EE-219 Project 5

Project 5: Popularity Prediction on Twitter

March 19, 2018

Ashish Shah (804946005)
Ayush Dattagupta (305024749)
Shrey Agarwal (004943082)
Varun Saboo (505028591)

Objective

In this project, we analyze a twitter data set and explore the activity on Twitter. The main objective behind this project was to use different regression techniques and models to predict the popularity of topics on Twitter. Specifically, we wanted to know the activity of different hashtags on Twitter. Adn from that, we tried to predict the activity of that hashtag in the future. We explored various features from the tweets such as number of followers, timestamp, number of tweets etc to find the features of most significance for the task of predictions of hashtag activity.

Dataset

Twitter data is collected by querying popular hashtags related to the 2015 Super Bowl spanning a period starting from 2 weeks before the game to a week after the game. The dataset was grouped in 6 hashtags which were, gohawks, nfl, sb49, gopatriots, patriots and superbowl. Each of these tweet data is stored in a separate file. Given this dataset, our objective was to predict the popularity of each hashtag in the future.

Part 1 - Popularity Prediction

Problem 1.1
Plot “number of tweets in hour” over time for #SuperBowl and #NFL (a histogram with 1-hour bins).

The objective of this part was to analyze the initial twitter dataset in order to calculate the statistics mentioned. This was done for each hashtag. The following statistics were computed.

- Average number of tweets per hour
- Average number of followers of users posting the tweets
- Average number of retweets.

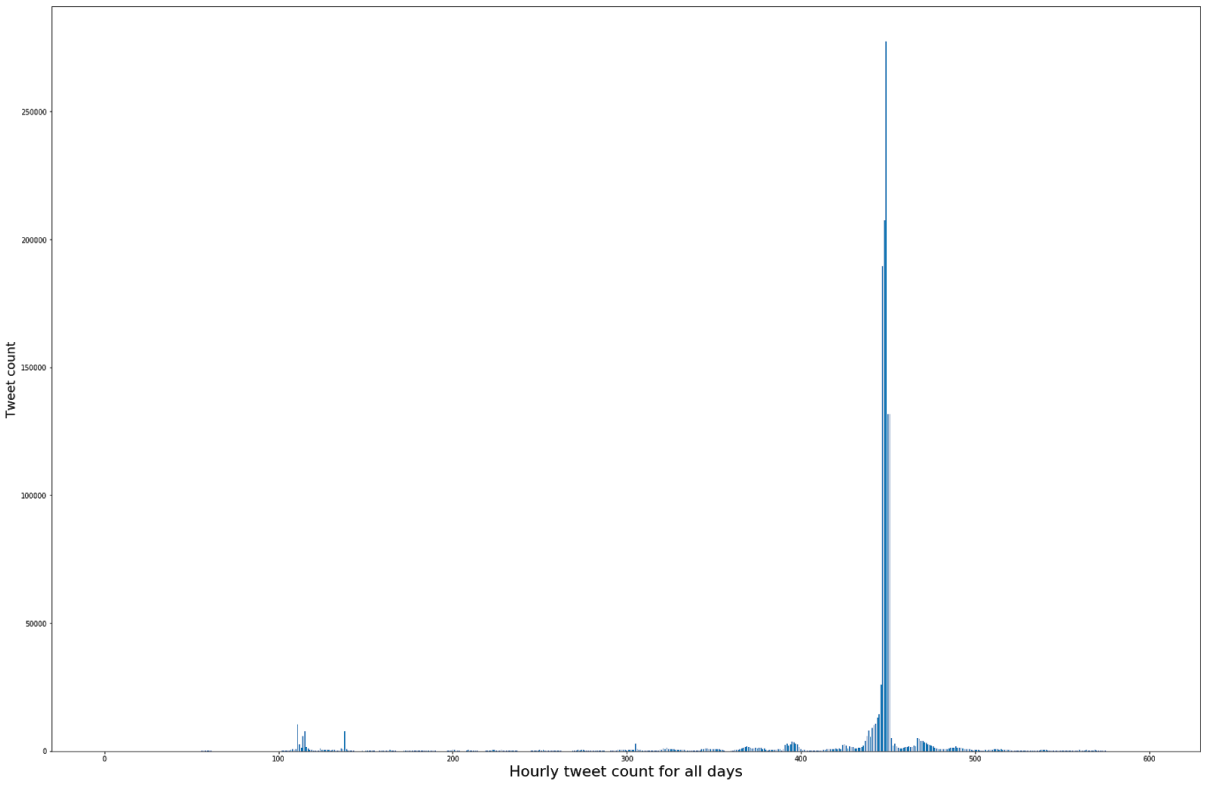
In the dataset, each tweet is described by a JSON string. We created an initial Panda dataframe object to store the tweet count, timestamp of tweet and retweet count for all the hashtags. After doing this, we grouped the data into hours to calculate the statistics. The following code represents the way we calculated each of them.

- Tweet_count += 1
- Date = Tweet_data['citation_date']
- Retweet_count = tweet_data['metrics']['citations']['total']
- Follower_count = tweet_data[‘author’][‘followers’]
- NOTE: For Average number of followers per user according to the answer given on piazza by the instructor we DO NOT follow the unique authors scheme and therefore it is assumed that tweets posted by the same author are considered to be different authors. (Therefore the average number of followers per tweet is equivalent to the average number of followers per tweet)

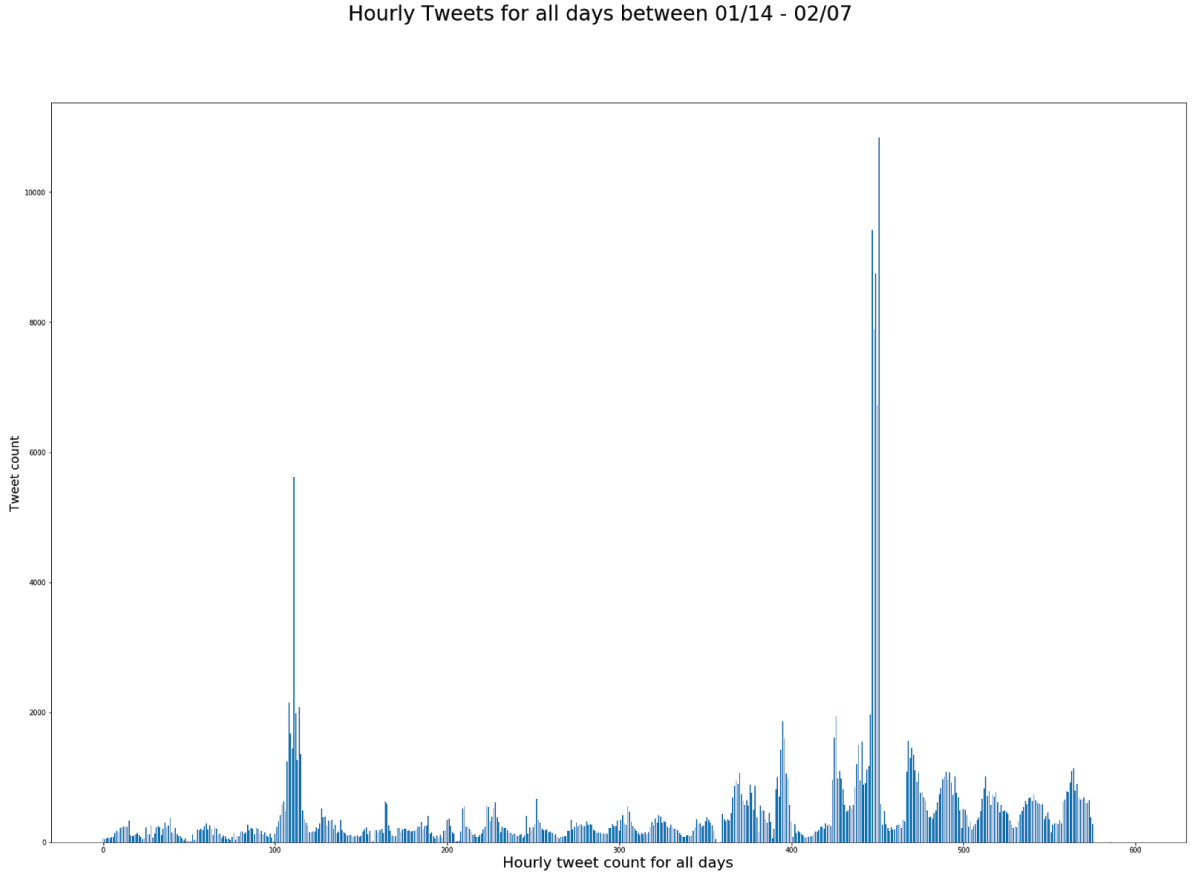
Hashtag	Average Number of tweets per hour	Average number of followers per tweet	Average number of retweets per tweet
#gohawks	325.4214	2203.931767	2.014617
#nfl	441.262	4653.252286	1.538533
#sb49	1417.3252	10267.316849	2.511149
#gopatriots	46.3712	1401.895509	1.400084
#patriots	835.3255	3309.978828	1.782816
#superbowl	2298.3281	8858.974663	2.388272

With the statistics above, we plotted a graph for the number of tweets in an hour for the two hashtags, #superbowl and #nfl.

Q1.1 Plotting the number of tweets per hour across all days of data with the **#superbowl**



Q1.1 Plotting the number of tweets per hour across all days of data with the **#nfl**



Observation:

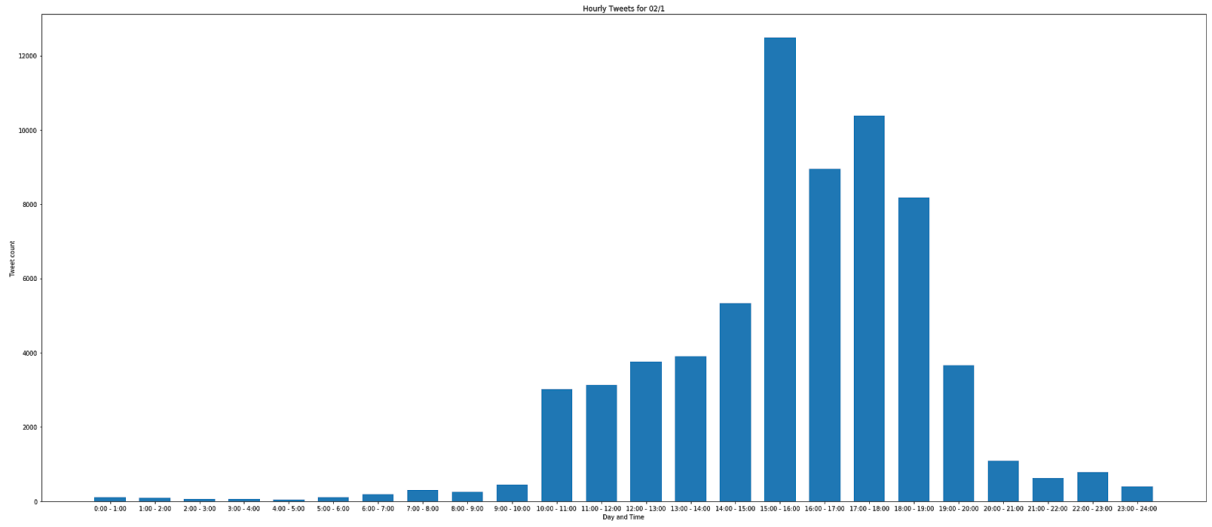
We can infer from above that the hashtag #superbowl was the highest trending hashtag out of all the 6. We confirm this from the graphs as well that there was a huge spike in the graph around the hour 450 for the #superbowl and #nfl. We see that this was superbowl day and hence there was this huge a spike in the hashtags on that day. We see another spike in the graph of #nfl which tells us another event which lead to the spike on the graph.

Some Interesting observations from the tweet count statistics of #goHawks and #patriots and it's trends with the actual game score!!

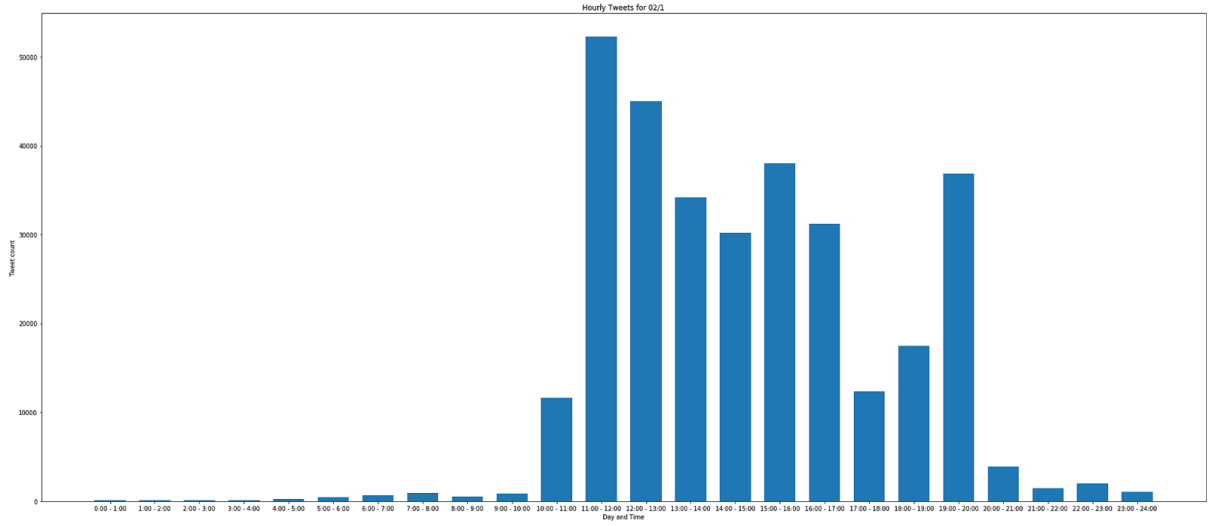
(Not asked as a part of the question)

We also plotted the tweet count graph for 1st Febuary (SUPERBOWL day) on an hourly basis for the two team hashtags #goHawks and #patriots here are the graphs

#goHawks



#goPatriots



If we analyze the general trend of increase and decrease in number of tweets for each time, it actually gives us a good sense of who was **winning** the game. As we can see above from the period between 15:00(start of the game) and 20:00 (end of the game). Initially both the teams have high tweet count as the game is just starting. In the first two quarter the game was tied and there is a similar trend with the tweet count decreasing slightly for both the teams over the first two quarters. In the third quarter the Hawks took the lead and if we look at the graph there is a sudden increase in tweet count for the hawks and a decrease in tweet count for the patriots from 17:00 - 19:00. In the last quarter the patriots equalized and eventually won the game in the last few moments and corresponding to that from 19:00 - 20:00 we can see a huge increase in patriots tweets and a huge decrease in hawks tweets. Therefore following this pattern and observing trends in the tweet counts for each team can also shed some light on which team was winning during any given quarter of the game!!

Problem 1.2

For each of your models, report your model's training accuracy and R-squared measure. Also, analyse the significance of each feature using the t-test and P-value. You may use the library statsmodels.api in Python.

Approach: For this part we train a linear regression model for each hashtag based on the following five features:

- Total number of tweets
- Total number of retweets
- Sum of Followers of users (tweets) posting the tweet.

(NOTE: As discussed in part 1.1 two tweets posted by the same author are considered to be different authors and therefore the metric really is sum of followers of tweets.)

- Maximum number of followers of a user(tweet) posting the tweet.
- Time of the day as an hour value from 0 - 23 (representing 23 hours in a day)

(NOTE: Using one hot encoding is a better approach to tackle the time of day feature as opposed to having a single numerical value as the feature with later time of day has more weight as opposed to feature of less time of day. This one - hot encoding approach has been used in part 1.3 as this part did not mention using the following approach)

The data for each hashtag was broken up into bins of 1 hour each starting from 14th January 12:00 A.M extending upto 7th february 11:59 P.M The date range was obtained using the **CITATION_DATE** parameter and for simplicity the calculations were done till the end of the last day i.e 11:59 pm rather than stopping at last hour of the day for which the data was recorded.

