Influencer Identification Using Clustering and Classification

Group 5

Introduction

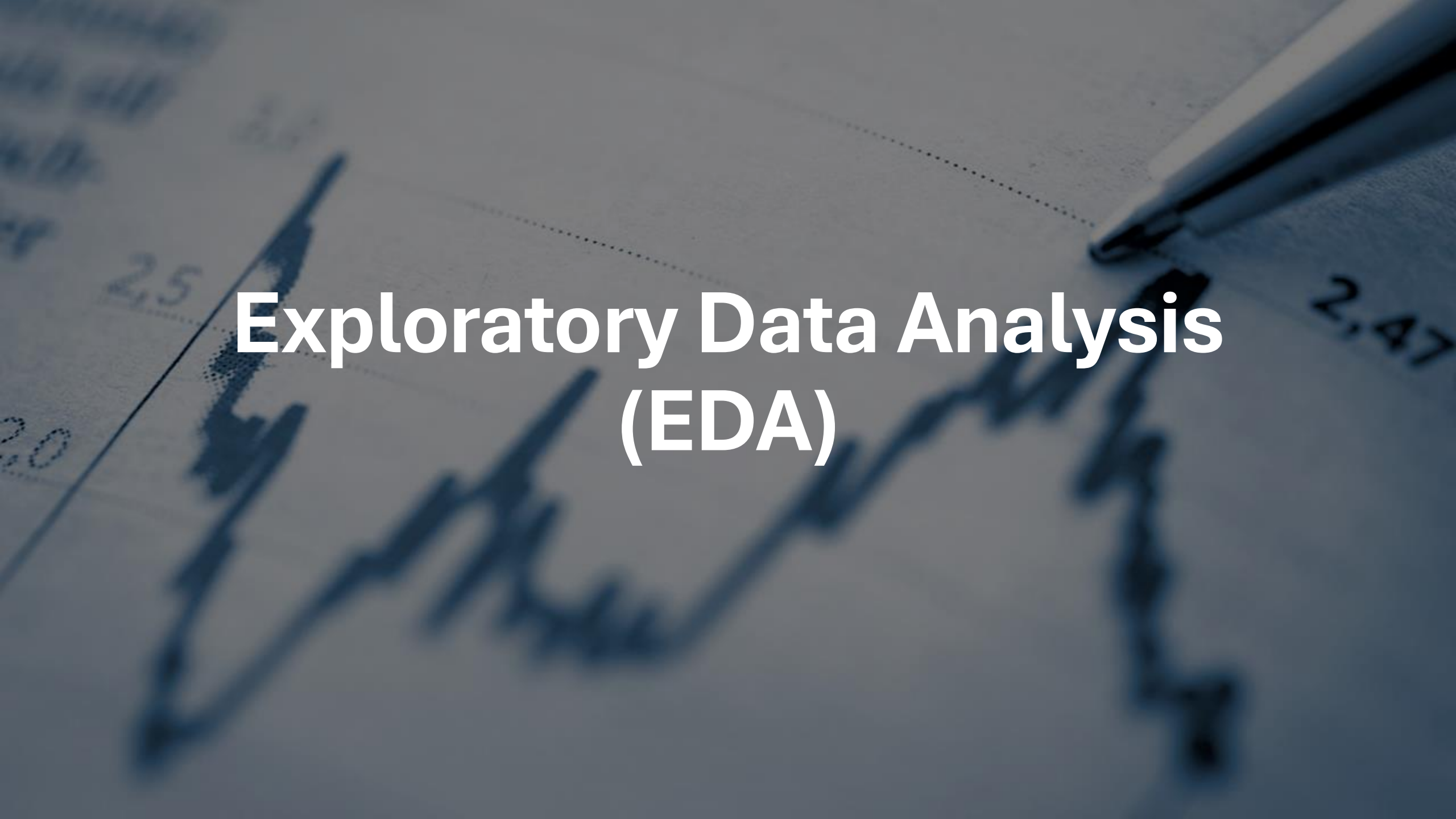
Objective: To identify potential influencers in the TikTok profiles dataset using machine learning techniques.

Approach: Utilize clustering algorithms to group similar profiles and classification models to predict influencer likelihood.

