EE219 Large Scale Data Mining Models and Algorithms

Winter 2018_Project5

Team member:

Dui Lin (504759948)

Xinyi Jiang (904818856)

Zhenli Jiang (304878235)

Content

1. Introduction

A useful practice in social network analysis is to predict future popularity of a subject or event. Twitter, with its public discussion model, is a good platform to perform such analysis. With Twitter's topic structure in mind, the problem can be stated as: knowing current (and previous) tweet activity for a hashtag, can we predict its tweet activity in the future? More specifically, can we predict if it will become more popular and if so by how much? In this project, we will try to formulate and solve an instance of such problems.

In this project, we try to predict the popularity of a topic on Twitter. More formally, knowing the previous and current tweet activity for a hashtag, we try to predict its tweet activity in the future and aim to determine whether it gets more or less popular and by how much. For this, we use Regression Models.

2. Problem Statement and Results

2.1 Popularity Prediction

2.1.1

In problem 1.1， our goal is to do some basic statistics calculation of the training tweet data for each hashtag. And also, we show histograms with 1-hour bins that show the number the tweets in hour over time for two hashtag groups, #SuperBowl and #NFL.

1. GoHawks

```
**************************************************
*
Info of #GoHawks
    Average number of tweets per hour: 325.371591304331
16
    Average number of followers per user: 1588.18866293
00582
    Average number of followers per tweet: 2203.9317674
44827
    Average number of retweets per tweet: 2.01461708551
2608

**************************************************
*
```

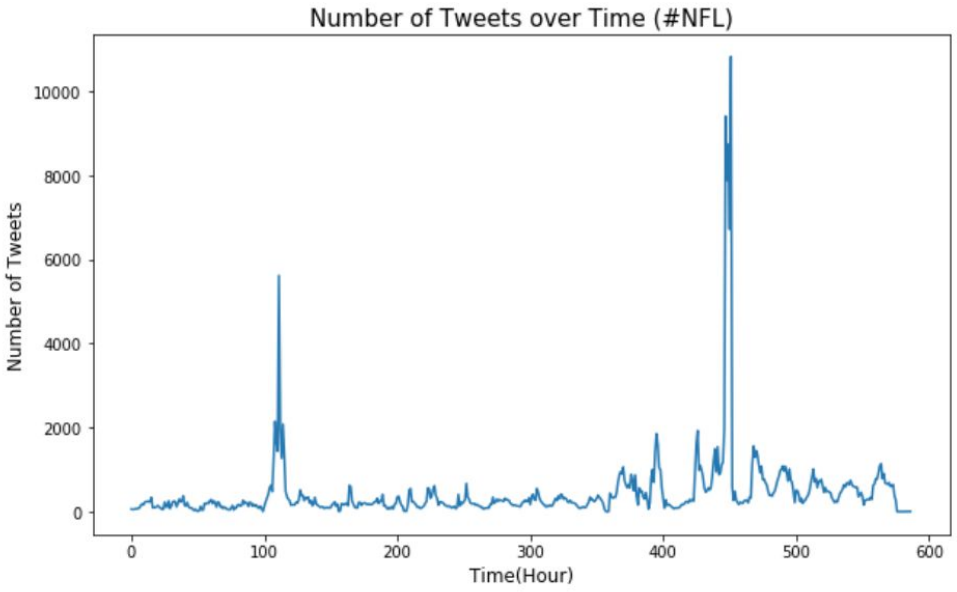
2. GoPatriots

```
**********************************************************
*
Info of #GoPatriots
    Average number of tweets per hour: 45.6945105735620
3
    Average number of followers per user: 1294.46936646
26748
    Average number of followers per tweet: 1401.8955093
016164
    Average number of retweets per tweet: 1.40008386703
26319

**********************************************************
*
```

3. NFL

```
**********************************************************
*
Info of #NFL
    Average number of tweets per hour: 441.323431137395
8
    Average number of followers per user: 4221.07698786
5717
    Average number of followers per tweet: 4653.2522855
02502
    Average number of retweets per tweet: 1.53853310890
11056
```




```
**********************************************************
*
```

4. Patriots

```
**********************************************************
*
Info of #Patriots
    Average number of tweets per hour: 834.555509164188
6
    Average number of followers per user: 1695.27106214
77224
    Average number of followers per tweet: 3309.9788284
15827
    Average number of retweets per tweet: 1.78281564916
59402

**********************************************************
*
```

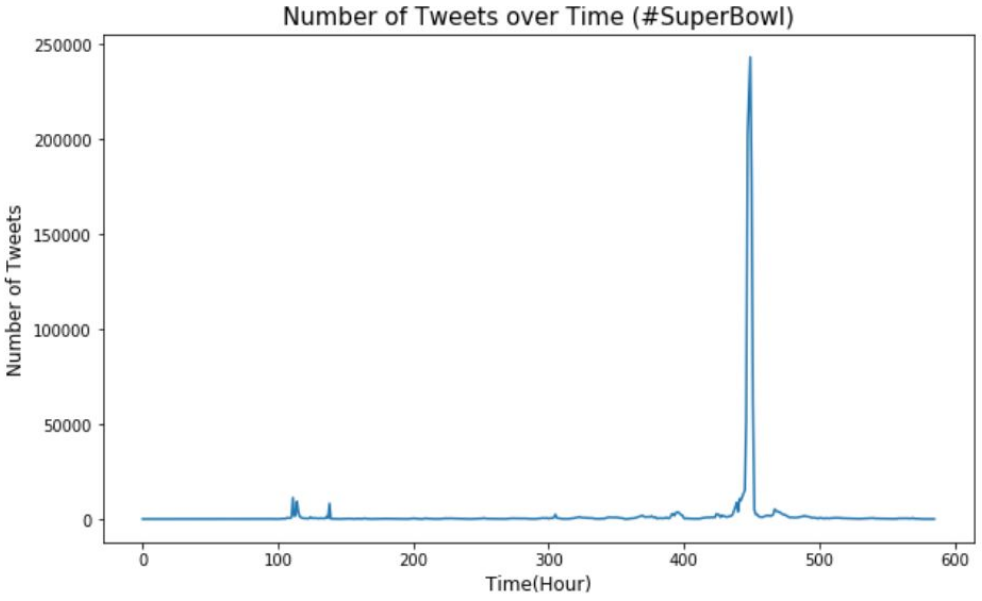
5. SB49

```
**********************************************************
*
Info of #SB49
    Average number of tweets per hour: 1419.88790748719
02
    Average number of followers per user: 2250.85025774
2912
    Average number of followers per tweet: 10267.316849
48685
    Average number of retweets per tweet: 2.51114878632
47035

**********************************************************
*
```

6. SuperBowl

```
**************************************************
*
Info of #SuperBowl
    Average number of tweets per hour: 2302.50040188332
74
    Average number of followers per user: 3798.67911708
5927
    Average number of followers per tweet: 8858.9746627
84603
    Average number of retweets per tweet: 2.38827239990
30224
```



2.1.2

Problem 1.2 asked us to predict the number of tweets in the nest hour by fitting a Linear Regression model for each hashtag. The p-value for each parameter tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that we can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to the model because changes in the predictor's value are related to changes in the response variable. Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response. On the other hand, the t-statistic is useful for making inferences about the regression coefficients. The hypothesis test on coefficient i tests the null hypothesis that it is equal to zero (meaning the corresponding term is not significant) versus the alternate hypothesis that the coefficient is different from zero. There we would want to consider features with high t-test values.

