Surprise Housing -Subjective Questions

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:

Ridge Regression:

➢ **Optimal alpha :** 500

a. R2 score (training) : 0.860442758854242
b. R2 score (test) : 0.8676796928079703
c. RMSE (train) : 0.38314872049896104
d. RMSE (test) : 0.3406468684742878
e. R2 diff : -0.007236933953728264

➢ **Important predictor variables:**

1. GrLivArea: 0.11717410822968215
2. OverallQual_Excellent: 0.10954205488560492
3. OverallQual_Very Excellent: 0.09114322452491164
4. Neighborhood_NoRidge: 0.0885302912247072
5. Neighborhood_NridgHt: 0.07255258784276349
6. GarageArea: 0.07183788955397798
7. 1stFlrSF: 0.06803806320487565
8. FullBath: 0.06575327878684548
9. OverallQual_Very Good: 0.06464423816700801
10. BsmtExposure_Gd: 0.0608968003978244

➢ **Double of optimal alpha :** 1000
a. R2 score (training) : 0.840260
b. R2 score (test) : 0.850351
c. RMSE (train) : 0.409919
d. RMSE (test) : 0.362266
e. R2 diff : -0.010091

➢ **Important predictor variables:**

1. GrLivArea: 0.0982951436226687
2. OverallQual_Excellent: 0.08659922719987112
3. OverallQual_Very Excellent: 0.07682827838459913
4. Neighborhood_NoRidge: 0.07142571976316771
5. 1stFlrSF: 0.06808411726139915
6. GarageArea: 0.06748960552384384
7. Neighborhood_NridgHt: 0.06458908103334599
8. FullBath: 0.05593487706795671
9. Fireplaces: 0.05421418703868723
10. BsmtExposure_Gd: 0.05334760691738084

Lasso Regression:

➢ **Optimal alpha :** 0.01

✧ R2 score (training) : 0.8732283681234707

✧ R2 score (test) : 0.8747406352830928

✧ RMSE (train) : 0.3651760042701473

✧ RMSE (test) : 0.33143338547448453

✧ R2 diff : -0.0015122671596220494

➢ **Important predictor variables:**

1. GrLivArea: 0.2998151593389595
2. OverallQual_Excellent: 0.18804235563847368
3. OverallQual_Very Excellent: 0.13485997309766845
4. OverallQual_Very Good: 0.12423189879518762
5. Neighborhood_NoRidge: 0.11117560970563088
6. YearBuilt: 0.09859013377711466
7. Neighborhood_NridgHt: 0.08187079985040875
8. BsmtExposure_Gd: 0.07326155731477879
9. GarageArea: 0.06871693108290554
10. Neighborhood_Crawfor: 0.06108543604134479

➢ **Double of optimal alpha :** 0.02

✧ R2 score (training) : 0.8576924361163204

✧ R2 score (test) : 0.8707015327280597

✧ RMSE (train) : 0.38690574992527366

✧ RMSE (test) : 0.33673467434920884

✧ R2 diff : -0.013009096611739324

➢ **Important predictor variables:**

1. GrLivArea: 0.3041046400967545
2. OverallQual_Excellent: 0.18576308798863186
3. OverallQual_Very Excellent: 0.12934215523232556
4. OverallQual_Very Good: 0.1146211488253411
5. Neighborhood_NoRidge: 0.10041521688000812
6. YearBuilt: 0.09950387961013338
7. GarageArea: 0.08266509415515132
8. BsmtExposure_Gd: 0.07715113556064601
9. Neighborhood_NridgHt: 0.07635999982352393
10. YearRemodAdd: 0.05774862937666737

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 Regression:

➢ **Optimal alpha :** 500

✧ R2 score (training) : 0.860442758854242
✧ R2 score (test) : 0.8676796928079703
✧ RMSE (train) : 0.38314872049896104
✧ RMSE (test) : 0.3406468684742878
✧ R2 diff : -0.007236933953728264

Lasso Regression:

➢ **Optimal alpha :** 0.01

✧ R2 score (training) : 0.8732283681234707
✧ R2 score (test) : 0.8747406352830928
✧ RMSE (train) : 0.3651760042701473
✧ RMSE (test) : 0.33143338547448453
✧ R2 diff : -0.0015122671596220494

Based on the above details I will choose Lasso regression as it has better R2 score, RMSE value on test data set and additionally it also provides default feature selection by pushing coefficient of predictor variables to 0

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:

1) 1stFlrSF: 0.21158065736391707
2) Foundation_PConc: -0.15711890599020367
3) HeatingQC_Gd: -0.15037006480868076
4) Foundation_Slab: -0.14848448833278363
5) BsmtHalfBath: 0.1408605988466331

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?

- **Regularization:**

Regularization prevents over-fitting of the model. To avoid over-fitting we can do Ridge Lasso regression.

- **Cross-Validation:**

Cross-validation evaluates the model's performance and robustness by training and testing the model on different subsets of the given data.

- **Metrics:**

Metrics such as R-squared, RMSE(Root Mean Squared Error) evaluates the model's performance on both the training and testing datasets to improve accuracy and make the model more robust.

- **Data preparation and EDA (exploratory data analysis):**

Data cleaning mechanism should be done before building any model like removing null values, handling zero values, categorical values.Highly correlated variables should be removed to simplify the model. Scalling of the features can improvise the linearity or normality of residuals. RFE (Recursive feature elimination) identifies the most important features that
contributes to the accurate modelling.

- **Multicollinearity:**

Multicollinearity among the predictor variables should be handled as it can can lead to unstable predictions. Variance Inflation Factor (VIF) can be used to identify the multicollinearity in the data set.

Generalized model bring in more robustness, makes less errors on the train data which in turn reduce bias. There should be a trade-off between bias and variance to optimize the accuracy as simple model performs well on new and unknown data thereby making it more accurate predictions.