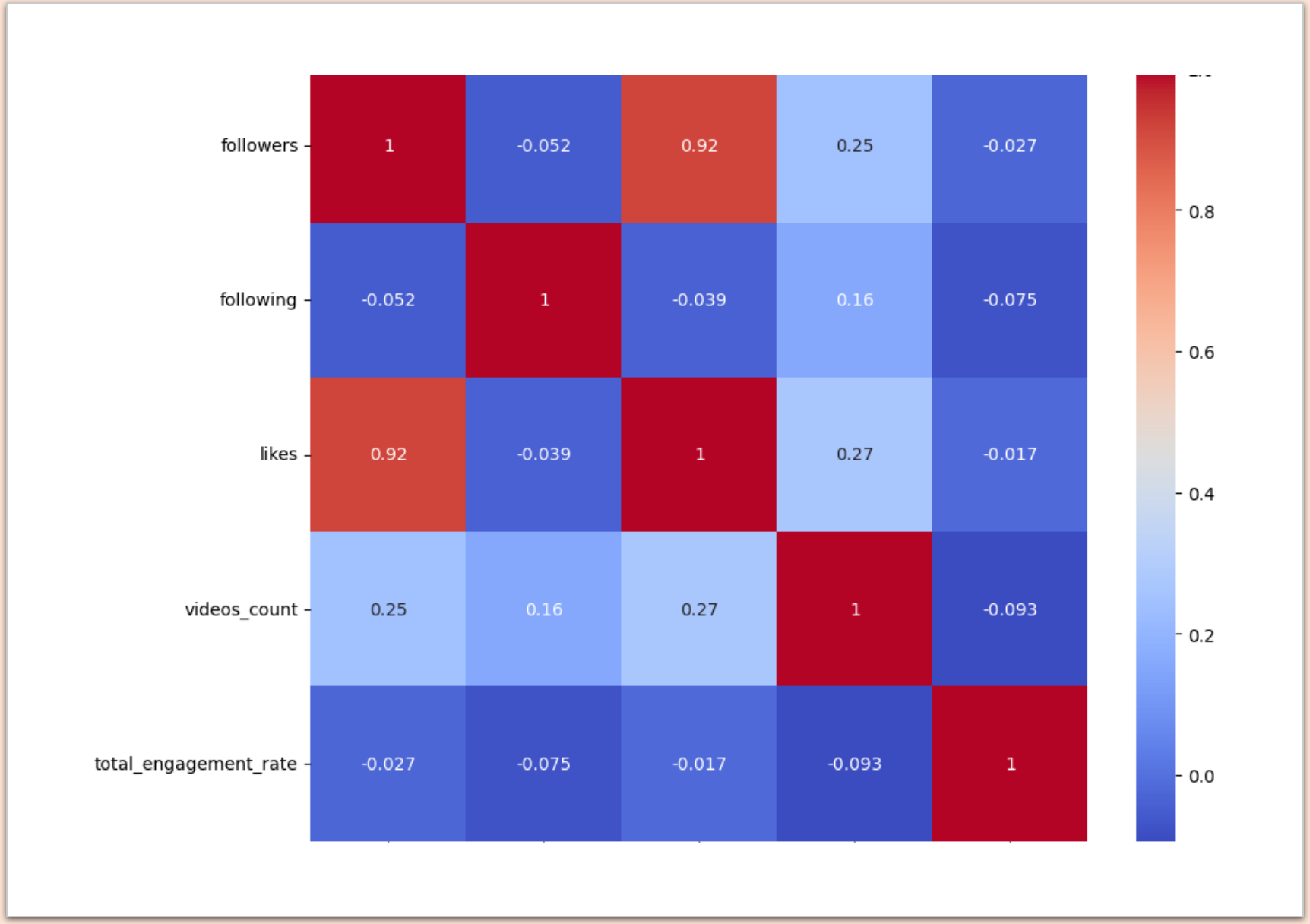
Importance: Helps in targeted marketing and understanding influencer impact.

Data Overview

- **Dataset:** TikTok profiles with features like followers, likes, videos count, and engagement rates.
- **Preprocessing:** Handling missing values and standardizing features.



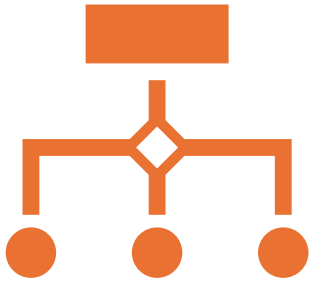
Exploratory Data Analysis (EDA)