The features we use are:

- X1：Maximum number of followers of the users posting the hashtag
- X2：Time of the day (which could take 24 values that represent hours of the day with respect to a given time zone)
- X3：Sum of the number of followers of the users posting the hashtag
- X4：Total number of retweets (hashtag of interest)
- X5：Number of tweets (hashtag of interest)

1. GoHawks

```
Details for Linear Regression Model for tweets_#gohawks.txt
##########################################################################
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.490
Model:                            OLS   Adj. R-squared:                  0.488
Method:                 Least Squares   F-statistic:                     186.0
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          8.81e-139
Time:                        01:08:32   Log-Likelihood:                -7817.0
No. Observations:                 973   AIC:                         1.565e+04
Df Residuals:                     967   BIC:                         1.568e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         39.0200     35.144      1.110      0.267       -29.947   107.987
x1             0.5719      0.121      4.716      0.000         0.334     0.810
x2            -0.1692      0.043     -3.909      0.000        -0.254    -0.084
x3             6.2179      3.122      1.992      0.047         0.091    12.345
x4             0.0004   8.15e-05      4.586      0.000         0.000     0.001
x5            -0.0007      0.000     -4.915      0.000        -0.001    -0.000
==============================================================================
Omnibus:                     1848.594   Durbin-Watson:                   2.336
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4377767.947
Skew:                          13.277   Prob(JB):                         0.00
Kurtosis:                     330.531   Cond. No.                     2.39e+06
==============================================================================
```

Best features found based on p and t-values are: **number of followers** and **number of tweets**. However, the R-squared value is relatively low (only 0.490), which means the model did not fit very well.

2. GoPatriots

```
Details for Linear Regression Model for tweets_#gopatriots.txt
######################################################################
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.662
Method:                 Least Squares   F-statistic:                     268.4
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          4.60e-158
Time:                        01:08:34   Log-Likelihood:                -4453.3
No. Observations:                 684   AIC:                             8919.
Df Residuals:                     678   BIC:                             8946.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const          4.2832      8.891      0.482      0.630       -13.174    21.740
x1            -0.5687      0.240     -2.369      0.018        -1.040    -0.097
x2             0.3815      0.262      1.456      0.146        -0.133     0.896
x3             0.7502      0.827      0.908      0.364        -0.873     2.373
x4             0.0011      0.000      5.432      0.000         0.001     0.002
x5            -0.0012      0.000     -6.359      0.000        -0.002    -0.001
==============================================================================
Omnibus:                      796.993   Durbin-Watson:                   2.103
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           452112.690
Skew:                           4.844   Prob(JB):                         0.00
Kurtosis:                     128.578   Cond. No.                     4.69e+05
==============================================================================
```

Best features found based on p and t-values are: **number of followers** and **number of retweets**. The R-squared value is 0.664, which means the model fit fairly well.

3. NFL

```
Details for Linear Regression Model for tweets_#nfl.txt
######################################################################
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.602
Method:                 Least Squares   F-statistic:                     281.7
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          9.19e-183
Time:                        01:08:54   Log-Likelihood:                -6999.4
No. Observations:                 927   AIC:                         1.401e+04
Df Residuals:                     921   BIC:                         1.404e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         33.7633     21.497      1.571      0.117        -8.425    75.952
x1             1.3297      0.110     12.078      0.000         1.114     1.546
x2            -0.1779      0.065     -2.722      0.007        -0.306    -0.050
x3             2.1759      2.036      1.069      0.286        -1.821     6.172
x4            -0.0001    2.5e-05     -5.600      0.000        -0.000 -9.11e-05
x5             0.0002    3.4e-05      5.622      0.000         0.000     0.000
==============================================================================
Omnibus:                     1053.806   Durbin-Watson:                   2.146
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1256393.880
Skew:                           4.531   Prob(JB):                         0.00
Kurtosis:                     183.127   Cond. No.                     3.91e+06
==============================================================================
```

Best features found based on p and t-values are: **maximum number of followers** and **number of tweets**. The R-squared value is 0.605, which means the model fit decently.

4. Patriots

```
Details for Linear Regression Model for tweets_#patriots.txt
##########################################################################
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.715
Method:                 Least Squares   F-statistic:                     492.1
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          1.06e-263
Time:                        01:09:31   Log-Likelihood:                -8761.2
No. Observations:                 981   AIC:                         1.753e+04
Df Residuals:                     975   BIC:                         1.756e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const         72.1464     83.417      0.865      0.387       -91.550   235.843
x1             1.7894      0.079     22.523      0.000         1.634     1.945
x2            -0.9539      0.073    -13.071      0.000        -1.097    -0.811
x3             6.7396      7.839      0.860      0.390        -8.644    22.123
x4             0.0003   4.28e-05      7.792      0.000         0.000     0.000
x5            -0.0003   9.01e-05     -2.844      0.005        -0.000 -7.94e-05
==============================================================================
Omnibus:                     1877.207   Durbin-Watson:                   1.694
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          4075004.536
Skew:                          13.560   Prob(JB):                         0.00
Kurtosis:                     317.577   Cond. No.                     7.10e+06
==============================================================================
```

Best features found based on p and t-values are: **number of followers** and **number of tweets**. The R-squared value is 0.716, which means the model fit well.

5. SB49

```
Details for Linear Regression Model for tweets_#sb49.txt
##########################################################################
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.821
Model:                            OLS   Adj. R-squared:                  0.819
Method:                 Least Squares   F-statistic:                     528.7
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          1.01e-212
Time:                        01:10:35   Log-Likelihood:                -5702.2
No. Observations:                 583   AIC:                         1.142e+04
Df Residuals:                     577   BIC:                         1.144e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        138.7825    323.784      0.429      0.668      -497.156   774.721
x1             1.1410      0.052     21.899      0.000         1.039     1.243
x2            -0.3677      0.043     -8.475      0.000        -0.453    -0.283
x3           -15.4646     25.008     -0.618      0.537       -64.582    33.653
x4             0.0002   2.96e-05      7.420      0.000         0.000     0.000
x5            -0.0003   6.92e-05     -4.086      0.000        -0.000    -0.000
==============================================================================
Omnibus:                     1163.174   Durbin-Watson:                   1.726
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2251333.588
Skew:                          14.042   Prob(JB):                         0.00
Kurtosis:                     306.134   Cond. No.                     5.73e+07
==============================================================================
```

Best features found based on p and t-values are: **number of followers** and **number of tweets**. However, the R-squared value is 0.821, which means the model fit very well.

6. SuperBowl

```
Details for Linear Regression Model for tweets_#superbowl.txt
##########################################################################
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     552.3
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          3.39e-279
Time:                        01:12:23   Log-Likelihood:                -9919.2
No. Observations:                 964   AIC:                          1.985e+04
Df Residuals:                     958   BIC:                          1.988e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        136.6962    318.892      0.429      0.668      -489.112   762.504
x1             1.6751      0.258      6.487      0.000         1.168     2.182
x2             0.0245      0.126      0.195      0.846        -0.222     0.271
x3             0.1737     31.361      0.006      0.996       -61.370    61.717
x4            -0.0004   2.58e-05    -13.814      0.000        -0.000    -0.000
x5             0.0013      0.000      9.530      0.000         0.001     0.002
==============================================================================
Omnibus:                     1889.238   Durbin-Watson:                   1.698
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          5789800.589
Skew:                          14.125   Prob(JB):                         0.00
Kurtosis:                     381.611   Cond. No.                     6.34e+07
==============================================================================
```

Best features found based on p and t-values are: **maximum number of followers** and **number of tweets**. The R-squared value is 0.742, which means the model fit well.

