EE219 Project 5
Popularity Prediction on Twitter
Winter 2017

Shuang Feng

Chenkai Ling

Tianyu Deng

Part 1)
In this project, 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 is readed into python using JSON and converted into a dictionary.

For each hashtag, the average number of tweets per hour, average number of followers of users posting the tweets, and average number of retweets were calculated.

Table 1.1 Average calculations for each hashtag

Hashtag	Average number of tweets per hour	Average number of followers of users posting the tweet	Average number of retweets
#Gohawks	193.56	2393.58	0.20916252073
#Gopatriots	38.40	1602.00	0.0268374504422
#Patriots	499.70	3641.68	0.0914617337093
#NFL	279.72	4763.32	0.0509373648774
#sb49	1420.87	10230.04	0.178012965702
#SuperBowl	1400.58	9958.11	0.136685580237

In addition, "number of tweets in hour" over time for #SuperBowl and #NFL plots were generated as histograms.

The following graph is number of #SuperBowl tweets per hour.

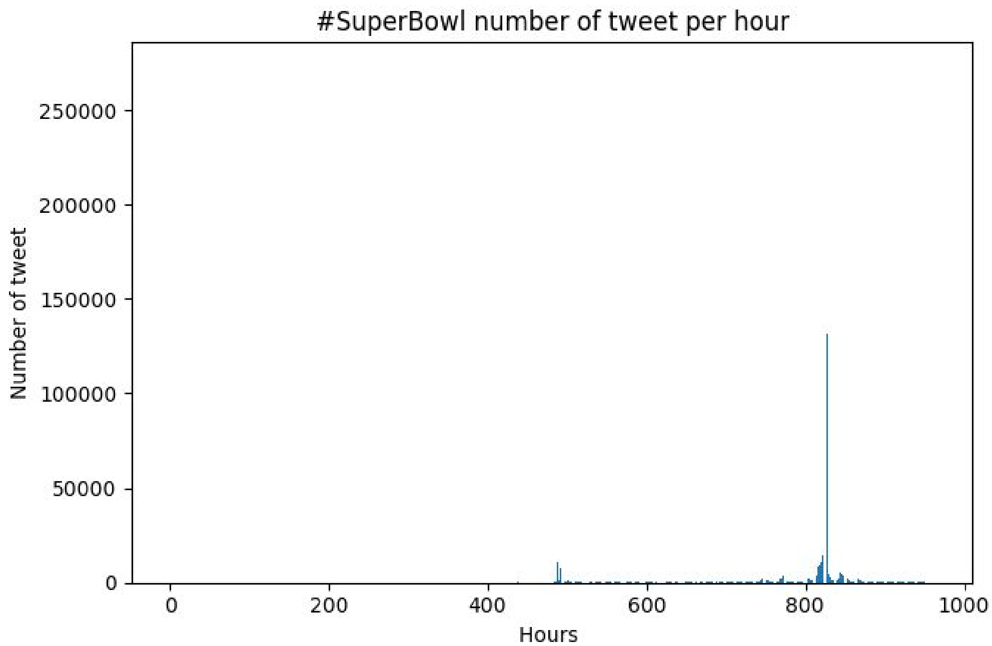


Figure 1.1 #SuperBowl number of tweet per hour

The following graph is number of #NFL tweets per hour.