Using these 5 features mentioned above for a particular hour, we predict the number of tweets in the next hour. I.e to predict the number of tweets on 14th January from 1:00 A.M to 1:59 A.M we use the attributes from the time frame 14th January 12:00 AM to 12:59 A.M as the feature vector for the predictions.

A linear regression model was fit with the following data and a T test was also conducted that gave resultant P-values for each feature. Using the data the RMSE as well as the importance of each of the features was computed and the top 3 most important features for each model are listed along with the R- squared value.

Question.

How to calculate significance of each feature using t-test and P-value?

When a predictor has a lower p-value, it means it is a valid and useful addition to our model since the changes and modifications in the predictor's value is directly related to the changes in the response variable. In the opposite case, when the predictor has a larger p-value, it tells us that it is not useful and any change in the predictor is not associated with the changes in the response.

The p-value can be interpreted as "the probability that this coefficient is actually 0 (while its estimator that we calculated might not be 0)".

R^2/Hashtag	#gohawks	#gopatriots	#NFL	#patriots	#sb49	#superbowl
R^2	0.474	0.632	0.566	0.670	0.805	0.802
RMSE	39127.487187	9338.701157	26759.235817	137554.168633	328104.893002	584906.332891

Results for #superbowl

```
 The regression results for the tweets with the hashtag #superbowl are
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.802
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     481.1
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          5.70e-206
Time:                        14:04:51   Log-Likelihood:                -6227.0
No. Observations:                 599   AIC:                         1.247e+04
Df Residuals:                     593   BIC:                         1.249e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          4.099e+04   1401.316     29.248      0.000    3.82e+04    4.37e+04
x2         -1.358e+04   1670.559     -8.132      0.000   -1.69e+04   -1.03e+04
x3          -1.57e+04   2213.603     -7.091      0.000      -2e+04   -1.13e+04
x4          2914.5314    539.507      5.402      0.000    1854.955    3974.108
x5          -157.3469    326.680     -0.482      0.630    -798.937     484.243
const       2251.6845    325.104      6.926      0.000    1613.190    2890.179
==============================================================================
Omnibus:                     1038.271   Durbin-Watson:                   2.316
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1958959.991
Skew:                          10.210   Prob(JB):                         0.00
Kurtosis:                     282.414   Cond. No.                         15.3
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Best features selected are:

tweetCount

retweetCount

followerCount
```

Best features are:

- No. of Tweets
- No. of Retweet

- No of followers

Results for #sb49

```
 The regression results for the tweets with the hashtag #sb49 are
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.805
Model:                            OLS   Adj. R-squared:                  0.803
Method:                 Least Squares   F-statistic:                     489.5
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          9.03e-208
Time:                        14:04:54   Log-Likelihood:                -5876.1
No. Observations:                 599   AIC:                         1.176e+04
Df Residuals:                     593   BIC:                         1.179e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          1.185e+04    936.078     12.662      0.000       1e+04    1.37e+04
x2         -4751.8030   1920.939     -2.474      0.014   -8524.474    -979.132
x3          1586.3879   1181.489      1.343      0.180    -734.024    3906.800
x4           431.2660    212.338      2.031      0.043      14.240     848.292
x5          -119.2228    182.097     -0.655      0.513    -476.856     238.411
const       1380.5526    180.958      7.629      0.000    1025.157    1735.949
==============================================================================
Omnibus:                     1223.720   Durbin-Watson:                   1.683
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2444265.127
Skew:                          14.899   Prob(JB):                         0.00
Kurtosis:                     314.522   Cond. No.                         23.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Best features selected are:

tweetCount

retweetCount

maxFollowers


RMSE : 328104.893002
```

Best features are:

- No. of Tweets
- No. of Retweet
- Max no. of Followers

Results for #patriot

```
 The regression results for the tweets with the hashtag #patriots are
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.670
Model:                            OLS   Adj. R-squared:                  0.667
Method:                 Least Squares   F-statistic:                     240.7
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          3.95e-140
Time:                        14:04:56   Log-Likelihood:                -5536.2
No. Observations:                 599   AIC:                         1.108e+04
Df Residuals:                     593   BIC:                         1.111e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          4004.3949    307.745     13.012      0.000    3399.991    4608.799
x2          -524.9808    350.224     -1.499      0.134   -1212.811     162.850
x3            -3.3114    224.492     -0.015      0.988    -444.208     437.585
x4           211.0105    128.642      1.640      0.101     -41.638     463.659
x5           -48.6896    103.504     -0.470      0.638    -251.969     154.590
const        817.5125    102.611      7.967      0.000     615.988    1019.037
==============================================================================
Omnibus:                      905.196   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           741809.555
Skew:                           7.899   Prob(JB):                         0.00
Kurtosis:                     174.675   Cond. No.                         7.48
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Best features selected are:

tweetCount

maxFollowers

retweetCount


RMSE : 137554.168633
```

Best features are:

- No. of Tweets

- No. of Retweet
- Max no. of Followers

Results for #gopatriot

```
 The regression results for the tweets with the hashtag #gopatriots are
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.632
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     203.9
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          3.06e-126
Time:                        14:04:57   Log-Likelihood:                -3964.3
No. Observations:                 599   AIC:                             7941.
Df Residuals:                     593   BIC:                             7967.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -24.1604     74.650     -0.324      0.746    -170.771     122.450
x2           194.2410     82.905      2.343      0.019      31.419     357.063
x3            96.0770     76.482      1.256      0.210     -54.131     246.285
x4           -56.7812     29.244     -1.942      0.053    -114.217       0.654
x5            -0.9127      7.499     -0.122      0.903     -15.640      13.815
const         43.7930      7.438      5.888      0.000      29.185      58.401
==============================================================================
Omnibus:                      539.270   Durbin-Watson:                   1.953
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           342835.077
Skew:                           2.849   Prob(JB):                         0.00
Kurtosis:                     120.063   Cond. No.                         25.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Best features selected are:

retweetCount

maxFollowers

followerCount


RMSE : 9338.701157
```

Best features are:

- Max no. of followers
- No. of Retweet
- No. of Followers

Results for #gohawk

```
 The regression results for the tweets with the hashtag #gohawks are
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.474
Model:                            OLS   Adj. R-squared:                  0.469
Method:                 Least Squares   F-statistic:                     106.7
Date:                Mon, 19 Mar 2018   Prob (F-statistic):           3.10e-80
Time:                        14:04:57   Log-Likelihood:                -4960.5
No. Observations:                 599   AIC:                             9933.
Df Residuals:                     593   BIC:                             9959.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1          1623.4544    219.550      7.394      0.000    1192.265    2054.644
x2          -301.6224    102.974     -2.929      0.004    -503.861     -99.384
x3          -479.0853    225.903     -2.121      0.034    -922.753     -35.418
x4            10.9198     72.959      0.150      0.881    -132.369     154.209
x5            13.0976     39.833      0.329      0.742     -65.133      91.328
const        314.0334     39.242      8.002      0.000     236.963     391.104
==============================================================================
Omnibus:                      956.244   Durbin-Watson:                   2.221
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           878638.612
Skew:                           8.829   Prob(JB):                         0.00
Kurtosis:                     189.795   Cond. No.                         14.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Best features selected are:

tweetCount

retweetCount

followerCount


RMSE : 39127.487187
```

Best features are:

- No. of Tweets

- No. of Retweet
- No. of Followers

Results for #nfl

```
 The regression results for the tweets with the hashtag #nfl are
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.566
Model:                            OLS   Adj. R-squared:                  0.563
Method:                 Least Squares   F-statistic:                     154.8
Date:                Mon, 19 Mar 2018   Prob (F-statistic):          4.62e-105
Time:                        14:04:59   Log-Likelihood:                -4656.6
No. Observations:                 599   AIC:                             9325.
Df Residuals:                     593   BIC:                             9352.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           603.8061    115.685      5.219      0.000     376.603     831.009
x2          -247.2031     93.645     -2.640      0.009    -431.119     -63.287
x3           347.6311    105.925      3.282      0.001     139.597     555.665
x4          -110.5336     44.863     -2.464      0.014    -198.643     -22.424
x5            -2.2889     23.898     -0.096      0.924     -49.225      44.647
const        432.3289     23.628     18.297      0.000     385.924     478.734
==============================================================================
Omnibus:                      630.671   Durbin-Watson:                   2.334
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           378483.016
Skew:                           3.881   Prob(JB):                         0.00
Kurtosis:                     125.900   Cond. No.                         11.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Best features selected are:

tweetCount

followerCount

retweetCount


RMSE : 26759.235817
```

Best features are:

- No. of Tweets

- No. of Retweet
- No. of Followers

Problem 1.3

For each of the top 3 features in your measurements, draw a scatter plot of predicted (number of tweets for next hour) versus value of that feature, using all the samples you have extracted, and analyze it.

For this question, we added 8 new features/attributes to our above model for better accuracy and regression model.

We added the following attributes.

- longTweet = number of tweets that are longer than 90 characters per hour interval
- impressionCount = sum of the impression count per hour
- rankingScore = sum of ranking scores of all tweets in an hour
- favoriteCount = sum of the favorite count per hour
- userID = number of unique users tweeting
- hashTag = sum of hashtags used in tweets per hour
- Url = sum of urls used in tweets per hour
- Mentions = sum of mentions of tweets per hour

We also modified the following attribute:

- Time of day = instead of having a single feature taking values from 0-23 where higher hours equal higher weight for the feature we performed a one hot encoding for the different possible hours where there are 24 values and a 1 at that index corresponds the the current hour and the rest are all 0.

After adding the above features to our feature set, we applied the above linear regression model.
We then perform t-test and use P-values to determine the top 3 attributes for each hashtag. We are comparing the values from parts 1.2 and 1.3. We can also see the plotted graphs of predicted vs value of the top 3 features for each hashtag for all extracted samples:

We can see that adding these new features improves the R^2 values for each hashtag.

R^2/Hashtag	#gohawks	#gopatriots	#NFL	#patriots	#sb49	#superbowl

R^2 from 1.2	0.474	0.632	0.566	0.670	0.805	0.802
R^2 from 1.3	0.694	0.915	0.780	0.820	0.882	0.905

From the above comparison, we can clearly see the increase in accuracy due to the addition of the attributes in our features. We can reason this by observing that the features are well defined and well distributed for the entirety of the match. They are not sparse.

We can also see the increase in accuracy can be because of this additional information that the model now has. By adding more features, the prediction can be done better. Hence we see a significant increase in accuracy for #gopatriots, #gohawks, #nfl and #patriots.

We have also plotted scatter plots for the top features of each hashtag in order to get a more visual inference of the prediction.

Results for #superbowl

The regression results for the tweets with the hashtag #superbowl are

OLS Regression Results

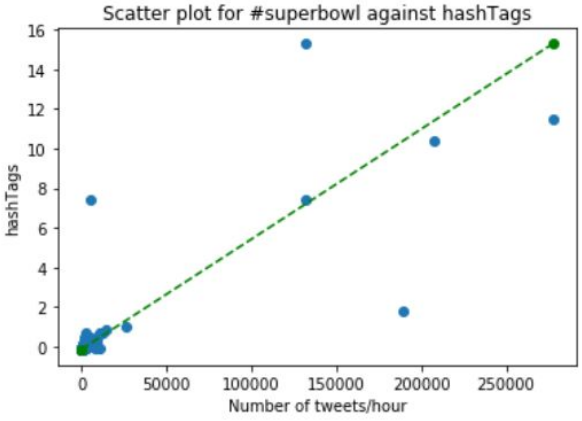
Dep. Variable:	y	R-squared:	0.905
Model:	OLS	Adj. R-squared:	0.899
Method:	Least Squares	F-statistic:	152.9
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	2.96e-262
Time:	14:39:52	Log-Likelihood:	-6008.0
No. Observations:	599	AIC:	1.209e+04
Df Residuals:	563	BIC:	1.225e+04
Df Model:	35		
Covariance Type:	nonrobust		

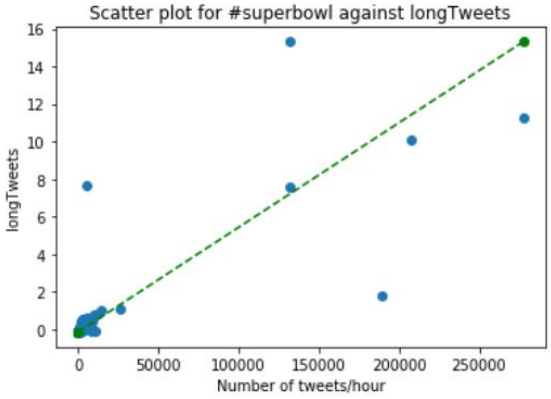
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-2.14e+05	1.29e+05	-1.654	0.099	-4.68e+05	4.01e+04
x2	-2.009e+04	4952.639	-4.056	0.000	-2.98e+04	-1.04e+04
x3	8256.7160	2.42e+04	0.341	0.733	-3.93e+04	5.58e+04
x4	-394.3228	455.800	-0.865	0.387	-1289.599	500.954
x5	50.6323	223.317	0.227	0.821	-388.005	489.269
x6	17.4926	222.842	0.078	0.937	-420.211	455.196
x7	17.8787	223.460	0.080	0.936	-421.039	456.797
x8	129.8947	224.066	0.580	0.562	-310.212	570.001
x9	38.9916	223.047	0.175	0.861	-399.114	477.098
x10	79.0765	222.385	0.356	0.722	-357.729	515.882
x11	30.9468	222.282	0.139	0.889	-405.657	467.551
x12	1.0727	223.487	0.005	0.996	-437.898	440.043
x13	42.9299	223.833	0.192	0.848	-396.720	482.580
x14	-38.8344	223.707	-0.174	0.862	-478.238	400.569
x15	-126.2612	223.585	-0.565	0.572	-565.425	312.902
x16	-103.8532	224.338	-0.463	0.644	-544.495	336.788
x17	-191.8663	224.173	-0.856	0.392	-632.185	248.452
x18	5.8632	224.077	0.026	0.979	-434.267	445.993
x19	1033.4911	222.926	4.636	0.000	595.623	1471.359
x20	-389.1347	227.236	-1.712	0.087	-835.469	57.199
x21	-71.1541	227.722	-0.312	0.755	-518.442	376.134
x22	97.4711	228.088	0.427	0.669	-350.536	545.478
x23	-113.1903	228.963	-0.494	0.621	-562.916	336.535
x24	-248.9572	231.386	-1.076	0.282	-703.442	205.528
x25	-108.8243	223.905	-0.486	0.627	-548.615	330.966
x26	-50.9688	222.818	-0.229	0.819	-488.624	386.687
x27	-89.4260	224.019	-0.399	0.690	-529.441	350.589
x28	-13.5327	223.331	-0.061	0.952	-452.197	425.131
x29	-1.067e+04	2.46e+04	-0.434	0.665	-5.9e+04	3.77e+04
x30	1.419e+05	1.22e+05	1.164	0.245	-9.77e+04	3.82e+05
x31	2.002e+05	1.95e+04	10.240	0.000	1.62e+05	2.39e+05
x32	-1213.5027	985.425	-1.231	0.219	-3149.062	722.057
x33	4.032e+04	1.1e+04	3.662	0.000	1.87e+04	6.19e+04
x34	1.229e+04	3275.367	3.753	0.000	5858.990	1.87e+04
x35	-1.959e+04	9026.179	-2.170	0.030	-3.73e+04	-1862.175
x36	-1.233e+05	1.77e+04	-6.970	0.000	-1.58e+05	-8.85e+04
const	2251.6845	231.487	9.727	0.000	1797.002	2706.367

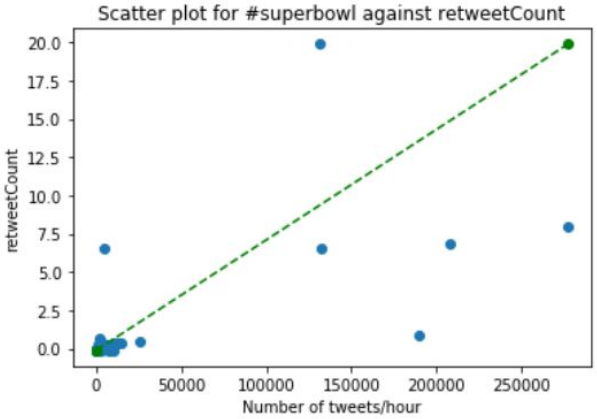
Omnibus:	1043.453	Durbin-Watson:	2.050
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1435883.895
Skew:	10.482	Prob(JB):	0.00
Kurtosis:	241.939	Cond. No.	6.23e+15

Best Features for #superbowl is:

- hashTags
- longTweets
- retweetCount







RMSE : 601320.945760

There is a small clustered region with a very small linear deviation for the majority of the points. The t- test suggests that these features are especially important and we can see that they are indeed correlated well with the predicted values.

