Thanh Luong

9th March 2023

**Perth Housing Data**

**Dataset Description**

Our dataset is named Perth Housing Data and was extracted from Kaggle. The dataset contains information about the house prices in Perth, Australia, and any factors that may influence it. It originally consisted of 33656 rows and 19 columns. The data represented in the columns were: Address, Suburb, Price, # of Bathrooms, # of Garage places, Land Area in m2 , Internal floor area in m2 , Build year, Distance to Centre of Perth, Nearest Station Distance, Month & Year property was sold, Postcode, Lat & Long, Nearest School location and distance, and the ranking of the nearest school.

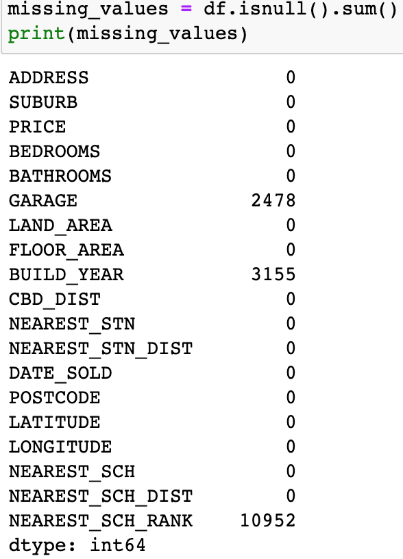
**Business Question**

What factors most influence the House Prices in Perth, Australia? This information can further be used to predict house prices in that specific area.

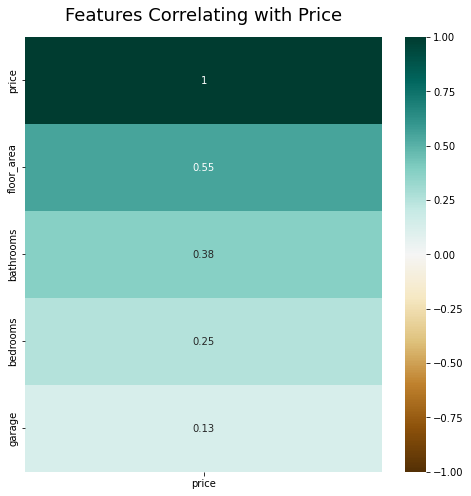
**Data Cleaning and EAD**

When exploring the data and checking for missing values we utilized the function isnull(). There was a considerable number of null values in the garage column, build year, and nearest school ranking.

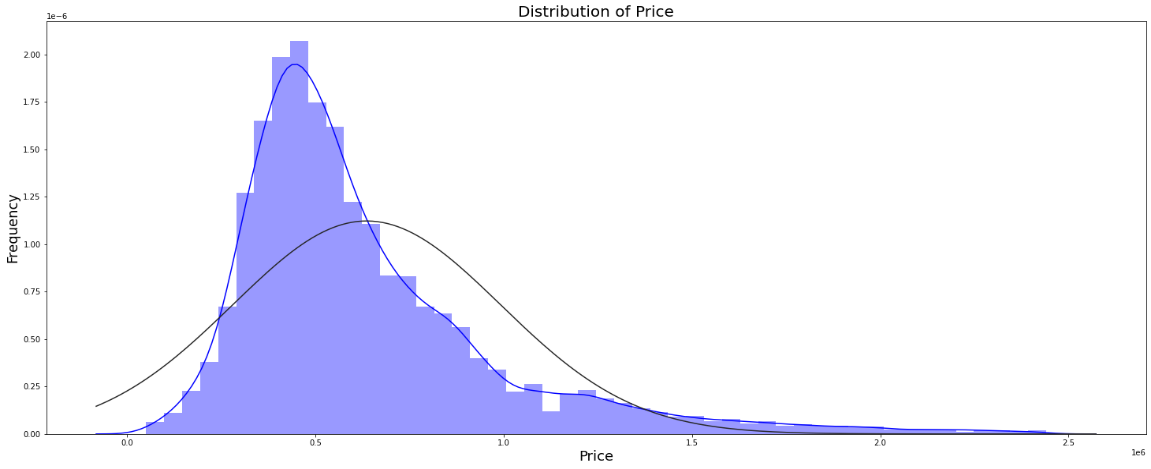
To handle missing values, we decided to impute the median for the attribute ‘Build\_Year’ and the mean for the field ‘Garage’. Due to the substantial number of missing values, the ‘nearest\_sch\_rank’ was dropped. Lastly, we removed duplicates from the ‘Address’ attribute. After we made all the previously described changes, we checked for missing values in our dataset and found none.

