Submitted By – Varun Laroiya

Assignment-based 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?

If you double the value of alpha for both ridge and lasso regression, it will increase the regularization strength. This means that the penalty for large coefficients will be higher, leading to more shrinkage of coefficient values towards zero

For Ridge Regression:

- With double the value of alpha, the penalty for large coefficients will be stronger, leading to even more shrinkage of coefficient values towards zero compared to the original value of alpha.
- This can lead to further reduction in model complexity and potentially better generalization performance on unseen data.

For Lasso Regression:

Doubling the value of alpha will also increase the penalty for large coefficients, encouraging even sparser solutions where more coefficients are driven to exactly zero.

This can lead to even more feature selection, with potentially fewer predictor variables being considered important by the model

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?

Choosing between Ridge and Lasso regression depends on the specific characteristics of your dataset and the goals of your analysis. Here are some factors to consider:

1. **Feature Importance**: If you believe that only a subset of features is truly important for prediction and want a sparse model that selects only those features, Lasso regression (with L1 regularization) would be more appropriate. On the other hand, if you believe that all features are relevant but some may be noisy or redundant, Ridge regression (with L2 regularization) may be more suitable as it will shrink the coefficients of less important features towards zero but not eliminate them entirely.
2. **Model Interpretability**: Lasso regression tends to produce sparse models by setting some coefficients exactly to zero, which can make the model more interpretable by highlighting the most important features.

3. **Computational Considerations**: Lasso regression with a large number of features can be computationally expensive, especially if the dataset is large. Ridge regression typically has a more straightforward computational implementation.
4. **Bias-Variance Tradeoff**: Ridge regression generally handles multicollinearity better than Lasso regression since it doesn't force coefficients to be exactly zero. Thus, if multicollinearity is a concern, Ridge regression might provide better predictions with less variance.
5. **Cross-Validation Performance**: Ultimately, the decision may come down to which model performs better on cross-validation or holdout datasets. You should evaluate both Ridge and Lasso regression using cross-validation and choose the one that yields the best predictive performance.

Q3. 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?

To ensure that a model is robust and generalizable, especially when using ridge and lasso regularization, you can follow these strategies:

Cross-validation: Utilize techniques like k-fold cross-validation to assess the model's performance on multiple subsets of the data. This helps in evaluating how well the model generalizes to unseen data and reduces the risk of overfitting.

Regularization Strength Tuning: Perform a grid search or similar techniques to tune the regularization parameter (alpha) for ridge and lasso regression. This helps in finding the optimal balance between bias and variance, leading to a more robust model.

Evaluate on Holdout Data: Reserve a portion of your dataset as a holdout set and evaluate the model's performance on this unseen data. This gives a more realistic estimate of how the model will perform on completely new data.

Feature Scaling: Ensure that all features are properly scaled, especially if using regularization techniques like ridge and lasso. This helps in preventing features with larger scales from dominating the regularization process.

Feature Selection: With lasso regularization, feature selection occurs naturally as some coefficients are driven to zero. However, it's essential to validate the importance of the selected features and ensure they are truly meaningful for prediction.

Q4. 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?

1st Iteration of Lasso

	index	Ridge	Lasso
13	GrLivArea	0.047546	0.186668
107	RoofMatl_WdShngl	0.042824	0.162880
7	BsmtFinSF1	0.034423	0.137216
3	OverallQual	0.068929	0.115722
11	1stFlrSF	0.047462	0.106601
...	...	...	...
179	KitchenQual_TA	-0.033191	-0.029007
178	KitchenQual_Gd	-0.038875	-0.033993
148	BsmtQual_TA	-0.034555	-0.034284
113	Exterior1st_ImStucc	-0.004807	-0.034557
147	BsmtQual_Gd	-0.036428	-0.036901

2nd iteration of lasso

	index	Ridge	Lasso
8	TotalBsmtSF	0.053124	0.440665
89	HouseStyle_2.5Fin	0.027257	0.193357
60	Neighborhood_NoRidge	0.081230	0.113306
16	TotRmsAbvGrd	0.068079	0.082829
67	Neighborhood_StoneBr	0.049704	0.082562
...	...	...	...
84	BldgType_Duplex	-0.018157	-0.040755
141	BsmtQual_Fa	-0.035368	-0.040992
173	KitchenQual_Gd	-0.047542	-0.044461
143	BsmtQual_TA	-0.042744	-0.046311
142	BsmtQual_Gd	-0.043201	-0.046416