Project Name: Housing Price Prediction Project

Author

PRASHANT KUMAR

Introduction

Machine Learning is a subfield of Artificial Intelligence (AI) that works with algorithms and technologies to extract useful information from data. Machine learning methods are appropriate in big data since attempting to manually process vast volumes of data would be impossible without the support of machines. Machine learning in computer science attempts to solve problems algorithmically rather than purely mathematically. Therefore, it is

based on creating algorithms that permit the machine to learn. However, there are two general groups in machine learning which are supervised and unsupervised. Supervised is where the program gets trained on pre-determined set to be able to predict when a new data is given. Unsupervised is where the program tries to find the relationship and the hidden pattern between the data. Here, we are using supervised technique.

Abstract

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.

Problem Statement

The goal of this statistical analysis is to help us understand the relationship between house features and how these variables are used to predict house price.

Objective

- Predict the house price
- Using two different models in terms of minimizing the difference between predicted and actual rating

Work Report

Step - 1:-

We know that this House Price Prediction problem is based on a supervised problem. So, first of all, we fetch all the essential libraries Like pandas, NumPy, for data visualization matplot, seaborn, here we have a price as target variable for this we import Linear Regression, Extra Trees Regressor, from sklearn we import model selection and import train_test_split, etc.

Step - 2:-

Now, we load data by using the pandas read CSV method. Here, CSV means comma separate value. Pandas change CSV into a data frame. Which shows every row and column in table form. Here, we can see the column's names and understand them.

Step 3:-

Now, we start doing Exploratory Data Analysis (EDA). It is also called feature engineering. In this, we are going to find data shape, columns, null values, information about data like data type, we work on null values means missing values, then we find describe which tell us about the minimum, maximum, inter-quartile values, etc.

Step 3.1:-

First, we check data shape with shape and sample function. It shows us different rows and columns. We check columns' name by columns function.This creates some understanding of our dataset.

Step 3.2:-

Now, we check data information by info function. It shows us the total values of data with missing values and their data type. Data type shows is the data integer, float, or object? This also makes some understanding for clearing or managing our data wisely.

Step 3.3:-

Now, we check missing values from our data by isnull function. Basically, isnull function gives a boolean value so, we use isnull().sum() which shows the total count of missing values. This is more clear than a boolean value. We also visualize these missing values through a graph. This shows a beautiful representation of missing values. Here, we use Heat map of seaborn libraries and use cmap means color map equals to viridis. This function of seaborn makes all missing value appear.

Step 3.4:-

Now, we see missing values of data through visualization. But we really want actual value so, we calculate the percentage of missing value inside our dataset. For this, we find how many features which have more than 1 missing value and taking show them. Here, we can see Alley, PoolQC, and MiscFeature have more than 90% missing values. We take may Encoding techniques or pandas functions to fill these missing values. Or we can drop all these but before doing this let's check their effect on our target column through visualization.

Step 3.5:-

Now, we check the effect of missing values on our target House Price. For visualization, we use jointplot function of seaborn. This represents beautifully our desired results. Here, we see that the above describe columns are not affecting our target.

So, we drop id, alley, poolqc, fence, and miscfeature and take inplace = True, which make a permanent change in our dataset.

Step 3.6:-

Now, we fill missing values by their mean. We use mean when values in continuous/integer form. Here we fill lotfrontage, masvnrarea, and garageyrblt columns. Now, let's check missing values are reduced from the past. Yes, it changed.

Step 3.7:-

Now, we use Label Encoder Technique to change object data into integer values and check missing values then we find zero. Our desired work is done now.

Step 4:-

Now, we check minimum, maximum, standard deviation, mean, and many more things of every columns by the function of describe. Here, we see the mean values which show average distribution of data values, standard deviation shows expansion of data values, and 25%, 50%, 75% are showing the inter-quartile of data values. Here, we can see if data values have larger gap between 75% and maximum values then it shows outliers. Now, we see there are some columns which have outliers. We will use z-score as a treatment for outliers.

Step 4.1:-

Now, we use a scatter plot for visualizing every column’s effect on our target.

Step 5:-

Now, we are calculating correlation. Correlation is the statistical analysis of the relationship or dependency between two variables.

Correlation allows us to study both the strength and direction of the relationship between two sets of variables. Studying correlation can be very useful in many data science tasks.

Step 5.1:-

Now, we show correlation by graph.Here, we see a clear view.

Step 6:-

Now, we show outliers by boxplot. It shows clear view of outliers. Then, we use z-score to remove outliers data. This z-score is imported from scipy.stats means scientific statistics.

Step 7:-

Now, we calculate skewness of data. This shows the distribution of data in left side or right side. If data shows middle skewness in bell curve graph means the data value is fair distributed.Here, we use square root technique to remove skewness of data.

Step 8:-

Now, we see our dataset has meny columns. Here, we use Principal Component Analysis (PCA) to reduce the dimension of our dataset. This technique is called dimension reduction. Now, data shape is 1168 rows and 15 columns.

Step 9:-

In Housing Price Prediction, we have test data to predict price. Now we are going to use test data. Here we also use the same technique to clean data, reshape, and many more which we used in the train dataset.

Step 10:-

Now, we use the feature selection technique. Here, we split our dataset into the target and independent features. For this, we use the train_test_split method.

Step 11:-

Now, we select a model for testing our dataset.For this, we use the Extra Trees Regressor model. Its result is shown through a graph.

Step 12:-

Now, we fit the model and use the Random Forest Regressor (rf). This gives predicted values and train score and rf score.Here, we get 97% rf score. We show this rf score through graph. We use a distribution plot for showing the difference of test and predicted value.

Step 13:-

Now, we import metrics from sklearn to find Mean absolute error, Mean squared error.

Step 14:-

Now, we are doing Hyperparameter Tuning. Hyperparameter Tuning is a process which creates a machine model after repeated testing of model. It search the best fit model which gives best results. Here, we use Randomized Search CV which find the best model who fulfill our desired result. Here, we use different estimators to fit the model. We create random grid which shows maximum depth, sample split, and sample leaf. Then we again fit the model. Now we get values.

Step 15:-

We find best parameters. We display difference of y_test and prediction through distribution plot. It shows a bell curve. We also shows this through scatter plot. We check Mean Absolute Error and Mean Squared Error.

Step 16:-

Now, we save the file by using the pickle method as the name of the House Price rf file. We get final r2 score 80% , this means our model is working well.

Conclusion:

Coefficient of determination also called as R2 score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s). Here, we get r2 score 80%.

Thank You