Assignment Part-2 – Subjective Questions

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 2:
As per the model built, the optimal values of alpha for **Ridge** is 5 and for **Lasso** it is 0.0001. These values of alpha helped in getting the R2 value of 89% and 88% respectively. Below is the metrics with these values:

Model Evaluation of Ridge Regression

- **R2:** 0.89
- **MSE:** 0.0019
- **Alpha:** 5

Model Evaluation of Lasso Regression

- **R2:** 0.88
- **MSE:** 0.0020
- **Alpha:** 0.0001

After doubling the values of alpha for Ridge model, the value of R2 has been decreased from 89% to 88% and MSE has been increased from 0.0019 to 0.0020. However, there is no significant change in R2 or mean squared error in case of Lasso model. There is slight change in the coefficient values which can be seen in the below screenshots:

Ridge Regression:

Original alpha (5)

```
Number of non-zero Coefficients 260
MSE Train 0.0011566161075053545
MAE Score Train 0.0232763559954822
R2 Score Train 0.9304397325174814

MSE Test 0.001964354041660126
MAE Score Test 0.029562015344246562
R2 Score Test 0.8873594720960091
```

	Ridge Coefficient
BsmtFinSF1	0.072452
BsmtUnfSF	0.069371
GrLivArea	0.062249
2ndFlrSF	0.058334
ExcellentQual	0.053601
TotRmsAbvGrd	0.051400
FullBath	0.047819
1stFlrSF	0.047108
GarageCars	0.043754
VeryGoodQual	0.039013

Double alpha (10)

```
Number of non-zero Coefficients 260
MSE Train 0.0013086639140292627
MAE Score Train 0.02469606423236621
R2 Score Train 0.9212953967060554

MSE Test 0.0020243583774635517
MAE Score Test 0.029790973769653107
R2 Score Test 0.8839186870246407
```

	Ridge Coefficient
BsmtFinSF1	0.058998
GrLivArea	0.052142
BsmtUnfSF	0.051138
TotRmsAbvGrd	0.050016
ExcellentQual	0.046719
2ndFlrSF	0.046519
FullBath	0.046368
1stFlrSF	0.040488
GarageCars	0.038713
VeryGoodQual	0.037697

Lasso Regression:

Original alpha (0.0001)

```
Number of non-zero Coefficients 134
MSE Train 0.000977229515953313
MAE Score Train 0.02188461149970975
R2 Score Train 0.9412282553559285

MSE Test 0.002088397487502077
MAE Score Test 0.028892938444301908
R2 Score Test 0.8802465388231153
```

Double alpha (0.0002)

```
Number of non-zero Coefficients 109
MSE Train 0.0012541336542627381
MAE Score Train 0.02400517501639994
R2 Score Train 0.9245749113441768

MSE Test 0.0020447338689162987
MAE Score Test 0.028903712079136373
R2 Score Test 0.8827503100086518
```

	Lasso Co-Efficient
GrLivArea	0.207470
BsmtUnfSF	0.189810
BsmtFinSF1	0.160746
BsmtFinSF2	0.121384
2ndFlrSF	0.064008
ExcellentQual	0.063227
GarageCars	0.058393
RH	0.051358
RL	0.049766
FV	0.049274

	Lasso Coefficient
GrLivArea	0.163125
BsmtUnfSF	0.113558
BsmtFinSF1	0.111940
BsmtFinSF2	0.068917
ExcellentQual	0.067283
2ndFlrSF	0.065215
GarageCars	0.062589
TotRmsAbvGrd	0.044685
FullBath	0.042462
VeryGoodQual	0.041809

As we can see in the above tables, there is slight change in the predicted features or the order has been changed due to the change in coefficient values. The change is not significant because the change in alpha is very small.

