Jingzhi Yu (UID: 604514516), Haiman Duan (UID: 404406184), Mengnan Jia(UID: 804186335)
March 20, 2015

Report for Project 3 of EE 239AS

Part 1

The statistical results for each hashtag are collected and shown as the Table 1 below.

Table 1.1 Statistical Results for each hash tag

Hashtags	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
Average # of tweets/hour	193.54	38.38	279.55	499.42	1419.89	1401.25
Average #of followers of users	2393.58	1602.01	4763.33	3641.69	10230.05	9958.12
Average # of retweets	0.21	0.03	0.05	0.09	0.18	0.14

And the histogram indicating the number of tweets per hour of #nfl and #superbowl are shown below as Figure 1.1 and Figure 1.2 respectively.

Figure 1.1 Number of Tweets Per Hour of #nfl

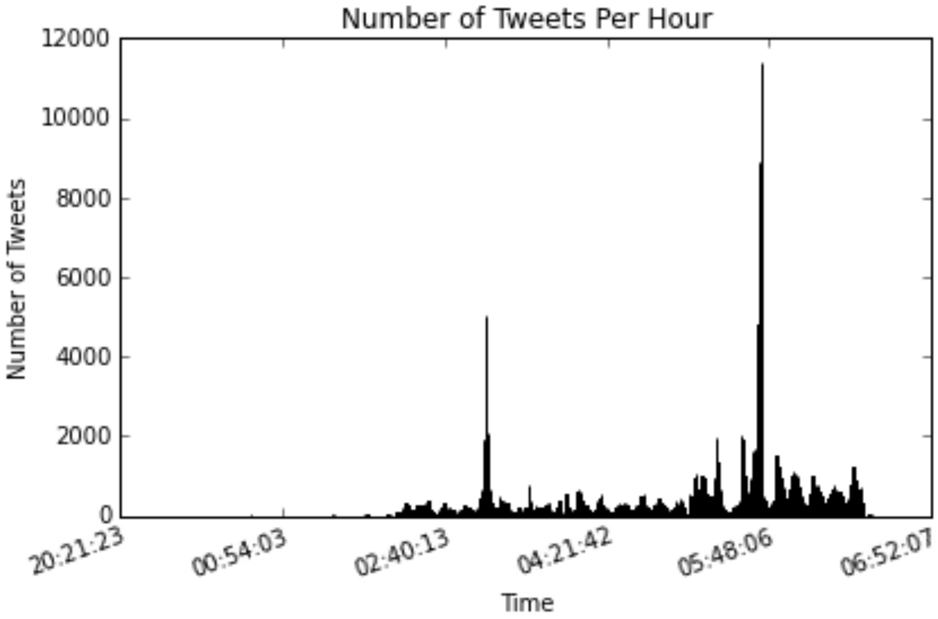
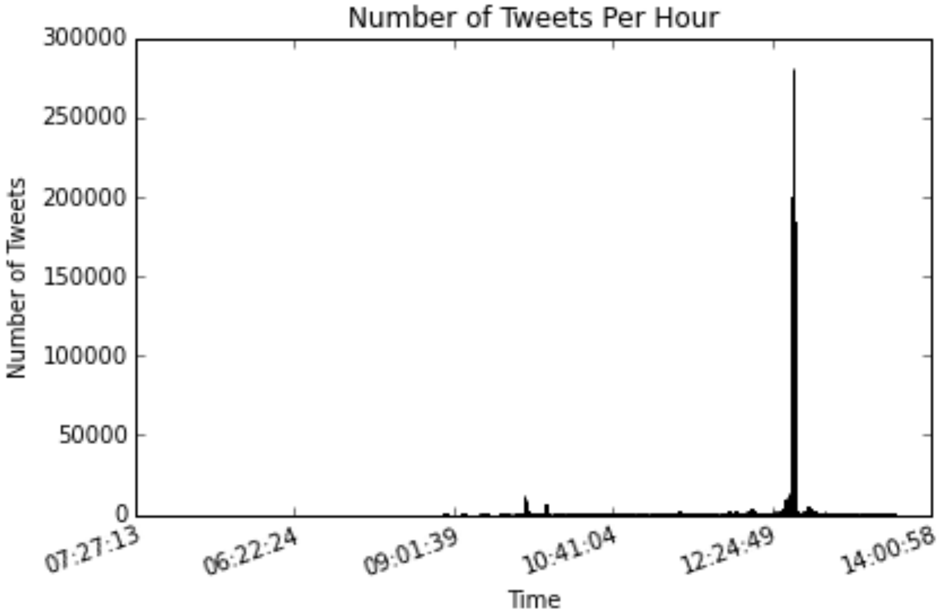


Figure 1.2 Number of Tweets Per Hour of #superbowl



Part 2

The linear regression model using 5 features to predict number of tweets in the next hour, with features extracted from tweet data in the previous hour (using features: the number of tweets, total number of retweet, sum of the number of followers posting the hashtag, maximum number of followers in users posting the hashtag and the time of the day) is constructed and evaluated in this part. All the results are stored in '3_2Linear_Regression_model.txt'. From the table we can use the parameter R-squared to evaluate the accuracy and use P>|t| to evaluate the significance of the features. we can conclude as below.

