SUBJECTIVE QUESTIONS

IITB, UPGRAD :

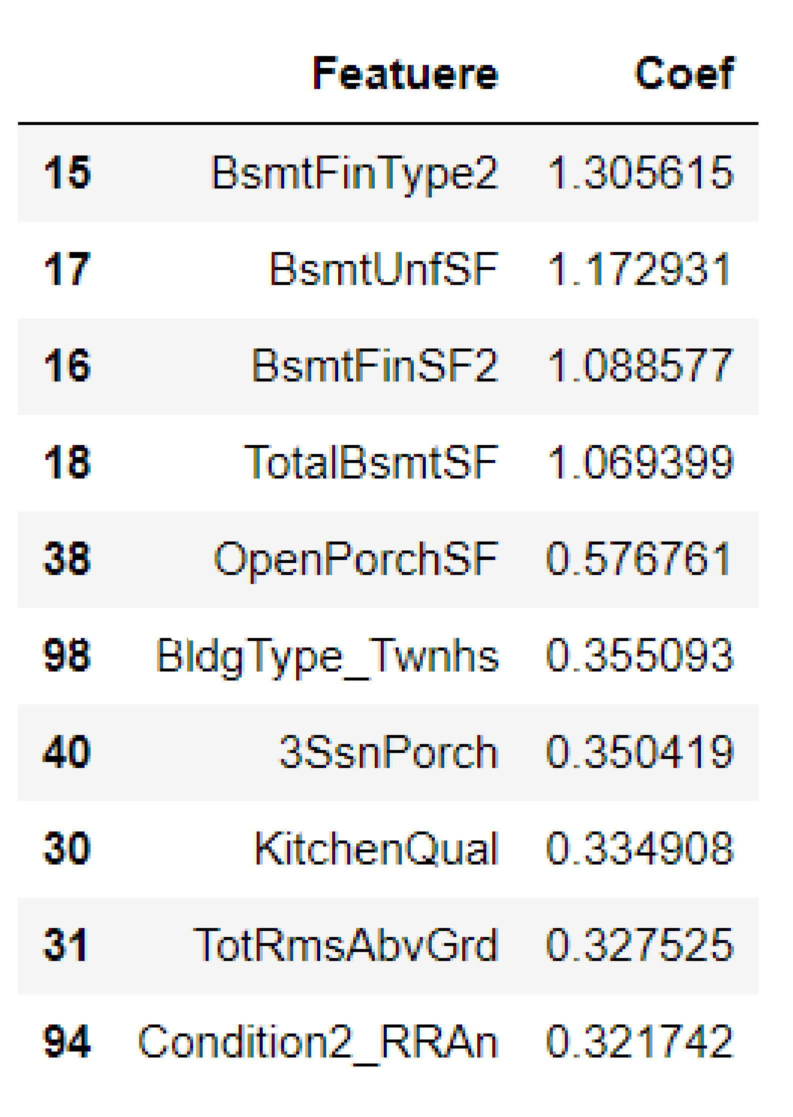
VAMSHI.KRISHNA.PRIME@GMAIL.COM

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?

The optimal value of alpha for lasso and ridge are 0.0001 and 1.0 respectively. If we increase the alpha(hyper parameter value) the accuracy of the model starts dropping gradually. It might increase a bit till the optimal hyper parameter value but the accuracy will dcvecrease with the increase in alpha and model will become more parse.

the most important predictor variables after the change is implemented are:



	Featuere	Coef
15	BsmtFinType2	1.305615
17	BsmtUnfSF	1.172931
16	BsmtFinSF2	1.088577
18	TotalBsmtSF	1.069399
38	OpenPorchSF	0.576761
98	BldgType_Twnhs	0.355093
40	3SsnPorch	0.350419
30	KitchenQual	0.334908
31	TotRmsAbvGrd	0.327525
94	Condition2_RRAn	0.321742

QUESTION 2

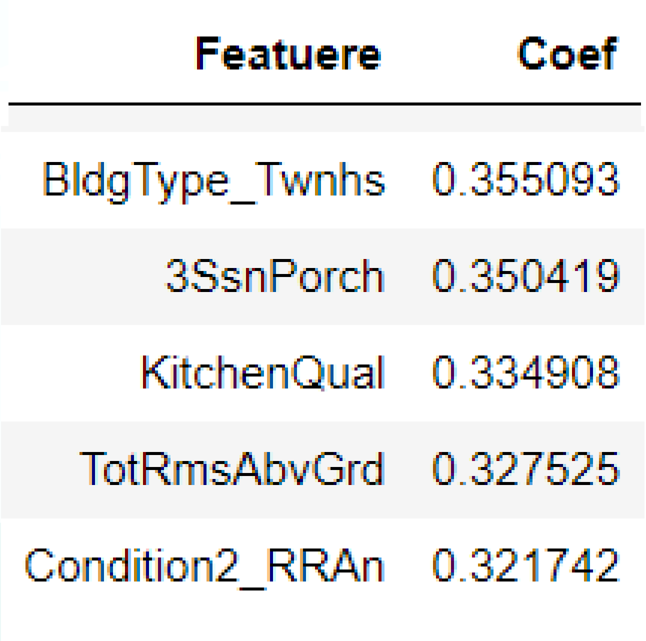
YOU HAVE DETERMINED THE OPTIMAL VALUE OF LAMBDA FOR RIDGE AND LASSO REGRESSION DURING THE ASSIGNMENT. NOW, WHICH ONE WILL YOU CHOOSE TO APPLY AND WHY?

After creating model in both Ridge and Lasso regressions, we can see that the R2 scores are similar for both Lasso and Ridge regression. However, as Lasso regression penalize more on the dataset, in-turn helping in more feature elimination, we will consider Lasso Regression model as the final model.

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?

The next best 5 variables that explain the model are:



Featuere	Coef
BldgType_Twnhs	0.355093
3SsnPorch	0.350419
KitchenQual	0.334908
TotRmsAbvGrd	0.327525
Condition2_RRAn	0.321742

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?

The model should be as simple as possible, though its accuracy will decrease but it will be more robust and generalisable. It can be also understood using the Bias-Variance trade-off. The simpler the model the more the bias but less variance and more generalizable. Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

Bias: Bias is error in model, when the model is weak to learn from the data. High bias means model is unable to learn details in the data. Model performs poor on training and testing data.

Variance: Variance is error in model, when model tries to over learn from the data. High variance means model performs exceptionally well on training data as it has very well trained on this of data but performs very poor on testing data as it was unseen data for the model.
It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.