Cars in the Middle East

Team 6 - Nabeel Bacchus

Models Used

In order to have the best prediction of how expensive a car in the Middle East would be based on its features, the following four different regression models have been fitted to the dataset: multiple linear regression, polynomial regression, ridge regression, and lasso regression. Through the use of these models, we would be able to find a model that can more accurately predict the price of the vehicle.

To increase the accuracy of the models, multiple features of the dataset have been edited. Outliers with a z-score greater than 3 have been removed. A robust scaler was used to handle any additional outliers and to normalize the remaining dataset. Finally to prevent heteroscedasticity and to keep the variance of the error terms constant, we found all the square-root values of the "price" variable. To create the training set and test set, the dataset was randomly split to have 75 percent of the rows in the training set and 25 percent of the rows in the test set.

Hyperparameter Tuning

By using grid search, the best hyperparameters were found for polynomial regression, ridge regression, and lasso regression. Grid search was used to find a quick and accurate estimate of the best hyperparameters. For polynomial regression, the only hyperparameter changed was the degree with values 2 and 3. The degree could not be raised higher due to hardware restrictions. The hyperparameter that scored the highest was when the degree was equal to 3. The table below shows the results:

Polynomial Regression Degree	R-Squared	MAE	MSE
2	0.871	23.188	1026.613
3	0.878	21.385	974.745

Grid search was also used for finding the best hyperparameters for ridge regression. It checks which is the best solver between svd, cholesky, lsqr, and sag. The alpha parameter goes through values: 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 5, 10, and 100. It also checks if the 'fit_intercept' parameter should be true or false and if the values should be normalized. The table below shows the comparison between three models that score the best for ridge regression:

Alpha	Solver	fit_intercept	normalize	R-Squared	MAE	MSE
1	sag	true	false	0.851	24.886	1190.605
5	sag	true	false	0.851	24.883	1192.211
.001	lsqr	true	true	0.851	24.887	1190.509

Based on the results, we find that the hyperparameter that has the best performance is the one in the last row. This is because it has the lowest MSE out of the three models and had the best score according to grid search.

For lasso regression, we check the same hyperparameters except for the 'solver' parameter, which does not exist for lasso. Additionally, we check to find what is the best amount of iterations to go through. The values checked for 'max_iter' are 100, 1000, and 10000.

Alpha	Max Iterations	fit_intercept	normalize	R-Squared	MAE	MSE
1	10000	True	False	0.841	25.367	1265.933
5	1000	True	False	0.823	27.167	1410.557
.001	1000	True	False	0.851	24.888	1190.323

Based on the results, we find that the hyperparameter that has the best performance is the one in the last row. When alpha is equal to .001, the R-squared is highest amongst the other three models and is strong at explaining the data. It also has the lowest MSE and MAE out of the other three models.

Performance Evaluation

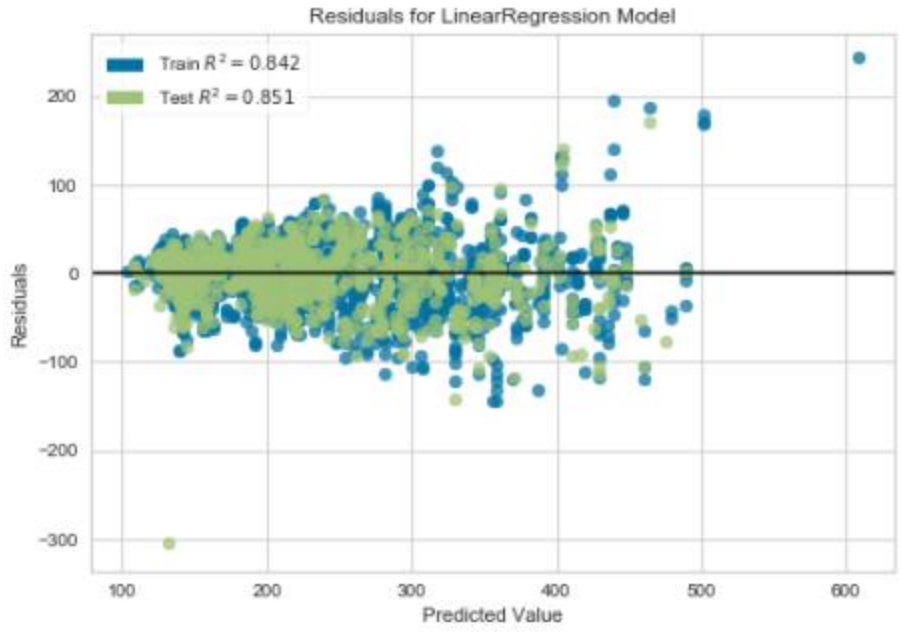
To measure the performance of all the models, The R-squared, mean squared error, and mean absolute error was calculated. The table below shows the comparisons between the four regression models that were made and performed the best among other models in their category:

