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 10.**

The optimal value of alpha for **Lasso regression is 0.01**.

If we choose double the value of alpha for both ridge and lasso regression,

a. The **coeffiecients** will **change** for both regressions and **adjusted r-square** will **decrease**, which means that the **model fit** has not improved.

b.**Mean squared error** will **increase** in both the regression models.

Mean squared error is the average of the square of the errors. The larger the number the larger the error. Error in this case means the difference between the observed values y1, y2, y3, … and the predicted ones pred(y1), pred(y2), pred(y3),

Mean squared error is a sum of the variance of an estimator and its bias.

Now a large mean squared error indicates the variance or the bias in your estimator is large.

The most important **feature variables** after we **double** the value of alpha are:

a. **OverallQual**

b. **GrLivArea**

c. **AgeOfHouse**

d. **GarageCars**

e. **MSZoning_RL**

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?

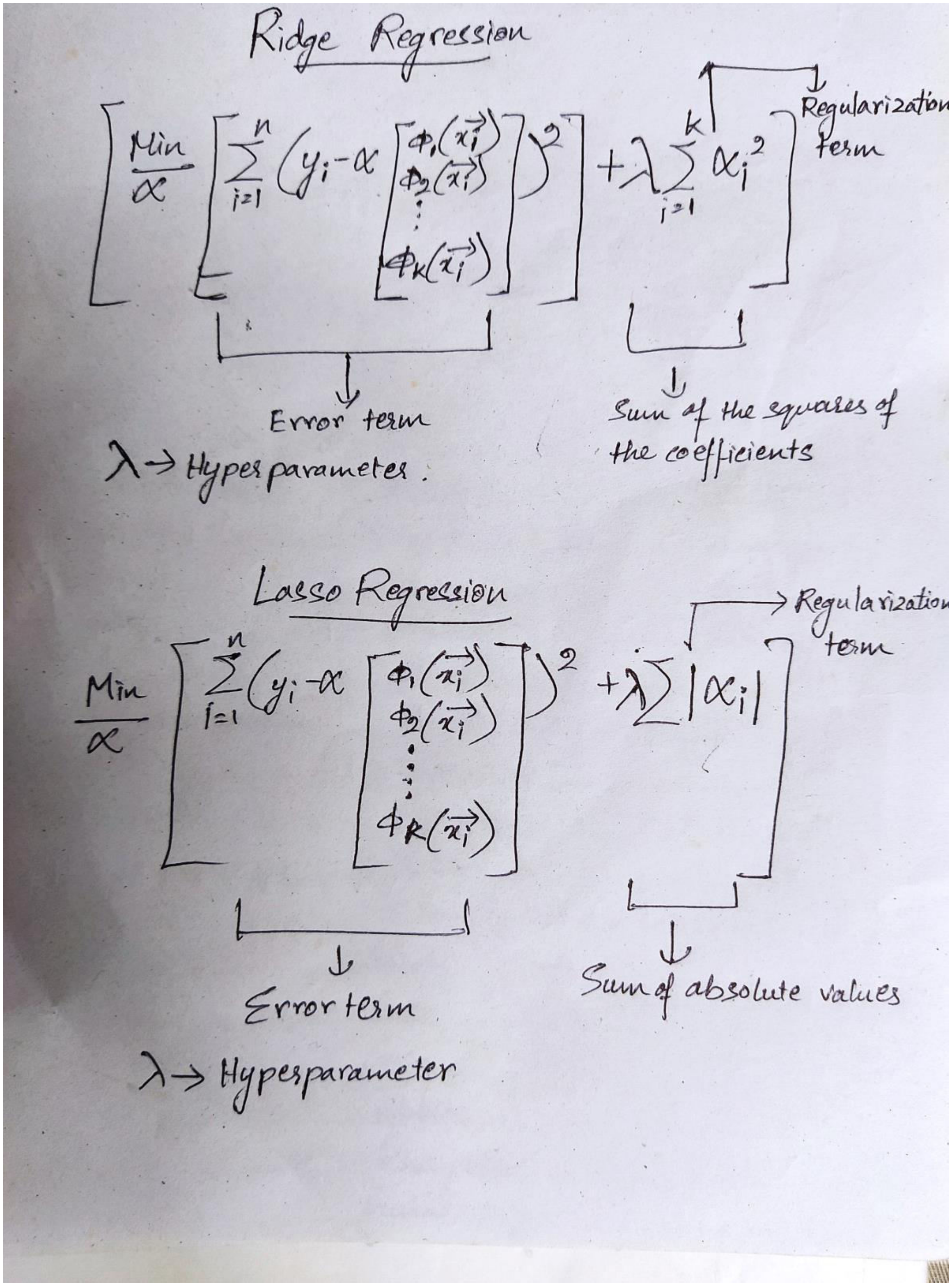
Answer.