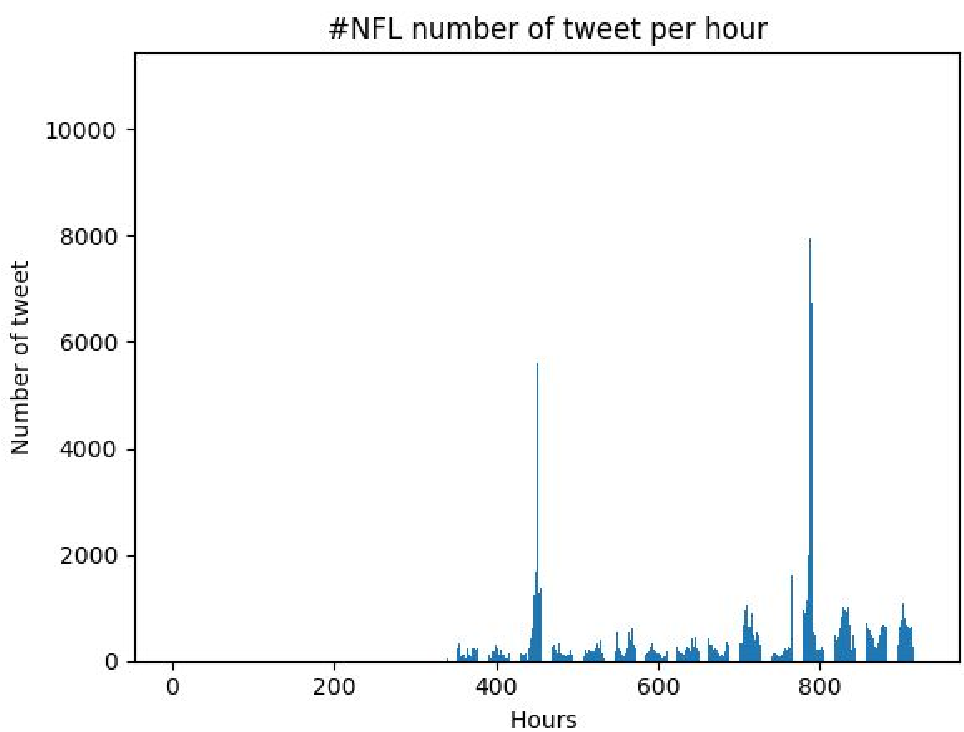


Figure 1.2 #NFL number of tweet per hour

Part 2)
In this part, we used number of tweets, total number of retweets, sum of the number of followers of the users posting the hashtag, maximum number of follower of the users posting the hashtag and time of the day as features for linear model. Each feature was extracted from tweet data in the previous hour and used to predict the tweets in the next hour.
The linear models are then be evaluated by calculating t-values and p-values for each model. Table 2.1 summarized the t-values and p-values for each linear model, as well as the coefficients for the linear models.

Table 2.1. Coefficient, t-Value and p-Value of linear model for #NFL

	t-value	p-value	coefficient
Number of tweet	12.828	0	0.8772
Number of retweet	-3.0215	0	-3.0215
Sum number of followers	5.275e-06	0.814	5.275e-06
Maximum number of followers	3.32e-05	0.038	3.28e-05
Time of the day	1.5578	0.128	1.5578

The mean error for #NFL tweet number prediction is calculated by the average sum of **|predict - actual|** and the result is 119.818.
The mean squared error for #NFL tweet number prediction is calculated by the average sum of **(predict - actual)^2** and the result is 150978.5112.
This shows that our model has poor accuracy because the mean squared error is large.

From t-value and p-value score we can summary that *number of tweet*, *number of retweet and time of the day* are the three top significant features for this linear regression model. These three all score fairly small p-value which indicates that the null hypothesis is rejected and there is a statistically significant difference and importance for these three features.

Table 2.2 Coefficient, t-value, p-value for linear model #Superbowl

	t-value	p-value	coefficient
Number of tweet	7.312	0	1.0690
Number of retweet	-25.931	0	-3.7118
Sum number of followers	2.780	0.006	0.0003
Maximum number of followers	0.566	0.572	1.187e-05
Time of the day	-0.281	0.778	-4.2360

The mean error for #Superbowl tweet number prediction is calculated by |predict - actual| and the result is 990.888.
The mean squared error for #Superbowl tweet number prediction is calculated by (predict - actual)^2 and the result is 37131116.0541.
This shows that our model has poor accuracy because the mean squared error is large.
Because of some high peak amount of tweets (part 1) the mean squared error is much higher compare to #NFL model.

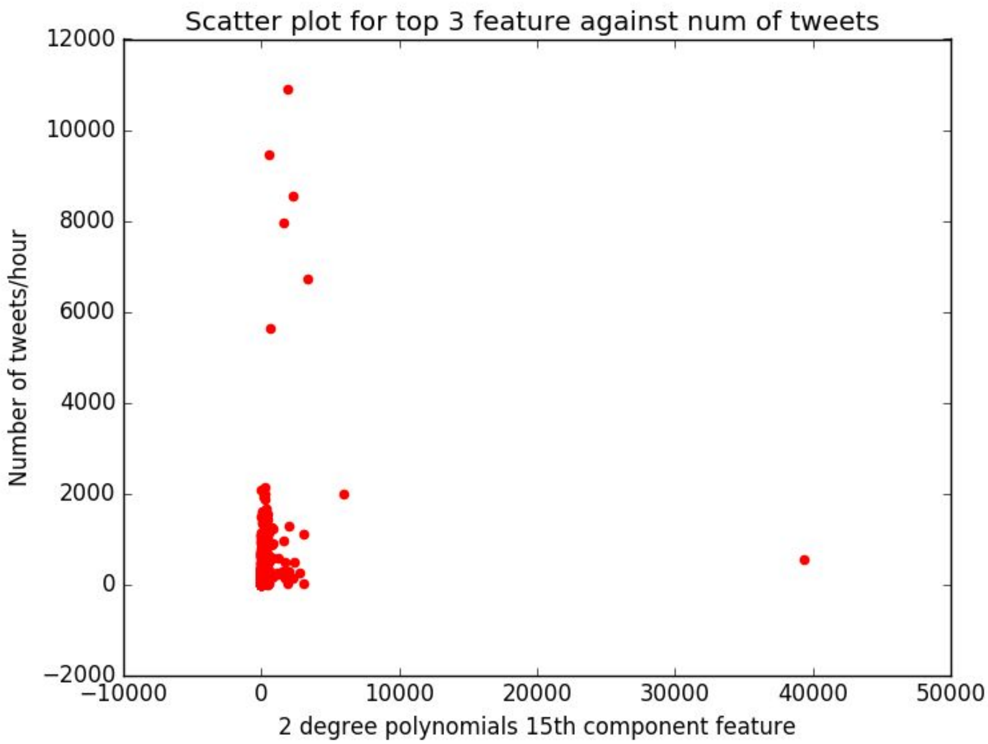
From t-value and p-value score we can summary that *number of tweet*, *number of retweet and Sum number of followers* are the three top significant features for this linear regression model. These three all score fairly small p-value which indicates that the null hypothesis is rejected and there is a statistically significant difference and importance for these three features.

