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: 9. **Optimal value of alpha for Lasso Regression**: 0.001

If we chose double the value of alpha:

Ridge Regression: The coefficient values of columns increase.

Lasso Regression: There is not much change in the coefficient of the columns if the alpha value is doubled.

Predictor Variables after the change is implemented:

When Alpha is NOT doubled:

Comparison between lasso and ridge regression model:

- Negative affect variables have remained the same overall. The only changes are: Neighborhood_MeadowV (Lasso) & SaleCondition_Partial (Ridge)

- There are changes in positive affect variables (top5- as per co-efficient):

Lasso variables (Positive):

- Fireplaces: Number of Fireplaces

- Exterior1st_BrkFace: Exterior covering on the house (Brick Face)

- BsmtExposure: Refers to walkout or garden level walls

- Neighborhood_BrkSide: Physical Location with Ames City Limits (Brookside)

- TotRmsAbvGrd: Total Rooms Above Grade

Ridge variables (Positive):

- BsmtFullBath: Basement Full bathrooms

- LotArea: Lot size in Square feet

- TotRmsAbvGrd: Total Rooms Above Grade

- 2ndFlrSF: 2nd Floor Square feet

- MSZoning_RH: General zoning classification of the sales(Residential High density)

When Alpha is doubled:

Comparison between Lasso and Ridge Regression Model:

Negative affect variables have remained the same overall. The only changes are: Neighborhood_MeadowV (Lasso) & SaleCondition_Partial (Ridge)

Lasso Variables (Positive):

MSZoning_RL: General zoning classification of sale (Residential Low density)

BsmtFinSF1: Type 1 finished square feet

Functional_Typ: Typical Home functionality

SaleCondition_Normal: Sale Condition (Normal)

GarageCars: Size of garage in car capacity

Ridge Variables (Positive):

GarageCars: Size of garage in car capacity

SaleCondition_Normal: Sale Condition (Normal)

Functional_Typ: Typical Home functionality

SaleType_New: Type of sale (home just constructed and sold)

MSZoning_FV: Zoning Classification (Floating Village Residential)

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 gives an additional advantage of feature selection. Multicollinearity variables are eliminated by co-efficient becoming zero. In the present model, after rfe with 50 variables, lasso automatically eliminated 10 of them by making coefficient to zero. I would choose to apply lasso regression, given computational cost is not an issue.

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:

We will choose the next 5 variables in the Lasso Model:

Positive effect variables:

LotArea: Lot size in Square feet
BsmtFullBath: Basement Full Bathrooms
Foundation_PConc: Type of Foundation (Poured Concrete)
BsmtFinSF1: Type 1 finished square feet
Neighborhood_Crawfor: Crawford

Negative effect variables:

SaleType_WD: Type of Sale(Warranty Deed- Conventional)
HouseStyle_2.5Fin: Style of Dwelling(Two and one-half story: 2nd level finished)
HouseStyle_2Story: Two Story
MSZoning_RM: Residential Medium Density
GarageType_NoGarage: Garage Location

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:

We can make sure that a model is robust and generalisable by considering a simple model over a complex model.
Factors determining the complexity of the model:

- No. of parameters required to specify the model completely. Lesser the parameters, simpler the model is.
- The degree of function required to specify the model completely. Lesser the degree of the function, simpler the model is.
- The depth or size of a decision tree. Lesser the depth, simpler the model is.

- Lesser the size taken by the best – possible representation of the model, lesser the complexity.

Simple models have low variance, high bias and complex models have low bias, high variance.
If model complexity goes up, the bias reduces while the variance increases, hence the trade- off is required and the phenomenon is referred to as the bias-variance trade-off.
The implications on the accuracy of the model:
Robust and generalizable models are simple models are not very accurate compared to the complex model. It is because simple models have low variance, high bias. Bias qualifies for the accuracy of the model - how accurate is the model likely to be on future (test) data. This is because the simple model doesn't memorize entire training dataset, so when the future (test) data appears – it will have some error.