Project 5: Application - Twitter data

Zeyu Bai, 904514671 Yi-Jui Chang, 804587644
Nan Liu, 004434221 Ruizhi Yang, 704514506

Winter 2019

Introduction

In this report, we discuss the prediction of future popularity of a subject or event using the twitter data. The data 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 hashtags in our dataset include #gohawks, #gopatriots, #nfl, #patriots, #sb49 and #superbowl.

Our goal is to predict the 'number of tweets next hour' for each individual hashtag or the aggregate hashtags. We train a regression model using several features that we find to be useful for the prediction. The regression model we used including both the linear and nonlinear model. And for the nonlinear model, we considered two ensemble methods (Random forest and Gradient Boosting), and also the Neural network method. After finding the best model through cross validation of the training data for each method, we also apply them to the test data to discuss the predictions performance.

Part 1: Popularity Prediction

1.1. A first look at the data

To have a first look at the training tweet data, we calculate some basic statistics and plot the histogram of "number of tweets per hour" for each hashtag. And the hashtag in the data include

#gohawks, #gopatriots, #nfl, #patriots, #sb49 and #superbowl.

Q1: Basic Statistics

We summarized some basic statistics for each hashtag in the table 1 below. The statistics include the "average number of tweets per hour", "average number of followers of users per tweet" and "average number of retweets per tweet.

Table 1: P-value of significance test

	Avg. # of tweets per hour	Avg. # of follows of users	Avg. # of retweets per tweet
#gohawks	292.599	2217.924	2.013
#gopatriots	40.889	1427.253	1.408
#nfl	397.648	4662.375	1.534
#patriots	751.913	3280.464	1.785
#sb49	1269.026	10374.160	2.527
#superbowl	2071.353	8814.968	2.391

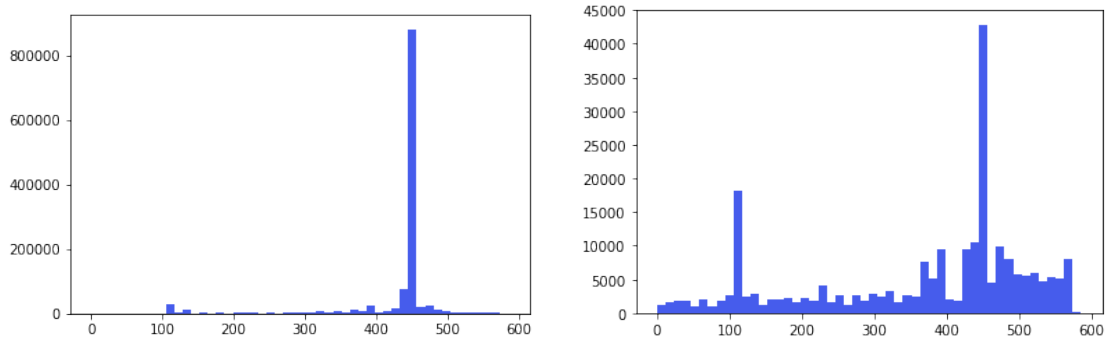
It is obvious that #sb49 and #superbowl have higher values for all the three statistics among the six hashtags. The reason might be that these two hashtags are more general and all the fans would like to tag them. But for hashtags like #gohawks, #gopatriots and #patriots, they would only be tagged by their supporters which included a smaller group of people.

In addition, since Seattle Seahawks was the defending champion, the #gohawks was more popular than #gopatriots which was reflected in the three statistics. But New England Patriots finally sealed the win, so the hashtag #patriots burst and #hawks did not.

Q2: Basic Statistics

In this question, we plotted the Histograms of "number of tweets in hour over time for all the six hashtags. And the histograms for #superbowl and #NFL are presented in figures 1(a) and

1(b) below.



(a) #SuperBowl

(b) #NFL

Figure 1: Histogram of number of tweets in hour

The histograms in 1 show the lifecycle of two hashtags #superbowl and #NFL. They two have different lifecycle patterns but it is clear that they both peak at around the 450th hours. Since the Twitter data is collected starting from 2 weeks before the game to a week after the game, the peak time is exactly corresponding to the actual time that Superbowl took place.

In addition, there are also some discussions about the #NFL during the other times and a small spike at around the 100th hour, which might corresponding to the NFL conference championships game time. But the discussion for #superbowl concentrated tightly around the actual time it have taken place.

1.2. Linear regression

In this section, we fit a linear regression model using several features to predict "number of tweets in the next hour" for each hashtag. The features name we use and their corresponding labels are shown in table 2.

For Q3, we only use the fist five features to do the prediction. And all the features are in 1-hour time window.

Table 2: Features and the label

Feature name	Label
tweets	Number of tweets
retweets	Total number of retweets
followers sum	Sum of the number of followers of the users posting the hashtag
followers max	Maximum number of followers of the users posting the hashtag
hour of day	Time of the day (24 values represent hours of the day)
mentioned	Number of mentioned user in this tweet
media	Number of media url attached
active	Active index defined by year
author	Author name
favourites_count	User's favourites count
title	Length of this tweet's title

Q3. Linear regression with 5 features

For each hashtag, we train a separate linear regression model using the first five features to predict the number of tweets in the next hour.

The MSE and R^2 measure for each model are reported in table 3.

Table 3: MSE and R^2 measure for each model

	MSE	R^2
#gohawks	759843.845	0.476
#gopatriots	27588.586	0.629
#NFL	270401.914	0.571
#patriots	5189695.981	0.668
#sb49	16107134.316	0.805
#superbowl	52573154.301	0.800

R^2 measures the proportion of the variance for the dependent variable "number of tweets in the next hour" that's explained by the five features. From table 3, the R^2 are moderate in the regression model for predicting #gohawks, #gopatriots, #nfl and #patriots, and are relatively high for #sb49 and #superbowl. The reason might be that the five features used in the model are general and do not cover team specific features. Thus they are more suitable to explain the variation for the general hashtags not the specific team's hashtags.

As showed in table 3, MSE are large for all the models. The reason is that linear regression model can not predict well for the peaks of the hashtags. To be specific, consider the prediction for #NFL, the plot of predicted values against the true values for " number of tweets of #NFL" is shown in figure 2. It is clear that the model predicts well for small values of the 'number of tweets'(concentrated at the 45^o line), but not for the large values (far from the 45^o line).

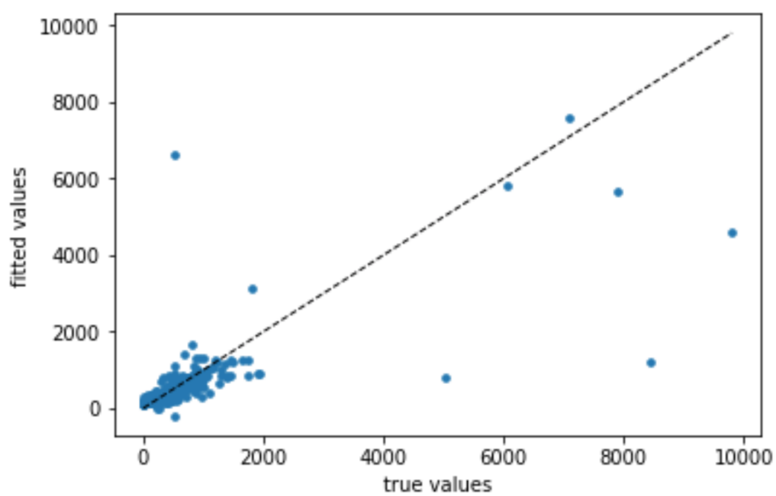


Figure 2: Fitted values V.S. true values for " number of tweets of #NFL"

In addition, for each models, we report the t-statisics and the corresponding p-value for the five features in the following figure 3. Features with high t-statistics or low p-values are considered to be significant different from zero. We choose the significance level to be 5%. Thus we say the feature is significance if p-value is smaller than 5%.

First, we notice that "hour of day" is not significant for all the six models. Since"hour of day" is a categorical variable which can not be linearly related to the "number of tweets in the next hour", it is reasonable that the feature is not significant. To solve this issue, we can use one-hot-encoding to transform it as an 24-dimensional vector, where only one entry is 1 and the rest are 0's. Then we can detect which periods have significant effect on the "number of tweets".

Next, consider the regression results for the three 'general' hashtags #nfl, #sb49 and #superbowl. The results are reported in 3(c), 3(e) and 3(f) respectively. In general, all the other four features are significance for explaining the 'number of tweets in next hour' for these three hashtags.

