EE219 Project 5 – Report
Popularity Prediction on Twitter

Fangyao Liu 204945018

Xuan Hu 505031796

Yanzhe Xu 404946757

Zhechen Xu 805030074

Content

Introduction

Twitter, with its public discussion model, is a good platform to predict future popularity of a subject or event. Based on the data of current (and previous) tweet activity for a hashtag, we are able to predict if such hashtag will be trendy and if so how much in the future.

In this project, we are given a set of data which is collected by querying popular hashtags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. The test data consists of tweets containing a hashtag in a specified time window. We use the test data to train a regression model, extract feature for each training set, and then the model is used for predicting popularity of each hashtag.

Part 1: Popularity Prediction

Problem 1.1

In this problem, we calculated the following statics for each hashtag: Average number of tweet per hour, average number of followers of users posting the tweets, average number of retweets

Hashtag	Average number of tweets per hour	Average number of followers of users posting the tweets	Average number of retweets
#gohawks	324.932642	2203.931767	2.014617
#gopatriots	45.620870	1401.895509	1.400084
#nfl	441.267462	4653.252286	1.538533
#patriots	834.264055	3309.978828	1.782816
#sb49	1418.440823	10267.316849	2.511149
#superbowl	2297.729131	8858.974663	2.388272

Table 1.1 Hashtag statistics

Bar plot for #nfl **and** #superbowl:

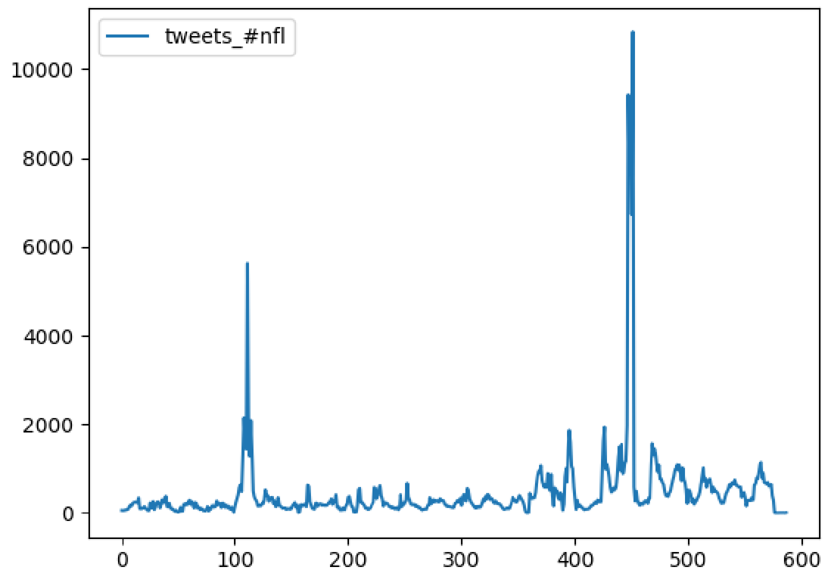

Figure 1.1 number of tweets in hour over a period for NFL

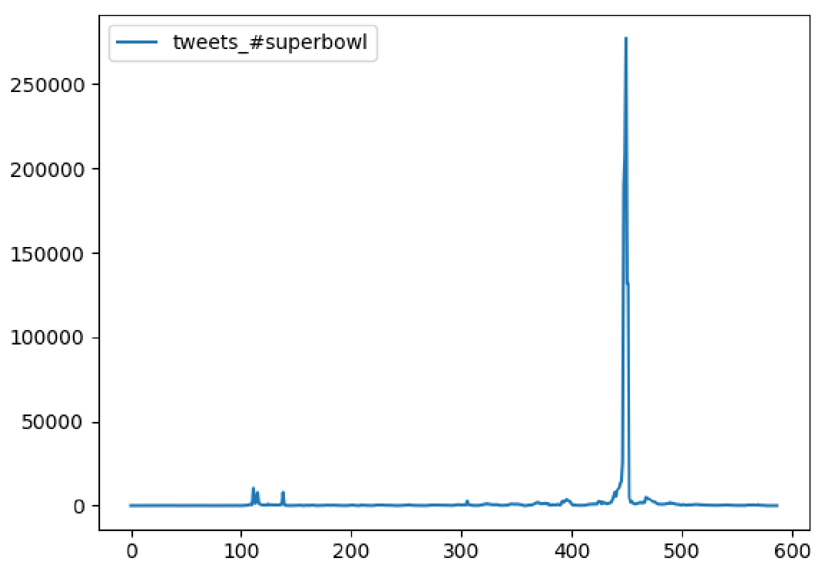

Figure 1.2 number of tweets in hour over a period for SuperBowl

Problem 1.2

In this problem, we fitted a linear regression model using five features to predict numbers of tweets in the next hour with features extracted from tweet data in the previous hour. The features we used are: number of tweets, total number of retweets, sum of the number of followers of the users posting the hashtag, maximum number of followers of the users posting the hashtag and time of the day.

Training accuracy and R-squared measure(in summary), p-value, t-test and summarized reports for each hashtag are as follows

1. #gohawks

```
rmse =  1923.953640
p_values:
[  1.32888306e-12   3.66024117e-03   3.99884136e-02   8.60890256e-01
   8.00905139e-03]
t_test :
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            79.2326     29.177      2.716      0.007      21.925     136.540
==============================================================================
summary
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.501
Model:                            OLS   Adj. R-squared:                  0.496
Method:                 Least Squares   F-statistic:                     114.8
Date:                Mon, 19 Mar 2018   Prob (F-statistic):           5.44e-84
Time:                        19:39:32   Log-Likelihood:                -4797.7
No. Observations:                 578   AIC:                             9605.
Df Residuals:                     573   BIC:                             9627.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.2319      0.170      7.253      0.000       0.898       1.566
x2            -0.1286      0.044     -2.918      0.004      -0.215      -0.042
x3            -0.0002   8.52e-05     -2.059      0.040      -0.000   -8.04e-06
x4           2.82e-05      0.000      0.175      0.861      -0.000       0.000
x5             8.8147      3.313      2.661      0.008       2.308      15.321
==============================================================================
Omnibus:                      910.878   Durbin-Watson:                   2.220
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           777761.522
Skew:                           8.574   Prob(JB):                         0.00
Kurtosis:                     181.887   Cond. No.                     2.33e+05
==============================================================================
```

Figure 1.3 rmse, r-squared measure(in summary), p-value, t-test and OLS Regression Results of #gohawks

2. #gopatriots

```
rmse =  68.412972
p_values:
[ 0.7645813   0.02569241  0.21021857  0.05534974  0.30144349]
t_test :
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0             0.7687      0.484      1.589      0.113      -0.182       1.719
==============================================================================
summary
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.640
Model:                            OLS   Adj. R-squared:                  0.637
Method:                 Least Squares   F-statistic:                     202.1
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          1.27e-123
Time:                        19:39:38   Log-Likelihood:                -3811.2
No. Observations:                 574   AIC:                             7632.
Df Residuals:                     569   BIC:                             7654.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.0765      0.255     -0.300      0.765      -0.578       0.425
x2             0.5009      0.224      2.237      0.026       0.061       0.941
x3             0.0003      0.000      1.254      0.210      -0.000       0.001
x4            -0.0004      0.000     -1.920      0.055      -0.001    8.76e-06
x5             0.7155      0.692      1.034      0.301      -0.643       2.074
==============================================================================
Omnibus:                      505.273   Durbin-Watson:                   1.951
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           300435.719
Skew:                           2.726   Prob(JB):                         0.00
Kurtosis:                     114.947   Cond. No.                     3.74e+04
==============================================================================
```

Figure 1.4 rmse, r-squared measure(in summary), p-value, t-test and OLS Regression Results of #gopatriots