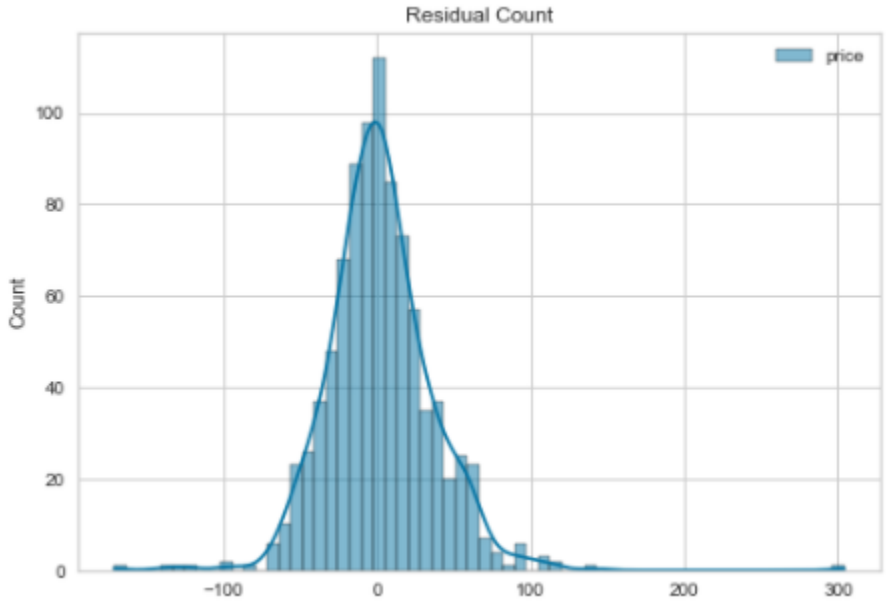
Regression Model	R-squared	MAE	MSE
Linear	0.851	24.888	1190.284
Polynomial (Degree = 3)	0.878	21.385	974.745
Ridge	0.851	24.887	1190.509
Lasso	0.851	24.888	1190.323