But for the other three team specific hashtags #gohawks, #gopatriots and #patriots, most of the features used here are not significant. For example, in the regression 3(b) for explaining the number of tweets for #gopatriots, only retweets is significant.

These results also indicate that the five features used here are more useful for explaining the number of tweets for the general hashtags but not the specific team's hashtags.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.476
Model:                            OLS   Adj. R-squared:                  0.472
Method:                 Least Squares   F-statistic:                     103.9
Date:                Tue, 19 Mar 2019   Prob (F-statistic):           7.03e-78
Time:                        01:13:46   Log-Likelihood:                -4725.3
No. Observations:                 577   AIC:                             9463.
Df Residuals:                     571   BIC:                             9489.
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            94.2103     70.292      1.340      0.181     -43.852     232.273
tweets            1.2824      0.164      7.822      0.000       0.960       1.604
retweets         -0.1365      0.044     -3.135      0.002      -0.222      -0.051
followers sum    -0.0002   8.01e-05     -2.425      0.016      -0.000   -3.69e-05
followers max  6.102e-05      0.000      0.409      0.683      -0.000       0.000
hour of day       1.7189      5.316      0.323      0.747      -8.722      12.159
==============================================================================
Omnibus:                      914.731   Durbin-Watson:                   2.216
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           778994.704
Skew:                           8.683   Prob(JB):                         0.00
Kurtosis:                     182.165   Cond. No.                     5.11e+06
==============================================================================
```

(a) #gohawks

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.629
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     192.9
Date:                Tue, 19 Mar 2019   Prob (F-statistic):          7.06e-120
Time:                        01:13:48   Log-Likelihood:                -3749.1
No. Observations:                 574   AIC:                             7510.
Df Residuals:                     568   BIC:                             7536.
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             6.0229     10.869      0.554      0.580     -15.325      27.371
tweets            0.3073      0.285      1.079      0.281      -0.252       0.866
retweets          0.4892      0.192      2.550      0.011       0.112       0.866
followers sum    -0.0001      0.000     -0.504      0.614      -0.001       0.000
followers max  -2.36e-05      0.000     -0.108      0.914      -0.000       0.000
hour of day       0.1144      0.938      0.122      0.903      -1.728       1.956
==============================================================================
Omnibus:                      485.015   Durbin-Watson:                   1.908
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           290846.979
Skew:                           2.515   Prob(JB):                         0.00
Kurtosis:                     113.161   Cond. No.                     6.00e+05
==============================================================================
```

(b) #gopatriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.571
Model:                            OLS   Adj. R-squared:                  0.567
Method:                 Least Squares   F-statistic:                     154.0
Date:                Tue, 19 Mar 2019   Prob (F-statistic):          7.45e-104
Time:                        01:14:23   Log-Likelihood:                -4488.6
No. Observations:                 585   AIC:                             8989.
Df Residuals:                     579   BIC:                             9015.
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const           123.7556     42.809      2.891      0.004      39.676     207.835
tweets            0.5667      0.135      4.196      0.000       0.301       0.832
retweets         -0.1653      0.064     -2.590      0.010      -0.291      -0.040
followers sum     0.0001    2.5e-05      4.577      0.000    6.53e-05       0.000
followers max    -0.0001   3.31e-05     -3.525      0.000      -0.000   -5.17e-05
hour of day       0.4475      3.148      0.142      0.887      -5.736       6.631
==============================================================================
Omnibus:                      668.832   Durbin-Watson:                   2.373
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           349115.930
Skew:                           4.594   Prob(JB):                         0.00
Kurtosis:                     122.324   Cond. No.                     8.57e+06
==============================================================================
```

(c) #NFL

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.668
Model:                            OLS   Adj. R-squared:                  0.666
Method:                 Least Squares   F-statistic:                     233.4
Date:                Tue, 19 Mar 2019   Prob (F-statistic):          3.35e-136
Time:                        01:14:56   Log-Likelihood:                -5352.8
No. Observations:                 585   AIC:                         1.072e+04
Df Residuals:                     579   BIC:                         1.074e+04
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const           181.2843    184.677      0.982      0.327    -181.435     544.003
tweets            0.9144      0.071     12.926      0.000       0.775       1.053
retweets         -0.0680      0.058     -1.177      0.240      -0.182       0.045
followers sum -1.098e-05   2.63e-05     -0.417      0.677   -6.27e-05    4.08e-05
followers max    0.0001    9.17e-05      1.338      0.182   -5.75e-05       0.000
hour of day      -5.9285     13.808     -0.429      0.668     -33.048      21.191
==============================================================================
Omnibus:                      885.848   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           686964.092
Skew:                           7.930   Prob(JB):                         0.00
Kurtosis:                     170.127   Cond. No.                     1.60e+07
==============================================================================
```

(d) #patriots

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     476.8
Date:                Tue, 19 Mar 2019   Prob (F-statistic):          1.45e-202
Time:                        01:15:52   Log-Likelihood:                -5684.1
No. Observations:                 585   AIC:                         1.138e+04
Df Residuals:                     579   BIC:                         1.141e+04
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const           171.9764    299.240      0.575      0.566    -415.752     759.705
tweets            1.1364      0.087     13.053      0.000       0.965       1.307
retweets         -0.1607      0.079     -2.046      0.041      -0.315      -0.006
followers sum  9.744e-06   1.25e-05      0.781      0.435   -1.48e-05    3.43e-05
followers max  9.544e-05   4.37e-05      2.184      0.029    9.62e-06       0.000
hour of day     -14.0619     23.216     -0.606      0.545     -59.660      31.537
==============================================================================
Omnibus:                     1186.403   Durbin-Watson:                   1.673
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2239749.831
Skew:                          14.620   Prob(JB):                         0.00
Kurtosis:                     304.715   Cond. No.                     1.43e+08
==============================================================================
```

(e) #sb49

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.798
Method:                 Least Squares   F-statistic:                     462.7
Date:                Tue, 19 Mar 2019   Prob (F-statistic):          1.51e-199
Time:                        01:17:24   Log-Likelihood:                -6030.1
No. Observations:                 585   AIC:                         1.207e+04
Df Residuals:                     579   BIC:                         1.210e+04
Df Model:                           5
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const          -150.5245    608.009     -0.248      0.805   -1344.696    1043.647
tweets            2.2766      0.080     28.512      0.000       2.120       2.433
retweets         -0.2543      0.046     -5.539      0.000      -0.344      -0.164
followers sum    -0.0001     2.2e-05    -6.259      0.000      -0.000   -9.47e-05
followers max     0.0007      0.000      4.884      0.000       0.000       0.001
hour of day     -20.4414     43.757     -0.467      0.641    -106.383      65.500
==============================================================================
Omnibus:                      971.880   Durbin-Watson:                   2.283
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1778202.995
Skew:                           9.264   Prob(JB):                         0.00
Kurtosis:                     272.460   Cond. No.                     2.22e+08
==============================================================================
```

(f) #superbowl

Figure 3

1.3. Feature analysis

In this section, we also fit the linear regression model but adding some new features which we find might be useful for predicting the 'number of tweets in next hour'. The newly added features and their labels are also listed in table 2. We discard the variable 'hours of day' which is shown to be insignificant for all the models.

Q4. Linear regression with 10 features

For each hashtag, we train a separate linear regression model using the 10 features listed in table 2 to predict 'the number of tweets in the next hour'.

The MSE and R^2 measure for each model are reported in table 4.

Table 4: MSE and R^2 measure for each model

	MSE	R^2
#gohawks	507327.120	0.650
#gopatriots	7491.296	0.899
#NFL	176690.487	0.720
#patriots	3702663.514	0.763
#sb49	13092915.122	0.841
#superbowl	25529353.588	0.903

Comparing with the results with 5 features in table 3, R^2 in table 4 increases for all the models. But R^2 can not be used as model selection tool since it will increase for sure by adding any features no matter useful or not.

The decreasing of MSE indicates that models with 10 features outperforms the models with 5 features in terms of prediction accuracy. But the MSE is still large which is the drawback of linear models for unable to predict the peaks of the time series data for 'number of tweets'.

In addition, for each models, we report the t-statistics and the corresponding p-value for the 10 features in the following figure 4. We still choose the significance level to be 5%.

First, consider the regression results for the three ‘general’ hashtags #nfl, #sb49 and #superbowl. From Q3 we know the original four features excluding ‘hours of day’ are almost all significant for these three models. The newly added features seem not be helpful for explaining the number of tweets under these cases since most of them are not significant.