3. #nfl

```
rmse =  1098.719217
p_values:
[  4.18847544e-08   5.46829187e-03   2.81994617e-03   4.40073263e-02
   6.72008220e-04]
t_test :
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            57.3777     16.582      3.460      0.001      24.809      89.946
==============================================================================
summary
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.647
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     213.4
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          5.59e-129
Time:                        19:40:25   Log-Likelihood:                -4565.4
No. Observations:                 586   AIC:                             9141.
Df Residuals:                     581   BIC:                             9163.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.7404      0.133      5.557      0.000       0.479       1.002
x2            -0.1785      0.064     -2.789      0.005      -0.304      -0.053
x3          7.889e-05   2.63e-05      3.000      0.003    2.72e-05       0.000
x4         -7.259e-05    3.6e-05     -2.018      0.044      -0.000   -1.95e-06
x5             7.5364      2.204      3.419      0.001       3.207      11.865
==============================================================================
Omnibus:                      566.966   Durbin-Watson:                   2.326
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           349036.837
Skew:                           3.275   Prob(JB):                         0.00
Kurtosis:                     122.382   Cond. No.                     4.26e+05
==============================================================================
```

Figure 1.5 rmse, r-squared measure(in summary), p-value, t-test and OLS Regression Results of #nfl

4. #patriots

```
rmse =  3257.032671
p_values:
[  1.45577505e-33   1.40438838e-01   9.63986171e-01   7.73088162e-02
   6.53276897e-01]
t_test :
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            16.2746     34.312      0.474      0.635     -51.117      83.666
==============================================================================
summary
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       4.245
Model:                            OLS   Adj. R-squared:                  4.273
Method:                 Least Squares   F-statistic:                    -152.0
Date:                Mon, 19 Mar 2018   Prob (F-statistic):               1.00
Time:                        19:41:59   Log-Likelihood:                -5422.9
No. Observations:                 586   AIC:                         1.086e+04
Df Residuals:                     581   BIC:                         1.088e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.9214      0.072     12.878      0.000       0.781       1.062
x2            -0.0871      0.059     -1.476      0.140      -0.203       0.029
x3         -1.185e-06   2.62e-05     -0.045      0.964   -5.27e-05    5.03e-05
x4             0.0002      0.000      1.770      0.077   -1.97e-05       0.000
x5             3.9266      8.736      0.449      0.653     -13.232      21.086
==============================================================================
Omnibus:                      878.850   Durbin-Watson:                   1.994
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           692543.929
Skew:                           7.765   Prob(JB):                         0.00
Kurtosis:                     170.697   Cond. No.                     7.66e+05
==============================================================================
```

Figure 1.6 rmse, r-squared measure(in summary), p-value, t-test and OLS Regression Results of #patriots

5. #sb49

```
rmse =  8858.740024
p_values:
[  7.14133776e-32   1.44772581e-02   1.83069811e-01   3.50380805e-02
   7.97742682e-01]
t_test :
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            18.0764     64.808      0.279      0.780    -109.211     145.364
==============================================================================
summary
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                     -16.789
Model:                            OLS   Adj. R-squared:                -16.944
Method:                 Least Squares   F-statistic:                    -108.9
Date:                Mon, 19 Mar 2018   Prob (F-statistic):               1.00
Time:                        19:44:02   Log-Likelihood:                -5717.9
No. Observations:                 582   AIC:                         1.145e+04
Df Residuals:                     577   BIC:                         1.147e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.1886      0.095     12.494      0.000       1.002       1.375
x2            -0.2151      0.088     -2.453      0.014      -0.387      -0.043
x3          1.869e-05     1.4e-05     1.333      0.183   -8.85e-06    4.62e-05
x4             0.0001   4.77e-05      2.113      0.035    7.09e-06       0.000
x5            -4.0764     15.900     -0.256      0.798     -35.304      27.152
==============================================================================
Omnibus:                     1181.809   Durbin-Watson:                   1.682
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2232492.092
Skew:                          14.651   Prob(JB):                         0.00
Kurtosis:                     304.998   Cond. No.                     7.51e+06
==============================================================================
```

Figure 1.7 rmse, r-squared measure(in summary), p-value, t-test and OLS Regression Results of #sb49

6. #superbowl

```
tweets_#superbowl.txt
rmse =  13775.045426
p_values:
[ 8.06131086e-115   4.38995558e-015   6.22750477e-012   7.18710969e-008
   1.91380719e-001]
t_test :
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0          1521.1962   1158.875      1.313      0.190    -754.899    3797.291
==============================================================================
summary
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                      93.136
Model:                            OLS   Adj. R-squared:                 93.929
Method:                 Least Squares   F-statistic:                    -117.5
Date:                Mon, 19 Mar 2018   Prob (F-statistic):               1.00
Time:                        19:47:43   Log-Likelihood:                -6098.3
No. Observations:                 586   AIC:                         1.221e+04
Df Residuals:                     581   BIC:                         1.223e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             2.3014      0.079     28.962      0.000       2.145       2.457
x2            -0.2899      0.036     -8.059      0.000      -0.361      -0.219
x3            -0.0001   1.87e-05     -7.020      0.000      -0.000   -9.46e-05
x4             0.0008      0.000      5.457      0.000       0.000       0.001
x5           -38.9335     29.765     -1.308      0.191     -97.394      19.527
==============================================================================
Omnibus:                     1012.648   Durbin-Watson:                   2.317
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1838518.315
Skew:                          10.124   Prob(JB):                         0.00
Kurtosis:                     276.656   Cond. No.                     1.09e+07
==============================================================================
```

Figure 1.8 rmse, r-squared measure(in summary), p-value, t-test and OLS Regression Results of #superbowl

From previous part, we can see other four attributes have relatively big p value and don't make a big different in prediction. However, number of tweets per hour play a important role in this procedure.

Problem 1.3

In this problem, the features we choose are "number of tweets", "sum of favorites count", "max number of favorite count", "ranking score" and "sum of friends count". Here we use p-value to evaluate the importance of each feature. We choose three attributes which have smaller p-value and plot their scatter plots relate to submitted tweets. It shows relatively linear relationship.

tweets_#gohawks.txt

```
rmse =  1839.422911
p_values:
[  1.83169955e-01   8.65740236e-17   1.63862576e-06   8.79122252e-01
   2.10870978e-02]
most three important features:
 sum of favourites_count    max number of favourite_count     sum of friends_count
```

Most three important features’ scatter:

