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?

Optimal value of alpha for Ridge Regression is 2.0
Optimal value of alpha for Lasso Regression is 0.0001
Created models by doubling the alpha values for both ridge and lasso, $R2$ score on test data remains almost same when compared with optimal values i.e; 0.89, But there is a slight change in the coefficients and important predictor variables. Here are some of the important predictors and values of coefficients for reference

	Ridge Co-Efficient
GrLivArea	0.1304
1stFlrSF	0.1089
GarageArea	0.0694
MSZoning_FV	0.0666
OverallQual_Ex	0.0626
TotalPorchArea	0.0514
TotalBsmtSF	0.0503
MSZoning_RL	0.0498
OverallCond_Ex	0.0492
SaleCondition_Alloca	0.0451
Neighborhood_Crawfor	0.0445
Neighborhood_StoneBr	0.0443
OverallQual_VeryGd	0.0415
LotArea	0.0404
FullBath	0.0396
Exterior1st_BrkFace	0.0391
BsmtFinSF1	0.0387
MSZoning_RM	0.0386
BedroomAbvGr	0.0385

	Ridge Double Co-Efficient
GrLivArea	0.1038
1stFlrSF	0.0970
GarageArea	0.0639
OverallQual_Ex	0.0568
TotalBsmtSF	0.0527
MSZoning_FV	0.0458
TotalPorchArea	0.0456
FullBath	0.0439
OverallCond_Ex	0.0429
Neighborhood_Crawfor	0.0418
BsmtFinSF1	0.0400
BedroomAbvGr	0.0387
LotArea	0.0387
OverallQual_VeryGd	0.0382
Neighborhood_StoneBr	0.0382
SaleCondition_Alloca	0.0353
SaleCondition_Normal	0.0349
Exterior1st_BrkFace	0.0347
MSZoning_RL	0.0330

	Lasso Co-Efficient
GrLivArea	0.3239
OverallQual_Ex	0.0793
GarageArea	0.0717
MSZoning_FV	0.0603
TotalBsmtSF	0.0568
OverallCond_Ex	0.0496
1stFlrSF	0.0493
Neighborhood_Crawfor	0.0486
OverallQual_VeryGd	0.0481
TotalPorchArea	0.0471
MSZoning_RL	0.0433
Neighborhood_StoneBr	0.0356
SaleCondition_Alloca	0.0355
SaleCondition_Normal	0.0345
SaleType_New	0.0342
LotArea	0.0338
Neighborhood_NridgHt	0.0336
Exterior1st_BrkFace	0.0332
BsmtFullBath	0.0308
OverallQual_VeryEx	0.0286

	Lasso Double Co-Efficient
GrLivArea	0.3434
GarageArea	0.0732
OverallQual_Ex	0.0711
TotalBsmtSF	0.0616
Neighborhood_Crawfor	0.0473
OverallQual_VeryGd	0.0391
TotalPorchArea	0.0383
1stFlrSF	0.0375
SaleType_New	0.0362
OverallCond_Ex	0.0355
SaleCondition_Normal	0.0328
Neighborhood_NridgHt	0.0316
MSZoning_FV	0.0295
Exterior1st_BrkFace	0.0282
BsmtFullBath	0.0270
BsmtExposure_Gd	0.0268
BsmtCond_TA	0.0261
BsmtFinSF1	0.0254
Neighborhood_StoneBr	0.0242
BsmtCond_Gd	0.0233

As the overall alpha value is small there are no much changes in the model even after doubling the alpha values. (For full coefficients and model values check the python 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?

Ridge and Lasso shrinks the coefficients depending on the hyper parameter lambda. In this process Lasso shrinks less important feature coefficients to 0 thus, removing features all together which leads to Feature Selection.
In this case, Optimal value of lambda for Ridge Regression is 2.0 and for Lasso Regression is 0.0001
Both Ridge and Lasso performs almost similar when compared with R2scores, RMSE.

