Key Factors Influencing Social Media Engagement Rates:

A Predictive Modeling Approach

Mariana Shannan July 22, 2025

**Abstract**

This study investigates the key factors that influence social media post engagement rates across four major platforms: Instagram, Face- book, Twitter, and LinkedIn. Leveraging a dataset of 100,000 posts, we developed a predictive modeling pipeline integrating post meta- data, audience demographics, temporal variables, and sentiment indi- cators. Both regression and classification approaches were employed to explore and forecast engagement. Regression models, including Gra- dient Boosting and Support Vector Regression (SVR), yielded weak predictive power (best *R*2 = 0*.*006), indicating that metadata alone is insufficient for modeling engagement as a continuous outcome. In con- trast, classification models, particularly the Random Forest classifier, outperformed logistic regression in terms of accuracy, F1 score, and re- call, especially in identifying low-engagement posts. Recursive Feature Elimination with Cross-Validation (RFECV) revealed five key predic- tors: *Impressions*, *Post Type*, *Sentiment*, *Audience Age*, and *Time Period*. These results underscore the critical role of visibility, con- tent format, emotional tone, and timing in shaping user engagement. Our findings highlight the limitations of metadata-only models and emphasize the need to incorporate textual and visual content through Natural Language Processing (NLP) and Computer Vision techniques. This work provides actionable insights for optimizing content strategy and enhancing engagement across diverse social media platforms.

# Introduction

In today’s digital landscape, social media platforms such as Instagram, Twit- ter, Facebook, and LinkedIn are not merely tools for communication—they have evolved into influential arenas for shaping public discourse, driving con- sumer behavior, and disseminating content at scale. These platforms play a pivotal role in how information spreads, how opinions form, and how brands engage with audiences.

Understanding what drives a post’s success on these platforms is of both academic and practical interest. Central to this question is the concept of engagement, commonly quantified through likes, comments, shares, and in- teraction rates. Among these, the engagement rate is widely accepted as a key metric of post effectiveness. Yet, the determinants of engagement remain complex and platform-specific.

At the heart of this question lies the concept of *engagement*, typically measured by metrics such as likes, comments, shares, and overall interaction rate. Among these, the *engagement rate* is widely used as a standard indi- cator of a post’s effectiveness. Yet, despite its importance, the factors that drive engagement remain complex and platform-dependent.

### This study addresses the following research question:

*What are the key factors that influence the engagement rate of social media posts across different platforms?*

Previous research has often been limited in scope, focusing on a single platform, relying heavily on textual features, or using simplistic linear mod- els. Few studies have explored cross-platform comparisons or combined both regression and classification frameworks on large datasets.

To address these gaps, we construct a predictive modeling pipeline us- ing a dataset of 100,000 social media posts, incorporating features such as post metadata, temporal patterns, audience demographics, and sentiment indicators. We evaluate both regression and classification models, including Random Forest, Gradient Boosting, Support Vector Regression (SVR), Lo- gistic Regression, and others. Feature selection is performed using Recursive Feature Elimination (RFECV) to identify the most informative variables.

This study contributes not only a comparative analysis of modeling strate- gies but also practical insights into the timing, structure, and audience tar- geting strategies that enhance post engagement. The methodology, dataset, and evaluation results are presented in the sections that follow.

# Literature Review

A growing body of research has investigated the prediction of post popularity on social media platforms using machine learning techniques. These studies vary in methodology, features used, and the platforms analyzed, yet they share a common goal: to identify what makes a post successful in attracting user engagement.

Carta et al. [[1]](#_bookmark17) studied the popularity of Instagram posts using XGBoost and Random Forest classifiers. Their results showed that XGBoost achieved a balanced accuracy of 64.72% when using the latest 50 posts per user. Textual features were among the most predictive variables, and distributed training accelerated the model’s runtime. Yet, the study did not include visual content (images or videos), which plays a crucial role on visual-first platforms like Instagram. The present research addresses this gap by incorporating richer metadata and exploring platform diversity.

Chaurasia and Jha [[2]](#_bookmark18) employed linear regression to predict audience awareness (reach and impressions) on Instagram based on post metadata and engagement metrics. Although the model achieved reasonable results (R² = 0.68 for reach), it was limited by its linear assumptions and narrow dataset. Unlike their study, we apply a wider range of models, including tree-based and neural models to capture nonlinearity and interaction effects. Doe and Smith [[3]](#_bookmark19) explored Twitter engagement using multiple classi- fiers, including logistic regression, decision trees, KNN, and Random Forest. They reported that sentiment and posting time were critical features, with Gradient Boosting and Random Forest achieving the best results (F1 = 0.78, AUC = 0.82). Importantly, adding user-related features (e.g., follower count) improved model performance, highlighting the need to consider audience con-

text in engagement prediction.

Kolhe [[4]](#_bookmark20) examined Instagram user engagement using linear models, Ran- dom Forest regressors, and neural networks. The study found that while neural networks captured variability, they were prone to overfitting. Lasso regression offered a more interpretable and reliable solution. However, no hybrid models or content-aware strategies were tested. Our project extends this line of inquiry by systematically comparing regression and classification approaches, and by using recursive feature elimination to identify the most relevant features.

Finally, Obadi´c et al. [[5]](#_bookmark21) introduced the NLPOP dataset to predict the popularity of NLP research shared on Twitter. Their work applied various models, including BERT (RoBERTa-large), GRU-based RNNs, and IDF- weighted word embeddings, finding that BERT-based classifiers significantly outperformed other models. Notably, thread-based posts and those rich in

BIO and PAPER metadata yielded higher prediction accuracy. However, their dataset was small (2,292 samples) and focused exclusively on Twit- ter, limiting generalizability. Moreover, not all tweets were authored by the researchers themselves, which reduced the relevance of engagement as a pop- ularity proxy.

While the reviewed studies each contribute valuable insights into predict- ing social media engagement, they vary considerably in scope, methodological rigor, and feature representation. Most notably, previous research has often focused on a single platform (e.g., Twitter or Instagram), limited feature sets (e.g., only text or basic metadata), or a narrow range of modeling techniques (e.g., only classification or linear regression). For instance, while Carta et al. [[1]](#_bookmark17) and Chaurasia and Jha [[2]](#_bookmark18) utilized textual and metadata features with tree-based and linear models, respectively, they did not explore nonlinearity in depth or integrate cross-platform perspectives. Similarly, while Obadi´c et al. [[5]](#_bookmark21) leveraged advanced language models, their work was constrained by dataset size and domain specificity.

By contrast, the current study aims to bridge these gaps by offering a uni- fied, multi-platform analysis that combines both regression and classification approaches on a large, balanced dataset. This dataset incorporates struc- tured metadata, sentiment analysis, temporal features, and demographic in- dicators, enabling a more comprehensive and generalizable exploration of engagement prediction. Furthermore, through the use of recursive feature elimination and cross-model comparison, our approach provides both inter- pretability and robustness, addressing methodological limitations seen in ear- lier works. Thus, this research addresses those challenges by applying both regression and classification frameworks to a large, balanced dataset that spans multiple platforms and includes diverse features such as sentiment, timing, demographics, and engagement metadata.

# Dataset

The dataset utilized in this study comprises approximately 100,000 social media posts collected between 2021 and 2024 from four major platforms: Instagram, Facebook, Twitter, and LinkedIn. Each post is accompanied by metadata describing its characteristics (e.g., content type, time of pub- lication), audience-related attributes (e.g., gender, age group, continent), sentiment classification, and platform-level engagement metrics such as im- pressions and reach.

## Structure and Features

The dataset consists of 24 variables spanning numerical, categorical, tempo- ral, and sentiment-based attributes. The target variable is the Engagement Rate, expressed as a percentage. It is calculated as the ratio of user interac- tions (likes, comments, shares) to the total reach of each post. For regression tasks, the engagement rate is treated as a continuous variable. For classifi- cation tasks, it was discretized into three ordinal categories: Low, Medium, and High, using quantile-based binning to ensure balanced class distribution.

