Housing Project

Submitted by:

HARSHIT REDDY

Introduction

Machine Learning (ML) by looking at the definition, we can say that it is a field of computer science. It is the learning and building of algorithms that can learn from it and make predictions on datasets. This model is built so that we can conduct an analysis and understand some of the important factors which are necessary to predict the actual value of the prospective properties and decide whether to invest in them or not.

Problem Statement:

We will use Linear Regression for this dataset to build the model. A US-based housing company named Surprise Housing has decided to enter the Australian market. The company will use data analytics to purchase at a cheap price and sell them at a relatively higher price. We need to predict the actual value of the prospective properties and decide whether to invest in them or not.

Exploratory Data Analysis:

EDA is a phenomenon under data analysis is used for gaining a better understanding of data aspects like a.) Main features of data. b.) Variable and relationships that hold between them. c.) Identifying which variables are important for our problem.

There are 1168 samples in this dataset. There are both categorical and numerical variables present in the dataset. We will be plotting the distribution plots, box plots, count plots etc. While plotting the distribution plots, if there is any skewness present in it then we remove it by applying some transforms like sqrt, cbrt, log and then we need to make sure that the skewness range is between -0.5 to +0.5 and we plot the distribution plot again without any skewness. We plot the boxplots to detect if there are any outliers present in them or not. Multiple libraries are available to perform basic EDA such as pandas, matplot library and seaborn.

Motivation for the Problem Undertaken:

The objective behind to make this project is to know what is the actual value of the properties and whether it is going to be profitable if we invest in them or not.

Analytical Problem Framing

Data Preprocessing Pipeline:

The pre-processing of data involves three steps namely data cleaning, feature selection and data transformation. The data cleaning involves missing data. We need to replace the missing values by using either mean, median or model. Before, training the model feature selection is one of the important factors that can influence the model's performance.

In the processing steps with the dataset, I have cleaned all the data with techniques like:

1. Handling the missing values which are necessary as there are some missing values present in the dataset.

2. Encoding the categorical variables which were done by using a label encoder as the categories are assigned starting from 0.

3. I have removed the skewness from some of the columns where there is some skewness present and I have made sure that the skewness range is in between -0.5 to +0.5.

Hardware and Software Requirements and Tools Used:

The software tool that I have used is Scikit Learn on my Windows 10 platform in Python. The libraries that I have used are numpy, pandas, matplotlib and seaborn.

The packages that I have used are: a.) from sklearn.linear_model import LinearRegression b.) from sklearn.model_selection import train_test_split

c.) from sklearn.metrics import mean_absolute_error, mean_squared_error

d.) from sklearn.tree import DecisionTreeRegressor

e.) from sklearn.neighbors import KNeighborsRegressor

f.) from sklearn.svm import SVR

g.) from sklearn.ensemble import RandomForestRegressor

h.) from sklearn.metrics import r2_score

Model Development and Evaluation

Testing of Identified Approaches (Algorithms):

The algorithms that I have used for training are Linear Regression, Decision Tree Regressor, K Neighbors Regressor, SVM and Random Forest Regressor. The same algorithms I have used to predict on the testing data.

Run and Evaluate selected models:

Firstly, I have imported all the algorithms which was required to build the model. Then, I have taken the Decision Tree Regressor model to check on which random state I am getting the best score such that the random state can be finalized. Then, by using the fit method, I have trained all the models. By using predict method, I have predicted the values for all the models. Predicted data is nothing but the answer given by the x_test model and y_test is the actual data.

The score for Linear Regression is 80.90% and r2 score is 85.81%, score for Support Vector Machine (SVM) is 68.49% and r2 score is 79.39%, Decision Tree Regressor is 100% and r2 score is 70.43%, score for K Neighbors Regressor is 72.56% and r2 score is 59.71%, and finally score for Random Forest Regressor is 97.12% and r2 score is 88.86%. After checking the accuracy scores for all the models, the Random Forest model is giving me the highest accuracy score of 90.63%.

Visualizations:

Machine learning data visualization is important to understand how data is used in a particular machine learning model it helps in analyzing it. We use visualizations because it is easy to understand the visualized data. Firstly, I have plotted the distribution plots for all the numerical columns and checked whether the plot is normally distributed or not. If the data is not distributed normally, then I have used some transforms like sqrt, cbrt and log to remove the skewness and make the plot normally distributed. Then, I have plotted the box plots to check whether there are any outliers present in them or not. Finally, I have also plotted the count plots for the categorical columns to make sure what is the count present in them.

Interpretation of the Results:

So, after getting all the scores for the models, I have done the cross validation for all the model to check whether the scores are accurate or not because the scores can be because of overfitting. The scores that I have got for the cross validation for Linear Regression is 73.96%, Decision Tree Regressor is 64.90%, K Neighbors Regressor is 60.26%, SVR is 69.19% and Random Forest Regressor is 84.34%. After comparing all the scores and cv scores for all the models, it is clear that Random Forest model is working better when compared to the other models.

Now, let's fine-tune the model by using hyper-parameter tuning. We use hyper parameter tuning technique to improve the accuracy score of the best model that we have chosen. Sklearn comes with Grid Search CV to do the search over specified parameter values for an estimator. It helps to optimize the model's performance. With the best parameters, the model will identify the patterns within the dataset in a better way. I used the parameters like criterion, max depth, max features and n estimators for the Random Forest Regressor model to find the improved accuracy of the model. After tuning the model, the score of the model with the best parameters is around 93.97%.

Conclusion

We have learned to build a complete machine learning project. We also learned to fine-tune our model and save it for further use. Finally, the hyper tuning score which was obtained 93.97% has been a good accuracy achieved so far.