	Lasso	Ridge
R2Train	0.9370	0.9409
R2Test	0.8965	0.8874
RSS Train	1.2894	1.2094

RSS Test	0.8821	0.9590
RMSE Train	0.0356	0.0345
RMSE Test	0.0450	0.0470

Here we consider Lasso Regression model as the final model as this will provide feature selection with good R2score.

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?

The top most predictor values for Lasso model are

- GrLivArea 0.3239
- OverallQual_Ex 0.0793
- GarageArea 0.0717
- MSZoning_FV 0.0603
- TotalBsmtSF 0.0568

After removing these features and build model using lasso, the important predictors are

1stFlrSF	0.2840
OverallCond_Ex	0.0627
LotArea	0.0619
FullBath	0.0531
TotalPorchArea	0.0530
HouseAge	-0.0647
OverallQual_Po	-0.1167

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?

As per Occam's Razor, a predictive model has to be as simple as possible but no simpler. To measure the simplicity, we can say that more complex the model is, less simple it is.
There are different ways to measure the complexity of the model. Few are

1. Number of coefficients
2. Degree of the function
3. Size of the best possible representation of the model. For e.g., precision, large numbers in the coefficients of the model.
4. Depth or size of a decision tree

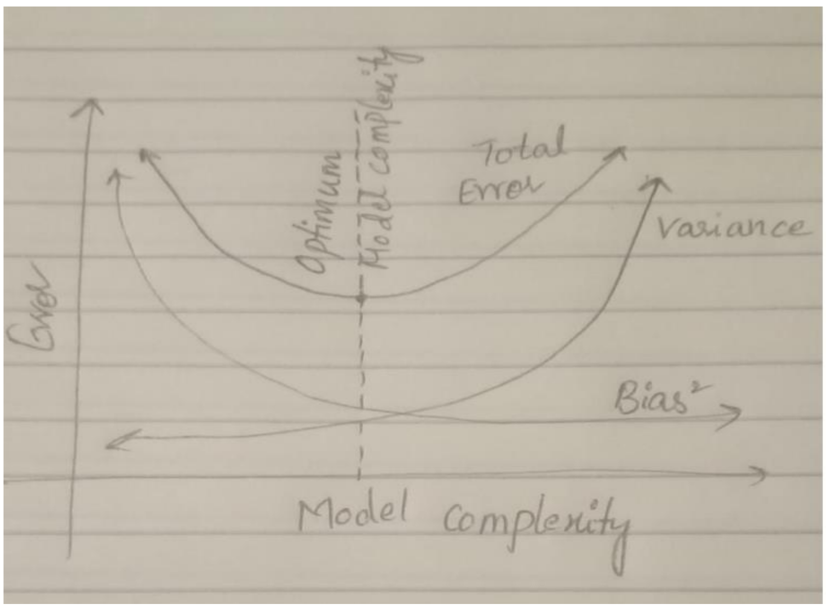
Given two models that show similar performance in the finite training or test data, we should pick the model that makes fewer assumptions about the unseen data due to following reasons

- Simpler models are usually more generic and more widely applicable.
- Simpler models require few training samples than the complex ones.
- Simpler models are more robust
- Simpler models make more errors in the training data set

So, to make the model more robust and generalizable, make the model simple but not simpler.

Now there is tradeoff between variance and bias to select the optimal model.
Bias quantifies how accurate is the model likely to be on test data. The Variance of a model is the variance in its output on some test data with respect to changes in the training data.
A complex model can do accurate predictions if there is enough training data. Models that are too simple or naive and which give same outputs to all the test data and makes no difference has large bias.

Thus, the accuracy of the model can be maintained by keeping the balance between Bias and Variance.



The figure explains the typical trade-off between bias and variance. Simpler model models have high bias and low variance whereas complex models have low bias and high variance.
The best model is the one that balances between both bias and variance without compromising too much on accuracy.

Regularization is the process used in machine learning to deliberately simplify models by achieving the balance between making simple model but not too naïve which is of no use.
This is a simplification done by the training algorithm to control the model complexity.
For Regression this involves adding regularization term to the cost that adds up the absolute values (Lasso) or squares of the parameters (Ridge) of the model.