**Data Analysis report**

**Introduction**

The objective of this analysis is to develop a prediction model for housing prices based on a data set which is Boston Housing. We will use different machine learning techniques, we try to understand the relationship between 13 housing features on the price of the house for ultimate price predictions for real estate transactions this was one of the data sets provided in which 506 observations and 13 features describing various aspects of the of this property in Boston. The model evaluated linear regression, ridge regression, polynomial regression

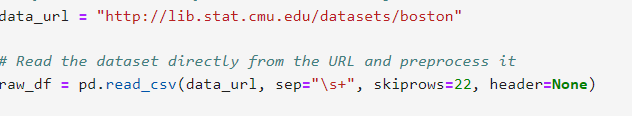
**Methodology steps**

In the preprocessing stage, we applied several techniques to clean and transform the raw data set in preparation for modeling our ultimate goal is data is suitable for training machine learning.

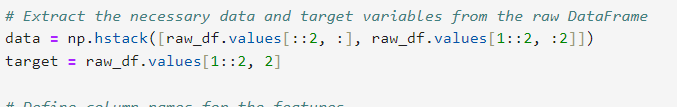
In the very beginning, i prepared a data model with linear regression and performed various evaluation parameters after performing the evaluation the result is shown below

To perform linear regression and its evaluation steps are as follows

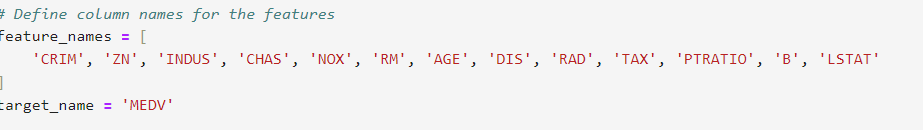
1. At the very first step I load the data from the URL which was mentioned using pd.read\_csv()



1. Preprocess the raw data all those 13 columns holding different features with this i extract the data feature and target feature



1. Now after loading next step is prepare data coloums need to mentioned for feature and target



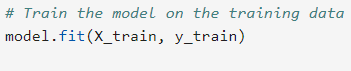
1. Next step data need to be split in two parts train and test in my case 80 % train data 20 % test for which data need to be split using train\_test\_spilt() method and (test \_size=0.2).



1. Then i create an instance of LinearRegression model



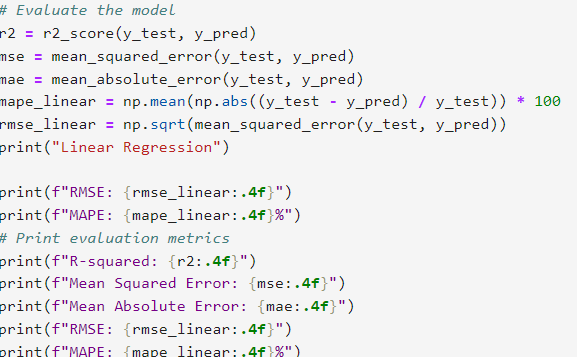
1. Train model fit the Linear\_regression model on a trainng data (x\_train, y\_train)using model.fit()



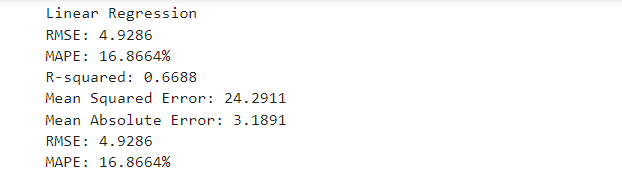
1. Make predictions y\_pred on the testing data x\_test with model.predict()



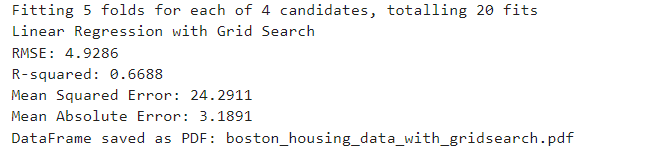
1. Then basic evalution have been performed such as R\_SQUARED , Mean Absolute Error, Root Mean Squared Error



