Social Media Analytics – Group Assignment

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 Reddit

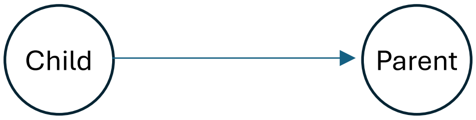
Select a subreddit from archival here and download both submission and comments. Use the following command to extract the file:

```
import pandas as pd

     df_submission=
     pd.read_json(filepath,compression=dict(method='zstd',
                  max_window_size=2147483648), lines=True)
```

Merge the files and construct the data such that:

Child ID	Parent ID	Link Type



For example, if comment B is to respond to comment A, then *parent* is comment A and *child* is comment B. If comment A is a submission, then *link type* is “respond to a submission.” Yet, if comment A is a comment, then *link type* is “respond to a comment.”

Most social network analysis tools (e.g., NodeXL, Gephi or UCINet) will take the first two columns and draw arrows from the left column to the one in the right – you can also use NetworkX in Python to draw networks.

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

Create a list of top 20 influencers from the subreddit.
Here is *one* way to do it.
Create a score for each author from your Reddit data:

Score = w1 × # submissions + w2 × # comments + w3 × # comments to the submissions + w4 × # comments to the comments + w5 × a network feature, where w1+w2+w3+w4+w5= 1

Note that you may play with the data to see if there is any information you can use to calculate the score. For instance, in the example, *ups* is the number of upvotes.

Choose the weights (it is subjective) such that bigger weights are given to factors that were more important. You should normalize your data before creating the overall scores. The examples of network feature are degree, betweenness and closeness.

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

Submit the following to myCourses:

1. Python scripts (.py AND .ipynb), separated by Part I and Part 2, with clear comments on the code (e.g., what is the objective of this code section).
2. Necessary input files
3. A PDF file with answer to Parts I and II