EE219 Project 5
Popularity Prediction on Twitter
Winter 2018

Shaoming Cheng, 505034686
Yao Xie, 105036239
Jiahui Li, 004356402
Ruiyi Wu, 304615036

I. Introduction

Social network analysis is the process of investigating social structures through the use of networks and graph theory. To predict the popularity of a subject or event is a useful practice in social network analysis. In this project, we have performed predictions of the popularity of 49th super bowl based on the data collected from Twitter, which is an excellent platform to perform such analysis. To be specific, we have tried to predict a tweet activity related to the 49th super bowl in the future based on the current and previous tweet activities for a hashtag. Moreover, we have used different models and compared the accuracy among them. Last, we have proposed a new project problem based on the analysis we learned from this project.

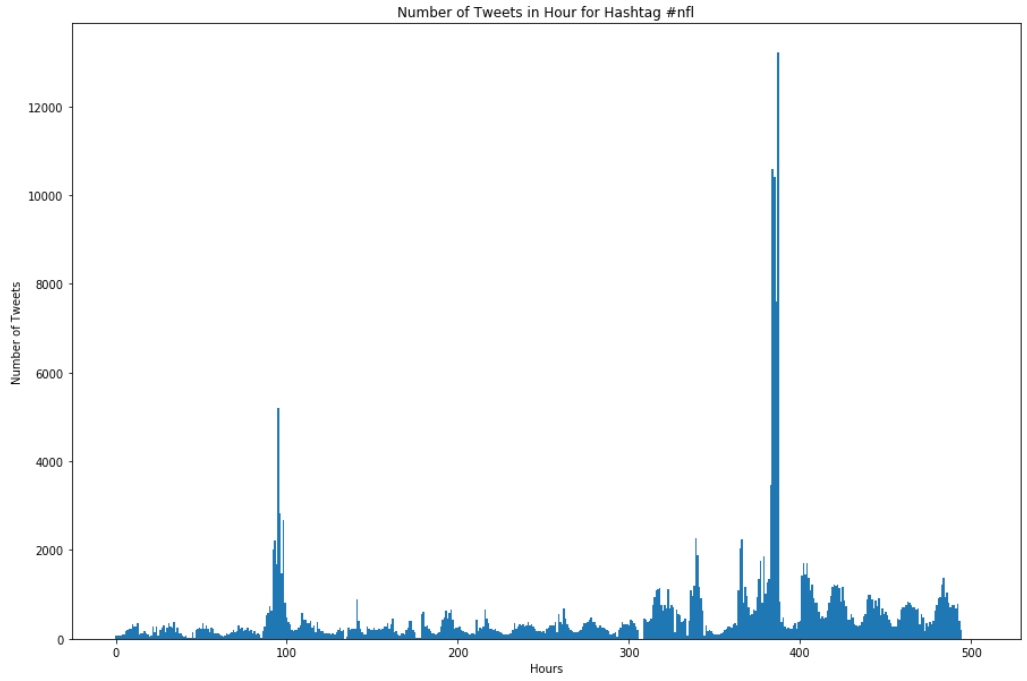
II. Part 1: Popularity Prediction

Problem 1.1

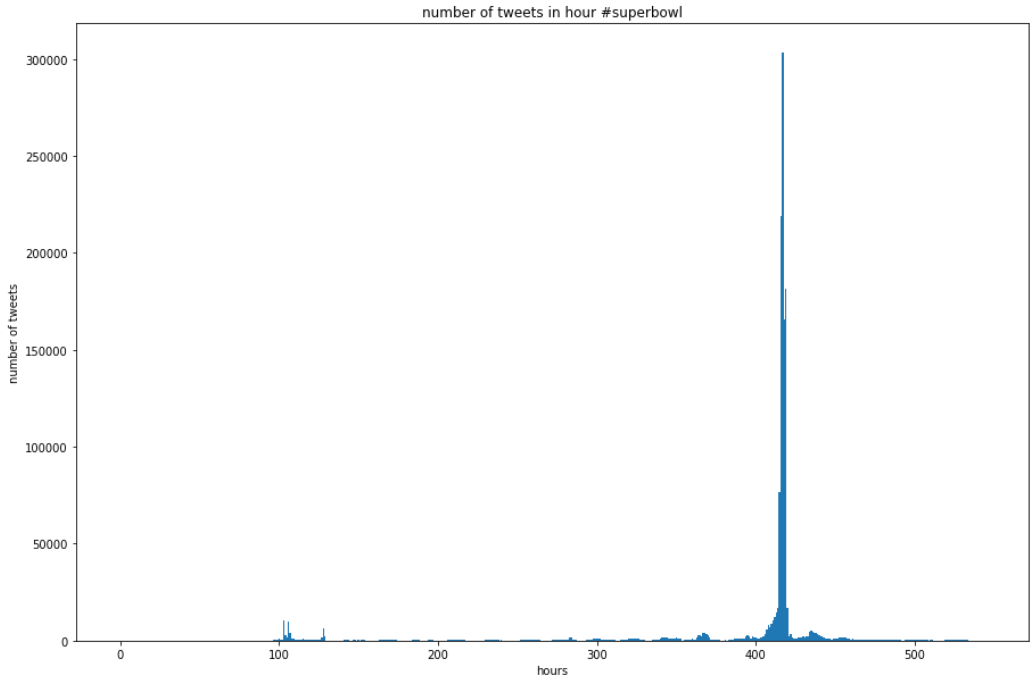
In this section, we first need to calculate the statistics for each hashtag, in terms of average number of tweets per hours, average number of followers of users posting the tweets and average number of retweet. The numbers are summarized as follows:

Hashtag	gopatriots	gohawks	nfl	patriots	sb49	superbowl
Average number of tweets /hour	46.35	380.07	513.93	845.79	1420.87	2472.06
Average number of followers of users posting tweets	1401.90	2203.93	4653.25	3309.98	10267.31	8858.97
Average number of retweets	1.40	2.01	1.54	1.78	2.51	2.39

The histogram of number of tweets in hour with 1-hour bins for hashtag #nfl is:



Similarly, the histogram of number of tweets in hour with 1-hour bins for hashtag #superbowl is:



Problem 1.2

In this part, we fit a linear regression model to each hashtag, and the five features extracted from data in the previous hour and used to predict the number of tweets in the next hour are: (1) number of tweets, (2) total number of retweets, (3) sum of the number of followers

of the users posting the hashtag, (4) maximum number of followers of the users posting the hashtag and finally (5) time of the day with 24 values representing hours of the day. The OLS regression results for each hashtag are shown as follows.

hashtag #gopatriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.776
Model:	OLS	Adj. R-squared:	0.774
Method:	Least Squares	F-statistic:	608.8
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	7.25e-283
Time:	16:25:05	Log-Likelihood:	-5483.3
No. Observations:	887	AIC:	1.098e+04
Df Residuals:	881	BIC:	1.101e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	6.5478	7.664	0.854	0.393	-8.493	21.589
x1	0.2468	0.572	0.431	0.666	-0.876	1.370
x2	0.4815	0.084	5.764	0.000	0.318	0.645
x3	-20.1878	0.926	-21.793	0.000	-22.006	-18.370
x4	0.0007	7.37e-05	9.431	0.000	0.001	0.001
x5	-0.0006	8.77e-05	-7.016	0.000	-0.001	-0.000

Omnibus:	1133.785	Durbin-Watson:	2.313
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1164649.702
Skew:	5.777	Prob(JB):	0.00
Kurtosis:	180.141	Cond. No.	7.40e+05

hashtag #gohawks

OLS Regression Results

Dep. Variable:	y	R-squared:	0.467
Model:	OLS	Adj. R-squared:	0.464
Method:	Least Squares	F-statistic:	154.7
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	5.17e-118
Time:	16:49:20	Log-Likelihood:	-7208.6
No. Observations:	890	AIC:	1.443e+04
Df Residuals:	884	BIC:	1.446e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

const	69.9156	51.973	1.345	0.179	-32.089	171.920
x1	1.2580	3.909	0.322	0.748	-6.413	8.929
x2	0.5885	0.118	5.003	0.000	0.358	0.819
x3	0.0122	0.039	0.313	0.755	-0.064	0.089
x4	7.452e-05	6.64e-05	1.123	0.262	-5.58e-05	0.000
x5	-0.0003	0.000	-2.085	0.037	-0.001	-1.54e-05

Omnibus:	1584.963	Durbin-Watson:	2.309
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3284266.661
Skew:	11.392	Prob(JB):	0.00
Kurtosis:	299.724	Cond. No.	4.58e+06

hashtag #nfl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.718
Model:	OLS	Adj. R-squared:	0.716
Method:	Least Squares	F-statistic:	453.3
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	5.48e-242
Time:	16:55:49	Log-Likelihood:	-6641.6
No. Observations:	898	AIC:	1.330e+04
Df Residuals:	892	BIC:	1.332e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	53.6779	26.088	2.058	0.040	2.477	104.879
x1	-1.7136	1.922	-0.892	0.373	-5.486	2.058
x2	0.8782	0.071	12.393	0.000	0.739	1.017
x3	-3.0111	0.151	-19.942	0.000	-3.307	-2.715
x4	3.082e-05	1.63e-05	1.896	0.058	-1.08e-06	6.27e-05
x5	7.267e-06	2.31e-05	0.315	0.753	-3.8e-05	5.26e-05

