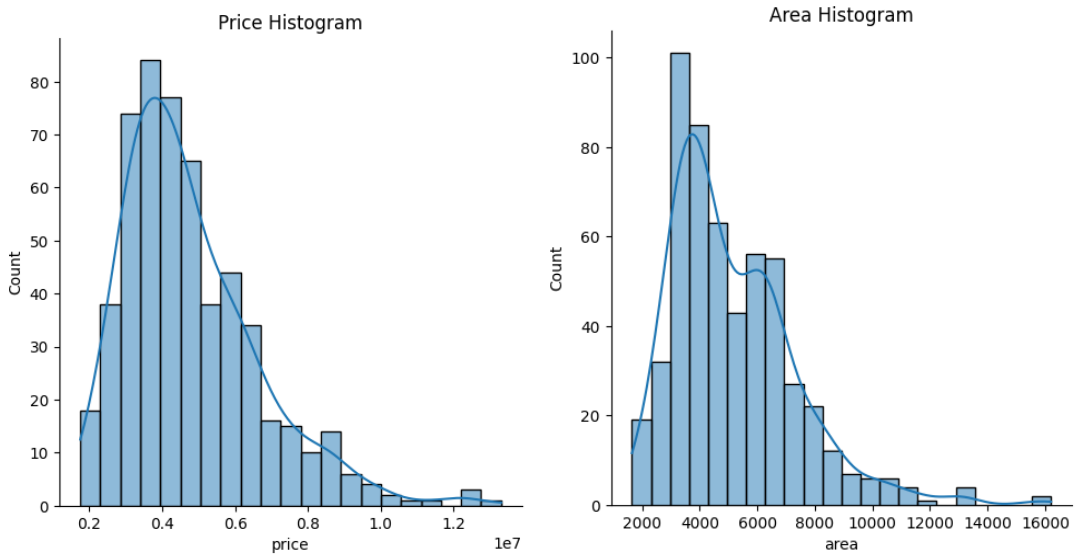
Housing Prices Prediction

Sang Uk Park
Brown University
https://github.com/sangupark?tab=repositories

Introduction:

As mentioned in the project proposal, this project was done to see what factors have the greatest impact on housing prices. This was extremely interesting to look at as housing has a great impact on an individual's life, whether they are an investor or a purchaser. It is not a decision people get to make often in their lives so making sure to look at the factors to accurately predict and possibly think holistically about the future of the house's prices is something that's extremely important.

EDA



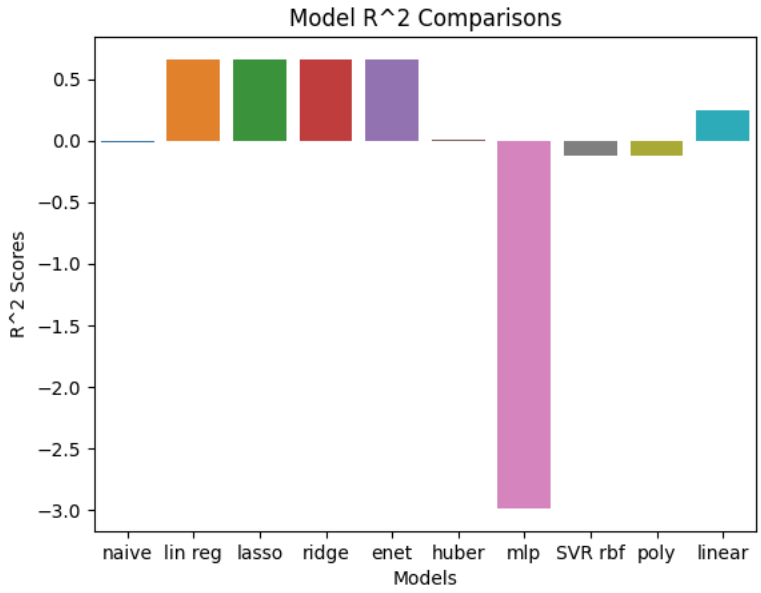
As we look at the distribution of the house, we can see that there is a slight skew, but this is due to the few extreme outliers that occur as a result of the luxury houses present in the dataset (houses with 6+ bedrooms, etc). If we exclude those data points, the distribution of the prices would be roughly normally distributed. It is interesting to see from the histogram of the area that the area is also distributed in a similar fashion. This makes sense as the price and the area should roughly be linearly correlated in my view.

