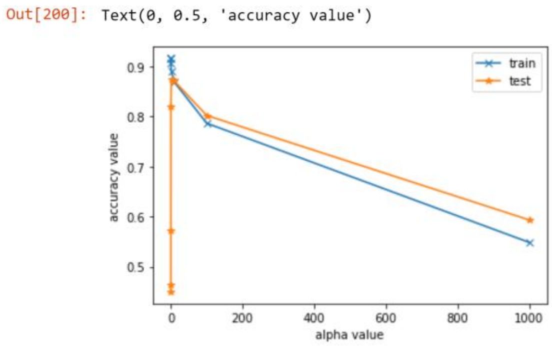
Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans: The alpha value for ridge and Lasso it will be 1 and 100 respectively as shown in the figure. Also, If we double the value for ridge and Lasso then the r2_score will go down slightly which can which can be also shown in the same figure below.



```
Out[200]: Text(0, 0.5, 'accuracy value')
```

Fig: accuracy at different alpha level for ridge

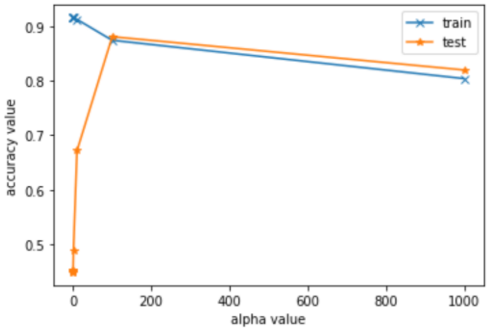
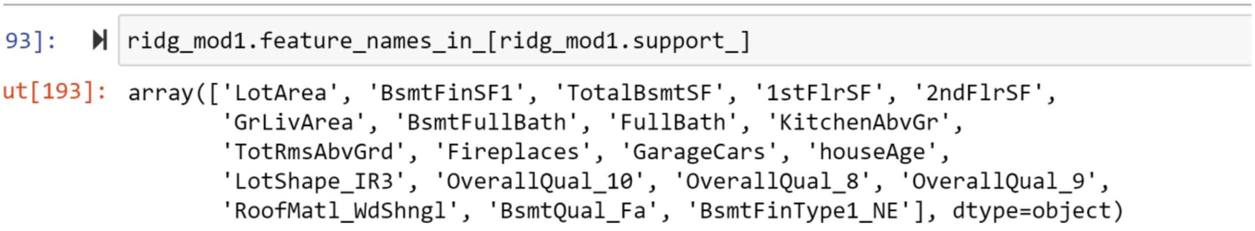


Fig: accuracy at different alpha level for lasso

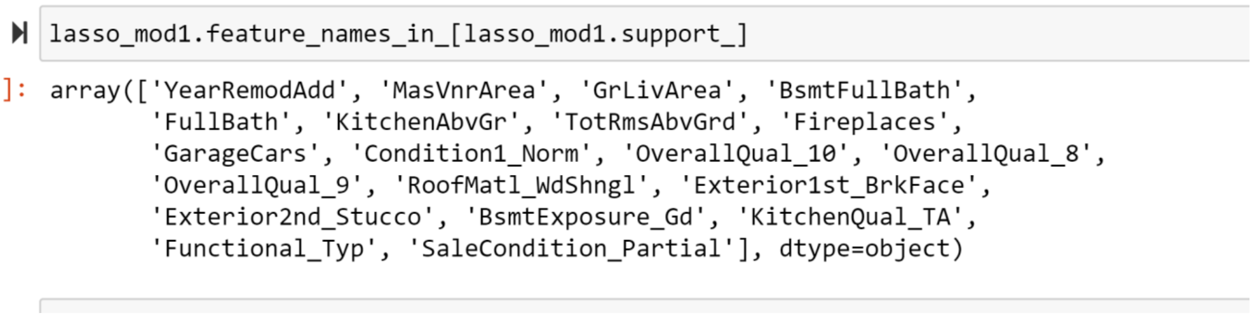
After the alpha value is doubled the top20 features in the model will be as :



```
93]:  ridg_mod1.feature_names_in_[ridg_mod1.support_]

ut[193]: array(['LotArea', 'BsmtFinSF1', 'TotalBsmtSF', '1stFlrSF', '2ndFlrSF',
                'GrLivArea', 'BsmtFullBath', 'FullBath', 'KitchenAbvGr',
                'TotRmsAbvGrd', 'Fireplaces', 'GarageCars', 'houseAge',
                'LotShape_IR3', 'OverallQual_10', 'OverallQual_8', 'OverallQual_9',
                'RoofMatl_WdShngl', 'BsmtQual_Fa', 'BsmtFinType1_NE'], dtype=object)
```

