|  |
| --- |
| DSC CAPSTONE 2 |
| Springboard |
| April 2019 |
| Rajesh V Subramanyam |

Introduction

This project involves analysing the trends in real estate market in a populous city such as Melbourne, Australia. Although this project attempts to use the data from a real-estate website such as DOMAIN.COM.AU, this project is not intended for particular organisation / entity. Goal of this project is to predict the real estate prices based on the various factors influencing the market. This project would help an individual or an entity in making informed decisions prior to purchasing or investing in Melbourne real estate market.

Background

The Australian Dream or Great Australian Dream is a belief that in Australia, home-ownership can lead to a better life and is an expression of success and security. Apart from the first home buyers, Investors are interested in building up their real estate portfolio due to a decrease in Banking interest rates.

Whether it is for first home buyers or Investors alike, this project will guide/ help them in analysing various trends that influence the price of properties in Melbourne

Dataset of the Features

There was an attempt to call the API's and use the sales history recorded in Domain.com.au. However, due to the relatively less number of records retrieved in the API call, we started using the historical sales figures available as a CSV file for the Melbourne city.

Listed below are the features of importance in the CSV file:

(1) Suburb

(2) Rooms

(3) Type

(4) Price (Target / Dependent feature)

(5) Bathroom

(6) Car

(7) Landsize

(8) BuildingArea

(9) CouncilArea

(10) Regionname

Features that need imputation to handle the NAN records

There are nearly 35K+ Data points available in the CSV file and some of the features have NAN records. As a first step, we are eliminating the outliers and this would bring down the number of NAN's. Listed below are the outliers removed from the dataset:

1) Land size more than 2000 m2

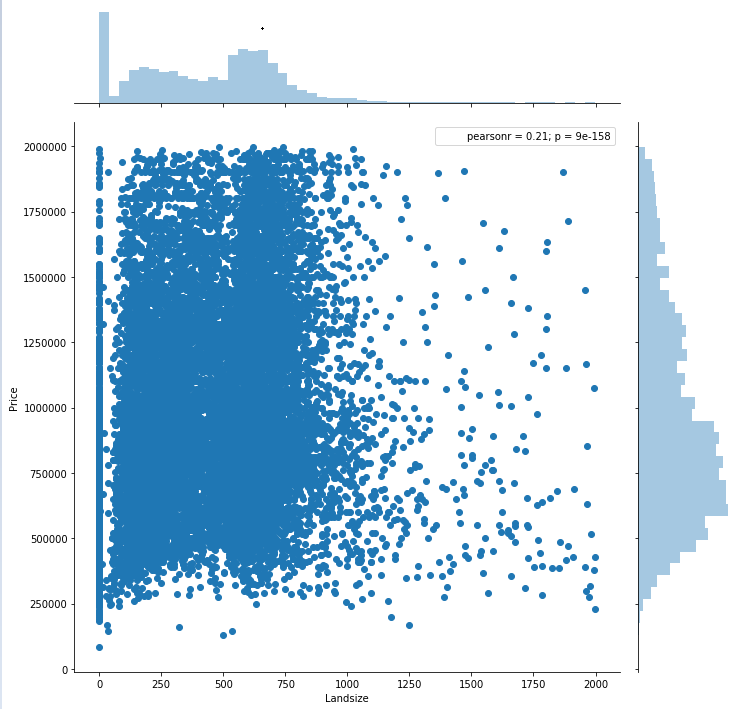
2) Price of the house is more than 2 million AUD

