INSY 670 Social Media Analytics

Professor: Taha Havakhor

Group Assignment

Students:

Yusen Tang:

Addison Ji: 261184567

Yvette Zhu

Yu Lu: 261150454

Yanhuan Huang: 260561335

**Part I: Find predictors of influence**

**Data acquisition**: The [dataset](https://www.kaggle.com/competitions/predict-who-is-more-influential-in-a-social-network/data?select=test.csv) from Kaggle consists of 5,500 observations, each representing a pair of individuals with 11 variables for each person based on their Twitter activity (like follower count, following count, listed count, mentions received, retweets received, etc.), making a total of 22 features plus the 'Choice' column, which indicates who is more influential (A > B is marked as “1”, B > A as “0”).

**Feature Engineering:** We perform targeted transformations on variables related to Twitter activity for pairs of individuals, A and B. This process includes calculating the sum (A+B), difference (A-B), ratio (A/B), and a normalized difference of each activity metric to capture various dimensions of relative influence. These transformations aim to enhance the model's ability to discern patterns of influence more effectively by leveraging both absolute and relative comparisons between individuals.

**Model Selection and Training:** We experimented with different machine learning models to see which performs best. The model performs the best is Random Forrest. By using the feature importance scores in the model, we have identified the top 20 features and use them to fine tune the model.

**Evaluation:** We evaluated the model using a confusion matrix to understand its performance, the results are as follows:

A diagram of a confusion matrix

Description automatically generated

Classification Report:

precision recall f1-score support

0 0.80 0.74 0.77 820

1 0.76 0.81 0.79 830

accuracy 0.78 1650

macro avg 0.78 0.78 0.78 1650

weighted avg 0.78 0.78 0.78 1650

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

The top 20 features from the “Random Forrest Classifier – Feature Importance Scores” are:

A screen shot of a computer

Description automatically generated

The model reveals that the best predictors of influence on Twitter involve a blend of engagement metrics and network characteristics, with a particular emphasis on normalized differences and raw differences in listed counts, follower counts, and mentions received. Notably, **listed\_count\_normalized\_diff** and **listed\_count\_diff** top the list, underscoring the importance of how often an individual is listed in comparison to others as a measure of their influence. The prominence of normalized differences in follower count and retweets received further highlights the significance of relative engagement over mere volume. Surprisingly, ratios and sums of network features also play a crucial role, suggesting the underlying structure and quality of one's network can greatly impact influence. For businesses, leveraging these insights can enhance marketing strategies by identifying and collaborating with influencers who exhibit high relative engagement and network quality, thereby optimizing their social media campaigns for greater reach and impact. The following part which we calculated the expected financial value of using this model for a retailer interested in utilizing influencers for product promotions will explains the value proposition in a more detailed manner.

**Financial Analysis:**

From the dataset, below factors can be calculated before the analysis:

Number of influencers: 5,500

Number of non-influencers: 5,500

Total followers of all influencers: 5,376,568,593 (avg 977,557 followers/influencer)

**Scenario 1 Without Using Analytics:** This scenario assumes no analytical model is used to identify influencers. Thus, the "Choice" column isn't directly used in calculations. Instead, this scenario reflects the approach of offering everyone (11,000) a chance to tweet for a fee without distinguishing between influencers and non-influencers.

**Scenario 2 With Analytics (Using Our Model):** Here the model's predictions determine who is considered an influencer. The performance (accuracy, precision, recall) of this model impacts the financial calculations, reflecting the real-world application of using analytics to make decisions. This approach will give the model identified influencer (True positive and False positive: 5,850) the chance to tweet twice.

**Scenario 3 Perfect Analytic Model:** This scenario would imply that we can perfectly identify influencers without error. Therefore, the actual influencers (5,500) based on the “Choice” column will be the only one has the chance to tweet twice.

The variation in net profits across the three scenarios primarily hinges on the costs incurred by the retailer for tweets and the revenue generated from purchases by followers influenced by these tweets. In Scenario 1, although each influencer is paid only once, the total cost matches Scenario 3 because twice the number of individuals is compensated. Scenario 2 incurs the highest costs, as the model overestimates the number of influencers. Revenue calculations consider the influencers' total follower count and the likelihood of these followers making a purchase. The critical distinction in revenue between Scenarios 1 and 3 lies in a modest 0.01% higher purchase probability in Scenario 3. Meanwhile, the gap between Scenarios 2 and 3 underscores the model's inefficiency, with an 81% recall rate for true influencers translating to a 19% loss in potential revenue due to incorrect influencer identification. The detailed metrics and results for these three scenarios are as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Number of individuals being paid | Cost Paid by retailer to tweet | Expected total followers who  made purchase | Expected Revenue | Expected  Net Profit |
| S1: Without Analytics | 11,000 | $ 55,000 | 1,075,313 | $ 10,753,130 | $ 10,698,130 |
| S2: Our Model | 5,885 | $ 58,850 | 1,306,506 | $ 13,065,060 | $ 13,006,210 |
| S3: Perfect Analytics | 5,500 | $ 55,000 | 1,612,970 | $ 16,129,700 | $ 16,074,700 |