Fig: top20 feature as per ridge after doubling alpha



```
lasso_mod1.feature_names_in_[lasso_mod1.support_]
```

```
]: array(['YearRemodAdd', 'MasVnrArea', 'GrLivArea', 'BsmtFullBath',
          'FullBath', 'KitchenAbvGr', 'TotRmsAbvGrd', 'Fireplaces',
          'GarageCars', 'Condition1_Norm', 'OverallQual_10', 'OverallQual_8',
          'OverallQual_9', 'RoofMatl_WdShngl', 'Exterior1st_BrkFace',
          'Exterior2nd_Stucco', 'BsmtExposure_Gd', 'KitchenQual_TA',
          'Functional_Typ', 'SaleCondition_Partial'], dtype=object)
```

Fig: top20 feature as per lasso after doubling alpha

Question 2

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

Ans: we will choose the value 1 and value 100 for the ridge and lasso alpha respectively because at this level of alpha the model seems to be able to predict more that 86% of variability in the data and after this the model seems to have overcome over the problem of overfitting which can also be verified using the residual analysis done after the prediction has been done for the test dataset.

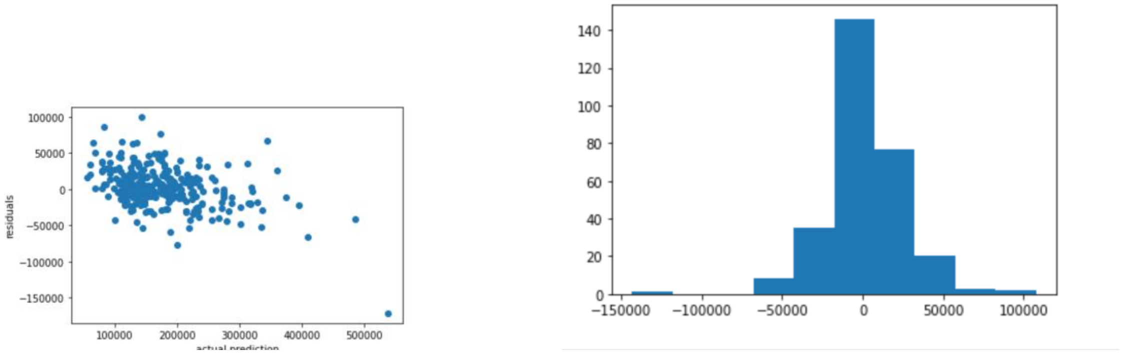


Fig: Analysing the residual plot which seems to satisfy the OLS assumption.

Question 3

After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

Ans: We can always select the important Variable for our model. Since the manual Feature selection will be difficult to decide on which value to choose and what to drop we can get help of the automation module of sklearn library called RFE (Recursive feature elimination) where we can calculate the rank for the different predictors and then can decide accordingly

For eg: In the above mentioned top 20 feature if 5 feature are missing then we can increase the feature selection count by 5 to incorporate 5 additional feature which are of high important on rank basis and then choose from them as :

RFE(estimator=lasso_model,n_features_to_select=25) then we will get 5 additional feature that are of high importance.

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Ans: In order to be sure on the robustness and the abilty to generalize the prediction for the model. We have to perform the residual analysis to see the following scenario :

a. if the model is overfitted or not,
b. If the Residual/Error is random or there are some pattern with respect to the prediction/predictors value or constant variation in error
c. If the Error are normally distributed around 0 or not
d. If there correlation among the predictor variables or not.

Once we are confirmed on the above checkpoints then we can say that model is robust and able to generalize well otherwise it might show the tendency of performing better on the data that it has already seen and might fail on unseen data to great extent affecting the decision making for whoever uses that particular model.