For #gohawks, 72.7% of outcome are explained by the model. The feature 'the number of tweets' and 'total number of retweet' can be considered as significant.

For #gopatriots, 63.6% of outcome are explained by the model. The feature 'the number of tweets', 'total number of retweet', 'sum of the number of followers posting the hashtag', and 'maximum number of followers in users posting the hashtag' can be considered as significant.

For #nfl, 75.9% of outcome are explained by the model. From the t-test and P-value it seems all 5 features can be considered as significant.

For #patriots, 73.3% of outcome are explained by the model. The feature 'the number of tweets', 'total number of retweet', 'sum of the number of followers posting the hashtag', and 'maximum number of followers in users posting the hashtag' can be considered as significant.

For #sb49, 86.0% of outcome are explained by the model. The feature 'the number of tweets', 'total number of retweet', 'sum of the number of followers posting the hashtag', and 'maximum number of followers in users posting the hashtag' can be considered as significant.

For #sb49, 89.2% of outcome are explained by the model. The feature 'the number of tweets', 'total number of retweet', 'sum of the number of followers posting the hashtag', and 'maximum number of followers in users posting the hashtag' can be considered as significant.

Part 3

We choose 6 features to train our model: 1. number of tweets 2. number of Retweets 3.maximum number of followers 4. favorite count. 5. number of friends 6. influential. The models are trained respectively according to 6 hashtags. The results are restored in txt files named as 'Linear_Regression_model_3#tag.txt'.

1) For #gohawks, 75.4% of outcome are explained by the model. The feature 'the number of tweets', 'maximum number of followers' and 'number of friends' can be considered as significant.

2) For #gopatriots, 60.5% of outcome are explained by the model. The feature 'the number of retweets', 'maximum number of followers' and 'number of friends' can be considered as significant.

3) For #nfl, 78.9% of outcome are explained by the model. The feature 'the number of tweets', 'number of retweets' and 'favorite count' can be considered as significant.

4) For #patriots, 72.7% of outcome are explained by the model. The feature 'the number of tweets', 'number of retweets' and 'favorite count' can be considered as significant.

