EE 219

Large-Scale Data Mining

Project 5

Popularity Prediction on Twitter

Winter 2018

By Xudong Li (804944940),

Tao Wu (504946672),

Yangyang Mao (504945234),

Di Jin (305026178)

March 19, 2018

PART 1: Popularity Prediction

Problem 1.1

Statistics:

```
Statistics For gohawks
Average numver of tweets per hour = 325.37
Average number of followers of users posting the tweets = 2203.93
Average number of retweets = 2.01
```

```
Statistics For gopatriots
Average numver of tweets per hour = 45.69
Average number of followers of users posting the tweets = 1401.90
Average number of retweets = 1.40
```

```
Statistics For nfl
Average numver of tweets per hour = 441.32
Average number of followers of users posting the tweets = 4653.25
Average number of retweets = 1.54
```

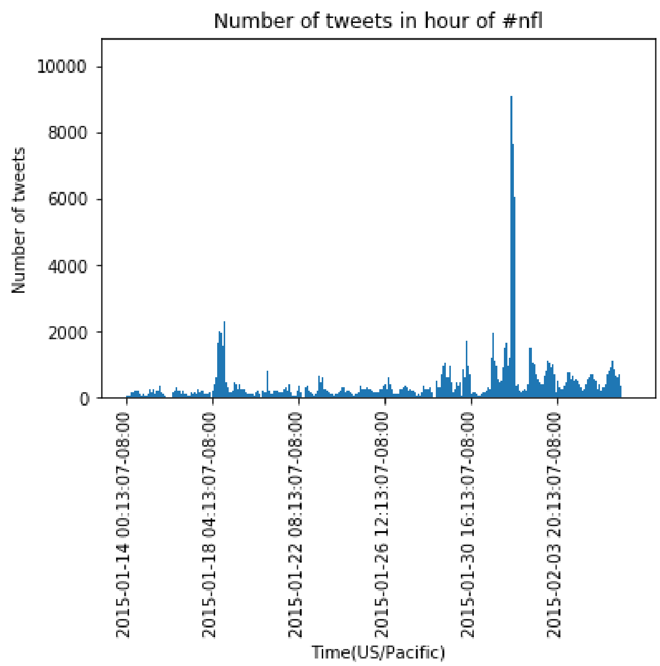
```
Statistics For patriots
Average numver of tweets per hour = 834.56
Average number of followers of users posting the tweets = 3309.98
Average number of retweets = 1.78
```

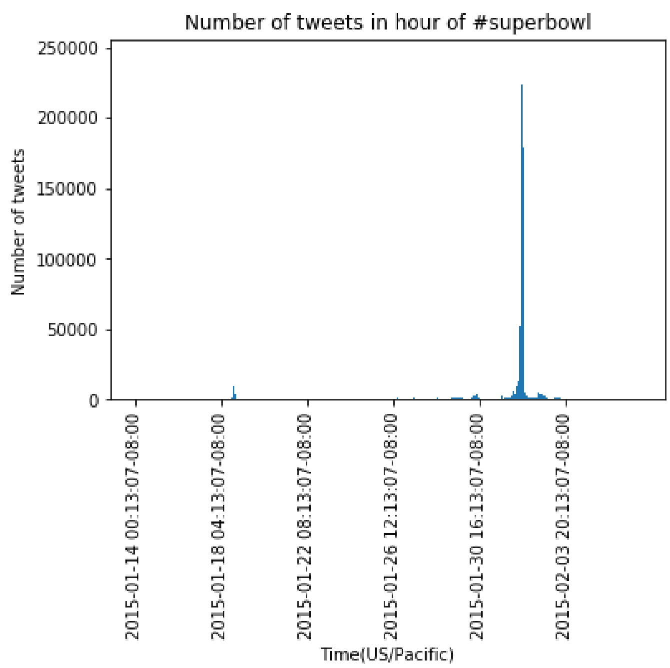
```
Statistics For sb49
Average numver of tweets per hour = 1419.89
Average number of followers of users posting the tweets = 10267.32
Average number of retweets = 2.51
```

```
Statistics For superbowl
Average numver of tweets per hour = 2302.50
Average number of followers of users posting the tweets = 8858.97
Average number of retweets = 2.39
```

Analysis

Above statistics show average number of tweets per hour, average number of followers of users posting the tweets and average number of retweets for each hashtag. We can see that super bowl is the most popular topics among this hashtags for it has the largest number of tweets per hour. And as for different teams, #gohawks has more tweets than #gopartriots. But for each hashtag, average numbers of retweets are very close.





Analysis

From above figures we can see that both #nfl and #superbowl experienced a peak in 2015-02-02 for super bowl held on that day and the tweets number for #superbowl even reached around 230000. And both hashtags had a small peak at 2015-01-18.

Problem 1.2

Model Analysis for #gohawaks

```
Model Analysis for gohawks
RMSE = 949.1656
R2_score = 0.4919
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.519
Model:                            OLS   Adj. R-squared:                  0.515
Method:                 Least Squares   F-statistic:                     123.9
Date:                Sat, 10 Mar 2018   Prob (F-statistic):           8.37e-89
Time:                        17:08:17   Log-Likelihood:                -4791.8
No. Observations:                 579   AIC:                             9594.
Df Residuals:                     574   BIC:                             9615.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.3846      0.165      8.378      0.000       1.060       1.709
x2            -0.1454      0.039     -3.749      0.000      -0.222      -0.069
x3            -0.0002   8.36e-05     -2.966      0.003      -0.000   -8.38e-05
x4             0.0003      0.000      1.514      0.130   -7.63e-05       0.001
x5             6.8062      3.261      2.087      0.037       0.400      13.212
==============================================================================
Omnibus:                      892.712   Durbin-Watson:                   2.223
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           831519.328
Skew:                           8.142   Prob(JB):                         0.00
Kurtosis:                     187.938   Cond. No.                     2.39e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.39e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Analysis for #gopatriots

```
Model Analysis for gopatriots
RMSE = 194.1643
R2_score = 0.6026
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.611
Model:                            OLS   Adj. R-squared:                  0.607
Method:                 Least Squares   F-statistic:                     178.8
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          3.05e-114
Time:                        17:08:28   Log-Likelihood:                -3845.8
No. Observations:                 575   AIC:                             7702.
Df Residuals:                     570   BIC:                             7723.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -0.4254      0.264     -1.614      0.107      -0.943       0.092
x2             0.4680      0.229      2.041      0.042       0.018       0.918
x3             0.0006      0.000      3.144      0.002       0.000       0.001
x4            -0.0007      0.000     -3.719      0.000      -0.001      -0.000
x5             0.7084      0.629      1.126      0.261      -0.528       1.944
==============================================================================
Omnibus:                      450.828   Durbin-Watson:                   2.086
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           346384.225
Skew:                           2.109   Prob(JB):                         0.00
Kurtosis:                     123.167   Cond. No.                     3.26e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.26e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Analysis for #nfl

```
Model Analysis for nfl
RMSE = 581.6922
R2_score = 0.5632
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.646
Model:                            OLS   Adj. R-squared:                  0.643
Method:                 Least Squares   F-statistic:                     212.8
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          7.71e-129
Time:                        17:10:09   Log-Likelihood:                -4573.4
No. Observations:                 587   AIC:                             9157.
Df Residuals:                     582   BIC:                             9179.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.7612      0.135      5.626      0.000       0.495       1.027
x2            -0.1736      0.066     -2.635      0.009      -0.303      -0.044
x3          7.177e-05   2.62e-05      2.741      0.006    2.03e-05       0.000
x4         -6.799e-05   3.59e-05     -1.894      0.059      -0.000    2.51e-06
x5             7.4472      2.201      3.383      0.001       3.124      11.771
==============================================================================
Omnibus:                      561.928   Durbin-Watson:                   2.328
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           352596.370
Skew:                           3.203   Prob(JB):                         0.00
Kurtosis:                     122.896   Cond. No.                     4.25e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.25e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Analysis for #patriots

```
Model Analysis for patriots
RMSE = 2368.8952
R2_score = 0.7064
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.716
Model:                            OLS   Adj. R-squared:                  0.714
Method:                 Least Squares   F-statistic:                     294.0
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          1.26e-156
Time:                        17:13:27   Log-Likelihood:                -5394.4
No. Observations:                 587   AIC:                         1.080e+04
Df Residuals:                     582   BIC:                         1.082e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.2159      0.079     15.395      0.000       1.061       1.371
x2            -0.3385      0.068     -4.945      0.000      -0.473      -0.204
x3          3.506e-05   2.62e-05      1.336      0.182   -1.65e-05    8.66e-05
x4             0.0002   9.48e-05      1.655      0.099   -2.93e-05       0.000
x5             7.7572      8.203      0.946      0.345      -8.354      23.868
==============================================================================
Omnibus:                     1019.164   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           973665.680
Skew:                          10.562   Prob(JB):                         0.00
Kurtosis:                     201.401   Cond. No.                     7.69e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.69e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Analysis for #sb49

```
Model Analysis for sb49
RMSE = 4006.2749
R2_score = 0.8405
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.844
Model:                            OLS   Adj. R-squared:                  0.842
Method:                 Least Squares   F-statistic:                     623.7
Date:                Sat, 10 Mar 2018   Prob (F-statistic):          3.65e-230
Time:                        17:19:21   Log-Likelihood:                -5663.6
No. Observations:                 583   AIC:                         1.134e+04
Df Residuals:                     578   BIC:                         1.136e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.2890      0.095     13.535      0.000       1.102       1.476
x2            -0.2955      0.087     -3.381      0.001      -0.467      -0.124
x3          2.873e-05   1.38e-05      2.077      0.038    1.56e-06    5.59e-05
x4             0.0002   4.24e-05      4.234      0.000    9.61e-05       0.000
x5           -16.1385     13.686     -1.179      0.239     -43.019      10.742
==============================================================================
Omnibus:                      959.740   Durbin-Watson:                   1.399
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           714899.977
Skew:                           9.508   Prob(JB):                         0.00
Kurtosis:                     173.494   Cond. No.                     7.06e+06
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.06e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model Analysis for #superbowl

```
Model Analysis for superbowl
RMSE = 6519.7941
R2_score = 0.8667
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.868
Method:                 Least Squares   F-statistic:                     769.9
Date:                Sat, 10 Mar 2018   Prob (F-statistic):         1.54e-253
Time:                        17:29:52   Log-Likelihood:                -5978.2
No. Observations:                 586   AIC:                         1.197e+04
Df Residuals:                     581   BIC:                         1.199e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             2.5465      0.107     23.765      0.000       2.336       2.757
x2            -0.1547      0.035     -4.387      0.000      -0.224      -0.085
x3            -0.0002   1.08e-05    -20.237      0.000      -0.000      -0.000
x4             0.0011      0.000     10.433      0.000       0.001       0.001
x5           -55.8291     24.146     -2.312      0.021    -103.254      -8.405
==============================================================================
Omnibus:                     1138.766   Durbin-Watson:                   1.845
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1944083.727
Skew:                          13.283   Prob(JB):                         0.00
Kurtosis:                     283.919   Cond. No.                     1.08e+07
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.08e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Analysis

In this part, we first created a dictionary to store one-hour data and extract the feature information, then used the linear regression model to predict the number of tweets in the next hour. From above results we can find that with data size increasing, it will have a higher RMSE but a higher R2_score. And among features we choose from the data, we can find that x1 and x4 always have a zero p-value which means those two are important in predicting the next hour's tweet number. Here x1 is total tweet number in current hour and x4 is maximum follower number of user posting the tweet. While in #superbowl, x1-x4 feature are all critical for predicting (x2 is total retweets in current hour and x3 is total followers).

