Data-Driven Insights: Analyzing Social Media Engagement **Report**

*University of Hartford, Barney School of Business*

*By Yamuna Devi Janapamala & Sameera Saumya Chittajallu*

*Business Analytics Capstone*

*Instructor: Dr. Xinying Qu*

TABLE OF CONTENTS:

[1. INTRODUCTION 3](#_Toc197432032)

[1.1 BUSINESS PROBLEM 3](#_Toc197432033)

[1.2 OBJECTIVES OF THE STUDY: 4](#_Toc197432034)

[2. BENCHMARK SOLUTIONS 5](#_Toc197432035)

[2.1 PROBLEM SOLUTION 7](#_Toc197432036)

[2.2 CHOSEN METHODOLOGY 7](#_Toc197432037)

[3. Data Collection and Preparation 8](#_Toc197432038)

[3.1 DATA SOURCES 8](#_Toc197432039)

[3.2 DATA PREPROCESSING AND DATA CLEANING 9](#_Toc197432040)

[3.2.1 OUTLIER DETECTION AND TREATMENT 9](#_Toc197432041)

[3.2.2 DATA TRANSFORMATION AND NORMALIZATION 10](#_Toc197432042)

[3.2.3 DATA SPLITING 11](#_Toc197432043)

[4. MODELING APPROACH 11](#_Toc197432044)

[4.1 MODEL DEVELOPMENT 11](#_Toc197432045)

[4.2 MODEL VALIDATION AND TESTING: 13](#_Toc197432046)

[5.RESULTS 14](#_Toc197432047)

[6.CONCLUSION 15](#_Toc197432048)

[7.FIGURES 16](#_Toc197432049)

[REFERENCES 17](#_Toc197432050)

# INTRODUCTION

Social media engagement refers to the interactions, responses, and connections users have with content across various social media platforms. It encompasses a range of behaviors that reflect the level of audience engagement and interest in a brand or its content, extending beyond merely counting likes and shares on posts. For businesses, this engagement is a crucial metric, as it indicates the effectiveness of their social media efforts in cultivating meaningful connections with their audience. It goes beyond simply amassing followers; it is about nurturing interactions that foster relationships and drive business results.

This project report investigates the effectiveness of social media engagement strategies by analyzing key metrics such as engagement rate, reach, and audience sentiment. Our analysis focuses on [specific platforms], examining [engagement metrics] over [time]. By utilizing data from [specific analytics platforms], we will evaluate how various engagement tactics influence Likes, Shares, Comments, and overall audience sentiment.

# 1.1 BUSINESS PROBLEM

In today's digital landscape, social media has become a vital tool for fostering customer connections, enhancing brand awareness, and achieving marketing success. However, businesses often face challenges in identifying which platforms—such as Facebook, Instagram, LinkedIn, or Twitter—and what types of content—whether text posts, videos, or images—yield the highest levels of engagement. Additionally, predicting advertising costs and managing ad budgets and predicting Return on Investment can be quite complex.

A low or inconsistent engagement rate frequently suggests a disconnect between audience interests and the content strategy. Without a thorough analysis of which platforms and content styles resonate most with their audience, companies risk continuing to share material that fails to engage. Conversely, by recognizing trends in engagement data, businesses can refine their content strategies, allocate resources more effectively, and create posts that drive higher interactions. Therefore, addressing this issue involves more than just improving metrics; it also means establishing a sustainable competitive edge in an increasingly crowded digital marketplace.

# 1.2 OBJECTIVES OF THE STUDY:

The primary objective of this study is to construct a robust and interpretable predictive model to estimate social media **engagement rates** based on a comprehensive set of platform-level, content-specific, and advertisement-related features. This research seeks to address the business challenge of optimizing digital engagement by identifying the key drivers that influence user interactions across different social media platforms.

