Machine Learning Project Report

Selected Project Idea:

“House Price Prediction”

Team Members:

1st Team Member Name:

Kareem Saeed Ragab

1st Team Member ID:

2018170282

2nd Team Member Name:

Kareem Sherif Fathy

2nd Team Member ID:

2018170283

3rd Team Member Name:

Abanoub Asaad Azab

3rd Team Member ID:

2018170001

4th Team Member Name:

Nada El Sayed Anies

4th Team Member ID:

2018170430

5th Team Member Name:

Nada Mohamed Abdelhamed

5th Team Member ID:

2018170434

Milestone # 1: "Preprocessing and Regression"

1- Preprocessing:

1.1 Drop unwanted Data

- Drop ID column because it only refer to row index
- For this dataset we are considering eliminating those variables which have or have more than 45% of total data missing or showing no values. PoolQC, MiscFeature, Fence, and FireplaceQu attributes for predicting house prices have very large numbers of missing values, which are, 99%, 96%, 81% and 47% respectively. Since it covers almost whole columns, eliminating these attributes won't cause any significant loss of values to our project. It is also important to be aware that right evaluation is required to deal with such missing values because there may be a chance of loss of important data.

1.2 Solve missing values

- LotFrontage attribute will be considered in our dataset. It has 211 missing values, which is a significant number, and we will replace it with the previous used value in the same column.
- Replace null values in the other columns by the most frequent value.

1.3 Label encoding

1.4 Feature Selection

- Get the top 50% correlation features with the SalePrice (target).
- From the top features correlation we noticed that there is features that have high correlation for that reason we will remove the least correlation with SalePrice For Example: 1stFlrSF and TotalBsmtSF both of them have very high correlation for that reason we will remove one of them, we chose to remove 1stFlrSF because it has least correlation with SalePrice.

1.5 Feature Scaling by using Normalization

2- Regression techniques used:

2.1 Gradient Boosting Regressor.

2.2 Ridge Regression.

3- Differences between each model:

3.1 Gradient Boosting Regressor: Builds an additive model in a forward stage wise fashion, it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

- Accuracy: 88.3%
- Mean Square Error: 807478598.5
- Training Time: 1.2 s

3.2 Ridge Regression: Estimates the coefficients of multiple-regression models in scenarios where independent variables are highly correlated.

- Accuracy: 82.4%
- Mean Square Error: 1095906187.7
- Training Time: 1.0 s

4- Training & Testing Sizes:

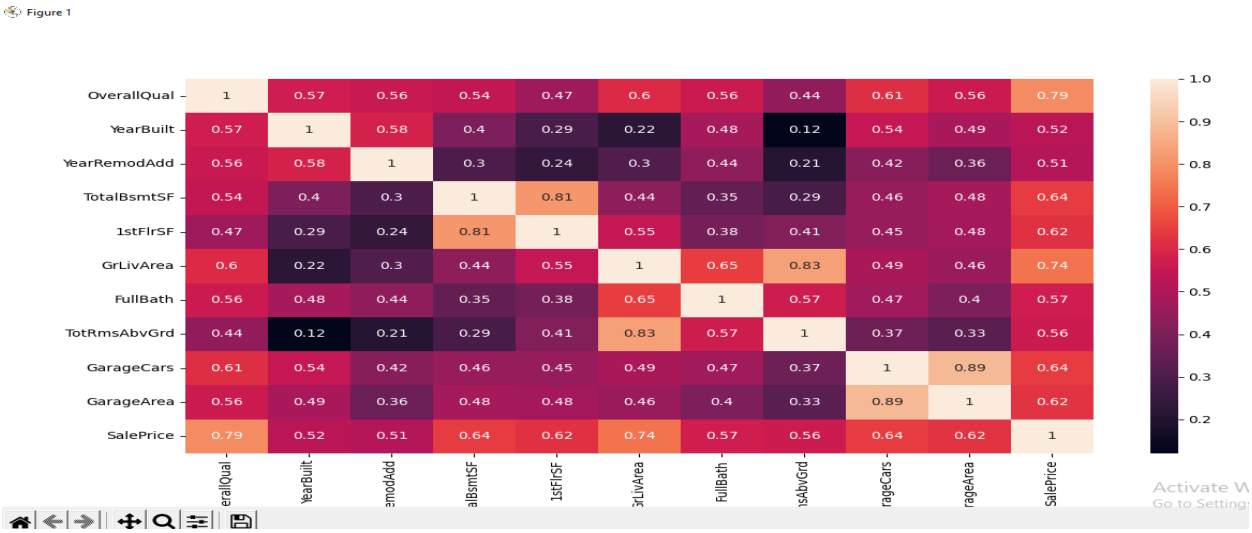
Training size is 80% of the total data size and the testing data size is 20%.

