

**Housing Project**

Submitted by:

Shraddha R. Waghmare

**ACKNOWLEDGMENT**

I would like to express my deepest gratitude to my SME (Subject Matter Expert) Khushboo Garg as well as Flip Robo Technologies who gave me the opportunity to do this project on Housing Project, which also helped me in doing lots of research which increases my analysis skills more.

Also, I have utilized a few external resources that helped me complete the project. I ensured that I learnt from the samples and modify things according to my project requirement. All the external resources that were used in creating this project are listed below:

1) <https://www.google.com/>

2) <https://www.youtube.com/>

3) <https://scikit-learn.org/stable/user_guide.html>

4) <https://github.com/>

5) <https://www.kaggle.com/>

**INTRODUCTION**

* Business Problem Framing
* Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the major contributors to the world’s economy. It is a very large market and there are various companies working in the domain. Data science comes as a very important tool to solve problems in the domain to help companies increase their overall revenue, and profits, improve their marketing strategies and focus on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, and recommendation systems are some of the machine learning techniques used for achieving the business goals of housing companies. Our problem is related to one such housing company.
* House price prediction can help the developer determine the selling price of a house and can help the customer to arrange the right time to purchase a house. House Price prediction is important to drive Real Estate efficiency. As earlier, House prices were determined by calculating the acquiring and selling price in a locality. Therefore, the House Price prediction model is very essential in filling the information gap and improving Real Estate efficiency.
* The aim is to predict efficient house pricing for real estate customers with respect to their budgets and priorities. By analysing previous market trends and price ranges, and also upcoming developments future prices will be predicted, cost of the property depends on the number of attributes considered.
* Now as data scientists our work is to analyse the dataset and apply our skills toward predicting house prices.
* Conceptual Background of the Domain Problem

Houses are one of the necessary needs of each and every person around the globe and therefore housing and real estate market is one of the markets which is one of the significant contributors to the world’s economy. It is a very large market, and various companies are working in the domain. Data science comes as an essential tool to solve problems in the domain to help companies increase their overall revenue, and profits, improve their marketing strategies and focus on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, and recommendation systems are some of the machine learning techniques used for achieving the business goals of housing companies. Our problem is related to one such housing company.

A US-based housing company named **Surprise Housing** has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company is looking at prospective properties to buy houses to enter the market. We are required to build a model using Machine Learning in order to predict the actual value of the prospective properties and decide whether to invest in them or not. This company wants to know:

* Which variables are important to predict the price of variables?
* How do these variables describe the price of the house?

**Business Goal:**

We are required to model the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

* Review of Literature

Machine learning has been used in prediction for some time now with increasingly better results that were put into practice and changed the economic landscape. Practically every economic domain now benefits from machine learning prediction models, and the current models are becoming more accurate given the computational power available for processing immense sets of data. In this research, the housing price problem is analysed using several machine learning techniques such as XGBoost, KNearest Neighbors, Random Forest, Decision Tree, Linear Regression, and Support Vector Regressor.

Park and Bae (2015) addressed the house price prediction problem considering the housing data available for Virginia's Fairfax County. To solve the problem, the authors have employed machine learning techniques such as Naive Bayesian, AdaBoost, and RIPPER to develop a house price classification model.

He and Xia (2020) studied the housing price problem stressing heterogeneous traders and a healthy urban housing market. Their paper covered the speculative investment effects on house prices and economic disturbance and proposed a dynamic stochastic general equilibrium model to solve the problem.

Nam and Seong (2019) studied stock market prediction problems by analysing media housing market information considering unstructured data and utilizing the asymmetric

relationships of firms.

The housing market is important for economic activities (Khamis & Kamarudin, 2014). Traditional housing price prediction is based on cost and sale price comparison. So, there is a need for building a model to efficiently predict the house price. Khamis compares the performance of predicting house prices between the Multiple Linear Regression model and Neural Network model in New York.

Díaz et. al. (2019) considered the prediction problem of Spanish day-ahead electricity prices. To solve the problem, a regression tree-based approach has been proposed. Moreover, in this problem the dataset, particularly, the model variables are obtained from publicly accessible energy consumption datasets.

In another research, the prediction of the daily bitcoin exchange rate was considered and the behaviour of financial markets was studied (Mallqui & Fernandes, 2019). The authors proposed a method including the machine learning features to solve the bitcoin exchange rate prediction problem.

In another study, Gu et. al. (2011) studied the housing price problem with the aim of forecasting a house price model. A hybrid of genetic algorithms and a support vector machine method was proposed to solve the model. The model dealt with a housing dataset that was collected in China, during the 1993-2002 period.

Plakandaras et. al. (2015) also addressed the U.S. real estate house price index problem. In their research, a novel hybrid forecasting method was proposed combining the ensemble empirical mode decomposition (EEMD) with Support Vector Regression (SVR). The obtained solutions of their proposed model are compared with Random Walk (RW), Bayesian Vector Autoregressive, and Bayesian Autoregressive models.

* Motivation for the Problem Undertaken
* In this project, I have to build a model that calculates the price of houses with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables..
* By doing this project I have got an idea about how to deal with data exploration & model building, where I used my analytic skills to predict the house price using ML models.

* Furthermore, the model will be a good way for the management to understand the pricing dynamics of a new market. The relationship between house prices and the economy is an important motivating factor for predicting house prices.

**Analytical Problem Framing**

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

This particular problem has two datasets, one is the train dataset and the other is the test dataset. I have built a model using the train dataset and predicted SalePrice for the test dataset. By looking into the target column, I came to know that the entries of the SalePrice column were continuous and this was a Regression problem so I have to use all regression algorithms while building the model.

