XGBOOST

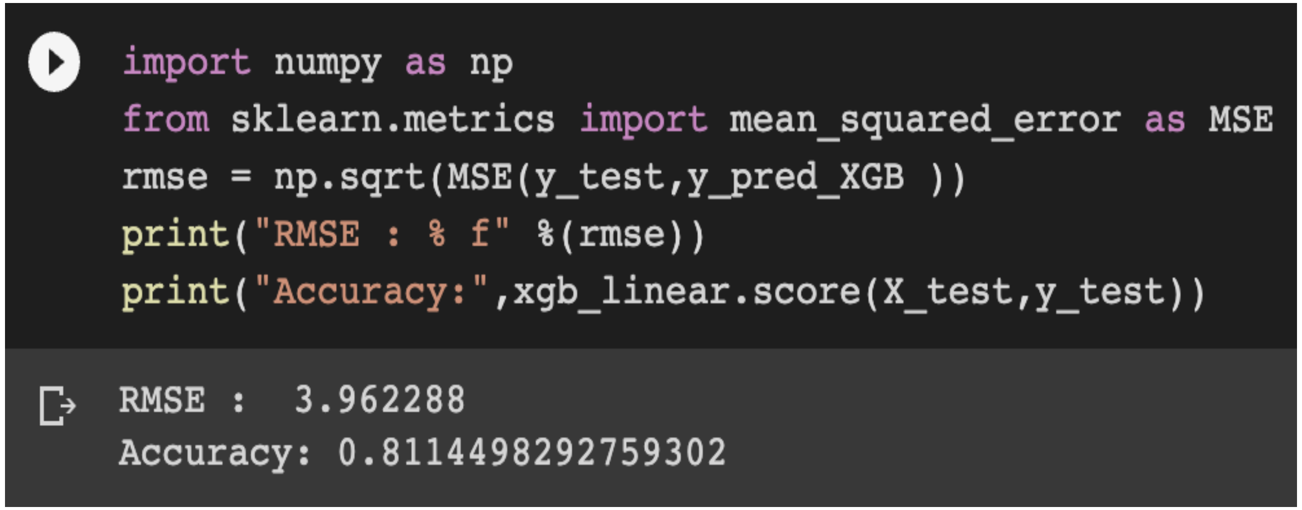
Raj Sarode (002762015)

Data Science Engineering Methods and Tools

Problem Statement & Approach

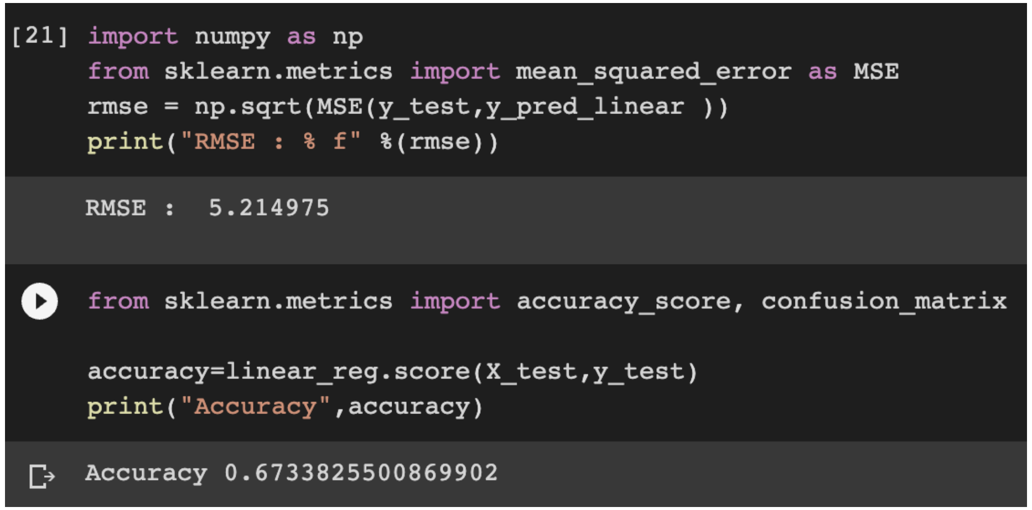
- Problem Statement: Predicting the median value of owner-occupied homes in the Boston Housing Dataset.
- Approach:
  - Split the data into a training set and a test set.
  - Trained an XGBoost model on the training set.
  - Evaluated the performance of the XGBoost model on the test set using the RMSE metric.
  - Compared the performance of the XGBoost model with that of a linear regression model using the same evaluation metric.
- Metric: RMSE (root mean squared error)
- Outcome:
  - If the XGBoost model outperforms the linear regression model on the test set, it may be a more accurate predictor of the median home value.
  - If the performance of the two models is similar, a simpler linear regression model may be sufficient for the task.