Figure 3.1.1 The Prediction vs. The Number of Tweets

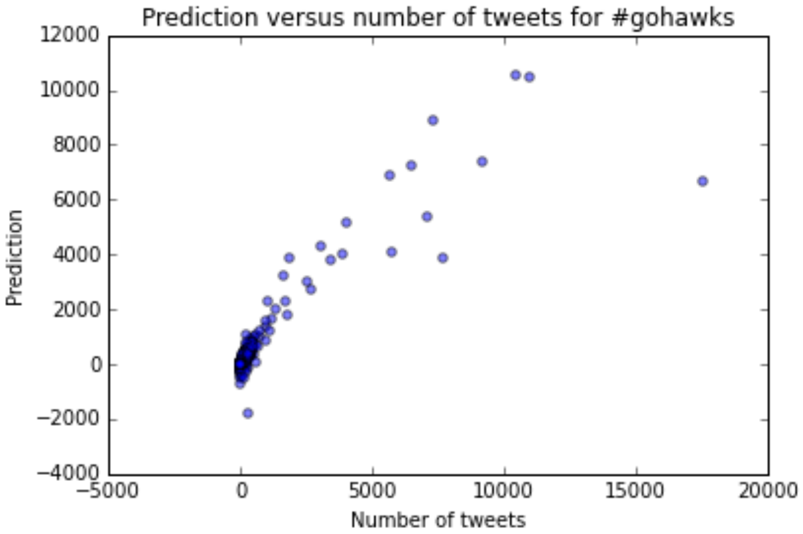


Figure 3.1.2 The Prediction vs. The Max Number of Followers

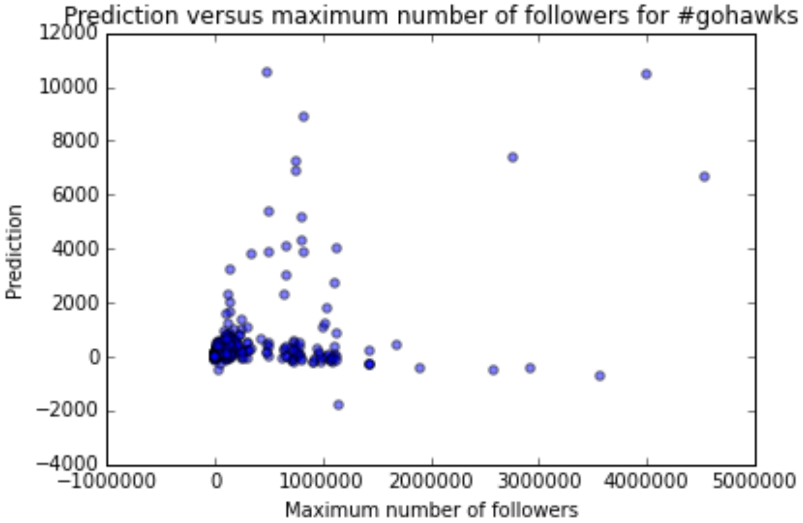


Figure 3.1.3 The Prediction vs. The Number of Friends

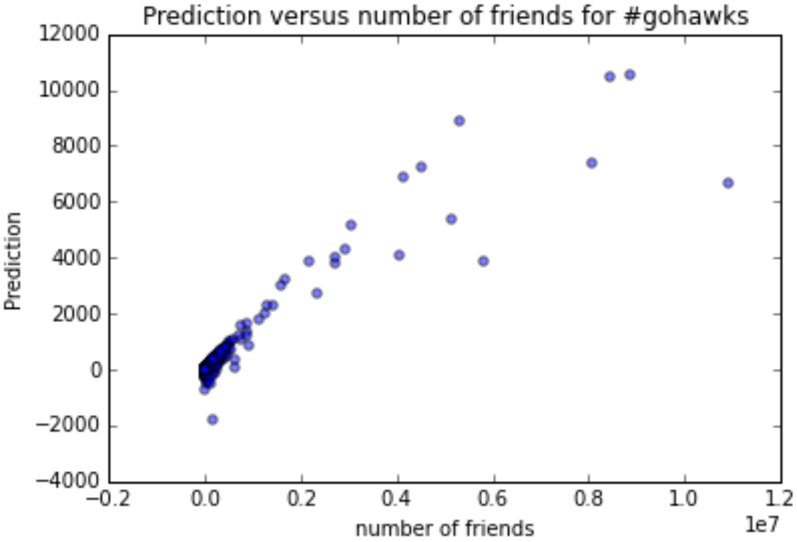


Figure 3.2.1 The Prediction vs. The Number of Retweets

