Subjective Type Question - Answers

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 1 **Ridge Regression:** The optimal value of alpha for the best model was **"4".** When we have doubled the value of alpha from 4 to 8, there was not much significant effect on model's performance. However, more number of variables having their coefficient tend towards the 0 making the model more regularized. Top 5 variables for the both the Ridge model were OverallQual, GrLivArea, OverallCond, TtlRmsAbvGrd, 1stFlrSF. However, rank of TtlRmsAbvGrd, 1stFlrSF was only interchanged for Both the models

Lasso Regression: The optimal value of alpha for the best model was **"0.001".** When we have doubled the value of alpha from 0.001 for best model to 0.002, there was not much significant effect on performance of the model. However, initially before doubling the value of alpha we had 70 features in the model, which decreased to 54 features when value of alpha was doubled. Model becomes more regularized. Top 5 variables for the best Lasso model were GrLivArea, OverallQual, OverallCond, GarageCars and TtlRmsAbvGrd. However, there was slight change in Top 5 variables when value of alpha was doubled. Top 5 variables after doubling the value of alpha were GrLivArea, OverallQual, GarageCars, TotRmsAbvGrd and OverallCond.

All the variables for both type of Lasso and Ridge Models were showing **positive relation** with the **SalePrice**.

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 2 **Ridge Regression** – Optimal value of alpha coming out to be **"4".**
Lasso Regression – Optimal value of alpha coming out to be **"0.001"**
We have chosen an evaluation metrics of Root Mean Squared Error. Ridge regression having RMSE value of 0.12 and 0.14 for train and test sets respectively and Lasso Regression is having an RMSE value of 0.13 and 0.14 for train and test sets respectively. Although Ridge Regression is performing slightly better but Lasso Regression is performing Feature Selection by making the coefficient of certain number features to 0. This will make the model much leaner.
In practical, it will be much easier to target less number of features. Therefore, in our case Lasso Regression model will be chosen as the final Model.

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 3 Although we were getting best Lasso model with value of alpha as **0.001.** But as our final Model we have chosen the value of alpha as **0.002** because as we have seen that when we double the value of alpha there was not much significant effect on model's performance, however, 24 number of features were reduced.
Top 5 features for the final Model were:

- GrLivArea
- OverallQual
- GarageCars
- TotRmsAbvGrd
- OverallCond

All Above features are positively related with the SalePrice

After realizing that incoming data is not having top 5 features. We have a built another model without the above features and it was observed that final model was having 60 features. There was increase in total number of features. However, Model's performance was slightly compromised. The new Top 5 features were:

- 1stFlrSF
- 2ndFlrSF
- GarageArea
- FullBath
- Age

All features were positively related with Target variable SalePrice except for Age which was negatively related with the SalePrice.

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 4 Model tries to learn all the data points provided to it. Which makes model more complex but with high accuracy. As a result, model will have high variance and low bias. This might result in model performing bad in the real world or on unseen data.

To tackle this problem, we can make the Model more robust and generalizable by reducing the variance or complexity of the model. By reducing the variance of the model, the model will become more stable and small changes in the data will not have much effect on the model.

By reducing the variance/complexity of the model, the **accuracy** of the Model **decreases** as model becomes more generalizable. Therefore, to achieve the best performance we have to tune the hyper parameters of the model in such a way to achieve a point where model have less complexity as well as less bias. This will make model more generalizable with good performance on training as well as testing dataset.