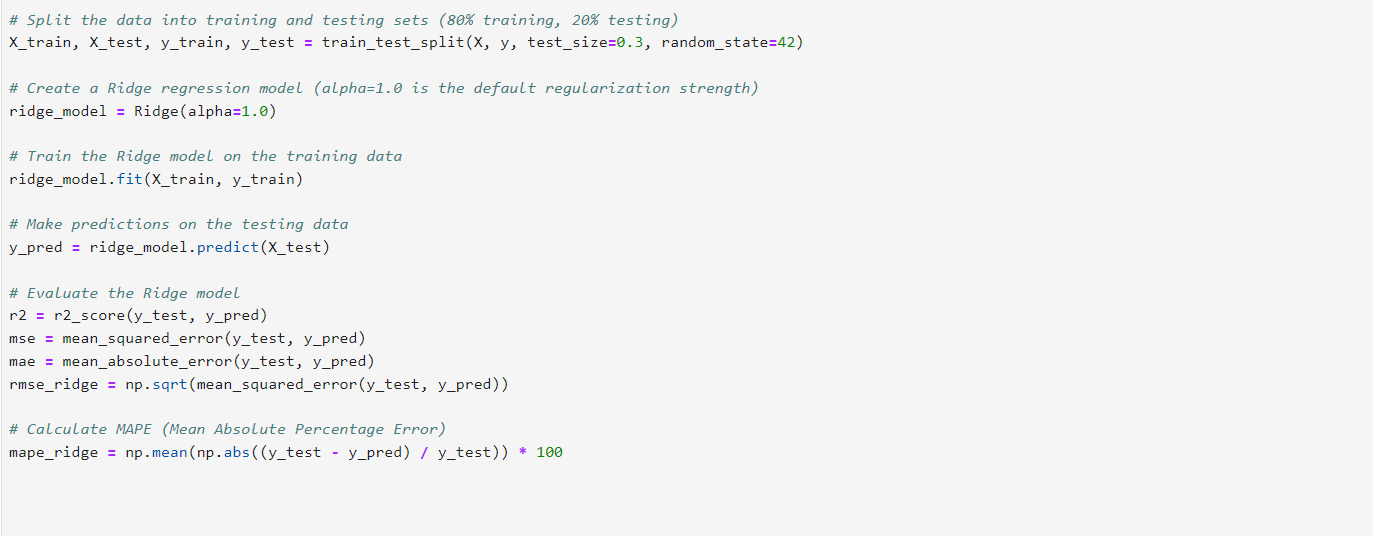
1. Result are shared below



1. In addition to this linear regression with grid search is also perform so can we auto tune our model and find the best model using grid search.



The second model we prepare usind the ridge regression and performed grid search also both results are shared below those steps which are common in both linear regression and ridge have not mentioned only the difference in both the model has been written below.

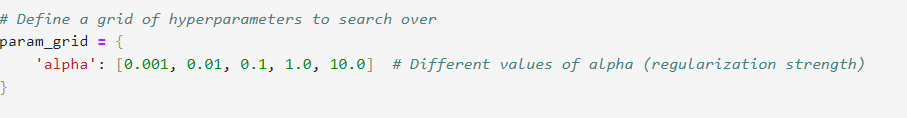
1. The steps which are similar such as loading dataset and spilting the data into train and test is same as we have in linear regression result are below at last we try to evaluate which datamodel is the best. 
2. We will initialize ridge regression which contains L2 regulations to control model complexity and reduce overfitting.



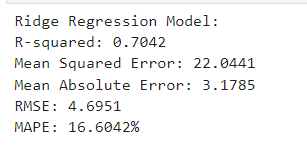
1. I define the various range of alpha values and result are used.