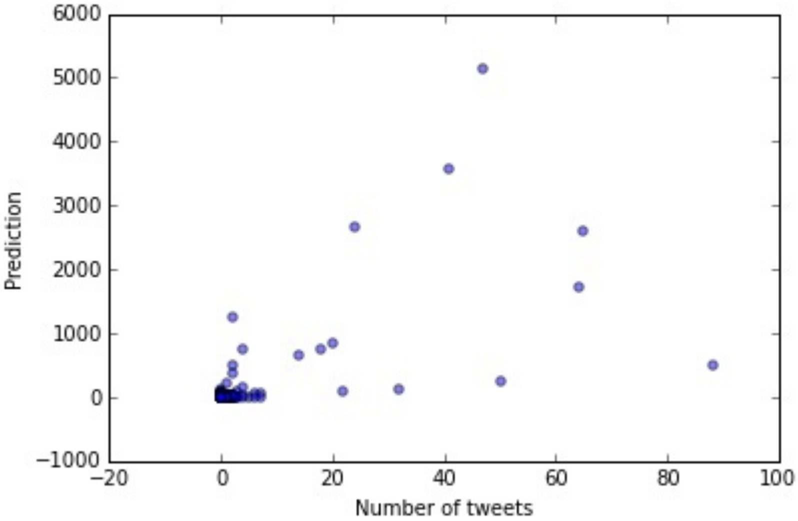


Figure 3.2.2 The Prediction vs. The Favorite Count

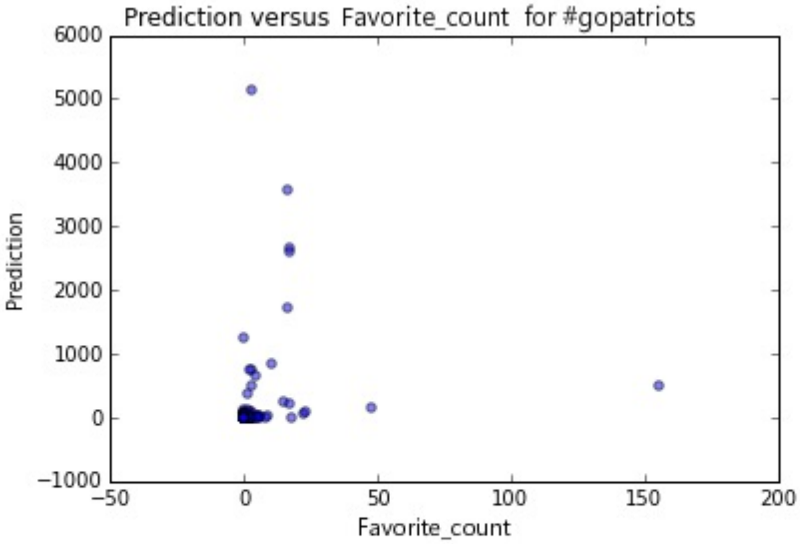


Figure 3.2.3 The Prediction vs. The Number of Friends

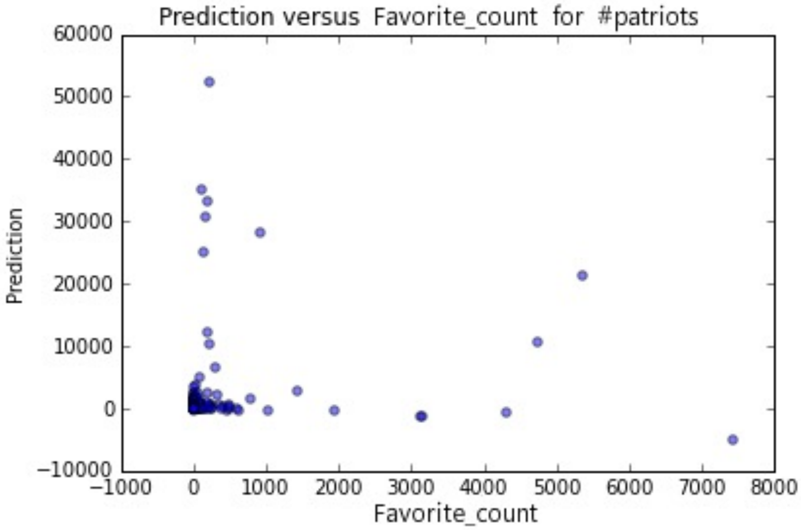


Figure 3.3.1 The Prediction vs. The Number of Tweets

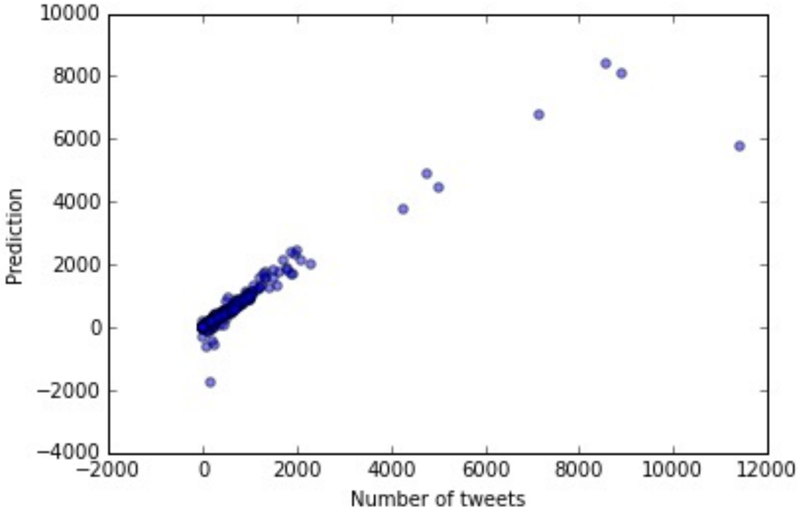


Figure 3.3.2 The Prediction vs. The Number of Retweets

