ECE219 Project 5

Application - Twitter data

Zhilai Shen, Yufei Hu, Zheang Huai, and Tianyi Liu
105023454, 404944367, 505222324, 705035425

March 19, 2019

Contents

1 Popularity Prediction

1.1 A first look at the data

Question 1:
The statistics for each hashtag are as follows.

Statistics for #GoHawks
Average number of tweets per hour: 292.488
Average number of followers of users posting the tweets per tweet: 2217.924
Average number of retweets per tweet: 2.0132

Statistics for #GoPatriots
Average number of tweets per hour: 40.955
Average number of followers of users posting the tweets per tweet: 1427.253
Average number of retweets per tweet: 1.408

Statistics for #NFL
Average number of tweets per hour: 397.021
Average number of followers of users posting the tweets per tweet: 4662.375
Average number of retweets per tweet: 1.534

Statistics for #Patriots
Average number of tweets per hour: 750.894
Average number of followers of users posting the tweets per tweet: 3280.464
Average number of retweets per tweet: 1.785

Statistics for #SB49
Average number of tweets per hour: 1276.857
Average number of followers of users posting the tweets per tweet: 10374.160
Average number of retweets per tweet: 2.527

Statistics for #SuperBowl
Average number of tweets per hour: 2072.118
Average number of followers of users posting the tweets per tweet: 8814.968
Average number of retweets per tweet: 2.391

Question 2:
The plots of number of tweets in hour over time for #SuperBowl and #NFL can be seen in Figure 1,Figure 2.

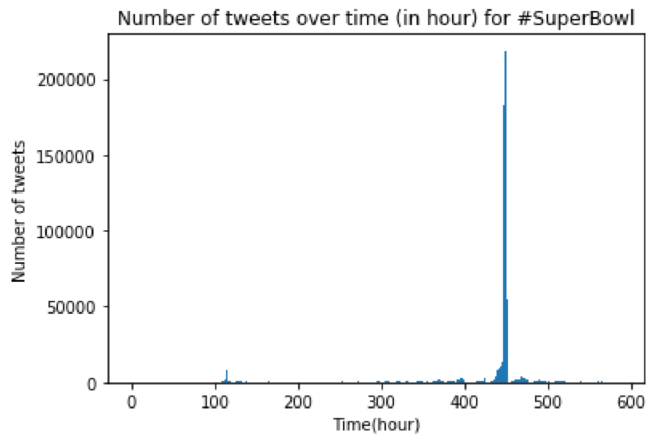


Figure 1: Number of tweets over time for #SuperBowl

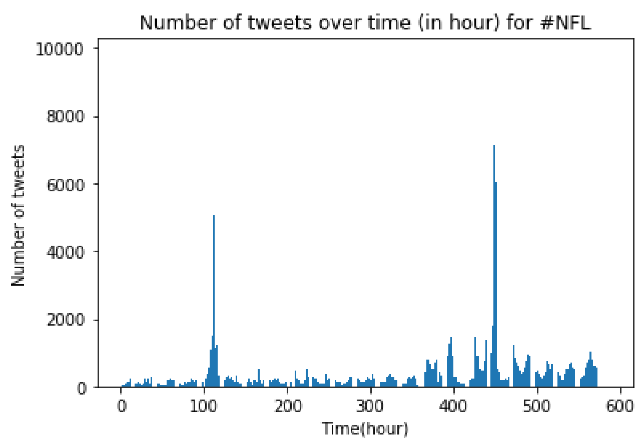


Figure 2: Number of tweets over time for #NFL

Table 1: MSE and R-squared measure

hashtag	MSE	R-squared measure
#GoHawks	758554.248	0.504
#GoPatriots	27583.582	0.637
#NFL	269962.153	0.652
#Patriots	5180890.103	0.679
#SB49	16180394.455	0.808
#SuperBowl	52483472.229	0.803

```
hashtag:  #GoHawks
mse:  758554.2484282234
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.504
Model:                            OLS   Adj. R-squared:                  0.500
Method:                 Least Squares   F-statistic:                     116.5
Date:                Sat, 09 Mar 2019   Prob (F-statistic):           7.10e-85
Time:                        02:13:54   Log-Likelihood:                -4733.9
No. Observations:                 578   AIC:                             9478.
Df Residuals:                     573   BIC:                             9500.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.2856      0.164      7.843      0.000       0.964       1.608
x2            -0.1378      0.043     -3.169      0.002      -0.223      -0.052
x3            -0.0002   8.01e-05     -2.434      0.015      -0.000   -3.76e-05
x4          7.145e-05      0.000      0.480      0.631      -0.000       0.000
x5             7.5919      2.956      2.569      0.010       1.787      13.397
==============================================================================
Omnibus:                      910.753   Durbin-Watson:                   2.214
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           771478.377
Skew:                           8.575   Prob(JB):                         0.00
Kurtosis:                     181.156   Cond. No.                     2.14e+05
------------------------------------------------------------------------------
P values:
[2.15583400e-14 1.60947513e-03 1.52369053e-02 6.31486219e-01
 1.04619105e-02]
```

Figure 3: t-test and p-values for #GoHawks

1.2 Linear regression

Question 3:

The models' Mean Squared Error(MSE) and R-squared measure for each hashtag can be seen in Table 1. The results of t-test and p-values can be seen in Figure 3, 4, 5, 6, 7, 8. x1-x5 represents Number of tweets, Total number of retweets, Sum of the number of followers of the users, Maximum number of followers of the users, Time of the day, respectively.

From p-values, we can get the significance of each feature for every hashtag, i.e., The feature with smaller p-value has higher significance. In general, number of tweets has low p-value for every hashtag and therefore it's a very significant feature.