Problem 1.3

In this part, we use 10 features: Number of Tweets, Number of Retweets, Number of Followers, Max Number of Followers, Total Number of Replies, Count of Impressions, Favorite Count, Ranking Score, user_id, Time of Day. We use linear regression model in this part. As a result, we report the RMSE and OLS regression results. We use p-values to select top 3 features and plot the figures of them.

#gopatriots

RMSE = 154.3784

R2_score = 0.7488

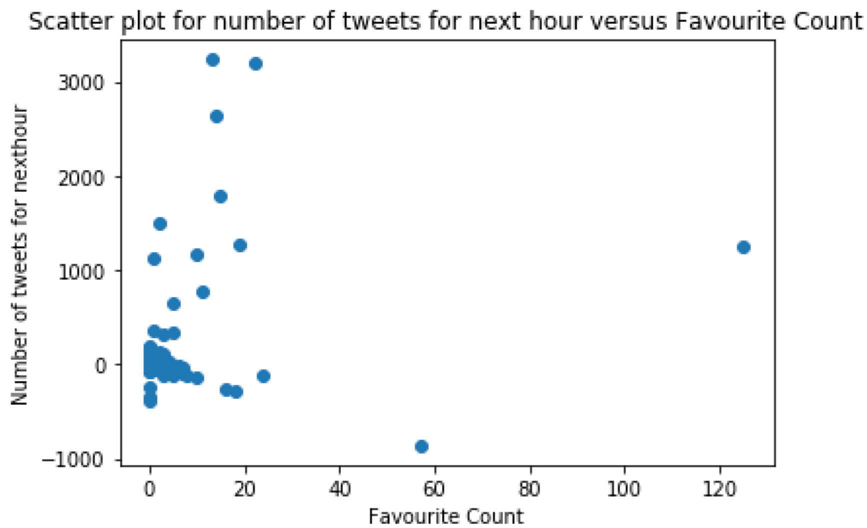
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.754
Model:                            OLS   Adj. R-squared:                  0.749
Method:                 Least Squares   F-statistic:                     172.9
Date:                Sun, 11 Mar 2018   Prob (F-statistic):         1.06e-164
Time:                        01:06:12   Log-Likelihood:                -3714.1
No. Observations:                 575   AIC:                             7448.
Df Residuals:                     565   BIC:                             7492.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             8.0801      2.456      3.290      0.001       3.257      12.903
x2             0.6225      0.202      3.079      0.002       0.225       1.020
x3             0.0031      0.000      7.946      0.000       0.002       0.004
x4            -0.0016      0.000     -7.934      0.000      -0.002      -0.001
x5           -19.1357      4.431     -4.319      0.000     -27.838     -10.433
x6            -0.0016      0.000     -5.581      0.000      -0.002      -0.001
x7           -17.5335      1.613    -10.868      0.000     -20.702     -14.365
x8            -1.3171      0.438     -3.009      0.003      -2.177      -0.457
x9         -3.987e-09   6.67e-10     -5.973      0.000    -5.3e-09   -2.68e-09
x10            1.0573      0.508      2.083      0.038       0.060       2.054
==============================================================================
Omnibus:                      453.089   Durbin-Watson:                   2.223
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            78528.802
Skew:                           2.550   Prob(JB):                         0.00
Kurtosis:                      60.024   Cond. No.                     1.61e+11
==============================================================================
```

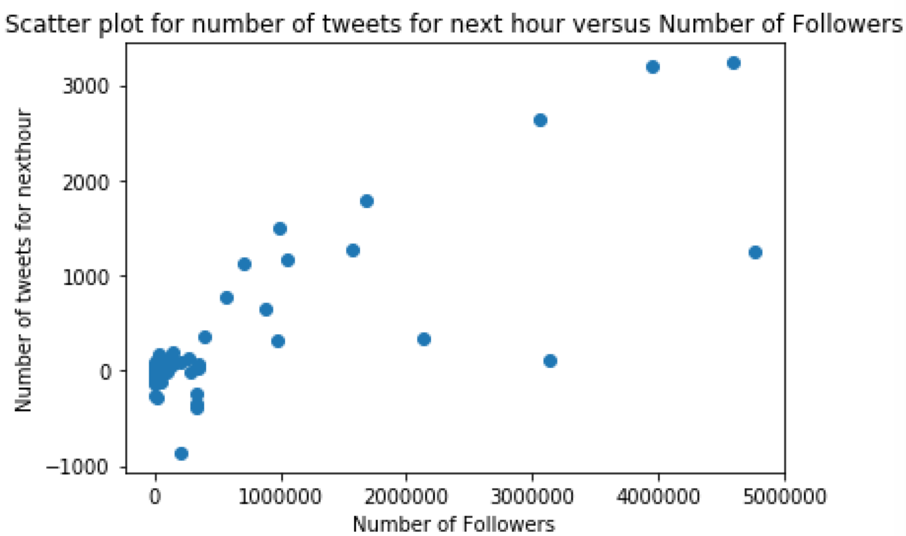
Top 3 features:

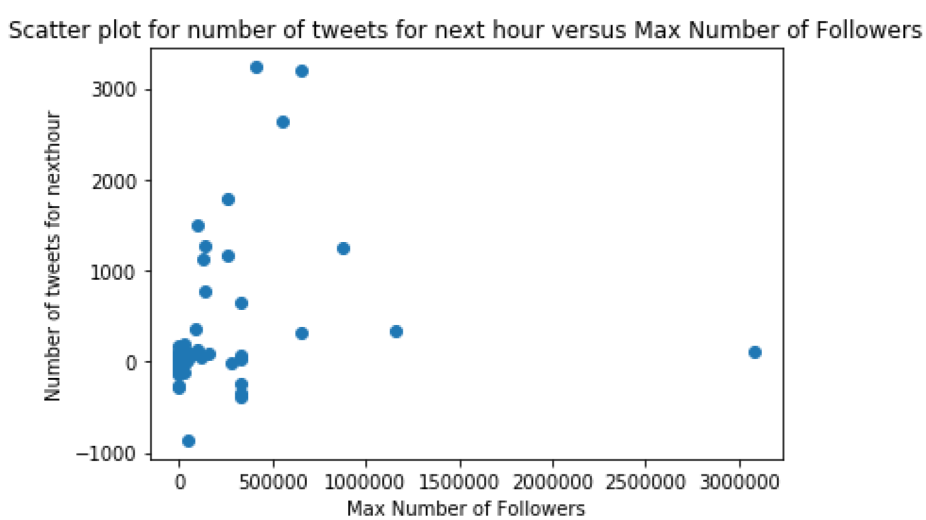
1.Favorite Count

2.Number of Followers

3.Max Number of Followers


Scatter plot for number of tweets for next hour versus Favourite Count
Number of tweets for nexthour
Favourite Count
3000
2000
1000
0
−1000
0
20
40
60
80
100
120


Scatter plot for number of tweets for next hour versus Number of Followers
Number of tweets for nexthour
Number of Followers
3000
2000
1000
0
−1000
0
1000000
2000000
3000000
4000000
5000000


Scatter plot for number of tweets for next hour versus Max Number of Followers
Number of tweets for nexthour
Max Number of Followers
3000
2000
1000
0
−1000
0
500000
1000000
1500000
2000000
2500000
3000000

#gohawks

RMSE = 888.5901

R2_score = 0.5547

OLS Regression Results

Dep. Variable:	y	R-squared:	0.579
Model:	OLS	Adj. R-squared:	0.571
Method:	Least Squares	F-statistic:	78.10
Date:	Sun, 11 Mar 2018	Prob (F-statistic):	5.73e-100
Time:	01:11:29	Log-Likelihood:	-4753.6
No. Observations:	579	AIC:	9527.
Df Residuals:	569	BIC:	9571.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-7.6243	4.172	-1.828	0.068	-15.818	0.569
x2	-0.7421	0.095	-7.822	0.000	-0.928	-0.556
x3	-0.0006	0.000	-5.229	0.000	-0.001	-0.000
x4	0.0005	0.000	2.963	0.003	0.000	0.001
x5	23.0600	8.993	2.564	0.011	5.397	40.723
x6	0.0002	8.14e-05	2.653	0.008	5.61e-05	0.000
x7	0.3267	0.048	6.834	0.000	0.233	0.421
x8	1.7121	0.772	2.218	0.027	0.196	3.228
x9	3.121e-09	1.05e-09	2.969	0.003	1.06e-09	5.19e-09
x10	4.8812	3.270	1.493	0.136	-1.541	11.304

Omnibus:	970.228	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	830069.370
Skew:	9.800	Prob(JB):	0.00
Kurtosis:	187.453	Cond. No.	2.34e+11

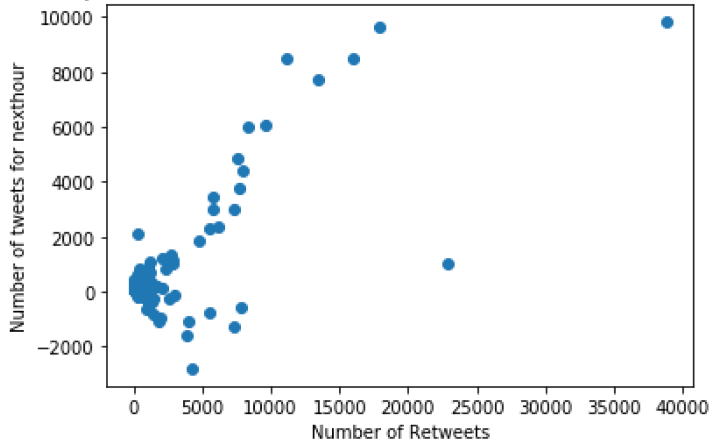
Top 3 features:

1.Number of Retweets

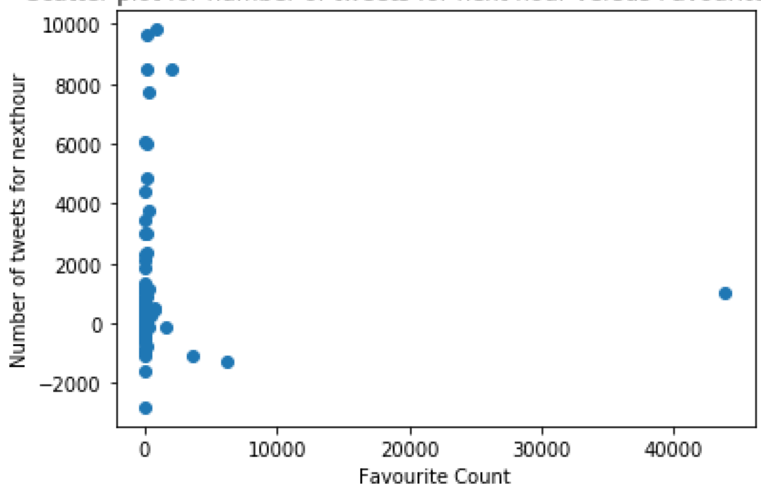
2.Favorite Count

3.Number of Followers

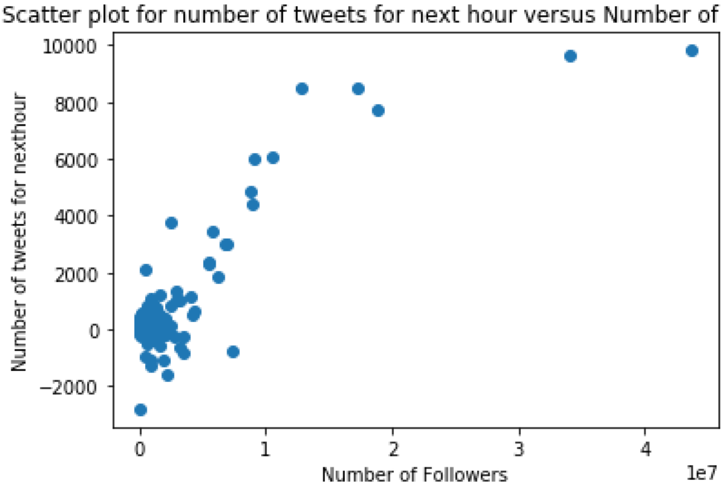
Scatter plot for number of tweets for next hour versus Number of Retweets



Scatter plot for number of tweets for next hour versus Favourite Count



Scatter plot for number of tweets for next hour versus Number of Followers



#nfl

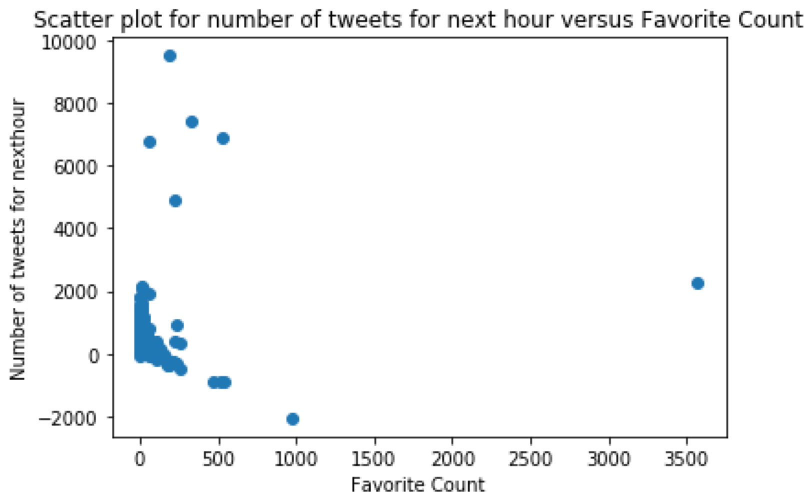
RMSE = 487.9852

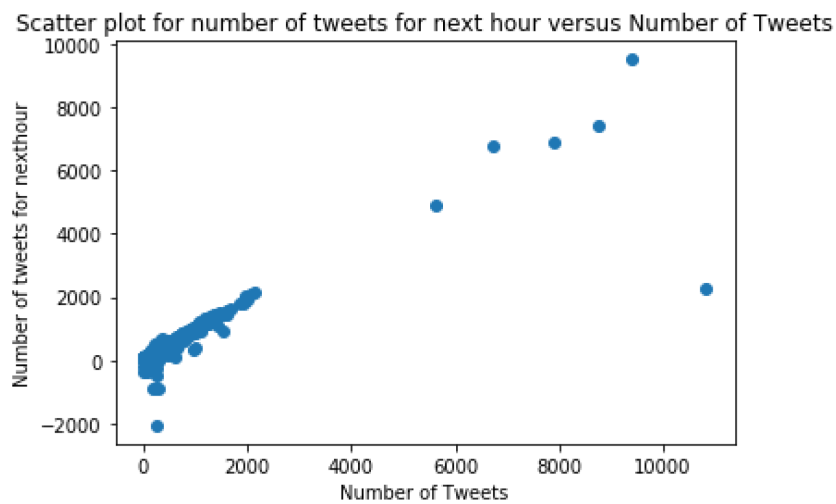
R2_score = 0.6926

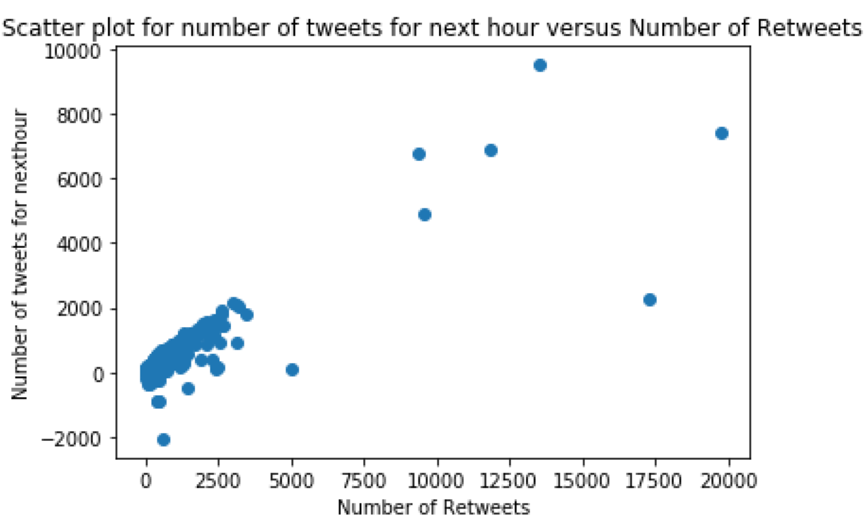
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.753
Model:                            OLS   Adj. R-squared:                  0.748
Method:                 Least Squares   F-statistic:                     175.6
Date:                Sun, 11 Mar 2018   Prob (F-statistic):          9.07e-168
Time:                        01:26:55   Log-Likelihood:                -4468.6
No. Observations:                 587   AIC:                             8957.
Df Residuals:                     577   BIC:                             9001.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             4.6241      1.792      2.581      0.010       1.105       8.143
x2            -0.1247      0.057     -2.169      0.031      -0.238      -0.012
x3         -1.019e-05   3.71e-05     -0.275      0.784    -8.3e-05    6.26e-05
x4          4.931e-05   3.25e-05      1.519      0.129   -1.45e-05       0.000
x5            -2.0996      3.462     -0.606      0.544      -8.899       4.700
x6          -6.66e-07   2.76e-05     -0.024      0.981   -5.49e-05    5.36e-05
x7            -2.4488      0.166    -14.779      0.000      -2.774      -2.123
x8            -0.7166      0.374     -1.914      0.056      -1.452       0.019
x9         -2.204e-10   2.03e-10     -1.085      0.278   -6.19e-10    1.78e-10
x10            2.2019      2.253      0.978      0.329      -2.222       6.626
==============================================================================
Omnibus:                      850.519   Durbin-Watson:                   2.411
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           272754.915
Skew:                           7.647   Prob(JB):                         0.00
Kurtosis:                     107.489   Cond. No.                     1.26e+11
==============================================================================
```

