SURPRISE HOUSING ASSIGNMENT

By: Hamza Tanveer

Assignment-based Subjective Questions

1. **What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will the most important predictor variables after the change is implemented?**

   The optimal value of alpha are:
   For Ridge Regression alpha is 0.9
   For Lasso Regression alpha is 0.0001

   If we double the alpha, then shrinking penalty will be higher, so the Ridge and Lasso will try to shrink the values of beta coefficients towards zero. That means it will increase the bias and variance will decrease.
   The following are the changes observed:
   RIDGE REGRESSION:

   - R2 score decreases slightly
   - More coefficients are pushed towards zero
   - Few of the predictor variables have changed the order of importance

   LASSO REGRESSION:

   - More coefficients become zero
   - Most of the important predictor variables retain the order of importance

   The most important predictor variables after the change is implemented are:

	Ridge
GrLivArea	0.350279
OverallCond	0.140497
GarageCars	0.109054
TotalBsmtSF	0.107470
ExterQual	0.103897
MSZoning_FV	0.076972
BsmtFinSF1	0.061910
Neighborhood_Crawfor	0.058247
LotArea	0.052375
Neighborhood_StoneBr	0.049236

	Lasso
GrLivArea	0.390559
OverallCond	0.149981
TotalBsmtSF	0.108725
GarageCars	0.102732
ExterQual	0.094674
BsmtFinSF1	0.060475
Neighborhood_Crawfor	0.053714
MSZoning_FV	0.044183
LotArea	0.035827
Neighborhood_StoneBr	0.034640

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?**
   As per Occam's razor a model should not be unnecessary complex.
   Model complexity depends on two main things:
   - Number of features or independent variables and
   - Magnitude of beta coefficients

   Normalization (Ridge and Lasso) already shrinks beta coefficients towards zero.

   Ridge and Lasso Regression have similar R2-scores and RMSE on test data. We choose Lasso as it does feature selection it gives similar R2-scores with lesser beta-coefficients.

```
Ridge Regression with 25 features and alpha =  0.9
=====================================================
R2 score (train) :  0.8973718121708262
R2 score (test) :  0.8708851708375369
RMSE (train) :  0.04358007763300673
RMSE (test) :  0.04769675988038313
```

```
Lasso Regression with 24 features and aplha 0.0001
=====================================================
R2 score (train) :  0.8957310720044349
R2 score (test) :  0.8708247308311595
RMSE (train) :  0.04392705861947114
RMSE (test) :  0.047707922251129534
```

   As both the models show similar performance on test dataset, I choose simpler model i.e LASSO MODEL

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?**
   Initially top 5 features in LASSO model are as below:

	Lasso
GrLivArea	0.391043
OverallCond	0.153862
TotalBsmtSF	0.107212
GarageCars	0.101704
ExterQual	0.092304

   After excluding the five most important predictor variables. Now 5 most important predictor variables are as follows:

LotArea	0.252480
LotFrontage	0.238960
Neighborhood_StoneBr	0.133847
MSZoning_FV	0.128681
Neighborhood_Crawfor	0.118843

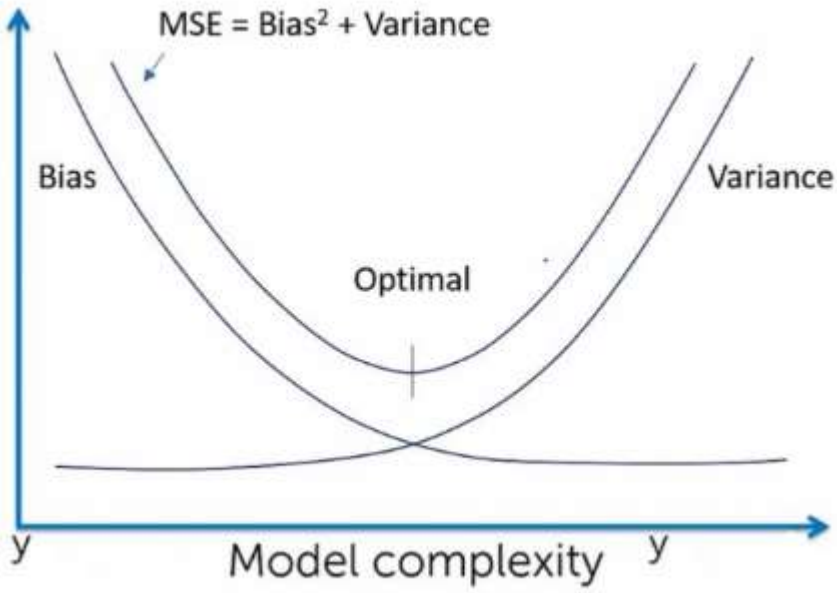
4. **How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?**

   A model should be complex enough so that it learns the data patterns in the training dataset but not too complex that it also learns noises in the training dataset. The model should be generalized enough and not so complex that it memorizes all data points in the training dataset.

   An under fitting model usually has high bias and low variance. It fails to understand data pattern in training dataset, so it performs bad both on training and testing dataset. Whereas an over fitting model usually show bias and high variance. It performs well on training dataset but performs badly on testing dataset or unseen data.

   A scenario of over fitting can be identified easily by comparing model performance in training and testing dataset. If there is a significant difference in model performance (r2 score, model accuracy, RMSE etc. other evaluation metrics) on training and testing dataset then it's a case of over fitting.

   A robust model should have low bias and low variance and it should not suffer from under fitting and over fitting. It can be achieved by doing a trade-off between bias and variance. One of the ways to remove over fitting to create a robust and generalizable model is to reduce model complexity



   Model complexity depends on two main things:

   - Number of features or independent variables and
   - Magnitude of beta coefficients

   Normalization (Ridge and Lasso) already shrinks beta coefficients towards zero.

   Accuracy of a robust and generalizable model should be almost same/closer on training and testing datasets.