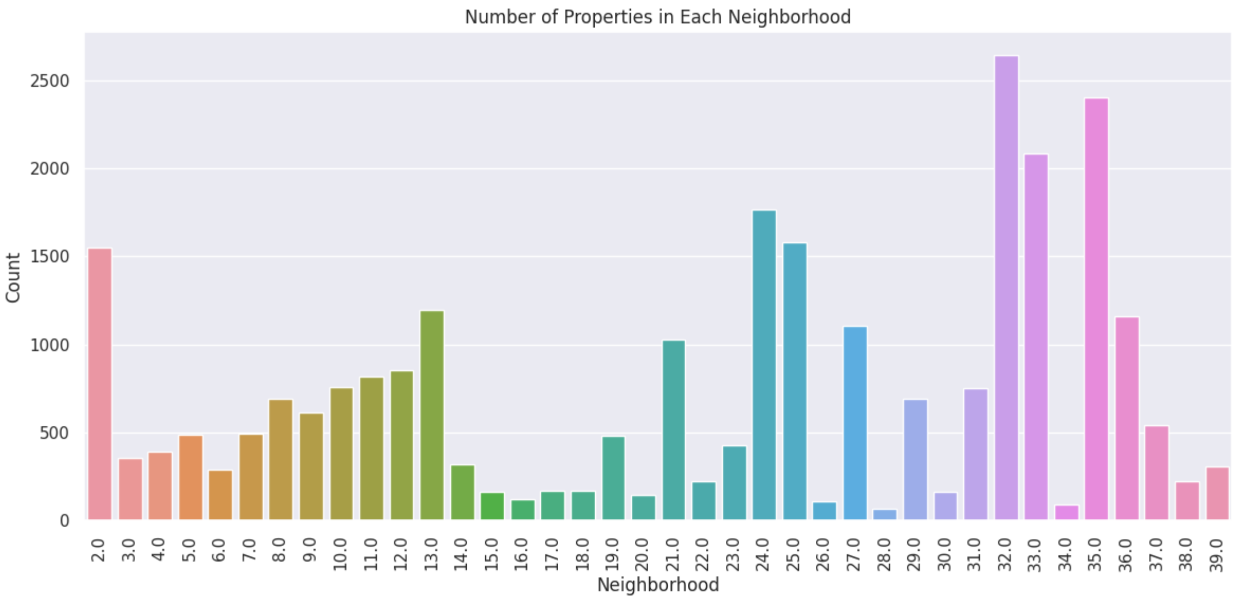
(Data description)

The neighborhood was mapped into numbers - neighborhood_mapping = {

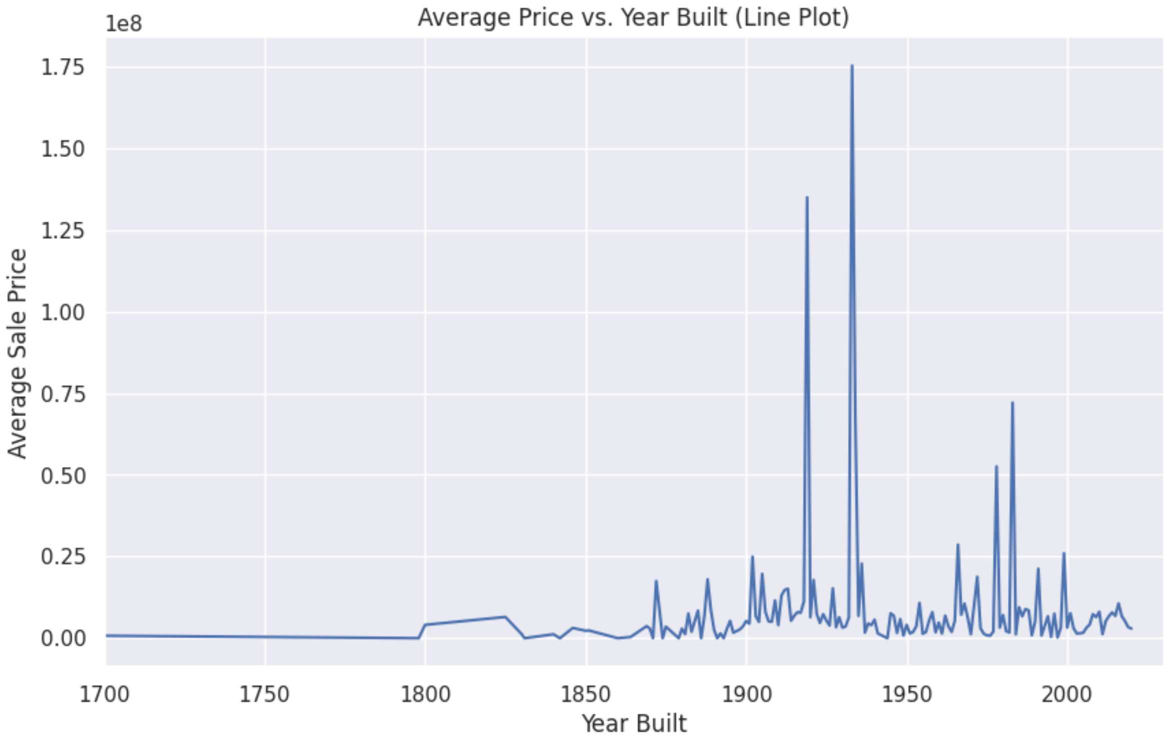
```
     'ALPHABET CITY': 1.0,
   'CHELSEA': 2.0,
   'CHINATOWN': 3.0,
   'CIVIC CENTER': 4.0,
   'CLINTON': 5.0,
   'EAST VILLAGE': 6.0,
   'FASHION': 7.0,
   'FINANCIAL': 8.0,
   'FLATIRON': 9.0,
   'GRAMERCY': 10.0,
   'GREENWICH VILLAGE-CENTRAL': 11.0,
   'GREENWICH VILLAGE-WEST': 12.0,
   'HARLEM-CENTRAL': 13.0,
   'HARLEM-EAST': 14.0,
   'HARLEM-UPPER': 15.0,
   'HARLEM-WEST': 16.0,
   'INWOOD': 17.0,
   'JAVITS CENTER': 18.0,
   'KIPS BAY': 19.0,
   'LITTLE ITALY': 20.0,
   'LOWER EAST SIDE': 21.0,
   'MANHATTAN VALLEY': 22.0,
   'MIDTOWN CBD': 23.0,
   'MIDTOWN EAST': 24.0,
   'MIDTOWN WEST': 25.0,
   'MORNINGSIDE HEIGHTS': 26.0,
   'MURRAY HILL': 27.0,
   'ROOSEVELT ISLAND': 28.0,
   'SOHO': 29.0,
   'SOUTHBRIDGE': 30.0,
   'TRIBECA': 31.0,
   'UPPER EAST SIDE (59-79)': 32.0,
   'UPPER EAST SIDE (79-96)': 33.0,
   'UPPER EAST SIDE (96-110)': 34.0,
   'UPPER WEST SIDE (59-79)': 35.0,
   'UPPER WEST SIDE (79-96)': 36.0,
   'UPPER WEST SIDE (96-116)': 37.0,
   'WASHINGTON HEIGHTS LOWER': 38.0,
   'WASHINGTON HEIGHTS UPPER': 39.0
}
```