Top 3 features :

1. Favorite Count
2. Number of Tweets
3. Number of Reweets


Scatter plot for number of tweets for next hour versus Favorite Count
Number of tweets for nexthour
Favorite Count
10000
8000
6000
4000
2000
0
−2000
0
500
1000
1500
2000
2500
3000
3500


Scatter plot for number of tweets for next hour versus Number of Tweets
Number of tweets for nexthour
Number of Tweets
10000
8000
6000
4000
2000
0
−2000
0
2000
4000
6000
8000
10000


Scatter plot for number of tweets for next hour versus Number of Retweets
Number of tweets for nexthour
Number of Retweets
10000
8000
6000
4000
2000
0
−2000
0
2500
5000
7500
10000
12500
15000
17500
20000

#patriots

RMSE = 2294.9456

R2_score = 0.7245

OLS Regression Results

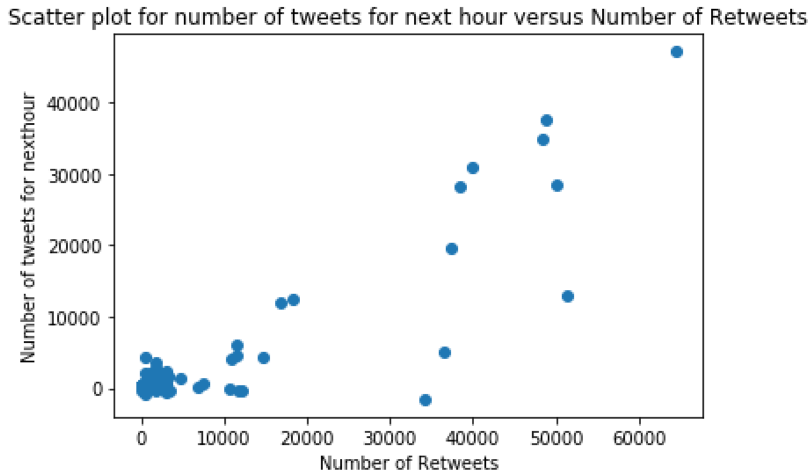
Dep. Variable:	y	R-squared:	0.724
Model:	OLS	Adj. R-squared:	0.720
Method:	Least Squares	F-statistic:	168.3
Date:	Sun, 11 Mar 2018	Prob (F-statistic):	4.78e-155
Time:	01:49:37	Log-Likelihood:	-5375.7
No. Observations:	587	AIC:	1.077e+04
Df Residuals:	577	BIC:	1.082e+04
Df Model:	9		
Covariance Type:	nonrobust		

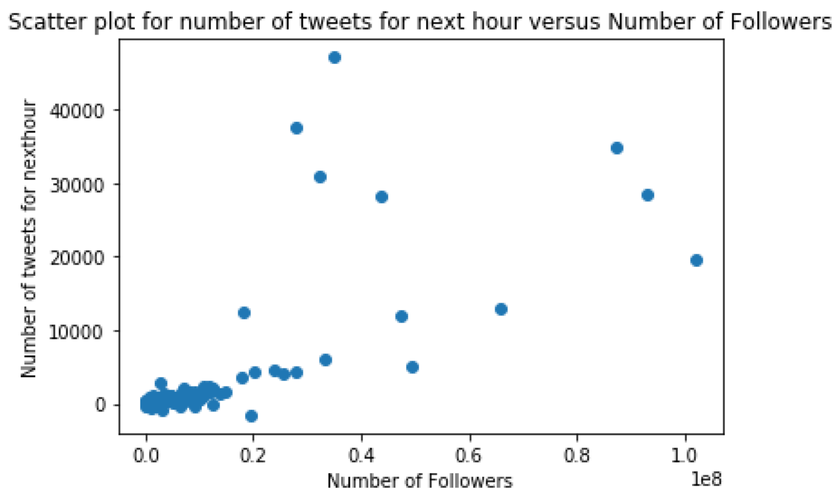
	coef	std err	t	P>|t|	[0.025	0.975]
x1	5.3304	5.115	1.042	0.298	-4.716	15.377
x2	-0.2119	0.098	-2.167	0.031	-0.404	-0.020
x3	0.0004	0.000	2.110	0.035	3.04e-05	0.001
x4	-0.0003	0.000	-1.842	0.066	-0.001	1.67e-05
x5	8.1891	6.112	1.340	0.181	-3.816	20.194
x6	-0.0001	0.000	-0.747	0.455	-0.001	0.000
x7	-0.0511	0.245	-0.208	0.835	-0.533	0.431
x8	-1.2512	1.052	-1.189	0.235	-3.317	0.815
x9	6.104e-10	7.44e-10	0.820	0.413	-8.52e-10	2.07e-09
x10	7.1349	8.212	0.869	0.385	-8.993	23.263

