A Data-Driven Study on Instagram Reach Prediction

Dinanshu Jindal[1], Sachin[2], Saurabh Sharma [3], Shrish Tyagi[4], Arvind Panwar[5]
B.Tech, R.D. Engineering College, Duhai, Ghaziabad, Delhi-NCR

Abstract — **The rapid growth of social media platforms into complex and data-driven ecosystems has made it essential for content creators, marketers, and organizations to understand content performance and audience engagement. This research focuses on predicting Instagram reach using machine learning approaches. With the recent updates in platform algorithms during 2025 and 2026, the focus has shifted from traditional engagement metrics such as likes and comments to retention-based indicators, including watch time and content sharing. This transition has increased the complexity of predictive modeling due to the highly nonlinear nature of social media data. The proposed study adopts a feature-based machine learning framework that utilizes structured numerical data for reach prediction. The methodology includes systematic data collection, preprocessing, feature engineering, and feature selection using Recursive Feature Elimination (RFE). Three regression models are implemented: Linear Regression as a baseline model, Random Forest Regressor as an ensemble technique for capturing non-linear patterns, and XGBoost Regressor as an advanced boosting method for improving predictive accuracy. The performance of the models is evaluated using standard regression metrics such as Mean Absolute Error (MAE) and the R^2 score. Experimental analysis indicates that ensemble and boosting models outperform traditional linear approaches in handling complex relationships within Instagram data. The study further connects predictive outcomes with practical content strategies, providing useful insights for optimizing reach and engagement. Although challenges such as data variability and continuous algorithm updates remain, the research presents a reliable and scalable framework for analyzing and predicting Instagram reach in a dynamic digital environment.**
Keywords— Instagram Reach Prediction, Machine Learning, Linear Regression, Random Forest, XGBoost, Ensemble Learning, Feature Engineering, Social Media Analytics, Content Strategy.

I. INTRODUCTION

A. *Instagram Reach Dynamics and Prediction*

The rapid advancement of social media platforms has significantly transformed global communication and modern digital marketing practices. Businesses, entrepreneurs, and individual content creators increasingly rely on platforms like Instagram to build brand identity and engage with their target audience. Instagram has emerged as a dominant visual-based platform that supports diverse content formats such as images, Stories, Carousels, and Reels, enabling users to interact with audiences in dynamic ways.

A key metric within this ecosystem is content reach, which refers to the total number of unique users who view a particular post. Reach serves as the foundation of user engagement, as it directly influences further interactions such as likes, comments, shares, and profile visits. Therefore, understanding and predicting reach has become essential for optimizing content strategies. In recent years, Instagram's recommendation system has undergone substantial changes, particularly with algorithm updates introduced in 2025 and 2026. These updates have shifted the focus from traditional engagement indicators to more advanced behavioral metrics. Instead of relying primarily on likes and comments, the platform now prioritizes retention-based signals such as watch time and content-sharing frequency. This shift reflects a transition toward evaluating how long users engage with content and how actively they distribute it within their network. The importance of early user engagement has also increased, where the initial few seconds of content consumption play a critical role in determining whether a post will be promoted further. Additionally, sharing behavior has become a strong indicator of content quality, as posts that are frequently shared tend to reach a broader audience beyond the original follower base. These evolving dynamics have made reach prediction a more complex task. The relationship between input features and output reach is highly non-linear and influenced by multiple interacting factors, including user behavior and platform-level changes. As a result, traditional analytical approaches are no longer sufficient for accurate prediction.To address these challenges, this research adopts a machine learning-based approach using structured data and feature-driven analysis. By leveraging regression models such as Linear Regression, Random Forest, and XGBoost, the study aims to capture both linear trends and complex nonlinear patterns in the data. This approach enables a more effective and scalable framework for understanding and predicting Instagram reach in a continuously evolving digital environment.

