MIS 184N
Barua
2/4/18

Social Media Analytics Assignment 1

By: Daxi Cheng, Corey Haines, Daniel Quijano, Jinru Su, Reece Wooten

Part I: Find Predictors of Influence

For this part of the assignment we aimed to determine those people with high social influence within a randomly sampled Twitter population. Each observation describes two individuals, A and B. For each person, 11 non-negative numeric features based on Twitter activity were provided, e.g., volume of interactions, number of followers, retweets, network characteristics, etc. Each observation shows whether A > B (Choice = "1") or B > A (Choice = "0") according to human judgement.

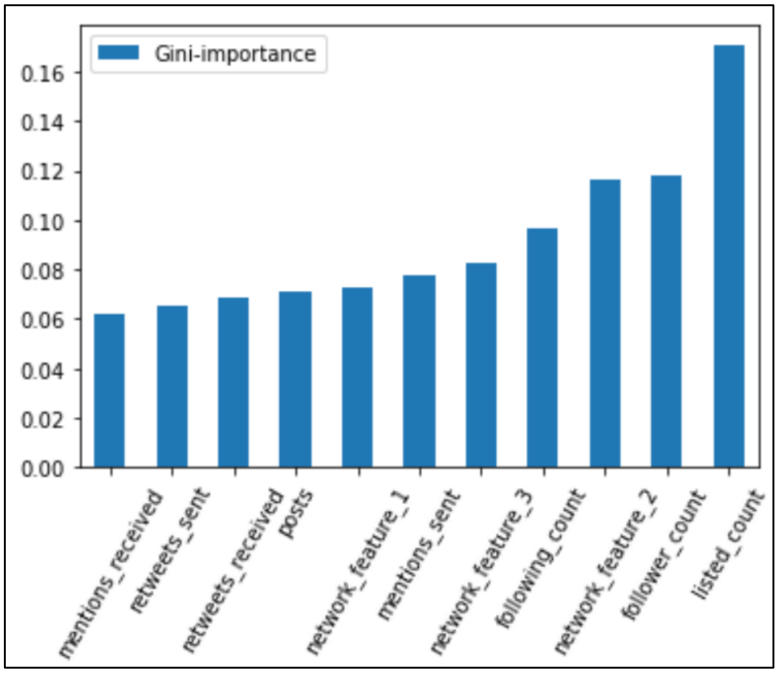
Before building our model, we first performed simple exploratory data analysis to determine the correlation structure amongst the features as well as whether the classes were balanced. When examining which features were highly correlated with one another, we found that number of mentions received is highly correlated with number of retweets (0.988363) and both number of mentions received, and number of retweets are highly correlated with network factor 1 (0.914479 and 0.920574 respectively). Because number of mentions and number of retweets was almost perfectly correlated, it was decided that they essentially provided the same information. However, we kept features in our final model because our sample size was sufficiently large and tree-based models are fairly robust to multicollinearity. After checking for multicollinearity amongst the features we made sure the classes were balanced for whether A or B was more influential, and it appears they were balanced.

Lastly, before building our model we took the difference between A and B for each variable, e.g., difference in retweets received between A_1 and B_1. The logic behind this transformation is to see how the net difference affects one social influence. Thus, it is more meaningful to look at the difference in a variable, such as number o mentions, because this can be viewed as an indication of popularity. Therefore, it follows that the difference should provide an indication as to which person is more popular/connected and thus more influential.

We trained prediction models using logistic regression, random forest, and XGBoost. Of these three models, XGBoost gave the best prediction results – 77.466% accuracy, 77.597% recall, and 77.329% specificity.

	Condition Positive	Condition Negative
Predicted Condition Positive	717	202
Predicted Condition Negative	207	689

