Assignment Part 2 – Subjective Questions – Advanced Regression

I have run the code in my jupyter notebook to get the answers for below questions, hence placing the screenshots of the piece of code that I ran.

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

- Optimal value of lambda for Ridge Regression = 10
- Optimal value of lambda for Lasso Regression = 0.001

```
In [99]:
## Doubling the value of alpha for ridge regression, alpha=20
ridge = Ridge(alpha=20)
ridge.fit(X_train, y_train) # fitting the model on training data
y_train_pred = ridge.predict(X_train) ##making predictions
y_pred = ridge.predict(X_test)
ridge_metrics = show_metrics(y_train, y_train_pred, y_test, y_pred) #getting the metrics

R-Squared (Train) = 0.92
R-Squared (Test) = 0.80
RSS (Train) = 10.93
RSS (Test) = 11.96
MSE (Train) = 0.01
MSE (Test) = 0.03
RMSE (Train) = 0.11
RMSE (Test) = 0.17
```

```
In [100]:
## Doubling the value of alpha for lasso regression, alpha=0.002
lasso = Lasso(alpha=0.002)
lasso.fit(X_train, y_train)  # fitting the model on training data
y_train_pred = lasso.predict(X_train) ##making predictions
y_pred = lasso.predict(X_test)
lasso_metrics = show_metrics(y_train, y_train_pred, y_test, y_pred)

R-Squared (Train) = 0.89
R-Squared (Test) = 0.79
RSS (Train) = 13.76
RSS (Test) = 12.89
MSE (Train) = 0.01
MSE (Test) = 0.03
RMSE (Train) = 0.12
RMSE (Test) = 0.18
```

In [101]:
```
# Again creating a table which contain all the metrics

lr_table = {'Metric': ['R2 Score (Train)','R2 Score (Test)','RSS (Train)','RSS (Test)',
                       'MSE (Train)','MSE (Test)', 'RMSE (Train)', 'RMSE (Test)'],
            'Ridge Regression' : ridge_metrics,
            'Lasso Regression' : lasso_metrics
        }

final_metric = pd.DataFrame(lr_table, columns = ['Metric', 'Ridge Regression', 'Lasso Regression'] )
final_metric.set_index('Metric')
```

Out[101]:

Metric	Ridge Regression	Lasso Regression
R2 Score (Train)	0.916284	0.894625
R2 Score (Test)	0.801449	0.785991
RSS (Train)	10.933509	13.762207
RSS (Test)	11.960819	12.891987
MSE (Train)	0.011607	0.014610
MSE (Test)	0.029533	0.031832
RMSE (Train)	0.107734	0.120870
RMSE (Test)	0.171851	0.178415

Based on doubling the value of alpha,

Ridge

- Train R2 score - Decreased from 0.92 to 0.91
- Test R2 score - Remains same at 0.80

Lasso

- Train R2 score - Decreased from 0.91 to 0.89
- Test R2 score - Decreased from 0.79 to 0.78

After finding the coefficients for both the models, below are the most important predictor variables.

```
Ridge Regression
 GrLivArea                1.121272
OverallQual              1.085764
Neighborhood_Crawfor     1.078536
SaleCondition_Normal     1.063361
Condition2_Norm          1.060517
OverallCond              1.052437
TotalBsmtSF              1.052417
Neighborhood_NridgHt     1.051649
Exterior1st_BrkFace      1.049290
CentralAir_Y             1.048751
Name: Ridge, dtype: float64


Lasso Regression
 GrLivArea                1.136251
OverallQual              1.102665
Neighborhood_Crawfor     1.063929
SaleType_New             1.056551
OverallCond              1.055999
TotalBsmtSF              1.054648
SaleCondition_Normal     1.043477
MSZoning_RL              1.040169
Neighborhood_NridgHt     1.038317
GarageCars               1.037334
Name: Lasso, dtype: float64
```

Therefore, the most important predictor variables after we double the alpha values are

- GrLivArea
- OverallQual
- Neighborhood_Crawfor
- SaleType_New
- OverallCond
- TotalBsmtSF
- SaleCondition_Normal
- MSZoning_RL
- Neighborhood_NridgHt
- GarageCars

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

Choosing of the model primarily depends on the use case that we are trying to solve. Lasso regression is mostly helpful for feature selection, in case if we don't want too many variables, then it can

eliminate most of them. On the other hand, Ridge regression helps when we want reduction of coefficient magnitude and want to get rid of large coefficients, then we go ahead with this.

