Report

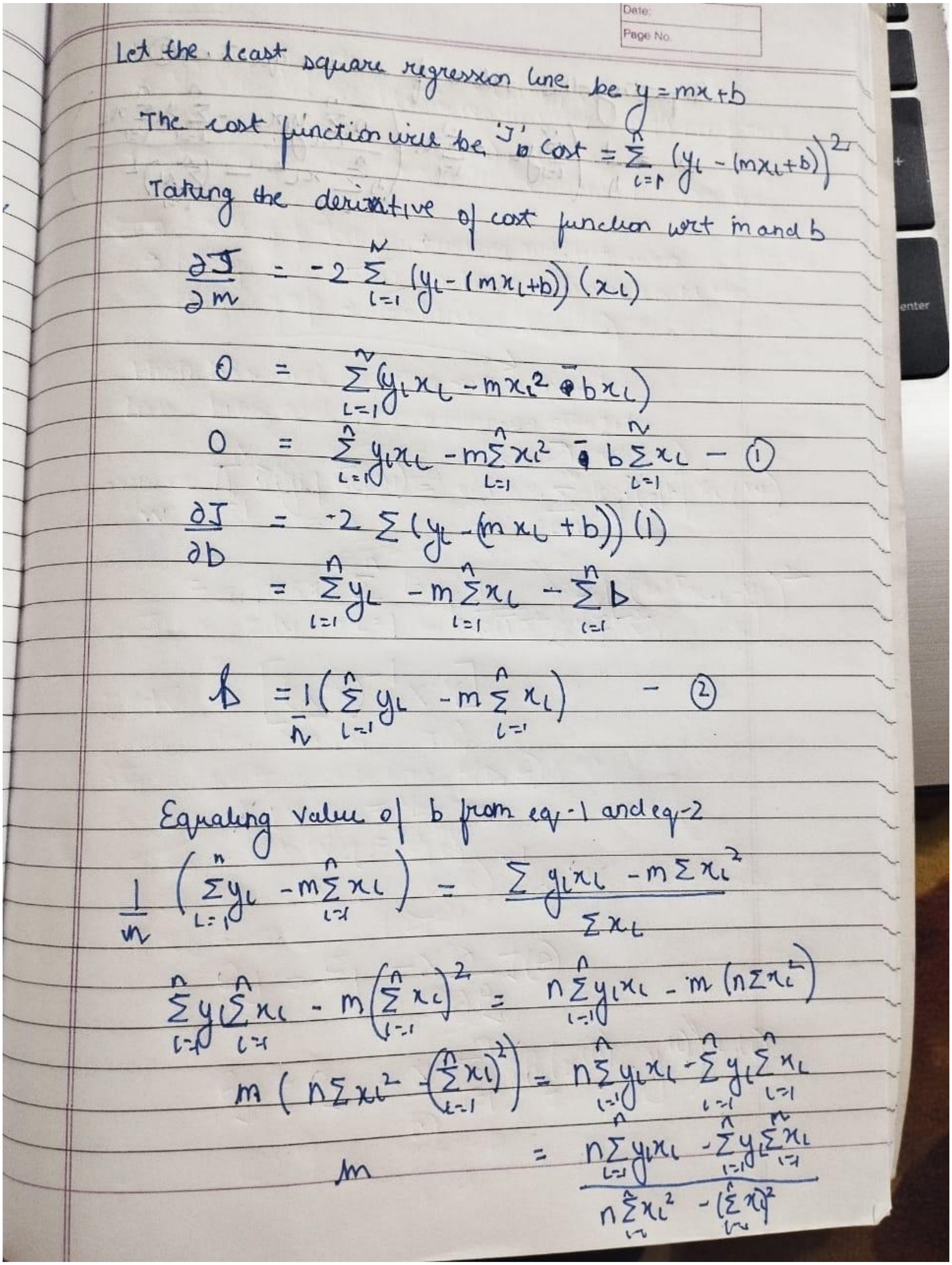
Section A :

A. No, if two variables exhibit a strong correlation with a third variable, it does not necessarily imply that they will also display a high degree of correlation with each other.
   Counter Example:
   a. In hot weather seasons like summers the electricity consumption increases hence they share a strong positive correlation with each other.
   b. In hot weather seasons like summers the sale of items like cold beverages and ice-cream increases hence they share a strong positive correlation with each other.
   c. But there is no strong correlation between the consumption of electricity and sale of items like cold beverages and ice-cream yet they share a strong positive correlation with hot weather/summers.

B. The criteria for a mathematical function to be considered as a logistic function is that its range should be bounded between lower and upper asymptotes, its graph should be a S-shaped curve and the function should be symmetric with respect to its point of inflection.
   a. Out of the given functions:
   b. Sinh(x) and Cosh(x) cannot be considered as logistic functions as their range is unbounded.
   c. Graph of the signum function is not S-shaped hence it cannot be regarded as a logistic function.
   d. Tanh(x) satisfies all the required conditions hence it can be regarded as a logistic function.