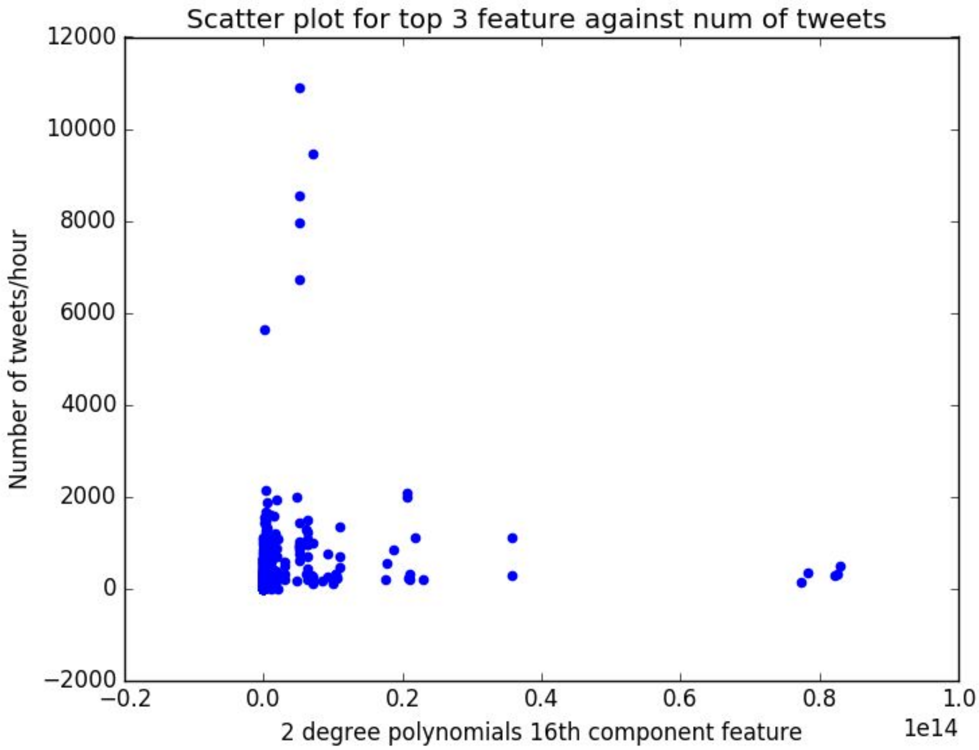
Part 3)
In this part we designed a 2-degree polynomial regression model to fit the data with same 5 features described in previous part. The evaluation results are in Table 3.1 for #Superbowl and Table 3.2 for #NFL.
Since we are using a 2-degree polynomial regression model, the original 5 features now expand to (5*(5-1))=20 features. The result shows that most of the features have low P-value which means these features are significant.

The mean error for #NFL tweet number prediction is calculated by the average sum of **|predict - actual|** and the result is 110.090932081.
The mean squared error for #NFL tweet number prediction is calculated by the average sum of **(predict - actual)^2** and the result is 112575.2007.

