Subjective Question Solution

Question-1:

Rahul built a logistic regression model having a training accuracy of 97% while the test accuracy was 48%. What could be the reason for the seeming gulf between test and train accuracy and how can this problem be solved.

Answer 1-

There might be various reasons for this -

1- He might have used very less features while doing testing phase. He might have drooped lots of features because of nan value, missing values etc. but he used all the features for training.
2- He might have not scaled all the features appropriately.
3- Regularisation has not been used properly. That is wrong hyper parameter have been selected.
4- He didn't separated data into training, validation and test in good ratio.
5- external cross validation has not been done appropriately.

Question-2:

List at least 4 differences in detail between L1 and L2 regularization in regression.

Answer 2-

Below are four differences -

1- Ridge and Lasso regression uses two different penalty functions. Ridge regression uses the square of the co-efficients while lasso uses the modulus.
2- Ridge is computationally less intensive than Lasso.
3- Ridge regression can't have zero coefficients. Here, we either select all the coefficients or none of them whereas LASSO does both parameter shrinkage and variable selection automatically because it zero out the co-efficients of collinear variables. Here it helps to select the variable(s) out of given n variables while performing lasso regression.
4- Ridge gives non-sparse output where as lasso gives sparse output.

L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection

Question-3:

Consider two linear models

L1: y = 39.76x + 32.648628

And

L2: y = 43.2x + 19.8

Given the fact that both the models perform equally well on the test dataset, which one would you prefer and why?

Ans 3-

I would prefer L2 model because of following reasons –

1- Its coefficients are simple i.e. it's a simpler model between two.
2- It is less computation intensive for any machine learning algorithm.
3- Its easily extendable because of simplicity.
4- It also saves storage at long run.
5- Easy to understand so easy to maintain.

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?

Ans 4-

There are various methods, metrices by which we can decide that our model is robust.

We can check below parameter to make our model more Robust.

1- RSS
2- TSS
3- ESS
4- Sensitivity
5- Specificity
6- TPR
7- FPR
8- Precision
9- Recall

Generalisability can be check using L1 and L2 methods. And then model can be validated using cross verification methods.

Implications –

It can cause sometime overfitting or underfitting the model which eventually leads to less model accuracy on test data set. Therefore we would not get higher prediction power.

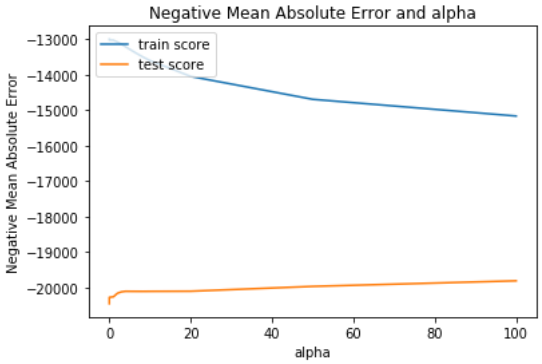
Question-5:

As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

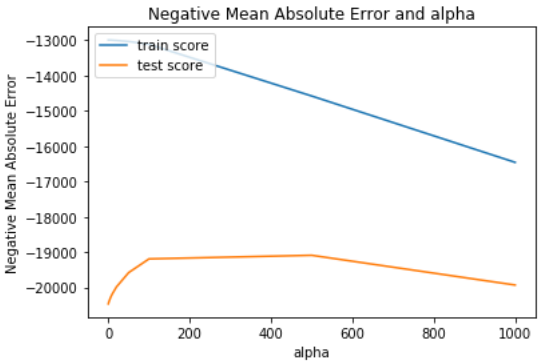
Answer 5-

I got the below values for hyper parameter –

1- Alpha for ridge- 8



2- Alpha for lasso – 100



Since we have more than 180 variables after treating the dummy variables.
I would like to use lasso so that I can eliminate the collinearly related features.