For our current use case, even though there is not much variation in terms of R2 score for train and test sets, Lasso regression would be a better option because after we added dummies, the number of features had gone up to 200+ which is a lot more than the given set of variables that we had around 80+ in our given dataset. Lasso helps us in choosing the right variables, does feature selection, hence it would make the model more robust and gives us more accurate predictions. There wpn't be much highly correlated variables which will bring the efficacy of the model down. Lasso regression would be an optimal choice.

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?

Answer

After dropping the earlier top 5 predictor variables and building a new model, we get the new top 5 predictor variables

- TotalBsmtSF 0.123439
- CentralAir_Y 0.116627
- SaleCondition_Normal 0.088726
- Neighborhood_NridgHt 0.083348
- Exterior1st_BrkFace 0.082107

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?

Answer

Robust Model - A model is said to be robust when the model is able to maintain its performance even when there are any variation in the data and it doesn't impact the model largely.

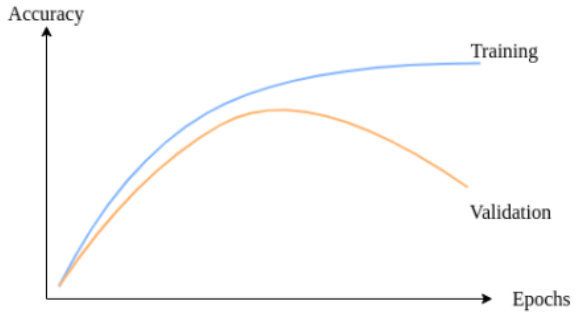
Generalizable Model - A model which has the capability to adapt itself to any unseen data, its ability to perform well on a new data which is of the same distribution used to create the model.

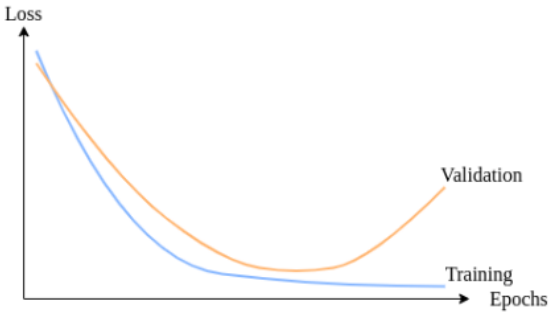
These 2 characteristics are important when a model is built, otherwise there would be more performance issues and we may not get accurate predictions.

The model should not be either overfit or underfit, that's when we come up with the term called regularization which helps in minimizing the adjusted loss function. When there is a high variance in the data, and if the predictions gets affected, then the model is said to be overfitted. The model may

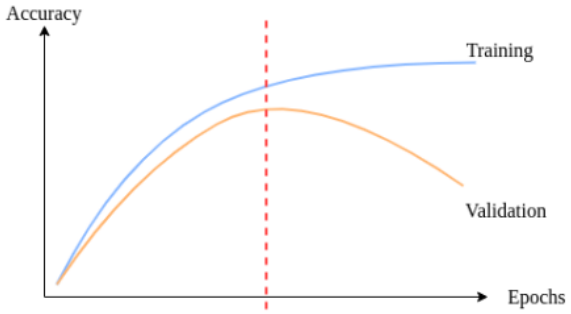
underfit if it has low variance but high bias. The model should neither be too complex (Overfit) nor too simple (Underfit).

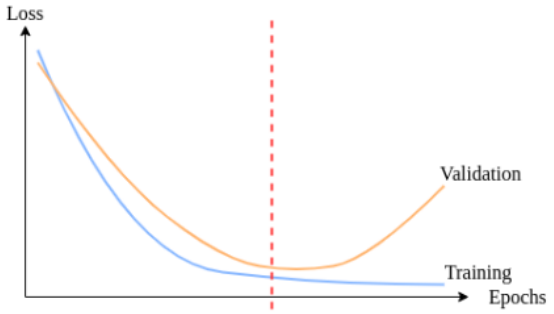
Implications of accuracy on Overfit model





Implications of accuracy on Underfit model





Underfitting	**Overfitting**
Model is too simple	Model is too complex
Not accurate neither on train or test set	Accurate on train but inaccurate in test set
Need to increase complexity to some extent	Need to reduce the complexity
Increase training	Reduce training
Increase the number of features	Reduce the number of features

It should not be impacted by outliers in the training data though we do all pre-processing and cleaning steps before building a model. In terms of accuracy, overfit models will have very high accuracy in training data but will have huge downfall when it comes to test dataset, which does not signify a good model. High bias may result in underfit model leading to an inaccurate model.

Hence, regularization comes into picture and this can be achieved by Ridge and Lasso regression.