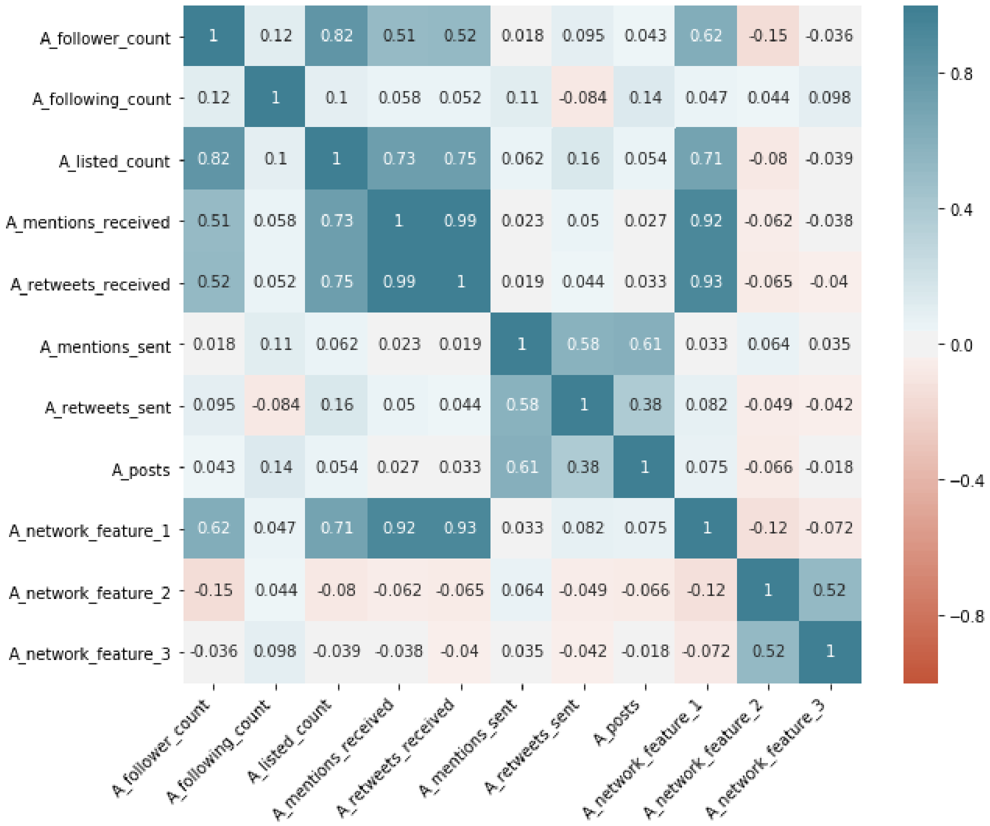
SMA Assignment 1

Group Members: Abhinav Singh, Catherine Miao, Eddie Eustachon, Elaine Wang, Thomas Bruce, Qinpei Zou

Link to our github for this assignment: https://github.com/catymiao/Social-Media-Analytics

Part I: Find predictors of influence

1. Check feature correlations with the correlation matrix



Correlation matrix indicates:

We should drop network_feature_1, mentions_received since these two features are both highly correlated with retweets_received, i.e. cor(network_feature_1,retweets_received)=0.928, cor(mentions_received, retweets_received)=0.99.

2. Feature Transformation

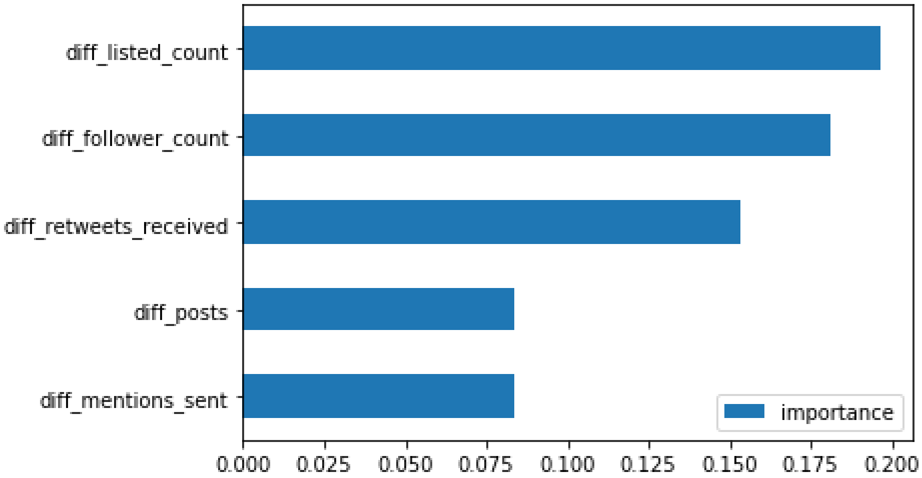
Use subtraction (-) to reduce the dimensionality of the dataset.

