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?

Kalyan's Answer:

Ridge Model:

After performing the cleansing and running the ridge regression model on the given data set, the optimal value of alpha achieved is 8.0 and this alpha value resulted into the ridge model with below values across the important metrics of Modelling.

R2 score for Training data: 0.85

R2 score for Test Data: 0.84

RMSE Training Data: 0.39

RMSE Test Data: 0.40

As per the ridge model built, the most important predictor variable of the model:

Neighborhood_NoRidge: 0.52

With alpha doubled and the ridge model is once again evaluated, the new metrics look like below:

R2 score on training data: 0.84

R2 score on test data: 0.84

RMSE on training data: 0.39

RMSE on test data: 0.41

Lasso Model:

After processing data set and performing the data cleaning activitiy, the optimal value of Alpha for Lasso is 0.001 and this outputted the lasso model with the below parameters as outcomes

R2 score on training data: 0.85

R2 score on test data: 0.84

RMSE on training data: 0.38

RMSE on test data: 0.40

The LASSO model that was generated during the assessment, the top predictor variables in the initial Lasso model:

Neighborhood_NoRidge: 0.67

The optimal value of alpha and the LASSO is re-evaluated, it is observed that the metrics are as per below

R2 score on training data: 0.85

R2 score on test data 0.83

RMSE on train data: 0.39

RMSE on test data: 0.41

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?

Kalyan's Answer:

If we compare the LASSO and RIDGE regression model, the R2 score for the training set is a little bit higher for the Lasso model compared to the Ridge.

The R2 score represents the proportion of the variance in the dependent variable that is predictable from the independent variables.

Lasso eliminates the non-significant features.

Both RIDGE and LASSO models have determined the optimal value of lambda, we can choose the LASSO model as it provides a little better fit to the data as per the R2 score and RMSE values.

The lasso model is simpler without much of the complexity.

Key metric comparison of both the models:

	R2(Training)	**RMSE(Train)**
RIDGE	0.850975851	0.386036461
LASSO	0.852471965	0.384093784

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?

Kalyan's Answer:

The Five most important predictor fields in the original Lasso model are:

Neighborhood_NoRidge: 0.6690544166467738

Neighborhood_NridgHt: 0.6105164809085014

2ndFlrSF: 0.37353198589266806

BldgType_Twnhs: -0.3339088014999202

Neighborhood_Somerst: 0.2953973727254121

And I have re-built the lasso model after dropping the above mentioned 5 top variables.

Top five predictor variables among the rest of the fields are:

OverallQual: 0.353154203253468

1stFlrSF: 0.1730974471335822

GarageArea: 0.08961515649491918

KitchenQual: 0.08116533812372087

GarageFinish: 0.026472489622987802

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?

Kalyan's Answer:

In order to build a robust and generalizable model, below points tend to be very important.

1) **Regularization:**
   To generalize the model, we need to prevent overfitting of the model. Regularization techniques (Ridge Regression / Lasso Regression) must be applied to prevent overfitting by penalizing large coefficients.

2) **Cross-Validation:**
   Cross-validation techniques (k-fold cross-validation etc.,) that evaluates the model's performance and robustness by utilizing the training and testing the model on different subsets of the given dataset.

3) **Performance Metrics:**
   Performance Metrics such as MSE(Mean Squared Error), RMSE(Root Mean Squared Error), MAE(Mean Absolute Error), R-squared to evaluate the model's performance on both the training and testing datasets to make them more generalized for broader usage.

4) **Handle Multicollinearity:**
   Identification of the multicollinearity is a critical step. If found, it has to be handled among the predictor variables. Multicollinearity can lead to unstable predictions or estimates of regression coefficients and affect the model's generalizability. Techniques like Variance Inflation Factor (VIF) are to be used to identify the multicollinearity in the data set.

5) **Data Cleansing, Pre-processing and Feature Engineering:**

Proper data cleansing mechanism is thoroughly needed before building any model. Because there could be null values, zero values, categorical values, non-numerical values, un-scaled values etc., all the different sorts of data need to be considered for cleansing before building the model. To handle the missing values, the common methods include mean/median imputation or removing non-significant columns. Drop the highly correlated data points as they are a sort of duplicates of one another. Dropping them would simplify the modelling.

Feature importance scores can do an excellent job in identifying the most impactful fields.
Transformation: transforming of the features (e.g., log transformation) can improvise the linearity or normality of residuals.
RFE (Recursive feature elimination) identifies the most important features that contributes to the accurate modelling.
Dropping the irrelevant fields may help in simplifying the model, reducing the overfitting of the model.

6) **Residual Analysis:**

In order to check if the assumptions of linear regression are met, we need to analyse the residuals. There should not be much deviation between the predictions and the residuals. Residual plots can help identify the patterns on the residuals.

Impact on the Accuracy of the modelling:

If we make the model more generalize and bring in the robustness, it makes less errors on the training data and lessens the biasing. Accuracy optimization is an equation of bias and variance. It is a trade-off between bias and variance to optimize the accuracy.

Prevents overfitting by making the model less sensitive to specific the given data points. A generalized model performs well on new and unknown data thereby making it more accurate predictions.

Robustness and generalizability make the model more reliable and trustworthy to make the future predictions on the test data and unknown data. Because the right relation is captured and not memorizing the every single data point that leads to overfitting of the data.