Advanced Regression Assignment – Surprise Housing

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?

Ridge Regression:

Is a method of estimating coefficients where independent variables are highly correlated. But does not make the coefficients zero, so feature selection is not direct.

Lasso Regression:

Is a method of estimating coefficients with regularization and feature selection, i.e., capability to make coefficients as zero

Optimal Values for Alpha	Ridge	Lasso
	0.1	0.0001

Changes:

We would have to rebuild model with the changed alpha to fit the model. Generate new r2 scores for train and test to evaluate.

If we choose to double the values for alpha for,

Ridge: then this will lower the coefficient values

Lasso: will make more coefficients to turn into 0 in case of Lasso.

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?

I would choose Lasso.

Because:

a. In my case, the r2 score for Train vs Test in Lasso is better than Ridge.

b. In ridge, coefficients of the linear transformation are normally distributed and in Lasso they are Laplace distributed.
c. Which gives the ability for some of the features to have zero coefficients.
d. Hence gives the ability to eliminate some features in Lasso, which wouldn't be the case with Ridge.

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?

For the case here for Lasso, the determined top 5 significant variables are the ones with high coefficient values which are as follows,

1. **OverallQual_Excellent:** 0.583255
2. **Neighborhood_NoRidge:** 0.406721
3. **Neighborhood_NridgHt:** 0.404969
4. **Neighborhood_Crawfor:** 0.367017
5. **Neighborhood_StoneBr:** 0.355975

If we drop these columns before building the model and re-generate, then the following are the top 5 most significant ones basing on high coefficients,

1. **OverallCond_Excellent:** 0.251107
2. **OverallQual_Very Good:** 0.208038
3. **MSSubClass_2-1/2 STORY ALL AGES:** 0.191887
4. **2ndFlrSF:** 0.184549
5. **Exterior2nd_MetalSd:** 0.172985

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?

To make sure the robustness and generalizability of Ridge and Lasso Regression models:

1. Cross-Validation: Using some cross validation like GridSearchCV to evaluate performance on various data subsets, to ensure the model's ability to generalize.

2. Hyperparameters: Tuning the regularization parameter (λ) through cross-validation, with a balance between bias and variance for a robust model.

3. Feature Scaling: Normalizing and then standardizing (scaling) to ensure consistency in scale, to prevent regularization sensitivity.

4. Feature Selection: In case of Lasso, automatic feature selection helps. In case of Ridge less important ones get lower coefficients.

5. Evaluate Metrics: Looking through metrics like RMSE, R-squared, to consider the bias-variance trade-off.

Implications for Accuracy:

1. Accuracy on the training set may decrease because of regularization, but it will improve generalization.

2. Regularization mitigates overfitting, which enhances the model's capability to make accurate predictions on new, unseen data.

3. Optimal hyperparameter, with the help of cross-validation, tuning and careful consideration of features can provide a more robust and model.

[End of Document]