Comparing RMSE values for XGBoost & Linear Regression



```
import numpy as np
from sklearn.metrics import mean_squared_error as MSE
rmse = np.sqrt(MSE(y_test,y_pred_XGB ))
print("RMSE : % f" %(rmse))
print("Accuracy:",xgb_linear.score(X_test,y_test))
```

```
RMSE :  3.962288
Accuracy: 0.8114498292759302
```



```
[21] import numpy as np
     from sklearn.metrics import mean_squared_error as MSE
     rmse = np.sqrt(MSE(y_test,y_pred_linear ))
     print("RMSE : % f" %(rmse))
```

```
RMSE :  5.214975
```

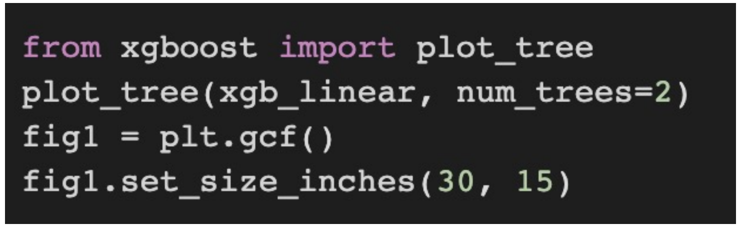
```
from sklearn.metrics import accuracy_score, confusion_matrix

accuracy=linear_reg.score(X_test,y_test)
print("Accuracy",accuracy)
```

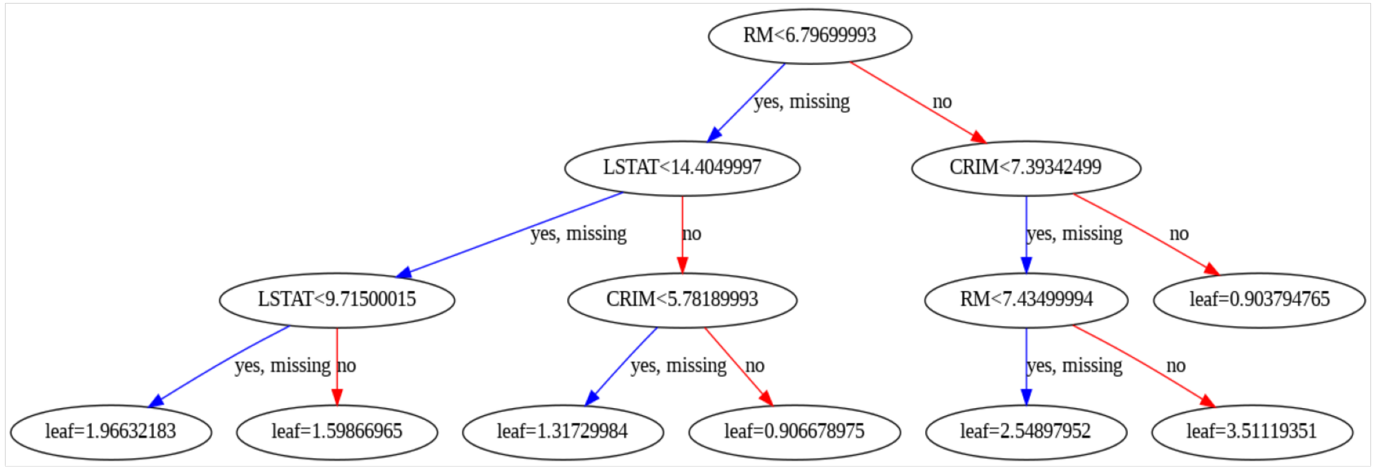
```
Accuracy 0.6733825500869902
```

As we can see, the RMSE value for XGBoost is lower than that of Linear Regression. This indicates that XGBoost can fit the data more effectively and is likely to produce more accurate predictions. Therefore, based on this metric, XGBoost appears to outperform Linear Regression in terms of predictive accuracy.

Gradient Boosting Decision Trees in XGBoost



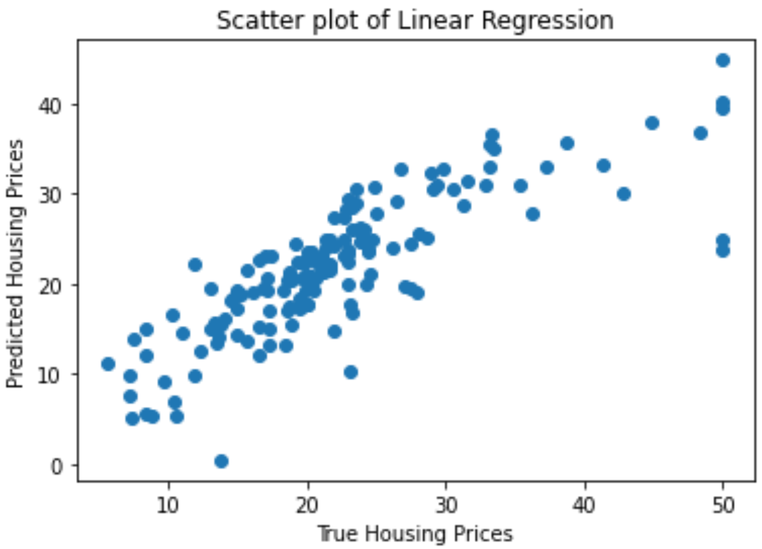
```
from xgboost import plot_tree
plot_tree(xgb_linear, num_trees=2)
fig1 = plt.gcf()
fig1.set_size_inches(30, 15)
```

