Submitted By: Bibhu Sundar Sahoo

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 value α is also sometimes called λ.
λ is called the tuning parameter. This hyperparameter controls how much to penalize the coefficients.
We have three cases

- **$\alpha=0$** - You get the same coefficients for linear regression as before
- **$0<\alpha<\infty$** - The coefficients are between 00 and the normal ones for linear regression
- **$\alpha=\infty$** - All coefficients are 0

As you increase α more coefficients go to zero because the ℓ_1 norm induces sparsity. It depends on what we're optimizing, the point is to get fewer coefficients or shrink the coefficients. And the model will be less complex.
For Ridge regression model optimal value was 3 and for lasso regression model it was 0.001.

On doubling alpha value, the penalty increases for cost function in both.
For Ridge:

- R2 changed from 92.02% to 90.53%

For Lasso:

- R2 changed from 88% to 83%

Top variables for Ridge:

('OverallQual', 0.082), ('GarageCars', 0.072), ('FullBath', 0.058), ('OverallCond', 0.055), ('BedroomAbvGr', 0.047), ('YearRemodAdd', 0.045), ('2ndFlrSF', 0.045), ('GrLivArea', 0.044)

Top Variables for lasso:

('OverallQual', 0.24), ('GarageCars', 0.12), ('FireplaceQu', 0.064), ('YearRemodAdd', 0.057), ('BedroomAbvGr', 0.04), ('HeatingQC', 0.025), ('BsmtFinType1', 0.022), ('KitchenQual', 0.02)

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:

Submitted By: Bibhu Sundar Sahoo

For the final model we looked into R2 and RMSE scores for both Ridge and Lasso. Comparing test and train set we looked into overfitting issues, below was our findings:

1. Model in Ridge had overfitting issues.
2. Model in Lasso with alpha 0.001 was the best fit(optimal alpha 0.005)
3. From residual analysis both test and train data seem to fit the assumptions of the Linear Regression.
    - Residuals have mean of zero and closely normally distributed.
    - Residuals do not have any pattern hence it has homoscedasticity.

Final Model equation

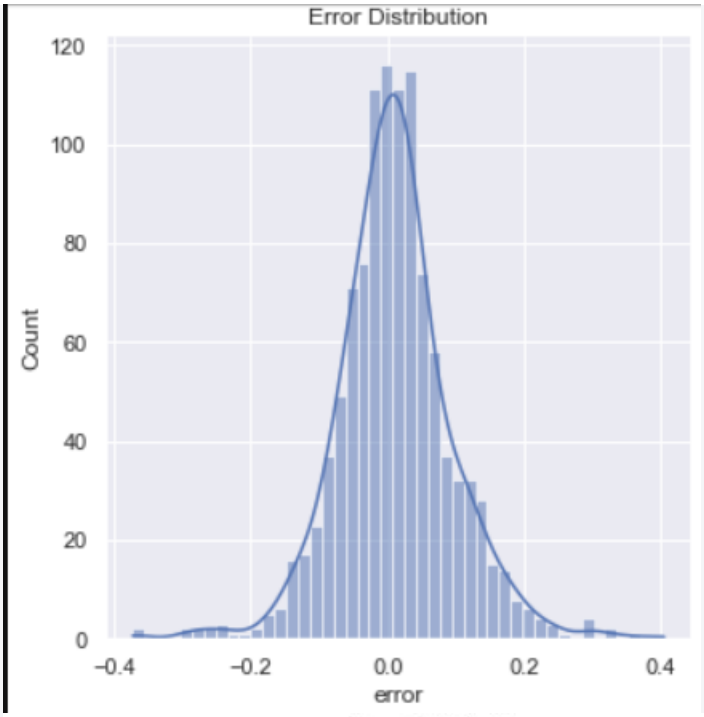
```
Sales Price = 0.095+ [ 0.246 X OverallQual + 0.115 X GarageCars + 0.082 X GrLivArea + 0.055 X FireplaceQu + 0.053 X BedroomAbvGr + 0.045 X YearRemodAdd + 0.044 X FullBath + 0.04 X OverallCond + 0.032 X BsmtQual + 0.029 X LotConfig_CulDSac + 0.026 X BsmtFinType1 + 0.018 X BsmtExposure + 0.018 X HeatingQC + 0.018 X KitchenQual + 0.015 X ExterQual + 0.014 X MSZoning_RL + 0.014 X Condition1_Norm + 0.008 X Neighborhood_Crawfor + 0.008 X Foundation_PConc + 0.007 X Neighborhood_NridgHt + 0.007 X HouseStyle_1Story + 0.006 X BsmtFullBath + 0.006 X Exterior1st_MetalSd + 0.006 X MasVnrType_BrkFace + 0.002 X 2ndFlrSF + 0.002 X GarageFinish + 0.001 X GarageType_Attchd + -0.002 X Neighborhood_SawyerW + -0.004 X SaleType_WD + -0.009 X GarageType_Detchd + -0.015 X Condition1_Feedr + -0.016 X MSZoning_RM + -0.02 X BldgType_Twnhs + -0.02 X Exterior1st_Stucco + -0.021 X MSSubClass + -0.035 X Neighborhood_Edwards + -0.057 X Neighborhood_IDOTRR ]
```

We got the best R2 score for train and test data from other models we tried.

