Artificial Intelligence and Data Science Mini Project Report

Gupta Kunal

ITA628

Details About Dataset

The Dataset that is used in the project is the Big Mart Sales data.

It has Total of 8523 Rows and 12 columns.

This Dataset is widely used for Time Series Analysis.

It has Total 12 features in which 7 are Categorical and 5 are Numerical

The Categorical Features are

Item_Identifier

Item_Fat_Content

Item_Type

Outlet_Identifier

Outlet_Size

Outlet_Location_Type

Outlet_Type

And the numerical features are

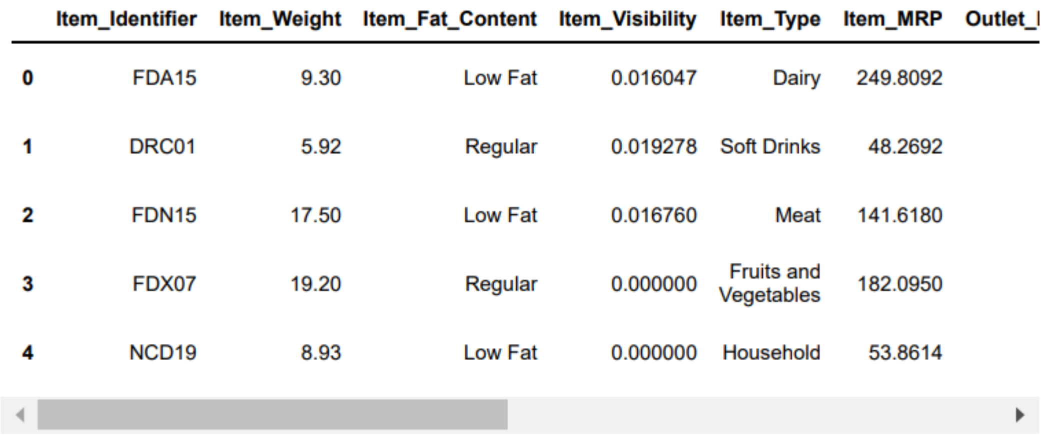
Item_Weight

Item_Visibility

Item_MRP

Outlet_Establishment_Year

Item_Outlet_Sales



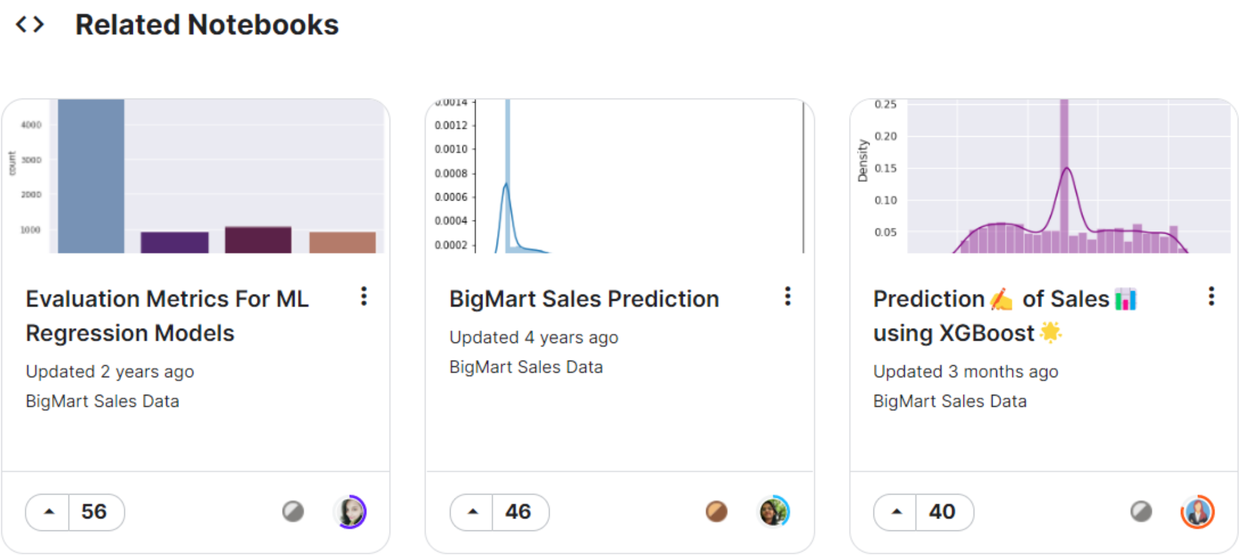
	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_I
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	

This is how the head of the dataset looks like

Work Previously Done on the Dataset

These are some of the previously done on the dataset available on Kaggle

Topic	Model Used	Learning Technique Used	Accuracy
1. Evaluation Metrics For ML Regression Models	Random Forest Regressor	Regression	54%
2. BigMart Sales Prediction	No Model Used Just had clustering which is used to model building	Clustering	N.A
3. Prediction✍ of Sales using XGBoost	XGBoost Regressor	Regression	Xtest- 83% Ytest-50%



Related Notebooks

Evaluation Metrics For ML Regression Models
Updated 2 years ago
BigMart Sales Data
56

BigMart Sales Prediction
Updated 4 years ago
BigMart Sales Data
46

Prediction ✍ of Sales 📊 using XGBoost 🌟
Updated 3 months ago
BigMart Sales Data
40

Work Done and Why

We have performed Linear regression on our dataset. Linear Regression is a part of Supervised learning. And as it is in supervised learning the required output will be present in the dataset.