Omnibus:	1527.771	Durbin-Watson:	2.379
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1309820.625
Skew:	10.726	Prob(JB):	0.00
Kurtosis:	188.866	Cond. No.	7.84e+06

hashtag #patriots

OLS Regression Results

Dep. Variable:	y	R-squared:	0.689
Model:	OLS	Adj. R-squared:	0.687
Method:	Least Squares	F-statistic:	394.7
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	3.60e-223

Time:	16:58:46	Log-Likelihood:	-8100.2
No. Observations:	898	AIC:	1.621e+04
Df Residuals:	892	BIC:	1.624e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	132.5815	130.876	1.013	0.311	-124.278	389.441
x1	-2.6816	9.720	-0.276	0.783	-21.759	16.395
x2	0.9491	0.031	30.621	0.000	0.888	1.010
x3	-1.0498	0.204	-5.143	0.000	-1.450	-0.649
x4	-6.584e-05	1.49e-05	-4.434	0.000	-9.5e-05	-3.67e-05
x5	0.0003	7.41e-05	3.514	0.000	0.000	0.000

Omnibus:	1675.046	Durbin-Watson:	1.965
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2573553.599
Skew:	12.848	Prob(JB):	0.00
Kurtosis:	263.999	Cond. No.	1.81e+07

hashtag #sb49

OLS Regression Results

Dep. Variable:	y	R-squared:	0.806
Model:	OLS	Adj. R-squared:	0.805
Method:	Least Squares	F-statistic:	740.9
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	1.35e-314
Time:	17:00:29	Log-Likelihood:	-8628.1
No. Observations:	898	AIC:	1.727e+04
Df Residuals:	892	BIC:	1.730e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	150.5008	234.756	0.641	0.522	-310.238	611.240
x1	-11.4223	17.506	-0.652	0.514	-45.780	22.935
x2	0.9601	0.027	36.162	0.000	0.908	1.012
x3	-0.2707	0.102	-2.647	0.008	-0.471	-0.070
x4	-1.147e-05	3.21e-06	-3.572	0.000	-1.78e-05	-5.17e-06
x5	0.0001	3.88e-05	3.297	0.001	5.18e-05	0.000

Omnibus:	1944.216	Durbin-Watson:	1.640
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7402477.325
Skew:	17.684	Prob(JB):	0.00
Kurtosis:	446.383	Cond. No.	1.38e+08

hashtag #superbowl

OLS Regression Results

Dep. Variable:	y	R-squared:	0.811
Model:	OLS	Adj. R-squared:	0.810
Method:	Least Squares	F-statistic:	767.7
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	3.75e-320
Time:	17:08:58	Log-Likelihood:	-9131.6
No. Observations:	898	AIC:	1.828e+04
Df Residuals:	892	BIC:	1.830e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-22.8032	416.142	-0.055	0.956	-839.535	793.928
x1	7.7673	30.659	0.253	0.800	-52.404	67.939
x2	1.0590	0.151	7.012	0.000	0.763	1.355
x3	-3.7151	0.148	-25.048	0.000	-4.006	-3.424
x4	1.358e-05	2.17e-05	0.627	0.531	-2.9e-05	5.61e-05
x5	0.0002	0.000	2.431	0.015	4.8e-05	0.000

Omnibus:	2021.315	Durbin-Watson:	1.843
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9758455.955
Skew:	19.319	Prob(JB):	0.00
Kurtosis:	512.227	Cond. No.	2.20e+08

p-value explanation:

In statistics, the p-value is a function of the observed sample results relative to a statistical model, and it measures how well of the observation is. In our case, a very low p-value for a feature indicates that there is substantial evidence against the null hypothesis to reject it, and we can thus claim that this feature plays a vital role in the performance of the model.

The p-value is the probability that the outcome would be at least as extreme or probably even more extreme than the result that was observed given a null hypothesis for the probability distribution of the data. As the p-value decreases, the evidence against this null hypothesis increases. Thus, we consider the features with the lowest p-values to be most significant.

t-value explanation:

The t-statistic is a ratio of the departure of an estimated parameter from its notional value and its standard error. The t-test is assessing how statistically significant a particular explanatory variable is.

It is important to find such significant explanatory variables to fit the regression model efficiently. In other words, the larger the absolute value of T is, the stronger the evidence against the null hypothesis that there is a significant difference will be. The closer T is to 0, the more likely there isn't a considerable difference. Thus, we consider the features with the highest absolute t-values to be most significant.

Here is a summary of the training accuracy (rmse) and R-squared measure for different hashtags:

Hashtag	gopatriots	gohawks	nfl	patriots	sb49	superbowl
training accuracy	226.98	769.30	511.16	2467.52	404.14	7111.82
R-squared measure	0.50	0.63	0.66	0.68	0.84	0.84

We also try to compare the essential features in each hashtag. In the following table, we highlight the top three most important elements as blue.

T-values:

Hashtag	gopatriots	gohawks	nfl	patriots	sb49	superbowl
X1	0.431	0.322	0.892	0.276	0.652	0.253
X2	5.764	5.003	12.393	30.621	36.162	7.012
X3	21.793	0.313	19.942	5.143	2.647	25.048
X4	9.431	1.123	1.896	4.434	3.572	0.627
X5	7.016	2.085	0.315	3.514	3.297	2.431

Discussion:

From training accuracy and R-squared measure table, the data labeled with "gopatriots" has the highest accuracy while the hashtag with "superbowl" has the lowest accuracy. We have expected this result since the first hashtag has the smallest data while the second hashtag has the most massive data.

On the second table, we can conclude that feature X4 (maximum number of followers of the users posting the hashtag) is the most significant feature compared to others. X1 (the total

number of tweets) is the least important feature with the popularity prediction for most of the hashtags.

Problem 1.3

In this part, we fit a linear regression model to each hashtag, and also we add more features extracted from data in the previous hour and used to predict the number of tweets in the next hour. Moreover, For each of the top 3 features in our measurements, we draw a scatter plot of predictants (number of tweets for next hour) versus value of that feature. We define 12 features in total, and the table for each of them is the following:

index	x	features
0	x1	Number of tweets
1	x2	Number of retweets
2	x3	Number of followers
3	x4	Number of hashtags
4	x5	Number of favorites tweets
5	x6	Number of friends
6	x7	Number of verified users
7	x8	Number of citations
8	x9	Ranking score
9	x10	Influence
10	x11	Mathings
11	x12	Number of statuses count

hashtag #gopatriots

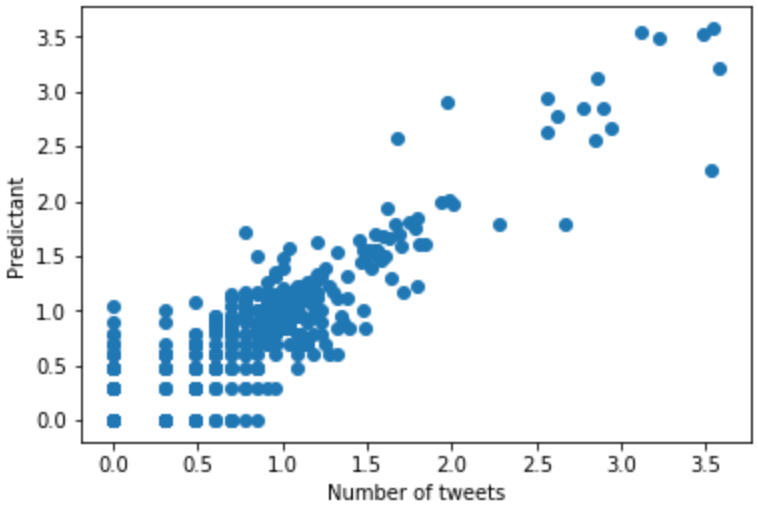
OLS Regression Results

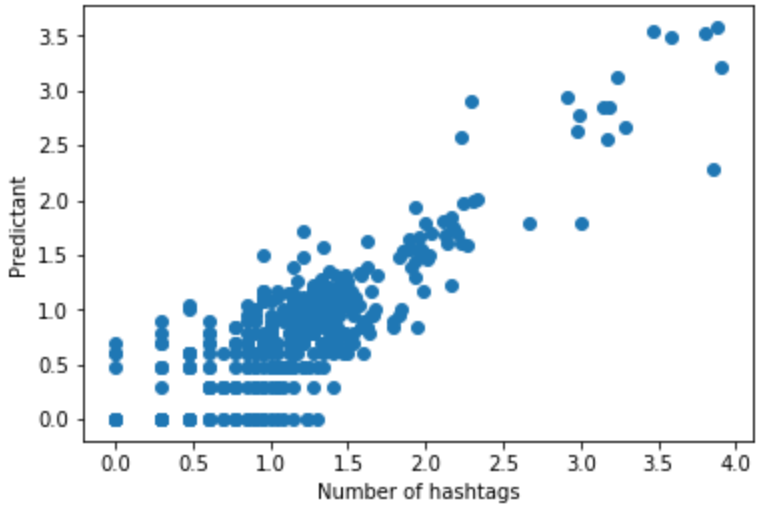
Dep. Variable:	y	R-squared:	0.881
Model:	OLS	Adj. R-squared:	0.879
Method:	Least Squares	F-statistic:	537.1
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	0.00
Time:	16:25:18	Log-Likelihood:	-5203.3
No. Observations:	887	AIC:	1.043e+04