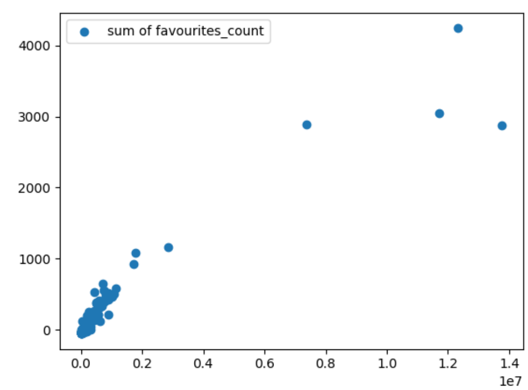

Figure 1.9 scatter plot of sum of favourites_count

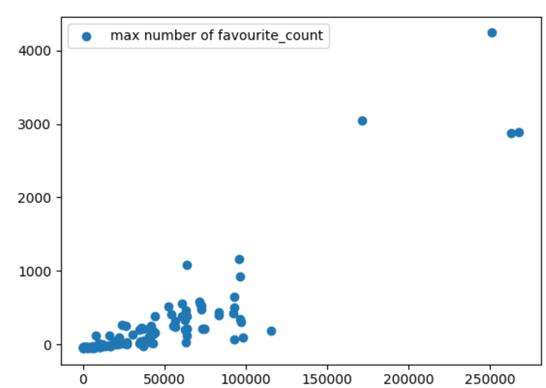

Figure 1.10 scatter plot of max number of favourite_count

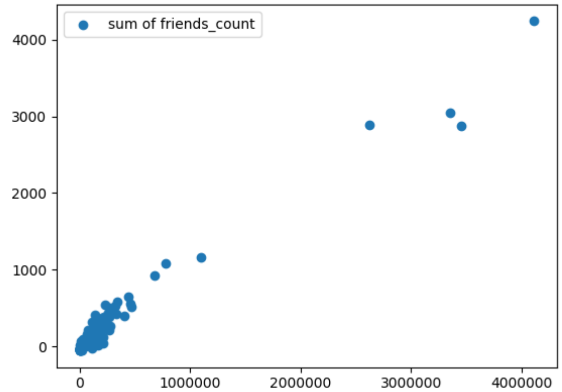

Figure 1.11 scatter plot of sum of friends_count

tweets_#gopatriots.txt

```
rmse =  109.972915
p_values:
[  5.69996104e-06   2.05291498e-84   7.09862397e-25   3.30060482e-01
   2.62159181e-11]
most three important features:
 sum of favourites_count    max number of favourite_count     sum of friends_count
```

Most three important features' scatter:

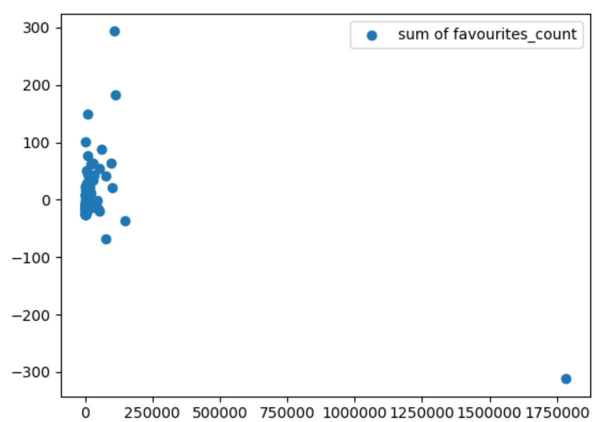

Figure 1.12 scatter plot of sum of favourites_count

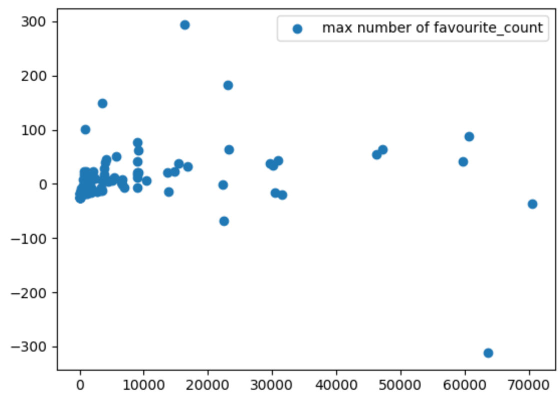

Figure 1.13 scatter plot of max number of favourite_count

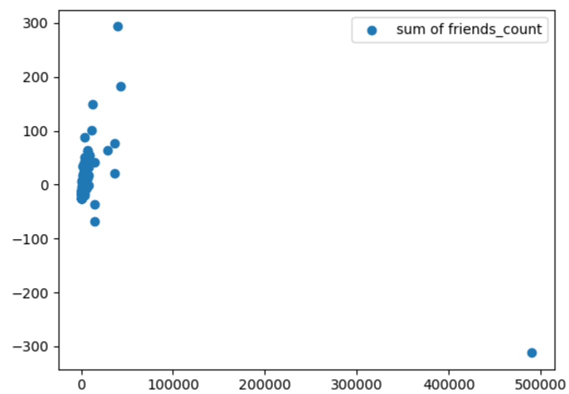

Figure 1.14 scatter plot of sum of friends_count

tweets_#nfl.txt

```
rmse =  1012.918621
p_values:
[  8.68559218e-05   1.05844221e-02   5.52847811e-12   9.40697072e-05
   1.05862706e-03]
most three important features:
 max number of favourite_count    number of tweets     ranking_score
```

Most three important features' scatter:

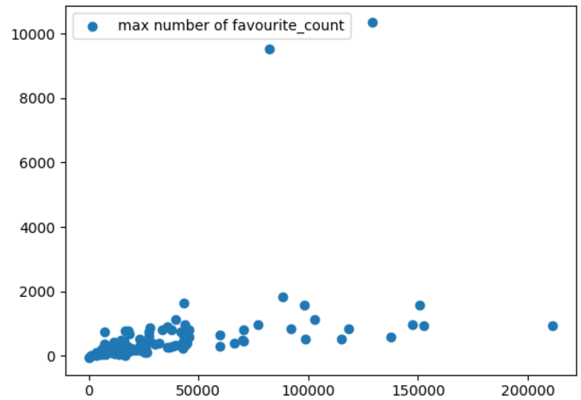

Figure 1.15 scatter plot of max number of favourite_count

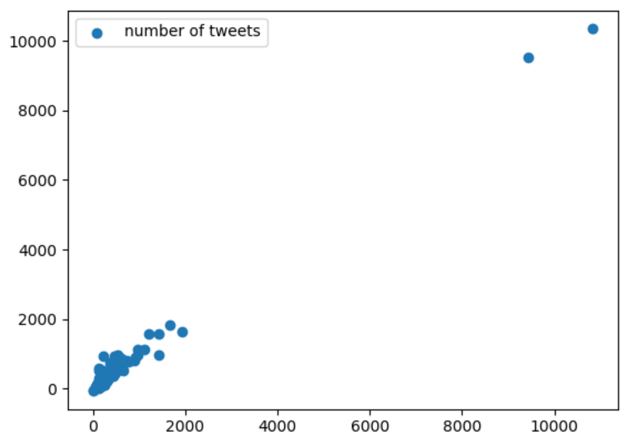

Figure 1.16 scatter plot of number of tweets

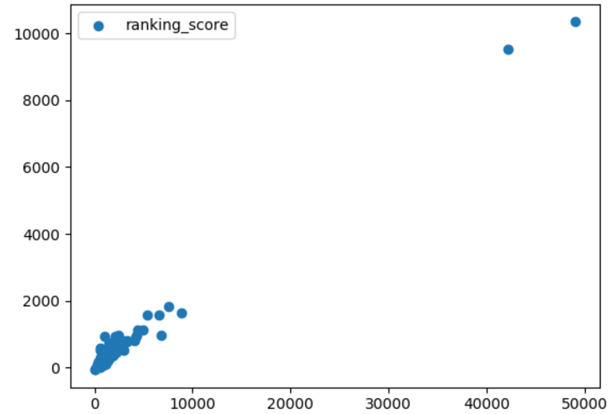

Figure 1.17 scatter plot of ranking_score

tweets_#patriots.txt

```
rmse =  2887.663069
p_values:
[  7.18859822e-01   4.16722990e-09   1.42575683e-07   9.93507402e-01
   2.27037513e-01]
most three important features:
 sum of favourites_count    max number of favourite_count     sum of friends_count
```

Most three important features' scatter:

