JSC370 Project Report:
Factors that affect the house price in USA

Yuxuan Wang

April 30, 2024

1 Introduction

1.1 Motivation

The real estate markets, present an interesting opportunity for data analysts to analyze and predict where property prices are moving towards. Prediction of property prices is becoming increasingly crucial, offering valuable insights into both the broader market's trajectory and the economic health of a nation. This project tackles the classic problem of house price prediction using machine learning and statistical methods, with a specific focus on identifying the key factors influencing housing prices in the United States. Furthermore, we dive into a case study of Seattle, comparing the city's housing market dynamics to those of the broader national market, and provides a more comprehensive understanding of the factors affecting property prices at both local and national levels.

1.2 Related Works

Many researchers tried to evaluate and compare the factors that affect house prices. House price not only shows the price of the land, but also reflects the economy, urban planning, and social policy.

Dennis, Patric, Charlotte and Christopher find out the serial correlation and reversion parameters are then shown to vary cross sectionally with city size, real income growth, population growth, and real construction costs. Serial correlation is higher in metro areas with higher real income, population growth and real construction costs. Mean reversion is greater in large metro areas and faster-growing cities with lower construction costs.

Peter Englund and Yannis M. Ioannides's research compares the dynamics of housing prices in 15 OECD countries. And the result reveal a remarkable degree of similarity across countries and suggest rich dynamics for the first-differenced real annual house prices, with a significant structure of autocorrelation.The contemporaneous GDP growth rate and the rate of change in real rate of interest are very significant, along with the first-order lag, whose coefficient remains at 0.45. Lagged GDP growth and the real rate of interest exhibit significant predictive power.

2 Methods

2.1 Data

Our models will perform a regression task on a collection of USA house prices. The first dataset contains factors such as the number of bedrooms, the number of bathrooms, the square footage of the living space, the lot, the places above the ground, the age of the house, the number of waterfront, the score of view, and house condition. The second dataset contains only houses in Seattle, and the factors are the number of bedrooms, the number of bathrooms, the total size of the house, and the lot size. Both datasets are from Kaggle and are listed in the Reference list.

I use read.csv to import the data into R. From dim(), I checked the shape of the data. Using the str() function, I checked the type of the variables. I use na.omit() to delete the rows with missing data, and the result shows that the data has no missing values. For variables ”waterfront”, ”view”, ”condition”, I turn them from int into type factor. Since the difference between 0,1 and 1,2 doesn't represent the same amount of difference. We perform EDA on the first and second dataset.

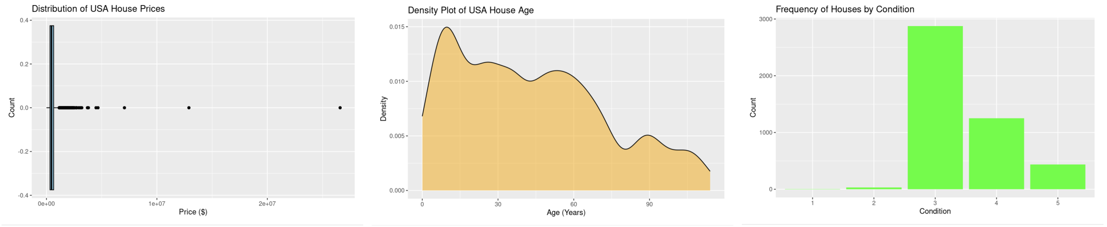


Figure 1: Exploratory Data Analysis on House Price of USA(Dataset1)

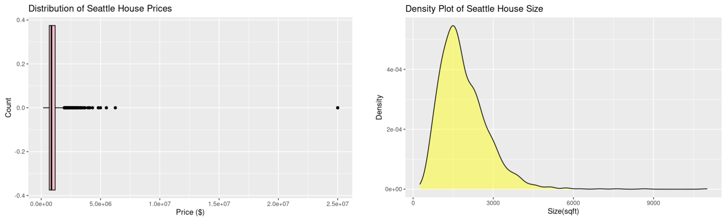


Figure 2: Exploratory Data Analysis on House Price of Seattle(Dataset2)

Then I combined the two dataset into a new dataset. The first dataset only contains rows that city is Seattle, and then combine with the second dataset, and only retains the column the number of bedrooms, the number of bathrooms, the size of the living space, and the size of the lot. With this new dataset, we focus on the house price of Seattle, and tried to find out its factors' similarity and difference compare with the whole country. The plot illustrates that, as a major United States city, Seattle's housing prices are higher than the national average.