Correlation Matrix

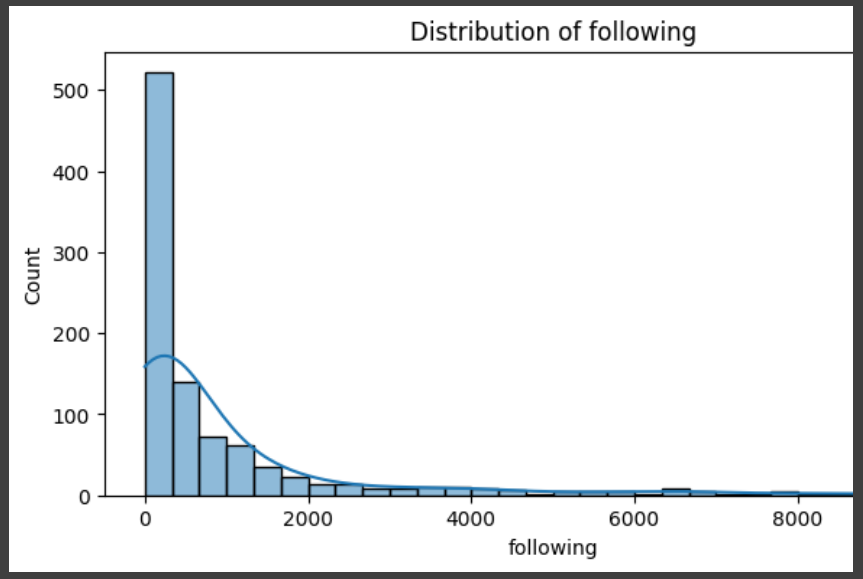
"The correlation matrix shows strong positive correlations between certain features, such as followers and likes, indicating that profiles with more followers tend to receive more likes. Understanding these relationships helps in building accurate models for influencer identification."

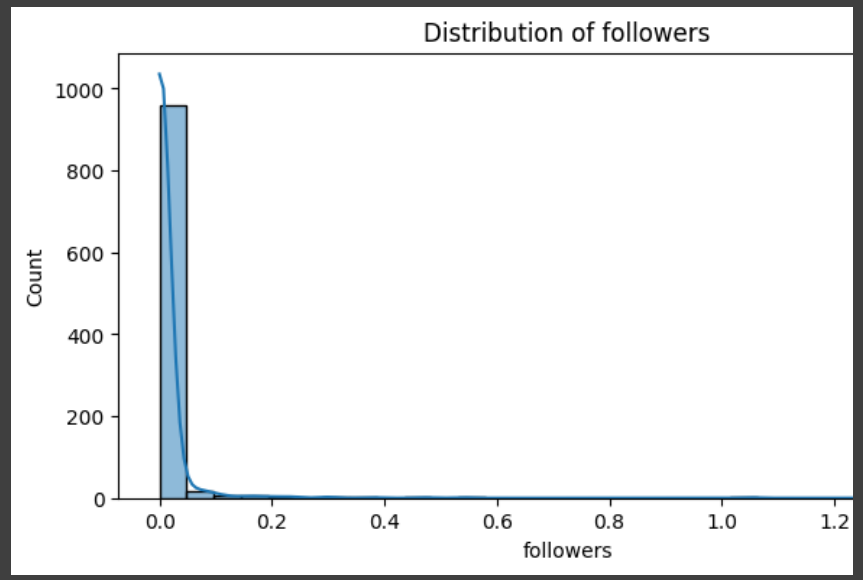
Handling Missing Values & Feature Engineering