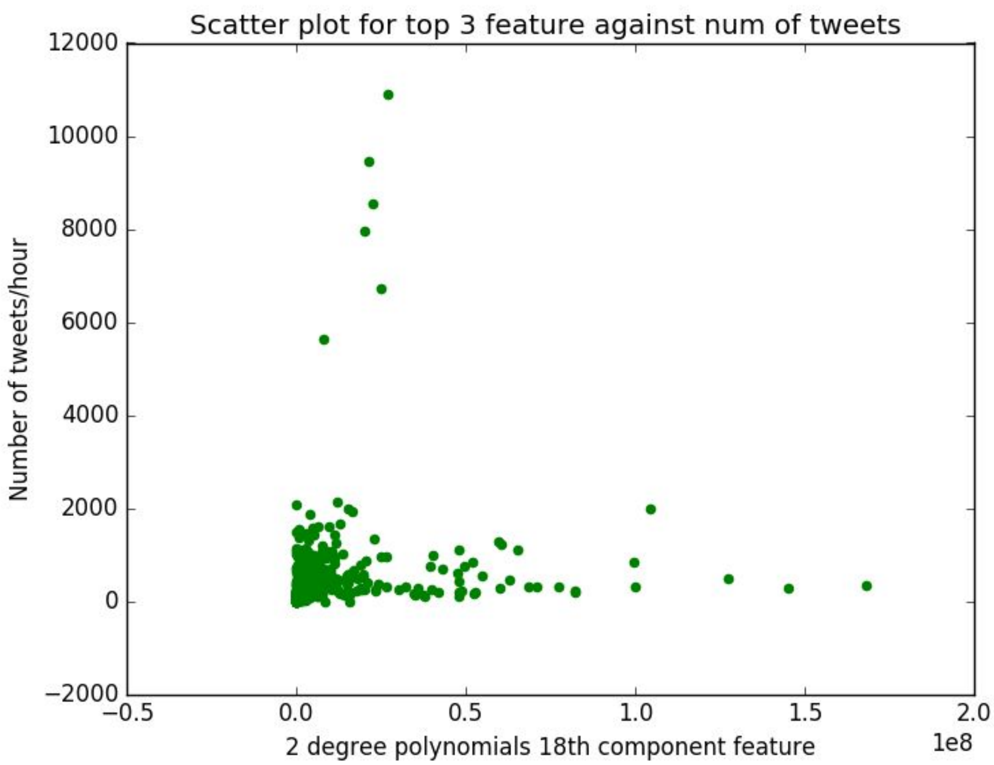
The mean error and mean error square is much lower compare to linear regression model which means that 2-degree polynomial model is a better model.

Since we are using 2-degree polynomial regression model, the feature doesn't have a real meaning. It is a combination of two of the five features to create a 2-degree feature. Three 2-degree features with lowest P-value are picked to represent the best feature and are plotted into three scatter plot.





Scatter plot for top 3 feature against num of tweets
Number of tweets/hour
2 degree polynomials 16th component feature
1e14



Scatter plot for top 3 feature against num of tweets
Number of tweets/hour
2 degree polynomials 18th component feature
1e8

Part 4)
In this part, the dataset is separated into 3 time periods and each uses one type of regression model. The feature data were splitted into 10 parts to perform 10-cross validation. Running the tests for 10 times and validate the prediction for one part at each time, the average prediction error can be calculated using the average sum of 10 test values of $|N_{predict} - N_{real}|$ over samples in remaining parts.
The results for each hashtag for each regression model during different time is reported in the table below.

Table 4.1 10-fold Cross validation error for each regression model during different time

Hashtag	Time	Regression model	Cross-validation error
NFL	Before Feb. 1, 8:00 a.m.	Lasso	73.0920817015
		Linear	73.092536311
		2 degree polynomial	87.3552500728
	Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Lasso	2602.91542869
		Linear	2630.74802911
		2 degree polynomial	8526.82151287
	After Feb. 1, 8:00 p.m.	Lasso	490.877260725
		Linear	490.883172074
		2 degree polynomial	1416.86647414

superbowl	Before Feb. 1, 8:00 a.m.	Lasso	201.220352698
		Linear	201.223049518
		2 degree polynomial	269.873331602
	Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Lasso	67335.5166259
		Linear	116824.8556
		2 degree polynomial	5023962.55976

	After Feb. 1, 8:00 p.m.	Lasso	3426.3007047
		Linear	3426.29143903
		2 degree polynomial	77448.0317539

gohawks	Before Feb. 1, 8:00 a.m.	Lasso	238.727534729
		Linear	238.729752614
		2 degree polynomial	1293.51220692
	Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Lasso	3407.41048114
		Linear	3407.56392637
		2 degree polynomial	19819.0171485
	After Feb. 1, 8:00 p.m.	Lasso	754.689136172
		Linear	762.138694728
		2 degree polynomial	14988070.5421

gopatriots	Before Feb. 1, 8:00 a.m.	Lasso	12.237420935
		Linear	12.242563713
		2 degree polynomial	156.285530725
	Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Lasso	1493.02025287
		Linear	1507.49398192
		2 degree polynomial	71069.6317489
	After Feb. 1, 8:00 p.m.	Lasso	83.3344092106
		Linear	75.6535133072
		2 degree polynomial	28388.089816