Note: Here I follow the same instruction as in Question1 "if a users posted twice, we count the user and the user's followers twice as well", when I calculate sum of the number of followers of the users.

```
hashtag:  #GoPatriots
mse:  27583.582295460703
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.637
Model:                            OLS   Adj. R-squared:                  0.634
Method:                 Least Squares   F-statistic:                     199.8
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          1.02e-122
Time:                        02:13:58   Log-Likelihood:                -3749.4
No. Observations:                 574   AIC:                             7509.
Df Residuals:                     569   BIC:                             7531.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.3071      0.285      1.079      0.281      -0.252       0.866
x2             0.5010      0.191      2.629      0.009       0.127       0.875
x3            -0.0001      0.000     -0.570      0.569      -0.001       0.000
x4         -9.038e-06      0.000     -0.042      0.967      -0.000       0.000
x5             0.3459      0.539      0.641      0.522      -0.714       1.405
==============================================================================
Omnibus:                      481.255   Durbin-Watson:                   1.908
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           290776.984
Skew:                           2.475   Prob(JB):                         0.00
Kurtosis:                     113.152   Cond. No.                     2.98e+04
==============================================================================
P values:
[0.28086202 0.00878988 0.56881223 0.96689613 0.52158747]
```

Figure 4: t-test and p-values for #GoPatriots

```
hashtag:  #NFL
mse:  269962.15281800315
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.652
Model:                            OLS   Adj. R-squared:                  0.649
Method:                 Least Squares   F-statistic:                     217.8
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          1.23e-130
Time:                        02:14:36   Log-Likelihood:                -4500.0
No. Observations:                 586   AIC:                             9010.
Df Residuals:                     581   BIC:                             9032.
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.6317      0.134      4.718      0.000       0.369       0.895
x2            -0.1811      0.064     -2.831      0.005      -0.307      -0.055
x3             0.0001    2.5e-05      4.256      0.000    5.73e-05       0.000
x4          -9.96e-05   3.28e-05     -3.038      0.002      -0.000   -3.52e-05
x5             7.5679      1.965      3.852      0.000       3.709      11.426
==============================================================================
Omnibus:                      619.607   Durbin-Watson:                   2.363
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           342008.050
Skew:                           3.927   Prob(JB):                         0.00
Kurtosis:                     121.091   Cond. No.                     3.91e+05
==============================================================================
P values:
[2.99044196e-06 4.79416843e-03 2.42865533e-05 2.48610844e-03
 1.30036264e-04]
```

Figure 5: t-test and p-values for #NFL

```
hashtag:  #Patriots
mse:  5180890.103265264
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.679
Model:                            OLS   Adj. R-squared:                  0.677
Method:                 Least Squares   F-statistic:                     246.3
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          5.98e-141
Time:                        02:15:43   Log-Likelihood:                -5361.9
No. Observations:                 586   AIC:                         1.073e+04
Df Residuals:                     581   BIC:                         1.076e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.9148      0.071     12.943      0.000       0.776       1.054
x2            -0.0675      0.058     -1.170      0.243      -0.181       0.046
x3         -1.156e-05   2.63e-05     -0.439      0.661   -6.32e-05    4.01e-05
x4             0.0001   9.08e-05      1.489      0.137   -4.31e-05       0.000
x5             5.2220      7.843      0.666      0.506     -10.182      20.626
==============================================================================
Omnibus:                      884.481   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           688343.951
Skew:                           7.876   Prob(JB):                         0.00
Kurtosis:                     170.163   Cond. No.                     6.81e+05
==============================================================================
P values:
[7.59459794e-34 2.42607082e-01 6.60651423e-01 1.37065456e-01
 5.05783085e-01]
```

Figure 6: t-test and p-values for #Patriots

```
hashtag:  #SB49
mse:  16180394.454846866
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.808
Model:                            OLS   Adj. R-squared:                  0.807
Method:                 Least Squares   F-statistic:                     486.4
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          3.15e-204
Time:                        02:17:33   Log-Likelihood:                -5656.5
No. Observations:                 582   AIC:                         1.132e+04
Df Residuals:                     577   BIC:                         1.134e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             1.1370      0.087     13.037      0.000       0.966       1.308
x2            -0.1615      0.079     -2.054      0.040      -0.316      -0.007
x3          9.832e-06   1.25e-05      0.786      0.432   -1.47e-05    3.44e-05
x4          9.889e-05    4.2e-05      2.356      0.019    1.65e-05       0.000
x5            -4.3893     13.259     -0.331      0.741     -30.431      21.653
==============================================================================
Omnibus:                     1177.660   Durbin-Watson:                   1.673
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2194090.157
Skew:                          14.537   Prob(JB):                         0.00
Kurtosis:                     302.387   Cond. No.                     6.31e+06
==============================================================================
P values:
[3.10714295e-34 4.04375643e-02 4.31982721e-01 1.87947699e-02
 7.40735299e-01]
```

Figure 7: t-test and p-values for #SB49

```
hashtag:  #SuperBowl
mse:  52483472.22917878
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.803
Model:                            OLS   Adj. R-squared:                  0.801
Method:                 Least Squares   F-statistic:                     473.8
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          2.80e-202
Time:                        02:20:39   Log-Likelihood:                -6039.9
No. Observations:                 586   AIC:                         1.209e+04
Df Residuals:                     581   BIC:                         1.211e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             2.2765      0.080     28.559      0.000       2.120       2.433
x2            -0.2553      0.046     -5.595      0.000      -0.345      -0.166
x3            -0.0001   2.19e-05     -6.278      0.000      -0.000   -9.44e-05
x4             0.0007      0.000      5.013      0.000       0.000       0.001
x5           -29.0126     26.714     -1.086      0.278     -81.480      23.455
==============================================================================
Omnibus:                      974.639   Durbin-Watson:                   2.285
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1789674.506
Skew:                           9.288   Prob(JB):                         0.00
Kurtosis:                     273.097   Cond. No.                     9.75e+06
==============================================================================

P values:
[9.67194039e-113 3.40873070e-008 6.71733985e-010 7.12296562e-007
 2.77906836e-001]
```

Figure 8: t-test and p-values for #SuperBowl

1.3 Feature analysis

Question 4:

The new features we find useful for this problem are:

-Url ratio. A url in Twitter can be a link of a picture, a song, a video, or a piece of news. High ratio of tweets with urls may indicate a topic about a good song, an interesting picture or video, or a piece of breaking news. In our project, we used "url count" to represent "url ratio".

