Advanced Regression Assignment

Name: Budhaditya Saha

Date: 17-02-2021

Assignment-based 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 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> The optimal value of alpha for ridge is 0.1 and for Lasso is 0.0001.

Following is the table of the metrices captured for alpha and 2*alpha to demonstrate the changes in the ridge and lasso models.

As we can observe from the table below, there are no significant changes for the metrics value other than the RSS values between Ridge Regression (for alpha) and Ridge Regression_2 (for 2*alpha). Also, similar is the case for Lasso Regression (for alpha) and Lasso Regression_2 (for 2*alpha).

The RSS values has increased for both Ridge and Lasso when the alpha is higher.

	Metric	Linear Regression	Ridge Regression	Ridge Regression_2	Lasso Regression	Lasso Regression_2
0	R2 Score (Train)	0.866939	0.848032	0.832905	0.799707	0.784254
1	R2 Score (Test)	0.753414	0.773528	0.781570	0.780122	0.787400
2	RSS (Train)	1.637317	1.869966	2.056114	2.464610	2.654755
3	RSS (Test)	1.340397	1.231059	1.187348	1.195217	1.155655
4	MSE (Train)	0.040045	0.042796	0.044876	0.049132	0.050992
5	MSE (Test)	0.055320	0.053015	0.052066	0.052238	0.051366

The comparison of the predictors for Ridge & Lasso for before and after the change is captured in the table below:

	Ridge	Ridge2	Lasso	Lasso2
LotArea	0.142824	0.137836	0.108932	7.009392e-02
OverallQual	0.289363	0.299267	0.319999	3.294737e-01
BsmtFinSF1	0.210838	0.192301	0.164018	1.487186e-01
TotalBsmtSF	0.246730	0.198390	0.088737	3.299499e-02
1stFlrSF	0.190310	0.196307	0.082993	1.171108e-01
2ndFlrSF	0.091859	0.083342	0.000000	0.000000e+00
LowQualFinSF	-0.033000	-0.034663	-0.047813	-3.755106e-02
GrLivArea	0.191167	0.192601	0.364314	3.382306e-01
MSSubClass_90	-0.019589	-0.018269	-0.026769	-2.237499e-02
Condition2_PosN	-0.537665	-0.474707	-0.427799	-2.980990e-01
BldgType_2fmCon	-0.035461	-0.034309	-0.028189	-2.151567e-02
BldgType_Duplex	-0.019589	-0.018269	-0.000449	-3.915159e-04
RoofMatl_CompShg	0.572179	0.390827	0.105851	2.990357e-02
RoofMatl_Membran	0.560986	0.362890	0.050939	0.000000e+00
RoofMatl_Metal	0.543066	0.346559	0.029342	0.000000e+00
RoofMatl_Roll	0.515678	0.321207	0.000000	0.000000e+00
RoofMatl_Tar&Grv	0.548480	0.365704	0.080785	0.000000e+00
RoofMatl_WdShake	0.533769	0.347554	0.050798	0.000000e+00
RoofMatl_WdShngl	0.663178	0.480103	0.205640	1.145397e-01
BsmtQual_none	0.017661	0.014436	0.017226	1.918569e-03
BsmtCond_none	0.017661	0.014436	0.000000	2.955205e-18
BsmtFinType1_none	0.017661	0.014436	0.000000	0.000000e+00
Heating_OthW	-0.147493	-0.136538	-0.068203	-0.000000e+00
BedroomAbvGr_8	-0.035184	-0.026839	-0.000000	-0.000000e+00
SaleType_New	0.022665	0.021760	0.040506	3.860928e-02
SaleCondition_Partial	0.022665	0.021760	0.000000	0.000000e+00

Top 5 predictors for Ridge before the change were:

RoofMatl_WdShngl	0.66
RoofMatl_CompShg	0.57
RoofMatl_Membran	0.56
RoofMatl_Tar&Grv	0.55
RoofMatl_Metal	0.54

Top 5 predictors for Ridge after the change are:

RoofMatl_WdShngl	0.48
RoofMatl_CompShg	0.39
RoofMatl_Membran	0.36
RoofMatl_Tar&Grv	0.37
RoofMatl_Metal	0.35

Hence, we can infer that there although there are no changes to the most important predictor variables when the alpha is doubled, however, the value of the coefficients has reduced in the model with 2*alpha for the Ridge Regression thereby reducing the model complexity for addressing the overfitting.