But for the other three team specific hashtags #gohawks, #gopatriots and #patriots, adding new features improves the significance results a lot. For example, in the regression 4(b) for explaining the number of tweets for #gopatriots, only two features are insignificant, compared with the poor results in the regression 3(b) which only has one significant feature.

Dep. Variable:	y	R-squared:	0.650
Model:	OLS	Adj. R-squared:	0.644
Method:	Least Squares	F-statistic:	105.3
Date:	Wed, 20 Mar 2019	Prob (F-statistic):	3.29e-122
Time:	00:37:40	Log-Likelihood:	-4608.7
No. Observations:	577	AIC:	9239.
Df Residuals:	566	BIC:	9287.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	100.2750	185.620	0.540	0.589	-264.313	464.863
tweets	-1.9045	0.523	-3.641	0.000	-2.932	-0.877
retweets	-0.0658	0.037	-1.770	0.077	-0.139	0.007
followers sum	-0.0004	7.15e-05	-6.263	0.000	-0.001	-0.000
followers max	0.0005	0.000	3.945	0.000	0.000	0.001
mentioned	-1.0554	0.449	-2.349	0.019	-1.938	-0.173
media	4.6705	0.820	5.699	0.000	3.061	6.280
active	-0.0049	0.010	-0.478	0.633	-0.025	0.015
author	0.2857	0.761	0.375	0.707	-1.209	1.781
favourites_count	0.0013	0.000	10.217	0.000	0.001	0.002
title	-0.8097	1.839	-0.440	0.660	-4.421	2.802

Omnibus:	1115.198	Durbin-Watson:	2.057
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1636650.994
Skew:	13.170	Prob(JB):	0.00
Kurtosis:	262.580	Cond. No.	2.46e+07

(a) #gohawks

Dep. Variable:	y	R-squared:	0.899
Model:	OLS	Adj. R-squared:	0.898
Method:	Least Squares	F-statistic:	503.1
Date:	Wed, 20 Mar 2019	Prob (F-statistic):	3.46e-273
Time:	00:38:54	Log-Likelihood:	-3374.9
No. Observations:	574	AIC:	6772.
Df Residuals:	563	BIC:	6820.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.0352	7.058	0.288	0.773	-11.829	15.899
tweets	7.5603	0.382	19.816	0.000	6.811	8.310
retweets	-0.9247	0.120	-7.727	0.000	-1.160	-0.690
followers sum	0.0005	0.000	3.044	0.002	0.000	0.001
followers max	-0.0005	0.000	-3.151	0.002	-0.001	-0.000
mentioned	4.8567	0.431	11.267	0.000	4.010	5.703
media	8.3967	0.981	8.557	0.000	6.469	10.324
active	0.0024	0.001	1.730	0.084	-0.000	0.005
author	-6.0159	0.384	-15.682	0.000	-6.769	-5.262
favourites_count	-0.0014	9.64e-05	-14.925	0.000	-0.002	-0.001
title	-0.1245	0.089	-1.395	0.164	-0.300	0.051

Omnibus:	262.173	Durbin-Watson:	2.389
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34760.722
Skew:	-0.937	Prob(JB):	0.00
Kurtosis:	41.077	Cond. No.	1.22e+06

(b) #gopatriots

Dep. Variable:	y	R-squared:	0.720
Model:	OLS	Adj. R-squared:	0.715
Method:	Least Squares	F-statistic:	147.3
Date:	Wed, 20 Mar 2019	Prob (F-statistic):	2.77e-151
Time:	00:44:00	Log-Likelihood:	-4364.1
No. Observations:	585	AIC:	8750.
Df Residuals:	574	BIC:	8798.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	203.4057	189.215	1.075	0.283	-168.232	575.044
tweets	1.4937	0.198	7.537	0.000	1.104	1.883
retweets	-0.0356	0.057	-0.627	0.531	-0.147	0.076
followers sum	-3.638e-05	2.48e-05	-1.469	0.142	-8.5e-05	1.23e-05
followers max	4.472e-05	3.09e-05	1.446	0.149	-1.6e-05	0.000
mentioned	1.3609	0.749	1.816	0.070	-0.111	2.833
media	7.7024	1.181	6.524	0.000	5.384	10.021
active	-0.0093	0.005	-2.018	0.044	-0.018	-0.000
author	-4.2541	0.305	-13.943	0.000	-4.853	-3.655
favourites_count	0.0014	0.000	9.938	0.000	0.001	0.002
title	0.4535	1.646	0.275	0.783	-2.780	3.687

Omnibus:	771.196	Durbin-Watson:	1.975
Prob(Omnibus):	0.000	Jarque-Bera (JB):	227062.410
Skew:	6.352	Prob(JB):	0.00
Kurtosis:	98.677	Cond. No.	4.75e+07

(c) #NFL

Dep. Variable:	y	R-squared:	0.763
Model:	OLS	Adj. R-squared:	0.759
Method:	Least Squares	F-statistic:	185.2
Date:	Wed, 20 Mar 2019	Prob (F-statistic):	2.14e-172
Time:	00:45:26	Log-Likelihood:	-5254.0
No. Observations:	585	AIC:	1.053e+04
Df Residuals:	574	BIC:	1.058e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5191.9213	1411.553	3.678	0.000	2419.483	7964.360
tweets	-4.2871	1.176	-3.645	0.000	-6.597	-1.977
retweets	-0.1128	0.063	-1.782	0.075	-0.237	0.012
followers sum	0.0006	5.15e-05	10.906	0.000	0.000	0.001
followers max	-0.0007	0.000	-7.025	0.000	-0.001	-0.001
mentioned	0.5922	0.192	3.089	0.002	0.216	0.969
media	-2.1462	3.016	-0.712	0.477	-8.070	3.778
active	-0.0085	0.016	-0.541	0.589	-0.039	0.022
author	3.6054	1.497	2.408	0.016	0.664	6.546
favourites_count	0.0002	0.000	2.071	0.039	1.12e-05	0.000
title	-43.9848	12.795	-3.438	0.001	-69.115	-18.855

Omnibus:	1024.197	Durbin-Watson:	1.892
Prob(Omnibus):	0.000	Jarque-Bera (JB):	831689.278
Skew:	10.845	Prob(JB):	0.00
Kurtosis:	186.440	Cond. No.	2.14e+08

(d) #patriots

Dep. Variable:	y	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.838
Method:	Least Squares	F-statistic:	304.0
Date:	Wed, 20 Mar 2019	Prob (F-statistic):	6.99e-222
Time:	00:47:37	Log-Likelihood:	-5623.4
No. Observations:	585	AIC:	1.127e+04
Df Residuals:	574	BIC:	1.132e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	136.5785	485.248	0.281	0.778	-816.501	1089.658
tweets	-0.8401	0.453	-1.855	0.064	-1.730	0.050
retweets	-0.0496	0.099	-0.501	0.616	-0.244	0.145
followers sum	2.666e-05	1.34e-05	1.986	0.048	2.91e-07	5.3e-05
followers max	-0.0001	4.6e-05	-3.187	0.002	-0.000	-5.63e-05
mentioned	1.7073	0.209	8.169	0.000	1.297	2.118
media	15.0630	2.133	7.062	0.000	10.874	19.253
active	0.0020	0.050	0.040	0.968	-0.096	0.100
author	-1.8321	0.687	-2.667	0.008	-3.181	-0.483
favourites_count	-8.329e-05	6.35e-05	-1.312	0.190	-0.000	4.14e-05
title	-3.2625	4.880	-0.669	0.504	-12.847	6.322

Omnibus:	1233.024	Durbin-Watson:	1.738
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2992069.950
Skew:	15.870	Prob(JB):	0.00
Kurtosis:	351.919	Cond. No.	2.62e+08

