Advanced Regression 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 be the most important predictor variables after the change is implemented??

Answer –

1) Based on the analysis done the most optimum value of alpha for:
   a) **Ridge**: 0.0100
   b) **Lasso**: 0.1000
2) With the features (predictor) variable selected for both ridge and lasso, doubling the alpha seems to have no impact. However, when the alpha goes above 10, we can see a significant drop in the R-squared. Given below are the analysis under various alpha for both ridge and lasso.

Ridge:

Alpha	R-Squared	MSE	RMSE
0.0001	0.871765	825.246083	28.7271
0.001	0.871765	825.246084	28.7271
0.01	0.871765	825.246134	28.7271
0.1	0.871764	825.250998	28.72718
1	0.871706	825.626071	28.73371
10	0.86958	839.309094	28.97083
100	0.849973	965.489473	31.07233
1000	0.810576	1219.02095	34.91448
10000	0.794085	1325.1516	36.40263

Figure 1 Ridge under various alphas

Lasso:

Alpha	R-Squared	MSE	RMSE
0.0001	0.871747	825.36063	28.72909
0.001	0.871747	825.361423	28.7291
0.01	0.871735	825.440689	28.73048
0.1	0.87094	830.555742	28.81936
1	0.853114	945.270842	30.74526
10	0.798353	1297.68238	36.02336
100	0.783196	1395.22107	37.35266
1000	0.726151	1762.3351	41.98018
10000	0.548374	2906.3998	53.91104

Figure 2 Lasso under various alphas

3) The most important variables after making the changes are:

a) **Ridge**:

Predictor variable	Coefficient
SaleType	27.845646
LotShape	19.500471
SaleCondition	18.256722
OverallQual	15.401151
MSZoning	13.683681
LandSlope	13.081616
RoofStyle	12.226264
Exterior1st	11.745546
HeatingQC	10.314656
GarageCars	8.813108
TotRmsAbvGrd	6.629611
Foundation	4.98183
CentralAir	3.653439
BsmtHalfBath	3.536097
PavedDrive	3.509219
Fireplaces	2.788057
YrSold	1.924351
BsmtFinType1	1.919657
YearRemodAdd	0.157323
ScreenPorch	0.070239
PoolArea	0.054333
WoodDeckSF	0.035856
GrLivArea	0.031607
MasVnrArea	0.02825
OpenPorchSF	0.025915
TotalBsmtSF	0.023892
2ndFlrSF	0.023502
BsmtFinSF1	0.02063
GarageArea	0.011112
BsmtFinSF2	0.010999
1stFlrSF	0.008891
3SsnPorch	0.007349
LowQualFinSF	-0.000768
EnclosedPorch	-0.001213
BsmtUnfSF	-0.00773
MiscVal	-0.009061
YearBuilt	-0.050206
MoSold	-0.098661
MSSubClass	-0.103937
GarageType	-0.59578
HouseStyle	-2.149937

Predictor variable	Coefficient
FullBath	-2.222125
HalfBath	-2.695104
Heating	-2.878484
GarageCond	-3.133723
GarageFinish	-4.299558
BsmtCond	-4.462299
GarageQual	-6.112035
BsmtFullBath	-6.307351
Electrical	-10.977383
BldgType	-12.070607
BedroomAbvGr	-13.801407
KitchenAbvGr	-21.164636

Figure 3 Most important predictors in RIdge

b) **Lasso**:

Predictor variable	Coefficient
SaleType	27.846598
LotShape	19.452451
SaleCondition	18.239368
OverallQual	15.369753
MSZoning	13.662058
LandSlope	13.20469
RoofStyle	12.226829
Exterior1st	11.874126
HeatingQC	10.273612
GarageCars	8.662425
TotRmsAbvGrd	6.63319
Foundation	5.000981
CentralAir	3.574482
BsmtHalfBath	3.477366
PavedDrive	3.324158
Fireplaces	2.799208
BsmtFinType1	1.938961
YrSold	1.912537
YearRemodAdd	0.158175
ScreenPorch	0.069506
PoolArea	0.054839
2ndFlrSF	0.048955
BsmtFinSF1	0.037196
WoodDeckSF	0.035559
1stFlrSF	0.034485
MasVnrArea	0.028598
BsmtFinSF2	0.027632