According to the results above, it can be concluded that top 3 important features are **Maximum number of followers, Number of retweets and Number of tweets.**

2.1.3

In this part we aim to train the model using features of our own. We selected the following features in addition to the ones mentioned in the previous part:

X1: 'totalTweets'
X2: 'retweets'
X3: 'time'
X4: 'followers'
X5: 'favorite_count'
X6: 'ranking_score'
X7: 'urls'
X8: 'user_count'
X9: 'impressions'

Here are the different top-3 features for each group from observation of t-test and p-values:

1. GoHawks

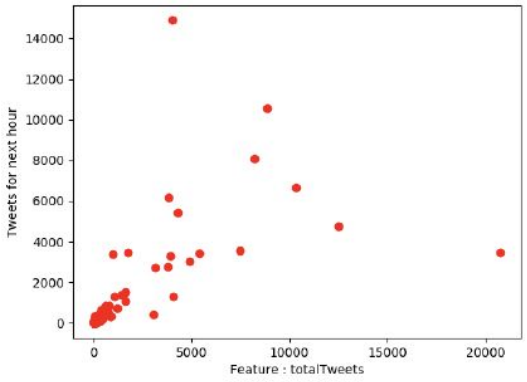
OLS Regression Results

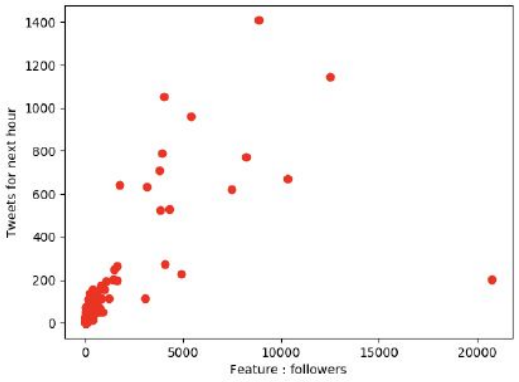
Dep. Variable:	y	R-squared:	0.639
Model:	OLS	Adj. R-squared:	0.636
Method:	Least Squares	F-statistic:	189.7
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	2.34e-206
Time:	18:58:41	Log-Likelihood:	-7648.6
No. Observations:	973	AIC:	1.532e+04
Df Residuals:	963	BIC:	1.537e+04
Df Model:	9		
Covariance Type:	nonrobust		

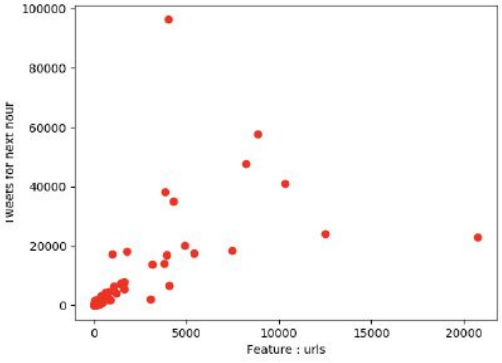
	coef	std err	t	P>|t|	[95.0%	Conf. Int.]
const	1.2665	28.654	0.044	0.965	-54.964	57.498
x1	-38.3959	2.963	-12.959	0.000	-44.210	-32.582
x2	-0.2011	0.055	-3.676	0.000	-0.308	-0.094
x3	2.9099	2.563	1.135	0.257	-2.120	7.940
x4	-0.0003	4.98e-05	-6.967	0.000	-0.000	-0.000
x5	0.0881	0.021	4.222	0.000	0.047	0.129
x6	7.7322	0.587	13.172	0.000	6.580	8.884
x7	9.1580	0.775	11.824	0.000	7.638	10.678
x8	4.5862	0.737	6.223	0.000	3.140	6.032
x9	-4.603e-10	1.95e-10	-2.364	0.018	-8.42e-10	-7.82e-11

Omnibus:	1962.042	Durbin-Watson:	2.216
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5530322.805
Skew:	15.179	Prob(JB):	0.00
Kurtosis:	371.089	Cond. No.	4.05e+11

The top-3 are: totalTweets, followers, urls; And here are scatter plots:


Tweets for next hour
Feature : totalTweets
0
2000
4000
6000
8000
10000
12000
14000
0
5000
10000
15000
20000


Tweets for next hour
Feature : followers
0
200
400
600
800
1000
1200
1400
0
5000
10000
15000
20000


Tweets for next hour
Feature : urls
0
20000
40000
60000
80000
100000
0
5000
10000
15000
20000

2. GoPatriots

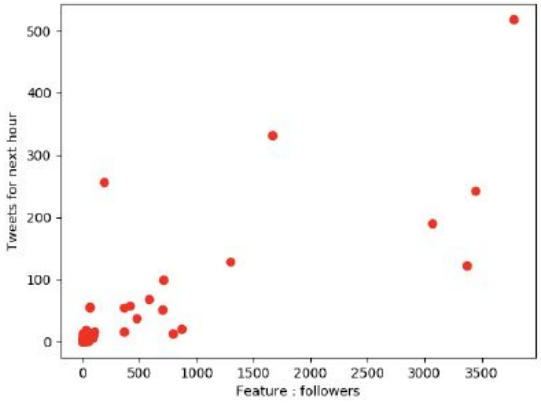
OLS Regression Results

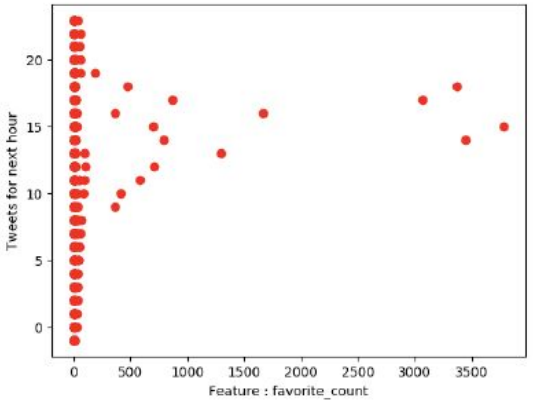
Dep. Variable:	y	R-squared:	0.792
Model:	OLS	Adj. R-squared:	0.789
Method:	Least Squares	F-statistic:	284.8
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	5.83e-223
Time:	18:58:43	Log-Likelihood:	-4290.0
No. Observations:	684	AIC:	8600.
Df Residuals:	674	BIC:	8645.
Df Model:	9		
Covariance Type:	nonrobust		

