House Prices Regression Using Python

Introduction

The study is to compare different methods of regression and decide which model will fit a particular dataset the best. The features in the dataset include bedrooms/house, bathrooms/bedroom, area of the house and lot, presence of a waterfront, views, condition of the house, the grade assigned by the county, built year, renovated year, and the location of the house. There are 2 categorical variables, 17 continuous variables, 1 variable to store house ID, and 1 variable to store the date the house sold. The models used for comparison will be a multiple linear regression model, a polynomial regression model with degree 2, and a random forest regression model.

Data Cleaning and Pre-Processing

After importing libraries, we will also import the dataset that will be used using pandas read_csv. For checking any null values, info and isnull().sum() are used, and it is found that there are no missing values detected from the dataset. In the case of preprocessing, the id column is dropped since it is unique, and the date column is transformed into year, month, and day. The dataset is split into 80% train and 20% test.

Visualizing the Data

Drawing charts and examining the data before applying a model is a very good practice because we may detect some possible outliers or decide to do normalization. To determine bedrooms, floors, or bathrooms/bedrooms vs. price, I preferred a boxplot because we have numerical data, but they are not continuous. From the charts, it can be seen that there are very few houses which have some features or price appears far from others like 33 bedrooms or price around 7000000. However, determining their possible negative effect will be time-consuming, and in real data sets, there will always be some outliers. For the price vs. some features and it seems that there is not a perfect linear relationship between the price and these features. The charts show that when the sqrt_living increases, sqrt_lot, and bedrooms or bathrooms/bedrooms increase. However, the floors, bedrooms, and bathrooms/bedrooms or sqrt_living do not have a similar relationship.

The boxplots show that grade and waterfront affect price visibly. On the other hand, the view seems to effect less, but it also has an effect on price. to determine the relationship between the view, grade, and year built, the chart shows that the newer houses have better grades, but we can not say much about the change in the view. Regarding collinearity, sqt_above and sqt_living are highly correlated. This can be estimated when you look at the definitions of the dataset and check to be sure by looking at the correlation matrix. However, this does not mean that you must remove one of the highly correlated features. For instance: bathrooms and sqrt_living. They are highly correlated, but I do not think that the relation among them is the same as the relation between sqt_living and sqt_above.

Insights and Findings

The multiple linear regression model with all the parameters except the ID was created, and the evaluation metrics for MSE with test set is 44951491944.93195 and train is almost similar to the test, and R^2 for both test and train is almost the same 0.7. The same process was repeated for polynomial regression, and MSE with the test is 30195015951.97894, which is lesser than the multiple linear regression the train MSE is almost similar to the test also, and the R^2 for both the test and train is almost the same at 0.8. In the case of random forest, MSE for the test is 21897480069.060936, which is lesser than the previous two models, but the difference between the test and train MSE is very high, as shown in the figure below. Also, the R^2 for the test is 0.85, and the train is 0.98. So even though the random forest has less error than the other two models, it has a major difference between the test and train error values and R^2. Thus the better model from this comparison is polynomial regression with degree 2 as it has an error that is less, and both the test and train evaluation is almost the same.

Figure 1

Model Evaluation Metrics

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	126929.173470	4.495149e+10	212017.668945	0.702656	0.6968
1	Polynomail Regression	104067.660103	3.019502e+10	173767.131391	0.800267	0.0000
2	Random Forest Regressor	72522.055112	2.147901e+10	146557.191274	0.857921	0.0000

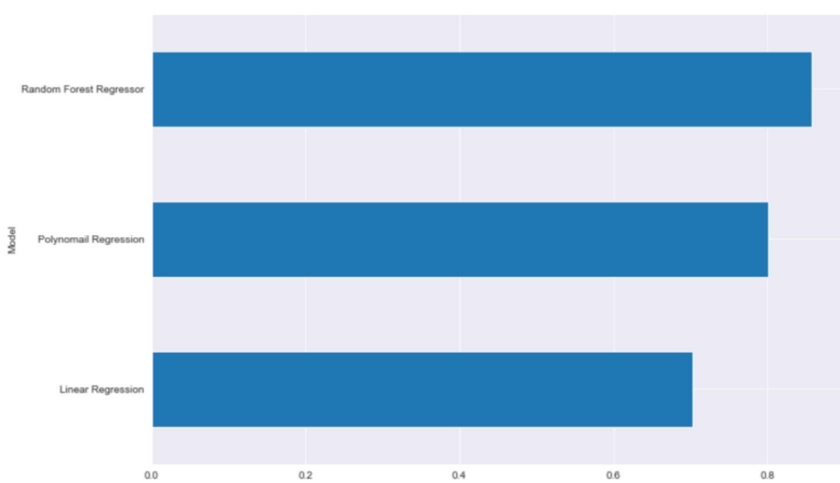
```
Test set evaluation Polynomial Regression:
_____________________________________
MAE: 104067.66010293778
MSE: 30195015951.97894
RMSE: 173767.1313913507
R2 Square 0.8002667507368718
__________________________________
====================================
Train set evaluation Polynomial Regression:
_____________________________________
MAE: 96278.70322440717
MSE: 21212437655.46112
RMSE: 145644.9026072012
R2 Square 0.837637483887351
__________________________________
```

```
Test set evaluation Randon Forest:
_____________________________________
MAE: 72623.47789382371
MSE: 21897480069.060936
RMSE: 147977.97156692256
R2 Square 0.8551530871245827
__________________________________
Train set evaluation Randon Forest:
_____________________________________
MAE: 25158.696711046847
MSE: 2167555823.053826
RMSE: 46557.016904585136
R2 Square 0.9834092704024971
__________________________________
```

Note. The screenshots are taken from the Jupyter Notebook of this analysis.

Figure 2

The Graph of R^2 for all the Models



Note. The screenshots are taken from the Jupyter Notebook of this analysis.

Conclusion

From the output, it is visible that the random forest algorithm is better at predicting house prices for the housing dataset since the values of MAE, RMSE, and MSE for the random forest algorithm are far less compared to the linear regression algorithm and polynomial regression. However, the test and train values of the errors and R-squared for random forecast has a huge difference. Thus, the Polynomial regression model gives us R-squared (testing) score of 0.8 and a training score of around 0.83, where both are almost the same and there seems no considerable difference in the evaluation of the test and training. From the above analysis, we can conclude that Polynomial regression for degree=2 is the best solution even though the random forest has fewer errors than the other two models.