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 Ridge Regression is 1.0 & the optimal value of Lasso Regression is 10.0

If we double the value of alpha for Ridge i.e. value of alpha as 2 then R2score on training data has decreased but it has increased on testing data & if we double the value of alpha for Lasso i.e. value of alpha as 20 then R2score on training data has decreased but it has increased on testing data.

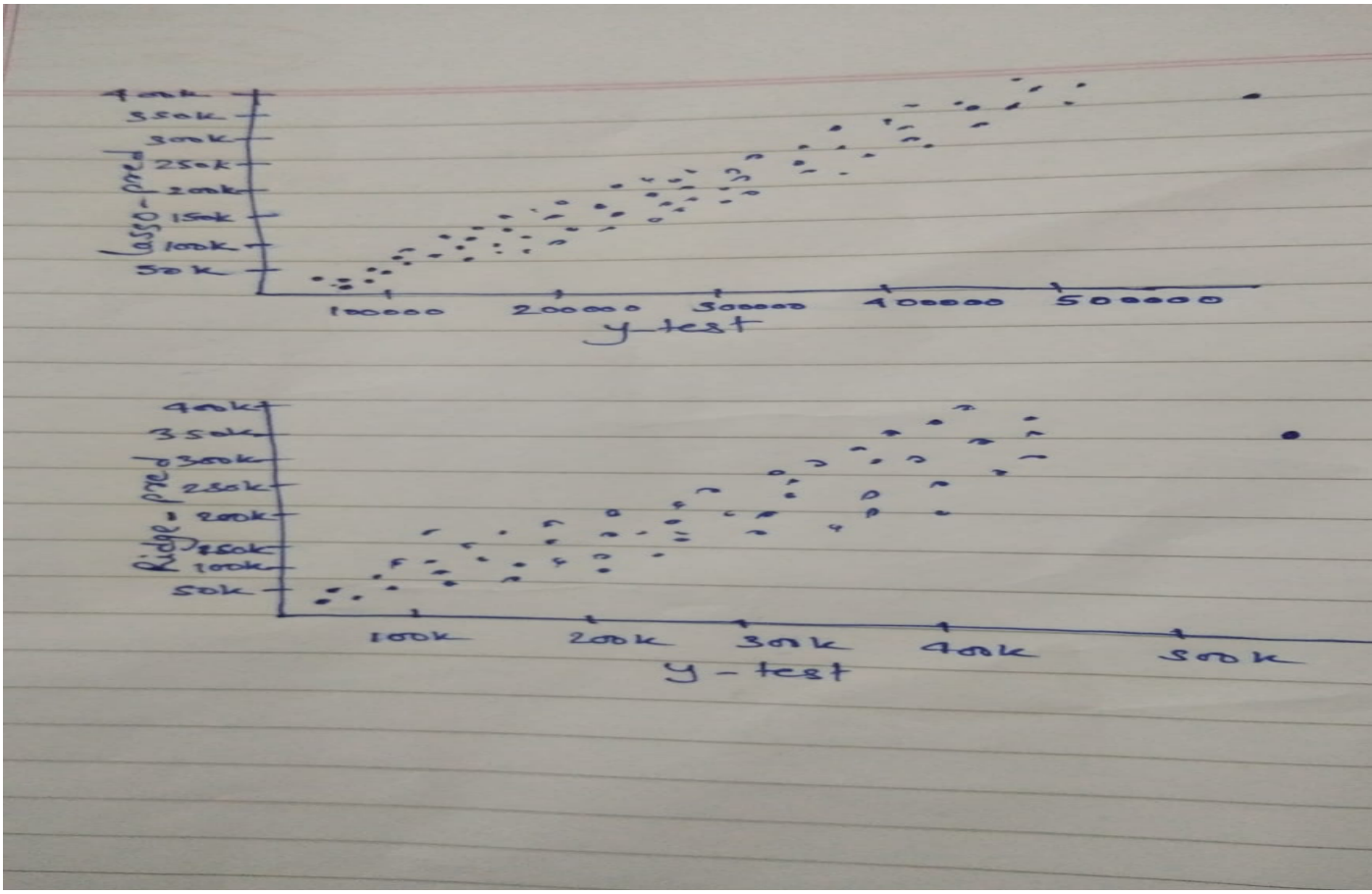
The most important predictor variable after the change is implemented are -

- LotArea
- OverallQual
- OverallCond
- YearBuilt
- BsmtFinSF1
- TotalBsmtSF
- GrLivArea
- TotRmsAbvGrd
- Street_Pave
- RoofMatl_Metal

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 - Even though the r2_score of Ridge is slightly higher for the test dataset, it's better to use Lasso as it assigns a zero value to insignificant features, thus helping us to choose the correct variables. Also its preferred to use simple yet robust model always.



