Surprise Housing: House Price Prediction Case Study

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: By theory, the optimal lambda value comes out to be 0.001 and will be used to build the ridge regression model. For regression, we plot the curve between negative mean absolute error and alpha to see how the value of alpha increases from 0 as the error term decrease. For our dataset, the train error is showing increasing trend when value of alpha increases.

Best Estimator function was used to find alpha value for Ridge and Lasso regression which is 10 and 0.0005 respectively.

When we double the value of alpha, for our ridge regression with alpha = 10, the model will apply more penalty on the curve and try to make the model more generalized that is making model simpler.

Same way for lasso regression, with increased alpha value, the model will be panelised more and coefficient of the variable will reduced to zero which in turn decreases the value of RMSE.

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: Regularization refers to techniques that are used to calibrate machine learning models to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

So it is important to regularize coefficients to improve the prediction accuracy ,decrease in variance to make the model interpretably easier.

In Ridge regression, we use a tuning parameter called lambda (λ) as the penalty which is square of magnitude of coefficients. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

Lambda value is identified by cross validation. With regularization. The RMSE decreases as we apply penalty with lambda on the coefficients that have greater values.

In Lasso regression we use a tuning parameter called lambda (λ) as the penalty which is absolute value of magnitude of coefficients. As lambda increases, shrinkage occurs so that variables that are at zero can be thrown away.

Lambda value is identified by cross validation for lasso regression as well. Lasso also does variable selection to identify which variable needs to be shrunk to 0.

For our dataset, the MSE for Lasso Regress is lesser than Ridge Regression so we will choose lasso as regularization method.

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: Those 5 most important predictor variables that will be excluded are: -

- GrLivArea
- OverallQual
- OverallCond
- TotalBsmtSF
- GarageArea

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?

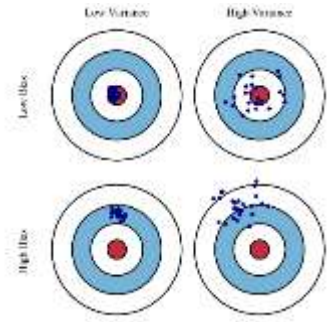
Answer:

A machine learning model is called robust and generalizable if they are not impacted by outliers in during model training on training data. A model should be generalisable so that the test accuracy is not lesser than the training score. A machine learning model should be able to perform well on unseen data with similar accuracy as on training data.

Since outliers are unavoidable in real time data, the outlier's analysis and treatment is needed to ensure the accuracy predicted by the model is good. To ensure this, unnecessary erratic data points need to be removed. This would help standardize the predictions, making it robust.

Simpler models are the best models, even though this impacts accuracy, but it will be more robust and generalisable.

During model evaluation, we need to focus on the **Bias-Variance trade-off** concepts.



Bias is the difference between the Predicted Value and the Expected Value. To explain further, the model makes certain assumptions when it trains on the data provided. When it is introduced to the testing/validation data, these assumptions may not always be correct.

Contrary to bias, the **Variance** is when the model takes into account the fluctuations in the data i.e. the noise as well. So, what happens when our model has a high variance?

The model will still consider the variance as something to learn from. That is, the model learns too much from the training data, so much so, that when confronted with new (testing) data, it is unable to predict accurately based on it.

The simpler the model, the higher the bias but less variance and more generalizable. A robust and generalisable model will perform equally well on both training and test data i.e. the accuracy does not change much for training and test data.

It is important to have balance in Bias and Variance to avoid overfitting and under-fitting of data.