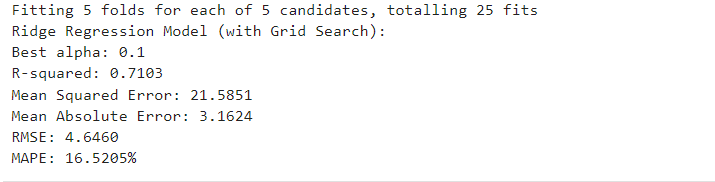


1. With this later on grid search is also performed result are shared for the analysis and i try to minimizes the mean squared error on the validaton set with values.



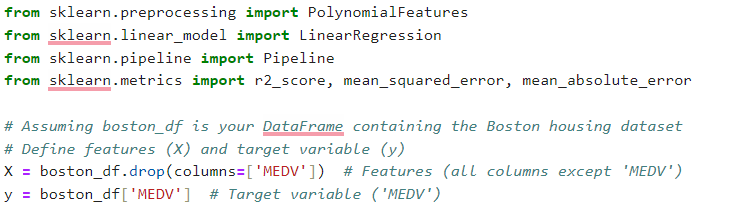
1. I try to otimizes the alpha value to reach the best data model.



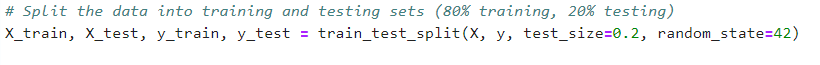


The third data model we use using the polynomial regression using linear regression pipeline.

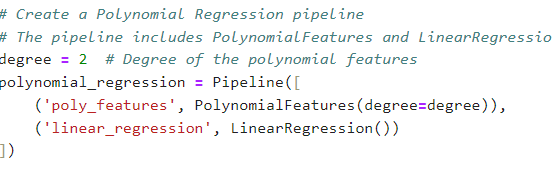
1. Same as linear regression load and preprocess the dataset to extract features (x) and target variable (y).



1. Then i performed spilit of the data into training and testing sets for model evalution.



1. Then i apply polynomial feature to generate high degree polynomial feature from original dataset to capture non linear relationship between variables.



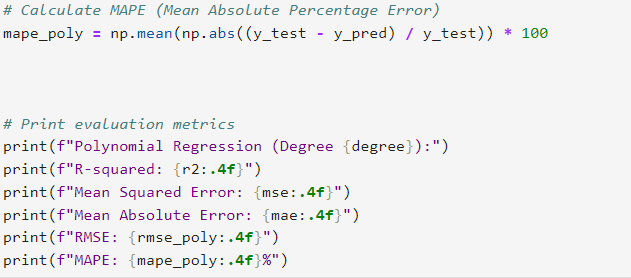
1. Then i will create a pipeline that combines polynomial features with linear regression to form a polynomial regression model.
2. With this we provide training to data model using x train and y train

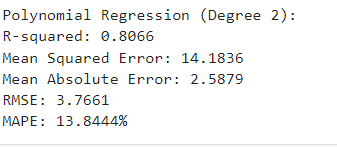


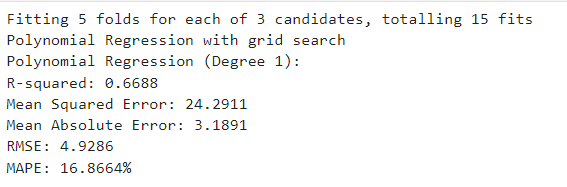
1. After training the data we will use this Polynomial Regression Model to make predictions on the testing data (x\_test)



1. We will evalute the polynomial regression model using the same metrics as linear regression such as R-squared, MSE , MAE, RMSE





