Social Media Analytics - Group Assignment

Krishan Gupta, Lakshya Agarwal,
Om Sangwan, Yash Joshi, Yiyi Yang

27th March, 2024

1 Section 1: Find predictors of influence

1.1 Introduction

Social networks significantly influence public opinion and consumer behavior, with certain individuals, known as *influencers*, playing a pivotal role due to their extensive reach and impact. Identifying these influencers is crucial for optimizing marketing strategies and maximizing engagement.

This report is based on a dataset from the Influencers in Social Networks Kaggle competition provided by PeerIndex, which includes pairwise comparisons of individuals' Twitter activities. The goal is to develop a machine learning model to predict which individual in each pair is more influential, based on pre-computed Twitter activity features and human judgments.

Through exploratory data analysis, feature engineering, and careful model selection, we aim to uncover the determinants of social influence and assess the potential financial impact of employing such analytics in marketing campaigns.

1.2 Data Description

The dataset underpinning our analysis comprises rows, each representing a pair of influencers, denoted as A and B. For each pair, various metrics detailing their Twitter activities are provided, such as follower counts, tweets made, and engagement metrics. Additionally, a binary label accompanies each pair, where '1' signifies that A is more influential than B, and '0' indicates the opposite. This streamlined dataset forms the basis of our predictive modeling efforts to discern the relative influence between pairs of individuals.

1.3 Methodology

Our structured approach to the analysis encompassed several critical stages:

- **Data Preprocessing:** Initial efforts were dedicated to cleaning the dataset for quality and consistency. This included addressing missing values, examining multicollinearity, and omitting network-specific features to streamline the data for subsequent analysis.
- **Feature Engineering:** At the forefront of our feature engineering efforts was the creation of a 'normalized engagement' feature. This was designed to encapsulate the overall interaction level of each individual, factoring in metrics like retweets, mentions, and replies, and then normalizing these values for uniformity. Building on this, we employed transformers to generate two distinct sets of features to further explore the influencers' dynamics. The first set utilized division to produce features representing relative differences, offering insights into the proportional disparities in influence metrics between pairs. Conversely, the second set applied subtraction, yielding features that elucidated the absolute differences in their Twitter activities. This dual approach allowed us to assess the impact of both relative and absolute metrics on influence prediction.

- **Model Selection and Training:** The evaluation of predictive models was conducted in a phased manner. Initially, models were trained on the raw data to establish a performance baseline. Subsequent phases involved the application of the engineered features, both subtractive and divisional, through the use of transformers. This iterative process, centered around accuracy as the key metric, facilitated the identification of the most efficacious model and feature set for discerning the more influential individual within each pair.

This comprehensive methodology, from meticulous data preprocessing to innovative feature engineering and strategic model evaluation, laid the groundwork for developing a predictive model adept at identifying influential individuals based on nuanced Twitter activity metrics.

1.4 Model Results & Financial Impact

Model Results

Our analysis yielded insightful outcomes, particularly regarding model performance and feature importance. Below, we summarize these findings and reference the relevant tables for detailed scores and metrics.

- **Model Performance:** Table 1 showcases the accuracy scores of different models using both subtraction and division feature sets. Notably, the Gradient Boosting model with division-based features (**76.36%** accuracy) outperformed other configurations, underscoring its efficacy in predicting influence based on Twitter metrics.
- **Feature Importance:** The significance of division-based features in the selected Gradient Boosting model is detailed in Table 2. The ratio of mentions received was identified as the most influential feature, highlighting the critical role of user engagement in determining online influence.
- **Comparison of Feature Sets:** The direct comparison between subtraction and division feature sets, as illustrated in Table 1, reaffirms our choice of division-based features for the final model. This choice is based on the nuanced understanding these features provide regarding the relative dynamics of influence.

Model	Subtraction Set Score	Division Set Score
Random Forest	75.55%	75.55%
Gradient Boosting	75.27%	**76.36%**
AdaBoost	75.91%	74.36%

Table 1: Model performance comparison using subtraction and division feature sets.

