# Assignment 1 Writeup
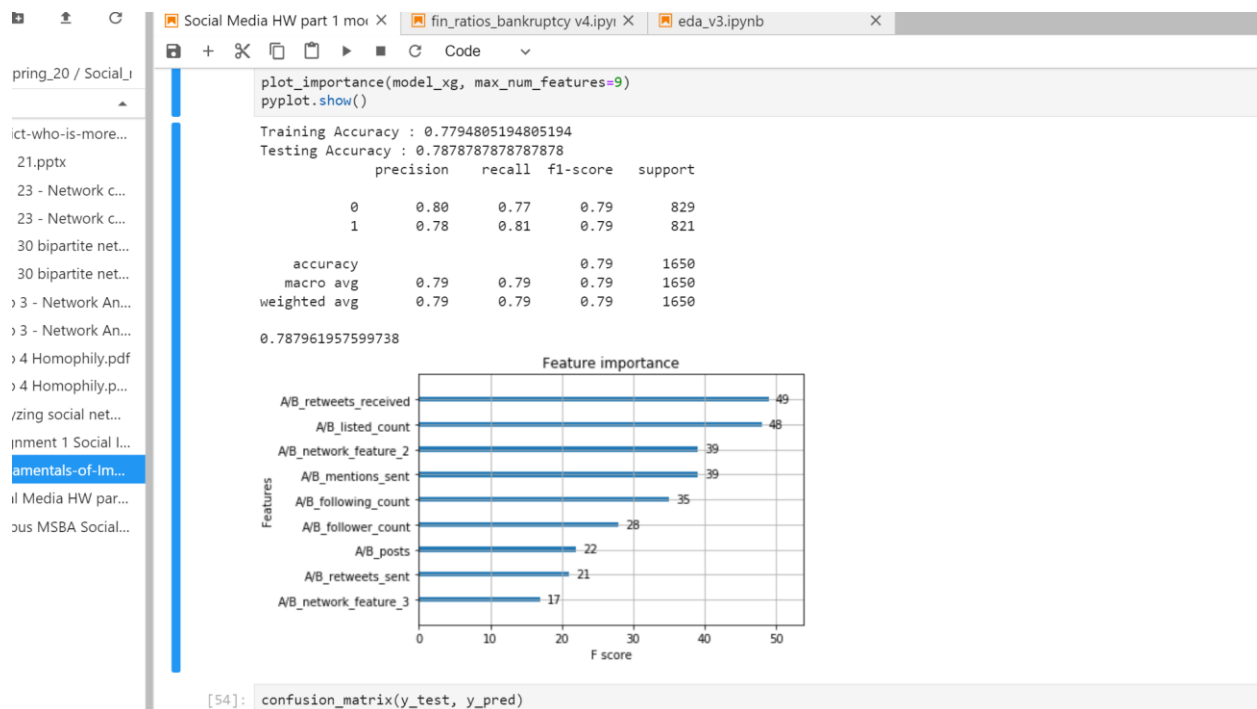### Hannah Ho, Hannah Warren, David Owen, Akankshi Mody, Candice Zuo, Ananya Garg

## Part I: Find predictors of influence

We tried three models: logistic regression, Random Forest and XGBoost. Among them, XGBoost has the highest accurate rate of 78.8%.

Confusion matrix of the XGBoost model and best predictors of influence:

```
confusion_matrix(y_test, y_pred)
```

```
array([[634, 195],
       [175, 646]])
```



The best 5 predictors of influence at **retweets_received, listed_count, network_feature_2, mentions_sent, and following_count**. Retweets indicate engagement whereas following count does not, so we expect retweets to be ranked higher. However, it is surprising that following count ranks higher than follower count, because we thought how many people followed you would affect others' perception of your influence more than how many people you follow.

A business can use the ranked features of what makes a person influential online as a guide to deciding who to partner with on social media.

**Calculating the *financial value* of the model**

(Assumption: each user appears only once in the data, hence each row has two new names. Hence out no. of users is twice the numbers of rows in the data)

Net profit without using analytical model:
We assume either A or B is an influencer and the other is not. The average number of followers is 667,686 for all A's and B's.