The independent variables include:

**Platform** (Instagram, Facebook, Twitter, LinkedIn)

**Post Type** (Image, Video, Link, Text)

**Temporal Features** (Post Hour, Day of Week, Month, Weekday/Weekend, Time Period)

**Audience Demographics** (Gender, Age Group, Age, Continent)

**Sentiment** (Positive, Negative, Neutral, Mixed)

Unique identifiers such as Post ID, Campaign ID, and Influencer ID were excluded from modeling due to their lack of predictive value and high levels of missingness (over 80%).

A full list of variables, including data types, allowed values, and modeling roles, is provided in Appendix [9.](#_bookmark22)

## Data Quality and Missing Values

A comprehensive missing value analysis revealed that 22 out of the 24 vari- ables were fully complete. The remaining two Influencer ID and Campaign ID exhibited more than 90% and 80% missing values, respectively. Due to their high degree of incompleteness and their function as unique identifiers rather than informative features, both variables were excluded from further analysis.

## Descriptive Statistics and Interpretation

Descriptive statistics provide valuable insight into post performance and au- dience behavior across social media platforms. Table [1](#_bookmark0) summarizes key nu- meric variables, including engagement metrics (likes, comments, shares), vis- ibility indicators (impressions, reach), the target variable (engagement rate), and a demographic attribute (audience age). All metrics are based on 100,000 social media posts.

Table 1: Descriptive statistics for key numeric variables (N = 100,000)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Median** | **Min** | **25%** | **75%** | **Max** |
| Likes | 499.40 | 500.00 | 0 | 249.00 | 750.00 | 1,000.00 |
| Comments | 249.70 | 250.00 | 0 | 124.00 | 375.00 | 500.00 |
| Shares | 100.12 | 100.00 | 0 | 50.00 | 150.00 | 200.00 |
| Impressions | 5,487.63 | 5,477.00 | 1,000 | 3,239.00 | 7,733.00 | 10,000.00 |
| Reach | 2,751.52 | 2,754.00 | 500 | 1,627.00 | 3,877.25 | 5,000.00 |
| Engagement Rate (%) | 43.41 | 30.77 | 0.49 | 20.03 | 52.37 | 312.55 |
| Audience Age | 41.51 | 42.00 | 18 | 30.00 | 54.00 | 65.00 |

The data exhibit moderate central tendencies but considerable variabil- ity. For example, while the average number of likes is approximately 499, the distribution is wide, spanning from 0 to 1,000. Similarly, engagement rates range from 0.49% to 312.55%, with a highly skewed distribution and a median of 30.77%. This suggests that while most posts achieve moderate engagement, a minority experience disproportionately high success.

Impressions and reach are centered around 5,487 and 2,751, respectively, with maximum values that likely reflect platform-imposed limits on exposure. The average audience age of 42 years indicates a mature user base.

Figures [1](#_bookmark1) and [2](#_bookmark2) illustrate these trends. The first shows a right-skewed dis- tribution of engagement rate, with most values between 20% and 40%. This suggests that high engagement is rare and most posts perform in a moder- ate range. The second graph shows that impressions are also skewed to the right, peaking around 5,000. Only a small subset of posts gains widespread visibility, while most remain within a typical exposure range.

These patterns underscore the importance of predictive models that ac- count for skewed distributions and identify rare high-performing posts.

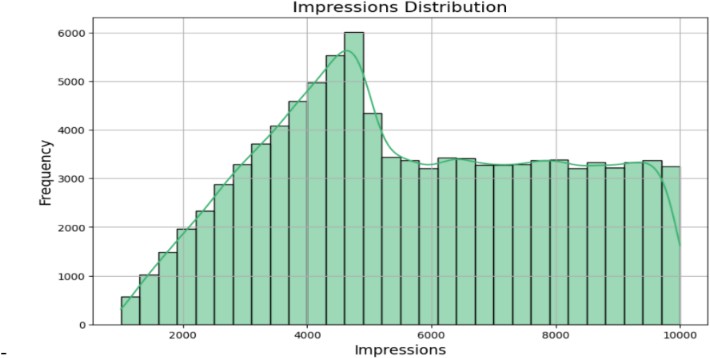
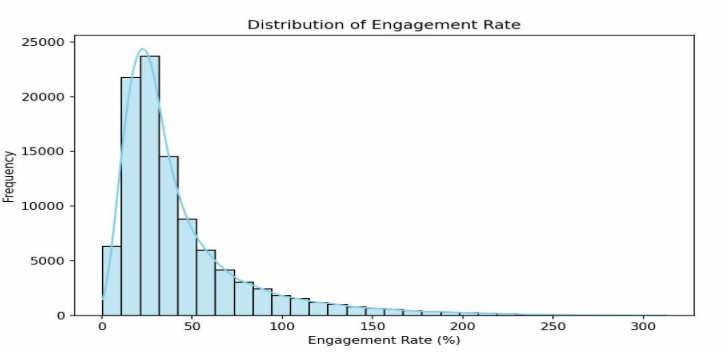


Figure 1: Distribution of Engage- ment Rate —most values fall be- tween 20% and 40%, with a long right tail.

Figure 2: Distribution of Impressions

— peak around 5,000 with few highly visible outliers.

Overall, these findings demonstrate that while most posts perform within typical ranges, a minority achieve exceptionally high engagement. This sup- ports the need for predictive models that can handle skewed distributions and capture outliers effectively.

## Categorical Distributions

The dataset includes several categorical variables capturing content type, platform usage, audience demographics, sentiment, and temporal aspects. These variables vary in cardinality and frequency distribution, as summarized in Table [2.](#_bookmark3)

Table 2: Frequency counts for selected categorical variables

|  |  |
| --- | --- |
| **Variable** | **Categories and Frequencies** |
| Post Type | Video: 33,384; Link: 33,338; Image: 33,278 |
| Platform | Twitter: 25,160; LinkedIn: 25,126; Facebook: 24,879; Insta-  gram: 24,835 |
| Audience Gender | Male: 33,476; Female: 33,385; Other: 33,139 |
| Sentiment | Mixed: 50,100; Positive: 16,738; Neutral: 16,645; Negative:  16,517 |
| Weekday Type | Weekday: 71,464; Weekend: 28,536 |
| Time Periods | Night: 33,199; Afternoon: 24,991; Morning: 24,955;  Evening: 16,855 |
| Age Group | Senior Adults: 41,717; Mature Adults: 31,251; Adolescent  Adults: 27,032 |
| Audience Continent | Africa: 24,416; Europe: 20,413; Asia: 20,368; North Amer-  ica: 15,885 |

Variables such as *Post Type*, *Platform*, and *Audience Gender* exhibit well-balanced distributions, supporting robust comparative analysis. For in- stance, the number of posts across the four platforms is nearly identical, ensuring no single platform dominates the dataset.

However, the *Sentiment* variable displays a pronounced class imbalance, with over 50% of posts labeled as “Mixed,” while the remaining classes: “Positive,” “Neutral,” and “Negative” each represent only 16.5% of the data. This imbalance should be considered during model training and evaluation, particularly in classification tasks involving sentiment prediction.

Several high-cardinality variables (e.g., *Audience Location*, *Audience In- terests*, *Post Content* ) exhibit long-tailed distributions. These are common in behavioral and demographic data and may require dimensionality reduction or embedding techniques in more advanced modeling.

Metadata fields such as *Post ID*, *Campaign ID*, and *Influencer ID* are excluded from modeling due to their identifier nature.

Figures [3](#_bookmark4) and [4](#_bookmark5) illustrate the distribution of posts across platforms and sentiment categories, respectively. The first graph confirms the balanced nature of the platform data, while the second reveals a significant skew, with over 50% of posts labeled as “Mixed.”