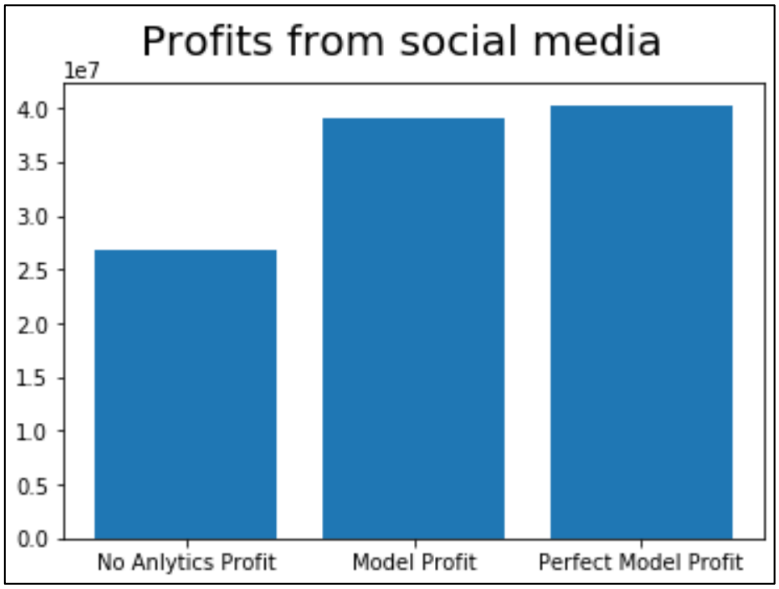
Using Gini coefficient as our metric of information, the difference in listed_count and difference in follower_count between users were the most powerful predictors of social influence. This makes logical sense because both metrics are indicators of one's outreach potential on Twitter. Those users that belong to more public lists and have greater follower counts will have their tweets seen by more people. Thus, they have a greater chance of influencing someone around them.



In addition to the previous two visibility metrics, one's network features were also important. This indicates that the strength of one's network has an impact on one's social influence. While it is not surprising that the difference in listed_count and difference in follower_count were highly important, it was somewhat surprising to find that the number of mentions received and number of retweets were two of the least important features. One would think that users who are mentioned a lot and have a high number of retweets are active within their networks and have high visibility. In conclusion, businesses should target those Twitter users with high follower counts, users who are part of multiple public lists, and users with strong network centrality measures. The potential financial impact from a model that correctly identifies social influencers can be seen through the use of a simple example:

> *Given an influencer tweets once, there is a 0.05% chance that his/her followers will buy one unit of a product. Assume the retailer has a profit margin of $10 per unit, and that one customer can buy only one unit. If an influencer tweets twice, the overall buying probability will be 0.075%. Without analytics, the retailer offers $5 to each person (A and B) to tweet once. With analytics, the retailer offers $10 to those identified as influencers by the model to send two tweets each. If the model classifies an individual as a non-influencer, s/he is not selected/paid by the retailer to tweet.*

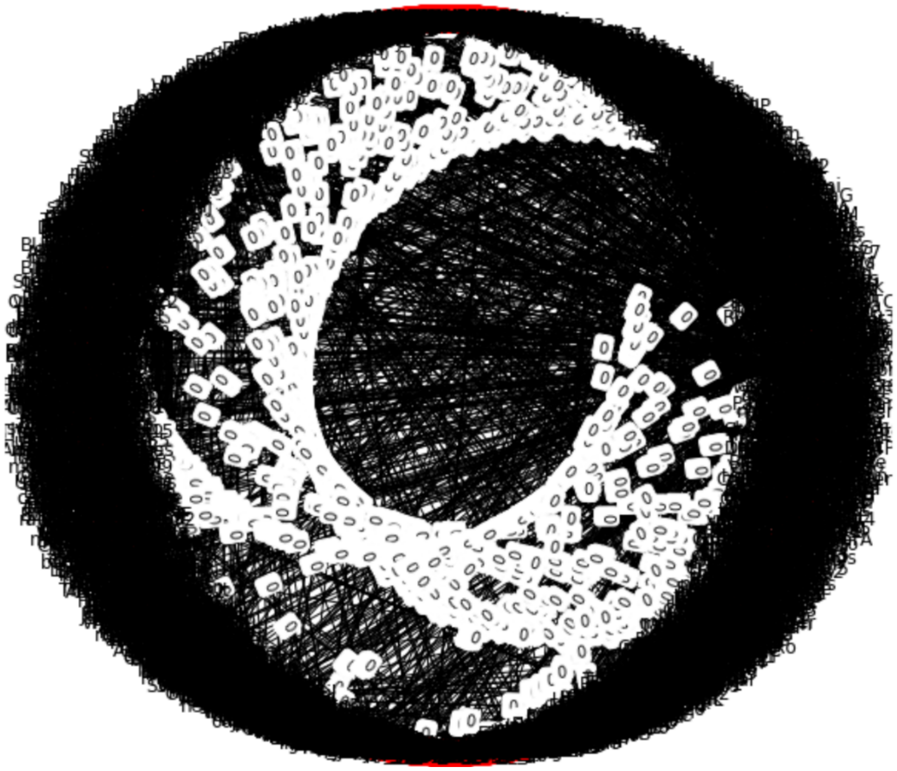
The theoretical profit gained from paying both influencers and non-influencers to tweet is $26,827,842.97. This is a substantial decrease from the theoretical profit of $38,993,274.70 (Δ = $12,165,431.73 or 45.35%) that could be gained using our classification model. A perfect model would result in theoretical profits of $40,269,264.45 ($\Delta_{NA}$ = $13,441,421.48 or 50.10% || Δ_{A} = $1,275,989.75 or 3.27%). ***Please see codebook for worked out solution