C. Leave one out validation is beneficial for very sparse datasets as in this technique we leave one sample and train on the remaining data-points, we repeat this for every data point and take the average of the errors whereas in K fold validation we divide the data in k folds and then train on k-1 folds, test on the kth fold and then take the average of the error. Leave one out validation can also be considered as a special case of K fold validation where K is equal to the number of data points.
   a. We use Leave one out validation because it is beneficial in case of sparse datasets as otherwise the size of the training data will be less which will lead to poor training of the model.

D.



Date:

Page No.

Let the least square regression line be $y = mx + b$

The cost function will be 'J' is Cost $= \sum_{i=1}^{n} (y_i - (mx_i+b))^2$

Taking the derivative of cost function wrt m and b

$$\frac{\partial J}{\partial m} = -2 \sum_{i=1}^{N} (y_i - (mx_i+b))(x_i)$$

$$0 = \sum_{i=1}^{N} (y_i x_i - m x_i^2 - b x_i)$$

$$0 = \sum_{i=1}^{n} y_i x_i - m \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{N} x_i \quad - ①$$

$$\frac{\partial J}{\partial b} = -2 \sum (y_i - (m x_i + b))(1)$$

$$= \sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b$$

$$b = \frac{1}{n} \left(\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i \right) \quad - ②$$

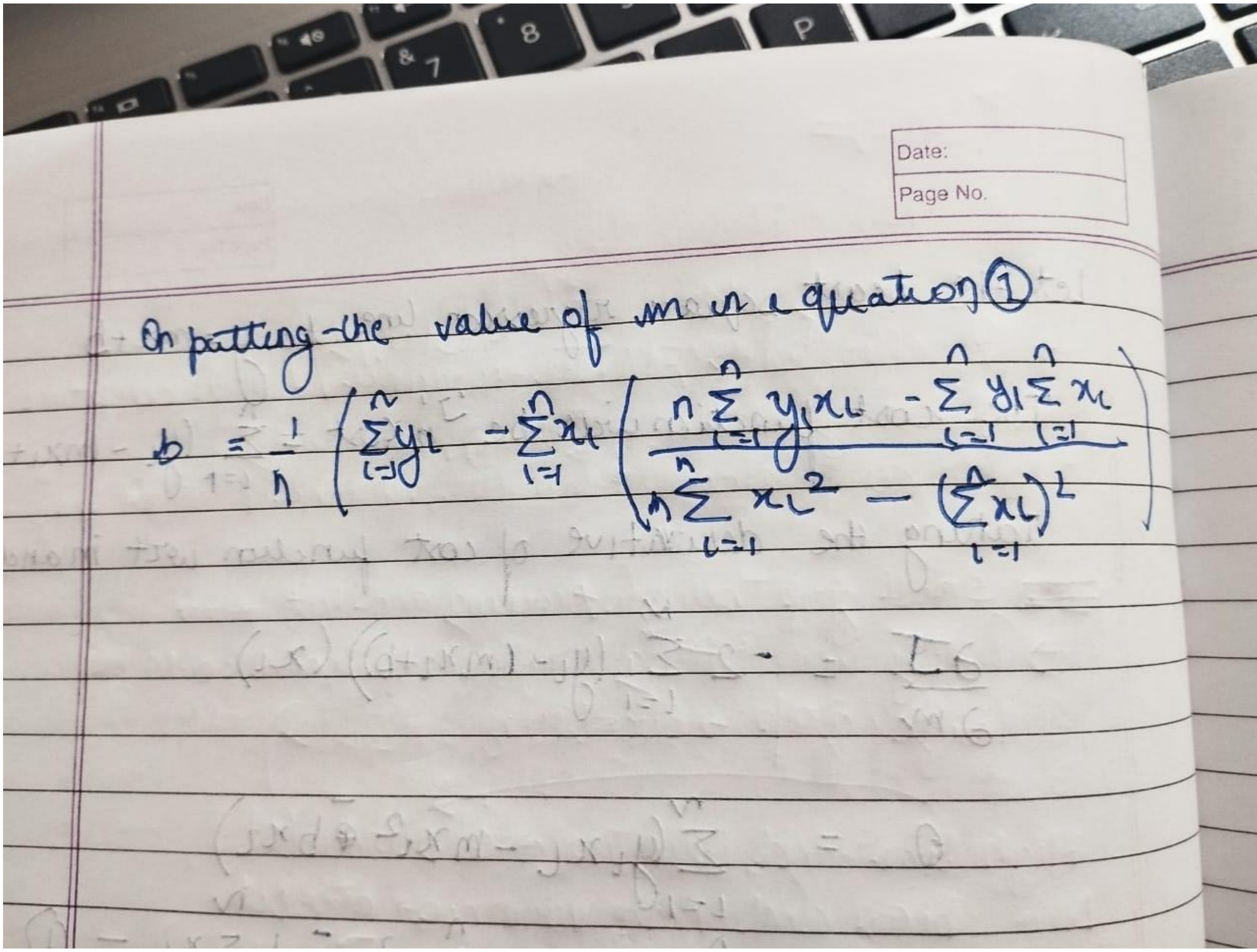
Equaling value of b from eq-1 and eq-2

$$\frac{1}{n} \left(\sum_{i=1}^{n} y_i - m \sum_{i=1}^{n} x_i \right) = \frac{\sum y_i x_i - m \sum x_i^2}{\sum x_i}$$

$$\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - m \left(\sum_{i=1}^{n} x_i \right)^2 = n \sum_{i=1}^{n} y_i x_i - m \left(n \sum x_i^2 \right)$$

$$m \left(n \sum x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right) = n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

$$m = \frac{n \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}$$



Date:

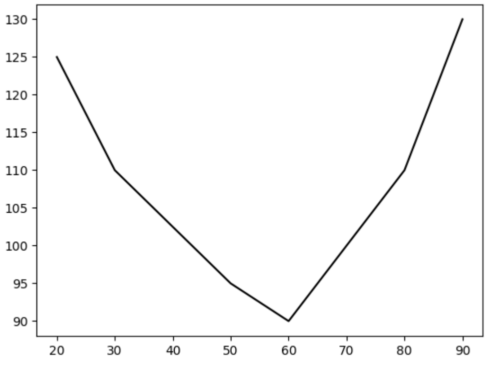
Page No.

On putting the value of m in equation ①

$$b = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \left(\frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\right)\right)$$

E. α, β, σ.

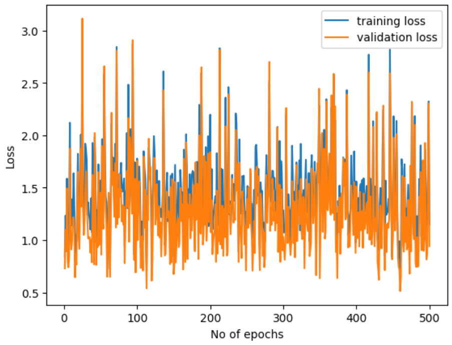
F. Y= α + β1x + β2x^2 + ε, β2 > 0 From the plot we can see that the graph is quadratic upwards therefore β2 has to be greater than 0.

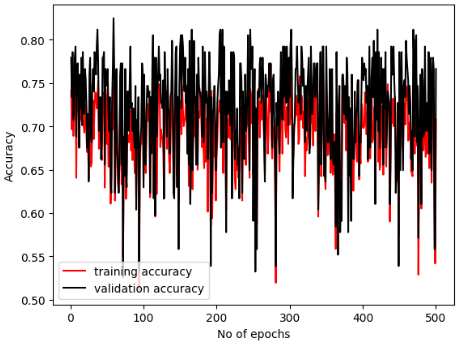


Section B:

Learning rate = 1

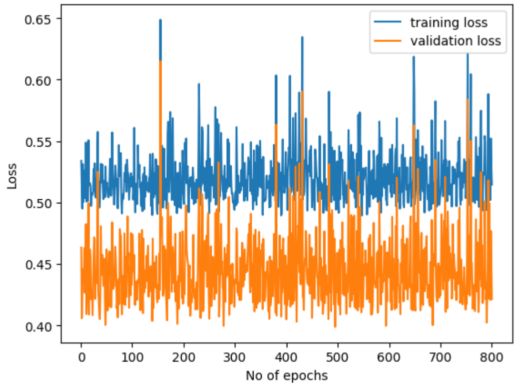
1. Confusion Matrix : [[39. 7.] [13. 18.]]
2. True negative is : 39.0
3. False positive is : 7.0
4. False negative is : 13.0
5. True positive is : 18.0
6. Recall is : 0.5806451612903226
7. Precision is : 0.72
8. F1 score is : 0.6428571428571428
9. Testing Accuracy obtained is 0.7402597402597403

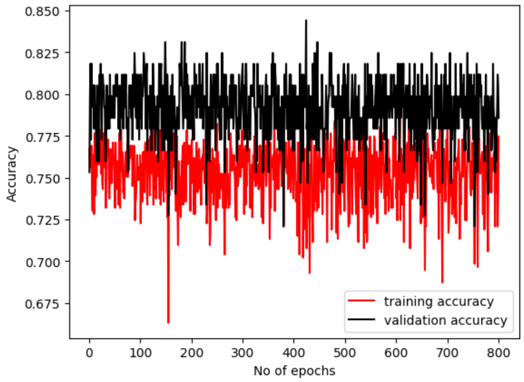




Learning Rate = 0.1

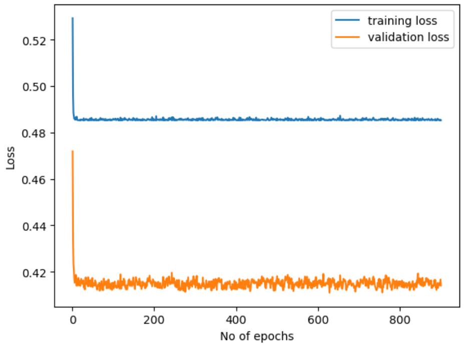
1. Confusion Matrix : [[42. 4.] [14. 17.]]
2. True negative is : 42.0
3. False positive is : 4.0
4. False negative is : 14.0
5. True positive is : 17.0
6. Recall is : 0.5483870967741935
7. Precision is : 0.8095238095238095
8. F1 score is : 0.6538461538461537
9. Testing Accuracy obtained is 0.766233766233766

