Advanced Regression Assignment

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 to 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 2.0 and the optimal value of alpha for lasso regression is 0.0006.

When we double the value of alpha for ridge, the value increases and hence the reduces the co-efficient value towards zero. This will lead the model to underfitting. Here the variance reduces with a slight compromise in terms of bias.

When we double the value of alpha for lasso regression, the alpha value decreases and hence this leads the model to overfitting. Hence we have low bias and high variance.

The important predictor variables remained same even after the change is implemented. Only the coefficients value changed**.**

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: Lasso regression is my preferred method over Ridge regression. Using a R2 score of 2 for ridge regression and 0.0006 for lasso regression, the R2 score for train and test data is good for lasso regression. There is a comparably small error deviation. Though lasso regression shows a zero co-efficient, in this assignment most of the predictor variables had non-zero values.

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:

The important predictor variables for lasso regression are

- SaleCondition_Partial
- SaleCondition_Others
- SaleCondition_Normal
- GarageFinish_Unf

- 'GarageFinish_RFn

After removing these predictor variables, the next most important predictor variables are as follows:

- 'GarageFinish_No Garage'
- 'GarageType_Others'
- 'GarageType_No Garage'
- 'GarageType_Detchd'
- 'GarageType_BuiltIn'

Question 4:

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

We need to consider the following things to make a model robust and generalizable.

Presence of data

For better model building, we need more data. Rather than relying on assumptions and weak correlations, more data allows for more accurate and reliable models.

Imputation of missing and Outlier values

The presence of missing and outlier values in the data reduces the accuracy of a model. It leads to inaccurate predictions and bias in the model. So, it is important to treat missing and outlier values.

1. Missing: In case of continuous variables, you can impute the missing values with mean, median, mode. For categorical variables, you can treat variables as a separate class or have the value either mode or most occurring value.
2. Outlier: You can delete the observations, perform transformation, binning and Imputation.

Feature Engineering

As a result of this step, more information can be extracted from existing data. This information is in the form of derived features. These features may be able to explain a higher percentage of the variance in the training data. As a result, the model will be more accurate.

In feature engineering, hypotheses generation is crucial. A good hypothesis leads to a good feature. Data can be normalized or regularized, skewness can be removed from data, and data can be discretized. Furthermore, we can derive new features and predict the results. By doing so, we can uncover the hidden relationships within a dataset. This will help improve the model.

Feature Selection

It is the process of identifying which subset of attributes best explains the relationship between independent variables and the target variable. Based on various metrics, Domain Knowledge, Statistical Parameters, Hyperparameters, and Visualization are all useful features

Multiple algorithms

Choosing the right machine learning algorithm is the best way to achieve higher accuracy. There are algorithms that are better suited to certain types of data sets than others. Hence, we should apply all relevant models and assess their performance.

Algorithm Tuning

Machine learning algorithms are driven by parameters. The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model.

Cross Validation

In data modeling, cross validation is one of the most important concepts. Try to leave a sample on which you do not train the model and test it on this sample before finalizing the model.