Housing Price Prediction Exercise

Problem Statement – Part II

Q1. 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?

Ans. After applying the Ridge and Lasso regression model independently, using Grid Search, cross validation has been applied using multiple values of Hyperparameter (lambda / alpha). This has been done in both the cases i.e., Ridge and Lasso.

In order to determine the best fit value of alpha, plot has been drawn between "Negative Mean Error" and "alpha" to see the trend for Test and Train score.

In case of Ridge and Lasso, alpha equals 500 and 0.03 have been selected respectively as for these values, r2 score for Test and Train data have been good and close to each other. As well as, value of Mean Square Error (MSE) has also been quite low for both regression methods.

If the value obtained for alpha get doubled then this will increase the weightage of penalty term in the model equation and model will become more simpler than the previous one and it can be generalized in the case that it can be applied over different dataset in which better outcomes can be expected. In this case, below are the important predictor variables which got affected when this change has been implemented –

- OverallQual: Rates the overall material and finish of the house
- 1stFlrSF: First Floor square feet
- GarageCars: Size of garage in car capacity
- 2ndFlrSF: Second floor square feet
- TotalBsmtSF: Total square feet of basement area
- OverallCond: Rates the overall condition of the house
- CentralAir: Central air conditioning
- LotArea: Lot size in square feet
- BsmtFullBath: Basement full bathrooms
- Foundation_PConc: Foundation: Type of foundation - PConc Poured Contrete

Q2. 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?

Ans. Once the optimal value of hyperparameter, lambda has been identified using Grid Search and cross validation for both Ridge and Lasso regression, r2 score for Test and Train data and Mean Square Error (MSE) for test data is calculated.

Once these values are obtained, among both models, one is selected for which **MSE score is lowest** and there **r2 score for Test and Train data should NOT has much difference.**

This consideration is important to take care of as MSE value shows slippage of predicted values from the actual values whereas low difference of r2 score for test and train data shows that model is neither overfitted nor underfitted i.e., model will give fair results when applied over other unseen test data.

Q3. 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?

Ans. Below are the five most important variables:

- OverallQual: Rates the overall material and finish of the house
- GarageCars: Size of garage in car capacity
- TotalBsmtSF: Total square feet of basement area
- OverallCond: Rates the overall condition of the house
- LotArea: Lot size in square feet

Q4. 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?

Ans. For a model has to be robust and generalizable, it is important that model should be kept Simple and not very much Complex. Here comes the concept of Bias – Variance trade off which has to be taken care off while building a model.

In case of a very simple model, Bias will be high which means that error made of model on both Test and Train data will be high. Variance will be low which means model performance will not be much impacted when put to perform on other unseen data.

Whereas, in case of Complex models, Bias is Low but Variance is High which means that model will perform very well on train data but will fail miserably on unseen test data. In this case, r2 score for Train data will be very high but for test data, r2 score will be low. This shows that model is overfitted and thus it will not give desired results for test data.

Thus, using r2-score, bias variance trade-off, it is required to fine tune model performance so that model can be robust and generalizable. This is also required in term of accuracy also, as simple and robust model accuracy will not vary much in case of test and train data whereas in case of complex model, accuracy can be very high in case of train data and it can be equally poor when case of unseen test data would come as complex models tend to leant the train data point by heart.