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 2:
Let's look at the metrics below to determine which one is important:

Model Evaluation of Ridge Regression

- **R2:** 0.89
- **MSE:** 0.0019
- **Alpha:** 5

Model Evaluation of Lasso Regression

- **R2:** 0.88
- **MSE:** 0.0020
- **Alpha:** 0.0001

Now, while building the model, we try to find the best fit model which is measured by R2 and MSE values. Combination of maximum R2 and minimum MSE is always the best fit model. In our case, there is no slight change in MSE but 1% difference is there in R2.

Ridge model is giving more R2 value and less MSE value. Hence, we would go with Ridge model.

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 3:
As per the current Lasso model, top 5 predictors are as follows:

1. GrLivArea
2. BsmtUnfSF
3. BsmtFinSF1
4. BsmtFinSF2
5. 2ndFlrSF

After dropping these features from the dataset, R2 value has been decreased from 0.88 to 0.87 and MSE value has been increased from 0.0020 to 0.0021. This is not a significant change in the measures, so we can say model is still performing well.

New 5 most important features

The new five most important predictor variables are:

	Lasso Coefficient
1stFlrSF	0.202955
TotRmsAbvGrd	0.085318
FullBath	0.072669
ExcellentQual	0.072404
GarageCars	0.068315

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 4:
Before knowing how to make model robust and generalize, we will understand what these terms means.

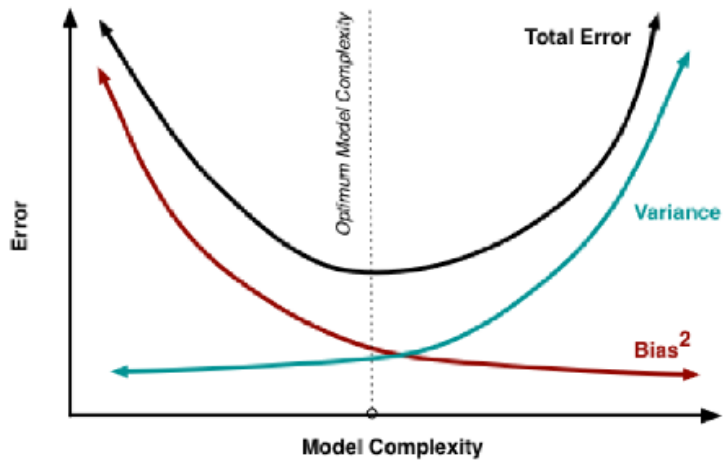
We can call a model robust if the output or predictions are accurate even if there is one or more of the input variables or assumptions are drastically changed due to unforeseen circumstances. Robustness evaluation estimates potential failure probabilities when the model is pushed to its limits.

Model is as generalizable model if the performance and adaptability of a model when applied to new conditions while maintaining the same basic set of explanatory variables. In another words, its performance on unseen test scenarios.

The model should be generalized so that the test accuracy is almost equal to the training score. The model should be accurate for datasets other than the ones which were used during training. The outliers analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be removed from the dataset. **If the model is not robust, it cannot be trusted for predictive analysis.**

How to make model robust and generalizable?

- **Occam's Razor:** A predictive model has to be as simple as possible, but no simpler. Often referred to as the Occam's Razor, this is not just a convenience but a fundamental tenet of all of machine learning.
    - A simpler model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen data sets.
    - A simpler model requires fewer training data points. This becomes extremely important because in many cases, one has to work with limited data points.
    - A simple model is more robust and does not change significantly if the training data points undergo small changes.
- **Bias-Variance Tradeoff:** Consider the bias-variance tradeoff while building the simple models. Simple models tend to lead bias-variance tradeoff issue. A model should not be high bias or high variance. We should find a point where both bias and variance is minimum. Such model will be more robust and generalized.



- **Use a model that's resistant to outliers:** Regression-based models serves such purpose. If you are performing a statistical test, try a non-parametric test instead of a parametric one.
- **Use a robust error metric:** Switching from mean squared error to mean absolute difference reduces the influence of outliers
- **Transform your data:** If your data has a very pronounced right tail, try a log transformation.