Name :- Chetan D. Wargantiwar

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:

- **Ridge** is **6**, with Mean Train Score of **-10784.713** & Mean Test Score of **-13966.979**
- **Lasso** is **100**, with Mean Train Score of **-11326.558** & Mean Test Score of **-13835.103**

After choosing double the value of Alpha for:

- **Ridge** is **12**, with Mean Train Score of **-11456.825** & Mean Test Score of **-14186.962**
- **Lasso** is **200**, with Mean Train Score of **-12283.738** & Mean Test Score of **-14190.755**

The 10 most important predictor variables after the change is implemented for Ridge:

- OverallQual (when: 9, 6, 4)
- Neighborhood (when: StoneBr, Crawfor)
- Functional (Typ)
- BsmtQual (Gd)
- KitchenQual (Gd)
- BsmtExposure (Gd)
- SaleCondition (Partial)

The 10 most important predictor variables after the change is implemented for Lasso:

- OverallQual (9, 8,
- Neighborhood (StoneBr,Crawfor, NridgHt)
- Functional (Typ)
- SaleCondition (Partial)
- BsmtQual (Gd)
- BsmtExposure (Gd)
- MSSubClass (160)

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 :-

The optimal lambda value for:

- Ridge - 6
- Lasso - 100

The Mean Squared error for:

- Ridge has Mean Test Score of **-13966.979**
- Lasso has Mean Test Score of **-13835.103**
- The Mean Squared Error of Lasso is slightly lower than that of Ridge that's why we choose to apply Lasso.
- Lasso helps in feature reduction as the coefficient value of least important features became 0.

That's how Lasso has a better edge over Ridge. Therefore, the variables predicted by Lasso can be applied to choose significant variables for predicting the price of a house.

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 :-

After building the model, as we realised that the five most important predictor variables in the lasso model are not available in the incoming data. So, we will now have to create another model excluding the five most important predictor variables. Therefore the five most important predictor variables now are:-

1) MSZoning
2) MSSubClass
3) KitchenQual
4) BsmtFinType1
5) OverallCond

Question 4

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

Answer :-

In machine learning, generalization usually refers to the ability of an algorithm to be effective across a range of inputs and applications and for making your model more robust to outliers following ways can be adopted:

1) We can use a model that's resistant to outliers. Like, Tree-based models which are generally not affected by outliers, while regression-based models are affected by outliers.
2) By Switching from mean squared error to mean absolute difference reduces the influence of outliers3.)
By Capping the data at some threshold.
4) If data has a right tail we can try a log transformation. We can try such other transformation techniques as well. To make them simple to understand by the model.

The implications of the same for the accuracy of the model

1) The model needs to be made robust and generalizable so that they are not impacted by outliers in thetraining data.
2) The model should also be generalizable so that the test accuracy is not lesser than the training score.
3) The model should be accurate for datasets other than the ones which were used during training.
4) Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high.
5) To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained.
6) Those outliers which does not make sense must be removed from the dataset.

This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust , it cannot be trusted for predictive analysis.