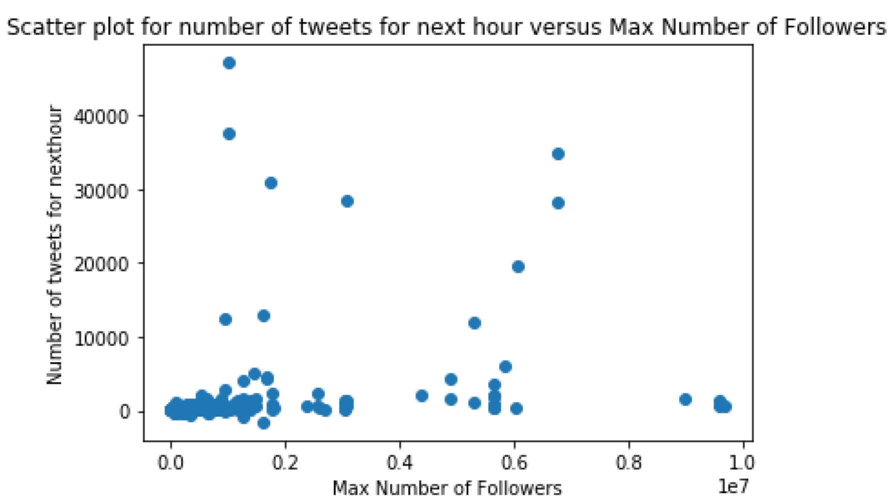
Omnibus:	1054.228	Durbin-Watson:	1.848
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1185261.532
Skew:	11.302	Prob(JB):	0.00
Kurtosis:	221.974	Cond. No.	4.11e+11

Top 3 features:

1. Number of Retweets
2. Number of Followers
3. Max Number of Followers


Scatter plot for number of tweets for next hour versus Number of Retweets
Number of tweets for nexthour
Number of Retweets
0
10000
20000
30000
40000
50000
60000


Scatter plot for number of tweets for next hour versus Number of Followers
Number of tweets for nexthour
Number of Followers
0
10000
20000
30000
40000
0.0
0.2
0.4
0.6
0.8
1.0
1e8


Scatter plot for number of tweets for next hour versus Max Number of Followers
Number of tweets for nexthour
Max Number of Followers
0
10000
20000
30000
40000
0.0
0.2
0.4
0.6
0.8
1.0
1e7