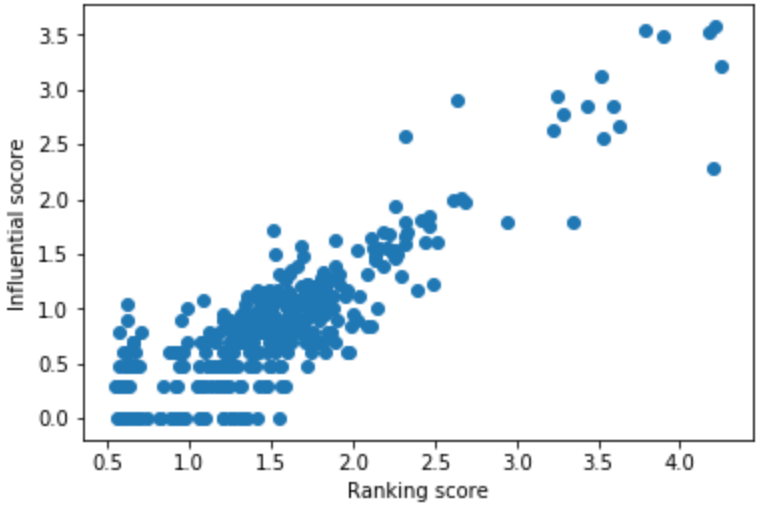
Df Residuals:	874	BIC:	1.049e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.3515	2.984	1.123	0.262	-2.506	9.209
x1	-24.9380	1.744	-14.297	0.000	-28.361	-21.515
x2	-30.9915	3.446	-8.994	0.000	-37.755	-24.228
x3	0.0003	2.48e-05	10.721	0.000	0.000	0.000
x4	5.0978	0.239	21.350	0.000	4.629	5.566
x5	10.2594	3.023	3.394	0.001	4.326	16.193
x6	0.0004	0.000	1.756	0.079	-4.63e-05	0.001
x7	-123.4702	13.617	-9.067	0.000	-150.197	-96.744
x8	-0.0796	0.168	-0.475	0.635	-0.408	0.249
x9	4.2777	0.345	12.404	0.000	3.601	4.955
x10	28.0679	6.155	4.560	0.000	15.988	40.148
x11	-3.1402	0.690	-4.554	0.000	-4.494	-1.787
x12	-0.0002	2.18e-05	-10.159	0.000	-0.000	-0.000

Omnibus:	553.779	Durbin-Watson:	2.225
Prob(Omnibus):	0.000	Jarque-Bera (JB):	165949.083
Skew:	1.631	Prob(JB):	0.00
Kurtosis:	69.929	Cond. No.	1.19e+07







hashtag #gohawks

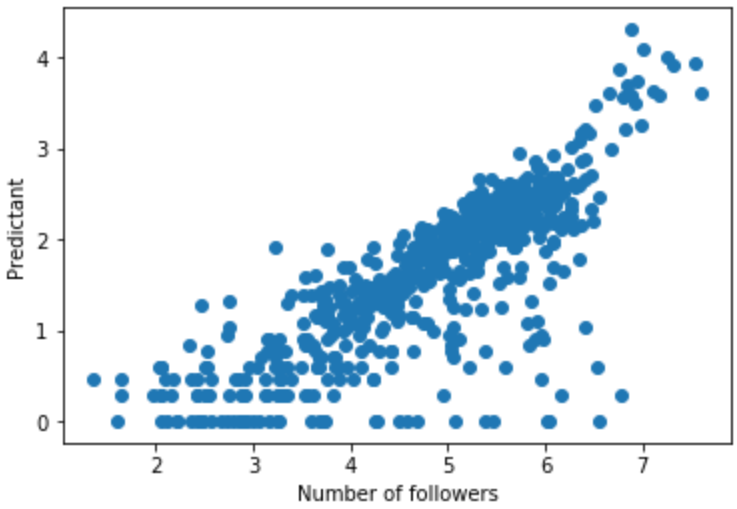
OLS Regression Results

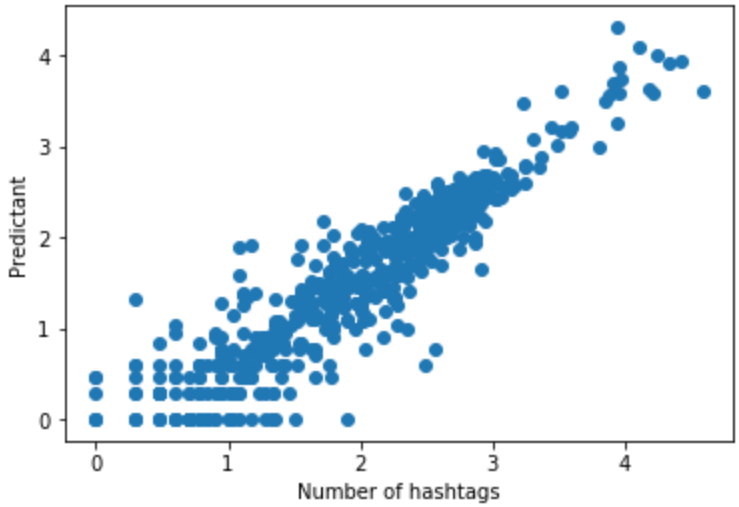
Dep. Variable:	y	R-squared:	0.627
Model:	OLS	Adj. R-squared:	0.621
Method:	Least Squares	F-statistic:	122.6
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	3.43e-178
Time:	16:49:58	Log-Likelihood:	-7049.9
No. Observations:	890	AIC:	1.413e+04
Df Residuals:	877	BIC:	1.419e+04
Df Model:	12		
Covariance Type:	nonrobust		

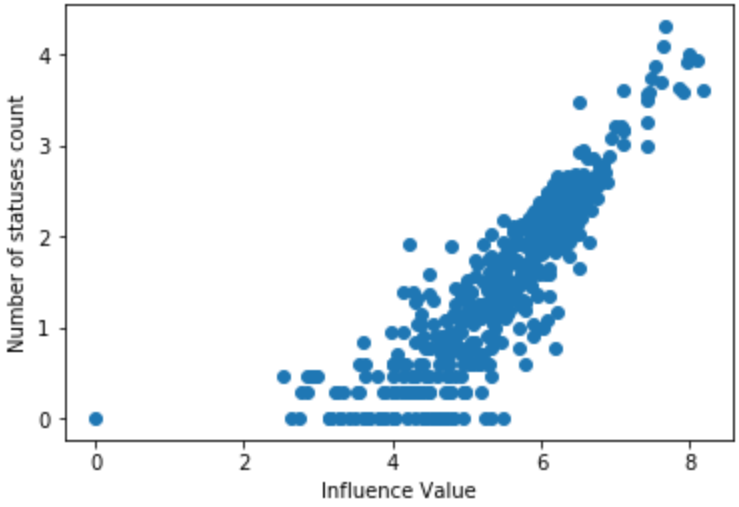
	coef	std err	t	P>|t|	[0.025	0.975]
const	-25.1905	25.105	-1.003	0.316	-74.464	24.083

x1	-9.7824	3.582	-2.731	0.006	-16.812	-2.753
x2	0.2696	0.215	1.255	0.210	-0.152	0.691
x3	-0.0003	4.63e-05	-6.139	0.000	-0.000	-0.000
x4	1.0740	0.255	4.215	0.000	0.574	1.574
x5	-0.0560	0.090	-0.623	0.534	-0.232	0.120
x6	0.0007	0.000	2.030	0.043	2.22e-05	0.001
x7	34.3170	10.290	3.335	0.001	14.120	54.514
x8	-0.3272	0.116	-2.832	0.005	-0.554	-0.100
x9	0.9117	0.621	1.469	0.142	-0.306	2.130
x10	14.3612	4.467	3.215	0.001	5.593	23.129
x11	2.7322	1.663	1.643	0.101	-0.532	5.996
x12	0.0002	2.69e-05	5.763	0.000	0.000	0.000

