Advanced Regression Surprise Housing 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

- **Optimal value of alpha:**
  - Optimal value of alpha for Ridge regression is 10
  - Optimal value of alpha for Lasso regression is 0.002

- **The changes in the model if you choose double the value of alpha for both ridge and lasso:**
  For Ridge it will reduce the coefficients of predictor variables. It will shrink coefficient and it will minimize features weight.

  For Lasso, along with shrinking coefficients it will also eliminate some more features. More features will have coefficient as 0 and more predictors will be eliminated and hence number of predictors will reduce.

  For Lasso, when alpha is doubled and set to 0.004
  - Lasso score on train set is 86.12
  - Lasso score on test set is 87.33
  - 120 fields got eliminated when alpha is doubled to 0.004
  - Before 93 fields were eliminated when alpha was 0.002

- **The most important predictor variables after the change is implemented**
  For Lasso – top 10 important variable remains the same however coefficients reduced further. Note - 120 fields got eliminated.

Lasso Alpha = 0.002		Lasso Alpha = 0.004	
AgeOfProperty	0.872	AgeOfProperty	0.812
SaleCondition_Partial	0.792	SaleCondition_Partial	0.575
SaleCondition_Normal	0.523	SaleCondition_Normal	0.413
SaleCondition_Family	0.341	SaleCondition_Family	0.288
SaleCondition_Alloca	0.329	SaleCondition_Alloca	0.281
SaleCondition_AdjLand	0.308	SaleCondition_AdjLand	0.279
GarageFinish_Unf	0.249	GarageFinish_Unf	0.237
GarageFinish_RFn	0.242	GarageFinish_RFn	0.223

For Ridge – top 10 important variable remains the same however coefficients reduced further. Note - No feature selection/elimination in Ridge.

Ridge Alpha = 10		Ridge Alpha = 20	
AgeOfProperty	0.512	AgeOfProperty	0.399
SaleCondition_Partial	0.403	SaleCondition_Partial	0.311
SaleCondition_Normal	0.362	SaleCondition_Normal	0.278
SaleCondition_Family	0.326	SaleCondition_Family	0.235
SaleCondition_Alloca	0.229	SaleCondition_Alloca	0.207
SaleCondition_AdjLand	0.225	SaleCondition_AdjLand	0.187
GarageFinish_Unf	0.223	GarageFinish_Unf	0.186
GarageFinish_RFn	0.218	GarageFinish_RFn	0.185
GarageFinish_NoGarage	0.193	GarageFinish_NoGarage	0.179
GarageType_NoGarage	0.192	GarageType_NoGarage	0.151

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

Considering this data set has large number of predictors/columns, I will choose Lasso regression as it helps in feature selection with optimum lambda.

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

Next top 5 most important variables in Lasso regression are

SaleCondition_AdjLand	0.308
GarageFinish_Unf	0.249
GarageFinish_RFn	0.242
GarageFinish_NoGarage	0.204
GarageType_NoGarage	0.172

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

We can make sure that a model robust and generalizable when they are not impacted by unseen data, outliers and in test data.

Data-Based Methods

- **Winsorizing:** Artificially cap your data at some threshold. This method involves setting the extreme values of an attribute to some specified value. For example, for a 90% Winsorization, the bottom 5% of values are set equal to the minimum value in the 5th percentile, while the upper 5% of values are set equal to the maximum value in the 95th percentile. This is more advanced than trimming where we just exclude the extreme values.
- **Transform your data:** If your data has a very pronounced right tail, try a log transformation. Log-Scale Transformation method is often used to reduce the variability of data including outlying observation. Here, the y value is changed to log(y). It's often preferred when the response variable follows exponential distribution or is right-skewed.
- **Remove the outliers.** This works if there are very few of them and you're fairly certain they're anomalies and not worth predicting. Outliers which are important for model should be retained else must be removed. We can review the outliers using box plot or z-score to ensure much weightage is not given to the outliers so that the accuracy predicted by the model is high.
- **Binning:** This refers to dividing a list of continuous variables into groups. We do this to discover sets of patterns in continuous variables, which are difficult to analyze otherwise. But, it also leads to loss of information and loss of power.

Model-Based Methods

- We can also confirm model is robust and generalizable by reviewing if test score is better than training score.
- Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model.
- Use a different model: Instead of linear models, we can use tree-based methods like Random Forests and Gradient Boosting techniques, which are less impacted by outliers. Use a model that's resistant to outliers.
- Use a more robust error metric. switching from mean squared error to mean absolute difference.