Assignment Part – 2

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:

(*Please refer the Jupiter note file for the code.)*

- Optimal value of alpha
  - Optimal value of alpha for ridge regression = 20.0
  - Optimal value of alpha for lasso regression = 0.001

- Double the value of alpha for both ridge and lasso regression:
  - Ridge Regression:
    - Original model alpha = 20, Doubled alpha model = 40

```
ridge  Regression with  20
====================================
R2 score (train) :  0.9082161165128747
R2 score (test) :  0.9028399605761104
RMSE (train) :  0.12127004854480684
RMSE (test) :  0.12374911605573792
```

```
Model Evaluation : Ridge Regression, alpha = 40
R2 score (train) :  0.9074
R2 score (test) :  0.9034
RMSE (train) :  0.1218
RMSE (test) :  0.1234
```

    - Observations:
      - The test set R square value of Ridge regression model with double alpha (alpha = 40) model is slightly higher in comparison to test set R square value of the ridge regression model alpha (alpha = 20) model.
      - The train set R square value of Ridge regression model with double alpha (alpha = 40) model is slightly lower in comparison to train set R square value of the ridge regression model
      - The test set RMSE value of Ridge regression model with double alpha (alpha = 40) model is slightly lower in comparison to test set RMSE value of the ridge regression model alpha (alpha = 20) model.

- Lasso Regression:
  - Original model alpha = 0.001, Doubled alpha model = 0.002

```
lasso  Regression with  0.001
================================
R2 score (train) :  0.9077451204811589
R2 score (test) :  0.9031186604504103
RMSE (train) :  0.12158080366266924
RMSE (test) :  0.12357150378155785
```

```
Model Evaluation : Lasso Regression, alpha=0.002
R2 score (train) :  0.9069
R2 score (test) :  0.9039
RMSE (train) :  0.1221
RMSE (test) :  0.1231
```

  - Observations:
    - The test set R square value of lasso regression model with double alpha (alpha = 0.002) model is slightly higher in comparison to test set R square value of the lasso regression model alpha (alpha = 0.001) model.
    - The train set R square value of Lasso regression model with double alpha (alpha = 40) model is slightly higher in comparison to train set R square value of the lasso regression model
    - The test set RMSE value of Lasso regression model with double alpha (alpha = 0.002) model is slightly lower in comparison to test set RMSE value of the ridge regression model alpha (alpha = 20) model.

- The most important predictor variables after the change:
  - Ridge Regression Model with double alpha (alpha = 40)

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. LotArea

	Ridge (alpha = 40)
1stFlrSF	0.126582
2ndFlrSF	0.102397
OverallQual	0.087176
OverallCond	0.042669
LotArea	0.036941

  - Lasso Regression Model with double alpha (alpha = 0.002)

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. LotArea

	Lasso (alpha = 0.002)
1stFlrSF	0.135743
2ndFlrSF	0.108855
OverallQual	0.092196
OverallCond	0.042953
LotArea	0.033809

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:

(*Please refer the Jupiter note file for the code.)*

- Optimal value of alpha for ridge regression = 40.0
- Optimal value of alpha for lasso regression = 0.002

```
Model Evaluation : Ridge Regression, alpha = 40
R2 score (train) :  0.9074
R2 score (test) :  0.9034
RMSE (train) :  0.1218
RMSE (test) :  0.1234
```

```
Model Evaluation : Lasso Regression, alpha=0.002
R2 score (train) :  0.9069
R2 score (test) :  0.9039
RMSE (train) :  0.1221
RMSE (test) :  0.1231
```

- Ridge regression: R2 score of test is 0.9034, RMSE of test is 0.1234
- Lasso regression: R2 score of test is 0.9039, RMSE of test is 0.1231
- Lasso Regression produced slightly high R2 score on test data than Ridge Regression test R2 score. And Lasso regression shows slightly lower RMSE value on test data than Ridge regression test RMSE value. Choosing Lasso as the final model.

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:

(*Please refer the Jupiter note file for the code.)*

- Top five features in Lasso final model (Alpha = 0.001)

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. LotArea

	Lasso (alpha=0.001)
1stFlrSF	0.138275
2ndFlrSF	0.111747
OverallQual	0.090696
OverallCond	0.044980
LotArea	0.033862

- Top five feature after excluding the previous five features.

1. GarageArea
2. FireplaceQu
3. KitchenQual
4. BsmtFinSF1
5. HalfBath

	Lasso
GarageArea	0.103335
FireplaceQu	0.073160
KitchenQual	0.058781
BsmtFinSF1	0.054331
HalfBath	0.048404

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:

- Model robustness is the model's ability to withstand uncertainties and perform accurately in different contexts. Ensuring the robustness and generalizability of a machine learning model involves several key steps:
  - **Diverse Training Data**: Use a large and diverse dataset that represents the problem space well. This helps the model learn various patterns and reduces overfitting.
  - **Regularization**: Apply regularization methods to prevent the model from becoming too complex and overfitting the training data.
  - **Cross-Validation**: Implement cross-validation techniques to evaluate the model's performance on unseen data. This helps in assessing the model's ability to generalize.
  - **Continuous Monitoring**: After deployment, continuously monitor the model's performance to detect any shifts in data distribution or declines in performance over time.

- **Implications for Accuracy:**
    - A robust and generalizable model may exhibit slightly lower accuracy on the training set due to its avoidance of overfitting and noise.
    - However, such a model is expected to perform better on real-world data, maintaining high accuracy when faced with new examples not seen during training.

  The important thing is to make sure the model learns from the training data effectively but also does well with new data it hasn't seen before. This way, the model can be accurate and reliable in different situations.
