APPLIED MACHINE LEARNING – ASSIGNMENT 1

INTRODUCTION

In this fast-paced world of technology data is growing at an astounding rate. For this assignment, I have worked on a social networking service data from Facebook to recognize the patterns in the user's behavior. The Facebook dataset has 54 variables and the Target Variable is the number of comments in next H hours. The given dataset has 5 variants of the training data, so I have taken 'Feature_Variant_1' for this assignment. I have combined the 'Feature_Variant_1' dataset and the 10 'test_case' datasets in one dataset named 'Fb_Data' (present in the zip folder). As required I have randomly partitioned the data into train and test set using 70/30 split.

Exploratory Analysis

Performed feature scaling to bring all the features to the same level of magnitude. To choose the features for the predictive model, I found correlation of all the variables with 'Target Variable'.

```
PageCheckin               0.018201
Pagetalking about         0.166511
PageCategory             -0.072051
MinCC1                    0.165276
MaxCC1                    0.229230
AverageCC1                0.338254
MedianCC1                 0.323717
SDCC1                     0.303257
MinCC2                    0.136127
MaxCC2                    0.217531
AverageCC2                0.352975
MedianCC2                 0.325066
SDCC2                     0.296415
MinCC3                    0.016874
MaxCC3                    0.216926
AverageCC3                0.309831
MedianCC3                 0.242777
SDCC3                     0.292286
MinCC4                    0.171789
MaxCC4                    0.224647
AverageCC4                0.338322
MedianCC4                 0.324047
SDCC4                     0.299378
MinCC5                   -0.194838
MaxCC5                    0.212759
AverageCC5                0.194769
MedianCC5                 0.073851
SDCC5                     0.305467
CC1                       0.328436
CC2                       0.503889
CC3                       0.076870
CC4                       0.342349
CC5                       0.367229
Basetime                 -0.215861
Postlength               -0.001781
PostShareCount            0.130822
PostPromotionStatus            NaN
HLocal                   -0.026526
PostpublishedSunday       0.009418
PostpublishedMonday      -0.003350
PostpublishedTuesday     -0.003767
PostpublishedWednesday    0.014085
PostpublishedThursday    -0.004047
PostpublishedFriday      -0.004123
PostpublishedSaturday    -0.008059
BaseDateTimeSunday       -0.006501
BaseDateTimeMonday        0.001374
BaseDateTTuesday         -0.004792
BaseDateTimeWednesday     0.010746
BaseDateTimeThursday      0.004371
BaseDateTimeFriday       -0.001923
BaseDateTimeSaturday     -0.003400
TargetVariable            1.000000
Name: TargetVariable, dtype: float64
```

Fig. 1

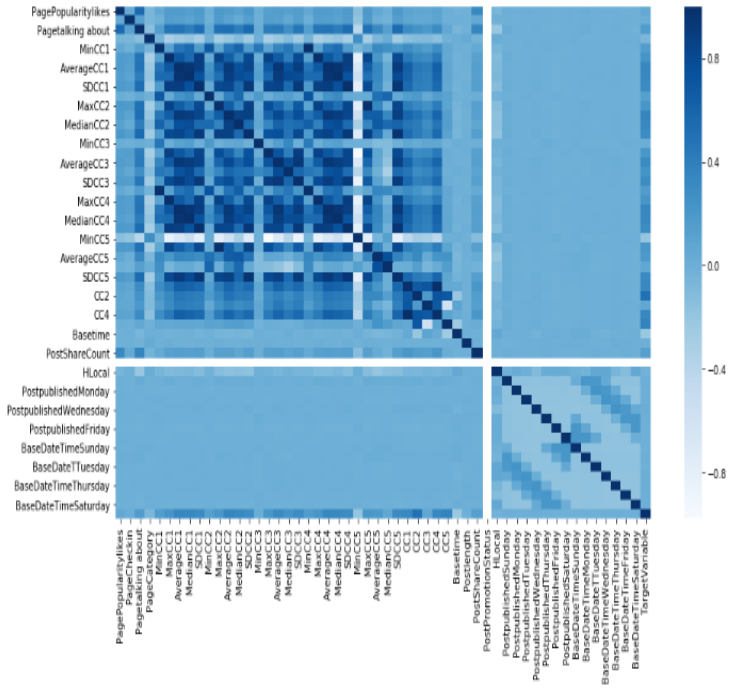


Fig. 2

Experiment 1

In this experiment I have taken various values of learning rate, please find the results below:

- **Learning Rate is equal to 1 –** If we choose learning rate to be large, gradient descent can overshoot the minimum. It may fail to converge or even diverge (as we can see from the graph below)

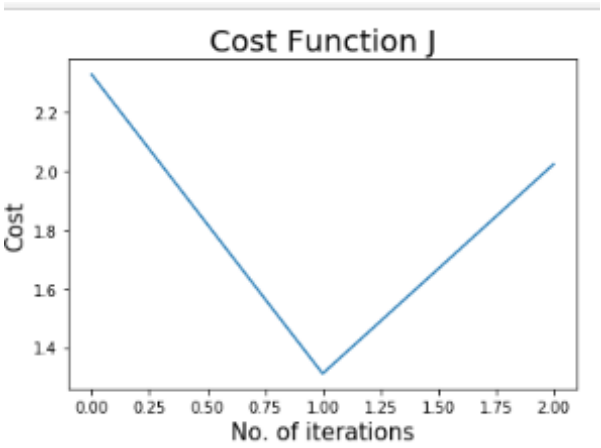


Fig. 3

Mean Square Error using equation of hyperplane for training set: 6.864
Mean Square Error from Gradient Descent prediction: 4.045
Mean Square Error using equation of hyperplane for test set: 4.987

- **Learning Rate is equal to 0.1 –**

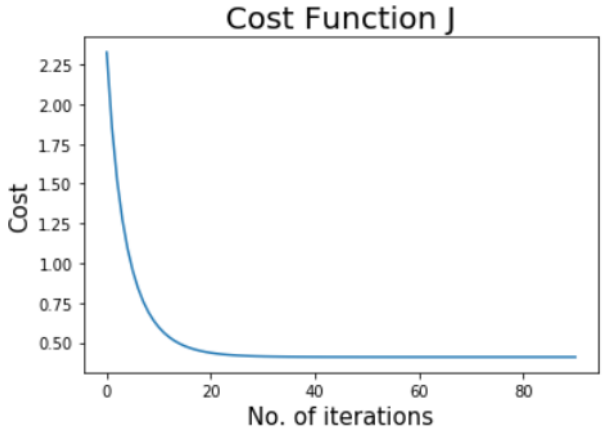
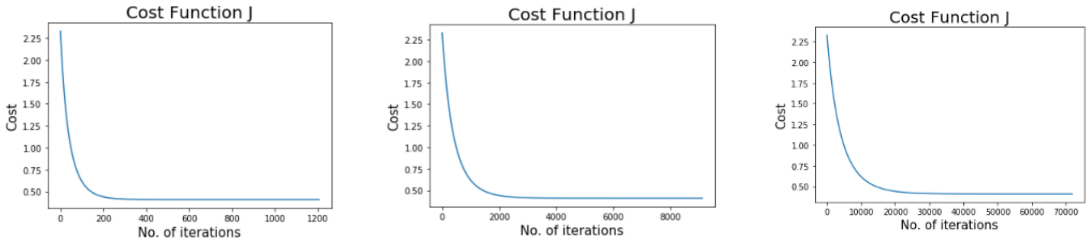


Fig. 4

Mean Square Error using equation of hyperplane: 0.861
Mean Square Error from Gradient Descent prediction: 0.819
Mean Square Error using equation of hyperplane for test set: 0.878

- **Learning Rate is equal to 0.01/0.001/0.0001 –** If we choose alpha to be very small, gradient descent will take small steps to reach minima and will take a longer time to

reach minima.