To this end, the study involves the systematic analysis of a multi-platform dataset comprising user engagement metrics such as likes, comments, shares, impressions, and reach. Advanced data preprocessing and feature engineering techniques including the derivation of financial indicators such as *Ad Budget*, *Cost per Ad*, and *Return on Investment (ROI)* were employed to enhance model performance and interpretability. Multiple machine learning algorithms, including **Random Forest**, **XG Boost**, and **Linear Regression**, were developed and compared based on standard evaluation metrics.

The overarching goal is to generate actionable insights that enable businesses to make data-informed decisions regarding platform selection, content strategy, and advertising investments, thereby maximizing user engagement and return on marketing spend.

# 2. BENCHMARK SOLUTIONS

1. **Prediction of Customer Engagement Behavior Response on social media to Marketing Posts Based on Machine Learning (2020)**[**2**](#_REFERENCES)**:**  
   • **Researchers:** Junhao Liu, Po-Hsuan Hsieh, and Hui-Ju Wu  
   **• Published in**: Journal of Business Research  
    **Summary**: This study explores the use of machine learning algorithms to predict how customers engage with marketing posts on social media. By analyzing various features like post content, timing, and customer behavior history, the model aims to predict the likelihood of different types of engagement (like, comment, share).  
   **Key Contribution:**

* Applied several ML models (Logistic Regression, Random Forest, XGBoost, SVM) to predict engagement behavior.
* Demonstrated that content-related features and customer interaction history significantly influence engagement prediction.
* Highlighted the effectiveness of using machine learning over traditional statistical models.  
  **Metrics Analyzed**: Precision, Recall, F1-score, Accuracy, ROC-AUC for different types of engagement (likes, comments, shares).

1. **Machine Learning for Predictive Analytics in Social Media Data (2023)**[**3**](#_REFERENCES)**:**

* Author: Girish Shen
* Platform: Kaggle

**Summary:** This project looks at how post type, platform, and post timing affect engagement on social media.

**Key Findings**:

* Images and videos get the most likes and shares.
* Evening posts have better engagement.
* Instagram is good for visual content; Facebook gets more comments.
* Young users (18–24) are the most active.

**Metrics Used**: Likes, Shares, Comments, Engagement Rate, F1-score, Accuracy, ROC-AUC, Cross Validation.

**Use as Benchmark:**

* Compare performance of posts on different platforms.
* Predict which posts will do well.
* Find the best time and type of content to post.

# 2.1 PROBLEM SOLUTION

The goal of many current solutions is to assist companies in comprehending and maximizing social media participation. Platform-native tools that track post-performance include Facebook Insights, Instagram Analytics, and Twitter Analytics. These tools include basic analytics like likes, comments, and shares. Furthermore, to improve comparison and scheduling insights, third-party solutions such as Hootsuite, Sprout Social, and Buffer provide more thorough dashboards that incorporate data from many platforms.  
Technology-wise, machine learning-based methods have been created to forecast user engagement based on factors including post content, publishing time, and user behaviour. These include methods for classifying images, sentiment analysis, and natural language processing.

# 2.2 CHOSEN METHODOLOGY

The chosen methodology combines **exploratory data analysis**, **feature engineering**, and **machine learning modeling** to address the engagement prediction problem. This research uses a data-driven strategy that combines statistical analysis and machine learning modelling to fill in the gaps in current solutions. The following justifications support the methodology:

Cross-Platform Comparison: By breaking down the siloed structure of platform-native technologies, analysing data from several platforms (such as Facebook, Instagram, and Twitter) allows for a more comprehensive understanding of where interaction is highest.  
  
Content-Type Evaluation: We may assess how various content formats affect engagement metrics like likes, comments, and shares by classifying postings (videos, photos, links, etc.).

Predictive Modelling: Using structured features like post type, platform, posting time, and Engagement metrics, supervised machine learning models (such **as Linear Regression, Random Forest, XG Boost and Polynomial Ridged Regression**) are used to predict engagement rate.  
This methodology provides a solid answer for enhancing social media engagement through data insights, successfully bridging the gap between academic rigor and real-world applicability.

# 3. Data Collection and Preparation

# 3.1 DATA SOURCES

The data sets were sourced from Kaggle.com. The Social Media Engagement Report comprises three distinct data files[1](#_REFERENCES):

1) Countries or Areas: This file contains a total of 270 rows and 7 columns.