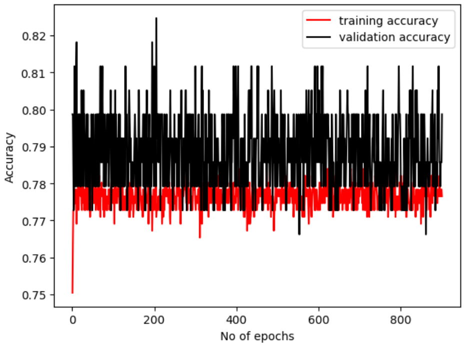




Learning Rate = 0.01

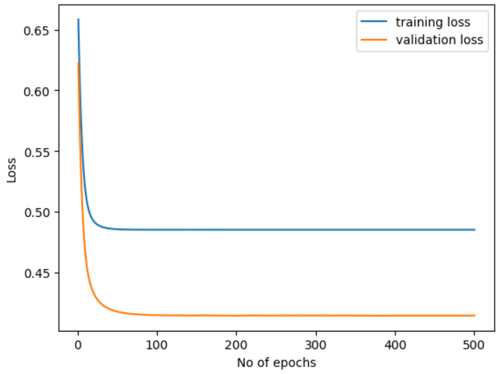
1. Confusion Matrix : [[43. 3.] [14. 17.]]
2. True negative is : 43.0
3. False positive is : 3.0
4. False negative is : 14.0
5. True positive is : 17.0
6. Recall is : 0.5483870967741935
7. Precision is : 0.85
8. F1 score is : 0.6666666666666665
9. Testing Accuracy obtained is 0.7792207792207793

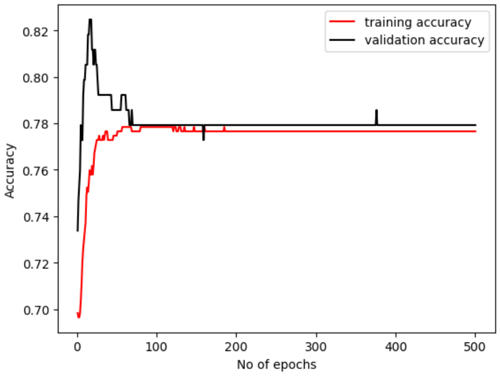




Learning Rate = 0.001

1. Model with learning rate 0.001
2. Confusion Matrix : [[43. 3.] [14. 17.]]
3. True negative is : 43.0
4. False positive is : 3.0
5. False negative is : 14.0
6. True positive is : 17.0
7. Recall is : 0.5483870967741935
8. Precision is : 0.85
9. F1 score is : 0.6666666666666665
10. Testing Accuracy obtained is 0.77922077922077



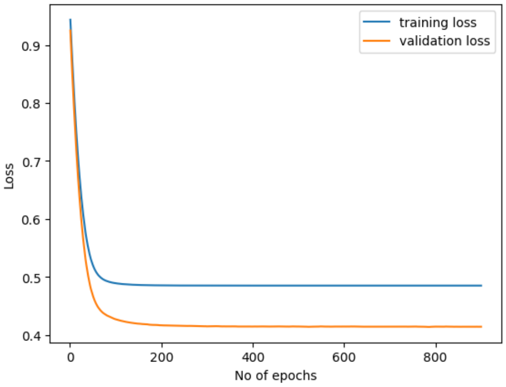


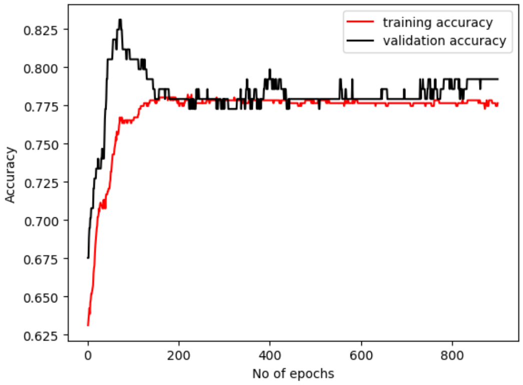
Conclusion:

Learning rates like 1 and 0.1 were too high hence the model was not able to converge properly. With learning rate 0.01 and 0.001 model provided smoother training and the model converged in fewer number of epochs with learning rate 0.01

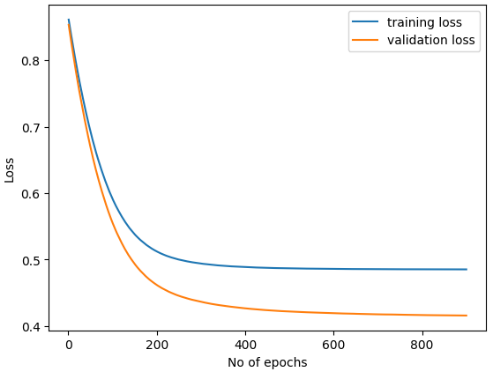
Mini-batch Gradient descent:

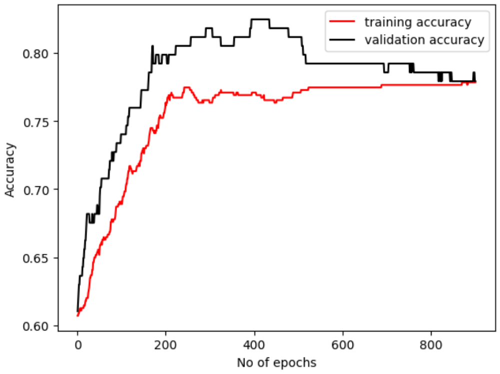
- Learning rate = 0.01 and batch size = 32