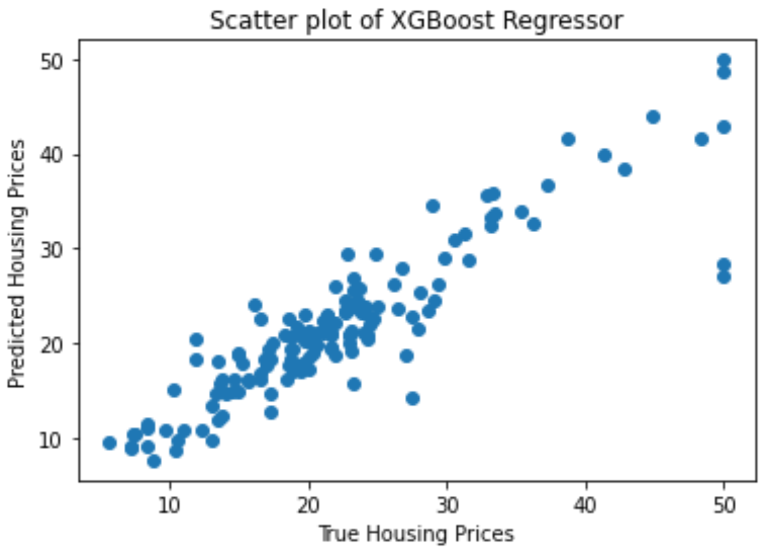


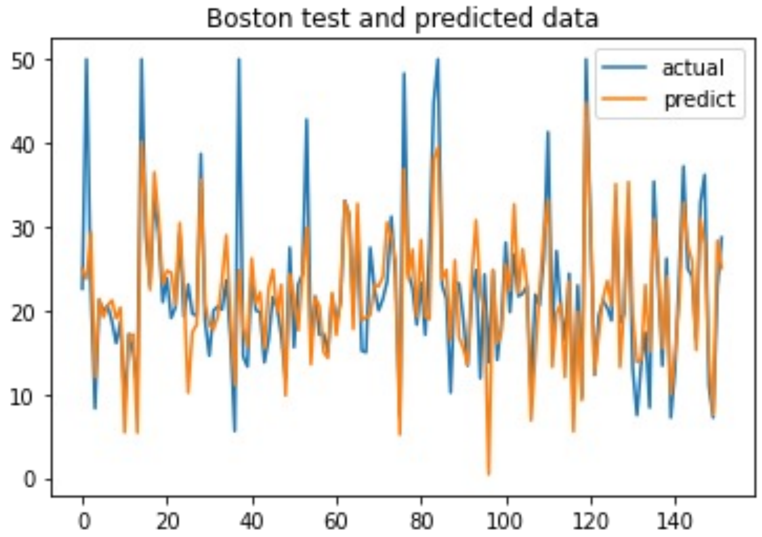
Here, RM is the root node, and the nodes at the bottom of the tree are the leaf nodes, which represent the final predicted value of the target variable. The split threshold values represents the point at which the data is splited into two groups based on the feature values-

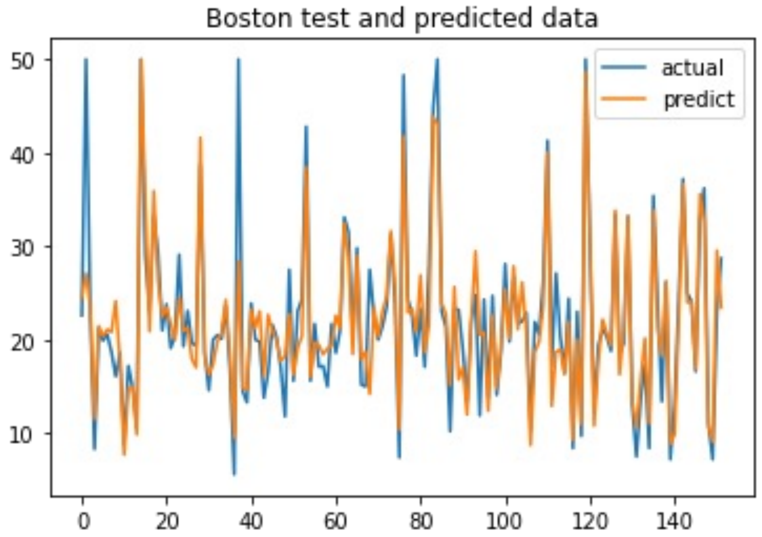
- One group for which the feature value is less than the threshold
- Another group for which the feature value is greater than or equal to the threshold.

Visualizations


Scatter plot of Linear Regression


Scatter plot of XGBoost Regressor


Boston test and predicted data


Boston test and predicted data

Conclusion

- XGBoost model outperforms linear regression in terms of RMSE score.
- XGBoost model has higher accuracy and can fit the data better than linear regression.
- XGBoost model provides better interpretation of the data through its decision tree visualization.
- Linear regression is a simpler model with fewer parameters and can provide a good baseline model for comparison.
- In complex datasets with high dimensionality, XGBoost can provide better results than linear regression

Thank You