Housing Price – Advanced Regression Part II

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.

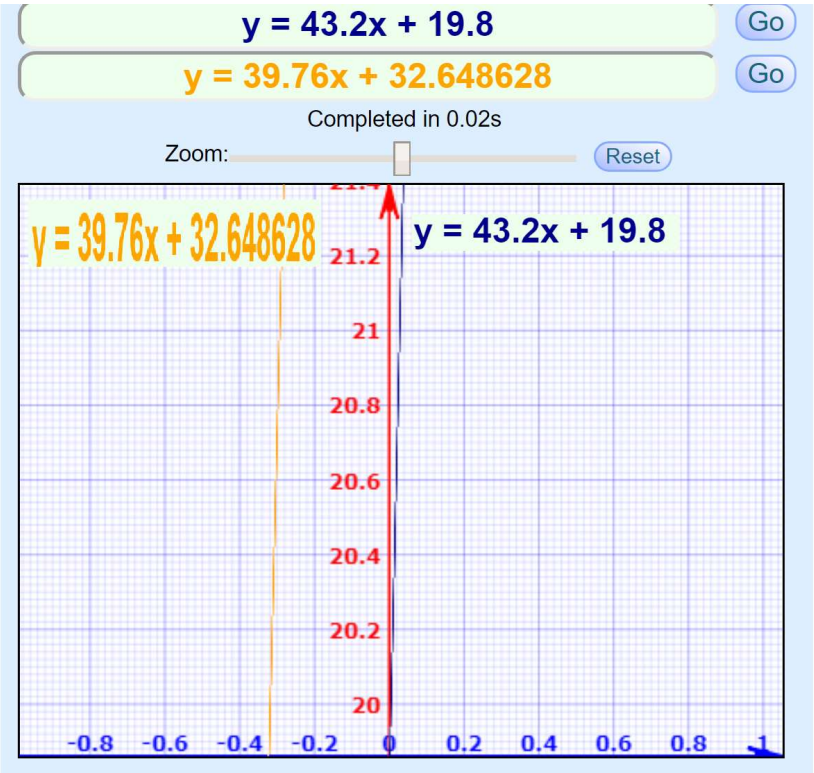
Answer:
The model that has been built is over-fitted on training data that's why test accuracy was so low and also model was not simple and was complex. This problem can be avoided if we do EDA properly and also follow all the rules of making model correct and model is simple.

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

Answer:
L1 regression is called lasso regression and L2 regression model is called ridge regression. Below are the difference:
a. Lasso trims down the coefficients of redundant variables to zero whereas Ridge, on the other hand, reduces the coefficients reduces not zero.
b. Lasso performs variable selection whereas Ridge doesn't do variable selection.
c. In lasso, regularization term of "sum of the absolute value of the coefficients" is added, whereas in ridge, an additional term of "sum of the squares of the coefficients" is added to the cost function along with the error term.
d. Lasso is used as a variable shrinkage method, whereas ridge doesn't.

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:
I will prefer to chose L2 model as this model is simpler than L2 model. And also if we plot both models on graph L2, L2 slope is better than L1 slope and graph also is best suited for L2 for regression as compared to L1. Refer below graph:



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 our model robust and generalisable by working on less training data and our model should have high bias and low variance.
If we make our model robust and generalizable, then there may be chance that our accuracy Is not that high for training model but accuracy for test data will be comparable to train dataset score.

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:
I will prefer to choose Ridge as my model is better in Ridge for both test and train dataset. Lambda value in Ridge for train and test is 10 and 100 while for lasso it came 0.001 and 20 for train and test data. Variation is less in ridge compared to lasso which is need to make model better robust and generalizable.