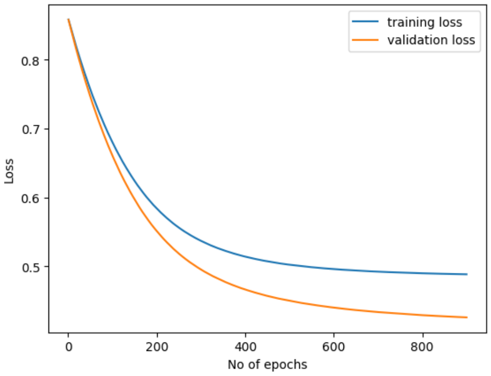


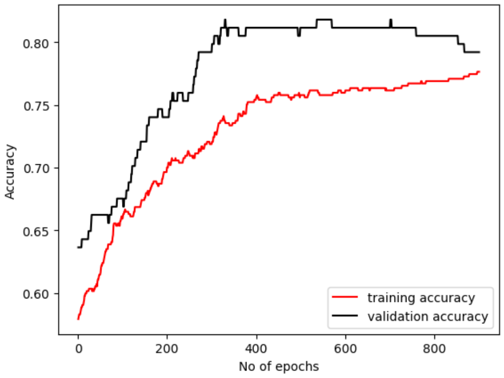
- Learning rate = 0.01 and batch size = 128





- Learning rate = 0.01 and batch size = 256



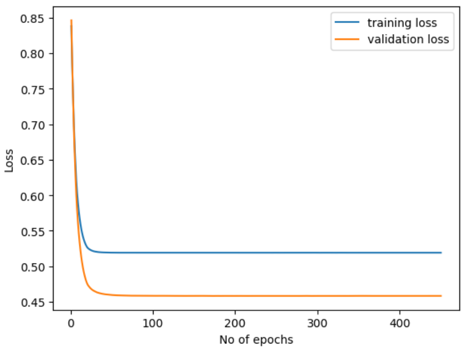


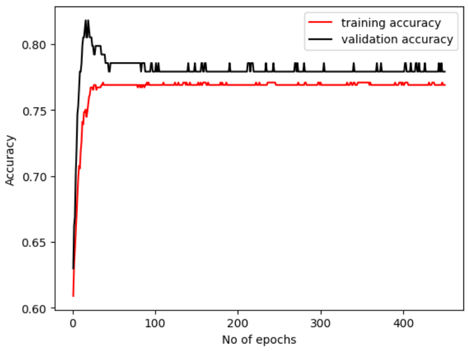
Conclusion:

As the batch size increased while keeping the learning rate same the number of epochs required by the model to converge increased subsequently which is evident by the loss curves. With SGD the model converged very quickly as compared to the Mini batch gradient descent.

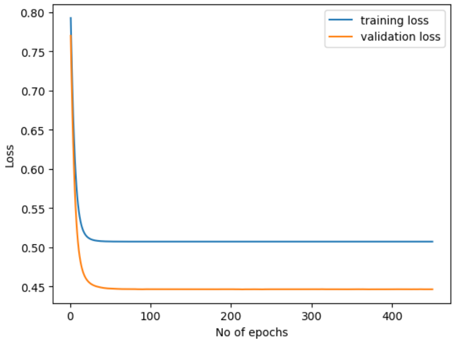
Regularization:

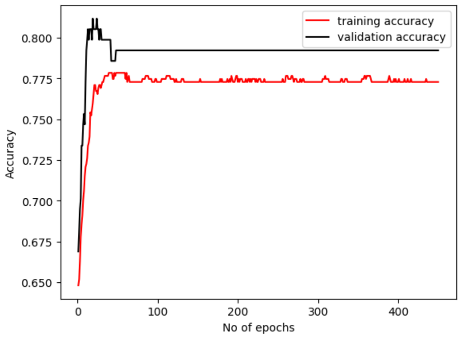
- L1 $with\ \lambda = 0.01$ and learning rate = 0.001
  1. Confusion Matrix : [[44. 2.] [13. 18.]]
  2. True negative is : 44.0
  3. False positive is : 2.0
  4. False negative is : 13.0
  5. True positive is : 18.0
  6. Recall is : 0.5806451612903226
  7. Precision is : 0.9
  8. F1 score is : 0.7058823529411764
  9. Testing Accuracy obtained is 0.805194805194805





- L2 $with\ \lambda = 0.01$ and learning rate = 0.001
  1. Confusion Matrix : [[43. 3.] [14. 17.]]
  2. True negative is : 43.0
  3. False positive is : 3.0
  4. False negative is : 14.0
  5. True positive is : 17.0
  6. Recall is : 0.5483870967741935
  7. Precision is : 0.85
  8. F1 score is : 0.6666666666666665
  9. Testing Accuracy obtained is 0.7792207792207793