-Author count. Besides tweet count for a hashtag, we also consider the unique number of authors who posted tweets containing the hashtag. This feature can be used to recognize those hashtags automatically posted by some fake accounts.

-Mention count. Mention is a directional sharing behavior in Twitter. Messages can be shared to a designated user using @ as the prefix of the user's name. If a user was mentioned in a tweet with a hashtag, he probably took part in the topic, especially when this mention came from his friends.

-Ranking score. Ranking scores are listed in each tweet to show its scores intuitively, which shows its spread ability.

-Number of hashtags. Sometimes, some hashtags are not used individually, but are used together with other hashtags, e.g. #boston#explosion. It's reasonable to guess the number of hashtag in tweets are critical to indicate the popularity of the topic.

After adding these five new features, the models' Mean Squared Error(MSE) and R-squared measure for each hashtag can be seen in Table 2. The results of

Table 2: MSE and R-squared measure(after adding new features)

hashtag	MSE	R-squared measure
#GoHawks	485098.424	0.684
#GoPatriots	8182.927	0.892
#NFL	163901.087	0.791
#Patriots	2922872.021	0.819
#SB49	12387685.9921	0.853
#SuperBowl	30825995.718	0.884

```
hashtag:  #GoHawks
mse:  485098.4239537786
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.684
Model:                            OLS   Adj. R-squared:                  0.678
Method:                 Least Squares   F-statistic:                     122.7
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          6.98e-135
Time:                        02:48:06   Log-Likelihood:                -4604.0
No. Observations:                 578   AIC:                             9228.
Df Residuals:                     568   BIC:                             9272.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -57.7831      4.578    -12.622      0.000     -66.775     -48.791
x2             0.0532      0.038      1.409      0.159      -0.021       0.127
x3            -0.0006    7.4e-05     -8.479      0.000      -0.001      -0.000
x4             0.0005      0.000      4.046      0.000       0.000       0.001
x5            -1.0543      2.597     -0.406      0.685      -6.156       4.047
x6             6.3747      1.484      4.297      0.000       3.461       9.289
x7             5.8239      0.833      6.992      0.000       4.188       7.460
x8             1.5927      0.495      3.219      0.001       0.621       2.565
x9            11.4319      0.916     12.480      0.000       9.633      13.231
x10            0.5026      0.329      1.526      0.128      -0.145       1.150
==============================================================================
Omnibus:                      961.012   Durbin-Watson:                   2.035
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           721502.361
Skew:                           9.699   Prob(JB):                         0.00
Kurtosis:                     174.995   Cond. No.                     4.24e+05
==============================================================================

P values:
[2.28315196e-32 1.59435976e-01 1.96622821e-16 5.93207264e-05
 6.84967590e-01 2.04071204e-05 7.62047898e-12 1.35894063e-03
 9.25140917e-32 1.27678600e-01]
```

Figure 9: t-test and p-values for #GoHawks(after adding new features)

t-test and p-values can be seen in Figure 9, 10, 11, 12, 13, 14. x1-x10 represents Number of tweets, Total number of retweets, Sum of the number of followers of the users, Maximum number of followers of the users, Time of the day, Url number, Author count, Mention count, Ranking score, Number of hashtags, respectively. From p-values, we can get the significance of each feature for every hashtag, i.e., The feature with smaller p-value has higher significance.

Question 5:

The features that we explored are "number of tweets", "sum of favorites count", "max number of favorite count", "ranking score" and "sum of friends count". P-values are printed for each feature. The three features with smallest p-values are chose and their scatter plots are plotted. They all exhibit a linear relationship with label. All the regression coefficients highly agree with the trends in the plots.

For hashtag of tweets_#gohawks, RMSE=869.425, p_values=[$3.25025270e-26$, $6.92524349e-01$, $7.21598480e-01$, $3.90685740e-26$, $8.97754600e-02$].

```
hashtag:  #GoPatriots
mse:  8182.92723648718
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.892
Model:                            OLS   Adj. R-squared:                  0.890
Method:                 Least Squares   F-statistic:                     466.4
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          3.43e-265
Time:                        02:48:10   Log-Likelihood:                -3401.3
No. Observations:                 574   AIC:                             6823.
Df Residuals:                     564   BIC:                             6866.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -10.8257      2.071     -5.227      0.000     -14.893      -6.758
x2            -1.7735      0.133    -13.343      0.000      -2.035      -1.512
x3            -0.0001      0.000     -0.656      0.512      -0.000       0.000
x4             0.0002      0.000      1.039      0.299      -0.000       0.001
x5            -0.2923      0.302     -0.967      0.334      -0.886       0.301
x6            10.0343      0.703     14.273      0.000       8.653      11.415
x7            -5.3318      0.437    -12.209      0.000      -6.190      -4.474
x8             5.2757      0.381     13.859      0.000       4.528       6.023
x9             3.0621      0.364      8.407      0.000       2.347       3.778
x10            0.8347      0.318      2.628      0.009       0.211       1.459
==============================================================================
Omnibus:                      380.028   Durbin-Watson:                   2.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            51351.092
Skew:                           1.936   Prob(JB):                         0.00
Kurtosis:                      49.175   Cond. No.                     2.15e+05
==============================================================================

P values:
[2.42531988e-07 1.71524635e-35 5.12274771e-01 2.99380437e-01
 3.33928751e-01 1.10860701e-39 1.39366481e-30 8.50135966e-38
 3.45531072e-16 8.83355142e-03]
```