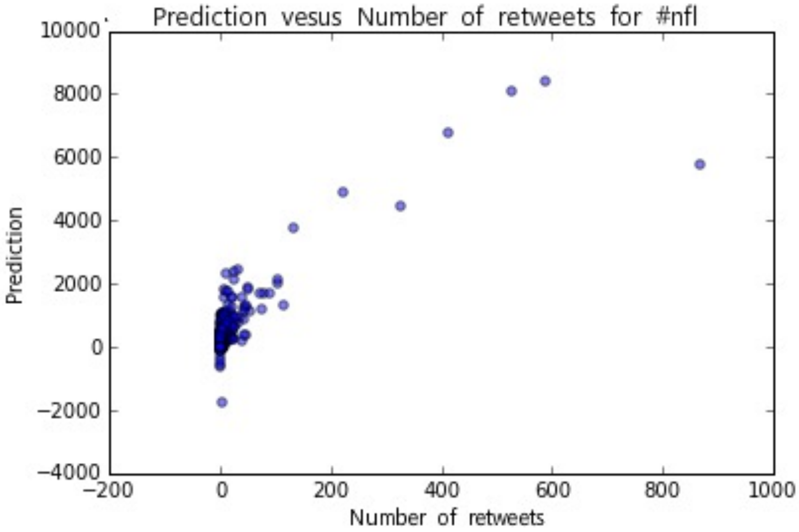


Figure 3.3.3 The Prediction vs. The Favorite Count

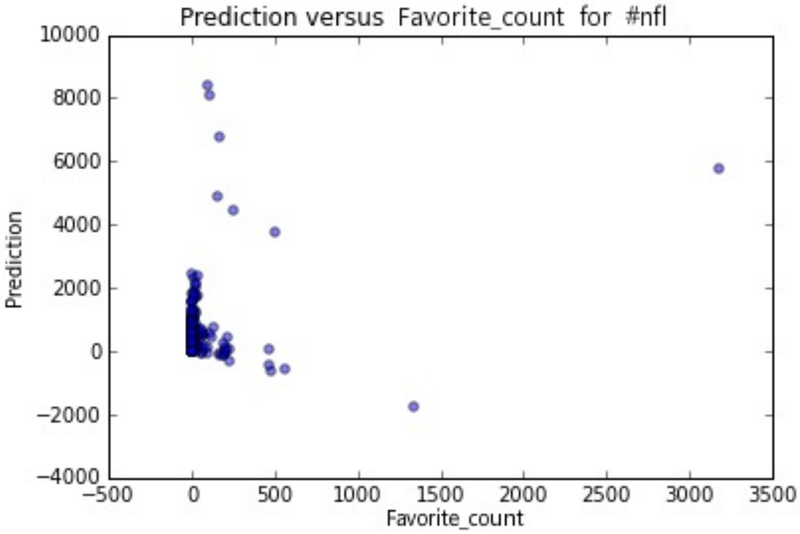


Figure 3.4.1 The Prediction vs. The Number of Tweets

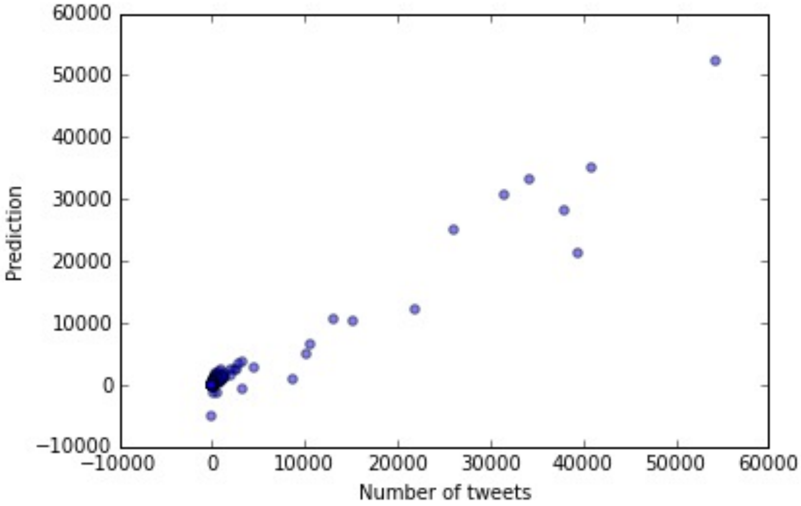


Figure 3.4.2 The Prediction vs. The Number of Retweets

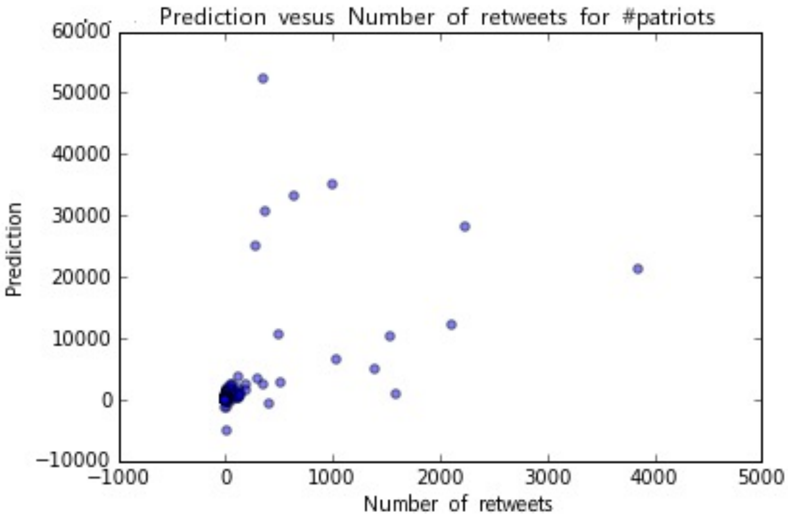
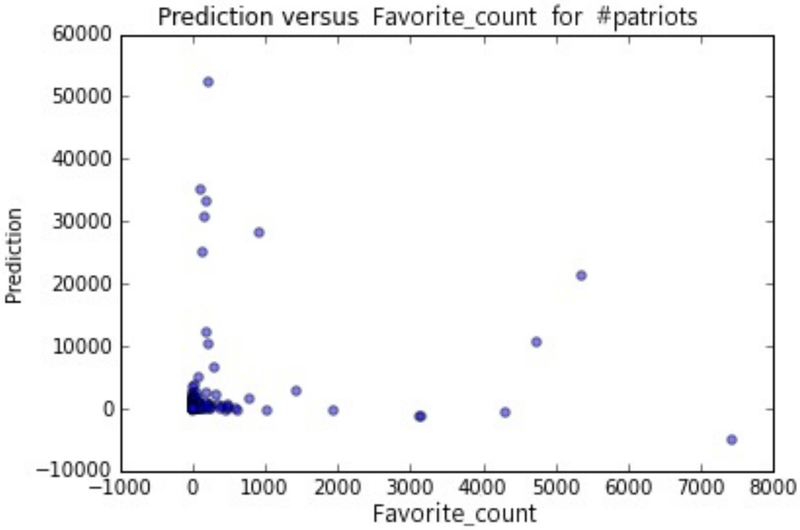