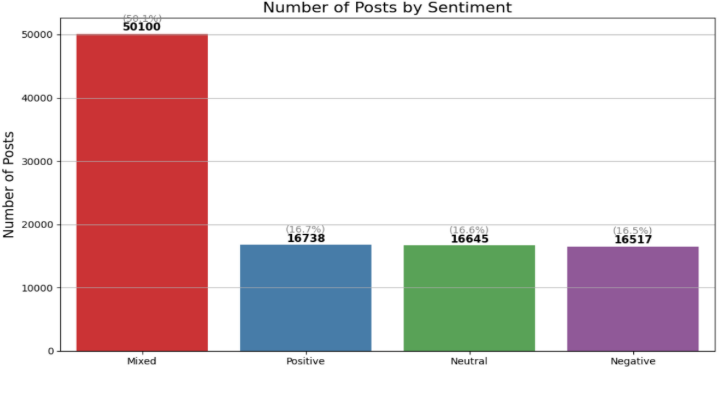
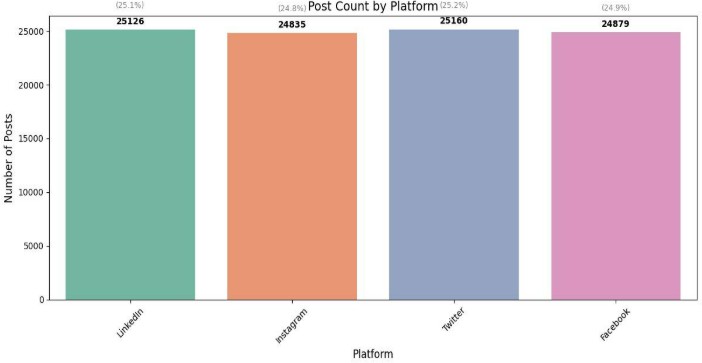


Figure 3: Post count by platform — balanced distribution across all four platforms

Figure 4: Post count by sentiment

— substantial imbalance with over- representation of “Mixed”

# Methodology

## Research Design

This study adopts a structured, data-driven, and quantitative research de- sign aimed at identifying the key post-related and audience-related features that influence engagement rates across multiple social media platforms. The central research question guiding this work is:

*Which post and audience attributes are most predictive of engage- ment rate, and how can machine learning models uncover these relationships?*

To address this question, two predictive modeling frameworks were imple- mented: a regression approach and a classification approach. In the regres- sion task, the objective was to model the *Engagement Rate* as a continuous target variable and assess how well it can be predicted from post metadata, audience demographics, and temporal attributes. In contrast, the classifica- tion task involved converting the engagement rate into an ordinal categorical variable (Low, Medium, High) and modeling the problem using multiclass classification algorithms.

By incorporating both perspectives, regression for precision and classi- fication for interpretability, the research design allows for a holistic under- standing of engagement behavior. This dual-approach supports both gran- ular analysis (e.g., estimating actual engagement percentages) and practi- cal decision-making (e.g., classifying posts into engagement tiers for content strategy). The overall modeling pipeline includes comprehensive data pre- processing, categorical encoding, feature engineering, recursive feature elimi- nation (RFECV), model training, and rigorous evaluation using appropriate performance metrics.

## Data Preprocessing

The dataset comprises approximately 100,000 p social media posts from four major platforms: Instagram, Facebook, Twitter, and LinkedIn. Each post is characterized by features such as platform, post format (e.g., image, video, link), timestamp of publication, audience demographics (age, gender, con- tinent), and sentiment classification. These features form the basis for pre- dicting the target variable: **Engagement Rate**, defined as the ratio of user interactions to reach, expressed as a percentage.

To support both regression and classification frameworks, two target vari- ables were created:

* For regression: the engagement rate was used in its raw, continuous form.
* For classification: the engagement rate was discretized into three or- dinal classes: Low, Medium, and High, using quantile-based binning (qcut) to ensure a balanced class distribution.

### Handling Missing Values

A detailed analysis of missing values revealed that 22 out of 24 variables were fully complete. The remaining two variablesInfluencer ID and Campaign ID, which had missingness rates of approximately 90% and 80%, respec- tively. As both fields functioned primarily as identifiers and lacked predic- tive relevance, they were excluded from all subsequent stages. This exclusion aligns with best practices in machine learning, as high-missingness or non- informative identifiers can introduce noise, lead to overfitting, or cause data leakage.

### Resolving Class Imbalance

**Regression Task:** In the regression task, the target variable *Engagement Rate* is a continuous numeric value representing the percentage of engagement on social media posts. Since regression models predict continuous outcomes rather than discrete categories, the concept of class imbalance is not applica- ble. Class imbalance refers to unequal representation of labels in classification tasks, which can bias learning. As no such labels are used in regression, no balancing or resampling techniques were required.

**Classification Task:** To convert the regression problem into a classifica- tion task, we transformed the continuous *Engagement Rate* variable into an ordinal categorical feature named *Engagement Class*, which represents three levels of engagement: **Low**, **Medium**, and **High**.

This transformation was performed using the qcut function from the pandas library with *q* = 3, which partitions the distribution into three quantile-based bins of approximately equal size. This method ensures a bal- anced class distribution that aligns with the underlying data characteristics.

Descriptive analysis revealed that the *Engagement Rate* variable follows a right-skewed distribution, with a median of 34.94% and a mean of 46.92%. In such cases, fixed-threshold binning (e.g., 0–30%, 30–60%, 60–100%) or median-based binary splits could result in imbalanced classes and reduced differentiation between engagement levels.

Quantile-based binning avoids these pitfalls by adapting to the empiri- cal distribution, resulting in three balanced categories. This also enhances model robustness and interpretability by preserving sample parity across en- gagement levels.

Table 3: Distribution of Engagement Classes after Quantile Binning

|  |  |  |
| --- | --- | --- |
| **Engagement Class** | **Count** | **Percentage** |
| Low | 33,333 | 33.33% |
| Medium | 33,334 | 33.33% |
| High | 33,333 | 33.33% |

As shown in Table [3,](#_bookmark6) the resulting distribution is nearly perfectly bal- anced. The one-sample difference in the *Medium* class is due to rounding in the quantile split and is statistically negligible. Therefore, no additional techniques such as oversampling, undersampling, or class weighting were nec- essary. Classification models were trained directly on this balanced target without any further adjustments.

### Preventing Data Leakage

To ensure the validity of model performance, variables used to compute the engagement rate, Likes, Comments, Shares, and Reach were excluded from model training. Including these would constitute target leakage, leading to artificially inflated performance metrics. By removing them, models were constrained to rely solely on metadata, demographics, sentiment, and tem- poral features.

### Encoding Categorical Variables

Categorical variables were transformed using encoding methods aligned with their semantic properties:

* + - * **One-Hot Encoding** was applied to nominal variables such as Plat- form, Post Type, Weekday Type, Time Periods, Audience Gender, Au- dience Continent, and Sentiment.
      * **Ordinal Encoding** was applied to Age Group, preserving the natu- ral order of life stages: Adolescent Adults *<* Mature Adults *<* Senior Adults.

### Temporal Feature Extraction

To capture time-sensitive behavioral patterns, timestamp fields were trans- formed into structured numeric features:

* + - * *Post Hour* (0–23) — hour of the day the post was published
      * *Day of Week* (0 = Monday to 6 = Sunday)
      * *Post Month* (1–12)

These engineered features enable the detection of temporal trends, such as optimal posting times or seasonal engagement fluctuations.

### Train-Test Splitting Strategy

For both regression and classification tasks, the dataset was split into training (80%) and testing (20%) subsets:

* + - * In the **regression task**, a random split was performed without strati- fication.
      * In the **classification task**, stratified sampling was used to maintain the equal distribution of engagement classes across both subsets.

To prevent data leakage, all data transformations (including scaling and encoding) were applied to the training set only, and then the learned trans- formations were applied to the test set.

### Numeric Variables Scaling

All numerical features were standardized using Z-score scaling (mean = 0, standard deviation = 1). This scaling technique was selected for its ability to preserve distributional characteristics and to ensure compatibility with ma- chine learning algorithms that are sensitive to scale. Notably, standardization was applied after filtering and encoding, in a strictly sequential pipeline to ensure data integrity.

## Feature Engineering and Selection

### Feature Engineering

No additional features were engineered beyond the preprocessing steps pre- viously described. The original dataset was inherently rich, encompassing a wide array of variables that captured platform-specific characteristics (e.g., type of social media platform and post format), audience demographics (e.g., age, gender, and continent), temporal aspects (e.g., post hour and day of the week), and sentiment classifications. These dimensions were deemed suffi- cient to comprehensively represent the problem space.

