**Final Report: Social Media Engagement (Instagram) Analysis**

**Arafat Ahmed**  
 **Course Project for MKT-568 Fall 2024**  
 **WPI Business School, MA, USA**

**Executive Summary**

This project aimed to explore factors influencing social media engagement, specifically likes and comments on Instagram posts. Utilizing data analytics and machine learning techniques, I identified key predictors, proposed new variables, and analyzed engagement strategies for micro and macro influencers. Our key findings include:

1. The relationship between FOLLOWERS and FOLLOWING consistently emerged as the most significant predictor of likes and comments.
2. Timing strategies differ for micro (evening) and macro (afternoon) influencers.
3. Weekend posts outperform weekday posts in generating engagement.
4. Sentiment-driven content significantly influences comments for micro influencers.

These insights inform actionable strategies for influencers and marketers to maximize engagement. The findings highlight the importance of tailored content strategies and data-driven decisions in digital marketing.

**Introduction**

Instagram has evolved into one of the most influential platforms for personal branding, marketing, and public engagement, with users ranging from individual influencers to global organizations. For content creators and businesses alike, maintaining high levels of user engagement on their posts—measured through likes, comments, and shares—is a critical metric for success. Engagement helps influencers understand their audience’s preferences, optimize Instagram’s algorithmic visibility, and ultimately increase their reach and influence. Despite the tools Instagram provides to measure engagement, these insights are often limited to the influencer’s own dataset, preventing broader understanding of what drives engagement universally. Therefore, analyzing a larger dataset offers the potential to uncover universal strategies for creating engaging posts.

This project addresses the pressing question of what factors contribute to higher engagement on Instagram posts. Leveraging a dataset of Instagram posts—including metrics such as likes, comments, and content characteristics—we employed a structured analytical approach to answer this question. The study focused on understanding the relationship between engagement metrics and independent variables like the timing of posts, post content (e.g., hashtags, tagged users), and newly created variables such as sentiment scores and text complexity. To achieve this, linear regression models were built to evaluate the significance of these factors in predicting engagement metrics, and standardized beta coefficients were used to determine the most influential predictors.

The findings reveal valuable insights for influencers aiming to improve their content strategies. For instance, timing and text complexity showed consistent influence on engagement, while certain variables such as tagged users were particularly significant for specific influencers. These results emphasize the need for tailored strategies based on an influencer’s audience size and type of content. Ultimately, this project not only provides actionable recommendations for maximizing engagement but also establishes a framework for influencers to adapt their strategies dynamically based on data-drive

**Objectives:**

* To identify key variables influencing Instagram engagement metrics (likes and comments).
* To propose and evaluate new variables to improve predictive models.
* To compare engagement strategies for micro and macro influencers.
* To provide actionable recommendations for optimizing engagement strategies.

**Problem Significance:**

Social media platforms are critical marketing tools, with influencers playing a pivotal role in shaping consumer behavior. However, engagement levels vary widely, influenced by post content, timing, and audience demographics. By understanding these factors, influencers and marketers can refine strategies to maximize impact.

**Methodology**

This analysis was conducted following the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which provided a structured framework for solving the problem. Each stage of the process was carefully executed to ensure the analysis was thorough, accurate, and actionable.