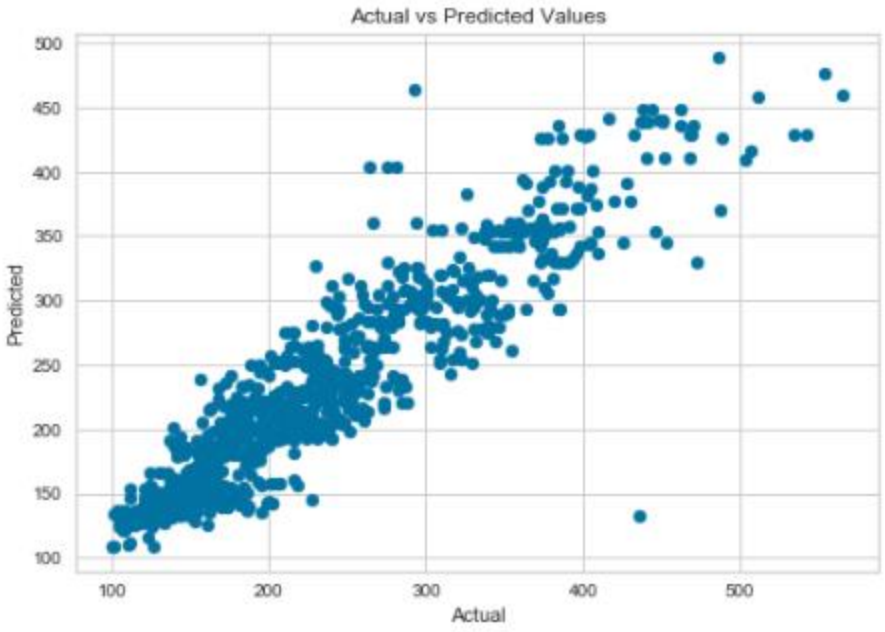
Based on the scores of all these models, the polynomial regression model has the highest R-squared and the lowest error rates. The linear, ridge, and lasso regression models all performed about the same or produced the same results for R-squared, MAE, and MSE.

For all of the regression models, they have high bias and low variance. This causes there to be some level of overfitting for all the models. To prevent this as much as possible, cross-validation was used.

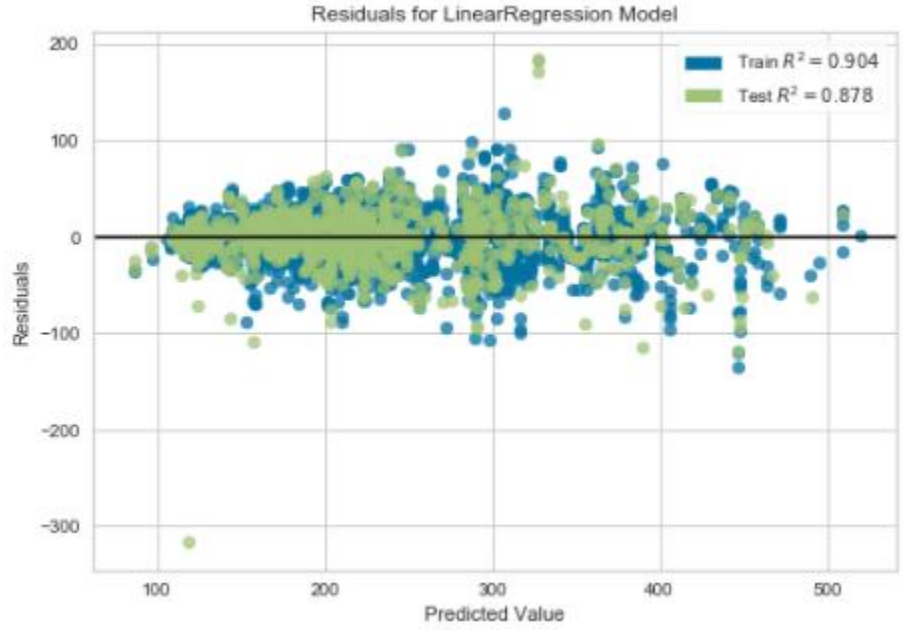
For linear regression, the residual graph is symmetrically distributed and there is a slight funnel pattern. The points are not clustered around lower digits for the residuals. The residuals are normally distributed. There is also a strong correlation between the actual values and the predicted value. There are a few outliers that exist that decrease the accuracy of the model.

