Assignment : Advanced Regression

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?

Answer:-

The optimal Value of alpha for
Ridge :- 50
Lasso :- 0.005

When we double the alpha value of alpha in ridge we get

Metrics	Ridge_Regression	Ridge_Regression_double_alpha	Difference
Train_r2_score	0.937247	0.934367	0.002880
Test_r2_score	0.893651	0.894082	-0.000431
Train_Mean_absolute_error	0.166105	0.185029	-0.018924
Test_Mean_absolute_error	0.184717	0.168178	0.016539
Train_Mean_squared_error	0.062753	0.071310	-0.008558
Test_Mean_squared_error	0.071601	0.065633	0.005968
Train_RMSE	0.250505	0.267040	-0.016535
Test_RMSE	0.267583	0.256189	0.011394

We are getting almost similar results when compared with original alpha with slight changes in values ,r2 score for train increases while decreases in test data ,similarly for MAE and MSE error on train decreases while increase on test data, for RMSE on test is increased and decreased on train data

can be seen in above table where red signifies decrease while green signifies increase in value.

When we double the alpha value of alpha in Lasso we get

Metrics	Lasso_Regression	Lasso_Regression_double_alpha	Difference
Train_r2_score	0.933907	0.925186	0.008720
Test_r2_score	0.906023	0.906300	-0.000277
Train_Mean_absolute_error	0.170544	0.181967	-0.011423
Test_Mean_absolute_error	0.175173	0.174696	0.000477
Train_Mean_squared_error	0.066093	0.074814	-0.008720
Test_Mean_squared_error	0.063271	0.063085	0.000187
Train_RMSE	0.257087	0.273521	-0.016434
Test_RMSE	0.251538	0.251166	0.000371

Here also for lasso we are getting similar results ,with slight change in values ,r2 score for train increases while decrease in test .Similarly for MAE and MSE error on train decreases while increase on test data,for RMSE on test is increased and decreased on train data

can be seen in above table where red signifies decrease while green signifies increase in value.

when Changes applied ,the top 5 important variables are

Ridge

Variables	Ridge_Regression
GrLivArea	0.278024
OverallQual	0.126333
TotalBsmtSF	0.118020
BsmtFinSF1	0.109186
LotArea	0.082545
YearBuilt	0.079545
2ndFlrSF	0.078328
BsmtExposure_Gd	0.071146
OverallCond	0.068115
Neighborhood_StoneBr	0.066658

Variables	Ridge_Regression_double_alpha
GrLivArea	0.237730
OverallQual	0.129413
TotalBsmtSF	0.119832
BsmtFinSF1	0.105557
LotArea	0.081225
2ndFlrSF	0.074639
BsmtExposure_Gd	0.070565
Neighborhood_NridgHt	0.067420
Neighborhood_StoneBr	0.064758
OverallCond	0.062286

Top 5 variables are same though but their coefficient value has changed.

Lasso

Variables	Lasso_Regression
GrLivArea	0.390465
OverallQual	0.136771
YearBuilt	0.121850
BsmtFinSF1	0.102961
SaleCondition_Partial	0.082665
TotalBsmtSF	0.081680
BsmtExposure_Gd	0.076671
LotArea	0.076054
OverallCond	0.069378
Neighborhood_NridgHt	0.067939

variables	Lasso_Regression_double_alpha
GrLivArea	0.384695
OverallQual	0.166769
YearBuilt	0.113368
BsmtFinSF1	0.101810
TotalBsmtSF	0.080768
SaleCondition_Partial	0.080456
BsmtExposure_Gd	0.073735
Neighborhood_NridgHt	0.072876
LotArea	0.070206
OverallCond	0.060579

Here , Top 4 variables are same though but 5th variable is changed 'TotalBsmtSF' comes inplace of 'SaleCondition_Partial' ,when we double the alpha in Lasso with slight change in coefficient values.

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?

Metrics	Ridge_Regression	Lasso_Regression
Train_r2_score	**0.937247**	**0.933907**
Test_r2_score	**0.893651**	**0.906023**
Train_Mean_absolute_error	0.166105	0.170544
Test_Mean_absolute_error	0.184717	0.175173
Train_Mean_squared_error	0.062753	0.066093
Test_Mean_squared_error	0.071601	0.063271
Train_RMSE	0.250505	0.257087
Test_RMSE	0.267583	0.251538

We will choose Lasso over ridge

Above Analysis shows difference between Ridge and Lasso Regression,here r2_Score for test is higher in lasso than ridge. Also the difference between train and test dataset is smaller in lasso than ridge

Ridge_Regression	4.359571 %
Lasso_Regression	**2.788347 %**

Lasso tends to produce sparse models by setting some coefficients to zero, which can be useful for feature selection in our dataset because we have large number of variables with many irrelevant features. Also the sum of residuals in test dataset is lower in lasso than ridge

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?

We rebuild the model after removing the lasso five most important variables, and calculated the new optimal alpha with GridsearchCV and find these results

Removed Variables :- 'GrLivArea', 'OverallQual', 'YearBuilt','BsmtFinSF1','SaleCondition_Partial'
Now with new model we have

Original Lasso

Variables	Lasso_Regression
GrLivArea	0.390465
OverallQual	0.136771
YearBuilt	0.121850
BsmtFinSF1	0.102961
SaleCondition_Partial	0.082665

After Rebuilding

Variables	lasso_new
TotalBsmtSF	0.340856
2ndFlrSF	0.233098
FullBath	0.100674
GarageArea	0.097575
LotArea	0.094669

Now after rebuilding the model Top 5 most Important Variables are :-

1) TotalBsmtSF

2) 2ndFlrSF

3) FullBath

4) GarageArea

5) LotArea

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?

Let’s Understand both terms first

Generalizability :- Generalizability is a measure of a model's ability to be successfully applied to data sets other than the one used for training and testing...i.e unseen data

Robustness :- The ability of a model to maintain its performance when unseen data has some uncertainties like noise, distribution shifts,variations and other challenges

Model are more generalizable and robust if:-

1) We Understand the Data :- this includes to conduct exploratory data analysis, understand the relationships between variables, identify potential outliers, and overall structure of your data. This initial exploration helps us to make informed decisions about which machine learning techniques might be most appropriate for our problem.

2) We preprocess the Data .To ensure this step we handle missing values, created dummy variables for categorical variables, Standardizing numerical variables. The goal of preprocessing is to make our data compatible with the machine learning algorithm we are using and to improve the algorithm's ability to uncover meaningful patterns.

3) We choose the model which depend on the problem at hand, the nature of the data, and the requirement of the task. Since we started with the base model of multiple linear regression but it fails ,so we choose regularization technique like Lasso and Ridge.

4) Cross-validation and hyperparameter tuning is a powerful technique that can help prevent overfitting, a common problem in machine learning where a model learns the training data too well and performs poorly on unseen data. By dividing our data into training and validation sets multiple times, we can ensure that our model generalizes well to unseen data and we choose optimal alpha value for ridge and lasso

5) We should always evaluate our model using appropriate metrics .

If a model fails to generalize well and lacks robustness, it will exhibit several undesirable characteristics. These include poor accuracy, potential overfitting, and inaccurate predictions made with high confidence when faced with noise, uncertainties, or challenges in unseen data. Such a model may also pose significant risks if deployed in critical domains such as healthcare, banking, or autonomous driving, potentially leading to costly errors and adverse consequences.