Figure 10: t-test and p-values for #GoPatriots(after adding new features)

```
hashtag:  #NFL
mse:  163901.0871233339
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.791
Model:                            OLS   Adj. R-squared:                  0.788
Method:                 Least Squares   F-statistic:                     218.3
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          1.30e-188
Time:                        02:48:45   Log-Likelihood:                -4350.3
No. Observations:                 586   AIC:                             8721.
Df Residuals:                     576   BIC:                             8764.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1            -3.6700      1.471     -2.495      0.013      -6.560      -0.780
x2            -0.0836      0.054     -1.535      0.125      -0.191       0.023
x3         -1.897e-05   2.26e-05     -0.838      0.402   -6.34e-05    2.55e-05
x4          1.996e-05   2.87e-05      0.694      0.488   -3.65e-05    7.64e-05
x5            -1.4648      1.918     -0.764      0.445      -5.232       2.302
x6             0.1602      0.139      1.156      0.248      -0.112       0.432
x7            -3.5006      0.311    -11.240      0.000      -4.112      -2.889
x8             3.0007      0.595      5.046      0.000       1.833       4.169
x9             0.6479      0.305      2.121      0.034       0.048       1.248
x10            1.1603      0.082     14.098      0.000       0.999       1.322
==============================================================================
Omnibus:                      739.263   Durbin-Watson:                   2.133
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           117499.143
Skew:                           6.109   Prob(JB):                         0.00
Kurtosis:                      71.286   Cond. No.                     4.93e+05
==============================================================================

P values:
[1.28891993e-02 1.25353022e-01 4.02386272e-01 4.87828794e-01
 4.45328518e-01 2.48123923e-01 1.22386555e-26 6.06489007e-07
 3.43140665e-02 5.46318968e-39]
```

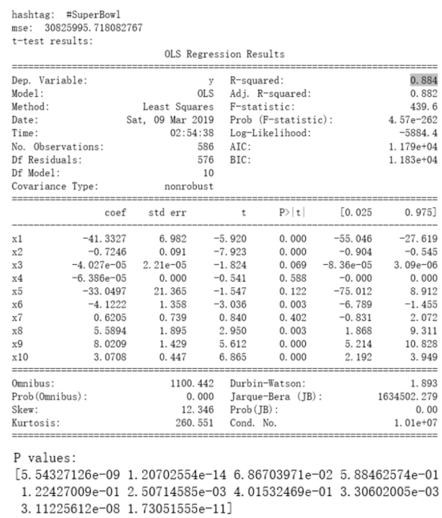
Figure 11: t-test and p-values for #NFL(after adding new features)

```
hashtag:  #Patriots
mse:  2922872.020846646
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.819
Model:                            OLS   Adj. R-squared:                  0.816
Method:                 Least Squares   F-statistic:                     260.4
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          2.69e-206
Time:                        02:49:46   Log-Likelihood:                -5194.7
No. Observations:                 586   AIC:                         1.041e+04
Df Residuals:                     576   BIC:                         1.045e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -60.7767      4.548    -13.365      0.000     -69.709     -51.845
x2            -0.2461      0.046     -5.395      0.000      -0.336      -0.157
x3             0.0002   5.85e-05      2.656      0.008    4.05e-05       0.000
x4            -0.0003   9.89e-05     -2.909      0.004      -0.000   -9.35e-05
x5            -8.5936      7.007     -1.226      0.221     -22.355       5.168
x6            -4.6197      1.575     -2.933      0.003      -7.713      -1.526
x7             1.1840      0.954      1.242      0.215      -0.689       3.057
x8             6.5543      0.849      7.720      0.000       4.887       8.222
x9            11.0991      0.857     12.949      0.000       9.416      12.783
x10            3.4480      0.382      9.038      0.000       2.699       4.197
==============================================================================
Omnibus:                     1079.383   Durbin-Watson:                   1.835
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1241634.834
Skew:                          11.953   Prob(JB):                         0.00
Kurtosis:                     227.233   Cond. No.                     8.20e+05
==============================================================================

P values:
[1.11975846e-35 1.00357085e-07 8.12435017e-03 3.76417503e-03
 2.20517141e-01 3.48929093e-03 2.14917809e-01 5.18295590e-14
 7.69195923e-34 2.41532463e-18]
```

Figure 12: t-test and p-values for #Patriots(after adding new features)

```
hashtag:  #SB49
mse:  12387685.991980104
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.853
Model:                            OLS   Adj. R-squared:                  0.851
Method:                 Least Squares   F-statistic:                     332.3
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          8.06e-231
Time:                        02:51:36   Log-Likelihood:                -5578.9
No. Observations:                 582   AIC:                         1.118e+04
Df Residuals:                     572   BIC:                         1.122e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -48.7024      8.068     -6.037      0.000     -64.549     -32.856
x2             0.4494      0.095      4.707      0.000       0.262       0.637
x3             0.0001   1.74e-05      7.673      0.000    9.92e-05       0.000
x4            -0.0002   4.85e-05     -4.914      0.000      -0.000      -0.000
x5            -9.4157     11.708     -0.804      0.422     -32.412      13.581
x6            -3.7549      1.397     -2.688      0.007      -6.498      -1.012
x7            -4.5307      1.000     -4.532      0.000      -6.494      -2.567
x8             7.4970      0.786      9.536      0.000       5.953       9.041
x9             8.4229      1.673      5.035      0.000       5.137      11.709
x10            3.8603      0.455      8.489      0.000       2.967       4.753
==============================================================================
Omnibus:                     1206.482   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          2386925.090
Skew:                          15.361   Prob(JB):                         0.00
Kurtosis:                     315.228   Cond. No.                     6.35e+06
==============================================================================

P values:
[2.83219840e-09 3.15931275e-06 7.27929809e-14 1.16707917e-06
 4.21621721e-01 7.38998313e-03 7.12437984e-06 4.19384529e-20
 6.40522325e-07 1.79154234e-16]
```