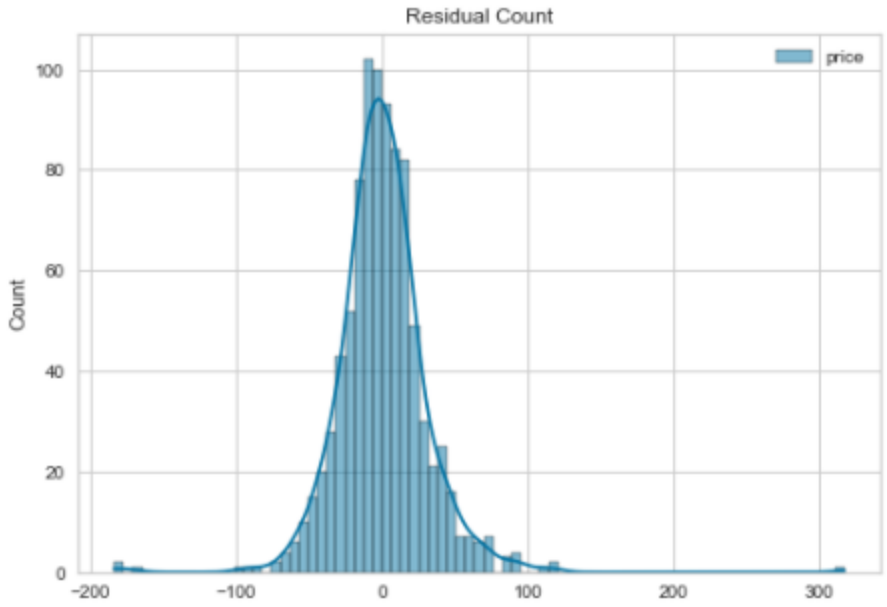

Residuals for LinearRegression Model
Train $R^2 = 0.842$
Test $R^2 = 0.851$
Residuals
Predicted Value
200
100
0
−100
−200
−300
100
200
300
400
500
600