Results for #sb49

The regression results for the tweets with the hashtag #sb49 are

OLS Regression Results

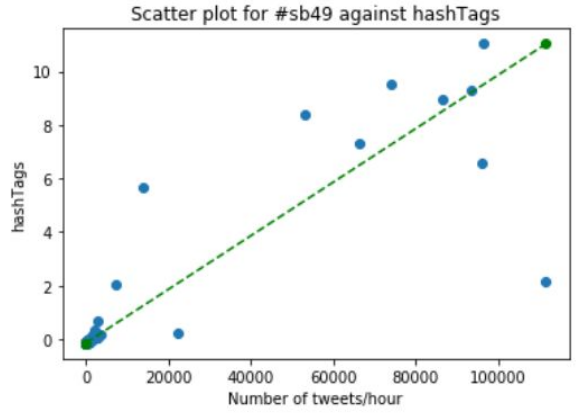
Dep. Variable:	y	R-squared:	0.882
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	119.9
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	7.63e-236
Time:	14:40:02	Log-Likelihood:	-5726.4
No. Observations:	599	AIC:	1.152e+04
Df Residuals:	563	BIC:	1.168e+04
Df Model:	35		
Covariance Type:	nonrobust		

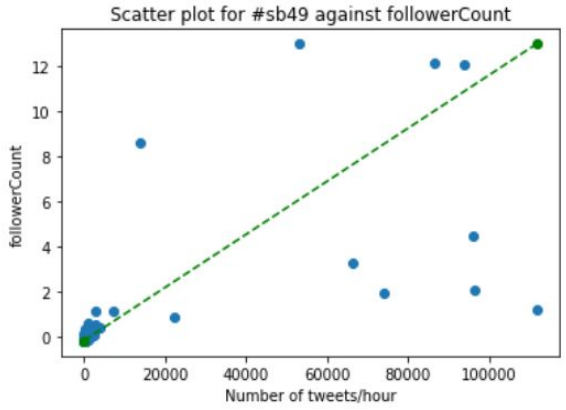
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-7.947e+05	8.54e+04	-9.310	0.000	-9.62e+05	-6.27e+05
x2	1.806e+04	2549.497	7.084	0.000	1.31e+04	2.31e+04
x3	5.974e+04	5904.339	10.119	0.000	4.81e+04	7.13e+04
x4	-1626.9980	253.905	-6.408	0.000	-2125.714	-1128.282
x5	138.8912	139.480	0.996	0.320	-135.074	412.856
x6	98.2037	139.319	0.705	0.481	-175.445	371.852
x7	51.5355	139.250	0.370	0.711	-221.978	325.049
x8	86.1336	139.302	0.618	0.537	-187.482	359.749
x9	30.4792	138.983	0.219	0.826	-242.509	303.467
x10	-11.3848	138.893	-0.082	0.935	-284.197	261.427
x11	39.6959	138.995	0.286	0.775	-233.316	312.707
x12	-68.7575	139.750	-0.492	0.623	-343.252	205.737
x13	-85.0806	139.569	-0.610	0.542	-359.220	189.059
x14	-134.7075	140.736	-0.957	0.339	-411.139	141.724
x15	578.5086	140.094	4.129	0.000	303.337	853.680
x16	-138.7393	141.849	-0.978	0.328	-417.357	139.879
x17	-213.9897	139.587	-1.533	0.126	-488.165	60.186
x18	-134.0744	140.792	-0.952	0.341	-410.615	142.467
x19	14.2831	140.821	0.101	0.919	-262.315	290.881
x20	-189.3432	142.007	-1.333	0.183	-468.272	89.585
x21	-64.7644	141.690	-0.457	0.648	-343.071	213.542
x22	-135.3080	141.865	-0.954	0.341	-413.958	143.342
x23	-61.9763	140.849	-0.440	0.660	-338.631	214.678
x24	23.3292	141.671	0.165	0.869	-254.940	301.598
x25	107.0422	141.542	0.756	0.450	-170.972	385.056
x26	-99.1079	139.357	-0.711	0.477	-372.831	174.615
x27	138.7526	140.437	0.988	0.324	-137.091	414.596
x28	30.9783	139.528	0.222	0.824	-243.080	305.037
x29	-4.482e+04	5492.297	-8.160	0.000	-5.56e+04	-3.4e+04
x30	5.756e+05	7.43e+04	7.748	0.000	4.3e+05	7.22e+05
x31	9.478e+04	8332.932	11.374	0.000	7.84e+04	1.11e+05
x32	-1089.8157	214.067	-5.091	0.000	-1510.283	-669.348
x33	2.003e+05	2.02e+04	9.932	0.000	1.61e+05	2.4e+05
x34	-6.647e+04	1.4e+04	-4.754	0.000	-9.39e+04	-3.9e+04
x35	-2.742e+04	7583.940	-3.615	0.000	-4.23e+04	-1.25e+04
x36	1.027e+04	2514.327	4.084	0.000	5331.078	1.52e+04
const	1380.5526	144.664	9.543	0.000	1096.405	1664.701

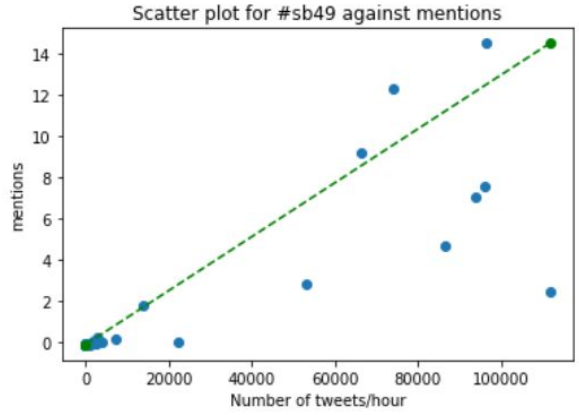
Omnibus:	1125.150	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1544222.510
Skew:	12.442	Prob(JB):	0.00
Kurtosis:	250.493	Cond. No.	6.71e+15

Best Features for #sb49 is:

- hashTags
- followerCount
- mentions







All three features seem to be linearly correlated. The first and third feature seem to be similarly proportional to the predicted values.

Results for #patriot

The regression results for the tweets with the hashtag #patriots are

OLS Regression Results

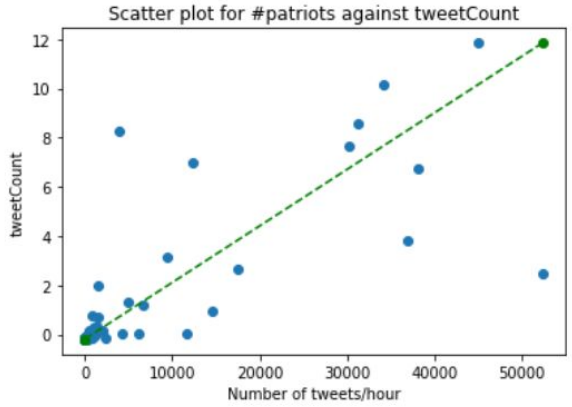
Dep. Variable:	y	R-squared:	0.820
Model:	OLS	Adj. R-squared:	0.808
Method:	Least Squares	F-statistic:	73.04
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	9.27e-185
Time:	14:40:08	Log-Likelihood:	-5355.4
No. Observations:	599	AIC:	1.078e+04
Df Residuals:	563	BIC:	1.094e+04
Df Model:	35		
Covariance Type:	nonrobust		

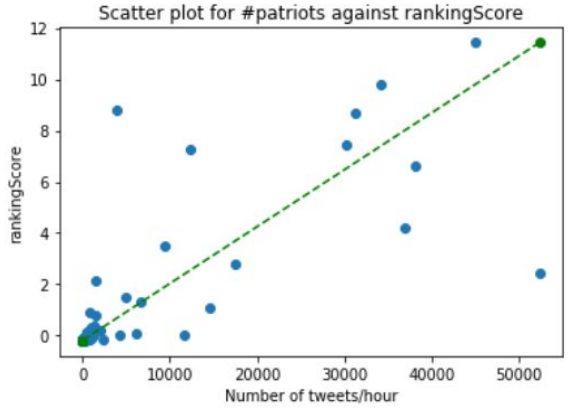
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-2.927e+05	2.62e+04	-11.172	0.000	-3.44e+05	-2.41e+05
x2	-1819.7813	415.708	-4.378	0.000	-2636.310	-1003.253
x3	3695.7020	1488.244	2.483	0.013	772.514	6618.890
x4	-385.7960	146.196	-2.639	0.009	-672.952	-98.640
x5	56.2797	75.364	0.747	0.456	-91.749	204.309
x6	34.8655	75.426	0.462	0.644	-113.286	183.017
x7	-7.6607	75.237	-0.102	0.919	-155.440	140.118
x8	18.2303	75.348	0.242	0.809	-129.768	166.228
x9	-3.3427	75.057	-0.045	0.964	-150.770	144.084
x10	6.3414	75.025	0.085	0.933	-141.021	153.704
x11	-40.4475	74.820	-0.541	0.589	-187.409	106.514
x12	-68.2522	74.988	-0.910	0.363	-215.542	79.037
x13	-84.2355	75.095	-1.122	0.262	-231.736	63.265
x14	-14.5148	75.611	-0.192	0.848	-163.028	133.999
x15	180.6900	75.802	2.384	0.017	31.801	329.579
x16	-64.6115	75.313	-0.858	0.391	-212.541	83.318
x17	-119.2092	75.581	-1.577	0.115	-267.664	29.246
x18	-11.1540	76.407	-0.146	0.884	-161.232	138.924
x19	6.9830	75.560	0.092	0.926	-141.430	155.396
x20	-65.9043	76.184	-0.865	0.387	-215.544	83.735
x21	-5.3287	76.695	-0.069	0.945	-155.971	145.314
x22	29.3012	76.095	0.385	0.700	-120.164	178.767
x23	85.2723	76.991	1.108	0.269	-65.952	236.496
x24	-27.5760	76.752	-0.359	0.720	-178.331	123.179
x25	57.0920	76.273	0.749	0.454	-92.722	206.905
x26	-6.3482	75.847	-0.084	0.933	-155.327	142.630
x27	-7.6663	75.252	-0.102	0.919	-155.475	140.142
x28	52.2066	75.243	0.694	0.488	-95.585	199.998
x29	-2371.0687	1429.054	-1.659	0.098	-5177.997	435.860
x30	2.295e+05	2.18e+04	10.509	0.000	1.87e+05	2.72e+05
x31	3.447e+04	4121.940	8.362	0.000	2.64e+04	4.26e+04
x32	123.5722	116.038	1.065	0.287	-104.348	351.493
x33	3.501e+04	6208.700	5.638	0.000	2.28e+04	4.72e+04
x34	-7598.9437	6122.969	-1.241	0.215	-1.96e+04	4427.710
x35	9892.6013	4158.527	2.379	0.018	1724.479	1.81e+04
x36	-1100.0174	821.614	-1.339	0.181	-2713.821	513.786
const	817.5125	77.871	10.498	0.000	664.560	970.465

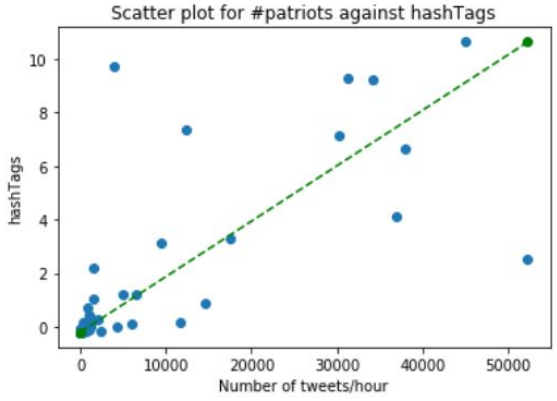
Omnibus:	1092.905	Durbin-Watson:	1.820
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1243150.224
Skew:	11.745	Prob(JB):	0.00
Kurtosis:	224.940	Cond. No.	6.27e+15

Best Features for #patriots is:

- tweetCount
- rankingScore
- hashTags







There a linear relationship with all three features. The first and third feature exhibit a similar correlation, suggesting that these two are of similar importance. The second feature also has a nice linear relation with the predicted values.

Results for #gopatriots

The regression results for the tweets with the hashtag #gopatriots are

OLS Regression Results

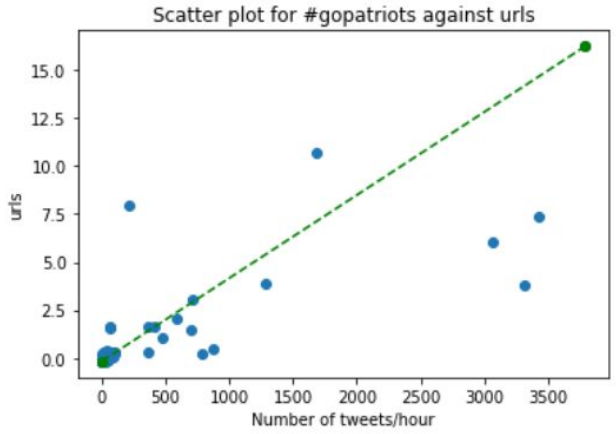
Dep. Variable:	y	R-squared:	0.915
Model:	OLS	Adj. R-squared:	0.910
Method:	Least Squares	F-statistic:	173.8
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	1.96e-276
Time:	14:40:08	Log-Likelihood:	-3524.5
No. Observations:	599	AIC:	7121.
Df Residuals:	563	BIC:	7279.
Df Model:	35		
Covariance Type:	nonrobust		