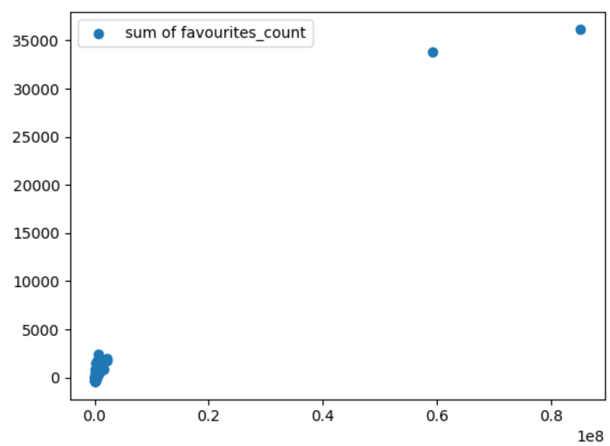

Figure 1.18 scatter plot of sum of favourites_count

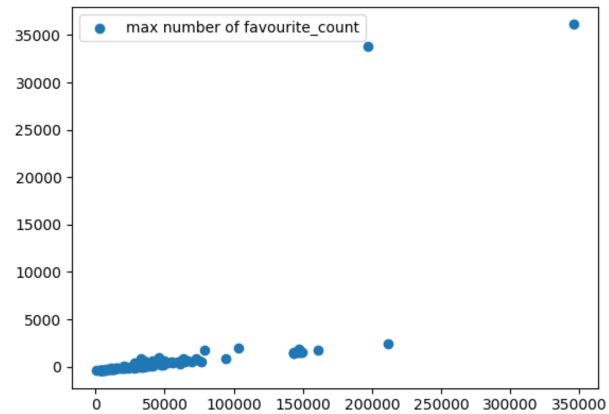

Figure 1.19 scatter plot of max number of favourite_count

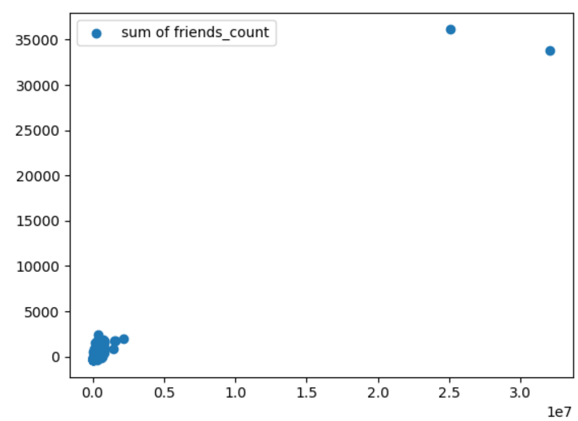

Figure 1.20 scatter plot of sum of friends_count

tweets_#sb49.txt

```
rmse =  9886.703299
p_values:
[  3.81660150e-18   3.28583049e-08   7.15254398e-01   1.59086795e-16
   8.08115666e-13]
most three important features:
 number of tweets    ranking_score     sum of friends_count
```

Most three important features' scatter:

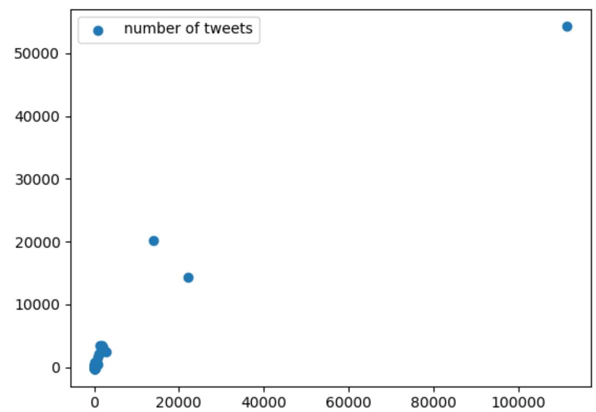

Figure 1.21 scatter plot of number of tweets

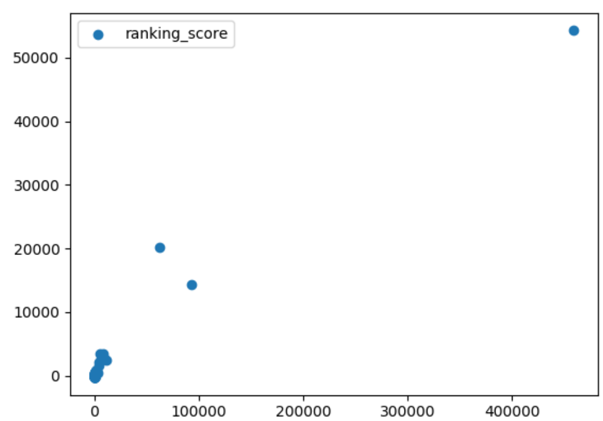

Figure 1.22 scatter plot of ranking_score

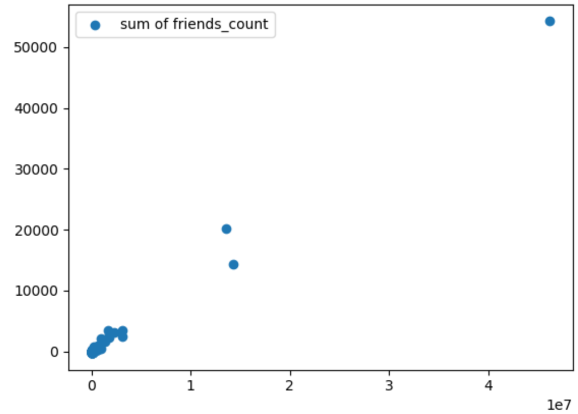

Figure 1.23 scatter plot of sum of friends_count

tweets_#superbowl.txt

```
rmse =  26242.268161
p_values:
[  2.30741534e-01   3.29813560e-18   1.94888696e-18   5.60380653e-01
   3.98563012e-18]
most three important features:
 max number of favourite_count    sum of favourites_count     sum of friends_count
```

Most three important features' scatter:

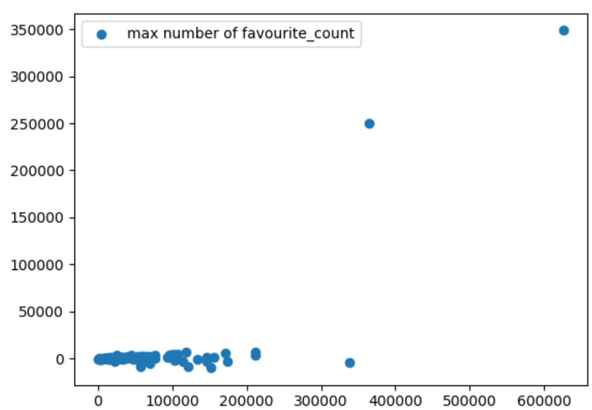

Figure 1.24 scatter plot of max number of favourite_count

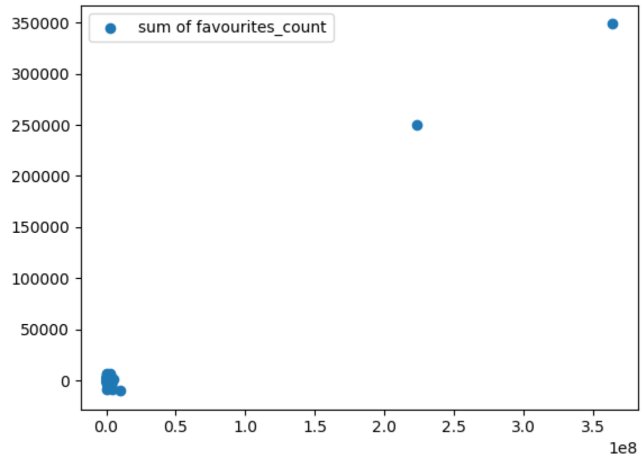

Figure 1.25 scatter plot of sum of favourites_count

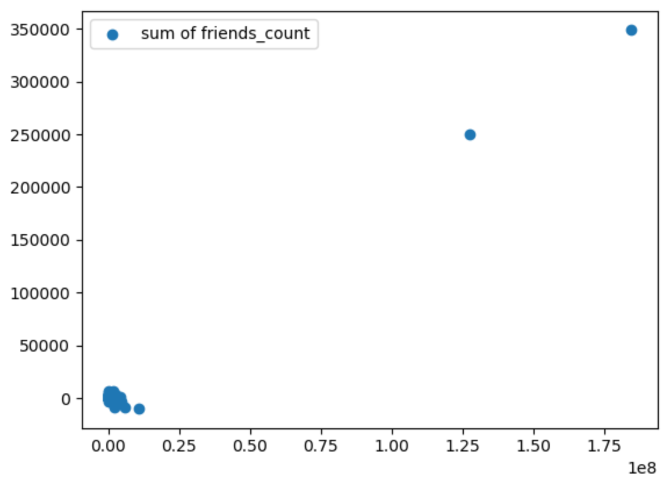

Figure 1.26 scatter plot of sum of friends_count

From previous part, we can find that for different hashtags, we obtain different important features. But in general, "sum of favorites count", "max number of favorite count", and "sum of friends count" are three most important attributes for prediction. Obviously, if a tweet is liked by a lot of people, it will retweet more compared with other tweets. Also, if a user has many friends in tweet, it will increase the probability of retweet. Number of tweets per hour and ranking score seems less import in these procedures. Although number of tweets per hour make a difference in Question1.2, it doesn't work well here. We think it's due to other three important attributes play a more important role in prediction. For some features, we can observe a clear relatively linear relation between itself and target value, which mean it really helpful for us to predict the number of tweets in next hour.

Problem 1.4

Question 1

In this question, we are asked to perform 10 fold cross validation to measure the prediction effect of different regression model in different period of time. We use RMSE as a metric to measure the effect of different models. We choose linear regression, polynomial regression and logistic regression as our models to do the regression. Combining the results of part 1.2 and 1.3, we select top seven features for each of the hashtag. The results are listed below：

Go_hawks:	**Go_patriots**	**NFL**	**Patriots**	**Sb49**	**Superbowl**
Sum of favorite counts	Sum of favorite counts	Sum of favorite counts	Sum of favorite counts	Number of tweets	Sum of favorite counts
Max of favorite counts	Max of favorite counts	Time of the day	Max of favorite counts	Ranking score	Max of favorite counts
Number of tweets	Number of tweets	Number of tweets	Number of tweets	Sum of favorite counts	Number of tweets
Sum of friends counts	Sum of friends counts	Sum of friends counts	Sum of friends counts	Max of favorite counts	Sum of friends counts
Number of followers	Number of followers	Number of followers	Number of followers	Sum of friends counts	Number of followers
Number of retweets	Number of retweets	Number of retweets	Number of retweets	Number of followers	Number of retweets
Ranking score	Ranking score	Ranking score	Ranking score	Number of retweets	Ranking score

Table 1.2 top seven feature for each hashtag

	the first beginning to 02/01/8:00	**02/01/8:00 to 8:00 PM**	**02/01/8:00 PM to end**
linear regression model error	928.675938624	3841.9567901	88.5465656843
polynomial mode error	8728.70253478	203595.726696	4107.5565657
logistic mode error	2198.4283365	4863.74393464	106.541369142

Table 1.3 statistics of tweets_#gohawks.txt

	the first beginning to 02/01/8:00	02/01/8:00 to 8:00 PM	02/01/8:00 PM to end
linear regression model error	69.4228842002	3218.1963846	4.08378667781
polynomial mode error	1195.19090562	14558.9284888	6266.77095634
logistic mode error	185.22356176	1849.23742127	13.8601670076

Table 1.4 statistics of tweets_#gopatriots.txt

	the first beginning to 02/01/8:00	02/01/8:00 to 8:00 PM	02/01/8:00 PM to end
linear regression model error	280.806579377	11542.7806919	149.904219156
polynomial mode error	500.844495875	138830.396783	367.615450095
logistic mode error	764.350759305	4751.19050787	442.777093386

Table 1.5 statistics of tweets_#nfl.txt

	the first beginning to 02/01/8:00	02/01/8:00 to 8:00 PM	02/01/8:00 PM to end
linear regression model error	701.041690508	21419.3642671	146.170584088
polynomial mode error	2771.05447083	85826.9181395	1270.51462416
logistic mode error	1055.00042348	16285.1855746	424.305594539

Table 1.6 statistics of tweets_#patriots.txt

	the first beginning to 02/01/8:00	02/01/8:00 to 8:00 PM	02/01/8:00 PM to end
linear regression model error	105.880582886	75294.616913	183.071064714
polynomial mode error	204.794964568	477909.194709	566.305501752
logistic mode error	967.66145733	27334.5250936	1114.42695923

Table 1.7 statistics of tweets_#sb49.txt

	the first beginning to 02/01/8:00	02/01/8:00 to 8:00 PM	02/01/8:00 PM to end
linear regression model error	879.053895943	307411.483586	294.808616944
polynomial mode error	2636.88528261	8675346.17783	1172.75877026
logistic mode error	2851.42263087	20293.7853564	1756.68435711

Table 1.8 statistics of tweets_#superbowl.txt

From the results above, it is obvious for us to find that:

1. Different models apply to different period of time. Before the event and after the event, it will be more appropriate to use a linear regression model. During the event, logistic regression has a better performance.
2. Among all the hashtag, the rmse of prediction during the event is much larger than the other two periods. It can explained that the number of tweets during the event is huge. Even a 1% prediction error could lead to a large absolute rmse. Also, 12 hours' training period is less than the first period and the third period. All the above reasons could lead to such a result.

Question 2:

According to results above, we choose linear regression model for first and third period. We choose logistic regression model for the second period. We get the result below:

	the first beginning to 02/01/8:00	**02/01/8:00 to 8:00 PM**	**02/01/8:00 PM to end**
linear regression model error	2290.72202556	/	593.469172798
polynomial mode error	/	/	/
logistic mode error	/	28416.1089583	/

Table 1.9 statistics of aggregate file

Comparing the large base number of the data, such absolute error is acceptable. It proves that our selection of model work for the aggregate data.

Problem 1.5

We take advantage of the results below: using the best seven features we find and apply best models we find for different period of time.

A change we made is that we use firstpost_date as the record to do the prediction because citation data didn't conform our problem description: using first five hour's data to predict the sixth hour's tweet number. Here, we concatenate five hour's feature to make a 35 dimension feature vector rather than averaging five days' feature vectors.

Here is the prediction result:

period 1 prediction	**real value**
776.41777386	178

Table 1.10 prediction and real value of sample1_period1.txt

period 2 prediction	**real value**
355487	82923

Table 1.11 prediction and real value of sample2_period2.txt

period 3 prediction	**real value**
1005.17360591	523

Table 1.12 prediction and real value of sample3_period3.txt

period 1 prediction	**real value**
110.51486798	201

Table 1.13 prediction and real value of sample4_period1.txt

period 1 prediction	**real value**
1505.21503612	213

Table 1.14 prediction and real value of sample5_period1.txt

period 2 prediction	**real value**
132031	37307

Table 1.15 prediction and real value of sample6_period2.txt

period 3 prediction	**real value**
202.01717491	120

Table 1.16 prediction and real value of sample7_period3.txt

period 1 prediction	**real value**
50.09052172	11

Table 1.17 prediction and real value of sample8_period1.txt

period 2 prediction	**real value**
15095	2790

Table 1.18 prediction and real value of sample9_period2.txt

period 3 prediction	**real value**
118.79656167	61

Table 1.19 prediction and real value of sample10_period3.txt

Comparing the prediction and real value, we can observe that our predictions are not as good as we expect. Most of the time, prediction is twice or three times or one half or one third of the real value. It does imply a fact that, it is impossible to utilize a single model to predict a huge amount of people's behavior. Maybe a combination of different model will work. Or maybe give exact number of prediction is impossible but predict a trend will be possible. All these guesses are likely to provide a solution to this prediction problem and future researches are required.