Categorical variables were encoded using appropriate techniques—either one-hot encoding or ordinal encoding—depending on the nature of the vari- able. Temporal features were extracted from timestamps to allow models to capture cyclical or seasonal posting patterns. Specifically, the variables *Post Hour*, *Post DayOfWeek*, and *Post Month* were derived to account for potential timing effects on engagement. As such, no handcrafted or domain- specific feature transformations were required. This minimalist engineering approach maintained a lean feature space while leveraging the informative richness of the original dataset.

### Feature Selection

To improve model interpretability, reduce overfitting risk, and optimize pre- dictive performance, a wrapper-based feature selection technique was em- ployed: Recursive Feature Elimination with Cross-Validation (RFECV). This

approach was applied independently to the regression and classification tasks, using task-specific estimators.

**Regression Task:** In the regression pipeline, RFECV was employed using a RandomForestRegressor as the base estimator. Random Forest was chosen for its ability to capture complex, non-linear relationships and for its robustness to multicollinearity and irrelevant variables. Its ensemble nature and built-in feature importance scores made it particularly effective in guiding the elimination of low-contributing features [[3,](#_bookmark19) [1]](#_bookmark17)

RFECV was configured with 5-fold cross-validation and used the *R*2 score as the evaluation metric. It iteratively eliminated the least important fea- tures, identifying an optimal subset of 24 features, including both original numeric variables and encoded categorical attributes. The most predictive features included:

* + - * Impressions
      * Age Group
      * Post Hour, Post DayOfWeek
      * Platform Instagram, Post Type Video
      * Audience Continent Europe
      * Several sentiment indicators

These selected features were then used to train a Support Vector Regres- sor (SVR). SVR was selected due to its theoretical foundation in margin- based optimization and its capacity to model non-linear relationships, even in high-dimensional spaces. However, despite careful tuning through grid search, the SVR model underperformed, achieving a training *R*2 of *−*0*.*08 and a test *R*2 of *−*0*.*11. This result suggests that the engagement rate may be influenced by latent or unobserved variables not present in the dataset.

**Classification Task:** In the classification pipeline, a similar RFECV approach was applied using a RandomForestClassifier as the estimator. The rationale behind this choice was consistent with the regression task: Random Forest’s intrinsic ability to assess feature importance and its strong performance on encoded, high-dimensional data made it ideal for guiding feature elimination in multiclass [settings.[3,](#_bookmark19) [1]](#_bookmark17)

This version of RFECV used stratified 5-fold cross-validation and accu- racy as the evaluation metric, ensuring class balance throughout the selection

process. The final model retained 26 features, drawn from diverse categories including post metadata, audience characteristics, temporal variables, and sentiment classifications. Key selected features included:

* + - * Impressions, Audience Age, Post Hour
      * Platform Instagram, Post Type Video, Time Periods Morning
      * One-hot encoded variables from Audience Continent and Sentiment These features contributed to improved performance and interpretabil-

ity in the subsequent classification models, particularly in multiclass settings where feature redundancy can obscure decision boundaries.

In summary, the feature engineering strategy employed in this study was deliberately minimal yet effective. Rather than introducing artificial vari- ables, the project capitalized on the existing richness of the dataset, enhanced through precise encoding and temporal decomposition. The use of RFECV for feature selection proved crucial in both pipelines, helping models focus on the most informative predictors. Although predictive performance was lim- ited in some scenarios, the selection process itself offered meaningful insights into the key drivers of engagement across social media platforms.

## Models

### Regression Task

The objective of the regression task was to predict the *Engagement Rate*, a continuous variable representing post performance, based on features such as platform, timing, audience demographics, and sentiment. Four diverse regression models were developed to capture various assumptions about the data:

* + - * Gradient Boosting Regressor (tree-based, ensemble)
      * Ridge Regression (linear, L2-regularized)
      * Lasso Regression (linear, L1-regularized)
      * Multi-Layer Perceptron (MLP) (nonlinear, neural network)

These models represent complementary algorithmic families and were se- lected to explore different assumptions about the structure of the data. All models were trained using standardized numeric features and evaluated using 5-fold cross-validation with the *R*2 score as the primary performance metric.

**Gradient Boosting Regressor (GBR):** Gradient Boosting Regression (GBR) is an ensemble learning method that builds additive predictive models by sequentially fitting decision trees to the residuals of prior models. This iterative refinement process allows the model to capture complex non-linear relationships and interactions among features, which is particularly valuable when modeling user engagement, a phenomenon influenced by multifaceted behavioral, temporal, and contextual v[ariables.[3]](#_bookmark19)

In our study, a pipeline was constructed incorporating a StandardScaler followed by a GradientBoostingRegressor. The model’s hyperparameters were optimized via GridSearchCV using five-fold cross-validation, evaluating combinations of n estimators, learning rate, max depth, and subsample. This rigorous tuning process aimed to strike a balance between bias and variance and to prevent overfitting.

Although the model achieved a relatively low *R*2 value on the test set (0.006), the feature importance analysis revealed substantive insights. No- tably, *Impressions* dominated the model’s predictions, followed by *Audience Age* and *Post Hour*, suggesting that visibility and temporal targeting play critical roles in determining post engagement. The explainability of GBR through its built-in feature importance metric further supports its relevance in this research context, even when predictive performance is modest.

## Feature Importance Analysis

To interpret the model and identify the most influential predictors, we ex- tracted the feature importances from the trained Gradient Boosting model. The following figure presents a horizontal bar chart ranking features by their relative importance in the prediction process.

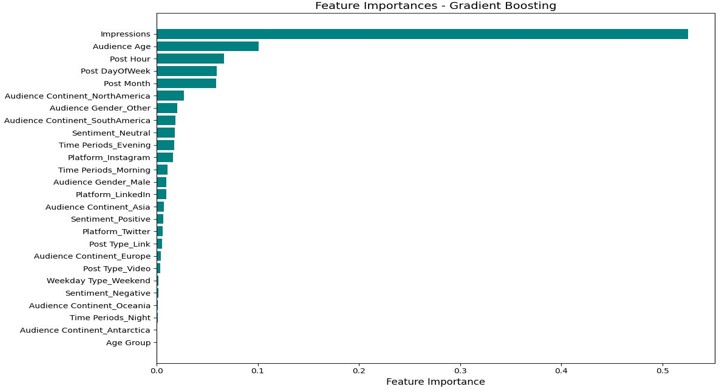


Figure 5: Feature Importance Ranking from Gradient Boosting Regressor As can be seen, Impressions overwhelmingly dominated the model, fol-

lowed by Audience Age, Post Hour, and temporal features like Day of Week and Post Month. These findings suggest that visibility and timing are core drivers of engagement.

**Ridge Regression:** Ridge Regression is a linear model with L2 regular- ization, designed to mitigate multicollinearity and reduce overfitting by pe- nalizing the magnitude of coefficients. This model is particularly suitable for datasets where predictors may be highly correlated, a common occurrence in social media analytics, given the presence of numerous temporally and demographically related variables.

In this study, Recursive Feature Elimination with Cross-Validation (RFECV) was applied using a Ridge estimator to systematically identify the most infor- mative features. Following this selection phase, hyperparameter optimization was conducted using GridSearchCV, focusing on tuning the regularization parameter *α*. The optimal configuration retained only a single feature *Im- pressions* with the best-performing model found at *α* = 10.

Despite its simplicity, the Ridge model’s performance on the test set was limited, achieving an *R*2 = *−*0*.*0039, indicating that the linear assumptions underlying the model may not sufficiently capture the variability in engage- ment rates. Nonetheless, Ridge Regression offers key advantages in terms of interpretability and robustness to noise, making it a valuable comparative baseline. Its capacity to emphasize dominant linear signals, such as the im- portance of post visibility via *Impressions*, reinforces its analytical utility,

especially in early-stage modeling or feature selection scenarios.

**Lasso Regression:** Lasso Regression (Least Absolute Shrinkage and Se- lection Operator) is a linear model that employs L1 regularization to perform both shrinkage and variable selection. By penalizing the absolute values of the coefficients, Lasso can drive some coefficients to exactly zero, effectively excluding noninformative features from the model. This property is partic- ularly advantageous in settings where the underlying relationship is sparse, or where only a subset of features is expected to have substantial predictive value.

