Subjective Questions

Assignment 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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

ANS:

Before Change

For Lasso:

optimal value of alpha: `0.001`

`For Ridge:`

optimal value of alpha: `20.0`

After Change:

For Lasso:

R 2 (Test Set): `0.8996605425007069`

for Ridge:

R 2 (Test Set): `0.9026750561898644`

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

For Lasso:

Feature Name	Coefficient Value
RoofStyle_Others	0.455
LowQualFinSF	0.318
PavedDrive_Y	0.268
RoofStyle_Hip	0.219
BsmtUnfSF	0.211
BsmtQual_TA	0.199
HouseStyle_2Story	0.186
BsmtCond_Others	0.185
GarageType_Detchd	0.181
Neighborhood_NridgHt	0.164
SaleType_WD	0.111
Neighborhood_NAmes	0.106
OverallQual_5.0	0.0955
Exterior1st_Wd Sdng	0.0922
Fireplaces	0.0893

For Ridge:

Feature Name	Coefficient Value
RoofStyle_Others	0.221
BsmtUnfSF	0.197
LowQualFinSF	0.184
BsmtCond_Others	0.156
GarageType_Detchd	0.155
BsmtQual_TA	0.145
1stFlrSF	0.14
HouseStyle_2Story	0.126
PavedDrive_Y	0.12
TotRmsAbvGrd	0.0978
Fireplaces	0.0975
RoofStyle_Hip	0.0915
Neighborhood_NridgHt	0.087
TotalBsmtSF	0.0867
SaleType_WD	0.0788

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?

ANS: Alpha value, first I used RFECV to gets best attributes and then Grid Search CV for best hyper parameters.

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?

For Lasso:

Feature Name	Coefficient Value
CentralAir_Y	0.448
Functional_Typ	0.242
OverallQual_Others	0.23
1stFlrSF	0.207
ExterQual_TA	0.206
BsmtExposure_Mn	0.202
Neighborhood_OldTown	0.201
BsmtFullBath	0.194
Heating_Others	0.194
LotConfig_Inside	0.187
GrLivArea	0.162
LotConfig_CulDSac	0.143
BldgType_Others	0.142
LotShape_Reg	0.126
OverallCond_Others	0.122

For Ridge:

Feature Name	Coefficient Value
CentralAir_Y	0.285
ExterQual_TA	0.19
BsmtExposure_Mn	0.184
BsmtFullBath	0.18
1stFlrSF	0.164
GrLivArea	0.162
Heating_Others	0.161
BldgType_Others	0.141
OverallQual_Others	0.119
LotConfig_Inside	0.114
GarageCars	0.106
WoodDeckSF	0.0998
OverallCond_Others	0.0986
BsmtFinType1_BLQ	0.0868
LotShape_Reg	0.0833

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?

Ans:

1. Make sure that train and test data are separated before standardization.
2. Check R 2 for train and test set --> They should hot have huge difference.
3. Check R 2 and adjusted R 2while training--> They should hot have huge difference.
4. Use RFE for feature elimination ie only one feature should be removed at a time
5. Use p-value (<= 0.05) and VIF (<5)