ADVANCED REGRESSION SUBJECTIVE ANSWERS

SUBMITTED BY : ANUSHKUMAR K

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:

The optimal value of alpha for ridge regression is **4** and the optimum value of alpha for lasso regression is **0.001**.

If we double up the value of alpha to 8 and 0.002 in Ridge and Lasso regression,

We could observe the following:

- The lasso regression is reducing the coefficients of many non significant features to 0.
- Thus Lasso regression does the feature selection automatically eliminating the non significant variables or features.
- The Ridge regression also shrinks the coefficient values of many features to nearly 0, but is resulting in the building of a model which is lesser complex.
- When the coefficient of the features is reducing towards 0, then variance is lowered and error value is lowered correspondingly.

Important predictor variables are remaining mostly the same such as:

- BedroomAbvGr
- BsmtUnfSF
- BsmtFinSF2
- MiscVal
- ExterCond
- BsmtFinSF1
- LotArea
- LotShape
- BsmtCond

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:

Based on the above models on Lasso and Ridge, we could observe that the R2 values of both regression models correspond to 0.844 and 0.85 respectively.

But considering the fact that the Lasso regression model penalizes more on the dataset for variables and it also helps in feature elimination. Thus number of variables is reduced and an efficient model can be built for the dataset.

So Lasso regression is preferred over the Ridge regression.

Test error and training error both are estimated to find the simplicity and fitting of the model. There are few metrics such as AIC, BIC, Adjusted_R2, Mallow's Cp based on which we can select the best suitable model which is fitting and simple.

We can calculate the above metrics as,

Mallow's Cp:

$C_p = 1/n(RSS + 2d\sigma^2)$

AIC:

AIC=$1/\mathrm{n}\sigma^2(\mathrm{RSS} + 2d\sigma^2)$

BIC:

BIC=$1/\mathrm{n}(\mathrm{RSS} + \ln(\mathrm{n})d\sigma^2)$

Adjusted r2:

Adjusted r2 =

$$\mathbf{1} - \left(\frac{\frac{\boldsymbol{RSS}}{\boldsymbol{n-d-1}}}{\frac{\boldsymbol{TSS}}{\boldsymbol{n-1}}}\right)$$

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?

ANSWER:

Initially we add the dummy variables to our dataset for building a model. Those dummy variables are meaningful features which may or may not be significant for the model.

After dummy variables are created, we drop the original variables from the dataset. This causes a lot of changes to the feature selection where a non-significant variable can become the most significant variable. Instead of dropping the variable which has the corresponding dummy variable all those variables will be taken into account for model feature selection.

From the model built in the assignment the below are most significant variables from lasso regression,

- BedroomAbvGr
- BsmtUnfSF
- BsmtFinSF2
- MiscVal
- ExterCond
- BsmtFinSF1
- LotArea
- LotShape
- BsmtCond

As per the current feature selection all the above variables are available in the dataset, but since we are neglecting the top five variables as per the question, the most significant variables would be as below,

- BsmtFinSF1
- LotArea
- LotShape
- BsmtCond
- GarageArea
- HalfBath

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:

- When the model created is fitting for different datasets and is yielding better results , then such model is said to be generalization. So if any model can work the similar way of effectiveness is called generalizable
- The model which can withstand any value for error terms and error complexity in the dataset is said to be robustness of a model.
- There is a bias-variance trade off which we need to identify in a model evaluation. When the model is consistent then the variance is lower and bias is higher.
- Hence normally when considering a model, we check for its generalizability and robustness.
- Usually the best model is the one which balances both such as lower variance and higher bias without compromising on the accuracy of the model.
- Some of the methods to check for the robustness and generalizability are
    - Hold-out Strategy
    - Cross validation
- **So a model which has higher bias is more accurate, but the model which has higher variance is more generalized model. Hence the bias-variance trade-off.**