The results of scatter plot show dots representing the relationship between 'NEIGHBORHOOD' and 'SALE PRICE'. Each dot's color corresponds to the 'NEIGHBORHOOD' it belongs to, and the position of the dot along the y-axis represents the 'SALE PRICE'. The plot gives you an overview of how 'SALE PRICE' varies across different 'NEIGHBORHOOD' values and if there are any patterns or trends in the data. As we can see the from the plot downtown areas and midtown have a large amount of sales and were the most expensive one at the same time.

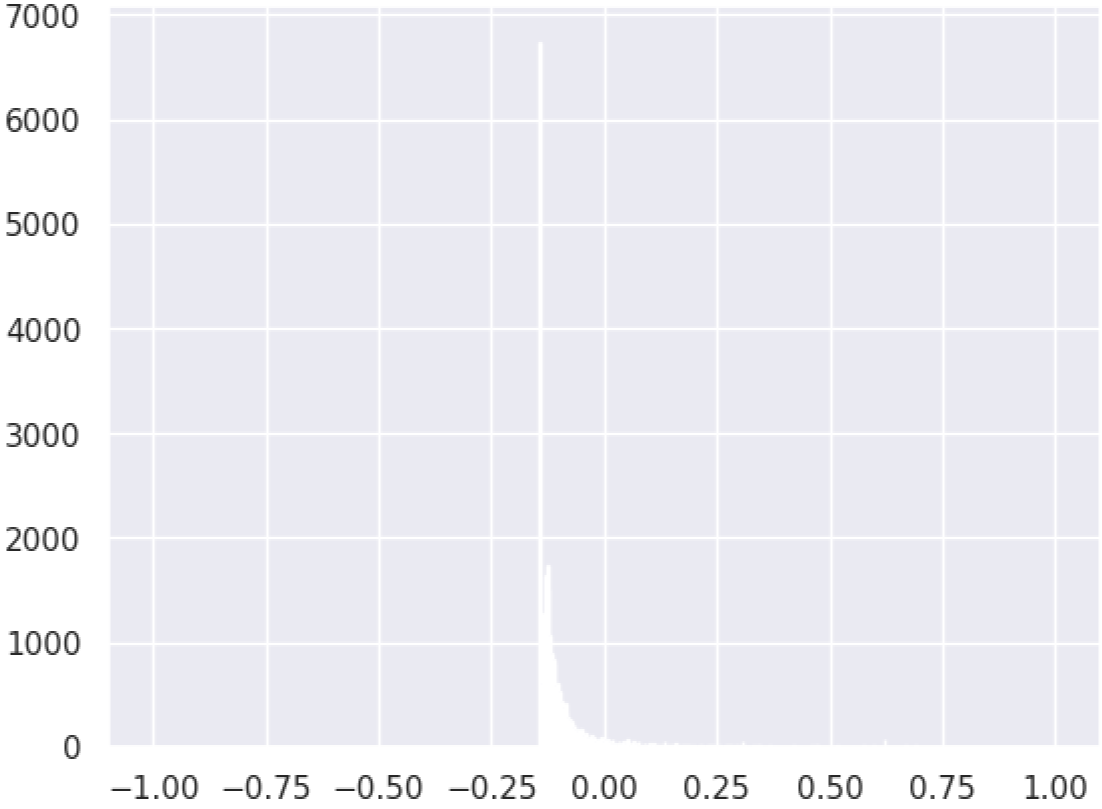


However this plot for Neighborhoods gave more insights on the amount of units were sold and we can clearly see that on the first place it's UPPER EAST SIDE (59-79) then UPPER WEST SIDE (59-79), then UPPER EAST SIDE (79-96). Let's note that it doesn't show the price but it seems like those areas are in demand. The results count plot shows a bar for each unique neighborhood along the x-axis, and the height of each bar represents the number of properties in that neighborhood. This visualization allows you to quickly understand the distribution of properties across different neighborhoods in your dataset.