Figure 3.4.3 The Prediction vs. The Favorite Count



5) For #sb49, 86.8% of outcome are explained by the model. The feature 'the number of tweets', 'number of retweets' and 'favorite count' can be considered as significant.

Figure 3.5.1 The Prediction vs. The Number of Tweets

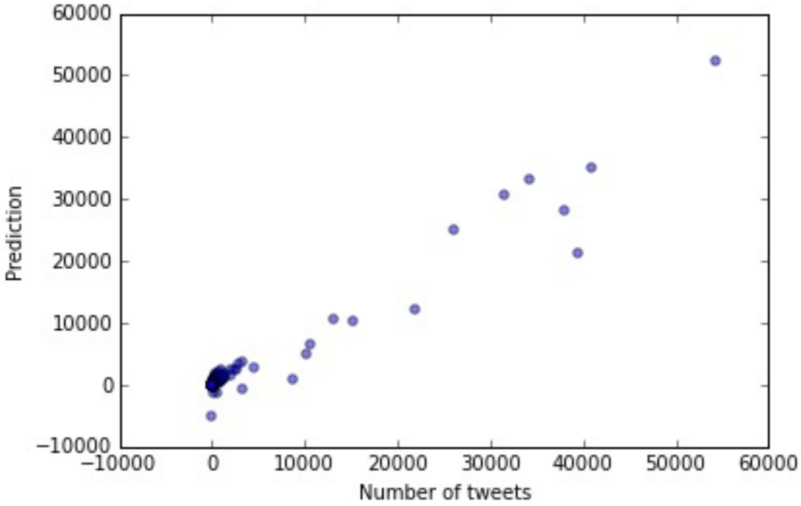


Figure 3.5.2 The Prediction vs. The Number of Retweets

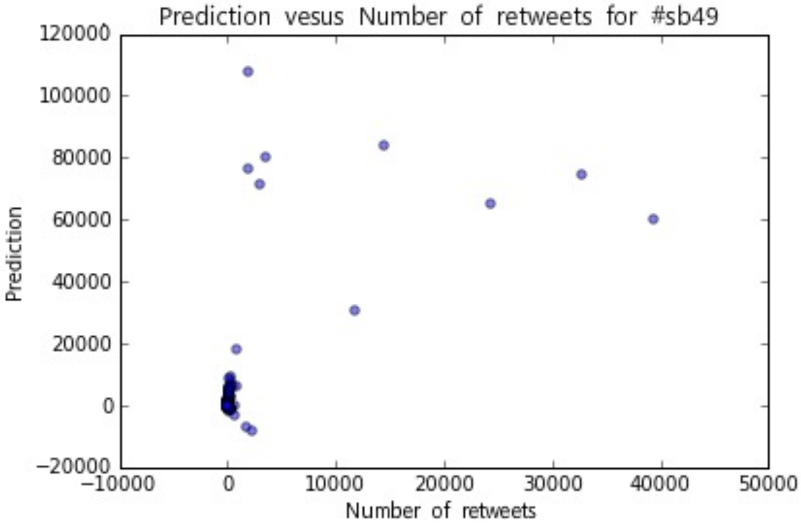
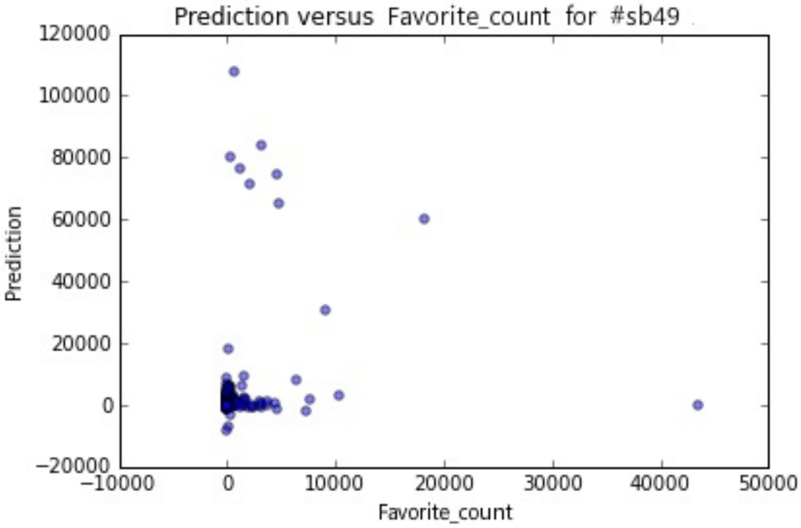