In our implementation, RFECV was employed with a Lasso estimator to identify the most relevant predictors. The feature selection process ulti- mately converged on a single variable, *Impressions*. Hyperparameter tun- ing via GridSearchCV identified the optimal regularization parameter as *α* = 0*.*0001. Although the resulting model yielded a modest test set per- formance of *R*2 = *−*0*.*004, it aligned with patterns observed across other regression models, underscoring the centrality of *Impressions* in explaining engagement outcomes.

Lasso’s capacity to simplify models by eliminating redundant or irrelevant features enhances interpretability and model efficiency. While it lacks the flexibility of nonlinear models, its strengths in variable selection and sparsity promotion render it an effective tool for early-stage modeling and exploratory analysis.

**Multi-Layer Perceptron (MLP) Regressor:** The MLPRegressor is a feedforward artificial neural network that models complex nonlinear functions via multiple layers of interconnected neurons. Each neuron applies a learned weight and bias, followed by a nonlinear activation function. MLP is well- suited for high-dimensional datasets with intricate patterns and nonlinearity, characteristics often observed in user behavior and social media data.

In our implementation, the model was trained using a grid search over var- ious hyperparameters: hidden layer sizes, *α*, learning rate init, and activation. The best configuration consisted of a single hidden layer of 100 neurons, *α* = 0*.*01, learning rate init = 0.001, and the tanh acti- vation function. Despite this tuning effort, the test set performance was poor (*R*2 = *−*0*.*0175).

The suboptimal result may stem from the limited dataset size or insuffi- cient feature expressiveness after preprocessing.

## Comparative Summary of Regression Models

To evaluate different approaches for predicting engagement rate, we com- pared four regression models representing diverse algorithmic strategies: Gra- dient Boosting, Ridge Regression, Lasso Regression, and MLP Regressor. The models were assessed based on test set *R*2 scores, feature selection meth- ods, and the top contributing features.

Table 4: Performance Summary of Regression Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | *R*2 **(Test)** | **Feature Selection** | **Top Feature** | **Notes** |
| Gradient Boosting | 0.0060 | None | Impressions | Captured non-linear  patterns but showed limited generalization. |
| Ridge Regression | –0.0039 | RFECV | Impressions | Robust to multi-  collinearity; weakest performance. |
| Lasso Regression | –0.0041 | RFECV | Impressions | Sparse model with au-  tomatic feature selec- tion. |
| MLP Regressor | –0.0175 | None | All features | Underperformed;  likely due to overfit- ting or data limita-  tions. |

**Interpretation:** Across all models, *Impressions* consistently emerged as the most important predictor, indicating that content visibility plays a cen- tral role in driving user engagement. This finding was supported by feature importance in GBR and selected features in both Ridge and Lasso regres- sions.

**Implications:** Despite methodological diversity, all models showed limited predictive performance (*R*2 *<* 0*.*01), suggesting that structured features alone (e.g., platform, timing, demographics) may not sufficiently explain the vari- ability in engagement rate. These results highlight the need for integrating richer feature sets in future work, including textual, visual, or behavioral signals.

## Classification Task

In the classification task, the objective was to predict the *Engagement Class* of each social media post, categorized as **Low**, **Medium**, or **High**, based on a variety of content, temporal, and audience-related features. To address this, we evaluated two widely used classification algorithms: **Logistic Re- gression** and **Random Forest Classifier**. Both models were trained on the subset of features selected through RFECV, as described in the feature selection section. Hyperparameter tuning was performed using 5-fold cross- validation to ensure generalizability and robustness.

**Logistic Regression:** Logistic Regression is a foundational linear classifi- cation model that is well-suited for multi-class tasks due to its probabilistic interpretation and strong baseline performance. It is valued for its simplicity, interpretability, and resistance to overfitting, particularly when regularized. In this study, it served as a benchmark model against which more complex classifiers could be evaluated.

The model was implemented within a pipeline that included Standard Scaler for feature standardization, followed by Logistic Regression with

the multi class =’ovr’ (one-vs-rest) strategy and class weights set to balanced to account for potential class imbalance. Hyperparameter tuning was con- ducted via GridSearchCV, with the regularization strength parameter *C* be-

ing varied while keeping the penalty fixed to L2 (Ridge penalty).

**Random Forest Classifier:** Random Forest is an ensemble learning algo- rithm that constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness. It is particularly effective at captur- ing nonlinear relationships and variable interactions due to its hierarchical tree-based structure. Additionally, Random Forest models naturally per- form embedded feature selection and are resilient to noise and overfitting when appropriately tuned.

The Random Forest model was incorporated into a pipeline for consis- tency across experiments. Although tree-based methods are scale-invariant, a StandardScaler was applied to maintain comparability with other mod- els. The classification was conducted using RandomForest Classifier, and hyperparameter tuning was performed via GridSearchCV.

The tuning process explored a range of values for three key parameters:

* n estimators – the number of trees in the forest
* max depth – the maximum depth of each tree
* min samples split – the minimum number of samples required to split an internal node

This grid search was conducted using 5-fold cross-validation to identify the configuration yielding the highest validation accuracy while mitigating overfitting. The use of stratified folds ensured class balance across training and validation sets.

# Evaluation

## Regression Task

To evaluate the performance of the regression models predicting the *Engage- ment Rate*, four complementary evaluation metrics were selected:

* **Mean Absolute Error (MAE):** Provides a straightforward measure of average prediction error in absolute terms. It is particularly use- ful when treating all deviations equally, without prioritizing extreme outliers.
* **Mean Squared Error (MSE):** Penalizes larger errors more heavily by squaring the residuals. It is valuable in contexts where large pre- diction deviations (e.g., underestimating viral posts) carry significant implications.
* **Root Mean Squared Error (RMSE):** Restores the unit scale of the target variable (engagement rate percentage), enhancing interpretabil- ity while retaining the benefits of MSE.
* **R2 (Coefficient of Determination):** Indicates the proportion of variance in the engagement rate that the model can explain. A higher value reflects better model fit and stronger explanatory power.

Collectively, these metrics offer a robust and balanced assessment of re- gression model performance. While MAE focuses on general accuracy, MSE and RMSE emphasize the cost of large deviations, and R2 captures overall model explanatory power. This multifaceted approach is well-aligned with the research objective: to understand the variability in social media post engagement and to develop predictive tools that are both accurate and in- formative.

We evaluated five regression models: Random Forest (used as the base estimator in the SVR pipeline), Gradient Boosting, Ridge, Lasso, and MLP Regressor. Table [5](#_bookmark7) presents their performance across four evaluation met- rics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R2).

Table 5: Performance of Regression Models Across Evaluation Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | **MAE** | **MSE** | **RMSE** | **R2** |
| Random Forest | 24.7959 | 1588.6585 | 39.8580 | -0.1112 |
| Gradient Boosting | 25.9116 | 1421.0135 | 37.6963 | 0.0060 |
| Ridge | 26.1148 | 1435.2620 | 37.8849 | -0.0039 |
| Lasso | 26.1150 | 1435.4681 | 37.8876 | -0.0041 |
| MLP Regressor | 26.7889 | 1454.7158 | 38.1407 | -0.0175 |

**Table** [**5**](#_bookmark7)summarizes the comparative outcomes. The Gradient Boosting model achieved the best R2 score (0.0060) and the lowest MSE and RMSE values among all models, indicating a slightly stronger ability to capture un- derlying data patterns. However, the overall low R2 values and relatively high error metrics across all models suggest that the available feature set lacks sufficient predictive power. This emphasizes the potential influence of latent, unobserved factors on social media engagement.

To better understand the nature of the regression problem, we analyzed the distribution of the target variable, Engagement Rate, prior to model training. As illustrated in Figure [6,](#_bookmark8) the distribution is positively skewed, with a long right tail indicating the presence of high-engagement outliers. The majority of posts exhibit engagement rates between 10% and 60%, while a smaller portion reach extreme values exceeding 200%.

