Exploring Models for Predicting Housing Values

Arnav Narain, William Gadala, Abhinav Bachu, Jinwoo Jung

DS 3000 Final Project

Abstract

New York City housing has always been known to be luxurious, but expensive. Buying a home is one of the most important life decisions someone can make and the number of home-owning residents in NYC has only been growing. Zillow offers a vast selection of housing options to choose from, but it can be quite difficult to find a good offer that gets you the best value for your money. This project aims to determine which features of a NYC home are the most influential in affecting a house's price in order to give prospective homeowners more information for their search.

Introduction

For centuries, New York City has been one of the most pivotal cities in the world. The city has immense economical power, and consequently, there are countless of jobs that so many desire. As a result, the demand for the houses are incredibly high, and the prices are more than what most would hope for.

With such a high price range, it is more important than ever for consumers to understand exactly how much they should invest on the homes that are in the market. While there are websites such as Zillow that provide information on prices, it is well known that there are many data that are rather inconsistent or missing, causing further confusion.

Our goal is to create a model that can successfully predict the house prices in New York City by analyzing various factors such as city, number of bedrooms and bathrooms, and even nearby schools.

Related Works

The housing prices in New York have been a popular area of interest for housing experts as New York City is the most expensive city in the US. As different factors continually change the housing market, it is important to explore and redevelop housing models.

"Why is Manhattan So Expensive?" explores the factors that make Manhattan cost more than neighboring areas in New York. It is similar to our model as it determines some factors that can be used to predict housing prices, however our model attempts to predict housing prices instead of identifying factors leading to expensive housing.

Methodology

Data Acquisition: Our data was obtained as a CSV file from Kaggle. The data was compiled using housing listings on Zillow.com using Zillow's API on 1/20/2021. The data consists of 75,629 listings, although this number slimmed to 18,573 when all null values and outliers were handled.

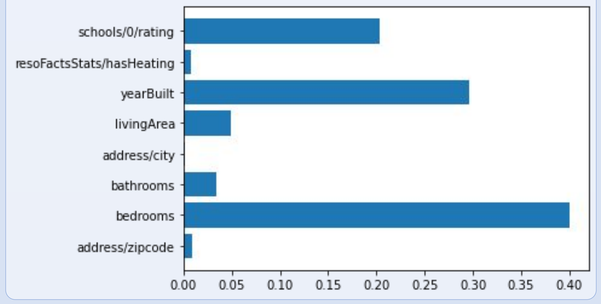
EDA: An important step of our EDA was to standardize locations. A feature we identified as potentially correlated with housing price was the borough location of the home. The dataset swapped between neighborhoods and boroughs in the dataset, so replacing neighborhoods with the corresponding borough was critical for standardized data.

Many observations in our dataset were not homes; office spaces and garages were listed in the dataset, potentially skewing our results. To isolate and remove these outliers, we included only observations that fell into a particular bedroom/bathroom range, and maintained a price threshold, reducing outliers in the both the upper and lower directions.

Two features, heating and location, were categorical and had to be encoded. The rest were numerical and were normalized.

Modeling: Due to the high number of features we included, we identified Random Forest Regression as the most effective model to model our data. We also identified kNN classifier as a potentially good model, and linear regression as a baseline. We tuned the n_estimators for the Random Forest Regression and the neighbors for kNN classifier and identified the models with optimal mean squared errors and scores.

We then identified the features most closely correlated with price and visualized them using a bar chart, as seen below.



Results and Evaluation

After filtering and cleaning the data, our final dataset had 18,573 rows and 9 columns. These 9 columns were the specific features that we believed would influence a house's price. They were: school ratings, heating, year built, living area, city, number of bathrooms, number of bedrooms, and zip code.

Of the three models we tested, Random Forest Regression, kNN classifier, and Linear Regression, we found Random Forest Regression produced the smallest root mean squared error, as shown from the table below.

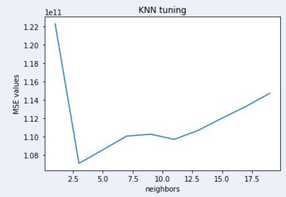
Model	RMSE ($)	Tuning
Random Forest	244867.82	n_estimators=81
K-Nearest	327295.25	neighbors=3
Linear Regression	421723.04	

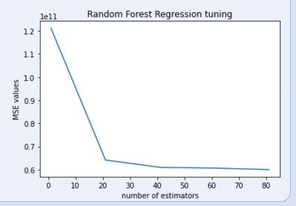
Table 1

Random Forest: As predicted, Random Forest was the most accurate model with a root mean squared error of 244,867.82. This means that, on average, this model was inaccurate by $244,867.82 when predicting price. Tuning the random forest showed diminishing returns, leveling off when n_estimators reached 20. Due to the EDA process, the dataset was slimmed down massively, thus reducing the available observations to train this model. With a larger dataset, random forest, and kNN, could certainly reach an RMSE of < 200,000.

K-Nearest: After finding the ideal number of k through KNeighborsRegressor, we were able to get a RMSE of 327,295.25. which was higher than our anticipated prediction.

Linear Regression: It makes sense that the linear regression has the highest mean error, as it is intuitively the least relevant model. Housing prices cannot be easily distinguished by forming regression lines for each feature, so a linear regression cannot fit the data as accurately.





Impacts

Since there are many factors that should be considered when evaluating New York City's housing prices, many consumers and sellers have a hard time acquiring the exact value of the homes and they end up paying too much for what they are purchasing or earning too little on what they are selling.

Our model can solve this problem as its role is to determine the adequate price for the houses in New York City by executing machine learning algorithms on the dataset of house prices, allowing both buyers and sellers to understand the value for houses on sale. The model could provide further positive impacts educating participants the top factors that influence the prices of houses in New York City.

Conclusion

Our goal for this project was to develop a model that can accurately predict New York City Housing prices, which we were able to accomplish with a fair level of accuracy. Through careful hyper-parameterization of the estimators in a random forest regression, our model can predict housing prices with an average error of $250,000 based on 8 features. Of these features, the most impactful included the number of bedrooms, school rating, and year built.

While our model is able to predict housing prices within decent bounds, more improvement could definitely be made to decrease the error. The dataset we used contained some non-housing data. An example was a 50 car garage with no other features. Our EDA made an attempt to remove such data and was able to delete a significant portion of it. It also decreased the amount of data in the dataset. Having a greater amount of data and improved data relevance would thus enhance the training of our models and lower the RMSE.

References

1. https://scholar.harvard.edu/glaeser/files/why_is_manhattan_so_expensive_regulation_and_the_rise_in_house_prices.pdf

2. https://www.kaggle.com/datasets/ericpierce/new-york-housing-zillow-api