Figure 3.5.3 The Prediction vs. The Favorite Count



6) For #superbowl, 92.6% of outcome are explained by the model. The feature 'the number of tweets', 'number of retweets' and 'favorite count' can be considered as significant.

Figure 3.6.1 The Prediction vs. The Number of Tweets

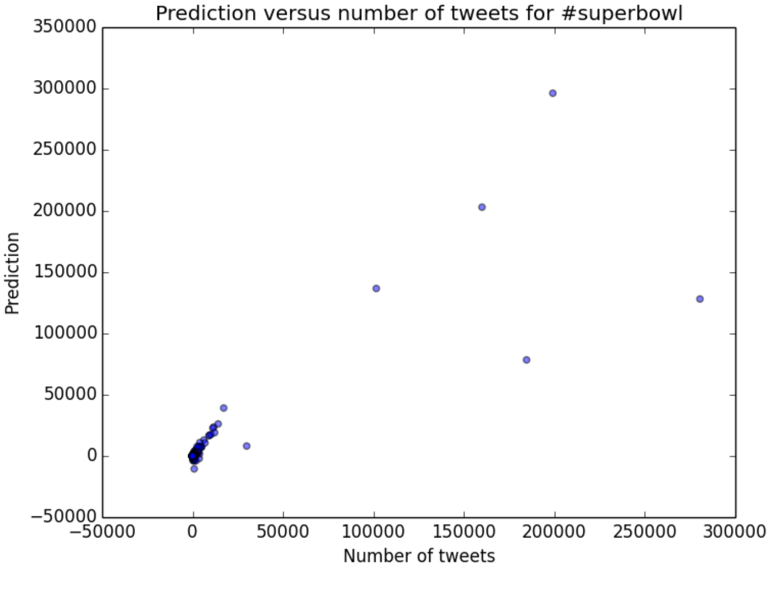


Figure 3.6.2 The Prediction vs. The Number of Retweets

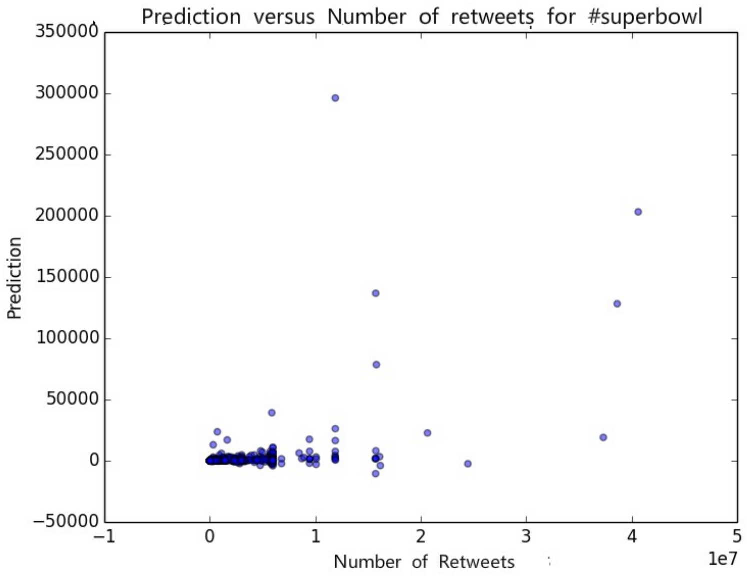
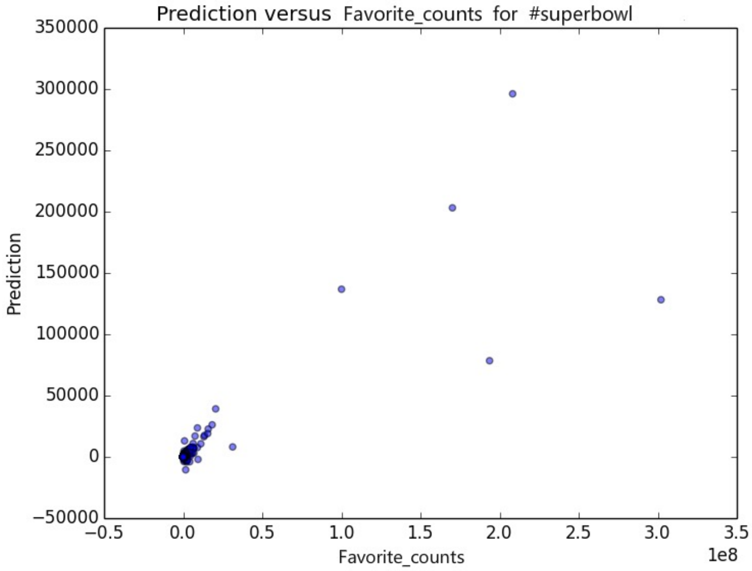


Figure 3.6.3 The Prediction vs. The Favorite Count



Part 4

The cross-validation results are stored in txt files named as 'cross_validation_err_#tag.txt' according to each hashtag respectively.