The results of line plot shows how the average sale price has changed over the years for the given range of 'YEAR BUILT' values. It helped us to observe trends, fluctuations, and potential patterns in the relationship between the average sale price and the year the property was built. We were surprised by the fact that 100 years old building were so expensive, we think this tells us that the year built doesn’t have a significant influence on market price.

For further analysis we decided to find z-score. A z-score, also known as a standard score, is a statistical measure that indicates how many standard deviations a data point is away from the mean of the dataset. It is a way to standardize and compare values that come from different distributions. The formula to calculate the z-score for a data point 'x' in a dataset with mean 'μ' and standard deviation 'σ' is: z =(x-p)/σ. The resulting histogram shows the distribution of z-scores of the 'SALE PRICE' values. This can help you understand how the 'SALE PRICE' values deviate from their mean in terms of standard deviations. If the distribution is centered around 0 and follows a normal distribution, most values will be within the range of -1 to 1. However, in this case the distribution has a long right tail, it might indicate the presence of outliers or a heavy-tailed distribution. This helped us to identify the outliers in the dataset before we processed it.



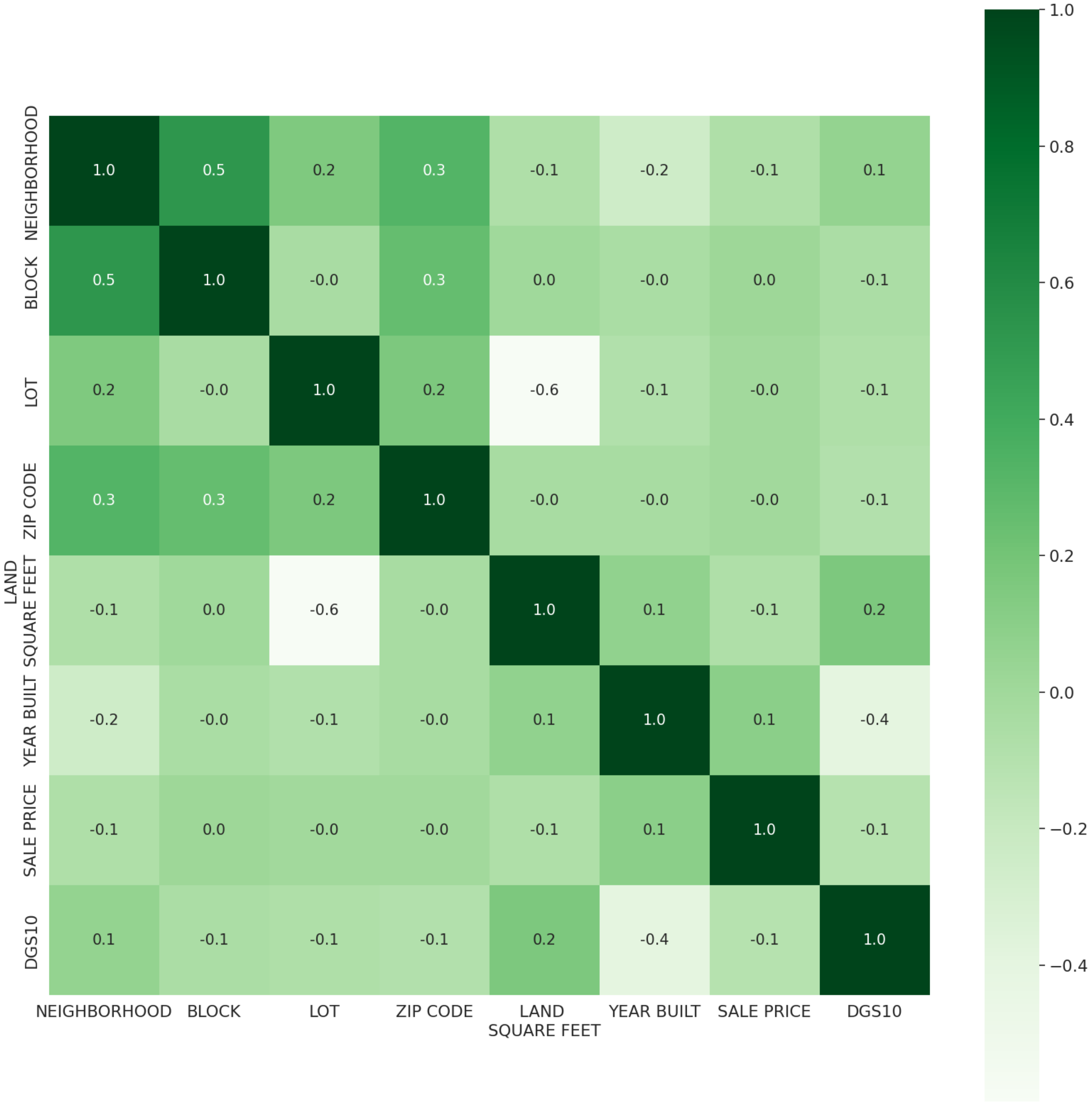
(z-score of the dataset)

C. Feature Selection

Now, covariance itself is harder to work with because the magnitude is unbounded and can be arbitrarily large depending on your data. To make covariance easier to work with, we introduce the idea of Pearson correlation, which standardizes the range of value for covariance to always be between -1 and 1. This is calculated by dividing the covariance by the product of the standard deviations of the two variables. The reason we choose Pearson is that it is the most widely used correlation formula.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