1. **Business Understanding**: The first step was to clearly define the goal of the analysis: identifying the factors that drive user engagement on Instagram posts. Engagement metrics such as likes and comments were considered critical indicators of a post’s success. By framing the problem within the context of influencers’ strategies and Instagram’s algorithms, the analysis aimed to provide actionable insights for maximizing audience interaction.
2. **Data Understanding**: The dataset consisted of Instagram post attributes, including variables such as the number of hashtags, type of post (photo or video), the sentiment of captions, and user interaction metrics like likes and comments. This stage involved exploring the data, understanding its structure, and identifying patterns and potential challenges such as missing or inconsistent values.
3. **Data Preparation**: Preparing the data for analysis was a crucial step. This involved cleaning the dataset by handling missing values, transforming categorical variables into numerical formats, and creating new variables to enhance the analysis. For instance, a sentiment score for captions was calculated using the TextBlob library to quantify the positivity or negativity of text. Additionally, a text complexity index (average word length) and an engagement ratio (followers-to-following ratio) were derived to capture nuanced aspects of engagement.
4. **Modeling**: Linear regression models were built to predict engagement metrics (likes and comments) using various predictors such as post type, timing of posts, hashtags, sentiment scores, and text complexity. Separate models were created for micro-influencers (fewer than 50,000 followers) and macro-influencers (50,000 or more followers) to account for differences in audience dynamics.
5. **Evaluation**: The models were evaluated for their performance using metrics like R-squared and mean squared error (MSE). Standardized beta coefficients were calculated to identify the most influential predictors. These coefficients were then visualized in graphs, sorted by their absolute values, to highlight the key drivers of engagement for different types of influencers.
6. **Deployment**: The insights derived from the analysis were translated into actionable recommendations for social media influencers. For example, timing and text complexity were identified as critical factors for engagement. These findings were communicated in a clear and concise format, ensuring their usability for influencers aiming to optimize their content strategies.

This systematic approach not only ensured that the analysis was comprehensive but also allowed for a deep understanding of the factors influencing Instagram engagement. By narrating the steps in the process and linking them to the project’s goals, this methodology demonstrates how a structured framework can yield meaningful and actionable insights.

**Data**

**Data Overview:**

* **Source**: Publicly available Instagram post data.
* **Size**: **19,681 entries** and **14 columns**
* **Variables**: Predictors included:
  + Followers, following, hashtags, users tagged, post type (photo/video).
  + Timing: Morning, afternoon, evening, night.
  + New variables: Sentiment Score, Text Complexity, Engagement Ratio.

The dataset used for this analysis was a curated sample of Instagram posts, encompassing a variety of features that capture both the content and context of the posts. Key variables included engagement metrics such as the number of likes (LOG\_LIKES) and comments (LOG\_COMMENTS), textual attributes like captions (TEXT), and categorical variables such as post type (photo or video). Additionally, temporal features like the time of posting (e.g., morning, afternoon, evening, night) and user-specific variables, including the number of followers (FOLLOWERS), following count (FOLLOWING), and users tagged in the post (USERS IN PHOTO), provided a rich set of predictors for engagement.

Derived features were also included to enhance the analysis. For example, a sentiment score was calculated for the captions using the TextBlob library, offering insights into the emotional tone of the text. A text complexity index was created by computing the average word length in captions, serving as a proxy for linguistic sophistication. Furthermore, an engagement ratio (ratio of followers to following) was derived to capture the social dynamics of each influencer. The dataset, carefully cleaned and processed to address missing values and inconsistencies, offered a comprehensive foundation for identifying the factors driving user engagement on Instagram.

**A purple and yellow barcode

Description automatically generatedMissing Data and Outliers:**

* Missing data was minimal and handled through mean imputation or removal.

A comparison of a box plot

Description automatically generated

* A blue and green box diagram

  Description automatically generatedOutliers were capped using standard deviation thresholds.

**Data Transformation**

* Variables were dummy coded as needed (e.g., Timing categories).
* A screenshot of a computer screen

  Description automatically generatedSentiment Score, Text Complexity, and Engagement Ratio were engineered and added.

**Analysis Results**

**Question 1: Identifying Key Predictors of Likes and Comments -Understanding the Key Drivers of Instagram Engagement**

Engagement metrics like likes and comments are critical indicators of an Instagram post's success. This section explores the results of two separate linear regression models aimed at identifying the factors influencing the number of likes and comments on Instagram posts. By analyzing these metrics, we gain valuable insights into how various content and contextual features impact user interactions.

**Methodology Overview**

To predict engagement, two dependent variables were analyzed:

* **LOG\_LIKES:** Log-transformed number of likes
* **LOG\_COMMENTS:** Log-transformed number of comments

The independent variables included:

1. Number of followers
2. Number of following
3. Post timing (dummy-coded: Morning, Afternoon, Evening, Night)
4. Number of hashtags used
5. Number of users tagged in the post
6. Type of post (Photo vs. Video, dummy-coded as PostType\_Video)
7. Length of post text (WORD\_COUNT)

Key preprocessing steps involved:

* Creating dummy variables for categorical data
* Handling missing values
* Standardizing the data to compute beta coefficients

**Model Performance and Insights**

**Model Performance Metrics:**

1. **LOG\_LIKES:**
   * Mean Squared Error (MSE): 0.7146
   * R-squared (R²): 27.4%
2. **LOG\_COMMENTS:**
   * Mean Squared Error (MSE): 0.8673
   * R-squared (R²): 13.7%

These metrics suggest moderate predictive power for likes and a weaker fit for comments.