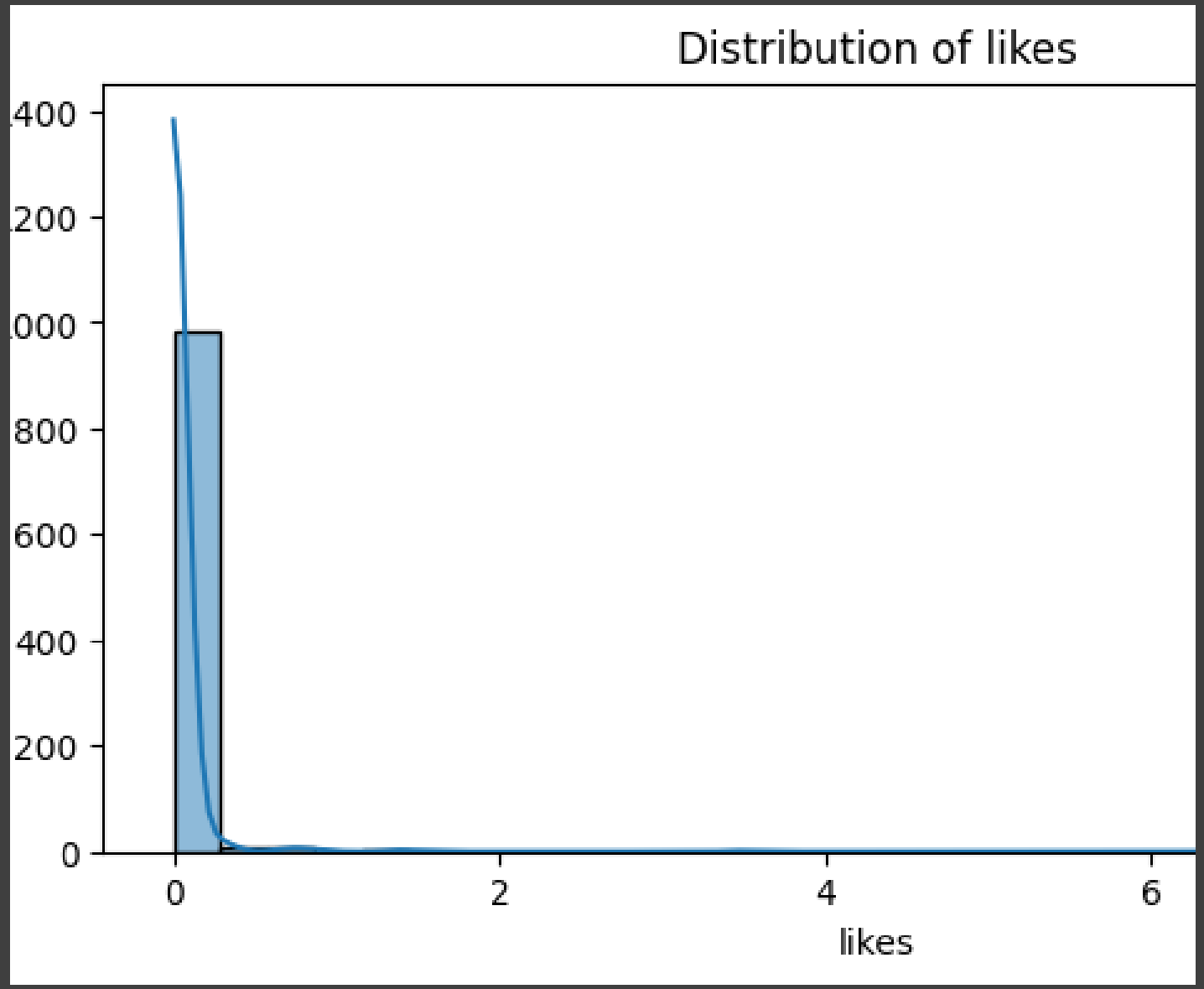


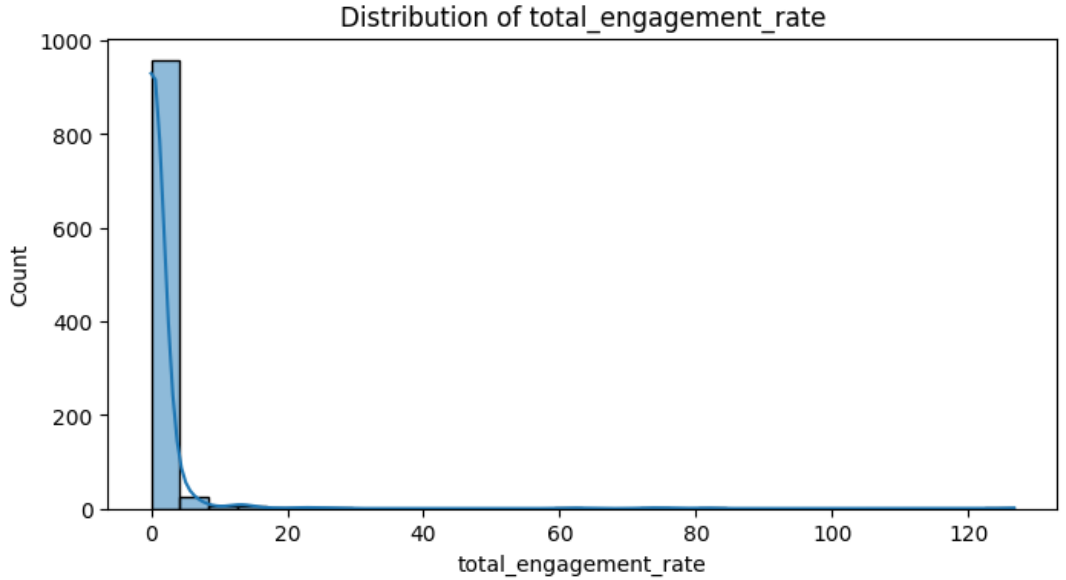
We imputed missing values using 'mean' for numerical features and also applied standard scalar to the features which is very important for clustering algorithm.