Omnibus:	1870.477	Durbin-Watson:	2.104
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5941367.019
Skew:	16.536	Prob(JB):	0.00
Kurtosis:	401.902	Cond. No.	1.11e+07







hashtag #nfl

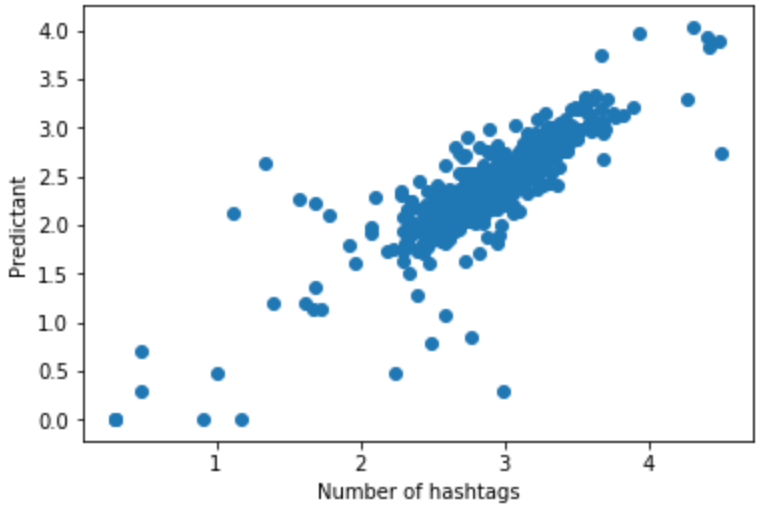
OLS Regression Results

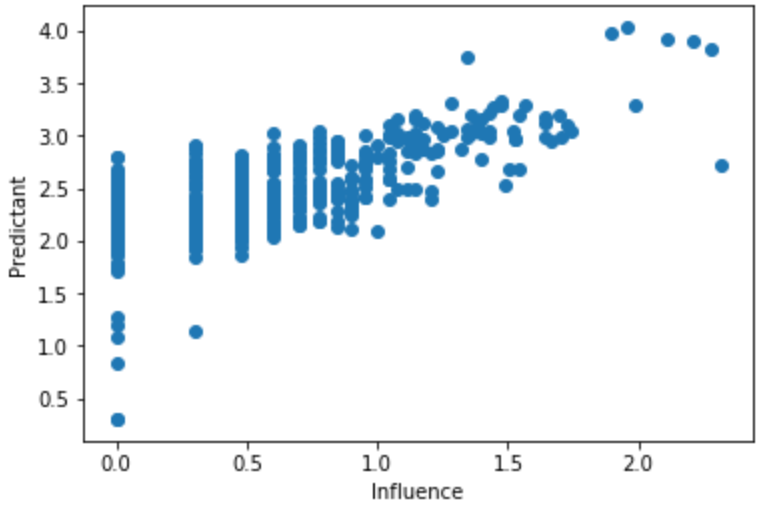
Dep. Variable:	y	R-squared:	0.757
Model:	OLS	Adj. R-squared:	0.754
Method:	Least Squares	F-statistic:	230.0
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	3.42e-262
Time:	16:56:37	Log-Likelihood:	-6573.7
No. Observations:	898	AIC:	1.317e+04
Df Residuals:	885	BIC:	1.324e+04
Df Model:	12		
Covariance Type:	nonrobust		

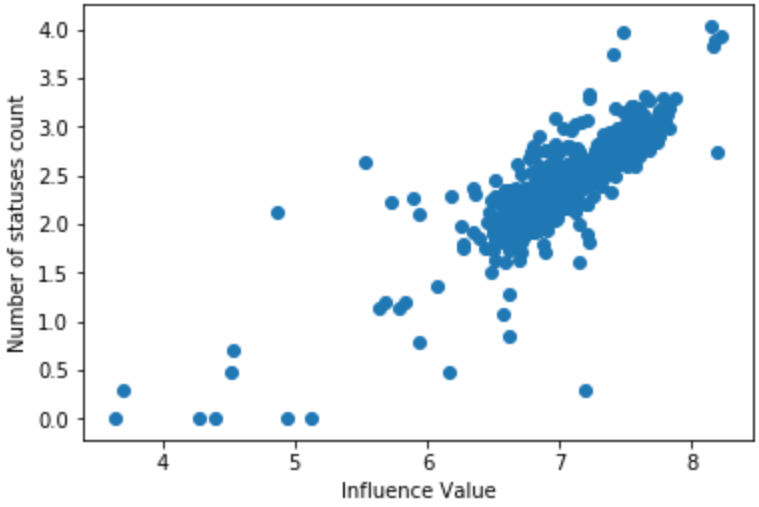
	coef	std err	t	P>|t|	[0.025	0.975]

const	-3.9641	15.957	-0.248	0.804	-35.283	27.355
x1	2.7499	1.532	1.795	0.073	-0.257	5.757
x2	-0.4659	0.813	-0.573	0.567	-2.062	1.130
x3	2.022e-05	1.07e-05	1.893	0.059	-7.46e-07	4.12e-05
x4	0.6705	0.076	8.781	0.000	0.521	0.820
x5	-1.1886	0.624	-1.903	0.057	-2.414	0.037
x6	-0.0004	0.000	-3.603	0.000	-0.001	-0.000
x7	6.5573	4.901	1.338	0.181	-3.061	16.175
x8	-0.1228	0.066	-1.862	0.063	-0.252	0.007
x9	-0.0865	0.258	-0.335	0.738	-0.593	0.420
x10	15.2444	3.675	4.149	0.000	8.033	22.456
x11	-3.6973	0.975	-3.791	0.000	-5.612	-1.783
x12	2.205e-05	2.77e-06	7.954	0.000	1.66e-05	2.75e-05

Omnibus:	1340.534	Durbin-Watson:	2.370
Prob(Omnibus):	0.000	Jarque-Bera (JB):	578654.399
Skew:	8.408	Prob(JB):	0.00
Kurtosis:	126.217	Cond. No.	2.56e+07







hashtag #patriots

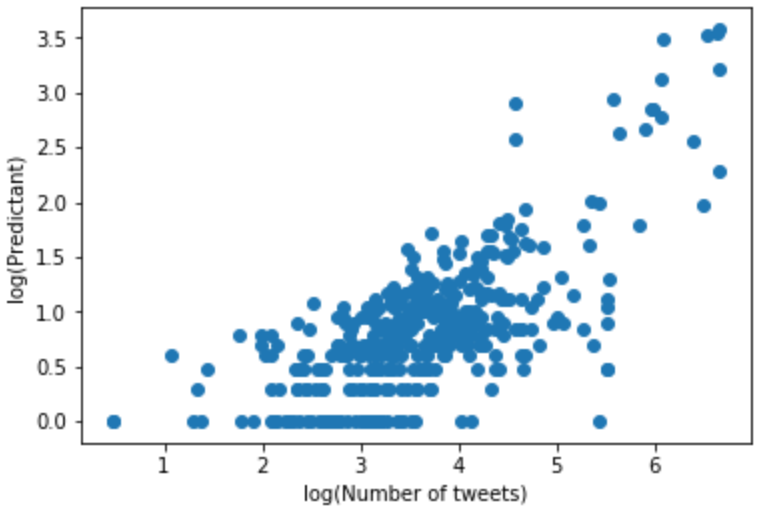
OLS Regression Results

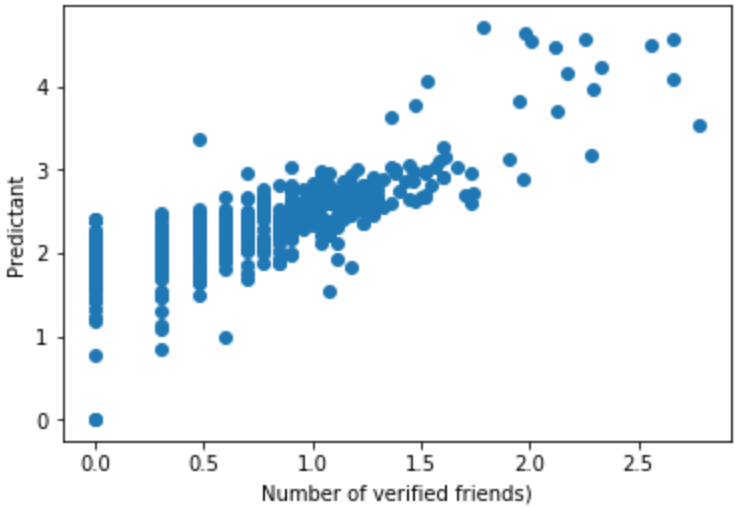
Dep. Variable:	y	R-squared:	0.749
Model:	OLS	Adj. R-squared:	0.746
Method:	Least Squares	F-statistic:	220.2
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	6.44e-256
Time:	17:00:23	Log-Likelihood:	-8003.4
No. Observations:	898	AIC:	1.603e+04
Df Residuals:	885	BIC:	1.610e+04
Df Model:	12		
Covariance Type:	nonrobust		