Top 5 predictors for Lasso before the change were:

GrLivArea	0.36
OverallQual	0.32
RoofMatl_WdShngl	0.21
BsmtFinSF1	0.16
LotArea	0.11

Top 5 predictors for Lasso before the change are:

GrLivArea	0.34
OverallQual	0.33
RoofMatl_WdShngl	0.11
BsmtFinSF1	0.15
LotArea	0.07

Hence, we can infer that there although there are no changes to the most important predictor variables when the alpha is doubled, however, the value of the coefficients has reduced in the model with 2*alpha for the Lasso Regression thereby reducing the model complexity for addressing the overfitting.

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>

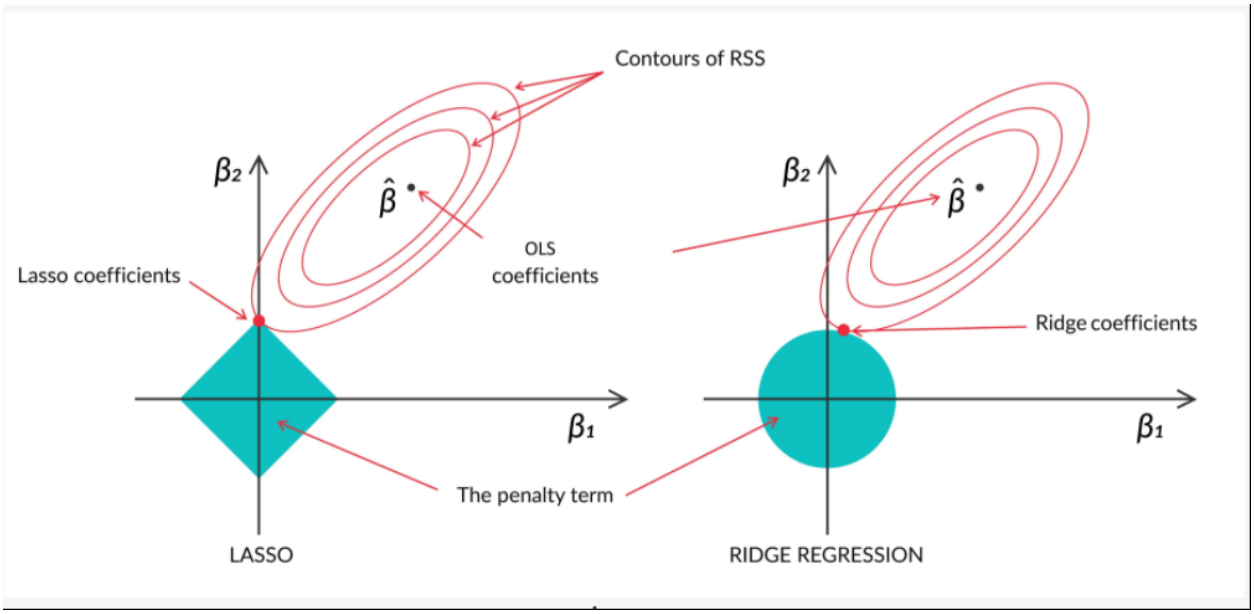
Some of the key differentiators between Ridge and Lasso is listed below:

- Ridge regression does have one obvious disadvantage. It would include all the predictors in the final model. This may not affect the accuracy of the predictions but can make model interpretation challenging when the number of predictors is very large.
- Like Ridge regression, the Lasso shrinks the coefficients estimate towards zero.

- Penalty in Lasso forces some of the coefficients estimates to be exactly equal to zero – hence it performs feature scaling
- Models generated from Lasso and generally easier to interpret than those produced by Ridge regression
- Hence the optimal selection of the lambda value is crucial for Lasso regression
- In Lasso also, as the lambda increases, the variance decreases, and the bias increases hence it moves from being overfitting to underfitting.