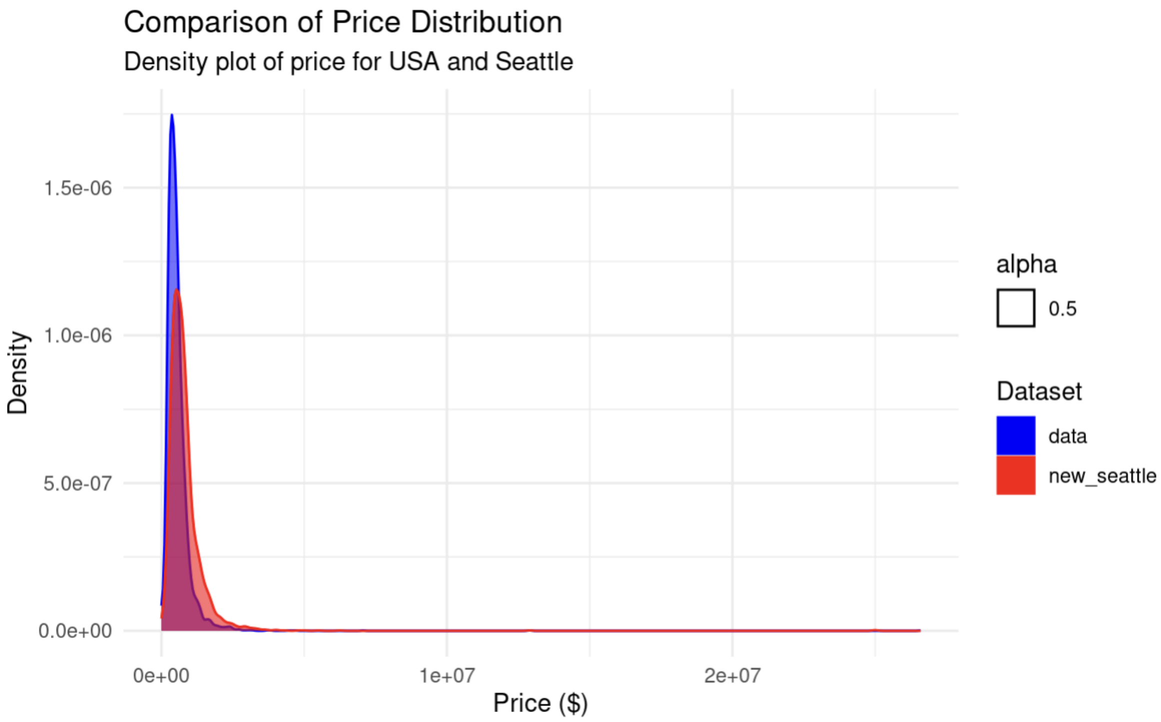


Figure 3: Comparison of Seattle house price and the Whole nation's house price

2.2 Model Structures

The following sections will describe the architecture implementation for the models. We will use two methods on all the three datasets. The first methods is statistical regression method, and the second is the machine learning random forest methods.

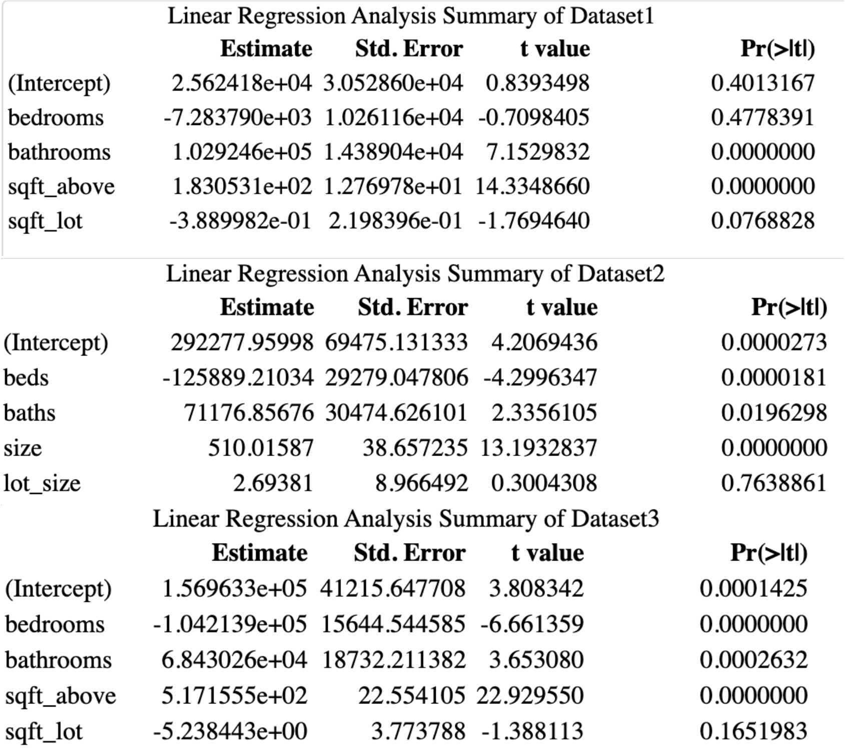
First, I use linear regression on the all three dataset, to check p-value of the models to conclude the factors that affects the house price. If the p-value of the variable is less than 0.05, we have evidence that the variable is statistical important to the prices. Since the basic linear regression is a relatively simple model, I will then use GAM model to capture the non-linear relationship in the dataset. The GAM model is particularly useful because it does not assume a strictly linear relationship between the variables and the outcome. Instead, it allows for flexibility in the shape of the relationship, capturing more complex dynamics in the data.