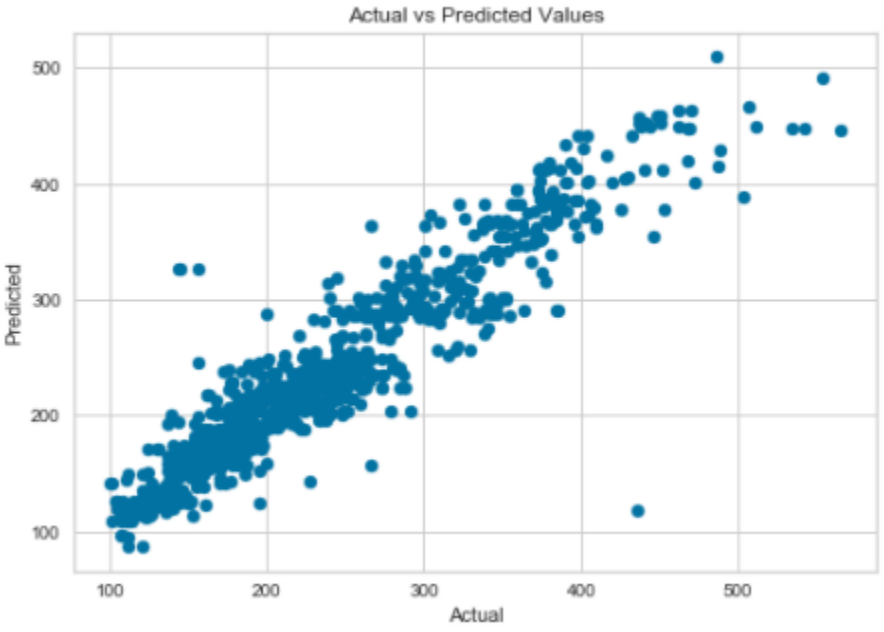

Residual Count
price
Count
100
80
60
40
20
0
−100
0
100
200
300



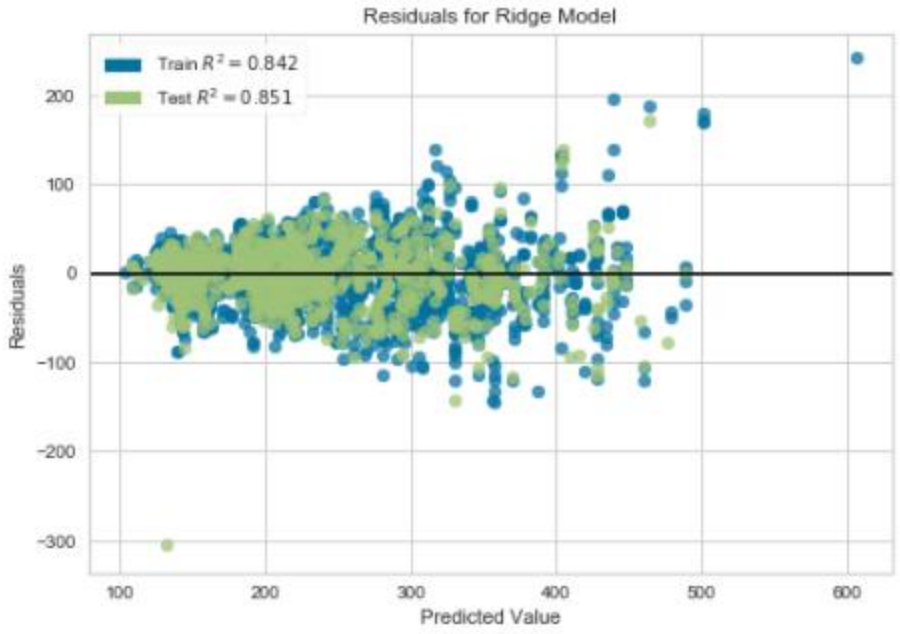
For polynomial regression where degree is equal to three, the residual graph is symmetrically distributed and there is not any clear pattern. However, the points are not clustered around lower digits for the residuals. The residuals are normally distributed. There is also a strong correlation between the actual values and the predicted value. There are a few outliers that exist that decrease the accuracy of the model.