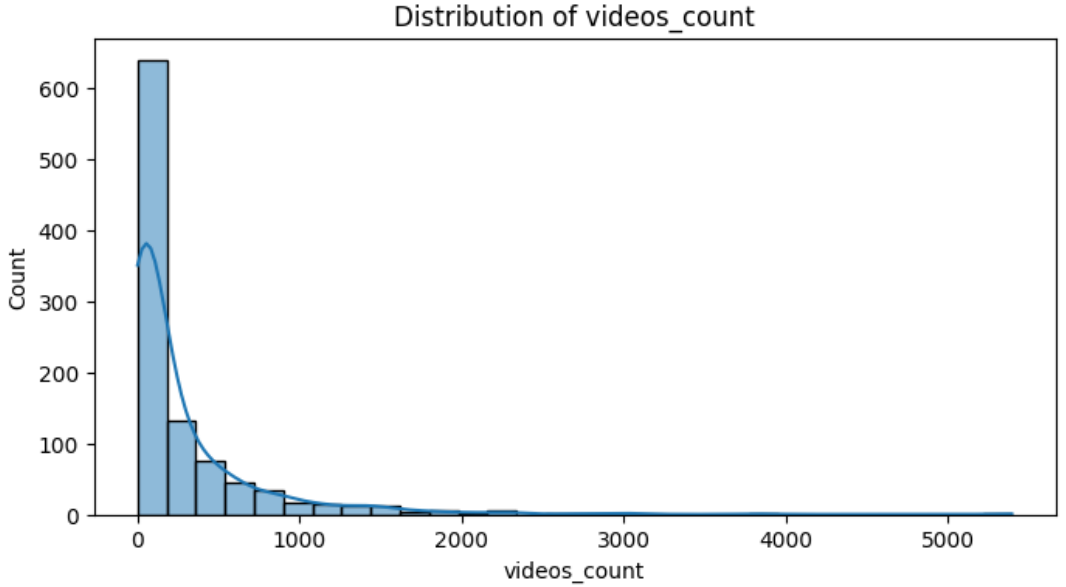
Also, we combined 'Avg_engagement', 'comment_engagement' and like engagement into 'total_engagement'.


Distribution of following
Count
following
0
100
200
300
400
500
0
2000
4000
6000
8000


Distribution of followers
Count
followers
0
200
400
600
800
1000
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Distribution of likes
likes
0
200
400
600
800
000
400
0
2
4
6


Distribution of total_engagement_rate
Count
total_engagement_rate
0
200
400
600
800
1000
0
20
40
60
80
100
120