Methods

The data was split into 80% training and 20% testing. I tried various Machine Learning algorithms. They are the following: linear regression, lasso, ridge, elastic net, huber rregressor, MLP, support vector regression (RBF, Polynomial and Linear). I used the R^2 score to explain as this is a regression model and we are measuring the goodness of fit.

Results

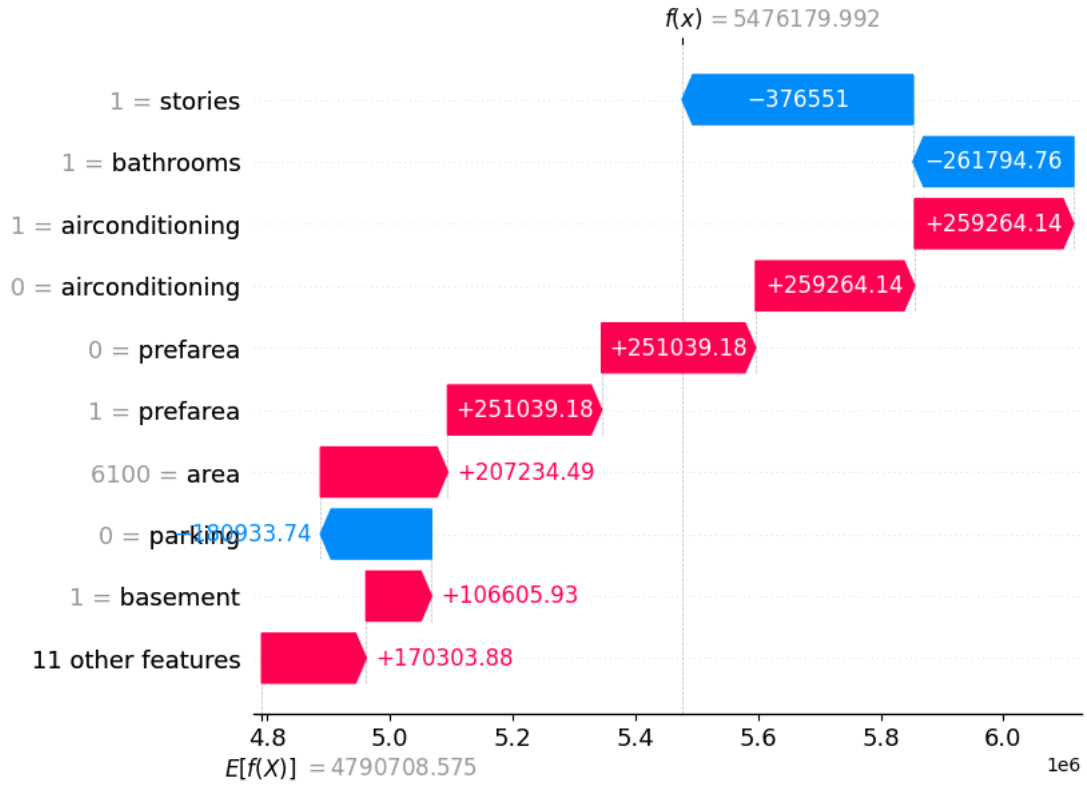
The baseline score was -.014 and linear regression, lasso, ridge, and elastic net all significantly outperformed the baseline with the scores of 0.654, 0.658, 0.658 and 0.658 respectively. However, the remaining algorithms were quite poor in comparison to the previously mentioned algorithms and if not worse to the baseline.



The features that were the most important were area, number of bathrooms, stories and the availability of air conditioning. This was largely expected as they seem to be the primary features people look for when looking to either selling or buying houses.

```
area                0.15996 +/- 0.03125
bathrooms           0.15258 +/- 0.02243
stories             0.07344 +/- 0.02531
airconditioning     0.03008 +/- 0.00878
basement            0.00775 +/- 0.00235
guestroom           0.00698 +/- 0.00238
```

One factor that was surprising was that the number of bedrooms was not in the top four factors. However, that might be chalked off fo the fact that it is redundant with area and also the number of bedrooms does not necessarily correlate to higher volume.



This was the SHAP plot.

Outlook

I believe that this data set was oversimplified as it overlooked the socioeconomic status of the neighbourhoods, regions, and other smaller factors that might have a greater impact on the housing price.

Reference

https://www.kaggle.com/datasets/yasserh/housing-prices-dataset