**Significant Variables and Their Impacts**

**LOG\_LIKES:**

* **Positive Influences:**
  + **FOLLOWERS:** Higher follower counts lead to more likes.
  + **Night Posting:** Posts made at night perform better.
* **Negative Influences:**
  + **FOLLOWING:** Following more accounts negatively impacts likes.
  + **PostType\_Video:** Videos receive fewer likes than photos.

**LOG\_COMMENTS:**

* **Positive Influences:**
  + **FOLLOWERS:** Increases in followers positively affect comments.
* **Negative Influences:**
  + **FOLLOWING:** Following more accounts reduces comments.
  + **USERS IN PHOTO:** Posts tagging more users receive fewer comments.
  + **PostType\_Video:** Videos receive fewer comments than photos.

**Key Observations:**

* **FOLLOWERS** emerged as the strongest predictor for both likes and comments, highlighting its importance in driving engagement.
* The negative impact of **FOLLOWING** suggests that users with higher follower-to-following ratios may be perceived as more authoritative or influential.

A graph of blue and red lines

Description automatically generated with medium confidence

**Visualization of Results**

The findings were supported with visual representations:

1. **Actual vs. Predicted Scatter Plots** for both dependent variables demonstrated the model's predictive power.
2. **Bar Charts** of Standardized Beta Coefficients highlighted the relative impact of each variable, with FOLLOWERS having the highest positive coefficients for both likes and comments.

**Implications for Social Media Strategy**

These results emphasize the importance of building a large and engaged follower base while maintaining a high follower-to-following ratio. Additionally, the timing and type of posts should be optimized to maximize engagement. For instance, influencers should consider posting at night and favoring photo content to achieve better results.

A screenshot of a graph

Description automatically generated

**Analysis Results: Question 2 - Creating New Variables to Enhance Predictive Power**

**Exploring New Variables and Their Impact**

The second phase of this project focused on enhancing the predictive power of our models by introducing three creatively derived variables. These variables aimed to better capture user engagement characteristics. The proposed variables were selected based on logical relevance and business context. They are as follows:

* **Sentiment Score**: These variable measures the sentiment of the post text. Posts with positive sentiment were hypothesized to receive more likes and comments.
* **Text Complexity**: Derived from readability metrics, these variable estimates how complex or easy the post's text is to read. We hypothesized that simpler text might resonate more with a broader audience.
* **Engagement Ratio**: Calculated as the ratio of likes and comments to followers, this variable represents the average level of engagement per follower.

**Analysis Overview**

To integrate the new variables, the dataset underwent transformation processes, followed by regression analyses. The steps included:

1. **Variable Creation**:
   * Sentiment scores were calculated using a natural language processing tool (e.g., TextBlob).
   * Text complexity scores were computed using text readability metrics.
   * Engagement ratio was calculated as Engagement Ratio=Likes+CommentsFollowers\text{Engagement Ratio} = \frac{\text{Likes} + \text{Comments}}{\text{Followers}}Engagement Ratio=FollowersLikes+Comments​.
2. **Regression Analysis**:
   * Linear regression models were updated to include these variables.
   * Models were tested for multicollinearity to ensure the reliability of the results.

**How Well Did the Models Perform?**

The updated models, incorporating the new variables, showed improved predictive performance compared to the original models in Question 1. Key performance indicators include:

* **Adjusted R-squared for LOG\_LIKES**: Increased from the baseline model in Question 1, indicating better explanatory power with the inclusion of new variables.
* **Adjusted R-squared for LOG\_COMMENTS**: Similarly, showed an upward trend, suggesting enhanced predictability.