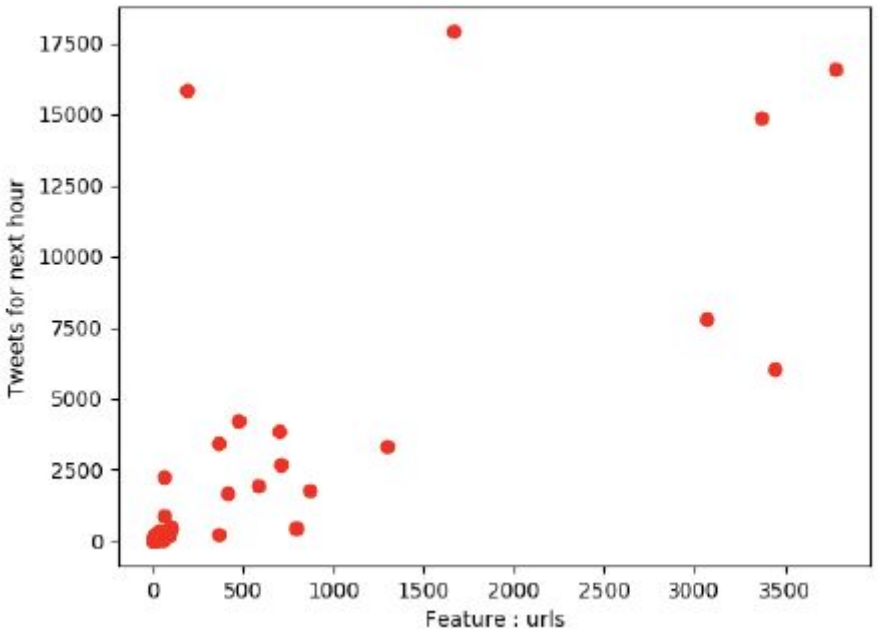
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	1.7719	6.733	0.263	0.792	-11.447	14.991
x1	-1.0377	1.981	-0.524	0.601	-4.927	2.852
x2	-0.6442	0.233	-2.761	0.006	-1.102	-0.186
x3	0.8724	0.625	1.395	0.163	-0.355	2.100
x4	-1.419e-05	4.42e-05	-0.321	0.749	-0.000	7.27e-05
x5	-7.1164	1.765	-4.031	0.000	-10.583	-3.650
x6	0.9229	0.359	2.572	0.010	0.218	1.628
x7	10.1974	0.795	12.835	0.000	8.637	11.757
x8	-2.9745	0.665	-4.472	0.000	-4.280	-1.669
x9	-2.664e-09	4.15e-09	-0.642	0.521	-1.08e-08	5.49e-09

Omnibus:	769.909	Durbin-Watson:	1.944
Prob(Omnibus):	0.000	Jarque-Bera (JB):	316337.494
Skew:	4.641	Prob(JB):	0.00
Kurtosis:	107.945	Cond. No.	1.06e+10

The top-3 are: followers, favourite_count, urls; And here are scatter plots:







3. NFL

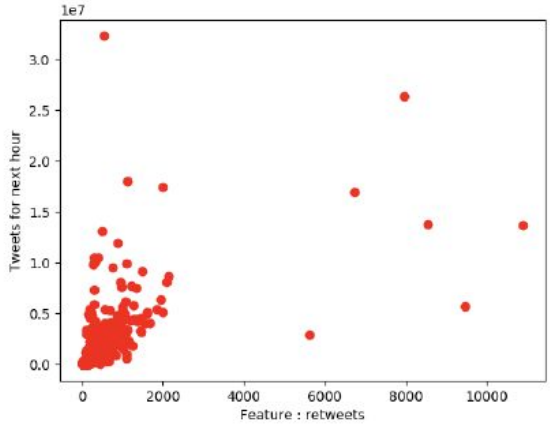
OLS Regression Results

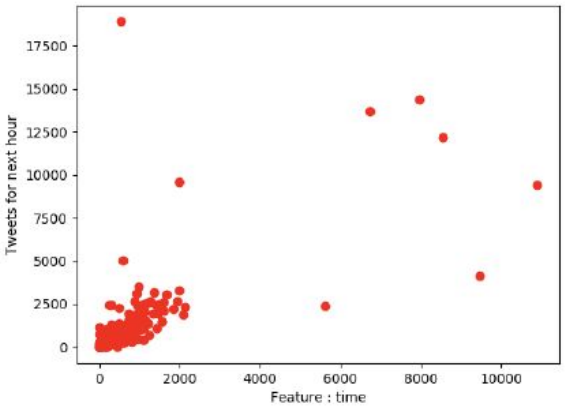
Dep. Variable:	y	R-squared:	0.720
Model:	OLS	Adj. R-squared:	0.717
Method:	Least Squares	F-statistic:	261.5
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	3.55e-246
Time:	18:59:06	Log-Likelihood:	-6840.1
No. Observations:	927	AIC:	1.370e+04
Df Residuals:	917	BIC:	1.375e+04
Df Model:	9		
Covariance Type:	nonrobust		

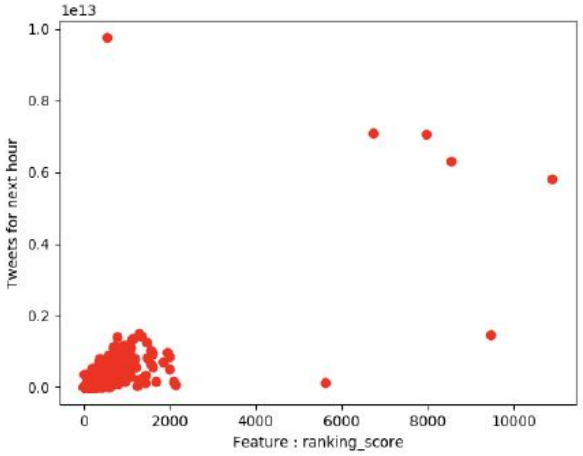
	coef	std err	t	P>|t|	[95.0%	Conf. Int.]
const	30.6816	18.207	1.685	0.092	-5.051	66.414
x1	1.9385	1.283	1.511	0.131	-0.579	4.456
x2	0.1184	0.059	1.990	0.047	0.002	0.235
x3	1.0294	1.834	0.561	0.575	-2.570	4.629
x4	1.572e-05	1.1e-05	1.431	0.153	-5.84e-06	3.73e-05
x5	-2.3785	0.162	-14.654	0.000	-2.697	-2.060
x6	-0.2160	0.266	-0.812	0.417	-0.738	0.306
x7	-0.0539	0.140	-0.384	0.701	-0.330	0.222
x8	-0.4702	0.325	-1.448	0.148	-1.107	0.167
x9	1.877e-10	7.82e-11	2.402	0.016	3.44e-11	3.41e-10

Omnibus:	1603.652	Durbin-Watson:	2.219
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1404624.689
Skew:	11.155	Prob(JB):	0.00
Kurtosis:	192.388	Cond. No.	8.79e+11

The top-3 are: retweets, time, ranking_score; And here are scatter plots:


1e7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Tweets for next hour
0
2000
4000
6000
8000
10000
Feature : retweets


17500
15000
12500
10000
7500
5000
2500
0
Tweets for next hour
0
2000
4000
6000
8000
10000
Feature : time


1e13
1.0
0.8
0.6
0.4
0.2
0.0
Tweets for next hour
0
2000
4000
6000
8000
10000
Feature : ranking_score