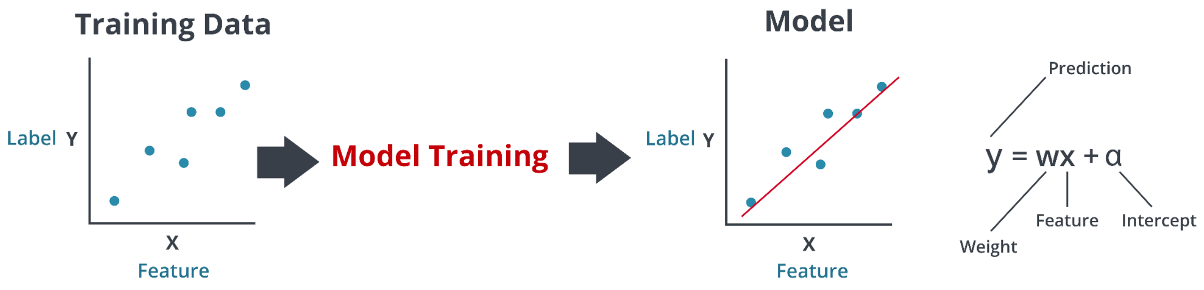
The resulting heatmap visualizes the correlation between different numeric features in the DataFrame. Each cell of the heatmap represents the correlation coefficient between two features. The color and intensity of the cell indicate the strength and direction of the correlation. Positive correlations are shown in green shades, while negative correlations are shown in cooler colors. A positive correlation between two variables means that as one variable increases, the other variable tends to increase as well. A negative correlation between two variables means that as one variable increases, the other variable tends to decrease. This visualization helps us to understand the relationships between variables and identify potential patterns or dependencies in the data.



(Heatmap of Feature Relevance)

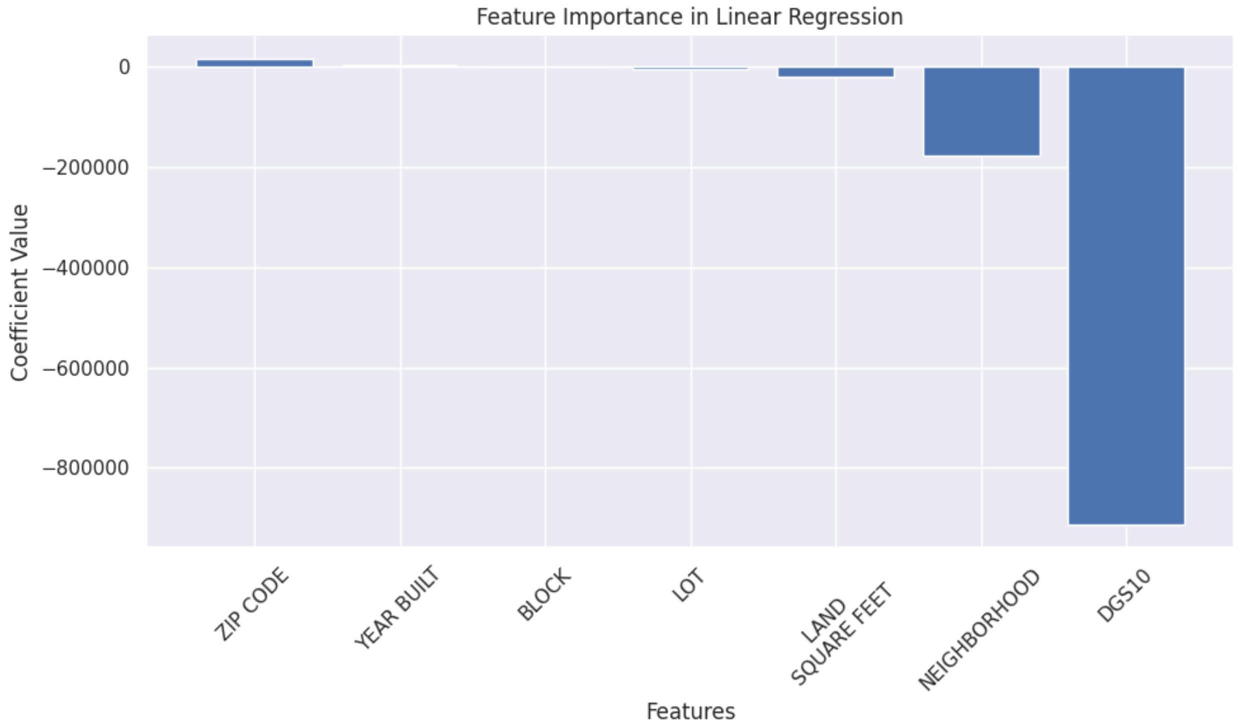
D. Model Selection and Deployment.

Linear regression - Linear regression is a popular supervised machine learning algorithm used for regression problems. Linear regression finds a linear relationship between one or more features and a label. Linear regression attempts to find the best-fit line that minimizes the least-squares error, that is, the squared difference between the actual training data's label and the model predicted label given by the equation above, summed over all examples. That is, linear regression produces a specific estimate for the model parameters .



(Linear Regression Algorithm illustration)

Since this is the most basic model we decided to start with that and created a sorted bar plot to visualize feature importance. Interestingly it focuses on DSG10 - which is the interest rate, the next highest level of importance feature happens to be Neighborhood which does make a lot of sense.



The model didn’t provide good results so we decided to try another model called Random Forest Regression.