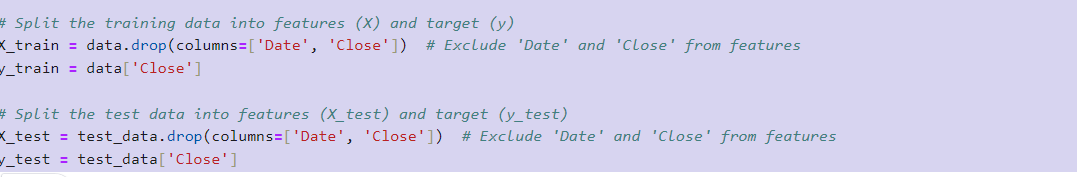


**Steps performed on other data set google \_stock\_flow**

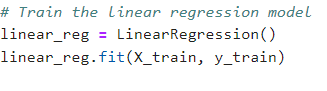
1. For this data set we have to download this data set which is in zip folder contain two data set inside test and train first we will build a model based on linear regression.



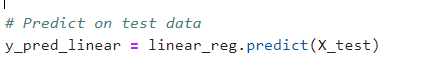
1. Then we will spilt the data into target and feature to train our model.



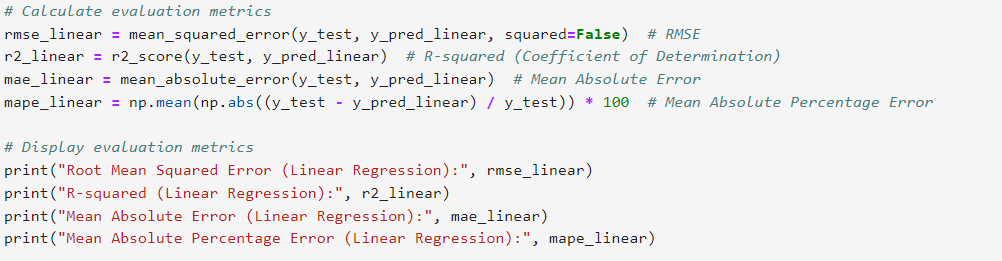
1. Then we will train our data model.

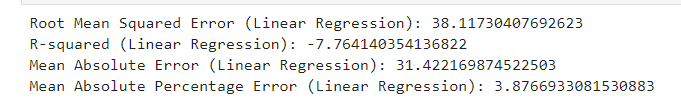


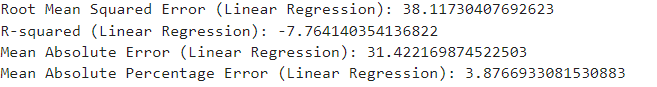
1. Then we will do predictions out of that data model.



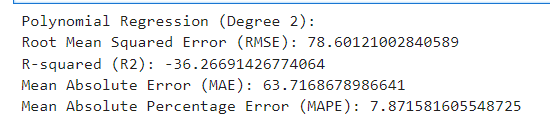
1. Then we will evaluate for best data model.

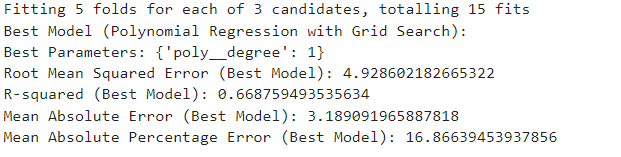


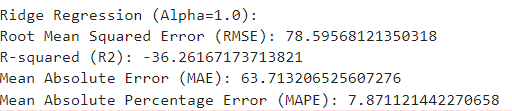


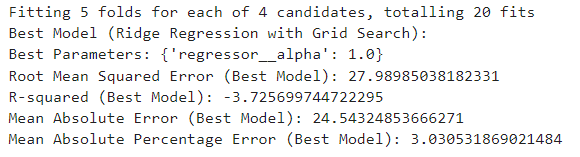


I performed the same steps for this data model polynomial regression and ridge regression first without grid search then with grid search for analysis of data in proper way here is the following result with both the regression.