Then, I use machine learning methods, particularly, random forest to figure out which factor is crucial to the house price. I will Fit a regression tree to predict price first, and prune it based on the optimal complexity parameter. And then use more complex model using bagging, random forest and boosting. From these complicated model, we are able to give a more concrete conclusion about whether the factors are important.

3 Results

3.1 Linear Model

For each dataset, I fit linear regression on the dataset. And for the first dataset, I fit two, one for all factors, and the second for factors only included in the dataset2 and dataset3. From the results, we are able to conclude that for the national dataset, the number of bedrooms is not statistical important, which is counterintuitive. For the second and third dataset, the house price of Seattle, the number of bedrooms, number of bathrooms, the house size are all crucial to the house price. The lot size is not important for all the datasets.



Linear Regression Analysis Summary of Dataset1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.562418e+04	3.052860e+04	0.8393498	0.4013167
bedrooms	-7.283790e+03	1.026116e+04	-0.7098405	0.4778391
bathrooms	1.029246e+05	1.438904e+04	7.1529832	0.0000000
sqft_above	1.830531e+02	1.276978e+01	14.3348660	0.0000000
sqft_lot	-3.889982e-01	2.198396e-01	-1.7694640	0.0768828

Linear Regression Analysis Summary of Dataset2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	292277.95998	69475.131333	4.2069436	0.0000273
beds	-125889.21034	29279.047806	-4.2996347	0.0000181
baths	71176.85676	30474.626101	2.3356105	0.0196298
size	510.01587	38.657235	13.1932837	0.0000000
lot_size	2.69381	8.966492	0.3004308	0.7638861

Linear Regression Analysis Summary of Dataset3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.569633e+05	41215.647708	3.808342	0.0001425
bedrooms	-1.042139e+05	15644.544585	-6.661359	0.0000000
bathrooms	6.843026e+04	18732.211382	3.653080	0.0002632
sqft_above	5.171555e+02	22.554105	22.929550	0.0000000
sqft_lot	-5.238443e+00	3.773788	-1.388113	0.1651983

Figure 4: The Linear Regression Coefficients for the 3 datasets

For the counterintuitive part in the national dataset, we conduct a GAM model to ensure the result. And the result from GAM shows that the number of bedrooms in the national dataset is not important to the house price, as we can see a horizontal line in the graph, indicating the number of bedrooms has no effect on the house price. This will be discussed in the limitation parts. The GAM models' result

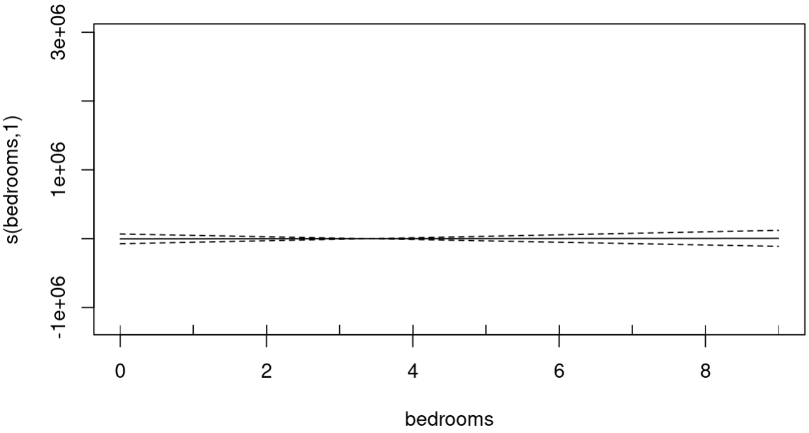


Figure 5: Effect of the number of bedrooms on house prices in national dataset's GAM model

also reflects non-linear relationships in the variables and the house price. The most obvious non-linear relationship is the number of bathrooms and house price in datasets3.

