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:

The optimal value of alpha for Lasso Regression is 0.001 and for Ridge Regression is 10.

The metrics for Lasso Regression for **alpha = 0.001** are:

- R-Squared of train data is: 0.9277
- R-Squared of test data is: 0.8950
- RMSE is: 0.01801

The metrics for Lasso Regression for **alpha = 0.002** are:

- R-Squared of train data is: 0.9106
- R-Squared of test data is: 0.8870
- RMSE is: 0.0193

With double the 'alpha' value, it is observed that for Lasso Regression, the R-squared for both training and test data decreased, and RMSE increased by very small value.

The metrics for Ridge Regression for **alpha = 10** are:

- R-Squared of train data is: 0.93822
- R-Squared of test data is: 0.8913
- RMSE is: 0.01865

The metrics for Ridge Regression for **alpha = 20** are:

- R-Squared of train data is: 0.9326
- R-Squared of test data is: 0.8918
- RMSE is: 0.0185

With double the 'alpha' value, it is observed that for Ridge Regression, the R-squared for both training data slightly decreased whereas for test data, there is hardly any change in the value. RMSE also has negligible difference.

After the ‘alpha’ value is doubled, the top 10 important predictor variables are:

Lasso Regression:

Feature	Coefficients
Total_Home_Quality	0.101673
TotalSF	0.097918
Neighborhood_Crawfor	0.077755
TotalBsmtSF	0.042896
Neighborhood_NridgHt	0.041101
SaleType_New	0.034671
GarageArea	0.034442
Condition1_Norm	0.029699
Foundation_PConc	0.029689
BsmtFinSF1	0.027474

Ridge Regression:

Feature	Coefficients
Total_Home_Quality	0.085125
Neighborhood_Crawfor	0.080956
Neighborhood_StoneBr	0.067701
Neighborhood_NridgHt	0.052687
SaleCondition_Normal	0.050375
TotalSF	0.048511
Condition1_Norm	0.046079
Functional_Typ	0.044309
Condition2_Norm	0.044308
Exterior1st_BrkFace	0.038804

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:

Both Lasso and Ridge Regression are powerful Regularization techniques used in the industry; and the choice of model depends on what is to be achieved as per the business problem. For example, there are 1000+ variables in the data, we can go for Lasso Regression because it has the capability to identify the insignificant features (by making some of the coefficients zero), thereby reducing the number of variables which are significant in predicting the dependent variable. If the aim of the exercise is to just reduce the coefficients of the features and not reduce the number of features, we can use Ridge Regression.

In this ‘SalePrice’ prediction exercise, as there are many columns (~250, considering the encoded categorical variables), it is better to identify the top significant features and limit the model building using those features and try to achieve the nearest possible R-Squared and Adjusted R-Squared values obtained when all variables are used in model building. For example, if we get an R-Squared which is 0.95 and Adjusted R-Squared which is 0.94 using all ~250 features, we need to identify the number of significant variables (For example: 50,100,150 etc. in various iterations) and check if the R-Squared and Adjusted R-Squared are close to 0.95 and 0.94. As Lasso Regression has the capability of penalizing the coefficients of insignificant variables and can make them zero, we can use Lasso to identify the significant variables for predicting 'SalePrice'.

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 5 most important predictor variables in Lasso model with optimum alpha value are:

- Neighborhood_Crawfor
- TotalSF
- Total_Home_Quality
- Neighborhood_StoneBr
- SaleType_New

After removing these features from the data and then training the Lasso model again, the 5 most important predictor variables are:

- 2ndFlrSF
- 1stFlrSF
- Neighborhood_NridgHt
- Functional_Typ
- Neighborhood_Somerst