3) Number of rooms in the house more than 8

Analysis of the dataset

Joint plot to show the variation of Price with Land size of the house

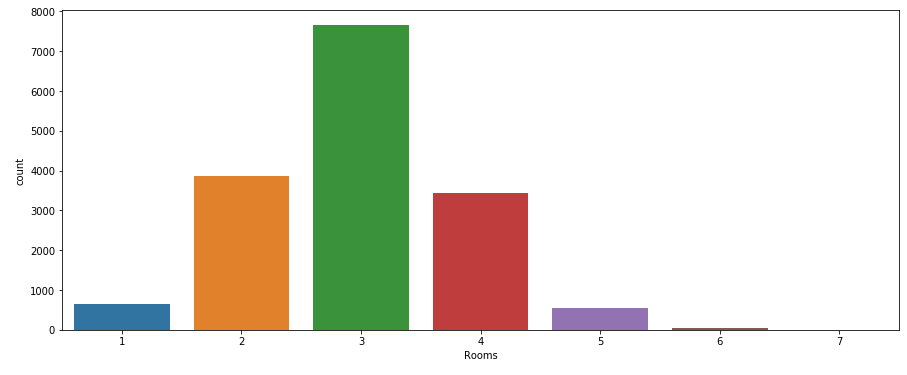
A joint plot was constructed to analyse the variation of price of the house with respect to Landsize. We find that there are more houses in the price range of 7,50,000 to 1 Million AUD



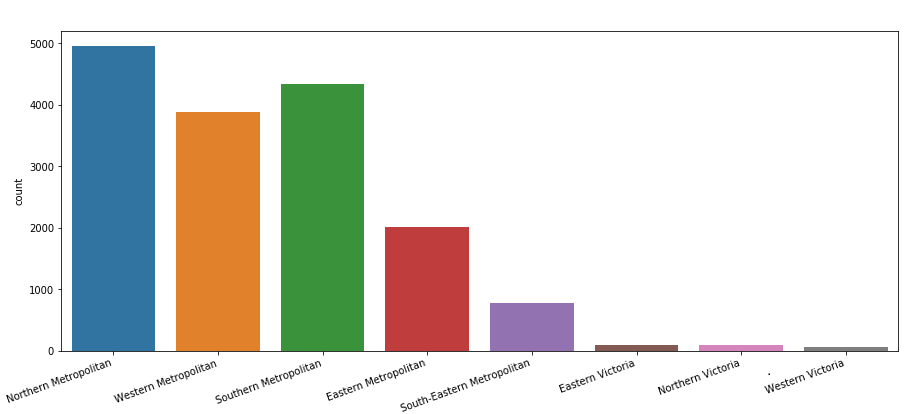
Influence of other features in the dataset

Bar chart was plotted to analyse the other features involved in the dataset such as:

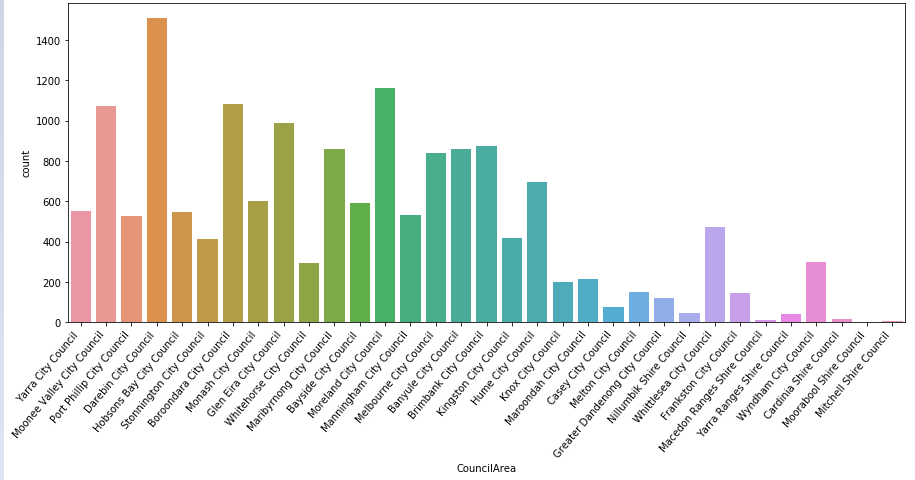
1. Number of Rooms



1. Council Area



1. Region Name



Based on the above plots, we could infer that:

1) DareBin, Moonee Valley, Boroondata, and Moreland City are the primary city councils with highest count of houses.