#sb49

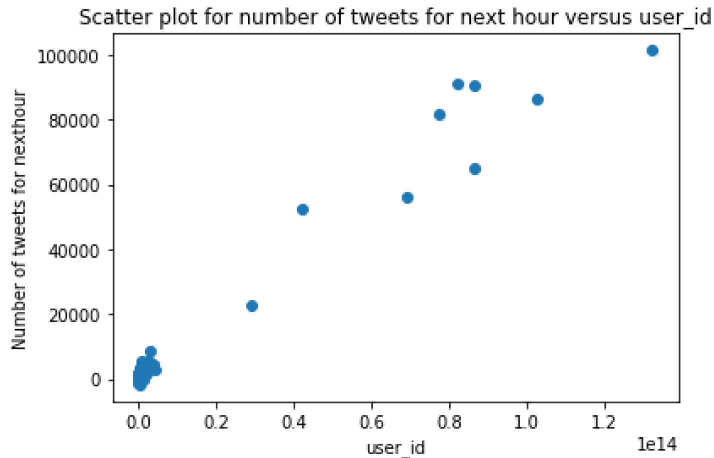
RMSE = 3681.1277

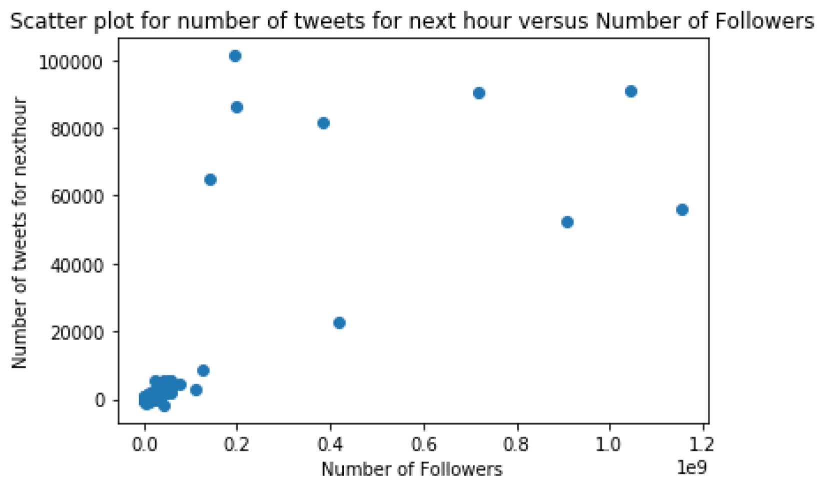
R2_score = 0.8654

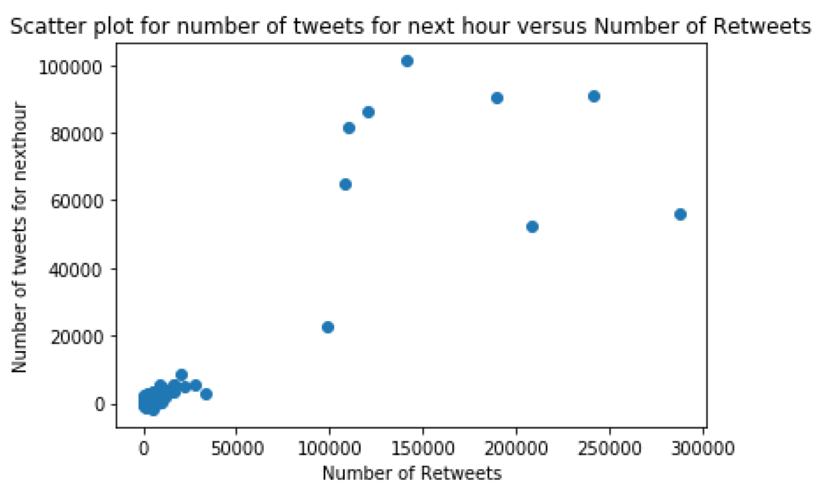
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.865
Model:                            OLS   Adj. R-squared:                  0.863
Method:                 Least Squares   F-statistic:                     408.9
Date:                Sun, 11 Mar 2018   Prob (F-statistic):         8.30e-243
Time:                        03:44:48   Log-Likelihood:                -5614.5
No. Observations:                 583   AIC:                         1.125e+04
Df Residuals:                     573   BIC:                         1.129e+04
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -5.5733      5.761     -0.967      0.334     -16.888       5.742
x2             0.2016      0.100      2.018      0.044       0.005       0.398
x3             0.0001   6.65e-05      2.192      0.029    1.51e-05       0.000
x4          6.723e-05   6.03e-05      1.115      0.265   -5.12e-05       0.000
x5            -9.3727      6.684     -1.402      0.161     -22.500       3.755
x6         -8.887e-05   6.55e-05     -1.357      0.175      -0.000    3.98e-05
x7            -0.1545      0.083     -1.856      0.064      -0.318       0.009
x8             0.6022      1.151      0.523      0.601      -1.659       2.863
x9          3.047e-09   8.74e-10      3.486      0.001    1.33e-09    4.76e-09
x10           -8.2009     12.864     -0.638      0.524     -33.467      17.065
==============================================================================
Omnibus:                     1084.733   Durbin-Watson:                   1.298
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1141554.279
Skew:                          12.273   Prob(JB):                         0.00
Kurtosis:                     218.386   Cond. No.                     8.75e+11
==============================================================================
```

Top 3 features:

1. user_id
2. Number of Followers
3. Number of Retweets


Scatter plot for number of tweets for next hour versus user_id
Number of tweets for nexthour
user_id
1e14


Scatter plot for number of tweets for next hour versus Number of Followers
Number of tweets for nexthour
Number of Followers
1e9


Scatter plot for number of tweets for next hour versus Number of Retweets
Number of tweets for nexthour
Number of Retweets

#superbowl

RMSE = 6252.8640

R2_score = 0.8774

OLS Regression Results

Dep. Variable:	y	R-squared:	0.877
Model:	OLS	Adj. R-squared:	0.875
Method:	Least Squares	F-statistic:	457.1
Date:	Sun, 11 Mar 2018	Prob (F-statistic):	1.13e-255
Time:	16:53:45	Log-Likelihood:	-5954.2
No. Observations:	586	AIC:	1.193e+04
Df Residuals:	576	BIC:	1.197e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	14.7687	5.004	2.951	0.003	4.940	24.597
x2	-0.0487	0.044	-1.112	0.267	-0.135	0.037
x3	-3.24e-05	0.000	-0.150	0.881	-0.000	0.000
x4	0.0013	0.000	11.915	0.000	0.001	0.002
x5	-67.0767	18.051	-3.716	0.000	-102.530	-31.623
x6	-0.0001	0.000	-0.612	0.541	-0.001	0.000
x7	-1.4970	0.261	-5.742	0.000	-2.009	-0.985
x8	-2.6149	1.056	-2.476	0.014	-4.689	-0.541
x9	-9.529e-10	4.87e-10	-1.956	0.051	-1.91e-09	4.17e-12
x10	-60.6348	23.651	-2.564	0.011	-107.088	-14.181

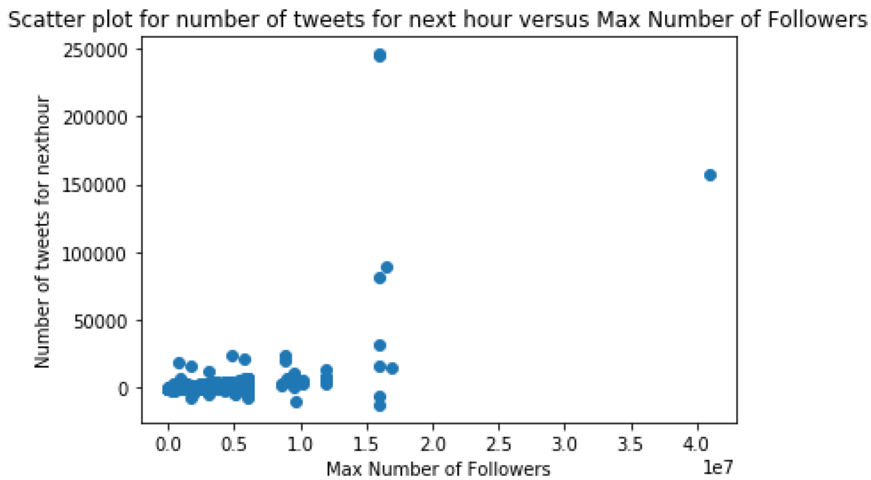
Omnibus:	1096.686	Durbin-Watson:	1.897
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1443463.269
Skew:	12.318	Prob(JB):	0.00
Kurtosis:	244.891	Cond. No.	1.10e+12

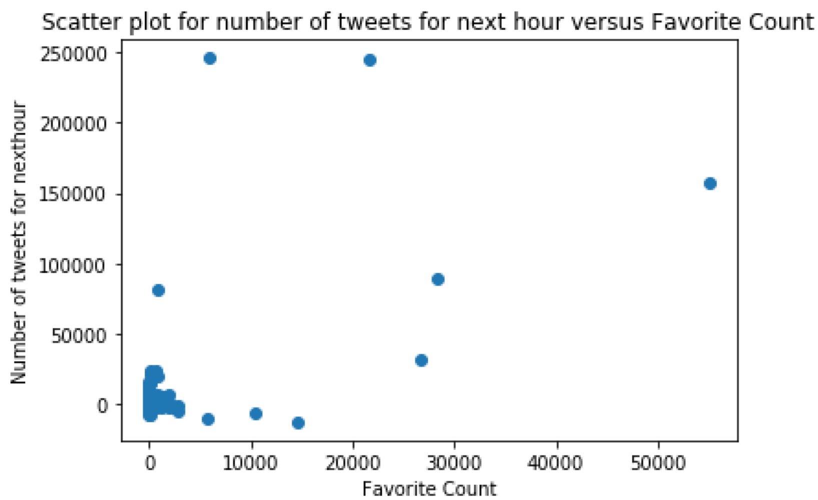
Top 3 features:

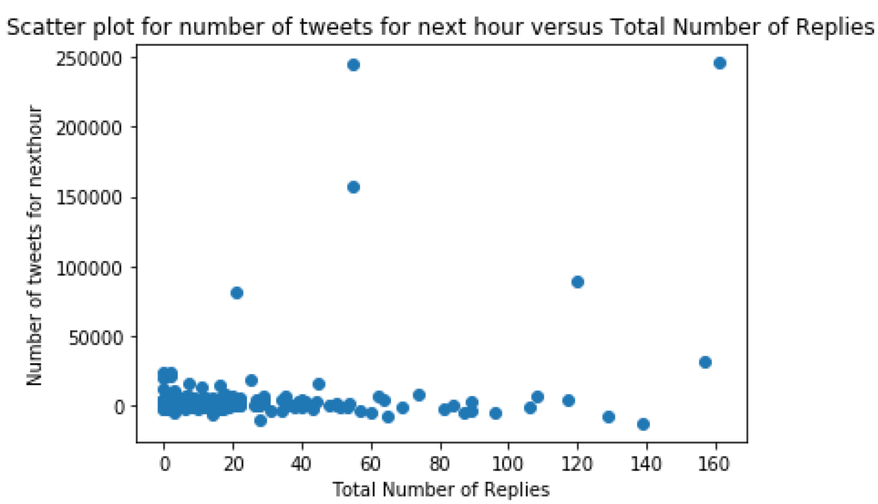
1.Max Number of Followers

2.Favorite Count

3.Total Number of Replies


Scatter plot for number of tweets for next hour versus Max Number of Followers
Number of tweets for nexthour
250000
200000
150000
100000
50000
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e7
Max Number of Followers


Scatter plot for number of tweets for next hour versus Favorite Count
Number of tweets for nexthour
250000
200000
150000
100000
50000
0
0
10000
20000
30000
40000
50000
Favorite Count


Scatter plot for number of tweets for next hour versus Total Number of Replies
Number of tweets for nexthour
250000
200000
150000
100000
50000
0
0
20
40
60
80
100
120
140
160
Total Number of Replies

Analysis

In problem 1.3, we reported RMSE value and OLS results. For each of the top 3 features in your measurements, draw a scatter plot of predictant (number of tweets for next hour) versus value of that feature. From results listed above, we can reach the concludes that:

1. RMSE as well as p-values is getting large with the increasing amount of data, with indicates that the fitting accuracy is decreasing with the increasing amount of data.
2. In some features and hashtags, we observe a relatively linear relationship between top features and target value. For example, in the hashtag of #gopatiots, and the feature of number of followers; in the hashtag of #gohawk and feature of number of retweet and number of followers; in the hashtag of #nfl and feature of number of tweet and number of retweet; and in the hashtag of #sb49 and all three top feature of user_id, Number of Followers and Number of Retweets. It demonstrated that we designed good features.

Problem 1.4

In this section, we are asked to train 3 types of regression models for 3-time intervals. Since we first want to predict for every hashtag, there are total 6*3*3 = 54 models. The accuracy of the model is evaluated by calculating average prediction error:

$$Average\ Prediction\ Error = |N_{predicted} - N_{real}|\ .$$

Results are shown below. The first three is predict by Linear Regression; the second three used K-Neighbors Regression; the third three used Random Forest Regression:

	# **gopatriots**	# **gohawks**
Linear Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [42.46813498]	Before Feb. 1, 8:00 a.m. Averaged error is: [304.57087378]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [5238.88288653]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [2535.50160568]
	After Feb. 1, 8:00 p.m. Averaged error is: [67.48551112]	After Feb. 1, 8:00 p.m. Averaged error is: [4320.08333116]
K-Neighbors Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [10.99378428]	Before Feb. 1, 8:00 a.m. Averaged error is: [125.43337335]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [957.87272727]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [2506.70909091]
	After Feb. 1, 8:00 p.m. Averaged error is: [4.06341463]	After Feb. 1, 8:00 p.m. Averaged error is: [27.86190476]
Random Forest Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [7.83074332]	Before Feb. 1, 8:00 a.m. Averaged error is: [75.88450472]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [722.17272727]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [2277.82727273]
	After Feb. 1, 8:00 p.m. Averaged error is: [3.15567751]	After Feb. 1, 8:00 p.m. Averaged error is: [24.44292328]

	# **nfl**	# **patriots**
Linear Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [83.0804976]	Before Feb. 1, 8:00 a.m. Averaged error is: [328.3411029]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [35792.05881429]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [29245.22188399]
	After Feb. 1, 8:00 p.m. Averaged error is: [125.46516149]	After Feb. 1, 8:00 p.m. Averaged error is: [843.37199021]
K-Neighbors Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [100.15840822]	Before Feb. 1, 8:00 a.m. Averaged error is: [139.21608643]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [1830.2]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [15287.69090909]
	After Feb. 1, 8:00 p.m. Averaged error is: [148.92985075]	After Feb. 1, 8:00 p.m. Averaged error is: [104.1641791]
Random Forest Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [76.92386792]	Before Feb. 1, 8:00 a.m. Averaged error is: [116.66830124]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [1473.80909091]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [18301.68181818]
	After Feb. 1, 8:00 p.m. Averaged error is: [140.20970149]	After Feb. 1, 8:00 p.m. Averaged error is: [111.85970149]

	# **superbowl**	# **sb49**
Linear Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [3399.33174011]	Before Feb. 1, 8:00 a.m. Averaged error is: [161.04890856]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [1287929.46952121]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [291186.16355755]
	After Feb. 1, 8:00 p.m. Averaged error is: [358.76516183]	After Feb. 1, 8:00 p.m. Averaged error is: [165.28865759]
K-Neighbors Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [218.76691176]	Before Feb. 1, 8:00 a.m. Averaged error is: [57.54712644]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [51279.50909091]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [39122.74545455]
	After Feb. 1, 8:00 p.m. Averaged error is: [268.94477612]	After Feb. 1, 8:00 p.m. Averaged error is: [121.0641791]
Random Forest Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [167.44461807]	Before Feb. 1, 8:00 a.m. Averaged error is: [53.74366411]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [42598.36363636]	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [39091.61818182]
	After Feb. 1, 8:00 p.m. Averaged error is: [284.12089552]	After Feb. 1, 8:00 p.m. Averaged error is: [135.15074627]

Analysis

From the result, we can see that the data is separated by three-time intervals: Before Feb. 1, 8:00 a.m.; Between Feb. 1, 8:00 a.m. and 8:00 p.m.; After Feb. 1, 8:00 p.m. Also, it is obvious that, compare to linear regression and K-Neighbor regressor, Random Forest Regressor is better to use. linear regression has the largest error. Its error value is about tripled compared to the other two. K-Neighbor regressor also did not as well as Random Forest Regressor. Also, although it has a little difference with random forest regressor (RFR), RFR still perform better generally. Besides, we can find that the error of the second period is much larger than the other two intervals. I think this is because it is time of a big event. A lot of people tweet in this time period. The amount of the data is much larger than the other period, so the relative accuracy may not change.

After the best regressor which is random forest regressor in this case is decided, we try to predict through an aggregated data. We first load each hashtag separately, and then add them to a same data frame. Next, predict the data through this three-time interval:

Aggregated Hashtags

	Aggregated
Random Forest Regression	Before Feb. 1, 8:00 a.m. Averaged error is: [395.03697344]
	Between Feb. 1, 8:00 a.m. and 8:00 p.m. Averaged error is: [102679.98181818]
	After Feb. 1, 8:00 p.m. Averaged error is: [430.20970149]

Analysis

We can see from the result that the value of error in each interval is much larger comparing to the random forest regression results from the previous part where the time is separated. The reason that this happens is the amount of data is becoming large since we put all hashtags together. This phenomenon is quite normal when we see that the averaged error is always greater between Feb. 1, 8:00 am and 8:00 pm than that of other times because the difference in amounts of tweets. In addition, due to the irregular events before, during and after the Super Bowl, the data is harder to predict. Therefore, we can conclude that when doing data analysis, it is better to split the dataset into different segments based on time if the data is time-variant. Such method will usually give us a better prediction than stacking everything in one model.

Problem 1.5

Results

```
Predict for sample1_period1
Now loading sample1_period1
Actual Tweet Num: [ 177.]
Predict Tweet Num [ 169.4]
MAE 7.6

Predict for sample2_period2
Now loading sample2_period2
Actual Tweet Num: [ 82890.]
Predict Tweet Num [ 3822.]
MAE 79068.0

Predict for sample3_period3
Now loading sample3_period3
Actual Tweet Num: [ 523.]
Predict Tweet Num [ 568.4]
MAE 45.4

Predict for sample8_period1
Now loading sample8_period1
Actual Tweet Num: [ 11.]
Predict Tweet Num [ 176.6]
MAE 165.6
```

```
Predict for sample4_period1
Now loading sample4_period1
Actual Tweet Num: [ 201.]
Predict Tweet Num [ 206.4]
MAE 5.4

Predict for sample5_period1
Now loading sample5_period1
Actual Tweet Num: [ 210.]
Predict Tweet Num [ 252.2]
MAE 42.2

Predict for sample6_period2
Now loading sample6_period2
Actual Tweet Num: [ 37278.]
Predict Tweet Num [ 3544.5]
MAE 33733.5
```

```
Predict for sample7_period3
Now loading sample7_period3
Actual Tweet Num: [ 120.]
Predict Tweet Num [ 60.7]
MAE 59.3