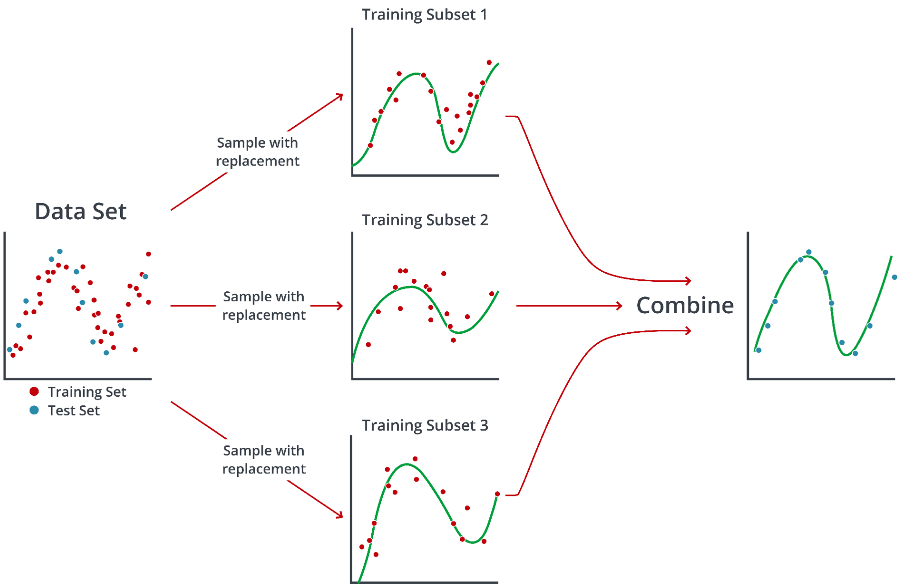
Random Forest Regression

Random forest is used to achieve the optimal balance of low bias and low variance. Random forest is an ensemble of decision trees that generalizes better than an individual decision tree. Although a random forest provides better performance than a single tree, it takes more time and resources to train. While a random forest tends to do a better job at generalization than a single tree alone, the training time as well as the prediction time are more costly than a single tree alone. For this reason, we cannot simply have an arbitrarily large number of trees. So in our case it wasn’t expensive.

PseudoCode:

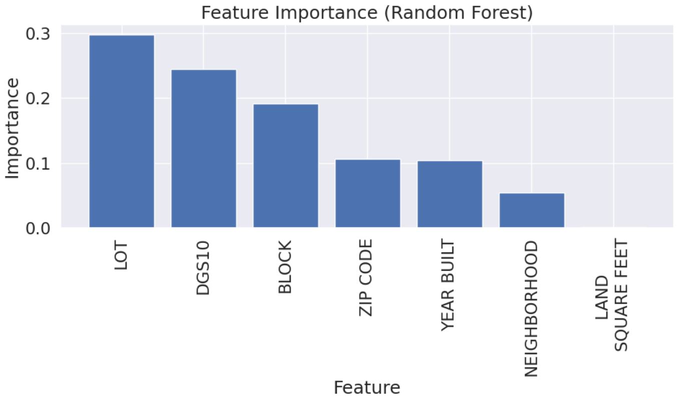
```
    For I → number of estimators:
```

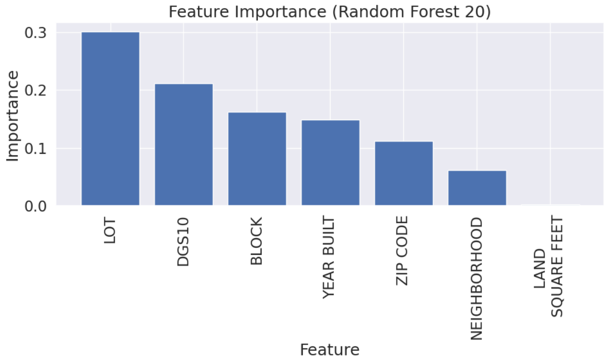
1. Bootstrap data = sample N examples randomly, with replacement
2. Randomly select a subset of features
3. Build a decision tree with bootstrap data and add it to the ensemble



(Random Forest Algorithm illustration)

After running the model we had better results and found a feature importance level .This was an interesting observation, the first one is lot(A tax Lot is a subdivision of a tax Block and represents the property's unique location) and the second is interest rate.   First model ran 100 trees in the forest and the second one did only 20. The difference in performance was slightly better.