Patriots	Before Feb. 1, 8:00 a.m.	Lasso	131.421458329
		Linear	131.427060544
		2 degree polynomial	118.84468738
	Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Lasso	13842.7422816
		Linear	13842.6015685
		2 degree polynomial	219652.840789
	After Feb. 1, 8:00 p.m.	Lasso	2018.53287704
		Linear	2037.05977511
		2 degree polynomial	11704.0883616

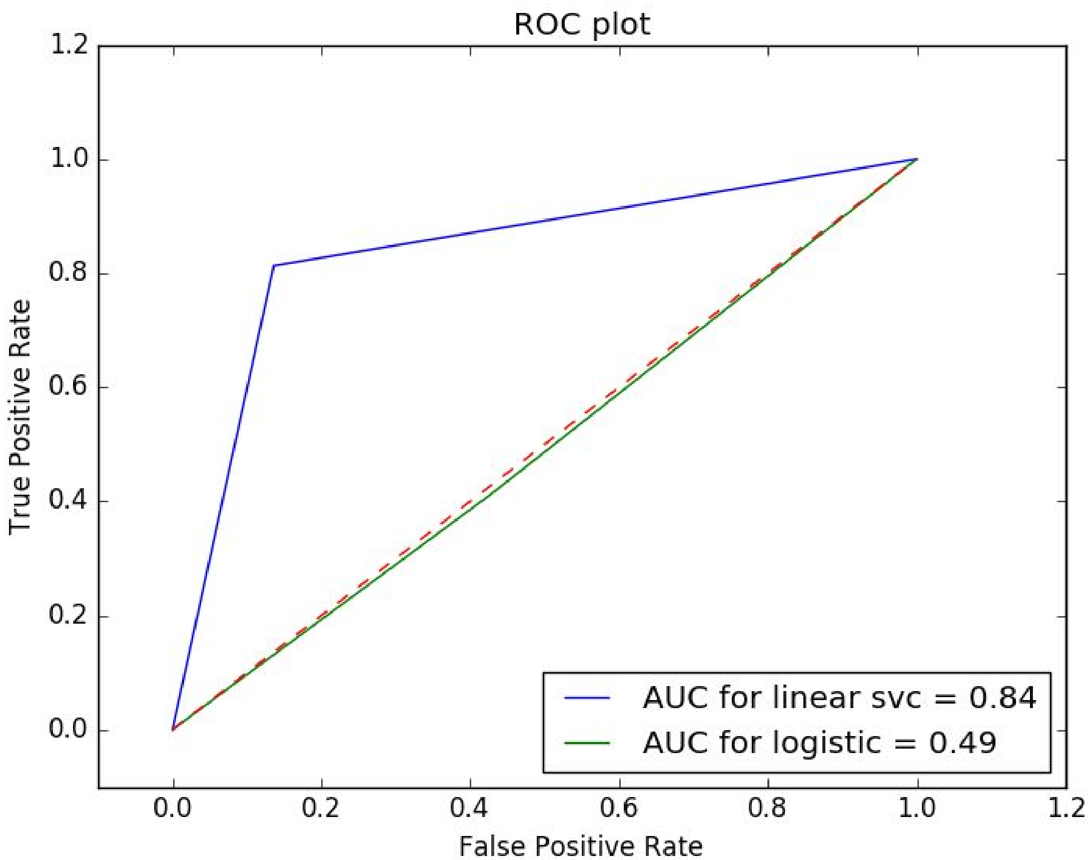
sb49	Before Feb. 1, 8:00 a.m.	Lasso	39.8855215544
		Linear	39.890592397
		2 degree polynomial	42.5512441142
	Between Feb. 1, 8:00 a.m. and 8:00 p.m.	Lasso	105285.099395
		Linear	105285.067259
		2 degree polynomial	1779463.19052
	After Feb. 1, 8:00 p.m.	Lasso	862.590887023
		Linear	862.429222041
		2 degree polynomial	44807.1718398

Part 5)
In this part, a new dataset was downloaded from the web. Each file in this dataset contains 6 hour data in the three time periods discussed in part 4. This new dataset is used to predict each following hour's number of tweets.
We use the best model from part 4, which is Lasso regression model, for each prediction. The prediction result is shown below.

	Prediction for next hour tweets number
sample1_period1.txt	81.37430633
sample2_period2.txt	-1217435.49498434
sample3_period3.txt	562.57268338
sample4_period1.txt	264.97011954
sample5_period1.txt	179.9888356
sample6_period2.txt	45775.24857894
sample7_period3.txt	35.83656331
sample8_period1.txt	81.00870557
sample9_period2.txt	1954.2352173
sample10_period3.txt	57.07136728

The negative prediction of sample2_period2 may be due to the outlier data point fed into the model.