Feature	Importance
Mentions Received Ratio	54.33%
Listed Count Ratio	13.28%
Retweets Received Ratio	9.39%
Follower Count Ratio	8.42%
Following Count Ratio	4.63%
Mentions Sent Ratio	4.32%
Engagement Ratio	3.55%
Retweets Sent Ratio	2.08%

Table 2: Feature importance in the selected Gradient Boosting model.

1.4.1 Financial Impact

A critical aspect of our analysis was to evaluate the financial benefits of implementing our analytic model in a real-world scenario. Specifically, we aimed to quantify the increase in net profit from using our model to identify influencers for a marketing campaign, as opposed to a non-analytic approach. The following points summarize our findings:

- **Without Analytics:** Traditionally, offering a flat fee to individuals for promotion results in a net profit of $7.29 million. This approach does not differentiate between influencers and non-influencers, potentially leading to suboptimal allocation of marketing resources.
- **With Our Model:** By employing our Gradient Boosting model to selectively engage influencers for more targeted promotions, the net profit significantly increases to $14.2 million. This nearly doubles the profitability by ensuring that marketing efforts are concentrated on individuals with a higher likelihood of influencing purchasing decisions.
- **Comparison to Perfect Model:** While our model captures 78.70% of the potential profit increase achievable with a hypothetical perfect model, it represents a substantial improvement over the non-analytic approach, enhancing net profits by 94.79%.

To further illustrate these financial outcomes, we present the calculations underlying our analysis:

Scenario	Net Profit ($)
Without Analytics	7,289,539
With Our Analytic Model	**14,198,245**
With a Perfect Model	16,074,705

Table 3: Net profit comparison across different scenarios.

1.5 Conclusion

This study's journey into Twitter's influence dynamics underscored the value of division-based features, revealing that relative metrics like engagement ratios are more indicative of influence than absolute numbers. Specifically, features such as the ratio of mentions received emerged as pivotal in distinguishing influencers, highlighting the essence of engagement over follower size.

Our financial impact assessment demonstrated the analytic model's capacity to nearly double marketing campaign profitability by employing a data-driven approach to influencer selection. This significant enhancement in net profit underscores the practical benefits of integrating machine learning into marketing strategies.

These findings not only affirm the model's efficacy but also offer a foundation for further exploration into refining influencer identification methods. The insights gained lay the groundwork for more sophisticated, targeted, and effective marketing endeavors in the evolving landscape of social media.

2 Section 2: Finding influencers from Reddit

2.1 Introduction

In Section 2 of our analysis, we focus on the practical application of social network analytics to identify influencers within a specific online community. This section builds on the theoretical foundations discussed in class and applies methodologies to real-world data from Reddit. Our goal is to leverage network analysis techniques to pinpoint key individuals in the subreddit "IndiaInvestments" who wield significant influence over discussions and user engagement.

2.2 Methodology

2.2.1 Data Collection and Preparation

We collected data on submissions and comments from the "IndiaInvestments" subreddit, ensuring to exclude any entries by deleted users to maintain the integrity of our analysis. The collected data were then transformed to highlight the relationships between authors, submissions, and comments, setting the stage for network construction.

2.2.2 Network Construction

Using the `NetworkX` library, we constructed a directed graph representing the interaction dynamics within the subreddit. Nodes in the graph correspond to users (authors of comments), while edges represent the direction of communication (e.g., a comment on a post or on a submission). This graph served as the basis for our subsequent analysis.

2.2.3 Centrality Measures

To understand the influence dynamics within the network, we calculated three centrality measures for each node:

- **Degree Centrality**: Reflects the number of connections each node has, indicating general activity and visibility within the network. For a graph $G = (V, E)$ with $|V|$ vertices, the degree centrality for a vertex v is defined as the fraction of nodes it is connected to:

  $$C_D(v) = \frac{deg(v)}{|V| - 1}$$

  where $deg(v)$ is the degree of vertex v, and $|V| - 1$ accounts for v itself not being included in its own degree count. This can be calculated in NetworkX using the `degree_centrality(G)` function.