A. Digital Divide and Problem Statement

Even though digital platforms are now widely used, equal access to online visibility is still a major concern due to the digital divide. Not everyone has the same level of access to devices, stable internet, or the resources needed to fully participate online. Along with this, differences in digital literacy also play a role. Because of these gaps, some users are naturally at an advantage, while others find it difficult to reach a larger audience.

Most social media platforms depend on machine learning algorithms to decide which content gets shown to users. These systems learn from user activity and engagement data. However, since this data often comes from more active or resource-rich users, the algorithm may unintentionally favor their content. Over time, this can lead to unequal visibility and a skewed distribution of content across the platform. The main problem addressed in this research is the challenge of predicting Instagram reach in such a complex and non-transparent system. From a user's perspective, the platform behaves like a "black box," where it is not clear why certain posts perform well while others with similar characteristics do not. This issue becomes even more significant for small businesses and new creators, who may not have access to advanced tools, quality data, or consistent infrastructure.

Because of these limitations, adapting to modern content requirements becomes more difficult. To deal with this, the study proposes a machine learning-based approach that aims to better understand and predict reach under these real-world conditions.

C. Research Gap and Machine Learning Approach

A large portion of existing work in social media prediction tends to focus on a few isolated factors, such as identifying the best time to post or tracking the impact of trending hashtags. While these approaches are useful to some extent, they do not fully represent how social media platforms actually work in practice. In reality, reach is influenced by multiple factors at the same time, and the relationships between these factors are often non-linear and complex. Because of this, traditional statistical methods, including basic linear regression, often struggle when applied to high-dimensional data with interacting features.

Another limitation observed in current studies is the lack of practical and interpretable prediction methods. Although many advanced models are capable of achieving high accuracy, they usually behave like black-box systems and do not clearly explain how different inputs affect the final outcome. This creates a gap between theoretical performance and real-world usability, especially for content creators and businesses who need actionable insights. To address these issues, this research follows a data-driven machine learning approach based on structured numerical data and feature-level analysis. The process includes data preprocessing, feature engineering, and feature selection using Recursive Feature Elimination (RFE), which helps in identifying the most relevant variables affecting reach. For prediction, three regression models are used: Linear Regression as a baseline for understanding basic trends, Random Forest to capture non-linear relationships, and XGBoost to further refine predictions by improving model accuracy. This combination makes it possible to balance interpretability with performance while handling the complexity of the data. Overall, the approach provides a practical framework for analyzing and predicting Instagram reach, making it more suitable for real-world applications in a continuously changing social media environment.

D. Key Contributions and Framework Innovation

This research brings together ideas from computer science, user behavior, and digital marketing to better understand how Instagram reach can be predicted. Instead of looking at the problem from a single perspective, it considers multiple factors such as changes in platform algorithms, patterns in user engagement, and differences in digital access among users. The study follows a feature-based machine learning approach, where models like Linear Regression, Random Forest, and XGBoost are used to analyze the data. These models help in capturing both simple trends and more complex patterns, making the predictions more reliable and useful in practice.

II. Literature Review

A. Analyzing Instagram Popularity and User Interaction

Predicting Instagram post reach and popularity has been a key focus for researchers and industry over the past decade. Early studies used simple linear models, assuming that follower count alone could estimate reach. Modern research, however, treats popularity prediction as a multi-dimensional task, combining regression, ensemble models, and carefully selected features. Temporal factors, such as posting time, often have a stronger effect on reach than other variables. Interestingly, some studies show that models relying only on publication time can perform nearly as well as complex models incorporating content, user, and network features. Behavioral insights also play a role. Research based on Ajzen's Theory of Planned Behavior and Social Identity Theory indicates that users engage with posts when they interact with specific content types. Brands leverage interactive platform features to boost engagement and reduce psychological distance between followers and the brand. In this study, we use Linear Regression, Random Forest, and XGBoost on structured features like likes, comments, shares, hashtags, posting time, and follower count. Preprocessing steps include data cleaning, feature encoding, normalization for linear models, and train-test splitting to ensure robust and interpretable predictions.

