Capstone project 2 Milestone Report -Predicting house prices based on Australian Real estate data

Client:

Any Stakeholders interested in predicting the housing prices knowing their features.

Description:

This capstone project deals with real time housing data provided by the Australian realestate web site domain.com.au.Although this website lists housing prices Australia wide, we would like to focus on the subset for this project.

Goal of this project is to build a machine learning model that predict prices of houses as a function of different features that characterise them.

Dataset:

Australian real estate website - Domain.com.au

API:

https://api.domain.com.au/v1/salesResults/Melbourne

The above URL calls the API to extract the sales results. Therefore, one needs to have relevant authorisation to access this data. 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.

Data Wrangling:

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:

(i) Land size more than 2000 m2
(ii) Price of the house is more than 2 million AUD
(iii) Number of rooms in the house more than 8

Listed below are some observations made during data wrangling (For more details, refer to 'Data wrangling' document of Capstone project 2).

1. DareBin, Moonee Valley, Boroondata, and Moreland City are the primary city councils with highest count of houses.
2. There is a preference for the landsize in the range of 500 to 700 m2
3. 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'
4. Baroondara city council has the highest median price whereas Moorabook shire council has the lowest.
5. One more interesting insight is that both 5 and 6 bedroom houses are competing in the same price range.
6. 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.

Inferential statistics:

As a part of inferential statistics, following have been analysed:

1. ECDF plot was created for non-categorical column 'PRICE' (Selling price of the house) and we could observe from the plot that 95% of the properties had a price tag of 2 million AUD or more. This proves the fact that housing prices had skyrocketed in the recent years - This could also be attributed to the recent housing rates decreased by RBA (Reserve bank of Australia).
2. Bootstrap replicates are drawn randomly and it is found that the target 'Price' is normally distributed.

3. 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).

Next Steps:

Moving forward, A machine learning model on linear regression will be created and analysis would be conducted with respect to the accuracy of the model and its performance.
Also, the linear regression model based on Sci-kit learn and XGBoost libraries will be created to observe the differences in metrics and explain the reason for such differences in the first place.

After completion of ML Model, a final Report and presentation deck will be submited to conclude the Capstone project 2. All the code Submissions are present under the GITHUB:
https://github.com/vsrajesh1/Capstone_Project-2