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?

Ans: The optimal value of alpha obtained for ridge is **0.7** and for lasso is **0.0001** in the given problem.

When the alpha is doubled for ridge the values are :

METRICS	**Ridge(alpha=0.7)**	**Ridge(alpha=1.4)**
Train R2	0.907351	0.906076
Test R2	0.901586	0.902155
MAPE R2	-0.635351	-0.432668
Train RSS	30143.858827	29999.361714
Test RSS	2107.795977	2098.196991
Train MSE	0.001104	0.001119
Test MSE	0.001313	0.001305
Train MAE	0.020811	0.021019
Test MAE	0.022087	0.021896

Ridge(alpha=0.7 is the previous observations) and alpha=1.4 is the doubled observation.

It is observed that there are only a minute changes in the metrics.

Top 10 Features	Ridge(alpha=0.7)	Ridge(alpha=1.4)
OverallQual_10	0.116150	0.104884
GrLivArea	0.101198	0.095991
2ndFlrSF	0.094090	0.087379
OverallQual_9	0.090075	0.081716
TotalBsmtSF	0.085206	0.076612
1stFlrSF	0.083225	0.079944
OverallCond	0.075873	0.072020
MSZoning_FV	0.056769	NaN
Neighborhood_StoneBr	0.056251	0.054252
YearBuilt	0.053689	0.050089
Neighborhood_NoRidge	NaN	0.049598

In the coefficients the top 5 features seem unchanged but few features in the old model isn’t there in the new model and vise versa like MSZoning FV and Neighbourhood no ridge.

Similarly for lasso

METRICS	Lasso(alpha=0.0001)	Lasso(alpha=0.0002)
Train R2	0.901423	0.897346
Test R2	0.900261	0.902710
MAPE R2	0.142748	0.324929
Train RSS	29817.944239	29431.921662
Test RSS	2082.772905	2054.559713
Train MSE	0.001175	0.001223
Test MSE	0.001298	0.001330
Train MAE	0.021367	0.021826
Test MAE	0.021434	0.021374

Similar metrics for both, but the older lasso is slightly better.

Top 10 features	Lasso(alpha=0.0001)	Lasso(alpha=0.0002)
GrLivArea	0.254713	0.274772
OverallQual_10	0.121841	0.116109
OverallQual_9	0.094370	0.093355
OverallCond	0.072405	0.062060
TotalBsmtSF	0.065788	0.059435
YearBuilt	0.059613	0.055925
BsmtFinSF1	0.053903	0.053392
Neighborhood_StoneBr	0.050274	0.041017
Neighborhood_NoRidge	0.046812	0.042547
OverallQual_8	0.040477	0.038164

The coefficient values haven’t changed much . Cos the alpha values where very small.

(The code for getting these tables are present in the jupyter notebook file)

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 : The following table shows the values for ridge and lasso

METRICS	Ridge(alpha=0.7)	Lasso(alpha=0.0001)
Train R2	0.907351	0.901423
Test R2	0.901586	0.902710
MAPE R2	-0.635351	0.142748
Train RSS	30143.858827	29817.944239
Test RSS	2107.795977	2082.772905
Train MSE	0.001104	0.001175
Test MSE	0.001313	0.001298
Train MAE	0.020811	0.021367
Test MAE	0.022087	0.021434

Ridge and lasso both show good values in the evaluation , but lasso has slightly better score due to lower rss , mae and less difference between train and test r2 score.

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: The top 5 predictors of lasso were : ['GrLivArea', 'OverallQual_10', 'OverallQual_9', 'OverallCond','TotalBsmtSF'], after removing them the evaluation metrics are as follows:

metrics	Lasso(alpha=0.0001)
Train R2	0.884917
Test R2	0.886066
MAPE R2	0.129901
Train RSS	29548.178014
Test RSS	2108.900019
Train MSE	0.001372
Test MSE	0.001520
Train MAE	0.024234
Test MAE	0.024046

And the current top 5 features are :

1stFlrSF	0.196159
2ndFlrSF	0.150368
BsmtFinSF1	0.082796
Neighborhood_StoneBr	0.059079
KitchenQual_Fa	0.056797

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: For a model to be reliable and effective, its robustness and generalizability must be ensured. To develop these qualities, the following are to be followed:

High-Quality Training Data: Begin with a training dataset that is broad and representative of the target domain and includes a wide range of scenarios, edge cases, and variations. As a result, the model is better able to generalise to new data and learn patterns.

Outlier detection: Locate outliers or anomalies in the data and deal with them accordingly. When outliers are not representative of the target distribution, they can have a major effect on model training. Effective preprocessing methods and outlier removal procedures can lessen their impact.

Training, Validation, and Test Sets: Divide the dataset into these three categories. The test set is saved for the model's final evaluation whereas the validation set is utilized to adjust hyperparameters and make judgements throughout model development. This division makes sure the model isn't overfitting to the validation set and gives a trustworthy assessment of how well it generalises.

Cross-Validation: Cross-validate your results, especially if the dataset is small. Using this method, the data is divided into various subsets, and the model is trained/evaluated on various combinations of these subsets. It offers a more thorough evaluation of the model's effectiveness and generalizability across various data splits.

Model Complexity: To balance the model's complexity, consider both the quantity of the training dataset and the model's capacity (the number of parameters). Overfitting can occur if the model is overly complex in comparison to the data that are available. On the other side, a model with limited capability would find it difficult to identify intricate patterns, which would lead to underfitting.

Regularization: Use regularisation methods like L1/L2 regularisation, dropout, or early halting to reduce the likelihood of overfitting. Regularisation improves the model's ability to generalise to new cases by preventing it from memorising the training data or depending too heavily on particular features.

Testing on unseen data: Analyse how the model performs in scenarios or adversarial instances that are intended to test its robustness. Adversarial testing can assist in identifying flaws and weaknesses in the decision-making process of the model, enabling the creation of countermeasures.

Implications for Model accuracy

There are frequently trade-offs between enhancing robustness and generalizability and maintaining model correctness. The model may not obtain the same level of accuracy on the training data by emphasizing these qualities as it would with overfitting. When used with fresh, untested data, a robust and generalizable model's overall performance is more trustworthy and significant. It lessens the possibility of unanticipated failures in real-world circumstances, resulting in an improvement in the predictability and efficacy of the model. Thus the accuracy of the model can be maintained by keeping balance between the bias and variance. The best model will always have low bias and low variance