4. Patriots

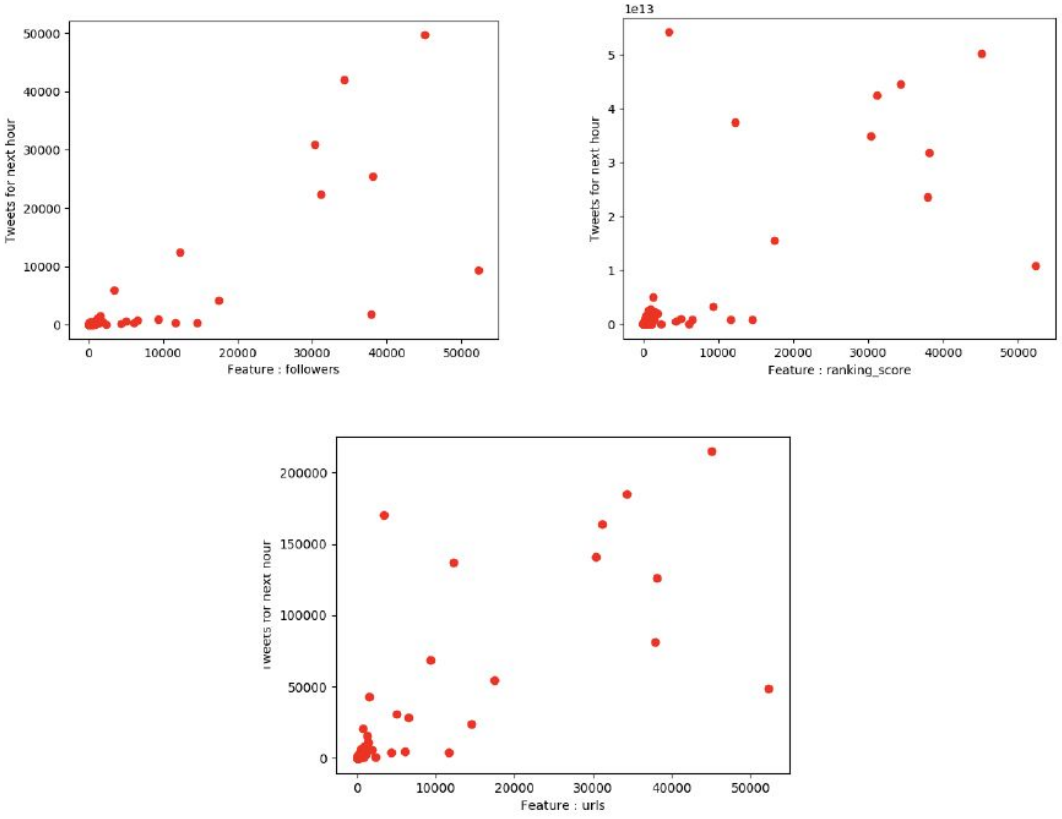
OLS Regression Results

Dep. Variable:	y	R-squared:	0.778
Model:	OLS	Adj. R-squared:	0.776
Method:	Least Squares	F-statistic:	377.2
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	1.03e-309
Time:	18:59:46	Log-Likelihood:	-8641.6
No. Observations:	981	AIC:	1.730e+04
Df Residuals:	971	BIC:	1.735e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-65.2812	71.623	-0.911	0.362	-205.836	75.273
x1	-49.8016	4.203	-11.848	0.000	-58.050	-41.553
x2	-0.4334	0.118	-3.672	0.000	-0.665	-0.202
x3	-10.3892	6.750	-1.539	0.124	-23.636	2.858
x4	7.794e-05	3.6e-05	2.167	0.031	7.35e-06	0.000
x5	-0.1429	0.180	-0.795	0.427	-0.496	0.210
x6	10.4596	0.825	12.672	0.000	8.840	12.079
x7	5.7093	0.376	15.203	0.000	4.972	6.446
x8	2.4032	0.797	3.017	0.003	0.840	3.966
x9	5.705e-10	9.82e-11	5.807	0.000	3.78e-10	7.63e-10

Omnibus:	1989.577	Durbin-Watson:	1.637
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5662268.142
Skew:	15.397	Prob(JB):	0.00
Kurtosis:	373.915	Cond. No.	5.19e+12

The top-3 are: followers, ranking_score, urls; And here are scatter plots:


Tweets for next hour
Feature : followers
1e13
Tweets for next hour
Feature : ranking_score
Tweets for next hour
Feature : urls

5. Sb49

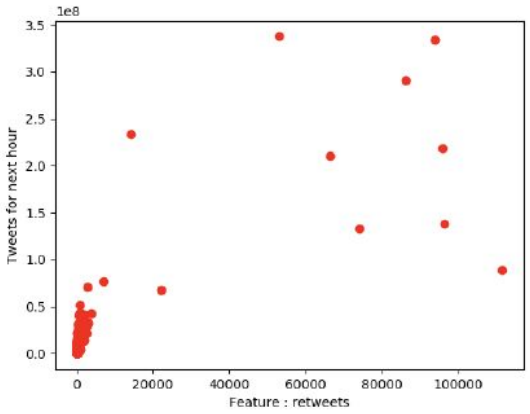
OLS Regression Results

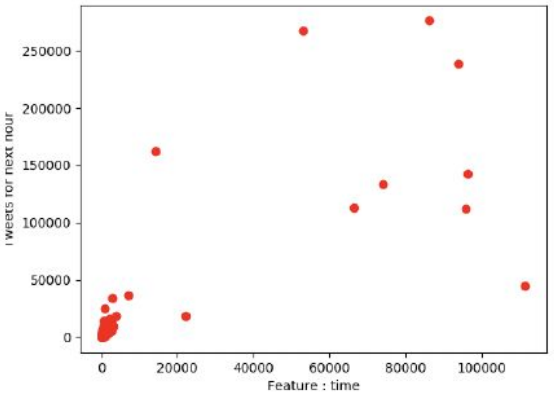
Dep. Variable:	y	R-squared:	0.863
Model:	OLS	Adj. R-squared:	0.861
Method:	Least Squares	F-statistic:	402.7
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	3.69e-241
Time:	19:00:55	Log-Likelihood:	-5623.0
No. Observations:	583	AIC:	1.127e+04
Df Residuals:	573	BIC:	1.131e+04
Df Model:	9		
Covariance Type:	nonrobust		

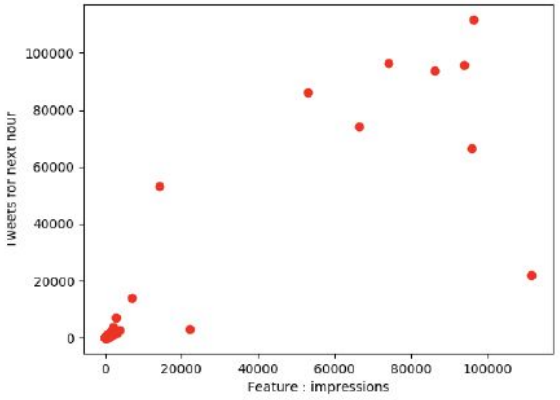
	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
const	-95.6047	281.680	-0.339	0.734	-648.856	457.647
x1	16.5686	8.743	1.895	0.059	-0.603	33.740
x2	0.3104	0.108	2.872	0.004	0.098	0.523
x3	-22.4972	21.573	-1.043	0.297	-64.870	19.876
x4	0.0001	2.33e-05	5.891	0.000	9.13e-05	0.000
x5	-0.2443	0.089	-2.755	0.006	-0.419	-0.070
x6	-3.2295	1.811	-1.783	0.075	-6.787	0.328
x7	-2.4354	1.030	-2.365	0.018	-4.458	-0.413
x8	0.3526	0.828	0.426	0.670	-1.273	1.978
x9	-4.415e-10	4.2e-11	-10.503	0.000	-5.24e-10	-3.59e-10

Omnibus:	1208.619	Durbin-Watson:	1.906
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2467337.890
Skew:	15.347	Prob(JB):	0.00
Kurtosis:	320.221	Cond. No.	7.13e+13

The top-3 are: retweets, time, impressions; And here are scatter plots:







6. Superbowl

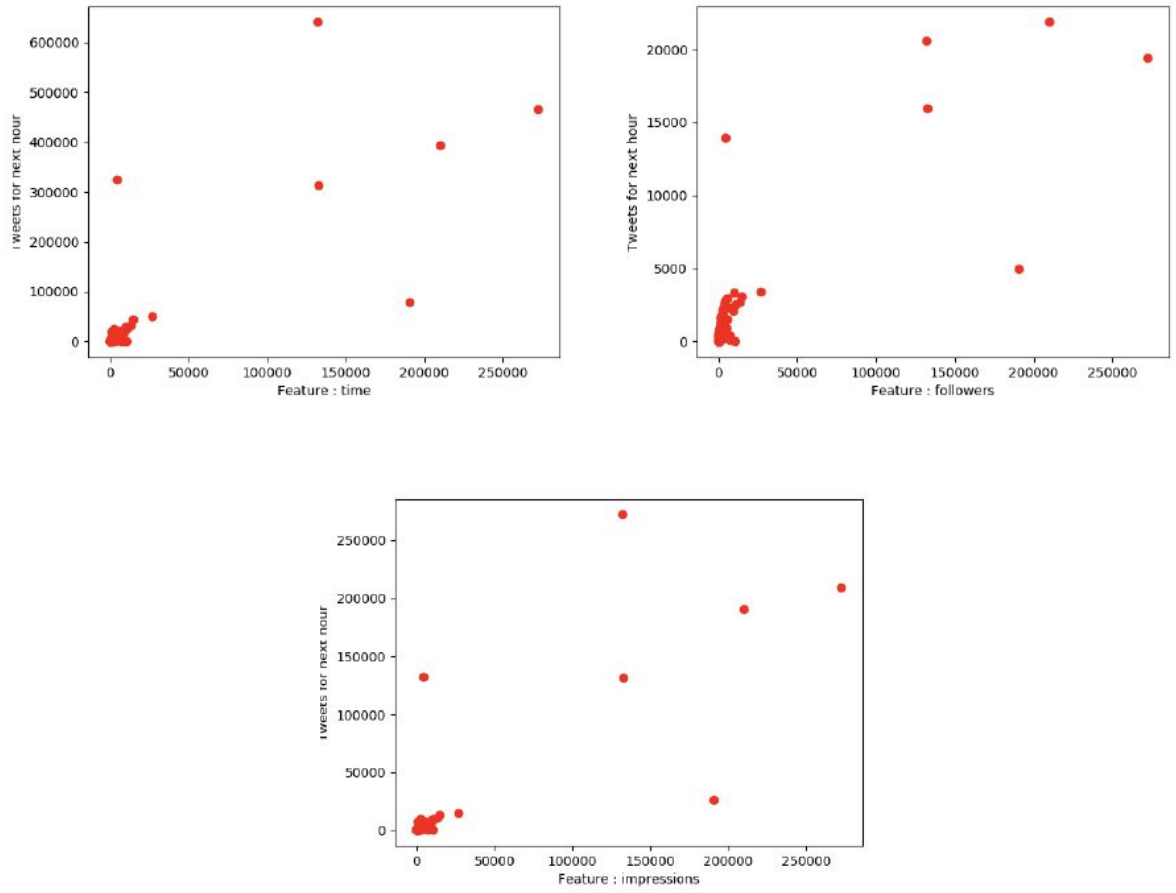
OLS Regression Results

Dep. Variable:	y	R-squared:	0.891
Model:	OLS	Adj. R-squared:	0.890
Method:	Least Squares	F-statistic:	863.1
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	0.00
Time:	19:02:48	Log-Likelihood:	-9506.3
No. Observations:	964	AIC:	1.903e+04
Df Residuals:	954	BIC:	1.908e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0%	Conf. Int.]
const	-140.0569	199.910	-0.701	0.484	-532.370	252.257
x1	17.3253	5.427	3.192	0.001	6.675	27.976
x2	1.6410	0.100	16.356	0.000	1.444	1.838
x3	-12.3165	19.397	-0.635	0.526	-50.383	25.750
x4	-0.0001	3.46e-05	-4.178	0.000	-0.000	-7.66e-05
x5	-2.1247	0.187	-11.384	0.000	-2.491	-1.758
x6	-3.9607	1.102	-3.594	0.000	-6.123	-1.798
x7	5.9035	0.955	6.184	0.000	4.030	7.777
x8	-1.3614	0.543	-2.505	0.012	-2.428	-0.295
x9	-1.558e-10	1.35e-11	-11.501	0.000	-1.82e-10	-1.29e-10

Omnibus:	1823.521	Durbin-Watson:	2.067
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4710972.664
Skew:	13.084	Prob(JB):	0.00
Kurtosis:	344.469	Cond. No.	1.28e+14

The top-3 are: followers, time, impressions; And here are scatter plots:



2.1.4

Q1. We ignore the feature “the number of tweets in the last hour” since we have found that it doesn’t show good properties when doing the prediction of the number of tweets in the next hour. The results of average |NPredicted-NTrue| using 10 cross validation are shown in the table below:

Before Feb. 1, 8:00 a.m:

SB49	150.5588870716524
GoHawks	282.5089304764196
NFL	210.14342695041668
SuperBowl	500.87892420934674
Patriots	309.9998655781271
GoPatriots	21.253749166950666

Average Error	245

Between Feb. 1 8:00 a.m and Feb.1 9 p.m.:

SB49	74829.18785219172
GoHawks	4897.045409810998
NFL	4705.738936971248
SuperBowl	217697.46646345916
Patriots	21968.50682678939
GoPatriots	2136.5506403565964
Average Error	54372

After Feb. 1. 9 p.m.:

SB49	449.473083640796
GoHawks	63.47644405764414
NFL	261.22056644489044
SuperBowl	654.7042266053791
Patriots	205.32316167691945
GoPatriots	19.56016215061979
Average Error	275

We can see from the tables that the errors between 8:00 am and 9:00 pm are extraordinarily high. Since we only have 13 data points to train the model, and the number of tweets during this period is very high, our result is reasonable.

Q2. We combined the data from all tags together. We tried three different methods to predict the number of tweets in the next hour: Linear Regression, KNN and Random forest.

Linear Regression:

279.0621562710869
34627.90262937886
388.199220305292

Random forest:
297.8232232741618
28394.08926388889
359.83636493370955

KNN:
316.5681262327416
30281.847222222223
369.557225433526

Compared with the average errors of models trained using different tags, we can see that the errors using the combined model before 8am and after 9pm are higher. It shows that the number of tweets with different tags will have different trends.

On the other hand, the prediction error between 8am and 9pm is lower using the combined model. The combined model from all tags will include more data points and the model is less likely to be over-fitted. Therefore, the combined model between 8am and 9pm has better performance.

2.1.5

The results of 1.4 shows that the Random Forest Regression will have the best performance. Therefore, we choose the random forest as our model.

The features here are: (1) number of re-tweets; (2) number of followers; (3) largest number of followers; (4) citation date.

We trained the model using the twitter items from the last 6 hours except the last one hour and did the prediction of number of tweets in the last one hour for each time period.

