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 4 and for Lasso is 0.0001.

The following are the metrics of ridge with optimal alpha value 4 and lasso regression with optimal alpha value 0.0001.

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.944215	0.942106
R2 Score (Test)	0.877915	0.878674
RSS (Train)	0.899583	0.933598
RSS (Test)	1.112895	1.105971
MSE (Train)	0.030023	0.030585
MSE (Test)	0.050933	0.050774

The following are the metrics of ridge with optimal alpha value 8 and lasso regression with optimal alpha value 0.0002.

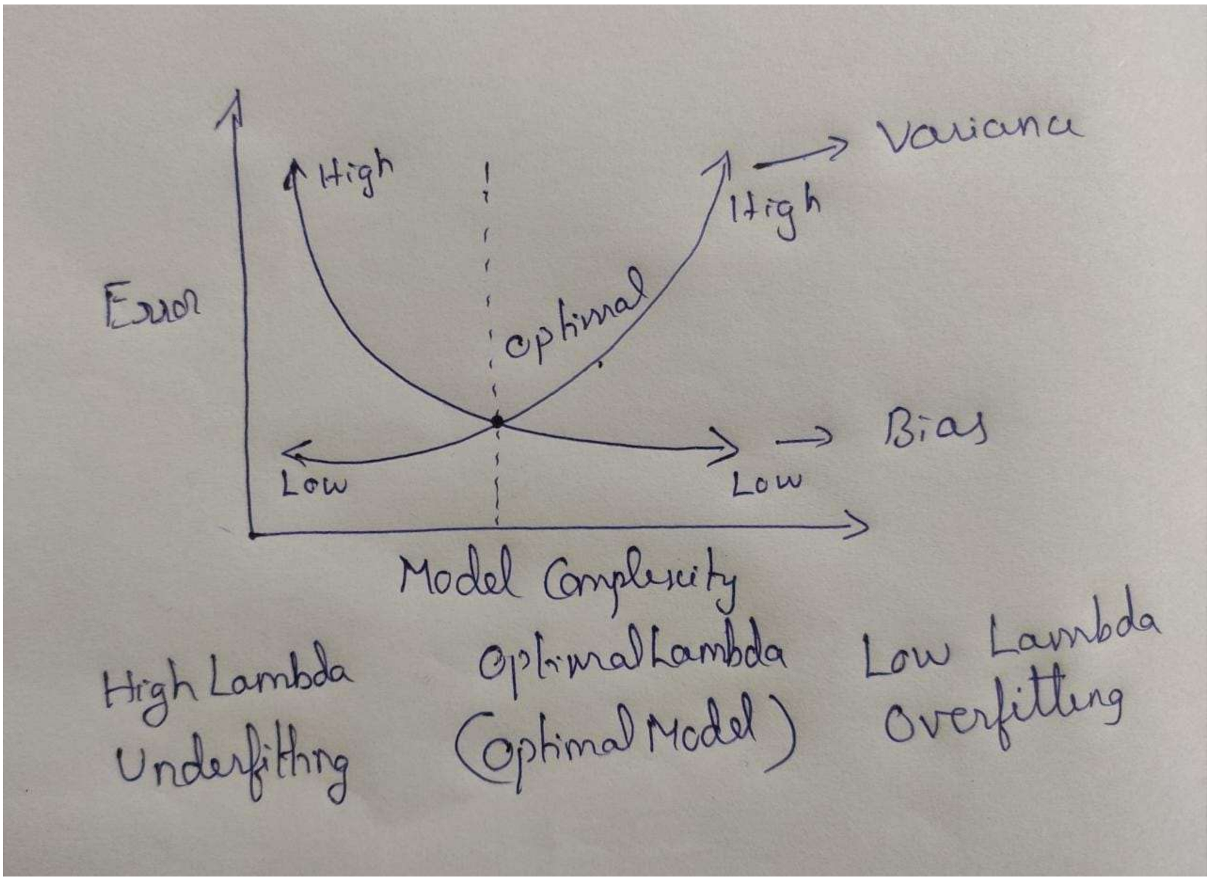
Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.938795	0.936313
R2 Score (Test)	0.868574	0.872212
RSS (Train)	0.986991	1.027011
RSS (Test)	1.198043	1.164877
MSE (Train)	0.031448	0.032079
MSE (Test)	0.052845	0.052109

Regularization helps with managing model complexity by essentially shrinking the model co-efficient estimates towards 0.Regularization helps to avoid overfitting, discourage the model to become too complexity by shrinking the co-efficient value towards 0. The lambda or alpha is the hyper parameter value. The optimal value of alpha we got by tuning, for the model to have low bias and low variance. If alpha is high then the co-efficient parameters of the model shrink towards 0 and avoid overfitting, more towards generalization.

By doubling the alpha value, the RSS and MSE value of both Ridge and Lasso regression also increased by little. Accuracy also reduced by little. R-square value of both Ridge and Lasso regression for training set and testing set is reduced. If we go on increasing the alpha value the model move towards underfitting.

So, it is very important to get the optimal value of alpha, for the model to have low bias and low variance.