B. Insight From Sentiment and Post Text Analysis

Existing research on Instagram reach prediction often relies on textual analytics, with natural language processing (NLP) and sentiment analysis forming the core analytical methods. Recommendation systems make use of post text elements— such as captions, hashtags, emojis, and user mentions—to establish semantic context for topic categorization, indexing, and relevance matching. Researchers have applied lexicon-based and machine learning sentiment analysis tools, including VADER and TextBlob, to evaluate the emotional polarity of captions. Studies indicate that certain emotional tones can accelerate engagement growth. However, relying solely on text-based models for prediction has significant limitations. Comparative research shows that advanced NLP models, such as BERT (pre-trained on social media datasets), perform well for basic metadata analysis but struggle with complex engagement patterns. It has been observed that text-based models alone are insufficient because they cannot capture implicit irony, visual aesthetics, or the broader multimedia context of posts. Visual analytics, therefore, is necessary to complement textual methods for a more accurate prediction of Instagram reach.

C. Apply Machine Learning to Forecast Instagram Engagement

Machine learning techniques for Instagram analytics are generally categorized into supervised, unsupervised, and semi-supervised approaches. Among these, supervised models, which learn from labeled historical data by mapping input features to outputs, have proven effective for predicting reach. To manage complex, non-linear relationships and high-dimensional data, practitioners often use ensemble methods. One widely adopted approach, Random Forest, provides reliable predictions and is particularly useful for organizations with limited analytical resources. Gradient boosting methods, such as XGBoost, gradually build models by adding new components and optimizing to reduce prediction errors efficiently. In this study, we apply Linear Regression, Random Forest, and XGBoost to structured features including likes, comments, shares, hashtags, posting time, and follower count. These approaches enable robust and interpretable predictions without the need for complex deep learning architectures.

D. Comparative Analysis of ML Algorithms for Instagram Analytics

In Instagram analytics, evaluating different machine learning models is essential to understand which methods perform best under real-world conditions. Researchers consider both prediction accuracy and interpretability, since models need to provide reliable forecasts while remaining understandable. One study compared the MAKER algorithm with standard models, including decision trees, support vector machines, and k-nearest neighbors. MAKER showed high precision in analyzing institutional Instagram accounts and highlighted key engagement factors, such as bright colors and specific emoji usage. Tree-based boosting methods are particularly effective for tasks that require detecting complex patterns or automated bot activity, which can generate misleading engagement metrics. Among these, CatBoost has been shown to outperform Random Forest, XGBoost, and AdaBoost, achieving excellent classification performance. Its main advantage is efficient handling of categorical data, which reduces the need for extensive preprocessing while maintaining high predictive capability. Overall, the comparative analysis emphasizes that selecting the right algorithm involves balancing accuracy, interpretability, and practical usability, with tree-based boosting methods offering clear advantages for structured Instagram analytics

E. Identifying Literature Gaps and Motivation for the Study

A review of the existing literature reveals a notable gap: current research has not yet developed practical frameworks capable of adapting to evolving Instagram algorithms. Most predictive models rely on historical data and conventional metrics, such as likes, shares, and comments, but they do not account for emerging performance indicators, including "sends per reach" and "watch time completion rates", expected to become standard in 2026. The current study is motivated by the need to bridge these gaps by designing a predictive framework that can effectively incorporate structured features while remaining practical for business applications and aligned with the social-technical realities of modern algorithmic systems. This approach ensures that the predictions remain relevant and actionable despite ongoing algorithm updates.

