SUBMITTED BY DR RANJEET SINGH MAHLA
SUBMITTED ON FEBRUARY 23, 2023
COURSE MSC IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Problem Statement - Part II

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?

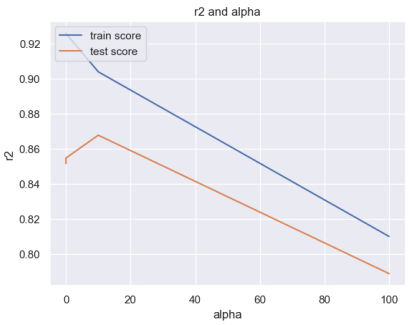
(a) What is the optimal value of alpha for ridge and lasso regression?

Ridge regression: optimum alpha 10
Top 5 predictors

```
('GarageFinish_No Garage', 0.048),
 ('GarageFinish_RFn', 0.052),
 ('GarageFinish_Unf', 0.052),
 ('SaleCondition_Normal', 0.052),
 ('SaleCondition_Others', 0.057)
```

Ridge R2 values

```
- Train Score: 0.9048825126844456
- Test Score: 0.8762875068524612
```



Lasso regression: optimum alpha 0.00006
Top 5 predictors

```
('GarageFinish_No Garage', 0.051),
 ('GarageFinish_RFn', 0.066),
 ('GarageFinish_Unf', 0.072),
 ('SaleCondition_Normal', 0.074),
 ('SaleCondition_Others', 0.302)
```

Lasso R2 score
- Train R2 Score: 0.9178195810624109
- Test R2 Score: 0.883834546544181

(b) ***What will be the changes in the model if you choose double the value of alpha for both ridge and lasso?***

Answer: The R2 value decreased further, and the top 5 predictors, as well for Ridge

```
- Train R2 Score: 0.8927204503770251
 - Test R2 Score: 0.8643973267373375
```

for Lasso

```
- Train R2 Score: 0.912810049272631
 - Test R2 Score: 0.8834132924110816
```

(c) What will be the most important predictor variables after the change is implemented?

Answer: The top 5 predictors remain the same, but their value decreases further as we increase the alpha value

Top 5 predictors for Ridge

```
('GarageFinish_No Garage', 0.048),
 ('GarageFinish_RFn', 0.052),
 ('GarageFinish_Unf', 0.052),
 ('SaleCondition_Normal', 0.052),
 ('SaleCondition_Others', 0.057)
```

Top 5 predictors for Lasso

```
('GarageFinish_No Garage', 0.051),
 ('GarageFinish_RFn', 0.066),
 ('GarageFinish_Unf', 0.072),
 ('SaleCondition_Normal', 0.074),
 ('SaleCondition_Others', 0.302)
```

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 regression model was created with Ridge and Lasso. The R2 value is not different (a little high for Lasso). It is observed Lasso will penalize the features more and be helpful in features selection. Some coefficients in Lasso are absolute zero with reduced computation by reducing the number of features. Considering the above three points, I would select Lasso.

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:
Following 5 Important predictors variables after removing the top 5 important predictors variables

For Ridge at alpha =10

```
('KitchenQual_TA', 0.034),
 ('GarageType_BuiltIn', 0.041),
 ('GarageType_Detchd', 0.043),
 ('GarageType_No Garage', 0.044),
 ('GarageType_Others', 0.045)
```

For Lasso at alpha =0.00006

```
('KitchenQual_TA', 0.04),
 ('GarageType_BuiltIn', 0.041),
 ('GarageType_Detchd', 0.045),
 ('GarageType_No Garage', 0.045),
 ('GarageType_Others', 0.051)
```

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: I have done the following to make the model robust and generalizable

- Removed skewed columns and processed them to normalize
- NA/null/missing values were replaced with medians for columns having outliers and mean for columns with normal distribution
- The variables with less correlation to the target variable were removed
- The best model from a 5-fold cross-validate the model using Ridge and Lasso regression
- The mean test train scores with alpha show a similar curve for train and test data.
- Also, the R2 (regression scores) are pretty close and are acceptable to consider this model a robust and generalizable model.
- I could observe a decent score for Ridge and Lasso Regression which are as follows-

Ridge Regression Score:

- Train Score: 0.9048825126844456

- Test Score: 0.8762875068524612

Lasso Regression Score:

- Train Score: 0.9178195810624109

- Test Score: 0.883834546544181

The R2 value of test and train are close, which indicate this model would work fine for incoming unknown data.

FEATURE OF A ROBUST MODEL

- Model is resistant to outliers: Outliers should not impact outcomes. A box plot can detect the outliers.
- Model is significant: can be determined by R2; always, a simple model is robust.

Creating a robust model

- Data: the amount of training data is always good for model robustness.
- Fix missing values: replace the missing value with median/mean or remove the variables with too many missing values.
- Data transformation: Remove data skew and remove outliers.

- Feature selection: The selection of variables and features can be made with statistical analysis.
- Algorithm: selection of the correct algorithm is essential (such as a regression model)
- Validation: the model should be randomly built against available data by test train split.