(e) #sb49

Dep. Variable:	y	R-squared:	0.903
Model:	OLS	Adj. R-squared:	0.901
Method:	Least Squares	F-statistic:	533.1
Date:	Wed, 20 Mar 2019	Prob (F-statistic):	5.61e-283
Time:	00:50:05	Log-Likelihood:	-5818.8
No. Observations:	585	AIC:	1.166e+04
Df Residuals:	574	BIC:	1.171e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3399.0244	3280.489	1.036	0.301	-3044.202	9842.251
tweets	2.1289	0.852	2.498	0.013	0.455	3.803
retweets	-0.3118	0.055	-5.681	0.000	-0.420	-0.204
followers sum	-0.0002	2.12e-05	-10.596	0.000	-0.000	-0.000
followers max	0.0002	0.000	2.110	0.035	1.62e-05	0.000
mentioned	2.3579	1.096	2.150	0.032	0.204	4.512
media	24.6573	2.039	12.092	0.000	20.652	28.662
active	0.0360	0.072	0.499	0.618	-0.106	0.178
author	-4.4657	0.633	-7.058	0.000	-5.708	-3.223
favourites_count	0.0003	0.000	1.147	0.252	-0.000	0.001
title	-36.5777	29.630	-1.234	0.218	-94.774	21.619

Omnibus:	1002.583	Durbin-Watson:	1.920
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1186253.525
Skew:	10.145	Prob(JB):	0.00
Kurtosis:	222.670	Cond. No.	1.77e+09

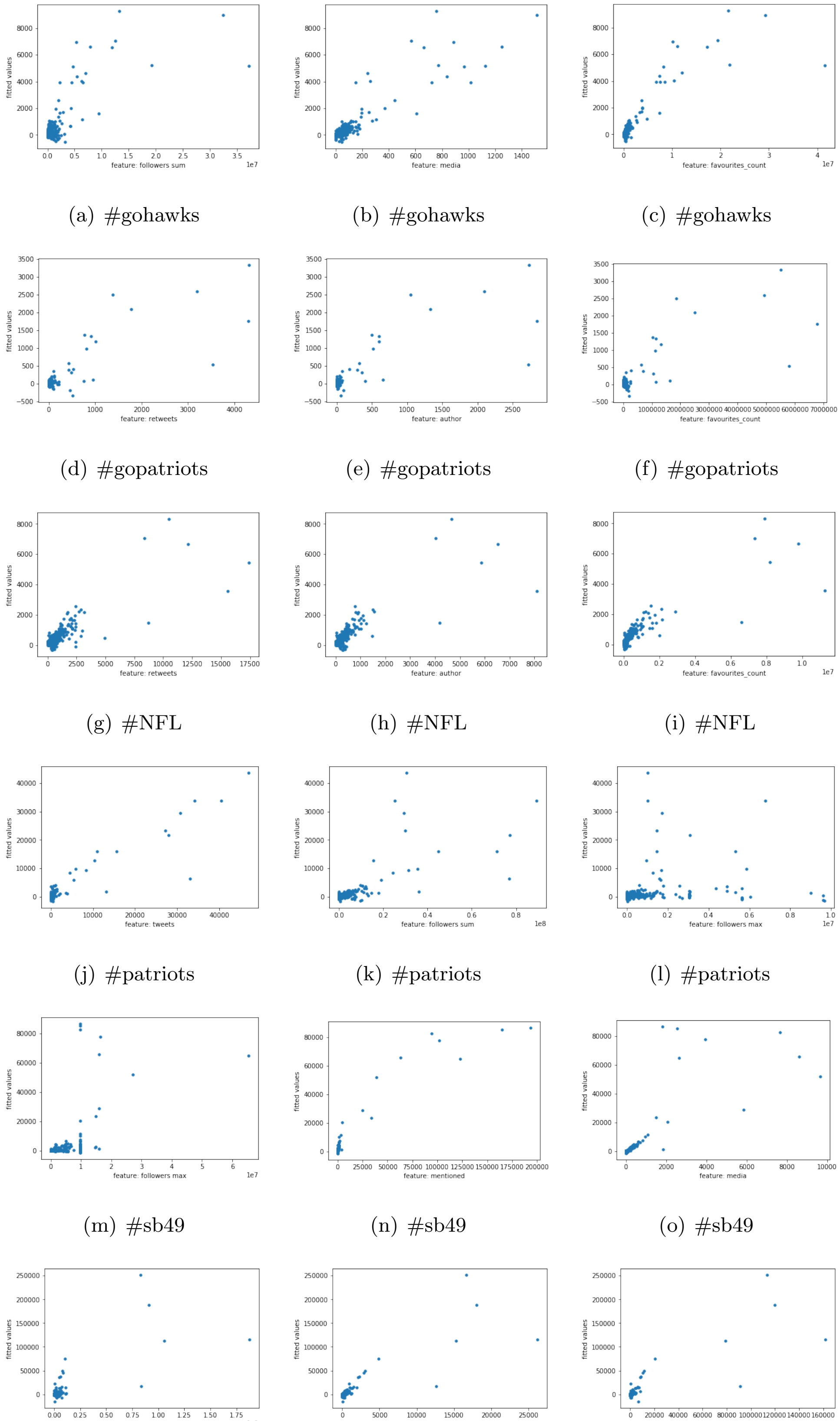
(f) #superbowl

Figure 4

Q5. Top features

In this section, we consider the relationship of the top 3 features with the predictant (number of tweets for next hour). Top features is defined to have the smallest p-value or the largest t-statistics. For each of the hashtag, we plot the predictant against the top 3 features respectively in figure 5.

For all the 18 plots in figure 5, features and the fitted value are positively correlated. But the regression coefficients are not all to be positive. For example, in the regression of #gohawks, the top 3 features are favourites_count, media and followers sum. Among them, the coefficient for 'followers sum' is negative which does not agree with the positive trend in the plot 5(a). The reason is because the two way plot of fitted value against 'followers sum' measures the correlation between these two variables. However, the regression coefficient captures the correlation between these two variables after partial out the effect of all the other features included in the regression model. Thus it is possible that the original correlation is positive, but after controlling the other features, the pure correlation left between them becomes negative. This phenomenon is related to the omitted variable bias problem.



(a) #gohawks (b) #gohawks (c) #gohawks

(d) #gopatriots (e) #gopatriots (f) #gopatriots

(g) #NFL (h) #NFL (i) #NFL

(j) #patriots (k) #patriots (l) #patriots

(m) #sb49 (n) #sb49 (o) #sb49

(p) #superbowl (q) #superbowl (r) #superbowl

Figure 5: Scatter plot of predictant V.S. top 3 features

4. Piece-wise linear regression

In this section, we further explore the linear regression model for three different periods, before active time, active period and after active period. We use 5-minute window for the active period and 1-hour window for the other two periods. And we still use the 10 features we selected in Q4-5 to predict the 'number of tweets next hour'.

Q6: Piece-wise LR for each hashtag

For each hashtag, we train the linear regression model for all the three periods. The MSE and R^2 for each case are reported in table 5

Table 5: MSE and R^2 measure for individual hashtags

hashtag	Period	MSE	R^2
#gohawks	before active	414263.048	0.598
	active	63160.148	0.555
	after active	919.848	0.896
#gopatriots	before active	1196.448	0.708
	active	13902.440	0.459
	after active	2.734	0.982
#NFL	before active	59097.840	0.562
	active	15448.260	0.866
	after active	15726.388	0.819
#patriots	before active	292559.762	0.626
	active	592589.065	0.744
	after active	5485.773	0.931
#sb49	before active	3967.929	0.873
	active	1061154.176	0.890
	after active	16093.080	0.952
#superbowl	before active	460135.946	0.467
	active	4188144.471	0.933
	after active	86282.69	0.878

From the result in table 5 we see the MSE is relatively smaller for the period before and after the active superbowl time. The reason is that during these two periods the time series data for 'the number of tweets' is more stationary. It is more likely to obtain a good prediction result if we just include one-period lag variables as we did here .

But for the active period during the superbowl, the number of tweets bursts and the time series data becomes non-stationary. Then our model with one-period lag terms can not do well in the prediction. Thus we have high MSE during the active period for all the cases.

Q7: Piece-wise LR for aggregated data

In this question, we consider the aggregate data for all hashtags, and still train a linear regression model for all the three periods. The MSE and R^2 are reported in table 6 below.

Table 6: MSE and R^2 measure for aggregate hashtags

Period	MSE	R^2
before active	3875744.890	0.483
active	13255242.359	0.884
after active	241287.507	0.921

The results for the aggregate data have the similar pattern with the individual hashtags case that has relatively lower MSE for periods before and after the active time. And the MSE for the aggregate data should be some weighted average of the MSE for the individual data.

1.5 Nonlinear regressions

Due to the drawback of linear regression model for unable to predict outliers, we consider to use nonlinear regression models to predict the 'number of tweets next period' in this section. We apply two ensemble methods (Random Forest, Gradient Boosting) and also Neural network method to do the prediction.

