Problem Statement - Part II

Subjective Questions Assignment:

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 1: For more details, please check the Python notebook.

Optimal value of alpha for Ridge and Lasso Regression is below:

- **Alpha value for Ridge Regression:** 1
- **Alpha value for Lasso Regression:** 0.0001

Increasing Alpha value of Ridge Regression from 1 to 2

**********Data after Ridge Regression Alpha Value = 2********** ********
Train R2 score: 0.897037606776905
Test R2 score: 0.8756015874789491
Train RSS score: 1.7479548545580181
Test RSS score: 0.952363033697311
Train MSE score: 0.0017120027958452675
Test MSE score: 0.00216939187630367
Train RMSE score: 0.04137635551671108
Test RMSE score: 0.04657673106073106 **********

Train R2 score reduced a little bit from 90.05% to 89.70% and the Test R2 score also reduced from 87.86% to 87.11%.

Increasing Alpha value of Lasso Regression from 0.0001 to 0.0002

**********Data after Lasso Regression with Value = 0.0002**********
Train R2 score: 0.8905875740768064
Test R2 score: 0.8765717458583819
Train RSS score: 1.8574546983093807
Test RSS score: 0.9449357445649256
Train MSE score: 0.0018192504390885217
Test MSE score: 0.0021524732222435662
Train RMSE score: 0.042652672121316404
Test RMSE score: 0.04639475425350981
****

Train R2 score reduced a little bit from 89.88% to 89.05% and Test R2 score also reduced from 88.20% to 87.65%.

Comparision metric of Ridge and Lasso regression with original and double the value:

SUBMITTED BY:
DEBASISH DEATY

Problem Statement - Part II

	Metrics	Ridge Regression_1	Lasso Regression_0.0001	Ridge Regression_2	Lasso Regression_0.0002
0	R2 Score (Train)	8.97e-01	8.99e-01	8.97e-01	8.91e-01
1	R2 Score (Test)	8.76e-01	8.82e-01	8.76e-01	8.77e-01
2	RSS (Train)	1.75e+00	1.72e+00	1.75e+00	1.86e+00
3	RSS (Test)	9.52e-01	9.03e-01	9.52e-01	9.45e-01
4	MSE (Train)	1.71e-03	1.68e-03	1.71e-03	1.82e-03
5	MSE (Test)	2.17e-03	2.06e-03	2.17e-03	2.15e-03
6	RMSE (Train)	4.14e-02	4.10e-02	4.14e-02	4.27e-02
7	RMSE (Test)	4.66e-02	4.54e-02	4.66e-02	4.64e-02

There is minor reduction in both train and test R2 when we double value of alpha

Below are the TOP 10 variables which are significant in predicting the Sale Price after doubling ALPHA values.

- GrLivArea: Above grade (ground) living area square feet.
- OverallQual: Rates the overall material and finish of the house.
- GarageCars: Size of garage in car capacity.
- OverallCond: Rates the overall condition of the house.
- FullBath: Full bathrooms above grade
- BedroomAbvGr: Bedrooms above grade (does NOT include basement bedrooms)
- MSZoning_RL: Identifies residential with Low Density zone.
- Neighborhood_NridgHt: hysical locations within Ames city limits Northridge Heights
- BsmtFullBath: Basement full bathrooms
- Neighborhood_Crawfor: Physical locations within Ames city limits is Crawford.

==========

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 2: For more details, please check the Python notebook.

The choice between Ridge and Lasso regression hinges on the specific context and modeling goals. If interpretability is paramount or feature selection is desired, Lasso regression with the determined optimal alpha might be preferable. However, if maintaining all features' influence while mitigating multicollinearity is crucial, Ridge regression with the determined optimal lambda is a better fit. Both techniques have their strengths: Ridge is robust against multicollinearity, while Lasso performs feature selection. Hence, the choice should align with the priorities: interpretability and feature selection (Lasso) or multicollinearity control (Ridge), determined by the dataset's characteristics and modeling objectives.

As per our House Price Data set prediction, the R2 Score of Lasso is better than Ridge for **Test Data**, so we will prefer to go for **Lasso regression**.

SUBMITTED BY:
DEBASISH DEATY

Problem Statement - Part II

	Metrics	Ridge Regression_1	Lasso Regression_0.0001	Ridge Regression_2	Lasso Regression_0.0002
0	R2 Score (Train)	8.97e-01	8.99e-01	8.97e-01	8.91e-01
1	R2 Score (Test)	8.76e-01	8.82e-01	8.76e-01	8.77e-01
2	RSS (Train)	1.75e+00	1.72e+00	1.75e+00	1.86e+00
3	RSS (Test)	9.52e-01	9.03e-01	9.52e-01	9.45e-01
4	MSE (Train)	1.71e-03	1.68e-03	1.71e-03	1.82e-03
5	MSE (Test)	2.17e-03	2.06e-03	2.17e-03	2.15e-03
6	RMSE (Train)	4.14e-02	4.10e-02	4.14e-02	4.27e-02
7	RMSE (Test)	4.66e-02	4.54e-02	4.66e-02	4.64e-02

==========

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 3: For more details, please check the Python notebook.

If the five most important predictor variables from the Lasso model are not available in the incoming data, the next five most important predictors should be identified to build a new model. Rebuilding the model without those unavailable variables necessitates reassessing feature importance. By rerunning the Lasso model on the available dataset and excluding the previously identified five variables, the new five most important predictors would emerge based on their non-zero coefficients. These newly identified predictors would hold significance in the updated model, contributing the most to the model's predictive capability in the absence of the previously unavailable variables.