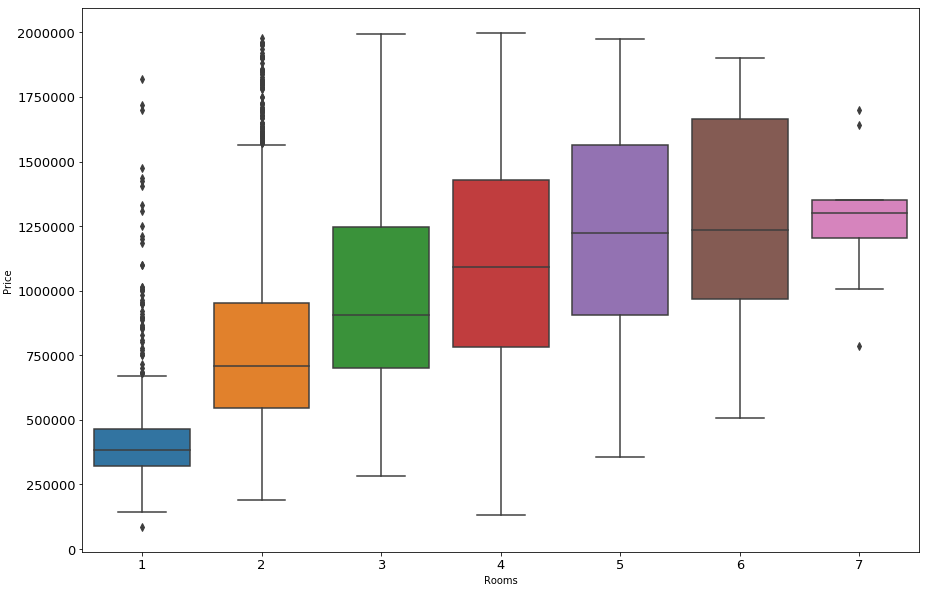
2) Number of 3 bedroom houses are double the number of 2/4 bedroom houses. So the preference when building/ buying a house seems to be 3 bedroom one.

Influence of 'Number of Rooms' on Price of the house.