Q8.Random Forest and Gradient Boosting

In this question, we apply the two ensemble methods (Random Forest, Gradient Boosting) to predict the number of tweets next period. We use the grid search with 5-folded CV to find the be best parameter set for these two methods respectively. The optimal parameters set and the corresponding test MSE from CV are summarized in table 7.

Table 7: Grid search result for Random Forest and Gradient Boosting

	Random Forest	Gradient Boosting
max_depth	10	20
max_features	sqrt	sqrt
min_samples_leaf	1	1
min_samples_split	2	2
n_estimators	1800	200
Test MSE	19791627.152	21199998.646

The test MSEs from 5-folded CV for both Random Forest and Gradient Boosting look poor. The reason might be that these two methods have over-fitting problem, thus the train MSEs are small but test MSEs are huge.

Q9. Comparison of OLS results and random forest visualization

We used the best random forest model we found in Q8 and visualize it using "graphviz" package. The tree is quite large so we only show the root node part of it in figure 6. OLS results on the entire dataset is shown in figure 7. We can see that the features used near root node are: media, mentioned, author. In OLS results, the most important 3 features with smallest p-values are media, retweets, mentioned, author. We can conclude that random forest regressor is pretty consistent with OLS analysis.

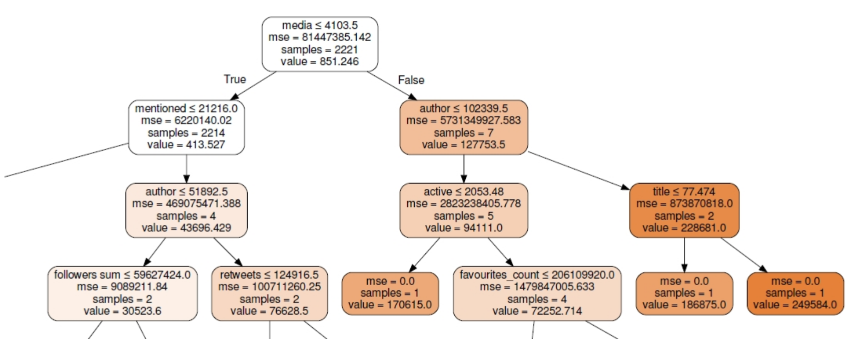


Figure 6: Random forest visualization

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.848
Model:                            OLS   Adj. R-squared:                  0.848
Method:                 Least Squares   F-statistic:                     1943.
Date:                Sat, 23 Mar 2019   Prob (F-statistic):               0.00
Time:                        18:07:38   Log-Likelihood:                -32962.
No. Observations:                3491   AIC:                         6.595e+04
Df Residuals:                    3480   BIC:                         6.601e+04
Df Model:                          10
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const              148.8645    193.099      0.771      0.441    -229.734     527.463
tweets               3.4201      0.140     24.442      0.000       3.146       3.694
retweets            -0.4726      0.014    -34.686      0.000      -0.499      -0.446
followers sum    -8.285e-05   4.42e-06    -18.760      0.000   -9.15e-05   -7.42e-05
followers max     3.259e-05   2.63e-05      1.240      0.215   -1.89e-05    8.41e-05
mentioned            1.1546      0.047     24.475      0.000       1.062       1.247
media               20.1908      0.541     37.354      0.000      19.131      21.251
active               0.0046      0.011      0.420      0.674      -0.017       0.026
author              -5.0160      0.150    -33.375      0.000      -5.311      -4.721
favourites_count  1.075e-06   4.23e-05      0.025      0.980   -8.19e-05    8.41e-05
title               -3.2298      2.014     -1.604      0.109      -7.179       0.719
==============================================================================
Omnibus:                     7841.347   Durbin-Watson:                   1.796
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         74341453.686
Skew:                          20.544   Prob(JB):                         0.00
Kurtosis:                     716.720   Cond. No.                     2.15e+08
==============================================================================
```

Figure 7: OLS results on the entire dataset

Q10. Piece-wise Gradient Boosting

In this section, we further explore Gradient Boosting method for three different periods described in Question 6. For each period, we perform the same grid search with 5-fold CV for Gradient Boosting. The optimal parameters set and the corresponding test MSE from CV under different cases are summarized in table 8.