As per our House Price Data set prediction, Below are the TOP 5 variables that are significant in predicting the Sale Price.

- GrLivArea: Above grade (ground) living area square feet.
- OverallQual: Rates the overall material and finish of the house.
- GarageCars: Size of garage in car capacity.
- OverallCond: Rates the overall condition of the house.
- BsmtFullBath: Basement full bathrooms

**********Data after Lasso Regression with Value = 0.0001 after removal top 5 variables**********
Train R2 score: 0.8420975808214957
Test R2 score: 0.802262126816478
Train RSS score: 2.680651561308232
Test RSS score: 1.513831543067687
Train MSE score: 0.0026255157309581115
Test MSE score: 0.0034483634238443896
Train RMSE score: 0.051239786601410736
Test RMSE score: 0.05872276750838971
****

Train data R2 score reduces drastically bit from 89.88% to 84.20% and Test R2 score also reduces from 88.20% to 80.22% after the removal of the Top 5 predictive variables.

SUBMITTED BY:

DEBASISH DEATY

Problem Statement - Part II

	Metrics	Lasso Regression_0.0001	Lasso Regression_drop_0.0001
0	R2 Score (Train)	8.99e-01	8.42e-01
1	R2 Score (Test)	8.82e-01	8.02e-01
2	RSS (Train)	1.72e+00	2.68e+00
3	RSS (Test)	9.03e-01	1.51e+00
4	MSE (Train)	1.68e-03	2.63e-03
5	MSE (Test)	2.06e-03	3.45e-03
6	RMSE (Train)	4.10e-02	5.12e-02
7	RMSE (Test)	4.54e-02	5.87e-02

Both Train and Test R2 score drastically decreases in Lasso Regression, after we remove the top 5 reductive variables.

Below are the TOP 5 variables that are significant in predicting the Sale Price after removing 5 predictive variables from the previous model.

- BedroomAbvGr: Bedrooms above grade (does NOT include basement bedrooms)
- LotArea: Lot size in square feet
- MSZoning_RL: Identifies residential with Low-Density zone.
- FullBath: Full bathrooms above grade
- BsmtFinSF1: Type 1 finished square fee

==========

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 4: For more details, please check the Python notebook.

Below are the pointers I learned from the course module lecture and note down as key points:

- To ensure model robustness and generalizability, utilize techniques such as train-test splitting, cross-validation, proper feature engineering, regularization, hyperparameter tuning, and out-of-sample validation. These practices help in assessing the model's performance on unseen data, minimizing overfitting, and enhancing its ability to generalize. The implications include a trade-off between bias and variance. Prioritizing robustness may slightly lower accuracy on the training set but significantly improves the model's ability to generalize to new data, ensuring more reliable real-world predictions.
- Robust refers to the model works for a broad range of inputs. If the model gets really good results at training time (it seems “more accurate”) but won’t generalize to out-of-sample data (i.e., it isn’t robust) then we call it overfitting.
- The model should be generalized so that the test accuracy is not less than the training score.
- Here in our case, based on all data and modeling both Ridge and Lasso performed well on Train and Test Data which shows our model with Alpha value "1" for Ridge and "0.0001" for Lasso is a Robust and more Generalized model.
    - Simpler models are more generic.
    - A simpler model requires fewer training samples.
    - A simpler model is more robust.

SUBMITTED BY:

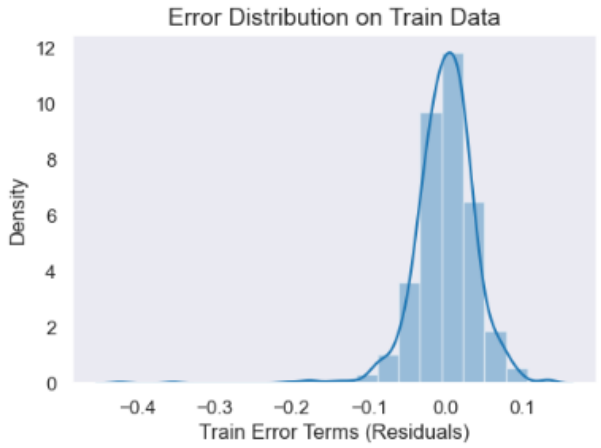
DEBASISH DEATY

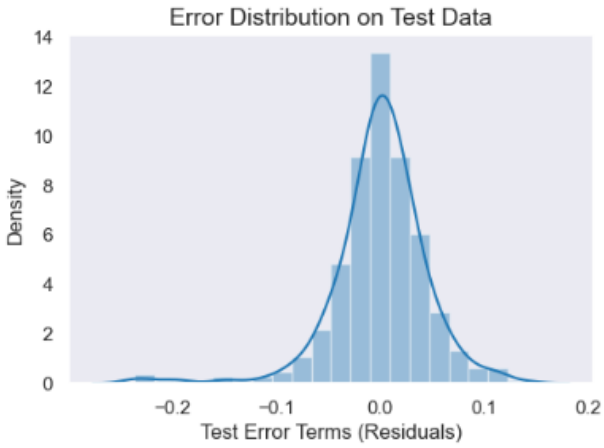
Problem Statement - Part II

- Too much importance should not be given to the outliers so that the accuracy predicted by the model is high. But outlier analysis needs to be done and only those which are relevant to the dataset need to be retained and the rest should be dropped.

If the accuracy of the Train and Test are the same, then that means the model is overfitted and it learned all the Train and Test data and the model is not robust and generalized. So, it will drastically fail and will not work on a broad range of unseen data.

Below you can see residual analysis on Train and test data:





SUBMITTED BY:
DEBASISH DEATY