Ridge Resgression

In Ridge regression, we add a penalty term which is equal to the square of the coefficient. The L2 term is equal to the square of the magnitude of the coefficients. We also add a coefficient lambda to control that penalty term. In this case if lambda is zero then the equation is the basic OLS else if lambda > 0 then it will add a constraint to the coefficient. As we increase the value of lambda this constraint causes the value of the coefficient to tend towards zero. This leads to both low variance (as some coefficient leads to negligible effect on prediction) and low bias (minimization of coefficient reduce the dependency of prediction on a particular variable).

Lasso Regression

It adds penalty term to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from 0 this term penalizes, cause model, to decrease the value of coefficients in order to reduce loss. The difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.



Ridge Regression

$$\left[\frac{\text{Min}}{\alpha}\ \underbrace{\left[\sum_{i=1}^{n}\left(y_i-\alpha\begin{bmatrix}\phi_1(\vec{x_i})\\ \phi_2(\vec{x_i})\\ \vdots\\ \phi_k(\vec{x_i})\end{bmatrix}\right)^2\right]}_{\text{Error term}}+\overbrace{\underbrace{\lambda\sum_{i=1}^{k}\alpha_i^2}_{\text{Sum of the squares of the coefficients}}}^{\text{Regularization term}}\right]$$

λ → Hyperparameter.

Lasso Regression

$$\frac{\text{Min}}{\alpha}\left[\underbrace{\sum_{i=1}^{n}\left(y_i-\alpha\begin{bmatrix}\phi_1(\vec{x_i})\\ \phi_2(\vec{x_i})\\ \vdots\\ \phi_k(\vec{x_i})\end{bmatrix}\right)^2}_{\text{Error term}}+\overbrace{\underbrace{\lambda\sum|\alpha_i|}_{\text{Sum of absolute values}}}^{\text{Regularization term}}\right]$$

λ → Hyperparameter

Limitation of Ridge Regression is that it **decreases the complexity** of a model but **does not reduce the number of variables** since it **never** leads to a **coefficient been zero** rather only minimizes it. Hence, this model is **not good for feature reduction**.

In our assignment, our **goal** is to **choose the variables** are **significant** in **predicting the price** of a house, and how well those variables describe the price of a house. This is **not possible** in **Ridge Regression**. Only by applying **Lasso regression**, we **can choose the value of alpha** such that the **least important co-efficients become 0** and we are able to **choose** the **significant variable**. So in this assignment we will **choose Lasso Regression** to predict **the feature variables** and the fit of the model.

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:

After creating the original model, the five most important variables are:

a. **OverallCond**

b. **GarageCars**

c. **MSZoning_RL**

d. **GarageType_Attchd**

e. **BsmtFinType1_Unf**

Now, if in the incoming data, the **above variables are missing**, and i need to create a new model with the rest of the feature, then the most **important variables in the new model** will be:

a. **OverallQual**

b. **GrLivArea**

c. **MSZoning_RM**

d. **BsmtExposure_Gd**

e. **AgeOfHouse**

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** is considered to be **robust** if its **output and forecasts** are consistently **accurate** even if **one or more of the input variables** or **assumptions** are drastically **changed** due to **unforeseen circumstances.**

To make our model **robust and generalisable**, the following changes have been made to our model:

a. **Transform the data**- Check the **target** variable and check whether it is **normally distributed** or **skewed**.It it is skewed, **remove the skewness** by **treating the data** either by log transformation or any other method.

b**. Remove the outliers**- **Extreme values** can be present in both dependent & independent variables, in the case of supervised learning methods. These outliers can easily be **found out** by **plotting** them on a **boxplot** or by checking the the data at various **percentile**.

The **same factors** can be considered also for the **accuracy** of the model.Besides that, having a **high(~80%) adjusted R-square** show the accuracy of the model. But in some cases, we should **consider robustness of the model before accuracy**, especially when the data is **skewed**.

Suppose we have a data where only 1% of population are suffering from a disease. In such a cse, if we consider accuracy as a metric to evaluate our model, it will be wrong because the model will be 99% accurate.In such cases we need to heck the robustness of the model.