Part 2: Fan Base Prediction

In this part, we want to use features in users' tweet to predict their location. Since the tweets are related to superbowl final, we assume those tweets belong to two locations – Washington and Massachusetts which are two sides of this match. We extract some key word about location like "Seattle" to determine their location ,then use this label to set up a training dataset. After that, use the hashtag as the features to train the classifier and then predict the rest dataset which is used as the testing dataset. There are several methods which can be used to predict this, we just use SVM, Adaboost, Random Forest and Neural Network algorithms to do this task.

2.1 SVM

By using SVM, we can get the performance as the following figure which shows the ROC curve of this classifier.

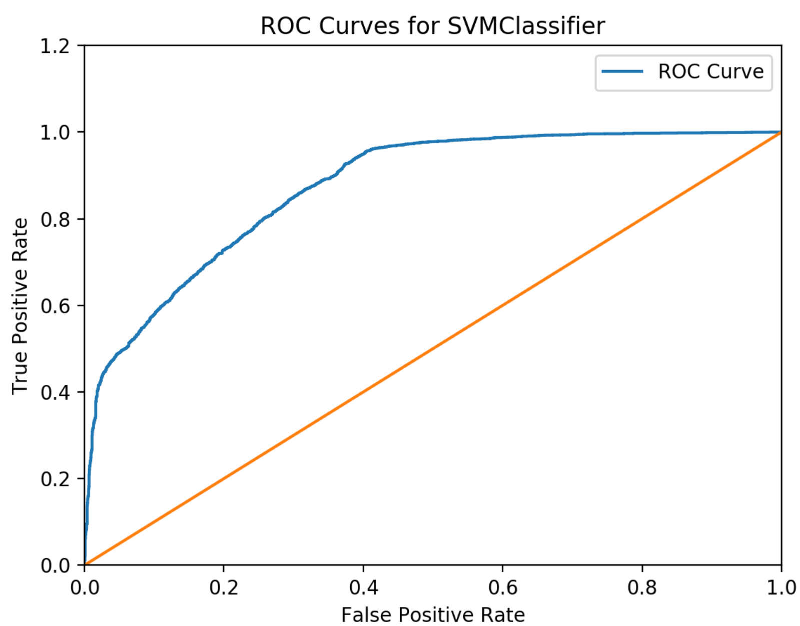

Figure 2.1 ROC curve for SVM classifier

The confusion matrix of this classifier is:

1641	1181
113	3156

And other performance parameters have been reported in the following table.

SVM	Accuracy	Recall	Precision
Value	0.7876	0.7735	0.8316

Table 2.1 SVM performance parameters

As the table and figure shows, this binary classifier-SVM fits the data pretty well, it has nearly 80% accuracy and it predicts Massachusetts very precisely.

2.2 Adaptive boosting

We also apply adaptive boosting algorithm to this dataset. And the performance is quite well. Here is the result of this classifier.

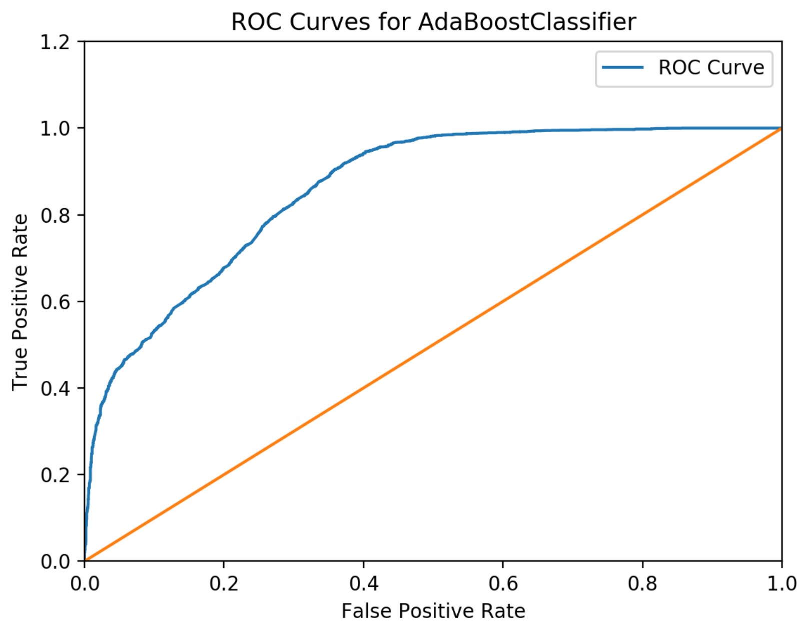

Figure 2.2 ROC curve for Adaptive boosting classifier

The figure above is the roc curve of Adaptive boosting classifier, it is looks the same with the roc curve of SVM classifier.

The confusion matrix of this classifier is:

1792	1030
311	2958

Other performance parameters are shown in table 2.2.

Adaptive boosting	Accuracy	Recall	Precision
Value	0.7798	0.7699	0.7969

Table 2.2 Adaptive boosting performance parameters

From the confusion matrix and performance parameter table, we can find that the performance of Adaptive boosting classifier is similar as the result of SVM. They both show great accuracy in class Massachusetts and has great overall precision.

2.3 Random Forest

In this part, we apply random forest algorithm to this dataset to find whether it can show better performance or not. The roc curve of this classifier shows in Figure 2.3.

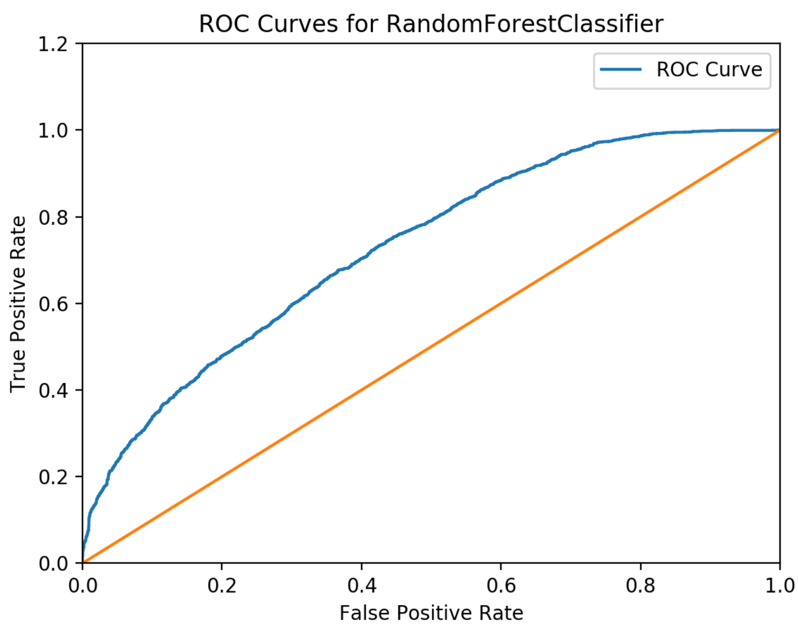

Figure 2.3 ROC Curve for Random Forest Classifier

And the confusion matrix is :

1049	1773
216	3053

The result of this classifier is shown in table 2.3.

Random Forest	**Accuracy**	**Recall**	**Precision**
Value	0.6735	0.6528	0.7309

Table 2.3 Random Forest performance parameters

As we can see from the above table and confusion matrix, random forest classifier has bad performance in this dataset. Although it shows high accuracy in class Massachusetts, it has awful performance in class Washington. The accuracy of class Washington is even less than 40% while the overall precision is 10% less than the other two method.

2.4 Neural Network

In neural network algorithm, the result is shown in the following table and figure.