Part 6)
In this part, a TFxIDF matrix is used to analysis the content of tweets from two different locations. The number of terms were truncated into 50 features using SVD. Then the data were fed into 2 different classification methods to predict user's location. The first classification method is linear SVC and second method is logistic regression classification. The target is user's location and the features are truncated TFIDF data from user's tweet messages. For each model, Washington and Massachusetts is labeled as 0 and 1.
The ROC curve for each classifier is shown below.



The ROC plot indicates linear SVC model is better since the curve is way above threshold. Therefore, we use linear SVC model for both this part and next part.

The result for linearSVC prediction is shown in the confusion matrix below.

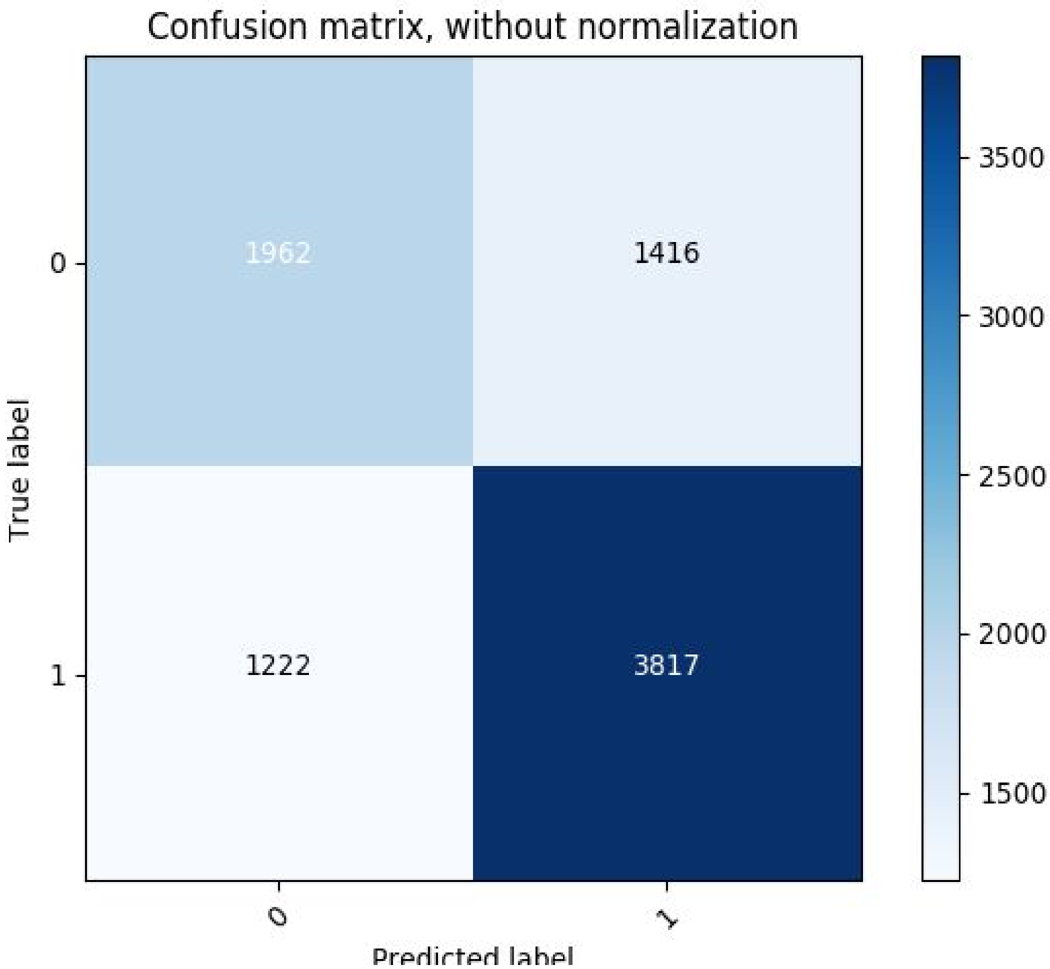


Figure6.1 confusion matrix for LinearSVC

Calculated from the confusion matrix, the precision for this model is 0.616 and the recall is 0.581. The accuracy is 0.5985.

The result for logistic classification prediction is shown in the confusion matrix below.

