Advanced Linear Regression

Assignment Part II - Subjective Question

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?

Optimal value of alpha for ridge is **10** and lasso is **0.001.**

Changes in Ridge Regression metrics:

- R2 score of train set remained same at 0.91
- R2 score of test set remained same at 0.91

Changes in Lasso metrics:

- R2 score of train set decreased from 0.91 to 0.89
- R2 score of test set remain same at 0.90

Most Important Predictor after the change for Ridge Regression and Lasso regression after the change are shown in the below table:

Ridge Regression	Lasso Regression
GrLivArea	GrLivArea
Functional_Typ	Functional_Typ
OverallQual_Very Good	OverallQual_Very Good
CentralAir_Y	CentralAir_Y
TotalBsmtSF	TotalBsmtSF
Neighborhood_Somerst	YearRemodAdd
Condition1_Norm	Condition1_Norm
OverallCond_Good	BsmtFinSF1
MSSubClass_1-STORY PUD (Planned Unit Developme...	Fireplaces
Exterior2nd_Wd Sdng	OverallCond_Good

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?

- The model we will choose to apply will depend on the use case.
- If our primary goal is to select features then we will use **Lasso Regression.**
- If we don't want to get too large coefficients and reduction of coefficient magnitude is one of our prime goals, then we will use **Ridge Regression**.

Question 3

After building the model, you realized 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?

Top 5 Model features before and after change

Old_top_5_features	**New_top_5_features**
GrLivArea',	2ndFlrSF
Functional_Typ',	1stFlrSF
OverallQual_Very Good',	TotalBsmtSF
CentralAir_Y',	OverallCond_Good
Condition1_Norm'	Neighborhood_Somerst

Model Performance after and before change

	old_model	**new_model**
R-Squared (Train)	**0.9**	**0.89**
R-Squared (Test)	0.91	0.91
RSS (Train)	**13.85**	**16.04**
RSS (Test)	**3.55**	**3.57**
MSE (Train)	0.01	0.01
MSE (Test)	0.01	0.01
RMSE (Train)	**0.11**	**0.12**
RMSE (Test)	0.11	0.11

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?

Here are some strategies to achieve model robustness and generalizability:

Cross-Validation: Use techniques like k-fold cross-validation to assess the model's performance on multiple subsets of the data. This helps in obtaining a more stable estimate of the model's accuracy and ensures that the model is not overfitting to the specific training data.

Train-Test Split: Split the dataset into separate training and test sets. Train the model on the training set and evaluate its performance on the unseen test set. This helps to assess how well the model generalizes to new, unseen data.

Regularization: Apply regularization techniques like L1 (Lasso) or L2 (Ridge) regularization to avoid overfitting and promote simpler models. Regularization helps in controlling the complexity of the model and prevents it from memorizing noise in the training data.

Feature Engineering and Selection: Carefully choose relevant features and perform feature engineering to extract meaningful information from the data. Removing irrelevant or highly correlated features can help the model focus on important patterns.

Data Preprocessing: Properly preprocess the data by handling missing values, scaling features, and encoding categorical variables. Consistent and appropriate data preprocessing ensures the model is less sensitive to variations in the data.

Hyperparameter Tuning: Optimize hyperparameters of the model using techniques like grid search or random search. Proper tuning helps in finding the best configuration that maximizes the model's performance on the test set.

Implications for Model Accuracy:

Robust and generalizable models tend to have more stable and reliable performance across different datasets.
A robust model is less likely to be affected by small changes in the training data and can make consistent predictions on unseen data.
Generalizable models have the ability to perform well on new, previously unseen data points, which is essential for real-world deployment.
By improving robustness and generalizability, the model's accuracy on the test set is more representative of its performance on new data, reducing the risk of overfitting.