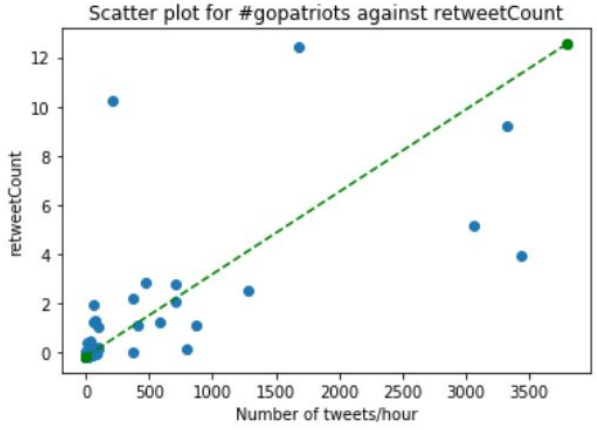
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-9933.8528	917.829	-10.823	0.000	-1.17e+04	-8131.066
x2	-813.8093	52.750	-15.428	0.000	-917.421	-710.198
x3	-105.2861	107.412	-0.980	0.327	-316.264	105.692
x4	285.3944	31.257	9.130	0.000	223.999	346.790
x5	0.8732	3.519	0.248	0.804	-6.038	7.785
x6	0.4241	3.520	0.120	0.904	-6.489	7.337
x7	0.5854	3.522	0.166	0.868	-6.332	7.503
x8	1.1358	3.518	0.323	0.747	-5.774	8.046
x9	-0.6694	3.516	-0.190	0.849	-7.576	6.238
x10	0.2935	3.517	0.083	0.934	-6.614	7.201
x11	-1.3489	3.520	-0.383	0.702	-8.262	5.564
x12	-2.4526	3.519	-0.697	0.486	-9.364	4.459
x13	-4.7643	3.516	-1.355	0.176	-11.671	2.142
x14	-2.5328	3.526	-0.718	0.473	-9.458	4.393
x15	-0.6257	3.523	-0.178	0.859	-7.546	6.294
x16	-4.7975	3.532	-1.358	0.175	-11.735	2.140
x17	-4.7521	3.533	-1.345	0.179	-11.691	2.187
x18	0.0342	3.606	0.009	0.992	-7.048	7.116
x19	10.7675	3.572	3.015	0.003	3.752	17.783
x20	2.3087	3.645	0.633	0.527	-4.851	9.469
x21	-2.9738	3.668	-0.811	0.418	-10.179	4.231
x22	8.6224	3.575	2.412	0.016	1.601	15.644
x23	-3.4560	3.638	-0.950	0.343	-10.601	3.689
x24	-0.8346	3.609	-0.231	0.817	-7.923	6.253
x25	-5.1868	3.556	-1.459	0.145	-12.171	1.797
x26	5.7315	3.542	1.618	0.106	-1.225	12.688
x27	0.4904	3.516	0.139	0.889	-6.416	7.397
x28	3.1897	3.525	0.905	0.366	-3.734	10.113
x29	-580.4917	67.721	-8.572	0.000	-713.508	-447.475
x30	1.042e+04	677.547	15.375	0.000	9086.350	1.17e+04
x31	884.3562	286.908	3.082	0.002	320.816	1447.896
x32	17.4423	7.640	2.283	0.023	2.436	32.448
x33	388.8288	36.626	10.616	0.000	316.888	460.770
x34	498.6153	27.503	18.129	0.000	444.594	552.636
x35	-460.6857	192.865	-2.389	0.017	-839.508	-81.863
x36	-192.7385	78.176	-2.465	0.014	-346.291	-39.186
const	43.7930	3.664	11.953	0.000	36.597	50.989

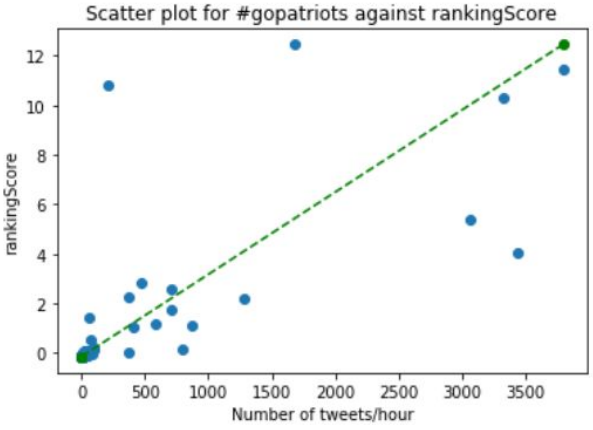
Omnibus:	299.440	Durbin-Watson:	2.004
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45040.545
Skew:	1.118	Prob(JB):	0.00
Kurtosis:	45.422	Cond. No.	4.69e+15

Best Features for #gopatriots is:

- urls
- retweetCount
- rankingScore







There is a linear relationship between the input and output for all the features but there is some spread in the distribution.

Results for #gohawks

The regression results for the tweets with the hashtag #gohawks are

OLS Regression Results

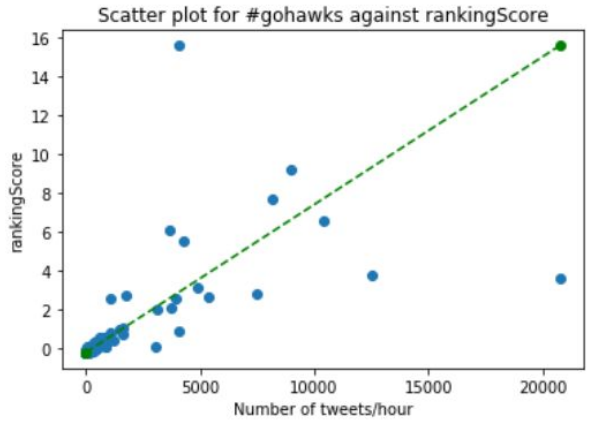
Dep. Variable:	y	R-squared:	0.694
Model:	OLS	Adj. R-squared:	0.675
Method:	Least Squares	F-statistic:	36.41
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	3.10e-121
Time:	14:40:11	Log-Likelihood:	-4798.4
No. Observations:	599	AIC:	9669.
Df Residuals:	563	BIC:	9827.
Df Model:	35		
Covariance Type:	nonrobust		

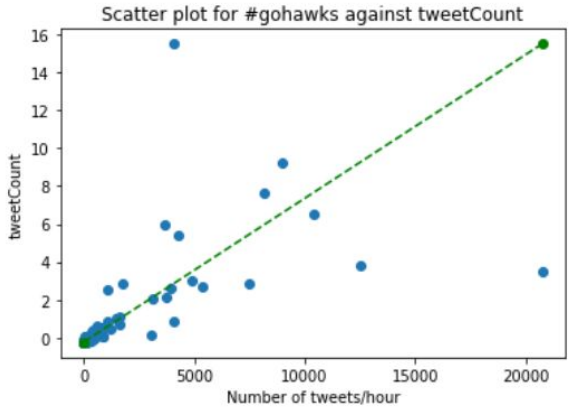
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-6.48e+04	6141.362	-10.551	0.000	-7.69e+04	-5.27e+04
x2	97.7499	126.641	0.772	0.441	-150.996	346.496
x3	-2062.0802	280.206	-7.359	0.000	-2612.458	-1511.702
x4	294.5490	67.816	4.343	0.000	161.346	427.752
x5	10.6119	29.722	0.357	0.721	-47.767	68.991
x6	5.5534	29.676	0.187	0.852	-52.736	63.843
x7	8.9599	29.675	0.302	0.763	-49.328	67.247
x8	8.7879	29.711	0.296	0.768	-49.570	67.145
x9	3.8494	29.658	0.130	0.897	-54.404	62.103
x10	3.8970	29.645	0.131	0.895	-54.331	62.125
x11	1.0087	29.579	0.034	0.973	-57.090	59.108
x12	0.9003	29.571	0.030	0.976	-57.182	58.983
x13	-0.9733	29.839	-0.033	0.974	-59.582	57.636
x14	-13.2743	29.727	-0.447	0.655	-71.664	45.116
x15	-12.5273	30.418	-0.412	0.681	-72.273	47.219
x16	-44.9854	30.122	-1.493	0.136	-104.151	14.180
x17	-69.3604	30.112	-2.303	0.022	-128.506	-10.215
x18	2.5780	30.054	0.086	0.932	-56.454	61.609
x19	141.1396	29.851	4.728	0.000	82.506	199.773
x20	-46.4848	30.390	-1.530	0.127	-106.176	13.207
x21	16.8983	29.671	0.570	0.569	-41.382	75.178
x22	34.0930	29.933	1.139	0.255	-24.700	92.886
x23	-41.0252	29.764	-1.378	0.169	-99.487	17.436
x24	-10.3578	30.722	-0.337	0.736	-70.702	49.986
x25	5.1599	29.954	0.172	0.863	-53.675	63.994
x26	-15.0454	30.262	-0.497	0.619	-74.485	44.394
x27	4.9067	29.688	0.165	0.869	-53.406	63.220
x28	5.8023	29.689	0.195	0.845	-52.512	64.116
x29	191.3944	190.657	1.004	0.316	-183.092	565.881
x30	6.052e+04	5687.286	10.642	0.000	4.94e+04	7.17e+04
x31	-101.2555	942.911	-0.107	0.915	-1953.309	1750.798
x32	42.5779	43.474	0.979	0.328	-42.813	127.968
x33	2146.6169	329.476	6.515	0.000	1499.465	2793.769
x34	1000.7235	194.762	5.138	0.000	618.175	1383.272
x35	5718.5575	806.732	7.089	0.000	4133.986	7303.129
x36	-1863.8646	352.229	-5.292	0.000	-2555.709	-1172.020
const	314.0334	30.725	10.221	0.000	253.683	374.384

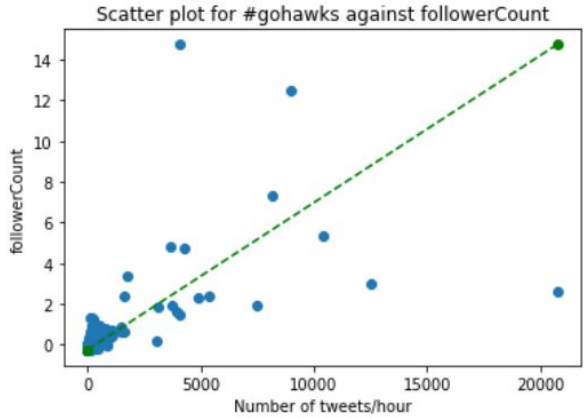
Omnibus:	924.839	Durbin-Watson:	1.951
Prob(Omnibus):	0.000	Jarque-Bera (JB):	564151.585
Skew:	8.412	Prob(JB):	0.00
Kurtosis:	152.401	Cond. No.	5.92e+15

Best Features for #gohawks is:

- rankingScore
- tweetCount
- followerCount







There is a linear relationship in the scatter plots reflecting a good relationship between the three features.

Results for #nfl

The regression results for the tweets with the hashtag #nfl are

OLS Regression Results

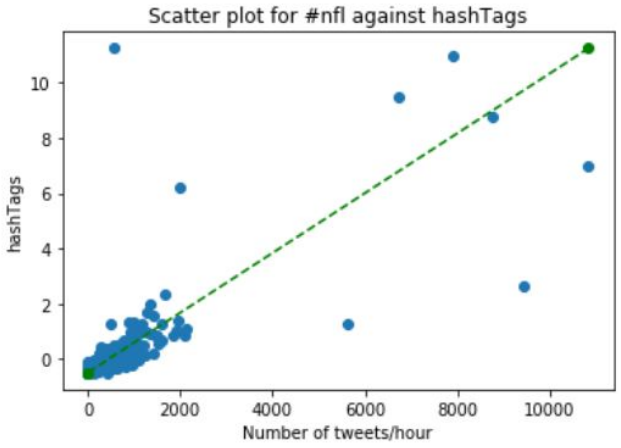
Dep. Variable:	y	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.766
Method:	Least Squares	F-statistic:	57.01
Date:	Mon, 19 Mar 2018	Prob (F-statistic):	7.08e-161
Time:	14:40:14	Log-Likelihood:	-4453.4
No. Observations:	599	AIC:	8979.
Df Residuals:	563	BIC:	9137.
Df Model:	35		
Covariance Type:	nonrobust		

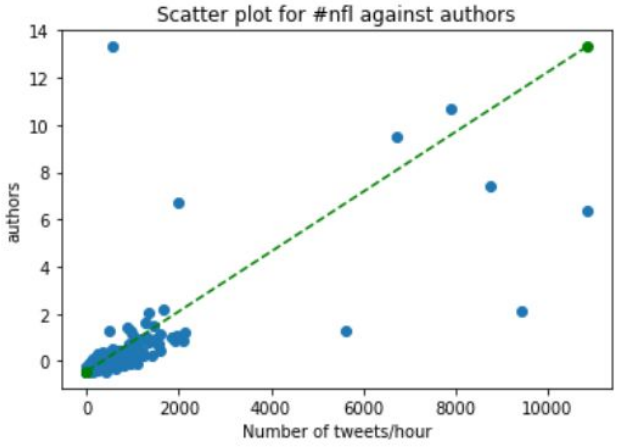
	coef	std err	t	P>|t|	[0.025	0.975]
x1	-1680.8943	1250.386	-1.344	0.179	-4136.887	775.098
x2	-64.9280	82.320	-0.789	0.431	-226.620	96.764
x3	-374.4514	143.435	-2.611	0.009	-656.185	-92.718
x4	114.7793	42.038	2.730	0.007	32.209	197.350
x5	-25.0116	16.952	-1.475	0.141	-58.308	8.285
x6	-24.1847	16.855	-1.435	0.152	-57.292	8.923
x7	-20.0622	16.885	-1.188	0.235	-53.228	13.103
x8	-8.1570	16.737	-0.487	0.626	-41.031	24.717
x9	0.0412	16.657	0.002	0.998	-32.677	32.759
x10	15.7586	16.696	0.944	0.346	-17.036	48.553
x11	26.7896	16.621	1.612	0.108	-5.857	59.436
x12	17.7093	16.758	1.057	0.291	-15.207	50.625
x13	12.1089	17.018	0.712	0.477	-21.319	45.536
x14	12.0021	16.884	0.711	0.477	-21.161	45.165
x15	21.3595	16.963	1.259	0.208	-11.959	54.678
x16	18.4304	17.063	1.080	0.281	-15.085	51.946
x17	-6.6134	17.151	-0.386	0.700	-40.300	27.074
x18	2.5218	16.808	0.150	0.881	-30.492	35.535
x19	84.1302	16.833	4.998	0.000	51.067	117.194
x20	-17.7716	17.313	-1.027	0.305	-51.777	16.234
x21	-4.6663	16.842	-0.277	0.782	-37.747	28.415
x22	12.9416	17.074	0.758	0.449	-20.595	46.478
x23	8.2873	16.801	0.493	0.622	-24.713	41.288
x24	-15.3547	16.878	-0.910	0.363	-48.507	17.798
x25	-19.7492	16.758	-1.178	0.239	-52.665	13.167
x26	-20.2913	16.838	-1.205	0.229	-53.364	12.781
x27	-41.4848	16.804	-2.469	0.014	-74.492	-8.478
x28	-29.3005	16.981	-1.725	0.085	-62.655	4.054
x29	86.1305	105.312	0.818	0.414	-120.722	292.983
x30	1891.4932	1173.020	1.612	0.107	-412.537	4195.524
x31	2642.1668	255.669	10.334	0.000	2139.986	3144.348
x32	-176.1530	29.210	-6.031	0.000	-233.527	-118.779
x33	556.1614	130.071	4.276	0.000	300.677	811.645
x34	249.2511	93.829	2.656	0.008	64.953	433.549
x35	-1502.4054	246.587	-6.093	0.000	-1986.748	-1018.063
x36	-1075.1501	257.418	-4.177	0.000	-1580.767	-569.533
const	432.3289	17.272	25.031	0.000	398.404	466.254

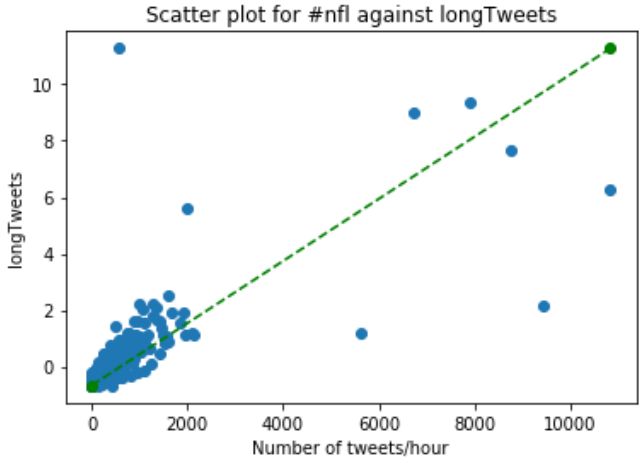
Omnibus:	769.419	Durbin-Watson:	2.309
Prob(Omnibus):	0.000	Jarque-Bera (JB):	132722.405
Skew:	6.304	Prob(JB):	0.00
Kurtosis:	74.825	Cond. No.	8.37e+15

Best Features for #nfl is:

- hashTags
- authors
- Long Tweets







A linear relationship can be seen for features and all three features seem to be proportional to the predicted values in a similar way. This suggests that these three features can be very strong when used to predict the number of tweets for this hashtag in the next hour.

