ASSIGNMENT: ADVANCED REGRESSION 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?

Solution 1

- The optimal value of alpha are as follows :-
    - For Ridge Regression – 1.0
    - For Lasso Regression – 0.0001

- If we choose to double the value of alpha, the following changes are observed for both Ridge and Lasso regression models :-
    - The value of Mean Squared Error increases.
    - The value of coefficients of features decrease.
    - The model accuracy decreases.

- The most important predictor variable after change in implemented is :-
    - For Ridge Regression – **OverallQual** (Rates the overall material and finish of the house)
    - For Lasso Regression – **GrLivArea** (Above grade (ground) living area square feet)

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?

Solution 2

I will choose to apply Lasso regression due to the following factors :-

- The Mean Squared Error of the Lasso model is less than the Ridge Model.
- The accuracy on Training and Test dataset of the Lasso model is higher than the Ridge Model.
- The Lasso model has the added advantage that it removed features which are not significant and therefore reduces the complexity of the model. It is thus likely to be less complex model than the Ridge Regression model, and thus more robust and generalizable.

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?

Solution 3

The five most important predictor variables in the Lasso Regression Model were as follows :-

- GrLivArea : Above grade (ground) living area square feet
- OverallQual : Rates the overall material and finish of the house
- OverallCond : Rates the overall condition of the house
- MSZoning : Identifies the general zoning classification of the sale.
    - More specifically (as MSZoning Feature is split into Dummy variables)
        - MSZoning_FV (Floating Village Residential)
        - MSZoning_RL (Residential Low Density)
        - MSZoning_RH (Residential High Density)
- TotRmsAbvGrd : Total rooms above grade (does not include bathrooms)

On excluding the above predictor variables, the accuracy of the model reduced as expected. The five most important predictor variables are now :-

- **1stFlrSF** : First Floor square feet
- **2ndFlrSF** : Second floor square feet
- **GarageArea** : Original construction date
- **FullBath** : Full bathrooms above grade
- **Fireplaces** : Number of fireplaces

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?

Solution 4

A model is considered "**Robust**" if the output variable of the model is consistently accurate in case of any drastic and unexpected change in the input feature variables.

A "**Generalisable**" model is one which continues to provide accurate results even when it encounters data which it has not seen during training.

Creating a robust and generalizable model involves trade-offs by the data scientist based on the business objective. If a model is very accurate, there is a danger of overfitting in which case the model will not generalize well to new data. Therefore, due to the Bias-Variance tradeoff, it is not possible to have a model which has both low bias and low variance.

There are many techniques available which can be used to make sure that a model is robust and generalizable. A few of these are as follows :-

- **Reduce Model Complexity** Reduce the complexity of the model by removing input features which do not have a significant impact on the output variable.

- **L1 and L2 regularization** Use regularization methods like Ridge, Lasso and Elasticnet.

- **Performance Metrics** Use metrics like “Adjusted R Squared”, AIC, BIC, AUC under ROC curve, etc to evaluate the model performance and remove features from the training dataset if required to reduce model complexity.

- **Cross Validation** Use k-fold Cross-validation to evaluate the model. A robust and stable model would have similar performance for different folds.

- **Remove Multicollinearity** Identify multicollinearity and remove predictor variables which are correlated. You may either keep the variable which explains the change or create a new derived variable as a new feature.

- **Train Test Data Split** Ensure that the Training dataset and Validation dataset are not similar and are diverse.

Implications on Accuracy. The implication of making a model more Generalisable and Robust is that the accuracy is likely to reduce. The accuracy of the model may be lower. This is a trade-off that the data scientist has to make based on how much accuracy is acceptable for the business case.