Pizon Shetu

Final Report

New York City Housing Price Analysis

Problem Statement

Capital Fortune is a real estate company who invest in diverse locations and builds residential homes and have set their sights on New York City. Their concerns are their inexperience in the current market as they hope to build residential units set to be completed in 2024 set across the five boroughs of New York City and are seeking our help on how to properly price their homes in comparison to current market conditions.

Utilizing Zillow's NYC Housing Dataset, I've created a predictive model that will help analyze which features of homes are highly valuable and key insights that lead to greater return in price of homes. Applying various exploratory data analysis, supervised machine learning and hyper-parameter tuning I have developed a XGBoost Regression predictive model which allows us to properly price homes.

Data Wrangling

Zillow's NYC raw dataset was 75,000 observations with 1507 features, many of which were redundant or repeated. While the dataset had expansive it needed much cleaning as many data entries were wrong or mislabeled. The dataset contained units which were lands, commercial use buildings, and corporate office buildings all which I have omitted. It also contained houses with inaccurate pricing mostly on the lower spectrum after further investigation many of these homes were sold to family and friends for far below retail value of the home and thus, I omitted all homes below 100K as well as homes above 10 million.

With so many features I had to reduce dimensionality, I first started by removing all columns with missing data above a threshold. Then I proceeded to remove redundant or repeated columns which have same information. Final step

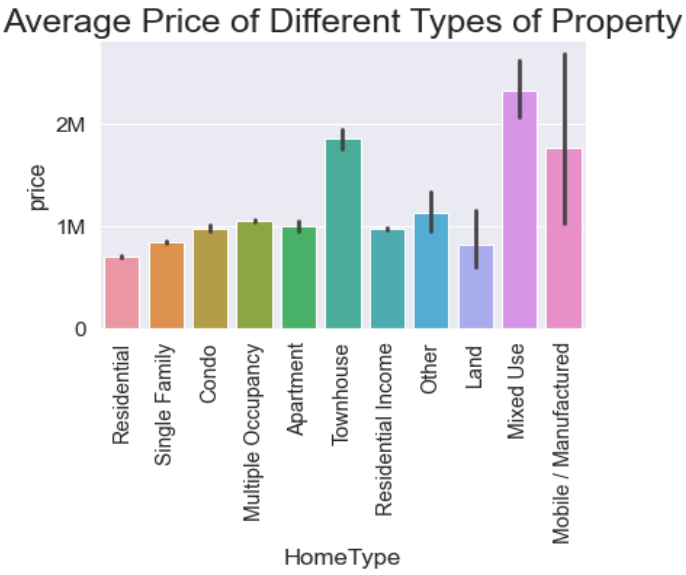
was to view the impact of the remaining feature set and trim of the unnecessary fat which have no impact on price. To achieve this, I utilized randomforest and XGBoost feature importance tool build in on scikit-learn library. I was left with a final dataset 62456 rows and 24 columns.

I was able to utilize the latitude and longitude in the dataset to reverse geolocate the neighborhoods and boroughs of each house. This led to key insights into pricing of different areas.

Exploratory Data Analysis

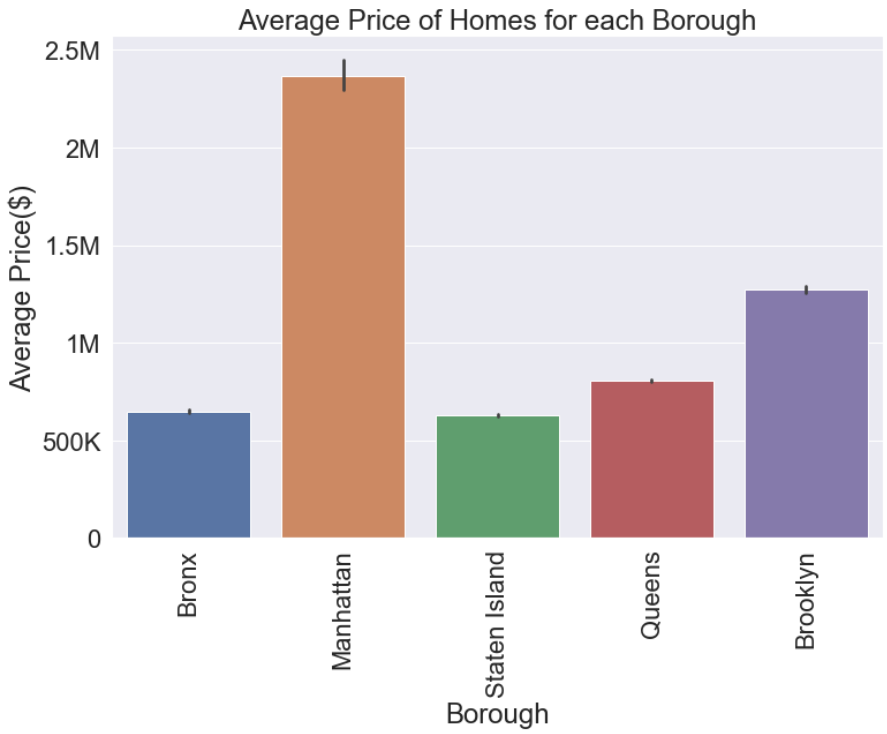
Here I really tackled the relationships between different features and its affects on price of a house, with such an abundance of features we saw that lot size, living space, bedrooms, bathrooms, and taxes paid for that property had the highest implications on the price of a house.