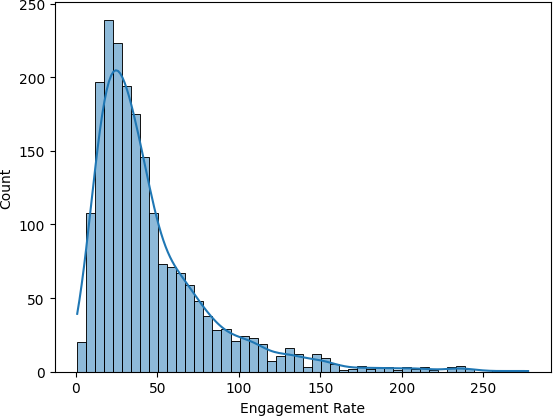


Figure 6: Histogram and density plot of the target variable (Engagement Rate).

This skewed distribution has several important implications:

* **Non-normality:** The target variable violates the normality assump- tion underlying linear models such as Ordinary Least Squares (OLS). This may lead to biased coefficient estimates and reduced predictive power.
* **Outlier sensitivity:** The long tail of rare high-performing posts can inflate error metrics like MSE and RMSE, potentially misrepresenting overall model performance.
* **Model selection:** The skewed distribution supports the use of robust metrics (e.g., MAE) and models that are resilient to non-Gaussian dis- tributions, such as tree-based methods.

In this study, no transformation (e.g., logarithmic) was applied to the tar- get variable to retain interpretability. However, future research may consider normalizing the target distribution to reduce variance and improve perfor- mance in linear models.

To gain deeper insights into model performance and its limitations, we conducted a residual and error analysis focused primarily on the best-performing model in the regression task—**the Gradient Boosting Regressor (GBR)**.

The goal of this analysis was to understand not only how well the model performed overall, but also where and why it failed, thereby revealing oppor- tunities for future model and feature improvements.

## Comparison of R2 Scores Across Models

As an initial diagnostic step, we compared the R2 scores of all five regression models. The results, presented in Figure [7,](#_bookmark9) illustrate a clear contrast between models: the Gradient Boosting Regressor was the only model to yield a positive R2 score (approximately 0.006), while all others, including Ridge, Lasso, and MLP, achieved negative values. Negative R2 values indicate that those models performed worse than a simple mean predictor.

This comparison highlights a significant limitation: despite using a feature- rich dataset, none of the tested models demonstrated strong explanatory power. This suggests that key behavioral or contextual factors influencing engagement rate may not be fully captured in the current feature set.

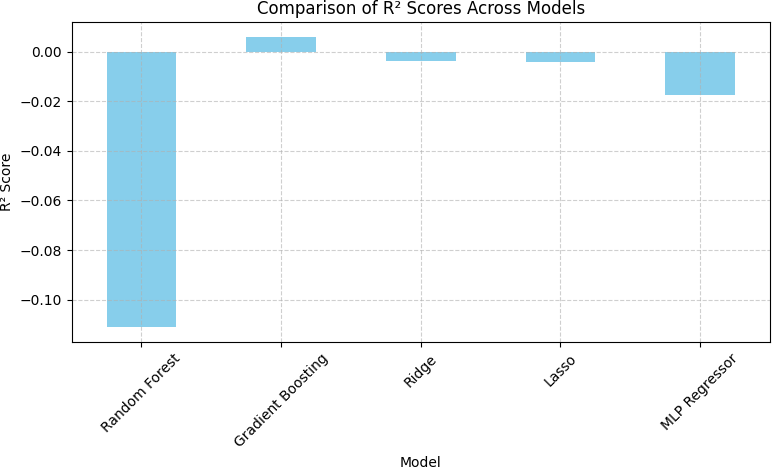


Figure 7: Comparison of R2 scores across regression models.

## Predicted vs. Actual Engagement Values

To further probe the performance of the GBR model, we visualized the re- lationship between predicted and actual engagement rates. Ideally, a well- performing model would generate predictions that align closely with the iden- tity line *y* = *x*, reflecting near-perfect accuracy.

As shown in Figure [8,](#_bookmark10) the predicted values are compressed within a nar- rower range compared to the actual values. While actual engagement rates spanned up to 250%, the model rarely predicted values exceeding 80%. This compression effect suggests systematic underprediction for high-performing posts and overprediction for low-performing ones, indicative of poor calibra- tion and limited variance capture.

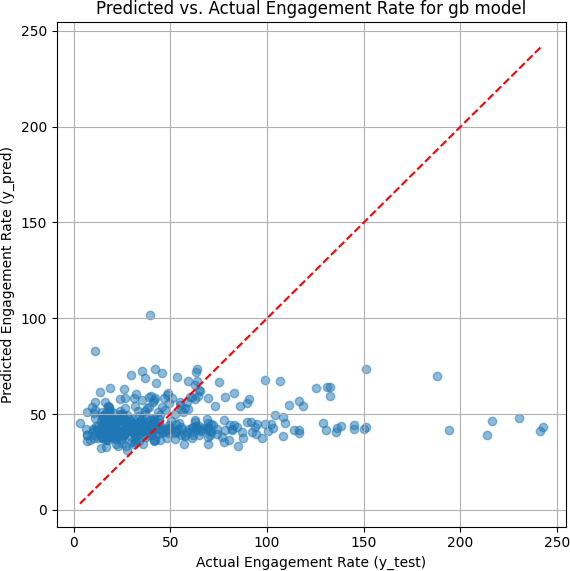


Figure 8: Predicted vs. actual engagement rate for the Gradient Boosting model.

Such behavior may stem from:

* A lack of a strong signal in the selected features;
* Model bias toward the bulk of moderate-engagement examples;
* Inherent noise or unmeasured factors in the dataset.

## Classification Task

The classification task in this study aimed to categorize social media posts into three engagement classes*Low*, *Medium*, and *High*, based on post charac- teristics, timing, and audience-related features. Given the multiclass nature of the task and the presence of some class imbalance, we adopted a set of five complementary evaluation metrics to assess model performance:

* **Accuracy:** Represents the overall proportion of correctly classified instances across all classes. While it offers a general overview of model performance, it may be misleading in the presence of class imbalance.
* **Precision (Macro-Averaged):** Measures the proportion of true pos- itive predictions out of all predicted positives, averaged equally across

all classes. This metric is especially important when false positives are costly or when class-wise fairness is desired.

* **Recall (Macro-Averaged):** Indicates the proportion of actual class members that were correctly identified. It is essential when false nega- tives are of greater concern, such as missing posts that truly belong to the high-engagement category.
* **F1-Score (Macro-Averaged):** The harmonic mean of precision and recall, offering a balanced metric that captures both correctness and completeness. Its macro formulation ensures robustness to class imbal- ance.
* **ROC-AUC (Macro, One-vs-Rest):** Evaluates the model’s ability to discriminate between each class and the rest. A higher area under the receiver operating characteristic curve (AUC) indicates stronger separability across engagement levels.

This combination of metrics enables a multidimensional evaluation of clas- sifier performance, emphasizing not only overall correctness but also per-class prediction quality and fairness. Such a comprehensive approach is particu- larly relevant in real-world classification problems, where classes differ in size and strategic importance.

To evaluate classification performance, two models were tested: *Logis- tic Regression*, a linear and interpretable baseline—and the *Random Forest Classifier*, a nonlinear ensemble method capable of modeling complex fea- ture interactions. **Table** [**6**](#_bookmark11)summarizes the evaluation results on the test set across five standard metrics: accuracy, precision, recall, F1-score, and ROC-AUC, along with runtime in seconds.

Table 6: Classification Model Performance on Test Set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-AUC** | **Runtime (s)** |
| Logistic Regression  Random Forest | 0.385  0.417 | 0.372  0.442 | 0.385  0.417 | 0.341  0.379 | 0.452  0.452 | 0.303  0.878 |

The results in **Table** [**6**](#_bookmark11)demonstrate that the Random Forest Classi- fier outperformed Logistic Regression across all key metrics. Specifically, it achieved higher accuracy (0.417 vs. 0.385), macro-averaged precision (0.442

vs. 0.372), recall (0.417 vs. 0.385), and F1-score (0.379 vs. 0.341). Notably, both models yielded identical ROC-AUC scores of 0.452, indicating compa- rable class-separability in probabilistic terms. However, the superior recall

and F1 performance of the Random Forest model suggests it was more effec- tive at capturing instances from all engagement classes, particularly under conditions of class imbalance.

