# Assignment 4 Report

**5.**

(a)

I implemented Logistic regression code from scratch declaring with initial values w = [0 0], b=0 and attained an an accuracy of 66.67% ( 4 out of 6 inputs of test data are correct )

(b)

(i)

The logistic function σ is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

The cross-entropy error function is :

$$\xi(t, y) = -t \log(y) - (1 - t) \log(1 - y)$$

(ii)

The updated model is implemented in the code attached.

(iii)

Accuracy of the model is : 50.0 %

Precision 0.5

Recall 0.5

**6.**

As the dataset has 55 million rows I considered the first 2 lakh rows and further done data cleaning and feature engineering on it.
Steps implemented are
Calculated distance travelled
Removed few columns using correlation matrix
Created new columns from existing datetime column
Plotted Scatter plots to remove outliers
Done Preprocessing

Models I developed with this processed data are
Linear regression
Random Forest Regression
GradientBoost Regression
Bayesian Regression

I have achieved RMSE less than 4 with Bayesian Regression, GradientBoost Regression and RandomForest Regression.

With Bayesian regression you can achieve a whole range of involved inner conclusions of data which helps to get a better score compared to other regressors.
GradientBoost regressor is an ensemble regressor which develops a strong model with weakly attached features by working on residuals in each step.This helps the model to turn out as a strong predictor compared with other models.
In the RandomForest regression model the model takes care more about misclassified samples by increasing more weight on them in each layer. This would make model to turn out as a strong predictor

Compared to all these regressors, the Linear regression model is a weak predictor. It is a very basic model as it doesn't care about the misclassified points, residuals and many more which are taken care more effectively with the above mentioned models.