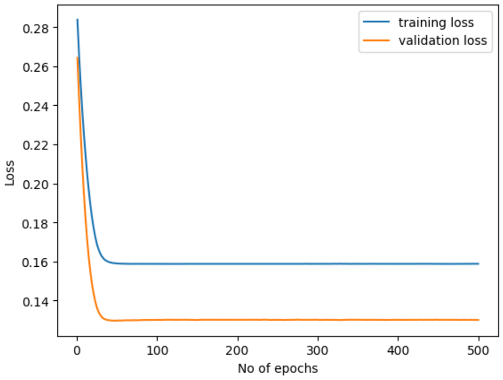


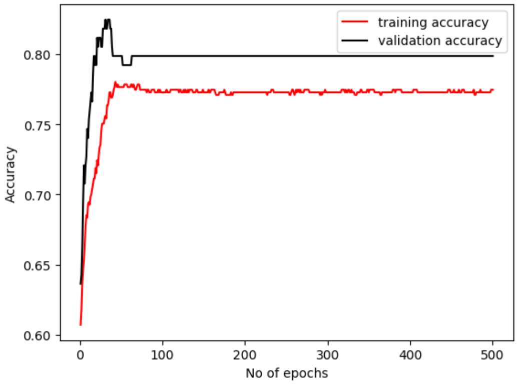
Conclusion:

The model provided best results with L1 regularization $with\ \lambda = 0.01$

Activation Function Tanh:

- With the activation function “Tanh” , learning rate = 0.001 and loss = “MSE” testing results are
    1. Confusion Matrix : [[43. 3.][13. 18.]]
    2. True negative is : 43.0
    3. False positive is : 3.0
    4. False negative is : 13.0
    5. True positive is : 18.0
    6. Recall is : 0.5806451612903226
    7. Precision is : 0.8571428571428571
    8. F1 score is : 0.6923076923076923
    9. Testing Accuracy obtained is 0.7922077922077922





Conclusion:

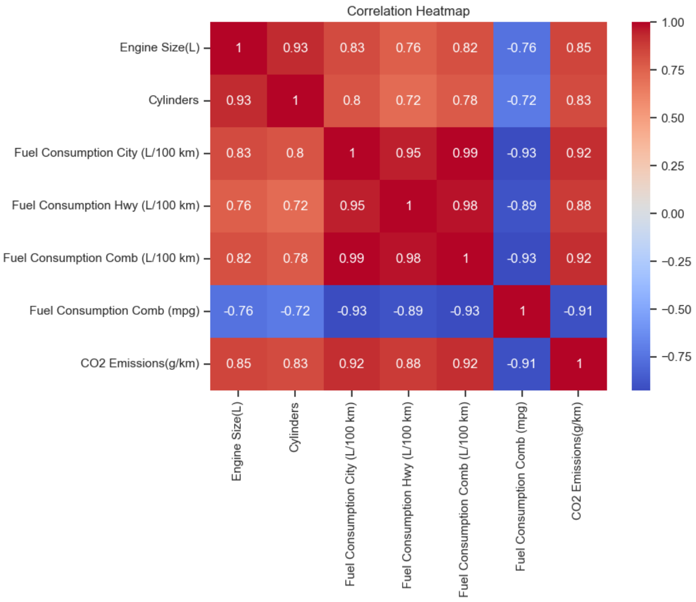
The results with tanh are almost similar to the results obtained with sigmoid function. The training and testing loss is relatively small as we have changed the loss function.

Section C :

Data Visualization Plots :

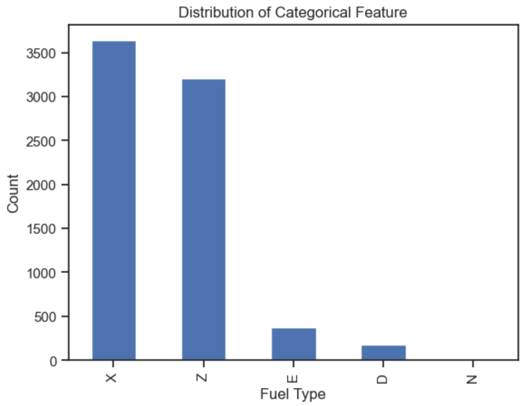
Correlation Heatmap-

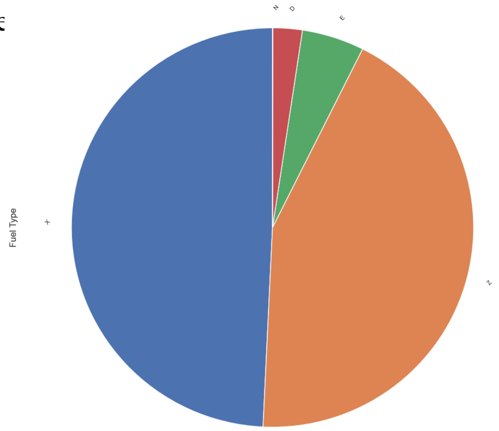
1. There is strong negative correlation between CO2 emission and Fuel consumption Comb(mpg)
2. There is strong positive correlation between CO2 emission and Fuel consumption Comb(L/100 km)