Figure 13: t-test and p-values for #SB49(after adding new features)



```
hashtag:  #SuperBowl
mse:  30825995.718082767
t-test results:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.884
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     439.6
Date:                Sat, 09 Mar 2019   Prob (F-statistic):          4.57e-262
Time:                        02:54:38   Log-Likelihood:                -5884.4
No. Observations:                 586   AIC:                         1.179e+04
Df Residuals:                     576   BIC:                         1.183e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1           -41.3327      6.982     -5.920      0.000     -55.046     -27.619
x2            -0.7246      0.091     -7.923      0.000      -0.904      -0.545
x3         -4.027e-05   2.21e-05     -1.824      0.069   -8.36e-05    3.09e-06
x4         -6.386e-05      0.000     -0.541      0.588      -0.000       0.000
x5           -33.0497     21.365     -1.547      0.122     -75.012       8.912
x6            -4.1222      1.358     -3.036      0.003      -6.789      -1.455
x7             0.6205      0.739      0.840      0.402      -0.831       2.072
x8             5.5894      1.895      2.950      0.003       1.868       9.311
x9             8.0209      1.429      5.612      0.000       5.214      10.828
x10            3.0708      0.447      6.865      0.000       2.192       3.949
==============================================================================
Omnibus:                     1100.442   Durbin-Watson:                   1.893
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1634502.279
Skew:                          12.346   Prob(JB):                         0.00
Kurtosis:                     260.551   Cond. No.                     1.01e+07
==============================================================================

P values:
[5.54327126e-09 1.20702554e-14 6.86703971e-02 5.88462574e-01
 1.22427009e-01 2.50714585e-03 4.01532469e-01 3.30602005e-03
 3.11225612e-08 1.73051555e-11]
```

Figure 14: t-test and p-values for #SuperBowl(after adding new features)

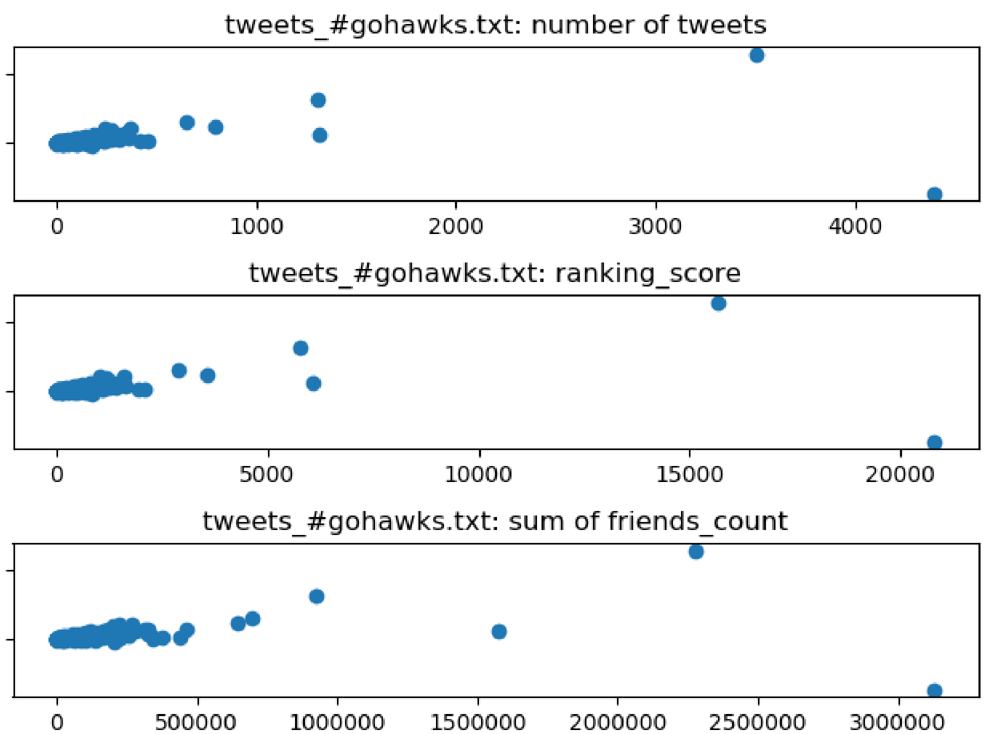


Figure 15: Predictant versus value of that feature (tweets_#gohawks)