Predictor variable	Coefficient
OpenPorchSF	0.025656
LowQualFinSF	0.025317
GarageArea	0.011035
BsmtUnfSF	0.008796
3SsnPorch	0.007341
TotalBsmtSF	0.007245
GrLivArea	0.006266
EnclosedPorch	-0.000872
MiscVal	-0.009113
YearBuilt	-0.045609
MoSold	-0.090672
MSSubClass	-0.104154
GarageType	-0.936312
HouseStyle	-2.139323
FullBath	-2.298964
HalfBath	-2.724867
Heating	-2.912238
BsmtCond	-4.132245
GarageFinish	-4.403695
BsmtFullBath	-6.334328
GarageQual	-8.505059
Electrical	-10.936979
BldgType	-12.118996
BedroomAbvGr	-13.8011
KitchenAbvGr	-21.126024

Figure 4 Lasso most important predictors in Lasso

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 – Choosing between ridge and lasso regression depends on the specific characteristics of the data and the goals of the analysis.

In general, Ridge regression is preferred when dealing with multicollinearity, as it shrinks coefficients towards zero without eliminating them entirely.

Lasso regression, on the other hand, performs variable selection by forcing some coefficients to zero, which can be useful when we have many predictors and want a sparse model.

Understanding we have a huge number of predictor variables, and an iterative modelling be inefficient, in our particular case, I would prefer choosing **Lasso** regression. We don't see a lot of difference in r-squared values between ridge and lasso, also as shown below the residual of errors also shows a proper Gaussian distribution.