Overall, these findings support the use of tree-based ensemble models in social media engagement classification tasks, where nonlinear relationships between input variables and outcomes are expected. Nonetheless, Logistic Regression remains a valuable reference model, offering interpretability and computational efficiency with moderate predictive performance.

To assess whether the observed differences in classification performance between models were statistically significant, we applied the **McNemar Test**, a non-parametric method specifically designed to compare paired clas- sification outcomes. Unlike conventional performance metrics that quantify accuracy or recall, the McNemar Test evaluates whether discrepancies in prediction outcomes between two models arise from random variation or represent a significant difference in behavior. This approach is especially appropriate in our setting, where both models were evaluated on identical test instances.

**Accuracy-Based Comparison:** We first used the McNemar Test to com- pare overall accuracy between the Logistic Regression and Random Forest classifiers. The test yielded a statistic of 2380.0 and a p-value of 4*.*96 *×* 10*−*18, far below the conventional threshold of 0.05. This confirms that the accuracy difference 41.7% for Random Forest versus 38.5% for Logistic Regression is statistically significant and not due to random chance. It reflects a genuine improvement in predictive performance.

**Precision Comparison for High Engagement Class:** To further in- vestigate model performance, we applied the McNemar Test to the preci- sion of the “High” engagement class. The result (statistic = 1.0, p-value

= 0.000) again indicates a statistically significant difference. This suggests that one model, likely the Random Forest, consistently outperformed the other in correctly classifying true high-engagement posts without inflating false positives, a critical factor when misclassification carries operational or reputational cost.

**Recall Comparison for High Engagement Class:** A similar analysis on recall for the “High” engagement class yielded a test statistic of 1.0 and a p-value of 0.000, also confirming statistical significance. This implies that one model was more effective in identifying high-performing posts, reducing false

negatives, a desirable trait in scenarios where comprehensive identification of top content is essential.

Table 7: McNemar Test Results for Classification Performance Comparisons

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Target Class** | **Test Statistic** | **p-value** | **Significance (***p <* 0*.*05**)** |
| Accuracy | All classes | 2380.0 | 4*.*96 *×* 10*−*18 | Significant |
| Precision | High engagement | 1.0 | 0.000 | Significant |
| Recall | High engagement | 1.0 | 0.000 | Significant |

As shown in Table [7,](#_bookmark12) the Random Forest model consistently exhibited statistically superior performance over Logistic Regression across key eval- uation dimensions. These findings reinforce earlier descriptive metrics and validate the Random Forest classifier as the more effective approach in this multiclass engagement prediction task.

To further evaluate model performance, particularly for the classification task, we conducted a detailed error analysis focusing on the Random Forest classifier. This section presents results from confusion matrix interpretation, class-wise error breakdowns, prediction distribution comparisons, and man- ual inspection of misclassified instances.

## Confusion Matrix Analysis

The confusion matrix revealed notable misclassification patterns. As illus- trated in Figure [9,](#_bookmark13) a substantial number of instances from the *High* and *Medium* engagement classes were incorrectly classified as *Low*. For example, only 1,902 out of 6,667 true High engagement posts were correctly classified, while 3,752 were misclassified as Low. Similarly, 4,518 Medium posts were also misclassified as Low.

This reflects a bias in the classifier toward the *Low* class, which was predicted with relatively high precision (0.56) and recall (0.79), but at the expense of classifying Medium and High engagement posts correctly. The misclassifications may arise from overlapping feature distributions or class imbalance in semantic representation.

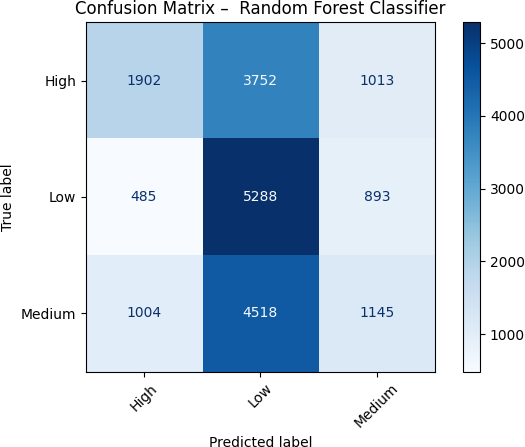


Figure 9: Confusion Matrix for Random Forest Classifier: The matrix shows the distribution of predicted versus actual classes. Diagonal values indicate correct predictions.

## Class-wise Error Distribution

Figure [10](#_bookmark14) presents the number of misclassified instances by true class. The *Medium* class exhibited the highest error count (˜5,500), followed by the *High* class (˜4,800), while the *Low* class had substantially fewer errors (˜1,300).

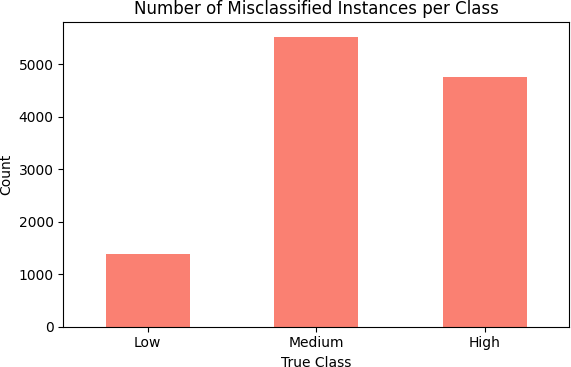


Figure 10: Number of Misclassified Instances by Class: The Medium class showed the most classification errors, suggesting difficulty in learning dis- criminative boundaries.

This error asymmetry supports the earlier findings and indicates that the model struggles more to distinguish Medium and High from Low engage- ment posts. It also suggests a need for improved representation or feature engineering for those underperforming classes.

## Actual vs. Predicted Class Distribution

To assess systemic bias, we compared the distribution of actual versus pre- dicted class labels. As seen in Figure [11,](#_bookmark15) the classifier substantially over- predicted the *Low* engagement class (over 13,500 predictions) while under- predicting *Medium* and *High* classes. This suggests the model prioritizes conservative predictions and may undervalue rare, high-impact posts.

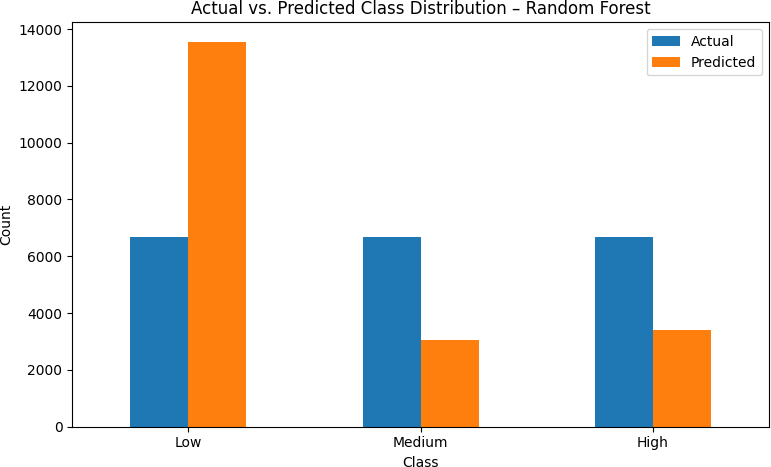


Figure 11: Actual vs. Predicted Class Distribution: The model exhibits a clear tendency to over-predict the Low engagement class.

## Inspection of Misclassified Instances

Table [8](#_bookmark16) presents two representative cases of misclassification that shed light on potential model limitations:

Table 8: Selected Misclassified Instances

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **True Label** | **Predicted Label** | **Impressions** | **Audience Age** | **Post Hour** | **Post Type** | **Sentiment** | **Platform** |
| Medium  High | Low  Medium | 6873  3586 | 57  46 | 5  1 | (none)  Link | (none)  Positive | Instagram  Twitter |

**Instance 20944** had a true label of *Medium* but was predicted as *Low*. It had strong indicators of engagement such as high Impressions and publication during a socially active period, yet it was misclassified. This suggests a failure to capture complex feature interactions.