Below in Table 4.1 are the average error in different period of each hashtag.

And all the cross-validation errors according to each hashtag are also shown in the tables below.

Table 4.1 The average error in different period of each hashtag

Hashtags	#gohawks	#gopatriots	#nfl	#patriots	#sb49	#superbowl
Before Feb, 1st, 8am	59.551	1.271	124.092	316.619	863.644	1606.758
Between Feb, 1st, 8am and 8pm	0.919	0.231	55.745	84.989	34..18	106.943
After Feb 1st, 8pm	6.59	0.486	58.154	246.546	81.884	155.277

Table 4.2 The cross-validation errors for #gohawks

set#	1	2	3	4	5	6	7	8	9	10
Before Feb, 1st, 8am	11.97	29.1	19.01	29.95	56.64	19.47	241.05	29.13	75.55	83.64
Betwee n Feb, 1st, 8am and 8pm	5.8	0	0	2.07	1.05	0	0.27	0	0	0
After Feb 1st, 8pm	1.24	9.92	10.32	4.2	4.01	1.3	1.47	13.67	14.89	4.88

Table 4.3 The cross-validation errors for #gopatriots

set#	1	2	3	4	5	6	7	8	9	10
Before Feb, 1st, 8am	0.32	4.7	0.88	0.42	0.64	1.04	2.48	0.51	0.49	1.23

set#	1	2	3	4	5	6	7	8	9	10
Between Feb, 1st, 8am and 8pm	0.58	0.03	0.84	0.02	0	0	0	0.19	0.58	0.07
After Feb 1st, 8pm	0.75	0.34	1.4	0.15	0.73	0.43	0.13	0.18	0.31	0.44

Table 4.4 The cross-validation errors for #nfl

set#	1	2	3	4	5	6	7	8	9	10
Before Feb, 1st, 8am	215.62	74.27	221.03	148.16	165.63	78.2	50.66	63.38	99.53	124.44
Between Feb, 1st, 8am and 8pm	37.15	398.58	1	1.03	49.5	19.06	14.9	36.23	0	0
After Feb 1st, 8pm	64.99	63.13	68.0	60.98	47.43	45.48	54.66	45.69	73.33	57.85

Table 4.5 The cross-validation errors for #patriots

set#	1	2	3	4	5	6	7	8	9	10

set#	1	2	3	4	5	6	7	8	9	10
Before Feb, 1st, 8am	259.0	304.02	126.72	507.23	1092.67	121.46	73.77	43.75	137.2	500.37
Between Feb, 1st, 8am and 8pm	76.34	0	147.78	0	295.55	34.67	0	0	0	295.55
After Feb 1st, 8pm	155.43	83.74	688.43	512.43	757.07	50.26	28.16	47.4	110.83	31.71

Table 4.6 The cross-validation errors for #sb49

set#	1	2	3	4	5	6	7	8	9	10
Before Feb, 1st, 8am	407.18	1695.47	85.46	2296.78	71.6	1725.08	690.13	1259.88	94.41	310.45
Between Feb, 1st, 8am and 8pm	106.35	2.11	39.16	3.74	90.69	1.44	35.95	4.65	28.6	29.11
After Feb 1st, 8pm	41.5	88.06	5.43	40.17	158.1	31.92	38.71	75.3	177.6	162.05

Table 4.7 The cross-validation errors for #superbowl

set#	1	2	3	4	5	6	7	8	9	10
Before Feb, 1st, 8am	463.24	460.68	420.41	4597.62	453.3	3546.79	3400.94	565.16	1717.48	441.96
Between Feb, 1st, 8am and 8pm	11.68	0	135.01	329.54	490.85	0	0	102.35	0	0
After Feb 1st, 8pm	55.45	25.03	186.1	289.89	318.81	58.23	190.0	153.67	183.54	92.05

Part (5)

All the prediction results are stored in 'Tweets_Prediction.txt'. And they are shown in the table below.

set#	S1P1	S2P2	S3P3	S4P1	S5P1	S6P2	S7P3	S8P1	S9P2	S10P3
Prediction	120	69366	833	132	138	31141	180	7	2330	91

Table 5.1 The Prediction of The Test Sets