5- Conclusion:

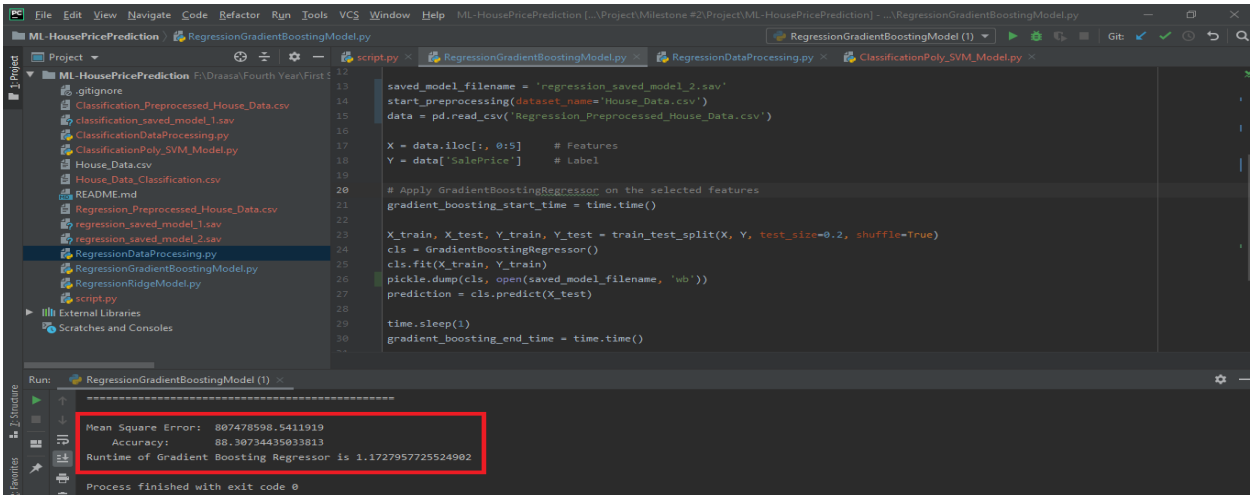
Prediction house prices are expected to help people who plan to buy a house so they can know the price range in the future, then they can plan their finance well.

6- Work Screenshots:

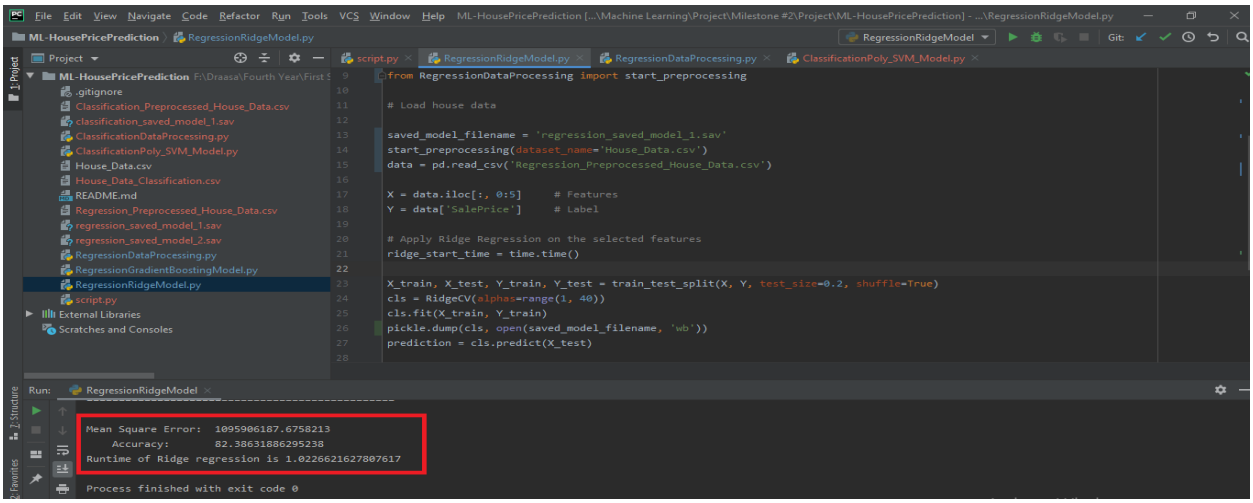
5.1 Dataset Features Correlation:



5.2 Gradient Boosting Regressor Model Output Information:



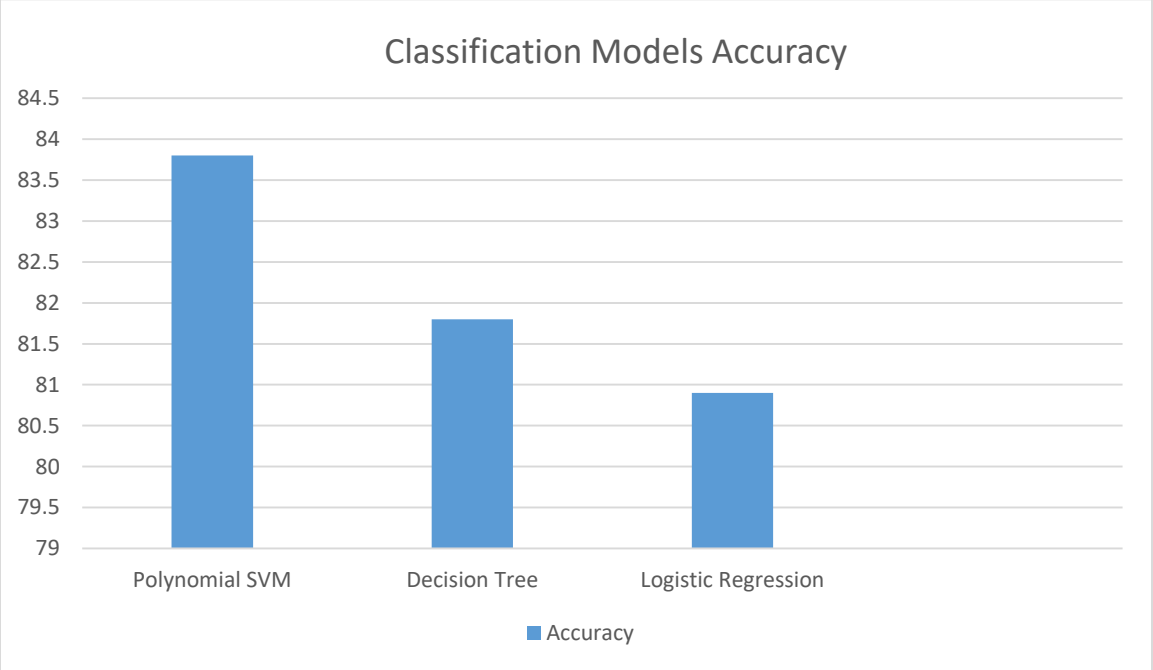
5.3 Ridge Regression Model Output Information:



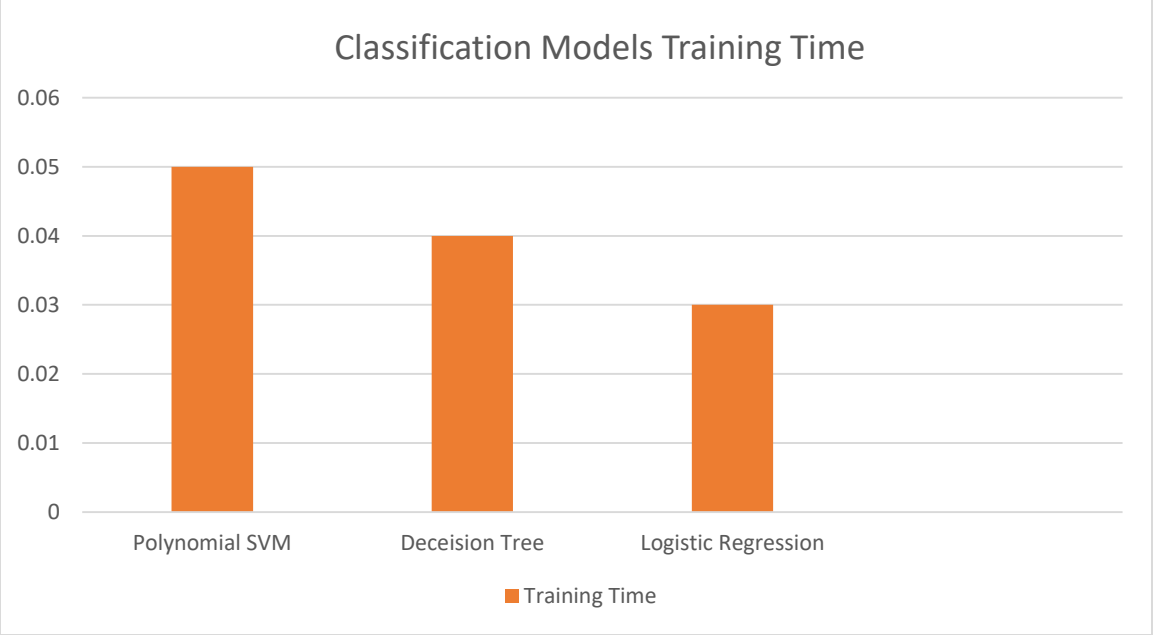
Milestone # 2: "Classification and Hyperparameter tuning"