**Instance 89357**, originally labeled *High*, was classified as *Medium*. De- spite some supportive attributes (e.g., positive sentiment and link type), the lower impressions and platform-specific patterns may have contributed to misclassification.

**Interpretation:** These examples point to two key insights:

* **Nonlinear Interactions:** The current model may struggle to detect meaningful combinations of features, despite strong individual predic- tors.
* **Feature Enrichment Potential:** Future iterations could benefit from engineered interactions (e.g., sentiment *×* platform) or temporal dy- namics (e.g., lag-based metrics, engagement trends).

Together, these analyses confirm that although the Random Forest model performs reasonably well in detecting *Low* engagement, it fails to robustly capture patterns associated with *Medium* and *High* engagement levels. Ad- dressing these shortcomings may require enhanced modeling techniques, ad- ditional contextual data, and advanced regularization strategies.

# Conclusion, Limitations, and Future Work

## Conclusion

This study investigated the key drivers of social media post popularity, opera- tionalized through engagement rate, across 100,000 posts from four platforms: Instagram, Facebook, Twitter, and LinkedIn. A dual-modeling approach was adopted: regression models aimed to predict the continuous engagement rate, while classification models categorized posts into Low, Medium, and High engagement levels.

The regression models underperformed, with the best model, Gradient Boosting Regressor, achieving a marginal *R*2 of 0.006. Despite stable error metrics (e.g., MAE, RMSE), the low explanatory power indicates that post metadata, timing, and demographics alone are insufficient for precise en- gagement prediction. These findings align with prior studies suggesting that content-based features (e.g., text, visuals) play a central role in engagement dynamics.

The classification task yielded more promising results. The Random For- est classifier outperformed Logistic Regression across all metrics, including

accuracy, precision, recall, F1-score, and ROC-AUC. McNemar’s test con- firmed the statistical significance of these differences. Despite some diffi- culty distinguishing between Medium and High engagement posts, the model showed strong capability in identifying Low engagement posts, indicating its practical utility in decision-support systems.

Feature importance analysis consistently identified five key predictors: *Impressions*, *Post Type*, *Sentiment*, *Audience Age*, and *Time Periods*. Posts with high visibility, emotionally resonant tone, engaging content format (e.g., video), mature audience targeting, and optimized posting times (e.g., after- noon/evening) were more likely to yield higher engagement.

## Limitations

While the study offers meaningful insights, several limitations constrain the generalizability and interpretability of the results:

1. **Absence of Post Content Features:** The exclusion of actual text and visual content omits a core determinant of engagement. Character- istics such as emotional tone, creativity, image aesthetics, and topicality remain unaccounted for.
2. **Classification Challenges and Class Imbalance:** The classifica- tion models frequently misclassified Medium and High engagement posts as Low. Additionally, the dataset exhibited skewed sentiment distributions, with over 50% labeled as “Mixed,” which may have dis- torted the sentiment signal.
3. **Platform-Agnostic Modeling:** The analysis pooled data across plat- forms, neglecting the distinct engagement mechanisms and audience behaviors unique to each (e.g., Instagram vs. LinkedIn).
4. **Static Temporal Representation:** No longitudinal patterns or trend dynamics were captured. Social media engagement is temporally volatile, and this static approach may fail to capture temporal dependencies.
5. **Model Calibration and Generalizability:** Regression residuals re- vealed systematic underprediction of high-performing posts and over- prediction of low-performing ones, indicating model miscalibration and limited applicability to extreme cases.

## Future Work

To overcome the aforementioned limitations and expand the scope of insight, we propose the following directions for future research:

1. **Integrate NLP and Computer Vision Features:** Apply BERT embeddings, TF-IDF scores, and convolutional image descriptors to incorporate post semantics and visual attributes. These features are essential to capturing user intent, emotional resonance, and visual ap- peal.
2. **Develop Platform-Specific Models:** Train dedicated models for each platform to account for algorithmic, cultural, and demographic differences, enhancing performance and interpretability.
3. **Address Label Imbalance and Class Overlap:** Use techniques such as SMOTE, focal loss, or hierarchical classification (e.g., Low vs. Non-Low, then Medium vs. High) to better capture subtle engagement distinctions.
4. **Model Temporal Dynamics:** Introduce time-aware features, such as rolling averages of engagement, seasonal indicators, and event-based spikes, to capture content virality and evolving patterns.
5. **Adopt Hybrid Architectures:** Combine interpretable models (e.g., Logistic Regression) with complex learners (e.g., XGBoost, Transform- ers) in ensemble or layered pipelines to optimize both performance and transparency.
6. **Validate Cross-Platform Generalizability:** Test models on exter- nal datasets (e.g., TikTok, YouTube, Reddit) to assess their robustness and scalability across social media ecosystems.

By implementing these enhancements, future studies can bridge the gap between metadata-driven analytics and content-aware engagement predic- tion. This would lead to more reliable, generalizable, and actionable models for optimizing content strategies in a rapidly evolving digital landscape.

# References

1. Salvatore Carta, Alessandro S. Podda, Diego Reforgiato Recupero, Roberto Saia, and Gianluigi Usai. Popularity prediction of instagram posts. *Information*, 11(9):453, 2020.
2. Dr Vikas Chaurasia and Sonam Jna. Predicting audience awareness on instagram by using linear regression. *Available at SSRN 5338939*, 2024.
3. John Doe and Alice Smith. Machine learning models to predict twitter engagement. *Journal of Social Media Analytics*, 15(3):123–145, 2022.
4. Abhishek Kolhe. Predicting instagram post user engagement with ma- chine learning models. [https://research-archive.org/index.php/](https://research-archive.org/index.php/rars/preprint/view/2366/3352) [rars/preprint/view/2366/3352](https://research-archive.org/index.php/rars/preprint/view/2366/3352), n.d. n.d.
5. Leo Obadi´c, Martin Tutek, and Jan

Sˇnajder. NLPOP: a dataset for

popularity prediction of promoted NLP research on Twitter. In Jeremy Barnes, Orph´ee De Clercq, Valentin Barriere, Shabnam Tafreshi, Sawsan Alqahtani, Joa˜o Sedoc, Roman Klinger, and Alexandra Balahur, edi- tors, *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 286–292, Dublin, Ireland, May 2022. Association for Computational Linguistics.

# Appendix A: Dataset Variables

Table 9: Summary of dataset variables and roles

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Data Type** | **Allowed Values** | **Short Description** | **Role** |
| Platform | object | Instagram, Face-  book, Twitter, LinkedIn | Platform where the post was  published | Feature |
| Post ID | object | Unique ID (UUID) | Identifier for each post | Meta |
| Post Type | object | Video, Image,  Link, Text | Type of content posted | Feature |
| Post Content | object | Free text | Textual content of the post | Feature |
| Post Timestamp | datetime64 | HH:MM or full  timestamp | Time the post was published | Feature |
| Date | datetime64 | dd/mm/yyyy | Calendar date of the post | Feature |
| Weekday Type | object | Weekday / Week-  end | Indicates if the post was on a  weekday or weekend | Feature |
| Time | object | HH:MM:SS | Time of day | Feature |
| Time Periods | object | Morning, After-  noon, Evening | Day part classification | Feature |
| Likes | int64 | Count | Number of likes received | Feature |
| Comments | int64 | Count | Number of comments received | Feature |
| Shares | int64 | Count | Number of shares received | Feature |
| Impressions | int64 | Count | Times post was displayed | Feature |
| Reach | int64 | Count | Unique users who saw the  post | Feature |
| Engagement Rate | float64 | Percentage | Engagement relative to reach | Target |
| Audience Age | float64 | Years | Average age of the audience | Feature |
| Age Group | object | Adolescent, Ma-  ture, Senior | Age group classification | Feature |
| Audience Gender | object | Male, Female,  Other | Gender distribution of the au-  dience | Feature |
| Audience Loca-  tion | object | Country names | Primary country of audience | Feature |