2) Working File: This file comprises 100000 rows and 24 columns.

3) Social Media Engagement Data: This file contains 100000 rows and 18 columns

The primary variable for this analysis is the Engagement Rate, which is calculated using the formula: Total likes, comments, and shares divided by Reach.

# 3.2 DATA PREPROCESSING AND DATA CLEANING

There are 79868 instances of missing values in the Campaign ID and 90006 in the Influencer ID. These values have been excluded from the analysis, as they do not provide any meaningful insights and are merely identifiers. To enrich the dataset with important financial metrics for marketing analysis, we have introduced three new columns: Ad Budget, Cost per Ad, and ROI (Return on Investment). The Ad Budget column reflects the financial allocation for each social media post, ranging from $100 to $1,000 based on synthetic campaign data. Cost per Ad is calculated by dividing the Ad Budget by the number of impressions, yielding a measure of advertising cost per view.

The ROI is calculated using the formula:

ROI = (Engagement Rate \* Reach) / Ad Budget.

This assessment evaluates the effectiveness of each post in generating engagement relative to its cost, which is essential for analyzing advertising strategies across various platforms.

# 3.2.1 OUTLIER DETECTION AND TREATMENT

Outlier detection was conducted using boxplot visualizations for several numerical features. These boxplots allowed for the visual identification of extreme values that lie beyond 1.5 times the interquartile range (IQR). While some outliers were present in variables such as Engagement Rate and Cost per Ad, the ROI variable exhibited a heavy-tailed distribution with a significant number of outliers, as shown in the before-and-after comparison ([see Figure 1](#_7.FIGURES)). After thorough examination, these outliers were retained in the dataset. The decision to keep them was based on their representation of real-world variability, as they highlight both high-performing and under-performing posts, which are crucial for accurately modeling ROI and engagement. Consequently, no trimming or adjustment was applied, ensuring the preservation of the natural distribution and signal within the data.

# 3.2.2 DATA TRANSFORMATION AND NORMALIZATION

Standard normalization or Min-Max scaling was deliberately avoided for features such as Ad Budget and Cost per Ad, as it distorted their interpretability and financial significance. However, the **log-transformed** is applied to numerical features to address its highly skewed distribution and reduce the impact of extreme values. Log transformation is crucial in stabilizing variance and improving model performance, particularly for regression tasks, by compressing the scale of large values and emphasizing relative differences. [Figure 2](#_7.FIGURES) shows the distribution of key numerical features after the log1p transformation (i.e., log (1 + x)) on positively skewed variables with extreme outliers.

Engagement Rate and ROI were normalized through the log transformation, yielding bell-shaped distributions suitable for regression modelling.

Cost per Ad remains right skewed but now has a more condensed and interpretable range.

# 3.2.3 DATA SPLITING

To ensure robust model evaluation and avoid overfitting, the dataset was partitioned into three subsets using a **70-10-20 split**:

* **70% Training Set**: Used for model training and hyperparameter tuning.
* **10% Validation Set**: Used to evaluate model performance during tuning and avoid overfitting.
* **20% Testing Set**: Held out to assess final model performance on unseen data.

This three-way split enables unbiased performance assessment and supports iterative model refinement, ensuring that results generalize beyond the training sample.

# 4. MODELING APPROACH

# 4.1 MODEL DEVELOPMENT

This project involved the implementation of multiple regression algorithms through Python's scikit-learn and machine learning libraries to develop robust predictive models for estimating social media engagement rates. The development process adhered to a structured pipeline, which included training and validation. To improve model accuracy and mitigate skewness, the target variable, Engagement Rate, a numerical variable, was log-transformed.

**1.** **Linear Regression:**

This algorithm served as the baseline model due to its simplicity and interpretability. It operates under the assumption of a linear relationship between the independent variables and the target variable, thereby allowing for a benchmark against which improvements from more complex models can be evaluated.

