Submission by Nitin Balaji Srinivasan – Cohort 58 – AI and ML

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 for alpha for ridge and lasso regression based on the models that I have developed is below:

Ridge alpha: 50.000

Lasso alpha: 0.001

The comparison summary based on R2 score and RMSE for these alpha values is below.

```
ridge  Regression with  50
====================================
R2 score (train) :  0.9020496865615709
R2 score (test) :  0.8759022434812159
RMSE (train) :  0.12525145087376433
RMSE (test) :  0.13998244238107563
```

```
lasso  Regression with  0.001
====================================
R2 score (train) :  0.9021510515706596
R2 score (test) :  0.8774506498742309
RMSE (train) :  0.12518662514533568
RMSE (test) :  0.1391063988512192
```

Lasso model performs better with higher R2 scores and lower RMSE value.

The most important (top 10) predictor variables for the existing models are below with their coefficients are below.

Ridge:

List	Ridge (alpha=50.0)	Lasso (alpha=0.001)
OverallQual	0.0916	0.1026
GrLivArea	0.0581	0.1091
OverallCond	0.0562	0.0594

Neighborhood_Crawfor	0.0472	0.0974
1stFlrSF	0.0416	0.0101
GarageArea	0.0383	0.0411
Condition1_Norm	0.038	0.0501
FullBath	0.0363	0.0326
SaleCondition_Normal	0.0352	0.0471
CentralAir_Y	0.0351	0.0474

Lasso:

List	Ridge (alpha=50.0)	Lasso (alpha=0.001)
GrLivArea	0.0581	0.1091
OverallQual	0.0916	0.1026
Neighborhood_Crawfor	0.0472	0.0974
SaleType_New	0.027	0.0702
Neighborhood_StoneBr	0.0299	0.0665
Neighborhood_NridgHt	0.035	0.0596
OverallCond	0.0562	0.0594
Condition1_Norm	0.038	0.0501
CentralAir_Y	0.0351	0.0474
SaleCondition_Normal	0.0352	0.0471

If we double the values for alpha in both ridge and lasso regression, then the accuracy of both the models reduce in both train and test sets. Ridge model performs slightly better with higher R2 score and lower RMSE value

Ridge alpha: 100.000

Lasso alpha: 0.002

The comparison summary based on R2 score and RMSE for these alpha values is below.

```
Model Re-evaluation : Ridge Regression, alpha=100.0
R2 score (train) :  0.8937
R2 score (test) :  0.874
RMSE (train) :  0.1305
RMSE (test) :  0.1411


Model Evaluation : Lasso Regression, alpha=0.002
R2 score (train) :  0.889
R2 score (test) :  0.8731
RMSE (train) :  0.1333
RMSE (test) :  0.1415
```

The most important (top 10) predictor variables for the updated models are below with their new coefficients.

Ridge:

Feature list	Ridge (alpha=50.0)	Lasso (alpha=0.001)	Ridge (alpha=100.0)	Lasso (alpha=0.002)
OverallQual	0.0916	0.1026	0.0887	0.1131
GrLivArea	0.0581	0.1091	0.0539	0.1048
OverallCond	0.0562	0.0594	0.0538	0.0588
GarageArea	0.0383	0.0411	0.0402	0.0423
1stFlrSF	0.0416	0.0101	0.0387	0.0144
FullBath	0.0363	0.0326	0.0365	0.0298
Neighborhood_Crawfor	0.0472	0.0974	0.0308	0.0681
2ndFlrSF	0.0334	0	0.0308	0
Condition1_Norm	0.038	0.0501	0.0299	0.0428
CentralAir_Y	0.0351	0.0474	0.0267	0.035

Lasso:

Feature list	Ridge (alpha=50.0)	Lasso (alpha=0.001)	Ridge (alpha=100.0)	Lasso (alpha=0.002)
OverallQual	0.0916	0.1026	0.0887	0.1131
GrLivArea	0.0581	0.1091	0.0539	0.1048
Neighborhood_Crawfor	0.0472	0.0974	0.0308	0.0681
OverallCond	0.0562	0.0594	0.0538	0.0588
Condition1_Norm	0.038	0.0501	0.0299	0.0428
GarageArea	0.0383	0.0411	0.0402	0.0423
SaleType_New	0.027	0.0702	0.0195	0.0414
Foundation_PConc	0.0313	0.0373	0.0266	0.0397
Neighborhood_NridgHt	0.035	0.0596	0.0247	0.0357
CentralAir_Y	0.0351	0.0474	0.0267	0.035

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: Between the two models, the final model will be lasso model at alpha value =0.001.

The comparison summary based on R2 score and RMSE for these alpha values is below.

```
ridge  Regression with  50
==================================
R2 score (train) :  0.9020496865615709
R2 score (test) :  0.8759022434812159
RMSE (train) :  0.12525145087376433
RMSE (test) :  0.13998244238107563
```

```
lasso  Regression with  0.001
==================================
R2 score (train) :  0.9021510515706596
R2 score (test) :  0.8774506498742309
RMSE (train) :  0.12518662514533568
RMSE (test) :  0.1391063988512192
```

From the comparison summary above, we can see that the lasso model outperforms the ridge model at alpha = 0.001

- it produced higher R2 score and lower RMSE score for both train and test datasets
- it reduces the coefficients to 0 for the least significant features.
- Out of 207 variables used, lasso has identified 78 strong predictor variables with accuracy

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 five most important predictor variables in the lasso model that are identified to be unavailable in the incoming data are as follows:

GrLivArea
OverallQual
Neighborhood_Crawfor
SaleType_New
Neighborhood_StoneBr

Creating another lasso model excluding the variables above, will yield the following results.

```
The updated optimal alpha value for lasso is 0.001000
lasso  Regression with  0.001
==================================
R2 score (train) :  0.8895716043941054
R2 score (test) :  0.8534488694382513
RMSE (train) :  0.13299038138709815
RMSE (test) :  0.1521199549607072
```

The top five predictors based on the new lasso model are below (with their coefficients).

Feature List	lasso_model_check (alpha=0.001)
1stFlrSF	0.1104
2ndFlrSF	0.1091
Neighborhood_NridgHt	0.0726
OverallCond	0.072
CentralAir_Y	0.0676

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: Regression models can be made more robust and generalisable using a combination of the following factors.

1. **Bias vs Variance Trade off** – We should always aim for Low Bias and Low variance, however to generalise the model, we may need to compromise to have a higher bias to achieve lower variance in testing, when the model predicts unknown data.
2. **Keeping the model simple**, subject to achieving less variance vs a overfitted model e.g. choosing lower number of predictor variables
3. Choosing the **correct value of alpha (hyperparameter) is critical –** a correct value of alpha helps to achieve bias vs variance trade off. Alpha parameter is introduced to address overfitting (i.e. where the model memorises the values).
    a. If alpha value is very low, it will not address overfitting
    b. As alpha value increases, the bias also increases
    c. If alpha value is very high, the model will start to "underfit"

In making the model more robust and generalisable (i.e. avoid overfitting and make it suitable to predict unknown data with a level of accuracy**), there will be implications on the accuracy.**

This means accepting a level of inaccuracy at the start (increase in bias) to ensure low variance.

i.e. While the generalised model will be slightly inaccurate during training (bias) this generalised model will be more accurate in predicting unknown data than a overfitted model