Social Media Analytics Assignment #1

Due 3/25 by 11:59 PM

The assignment has two parts. In Part I, you will use training data on social influence to build a model predicting influencers, to find out the **important predictors of influence**, and to **quantify the financial value of influence**. In Part II, you will collect tweets, and use the predictors from Part I to **identify top 100 influencers** in a domain of your choice.

Part I: Find predictors of influence

The dataset for Part I is here. Each observation describes two individuals, A and B. There are 11 variables for each person based on Twitter activity, e.g., number of followers, retweets, network characteristics, etc. Each observation shows whether A > B (Choice = "1") or B > A (Choice = "0").

Using the training data set (`train.csv`), create an analytic model for pairs of individuals to classify who is more influential

- Check if you should use all variables
- Perhaps a transformation of (A / B) or (A – B) variables will be better than using A and B variables separately. This may also be easier to interpret.
- Report the confusion matrix of your "best" model

From your model, which factors are best predictors of influence? Are there any surprises here? How can a business use your model/results?

Calculate the *financial value* of your model

A retailer wants influencers to tweet its promotion for a product. If a non-influencer tweets, there is no benefit to the retailer. If an influencer tweets once, there is a **0.02%** 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 **0.03%**. 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.

What is the boost in **expected net profit** from using your analytic model (versus not using analytics)? Show all calculations. What is the boost in **net profit** from using a perfect analytic model (versus not using analytics)?

***Assumption: Each user appears only once in the data**

A	B	A>B?
John	Ted	Yes

Sue	Ron	Yes
Fred	Sandy	No
Alex	Moe	No

The Influencers in the above table are John, Sue, Sandy & Moe, but no ordered ranking is possible (or needed in this case).

Part II: Finding influencers from Twitter

Collect about **5,000 tweets** on any topic (e.g., politics, sports, current events, etc.). **Write a script that parses through the tweets and does the following for each tweet:**

Any **retweet**, **mention** or **reply** should result in an edge from the person retweeting to the person retweeted, mentioned or replied to. So create a three-column CSV file as follows: If @XYZ retweets a tweet by @ABC, then put the following in the CSV file:

Column 1	Column 2	Column 3 (type of content)
@ABC	@ABC	Tweet
@XYZ	@ABC	RT

Most social network analysis tools (e.g., NodeXL, Gephi or UCINet) will take the first two columns and draw arrows from the user in the left column to the one in the right – you can also use `NetworkX` in Python to draw networks. Note that in most cases the set of tweets you may fetch will not have the original tweet that is being retweeted by someone else. E.g., a tweet in your data (tweeted by, say, @XYZ) may be: "RT @ABC Working on my social media assignment." It is quite possible that you will not have the original tweet by @ABC in your data. Still an arrow should go from @XYZ to @ABC. Therefore, even if you have fetched 5,000 tweets by 5,000 unique users, your network may consist of a much larger set of users.

Calculate the degree, betweenness and closeness of each node in the above network.

Using the results from Part I, create a list of top 100 influencers from the tweets. Here is *one* way to do it. Suppose four factors – retweets, listed count, # followers and network feature 1 turned out to be the most important indicators of influence in Part I. Now create a score for each author from your Twitter data:

Score = $w_1 \times$ retweets + $w_2 \times$ listed_count + $w_3 \times$ #followers + $w_4 \times$ network_feature_1, where $w_1+w_2+w_3+ w_4 = 1$.

Choose the weights (it is subjective) such that bigger weights are given to factors that were more important (as judged by, for example, coefficients and *p* values in Part I). You

should normalize your data before creating the overall scores. Note that the Kaggle data doesn't show what each network characteristic means. However, generally such metrics are presented in the following sequence: **degree**, **betweenness** and **closeness**.

Finally, provide a **network visual** of the 100 influencers you selected.

Submit the following to *myCourses*:

1. Python scripts with necessary input files (e.g., three column CSV file)
2. A PDF file with answer to Parts I and II
3. A Python notebook can be submitted in lieu of 1. and 2.

NetworkX

Once you have a three-column CSV file, you can use the NetworkX package in Python to calculate centrality scores. Example:

```
import networkx as nx

G = nx.DiGraph()

lst=list()
for (a,b) in zip(column1, column2):
    lst.append((a,b))

G.add_edges_from(lst)

closeness_centrality=pd.DataFrame.from_dict(nx.closeness_centrali
ty(G), orient='index').reset_index()
```

You can also see here for more information.