

Subjective questions Advanced Regression:

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 Regression is 1 and for Lasso Regression is 0.001.

Below are the R-squared values of the train and test datasets, RMSE values for the Ridge and Lasso Regression for the optimum alpha and when the alpha is doubled (scenario 2). In the below screenshot, scenario 2 refers to the results fetched after the alpha values of Ridge and Lasso regressions are doubled:

```
R-squared value of train data for Ridge: 0.957908126441241
R-squared value of test data for Ridge: 0.8997785009758426
RMSE of Ridge Regression is: 0.014609513745403948
```

```
R-squared value of train data for Ridge scenario2: 0.954525309030501
R-squared value of test data for Ridge scenario2: 0.8986029906331785
RMSE of Ridge Regression for scenario2 is: 0.014780870536873162
```

```
R-squared value of train data for Lasso: 0.9379007267523319
R-squared value of test data for Lasso: 0.8872467463634593
RMSE for Lasso Regression is: 0.016436295853497457
```

```
R-squared value of train data for Lasso scenario2: 0.9195454984361487
R-squared value of test data for Lasso scenario2: 0.8696001408935818
RMSE for Lasso Regression scenario2 is: 0.019008681296562458
```

We can observe that the R-square values have dropped and the RMSE values have increased after the alpha value is doubled. We can also see a drop in the coefficients for the two models after the alpha value is doubled. The higher the value of alpha, the lower the value of the model coefficients and more is the regularization.

The most important predictor variables after the changes were implemented are:

For Ridge Regression:

'GrLivArea', '1stFlrSF', 'LotArea', 'BsmtFinSF1', 'OverallQual_9', 'TotalBsmtSF'

For Lasso Regression:

'GrLivArea', 'TotalBsmtSF', 'BsmtFinSF1', 'OverallQual_9', 'LotArea', 'GarageArea'

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 value of lambda for ridge regression is 1 and for lasso regression is 0.001.

Below are the R-squared value of train/test data, RMSE values for Ridge and Lasso Regression:

R-squared value of train data for Ridge: 0.957908126441241

R-squared value of test data for Ridge: 0.8997785009758426

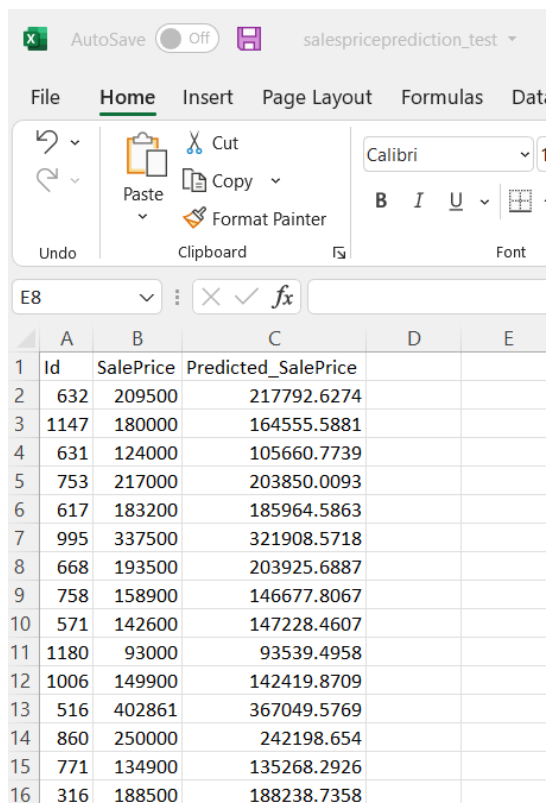
RMSE of Ridge Regression is: 0.014609513745403948

R-squared value of train data for Lasso: 0.9379007267523319

R-squared value of test data for Lasso: 0.8872467463634593

RMSE for Lasso Regression is: 0.016436295853497457

We would choose Ridge Regression since the R-squared value is high and the RMSE value is lower comparatively. We have generated a new excel with the actual and the predicted Sales Prices for the test dataset using the results of the Ridge regression. Below excel screenshot shows the data of the predicted Sales prices:



The screenshot shows an Excel spreadsheet titled 'salespriceprediction_test'. The 'Home' tab is active, displaying the 'Clipboard' and 'Font' sections of the ribbon. The formula bar shows 'E8'. The spreadsheet contains a table with 16 rows of data. The columns are labeled 'Id', 'SalePrice', and 'Predicted_SalePrice'. The 'Predicted_SalePrice' column contains numerical values with two decimal places, representing the model's predictions for each row's 'SalePrice'.

	A	B	C	D	E
1	Id	SalePrice	Predicted_SalePrice		
2	632	209500	217792.6274		
3	1147	180000	164555.5881		
4	631	124000	105660.7739		
5	753	217000	203850.0093		
6	617	183200	185964.5863		
7	995	337500	321908.5718		
8	668	193500	203925.6887		
9	758	158900	146677.8067		
10	571	142600	147228.4607		
11	1180	93000	93539.4958		
12	1006	149900	142419.8709		
13	516	402861	367049.5769		
14	860	250000	242198.654		
15	771	134900	135268.2926		
16	316	188500	188238.7358		

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: The **initial five most important predictor variables** for the optimum value of $\alpha = 0.001$ are

'GrLivArea', 'TotalBsmtSF', 'OverallQual_9', 'LotArea', 'BsmtFinSF1'

Above initial important predictor variables are excluded for building the new model, the new model has generated the same optimum α value of 0.001:

```
L3_X_train = L3_X_train.drop(['GrLivArea', 'TotalBsmtSF', 'OverallQual_9', 'LotArea', 'BsmtFinSF1'], axis=1)
L3_X_test = L3_X_test.drop(['GrLivArea', 'TotalBsmtSF', 'OverallQual_9', 'LotArea', 'BsmtFinSF1'], axis=1)
```

For Lasso Regression after the initial five predictor variables are excluded below are the R-squared value of train/test data, RMSE values, there is a small reduction in the R-squared value and a slight increase in the RMSE value:

R-squared value of train data for Lasso scenario 3: 0.9285228992960912

R-squared value of test data for Lasso scenario 3: 0.8816911456687796

RMSE for Lasso Regression scenario 3 is: 0.017246148285392695

The **new five most important predictor variables** after excluding the initial 5 predictor variables are

'1stFlrSF', '2ndFlrSF', 'GarageArea', 'BsmtQual', 'Neighborhood_StoneBr'

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 model has to be made robust and generalisable so that there is no impact of outliers and the test data accuracy and score is maintained well. The model has to be designed in such a way that it will perform well on the unseen data similar to the best fit model designed using the training dataset. Outlier treatment has to be taken care at the initial stage of data preparation and only the relevant data has to be included for the model building. Too many records also should not be eliminated as apart of outlier treatment and missing values. The missing values can be imputed so that we don't lose out on the relevant records in the process. This would help in increasing the accuracy of the predictions.

The accuracy of the regression model is not used for predicting the exact value but to know how close the predictions were to the expected values. One of the simplest methods for calculating the correctness of a model is to use the error between predicted value and actual value.

The below metrics can be used to understand the accuracy of a model:

1. Mean Absolute Error (MAE)
2. Root Mean Squared Error (RMSE)
3. Mean Absolute Percentage Error (MAPE)
4. R-Squared Score

We have used RMSE and R-square in the assignment to understand the accuracy of the model

Formula for RMSE:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

R-square formula: It helps in determining how well the model captures the variance in data.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \epsilon_i^2$.