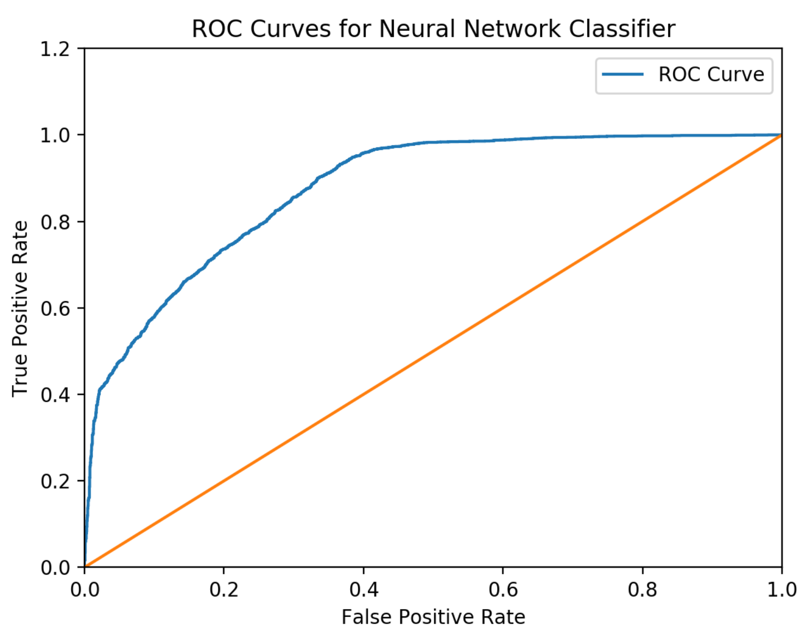

Figure 2.4 ROC Curve for Neural Network Classifier

The confusion matrix is:

1678	1144
108	3161

And the performance parameter is shown in table 2.4.

Neural Network	**Accuracy**	**Recall**	**Precision**
Value	0.7945	0.7808	0.8369

Table 2.4 Neural Network performance parameters

Neural Network algorithm shows great performance in this problem. It has high accuracy as SVM and Adaptive boosting classifier and it also shows great precision in predicting class Massachusetts.

2.5 Conclusion

We use 4 algorithms in predicting the location of user who wrote some specific tweet and SVM, Adaptive boosting and Neural Network performs pretty well while the Random Forest shows bad result of prediction. However, all of these algorithms show great accuracy in predicting user in Massachusetts and bad accuracy in predicting user in Washington. The reason of this phenomenon may be that the key words we select to distinguish different locations are not accurate enough or haven't covered all situations. Thus, one way to improve that is to use real label of the tweet rather than put label on the dataset on our own.

Part 3: Define your own project

In question3, we divided our own project into two parts.

3.1 Part 3a

For 3a part, we try to adopt several features of user and tweet to predict whether a tweet will be repost. Besides that, we also define tweet which has been repost equal or more than 5 times as hot event. We built Logistic Regression to predict whether a tweet will be repost or whether a tweet represent a hot event. Here we choose three related variables as training features. They are followers of the tweet's author, number of friends of tweet's user and favorites counts of user. The accuracy of prediction are shown below.

Here are two important points we should mention. First, "same" tweet may have different citation. Because, we see tweet repost by another one as different message compared with original one. In other words, we assume there is no relation between tweets and their retweets. Second, since most tweets are not repost, we need to do downsampling while build Logistic Regression model. If we miss this procedure, we will get high accuracy model since classifier automatically divide all points into 0 value.

Hashtag	**Accuracy for Retweet Prediction**	**Accuracy for Hot Tweet**
tweets_#gohawks.txt	0.788402	0.743363
tweets_#gopatriots.txt	0.856108	0.587719
tweets_#nfl.txt	0.874163	0.683761
tweets_#patriots.txt	0.861072	0.714456
tweets_#sb49.txt	0.875543	0.714844
tweets_#superbowl.txt	0.824136	0.786700