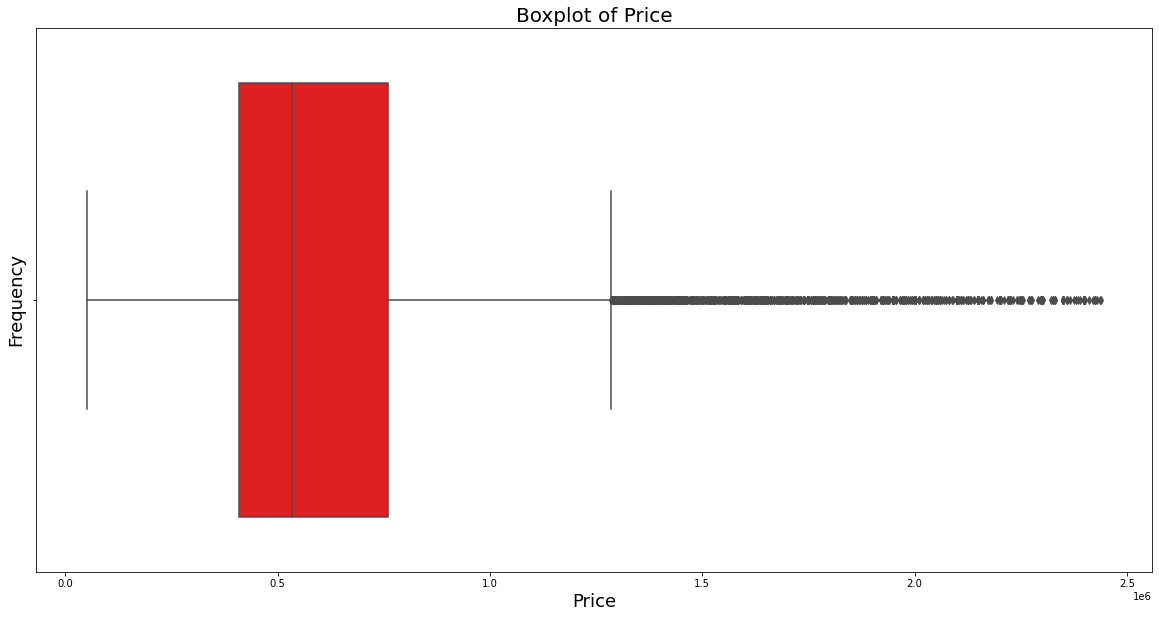


For EAD, we created a heatmap to evaluate the correlation coefficients among the predictors and the price attribute. There was a high correlation with floor area, # of bathrooms, and # of bedrooms. (See below)



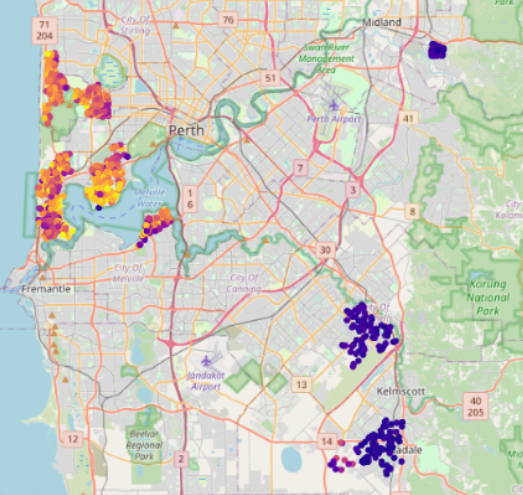
When exploring the distribution of price, we discovered that the attribute did not present a normal distribution. Instead, it was skewed to the left and was leptokurtic with a kurtosis value of 3.998462. All of these metrics indicated that there were outliers present in our dataset; to further confirm this information we created a boxplot.