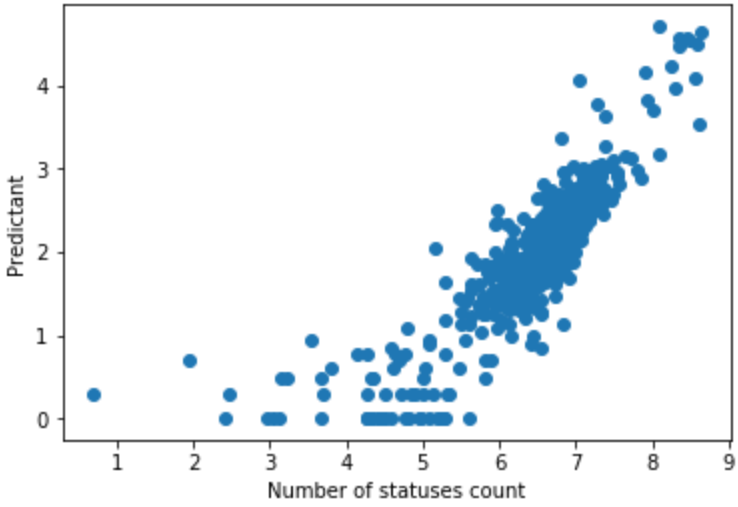
	coef	std err	t	P>|t|	[0.025	0.975]
const	-58.6457	72.473	-0.809	0.419	-200.885	83.594

x1	4.1782	1.979	2.112	0.035	0.295	8.062
x2	0.8233	0.899	0.916	0.360	-0.940	2.587
x3	-0.0002	4.41e-05	-4.098	0.000	-0.000	-9.42e-05
x4	0.0761	0.156	0.488	0.625	-0.230	0.382
x5	-0.8812	0.793	-1.111	0.267	-2.438	0.675
x6	-0.0016	0.000	-4.086	0.000	-0.002	-0.001
x7	92.5649	12.351	7.495	0.000	68.325	116.805
x8	-0.2581	0.158	-1.633	0.103	-0.568	0.052
x9	-0.8973	0.540	-1.660	0.097	-1.958	0.163
x10	-15.9668	8.200	-1.947	0.052	-32.060	0.127
x11	0.2078	1.712	0.121	0.903	-3.153	3.568
x12	0.0001	1.4e-05	9.294	0.000	0.000	0.000

Omnibus:	1762.652	Durbin-Watson:	1.751
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3850609.492
Skew:	14.258	Prob(JB):	0.00
Kurtosis:	322.528	Cond. No.	4.27e+07







hashtag #sb49

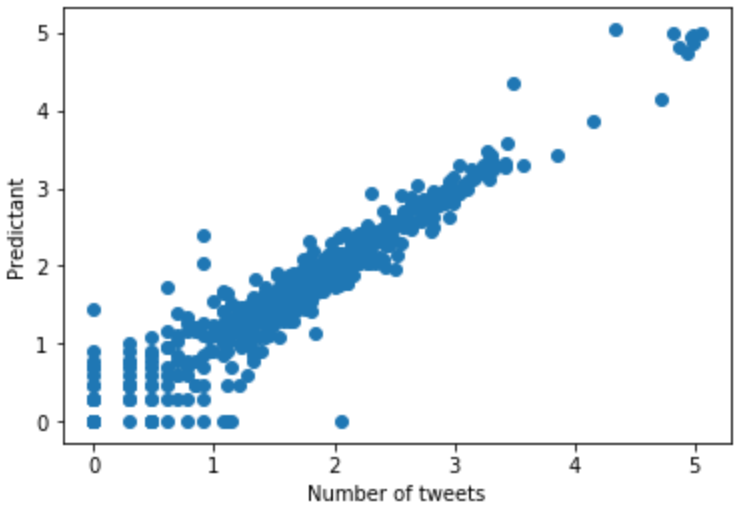
OLS Regression Results

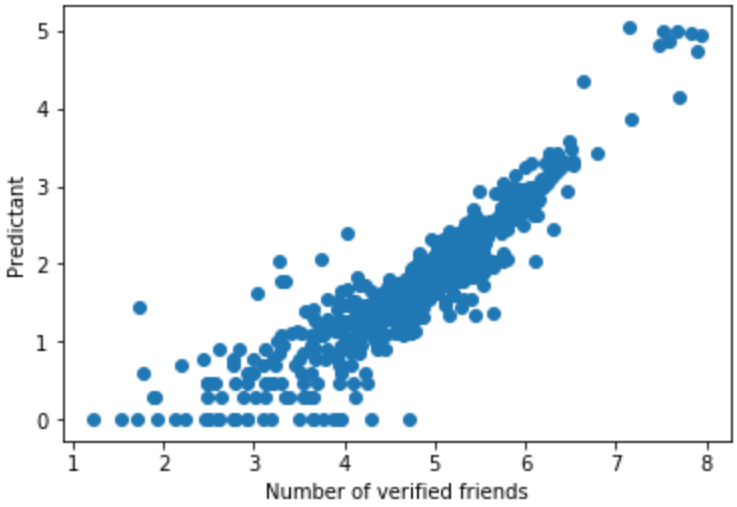
Dep. Variable:	y	R-squared:	0.873
Model:	OLS	Adj. R-squared:	0.871
Method:	Least Squares	F-statistic:	506.9
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	0.00
Time:	17:03:01	Log-Likelihood:	-8437.8
No. Observations:	898	AIC:	1.690e+04
Df Residuals:	885	BIC:	1.696e+04
Df Model:	12		
Covariance Type:	nonrobust		

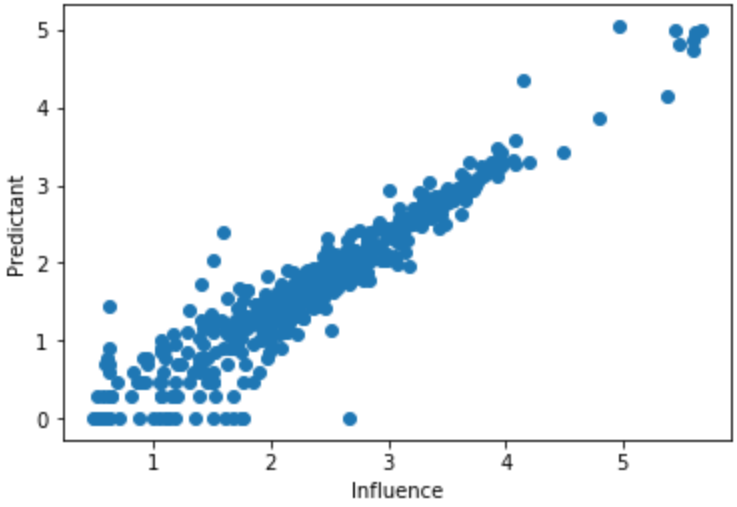
	coef	std err	t	P>|t|	[0.025	0.975]
const	-86.5588	104.879	-0.825	0.409	-292.399	119.282

x1	46.5032	2.318	20.066	0.000	41.955	51.052
x2	0.7570	0.435	1.741	0.082	-0.096	1.610
x3	3.08e-05	1.92e-05	1.605	0.109	-6.87e-06	6.85e-05
x4	-1.7383	0.183	-9.490	0.000	-2.098	-1.379
x5	-0.9375	0.310	-3.022	0.003	-1.546	-0.329
x6	0.0072	0.001	13.819	0.000	0.006	0.008
x7	66.4345	10.525	6.312	0.000	45.778	87.091
x8	1.1330	0.125	9.097	0.000	0.889	1.377
x9	-9.7462	0.497	-19.610	0.000	-10.722	-8.771
x10	-92.0671	7.386	-12.464	0.000	-106.564	-77.570
x11	-6.3864	0.648	-9.852	0.000	-7.659	-5.114
x12	1.903e-05	1.53e-05	1.241	0.215	-1.11e-05	4.91e-05

