Advanced Regression – Problem Statement Part-II

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 and lasso regression

Ridge Alpha 20

Lasso Alpha 0.001

Ridge Regression Results

```
# Model Evaluation
print("Ridge Regression with ",model_ridge_cv.best_params_)
print("====================================")
print('R2 score (train) : ',r2_score(y_train,y_train_pred))
print('R2 score (test) : ',r2_score(y_test,y_test_pred))
print('RMSE (train) : ', np.sqrt(mean_squared_error(y_train, y_train_pred)))
print('RMSE (test) : ', np.sqrt(mean_squared_error(y_test, y_test_pred)))
```

```
Ridge Regression with  {'alpha': 20}
====================================
R2 score (train) :  0.919237364415656
R2 score (test) :  0.8737208706744777
RMSE (train) :  0.1112925161856442
RMSE (test) :  0.15192810165635734
```

Lasso Regression Results

```
# Model Evaluation
print("Lasso Regression with ",model_lasso_cv.best_params_)
print("====================================")
print('R2 score (train) : ',r2_score(y_train,y_train_pred))
print('R2 score (test) : ',r2_score(y_test,y_test_pred))
print('RMSE (train) : ', np.sqrt(mean_squared_error(y_train, y_train_pred)))
print('RMSE (test) : ', np.sqrt(mean_squared_error(y_test, y_test_pred)))
```

```
Lasso Regression with  {'alpha': 0.001}
====================================
R2 score (train) :  0.9179125447640548
R2 score (test) :  0.8752467036915912
RMSE (train) :  0.1122016170566611
RMSE (test) :  0.15100743709430986
```

When we double the value of alpha for both Ridge and Lasso regression, below values are obtained.

Ridge Alpha 40, Lasso Alpha 0.002

Ridge Regression with alpha = 40	
R2 Score for Train	0.9184
R2 Score for Test	0.8748

Lasso Regression with alpha = 0.002	
R2 Score for Train	0.9161
R2 Score for Test	0.8764

If we increase alpha for Lasso Regression, R2 Score on Train decreases slightly but R2 Score on Test increases slightly.
If we increase alpha for Ridge Regression, R2 Score on Train decreases slightly but R2 Score on Test increases slightly.

The most **important predictor** variables after the change is implemented:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. YearBuilt

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?

It is important to regularize coefficients and improve the prediction accuracy also with the decrease in variance and making the model interpretably.
Ridge regression has a particular advantage over OLS when the OLS estimates have high variance, i.e., when they overfit. Regularization can significantly reduce model variance while not increasing bias much. The tuning parameter lambda helps us determine how much we wish to regularize the model.

The higher the value of lambda, the lower the value of the model coefficients, and more is the regularization. Choosing the right lambda is crucial so as to reduce only the variance in the model, without compromising much on identifying the underlying patterns, i.e., the bias.
Ridge regression does have one obvious disadvantage. It would include all the predictors in the final model.

This may not affect the accuracy of the predictions but can make model interpretation challenging when the number of predictors is very large.
The behavior of Lasso regression is similar to that of Ridge regression. With an increase in the value of lambda, variance reduces with a slight compromise in terms of bias. Lasso also pushes the model coefficients towards 0 in order to handle high variance, just like Ridge regression.

But, in addition to this,
Lasso also pushes some coefficients to be exactly 0 and thus performs variable selection.
This variable selection results in models that are easier to interpret.
Hence, I will choose Lasso Regression in our assignment for final modelling.

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?

Earlier, the five most important predictor variables were:

1. 1stFlrSF
2. 2ndFlrSF
3. OverallQual
4. OverallCond
5. YearBuilt

After excluding the above variables, the five most important predictor variables:

1. BsmtFinSF1
2. TotRmsAbvGrd
3. BsmtUnfSF
4. FullBath
5. LotArea

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?

To make model robust and generalisable 3 features are required:

1. Model accuracy should be > 70-75%: In our case its coming 88%(Train) and 87%(Test) which is fine.
2. P-value of all the features is < 0.05
3. VIF of all the features are < 5 Thus we are sure that model is robust and generalisable.

Implications of Accuracy of a model:

1. **Gain more data as much as you can:** Having more data allows the data to train itself, instead of depending on the weak correlations and assumption, it is good to have more data.
2. **Fix missing values and outliers:** If the data has missing values and outliers it can lead to inaccurate model. Outliers can affect the mean, median that we are imputing to

continuous variables You can get the outlier values using a boxplot, treating the outliers in the data will make our mode more accurate.

3. **Featuring Engineering or newly derived columns/Standardize the values:** We can extract the new data from the existing data.

   Ex: from DOB we can get the Age of the person, after extracting the new data required we can drop the existing features.

4. **Scaling the values:** Ex: One value is in meters, the other is Kilo meters, it is important to scale these features into one standardized unit. If we do this we can get accurate model.
5. **Feature Selection:** It is purely based on the domain knowledge, so that we can select important features that have good impact on the target variable. Data visualization also helps in selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.
6. **Applying the right algorithm:** Choosing the right machine learning algorithm is very important to get accurate model. This will come with experience
7. **Cross validation:** Sometimes more accuracy will cause overfitting, then we can use cross validation technique, i.e. leave a sample on which you do not train the model & test the model on this sample before going to the final model.