Revenue - costs = profit

[Profit margin per unit * .01% chance buy * no of average followers across all users - payment to each user]

=> [$10 * .01% * 667,686 - $5] = **$662.69 per paid user**

Net profit using our analytical model:
The inclusion of 10 is the expected value of a follower viewing the paid person's Tweeting accuracy multiplied by $10 is the expected value of a follower viewing the paid person's Tweeting accuracy

```
Training Accuracy : 0.7794805194805194
Testing Accuracy : 0.7878787878787878
            precision   recall  f1-score   support

        0       0.80      0.77      0.79       829
        1       0.78      0.81      0.79       821

 accuracy                           0.79      1650
macro avg       0.79      0.79      0.79      1650
weighted avg    0.79      0.79      0.79      1650
```

Our expected profit is the weighted average of the Profit from predicted influencers + loss from false influencers using the average followers count of our predicted influencers.

Expected value = .78* [ $10 * .015% * 1,087,897 - $10] + (-$10)*(1-.78)  = **$1,262.84 per user**

Lift in expected net profit using analytical model:
$1,262.84 - $662.69 = $600.15 per user

Net profit using a perfect analytic model:
Lift in expected net profit using perfect analytical model: is 100% of followers of influencers that
$1,262.84 - $662.69 = $595.50 per user

# Part II: Finding influencers from Twitter
We collected tweets related to the flat earth conspiracy theory from Twitter using Tweepy.
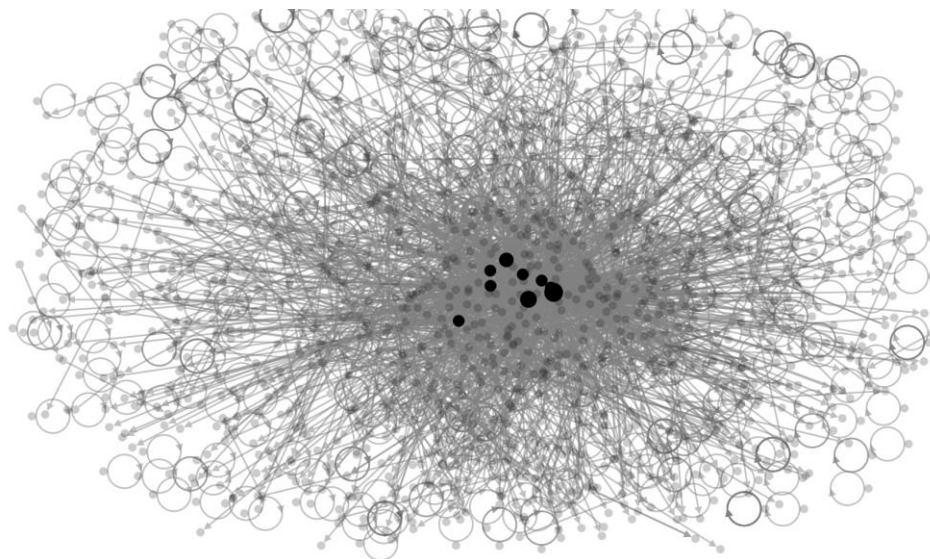We then use the tweets to construct a directed graph using networkx, we get a graph as follows (Tweet/

1. Number of nodes: 1499
2. Number of edges: 2419
3. Average in degree:   1.6137
4. Average out degree:  1.6137

We then calculate centrality metrics - degree, betweenness and closeness using networkx.

By using NodeXL to visualize the graph, we get the following graph:

*Figure 1: NodeXL Graph visualization*

We can clearly see that the nodes in black act as the key influencers. The graph was constructed by considering the in-degree as the cluster measure.

## List of Top 50 Influencers

In order to create a list of the top 50 influencers, I had to refer to the features that Part 1 of the presentation deemed to be the most important.

```
Out[20]:
```

|  | importance |
| --- | --- |
| A/B_listed_count | 0.212132 |
| A/B_follower_count | 0.193142 |
| A/B_retweets_received | 0.109217 |
| A/B_posts | 0.093201 |
| A/B_mentions_sent | 0.088297 |
| A/B_network_feature_2 | 0.084763 |
| A/B_network_feature_3 | 0.074815 |
| A/B_following_count | 0.074040 |
| A/B_retweets_sent | 0.070393 |

However, not all of these features existed in our dataset. A/B could be calculated. Therefore, we simply summed up the scores of each interaction between 2 users occurred multiple times. The top ten Influencers were as follows:

|  | Score |
| --- | --- |
| **A_handle** | |
| PeaceWTF | 1268.009662 |
| Constitution_NH | 1107.255669 |
| WeenieLinguini | 1104.249647 |
| AngelsFreak7 | 875.348617 |
| Royal_Time | 656.684665 |
| ConstitutionNd | 538.123950 |
| nanaof47 | 495.100137 |
| MEConstitution | 488.961483 |
| BotSiduri | 385.253428 |
| news2health | 325.749672 |

To see the full list of all 50 influencers, please refer to the code.