MIS 184N
Barua
2/4/18

Part II: Finding Influencers from Twitter

For this part of the project we sought to determine who the influential users are on Twitter within the Bitcoin community. To do this we scraped approximately 5000 tweets with any word containing bitcoin and collected user information associated with each tweet. Then, for each tweet we documented whether documented whether the tweet was a retweet, contained a mention or reply, or was just an "isolated" tweet. For retweets and those tweets containing a mention or reply an arrow was created from the person retweeting to the person retweeted, mentioned or replied to. From this we were able to construct a network and calculate the three main centrality measures – degree, betweenness, and closeness – for the total network as well as each person within the network.



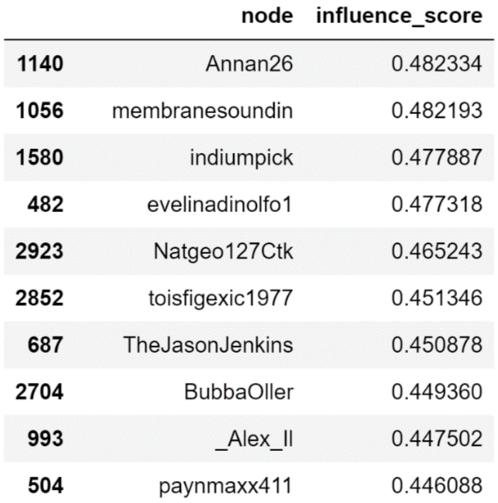
For the network, degree centrality was measured to be 0.116121, betweenness centrality was measured to be 0.072322, and closeness centrality was measured to be 0.082897.

With these metrics in hand we then determined who the most influential users were within the network. To do this an "influence score" was developed by assigning weights to the top three non-network features [as determined by gini information importance] from Part 1 and the sum of the network centrality measures. To ensure the weights added to one we took the absolute value of the coefficients from the best model in Part 1 and passed them through the softmax function. Thus, our influence score was calculated by the following equation:

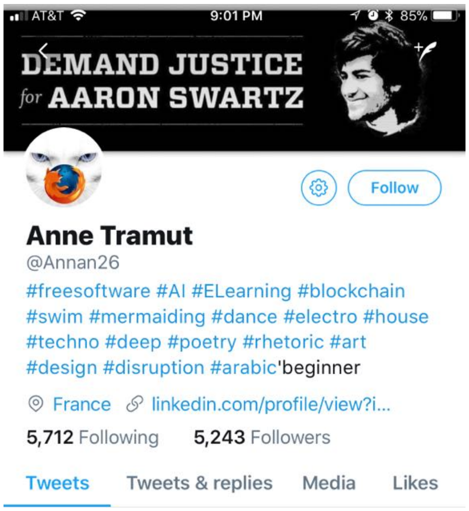
$$Score = \ w_1 \, \beta_{listed\_count} \, + \, w_2 \, \beta_{follower\_count} \, + \, w_3 \, \beta_{retweets} \, + w_4 \, \beta_{network\_effects} \, ,$$

where $\beta_{\text{network\_effects}} = \beta_{\text{degree}} + \beta_{\text{betweenness}} + \beta_{\text{closenesss}}$

Using this metric, we determined the top 50 influencers within the Bitcoin Twitter community. Below are the top 10 most influential users and home Twitter page for the top user:



	node	influence_score
1140	Annan26	0.482334
1056	membranesoundin	0.482193
1580	indiumpick	0.477887
482	evelinadinolfo1	0.477318
2923	Natgeo127Ctk	0.465243
2852	toisfigexic1977	0.451346
687	TheJasonJenkins	0.450878
2704	BubbaOller	0.449360
993	_Alex_ll	0.447502
504	paynmaxx411	0.446088



Anne Tramut
@Annan26
#freesoftware #AI #ELearning #blockchain #swim #mermaiding #dance #electro #house #techno #deep #poetry #rhetoric #art #design #disruption #arabic'beginner
France linkedin.com/profile/view?i...
5,712 Following 5,243 Followers

As can be seen by Annan26's Twitter page, his third hashtag in his bio is blockchain. An exploration of his Twitter feed unveiled that he retweeted many tweets regarding Bitcoin and blockchain in general, thus giving us confidence that our model from Part 1 and influence score yielded fairly reliable results regarding social influence within the Bitcoin Twitter community.