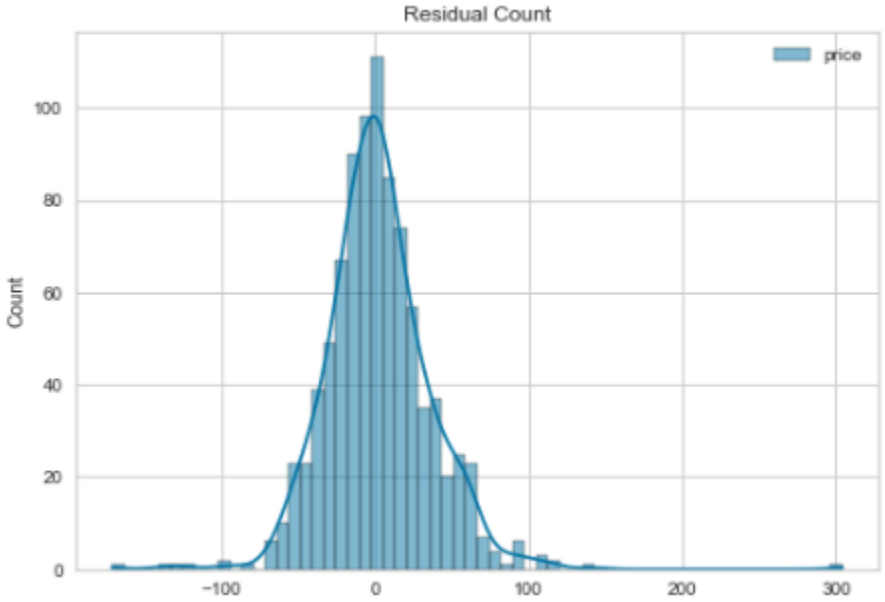


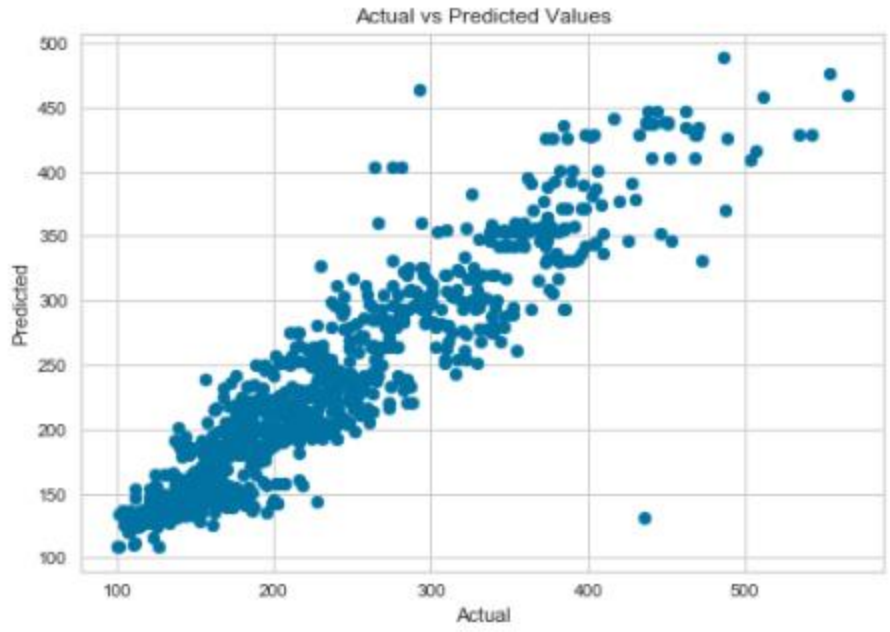




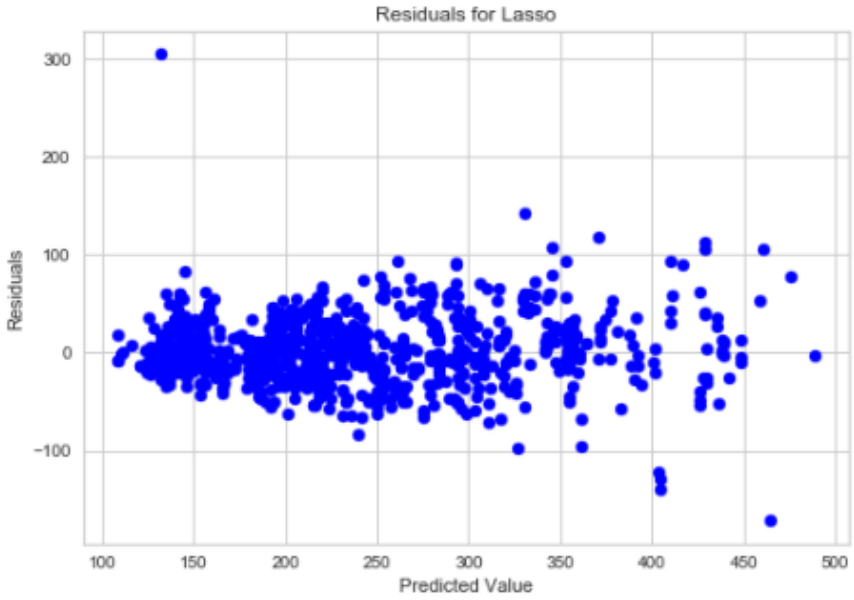
For the ridge regression model, the residual graph is symmetrically distributed and there is a slight funnel pattern. The points are not clustered around lower digits for the residuals. The residuals are normally distributed. There is also a moderate correlation between the actual values and the predicted value. There are a few outliers that exist that decrease the accuracy of the model.

