Assignment Part -II - Subjective Questions and Answers

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 alpha:

1. Ridge = 20
2. LASSO = 0.001

Changes caused by doubling alpha:

As summarized in the table below:

1. Ridge: R^2 for both train and test set reduces and MSE increases.
2. LASSO: R^2 for both train and test set reduces and MSE increases.

Since greater alpha penalizes the coefficients more and shrinks them to zero we also get less number features with non zero coefficients for LASSO. So the number of features for LASSO changes from 20 to 19 by doubling its alpha, but it wont change for Ridge.

Table1:

Metric	Linear Regression	Ridge Regression	Ridge Regression (Double Alpha)	Lasso Regression	Lasso Regression (Double Alpha)
R2 Score (Train)	9.356429e-01	0.895571	0.892368	0.902790	0.896901
R2 Score (Test)	-1.713781e+23	0.864620	0.863770	0.866751	0.865969
RSS (Train)	2.443430e+01	39.648201	40.864201	830927.603163	825680.897814
RSS (Test)	1.581777e+25	12.495225	12.573719	52668.274874	52331.380868
MSE (Train)	1.451972e-01	0.184957	0.187772	0.178450	0.183776
MSE (Test)	2.335466e+11	0.207574	0.208225	0.205934	0.206537

10 Most important features after doubling the alpha:

1. Ridge:

   There are 227 features with non zero coefficient for Ridge

   ```
   ['GrLivArea', 'OverallQual', 'Neighborhood_Crawfor',
   'SaleType_New', 'Age', 'Exterior1st_BrkFace',
   'Condition1_Norm', 'OverallCond',
   'SaleCondition_Normal', 'Neighborhood_StoneBr']
   ```

After RFE feature selection we will have 20 features with non zero coefficient for Ridge that the 10 most important ones are as follow:

```
['GrLivArea', 'OverallQual', 'Age',
'Neighborhood_Crawfor', 'LotArea',
'Neighborhood_NridgHt', 'OverallCond','SaleType_New',
'Exterior1st_BrkFace', 'BsmtFinSF1']
```

2. LASSO:
   There are 67 non zero coefficients for LASSO when all the features are being used

```
['GrLivArea', 'OverallQual', 'Age', '1stFlrSF',
'Neighborhood_Crawfor', 'Condition1_Norm',
'OverallCond', 'SaleCondition_Normal',
'Exterior1st_BrkFace', 'Neighborhood_StoneBr']
```

   After RFE selection of 20 features LASSO show 20 features as being important (nonzero coefficient) 10 of them are as follow:

```
['GrLivArea', 'OverallQual', 'SaleType_New',
'Neighborhood_Crawfor', 'Age',
'Neighborhood_NridgHt', 'OverallCond',
'Neighborhood_StoneBr', 'LotArea',
'Foundation_PConc']
```

   Due to lack of space I just mentioned 10 most significant features here but the rest are shown at the end of the jupyter notebook.

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?

1. I would use the same alpha that detected as optimal alpha determined by grid search for each model:
   a. Ridge=20
   b. LASSO=0.001

2. LASSO gives us lower test MSE and higher R^2 for both test and train set as mentioned in Table1. In comparison with Ridge since LASSO also gives us a less variance between R^2 train and test. so it's less prone to overfitting and regarding LASSOs' ability to reduce the coefficients of unimportant features to zero and selecting important features, LASSO seems a better model for our data.

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?

```
['1stFlrSF', 'OverallQual', 'SaleCondition_Partial', '2ndFlrSF',
'Neighborhood_NridgHt']
```

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?

In order to make sure that a model is generalizable and robust we need to check the model performance on unseen data. In nutshell it is managed through balancing Bias and variance tradeoff. In this context we decrease the variance and chance of overfitting at the cost of increasing bias (decrease accuracy) by reducing model complexity. It can be managed and controlled at different stages of the model training and development. Some of them are as follows:

1. Check diversity and quality of data:
    a. Clean and validate data: remove errors, outliers, missing values and biases
    b. Make sure the data covers a wide range of scenarios and context, and make sure that it reflects the reality and complexity of the problem we are trying to solve.

    c. If the data is skewed and noisy models are not able to generalize well and may face overfitting or underfitting problems.

2. Extract proper data, features and engineer or create proper features
    a. Quality and quantity of the features have a significant impact on our model performance and interpretability. So we need to extract relevant and informative features that properly capture useful patterns and relationships in data for modeling, using techniques such as; Transformation, aggregation, encoding, scaling, normalization, dimensionality reduction (feature extraction or feature selection), feature engineering.

3. Model Evaluation and model selection
    a. There are a variety of models that can be used to solve a problem. Each model has its own application, requirements, and advantages and disadvantages. We need to select the best model based on the nature and complexity of our data. We also need to select the proper evaluation metrics based on the task in hand the selected model, and evaluate, validate or compare the model performance.
    b. Check for possible overfitting or underfitting using train-test split and cross-validation.
4. Hyperparameters tuning and optimization:
    a. Improve model robustness and generalizability through hyperparameters tuning and finding optimal parameters that maximize model performance and minimize its error. Techniques such as; grid search, random search, bayesian optimization, gradient descent, and genetic algorithms are usually being used at this stage.
    b. There are also regularization techniques such as; lasso, ridge, elastic net, dropout, batch normalization that are being used to prevent overfitting by reducing model variance and complexity.
5. Testing Model on unseen data
    a. Evaluate model performance on unseen data and real world data.
    b. Update and retrain the model regularly as new data nad information become available.

6. Model interpretation:

a. Understand how your model work, what it is doing and why it's doing that
b. Explain model prediction and recommendations
c. Identify factors or features that influence the model
d. Define uncertainties and limitations using techniques such as: Feature importance, partial dependence plot, shap values, lime and counterfactuals.