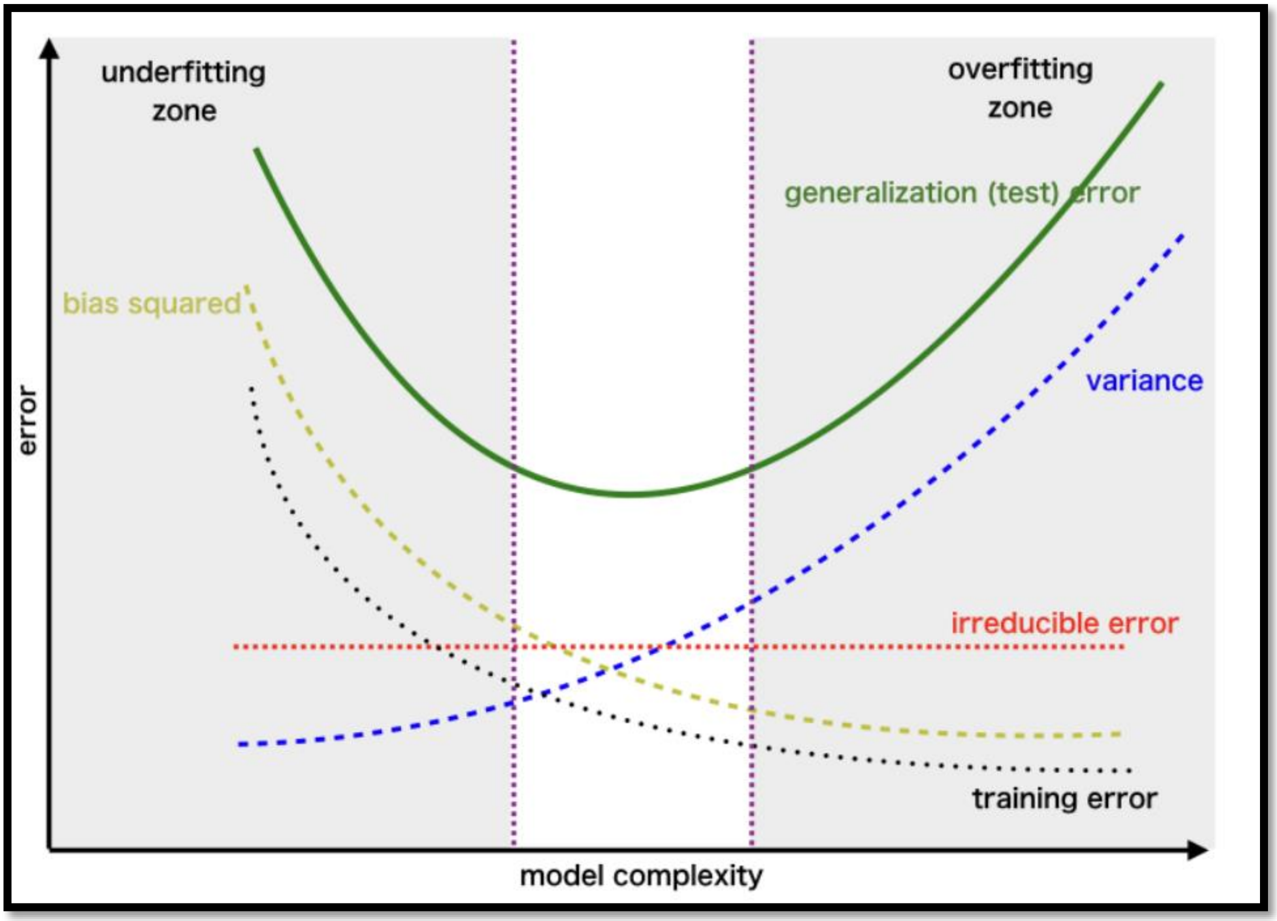
Question 4

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:

Robustness or generalizability is the property that signifies how well the model is adapting to new or unseen data. It can also be defined as the change in model's performance with unseen data, as compared to that of the training data. This usually happens when the model is overfit (also learns from the noise in the data). To avoid overfitting, there are many techniques which can be used. Regularization techniques, Hyperparameter tuning, RFE (Recursive Feature Elimination) etc. are some of them.

Implications for accuracy: In order to make sure that our model is robust and generalizable, we need to make sure it doesn't overfit or underfit. If the model is simple, chances are high that the accuracy is less, and the model is biased. If the model is complex, it may overfit and will be more sensitive to unseen data, that is the variance will be high. So, it is important to make a balance between the bias and variance of a model.



The image below shows how the error varies with respect to the model complexity. We can infer from the above image that the variance increases as bias decreases and vice-versa. So, we need to identify the optimal point or the sweet spot where both bias and variance are close to each other.