1- Three Classification Models Summary:

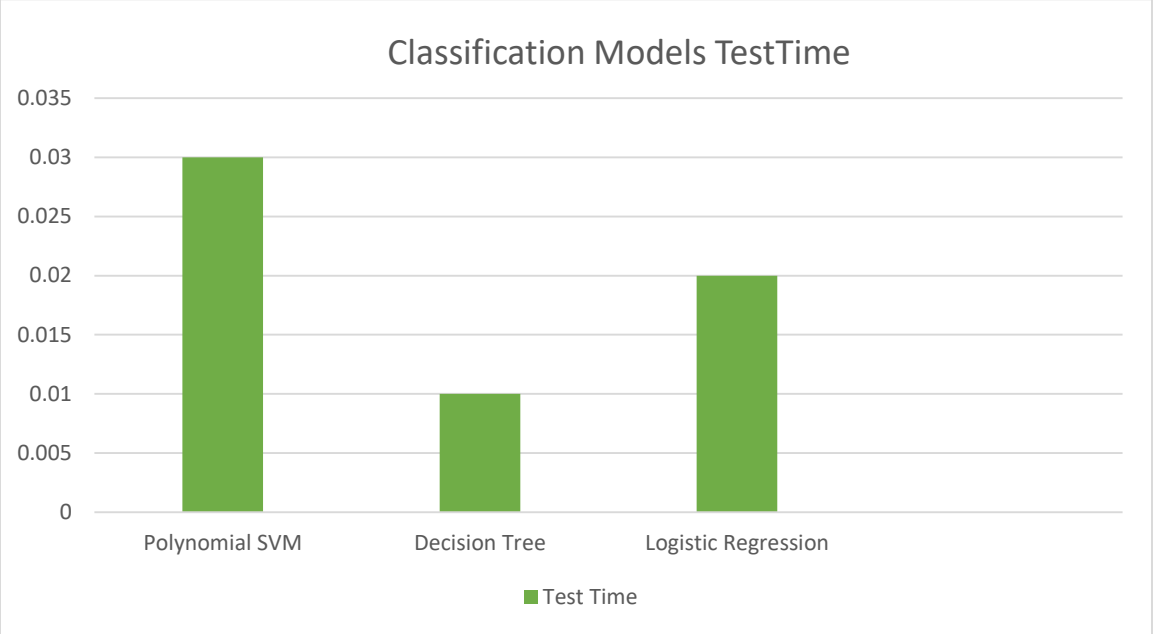
1.1 Accuracy



1.2 Training Time



1.3 Test Time



2- Feature Selection:

In this phase the feature selection differ from the previous phase So, Get the top 50% correlation features with the PriceRate (target).
HouseStyle and BsmtHalfBath features have the high correlation with PriceRate (target).

3- Hyperparameter Tuning:

3.1 According to Polynomial SVM Model. There are two hyperparameters Regularization parameter (C), Polynomial Degree (Degree)

C	Degree	Accuracy
0.1	3	78.1%
10	3	81.4%
50	3	84.7%

C	Degree	Accuracy
0.1	2	81.8%
0.1	3	78.1%
0.1	4	81.4%

3.2 According to Decision Tree Model. There are two hyperparameters Max Tree Depth (max_depth), Max Leaf Nodes (max_leaf_nodes)

max_depth	max_leaf_nodes	Accuracy
3	6	80.9%
4	6	81.8%
5	6	81.9%

max_depth	max_leaf_nodes	Accuracy
3	2	78.1%
3	3	78.9%
3	5	80.9%

3.3 According to Logistic Regression Model. There are two hyperparameters Regularization parameter (C), number of iterations (epochs)

C	epochs	Accuracy
0.1	1000	73.5%
10	1000	79.7%
50	1000	83.8%

C	epochs	Accuracy
0.5	1000	76.1%
0.5	3000	79.7%
0.5	5000	80.5%

4- Conclusion:

The correlation in this phase between the features and PriceRate (target) is low because the target has value in (cheap, moderate, expensive) So, the three values doesn't have strong correlation with other features that have various values. This problem solved when we begin to scale feature values between 0 to 2 then make the correlation. In this case the Accuracy updated to high and correlation became strong.