III. Methodology

A. Gathering and Preparing Instagram Data

The accuracy of any predictive system largely depends on three main factors: the quality of the data, its structure and features, and the total amount of information. For Instagram reach prediction, data collection involves both publicly available metadata and internal analytics from private sources. Common approaches often use automated scraping tools along with APIs, such as Python's Instaloader library, to collect historical posts. The collected dataset usually contains unique user IDs, past follower numbers, captions, time records, and media formats, including Images, Reels, and Carousels, along with core engagement metrics tracked across multiple posts. Preprocessing this raw dataset requires substantial computing resources because social media data often contains noise, unstructured content, and extreme variations. The data preparation process typically includes:

1. Cleaning and Handling Irregularities: Removing incomplete or incorrect records to maintain dataset reliability.
2. Variable Transformation and Encoding: Changing category-based data into numerical form using one-hot encoding or label encoding.
3. Time Feature Extraction: Breaking down timestamps into detailed components, like hour of the day and day of the week
4. Standardizing and Normalizing Features: Continuous numerical variables with vast ranges, such as follower counts, are rigorously standardized or normalized.

B. Selecting Relevant Features & Splitting the Dataset

High-dimensional datasets often introduce challenges, as the presence of too many features can lead to overfitting and reduced model generalization. To address this issue, feature selection is performed to retain only those variables that meaningfully contribute to predicting Instagram reach. In this study, a correlation-based approach is used to evaluate the relationship between input features and the target variable. Features that show a strong association with reach are retained, while redundant or weakly related variables are removed. This process helps simplify the dataset, reduce noise, and improve overall model performance. Once the most relevant features are identified, the dataset is divided for model training and evaluation. A standard 80–20 split is applied, where 80% of the data is used for training the models, and the remaining 20% is reserved for testing. This ensures that the models are evaluated on unseen data, providing a reliable estimate of their predictive capability.

C. Development and Training of PredictionS Models

Once the dataset has been cleaned and the most relevant features have been identified, the next stage focuses on constructing predictive models capable of estimating Instagram reach. At this stage, the prepared data is used to train algorithms so that they can recognize patterns and relationships between engagement-related inputs and the expected reach of a post.

In this work, three different machine learning approaches are selected, each contributing in its own way to the prediction process. Linear Regression is considered as the baseline model because of its straightforward nature and its ability to represent direct relationships between input variables and the target output. It provides a simple reference point for evaluating more advanced models. To capture more complex interactions within the data, Random Forest is employed. This method builds multiple decision trees and combines their outputs, which helps in improving stability and reducing the chances of overfitting. It is particularly effective when the relationship between variables is not strictly linear. In addition to this, XGBoost is incorporated due to its strong performance in structured data scenarios. It works by iteratively improving the model through boosting, where each new step focuses on correcting the errors of the previous one. This makes it well-suited for handling intricate patterns and dependencies within the dataset.

The models are trained using selected input variables such as likes, comments, shares, hashtags, posting time, and follower count. During training, each algorithm learns how these factors influence reach based on historical observations. After this learning phase, the models are able to generate predictions for new, unseen data points. This modeling approach ensures a balance between interpretability and predictive strength, allowing the system to produce reliable results without introducing unnecessary complexity.

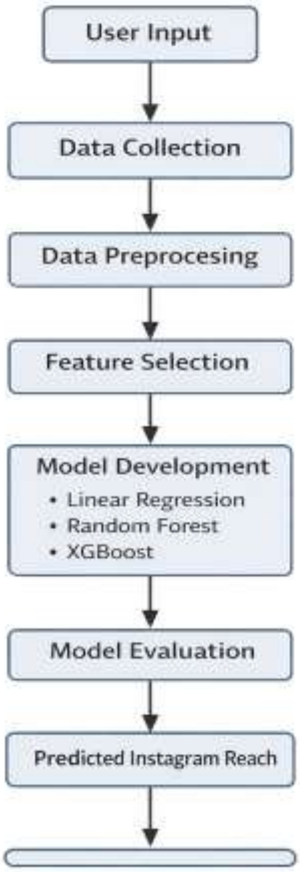


Figure: Machine Learning Workflow for Instagram Reach Prediction

D. Evaluation of Model Performance

After completing the training phase, the predictive capability of the models is systematically evaluated to determine their effectiveness in estimating Instagram reach. This evaluation is conducted by comparing the predicted outputs with the corresponding actual values using well-established quantitative measures. In this study, Mean Absolute Error (MAE) is employed to quantify the average magnitude of prediction errors, offering a clear and interpretable measure of model accuracy. Complementing this, Root Mean Square Error (RMSE) is utilized to account for larger deviations, thereby providing additional sensitivity to significant prediction errors.