Omnibus:	1911.919	Durbin-Watson:	1.849
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7003387.069
Skew:	16.996	Prob(JB):	0.00
Kurtosis:	434.297	Cond. No.	1.13e+08







hashtag #superbowl

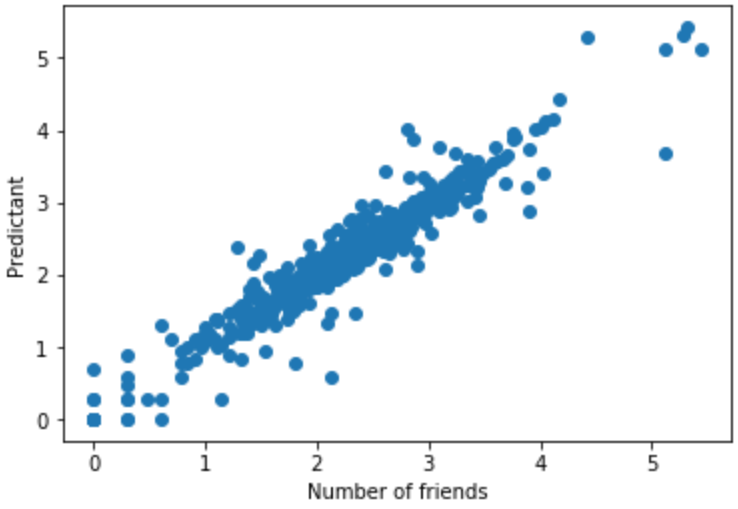
OLS Regression Results

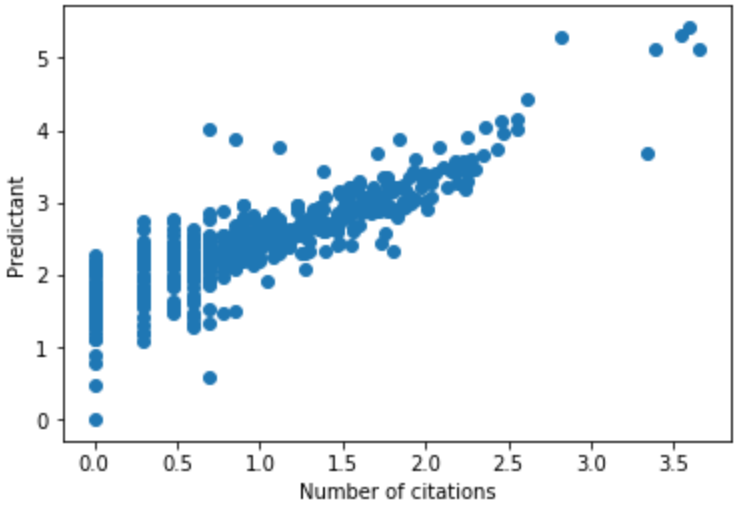
Dep. Variable:	y	R-squared:	0.936
Model:	OLS	Adj. R-squared:	0.935
Method:	Least Squares	F-statistic:	1075.
Date:	Sun, 18 Mar 2018	Prob (F-statistic):	0.00
Time:	17:15:15	Log-Likelihood:	-8647.6
No. Observations:	898	AIC:	1.732e+04
Df Residuals:	885	BIC:	1.738e+04
Df Model:	12		
Covariance Type:	nonrobust		

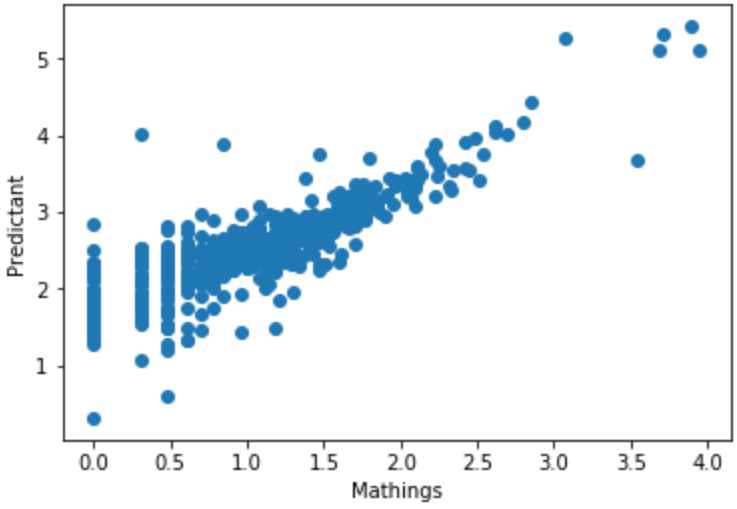
	coef	std err	t	P>|t|	[0.025	0.975]

const	-377.9644	137.067	-2.758	0.006	-646.978	-108.950
x1	99.1156	5.387	18.399	0.000	88.543	109.688
x2	-7.7311	0.964	-8.024	0.000	-9.622	-5.840
x3	-0.0002	2.07e-05	-9.824	0.000	-0.000	-0.000
x4	0.5433	0.229	2.368	0.018	0.093	0.994
x5	5.5847	0.791	7.061	0.000	4.032	7.137
x6	-0.0072	0.000	-16.998	0.000	-0.008	-0.006
x7	233.7714	9.612	24.322	0.000	214.907	252.636
x8	-0.1228	0.125	-0.986	0.324	-0.367	0.122
x9	-14.7935	0.956	-15.473	0.000	-16.670	-12.917
x10	98.5205	4.648	21.195	0.000	89.398	107.643
x11	-34.9757	2.358	-14.835	0.000	-39.603	-30.349
x12	0.0005	3.36e-05	13.737	0.000	0.000	0.001

Omnibus:	1070.960	Durbin-Watson:	1.919
Prob(Omnibus):	0.000	Jarque-Bera (JB):	493824.165
Skew:	5.312	Prob(JB):	0.00
Kurtosis:	117.390	Cond. No.	2.35e+08







Here is a summary of the training accuracy and R-squared measure for different hashtags:

Hashtag	gopatriots	gohawks	nfl	patriots	sb49	superbowl
training accuracy	169.72	533.31	345.27	1975.34	358.12	5821.64
R-squared measure	0.480	0.57	0.60	0.62	0.78	0.78

T-values:

Hashtag	gopatriots	gohawks	nfl	patriots	sb49	superbowl
X1	14.297	2.731	1.795	2.112	20.066	18.399
X2	8.994	1.255	0.573	0.916	1.741	8.024
X3	10.721	6.139	1.893	4.098	1.605	9.824
X4	21.350	4.215	8.781	0.488	9.490	2.368
X5	3.394	0.623	1.903	1.111	3.022	7.061
X6	1.756	2.030	3.603	4.086	13.819	16.998
X7	9.067	3.335	1.338	7.495	6.312	24.322
X8	0.475	2.832	1.862	1.633	9.097	0.986
X9	12.404	1.469	0.335	1.660	19.610	15.473
X10	4.560	3.215	4.149	1.947	12.464	21.195
X11	4.554	1.643	3.791	0.121	9.852	14.835
X12	10.159	5.763	7.954	9.294	1.241	13.737

Discussion:

From training accuracy and R-squared measure table, it shows that adding new features helps us to build a better regression model in most of the cases. And the trends of the error as similar where a more extensive dataset is hard to predict, while predicting a small dataset is more accurate.

On the second table, the top features now are different for each hashtag and we cannot conclude which features are the most important to all hashtags. However, we can reach a result that some of the features are less important than others, such as X2, X5, X8 and X11, and none of them appears to be the top three features in all the hashtags.

Last but not least, we draw a scatter plot of predictant versus the value of top three features, and we get a relatively linear relationship between our prime elements and our target values. Thus, we can conclude that our features are well designed.

Problem 1.4

In this part, for each hashtag, we train three regression models for three different time periods, and we perform 10-fold cross validation for the three different models for each hashtag. We define the three time periods as the following:

```
Time1: Before Feb. 1, 8:00 a.m.
Time2: Between Feb. 1, 8:00 a.m. and 8:00 p.m.
Time3: After Feb. 1, 8:00 p.m.
```

Then, we aggregate the data for all hashtags and train three models for the time intervals mentioned above to see how the prediction is after we combine all the hashtags and aggregate the data. Again, the same evaluations are performed as before for the combined model, which is then compared with models trained for individual hashtag.

Here is a summary of the results for three models for each hashtag: (1) linear regression, (2) RandomForest and (3) K nearest neighbor:

#gohawks

Model	**Linear**	**RandomForest**	**KNeighbors**
Time1	1173.4968218378565	634.0600558475293	804.2340663910537
Time2	77499.57334265517	4087.8909606299435	3581.8054869018224
Time3	582.4558967268821	78.15511730433715	102.2995725233568

#gopatriots

Model	**Linear**	**RandomForest**	**KNeighbors**
Time1	49.13674446518472	49.61253820471025	58.23872267534049
Time2	65794.42955988712	958.1497769138185	1249.5188249882433
Time3	9.10416922951089	9.93727765915867	11.699920008492716

#nfl

Model	**Linear**	**RandomForest**	**KNeighbors**
Time1	241.16188169003343	229.06756420464734	279.636909518102
Time2	26189.972048220006	2839.493118850616	3356.9990125110257
Time3	157.91836699830068	202.16810522243802	201.61990967665366

#patriots

Model	**Linear**	**RandomForest**	**KNeighbors**

Time1	610.7295029502072	671.3943943149602	691.4366702216025
Time2	136601.2215009904	19936.486211667292	17696.35031067141
Time3	212.84589902429693	253.6703831465437	290.1033621027111

#sb49

Model	**Linear**	**RandomForest**	**KNeighbors**
Time1	107.44757759610891	135.75005292223975	165.44717293996484
Time2	166621.5307068395	37056.226089282216	42303.2171929039
Time3	346.7104236331856	331.076746456099	467.98786154071405

#superbowl

Model	**Linear**	**RandomForest**	**KNeighbors**
Time1	645.0439503201999	698.2976925803789	766.9877367914282
Time2	165518.8950515691	72556.44733851707	85425.4580802702
Time3	593.4084238528479	604.0354304068704	654.9769957205191

#aggregate the data of all hashtags

Model	**Linear**	**RandomForest**	**KNeighbors**
Time1	2068.4410773514073	1885.6833750875996	2489.7495280306293
Time2	3880609.9905768223	101039.95485940698	118672.92632181107
Time3	1091.064020170654	729.5184218199876	832.6019908352712

Discussion:

In above table, we highlight the cell in each row to have a better look on which model is the best. In general, random forest yields a better result in most of the cases, especially for period Time2 during which all the hashtags contain a denser data. In general, Time2 is the hardest to predict as we discuss before since it has the most extensive dataset. Last, we can also see that if we aggregate the data of all hashtags, it yields the worst results no matter what models we used, which is what we expected.