Variables that we created are:
Diff_follower_count
Diff_posts
Diff_following_count
Diff_listed_count
Diff_retweets_received
Diff_mentions_sent
Diff_retweets_sent
Diff_posts
Diff_network_feature_2
Diff_network_feature_3

Analytics Model Creation and Predictions:

Model	Prediction Accuracy (Baseline: ~ 51%)
Logistic Regression	~ 75%
Random Forest	~ 79%
XGBoost	~ 79%

Variable Importance from Random Forest:



The top three most important features are:
1. Diff_listed_count 20%
2. Diff_follower_count 18%
3. Diff_retweets_received 15%

Therefore, both the Random Forest and the XGboost model achieve an accuracy ~ 79%.
From the random forest model, we determine that diff_listed_count, diff_follower_count, diff_retweets_received are three important predictors of influencers.

Diff_listed_count is the most important variable which accounts for about 21%. Listed_count measures that the number of people that are in this users' list. A List is a curated group of Twitter accounts. The higher listed_count, the more people follow with interest.

Diff_follower_count is the second important variable which accounts for 16.6%. Follow_count is the number of followers this account currently has. So diff_follower_count shows how many more followers A have than B. The more diff_follower_count, the more likely that A's influence power is higher than B's.

Diff_retweets_received is the third important variable which accounts for 15.1%. Retweets_received is the number of retweets this user received. The more retweets he/she received, the more people pay attention to the tweets he/she sent. So diff_retweets_received is another good predictor to measure the influential difference between A and B.

Business Implication of our model:

If a business is having a marketing promotion and in need of allocating limited funding to pay selected media influencers, then this model would help their decision making on who is the better influencer i.e. which media influencer will bring higher profit for the business, and thus the individual to pay.

In addition, it is a great model for businesses to strategize their advertising campaign. For example, when new product is launched, the business needs to advertise the new products on social media platform. It would be an excellent way to employ this predictive model so as to target right influencers, and in turn, make the advertising campaign more effective.

Calculate the financial value of our model

Without analytics:

Profit_1 = 10 ✖ 0.01% ✖ (A_followers_count + B_followers_count) - 5 ✖ (# of A + # of B)

With analytics:

Profit_2 = 10 ✖ 0.015% ✖ prediction accuracy ✖ Influencers_followers_count - 10 ✖ # of Influencer

Lift in net profit: Profit_2 - Profit_1

Percentage lift in net profit:

(Profit_2 - Profit_1)/Profit_1x100%

Profit under Each Scenario

	No Analytics	With analytics (79.5% accuracy)	Perfect Analytics (100% accuracy)
Profit ($)	5,321,569	6,352,526	8,009,853
Change in profit vs base case	NA	1,030,957 (19.4%)	2,688,284 (50.5%)

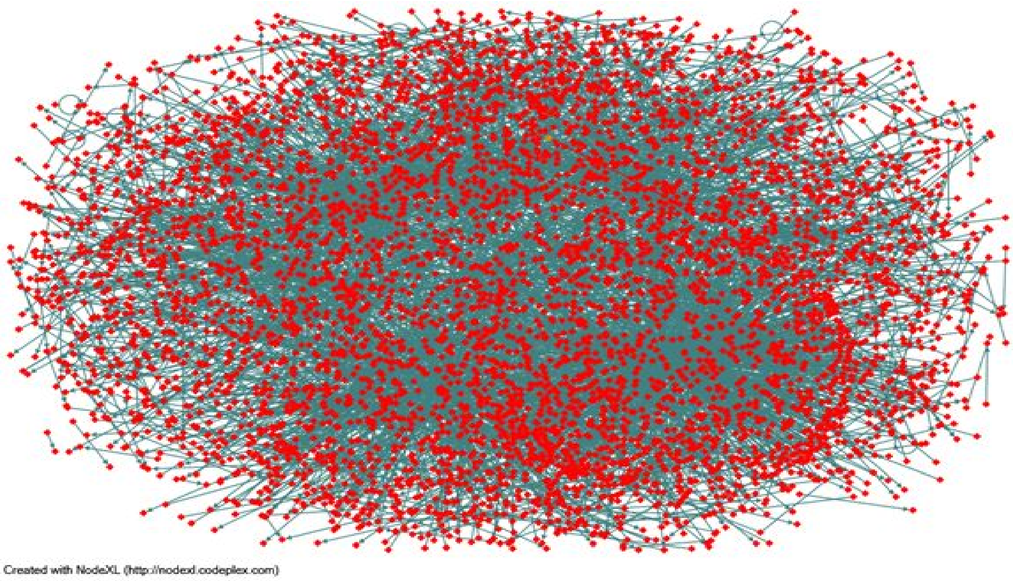
In conclusion, using our analytic model above would increase profit for the retailer by around 20% compared to not using analytics, while using a perfect model would increase profit by around 50%. Therefore, we determine that the financial value of analytics is substantial