The primary difference between Lasso and Ridge regression is their penalty term. The penalty term here is the sum of the absolute values of all the coefficients present in the model. As with Ridge regression, Lasso regression shrinks the coefficient estimates towards 0. However, there is one difference. With Lasso, the penalty pushes some of the coefficient estimates to be exactly 0, provided the tuning parameter, λ, is large enough.

Hence, Lasso performs feature selection. Choosing an appropriate value of lambda is critical here as well. Because of this, it is easier to interpret models generated by Lasso as compared with those generated by Ridge. Also, just like with Ridge regression, standardization of variables is necessary for Lasso as well.



For this assignment, using the grid search the optimal alpha values obtained for Ridge and Lasso are

- Ridge Regression: 0.1
- Lasso Regression: 0.0001

The R2 square, RSS and RMSE values obtained for the different models are listed in the table below.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.866939	0.848032	0.799707
1	R2 Score (Test)	0.753414	0.773528	0.780122
2	RSS (Train)	1.637317	1.869966	2.464610
3	RSS (Test)	1.340397	1.231059	1.195217
4	MSE (Train)	0.040045	0.042796	0.049132
5	MSE (Test)	0.055320	0.053015	0.052238

The following table provides the value of coefficients for the predictor variables.

	Linear	Ridge	Lasso
LotArea	0.146960	0.142824	0.108932
OverallQual	0.258376	0.289363	0.319999
BsmtFinSF1	0.260282	0.210838	0.164018
TotalBsmtSF	0.402327	0.246730	0.088737
1stFlrSF	0.158413	0.190310	0.082993
2ndFlrSF	0.114345	0.091859	0.000000
LowQualFinSF	-0.029342	-0.033000	-0.047813
GrLivArea	0.189714	0.191167	0.364314
MSSubClass_90	-0.023549	-0.019589	-0.026769
Condition2_PosN	-0.649300	-0.537665	-0.427799
BldgType_2fmCon	-0.038184	-0.035461	-0.028189
BldgType_Duplex	-0.023549	-0.019589	-0.000449
RoofMatl_CompShg	1.111425	0.572179	0.105851
RoofMatl_Membran	1.158233	0.560986	0.050939
RoofMatl_Metal	1.137692	0.543066	0.029342
RoofMatl_Roll	1.107797	0.515678	0.000000
RoofMatl_Tar&Grv	1.092990	0.548480	0.080785
RoofMatl_WdShake	1.091415	0.533769	0.050798
RoofMatl_WdShngl	1.205003	0.663178	0.205640
BsmtQual_none	0.027609	0.017661	0.017226
BsmtCond_none	0.027609	0.017661	0.000000
BsmtFinType1_none	0.027609	0.017661	0.000000
Heating_OthW	-0.156225	-0.147493	-0.068203
BedroomAbvGr_8	-0.056839	-0.035184	-0.000000
SaleType_New	0.025116	0.022665	0.040506
SaleCondition_Partial	0.025116	0.022665	0.000000

Inference:

1. Looking into the metrics table – R2 square for Lasso & Ridge regression in the train and test data suggests a better fit of Lasso as compared to Ridge which is an overfit.
2. The RSS & RMSE values are also better for Lasso as compared to Ridge in the test set
3. Looking into the coefficients table, the Lasso has performed feature selection and helped identified a lesser set of the most important predictor variable compared to Ridge.

Given the above, we can conclude that Lasso model is working better and hence I will choose this model with the optimal value of alpha i.e. 0.0001 which was obtained using the GridSearch.

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?

Answer 3>

The following table provides the coefficient values of the Predictor variables from the Lasso model.

List of Features	Coefficients
GrLivArea	0.364
OverallQual	0.320
RoofMatl_WdShngl	0.206
BsmtFinSF1	0.164
LotArea	0.109
RoofMatl_CompShg	0.106
TotalBsmtSF	0.089
1stFlrSF	0.083
RoofMatl_Tar&Grv	0.081
RoofMatl_Membran	0.051
RoofMatl_WdShake	0.051
SaleType_New	0.041
RoofMatl_Metal	0.029
BsmtQual_none	0.017
2ndFlrSF	0.000
RoofMatl_Roll	0.000
BsmtCond_none	0.000
BsmtFinType1_none	0.000

BedroomAbvGr_8	0.000
SaleCondition_Partial	0.000
BldgType_Duplex	0.000
MSSubClass_90	-0.027
BldgType_2fmCon	-0.028
LowQualFinSF	-0.048
Heating_OthW	-0.068
Condition2_PosN	-0.428

We observe that the top 5 features predicted by the model are the ones listed below.