Problem 1.5

In this part, we aim to apply our trained model to a test dataset. We find that the Random Forest Model has the best performance for fitting our data in problem 1.4. Therefore, we use the same model to predict the test data to get the best result. The results of predicted and true values for each test file are shown as:

Results:

```
sample1_period1:
predict value = 507.8646761617787
true value = 177.0

sample4_period1:
predict value = 560.6267008397111
true value = 201.0

sample5_period1:
predict value = 484.0790834812191
true value = 215.0

sample8_period1:
predict value = 345.4584684684563
true value = 11.0

sample2_period2:
predict value = 249004.2
true value = 83440.0

sample6_period2:
predict value = 175661.4
true value = 37199.0

sample9_period2:
predict value = 237987.2
true value = 2788.0

sample3_period3:
predict value = 1296.6450157475554
true value = 523.0

sample7_period3:
predict value = 121.20988042599456
true value = 120.0

sample10_period3:
predict value = 97.14126556112969
true value = 61.0
```

Discussion:

From the above results, we can observe that most of the predictions are not very accurate except the last two cases for period 3. We have tried different methods to improve the results, but we are not able to have much gain. We discuss this problem within our group members as well as other teams, and it seems that it is a common problem to most of the groups. Here is our explanation why this prediction doesn't work well.

First of all, a deviation may occur due to the personal selection of features. The result may be improved if we discover more features other than what we have right now. Secondly, as required in the documents, we fit a model on the aggregate of the training data for all hashtags and predict the number of tweets in the next hour for each test file. We have discussed in 1.4 that aggregate of the training data for all hashtags would yield terrible results due to the massive density of the data, which would be the main reason that causes the deviation of the prediction. After all, in this problem, we have learnt that it is not easy to perform popularity prediction on a large scale since a more extensive dataset would have a significant deviation.

III. Part 2: Fan Base Prediction

In this part, we are asked to train a binary classifier to predict the location of the author of a tweet with at least three different classification algorithms. In our implementation, we choose and run the Logistic Regression, Naive Bayes Classifier and Passive Aggressive. We extract the location of each tweeter, and we assign 0 if the place is Washington, assign 1 if the area is Massachusetts and split the dataset into 80 percent for training and 20 percent for the test.

To be specific, we first segregate the tweets based on location. Location, in this case, is a type of metadata. After classifying the tweet data based on metadata, we could observe better results. Tweets containing the following substrings in the location field are taken for Washington. They are:

- Seattle, Washington
- Seattle, WA
- Seattle
- Washington
- WA
- Kirkland, Washington
- Kirkland, WA

Note: “Washington DC” was excluded from the list.

Tweets containing the following substrings in the location field are taken for "Massachusetts", and they are:

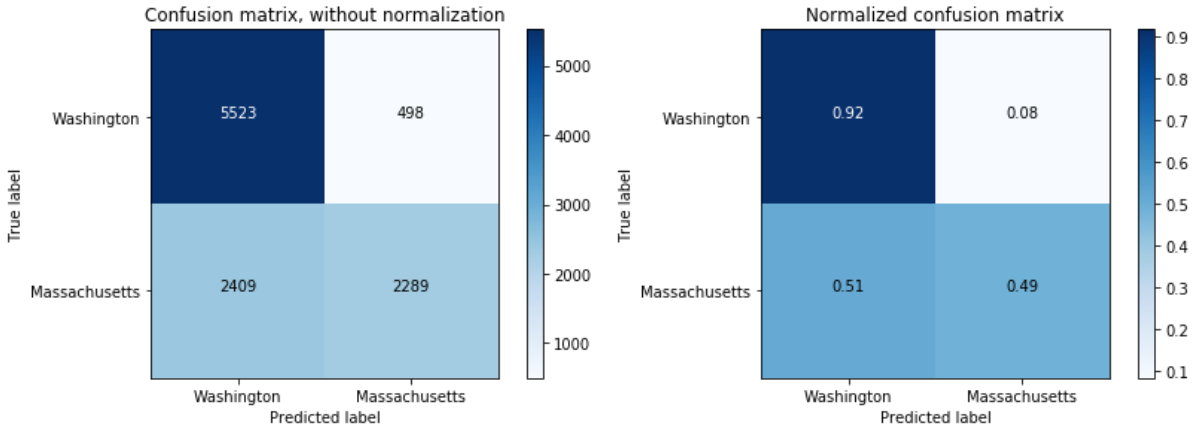
- Boston
- Boston, MA
- Boston, Massachusetts
- Worcester
- Worcester, MA
- Worcester, Massachusetts
- Massachusetts
- MA

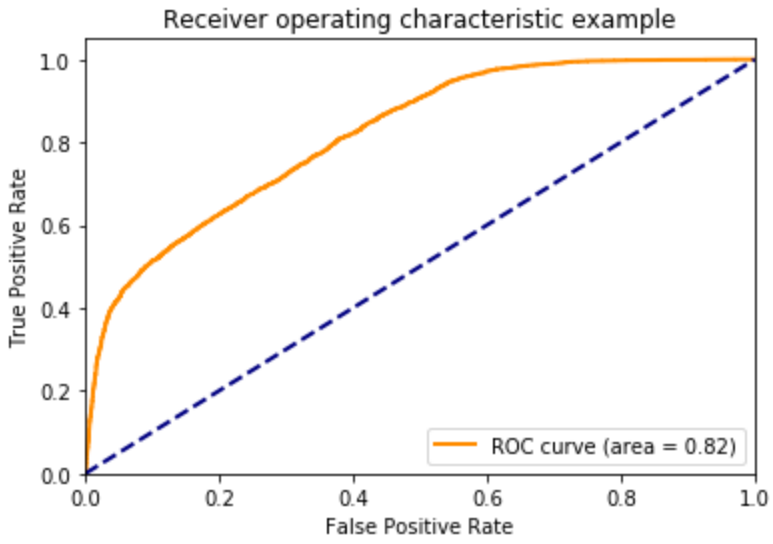
Here is the summary of the results for three models:

Results:

Logistic Regression:

```
precision = 0.8213132400430571
accuracy = 0.7287993282955499
recall = 0.4872286079182631
```



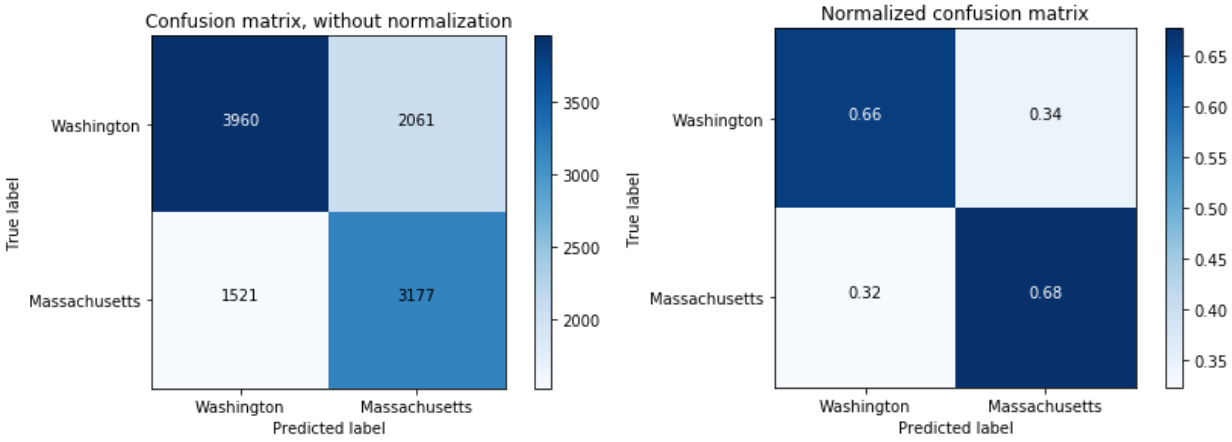


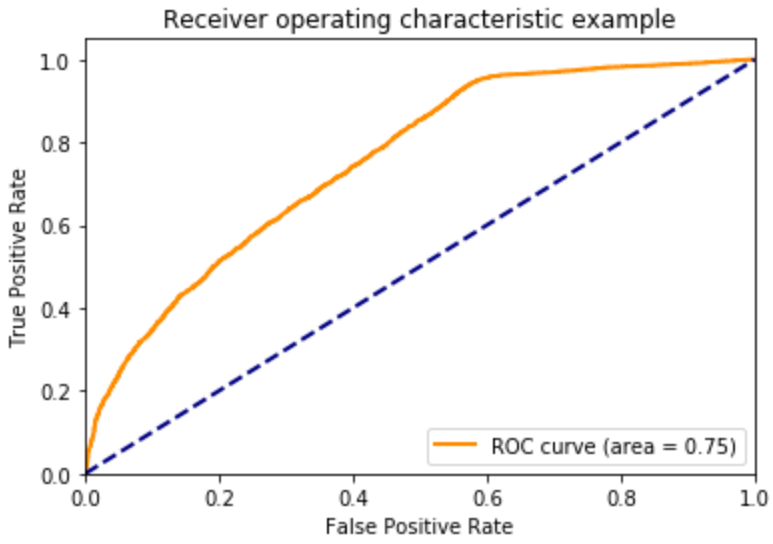
Naive Bayes:

