Advanced Regression Assignment

MAY 11

UpGrad Submission – C36
Authored by: Anjali Vashisth

Gist for Part -1 Assignment

The link for the Jupyter Notebook – Click Here

The content of the part-1 assignment covers the following topics:

1. Importing Libraries
2. Reading and Understanding data
    . reading data
    A. dimensions check
    B. info check
    C. null check
    D. describing the statistical Summary of the DS
3. Data Cleaning
    . Checking for the nulls
        - 1st check: entire row nulls
        - 2nd check: entire column nulls
        - 3rd check: %of nulls
    A. Delting the cols with more than 15% of nulls
    B. Check for unique values less than 2
    C. droping 'ID' of the dataset
    D. dropping records containing the missing value
4. Data Preparation
    . Creating derived features from YearBuilt, YearRemodAdd and GarageYrBlt
    A. Splitting the columns into numerical and categorical data fields
5. Perform EDA
    . creating scatter plots (relationship b/w independent and dependent var)
    A. box plot for categorical vars
    B. check correlation
6. Create Dummy Variables
7. Data Prep and Modelling
    . Outlier handling
    A. Test and train data - standard scalar
    B. Model building - simple regression model
    C. Model building - Ridge regression
    D. Model building -Lasso regrssion
8. Coming to Conclusions

major features from the initials columns that influence the Sales Price Some of the major influencers are: OverallQual; OverAllCond; Neighborhood; KitchenQual; BsmtQual; SaleCondition; GarageCars

Part 2 – Assignment – Answers to Questions

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?

The optimal values of lambda are:

- Ridge: 10
- Lasso: 100

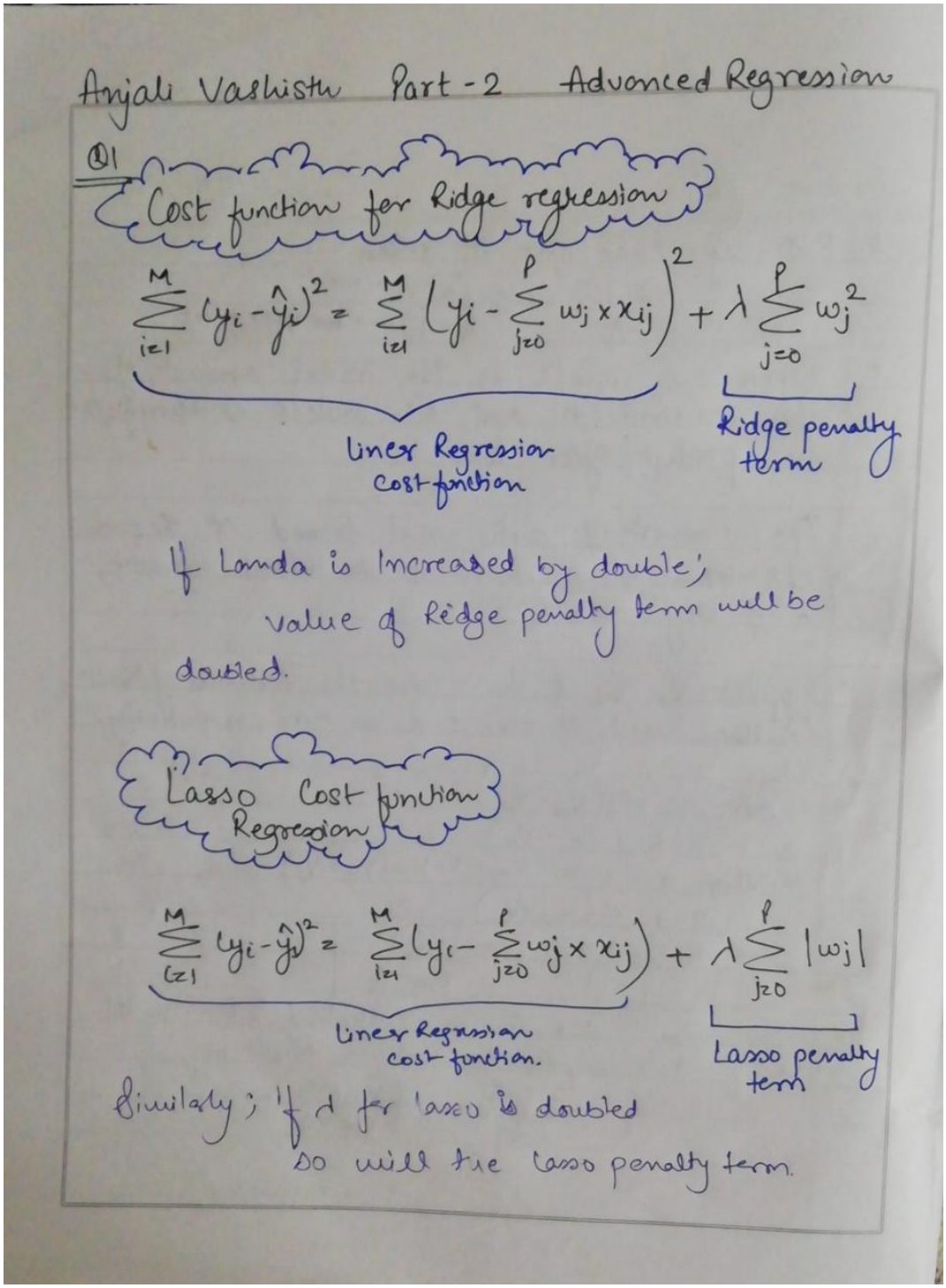
The r2 score for the optimal values of lambda comes out to be as:

	R2 SCORE TRAIN	R2 SCORE TEST
RIDGE (ALPHA=10)	0.9102708640356322	0.8291066996854972
LASSO (ALPHA=100)	0.9127400949466069	0.8226115465655931

When the values of lambda are doubled:

	R2 SCORE TRAIN	R2 SCORE TEST
RIDGE (ALPHA=20)	0.8992839099644752	0.8289147747773998
LASSO (ALPHA=200)	0.9001663484100466	0.8255244923413052

When the value of lambda is doubled, the r2 scores for both ridge and lasso change slightly. After the value of alpha is doubled, the following factors become important:
• OverallQual • RoofMatl • Neighborhood • SaleCondition • GarageCars • Functional



And same goes for Lasso.

- Model change is:
  - Increased in Root mean squared error value
  - Decreased R2 values.

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?

Lambda values of 10 and 100 are ideal for ridge and lasso, respectively. There is no noticeable difference in performance between the two models when built with these parameters. There is no significant change in the r2 score for the test data.
Ridge, on the other hand, does not make the coefficients zero when looking at the model parameters, but lasso does make the coefficients of a number of variables zero, assisting in feature selection. As a result, it is preferable to utilize lasso regression with lambda set to 100.

	Alpha	*R-squared Train*	*R-squared Test*
Ridge	10	0.9102708640356322	0.8291066996854972
Lasso	100	0.9127400949466069	0.8226115465655931

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?