Predict for sample9_period2
Now loading sample9_period2
Actual Tweet Num: [ 2789.]
Predict Tweet Num [ 2888.2]
MAE 99.2

Predict for sample10_period3
Now loading sample10_period3
Actual Tweet Num: [ 61.]
Predict Tweet Num [ 58.5]
MAE 2.5
```

File Name	Actual Tweets	Predict Tweets	MAE
Sample1_period1	177	169	8
Sample2_period2	82890	3822	79068
Sample3_period3	523	568	45
Sample4_period1	201	206	5
Sample5_period1	21	252	42
Sample6_period2	37278	3544	33734
Sample7_period3	120	61	59
Sample8_period1	11	177	166
Sample9_period2	2789	2888	99
Sample10_period3	61	59	2

Analysis

In this part we want to use previous 5 hour's features to predict the tweets number of the 6th hour. Here we used random forest to predict our data due to the previous work.

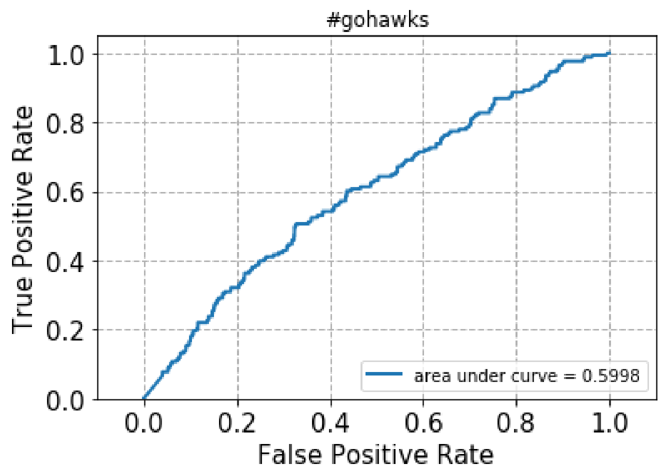
Here what we need to do first is to get the feature every 5 hours. Like in problem 1.2, we choose the tweet number, total retweets, total followers, maximum followers and time of the day to be 5 features. And one important thing here is when loading the data from the text file, we need to deal with some hours which do not have any tweets and replace the data with 0. And since we need to train the model due to different periods, we also need to extract feature for different period and use those data to train the models. Also, consider there is a test file only have 5 hours rather than 6 hours, we make the window size to be a hyperparameter and train the model with a 4-hour window and a 5-hour window.

Above result shows the prediction of our model. We can find that in most time the random forest model can have a great prediction with MAE smaller than 100. While in the predictions we made there are 2 bad predictions and both of them is from period 2 and the largest MAE can be about 80000. This is probably because the number of the train data for the period 2 is smaller than others so it may not have a great performance as period 1 and period 3 do.

PART 2: Fan Base Prediction

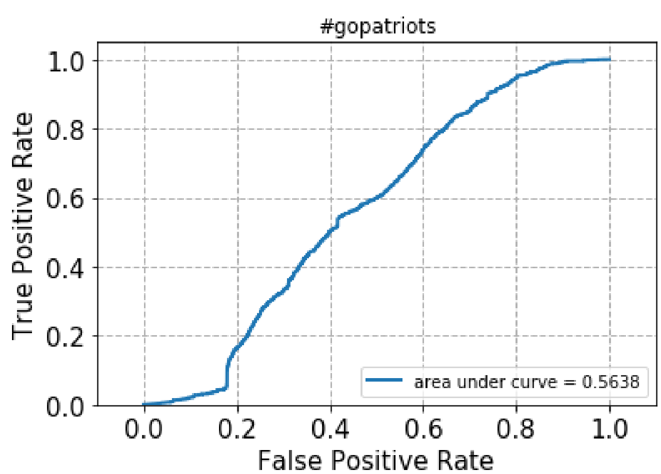
#gohawks

```
Fan Base Prediction for gohawks
/Users/oliviajin/anaconda3/lib/python3.6/site-packages/sklearn/metrics/
classification.py:1135: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
Confusion Matrix:
 [[11380     0]
 [  168     0]]
Accuracy:  0.985452026325
Recall: 0.985452026325
Precision: 0.971115696188
```



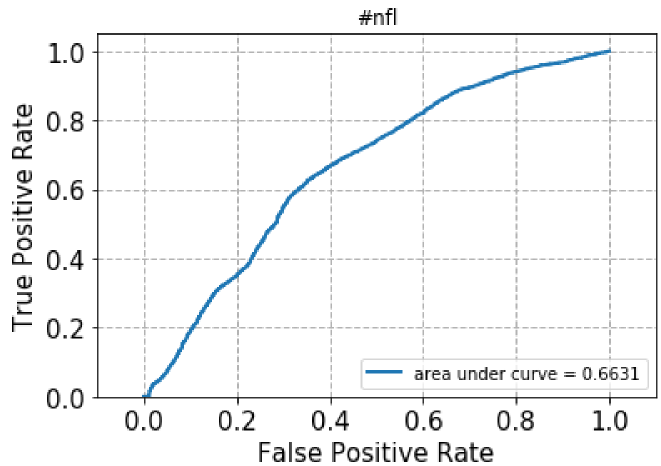
#gopatriots

```
Fan Base Prediction for gopatriots
/Users/oliviajin/anaconda3/lib/python3.6/site-packages/sklearn/metrics/
classification.py:1135: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
Confusion Matrix:
 [[10064     0]
 [ 1840     0]]
Accuracy:  0.845430107527
Recall: 0.845430107527
Precision: 0.714752066713
```



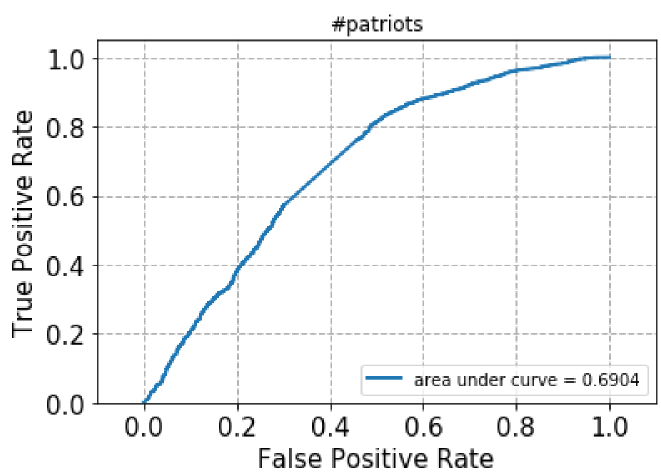
#nfl

```
Fan Base Prediction for nfl
/Users/oliviajin/anaconda3/lib/python3.6/site-packages/sklearn/metrics/
classification.py:1135: UndefinedMetricWarning: Precision is ill-defined and
being set to 0.0 in labels with no predicted samples.
  'precision', 'predicted', average, warn_for)
Confusion Matrix:
 [[5969    0]
 [8239    0]]
Accuracy:  0.420115427928
Recall: 0.420115427928
Precision:  0.176496972783
```



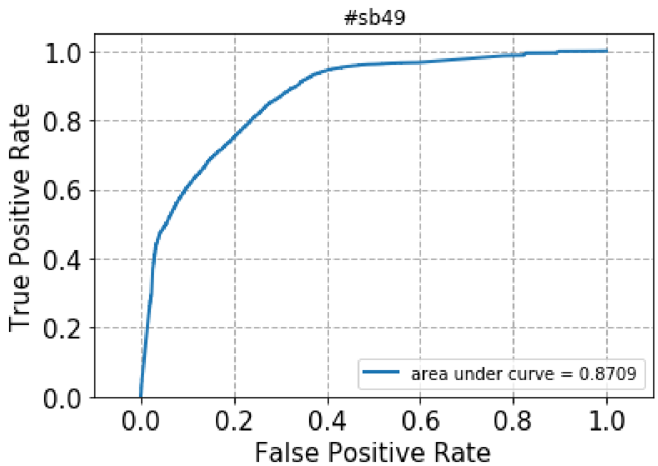
#patriots

```
Fan Base Prediction for patriots
Confusion Matrix:
 [[  309  1283]
 [  779 20509]]
Accuracy:  0.909877622378
Recall: 0.909877622378
Precision: 0.895402649186
```



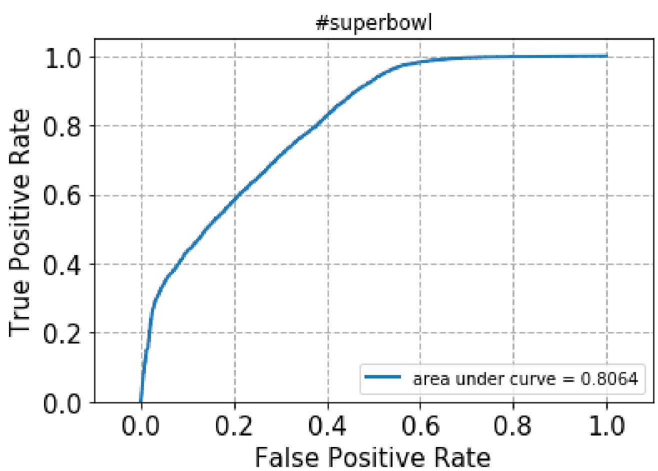
#sb49

```
Fan Base Prediction for sb49
Confusion Matrix:
 [[13887  1330]
 [ 6506  9079]]
Accuracy:  0.745600935004
Recall: 0.745600935004
Precision: 0.777739959209
```



#superbowl