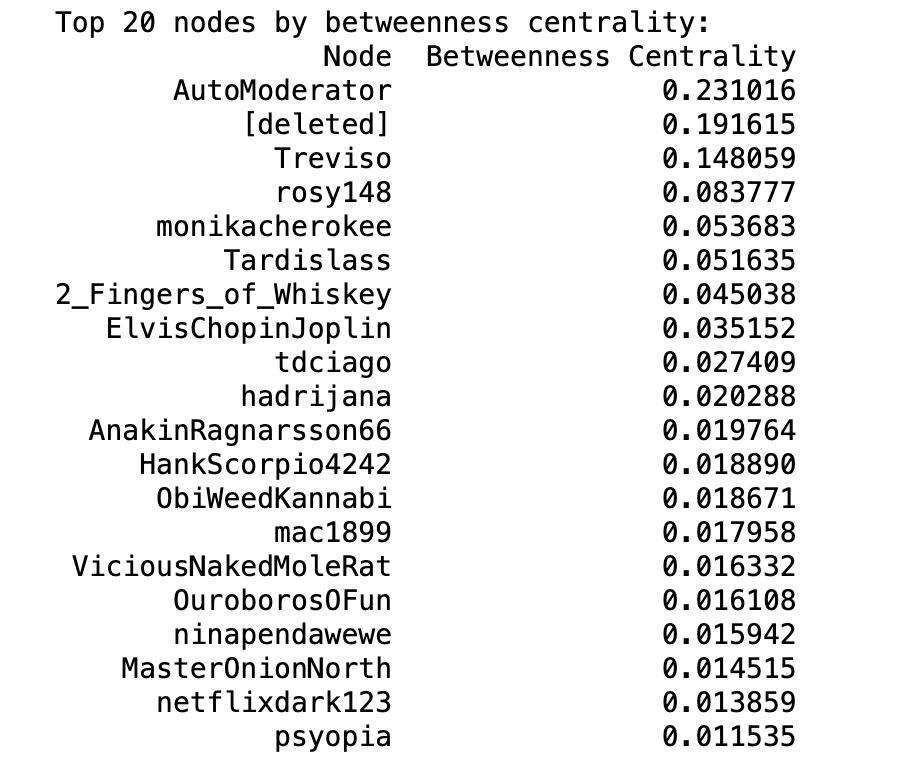
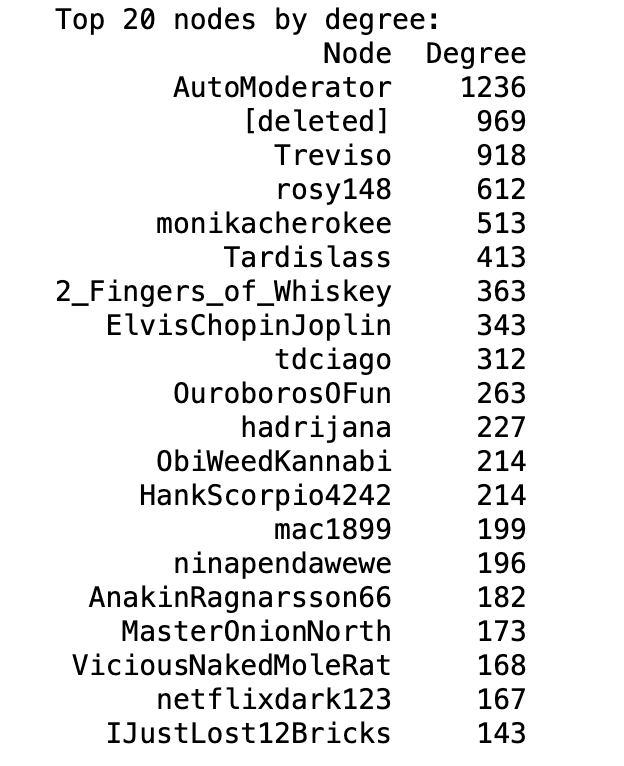
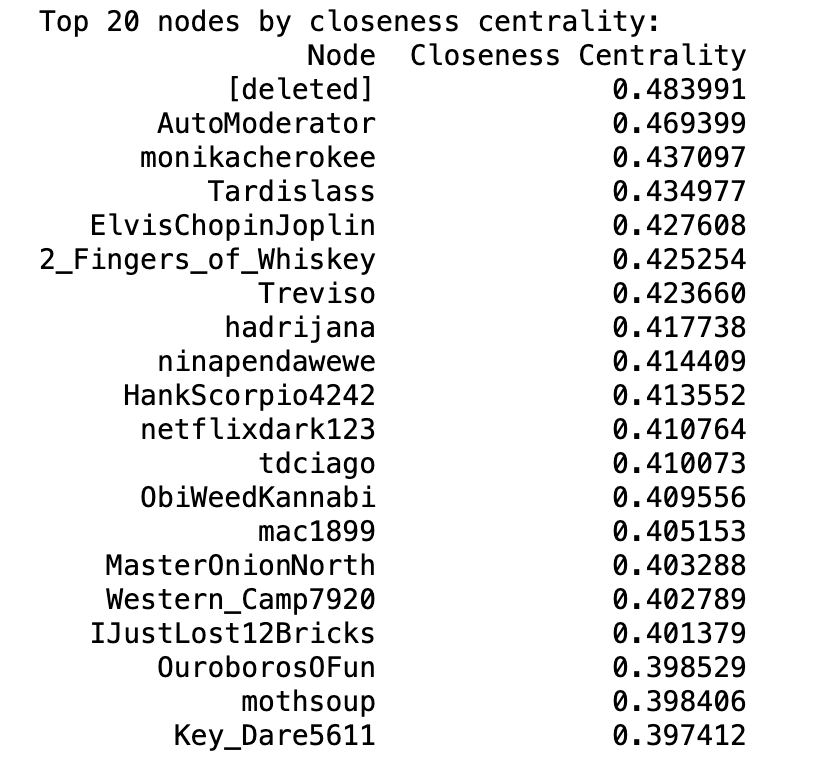
The transition from a non-analytical approach (Scenario 1) to a data-driven strategy (Scenario 2) enhances net profit by approximately $2,308,080, showcasing the substantial value added by employing analytics to identify and leverage influencers more efficiently. The leap to a Perfect Analytics (Scenario 3) further escalates net profit with additional $3,068,490, underlining the paramount importance of accurate influencer identification. This progression underscores the transformative impact of precise analytics in optimizing marketing strategies, highlighting the potential for businesses to significantly boost their marketing ROI by investing in advanced data analysis and modelling techniques to pinpoint and engage with key influencers.