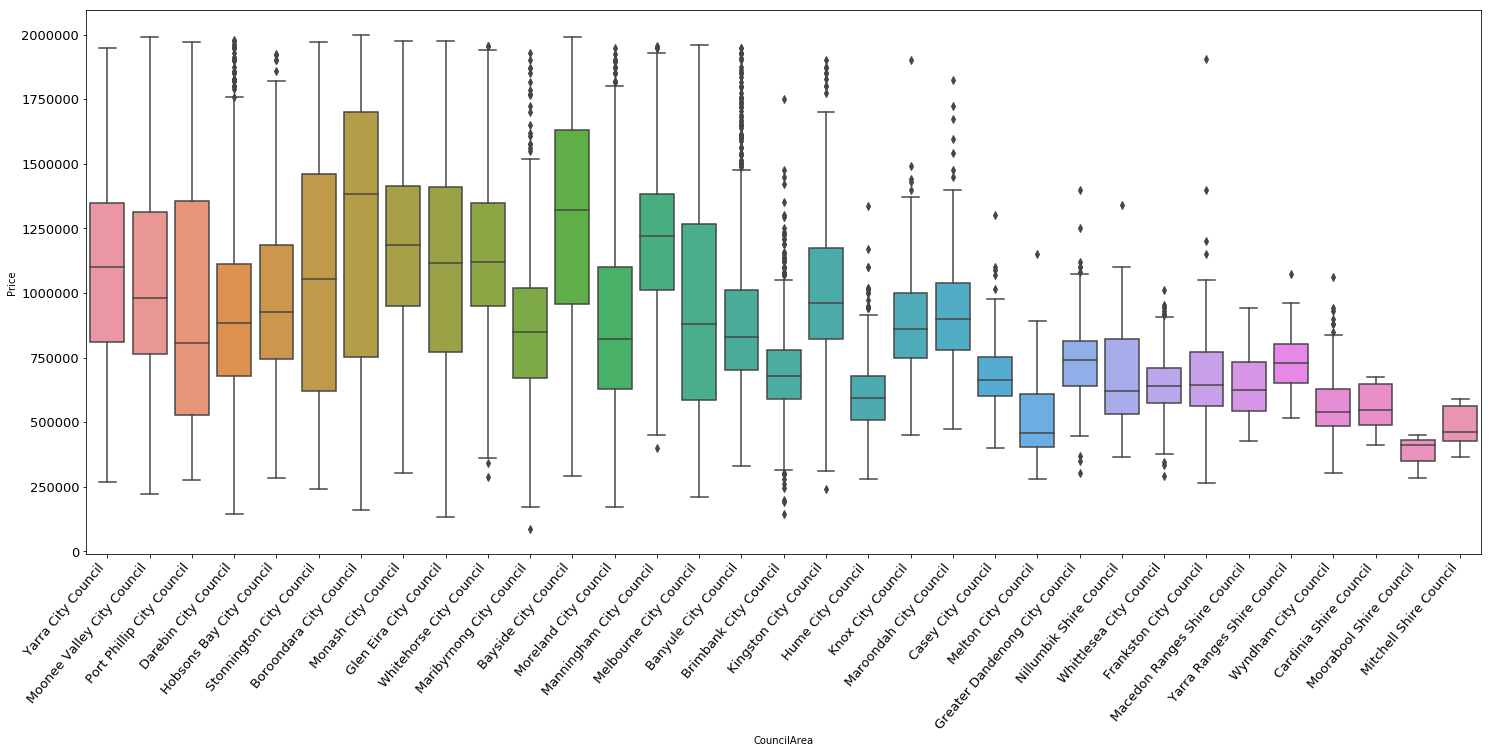
According to the plot shown below

1) It looks like the feature 'Number of Rooms' varies directly with the price of the house.

2) One more interesting insight is that both 5 and 6 bedroom houses are competing in the same price range



Valuation of CouncilArea with respect to Price



On analysing the above plot, we could infer that:

1) Baroondara city council has the highest median price whereas Moorabook shire council has the lowest.

2) Both Whitehorse city and Maribrynong city council have the same price range. By removing some extreme values, we can get better insights on these features

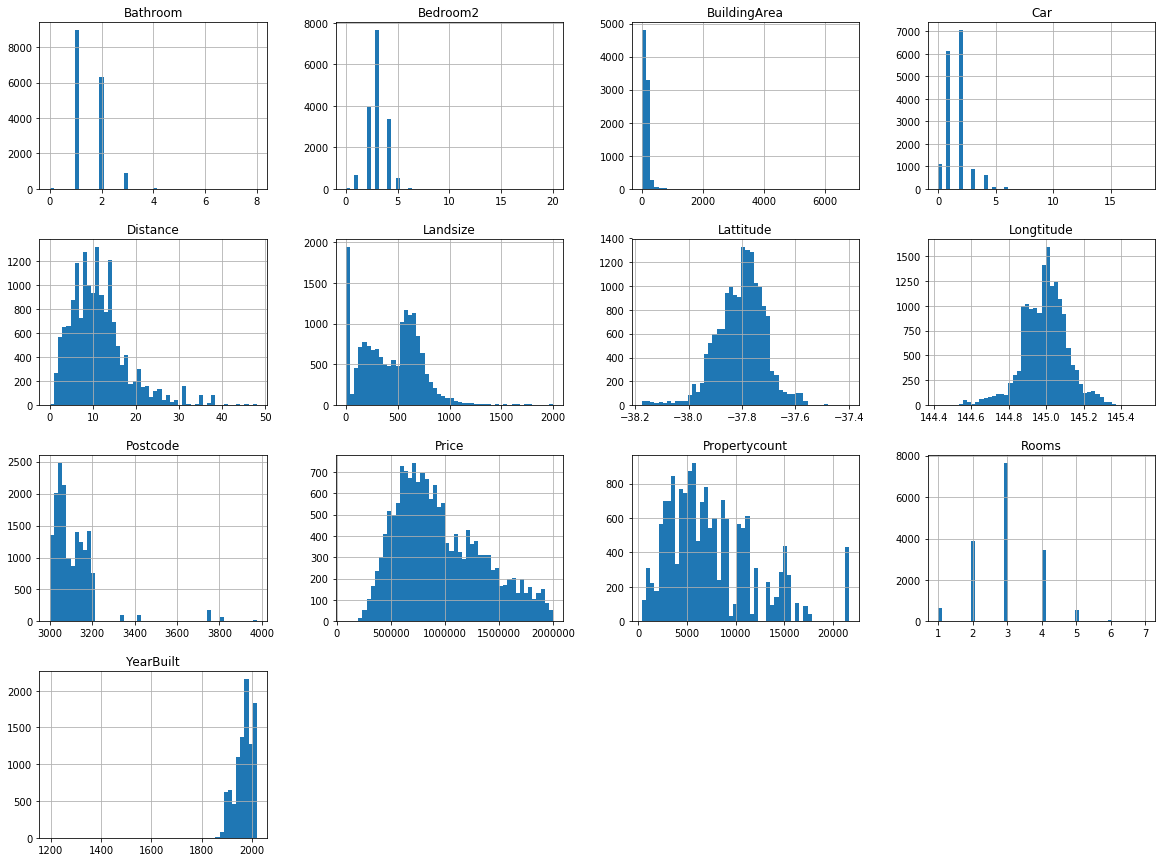
Analyse and Compare the histogram Results