- **Betweenness Centrality**: Measures the extent to which a node lies on the shortest path between other nodes, highlighting those who play a crucial role in information flow. It quantifies the number of times a node acts as a bridge along the shortest path between two other nodes and is given by:

  $$C_B(v) = \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

  where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. NetworkX provides the `betweenness_centrality(G)` function for this purpose. To speed up calculation, we took a subset of $k = 500$ nodes.

- **Closeness Centrality**: Captures how close a node is to all other nodes in the network, reflecting the ability to quickly interact with the entire community. It is calculated as the inverse of the average shortest path length from a given node to all other reachable nodes:

$$C_C(v) = \frac{|V| - 1}{\sum_{u \in V \setminus \{v\}} d(v, u)}$$

  where $d(v, u)$ is the shortest-path distance between v and u, and $|V|-1$ is the normalization factor. The `closeness_centrality(G)` function in NetworkX can be used to calculate this measure.

2.2.4 Calculating User Engagement Metrics

Understanding the dynamics of user engagement within the "IndiaInvestments" subreddit necessitated a granular analysis of user-generated content and interactions. This section elucidates the methodologies employed to quantify four key metrics: Number of Submissions, Number of Comments, Response to Comments, and Response to Submissions. These metrics serve as foundational pillars for our subsequent analysis, providing insight into the multifaceted nature of user influence and participation.

- **Number of Submissions**: This metric represents the total count of submissions made by each user. This metric was derived by aggregating the submissions dataset based on the author attribute, thereby quantifying each user's contribution to the subreddit in terms of submitted content. This process not only highlights the most active users in terms of content creation but also underscores the diversity and volume of topics introduced to the community.
- **Number of Comments**: Parallel to submissions, this metric encapsulates each user's engagement through the lens of comment activity. By tallying the comments attributed to each user, we gauged the extent of their participation in discussions across the subreddit. This metric offers a window into the interactive aspect of user engagement, reflecting the vibrancy of community dialogues and exchanges.
- **Response to Comments**: This metric was calculated to assess user responsiveness within threads. This involved analyzing the comments dataset to identify instances where a user's comment directly responds to another comment. By distinguishing these interactions, we obtained a measure of direct user-to-user engagement, highlighting individuals who actively contribute to deepening discussions and fostering a collaborative community environment.
- **Response to Submissions**: Similarly, this metric was determined by identifying comments that serve as initial responses to submissions. This metric illuminates the interface between content creation and community response, showcasing the propensity of users to engage with newly introduced topics. It reflects the initial wave of community interaction elicited by submissions, underscoring the role of submissions in sparking discussion.

Together, these metrics paint a comprehensive picture of user engagement within the "IndiaInvestments" subreddit. By quantifying both content creation and interaction, we derived a holistic understanding of user influence, laying the groundwork for identifying the most influential nodes within the community.

2.3 Identifying the Most Influential Nodes

In our quest to identify the most influential nodes within the "IndiaInvestments" subreddit, we employed a comprehensive analytical strategy that integrated both network-centric metrics and user engagement indicators. This approach necessitated a detailed aggregation of user-generated content metrics — specifically, the number of submissions and comments, alongside the user's engagement through responses to comments and submissions. By adopting this multifaceted perspective, we gained a nuanced understanding of user influence that extends beyond mere network position to include active participation within the subreddit's discourse.

A cornerstone of our analysis was the application of Principal Component Analysis (PCA). PCA proved instrumental in condensing the aforementioned metrics into a cohesive analytical framework, enabling us to isolate the principal components that encapsulate the core attributes of influence within the community. This reduction technique significantly streamlined our dataset while preserving its essential informational value. Prior to conducting PCA, we utilized a `StandardScaler` to normalize our data, ensuring each variable contributed equitably to the analysis and mitigating potential biases arising from variable scale differences.

The calculation of the final composite score for each node was meticulously executed as follows:

$$Score(v) = \sum_{i=1}^{n} \text{PCA}_{\text{weight}}^{i} \cdot \text{Metric}_{\text{value}}^{i}$$

Here, $\text{PCA}_{\text{weight}}^{i}$ denotes the weight assigned to each metric as determined by PCA, and $\text{Metric}_{\text{value}}^{i}$ represents the value of the metric for node v. This formula incorporates all metrics analyzed in the PCA, including:

- Closeness Centrality (*closeness_centrality*)
- Number of Submissions (*number_of_submissions*)
- Number of Comments (*number_of_comments*)
- Response to Comments (*respond_comment*)
- Response to Submissions (*respond_submission*)

The weights assigned to these components by the PCA reflect their relative importance in defining influence within the subreddit. This scoring mechanism emphasizes that influence is inherently multifaceted, stemming not only from one's position within the network but also from their engagement and the community's response to such activities.