```
Fan Base Prediction for superbowl
Confusion Matrix:
 [[ 6366  7488]
 [ 1063 26444]]
Accuracy:  0.793259350596
Recall: 0.793259350596
Precision: 0.805311953531
```



Analysis:

In part 2 we want to use a binary classifier to predict the location of the author of a tweet. In this part we separate the data according to their location from MA or WA. And when we deal with the location of Washington we need to extract the location from Washington D.C. After we get the tweets from MA and WA we choose 80% of data to be train data and the rest to be test data. Then we do the same job as we have done in project 1 to see the accuracy of the binary classifier by using CountVectorizer, TfidfTransformer and SVD. Here we just extract the stop words for a shorter running time. If we add stemmer in the analyzer we may get a better result.

In the result we get we can find that a lager data will have a higher AUC while accuracy and recall scores are not in that case. And we can find that people will prefer to support the team in their city or state, for example, people who live in WA posting tweet for #gohawks much more than people in WA and vice versa. In some hashtags the model will get a high accuracy and recall score but actually it does not performance so well. For example, in #gohawks it has an accuracy more than 90% but in confusion matrix we can see that all tweets posted in MA is predicted as WA. But because of low proportion of MA it doesn't influence the accuracy too much, but the model still need to be optimized.

PART 3: Design Your Own Project

The dataset in hands is rich as there is a lot of metadata to each tweet. It is a great idea to do a sentiment analysis of the fans for both teams to see how tweets reflect their emotions. Sentiment analysis is a process of determining whether a piece of writing is objective or subjective, and if subjective, then whether the text is positive or negative. This analysis is a type of opinion mining by deriving the attitude of the author.

Part I – Subjectivity Analysis

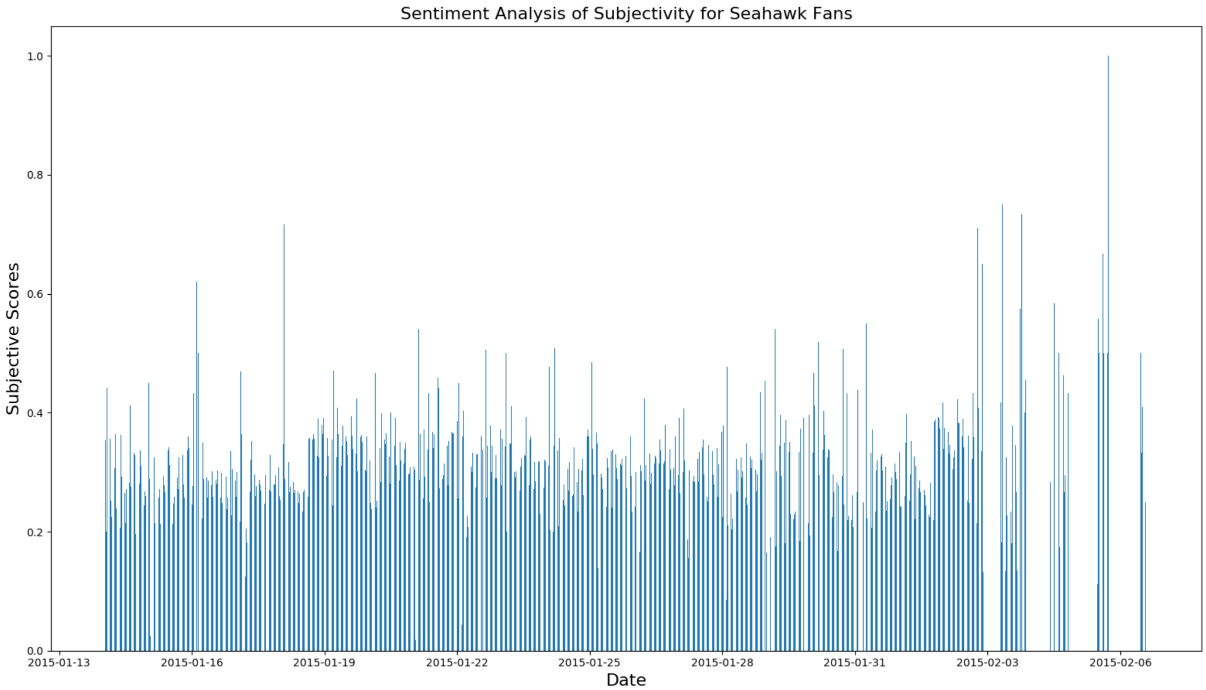
In the first part of the project, we want to do a sentiment analysis on subjectivity, which means to see how objective/subjective the tweets from both fans are. We use a toolkit called TextBlob, which is a Python library for processing textual data. It provides a simple API for diving into common natural language processing tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

Link: http://textblob.readthedocs.io/en/dev/

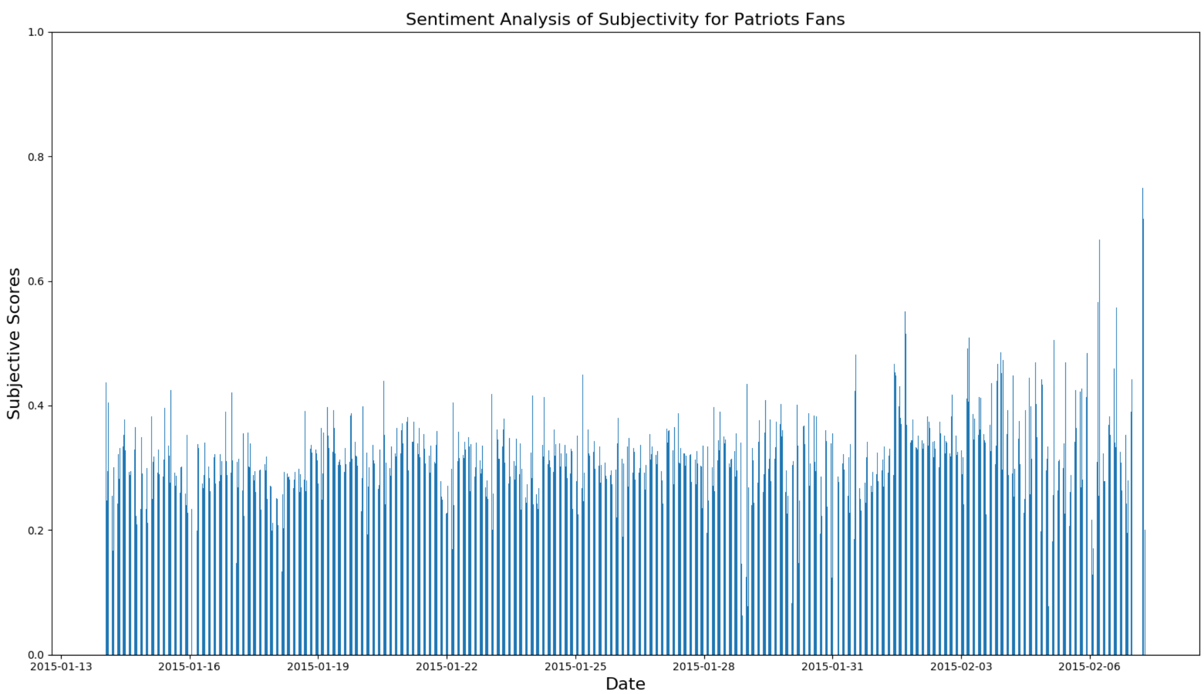
We use a class called Sentiment, which can calculate the subjective scores when passing a piece of literature. The subjective score is a float within the range [0.0, 1.0], where 0.0 means the writing is very objective without any emotion, and 1.0 indicates very subjective with strong emotion.

We begin the analysis by calculating the subjective scores for Seahawk fans using the hashtag #gohawks. We assume that anyone uses the #gohawks is a fan because the phrase "go" indicates that they cheer for the Seattle Seahawks. The subjective scores are calculated for all tweets under this hashtag, and since there more than 188,000 tweets, we divided the tweets into 30-minute sections and then average the subjective score in each section. This will reduce the effects of outliers and extremes. So, for each bar in the figures below, it is the average subjective scores of the fans during a 30-miniute period.

Seahawk Fans



Patriots Fans

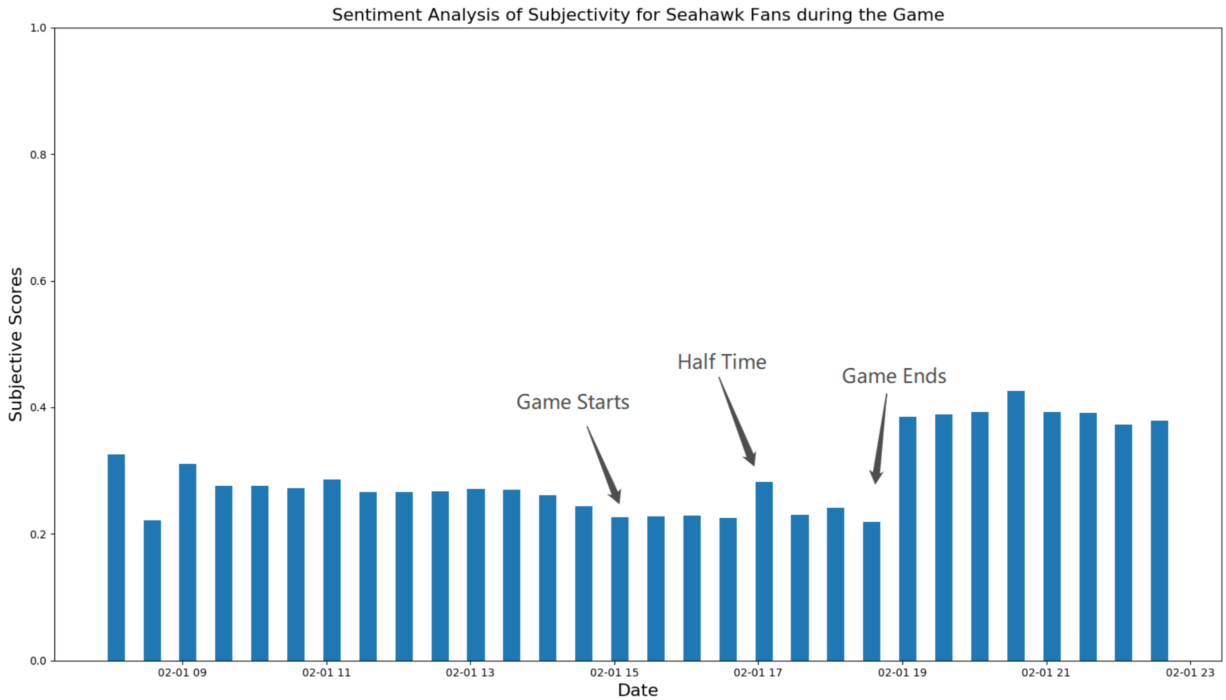


Analysis

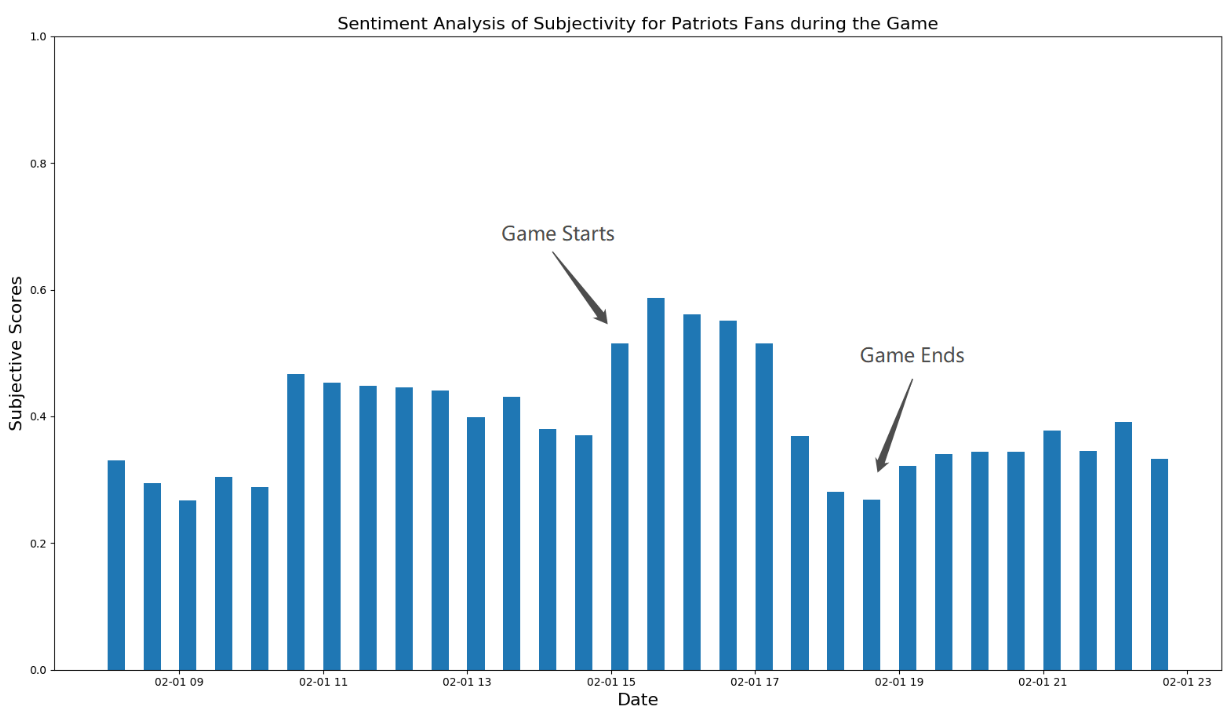
From the above figures, we see that the tweets from the Patriots fans are more subjective than the Seahawk fans. The average subjectivity scores for Patriots fans is 0.459, while that of Seahawk is only 0.282. The subjective scores are also more uniformly distributed for Patriots fans than that of the Sea Hawk fans. It is easy to see that there are many spikes (i.e. high scores) for Seahawks for a certain time, especially several days after the game. The average subjective scores increase significantly to 0.65 and above, and even to 1.0 in an extreme case. The same thing is applied to the Patriots fans but not as extreme. The reason that Patriots fans have such high overall subjectivity scores is the Patriots won the game. Many fans are expressing their emotions after a great victory. Their tweets will contain more certain subjective key words like “great”, “love”, “happy”, etc. On the other hand, since Seahawk lost game, although its fans are sad, people usually do not want to express their frustration on social networks, so they admitted the failure by using words that are more objective. An interesting observation is that even after days of the game, the subjective scores become very high. A possible explanation is that these people who tweeted are the true fans of the team. When majority of the bystanders have moved their focus on other topics, the comments of the true fans take more weights in the average scores. Since the true fans love their teams a lot, the average subjective scores will definitely increase. Another interesting observation is that days before the game, the subjective scores for the Seahawks fans have many spikes, which may due to the reason that they have won the Super Bowl last year, and the fans seek to let the Seahawks to be the first team that wins two consecutive Super Bowl after the Patriots.

