

HOUSE SALE PRICE PREDICTION

Submitted by:

KAVITA JHA

**ACKNOWLEDGMENT**

I first would like to thank god for giving us the opportunity to be here and make this project. Secondly, we thank our families for their unwavering support and great sacrifices in order to help us reach this moment. We would also like to thank the faculty of DataTrained for their work and efforts to give us the best education possible in order to reach our full potential, especially our supervisor Prof. Shankar for his guidance and assistance in not only this book but throughout our entire journey. We would also like to extend our thanks to Fliprobo technologies for help in the writing of this project.

**INTRODUCTION**

* Business Problem Framing

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 improve Real Estate efficiency. With this model, we would be able to better predict the prices.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

* Review of Literature

Most of the literature study is based on articles with full text online, open access articles. The literature study endeavours to construct a robust basis on regression techniques, regularisation in machine learning and on how it can precisely be applied to house prices prediction. The literature study gives an overview of the articles that are related to this study, the feature engineering methods that have been used in this study. As well as evaluation metrics that are used to measure the performance of the algorithms. In addition, the factors that have been used

* Motivation for the Problem Undertaken

Today buying a house a emotion to everyone because it’s a once in a lifetime work for everyone’s life no one buys a house on a regular basis so its important to this task a easier one for those who worked really hard to earn money and don’t waste time on this hectic task to find the price of the house on every locality it’s the basic thing on the one go the client would get the sale price of the house in every locality .

**Analytical Problem Framing**.

* Data Sources and their formats

Data sources are provided by the Fliprobo Techmologies to work on that. In the data source there are two different datasets

1. Train dataset:- on which we have build our model to predict the sale price of the dataset
2. Test model:- On which we tested our model that we build on training dataset to get the sale price of the house

* Data Preprocessing Done

Feature Engineering is the technique of improving the performance on a dataset by transforming its feature space, and it is the practice of constructing suitable features from given features of the dataset, which leads to improving the performance of the prediction model. However, several techniques should be implemented for better performance and a prediction result

1. Imputation :-- Missing value imputation is one of the biggest challenges encountered by the data scientist. In addition, most machine learning algorithms are not powerful enough to handle missing data. Missing data can lead to ambiguity, misleading conclusions, and results.
2. Outliers:--- Outliers are noisy data that they do have abnormal behaviour comparing with the rest of the data in the same dataset. Outliers can influence the prediction model and performance due to its oddity. There are three types of outliers, which are point, contextual, and collective outliers . Point outlier is an individual data instance that can be considered as odd with respect to the rest of the data. The contextual outlier is an instance of data that can be regarded as odd in a specific context but not otherwise. An example of contextual is the longitude of a location. A collective outlier is a collection of related data instances that can be considered as abnormal with respect to the entire dataset. In supervised, the detection of outliers can be accomplished visually, where a predictive model is built for normal against outliers' classes. Dean De has investigated the public dataset and he suggests to remove certain outliers from the public data when he said "I would recommend removing any houses with more than 4000 square feet from the data set". Another example of detecting outliers is by using Isolation forest, which has two stages, training, and testing. The training is to create the isolation trees and then to record the anomaly score of each entry in the testing stage. This method has shown a promising result, according to
3. Feature Selection:-- Feature Selection is an important technique that is used to handle high-dimensional input data and overfitting caused by a curse of dimensionality by selecting a relevant feature subset based on mutual information criterion . Moreover, feature selection has many advantages, such as improve the prediction performance by reducing dimensionality in the dataset. It speeds up the learning process and leads to a better understanding of the considered problem. However, there are many useful methods for feature selection, such as Mutual Information (MI) and Conditional Mutual Information (CMI)..

* Data Inputs- Logic- Output Relationships

We have divided the input and output as the training and testing dataset phase. Where we send the dependent variables as the input the and gets the independent variable as the output.

* Hardware and Software Requirements and Tools Used

We have worked in Python Jupyter in order to work on this problem.

Import libraries like

1. Numpy
2. Padas
3. Matplotlib
4. Seaborn
5. Importing the regression from skleran
6. Import train test split method

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

We have used three algorithm on this project

1. Linear Regression
2. Decision tree regression
3. Random forest regressor

* Run and Evaluate selected models

1.Linear Regressson model:- 78% accuracy

2. Decision Tree regression model gives:- 67%

3. Random Forest Regression model gives 92 % accuracy.

* Key Metrics for success in solving problem under consideration

Several evaluation metrics measure the performance of machine learning algorithms such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared, and Mean Absolute Error (MAE). However, in this study, the performance of the algorithms is measured by using RMSE and R-Squared. Root Mean Square Error (RMSE) is used as an evaluation metric in machine learning to measure the performance of the model. However, RMSE is similar to the Mean Square Error (MAE). Where all errors in MAE have the same weight, but RMSE penalises the variance, which means it gives more weight to the errors that have large absolute values than that have small absolute values. Therefore, when RMSE and MAE are calculated, RMSE is always bigger than MAE. RMSE is more sensitive to the errors than MAE; therefore, using RMSE for measuring the performance is better than MAE [38].

* Interpretation of the Results

Many machine learning algorithms are used to predict. However, previous researches have shown a comparison between them alongside Artificial neural network in different datasets. Therefore, using these algorithms is beneficial so that the result can be as near to the claimed results. However, the prediction accuracy of these algorithms depends heavily on the given data when training the model. If the data is in bad shape, the model will be overfitted and inefficient, which means that data pre-processing is an important part of this experiment and will affect the final results. Thus, multiple combinations of pre-processing methods need to be tested before getting the data ready to be used in training.

**CONCLUSION**

The study shows a comparison between the regression algorithms and artificial neural network when predicting house prices in Ames, Iowa, United States and Malmö, Sweden. The results were promising for the public data due to it being rich with features and having strong correlation, whereas the local data gave a worse outcome when the same pre-processing strategy was implemented due to it being in a different shape compared with the public data in terms of the number of features and the correlation strength.

* Learning Outcomes of the Study in respect of Data Science

We have used three algorithm on this project

1. Linear Regression
2. Decision tree regression
3. Random forest regressor

On all these algorithm random forest regressor works best on the project and gives us the 92% accuracy with the hyperparameter tuning and gridsearchcv which is to good that model is not underfit or overfit

* Limitations of this work and Scope for Future Work

The request contains a list of features, that matches the public dataset's features, that is desired to be available when the data is sent. There is no guarantee that the data will be available in time nor contains the exact requested list of features. Thus, there might be a risk that the access will be denied or delayed. If so, the study will be accomplished based only on the public dataset. Moreover, this study will not cover all regression algorithms; instead, it is focused on the chosen algorithm, starting from the basic regression techniques to the advanced ones. Likewise, the artificial neural network that has many techniques and a wide area and several training methods that do not fit in this study.