Fuel Type Histogram-

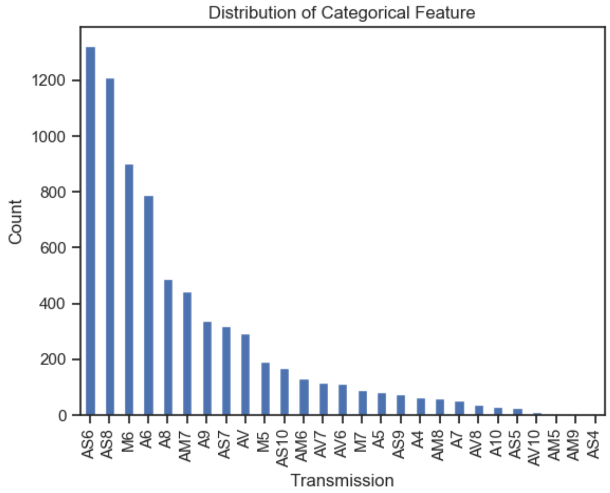
1. The histogram and pie chart shows that two fuel types ‘X’ and ‘Z’ are the most dominant fuels in the

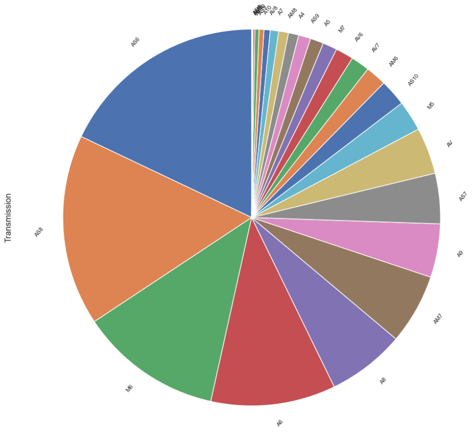




Transmission Type Histogram-

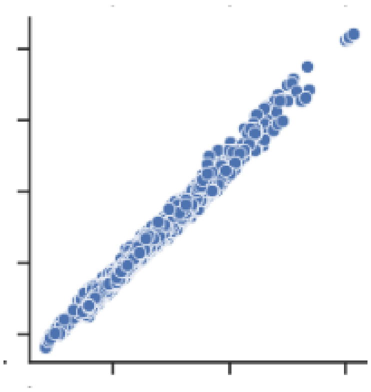
1. The histogram and pie chart shows that three to four transmission types comprise more than 50 percent of the transmissions found in the dataset.



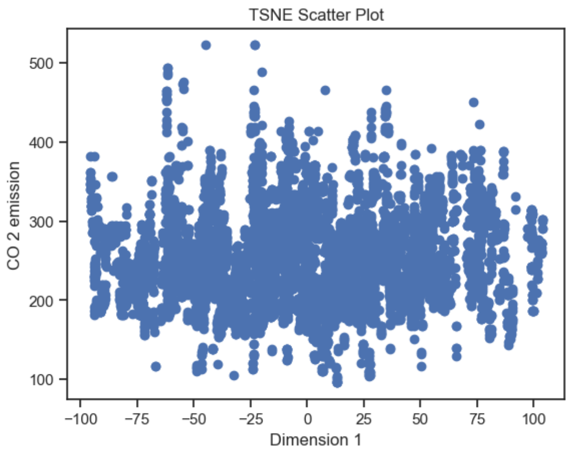


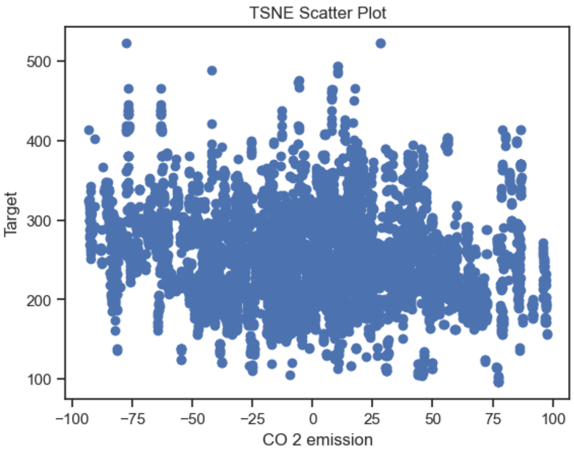
Scatter plot between Consumption Comb(L/100km) and Fuel Consumption City(L/100Km)-

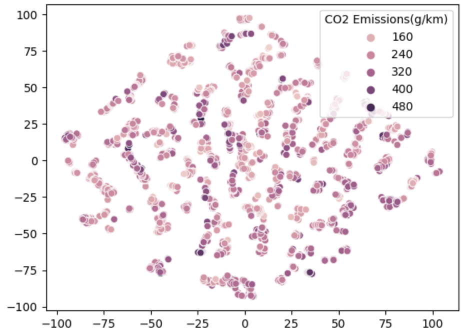
1. Shows strong positive correlation



TSNE Plots :







1. Based on the plots we can see that data is not much separable.

Label Encoder Results :

Training Results-

1. Mean squared error : 285.94899467668586
2. Root mean squared error : 16.910026454050445
3. R2 error : 0.9164537815525722
4. Adjusted R2 : 0.9162837129145113
5. Mean Absolute Error : 11.07130080002828

Testing Results-

1. Mean squared error : 295.4989301017663
2. Root mean squared error : 17.190082318062537
3. R2 error : 0.9137385977703066
4. Adjusted R2 : 0.9130315370962927
5. Mean Absolute Error : 10.98098166309119

Label Encoder + PCA Results :