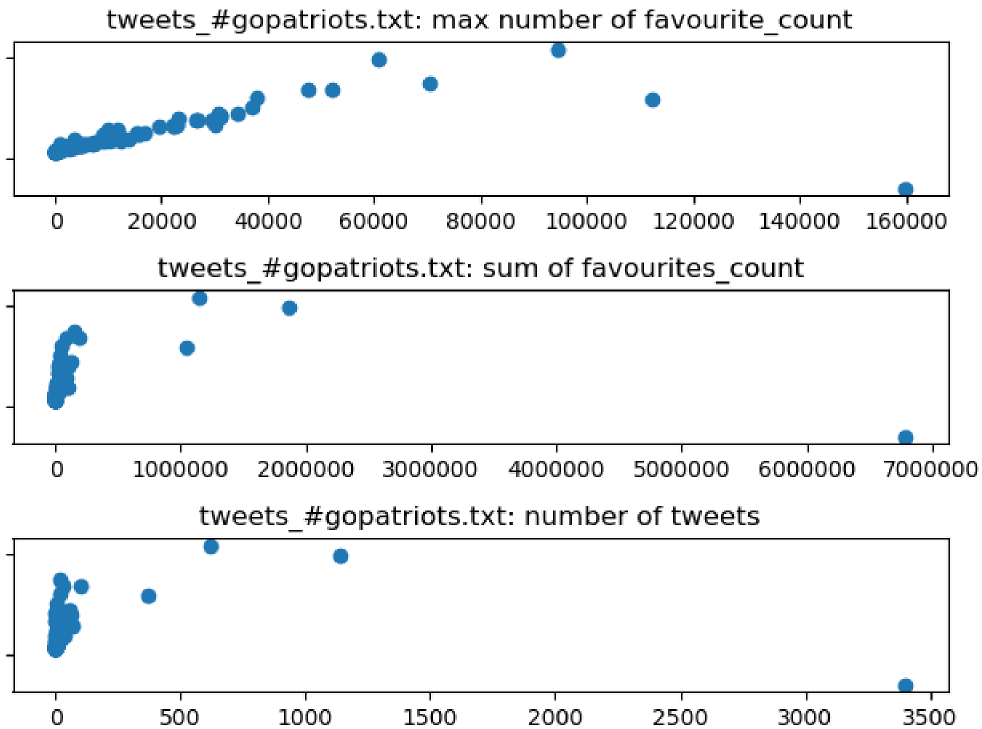


Figure 16: Predictant versus value of that feature (tweets_#gopatriots.txt)

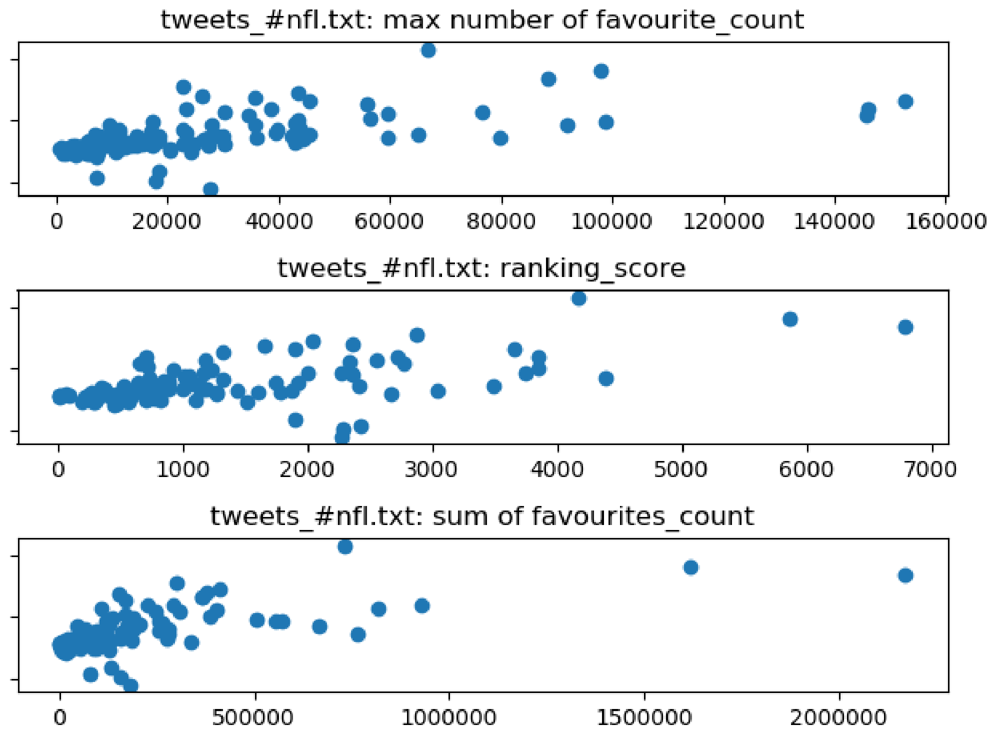


Figure 17: Predictant versus value of that feature (tweets_#nfl.txt)

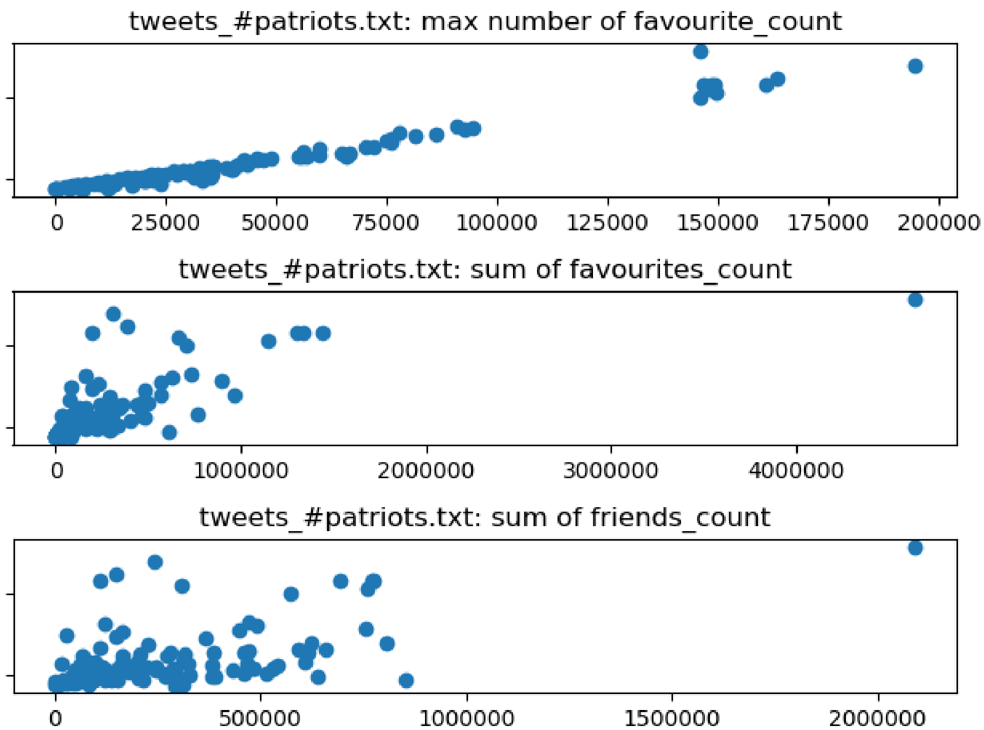


Figure 18: Predictant versus value of that feature (tweets_#patriots.txt)