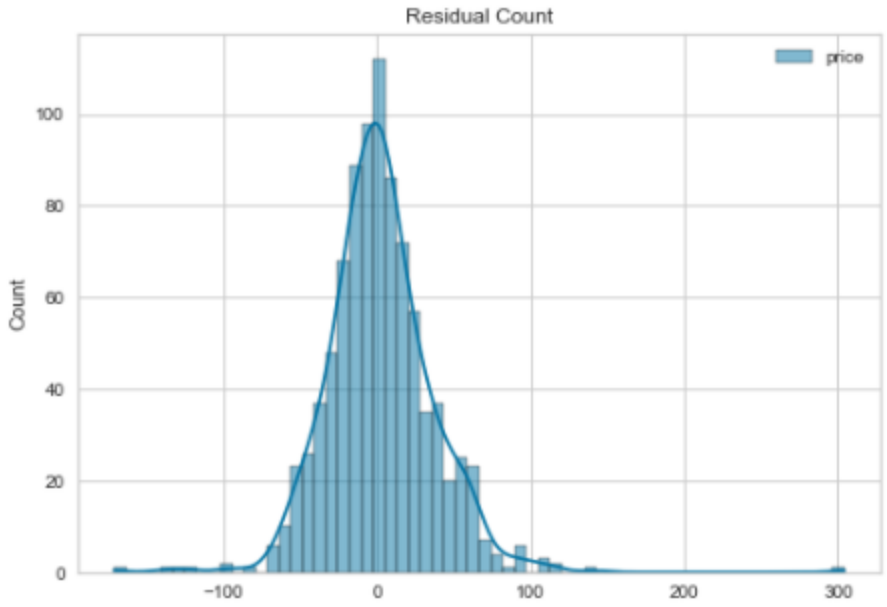

Residuals for Ridge Model
Train $R^2 = 0.842$
Test $R^2 = 0.851$
Residuals
Predicted Value
200
100
0
−100
−200
−300
100
200
300
400
500
600

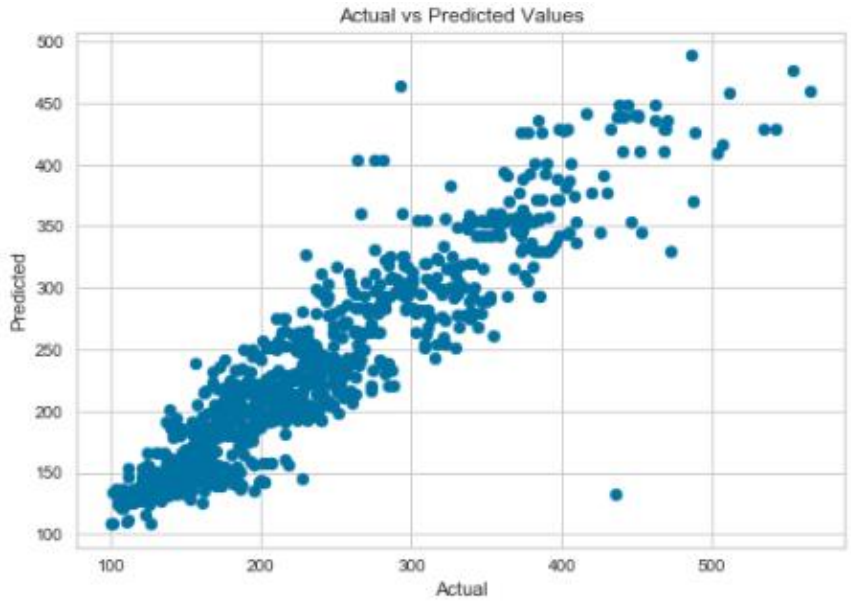

Residual Count
price
Count
100
80
60
40
20
0
−100
0
100
200
300



For lasso regression, the residual graph is symmetrically distributed and there is a pattern. The points are not clustered around lower digits for the residuals. The residuals are normally distributed. There is also a moderate correlation between the actual values and the predicted value. There are a few outliers that exist that decrease the accuracy of the model.







Based on the graphs and the performance of all the models, we found that the polynomial regression model where degree is equal to three has the best performance and is the most accurate for predicting the price of a car based on its features.