The relationship between the number of bathrooms and house prices is non-linear. Initially, as the number of bathrooms increases from 0 to around 2, the price effect increases. This suggests that additional bathrooms significantly increase the value of a house up to a certain point.Beyond 2 bathrooms, the curve peaks and then fluctuates. This could indicate diminishing returns; as the number of bathrooms continues to increase beyond 2, the additional value added to the house price becomes less pronounced and even decreases slightly at certain points.The confidence intervals widen as the number of bathrooms increases, especially beyond 4 bathrooms. This widening suggests greater uncertainty in the model's estimates for houses with many bathrooms, possibly due to fewer data points for houses with more than 4 bathrooms.

From a practical perspective, when investing in or building properties, it might be beneficial to focus on properties with up to 2 or 3 bathrooms for the best return on investment, as additional bathrooms beyond this point do not significantly increase house prices according to the model.

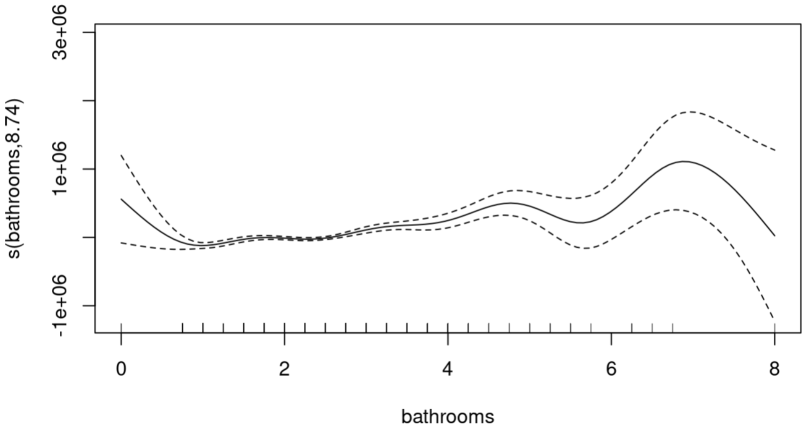


Figure 6: Effect of the number of bathrooms on house prices in Seattle dataset's GAM model

3.2 Random Forests

We first built a random tree based on the national dataset, and the result shows bedrooms is not an important factor to the house price in national dataset, which is coherent to the result of linear regression. To confirm the result, I also conducted bagging and boosting random forests models. In the random forest model, the number of bedrooms contributes less to node purity compared to the size of the lot, total size, and the number of bathrooms. This suggests that while the number of bedrooms is not significantly important, it does not differentiate prices as strongly as the size of the lot, the living area above ground, or the number of bathrooms.