Hyper-parameter Tuning:



The important predictor variables which help in predicting the SalePrice of property are as follows:

1. GrLivArea → Above grade (ground) living area square feet
2. OverallQual_Excellent → Rates the overall material and finish of the house -Excellent.
3. AgeOfHouse → Age of the house, we derived this variable by using Year of built and Year of Sold.
4. 'BsmtFinSF1' → Type 1 finished square feet. If the basement area has high rating and large the price of the property will increase.
5. 'TotalBsmtSF' →Total square feet of basement area
6. '1stFlrSF' → 1stFlrSF: First Floor square feet. If the house has large area in first floor the SalePrice will increases.
7. ExterQual_TA →Evaluates the quality of the material on the exterior - Avarage or Typical.
8. GarageArea → Size of garage in square feet. Garage area is positively impacting the SalePrice. If the garage area is large the SalePrice also high.
9. ExterQual_Fa → Evaluates the quality of the material on the exterior – Fair

10. OverallQual_Very Good → Rates the overall material and finish of the house – Good
11. BsmtExposure_Gd →Refers to walkout or garden level walls.
12. Neighborhood_Crawfor → Physical locations within Ames city limits - Crawford
13. ExterQual_Gd → Evaluates the quality of the material on the exterior - Good
14. OverallQual_Very Excellent
15. LotArea → Lot size in square feet

1. The Ground living area, Basement finished area, Total basement area, Area of the first floor, Garage area, Lot Area are the important predictors positively impacting the price of the property.
2. Overall Quality of the house - If the rating of Overall Quality, material and finish of the house is excellent, good, then the price of the property will increases.If it is below average it is negatively impacting the price of the property.
3. Age of the house,The age of the house is negatively impacting the SalePrice.
4. If the quality of the material used on the exterior is Poor , average of typical this will negatively impact the Sale price of the property.
5. If the property has walkout or garden level walls this will increases the price of the property.
6. If the property is near Crawford and NorthRidge then it has high price.

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?

The optimal value of lambda for Lasso Regression is 0.0001 and for Ridge regression is 4.

We will choose the Lasso Regression because the R-square value for Lasso Regression for test set is slightly higher and the error terms RSS and MSE is little low compare to Ridge regression.

1. Both Ridge and Lasso regression gives good result. The Ridge regression gives 87.79% of accuracy for the test set, the Lasso regression gives 87.86% accuracy for test set.
    - Lasso regression shows little high accuracy compare to Ridge regression
2. Both Ridge and Lasso regression satisfies all the assumption of linear regression.i.e,
    - Linearity of the independent variables with the target variable
    - Error terms shows Normal distribution with mean 0
    - Homoscedasticity of error terms
    - Error terms are independent to each other
3. Error terms should be least
    - RSS and MSE of Lasso regression has least value compare to the Ridge regression
4. Optimal alpha value of Ridge regression is 4 and optimal alpha value of Lasso is 0.0001

Lasso Regression shows little better result compare to Ridge Regression. Lasso regression also has the advantage of feature selection.

Ridge regression also has some disadvantages,

1. Ridge regression would include all thepredictors in the final model
2. The accuracy of the prediction will not get affected but model interpretation is challenging when number of predictors is very large.

Advantage of Lasso Regression:

1. Penalty in Lasso forces some of the co-efficient estimates to be exactly equal to 0. So, Lasso performs variable selection. (Feature selection).
2. Model generated from the Lasso are generally easier to interpret than those produced by Ridge regression.

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?

The top 5 predictors of the lasso model are as followes,

1. GrLivArea → Above grade (ground) living area square feet
2. OverallQual_Excellent → Rates the overall material and finish of the house -Excellent.
3. AgeOfHouse → Age of the house, we derived this variable by using Year of built and Year of Sold.
4. 'BsmtFinSF1' → Type 1 finished square feet. If the basement area has high rating and large the price of the property will increase.
5. 'TotalBsmtSF' →Total square feet of basement area

After removing the top 5 predictor of the lasso model, we got the new lasso model with following top 5 predictors.

1. '1stFlrSF' →First Floor square feet. If the house has large area in first floor, the Sale Price will increase.
2. '2ndFlrSF'→Second floor square feet
3. 'ExterQual_TA', 'ExterQual_Fa', 'ExterQual_Gd'→ Evaluates the quality of the material on the exterior. If the quality of the material on the exterior is average or typical then it decreases the price of the property.
4. 'GarageArea'→ Size of garage in square feet. Garage area is positively impacting the SalePrice. If the garage area is large the SalePrice also high.

5. 'BsmtExposure_Gd'→ Refers to walkout or garden level walls. If the property has walkout or garden level walls this will increases the price of the property.

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?

After building the model check the metrics like, R-square value, RSS(Residual Square Error), MSE (MeanSquare Error) for both training and test dataset. The model fits so well on the training data that it negatively impacts the model performance on unseen data – Overfitting (Low bias and high variance). If the accuracy model for test dataset is high then we can say the model is generalized.

1. R-square value of testing data set should be high.
2. RSS and MSE value should be least.
3. The model should satisfy all the assumptions of linear regression. Such as ,

    a. Linearity of the independent variables with the target variable
    b. Error terms shows Normal distribution with mean 0
    c. Homoscedasticity of error terms
    d. Error terms are independent to each other

If the model satisfies all the above conditions, then we can say the model is robust and generalizable.

The accuracy of the model is analysed by checking the R-square value and error terms. If the R-square value of test set is high then it indicates that the model is predicting target variable with high accuracy.

If the model is not predicting the target variable accurately then the error terms will increase and R-square value of unseen data get decreases.