Table 8: Grid search result for Gradient Boosting in three periods

	before active	active	after active
max_depth	20	20	20
max_features	auto	sqrt	sqrt
min_samples_leaf	2	4	4
min_samples_split	1	2	2
n_estimators	200	1200	200
Test MSE	3338958.538	19185618.506	2980.987

The test MSE for all three periods drops compared with the test MSE for the whole period shown in the table 6. The best parameter set in each period varies from each other and also different with those for the entire period. This flexibility in choosing parameters for each period can explain the better performance of the piece-wise Gradient Boosting method.

Q11. Neural network

In this part, we apply the neural network to predict 'the number of tweets next period'. We tried several different architectures with various numbers of layers and layer sizes and apply 5-folded CV to select five of them to apply to the entire aggregated data. The MSE of fitting the entire data under different architectures are reported in table 9

Table 9: MSE for neural net under different architectures

numbers of layers	layer sizes	fitting MSE
1	50	1493991680.303
2	(50,300)	2269056843.830
2	(50,600)	1435417557274.456
3	(50,600,300)	357390845.837
3	(50,600,500)	5562240254.907

The best architectures in terms of fitting MSE should contain 3 numbers of layers with Layer sizes for each layer to be $(50, 600, 300)$

Q12. Neural network with scaled data

In this question, we use the best architectures obtained in Question 11 which has layer sizes $(50, 600, 300)$ and apply to the scaled data. The fitting MSE under this case is

$$MSE = 9277604.124$$

Compared with the MSE with original data, we see a improvement in the performance for the scaled data.

Q13. Piece-wise Neural network with scaled data

In this section, we further explore Neural network method for three different periods described in Question 6. For each period, we try different architectures and select the best one for number of layers to be 1,2 and 3 using grid search with 5-folded CV. The MSE of fitting the entire data under different architectures and three periods are reported in table 10.

Table 10: MSE for neural net under different periods

Period	numbers of layers	layer sizes	fitting MSE
Before active	1	600	5110463.010
	2	(600,600)	3836845.647
	3	(600,600,500)	2180238.284
Active	1	600	298075960.221
	2	(600,600)	42518954.378
	3	(600,600,600)	15032521.862
After active	1	600	5074912.025
	2	(600,500)	557156.962
	3	(600,600,600)	199848.452

For all the three periods, the best architecture should contain 3 layers and with 600 sizes for each layer.

1.6. Using 5x window to predict

Q14. 5x window prediction

In the previous question, we only use the one period lag features to predict 'the number of tweets next period'. However, the time series data of 'number of tweets' might have dependence structures that can not be captured by just used one-period lag terms. In this part, we include all previous 5 hours/25 minutes lag features for making more accurate predictions. The test data consists of 3 set of examples with 3 time periods for each. We used aggregated data separated into 3 time periods for training, 3 types of regression including linear regression, random forest and gradient boosting are used for prediction. Predicted results are shown in table 11. Although the predicted results are not quite ideal, the order of magnitude is comparable to true values. For the first testing sample, linear regression gives the best prediction, while for the second and third testing sample, random forest gives the best prediction.

Table 11: Prediction for number of tweets in the next time window