List of Features	**Coefficients**
GrLivArea	0.364
OverallQual	0.320
RoofMatl_WdShngl	0.206
BsmtFinSF1	0.164
LotArea	0.109
RoofMatl_CompShg	0.106

Also, the following metric provides the key scores which help us understand how the model is performing.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.866939	0.848032	0.799707
1	R2 Score (Test)	0.753414	0.773528	0.780122
2	RSS (Train)	1.637317	1.869966	2.464610
3	RSS (Test)	1.340397	1.231059	1.195217
4	MSE (Train)	0.040045	0.042796	0.049132
5	MSE (Test)	0.055320	0.053015	0.052238

Now, as per the question, the top 5 variables are not available in the incoming data and we need to identify what will be the 5 top variables if the initial top 5 variables are missing.

We build another model after dropping the top 5 variables from the predictor variable data frame and performed the analysis and following are the observations.

The coefficients obtained for the predictor variables after dropping the top5 variables from the earlier Lasso model is listed below.

List of Features	Coefficients
1stFlrSF	0.521
TotalBsmtSF	0.444
2ndFlrSF	0.238
SaleType_New	0.062
BsmtQual_none	0.029
RoofMatl_CompShg	0.020
RoofMatl_Membran	0.000
RoofMatl_Metal	0.000
RoofMatl_Roll	0.000
RoofMatl_Tar&Grv	0.000
RoofMatl_WdShake	0.000
BsmtCond_none	0.000
BsmtFinType1_none	0.000
BedroomAbvGr_8	0.000
SaleCondition_Partial	0.000
BldgType_Duplex	-0.002
Heating_OthW	-0.003
LowQualFinSF	-0.053
BldgType_2fmCon	-0.057
MSSubClass_90	-0.069
Condition2_PosN	-0.480

The new top5 predictor variables predicted by the new model is as listed below.

List of Features	Coefficients
1stFlrSF	0.521
TotalBsmtSF	0.444
2ndFlrSF	0.238
SaleType_New	0.062
BsmtQual_none	0.029

What we have observed is that the sequencing of the new top5 variables has changed as compared to the previous model i.e. they are not the same variables from 6 to 10 from the previous model.

Also, from the following metric we observed that the model accuracy of the new model has significantly reduced as compared to the previous model.

```
R2 score_train:  0.6726215321554745
R2 score_test:  0.6494544531225319
RSS score_train:  4.028401363174951
```

```
RSS score_test:  1.905502920459972
RMSE score_train:  0.003945544919857934
RMSE score_test:  0.004350463288721397
```

This does confirm that the top 5 variables from the earlier model which are not available in the incoming data were the key indicators of the sale price.

Question 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 > A common problem in machine learning is overfitting i.e. when the complexity of the model is very high. These kinds of models will perform very well in the training set but when given an unseen data the model fails to perform well. These models are said to have low bias and high variance – low bias as it will not make much error in the training set and hence very robust. High variance because the because the errors will be very high in unseen data and hence not generalizable.

The opposite also holds true when the model is overtly simple, and this is the scenario of underfitting. In this scenario the errors will be very high in the training set i.e. high bias, but the model will not do much in an unseen data set as it has not learnt the pattern from the training set and hence low variance. This is case when the model is generalizable.

In order to address this issue, it is important that we can strike the right balance between overfitting and underfitting to obtain and robust and generalizable model.

Here comes regularization in which a penalty term is added to the algorithm's loss function. This changes the model's weights which result from minimizing the loss function.

The most popular regularization techniques are Lasso, Ridge and Elastic Net.

Hence, we can conclude that regularization is the process by which we can make the model robust and generalizable and prevent overfitting of the train dataset.

Regularization significantly reduces the variance of the model, without substantial increase in its bias. So, the tuning parameter λ, used in the regularization techniques described above, controls the impact on bias and variance. Hence it will finally help in improving the overall accuracy of the model by balancing out the accuracy of the train and test data sets