After the top 5 features are dropped the next most important predictors become

- MSSubClass,
- OverallQual,
- KirchenQual,
- BsmtQual,
- LotShape

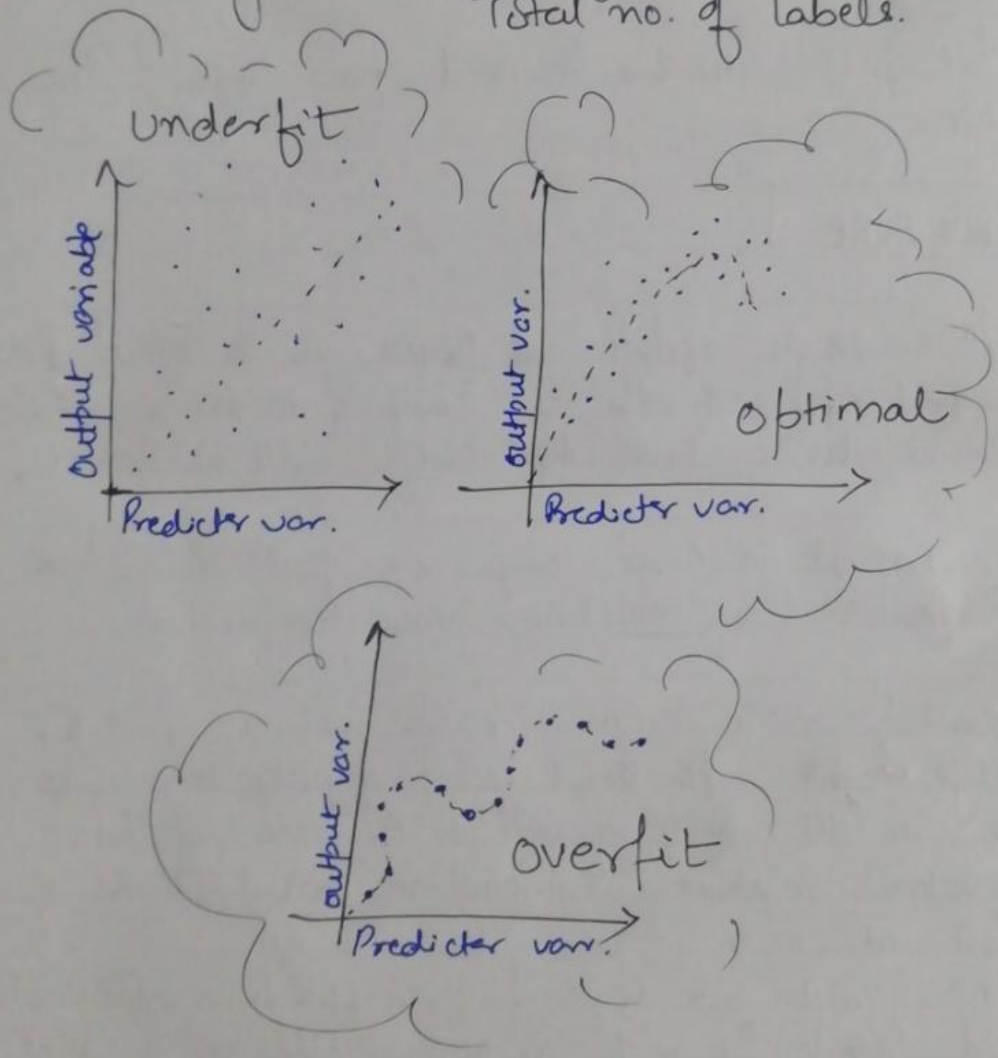
QUESTION 4- HOW CAN YOU MAKE SURE THAT A MODEL IS ROBUST AND GENERALIZABLE? WHAT ARE THE IMPLICATIONS OF THE SAME FOR THE ACCURACY OF THE MODEL AND WHY?

When the training set is altered, a model is called robust and generalizable if it does not demonstrate a significant change in performance, i.e. the model does not overfit on the training data and can handle new/unseen data adequately. A robust and generalizable model should perform equally well on training and test data when it comes to accuracy.

A more generalised model is a model which does not overfit the data. To ensure our model is not overfitting the data we must regularise our model using hyperparameters. These hyperparameters will help in reducing the complexity of the model by penalising features contributing to the complexity. We add a penalty term to the cost function that increases with increasing model complexity. So we try to bring it down and control model complexity Having a simple model will help ensure that it is robust and more generalised. But the model should not be over simplified otherwise it will underfit the data and this model will be too naive to give us a valid or accurate output. This is the scenario where model is underfitting. Accuracy of the model is defined as the ratio of number of correct predictions to the total number of input samples.

The following are the implications of making a model resilient and generalizable on model accuracy:
The accuracy of the model will be more steady if we make it more robust, that is, less vulnerable to outliers or changes in test data. This means that slight modifications to the test data set will not result in significant changes in accuracy values.
When trying to make the model simpler by penalising the model's complexity, the accuracy on the test set will increase in the beginning. When we have made the model sufficiently basic, the accuracy will stabilise on the test data set.
If the accuracy is not maintained, then the model can be underfitted or overfitted.

Accuracy = $\frac{\text{Correctly predicted labels}}{\text{Total no. of labels.}}$



Anjali Vashisth
Part 2
Advanced Regression
Assignment.


Thank
you