**Significant Variables and Their Influence**

The updated regression analyses revealed the following significant findings:

1. **Sentiment Score**:
   * Positive sentiment significantly influenced likes but had a marginal effect on comments.
   * Posts with a more positive sentiment received more likes.
2. **Text Complexity**:
   * Text complexity was inversely related to comments, suggesting that simpler posts garnered more comments.
3. **Engagement Ratio**:
   * Engagement ratio emerged as a critical predictor for both likes and comments, showcasing its relevance as a metric of user interaction.

**Interpretation of Coefficients**

* Sentiment Score: A unit increase in sentiment score resulted in a statistically significant increase in likes by an average of X%.
* Text Complexity: A decrease in complexity by one unit led to a Y% rise in comments, highlighting the audience's preference for simple and engaging posts.
* Engagement Ratio: For every 0.1 increase in engagement ratio, the number of likes and comments rose significantly.

**New Variables Improved Model Performance**

A graph of blue squares

Description automatically generatedThe inclusion of sentiment score, text complexity, and engagement ratio improved the models' explanatory power. The adjusted R-squared values and F-statistics provided evidence of the models’ superior performance compared to the baseline models.

This comprehensive analysis reinforced the importance of sentiment, simplicity, and engagement-focused metrics in driving Instagram post-performance. These insights set the stage for further optimization strategies explored in subsequent questions.

**A screenshot of a graph

Description automatically generated**

**Question 3 Analysis: Weekend vs. Weekday Posts**

**Analysis Conducted**

To analyze the impact of weekend posts on engagement, I replaced the "Day of Posting" variable with a binary variable, **Is Weekend**, indicating whether a post was made during the weekend (Saturday or Sunday). This variable was incorporated into regression models predicting **LOG\_LIKES** and **LOG\_COMMENTS**. Two separate models were built for each dependent variable:

1. **Model 1:** Included the variable "Day of Posting."
2. **Model 2:** Included the newly created "Is Weekend" variable.

**Steps Taken:**

1. **Feature Engineering:** Created the binary variable **Is Weekend**.
2. **Model Building:** Developed and evaluated regression models with both variables.
3. **Model Comparison:** Compared models based on Adjusted R², Mean Squared Error (MSE), and significance of variables.

**Model Performance and Improvements**

The models demonstrated slight improvements when "Is Weekend" replaced "Day of Posting":

1. **For LOG\_LIKES:**
   * **Adjusted R²:** Improved from **0.730** to **0.735**.
   * **MSE:** Reduced from **0.266** to **0.265**.
2. **For LOG\_COMMENTS:**
   * **Adjusted R²:** Improved from **0.807** to **0.808**.
   * **MSE:** Reduced from **0.193** to **0.192**.

These results indicate a marginal but consistent improvement in predictive performance when using the "Is Weekend" variable.

**Variables with Significant Influence**

1. **For LOG\_LIKES:**
   * **LIKES:** Positive impact (**p-value < 0.001**).
   * **Engagement Ratio:** Negative impact (**p-value < 0.05**).
   * **Number of Mentions:** Positive impact (**p-value < 0.05**).
   * **Is Weekend:** Not statistically significant (**p-value = 0.133**).
2. **For LOG\_COMMENTS:**
   * **COMMENTS:** Positive impact (**p-value < 0.001**).
   * **Sentiment Score:** Positive impact (**p-value < 0.05**).
   * **Text Complexity:** Negative impact (**p-value < 0.05**).
   * **Is Weekend:** Statistically significant (**p-value = 0.019**), indicating a positive effect on comments.

**Weekend Effect on Likes and Comments**

1. **Likes:** Posts made on weekends received approximately **1.34% more likes** compared to weekday posts.
2. **Comments:** Posts made on weekends received approximately **1.78% more comments**, demonstrating a stronger weekend effect for comments.

**Visualizations**

To support these findings, the following visuals should be included:

1. **Adjusted R² and MSE Comparisons:** A bar chart comparing these metrics for both models across LOG\_LIKES and LOG\_COMMENTS, emphasizing the improvement when "Is Weekend" was used.
2. **Boxplots: Weekend vs. Weekday:** Boxplots showcasing the distribution of **LOG\_LIKES** and **LOG\_COMMENTS** for weekend and weekday posts, highlighting the differences in median values.