Alpha = 0.01 **Alpha = 0.001** **Alpha = 0.0001**

Fig. 5

We can take alpha = 0.1 or 0.01 for this assignment as we can see from the graph that the value of cost is decreasing, and gradient descent is not taking a long time or large number of iterations to reach the minima. I have taken learning rate equal to 0.01

Experiment 2

As we can see from the Figure below (Fig. 6) as we increase the precision the error for both train and test set decreases at first and then it stabilizes after a certain threshold value. I have taken 1e-7 as the threshold value for this assignment as the error tend to stabilize after that value.

Threshold Values	Train Error	Test Error
1.E+00	2.568	2.236
1.E-01	1.278	1.268
1.E-02	0.91	0.933
1.E-03	0.895	0.917
1.E-04	0.868	0.887
1.E-05	0.861	0.879
1.E-06	0.861	0.878
1.E-07	0.861	0.878
1.E-08	0.861	0.878
1.E-09	0.861	0.878
1.E-10	0.861	0.878

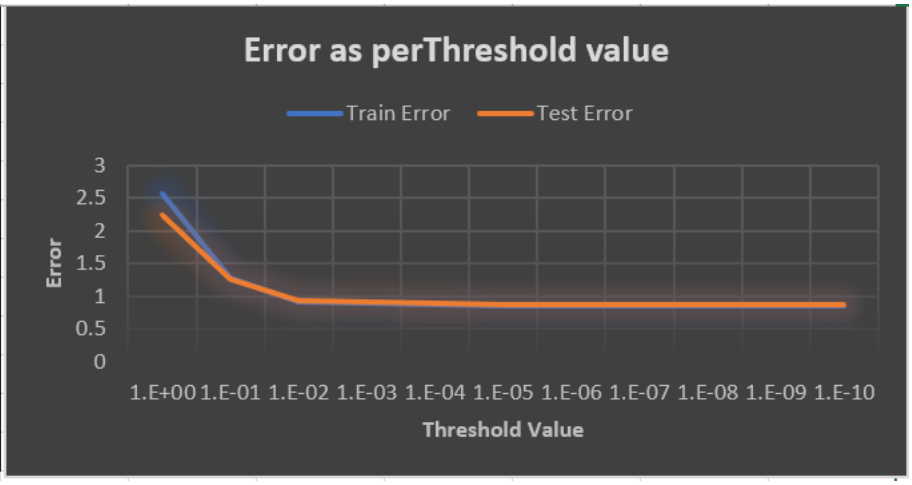


Fig. 6

Below is the graph for the test and train error as a function of number of gradient descent iterations:

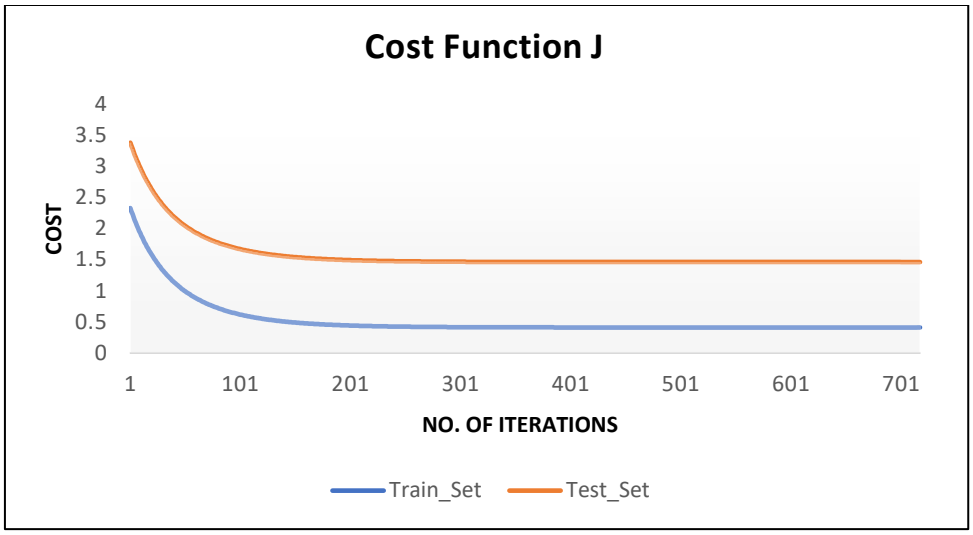


Fig. 7

Experiment 3

In this experiment I have randomly chosen five features and retrained the model and calculated the error for the train and test set. I selected the following five features randomly: Postlength, PageCategory, PagePopularitylikes, BaseDateTTuesday, PostpublishedThursday.

