Section 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.

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.

Methodology

Our methodology comprised a structured approach, focusing on the following key aspects:

- **Data Preprocessing:** The initial phase involved cleaning the dataset to ensure its quality and consistency. We checked for missing values and assessed multicollinearity among the features to prepare the data for analysis.
- **Feature Engineering:** We developed two sets of features to evaluate the influence of individuals on Twitter. The first set was created by subtracting metrics of individual B from individual A (e.g., follower counts, tweet frequencies), aiming to highlight absolute differences. The second set involved dividing A's metrics by B's to capture relative differences, providing insight into the proportionality of influence.
- **Model Selection and Training:** In our pursuit of the most effective predictive model, we conducted a thorough examination of several machine learning algorithms, applying them to both subtractive and divisional feature sets. The evaluation process was guided by accuracy as the primary metric, allowing us to gauge each model's performance in discerning the more influential individual within pairs.

This streamlined process enabled us to develop a model that effectively predicts the more influential individual in a pair, leveraging nuanced insights derived from Twitter metrics.

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.

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.

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.