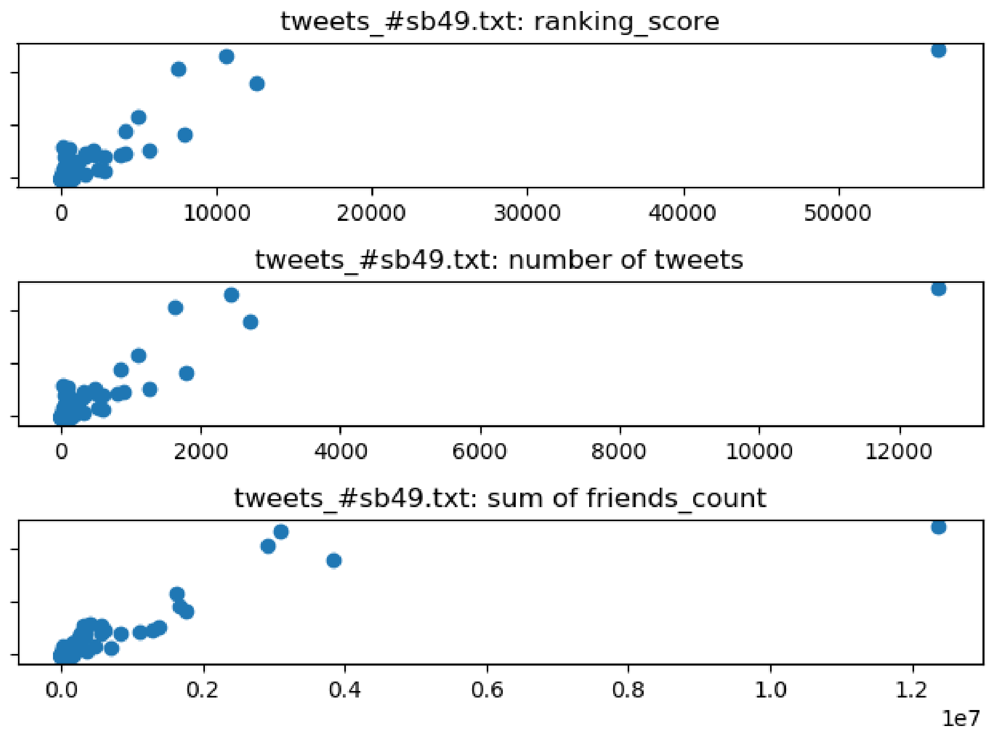


Figure 19: Predictant versus value of that feature (tweets_#sb49.txt)

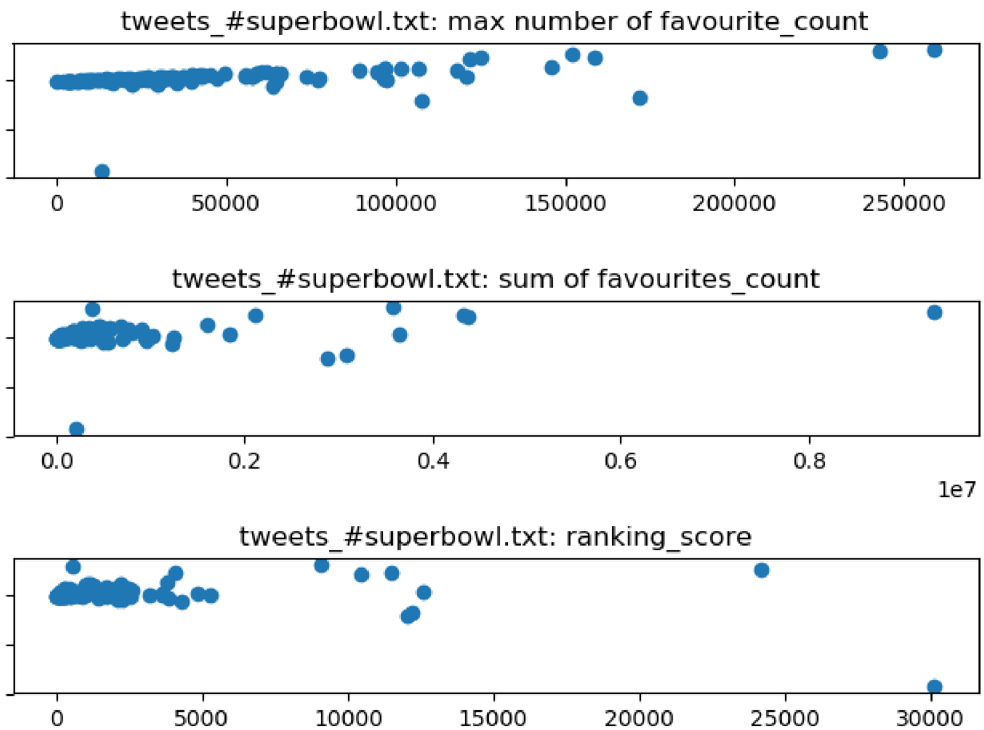


Figure 20: Predictant versus value of that feature (tweets_#superbowl.txt)

The three most important features are: number of tweets, ranking_score, sum of friends_count. The scatter plot can be found at 15.

For hashtag of tweets_#gopatriots, RMSE=331.184, p_values=[$6.71573022e-01$, $9.22773132e-03$, $9.01922321e-07$, $9.42369949e-01$, $8.95680420e-01$]. The three most important features are: max number of favourite_count, sum of favourites_count, number of tweets. The scatter plot can be found at 16.