Table 3.1 Accuracy for Retweet prediction and Hot Tweet

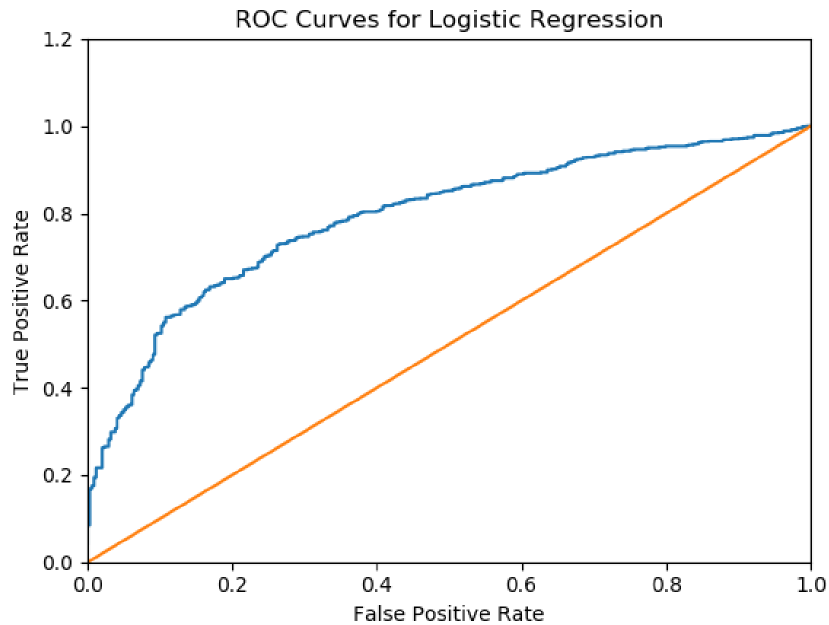

Figure 3.1 ROC Curve for gohawks

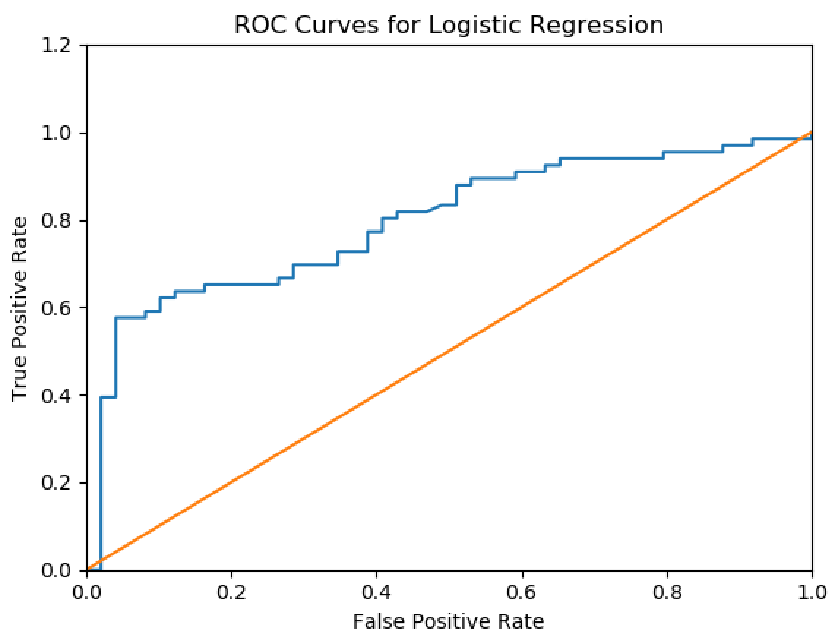

Figure 3.2 ROC Curve for gopatriots

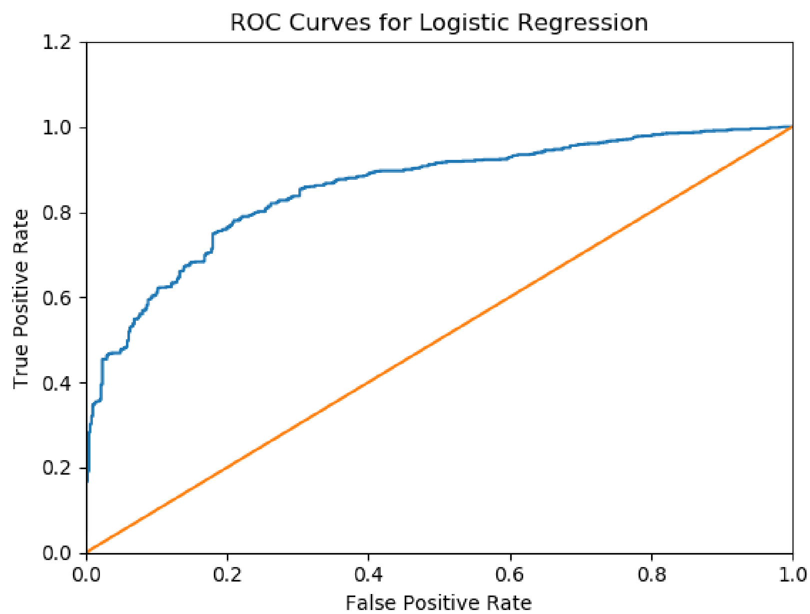

Figure 3.3 ROC Curve for nfl

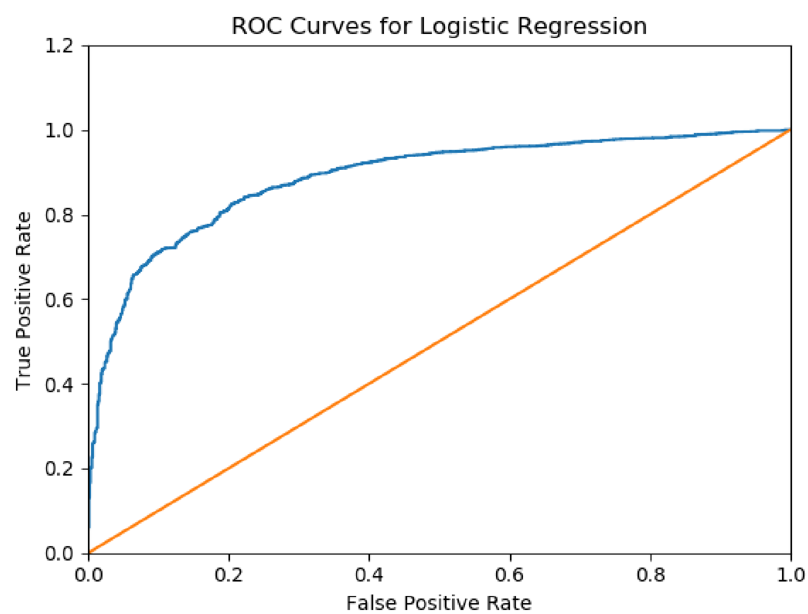

Figure 3.4 ROC Curve for patriots

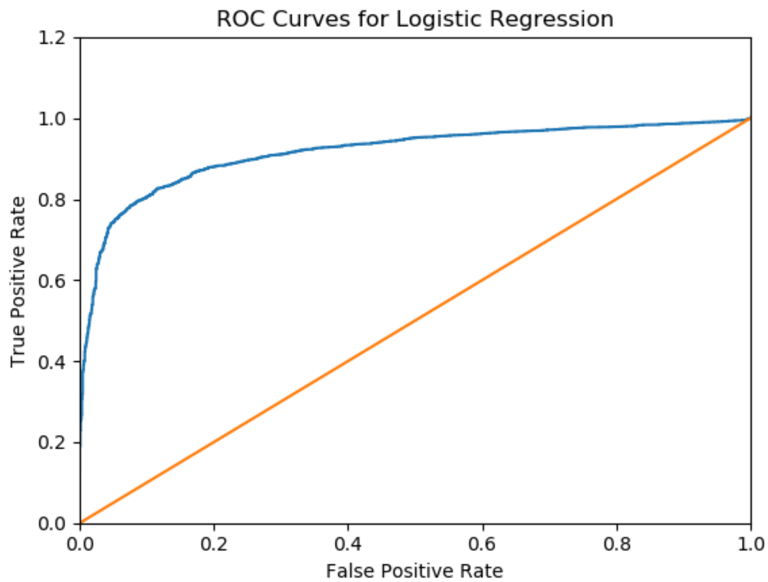

Figure 3.5 ROC Curve for sb49

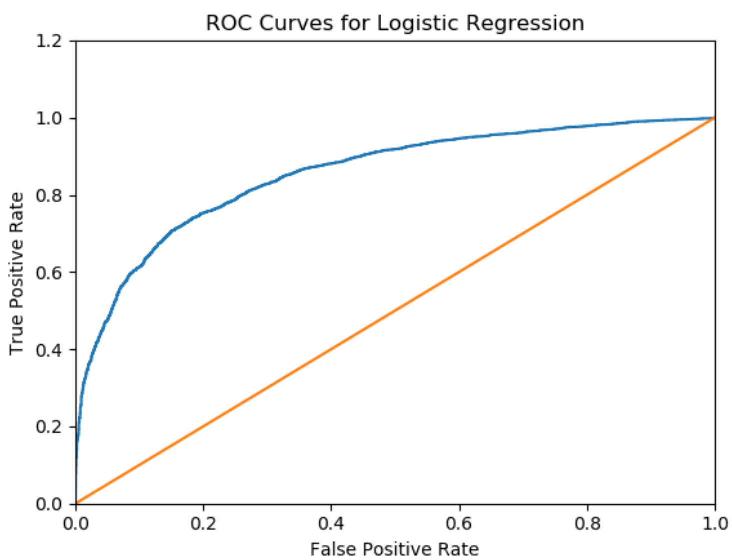

Figure 3.6 ROC Curve for Superbowl

3.2 Part 3b:

For part 3b, we retrieve the most frequent key words from those tweets and want to analyze those tweets which covers before and after the superbowl final to check what are frequent words people use in their expressions. It is obvious we use tf-idf to evaluate the frequency of those text and then pick top 10 frequent words out of these.

After we apply our methods on those data, we get the following 10 words:

[u'aaaaaaaaaah', u'aaaaaaaaand', u'aaahhhh', u'aa', u'aaannndd', u'aaaand', u'aaah', u'aaaaaaahhhh', u'aaannd', u'aaahhhhhhhhh']

As we expand the scale of frequent words to be 30, the result is pretty similar to the scale of 10. The results are:

[u'aaaaaaaaaah', u'aaaaaaaaand', u'aaahhhh', u'aa', u'aaannndd', u'aaaand', u'aaah', u'aaaaaaahhhh', u'aaannd', u'aaahhhhhhhhh', u'aaahhhhh', u'aagh', u'aahhhhhhhh', u'aac', u'aaaaaaahhhhhhhh', u'aaaaaa', u'aahh', u'aaaallll', u'aah', u'aaaaaand', u'aaaaaaa', u'aaaaaaahhhhhh', u'aaahhhhhhh', u'aalst', u'aaaaw', u'aaaayyy', u'aaaahhhhhh', u'aaaww', u'aaaamaz']

The result is very interesting since all the most frequent words are just some words to express users’ excitement. The last 20 words are also some words just like the top 10. They just use another expression and all of these words are related to their emotions of excitement. It is easy to understand that in such a big festival, people are so thrilled and then their tweet might not be normal expression. Instead of that, they are more likely to use some easy words to show their enthusiasm towards the superbowl final.