KHAID MOHAMAD ALADIN HAFEZ OMAR EMAIL: KHALIDOMARASU@GMAIL.COM

Subjective Questions: Advanced Regression Assignment

Problem Statement - Part II

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

Answer 1:

Optimal value of alpha for Ridge Regression Model :0.2
Optimal value of alpha for Lasso Regression Model :0.001
Doubling the value of alpha for both ridge and lasso then models would cause
The values of feature coefficient will be decreased slightly as some features would change their position in the top 10 features based on their coefficients while some features will change. If the value of alpha is not optimal for model so the coefficient of feature decreases, In lasso regression model the R-Squared value of train and test data declined due to doubling the alpha value, thus the bias of model would be high. Lasso model some of the coefficient value reached 0, but in case of ridge, the coefficient become close to 0 but not 0.

Ridge Doubling Alpha Effect on most important predictors:
Effect of doubling the value of alpha on top 10 features in Ridge Regression Model

1.	OverallQual	11.501259
2.	MSZoning_RH	0.400664
3.	MSZoning_RL	0.346604
4.	MSZoning_RM	0.345172
5.	LotConfig_FR3	0.271928
6.	SaleType_WD	0.198507
7.	Neighborhood_Veenker	0.153107
8.	SaleType_Oth	0.129255
9.	GarageCars	0.127775
10.	Neighborhood_Mitchel	0.123564

Lasso Doubling Alpha Effect on most important predictors:
Effect of doubling the value of alpha on top 10 features in Lasso Regression Model

1.	OverallQual	11.970515
2.	OverallCond	0.124028
3.	2ndFlrSF	0.112522
4.	GarageCars	0.109783
5.	HouseAge	0.067643
6.	Neighborhood_Mitchel	0.066710
7.	BsmtFinType1	0.063261
8.	Condition1_PosA	0.043119
9.	BsmtFinSF2	0.040854
10.	Neighborhood_StoneBr	0.037218

Doubling Alpha on both models Effect on most important predictors:
1. OverallQual
2. MSZoning
3. LotConfig
4. SaleType
5. Functional
6. Neighborhood
7. GarageCars
8. OverallCond
9. 2ndFlrSF
10. HouseAge

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 2:

Both ridge and lasso models show important predictive features are almost the same but on the other hand, the R-Squared score for Train data is higher in Ridge regression and R-Squared score for Test data is higher in Lasso regression model, Since Lasso regression model is performing great on unseen data it will be used while Lasso also helps feature selection as the optimum value of LASSO is smaller than RIDGE thus I will apply LASSO Regression 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 3:

After:			**Before**
1.	BsmtFinType1	11.752551	1. OverallQual
2.	RoofStyle_Shed	0.258873	2. GarageCars
3.	GarageType_Detchd	0.246021	3. OverallCond
4.	Neighborhood_NWAmes	0.217751	4. 2ndFlrSF
5.	Neighborhood_Veenker	0.192540	5. Neighborhood_Mitchel

KHAID MOHAMAD ALADIN HAFEZ OMAR EMAIL: KHALIDOMARASU@GMAIL.COM

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 4:

- All models shall be robust and generalizable to prevent impact of outliers in the training data.
- Models is said to be robust if its stable thus not changing hugely when changing the training set.
- Models are generalizable if they do not overfit the training data have high variance and runs well on another training data.
- Outliers weight shall not be high in order not to affect model prediction accuracy only those relative to the dataset shall be retained.
- Outliers shall be treated well and removed since it would affect the model accuracy if not treated it might increase the accuracy of predictions made by the model.
- Non-robust models cannot be trustable for predictive analysis.
- Model shall act slightly the same when exposed to new dataset other than the training dataset.
- Models built shall be generalizable to output a test accuracy not lower than the training score.

Implication in terms of accuracy:

- A robust and generalizable model will perform equally fine on both training and test data in other words the accuracy does not change much for training and test data.

Changes in model to make it more Robust:

1. Use a model that's resistant to outliers
2. Use a more robust error metric.

Changes to data for more robust:

❖ Transform your data and Remove the outliers.