Assignment Part-II (Subjective Questions)

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:
The problem with the model is that it is **overfitting** the training data. This is why it is performing well on training data, and not on the test data.
One way to solve this problem could be to use **regularized regression**. Regularization is a process used to create an optimally complex model, thus we would get a model which would not overfit the training data.

Question-2:
List at least 4 differences in detail between L1 and L2 regularization in regression.

Answer:

L1 Regularization (Least Absolute Error)	L2 Regularization (Least Square Error)
It shrinks some of the variable coefficients to 0	It does not shrink the variable coefficients to 0
It is robust but the solutions are not stable	It is not robust but the solutions are stable
It is computationally inefficient	It is computationally efficient
It does perform feature selection as well	It does not perform any 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?

Answer:
Given the fact that both the models are performing equally good on the test data, we would prefer the model based on simplicity.

Upon observing the precision of coefficients, we find that the model L2 is having lesser precision for the coefficients as compared to model L1, hence, L2 is simpler as compared to L1.
Thus, **we would prefer L2 over L1 for simplicity**.

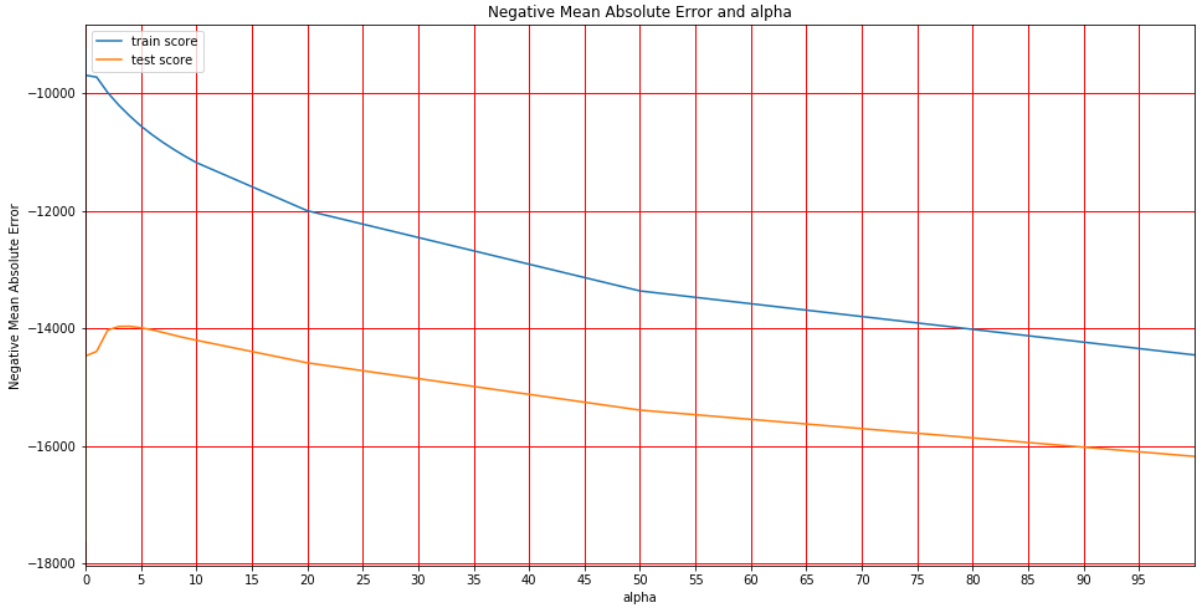
Question-4:

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:
A model is considered to be robust if the model is stable, i.e. does not change drastically upon changing the training set. The model is considered generalisable if it does not overfits the training data, and works well with new data.
Its implication in terms of accuracy is that a robust and generalisable model will perform equally well on both training and test data i.e. the **accuracy does not change much for training and test data**.

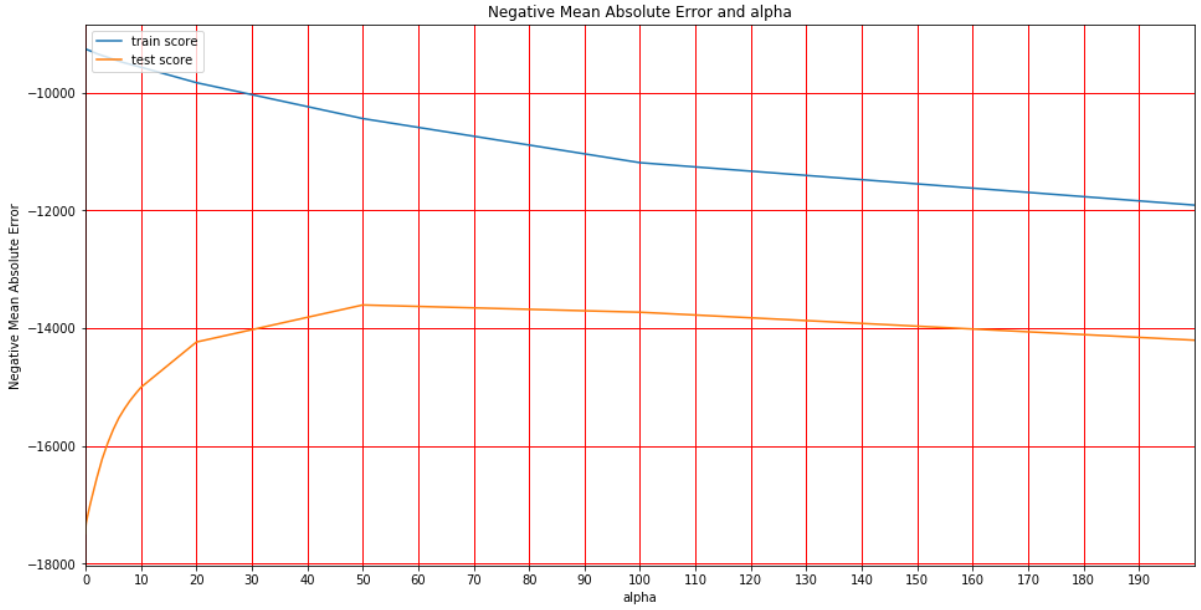
Question-5:
As you have determined the optimal value of lambda for ridge and lasso regression during the assignment, which one would you choose to apply and why?

Answer:
We would decide that on the basis of plots and chose a value of alpha where we have good training as well as the test score.
Ridge regression plot:



Based on the plot, we choose 4 as the value for lambda for Ridge Regression, since it has the best train as well as the test score.

Lasso Regression Plot:



Based on the plot, we choose 50 as the value for lambda for Lasso Regression, since it has the best train as well as the test score.