Observations

From the graphs above, we can clearly see that each hashtag, the patterns are similar within each hashtag. This tells us that the features have high importance and are similar to one another in the regression model. We also see that the pattern for the attributes is linear and hence using a linear regression is logical and useful. We were able to achieve a good R^2 score to confirm this.

Problem 1.4

For each hashtag, report the average cross-validation errors for the 3 different models. Note that you should do the 90-10% splitting for each model within its specific time window. I.e. Only use data within one of the 3 periods above for training and testing each time, so for each period you will run 10 tests.

Also, aggregate the data of all hashtags, and train 3 models (for the intervals mentioned above) to predict the number of tweets in the next hour on the aggregated data.

The below table shows the 2 cross validation RMSE and MAE scores.

Table for Before the Superbowl Match

Model\ Hashtags	**Error type**	**#Superbowl**	**#GoPatriots**	**#GoHawks**	**#NFL**	**#Patriots**	**#sb49**
1. Linear Regression	RMSE	22.84	5.75	19.41	11.85	19.41	9.09
	MAE	204.51	13.11	160.89	71.20	134.43	44.73
2. Neural Network Regression	RMSE	23.61	11.44	20.98	18.48	22.29	245.71
	MAE	273.29	28.17	155.35	186.31	203.58	226.42
3. Random Forest Regression	RMSE	22.32	4.923	16.97	12.95	17.76	8.72
	MAE	175.93	7.83	74.36	75.42	101.32	47.41
4. SVR	RMSE	22.94	6.25	19.86	14.75	19.48	12.87
	MAE	264.42	14.76	138.28	127.94	165.53	137.31
Order of MAE		3>1>4>2	3>1>4>2	3>4>2>1	1>3>4>2	3>1>4>2	1>3>4>2

Normalized MAE for all hashtags for the best performing model (Random Forest): 114.74

Table for during the Superbowl Match

Model\ Hashtags	Error type	#Superbowl	#GoPatriots	#GoHawks	#NFL	#Patriots	#sb49
1. Linear Regression	RMSE	852.34	73.34	79.41	87.25	164.49	225.21
	MAE	723874.81	4995.28	6310.79	7597.39	27035.51	50944.37
2. Neural Network Regression	RMSE	317.15	42.89	79.41	71.39	175.22	266.132
	MAE	101204.80	1854.41	6175.20	5006.19	30512.39	70451.47
3. Random Forest Regression	RMSE	199.14	29.86	51.55	44.19	127.31	181.16
	MAE	39911.95	904.32	2547.44	1831.66	15982.0	33101.94
4. SVR	RMSE	348.59	36.56	57.32	71.326	110.29	177.74
	MAE	120084.889	1412.46	3217.31	5128.5	12092.54	31525.49
Order of MAE		3>2>4>1	3>4>2>1	3>4>2>1	3>2>4>1	4>3>1>2	4>3>1>2

Normalized MAE for all hashtags for the best performing model (Random Forest): 28678.24

Table for after the Superbowl match

Model\Hashtags	Error type	#Superbowl	#GoPatriots	#GoHawks	#NFL	#Patriots	#sb49
1. Linear Regression	RMSE	21.46	2.31	7.85	13.95	12.3	15.97
	MAE	279.13	4.22	28.49	134.12	89.39	194.82
2. Neural Network Regression	RMSE	29.14	7.72	12.63	23.85	15.21	20.73
	MAE	738.12	6.21	73.91	536.41	186.33	395.14
3. Random Forest Regression	RMSE	19.96	3.61	5.86	14.79	11.168	13.146
	MAE	257.81	2.822	23.95	161.88	97.935	134.88

4. SVR	RMSE	21.71	2.823	8.67	16.27	13.62	22.67
	MAE	638.68	8.25	59.62	283.66	208.91	446.59
Order of MAE		3>1>4>2	3>1>2>4	3>1>2>4	1>3>4>2	1>3>2>4	3>1>2>4

Normalized MAE for all hashtags for the best performing model (Random Forest): 181.23

Normalized MAE = Summation over all (MAE for # * tweets for #)/Total tweets for #

We tried 4 different regression models for this part, namely, Linear Regression, Neural Networks, Random Forest and SVR.
We can see Random Forest is performing the best out of the 4 because

We can see that for some hashtags(patriots and nfl), linear regression also performs decently well when compared with Random Forest which tells us that the data for these hashtags are linear separable. In addition to this, the fact that data for time section 1 and 3 is less when compared with time section 2, Random forest is not able to generalize this data and doesn't give good results but instead, linear regression performs better hence it is linearly separable. But time section 2 has lot of data so RF performs well.

Problem 1.4b
Perform the same evaluations on your combined model and compare with models you trained for individual hashtags.

Before Feb. 1, 8:00 a.m.
'Mean absolute error: ', 211.835
'RMSE: ', 782.496

Between Feb. 1, 8:00 a.m. and 8:00 p.m.
'Mean absolute error: ', 29725.951
'RMSE: ', 37239.824

After Feb. 1, 8:00 p.m
'Mean absolute error: ', 176.0413
'RMSE: ', 303.729

Observations
Some observation we'll like to make is that when comparing the MAE values for the combined dataset vs the individual hashtags, we observe that the MAE values are lower for specific dataset than for combined dataset. This is because the combined dataset leads to a model which is more general and is unaware of the hashtag associated with the data point when we test it. This tells is that the importance of hashtag to predict tweet count (model trained on specific hashtag) performs better than the combined training model.

Problem 1.5

Report the model you use. For each test file, provide your predictions on the number of tweets in the next hour.

For this part, we used the trained aggregate model of part 1.4.2. We first sliced it hourly and then combined it vertically for the last 5 hour periods. We tried to test this with different samples from different time sections. We arrived at 3 different combined models for different time sections for 1,2 and 3.

For this section we created a vector for the 5 hourly period having 11*5 features. (Basically, 11 features for each hour section). This helps us understand how the tweet count will be for the next hour. This is done to see how the previous 5 hour sections can help predict the tweet count for the next hour.

Sample 8 was only file which did not have 5 hour data so we performed oversampling by replicating the last hour data and introduce more weight in it.

Sample and Period	Predicted tweets in next hour
Sample 1 Period 1	244.831
Sample 2 Period 2	165535.241
Sample 3 Period 3	877.323933
Sample 4 Period 1	362.0899531
Sample 5 Period 1	322.74213
Sample 6 Period 2	147211.824
Sample 7 Period 3	75.129
Sample 8 Period 1	28.426
Sample 9 Period 2	156374.812
Sample 10 Period 3	75.135

We can see from the above table that the next hour tweet predictions follow very closely to the tweet count we get in the 5 hour data. We compared the tweet count of the previous hour with true value and saw it matches well. Hence our model was able to predict the number of tweets in next hour.

We know from the data that time period 2 is most significant as it is during the match. Our data is larger in 2 when compared to 1 and 3. Hence the samples which are corresponding to period 2 have higher prediction which tells us that the model is trained well. Hence the higher tweet count of period 2 when compared with 1 and 3 is justified.

Order of number of tweets
Time period 2 > Time period 1 > Time period 3
The above can be see from the table.

Part 2 - Fan Base Prediction

In this part, we try to classify the location of a tweet based on it's content.
Instead of directly using the dataset for this part, we first need to create our training and test set by manually checking the location field for each tweet in the #superbowl tweet file.
In order to identify the location, we used the following terms to determine whether the tweet is from **Washington**:

- Seattle
- WA
- Washington
- Kirkland

Note that Washington does not guarantee that the tweet was from Washington state, it could have also originated from Washington DC. We ensured that we handled such boundary cases while labelling the location. For **Massachusetts**, the set of cities were as follows :

- MA
- Massachusetts
- Boston
- Worcester
- Springfield
- Lowell
- Arlington
- Bedford
- Brockton
- Quincy
- Lynn
- Northampton
- Cambridge

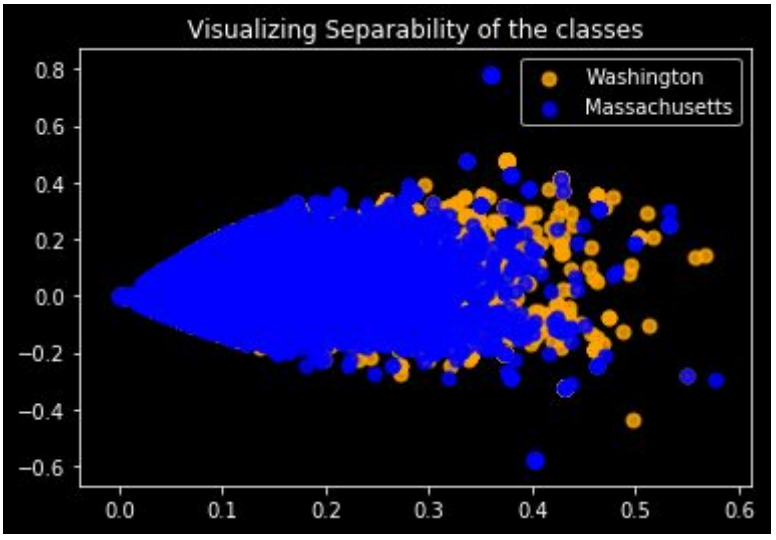
Even for Massachusetts there is an edge case. Boston is also a city in Ohio. We ensured that tweets which contained 'Boston, Ohio' as location were not considered in the dataset.
After labelling the data, we **preprocessed the tweets** to remove the following from any tweet:

- Mentions (@twitter_handle)
- URL
- HTML tags
- Emoticons
- Numbers
- 'RT' indicating the tweet was a retweet.

We also performed stemming to bring the words to their root form and restrict the vocabulary of the dataset.
Ex :
Original Tweet : Terminator Genisys: He's back, and we're lovin' it, @ParamountPics. RT to try to win movie tix for a year https://cards.twitter.com/cards/16ac3u/bmun
Preprocessed Tweet : terminator genisys he s back and we re lovin it to try to win movie tix for a year

After preprocessing, we **split the data** into train and test set.
On the train set we performed **TF-IDF** computations and reduced the dimension of the resulting sparse matrix using **Truncated SVD**.
On the test set, we transformed it using the previously learned TF-IDF transformer and the SVD model.

To visualize the linear separability of the problem, we visualized the dataset by projecting the training data onto a 2D plane. This visualization is observed as follows:



From this we can infer that the separating the tweet location just from the tweets is a very difficult and challenging task for all machine learning algorithms because there is a large overlap between the two classes. However, note that this visualization is only when we reduce the data to 2D, and it may be separable when we increase the dimensions.

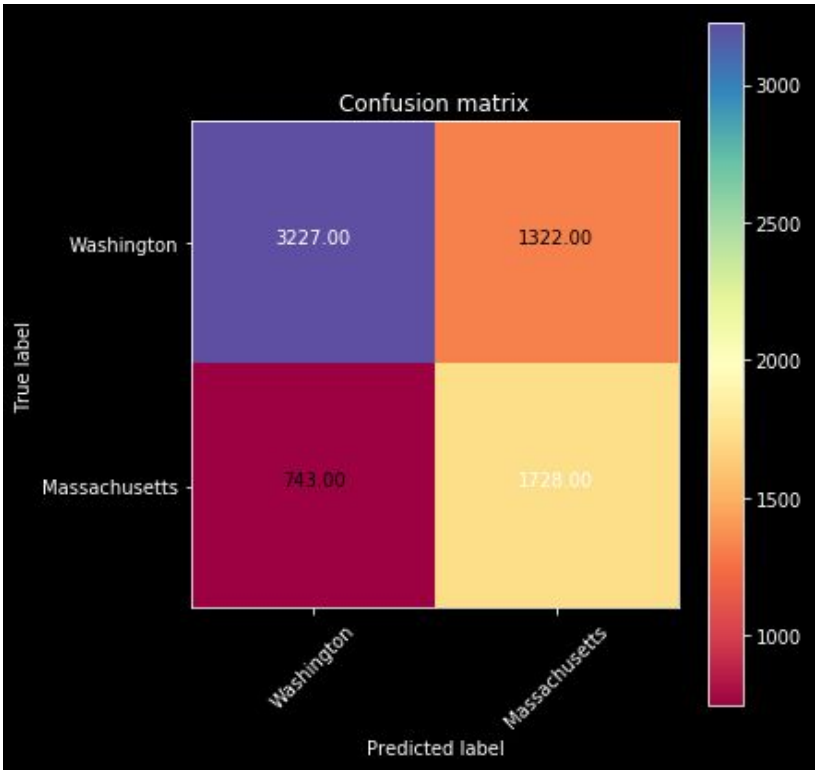
We tested the performance of many models on this binary classification task and observed the following results:

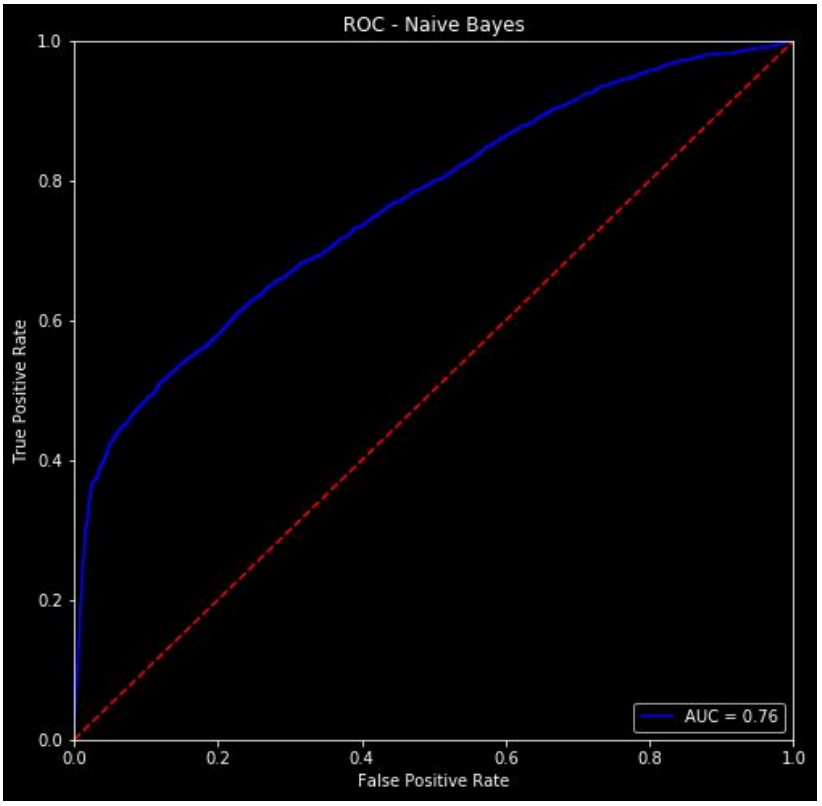
Classifier	Accuracy	Precision	Recall	AUC

Naive Bayes	0.7058	0.6897	0.7043	0.76
SVM	0.7723	0.7443	0.8165	0.84
Random Forest	0.7633	0.7356	0.8027	0.81
Adaboost	0.7628	0.7408	0.7814	0.82
Multilayer Perceptron	0.7465	0.7394	0.7424	0.83
Logistic Regression	0.7767	0.7506	0.8141	0.85

As we can observe from the above table, SVM and logistic regression were the best models to classify the tweet coming from a particular location. AUC (Area under the curve) is a popular metric to measure the performance of a binary classifier. From the table again, it is evident that these two classifiers perform better than the rest.