According to the histogram plot results shown below:

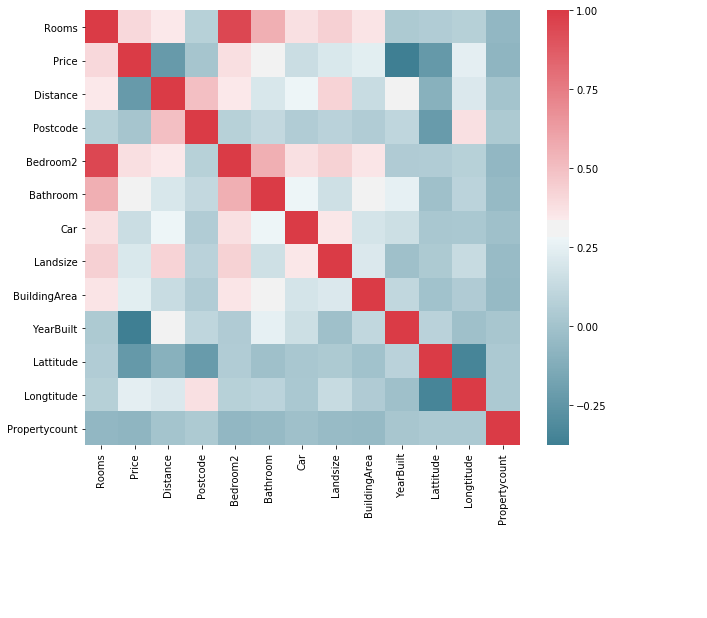
1) 3 bedroom houses with the provision to park 2-3 cars are mostly preferred.

2) Houses nearer to the city are mostly preferred (or) Most houses are built near the city based on the results in the plot of 'Distance'

3) There is a preference for the landsize in the range of 500 to 700 m2



Heatmaps of the features and their correlation to the Target



Based on the heatmap results:

1) Features such as 'No. of rooms', 'No. of Bathrooms', 'Building area' and 'PostCode' (location) has a positive correlation with the price.

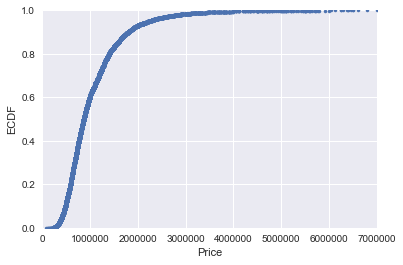
2) Feature 'Number of Rooms' has the highest percentage of correlation with Target among other features.

