Question 1

Rahul built a logistic regression model with a training accuracy of 97% and a test accuracy of 48%. What could be the reason for the gap between the test and train accuracies, and how can this problem be solved?

Ans: This is problem of overfitting, where the Model performs well on the Training data but fail on the test/unseen data. Problem of overfitting can be solved using **Regularization.**
Regularization discourages the model from becoming too complex
Regularization helps to solve over fitting problem in machine learning. Simple model will be a very poor generalization of data. At the same time, complex model may not perform well in test data due to over fitting. We need to choose the right model in between simple and complex model. Regularization helps to choose preferred model complexity, so that model is better at predicting. Regularization is nothing but adding a penalty term to the objective function and control the model complexity using that penalty term.

Question 2

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

Ans:

L1(Lasso) and L2(Ridge) regularization are used to avoid overfitting of data. L1 and L2 penalized estimation methods shrink the estimates of the regression coefficients towards zero relative to the maximum likelihood estimates. The purpose of this shrinkage is to prevent overfit arising due to either collinearity of the covariates or high dimensionality. Although both methods are shrinkage methods, the effects of L1 and L2 penalization are quite different in practice.

Main difference between L1 and L2 regularization are:

1. L2 regularization produces dense models while L1 regularization produces sparse models.
2. L2 regularization is computationally efficient while L1 regularization is computationally inefficient.
3. L2 penalty tends to result in all small but non-zero regression coefficients, whereas applying an L1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage.
4. L2 regularization is the sum of the square of the weights, while L1 regularization is just the sum of the weights

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 data set, which one would you prefer and why?

Ans:

I will choose L2 as this is simpler model and have smaller coefficient(upto. 1 Decimal place) which require less memory space.

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?

Ans: This can be made sure by keeping/choosing a simple model Advantage of simpler model are as below:

1. A simpler model is usually more generic than a complex model. This becomes important because generic models are bound to perform better on unseen datasets.
2. A simpler model requires less training data points. This becomes extremely important because in many cases one has to work with limited data points.
3. A simple model is more robust and does not change significantly if the training data points undergo small changes.

Implications on accuracy of model:

A simple model may make more errors in the training phase but it is bound to outperform complex models when it sees new data. This happens because of overfitting.

Question 5

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?

Ans: I would choose Lasso regression, As the model has feature reduction capability and Features reduced drastically from 126 to 38 by making coefficient 0. whereas r2 score for both ridge and lasso is same. Hence lasso is a better and **simpler** model with same accuracy as ridge.

- Ridge Regression
- Train r2 score 0.8864029286596737
- Test R2 score 0.8506945010633415
  - No. of Feature in final Model 126
  - Lambda = 100

- Lasso Regression

- Train R2 score 0.882959509531136
- Test R2 score 0.8529477147279473
  - No. of Feature in final Model 38
  - Lambda = 500 Lasso regression with below 43 feature defines the sale price