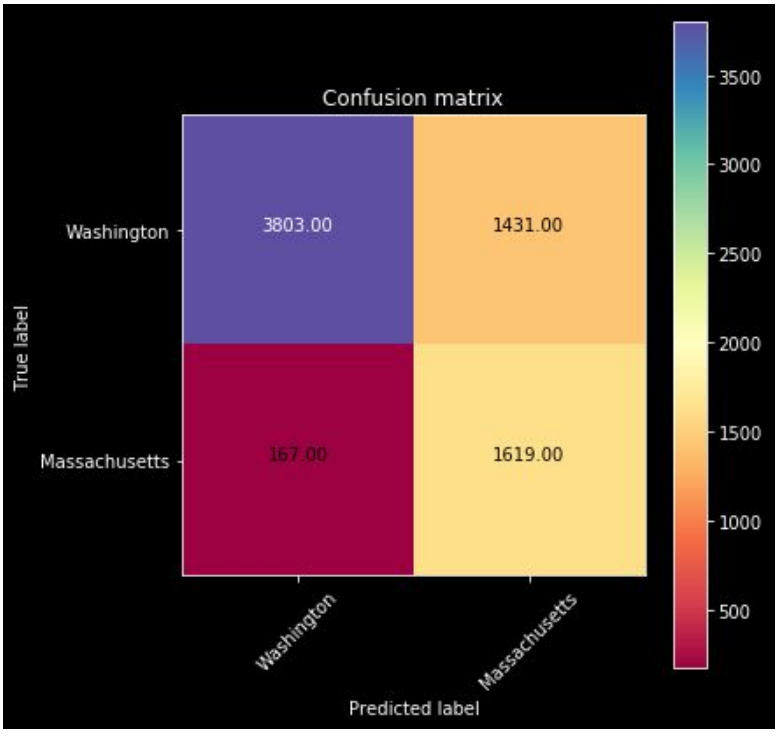
Naive Bayes

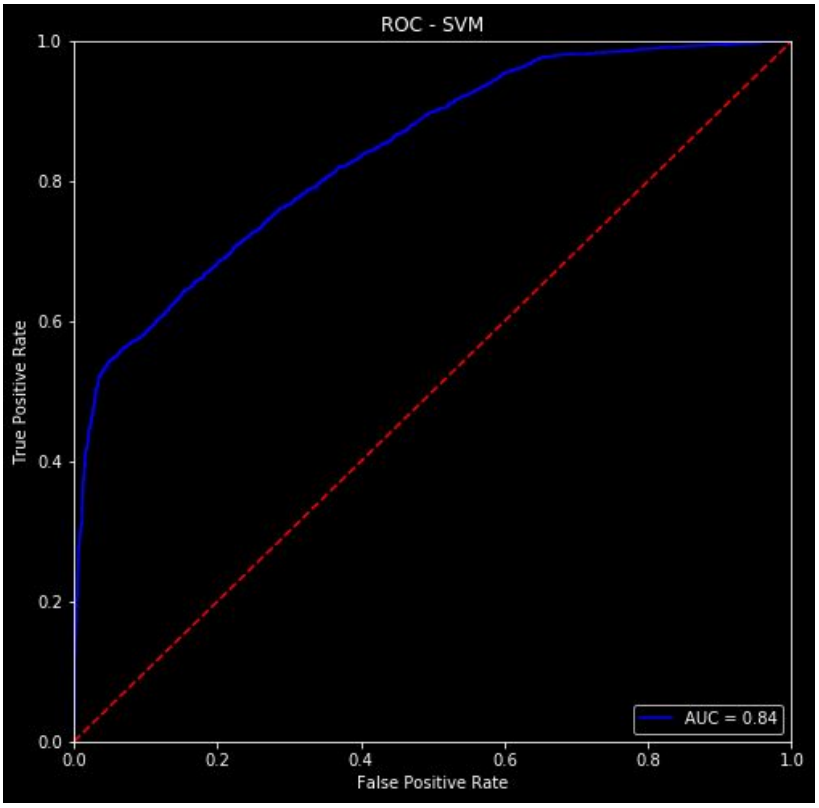




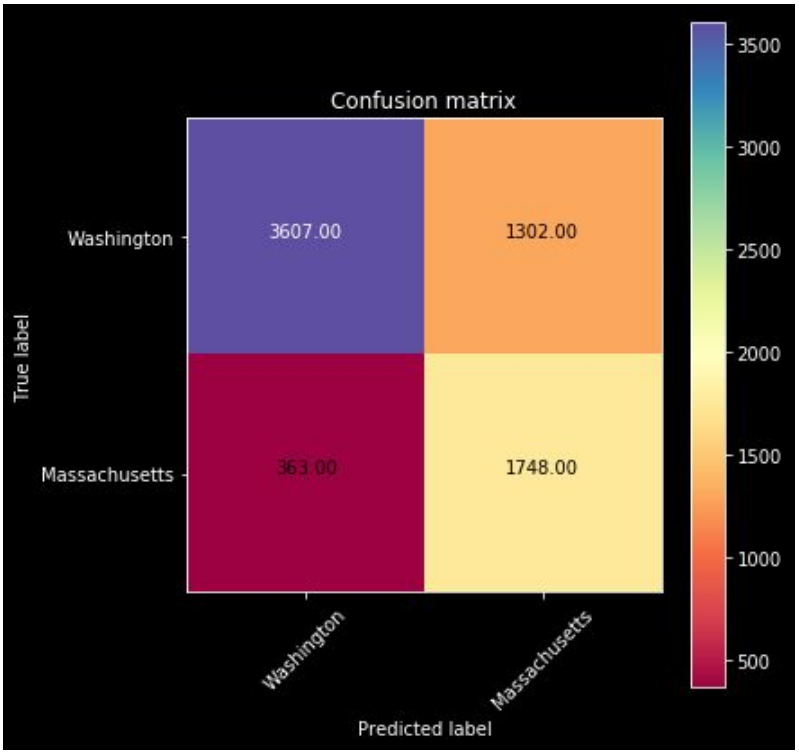
Amongst all the classifiers, Naive Bayes has the worst performance due to the independence assumption. It was also noticed in Project 1 that SVM performs better than Naive Bayes and the same holds true for this task also.

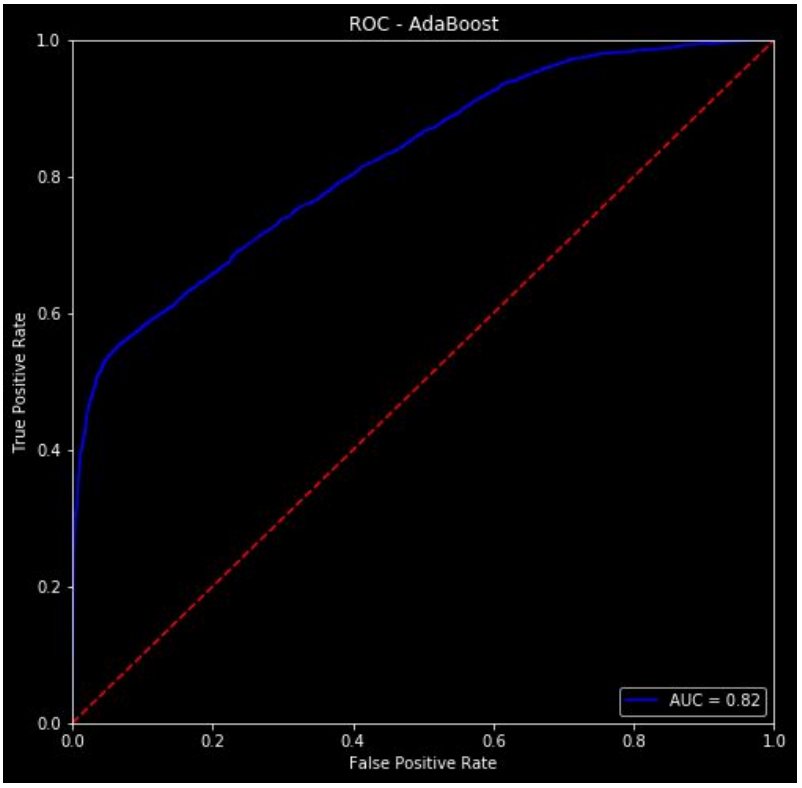
SVM



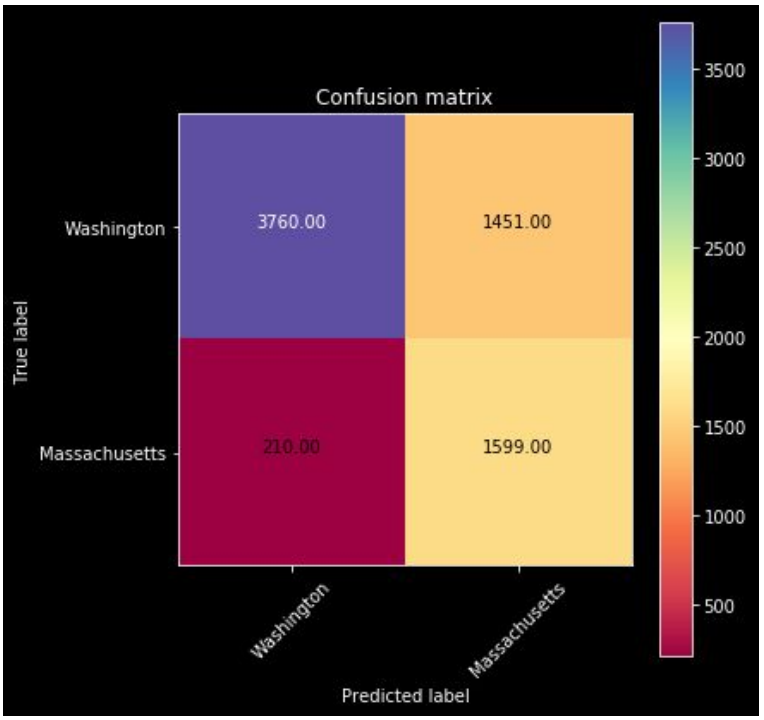

ROC - SVM
True Positive Rate
False Positive Rate
AUC = 0.84
0.0
0.2
0.4
0.6
0.8
1.0

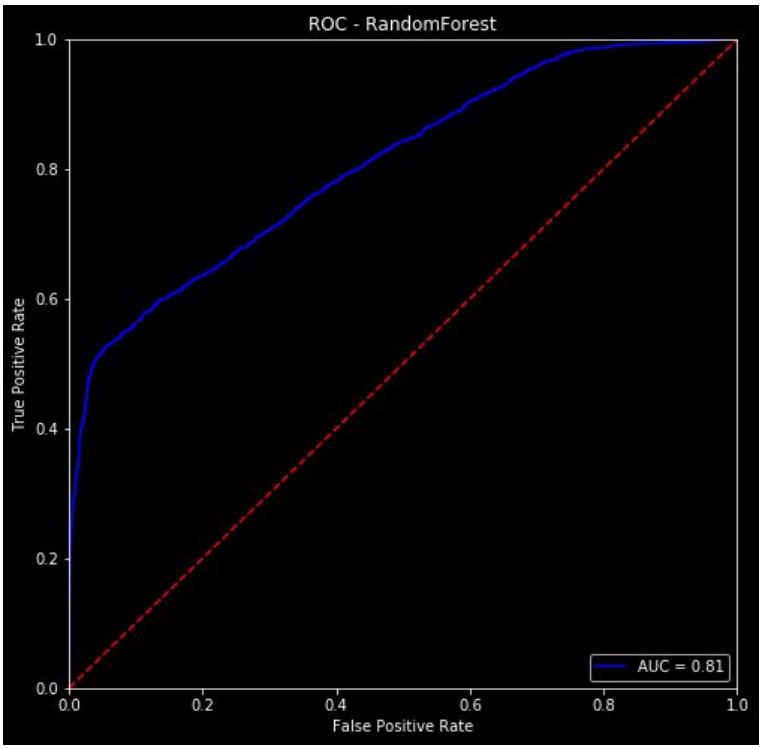
AdaBoost


Confusion matrix
Washington
Massachusetts
3607.00
1302.00
363.00
1748.00
True label
Predicted label
Washington
Massachusetts
3500
3000
2500
2000
1500
1000
500