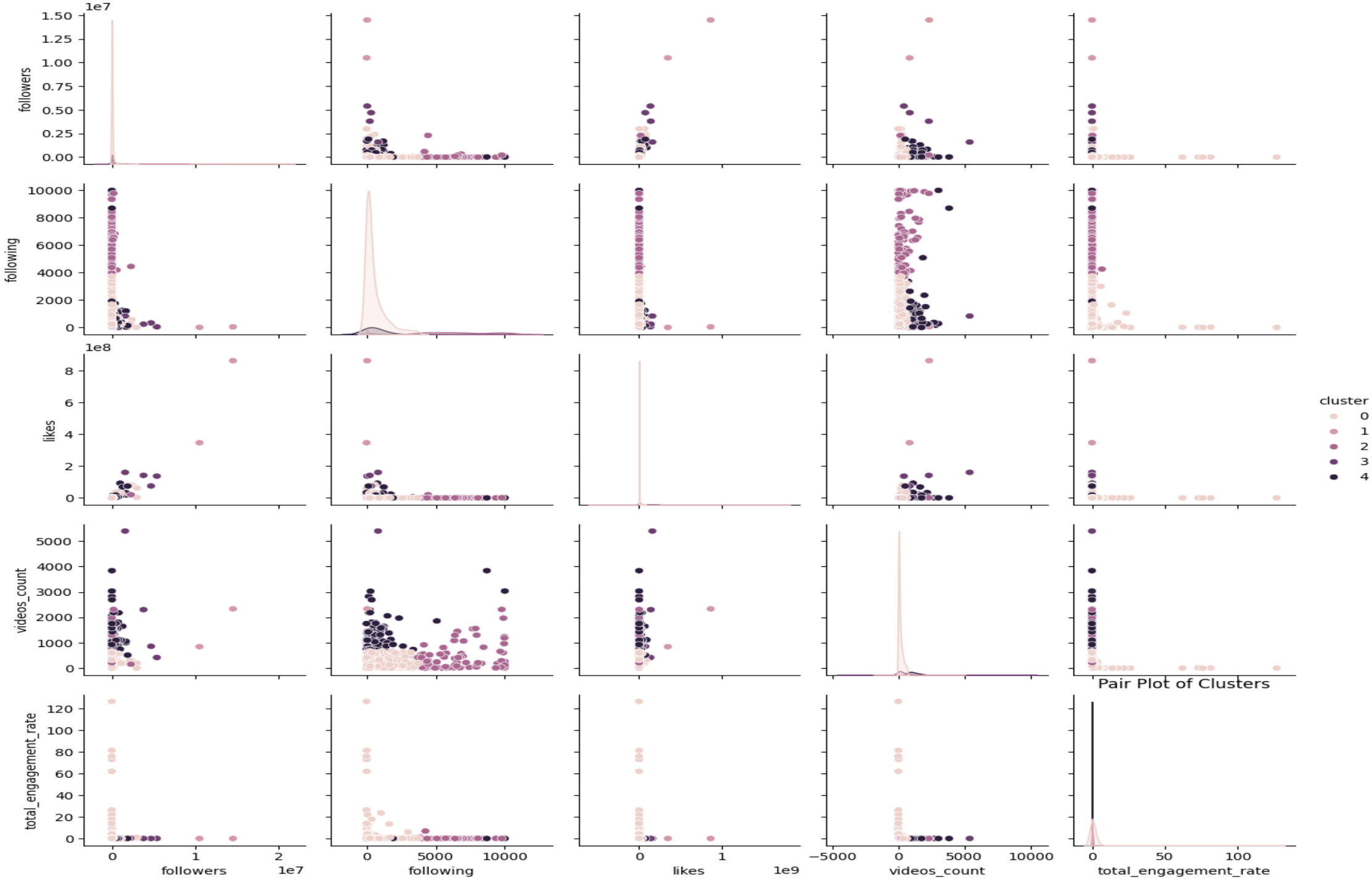

Distribution of videos_count
Count
videos_count
0
100
200
300
400
500
600
0
1000
2000
3000
4000
5000

Clustering

Algorithm: K-Means Clustering

Objective: Group similar profiles

Evaluation: Silhouette Score of 0.598 indicates well-defined clusters.


Pair Plot of Clusters
followers
following
likes
videos_count
total_engagement_rate
1e7
1e8
1e9
cluster
0
1
2
3
4