By leveraging PCA in this manner, we accurately quantified the various dimensions of user influence, anchoring our findings in a statistically sound interpretation of the data. This enabled us to rank nodes according to their overall influence, pinpointing those individuals who significantly shape the community's discourse and interactions.

2.4 Results and Visualization

Our analysis uncovered key influencers in the "IndiaInvestments" subreddit, with "AutoModerator," "crimelabs786," and "srinivesh" standing out as prominent figures. Notably, "AutoModerator" and "crimelabs786" are moderators of the subreddit, playing a critical role in content curation and community management. In contrast, "srinivesh" emerges as a highly influential member, contributing significantly through active engagement despite not holding a moderator role. These distinctions underline the diverse ways in which individuals can wield influence within online communities, whether through formal governance roles or through high levels of participation.

To depict the dynamics of influence within the subreddit visually, we crafted a plot that highlights the top 100 most influential users, as determined by our composite score blending centrality measures with user activity data. This graphical representation offers a vivid snapshot of the network's structural intricacies, emphasizing the pivotal roles that the top influencers play in bridging different segments of the community.

The ensuing plot underscores the interconnectedness prevalent among users in the "IndiaInvestments" subreddit. Node sizes within the plot are scaled according to the degree of influence, with larger nodes denoting users who boast higher centrality scores and are more actively engaged:

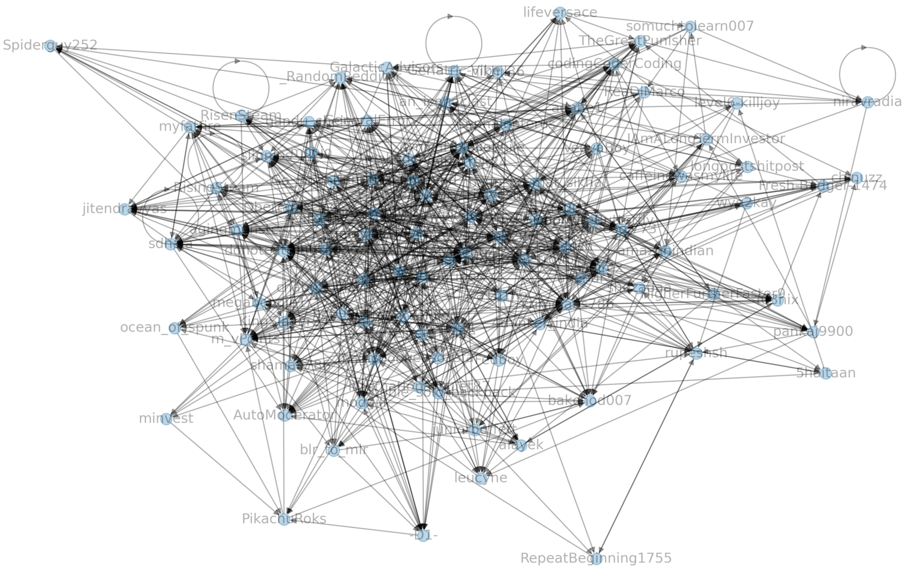

Figure 1: Top 100 Influential Users in the "IndiaInvestments" Subreddit

To view an interactive version of the network, please open the HTML file submitted alongside this paper.

2.5 Implications for Business and Research

Understanding the roles these influencers play can be valuable for businesses looking to engage with communities or for researchers studying information dissemination in social networks. By targeting influencers for outreach or monitoring their interactions for insights, stakeholders can more effectively navigate and leverage the dynamics of online communities.

2.6 Conclusion

The application of social network analysis techniques to Reddit data demonstrates the power of centrality measures and PCA in uncovering the underlying structure of online communities.