Project

Housing Price Prediction

Submitted by:

Mallesh R

**INTRODUCTION**

* **Business Problem:**
* Houses are one of the necessary need 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 in 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 the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in house sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for 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 data is provided in the CSV file below.
* The company is looking at prospective properties to buy houses to enter the market. You 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. For this company wants to know:

1.Which variables are important to predict the price of variable?

2.How do these variables describe the price of the house?

### 

### 

### **Business Goal:**

You 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.

* **Our Goal:**

House price can be determined based on many factors like garage, number of bedrooms and bathroom, access to alley extra.

We need to find the features that mostly influence the price of the house, and their relationship with the Sales Price of the house.

**Analytical Problem Framing**

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

This model is a supervised model where the target variable is continuous datatype. Thus we will be using all the best Regression models to find the price of the house. We will determine the best regression model based on the score value and the root mean square error.

* **Data Sources and their formats:**

The dataset has all the information about the house collected by the company which are real world data with lots of observations. The description of each variable are in the below link. [Click here.](https://docs.google.com/document/d/1IvaT9_SCblYoe-JiizK-loUqToCv5coI4rxJY3pL9rE/edit?usp=sharing)

* **Data Preprocessing Done:**

Data Preprocessing is done by starting to clean the missing values in the dataset. There are lots of missing data in each features some of them are replaced with the respective datatype and others where filled based on the data description. Then the skewness of the numerical dataset is reduced by some power transformation technique and the categorical dataset has been encoded for building the model.

* **Data Inputs- Logic- Output Relationships:**

House price can be determined based on many factors like garage, number of bedrooms and bathroom, access to alley extra. We need to find the features that mostly influence the price of the house, and their relationship with the Sales Price of the house Those insights are in the notebook.

* **State the set of assumptions (if any) related to the problem under consideration:**

No, presumptions are taken all are based on the data description and knowledge on the respective field.

* **Tools Used:**

There are some tools used for specific purpose Pandas, Numpy, matplotlib are used for building dataframe and making them to a structural dataset and doing complex mathematical steps and to visualise the insights. SKlearn library is used for creating different models to train and test the dataset.

**CONCLUSION**

**Model/s Development and Evaluation**

The Model building and their performance are evaluated using Root mean square error. The results and the visualisation of relationship between each features and the target variable are shown in the notebook