# HOUSING: PRICE PREDICTION(Surprise Housing)

## INTRODUCTION

### • Business Problem Framing

**• A US bases company wants to enter in Australian real estate market.**

Company uses data analysis currently to buy property and sell at high price. They have decided to build a machine-learning model for the same and in the same context and they have collected some data point in the csv file. Basis on collected data, a machine-learning model to be built on and need below important things to be revealed. 1. Which variables are important to predict the price of variable? 2. How do these variables describe the price of the house? Once model is built, the management to see how prices vary with the variables so that we can make their strategy and make profit will use it. Model will also make them help to understand the market dynamics

**• Conceptual Background of the Domain Problem**

One should have good understanding about the component, which are used to see property values. Many features are derived from other features. Understanding about the domain knowledge will provide better understanding on the data

**• Motivation for the Problem**

Undertaken Since home is one the important key for everyone. Understanding various component, which involve deciding the price is interesting and it helps to make good decision.

## Analytical Problem Framing

**• Mathematical/ Analytical Modeling of the Problem**

Have applied curtain statistical method to check the data distribution, outlier and making decision accordingly. I have used statistical method to check the skewness in the data and corrected by using some mathematical method like sqrt and cube and boxcox function. Used correlation function to check the correlation among the variable and relation between target and individual feature.

- **Data Sources and their formats**

The company collected data and it is stored in csv file, training and test data was given separately

- **Data Preprocessing**

Done While checking the data, it is found that some of the columns have missing values more of than 50% and some have somewhat. I have decided to drop those columns that are having more than 20% missing values and rest are imputed with MODE, MEDIAN. I have dropped columns that had no variance or single categorical values across. Some columns are transformed like builtyear and columns that have value like year into number of years Columns are labelled into categories as we have many columns that are having categories. Skewness has been corrected by using sqrt,cbrt and boxcox metthod I have seen some outlier in the data but left them as it is in the data As I made assumption that it is nature of the columns

- **State the set of assumptions (if any)**

Related to the problem under consideration According to the data, property plot will be bigger and smaller size Those cannot be outlier, I have taken this as assumption and didn't work on correcting ourlier.

## • Hardware and Software Requirements and Tools Used

I have used jupyter notebook and below library for working on data until modelling

```
import pandas as pd
 import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings as war war.filterwarnings("ignore")
pd.set_option("max_rows",100)
 pd.set_option("max_columns",150)
pd.set_option("display.precision",2)
pd.set_option('display.float_format', '{:.3f}'.format) from scipy.stats
 import boxcox from sklearn.model_selection
import GridSearchCV, cross_val_score, train_test_split from sklearn.metrics
import mean_squared_error,
r2_score from sklearn.ensemble
import RandomForestRegressor,
```

GradientBoostingRegressor,

AdaBoostRegressor

from sklearn.svm import SVR

from sklearn.tree

import DecisionTreeRegressor from sklearn.linear_model

import LinearRegression from sklearn.preprocessing

import RobustScaler

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

# Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

- **Testing of Identified Approaches (Algorithms)**

Listing down all the algorithms used for the training and testing.

- **Run and Evaluate selected models**

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

- **Key Metrics for success in solving problem under consideration**

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

- **Visualizations** Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail. If different platforms were used, mention that as well.

- **Interpretation of the Results**

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

# CONCLUSION

## • Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

## • Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

## • Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

Submitted by: Madhurima Srivastava