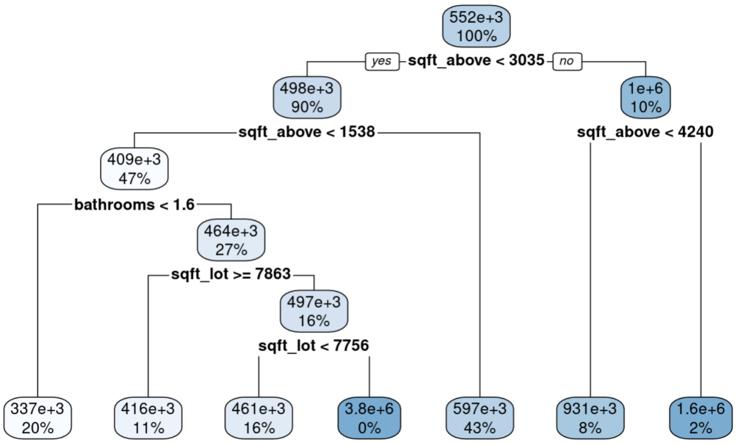


Figure 7: The Random Tree for the national dataset

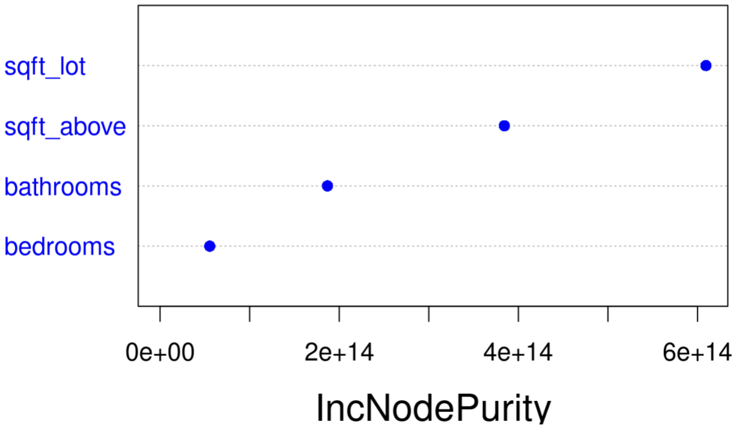


Figure 8: The Result for random forest bagging method

Then I built a random tree for the Seattle house price dataset, the dataset combined by the other two datasets. The tree is different from the tree above, indicating the number of bedrooms is important for the house price in Seattle house market, which is also coherent with our linear regression model. In the bagging model, the result shows the rank of the importance is the total area, the size of the lot, the number of bedrooms, and finally the number of bathrooms.

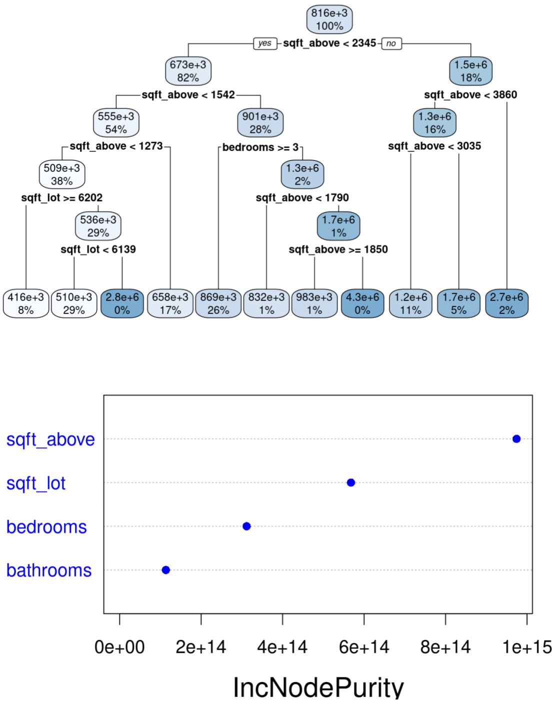


Figure 9: Random forest model for Seattle house price dataset

4 Discussion

4.1 Limitations

There are several general limitations to the project. First of all, the datasets are from Kaggle, and they may have disadvantages and error themselves. So our result may have bias from the beginning. There are also limitations in the model, we can conduct a more complicated random forest model if we have better hardware, and we can reduce the time we run it every time, so we can tune more hyperparameters.

4.2 Conclusion

The results indicate that the number of bathrooms and the size of the house are the two most significant factors affecting house prices in the U.S. housing market. Additionally, in the Seattle market, the number of bedrooms also becomes important for housing prices. This shows the difference between the Seattle house market and the national house market.

However, as mentioned above, the models have limitations in complexity, so the performance might be changed if larger models were used or the models were tuned differently. It may be of interest to discuss how scaling and tuning the models changes their performance.

To take one more step further, the comparison can be extended to factors that affect the house price in different states in the United States, and even compare between different country. And we can research more about the culture and the economy from the factors.

5 References

- Capozza, D. R., Hendershott, P. H., Mack, C., Mayer, C. J. (2002). Determinants of real house price dynamics.

  https://www.nber.org/papers/w9262

- Englund, P., Ioannides, Y. M. (1997). *House price dynamics: an international empirical perspective..* Journal of housing economics, 6(2), 119-136.
  https://www.sciencedirect.com/science/article/abs/pii/S1051137797902102

- SHREE (2018). *House price prediction.* Kaggle.
  https://www.kaggle.com/datasets/shree1992/housedata/data

- SAMUEL CORTINHAS, H. (2023). *HOUSE PRICE PREDICTION - SEATTLE-Real house price data from Seattle, Washington.* Kaggle.
  https://www.kaggle.com/datasets/samuelcortinhas/house-price-prediction-seattle