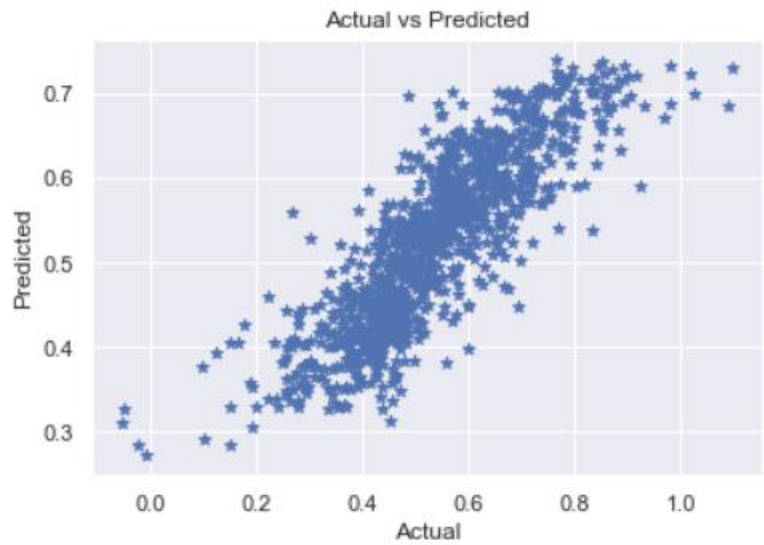
```
train R2 score is 0.88
test R2 score is 0.83
```

Assumptions of Linear Regression

Submitted By: Bibhu Sundar Sahoo



Submitted By: Bibhu Sundar Sahoo



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:

In Final Lasso model the top 5 most predictor variables are GrLivArea, GarageCars, FireplaceQu, BedroomAbvGr and FullBath.

After removing the above variables and retraining the model, the optimal alpha is 0.0005.

And the new predictor new variables are-

```
Fitting 5 folds for each of 25 candidates, totalling 125 fits
{'alpha': 0.0005}
                Variable  coeff_lasso_q4
15           OverallQual        0.293151
8               2ndFlrSF        0.117868
25           Fireplaces         0.072147
16           OverallCond        0.050962
2           YearRemodAdd        0.050732
77     HouseStyle_1Story        0.045357
```

Question 4

Submitted By: Bibhu Sundar Sahoo

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?

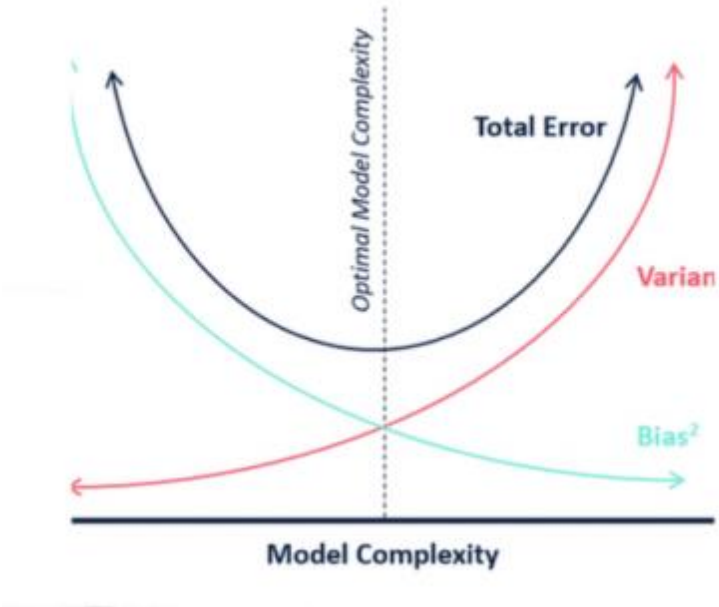
Answer:

In model selection we need a balance between Accuracy and Generability. Model selection technique are key for selecting the best model after the individual models are evaluated based on the required criteria.

Bias occurs when a model is strictly ruled by assumptions – like the linear regression model assumes that the relationship of the output variable with the independent variables is a straight line. This leads to underfitting when the actual values are non-linearly related to the independent variables.
How much error the model likely to make in test. Or correctness of the model

Variance is high when a model focuses on the training set too much and learns the variations very closely, compromising on generalization. This leads to overfitting.
How sensitive is the model to input. Or the consistency of the model

An optimal model is one that has the lowest bias and variance and since these two attributes are indirectly proportional, the only way to achieve this is through a tradeoff between the two. Therefore, the model selection should be such that the bias and variance intersect like in the image below.



Submitted By: Bibhu Sundar Sahoo

Models can be evaluated in multiple metrics in terms of

1. Accuracy
2. Stability

In case of Regression metrics-

As Regression models provide a continuous output variable, unlike classification models that have discrete output variables. Therefore, the metrics for assessing the regression models are accordingly designed looking into

1. R2
2. Mean Squared Error or MSE
3. Root Mean Squared Error or RMSE

In case of Classification metrics-

A confusion matrix can be constructed which demonstrates the number of test cases correctly and incorrectly classified.
It looks something like this (considering 1 -Positive and 0 -Negative are the target classes):

	Actual 0	Actual 1
Predicted 0	True Negatives (TN)	False Negatives (FN)
Predicted 1	False Positives (FP)	True Positives (TP)

So, the metrics for assessing these models include-

1. Accuracy
2. Precision
3. Recall
4. AUC-ROC