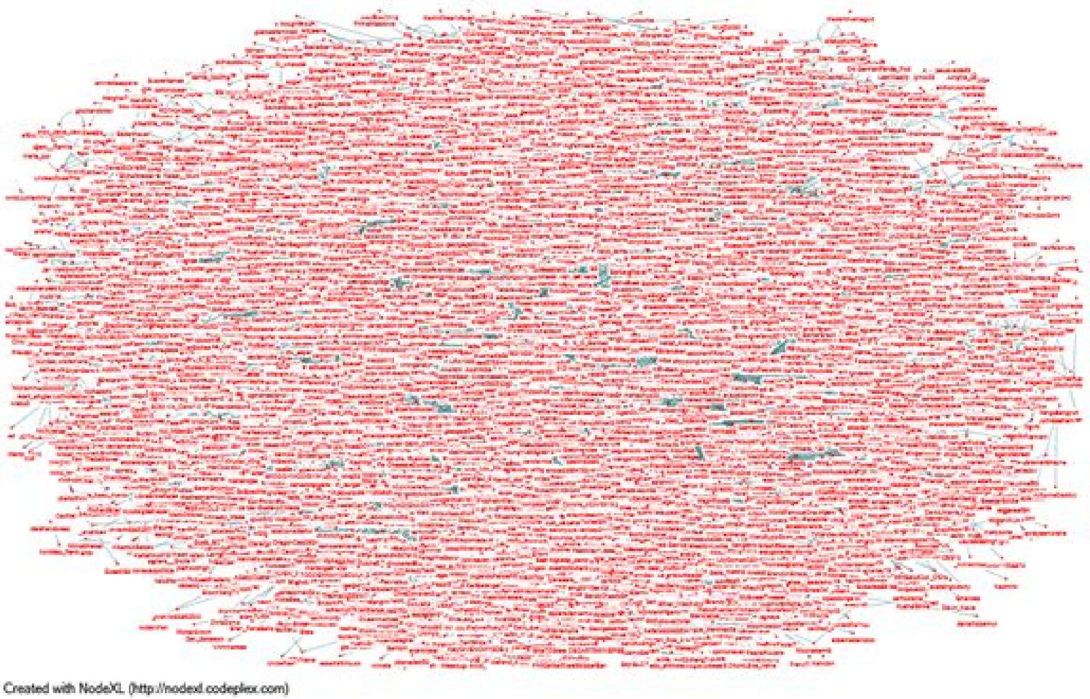
Part II: Finding influencers from Twitter

Interest topic:

Identify media influencers on twitter of the ongoing coronavirus

Network analysis from nodeXL:





Degree, betweenness and closeness calculation:
We created a column that contained the sum of degree, between, and closeness scores for each account name. This was then used as the fourth weight in our formula, the other three being the weights of the most important variables from the random forest (Diff_listed_count, Diff_follower_count, & Diff_Retweets_received). The degreeness + betweenness + closeness metric accounted for 53 percent of the score formula and the remaining three values comprised the remaining 47 percent, with the specific amount being determined by the variable importance. After this, the variables were multiplied by the weights/feature importances then ranked by top 50 influencers.

Top 50 influencers (with results from Part I)

The top 50 influencers are predominantly international news sources (the top influencer is the UK's Daily Mail, followed by the Chilean newspaper La Tercera and the New York Post). This makes sense, as coronavirus has dominated the news globally through the past month, and news sources are likely to be very active in social network conversations relating to the disease as it spreads throughout the world.

Calculating score for each author:

From the random forest variable importance, we obtained the top 3 variables, and their respective importance score:

1. Listed_count	20%

2.	Follower_count	18%
3.	Retweets_received	15%

Thus, the top three variables account for 53% of the total importance.

As we have 6 variables in total (Listed_count, Follower_count, Retweets_received, Degree, Betweenness, Closeness), we decide to assign approximately 50% weight for the first three variables (Listed_count, Follower_count, Retweets_received) and the rest of approximately 50% weight for degree, betweenness, and closeness. Then, our score function is:

Score = 0.2 * diff_listed_count + 0.18 * diff_follower_count + 0.15*diff_retweets_received + 0.47 (scaled degree + betweenness + closeness)