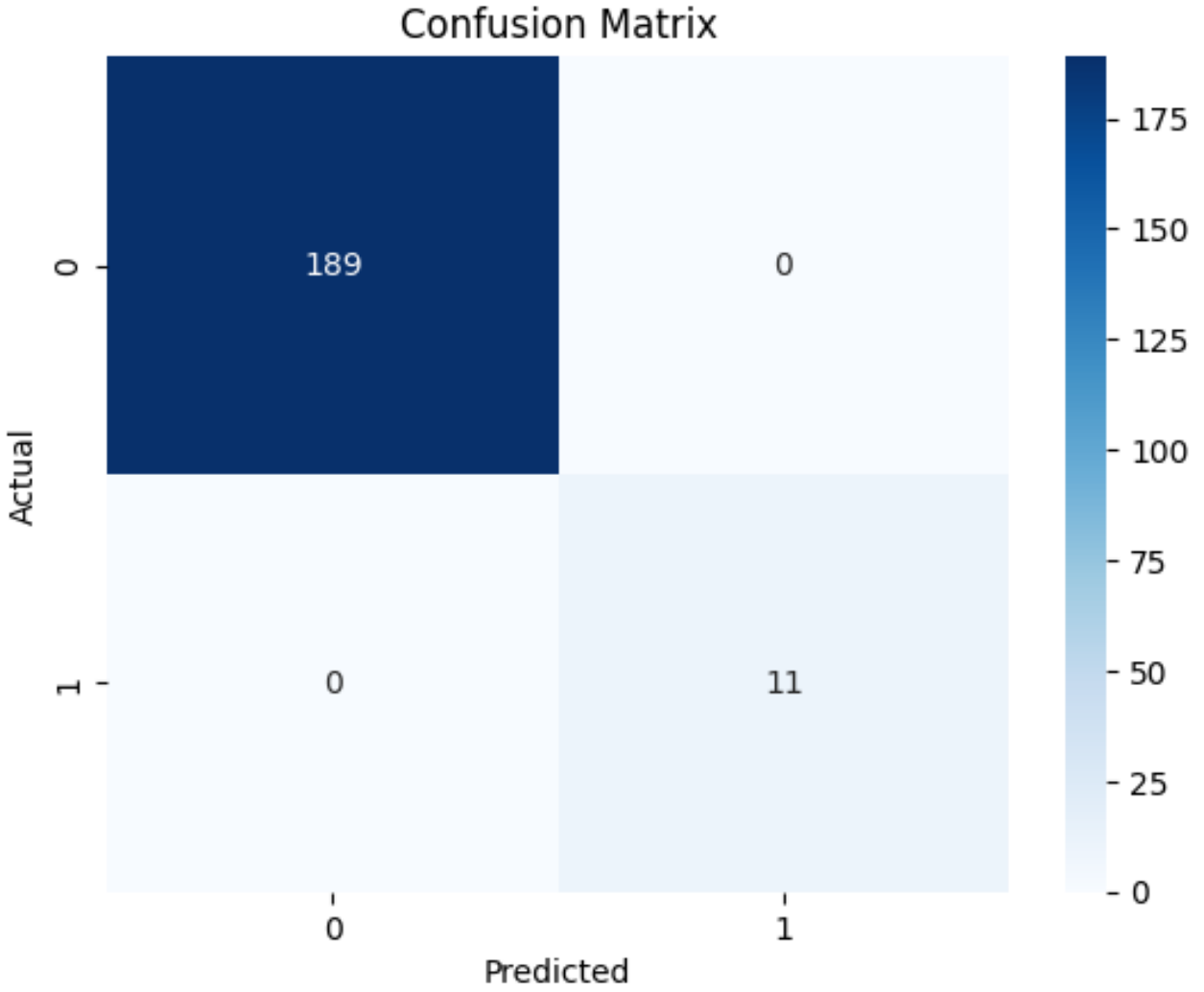
Classification

Model: Random Forest Classifier

Features: Followers, likes, videos count, engagement rate

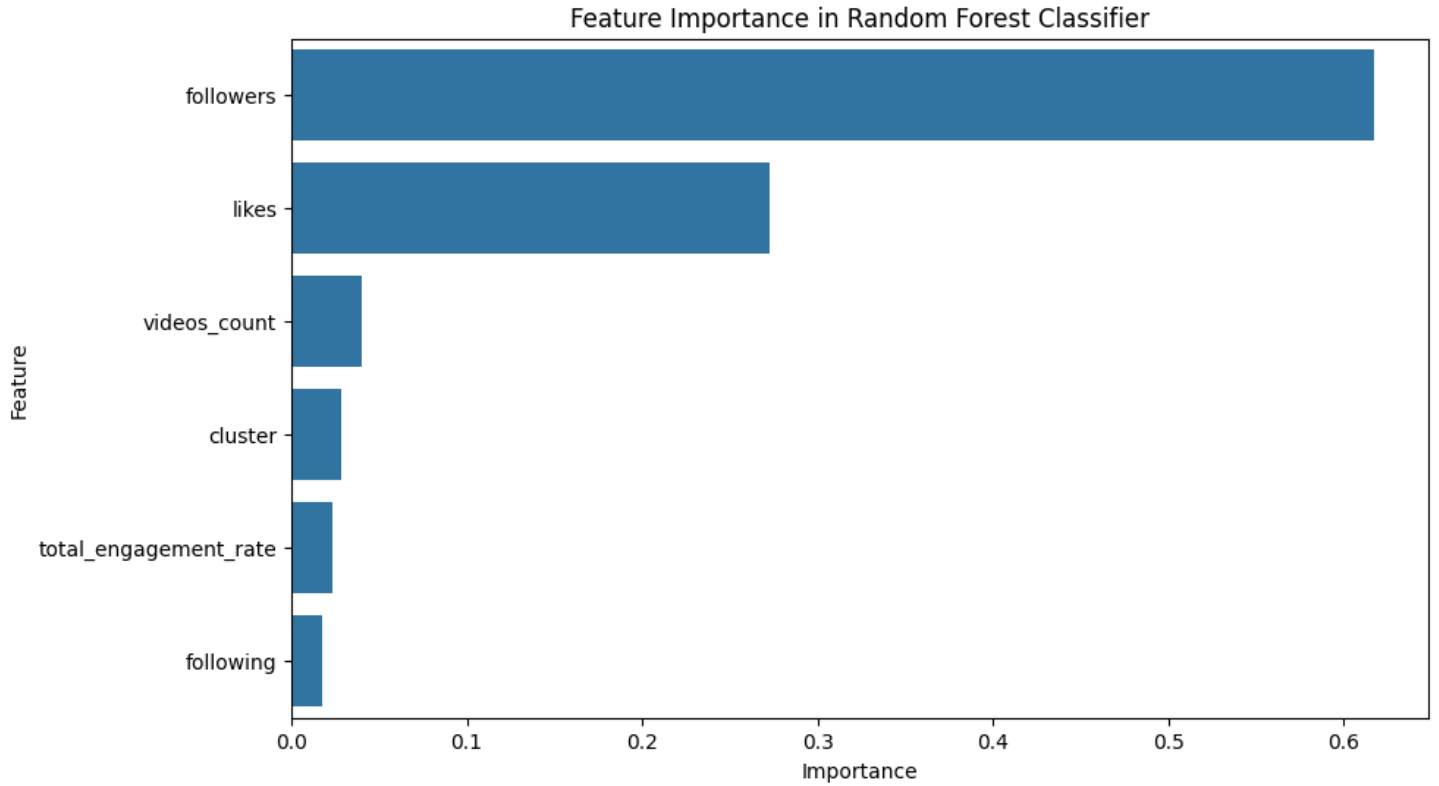
Target: Identifying influencers (top 5% followers)

"The confusion matrix shows perfect classification performance with all actual values matching predicted values, indicating the effectiveness of the Random Forest classifier in identifying influencers."



Feature Importance

"The feature importance plot shows that follower count and likes are the most significant features in identifying influencers, providing valuable insights for targeted marketing strategies."



Business Decisions

- **Target High-Engagement Profiles**:

Prioritize marketing on profiles with high engagement and follower counts to maximize ROI.