Testing Results with PCA = 4 -

1. Mean squared error : 471.51203267490234
2. Root mean squared error : 21.71432782001097
3. R2 error : 0.8623572373250127
4. Adjusted R2 : 0.861889382931148
5. Mean Absolute Error : 13.923452068225142

Testing Results with PCA = 6 -

1. Mean squared error : 386.49065544035096
2. Root mean squared error : 19.659365590993797
3. R2 error : 0.8871764920587824
4. Adjusted R2 : 0.886638871530812
5. Mean Absolute Error : 11.241390802935593

Testing Results with PCA = 8-

1. Mean squared error : 301.48245429994444
2. Root mean squared error : 17.36325010762514
3. R2 error : 0.9119919004559293
4. Adjusted R2 : 0.9114519734648614
5. Mean Absolute Error : 11.326897943292986

Testing Results with PCA = 10-

1. Mean squared error : 295.87914369266986
2. Root mean squared error : 17.201137860405336
3. R2 error : 0.9136276066493343
4. Adjusted R2 : 0.9129790767333907
5. Mean Absolute Error : 11.189131669198712

Conclusion:

The model performed best with the number of features = 10.

One hot Encoder Results :

Training Results-

1. Mean squared error : 9.442829225289232
2. Root mean squared error : 3.0729186818543104
3. R2 error : 0.9972410720516441
4. Adjusted R2 : 0.9956679990986342
5. Mean Absolute Error : 1.987334823268449

Testing Results-

1. Mean squared error : 65.35694467368086
2. Root mean squared error : 8.084364209613572
3. R2 error : 0.9809211434672247
4. Adjusted R2 : 1.042093261946751
5. Mean Absolute Error : 6.026219870674243

Conclusion:

One hot encoding has better results than label encoding

One hot Encoder + PCA Results :

Training Results with PCA = 300-

1. Mean squared error : 21.888900455692358
2. Root mean squared error : 4.678557518690174
3. R2 error : 0.9936046816282289
4. Adjusted R2 : 0.9932613011733764
5. Mean Absolute Error : 2.8586923425202015

Testing Results with PCA = 300-

1. Mean squared error : 21.569339955435982
2. Root mean squared error : 4.644280348497061
3. R2 error : 0.9937035253931915
4. Adjusted R2 : 0.99209055615349
5. Mean Absolute Error : 3.105308716391837

Training Results with PCA = 800-

1. Mean squared error : 18.248099074321935
2. Root mean squared error : 4.271779380342802
3. R2 error : 0.9946684209425622
4. Adjusted R2 : 0.9938320333935987
5. Mean Absolute Error : 2.686278289433396

Testing Results with PCA = 800-

1. Mean squared error : 20.76571310750946
2. Root mean squared error : 4.556941200795711
3. R2 error : 0.9939381183873106
4. Adjusted R2 : 0.9867446855402525
5. Mean Absolute Error : 3.1395780358243375

Training Results with PCA = 1000-

1. Mean squared error : 16.475258144350015
2. Root mean squared error : 4.058972547868489
3. R2 error : 0.9951863949811681
4. Adjusted R2 : 0.9942042468719445
5. Mean Absolute Error : 2.5905482799850117

Testing Results with PCA = 1000-

1. Mean squared error : 20.508018601547107
2. Root mean squared error : 4.528577988899729
3. R2 error : 0.99401334400462
4. Adjusted R2 : 0.9813972542122509
5. Mean Absolute Error : 3.123252883870307

Training Results with PCA = 1500-

1. Mean squared error : 13.837011520630417
2. Root mean squared error : 3.719813371747354
3. R2 error : 0.9959572161165692
4. Adjusted R2 : 0.9945799536088457
5. Mean Absolute Error : 2.2929696948787432

Testing Results with PCA = 1500-

1. Mean squared error : 21.012838002148133
2. Root mean squared error : 4.583976221813125
3. R2 error : 0.9938659782278519
4. Adjusted R2 : 1.3621526454276212
5. Mean Absolute Error : 3.0702433757425784

Conclusion:

With pca = 300 best results were achieved as the training and testing MSE were close to each other and had desirable values.

Regularization :

Testing Results with L1 (λ= 0.02)-

1. Mean squared error : 20.070030000456335
2. Root mean squared error : 4.479958705217754
3. R2 error : 0.9941412006803708
4. Adjusted R2 : 1.0129454907122346
5. Mean Absolute Error : 2.897265795886767

Testing Results with L1 (λ= 0.02)-

1. Mean squared error : 21.053579588384608
2. Root mean squared error : 4.588417983181634
3. R2 error : 0.9938540850329879
4. Adjusted R2 : 1.0135798959450746
5. Mean Absolute Error : 3.0299144244163547

Conclusion:

On applying regularization techniques like L1 and L2 the MSE significantly dropped from 290 in part(c) to 20-21 indicating that regularization helped in better training of the model.

SGDRegressor :

Testing Results-

1. Mean squared error : 1.2915086651704689e+18
2. Root mean squared error : 1136445627.8988752
3. R2 error : -377014388549686.3
4. Adjusted R2 : 833043768711585.6
5. Mean Absolute Error : 792468877.7844507

Conclusion:

Very high testing loss reported indicating that training was not efficient.