For hashtag of tweets_#nfl, RMSE=452.804, p_values=[$1.97628176e-04$, $8.38477484e-05$, $9.40719365e-09$, $5.34746493e-05$, $8.63316750e-01$]. The three most important features are: max number of favourite_count, ranking_score, sum of favourites_count. The scatter plot can be found at 17.

For hashtag of tweets_#patriots, RMSE=841.246, p_values=[$8.13914091e-02$, $7.91170792e-08$, $6.03920658e-10$, $1.21197912e-01$, $6.33201420e-02$]. The three most important features are: max number of favourite_count, sum of favourites_count, sum of friends_count. The scatter plot can be found at 18.

For hashtag of tweets_#sb49, RMSE=4561.116, p_values=[$6.29387464e-16$, $7.17424978e-01$, $5.46720292e-02$, $2.99497856e-16$, $4.33849037e-13$]. The three most important features are: ranking_score, number of tweets, sum of friends_count. The scatter plot can be found at 19.

For hashtag of tweets_#superbowl, RMSE=17719.564, p_values=[$1.59756651e-03$, $5.55823802e-05$, $1.37347540e-07$, $1.50896890e-03$, $5.02659534e-01$]. The three most important features are: max number of favourite_count, sum of favourites_count, ranking_score. The scatter plot can be found at 20.

Based on the observations above, it is found that for different hashtags, we obtain different important features. But in general, "sum of favorites count", "max number of favorite count", and "sum of friends count" are the three most important attributes for prediction. Obviously, if a tweet is liked by a lot of people, it will retweet more compared with other tweets. Also, if a user has many friends in tweet, it will increase the probability of retweet. Number of tweets per hour and ranking score seems less important in these procedures.

1.4 Piece-wise linear regression

Question 6:

Table 3: MSE and R2 Score for tweets_#gohawks

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	3778766.452	296932.872	36309.293
R2 Score	-356.419	-3.276	0.215

Table 4: MSE and R2 Score for tweets_#gopatriots

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	5026.292	27119.285	217.526
R2 Score	-0.645	-1.811	-0.327

Table 5: MSE and R2 Score for tweets_#nfl

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	19998.186	84296.536	16554.360
R2 Score	0.522	-1.207	0.722

All the results can be found at 3, 4, 5, 6, 7, 8.

Among all the hashtags, the MSE of predictions during the event is much larger than that of the other two periods. It can explained by the fact that the number of tweets during the event is huge. Even a minor 1% prediction error could lead to a large absolute MSE. Also, 12 hours' training period is less than the first period and the third period. All the above reasons could lead to such a result.

Question 7:
The result can be found at 9. Comparing the large base number of the data, such absolute error is acceptable. It is proved that a linear-wise model is a better fit for training such kind of data who has different shapes of distributions over certain periods.

1.5 Nonlinear regressions

1.5.1 Ensemble methods

Question 8:
The best parameters set found for RandomForestRegressor is: max_depth=200, max_features=sqrt, min_samples_leaf=2, min_samples_split=2, n_estimators=200. Its mean testing square error is 364321235.359.

The best parameters set found for GradientBoostingRegressor is: max_depth=10, max_features=sqrt, min_samples_leaf=1, min_samples_split=10, n_estimators=1000. Its mean testing square error is 439464153.590.

It seems that both models have smaller testing MSE comparing to that of the linear regression model. However, their errors are still quite large compared to that of the piece-wise linear regression model. The possible reason for this performance is probably due to the fact that the data has a different distribution over the three periods. Also, RandomForestRegressor exhibited a better performance than GradientBoostingRegressor.

Table 6: MSE and R2 Score for tweets_#patriots

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	270688.971	481731.719	7648.002
R2 Score	-2.659	0.476	0.721

Table 7: MSE and R2 Score for tweets_#sb49

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	8540.093	6073893.915	474622.911
R2 Score	0.852	-0.534	0.345

Question 9:

Ensemble methods have smaller testing MSE in comparson to that of the orginal linear regression model.

Question 10:

The best parameters set found on development set before $02/01/8:00AM$ is :max_depth=200, max_features=sqrt, min_samples_leaf=2, min_samples_split=2, n_estimators=400. Its mean testing square error is 7858906.446.

The best parameters set found on development set in between is: max_depth=10, max_features=auto, min_samples_leaf=1, min_samples_split=10 and n_estimators=200. Its mean testing square error is 22326386151.286

The best parameters set found on development set after $02/01/8:00PM$ is : max_depth=None, max_features=sqrt, min_samples_leaf=1, min_samples_split=2, n_estimators=1000. Its mean testing square error is 616647.490.

Both the cross-validation test error and the best parameter set have changed in comparison to those we found above. Each time period data has its own best parameter set and the performance seems better than before.

1.5.2 Neural network

Question 11:

Now we try to regress the aggregated data with MLPRegressor. We choose five different neural network architectures and the MSE of fitting the data is shown in Table 10. The best architecture we find among them is two hidden layers with 50 and 100.

Question 12:

This time we use StandardScaler to scale the data before feeding it to the best MLPRegressor. The MSE of fitting the data is 440656113.42903936 in comparison to the original 2121168208.4143672. It shows that normalization of data can improve the performance.

Question 13:

Table 8: MSE and R2 Score for tweets_#superbowl

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	996918.231	33362527.309	25247.791
R2 Score	-4.685	-0.090	0.933

Table 9: MSE and R2 Score for all aggregated data

	Before 02/01/8:00	02/01/8:00 to 8:00 PM	After 02/01/8:00 PM
MSE	7659485.231	18584408975238.746	1659689.035
R2 Score	-0.818	-0.595	0.386

1.6 Using 6x window to predict

Question 14:

Table 10: MSE of different architectures

Hidden layers	MSE
100	5648754724.645317
300	14387650189.417015
100:50	10574929262.096115
50:100	2121168208.4143672
100:100:100	4197681980.115976

2 Fan Base Prediction

Question 15:

3 Define Own Project

Question 16: