Milestone 2: Draft White Paper

Felipe Rodriguez

Bellevue University

DSC 680 Applied Data Science

Professor Amirfarrokh Iranitalab

March 29, 2024

Project Overview/Background

As the real estate market is consistently changing, it is important to consider the factors that contribute to prices. Understanding some of these factors and trends can help the consumer make informed decisions on their next sale. A real estate agency wants to understand their real estate listings cross-country. Additionally, the company wants to build a model to accurately predict prices based on zip code and size, specifically for the state of New York. The data provided contains historical listing information that will aid in the analysis and modeling. The model and predictions will be used to provide more accurate pricing recommendations to clients.

Data Overview

The data is being obtained from Kaggle and is sourced from Realtor.com. The data contains the following information: status, bed, bath, acre lot, city, state, zip code, house size, previous sold date, and price. The data will allow for the analysis of price based on the features provided. The fields in the data contain the following information:

Data Dictionary

column	description	data_type
status	Current standing of the home (for sale or ready to build)	object
bed	Number of beds in the home	float64
bath	Number of baths in the home	float64
acre_lot	Size of the lot	float64
city	City where the home is located	object
state	State where the home is located	object
zip_code	Zip code of the home	float64
house_size	Square fottage of the home	float64
prev_sold_date	Date when the home was previously sold	object
price	Current sale price or previously sold price if the house is not for sale	float64

The data preparation phase requires cleansing and prepping the data so that it is fit for use. The first portion involves removing null values. This piece will ensure that when creating models there are no errors. Prepping the data for modeling also involves removing the other states from the main data set. Also, to plot the prices by zip code, the fields latitude and longitude need to be appended to the data set.

Methods

The methods used in this project will include various visualizations of prices, correlation matrices of the variables, and a linear regression model to predict prices. The visualizations will include various scenarios of the prices, which will include geographical analysis and charts illustrating the prices. The correlation matrices that will be created will help understand the correlation between the features and the target - price.

The initial model chosen for this project was a linear regression. Linear regression is used because “Linear regression analysis is used to predict the value of a variable based on the value of another variable” (IBM, n.d.). In the first model created, the columns used were zip code and price. However, the R-Square score was very low, which indicated poor performance in the model (see Appendix A). R-square is used as the metric for determining model accuracy because R-square shows how well the data fits the regression model (Taylor, 2023). Since this model gave low results, the adjustment was to use the other variables in the data to see if there was improvement. This resulted in a better score; however, it is still low.

Since the linear regression model was providing low R-squared scores, a different model was selected to attempt to get a better fit. The model selected is a Random Forest Regressor Model. Random forest regression “is a supervised learning algorithm and bagging technique that uses an ensemble learning method for regression in machine learning” (Makhijani, 2023). Because the model creates multiple decision trees and can handle non-linear relationships in the data, it was the next model applied to the data. This model provided a high R-squared score and indicates that the model is accurately capturing the relationships in the data.

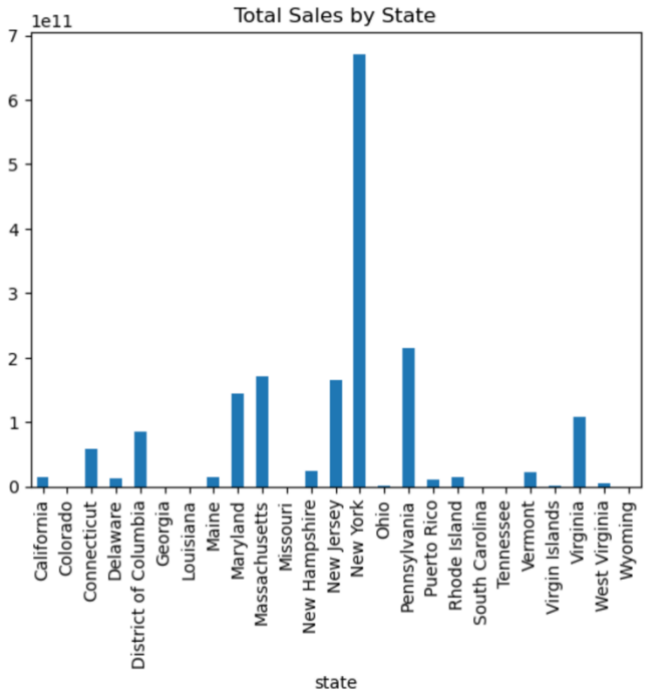
Assumptions

Prior to analyzing the data, the following assumptions are made:

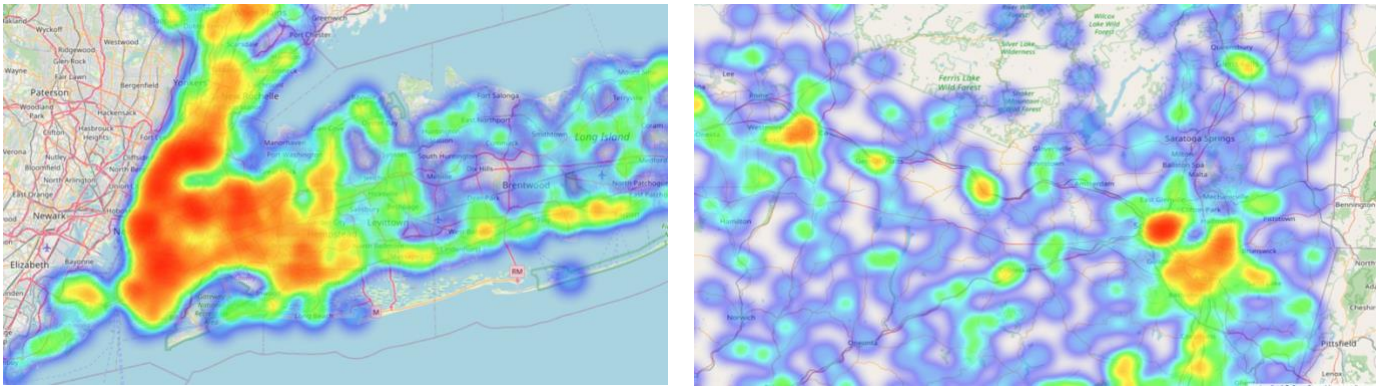
- The data provided and gathered reflects the housing market in New York.
- The model assumes there are no external factors that are impacting price.
- The model assumes that the relationships between the features and home prices is consistent across the different zip codes.

Analysis

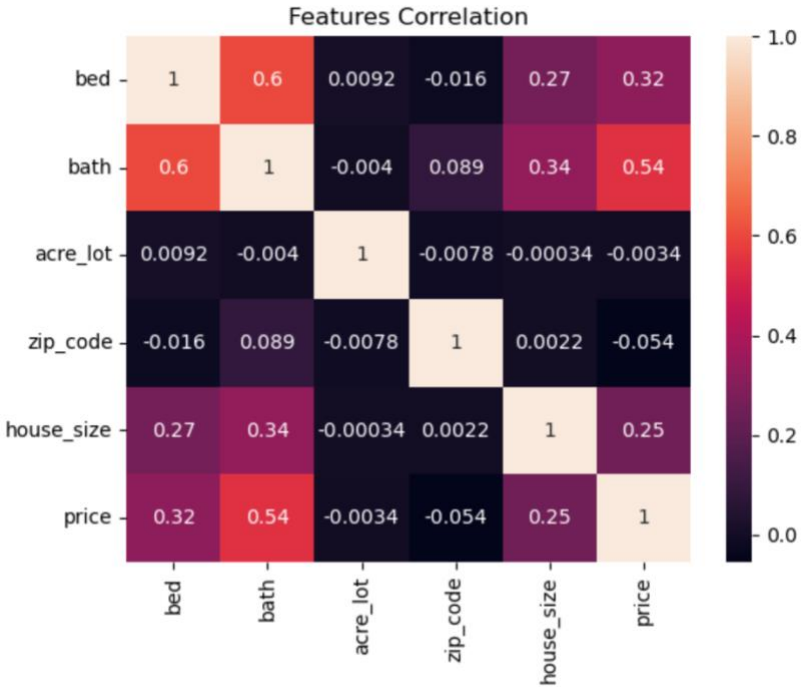
The data explored gives insights in how the prices are distributed. The first plot gives insight in where home prices are the highest. The chart below shows that the sum of all prices is highest in New York. The other states do not have as high of total prices. Since this goal of this project is to analyze prices in New York, having a wide variety of data gives confidence in the model being created.



The next set of visuals give an idea of how the prices are distributed. The view on the left is of New York City and on the right is Albany and Utica. These two visuals show that the listings are concentrated in bigger cities.



To understand further how the variables relate to each other, a correlation matrix was created on the different features of the data. This analysis is done on the numeric features of the dataset. Below are how the variables relate to each other.



The initial assumption on this data set was zip code and house size were going to play the biggest role in price. In a correlation matrix, the higher the value is to 1 the higher the correlation is. This analysis shows that the number of beds and baths plays a higher role in the price of the home.

Challenges/Limitations/Ethical Considerations

The challenges in this data were found when creating the model. When using the two proposed features zip code and house price the model performed poorly. Because of this finding, there was a decision to add the other numeric features to improve performance. However, that improved the model slightly. To get a higher performing model, it was decided to change the model from a Linear Regression to a Random Forest Regression. This method significantly improved the model.

A limitation of the Random Forest Regression is that this method will not be able to handle the data appropriately if there is missing data in a time series. If this model is explored in the future to predict prices over time as opposed to historical records, this will need to be considered as missing time data will result in the model to under or over predict (Thompson, 2019)

Ethical considerations for this data and models are to ensure that this data is secured from unauthorized uses and distribution. This model is meant to aid this organization and should be carefully reviewed before sharing. Additionally, since this data does not include any ethnic information is does not discriminate against any specific groups of people.

Future Uses/Recommendations/Implementation Plan/Conclusion

While the Linear Regression model did not perform as expected, if there is a future need for analysis over time, the Linear Regression model can be applied. This can also be done with

the Random Forest Regression model. Additionally, it was identified that there are 25 total states that there is pricing data for. This model can be applied to the other states to aid in pricing determination.

The Random Forest Regression Model performed well and should be refreshed when new homes are listed and sold. It is recommended that the model be refreshed every 6 months. By refreshing the model every 6 months, it can be maintained up to date with the most recent data and continuously monitored to give accurate prices.

The best way to implement this model is to develop a web application that will allow a user or agent to input data and retrieve predictions based on the model that was developed. An example of a program that will allow for this ability is Streamlit. To use this, a model needs to be created, an API, and lastly the Streamlit server (Makhijani, 2023). This method does not need knowledge of web application programming languages.

In this project, real estate prices were analyzed to understand which factors correlate and influence the price of a home. The main factors that resulted in a high correlation for price were bathrooms and bedrooms. The analysis also uncovered that the prices are elevated in bigger cities. Additionally, a model was created to predict prices in New York to give consumers more accurate prices when they list their home. This model performed well and is ready for deployment.

Appendix A

This appendix includes the results and findings of the three models created. All the models contain New York data and are limited to that state. The first two models are Linear Regression. The features use on the first model are on the left and the features on the right are for the second and third model.

	zip_code	price
0	10001	213750000.0
1	10002	689000.0
2	10003	275915000.0
3	10004	100000000.0
4	10005	14795000.0

Model 1

	bed	bath	acre_lot	zip_code	house_size	price
54248	3.0	2.0	2.02	12521	1600.0	425000.0
54258	4.0	2.0	0.24	12521	1239.0	225000.0
54267	4.0	1.0	4.20	12516	1500.0	299999.0
54268	3.0	2.0	2.90	12529	1404.0	374900.0
54278	3.0	2.0	1.20	12546	1350.0	375000.0

Model 2 and 3

For all models, there is an 80/20 split with training and testing data. The third model conducted is a Random Forest Regression. The data between models 2 and 3 are the same, the only difference is the method used. The target value for all models is price and the features are all the other variables. Below is the performance for all three models:

```
Model 1 Results
Mean Squared Error: 1.0095666005774102e+17
Mean Absolute Error: 127857450.28387067
R-squared score: 0.005190385914512263

Model 2 Results
Mean Squared Error: 1341708654258.3293
Mean Absolute Error: 436656.6665262822
R-squared score: 0.3201483824250393

Model 3 Results
Random Forest Regressor - Mean Squared Error: 29911620514.03588
Random Forest Regressor - Mean Absolute Error: 8707.15125422424
Random Forest Regressor - R-squared score: 0.9848436070482107
```

The first model performed poorly with a low R-Squared score. Because of this, the next model includes more features to significantly affect the model performance. The second model performed better, but not well enough to justify deployment. The last model is Random Forest Regression which performs the highest out of all three.

Questions:

1. Why did Random Forest Regression Perform better than linear regression?
2. Are there other models that can be explored?
3. Why isn't there a model using house size and price and just zip code?
4. What states do you recommend modeling next?
5. Would containing more features improve the performance in the model?
6. Can your team aid in model deployment?
7. What resources are needed for model deployment?
8. Is there any additional validation work needed for this?
9. We have a data set that contains date stamps, can this be applied to that set?
10. What are the resources needed for refreshing the model every 6 months?

Reference:

Makhijani, C. (2023, November 7). *Machine learning model deployment as a web app using Streamlit*. Medium. https://charumakhijani.medium.com/machine-learning-model-deployment-as-a-web-app-using-streamlit-4e542d0adf15

Taylor, S. (2023, November 22). *R-squared*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/r-squared/

Thompson, B. (2019, December 20). *A limitation of random forest regression*. Medium. https://towardsdatascience.com/a-limitation-of-random-forest-regression-db8ed7419e9f

What is linear regression?. IBM. (n.d.). https://www.ibm.com/topics/linear-regression