* I observed some unnecessary entries in some of the columns & in some columns, I found more than 80% null values and more than 85% zero values so I decided to drop those columns. If I choose to keep those columns as it is, it will create high skewness in the model.
* While checking the data for null values, I found many columns with NaN values and I replaced those NaN values with suitable entries like mean for respective numerical columns and mode for respective categorical columns.
* To get a better insight into the features I used a heat map wherein I was able to understand the null values which I was unable to see when in a table form due to a large dataset.
* I’ve done EDA on both train and test datasets and merged them to get the best model.
* I’ve used the XGBoost Regression model to get the best model and saved it.
* **Data Sources and their formats**

A US-based housing company named Surprise Housing has collected the dataset from the sale of houses in Australia the data is provided by Flip Robo Technologies and it is in Excel format(which I converted into CSV format).

**There are 2 data sets:**

1. Train dataset

2. Test dataset

* Train dataset will be used for training the machine learning models. The dataset contains 1168 rows and 81 columns, out of 81 columns, 80 are independent variables and the remaining 1 is a dependent variable (SalePrice).
* The Test dataset contains all the independent variables, but not the target variable. We will apply the trained model to predict the target variable for the test data. The dataset contains 292 rows and 80 columns.
* The dataset contains both numerical and categorical data. Numerical data contains both continuous and discrete variables and categorical data contains both nominal and ordinal variables.
* I can concatenate both train and test data, but this may cause data

leakage so I decided to process both the data separately.

* Data Pre-processing Done
* To begin with, I have imported required libraries and both the datasets in CSV format.
* Then I did all the statistical analysis like checking shape, unique, value counts, info etc.
* While checking the info of the datasets I’ve found some columns with more than 80% null values, so these columns will create skewness in the datasets so I decided to drop those columns.
* Then while looking into the value counts I found some columns with more than 85% zero values this also creates skewness in the model and there are chances of getting model bias so I have dropped those columns with more than 85% zero values.
* While checking for null values I found null values in most of the columns and I used an imputation method to replace those null values (mode for categorical columns and mean for numerical columns).
* In the Id and Utilities columns, the unique counts were 1168 and 1 respectively, which means all the entries in the Id column are unique and ID is the identification number given for a particular asset and all the entries in Utilities column were the same so these two columns will not help us in model building. So I decided to drop those columns.
* And all these steps were performed on both train and test datasets simultaneously
* Data Inputs- Logic- Output Relationships

To analyse the relation between features and target I have done EDA.

* Features having high Positive correlation with labels: OverallQual, GrLivArea, ExterQual, KitchenQual, BsmtQual, GarageCars, GarageArea, TotalBsmtSF, 1stFlrSF, FullBath, TotRmsAbvGrd.
* Features having high Negative correlation with label: Heating, MSZoning, LotShape, BsmtExposure, GarageType, Year\_SinceRemodAdded, GarageAge, Year\_SinceBuilt, GarageFinish.
* Hardware and Software Requirements and Tools Used

To build the machine learning projects it is important to have the following hardware and software requirements and tools.

Hardware Required:

* Processor: macOS version 12.4
* RAM: 8 GB or above
* ROM/SSD: 256 GB or above

Software required:

* Anaconda - language used Python 3

Libraries Used:

* import numpy as np
* import pandas as pd
* import seaborn as sns
* import matplotlib.pyplot as plt
* from xgboost import XGBRegressor
* **Interpretation of the Results**
* This dataset was very special as it had a separate train and test datasets. We had to work with both datasets simultaneously. First, the datasets contained null values and zero entries in a large number of columns, so we had to be careful while going through the statistical analysis of the datasets.
* Furthermore, going feature by feature helped us to get a better insight into the data. There were a lot of numerical continuous columns in a linear relationship with the target column.
* We have to use regression algorithms while building various models using a train dataset to get the best model.
* Finally, I have got the important features which affect the price.

**CONCLUSION**

* Key Findings and Conclusions of the Study

In this project report, we have used machine learning algorithms to predict house prices. We have mentioned the step-by-step procedure to analyse the dataset. Thus, helping us select the features which are not correlated to each other and are independent in nature. These feature sets were then given as input to the algorithms. Then we also saved the data frame of predicted prices of the test dataset.

* **Learning Outcomes of the Study in respect of Data Science**

Data cleaning is one of the most important steps to removing missing values and replacing null values and zero values with their respective mean, median or mode. This study is an exploratory attempt to use machine learning algorithms in estimating housing prices and then compare their results.

The application of machine learning in property research is still at an early stage. We hope this study has helped move a small step ahead in providing some methodological and empirical contributions to property appraisal and presenting an alternative approach to the valuation of housing prices.

* **Limitations of this work and Scope for Future Work**

**LIMITATIONS:**

* To begin with, the first drawback is the data leakage when we merge both train and test datasets.
* The second drawback was the null and zero values, which were dealt with using imputation techniques.
* The dataset has many limitations, the main limitation is that we have no information about potential buyers and the environment of the sale. The factors such as auctions can have an influence on the price of the house.
* The dataset does not capture many economic factors. Collecting more accurate and important details about the houses from the buyers will help to analyse the data more clearly.
* Also, this model doesn't predict the future prices of the houses mentioned by the customer.
* Due to this, the risk of investing in an apartment or an area increases considerably. To minimize this error, customers tend to hire an agent which again increases the cost of the process.

**FUTURE WORK:**

* A major future scope is adding an estate database of more cities, which will allow the user to explore more estates and reach an accurate decision.