ASSIGNMENT PART II

Deepak Mohanta

SLIDE SUMMARY

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?

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?

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?

QUESTION 4

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

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?

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?

- What is the optimal value of alpha for ridge and lasso regression?
  - The optimal value of alpha for ridge regression is 500
  - The optimal value of alpha for lasso regression is 1100
- What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?
  - Doubling the value of alpha in ridge and lasso regression will increase the regularization strength. Higher alpha values result in more aggressive shrinkage of the coefficients.
  - For Ridge Regression, all coefficients will be penalized more strongly, and the resulting coefficients will be closer to zero.
  - For Lasso Regression, doubling alpha will increase the sparsity of the model, potentially leading to more coefficients being exactly zero.
- What will be the most important predictor variables after the change is implemented?

Optimal Alpha

	Ridge	Lasso
GrLivArea	7068.770085	24920.916490
OverallQual_9	6947.658965	13321.842265
OverallQual_10	7025.585927	12425.540450
GarageCars	5744.227104	9615.639660
OverallQual_8	3328.503271	9105.895766
Neighborhood_NridgHt	5797.545657	6680.739387
Neighborhood_NoRidge	5478.928873	6092.675666
BsmtExposure_Gd	4286.753231	5261.195795
RoofMatl_WdShngl	4940.136618	5146.349344
SaleType_New	1592.450387	4611.439606
Neighborhood_Crawfor	2733.580629	3962.644007
FullBath	4297.676010	3926.074437
YearRemodAdd_2010	2968.939383	3758.345139
BsmtFullBath	2430.357947	3063.831383
BsmtFinType1_GLQ	2966.483511	3048.273789
LotArea	2805.091124	2944.455631
YearBuilt_1996	2288.580747	2792.073312
OverallQual_7	-300.271944	2491.758070
Functional_Typ	1376.588471	2399.087802

Observation/s

The top five predictors remained unchanged but the many variables become zero in Lasso regression and many approached zero in Ridge regression.

Double Alpha

	Ridge	Lasso
GrLivArea	6031.469555	24920.916490
OverallQual_9	5728.077417	13321.842265
OverallQual_10	5728.402204	12425.540450
GarageCars	4756.664721	9615.639660
OverallQual_8	2879.290205	9105.895766
Neighborhood_NridgHt	4970.050420	6680.739387
Neighborhood_NoRidge	4675.226969	6092.675666
BsmtExposure_Gd	3826.548809	5261.195795
RoofMatl_WdShngl	3653.279672	5146.349344
SaleType_New	1463.862274	4611.439606
Neighborhood_Crawfor	2119.386250	3962.644007
FullBath	3697.738503	3926.074437
YearRemodAdd_2010	2281.053470	3758.345139
BsmtFullBath	2055.997277	3063.831383
BsmtFinType1_GLQ	2817.801176	3048.273789
LotArea	2341.633771	2944.455631

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?

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?

- Factors to consider in making a choice:
  - **Feature Interpretability**: If you want a more interpretable model with a reduced set of important features, Lasso might be preferred due to its feature selection property.
  - **Predictive Performance**: If your primary goal is prediction, you might start with Ridge regression, especially if multicollinearity is a concern. Ridge can often lead to better predictive performance when many predictors contribute to the outcome.
  - **Trade-off Between Sparsity and Shrinkage**: If you are looking for a balance between sparsity (fewer predictors) and shrinkage (smaller coefficients), Elastic Net regression, which combines L1 and L2 penalties, might be a good compromise.
- The objective of our analysis to predict price of houses
  - Hence, I'll be using Ridge Regularization for our model and use this to predict Prices

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?

QUESTION 3

- After building the model, you realized 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?

- The top five predictors are
  - RoofMatl_CompShg
  - 2ndFlrSF
  - 1stFlrSF
  - RoofMatl_WdShngl
  - RoofMatl_Tar&Grv

	Ridge	Lasso
RoofMatl_CompShg	1530.559741	22671.234586
2ndFlrSF	5841.487830	22361.220562
1stFlrSF	7410.492356	18076.876463
RoofMatl_WdShngl	5624.202653	16664.606022
RoofMatl_Tar&Grv	206.226412	12671.230727
Neighborhood_NridgHt	7674.168055	10296.738676
RoofMatl_WdShake	923.931488	9428.785927
Neighborhood_NoRidge	6442.658418	8118.040898
BsmtExposure_Gd	5287.240137	6571.080206
TotalBsmtSF	5072.701005	6341.765252
FullBath	5416.427576	4881.667369
RoofMatl_Membran	437.040241	4752.618719
YearRemodAdd_2010	3157.603200	4360.429088
RoofMatl_Metal	462.458062	4325.656864
BsmtFinType1_GLQ	3554.941124	3970.943085
RoofMatl_Roll	-28.308085	3891.793927
Neighborhood_Somerst	1430.863769	3749.459843
MasVnrArea	5104.097916	3525.031093

QUESTION 4

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

QUESTION 4

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

- Ensuring that a model is robust and generalizable is crucial for its real-world applicability. Here are several key practices to make sure a model is robust and generalizable:
  - Cross-Validation, Train-Test Split, Holdout Validation Set, Feature Engineering, Regularization, Hyperparameter Tuning,
- Implications for Accuracy:
  - Training Accuracy vs. Testing Accuracy:
    - The training accuracy measures how well the model fits the training data, while testing accuracy assesses the model's performance on new, unseen data.
    - Ideally, the model should achieve high accuracy on both training and testing sets. If the training accuracy is much higher than the testing accuracy, it may indicate overfitting.
  - Overfitting and Underfitting:
    - Overfitting occurs when the model learns noise in the training data and performs poorly on new data. Underfitting occurs when the model is too simple to capture the underlying patterns. A balance between the two extremes is necessary for good generalization.
  - Robustness:
    - A robust model maintains consistent performance across different datasets and is less sensitive to variations in the input data.