|  |  |
| --- | --- |
| Feature | Co-relation to Dependent feature ‘PRICE’ 1.0 |
| Rooms | 0.402590 |
| Bedroom2 | 0.385279 |
| Bathroom | 0.302126 |
| Longtitude | 0.242001 |
| BuildingArea | 0.238583 |
| Landsize | 0.207768 |
| Car | 0.145257 |
| Postcode | 0.009676 |
| Propertycount | -0.075756 |
| Distance | -0.227055 |
| Lattitude | -0.232182 |
| YearBuilt | -0.378560 |

Inferential Statistics

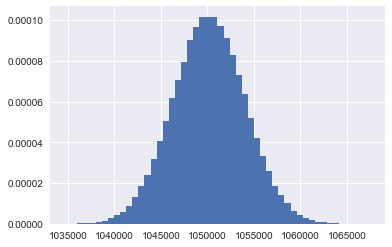
Plotting an ECDF on Price

Based on the ECDF/Price plot, we can infer that 95% of the properties have a price tag of 2 Million or more



Draw bootstrap replicates and check whether the price is normally distributed.

Plot based on bootstrap replicates appears to be normally distributed



C:\Users\Lavanya\AppData\Local\Temp\Temp2_Archive.zip\Cap2_Img11.png

Null Hypothesis: Median housing prices in Melbourne is greater than 1 million AUD?

One Sample T-test and Z-test is conducted to check whether the median housing price exceeds one million AUD. Based on results, we are rejecting the Null hypothesis (as P-Value is less than 0.05 or signiﬁcantly lesser) for both these tests and accepting the alternative hypothesis that Median value of housing price may not be more than 1 Million AUD (in Melbourne).

Machine Learning: Building a Linear Regression Model

As a part of this exercise, we are building a linear regression model using Simple stats, sci-kit learn libraries and compare the results with those of XGBoost Machine learning model.

Listed below are the features used to construct this model:

1) Rooms

2) Price

3) Bedroom

4) Bathroom

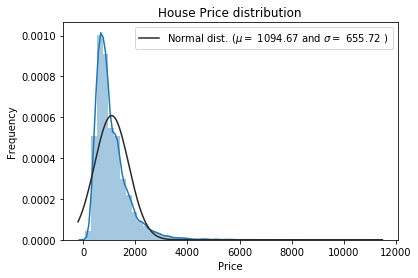
5) Car

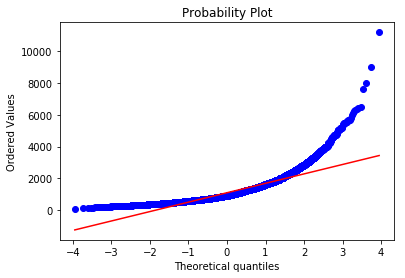
6) Landsize

7) Distance

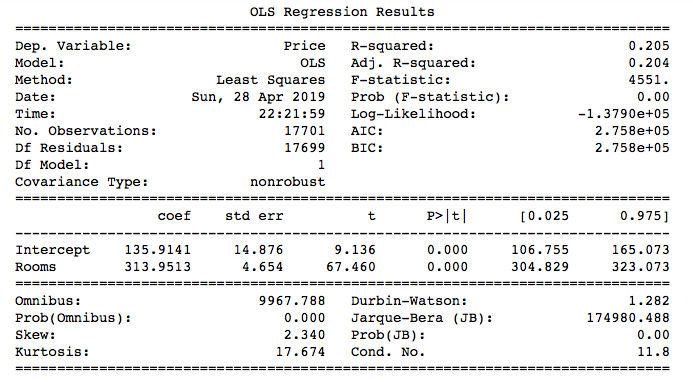
Variation of the Target

The target feature - Price is right skewed. As linear models like normally distributed data, we will transform SalePrice and make it more normally distributed by considering log(Price).





Simple Stats Linear model



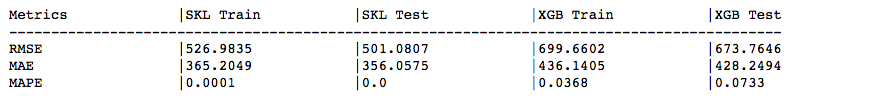
Based on Simple stats model Linear regression results, we ﬁnd that:

1) The coefficient of 313.95 means that as the RM variable increases by 1 i.e. predicted value of Price increases by 313.95(multiplied by 1000) i.e. 313950. So for every addit ional room added to the house, the house price increases by 313950

2) 95% confidence interval varies from 304.829 to 323.073 (multiply by 1000 for actual Price value).