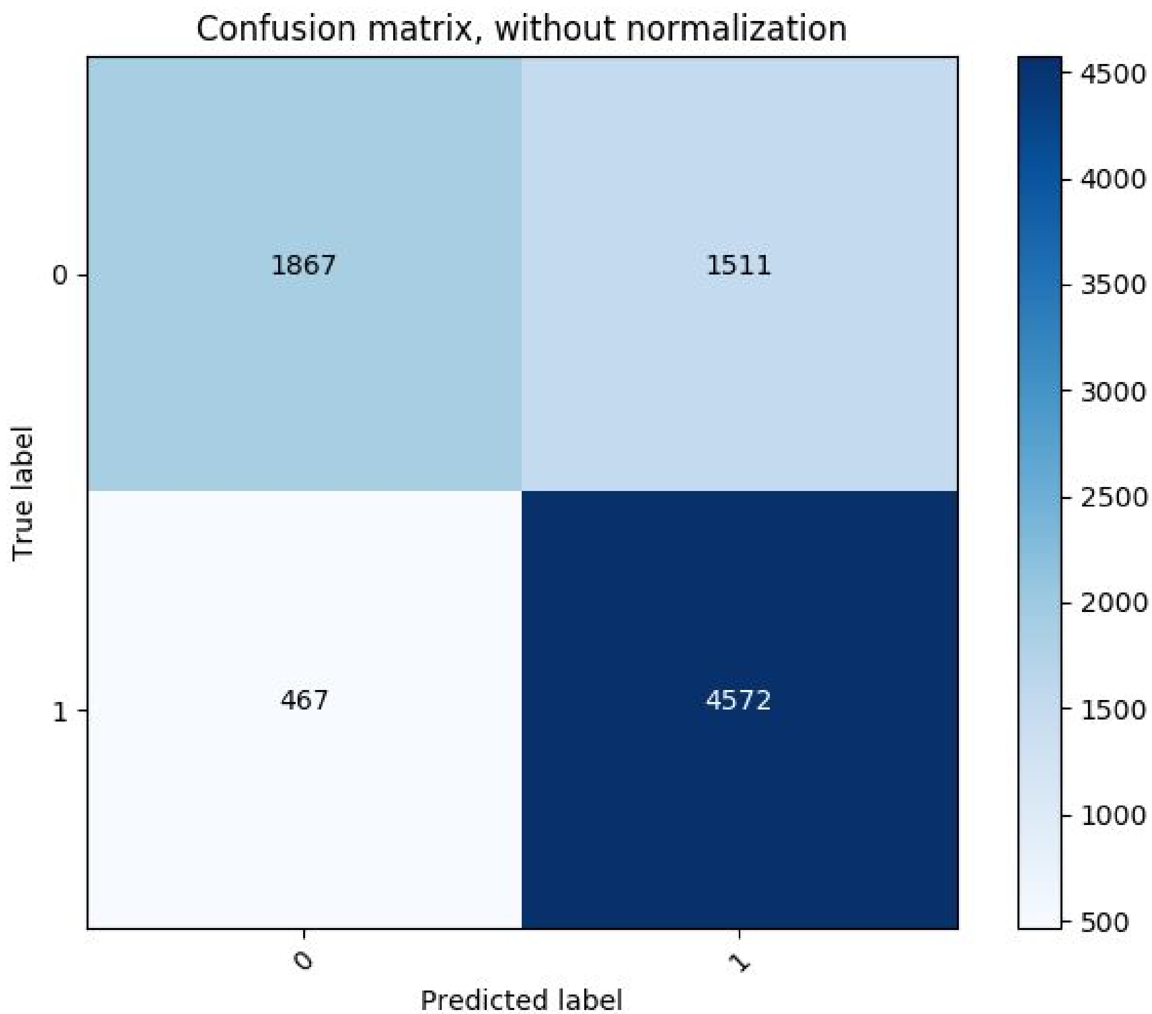


Figure 6.2 confusion matrix for Logistic classification

Calculated from the confusion matrix, the precision for this model is 0.799 and the recall is 0.553. The accuracy is 0.676.

Comparing the two models, logistic classification has better performance.