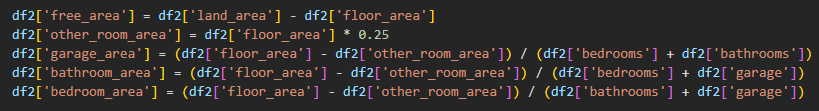
We were able to identify that the median price was 535,000 and that all prices above 1,285,000 were considered an outlier. When conducting our analysis, we found 2,100 outliers. The outliers were not excluded from out dataset in order to have an accurate representation of the real prices in the area of Perth, Australia and avoid bias in our analysis, as well as reveal non-linear patterns within our dataset.

We proceeded to create a choropleth map to have a better visualization of the house prices in the area of Perth. The yellow dots indicate the houses with the higher prices while the purple dots indicate the houses with lower prices in the city. As we can see in the image below, most of the lower priced houses are located on the outskirts of the city while the higher priced houses are located near the coast. Usually this is a pattern that can be observed in several popular cities. Properties located near the ocean tend to have higher prices for both insurance and property value due to their location.



**Feature Engineering**

New variables were created from the already existing ones to gain better insights of our data. ‘Free\_area’ represents the lot of land where the house is not included (e.g., backyard). ‘Other\_area’ represents the space of the house where the bedrooms and bathroom are not included (e.g., living room, kitchen, dining room, closet). ‘Garage\_area’, ‘Bathroom\_area’, and ‘Bedroom\_area’ were computed from the actual number of rooms and the available area of the house.

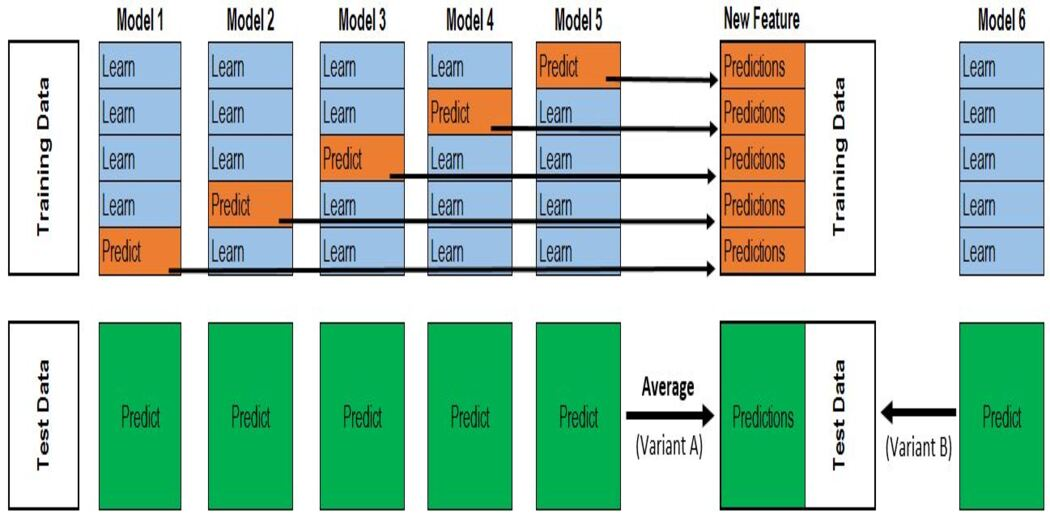


Then, we utilized categorical encoding in the attributes ‘ nearest\_stn\_’, ‘suburb’, and ‘nearest\_sch’ to be able to convert categorical data into a numerical format that machine learning models can handle.