	true value	linear regression	random forest	gradient boosting
sample 0 period 1	120	417	372	448
sample 0 period 2	1123	1939	4979	5460
sample 0 period 3	87	148	121	331
sample 1 period 1	846	1528	780	925
sample 1 period 2	903	5063	4367	4933
sample 1 period 3	46	-336	105	318
sample 2 period 1	61	1021	292	347
sample 2 period 2	28	9200	718	1157
sample 2 period 3	43	334	173	375

Part 2: Fan Base Prediction

First we find the user location for every tweet that includes #superbowl and classify the their locations into either "MA" (for Massachusetts) or "WA" (for Washington), see table 12. If the user's location string contains "Boston" for example, we classify it to the "MA" type. Those tweets not including keyword locations are not used for this part. We found there are 67978 tweets whose author locates in MA and 119391 in WA. We used binary label to denote the two locations, and we used 80% data as training data, 20% data as testing data.

Table 12: Keyword location used

MA	MA, Boston, boston, Massachusetts, massachusetts, Foxborough, "foxborough"
WA	WA, Seattle, seattle, Washington, washington, Kirkland, kirkland

Then we did classification analysis on the textual data collected. Feature extraction is done to construct the TF-IDF matrix, followed by dimension reduction through truncated SVD. Three kinds of classification algorithms are implemented, including logistic regression with L-2 regularization, random forest classifier and adaptive Boosting classifier. For logistic regression with L-2 regularization, ROC curve is reported in figure 8 and confusion matrix is reported in figure 9. For random forest classifier, ROC curve is reported in figure 10 and confusion matrix is reported in figure 11. For adaptive Boosting classifier, ROC curve is reported in figure 12 and confusion matrix is reported in figure 13. The accuracy, recall, precision and F-1 score for each classifier is reported in table 13. We observed that the random forest classifier has the best performance in all metrics.

Table 13: Keyword location used

	logistic regression	random forest	adaptive Boosting
accuracy	0.860	0.884	0.815
recall	0.928	0.940	0.913
precision	0.863	0.885	0.818
F-1 score	0.894	0.912	0.863

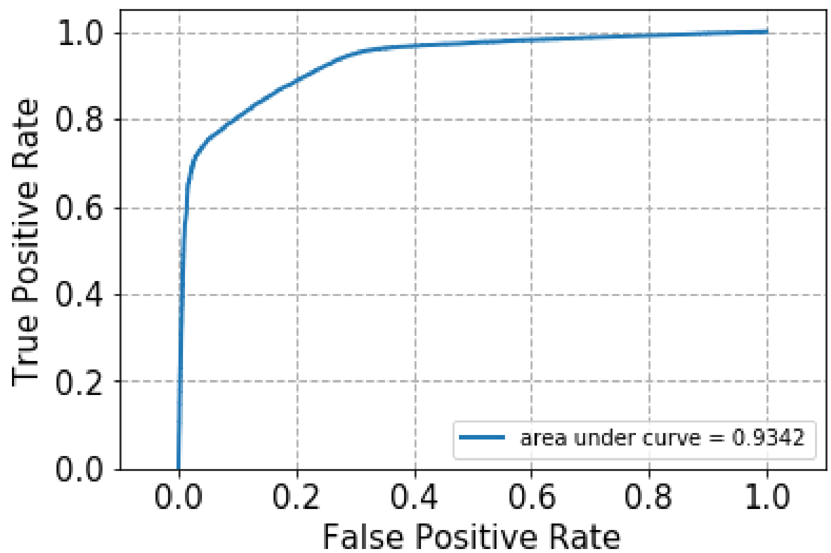
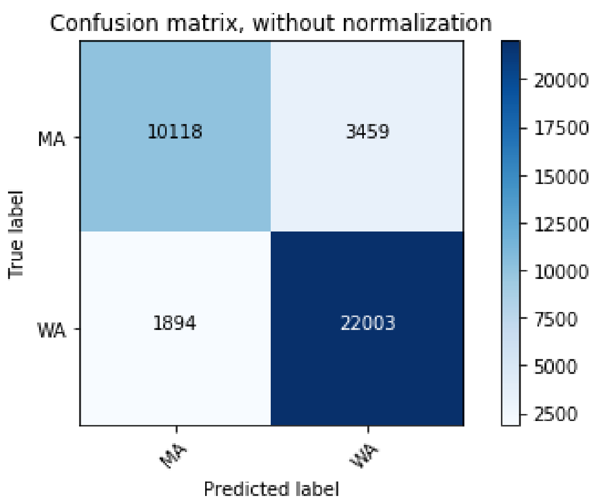
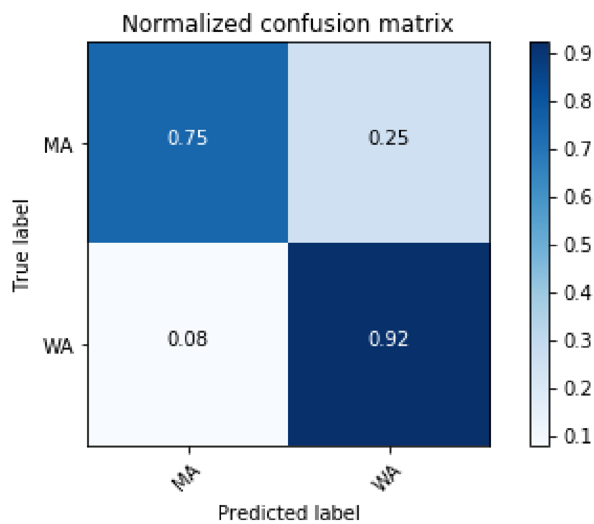


Figure 8: ROC curve of logistic regression classifier



(a)



(b)

Figure 9: Confusion matrix of logistic regression

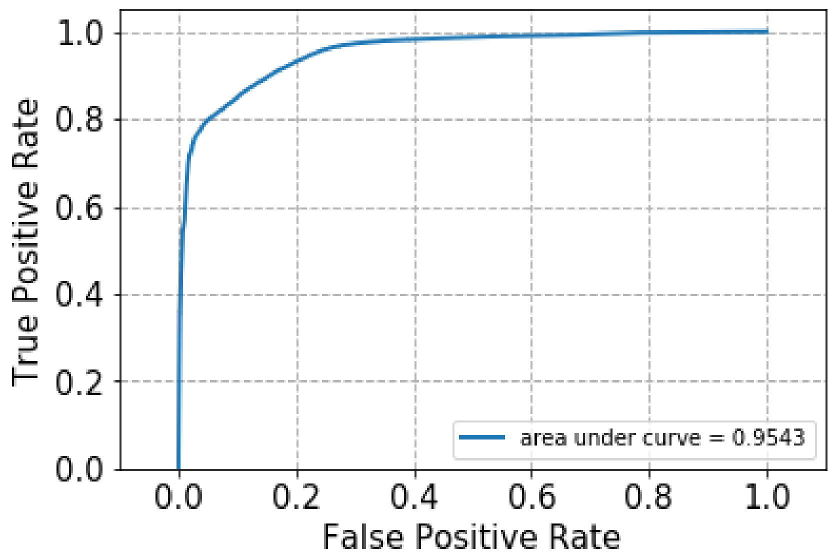
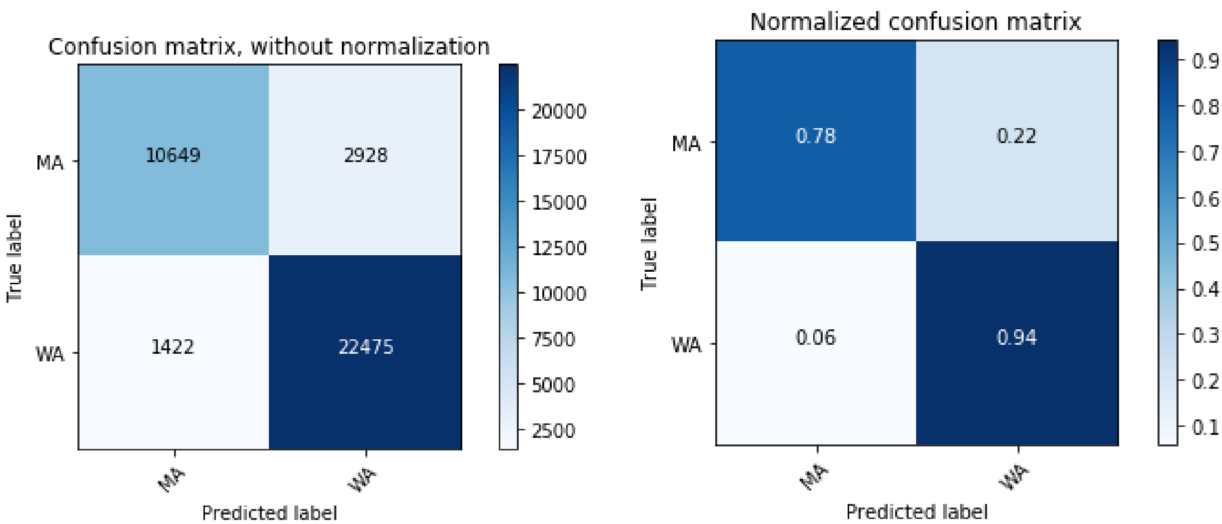


Figure 10: ROC curve of random forest classifier



(a)

(b)

Figure 11: Confusion matrix of random forest classifier

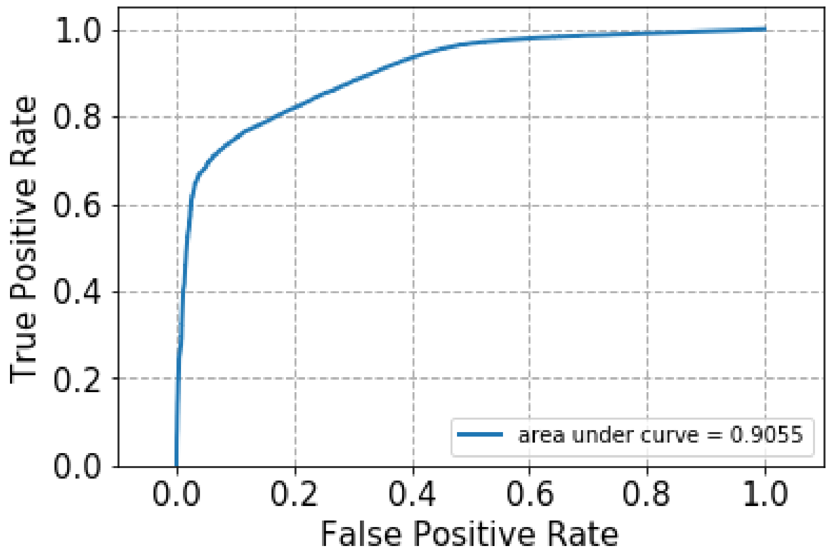
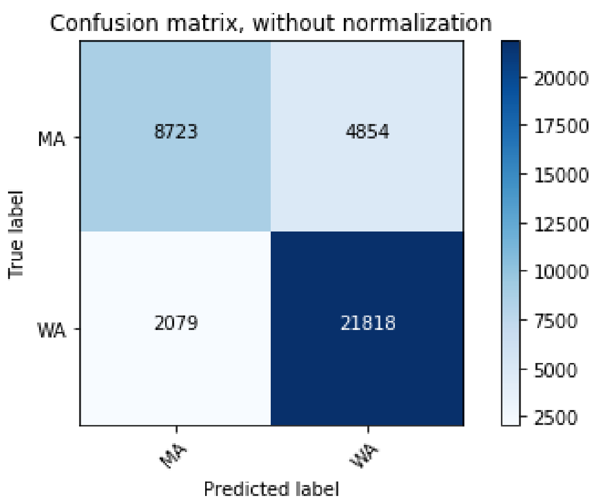
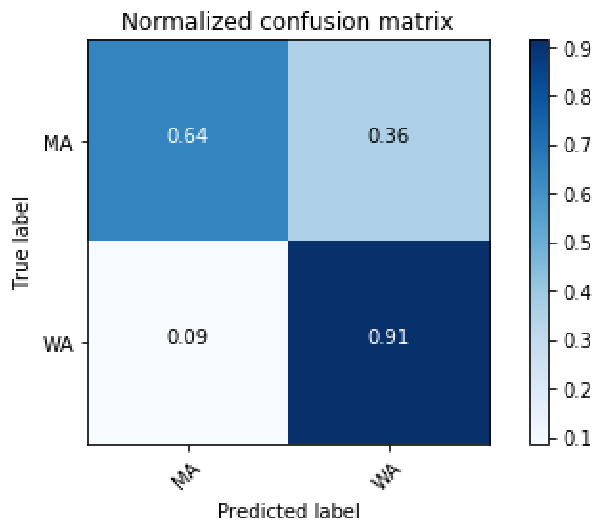


Figure 12: ROC curve of adaptive Boosting classifier



(a)



(b)

Figure 13: Confusion matrix of adaptive Boosting classifier

Part 3: Define your own project

Idea 1:Prediction of retweets number in next hour

The first idea is simple and directly inspired by the previous questions. We change the predictive value from the number of tweets number in next hour to the number of retweets number in next hour. The application here is that the number of retweets number could be served as an important indicator of web advertising. Imagine that, a company can utilize embedded advertising strategy in one tweet and then distribute the advertisement to more audiences via high number of retweets. We followed the previous solving procedure, using five features including number of tweets,total number of retweets,follower numbers, maximum follower number and time of the day to predict the retweets number in next hour. We tried gradient boosting method and neural networks and performed grid search