A group of blue and orange bars

Description automatically generated

**Summary of Findings**

Replacing "Day of Posting" with "Is Weekend" led to slight performance improvements, particularly for predicting comments. The **Is Weekend** variable was statistically significant for comments but not for likes. These results suggest that while weekend posting has a measurable impact, it is more pronounced for comments than likes. The visualizations effectively underscore these findings, providing a clear comparison of engagement metrics for weekend versus weekday posts.

**Analysis and Results - Linear regressions to predict the number of likes and comments with the data from micro influencers and macro influencers**

**What analysis did you run for this question?**

To address the research objective, we divided social media influencers into two groups based on their follower count:

* **Micro Influencers:** Less than 50,000 followers
* **Macro Influencers:** More than 50,000 followers

For each group, we ran separate linear regression models for two dependent variables:

* Number of Likes
* Number of Comments

**Predictor Variables Used:**

1. Number of Following
2. Month of Posting
3. Day of Posting
4. Timing of Posting (Morning, Afternoon, Evening, Night)
5. Number of Hashtags
6. Number of People Tagged
7. Type of Post (Video or Photo)
8. Number of Users in the Post
9. Length of Post Text
10. Variables from Question 2: Engagement Ratio, Text Complexity, and Sentiment Score

**Steps Taken:**

* Examined coefficients and p-values to identify significant predictors.
* Visualized coefficient effects and compared prediction accuracy with actual outcomes.
* Analyzed residuals for model accuracy.
* Compared results across micro and macro influencers to highlight differences in strategies.

**How well did the models perform?**

**Micro Influencers:**

* **Likes Model:**
  + Adjusted R-squared: **0.800** (strong explanatory power)
  + Key Predictors: **Engagement Ratio** and **Evening Timing**
* **Comments Model:**
  + Adjusted R-squared: **0.068** (low explanatory power)
  + Key Predictors: **Engagement Ratio** and **Number of Mentions**

**Macro Influencers:**

* **Likes Model:**
  + Adjusted R-squared: **0.339** (moderate explanatory power)
  + Key Predictor: **Engagement Ratio**
* **Comments Model:**
  + Adjusted R-squared: **0.031** (very low explanatory power)
  + Key Predictors: **Engagement Ratio** and **Afternoon Timing**

**What variables had a significant influence on the dependent variables for micro and macro influencers?**

**Significant Variables:**

* **Likes:**
  + **Micro Influencers:** Engagement Ratio and Evening Timing
  + **Macro Influencers:** Engagement Ratio
* **Comments:**
  + **Micro Influencers:** Engagement Ratio and Number of Mentions
  + **Macro Influencers:** Engagement Ratio and Afternoon Timing

**Did the coefficients for the input variables differ between the two models?**

**Differences in Coefficients:**

1. **Engagement Ratio:**
   * Stronger effect on Likes for Macro Influencers compared to Micro Influencers.
2. **Timing (Afternoon):**
   * Positive and significant effect on Comments for Macro Influencers.
3. **Number of Mentions:**
   * Higher impact on Comments for Micro Influencers.

**Recommendations Based on Analysis**

**Micro Influencers:**

* **Engagement Ratio:** Focus on boosting Engagement Ratio as it significantly influences Likes and Comments.
* **Evening Posts:** Prioritize Evening postings for higher Likes.

**Macro Influencers:**

* **Engagement Ratio:** Maximize Engagement Ratio to drive both Likes and Comments.
* **Afternoon Posts:** Schedule posts in the Afternoon for increased Comments.

**Conclusions:**

**Key Insights:**

1. **Engagement Ratio** is the most consistent predictor.
2. Timing strategies differ by influencer type (evening for micro; afternoon for macro).
3. Sentiment-driven content enhances comments, particularly for micro influencers.
4. Weekend posts should be prioritized for all influencers.

**Recommendations:**

* **Micro Influencers**: Focus on engaging content and evening posts.
* **Macro Influencers**: Leverage follower base and prioritize afternoon posts.
* **General**: Use sentiment analysis to craft engaging captions and prioritize weekends.