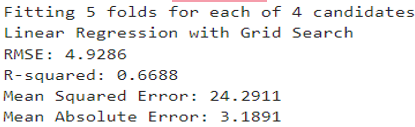


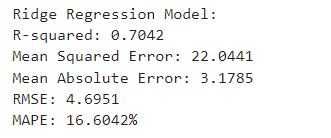


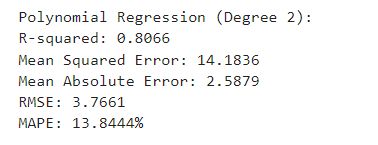


**Based on the evaluation made on the first data set to find which data model is best with the reason why?**

1. **Boston housing evaluation data analysis**



****

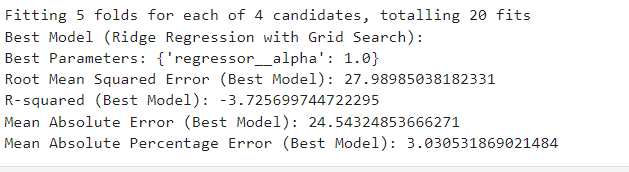
****

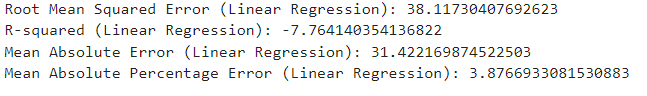
based on evaluation metrics which indicates the proportion of the variance in the dependent variable which we predict from the independent variables. The higher R-squared values suggest a better fit of the model to the data. In our case, the polynomial regression with the degree 2 model has the highest R squared value of 0.8066, including the best overall fit among the models.

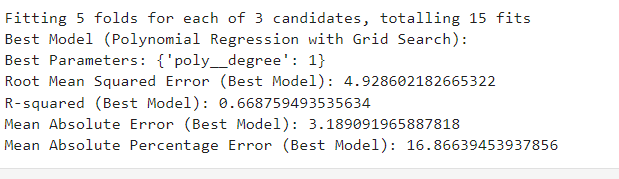
RMSE which stands for root mean squared error it is another parameter lower RMSE values indicate better model performance so here also polynomial regression degree 2 model has the lowest RMSE 3.7661 it suggests that it will give more accurate results as compared to other models.

MAPE(Mean Absolute Percentage Error):- this metric measures the accuracy of predictions as a percentage and lower values indicate better accuracy here in our case lower is 13.844%, which indicates superior accuracy in predicting target variables.

1. **Result analysis for 2nd data set google\_stock\_flow for best data model and why?**

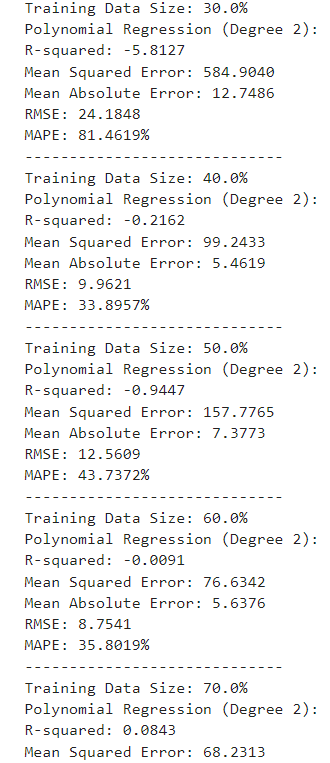


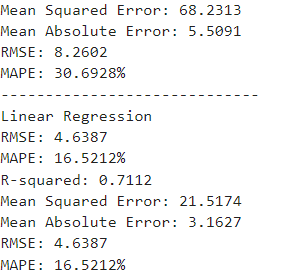


Root mean Squared Error(RMSE): lower value indicated better model performance in terms of prediction accuracy. The “Best Model” (Ridge regression with grid search significantly lower RMSE of 27.9898 compared to others

Mean absolute Error(MAE) and Mean Absolute Percentage ERROR(MAPE). The “best model” also exhibits lower which signalises the better model in our case Ridge Regression is better.