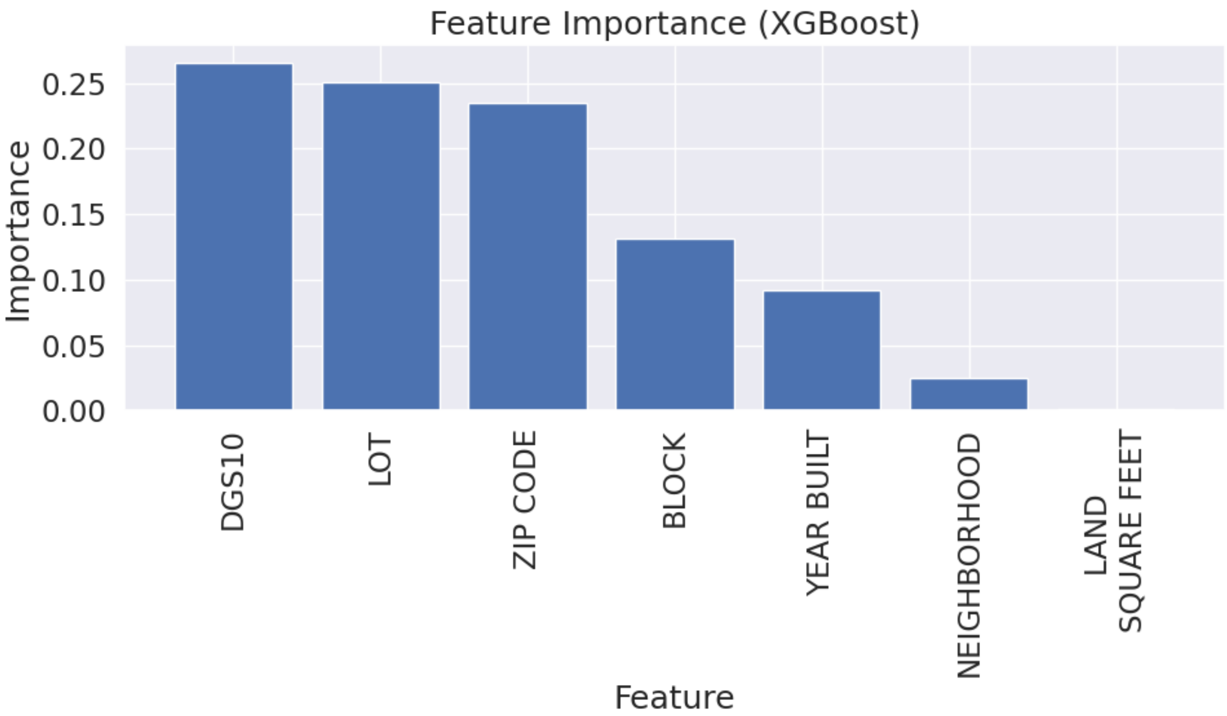




As you can see the feature importance has a different order of year built and zip code.

Gradient Boosting Regression

XGBoost - extreme gradient boosting, belongs to the gradient boosting family of algorithms. Gradient boosting is an ensemble learning method where multiple weak learners (usually decision trees) are combined to create a strong predictive model. XGBoost improves upon traditional gradient boosting by introducing several enhancements and optimizations. It is useful to be applied to predict regression problems. Even though the results were worse than Random Forest it had an interesting insight to the features. Interest rates are on top and LOT, Zip codes are almost at the same level, that means the 3 features were almost equally important in XGboosting.



E. Evaluation

To evaluate our linear regression model, we computed the RMSE (root mean square error) on the test set. RMSE is a metric used to evaluate Regression models.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{n}},$$

$$R^2 = 1 - \frac{\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2},$$

$$Adjusted\ R^2 = 1 - \left[\frac{\left(1 - R^2\right) \times (n-1)}{n-k-1}\right],$$

Root Mean Square Error (RMSE) finds the differences between the predicted values and the actual values. To compute the RMSE, we used the scikit-learn mean_squared_error() function, which computes the MSE between y_test and prediction. We will then take the square root of the result to obtain the RMSE. Finally, we will use the coefficient of determination, also known as R^2. R^2 is a measure of the proportion of variability in the prediction that the model was able to make using the input data. An R^2 value of 1 is perfect and 0 implies no explanatory value. We used scikit-learn's r2_score() function to compute it.

Random Forest Regression model with 100 trees got:

Mean Squared Error: 1115776723387660.75

R-squared: 0.43

Random Forest Regression model with 20 trees got:

Mean Squared Error: 1304235595353860.00

R-squared: 0.33

Linear Regression model:

Mean Squared Error = 44974976.14

R-squared = -0.04

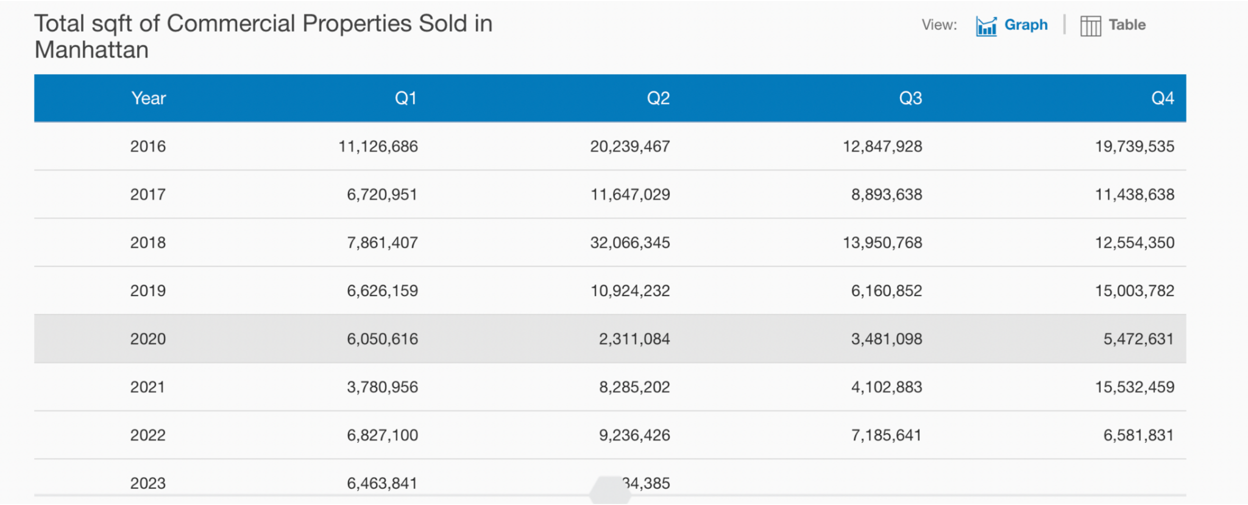
Extreme Gradient Boosting model got:

Mean Squared Error: 1186586957890756.50

R-squared: 0.39

The random forest models seem to have higher MSE values, indicating that they have larger prediction errors compared to XGBoost and linear regression models. The linear regression model's negative R-squared suggests that it's not capturing the relationships in the data effectively. XGBoost appears to strike a balance between the random forest models and the linear regression model in terms of both R-squared and MSE.