Text file	**Predicted number**	$\|N_{Predicted}-N_{True}\|$
sample1_period1	112.519	64.481
sample2_period2	1478.161	18318.839
sample3_period3	256.378	361.622
sample4_period1	992.762	574.762
sample5_period1	237.811	103.189

sample6_period2	353.48	207.48
sample7_period3	235.23	133.237
sample8_period1	66.38	55.38
sample9_period2	338.257	1516.743
sample10_period3	66.33	5.33

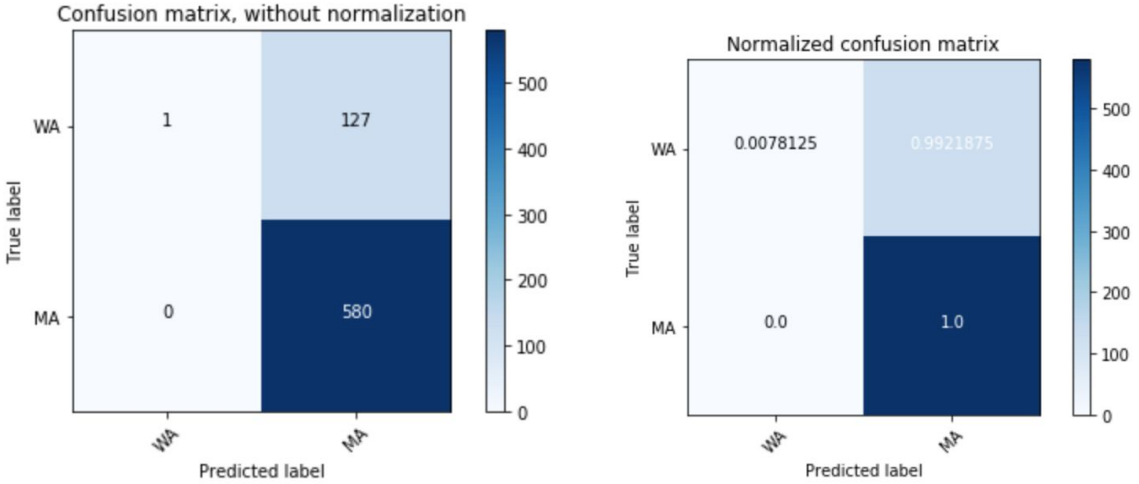
2.2 Fan Base Prediction

In this section, we trained three kinds of classification algorithms to predict the location of the author of a tweet. In order to make the problem more specific, let us consider all the tweets posted by the users whose specified location is either in the state of Washington or Massachusetts. As a preprocessing step, we converted the tweet's text into numerical features by applying a TF-IDF based vectorizer.

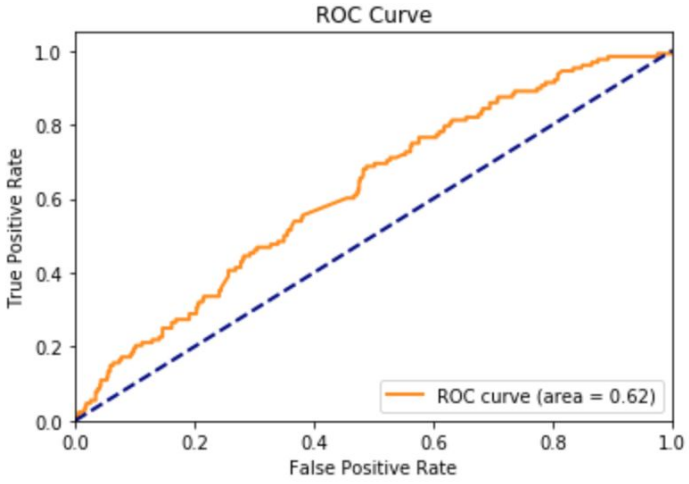
Here are the location classes:

WA: 'Seattle, Washington', 'Washington', 'WA', 'Seattle, WA', 'Kirkland, Washington'
MA: 'massachusetts', 'ma','mass','boston, massachusetts','boston, ma'

1. **Basic logistic regression model**

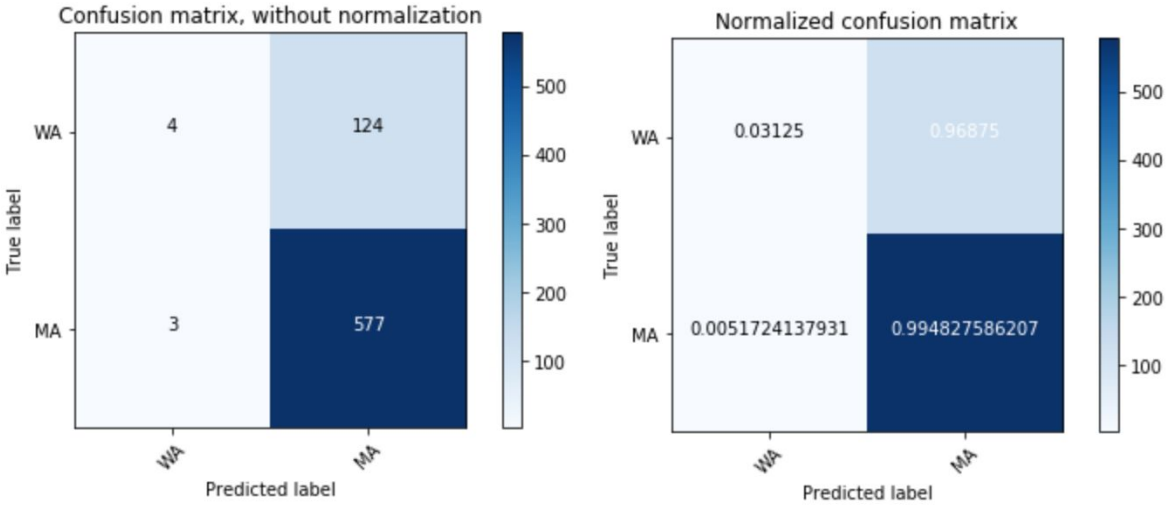



```
Confusion matrix, without normalization
[[  1 127]
 [  0 580]]
Normalized confusion matrix
[[ 0.01  0.99]
 [ 0.    1.  ]]
```

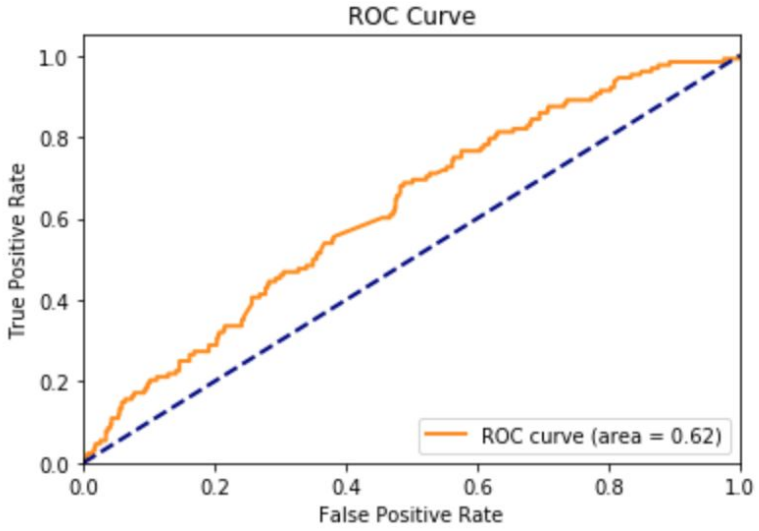



```
Classification report:
============================================================
Accuracy = 0.821, Precision = 0.820, Recall = 1.000
============================================================
```