Note Mixed Use and Mobile/Manufactured buildings were mostly corporate offices, and warehouses not residential housing.



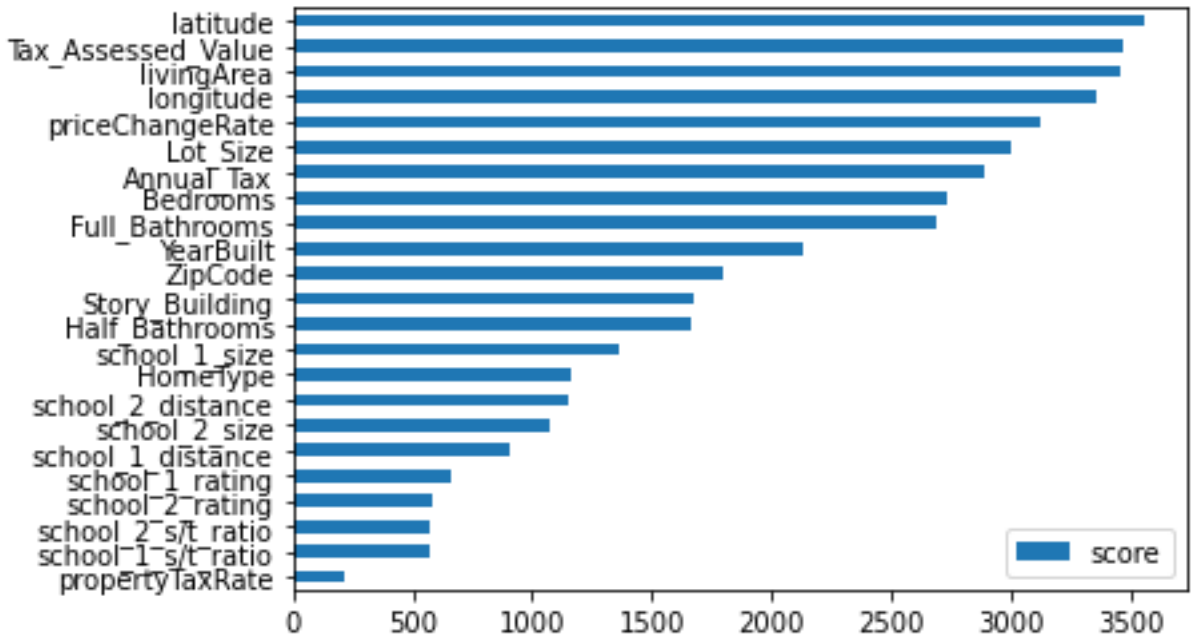
While normally the type of Housing would show a significance in the house pricing in this case, I believe it is very difficult to pinpoint if home type influences price. NYC prices tends to vary heavily on location and so while a Condo might be only 250k in Staten Island it can be over 2 million in Manhattan. Generally, in most areas Townhouses tend to be cheaper to purchase than detached houses but it seems in Manhattan and Brooklyn Townhouses are more expensive than Multiple Occupancy homes and Residential Income homes but the other 3 Borough it seems it follows the trend of normal USA houses prices where Residential Income and Multiple Occupancy housing is more expensive than Townhouses.

Most significant takeaway was how important the location of a house is more so than its underlying features. Manhattan dominates all boroughs for the most expensive homes on the market. Worth while mention is land price is much more expensive on Manhattan, but construction costs are nearly the same across the boroughs.



Curious about the affect the build year would have on price we saw that houses price prior to 1900's was much more expensive than those after. Looking into homes build before early 1900's reveals, 70% of all homes were in either Brooklyn or Manhattan. This further indicates how much value locations brings into the price of a house particularly Manhattan and Brooklyn. But looking into more recent trends we saw and uptick of average price increase to newer homes and 53% of these homes were in Queens and Staten Island.