The ten features in the original model are: AverageCC2, PageCheckin, PagePopularitylikes, PostShareCount, PostpublishedSunday, BaseDateTimeFriday, PostpublishedWednesday, Pagetalking about, PageCategory, Basetime

Below is the mean squared error for the randomly selected model features:

Mean Square Error using equation of hyperplane for train set: 0.971
Mean Square Error from Gradient Descent prediction for train set: 0.971
Mean Square Error using equation of hyperplane for test set: 1.038

Below is the mean squared error for the original model having 10 features:

Mean Square Error using equation of hyperplane: 0.861
Mean Square Error from Gradient Descent prediction: 0.819
Mean Square Error using equation of hyperplane for test set: 0.878

Experiment 4

In this experiment I have chosen top five features and retrained the model and calculated the error for the train and test set. I selected the following five features randomly: Basetime, CC2, CC1, Pagetalking about, CC4.

Below is the mean squared error for the top five selected model features:

Mean Square Error using equation of hyperplane: 0.707
Mean Square Error from Gradient Descent prediction: 0.707
Mean Square Error using equation of hyperplane for test set: 0.78

Below is the comparison between the train and test errors for the three models that I used for experimentation:

Experiment	MSE Train	MSE Test
10 features	0.819	0.878
5 Randomly selected Features	0.971	1.038
Top 5 features	0.707	0.78

Fig. 8

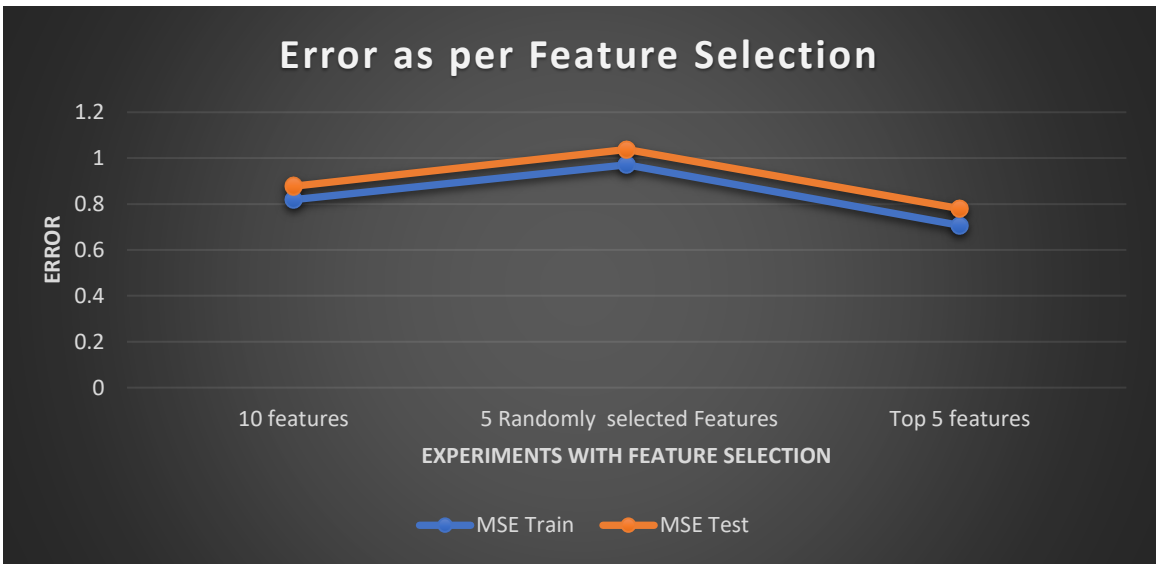


Fig. 9

As we can see from the graph (Fig. 9) and the table (Fig. 8) above the error is minimum when I consider top 5 features in the model as compared to when I take 5 randomly selected features. This is because in case of top 5 features I have taken the variables that have high correlation with the Target Variable, as a result the predicted values for the Target Variable are closer to the actual values and the error is less.