**Part II: Finding Influencers from Reddit**

**Data acquisition**: The [subreddit](https://old.reddit.com/r/1899/) selected consists 40,489 comments and 3950 submissions, indicating users contributed 3.95K posts and more than 40K individual comments to the subreddit “1899 Netflix Series”.

**Pre-processing:** We’ve adjusted the ID formats for both comments and submissions datasets to facilitate the calculation of various network metrics. The *'parent\_id'* column, inherent to the comments dataset, and a constructed *'child\_id'* based on the comment's unique ID enable us to map the interactions. Reddit's prefixing rule, where 't3\_' denotes a submission and 't1\_' indicates a comment, was used to create *'link\_type'* column. This approach also allowed us to redefine the 'id' field in both submissions and comments datasets as *'full\_id'* with the necessary prefixes. We then complied a comprehensive mapping of authors to their posts and comments by unifying both datasets into a singular authorship directory. This allowed for the identification of the original author for any given comment, filling in any gaps in authorship as 'none'. Consequently, *'parent\_id'* from the comments dataset was used to map the original poster of a comment.

**Degree, Betweenness, Closeness:** The top 20 nodes with highest value in each metric are

**Influencer score calculation:** Before calculating influencer score, we excluded records with value ‘[deleted]’ or ‘AutoModerator’ in *‘author’* as the previous represents a deleted account and the latter a system built in bot. Weights are given as

In assessing the impact of influencers, we prioritize their ability to generate engagement, specifically the quantity of comments on both submissions and comments. While degree centrality can shed light on an influencer's network position, it provides a limited perspective and is thus emphasized less. Additionally, reciprocal communication tends to diminish in larger networks, thus we did not anticipate influencers to frequently post submissions or comments. These activities are consequently assigned the least. Based on this method, the top 20 influencers we identified are:



**Top 100 Influencers Network:** The top 100 influencer network visualization is