ROC - AdaBoost
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.82

Random Forest Classifier

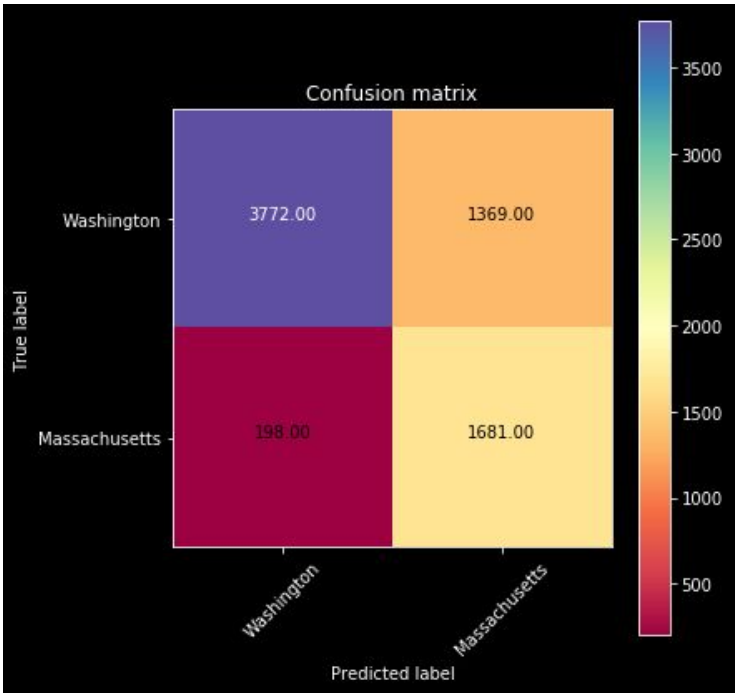


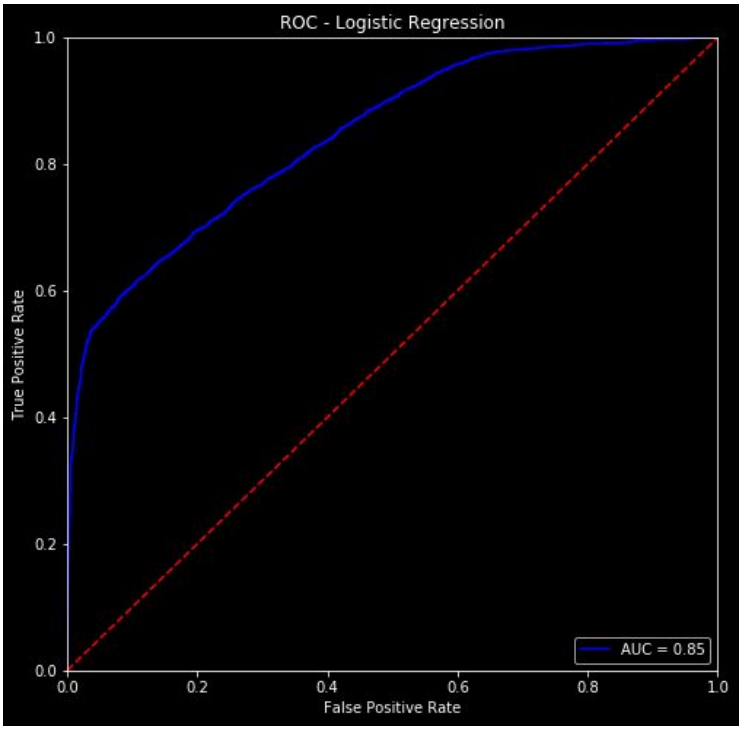


Max Depth = 5

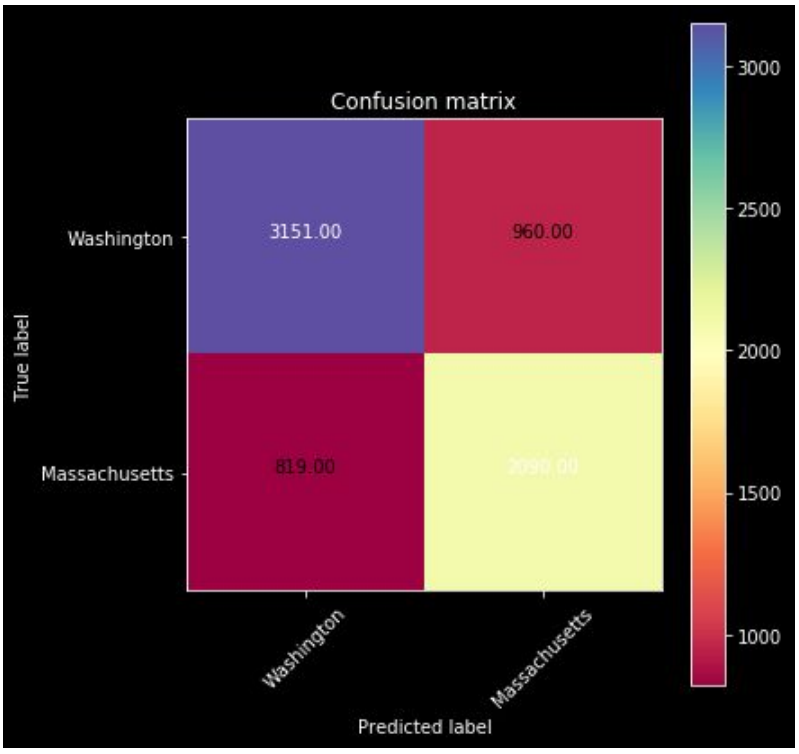
Number of estimators = 50

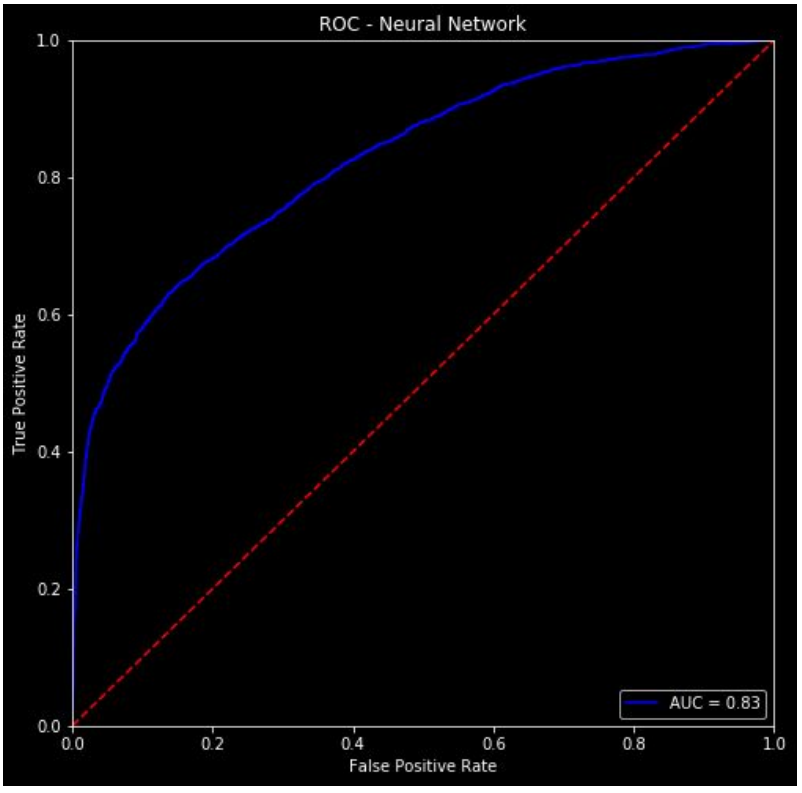
Logistic Regression





Multilayer Perceptron

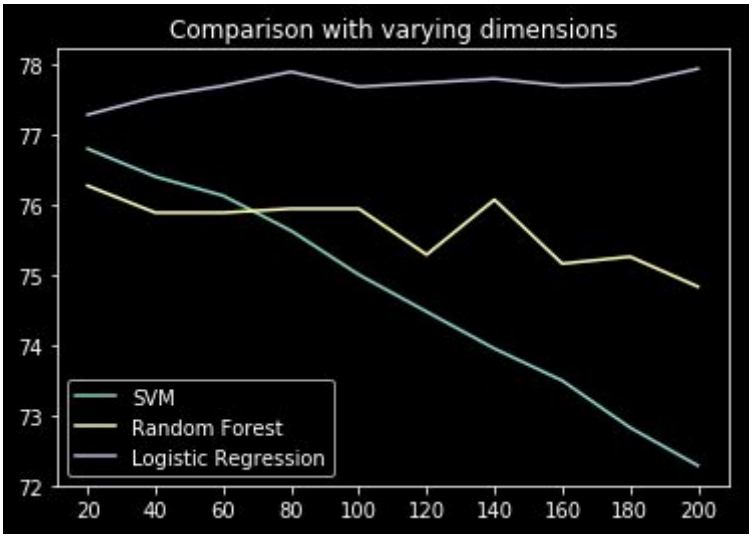




Hidden units = 100
Hidden layers = 2
Activation function = relu

From the graphs and the table, we can draw some of the following inferences:

1. SVM and Logistic Regression gave equivalent results for this task with accuracies of 77%.
2. Naive Bayes was not a suitable classifier for this task due to the linear independency.
3. For multi-layer perceptron, we can try various other hyperparameter settings (We tested it on a 2 layer with 100 units each). But given that there are nearly 40,000 tweets, this algorithm may require more data to perform better due to the large number of parameters.
4. We also found that on increasing more cities from Washington, reduced the performance of all algorithms by almost 10% for each classifier. In Washington, we added the following additional cities - Spokane, Bellevue, Everett, Yakima, Redmond, Kent and Tacoma. A possible reason for this could be that these city names are common across the world, for example, Kent is a county in Michigan, Ohio, UK, Washington, Connecticut and even Australia! Thus, adding these additional cities could have resulted in a larger, more incorrect dataset. The key factor in this particular task is to have a correctly labeled dataset, and thus handling these possible incorrect labels is crucial to the performance of the system.
5. We also tested the performance of the models by varying the dimension of the data.



As we can see from this graph, on increasing the dimension, the accuracy for Random Forest and SVM decreases, whereas the accuracy for logistic regression remains the same.

In general, we can say that increasing the dimensions reduces the performance of the algorithms due to increasing sparsity of the data. Also, since the tweets are short, increasing dimensions should not improve the performance of the system as most of the important features to distinguish between the location of the tweet should already be captured within the first few dimensions.

In general, neural networks worked a bit better. Logistic regression maybe worked well as seen in previous parts, since features are linearly correlated to predicted values. Neural network is able to fit the model well given the data. Also, random forests sometimes don't work well for textual data as text data has many terms or features which are not contributing. As a result, weak trees will be created and the forest has larger likelihood to make wrong decision which mainly results from those weak trees.

Part 3 - What's Trending?

One of the most important features in Twitter is it's capability to find who/what is trending at any given time. Analysing the twitter data to find out the trends is extremely vital to marketing officials for brand popularity. Not limited to just companies, by analysing the tweets, we can also try and find out key moments and popular and influential people. If a person is more influential, his/her brand value increases and thus they can monetize their services accordingly.

For this part, we propose the objective of trying to find out the trending topics and analysing the Super Bowl tweets for advertising companies and player performances. The part can be split up into 3 specific sub parts -

I. Analyzing advertisements
II. Analyzing celebrity importance at Super Bowl
III. Analyzing player impact for each team

Approach

- The approach for all parts is similar in structure. We considered all tweets generated on the Super Bowl day and split them up into 15 minute windows for analysis.
- After preprocessing the tweets by removing some punctuations and stop words, we get all hashtags and non-hashtag words from the tweet.
- We analyze the processed text for each tweet to get keywords with respect to ads, celebrities and players.
- The data is then classified into a respective advertisement/celebrity/player.
- We analyze all tweets in the 15 minute window based on tweet count based on the keywords to find out the most trending advertisement/celebrity/player for that window and to interpret the results by comparing it to the real world information avaliable.

Advertisements

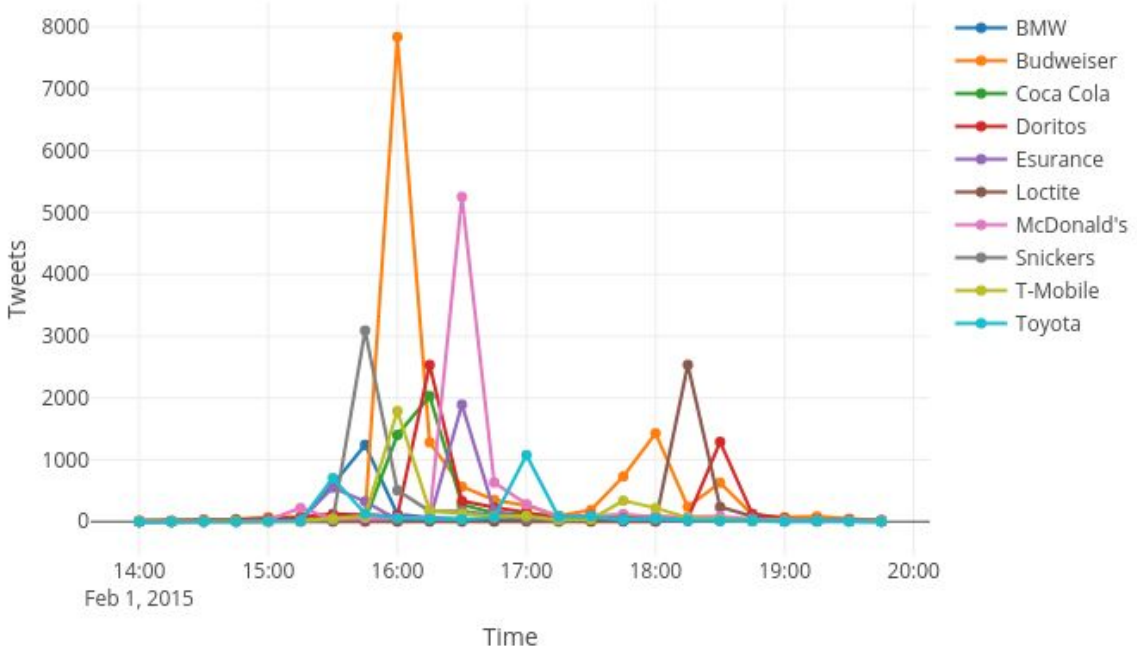
Super Bowl is one of the biggest events of the year creating a huge amount of interaction on the social media, buzz among brand advertisements and the celebrity halftime show. Over the years, it has become less about football and more about the entertainment surrounding the event. It has become an opportunity for companies to sell their lucrative products and connect with over 111 million of their customers. These companies spend millions of dollars for their 30-45 second commercials. We can use the Twitter data generated by the event to help brands understand their customers and see what is trending in that time.

For our analysis, we considered the following brands that had advertised during the Super Bowl :

- T-Mobile
- Budweiser
- BMW
- Coca Cola
- Doritos
- Esurance

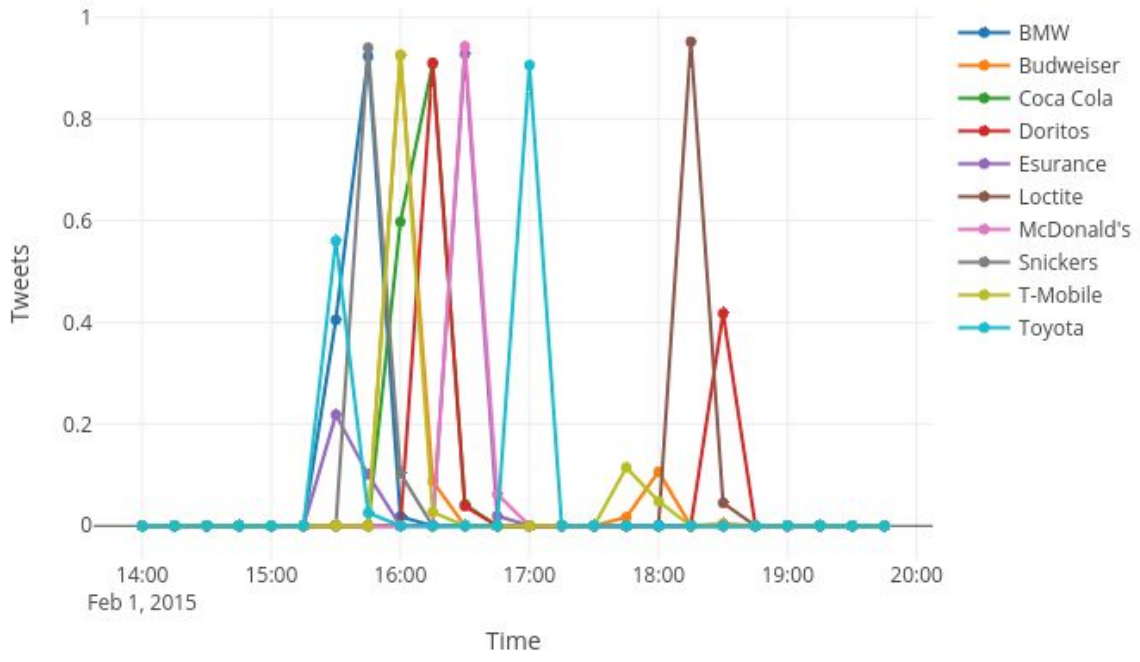
- Loctite
- Mcdonald's
- Snickers
- Toyota

To handle brands with 2 words in their names, we also computed a bigram count to find whether the tweet contained a reference to that company. Aside from just checking for the company name, we also checked for the tagline of the advertisement being used as a hashtag. This is a very popular way of promoting tweets by companies and we often see them happening in Twitter or any other social media platform.



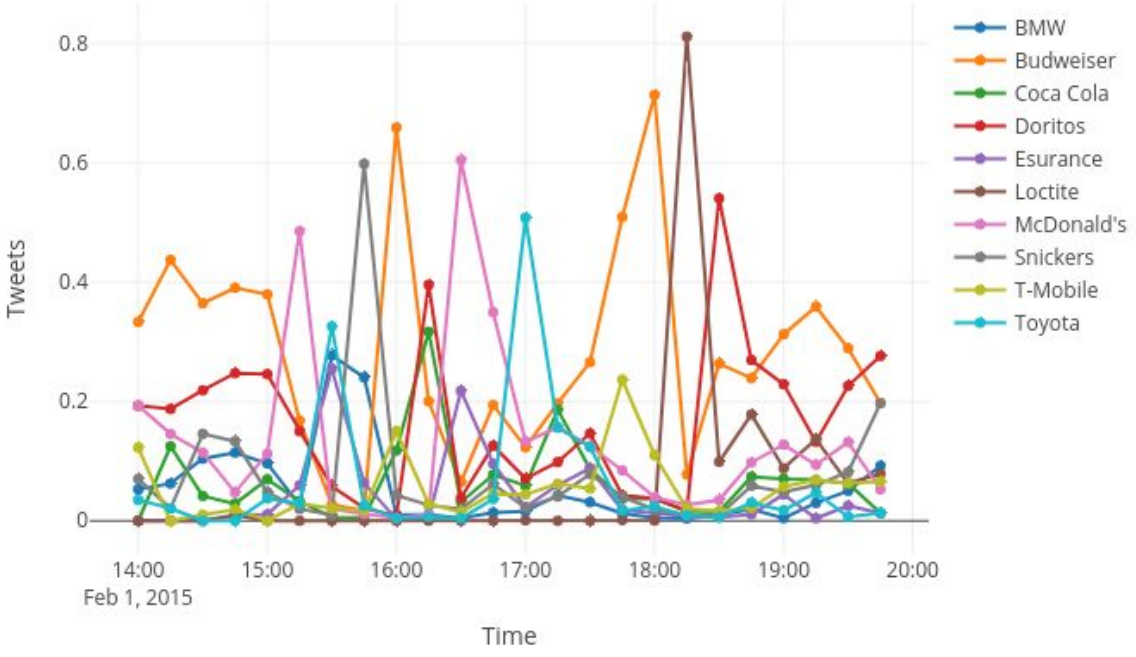
From this graph, we can see how popular each company's advertisement was. Majority of the peaks are between 15:00 and 17:00 hours. Budweiser had the most number of tweet references followed by McDonald's, Snickers and Loctite. While this graph shows the popularity for each company, we can normalize the tweets per company per window, and notice a key observation.