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 – The five most important predictor variables are -

- 1stFlrSF
- GrLivArea
- Street_Pave
- RoofMatl_Metal
- RoofStyle_Shed

Question 4

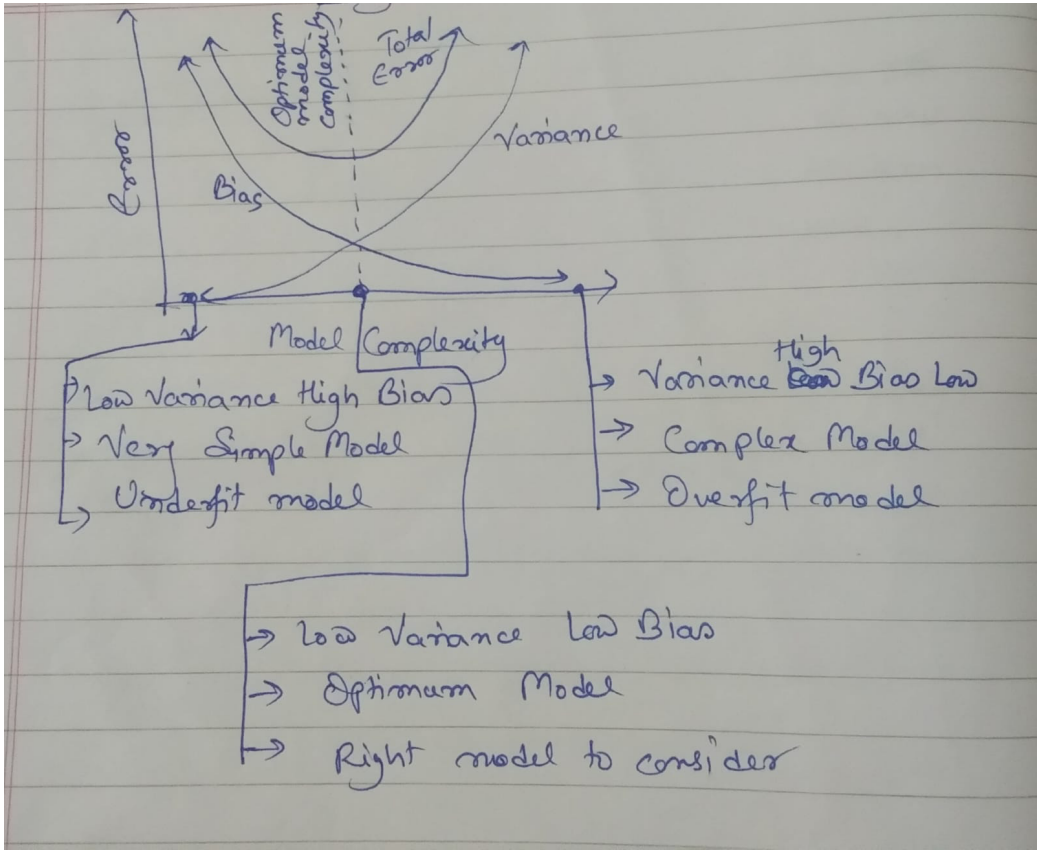
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer – The model needs to be as simple as possible. Even if the accuracy is compromised a little bit, it will be more robust & generalisable. This concept can be understood by the bias-variance trade off. Simpler model will have high bias , less variance & more generalisable . It's implication in terms of accuracy is ,that a robust & generalisable model will perform equally well on both training & test data sets.

Bias- Bias is error in model when it is weak in learning from data. High bias model performs poor on train & test datasets

Variance- Variance is error in model when it overlearn from the training data. High variance means model has memorised all data points in training dataset however performs low in test data set which is unseen.

A balance between the bias & variance is very much important to avoid overfit & underfit of the model. Hence a model with low bias & low variance is always a good preferred model.


Error
Optimum model complexity
Total Error
Variance
Bias
Model Complexity
Low Variance High Bias
→ Very Simple Model
→ Underfit model
Variance High Bias Low
→ Complex Model
→ Overfit model
→ Low Variance Low Bias
→ Optimum Model
→ Right model to consider