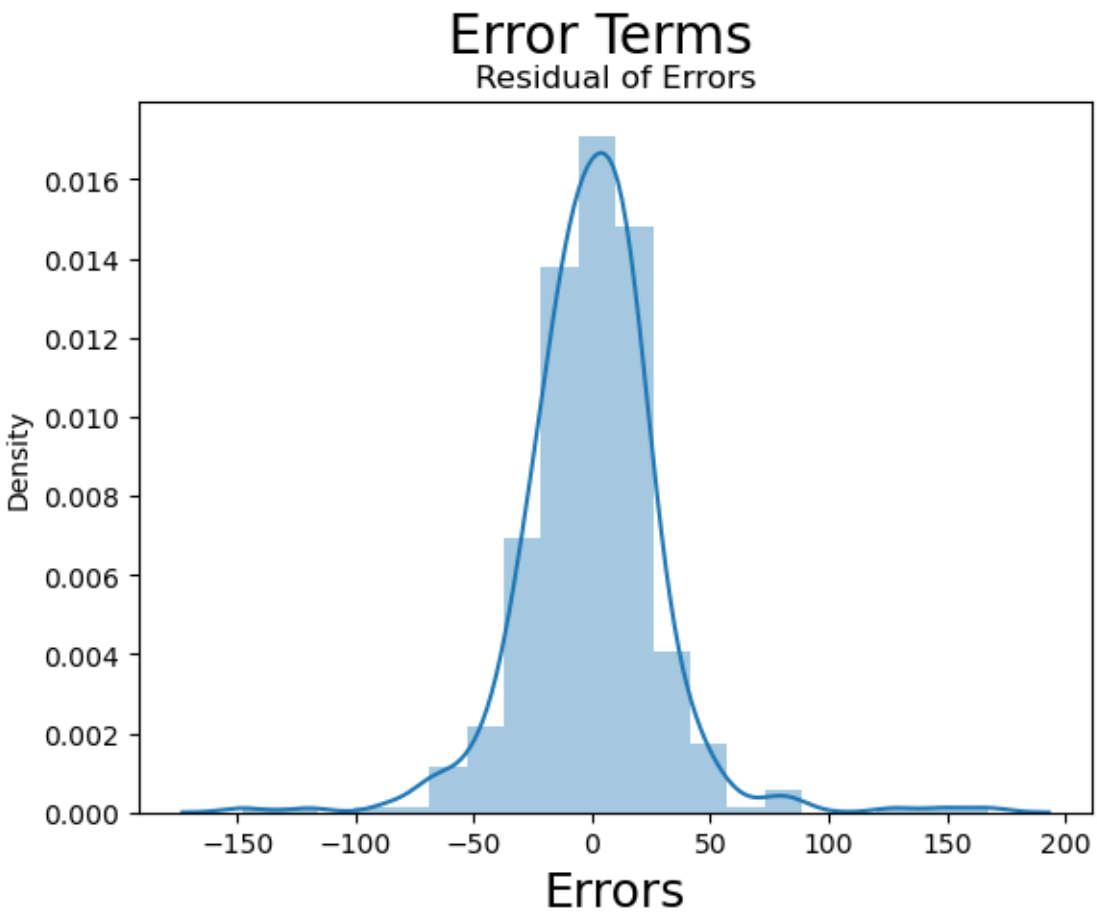

Figure 5 Distribution of residual of errors under Lasso regression

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?

Answer – If we go by the **Lasso** regression, the top 5 most important predictor terms are:

Predictor variable	Coefficient
SaleType	27.846598
LotShape	19.452451
SaleCondition	18.239368
OverallQual	15.369753
MSZoning	13.662058

Figure 6 Initial Top 5 predictors

Although they have a moderately high coefficients, they are contributing to achieve a ~87% of r-squared value on our test dataset.

If for some reason, these predictor variables are unavailable, the next top 5 most important predictor variables will be:

Predictor variable	Coefficient
Heating	20.112689
HeatingQC	18.955526
RoofStyle	15.148969
GarageCars	12.019159
TotRmsAbvGrd	8.798428

Figure 7 Next top 5 predictors

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?

Answer – Ensuring that a model is robust and generalizable involves several steps throughout the modeling process, from data preprocessing to model evaluation.

1) **Data Quality and Preprocessing**: We start by ensuring that our data is clean, free of errors, and representative of the problem we are trying to solve. This involves handling missing values, dealing with outliers, and encoding categorical variables appropriately. If we python source code submitted as part of this assignment, we have taken several steps to cleanse the data before beginning modeling process.

2) **Feature Selection and Engineering**: Choosing relevant features that contribute meaningfully to the predictive power of the model is extremely important. This might involve feature selection techniques like correlation analysis, or techniques like RFE that capture important attributes the data. In the python source code, we can see before I begun using ridge or lasso regression, I have used the general RFE technique and used linear regression modelling to determine the most important predictors.

3) **Model Selection and Evaluation**: One should experiment with different algorithms and model architectures to find the one that best fits their data and problem.

4) **Hyperparameter Tuning**: Fine-tuning the hyperparameters of our chosen model is important to optimize its performance.

5) **Regularization**: Incorporating regularization techniques like ridge regression or lasso regression helps to prevent overfitting and improve the model's generalization ability. In the python source code, we can see I have used various alpha parameters to check the optimal value of the R-squared.

Implications for Model Accuracy:

- Building a robust and generalizable model often involves striking a balance between bias and variance.
- A model with high bias (underfitting) may not capture the underlying patterns in the data, leading to poor accuracy on both the training and test sets.
- Conversely, a model with high variance (overfitting) may perform well on the training set but generalize poorly to new data, resulting in a drop in accuracy on the test set.
- By following best practices for model robustness and generalization, we can improve the chances of building a model that achieves high accuracy on unseen data, leading to more reliable predictions in real-world scenarios.