1. **Result as we increase or decrease the volume of data set we will have the following data result.**





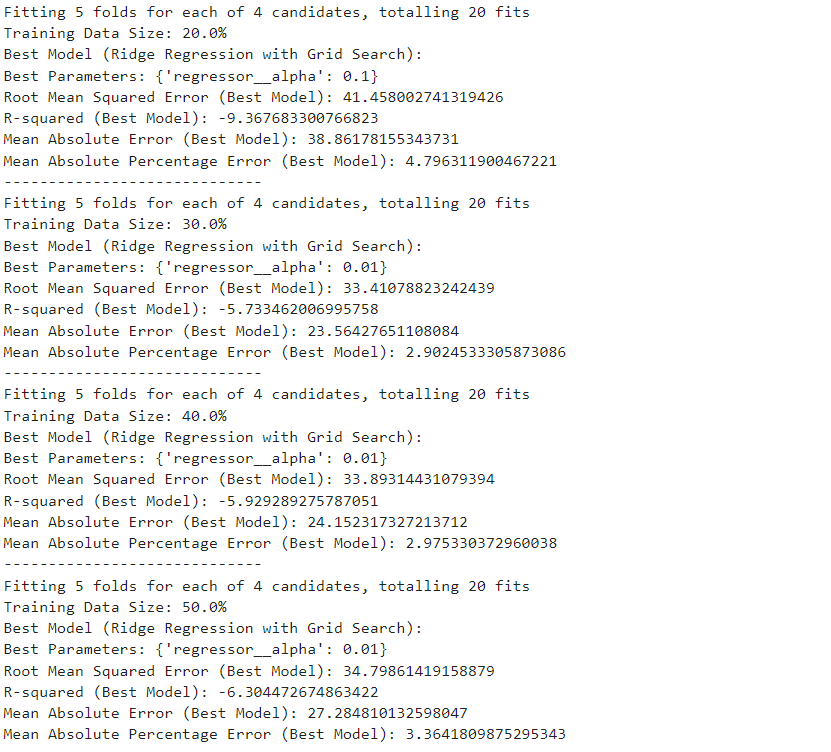
Analysis from the above data on the best model that was selected.

As we increase the data size the performance of the data model deteriorates, with a negative value in R-squared values indicating poor fit.

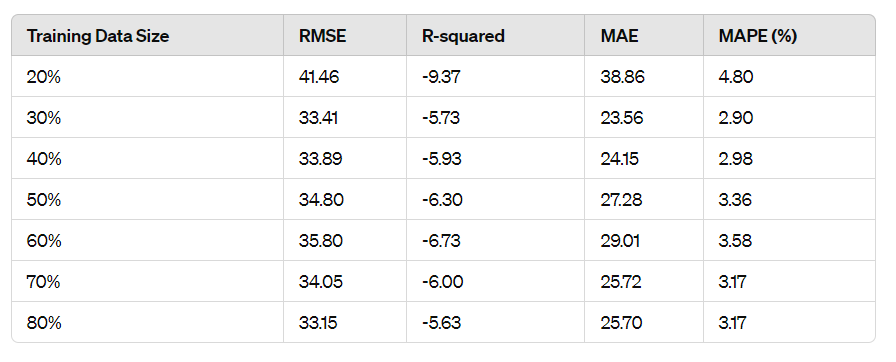
When we increase the data size it is not consistent

High RMSE and MAPE percentages suggest that the model predictions are significantly off from actual value.

1. In the same way we found that the best model for data set 2 is ridge regression we analyze the result from increasing and decreasing the data volume and its impact.







Summary table for better analysis.

Model performance improves with larger training data sizes, as seen in decreasing RMSE and increasing R-squared values.

The best model 30 %training data achieved an RMSE of 33.41 and an R-squared of -5.73, indicating moderate predictive accuracy.

Hyperparameter tuning optimal alpha =0.01 significantly influenced model performance.

Overall we analyzed that there is an impact of the data volume is there on the data set and also there are different models to calculate best.