- **Data-Driven Influencer Partnerships**:

Collaborate with top influencers identified through data analysis to ensure effective partnerships.

- **Optimize Content Strategy**:

Guide content creation by analyzing successful content types from top influencers.

- **Resource Allocation**:

Focus resources on high-impact clusters to ensure efficient use of marketing budgets.

Marketing Strategies

- **Personalized Marketing Campaigns**:

Develop campaigns tailored to the audience of high-engagement profiles for increased effectiveness.

- **Leverage Micro-Influencers**:

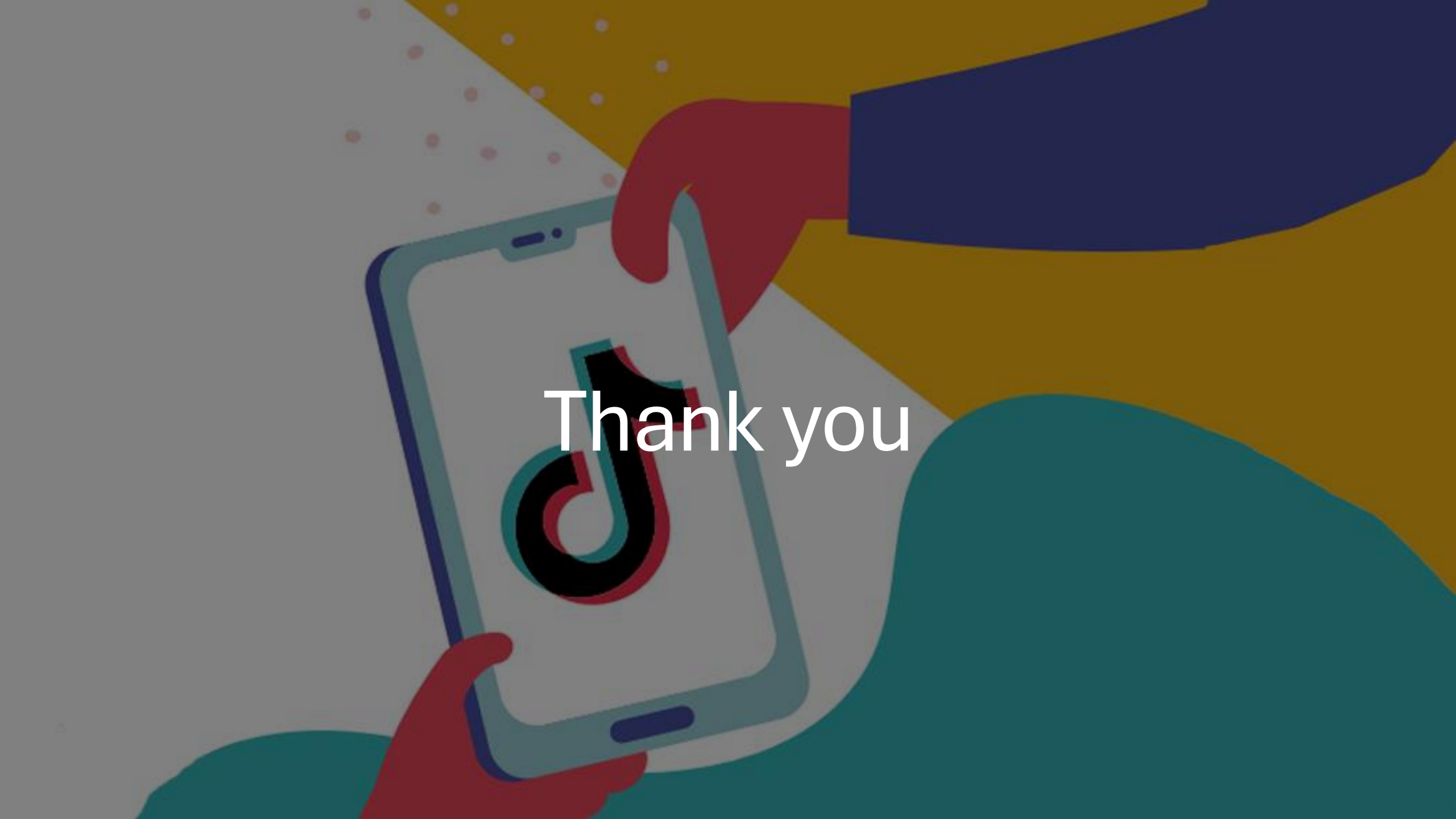
Collaborate with micro-influencers in niche markets to build authentic connections at a lower cost.

- **Continuous Monitoring and Adaptation**:

Monitor performance and adjust strategies based on real-time data to maintain campaign relevance.

- **Expand Influencer Network**:

Discover and engage with emerging influencers to stay ahead of the competition.


Thank you