# Subjective Questions

## Problem Statement - Part II

### 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?

Ans :

- The optimal alpha values are
  - Ridge : 5.0
  - Lasso : 200.0
- On doubling the alpha values :
  - Ridge the R2 score on train increase and test decreases
  - On doing the same for Lasso the R2 score has decreased on both train and test
- Most important predictors are:
  - Ridge:
    - 

| | Features | Coefficients |
|---|---|---|
| 11 | Neighborhood_NridgHt | 33419.3 |
| 7 | GrLivArea | 25257.1 |
| 1 | BldgType_Twnhs | -14819.8 |
| 14 | OverallQual | 12893.6 |
| 2 | BsmtFinSF1 | 10039.2 |

  - Lasso:
    - 

| | Features | Coefficients |
|---|---|---|
| 11 | Neighborhood_NridgHt | 33254.8 |
| 7 | GrLivArea | 25433.6 |
| 14 | OverallQual | 13398.3 |
| 1 | BldgType_Twnhs | -11815.1 |
| 2 | BsmtFinSF1 | 9918.46 |

## 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?

Ans: The below table describes the results of the models

| | Metric | Linear Regression RFE | Ridge Regression RFE | Lasso Regression RFE |
|---|---|---|---|---|
| 0 | R2 Score (Train) | 0.861 | 0.8609 | 0.8608 |
| 1 | R2 Score (Test) | 0.8629 | 0.8641 | 0.8651 |
| 2 | RSS (Train) | 944833529003.231 | 945513297644.4008 | 946582105641.886 |
| 3 | RSS (Test) | 329535235094.6472 | 326619972681.7866 | 324254531462.607 |
| 4 | MSE (Train) | 30420.39 | 30431.3311 | 30448.526 |
| 5 | MSE (Test) | 27429.2463 | 27307.6491 | 27208.586 |
| 6 | No of Features | 17.0 | 17.0 | 17.0 |
| 7 | Alpha | 0.0 | 5.0 | 200.0 |

On comparing R2_scores :

- Both the models has test scores higher than train
- Lasso has lower train score than Ridge but higher test lower scores

On Comparing MSE :

- Lasso has least MSE on test data when compared to Ridge
- Ridge has lower MSE on train data when compared to Lasso

Lower the error in Test data the model performs better on unseen data. And the R2 Score for test in lasso is significantly higher than other models.

Thus I will pick Lasso when I compare to 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?

Ans : After removing the top five features in Lasso, the new top five important features are

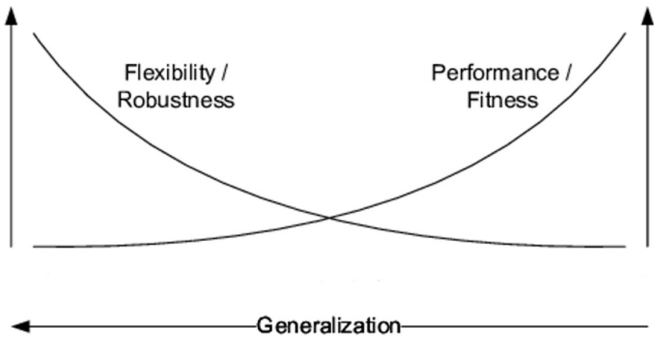
| | Features | Coefficients |
|---|---|---|
| 15 | SaleType_CWD | 51472 |
| 1 | 2ndFlrSF | 32032.6 |
| 0 | 1stFlrSF | 30615.2 |
| 14 | SaleCondition_Partial | 21821.1 |
| 3 | BsmtExposure_Gd | 21634.4 |

## 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?

The robust and generalized performance of the model is the one which does not deteriorate too much when training and testing with slightly different data (either by adding noise or by taking other dataset). The robustness property is also known as algorithmic stability.

Generalization refers to your model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model.




The accuracy of the model depends on n how well the model is fit and the complexity of the model. Lower bias gives more accuracy to the model. If the model goes underfit with less features are

overfit with complex model, then accuracy will decrease for the unseen data. So it is important to have a balanced model which should be robust and generalisable to have a good accuracy.