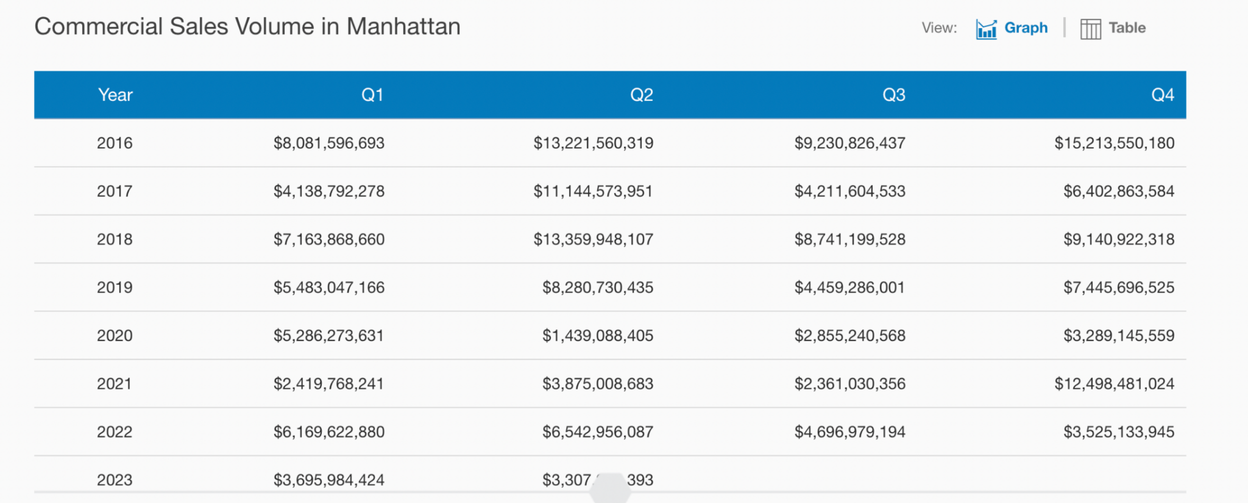
The common thing all these algorithms had was the feature importance, the top 2 features were: an interest rate and a tax lot. Therefore it proves the point that interest rates have a high influence on the market price of commercial real estate and can significantly slow the market. According to PropertyShark - a real estate agency in New York, since 2020 when the Covid-19 shutdowns happened and commercial property sales went slow the market started recovering in 2021, the highest amount of sales



Total sqft of Commercial Properties Sold in Manhattan

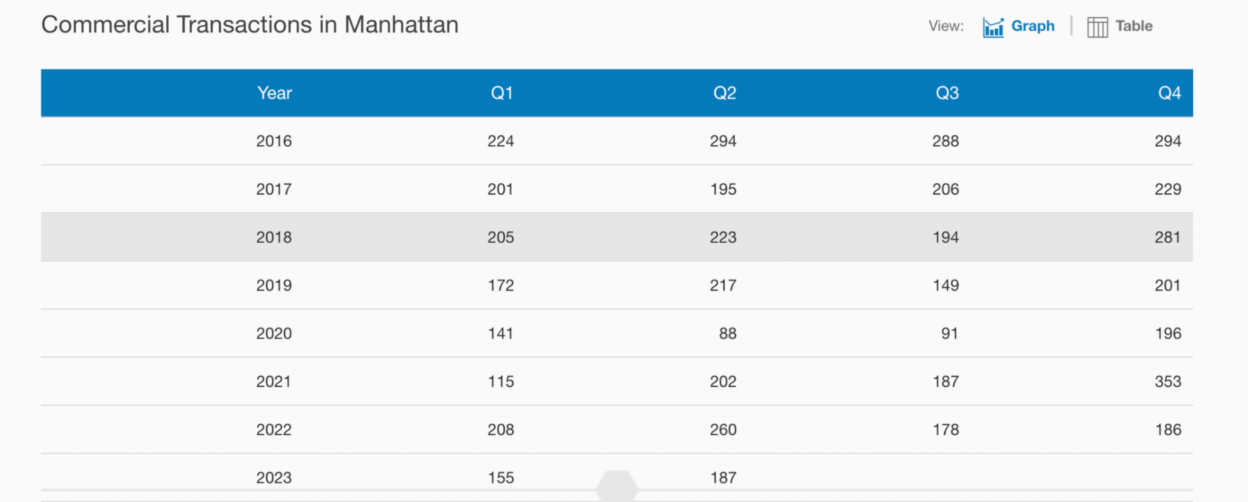
Year	Q1	Q2	Q3	Q4
2016	11,126,686	20,239,467	12,847,928	19,739,535
2017	6,720,951	11,647,029	8,893,638	11,438,638
2018	7,861,407	32,066,345	13,950,768	12,554,350
2019	6,626,159	10,924,232	6,160,852	15,003,782
2020	6,050,616	2,311,084	3,481,098	5,472,631
2021	3,780,956	8,285,202	4,102,883	15,532,459
2022	6,827,100	9,236,426	7,185,641	6,581,831
2023	6,463,841	34,385		

were at the fourth quarter when the sales went up to 353 units - 15 million sq ft which amounted 12.48 trillion dollars. But when the interest rates increased the sales significantly slowed in 2022, the highest that year was second quarter which was 260 units sold - 9.2 million sq ft - 6.54 trillion dollars. Not only the number of transactions decreased but also the price has increased from 688 dollars per sq ft in 2021 to 711 dollars per sq ft in 2022. This proves the point that the interest rates increased the price and made it harder to buy a property enough for the sales to go down 12.93%.