These results are not available in the pdf as we used Plotly library to visualize interactive graphs. These graphs can be found at - **https://plot.ly/~varunsaboo.**



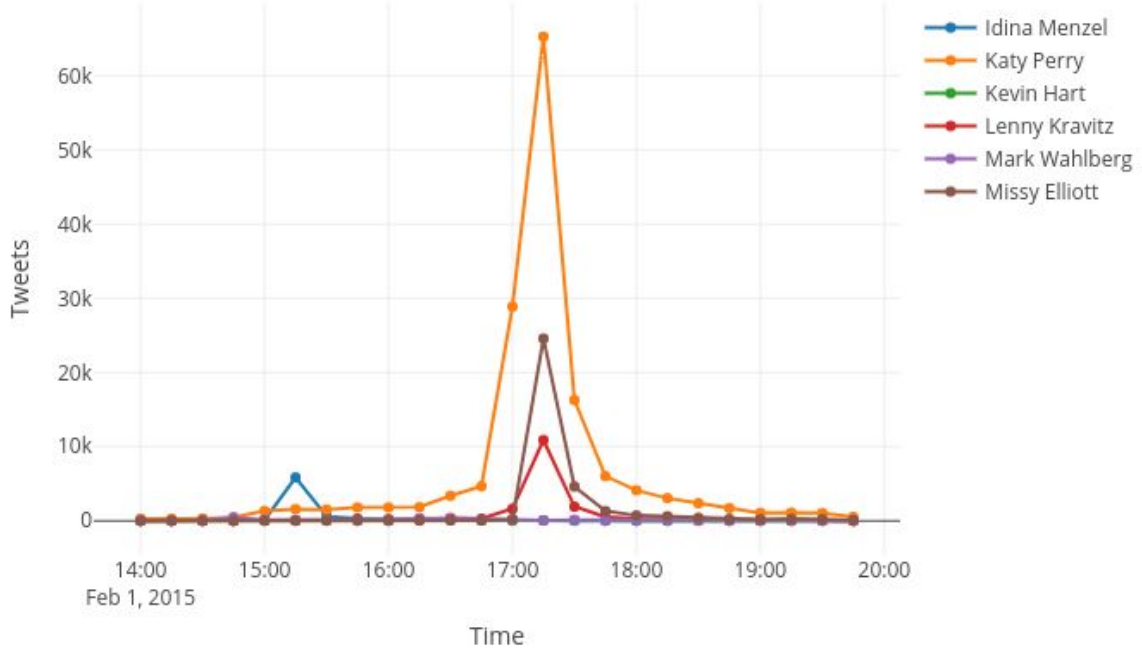
From this graph, it is evident that each company had a strategy to air their ads at specific times. We see that all companies peak at a specific time and then drop off. This is usually the case because the cost of advertisements is extremely expensive at Super Bowl. While majority of the companies made their efforts before half-time (17:00), Loctite was the leading company for advertisement chatter post half-time. Most of the focus on advertisements was before the half-time due to 'competition' amongst companies to have the biggest impact when the game is in its infancy. During the end, people tend to tweet lesser about ads as more importance is given to players and the outcome of the final.

Another visualization of this same graph is viewing it based on normalized number of tweets received per hour per company. The graph below demonstrates the ratio of tweets being addressed to a particular company per hour. For instance, at 18:15, Loctite received 80% of the tweets. This graph shows how each company fared against every other company during every window. The higher count on number of tweets for a company reflects that an ad for that company was recently played, and more people are tweeting about it. Certain large companies have more than one spike in this graph. This indicates that they probably spent more money to show their ads more times than other companies. Another factor for this could be due to the fact that the company had more than one advertisement. In this case, the first ad was aired during the first peak whereas the second ad was aired during the second peak. Such is the case for Budweiser. We identify that the company had 2 different hashtags by doing our own research online and finding out the different ads made by a company for Super Bowl 49.

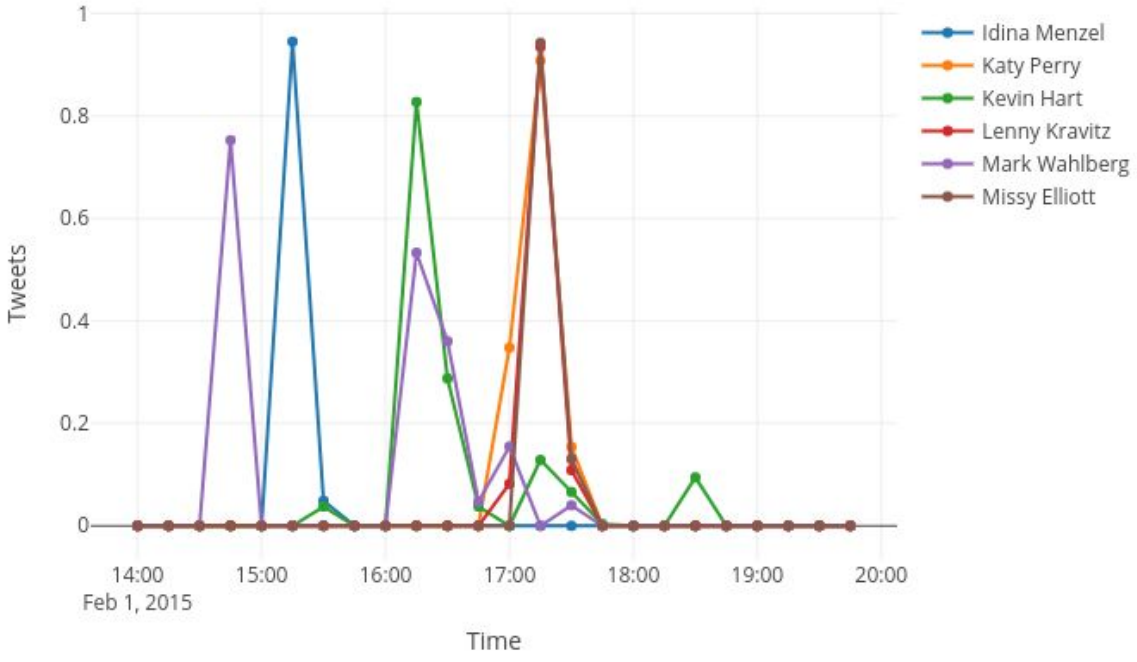


Celebrities

No major sporting event is ever complete without celebrity presence. Super Bowl is no different. In fact, it is opposite! Super Bowl attracts many top celebrities from Hollywood and the music industry. Celebrities flock out in large numbers to support their favorite teams, and entertain the fans by special half-time performances. During Super Bowl, each celebrity is present for one of 3 reasons - to perform at half-time, to show support to their team, or to sing the national anthem prior to kick off. In Super Bowl 49, Katy Perry was the headliner at the Pepsi halftime show and was supported by fellow artists Lenny Kravitz and Missy Elliott. The national anthem was sung by Idina Menzel. To pick all 3 types of celebrities at Super Bowl, we considered the aforementioned 4 celebrities as well as Mark Wahlberg and Kevin Hart who were present at the stadium to cheer their favored teams respectively. By performing steps similar to detecting popularity of Advertisements, we generated the following graphs -

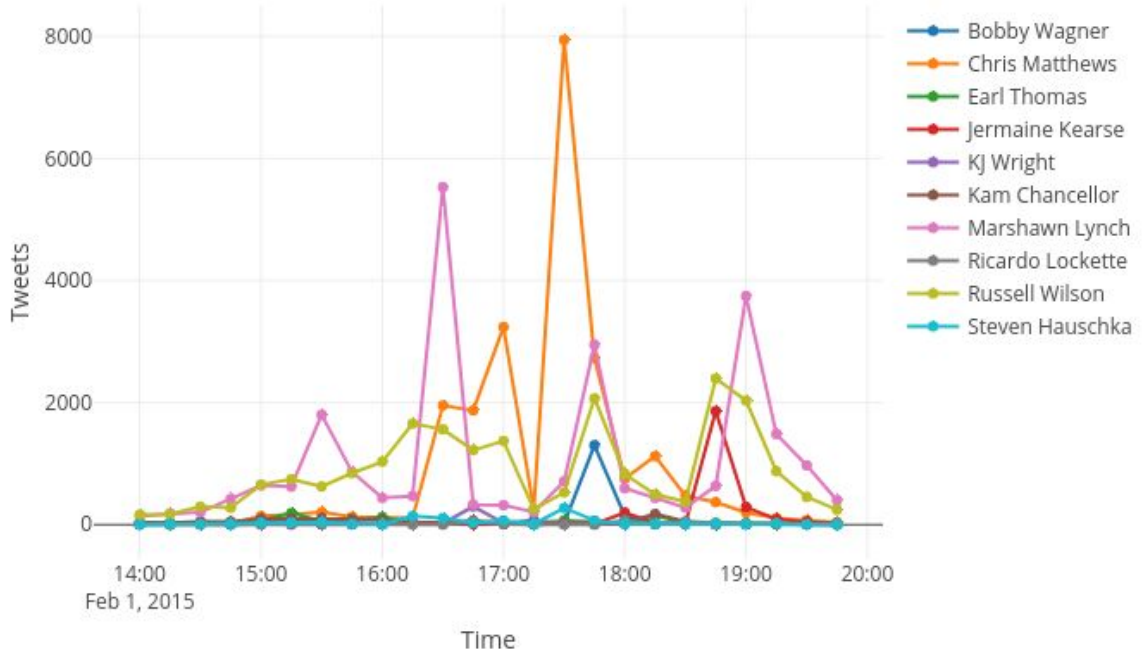


As we can see from this graph, it is evident that the halftime show took place at 17:15 and the national anthem was sung at 15:15. There was a lot of buzz and excitement for Katy Perry's performance and is indicated by the consistent increase in her tweets from 16:00. She was the headline for the show and got more than sixty thousand tweets during her performance, which was more than the combined sum of the tweets for all other performers. Idina Menzel was trending at 15:15 when she sang the national anthem. We also see that there were barely any tweets for Mark Wahlberg who was only a spectator, and hence didn't trend compared to the other celebrities. However, by viewing the next graph, we notice that he was probably shown on TV and hence received certain tweets before the match started.


1
0.8
0.6
0.4
0.2
0
Tweets
14:00
Feb 1, 2015
15:00
16:00
17:00
18:00
19:00
20:00
Time
Idina Menzel
Katy Perry
Kevin Hart
Lenny Kravitz
Mark Wahlberg
Missy Elliott

Player Impact

While it is fun to see which celebrities, companies were trending during a sporting event, the crux of any sporting event is the players. We also performed analysis on the players for each team to identify which players produced crucial impact during the game.



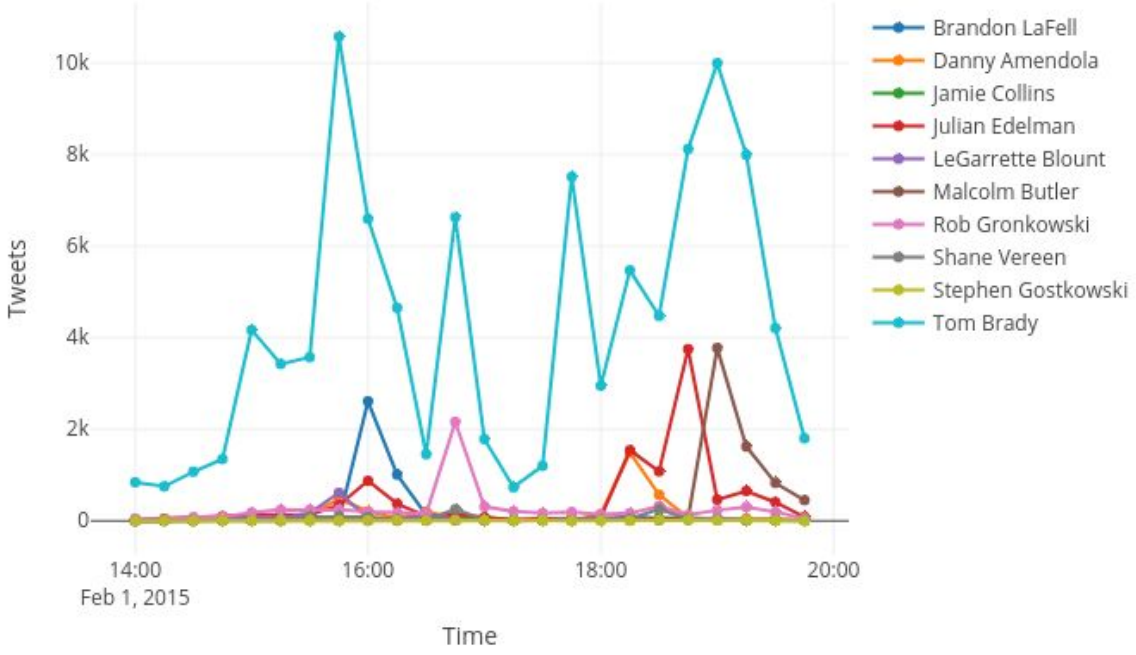
The above graph shows the trending players for Seahawks. According to our analysis, the 3 most influential/popular players during the Super Bowl for the Seattle Seahawks were :

1. Chris Matthews
2. Marshawn Lynch
3. Russell Wilson

The Seahawks players were trending during the 2nd and 3rd quarters.

In the 2nd quarter, the teams tied 14-14 and during the 3rd quarter Seahawks won 10-0.

In the second quarter, Lynch scored a touchdown for Seahawks. This is highlighted by the corresponding peak at 16:30.



This graph shows the trending players for New England Patriots during the Super Bowl. The popularity for Tom Brady is constant throughout the match because of his popular nature amongst other celebrities compared to all other athletes. For the PAtriots, the following 3 players were the most influential in the game -

1. Tom Brady
2. Julian Edelman
3. Malcolm Butler

Tom Brady was arguably the best player of the match and received the MVP trophy after the match. He made 2 crucial passes for touchdowns in the 4th quarter which resulted in Patriot's turning the game around on its head. The two touchdowns were for Edelman and Amendola, both of whom have a peak during the 4th quarter of the match. Additionally, at the end of the 4th quarter, Malcolm Butler made a crucial and mind blowing interception to prevent Lynch from scoring a touchdown for the Seahawks. This resulted in both players having a spike in their tweets in the 4th quarter.

Aside from just our analysis, Twitter also conducted their own analysis. Their results were -

Our results match all but one of Twitter's analysis. Hence, we feel confident that our analysis of the twitter data was correct.

An additional task that we could have implemented would be to use sentiment analysis of the tweets to find out whether the tweet for each ad, celebrity or player was positive or negative. With the help of this, we could have assigned a score to identify the most liked ad, player and celebrity performance.

Conclusion:

In this project, we analyze and explore the Twitter dataset and activity for the Superbowl event. We used various regression models to first predict the popularity of different hashtags. We tried to see the tweet activity of different hashtags going on and in future. We also saw the best features contributing the most for this task. We also predicted the number of tweets in next hour using previous hours for each hashtag. We also proposed additional ideas that can be used to help companies utilize the twitter activity using the vast and rich data.