to find the best fitting model from cross validation and reported the MSE here. The best model here for gradientboosting regressor is listed at table 14 and the best model for neural network is three hidden layers with each layers has five hundred hidden unit with MSE in cross validation test $MSE = 1201219940.3$. We can further use sample data in Q14 that to verify our model by predicting untrained data. In this case, the MSE for gradient boosting is $MSE = 139414602.17$ and for the neural network is $MSE = 49221304.908$, which means the neural network model is better than gradient boosting in predicting retweets number.

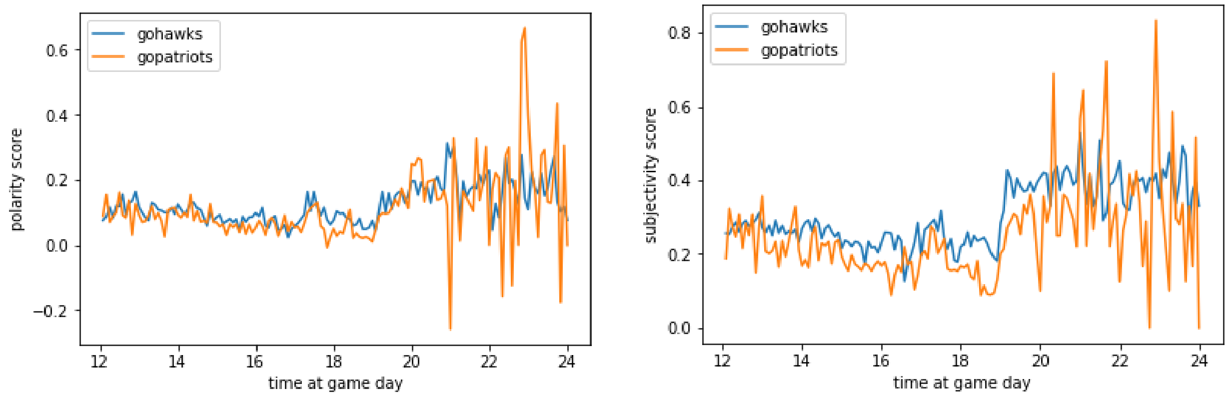
Table 14: Grid search result for Gradient Boosting

max_depth	20
max_features	auto
min_samples_leaf	2
min_samples_split	2
n_estimators	10
CV Test MSE	1595567074.7

Idea 2: Sentiment analysis on game day

In this part, we analyzed the sentiment of Twitter data using the toolbox “TextBlob”. We can guess that the fan’s will be most emotional on the game day, especially within several hours before and after the game. In this part, we first filter out all the tweets from 2015-02-01 12:00pm to 2015-02-02 12:00am. According to the record on the Internet, the super bowl XLIX started at 3:30pm and lasted about 3 hours and 37 minutes. Therefore, the 12 hours we select is a period of great interest. For each tweet, we use the text as the input of TextBlob, which returns a polarity and subjectivity score for each tweet. The polarity score is a float within the range [-1.0, 1.0], which describes how negative or positive the sentiment is. The subjectivity is a float within the range [0.0, 1.0] where 0.0 is very objective and 1.0 is very subjective. To reduce the variance, the mean score in 5 minutes are used for analysis.

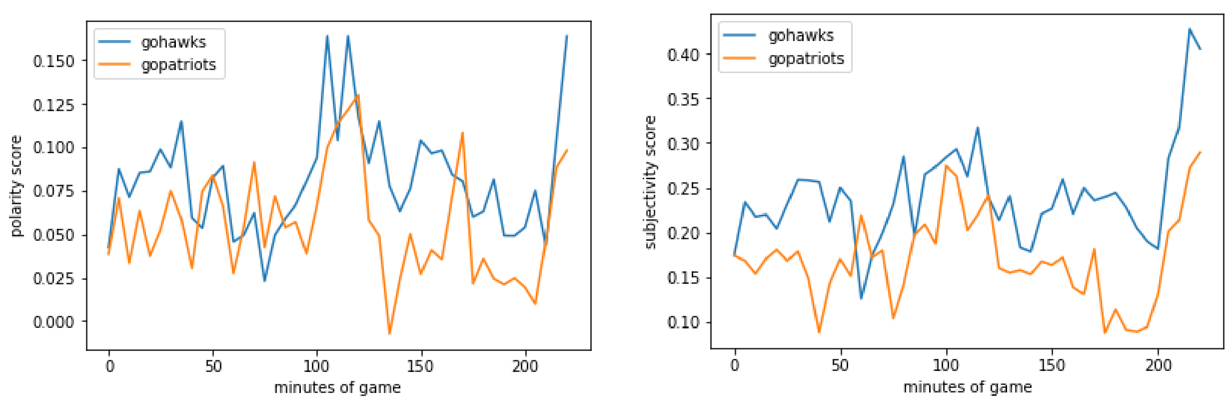
To compare the sentiment for fans of the two teams, we use the 2 files ”gohawks” and ”gopatriots”. Figure 14 compares the polarity and subjectivity for fans on the game day. Before the game started, the sentiments for fans of these 2 teams are very similar and the scores are stable. After the game started at 15:30, the scores showed more and more oscillations which are reasonable because the fans’ emotion changed dramatically. In evening of the game day, the sentiment of the Patriots’ fans showed much stronger oscillation than the Hawks’ fan, which probably related to the fact that they won the championship with a huge comeback. Figure 15 compares the polarity and subjectivity during the game. The polarity shows that the Hawks’ fans are more positive and subjective during the game because they had an advantage in most of the game.



(a) (b)

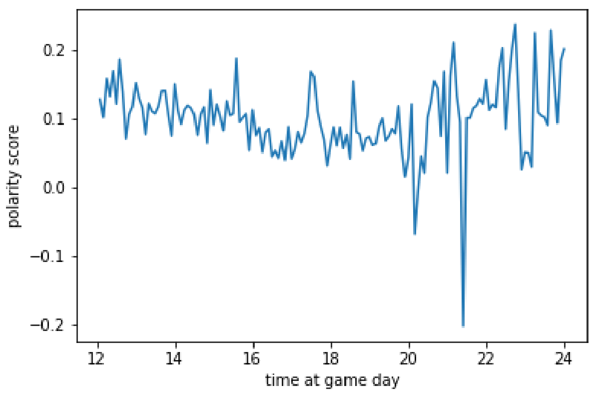
Figure 14: Comparison of fan's sentiment from 12pm to 12am

Figures 16 to 19 show the sentiment for tweets in other files. A interesting observation is that the tweets in "sb49" and "patriots" are both very positive and subjective before the game. This is because the fans are probably have a very high expectation on either the game or the Patriots. Therefore, they try to give strong declarations to show their views. After the game started, they became less positive and more objective.

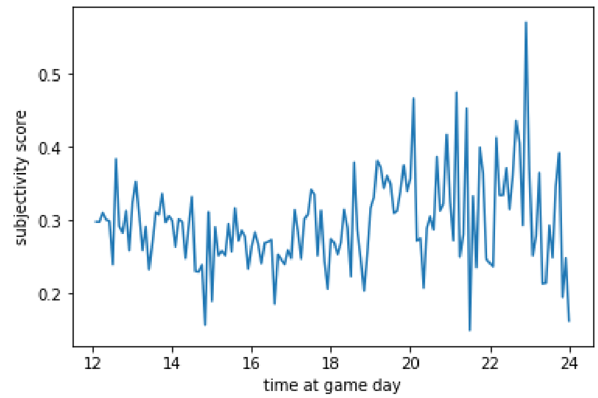


(a) (b)

Figure 15: Comparison of fan's sentiment during the game

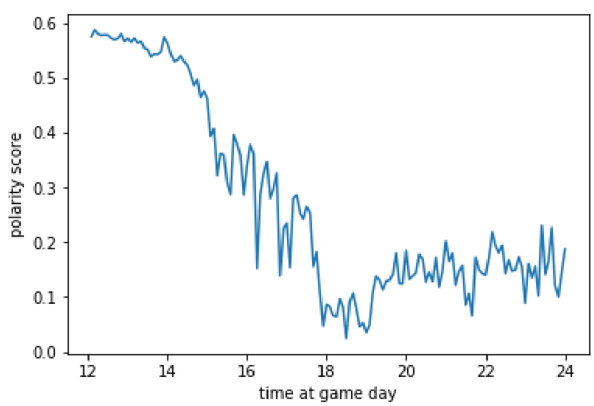


(a)

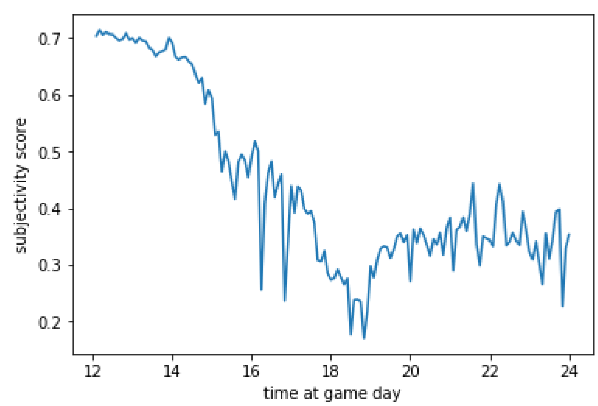


(b)

Figure 16: Sentiment of tweets during the game (#nfl)

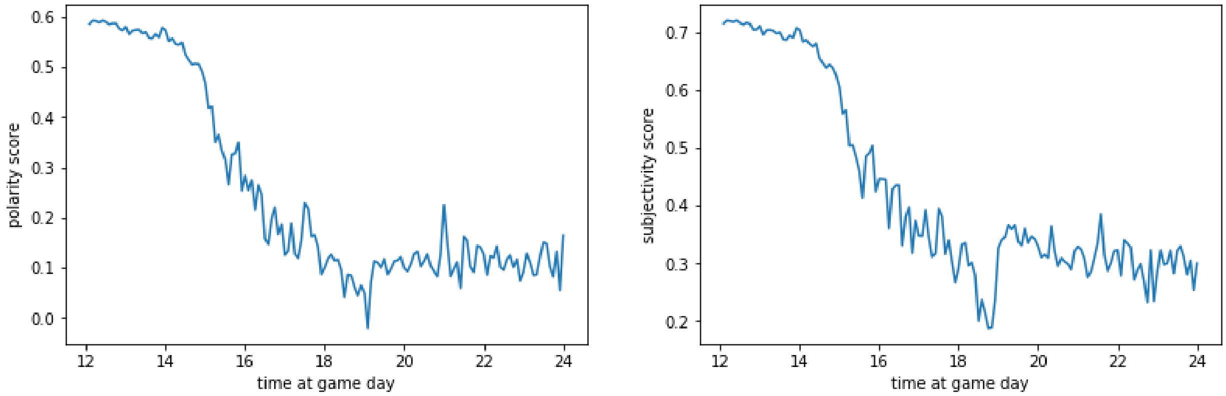


(a)



(b)

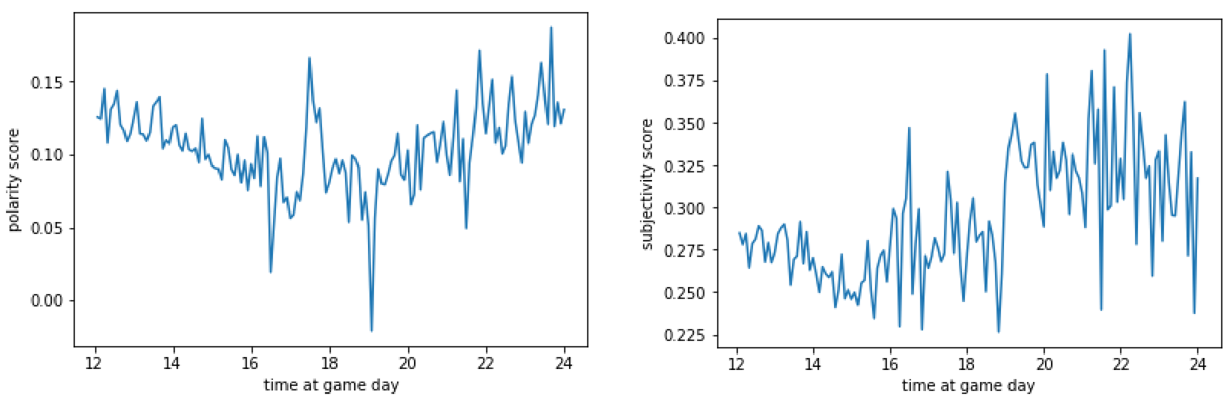
Figure 17: Sentiment of tweets during the game (#patriots)



(a)

(b)

Figure 18: Sentiment of tweets during the game (#sb49)



(a)

(b)

Figure 19: Sentiment of tweets during the game (#superbowl)