3) The percentage of variance our model explains i.e. R-squared value i s 0.205

Compute Mean absolute error, Mean absolute percentage error & Root mean Squared error using sci-kit learn and XGBoost

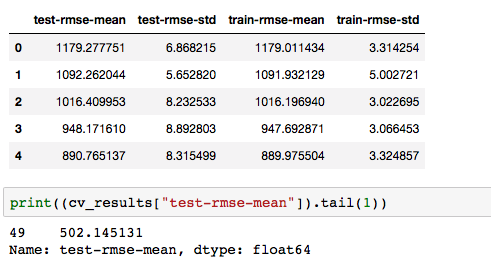


In the above table, SKL – Scikit learn Linear regression & XGB – XGBoost. Based on the results:

1) We could notice signiﬁcant change in the metrics between SKLearn and XGBoost results.

2) Mean absolute percentage error is showing 7% of error for the XGB-test dataset. However, we had a insigniﬁcant result for SKLearn test. We can try to improve the metrics(such as RMSE) by performing KFold cross validation

K Fold cross-validation using XGBoost

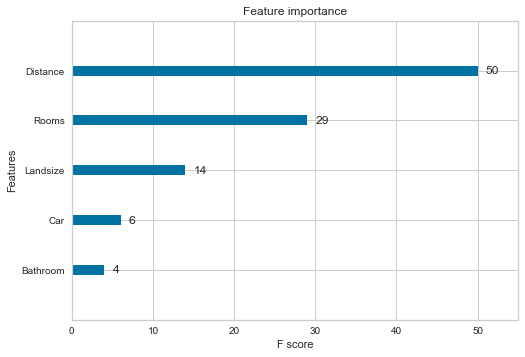


Based on the above observations, we can ﬁnd that Test RMSE has reduced from 673.76 to 502.14. RMSE can be further improved by:

1) Changing the hyper-parameters

2) Applying techniques such as Grid search, Random search and Bayesian optimization to reach optimal set of hyper-parameters.

Visualizing the importance of features in our dataset



Importance of features is analysed using the XGBoost algorithm. Above plot conﬁrms the fact "Distance" (a feature related to Location of the house) is much more important than any other feature in the dataset. Also, the feature "Distance" has the potential to override the importance of other features in price determination.

Conclusion & Recommendations to the client

We had constructed Linear regression models using SKlearn & XGBoost. Listed below are the observations based on the results.

1) SkLearn results showed that there is a room for improving the model further. Error metrics such as 'Root Mean Squared error', 'Mean absolute percentage error' etc proved that model is 70 - 75% accurate although there is room for improvement

2) Although the plot for Target 'House price' distribution appears to be right skewed and not linearly distributed. This can be corrected by taking logarthmic scale of target 'House price'.

3) Linear regression using XGBoost showed better results in terms of error metrics - Particularly after the usage of K-Fold cross validation techniques.

4) Using the XGBoost technique, we were able to compare the F-Scores of all independent features and understand the importance of each and every features with respect to the Target (Price of the property)

5) This linear regression model can be further improved by hyper-parameter tuning i.e. by applying techniques such as Grid search, Random search and Bayesian optimization to reach optimal set of hyper-parameters.

As a part of client recommendations, it would be better if we have more data points for analysis and care should be taken to ensure that there are minimal NANs in the features of new data points (although imputation can be performed to correct the features, it would be better to not have the NANs in the ﬁrst place). Although this model is good at predicting the prices, 'New' features can be added to the dataset such as 'Crime Rate', 'Presence of Railway routes/ ease of transportation', 'School Rankings' etc. These features are as inﬂuential as 'Location/ Distance from CBD' and if such features are added, there will be more than 15% improvement in the model and thereby predicting house prices more accurately.