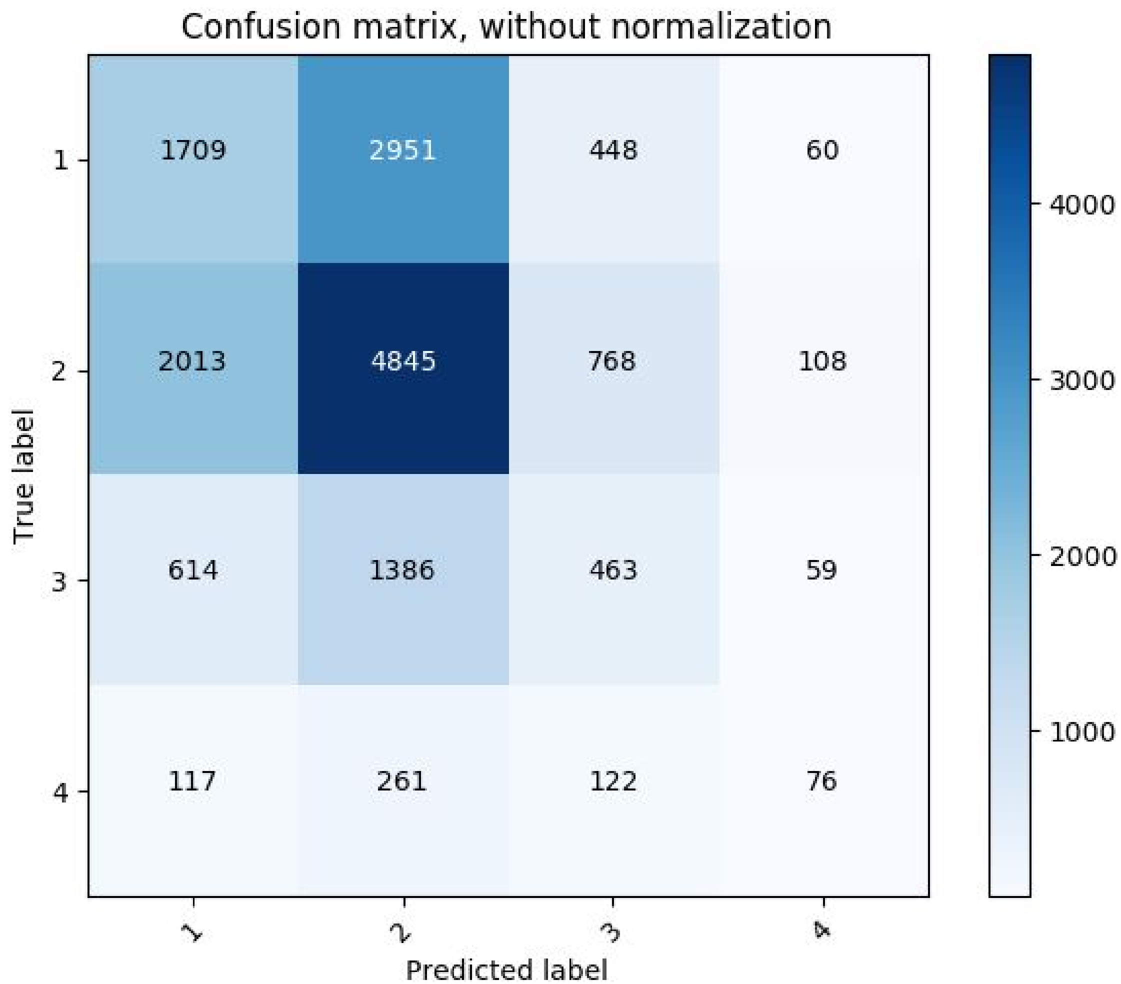
Part 7)

A person’s popularity can be determined by the person’s friend_count. Our goal is to predict a person’s popularity into following categories:

1. Friends < 200 : Inactive.
2. Friends 200<= x <1000: Active
3. Friends 1000<=x <3000: Popular
4. Friends 3000<=x : Famous

The linear SVC regression model is used to try to predict the popularity level of the person. From the confusion matrix we calculate the precision and recall of the predictions for each classes.

NO	Class	Precision	Recall
1	Inactive	0.384	0.33
2	Active	0.51	0.6265
3	Popular	0.257	0.184
4	Famous	0.251	0.132



The precision and recall are generally low. Tweet message is only a moderate indicator to see if the person has friend between 200 and 1000. However, beyond this range, tweet message becomes a poor indicator.

A linear regression model is used to estimate the linear relationship between friend_count and number of follower. The result shows a low R value at 0.009. This means that there is little linear relationship between friend_count and number of follower. The average prediction error is 0.588798096116. This means that the prediction is still accurate as each prediction is below 1 rank different from actual rank.

A linear regression model is also used to estimate the linear relationship between friend_count and 24-hour time of the day user tweets. The result shows an fair R value at 0.779. This means that there is a fair positive linear relationship between friend_count and 24-hour time of the day user tweets. The average prediction error is 0.589674799575. This means that the prediction is still accurate as each prediction is below 1 rank different from actual rank.

The analysis shows that tweet message can be a poor source to predict a person's popularity. This means that we can not guess if a person has lots of friends from merely the tweet message the person creates. However, time of the day a person tweets can be a good indicator to a person's friend count. Since there is a positive linear relationship between a person's friend count and time of day the person tweet, we can conclude that if a person who tweets at later time of the day generally has more friend. This can be explained by a popular person usually stay up late partying with their friends, or more devoted social network app user tend to stay up late online and tend to have more online friends. A person's number of followers can also be a good indicator to that person's popularity.