**2.** **Polynomial Ridge Regression:**

This model introduced polynomial features to capture non-linear relationships while simultaneously applying L2 regularization (ridge) to prevent overfitting. Polynomial features up to degree 2 were employed based on visual inspections and error metrics.

**3. Random Forest Regression:**

As a tree-based ensemble method, Random Forest constructs multiple decision trees and aggregates their outputs, demonstrating robustness to outliers and multicollinearity. The model was fine-tuned using the following hyperparameters:

- n\_estimators: The number of trees in the forest (tuned within a range of 100 to 500)

- max\_depth: The maximum depth of each tree (tuned between 5 and 20)

- min\_samples\_split: The minimum number of samples required to initiate a node split

**4. XGBoost Regression:**

This advanced gradient boosting algorithm is recognized for its high accuracy and performance. The following tuning parameters were evaluated:

- n\_estimators: The number of boosting rounds

- learning\_rate: The step size control during optimization (tested within the range of 0.01 to 0.2)

- max\_depth: The maximum tree depth (tuned from 5 to 10)

- subsample: The proportion of observations utilized for each tree (ranging from 0.8 to 1.0)

# 5.RESULTS

# 5.1 MODEL PERFORMANCE COMPARISON

The performance of each regression model was evaluated using standard metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R² Score calculated on the test dataset. The results are summarized below also shown in [fig3](#_7.FIGURES):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MODEL | MAE | |  | | --- | |  |   RMSE | R2 |
| Linear Regression | 14.36 | 20.61 | 0.70 |
| Polynomial Ridge | 2.94 | 6.09 | 1.00 |
| Random Forest | 1.02 | 1.84 | 1.00 |
| XG Boost | 1.39 | 2.07 | 1.00 |

Benchmark Solution: Random Forest demonstrated the best performance with minimal errors and a perfect R², effectively reducing the risk of overfitting while allowing for hyperparameter adjustments.

# 5.2 CROSS VALIDATION

To evaluate the robustness and generalizability of the models, we performed 5-fold cross-validation on both the training and test datasets:

Cross-Validation on Training Set:

Model CV RMSE ± Std

Random Forest 0.03 ± 0.00

XG Boost 0.04 ± 0.00

Cross-Validation on Test Set:

Metric Score

Test CV RMSE 1.80 ± 0.07

Test CV MAE 1.01 ± 0.02

Test CV R² 1.00 ± 0.00

These scores suggest that **Random Forest slightly outperformed XG Boost** in predictive accuracy on the training data, with less variability across folds.

After validation, the models were retrained on the combined training and validation sets. Final evaluation was done using the holdout test set, using the original scale (after inverse log transformation) with the following metrics:

* Mean Absolute Error (MAE)
* Root Mean Squared Error (RMSE)
* R² Score

These evaluation results confirmed that the Random Forest model consistently demonstrated high performance, making it the most suitable candidate for real-world deployment in predicting social media engagement based on platform, content, and ad metrics.

# 6.CONCLUSION

.

# 7.FIGURES

Fig 1 Outlier Boxplot Comparison

A screenshot of a computer program

AI-generated content may be incorrect.

Fig 2 EAD of Numeric Values

A group of blue lines

AI-generated content may be incorrect.

Fig 3 Model Performance Comparison

A graph of a bar chart

AI-generated content may be incorrect.

# REFERENCES

1. Data File: <https://www.kaggle.com/datasets/aliredaelblgihy/social-media-engagement-re>
2. Prediction of customer engagement behavior response to marketing posts based on machine learning-<https://www.tandfonline.com/doi/epdf/10.1080/09540091.2021.1912710?needAccess=true>
3. Machine Learning for Predictive Analytics in Social Media Engagement Analysis-99per-

<https://www.kaggle.com/code/girishshen/social-media-engagement-analysis-99per/notebook>

1. <https://www.kaggle.com/datasets/aliredaelblgihy/social-media-engagement-report?select=Social+Media+Engagement+Report.pbix>
2. <https://www.kaggle.com/code/nigarali/social-media-engagement>