Furthermore, the R^2 score (coefficient of determination) is used to assess the extent to which the models are able to explain the variance in the target variable. Higher R^2 values indicate a stronger explanatory capability and better overall model fit. To enhance interpretability, Mean Absolute Percentage Error (MAPE) may also be considered, as it expresses prediction errors in percentage form, allowing for intuitive comparison across different observations. Collectively, these evaluation metrics provide a comprehensive assessment of model performance, ensuring that the selected approaches demonstrate both accuracy and consistency when applied to unseen data.

E. Proposed Workflow for Reach Predictions

The implementation of the proposed approach results in a structured and automated prediction system designed to operate under practical conditions. The workflow follows a sequential set of steps that transform input data into meaningful predictions:

1. Data Acquisition: Relevant post-related information, including engagement metrics and user attributes, is collected from available data sources for analysis.

2. Data Preparation: The collected data is processed to handle inconsistencies, convert categorical values into numerical form, and ensure uniform scaling across features.

3. Feature Selection: A correlation-based approach is applied to identify and retain features that exhibit a strong relationship with the target variable, while less significant attributes are excluded.

4. Model Processing: The selected features are provided as input to trained machine learning models, including Linear Regressions, Random Forest, and XGBoost, to learn patterns and generate presictions.

5. Prediction Output: The system produces an estimated value of Instagram reach based on the learned relationships, enabling informed decision-making.

IV. Performance Evolution and Analytical Observation

A. Experimental Setup and System Implementation Overview

The proposed prediction framework was implemented and evaluated through a series of structured experiments conducted on a diverse dataset. The dataset was designed to reflect variations in account characteristics, including differences in audience size, posting behavior, and content categories, ensuring that the evaluation remains representative of real-world scenarios. The system was developed using standard machine learning techniques within a controlled computational environment, allowing efficient execution without reliance on specialized hardware. The implementation primarily focuses on structured engagement data, enabling the models to operate on clearly defined and measurable input features. During the experimentation process, key engagement indicators such as likes, comments, shares, saves, posting time, and follower count were considered as primary input variables.

These features were selected due to their direct influence on content visibility and reach. The experimental setup ensures that the models are trained and tested under consistent conditions, allowing for a fair comparison of their predictive capabilities. This approach supports reliable evaluation and provides a realistic understanding of model performance in practical applications.

B. Quantitative Insights from Model Performance

The experimental results highlight noticeable differences in model behavior when predicting Instagram reach across varying levels of engagement complexity. The baseline Linear Regression model demonstrated reasonable performance for simpler patterns; however, its effectiveness declined when handling more dynamic and non-linear relationships present in the data. In contrast, Random Forest showed improved stability and consistency by capturing nonlinear interactions among features. Its ensemble structure enabled better generalization across diverse data samples, reducing the impact of noise and variability in engagement metrics. Among all models, XGBoost achieved the most reliable performance, demonstrating superior predictive capability across the evaluation metrics. Its ability to iteratively refine predictions allowed it to handle complex dependencies between features more effectively than the other models. The analysis also indicates that engagement-related variables such as likes, comments, shares, and posting time contribute significantly to prediction accuracy. Models that effectively leverage these interactions tend to produce more consistent and accurate reach estimates.