Next, we zoom-in the plots to focus on the subjective scores around the game time. The following plots are shown:

Seahawk Fans



Patriots Fans



Analysis

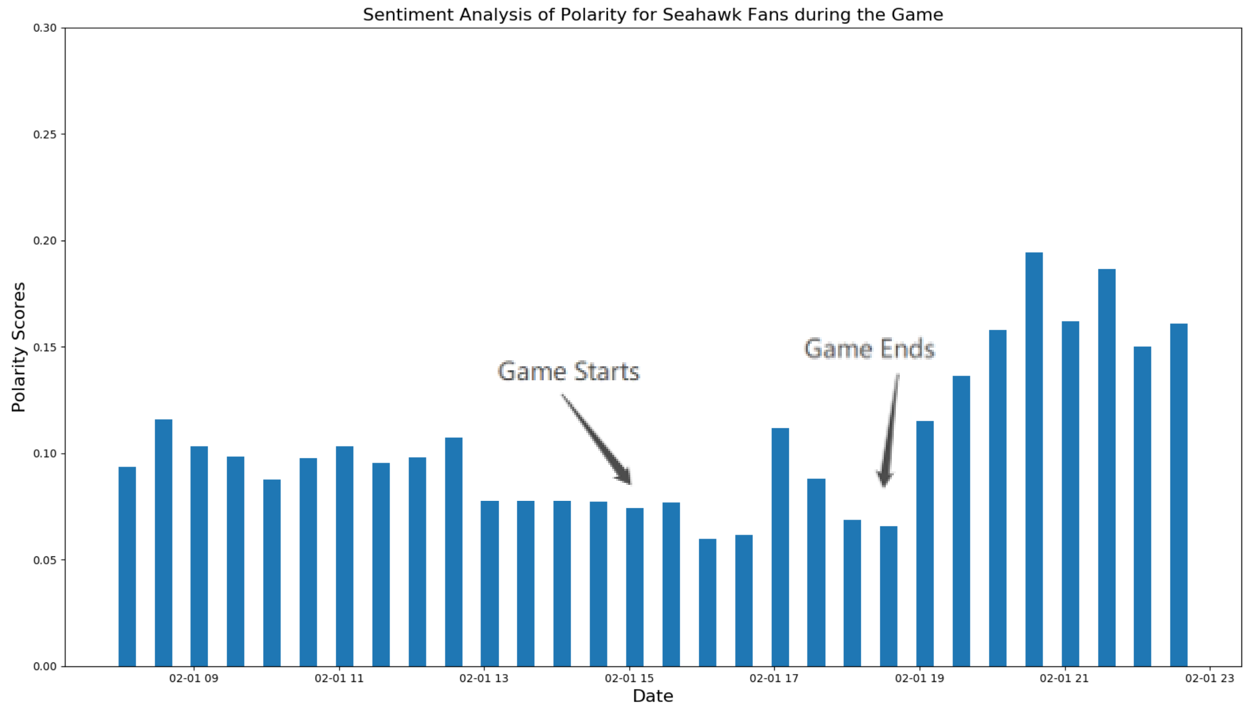
After we focus on the subjective scores during the game time instead of the overall dataset, we have more insights on the fans based on their tweets. Right before the game starts, we see that the Patriots fans are more subjective than the Seattle Seahawks fans. This is because based on the pre-game analysis, the Patriots are more likely to win the game than its opponents. There are more cheers for the Patriots and the fan base is significantly larger than Seahawks as well. When the game starts, the Patriots scored the first touchdown and their scores are always equal or ahead of the Seahawks in the first half of the match. As we can see, the fans are very excited and happy, hoping the Patriots to secure the game.

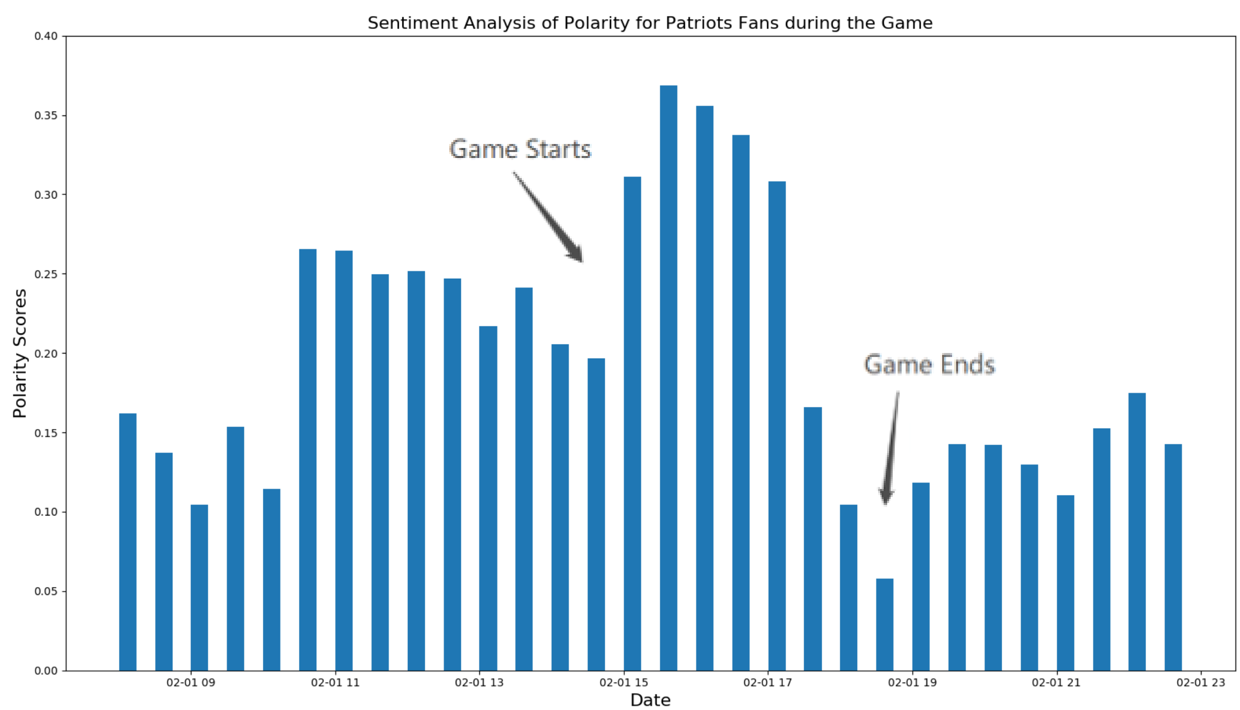
Right before the half time, when there are 2 seconds left for the first half, the Seahawks scored the second touchdown by Chris Matthews, and the fans are amazed by such achievement. Therefore, during the half time, there is a sudden increase in the subjective scores from the Seahawks fans because they are cheering. Another key observation is that in the second half of the game, the Seahawks started to take over the lead. They are ahead of the score board for the entire second half. This situation is reflected on the subjective score of the Patriots fans. As it is illustrated on the figures, the subjectivity scores decrease significantly compare to the first half, because the fans are nervous and afraid to lose the game, so they stopped cheering and watch the game with more objectivity.

Now, right after the game ends, the subjectivity scores increased for both fans in the next few hours. This is because everyone is discussing the game results with emotions. The Patriots fans could be very happy about the results because they have won the game, while Seahawk fans expressed their frustration and sadness on Twitter. So, the fans from the both teams expressed their attitude toward the game. However, the Seahawk were ahead of the game for the entire second half and only lost in the last 2 minutes of game, so the sudden change of the game led to a huge emotional explosion on Twitter, and therefore their subjective score is higher than the Patriots fans after the game.

Part II – Polarity Analysis

In this part of the project, we analyze the attitude polarity of the fans from both teams. The polarity score is a float within the range [-1.0, 1.0], which a score of -1.0 means the attitude is very negative, while a score of 1.0 means the attitude is very positive.





Analysis

From the above figures, it is easy to notice that they have similar trend as the subjective score, but there are more observations to analyze. The polarity score of the Patriots fans is significantly greater than the polarity score of the Seahawk fans. For example, the polarity score before the game for the Patriots are around 0.25 while that of the Seahawks is only around 0.1. As the game starts, the two teams show a complete opposite trend on polarity because of the scores. New England scored the first touchdown in the second quarter, and as a result, there is a sudden increase of positive attitude from 0.2 to 0.35 of the fans. On the other hand, since the Seahawks are losing the game, the positive attitude started to drop until the end of the first half where they scored a touchdown in the last 2 seconds. Therefore, during the half time, a sudden increase of positive attitude is shown from the Seahawks fans.

In the second half of the game, the Seattle Seahawks started to take the lead, and we can see that the positive attitude of the Patriots fans dropped significantly, nearly half of the score as before. Such positive attitude continued to decrease because the Seahawks kept increasing the score lead. At nearly the end of the match, the positive attitude reached all time low, a score of 0.05, because there is not much time left the Patriots while the score difference is 10. However, the Patriots are able to score two touchdowns consecutively in the last several minutes of game, which the fans are suddenly become very positive again.

After the game, the fans from both teams are quite positive in their attitude. This is because Super Bowl 49 is a great game which two teams scored back and forth. People cannot make a judgement on victory until the end of the game since the score is so close. So, they celebrate such a good match together which the fans are appreciating their team’s performance.