**Models**

We used a pipeline approach, meaning that we fed and ran all models with the same data in order to be able to identify the outperforming model. We conducted all the following models: Linear Regression, Decision Tree, Random Forest, KNN Neighbors, XGBR Regressor, and Stacking CV Regressor, LASSO, Elastic Net.

The advantage of using stacked cross-validation is that it accounts for both variance and bias in the model providing a more accurate estimate of its performance. (See below)

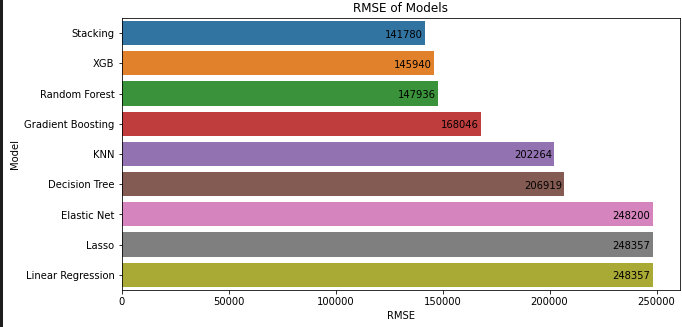


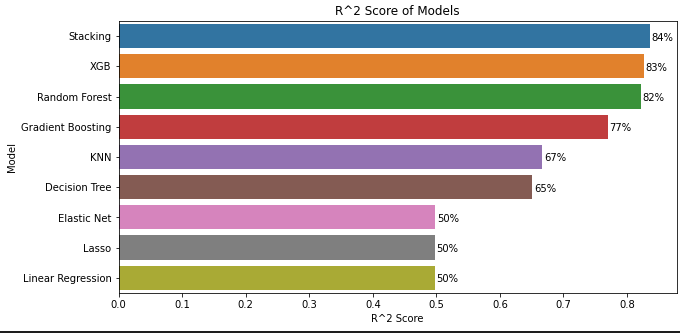
We decided to split out data 80% training and 20% test due to the size of the dataset and provided us with the most optimal end results.

**Results**

Our results were evaluated based on two metrics: RMSE and R^2 scores. We decided to use these metrics since this is a regression problem and the data is continuous.

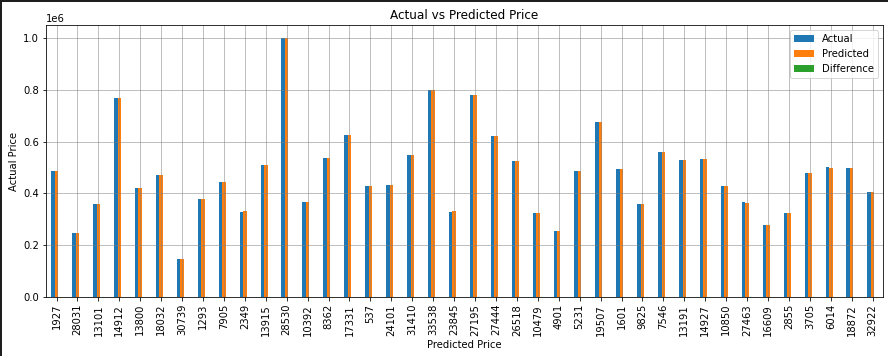
When comparing the performance of all models, stacking cv outperforms its competitors by having the lowest RMSE (141780) and the highest R^2 scores (84%). On the other hand, the performance of linear regression was deficient. The model has the highest RMSE score (248,357) and the lowest R^2 (50%).





**Conclusion**

By using the stacking model, we can adequately answer our business questions. 84% of the variance in price can be explained by the independent variables with this specific model, yielding a small difference between the actual and predicted price value. (See below)



However, we are aware that every machine-learning model has its own pitfall. The stacking model can take significantly longer time to train and require more memory, which can be computationally more expensive.

A few areas for improvement to the models that we were able to identify are the addition of more relevant features to the dataset, hyperparameter tuning, and evaluating other potential models to perform linear regression.