**Name :- Rajat Mandaniyan**

**EmailD :-** rajatkajob@gmail.com

**Batch MLC-46 Batch ( AI and ML )**

**Assignment-based Subjective**

## 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:-**

**Optimal value of alpha:**

**For Ridge regression :1.0**

**For Lasso Regression :0.0001**

**if you choose double the value of alpha for Ridge Regression :-**

```
Ridge Regression train r2: 0.9150662492961033
Ridge Regression test r2: 0.7371351945366085
```

**if you choose double the value of alpha for Lasso Regression :-**

```
Lasso Regression train r2: 0.9163334784634766
Lasso Regression test r2: 0.7350240220130699
```

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

| | Features | Coefficient | Mod |
|---|---|---|---|
| **0** | LotFrontage | 10.116167 | 10.116167 |
| **3** | OverallCond | 0.65504 | 0.65504 |
| **13** | BsmtFinSF2 | 0.426172 | 0.426172 |
| **10** | BsmtFinType1 | 0.36311 | 0.36311 |
| **11** | BsmtFinSF1 | 0.339782 | 0.339782 |
| **27** | KitchenAbvGr | 0.331734 | 0.331734 |
| **2** | OverallQual | 0.326209 | 0.326209 |
| **61** | MSZoning_FV | -0.294202 | 0.294202 |
| **8** | BsmtCond | 0.273548 | 0.273548 |
| **29** | TotRmsAbvGrd | 0.268702 | 0.268702 |

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

| | Feature | Coef | mod |
|---|---|---|---|
| 0 | LotFrontage | 10.042593 | 10.042593 |
| 13 | BsmtFullBath | 0.859076 | 0.859076 |
| 3 | OverallCond | 0.702908 | 0.702908 |
| 61 | Exterior1st_BrkFace | -0.554576 | 0.554576 |
| 2 | OverallQual | 0.483327 | 0.483327 |
| 27 | MSZoning_RH | 0.421919 | 0.421919 |
| 8 | CentralAir | 0.406363 | 0.406363 |
| 29 | MSZoning_RM | 0.345059 | 0.345059 |
| 30 | Street_Pave | 0.279916 | 0.279916 |
| 18 | GarageArea | 0.247055 | 0.247055 |

## 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:-**

Lasso regression would be a better option it would help in feature elimination and the model will be more robust.

Because § In the ridge, the coefficients of the linear transformation are normal distributed and in the lasso they are Laplace distributed.

In the lasso, this makes it easier for the coefficients to be zero and therefore easier to eliminate some of your input variable as not contributing to the output.

§ Ridge regression can't zero out coefficients; thus, you either end up including all the coefficients in the model, or none of them.

In contrast, the LASSO does both parameter shrinkage and variable selection automatically.

Lasso regression can produce many solutions to the same problem. § Ridge regression can only produce one solution to one problem.

## 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:-**

I have excluded the five most important variable I have got prior. Those are

| |
|---|
| LotFrontage |
| BsmtFullBath |
| OverallCond |
| Exterior1st_BrkFace |
| OverallQual |

I have created a new model after removing these columns code is mentioned the Python notebook. After the Lass Regression I have got the other important predictors are :-

| |
|---|
| MSZoning_RH |
| CentralAir |
| MSZoning_RM |
| Street_Pave |
| GarageArea |

## 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:-**

**The mode is robust and generalizable when:-**

1. Test accuracy is not much lesser than the training score

2. The model should not impacted by the outliers: Outlier treatment is most important to get the robust model. We can detect outliers in the dataset using box plots, Z score etc. Treating the outliers will not affect mean, median etc. so that we can impute correct values to missing values. , the outlier analysis needs to be done and only those which are relevant. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis

3. The predicted variables should be significant. **Model significance can be determined the P-values, R2 and adjusted R2.**

**Always a simple model can be more robust Implications of Accuracy of a model:**

1. **Gain the more data as much 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 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. Scaling the values: ex: one value is in meters, the other is Kilo meters, it is important to scale these feature into one standardized unit. If we did this we can get accurate model.

4. **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 the selecting the features. Statistical parameters like p-Values, VIF can give us significant variables.

5. **Applying the right algorithm** Choosing the right machine learning algorithm is very important to get accurate model. This will comes with experience

6. **Cross validation:** Some times 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 got to the final model.