Problem Statement - Part II

Subjective Questions

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?

Answer:

The optimal value of alpha for ridge and lasso regression is 8 and 0.0001 respectively.

RMSE: Ridge -> 0.125791 Lasso -> 0.125515

If I double the value of Ridge and Lasso to 16 and 0.0002 then the RMSE value changes.

RMSE: Ridge -> `0.131046` Lasso -> `0.130526`, and of the value order changed

Initial Ridge Feature Order:	Updated Ridge Feature Order:	Initial Lasso Feature Order:	Updated Lasso Feature Order:
MSZoning_RL	OverallQual	MSZoning_RL	MSZoning_RL
OverallQual	GrLivArea	MSZoning_RM	OverallQual
GrLivArea	MSZoning_RL	OverallQual	GrLivArea
MSZoning_RM	2ndFlrSF	GrLivArea	MSZoning_RM
2ndFlrSF	OverallCond	MSZoning_FV	MSZoning_FV
MSZoning_FV	TotalBsmtSF	2ndFlrSF	OverallCond
OverallCond	MSZoning_FV	OverallCond	2ndFlrSF
TotalBsmtSF	MSZoning_RM	TotalBsmtSF	TotalBsmtSF
1stFlrSF	1stFlrSF	1stFlrSF	1stFlrSF
BsmtFinSF1	BsmtFinSF1	BsmtFinSF1	BsmtFinSF1

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?

Answer: Compared to Ridge the model performance by Lasso Regression was better in terms of R2 values of Train and Test. It is better to use Lasso, as it brings a zero value to insignificant features which allow us to choose the predictive variables.

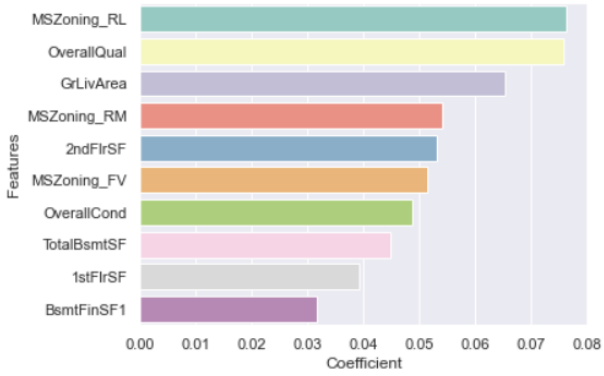
Question 3

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?

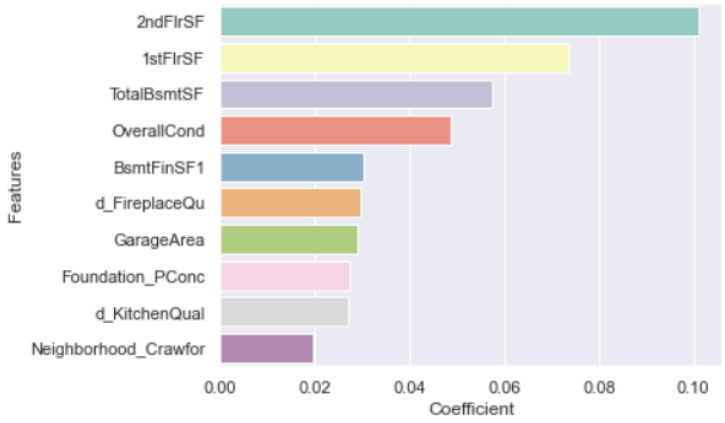
Answer: The new Features are:

2ndFlrSF
1stFlrSF
OverallCond
TotalBsmtSF
BsmtFinSF1
Foundation_PConc
d_FireplaceQu
d_KitchenQual
Neighborhood_Crawfor
d_SaleCondition

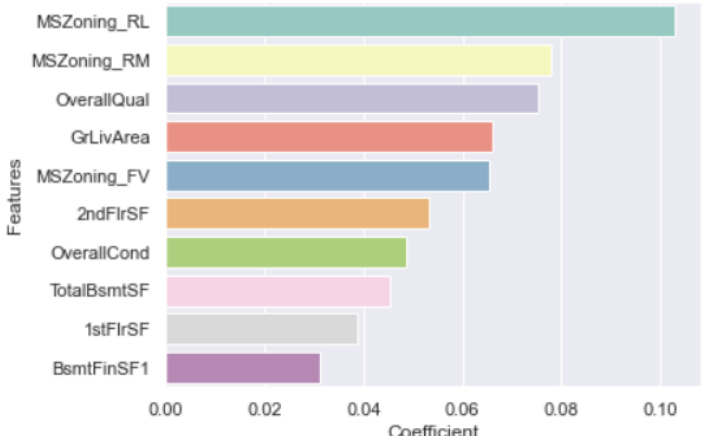
Ridge Original:



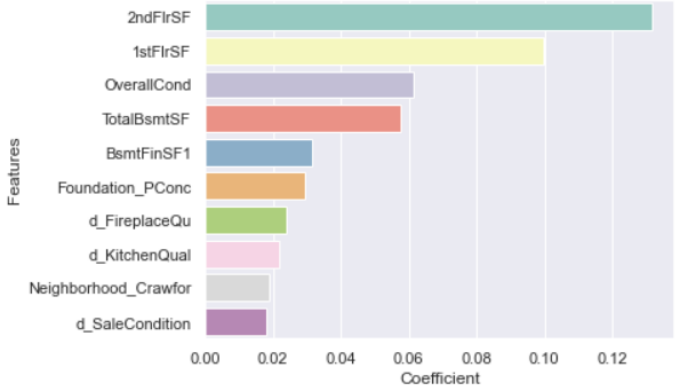
Ridge Feature Order after dropping:



Lasso Original:



Lasso Feature Order after dropping:



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?

Answer:

Robust model implies that when any variation in the data does not affect its performance, example Outliers in the training data. A **generalisable** model is able to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model, so that the test accuracy is not lesser than the training score. This can be **take care by fixing the overfit issue in the model**. An overfitting model has very high variance and a smallest change in data affects the model prediction heavily. Such a model will identify all the patterns of a training data, but fail to pick up the patterns in unseen test data. The Model should be simple and no under fit or overfit

When we say **Accuracy**, a too complex model will have a very high accuracy. So, to make our model more robust and generalizable, we will have to decrease variance which will lead to some bias. Addition of bias means that accuracy will decrease. We need to maintain a balanced bias and variance value. Where Regularization techniques like Ridge Regression and Lasso can help to main these values by means of lambda values or shrinkage penalty values.