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?

Answer:

The Optimal Value of Alpha with Ridge & lasso are:

Ridge = 9.0

Lasso = .0001

	Ridge		**Lasso**	
Alpha	**9.0**	**18.0**	**0.0001**	**0.0002**
R2 score (train)	0.9154	0.9152	0.9154	0.9154
R2 score (test)	0.8775	0.8778	0.8771	0.8773
RMSE (train)	0.1139	0.114	0.1139	0.1139
RMSE (test)	0.1497	0.1495	0.1499	0.1498

I have got very similar values for R2 Score after doubling the value of alpha for both ridge & lasso. In fact, values for R2 Score and RMSE in case of lasso are almost similar values.

Most Important predictor variables after the change implemented are:

Ridge:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. SaleCondition_Partial

Lasso:

- 1stFlrSF
- 2ndFlrSF
- OverallQual
- OverallCond
- MSZoning_RL

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?

Answer:

Although I have got similar nearly similar values for both Ridge and Lasso for the alpha values 9.0 and 0.0001 respectively however Lasso Model have zeroed one or two coefficients in the selected features. So Lasso is better option as it helps in eliminating few more features. So I would choose Lasso.

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?

Answer:

Below are five most important predictor after excluding initial five important predictors.

Lasso	
FullBath	0.080596
GarageArea	0.059863
Fireplaces	0.055031
LotArea	0.053244
MSZoning_RL	0.052637

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?

Answer:

The mode is robust and generalizable when

1. The R square value has no significant difference when fit model one both Train & Test Set of data
2. Outlier treatment is most important to get the robust model. We can detect outliers in the dataset using percentile (25,50,75), data visualization using Scatter Plot, box plots, Z score etc. Treating
3. Treating missing values in a right manner so that we can impute correct values to missing values.
4. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis
5. Model significance can be determined the P-values, R2 and adjusted R2.

Implications of Accuracy of a model:

1. Gain the more data as much you can: Having more data allows the data to train itself, instead of depending on the weak correlations and assumption, it is good to have more data.

2. Fix missing values and outliers: If the data has missing values and outliers can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to continuous variables

3. You can get the outlier values using a boxplot, treating the outliers in the data will make our mode more accurate.

4. Featuring Engineering or newly derived columns/Standardize the values: We can extract the new data from the existing data ex: from DOB we can get the Age of the person, after extracting the new data required we can drop the existing features.

5. Scaling the values: ex: one value is in meters, the other is Kilo meters, it is important to scale these feature into one standardized unit.

6. Data visualization also helps the selecting the features.

7. Statistical parameters like p-Values, VIF can give us significant variables.
8. Choosing the right machine learning algorithm is very important to get accurate model. This will comes with experience

9. Cross validation: Some times more accuracy will cause overfitting, then we can use cross validation technique, i.e. leave a sample on which you do not train the model & test the model on this sample before got to the final model.