Other features which played a key role were annual tax, the tax assessed value of the property, here is a graph of the most impactful features.



Normally you would not expect latitude and longitude to play such a high factor but due to a local market phenomenon homes clustered together with similar latitude and longitude tend to indicate the price better than others, I've tested multiple times to check this phenomenon and tried predictions without the inclusion of latitude and longitude all those models performance were very poor and some instances I ended up with results as bad as linear regression.

Preprocessing and Training Data

Missing data was a nuisance as many of the key features such as number of bedrooms and bathrooms had many null values. Initially I had implemented a mean and median imputation of all features, but this was too simple. I later implemented the MICE imputation technique alongside gradient boosting regressor to impute all missing values, this was a costly computational step but yielded the best results. I also scaled all my data utilizing sci-kit learn standard scaler function which standardizes a feature by subtracting the mean and then scaling to unit variance. This would lead to better prediction results. For the

training and testing split I used train-test-split function via sci-kit learn to create a 80-20 train-test split.

Model Selection

I've tested 4 models, linear regression, nearest neighbors, randomforest, and gradient boosting via XGBoost. Before we can evaluate and compare the models, I had to choose a metric for performance, I observed the following R^2(shows how well terms (data points) fit a curve or line.), MAE(Mean absolute error), MSE(Mean squared error), RMSE(Root mean squared error), and MAPE (Mean Absolute Percentage Error). From these 5 metrics I decided MAE would the main metric as this explained on average how far off the price of a house was to the actual price of the house (lower the better).

When it came to the 4 models, XGBoost performed the best with a MAE of 220K, randomforest was a close second with an MAE of 222K, and the rest were far too large as the linear model came in last with the largest MAE of 380K.

After selecting XGBoost model I proceeded to hyper-tune my parameters for the model. I also took the log transformation of my train and test split data as this yield better results and I made sure to transform those price results back to standard form. I used MAE as a metric to 'tune' my hyper-parameters, I tuned the following parameters, max_depth, min_child_weight, eta, subsample, and colsample_bytree. Max_depth controls how long our decision tree will be the more depth the more complex the model becomes making it more likely to overfit. Min_child_weight controls when we stop splitting our sample size in a node depending on our threshold in the case for regression, the larger min_child_weight is the more conservation our algorithm becomes. Eta shrinks the feature weights to make the boosting process more conservative. Subsample controls the number of observations to be randomly sampled if subsample is .5 then half of the observation is randomly sampled for each tree, col_sample similar as sub_sample but for columns.

I cross-validated each of these parameters to find the best values without over-fitting or under-fitting, and this resulted in a reduction of nearly 30K in MAE on testing data and 65.7K in training data.

Takeaways

We found the best results with XGBoost as this led to the most accurate price prediction compared to other models. Coupled our EDA we have found the most important factor of price of a home is location in New York City. My suggestion to Capital Fortune, if given a choice between the 5 boroughs I would choose Manhattan, then Brooklyn followed by Queens. While it is much more expensive to buy land for construction in Manhattan, we can see that return on investment is also much higher in this location, while Queens has the largest opportunity as land is far cheaper and cost of construction is relatively equal. I also suggest opting for townhouses and homes with multiple families as these saw the highest return on investment.

We saw that more expensive houses are older houses but upon a closer look this is because these are homes build in the 1800's and early 1900's in Manhattan and Brooklyn, leading to location being the main driving factor of the price, and many of these houses were remodeled even though their listed built year is older. But notice after the 1950's the average price of houses increases as more recent it is. This puts our clients houses at a premium compared to an older or remodeled old home.

On average we can expect homes build in Manhattan to run between 1 to 3 million more compared to other boroughs with an average price of 3 million for a newly build home. Next would be Brooklyn with an average price of 1.2 million for a newly build home and then Queens with 800K. Bronx and Staten Island might not be as lucrative as its neighbors, but they can be much to be desired as a newly build home on Bronx can be around 600K and for Staten Island 500K.

Future Research

This was a tedious and interesting project to work on, working with such dirty raw data really gave me and appreciation for data collection and integrity. It really changed my perspective and learning experience on how to properly clean and wrangle data. I would like to see how different results would’ve been in different markets as NYC is a very sought after and the gap between Manhattan compared to Bronx or Staten Island is staggering.

I believe there is more data we can compile for better predictions for instances the construction costs of homes, or prices of goods and services in the proximity of the house. I would love to continue this and look into the whole United States housing market and how different cities or states compare and how different features impact the overall market compared to local markets.