2. SVM classifier

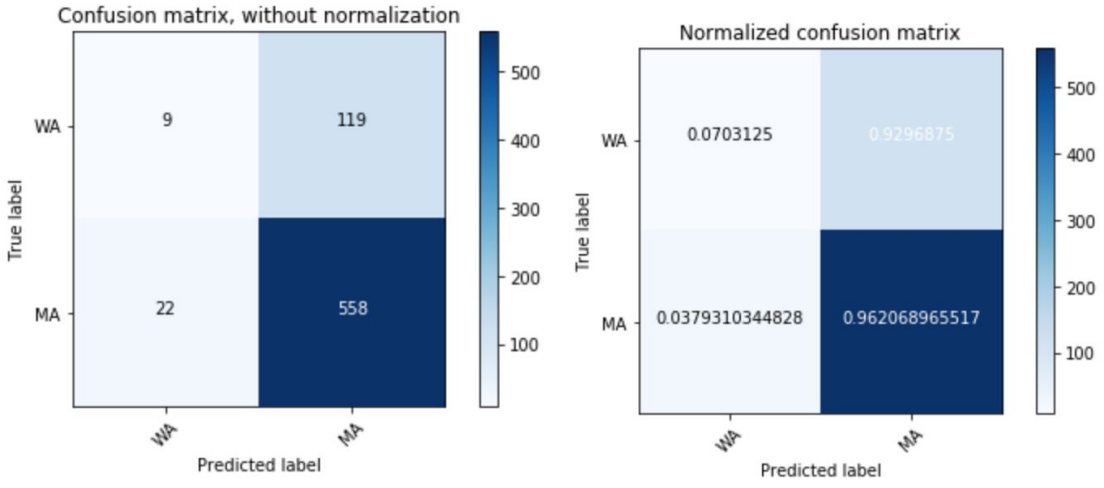



```
Confusion matrix, without normalization
[[  4 124]
 [  3 577]]
Normalized confusion matrix
[[ 0.03  0.97]
 [ 0.01  0.99]]
```

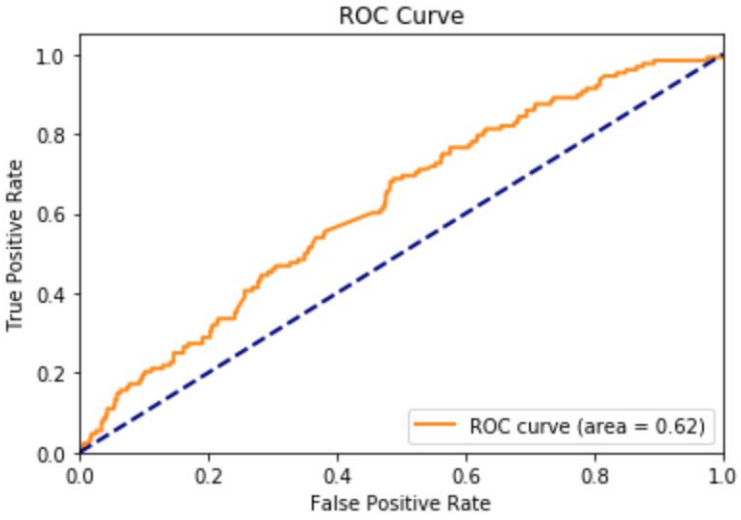



```
Classification report:
============================================================
Accuracy = 0.821, Precision = 0.823, Recall = 0.995
============================================================
```

3. KNN Classifier




```
Confusion matrix, without normalization
[[  9 119]
 [ 22 558]]
Normalized confusion matrix
[[ 0.07  0.93]
 [ 0.04  0.96]]
```




```
Classification report:
=======================================================
Accuracy = 0.801, Precision = 0.824, Recall = 0.962
=======================================================
```

2.3 Game Result Prediction with Semantic Mining and Analyses

We extract the tweets from Feb.1st 8:00 a.m. to Feb.1st 8:00 p.m. And split them into 12 groups according to their citation_date.

Then we extracted the content of the tweet, the citation date of the tweet and the location of the user. We applied the tf-idf matrix to format the content of each tweet.

The steps of our analysis:
(1) Did the NMF transformation for the tf-idf matrix in each time period. We set the number of potential factors to be 2
(2) For each line of the U matrix after NMF transformation, which represents an item of twitter, we checked the location of the user of this tweet and the type for this tweet. Then we generated a matrix(confusion matrix for location)
(3) For each column of the V matrix after NMF transformation, which represents an specific word, we sorted it according to the value.

We are able to have the insight into the trend during the hour according to the top words.

Our Results:
(For confusion matrix, the col1 means tweets come from MA, the col2 means tweets come from WA)

2015-02-01 08:00:00-08:00
68 22
76 136
superbowl http sunday happy ready nfl today game football day
sb49 seahawks patriots super bowl gohawks nfl superbowlxlix today http

2015-02-01 09:00:00-08:00
59 295
143 73
sb49 seahawks patriots gohawks game http day superbowlxlix today win
superbowl http sunday nfl super bowl happy today ready football

2015-02-01 10:00:00-08:00
80 637
505 434
seahawkswin seahawks winning ve got nfl sb49 http superbowlxlix superbowl
patriotswin patriots winning ve got nfl sb49 http superbowl superbowlxlix

2015-02-01 11:00:00-08:00
1462 514
111 1038
patriotswin patriots ve winning got nfl sb49 http superbowlxlix superbowl
seahawkswin seahawks ve winning got nfl sb49 http gohawks superbowl

2015-02-01 12:00:00-08:00
1315 542
80 1146
patriotswin patriots ve winning got nfl sb49 http superbowlxlix superbowl
seahawkswin seahawks ve winning got nfl sb49 http superbowlxlix superbowl

2015-02-01 13:00:00-08:00
79 907
911 591
seahawkswin seahawks winning ve got nfl sb49 http superbowlxlix superbowl
patriotswin patriots winning ve got nfl sb49 http superbowlxlix superbowl

2015-02-01 14:00:00-08:00
123 920
895 794
seahawkswin seahawks winning ve got nfl sb49 http superbowlxlix superbowl
patriotswin patriots winning ve got nfl sb49 http superbowlxlix superbowl

2015-02-01 15:00:00-08:00
1046 105
1647 3318
patriotswin patriots winning ve got nfl http sb49 superbowl superbowlxlix
seahawkswin seahawks winning ve got nfl http sb49 superbowl superbowlxlix

2015-02-01 16:00:00-08:00
253 181
3084 2976
winning ve got nfl http patriotswin sb49 patriots seahawkswin seahawks
touchdown superbowlxlix superbowl patriots seahawks sb49 http gohawks gronk commercial

2015-02-01 17:00:00-08:00
698 1250
2219 2746
winning got ve nfl http sb49 patriots patriotswin seahawks seahawkswin
superbowl superbowlxlix halftime katyperry katy perry http missy sb49 halftimeshow

2015-02-01 18:00:00-08:00
37 37
2227 2530
topspot2015 voted likeagirl year http sb49 dodgewisdom jeepplays dodge katyperry
superbowl patriots superbowlxlix http sb49 seahawks touchdown game nfl catch

2015-02-01 19:00:00-08:00
2187 336
1071 1019
patriots superbowlxlix win congrats champions sb49 game http super bowl
superbowl http patriotswin game sb49 seahawks nfl play wow fight

Conclusions:

From the words after 7pm, people were talking about Patriots and champions. We can see that Patriots has won the game.

Around 5pm, one key word with great weight in the second type is Katy Perry. It seems that Katy Perry gave a show at that time.

From 10a.m. to 3p.m., the top words in the V matrix reveals that the most significant potential factor should be the name of two teams. The people talking about super bowl at that time are mainly the supporters for each team.

One interesting fact: it seems that Patriots has more supporters. Even many people in WA support Patriots. However, few people in MA support SeaHawks.