

Advanced Regression – Assignment 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?

Answer 1:

The Optimal value of alpha for ridge = 2 and for lasso = 0.01

If you choose double the value of alpha for both Ridge and Lasso, the regularization strength will increase, leading to certain changes in the models. Specifically:

Ridge Regression:

Original Optimal Alpha: 2

New Alpha: $2 * 2 = 4$

The model will be more regularized.

The Ridge regularization term, which penalizes the sum of squared coefficients, will be more influential. The coefficients will be pushed more towards zero, but none will be exactly zero. The magnitudes of the coefficients will decrease compared to the original model.

Lasso Regression:

Original Optimal Alpha: 0.01

New Alpha: $0.01 * 2 = 0.02$

The model will be more regularized. The Lasso regularization term, which penalizes the absolute values of coefficients, will have a stronger effect. More coefficients will be driven exactly to zero.

The sparsity of the model will increase; fewer predictors will have non-zero coefficients.

In summary, doubling the value of alpha for both Ridge and Lasso will lead to more regularization, resulting in smaller coefficients and potentially sparser models. It's important to note that the specific impact depends on the dataset and the relationships within it. Regularization helps prevent overfitting and can be adjusted based on the trade-off between model simplicity and accuracy on new data.

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?

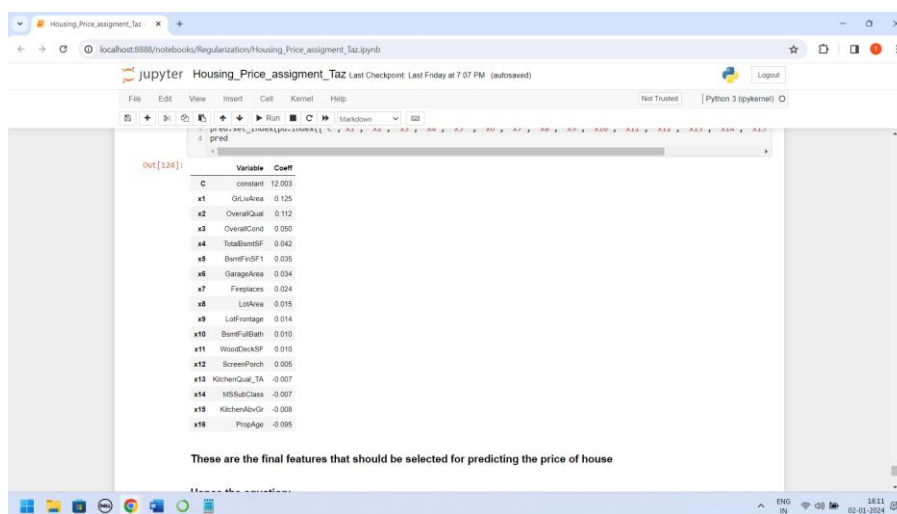
Answer 2:

Though the model performance by Ridge Regression was better in terms of R2 values of Train and Test, it is better to use Lasso, since it brings and assigns a zero value to insignificant features, enabling us to choose the predictive variables. It is always advisable to use simple yet robust model.

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:



	Variable	Coeff
0	constant	12.003
x1	GrLivArea	0.125
x2	OverallQual	0.112
x3	OverallCond	0.090
x4	TotalBsmntFt	0.042
x5	BsmntFndFl	0.035
x6	GarageArea	0.034
x7	Fireplaces	0.024
x8	LotArea	0.015
x9	LotFrontage	0.014
x10	BsmntFullBath	0.010
x11	WoodDeckSF	0.010
x12	ScreenPorch	0.005
x13	KitchenQual_TA	-0.007
x14	MSSubClass	-0.007
x15	KitchenAbvGr	-0.008
x16	ProptAge	-0.005

These are the final features that should be selected for predicting the price of house

The above 15 are key predictor variables, if we drop the top 5 variables then Fireplaces, Lotarea, LotFrontage, basementfullbath become the next important predictor variables. We will have to validate this by creating a new model.

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:

Some of the key practices to ensure a model is robust and generalisable are as follows:

- 1) Cross Validation
- 2) Train-Test Split
- 3) Feature Scaling
- 4) Feature Engineering
- 5) Regularization
- 6) Hyperparameter Tuning
- 7) Ensemble Methods like Random Forest or Gradient Boosting

Implications for Accuracy of the model are as follows:

- 1) Overfitting Vs Underfitting
- 2) Bias-Variance Trade-off

In summary, robust and generalizable models are essential for real-world applications. Practices such as cross-validation, appropriate data preprocessing, regularization, and hyperparameter tuning contribute to building models that perform well on new, unseen data. The goal is to create models that capture underlying patterns without being overly influenced by noise in the training data.