We can also see that the error for the model with top 5 features is less than the model with 10 features because the features in the 10-feature model are not that strongly corelated with the Target Variable as compared to the features in the top 5 priority model as a result that model fits the data more accurately with less training and testing error.

Model Interpretations

Top 5 feature model: Below is the final model equation with all the parameters:

Target Variable = 0.0007 -0.120*Basetime + 0.492*CC2 + 0.460*CC1 +
0.041*Pagetalking about -0.481*CC4

- The intercept term is almost equal to zero in this case.
- For every one unit increase in the Basetime the number of comments in the next H hour decreases by 0.12 units keeping other factors constant.
- For every one unit increase in the number of comments in last 24 hours, relative to the base date/time (CC2) the number of comments in next H hour increases by 0.492 keeping other factors constant.
- For every one unit increase in the total number of comments before selected base date/ time (CC1) the number of comments in next H hour increases by 0.46 keeping other factors constant.
- For every one unit increase in the daily interested individuals (Pagetalking about) the number of comments in next H hour increases by 0.041 keeping other factors constant.
- For every one unit increase in the number of comments in the first 24 hours after the publication of the post but before base date/time (CC4), the number of comments in next H hour decreases by 0.481 keeping other factors constant.

10 Feature Model - Below is the final model equation with all the parameters:

Target Variable = 0.0012 + 0.352* AverageCC2 - 0.016*PageCheckin - 0.002*PagePopularitylikes + 0.086*PostShareCount + 0.006*PostpublishedSunday - 0.008*BaseDateTimeFriday + 0.014*PostpublishedFriday - 0.057*Pagetalking about - 0.007*PageCategory - 0.203*Basetime

- The intercept term is almost equal to zero in this case.
- For every one unit increase in the average number of comments in the last 24 hours, relative to the base date/time (AverageCC2) the number of comments in the next H hou rs increase by 0.352 units keeping other factors constant.
- For every one unit increase in the number of individuals who so far visited the place (PageCheckin), the number of comments in the next H hours decreases by 0.016 units ke eping other factors constant.
- For every one unit increase in the popularity or support for the source of the document (PagePopularitylikes), the number of comments in the next H hours decreases by 0.002 units keeping other factors constant.
- For every one unit increase in the number of people who shared the post (PostShareCou nt), the number of comments in the next H hours increase by 0.086 units keeping other factors constant.
- For every one unit increase in the number of posts published on Sunday (PostpublishedSunday), the number of comments in the next H hours increase by 0.006 units keeping other factors constant.
- For every one unit increase in the number of Fridays selected as base date/time (BaseDateTimeFriday), the number of comments in the next H hours decreases by 0.008 units keeping other factors constant.
- For every one unit increase in the number of posts published on Friday

(PostpublishedFriday), the number of comments in the next H hours increases by 0.014 units keeping other factors constant.

- For every one unit increase in the number of daily interested individuals (Pagetalking about), the number of comments in the next H hours decreases by 0.057 units keeping other factors constant.
- For every one unit increase in the value of the category of page (PageCategory), the number of comments in the next H hours decreases by 0.007 units keeping other factors constant.
- For every one unit increase in the Basetime, the number of comments in the next H hou rs decreases by 0.203 units keeping other factors constant.

CONCLUSION

Some of the features important for predicting the number of comments in the next H hours are : The total number of comments before selected base date/time(CC1), The number of comment s in last 24 hours, relative to base date/time (CC2), The number of comments in last 48 to last 2 4 hours relative to base date/time (CC3), The number of comments in the first 24 hours after th e publication of post but before base date/time (CC4), Page talking about, Post Share Count. Several factors if considered can help to improve the model performance:

- Better feature selection – Can do more research by exploratory analysis and considering the collinearity between the independent variables to have more accurate predictions.
- Flexible Model – Increasing the flexibility of the model might improve the model performance.
- Considering more variants of data or gathering more data from different sources.
- Resampling or cross validation might help to create a more generalized model.