Automated House Price Prediction Using Machine Learning Model

By: Hamza Waheed

Date: October 31, 2024

Introduction

House is a basic need for every one now a days. With the increase in population day by day demand of house will increase. So, every one need a house for this they want to purchase house. But everyone does not know the price of house. As we know that every month or in a couple of months housing prices increases. This problem occurs due to many reasons like demand increase due to increase in population day by day.

This housing price is a serious problem all over the world. Predicting house price is difficult for anyone. The propose model automatically **AHPP (Automated House Price Prediction)** predict price of the house based on several factors. By this anyone can easily predict price of house by giving minimum requirements. By using proposed model, they will have a rough idea about the price of house that they want to buy. Although, proposed system will be used by real estate field by using this model they can easily predict the price of the respective house. But this model is trained on only five feature that are not enough for predicting price of house.

Dataset and feature

Proposed model AHPP trained on a predefined 'housing_price_dataset.csv' dataset that is present on seaborn library repository. Dataset contains 6 columns and 50K rows in which there are input features like 'SquareFeet', 'Bedrooms', 'Bathrooms', 'Neighborhood' and 'YearBuilt' with one label feature 'Price'. In this dataset there is no missing value but there are 59 outliers in this dataset.

Dataset contain a one categorical feature will ordinal values. I've applied label encoding technique to handle ordinal values. After applying encoding I've applied three different scalers like 'MinMaxScalar', 'StandardScalar()' and 'RobustScaler()' to normalize data to get different outputs. But generalized output is received when I applied MinMaxScalar().

Exploratory Data Analysis

I've plotted a pair plot using seaborn library which show the relation ship between all features. By visualizing pair plot 'SquareFeet' and 'Price' have some relationship because data is linear between them on the other hand data is uniform. Also, data is normally distributed in the 'Price' feature. I've

used box plot for displaying outliers presented in the dataset. After analyzing the boxplot I've come to know that price feature having almost 59 outliers.

After that I've use scatter plot display the distribution of the data. After visualizing the graph, I've come to know that data is linear. Also, I've used a heatmap to display the correlation of features to check the relationship of the features between them and with label. One feature 'SquareFeet' has the high correlation with the label which is around 0.75 as compared to other features.

I choose 'SquareFeet' feature for prediction 'Price' because this feature is highly colinear and relationship is high as compared to other. If we choose other feature there is a not a big change in the output. That's why I ignored other.

Model Approach

Proposed AHPP model is trained on predefined Linear regression model. I implement it using sciket-learn library. Because model need to predict a numeric value like 'Price' that's why this is a regression problem so I choose regression algorithm. I use linear regression algorithm because data is linear in the dataset. I choose 'SquareFeet' feature for prediction 'Price'. So, data is linearly distributed between them.

I've spited the data 80% for training the data and 20% data is for testing. Also, I compute it for 1 and 2 polynomial degrees. First, I generated polynomial feature and fit the linear regression model for training. Performance measure for this models used are **'root_mean_squared_error()', 'mean_absolute_error()' and 'pearsonr()'.** Root mean square will give us information that there is a error. Smaller the value of root mean square means the error is small and performance is good. Pearsonr will show the relationship of input feature with label feature.

Implementation

1. **Tools:**
    - VsCode
    - Jupyter Notebook
    - Python
2. **Libraries:**
    - Pandas
    - Sciket-learn
    - Seaborn
    - Matplotlib
    - Scipy

Results

I use **MinMaxScaler**() to transform data and trained model for degree 1 and 2. Following are the performance measure for polynomial degree 1 and 2 respectively. **Root mean square error (RMSE)** value is small for 'squarefeet' feature as compared to other at polynomial degree 1. So, this is generalized model as compared to other. If we took other for training then they are not contributing to predict the output because as talk above their correlation is low.

MinMaxScalar()					
Degree 1					
Features	**SquareFeet**	**Bedrooms**	**Bathrooms**	**SqurareFeet + Bathrooms + Bedrooms**	**All 5 features**
RMSE↓	0.0939	0.3782	0.3785	0.0934	0.0934
MAE↓	0.0750	0.1166	0.1169	0.0746	0.0746
Corr↑	0.7552	0.0600	0.0178	0.7585	0.7586
Degree 2					
RMSE↓	0.0939	0.3782	0.3785	0.0934	0.0934
MAE↓	0.0750	0.1166	0.1169	0.0746	0.0746
Corr↑	0.7552	0.0611	0.0186	0.7584	0.7584

Table 1 MinMaxScalar()

I use **StandardScaler**() to transform data and trained model for degree 1 and 2. The error is high If we compared it with **MinMaxScalar() and RobustScalar()** output.

StandardScaler()					
Degree 1					
Features	**SquareFeet**	**Bedrooms**	**Bathrooms**	**SqurareFeet + Bathrooms + Bedrooms**	**All 5 features**
RMSE↓	0.6522	0.9934	0.9950	0.6485	0.6484
MAE↓	0.5211	0.8099	0.8118	0.5180	0.5179
Corr↑	0.7552	0.0600	0.0178	0.7585	0.7586
Degree 2					
RMSE↓	0.6523	0.9933	0.9949	0.6485	0.6485
MAE↓	0.5211	0.8099	0.8115	0.5180	0.5181
Corr↑	0.7552	0.0611	0.0186	0.7584	0.7584

Table 2 StandardScalar()

I use **RobustScalar**() to transform data and trained model for degree 1 and 2. The error is high If we

compared it with **MinMaxScalar()** output but the output is good if we compared it with **StandardScalar()**.

RobustScaler()					
Degree 1					
Features	**SquareFeet**	**Bedrooms**	**Bathrooms**	**SqurareFeet + Bathrooms + Bedrooms**	**All 5 features**
RMSE↓	0.4539	0.6913	0.6924	0.4535	0.4534
MAE↓	0.3626	0.5636	0.5647	0.3622	0.3623
Corr↑	0.7552	0.0600	0.0178	0.7557	0.7558
Degree 2					
RMSE↓	0.4539	0.6912	0.6924	0.4535	0.4535
MAE↓	0.3626	0.5636	0.5647	0.3623	0.3623
Corr↑	0.7552	0.0611	0.0186	0.7556	0.7556

Table 3 RobustScalar()

Conclusion

The proposed AHPP model predict the price of house based on a square foot of the house. Proposed model solves the problem by predicting the price of house. But this model is trained on only one feature that are not enough for predicting price of house. So, in future, proposed model can be improved by increasing the features or using different datasets. By this AHPP can be used in the real estate market to automate the price prediction of house.