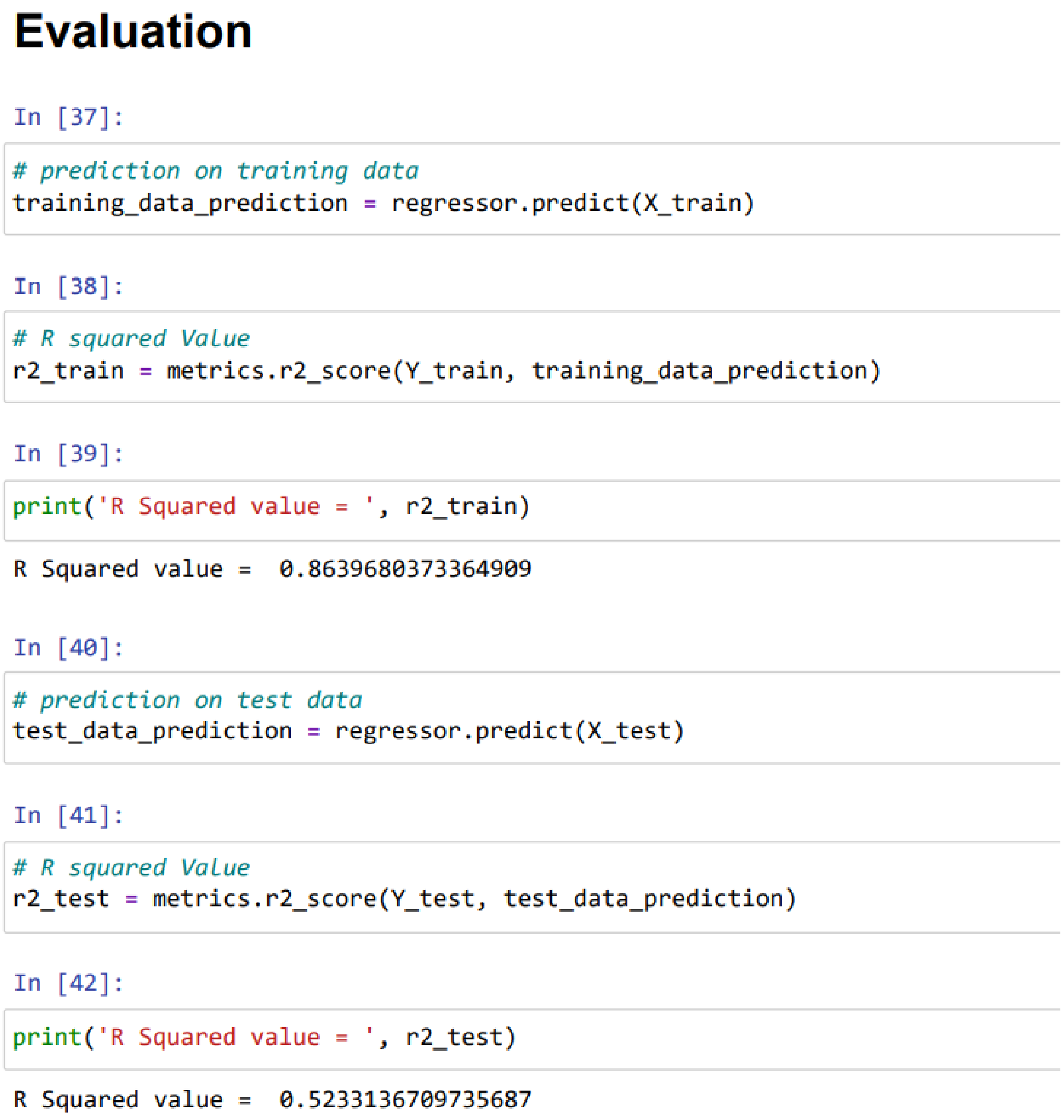
We have a attribute named ‘Item_outlet_sales’ which we are predicting in our project using Linear regression xgboost model. And the values that we are predicting will be compared to the real value present and we will rate this comparison using rsquare function which will give us the accuracy of the model.

We are predicting sales for the time series analysis which is used for market stocking and store inventory management. As we can predict the sales we can keep the extra stocks for that particular month accordingly.This can be very beneficial in business management.

We had many outliers in the dataset. There were outliers present in two columns in which one had numerical value and other had categorical value so in order to clean the outliers we replaced the numerical outliers with the median of the column and categorical values by the mode of the column.

Conclusion:

We performed Linear Regression on our Dataset using the XGBoost Function Successfully



Evaluation

In [37]:

```
# prediction on training data
training_data_prediction = regressor.predict(X_train)
```

In [38]:

```
# R squared Value
r2_train = metrics.r2_score(Y_train, training_data_prediction)
```

In [39]:

```
print('R Squared value = ', r2_train)
```

```
R Squared value =  0.8639680373364909
```

In [40]:

```
# prediction on test data
test_data_prediction = regressor.predict(X_test)
```

In [41]:

```
# R squared Value
r2_test = metrics.r2_score(Y_test, test_data_prediction)
```

In [42]:

```
print('R Squared value = ', r2_test)
```

```
R Squared value =  0.5233136709735687
```

We successfully predicted the item outlet sales and we got corresponding accuracy by the r square function

1.X_train Accuracy= 86.39%

2.Y_test Accuracy = 52.331%