Advanced Linear Regression
Submitted By: Kanika Khattar Ahuja

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?

Solution:

Ridge regression:
Optimal Value of alpha for is 500

Training r2: 0.915
Training RMSE: 0.116
Testing r2: 0.872
Testing RMSE: 0.145

Difference in training and testing RMSE: 0.029

The most important predictor that have impact on price:

- OverallQual 0.043460
- GrLivArea 0.039561
- 1stFlrSF 0.031051
- GarageArea 0.024856
- FullBath 0.023327
- Fireplaces 0.022426

Lasso regression:
Optimal Value of alpha for is 0.01

Training r2: 0.896
Training RMSE: 0.128
Testing r2: 0.86
Testing RMSE: 0.152

Difference in training and testing RMSE: 0.024

The most important predictor that have impact on price:

- GrLivArea 1.191096e-01

- OverallQual 1.086919e-01
- GarageArea 3.887910e-02
- OverallCond 2.247840e-02
- Fireplaces 2.181280e-02
- BsmtFullBath 2.090881e-02

With increase in alpha, the penalization will be greater for the variables.
As the coefficients of variables will be closer to zero with increase in alpha, the model will become more simpler and have a **high bias and low variance.**

Changes in model with change in alpha

For Ridge:

Training r2: 0.896
Training RMSE: 0.128
Testing r2: 0.862
Testing RMSE: 0.151

Important Predictor variables:

- OverallQual 0.036236
- GrLivArea 0.033544
- 1stFlrSF 0.026973
- GarageArea 0.022370
- Fireplaces 0.021358
- FullBath 0.021073

As the alpha increases, the predictor variables are the same but the coefficients are now closer to 0 due to higher penalisation.

As the model is biased, we see that there is a bigger difference

For Lasso:

Training r2: 0.86
Training RMSE: 0.148
Testing r2: 0.845
Testing RMSE: 0.16

Important Predictor variables:

- OverallQual 0.127570
- GrLivArea 0.103429
- GarageArea 0.043195
- Fireplaces 0.025063
- TotalBsmtSF 0.021433
- BsmtFullBath 0.015666

With high penalisation, Lasso reduced the number of non zero coefficients. The top predictors also changed .

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?

Solution:

We will go with alpha = 0.01 for lasso regression.

Difference in training and testing RMSE for Ridge Regression model: 0.029
Difference in training and testing RMSE for Lasso Regression model: 0.024

With Lasso Regression, the model seems to have low bias and low variance as compared to ridge. We can see that the difference between the training and test score is minimal as compared to the difference in ridge regression model, the lasso model performs better and is more generalized.

Considering the r2 score, Lasso gives an r2 score of 86% on the testing data which is good enough to proceed.

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?

After removing the important predictor variables:
Best alpha : 0.01

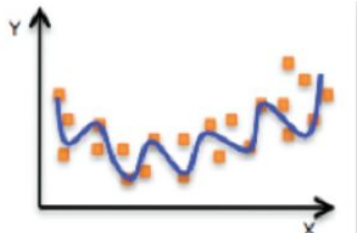
- 1stFlrSF 1.233862e-01
- 2ndFlrSF 1.100395e-01
- Neighborhood_NridgHt 3.258390e-02
- TotalBsmtSF 2.763967e-02
- FullBath 2.642513e-02

Question 4

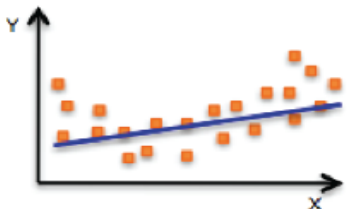
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Robustness of the model characterizes how effective the model is on the test data. A robust model should allow minimum change in the model when there is a slight modification in the input data. The model should have a low variance.

The model robustness can be confirmed with the difference in the performance on the train and test data. If the model is overfitted, as shown in the figure below, the model will have a high accuracy on the training data but will have a very low accuracy on testing data.
Because of the high variance in the model, it will not be flexible for a new data and thus it will be not robust.

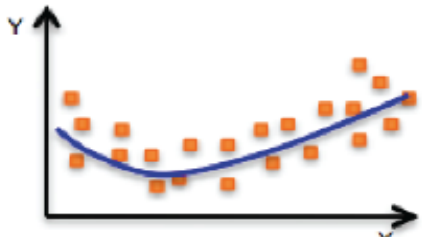


A model also needs to be generalized enough to perform well and learn to work on unseen data.
If the model is too much generalized or simple, it might underfit the data as shown in the figure.



In case of underfitting, the model will not perform well on training and testing datasets. If the model is too simple, it will have a high bias. And if the model is too complex and not generalizable, it will lead to overfitting as discussed above.

Therefore, it is very important to identify when to stop training the model and ensure that model has a perfect combination of bias and variance and perform equally when on training and testing data. The balanced model is as shown below.



Implications on accuracy: To make the model robust and generalisable, the accuracy on the training dataset might reduce but as the same time, the predictive power increases. What matters the most is, the model should be as accurate on unseen dataset as it was on the training dataset.
If the model is 90% accurate on the training dataset and it has a 60% accuracy on testing, it does not mean that it is a good model. In such scenarios, there are chances that the model is overfitted and complex and we need to generalize the model. Therefore, only accuracy of the model should not be used as evaluation metrics, we should also take into account the complexity of the model.