precision = 0.6065292096219931

accuracy = 0.6658270361041142

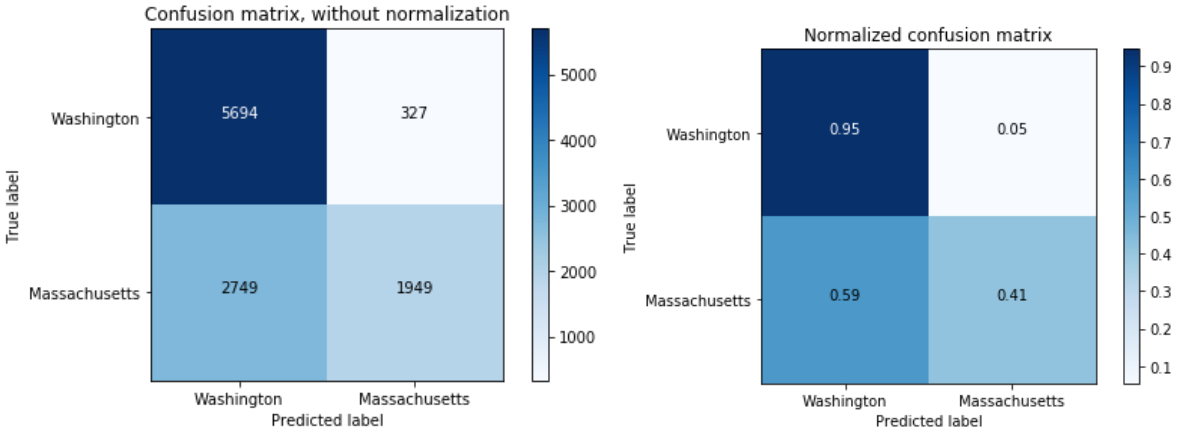
recall = 0.6762452107279694

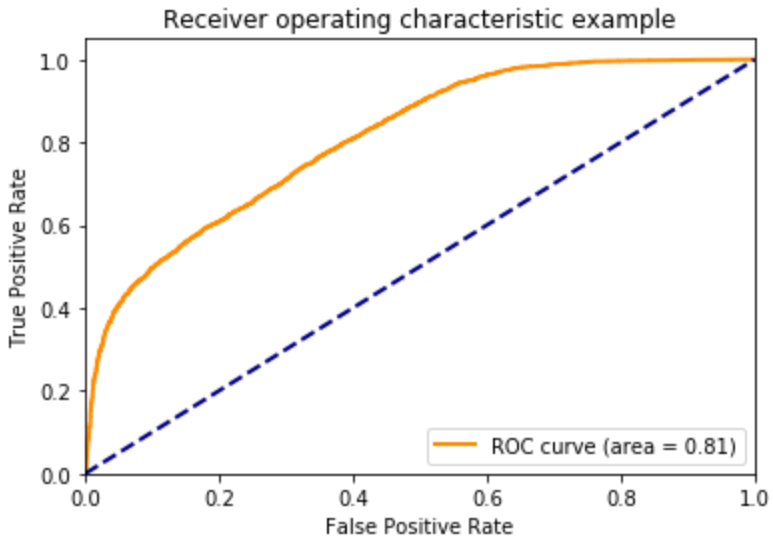




PassiveAggressive:

precision = 0.8563268892794376
accuracy = 0.713032932176509
recall = 0.41485738612175393





Discussion:

From the results above, we have found that Passive Aggressive has the highest precision which is 0.86, Logistic Regression has the most top accuracy which is 0.73, and Naive Bayes Classifier has the highest recall which is 0.676.

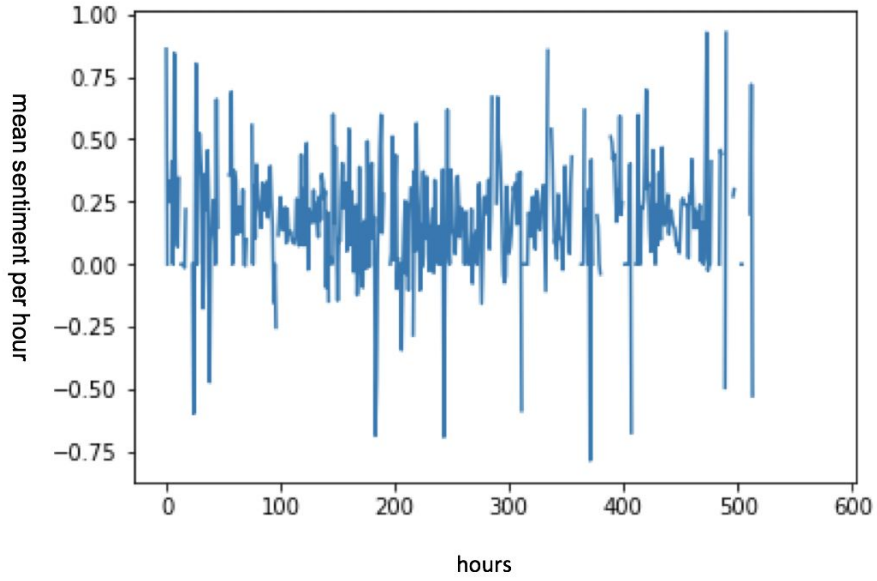
Overall, we think Logistic Regression is the best classifier algorithm to predict the location of the author of a tweet since the ROC curve has the largest area.

IV. Part 3: Define Your Own Project

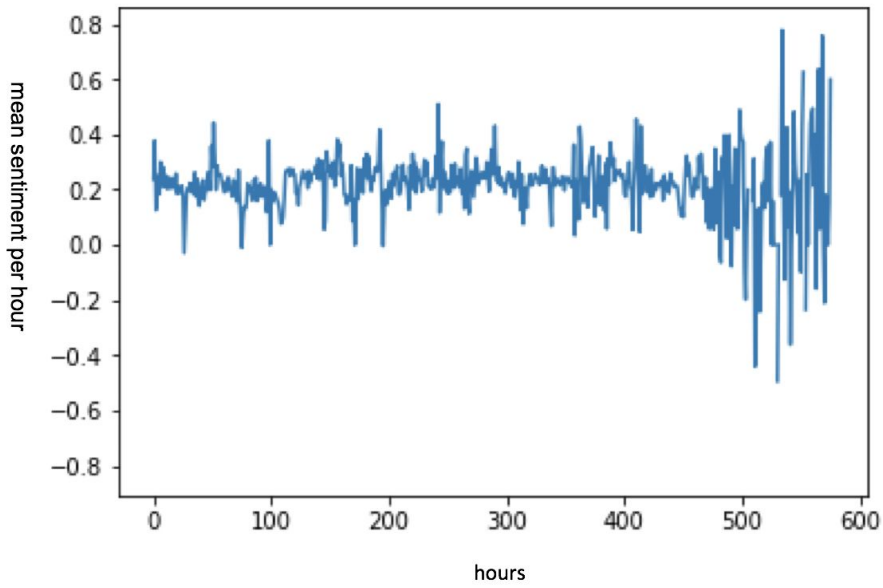
In this part, we have defined a new problem as how to use sentiments of tweets to predict the result of the game.

In the super bowl game of patriots against hawks, we think that supporters of the winning team would have higher sentiments and lower sentiment fluctuation. We calculate the mean sentiments per hour for both #gopatriots and #gohawks hashtags and then plot their values against hours as shown:

#gopatriots:



#gohawks:



Discussion:

We find that sentiments of hashtag #gopatriots has large fluctuation while sentiments of hashtag #gohawks has very small fluctuation except for the last several hours. The large fluctuation of #gopatriots may due to the small amount of tweeters during that period, but the large fluctuation of #gohawks at last several hours may indicate that hawks leads for a long time but finally loses the game.