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:

According to my models, the optimal value for alpha (hyperparameter) were:

Ridge Regression : $\alpha = 70$
Lasso Regression : $\alpha = 0.001$

If I double the values of alpha for the same models; That is
Ridge Regression : $\alpha = 70 \times 2 =$ **140**
Lasso Regression : $\alpha = 0.001 \times 2 =$ **0.002**

Then model parameters will change as follows:

For Ridge Regression:

Metric	Set 1 ($\alpha = 70$)	Set 2 ($\alpha = 0.002$)
R-squared (train)	0.8594	0.8489
R-squared (test)	0.8531	0.8488
RSS (train)	142.18	152.85
RSS (test)	64.51	66.42
MSE (train)	0.1401	0.1506
MSE (test)	0.1479	0.1523

For Lasso Regression:

Metric	Set 1 ($\alpha = 70$)	Set 2 ($\alpha = 0.002$)
R-squared (train)	0.8934	0.8815
R-squared (test)	0.8510	0.8563
RSS (train)	107.85	119.82
RSS (test)	65.44	63.11
MSE (train)	0.1063	0.1180
MSE (test)	0.1501	0.1447

The change is very negligible, however you can notice a very minute decrease in r2 scores indicating further generalization.

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:

According to my models, both Ridge and Lasso regressions provided almost similar results. The difference is as follows:

Metric	Ridge	Lasso	Difference (Lasso - Ridge)
R-squared (train)	0.8594	0.8933	0.0339
R-squared (test)	0.8531	0.8510	-0.0021
RSS (train)	142.18	107.85	-34.33
RSS (test)	64.51	65.44	0.93
MSE (train)	0.1401	0.1063	-0.0338
MSE (test)	0.1479	0.1500	0.0021

Observations:

- Although Lasso has a higher training R-squared, the difference is small (0.0339).
- The test R-squared is marginally better for Ridge (-0.0021 difference).
- Lasso achieves a significant reduction in training RSS (-34.33) but slightly higher test RSS (0.93).
- Similar trends are observed with MSE, where Lasso has a lower training MSE (-0.0338) but slightly higher test MSE (0.0021).

Hence, **I would choose Lasso here (marginally)**. However, the reasons other than the observation above would be because:

1. **Prioritize generalizability:** Predicting on unseen data is crucial, Lasso might be a better choice due to its slightly lower test MSE and similar test R-squared compared to Ridge.
2. **Reduce model complexity:** Since interpretability and reducing model complexity are important, Lasso is preferable as it performs feature selection.

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 top 5 important variables based on lasso beta-coefficients are :

1. RoofMatl_WdShngl : 0.655780
2. Neighborhood_NridgHt : 0.499205
3. Neighborhood_NoRidge : 0.466566
4. Neighborhood_StoneBr : 0.381173
5. GrLivArea : 0.369273

The next 5 important features based on beta-coefficients are :

1. RoofMatl_CompShg9 : 0.345758
2. Exterior2nd_ImStucc : 0.299241
3. Neighborhood_Crawfor : 0.221197
4. OverallQual : 0.199574
5. SaleType_New : 0.182588

Hence the answer is : **RoofMatl_CompShg9, Exterior2nd_ImStucc, Neighborhood_Crawfor, OverallQual, SaleType_New**

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?

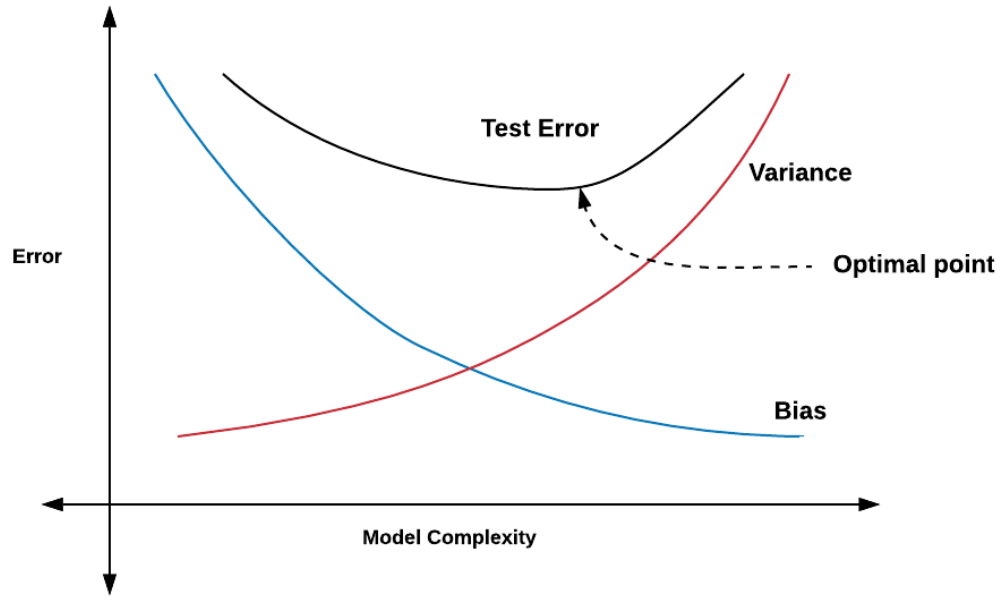
Answer:

There are many ways in which we can make sure the model is robust and generalisable:

1. **Data Quality and Diversity:** The quality of dataset plays a very important role. The collection of large diverse dataset can improve the model performance onto real-world unseen data.
2. **Regularisation:** We can use techniques like Lasso and Ridge to perform regularisation and reduce complexity of models, leading to simpler and more generalisable models.
3. **Cross-validation:** Evaluating your model's performance on unseen data using techniques like k-fold cross-validation or hold-out validation. This helps avoid overfitting and provides a more realistic estimate of how well your model will perform in practice.

Implication on Accuracy:

Robustness and generalizability come at the cost of accuracy. You are able to slightly sacrifice accuracy on the training data by simplifying your model and making it more focused on learning generalizable patterns.
Thus, this trade-off often leads to better performance on unseen data. Having high accuracy does not guarantee the same performance in real-world data. Hence, more generalised models tends to perform robustly in real-world scenarios.



As you can see from the above graph, test error(unseen data) increases with higher complexity. Hence, the Optimal model performance can be achieved by sacrificing accuracy (complexity) of the model.