Commercial Sales Volume in Manhattan

Year	Q1	Q2	Q3	Q4
2016	$8,081,596,693	$13,221,560,319	$9,230,826,437	$15,213,550,180
2017	$4,138,792,278	$11,144,573,951	$4,211,604,533	$6,402,863,584
2018	$7,163,868,660	$13,359,948,107	$8,741,199,528	$9,140,922,318
2019	$5,483,047,166	$8,280,730,435	$4,459,286,001	$7,445,696,525
2020	$5,286,273,631	$1,439,088,405	$2,855,240,568	$3,289,145,559
2021	$2,419,768,241	$3,875,008,683	$2,361,030,356	$12,498,481,024
2022	$6,169,622,880	$6,542,956,087	$4,696,979,194	$3,525,133,945
2023	$3,695,984,424	$3,307,[illegible]393		



Commercial Transactions in Manhattan

Year	Q1	Q2	Q3	Q4
2016	224	294	288	294
2017	201	195	206	229
2018	205	223	194	281
2019	172	217	149	201
2020	141	88	91	196
2021	115	202	187	353
2022	208	260	178	186
2023	155	187		

4. Conclusion:

In this paper our goal was to show the evidence that interest rates have enough power to affect people's lives. We have found a correlation between interest rates and its impact on individual’s, property developers, companies, businesses, investor’s, etc. ability to invest in commercial real estate. We have used open source dataset to predict the price of commercial properties with machine learning models such as Random Forest, Extreme Gradient Boosting, Linear Regression by preparing and visualizing the data. We trained and tested regression algorithms and found that the use of Interest Rates as a feature increases the prediction accuracy of the model.

Commercial properties provide spaces for businesses, creating jobs and economic growth. Mixed-use developments contribute to vibrant neighborhoods by offering a mix of residential, commercial, and recreational spaces. Therefore it is important to remember that increasing interest rates can cause a decrease in quality in daily lives of regular people.

We hope that this research helps to increase an accuracy in predicting the price of properties and the insights gained from this research contribute not only to the realm of real estate economics but also to the broader discussions on financial markets and their interconnectedness. As we navigate an era of uncertainty and change, this study underscores the significance of data-driven insights in guiding decision-making processes and shaping a more resilient and adaptive real estate market. The work can be found through this link: https://github.com/aidskiy/Research2023IR

References:

1. Qihang Yi, Yi Zuo, Tieshan Li, Yuhao Mao, Yang Xiao, "Forecasting of Vessel Traffic Flow Using BPNN Based on Genetic Algorithm Optimization", *2021 International Wireless Communications and Mobile Computing (IWCMC)*, pp.1059-1063, 2021.
2. Ouyang Jiantao, "The application for real estate investment price by nonlinear gray forecast model", *Industrial Technology & Economy*, vol. 24, no. 5, pp. 78-80.
3. P. D. Reddy and L. R. Parvathy, "Prediction Analysis using Random Forest Algorithms to Forecast the Air Pollution Level in a Particular Location," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1585-1589, doi: 10.1109/ICOSEC54921.2022.9952138.
4. C. G. Raju, V. Amudha and S. G, "Comparison of Linear Regression and Logistic Regression Algorithms for Ground Water Level Detection with Improved Accuracy," 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICONSTEM56934.2023.10142495.
5. Y. Zhao, R. Ravi, S. Shi, Z. Wang, E. Y. Lam and J. Zhao, "PATE: Property, Amenities, Traffic and Emotions Coming Together for Real Estate Price Prediction," 2022 IEEE 9th International Conference on Data Science and Advanced Analytics (DSAA), Shenzhen, China, 2022, pp. 1-10, doi: 10.1109/DSAA54385.2022.10032416.
6. Shi, Donghui, et al. “Deep Learning in Predicting Real Estate Property Prices: A Comparative Study.” *ScholarSpace*, 3 Jan. 2023, scholarspace.manoa.hawaii.edu/items/5df9d756-67ba-453f-ae66-11fec7a343b2.

7. S. Li, "Research on Evolutionary Optimization Algorithm of Real Estate Pricing Based on Data Mining," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 793-796, doi: 10.1109/I-SMAC52330.2021.9640695.
8. Tao, Fanghong, and Lili Jiao. “Coastal and Port Real Estate Forecasting Model Based on Fast-Adaptive Algorithm for Large Data Sets.” *Allen Press*, Allen Press, 1 Nov. 2019, meridian.allenpress.com/jcr/article-abstract/97/SI/35/428371/Coastal-and-Port-Real-Estate-Forecasting-Model.
9. Deghi, Andrea, et al. “Commercial Real Estate Sector Faces Risks as Financial Conditions Tighten.” *IMF*, 22 Sept. 2022, www.imf.org/en/Blogs/Articles/2022/09/21/commercial-real-estate-sector-faces-risks-as-financial-conditions-tighten.
10. *Rising Interest Rates and the Future of U.S. Commercial ... - Nyu Stern*, www.stern.nyu.edu/sites/default/files/assets/documents/Antell_Glucksman%20Paper_0.pdf. Accessed 31 Aug. 2023.