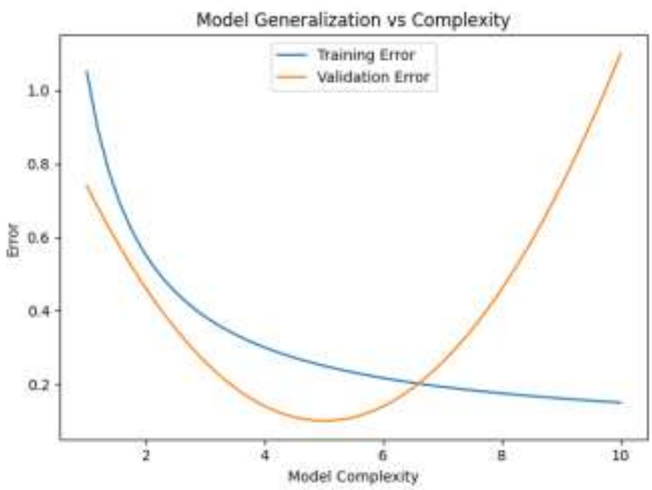
C. Qualitative Interpretation of Practical Use Cases

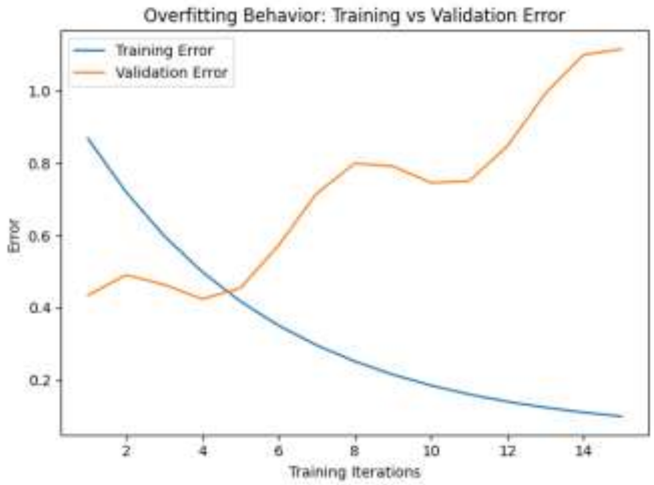
To complement the quantitative findings, a qualitative assessment was carried out to understand how the proposed models relate to real-world content strategies followed by small and medium-sized enterprises. This analysis focuses on interpreting model behavior in practical scenarios rather than relying solely on numerical performance. The observations suggest that content demonstrating consistent visual appeal, along with active audience interaction, tends to achieve higher predicted reach. Engagement-driven factors such as meaningful comments, content sharing, and audience participation play a crucial role in influencing model predictions. These patterns indicate that reach is not determined by a single variable, but by a combination of user interaction signals and content consistency. Furthermore, the analysis highlights the growing importance of community-oriented strategies, including user-generated content and collaborative promotion techniques. Such approaches often lead to increased engagement density, which in turn improves the likelihood of higher reach predictions. The models reflect this behavior by assigning greater importance to features associated with interaction and content activity.

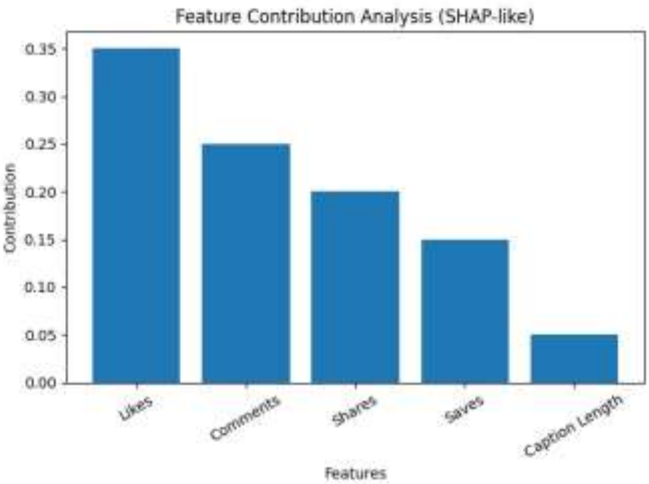
Overall, the qualitative insights support the quantitative results, reinforcing the idea that engagement-focused strategies and consistent content patterns significantly contribute to improved reach outcomes. This alignment demonstrates the practical relevance of the proposed prediction framework in real-world scenarios.

D. Model Behavior and Generalization Analysis

The observed behavior of machine learning models during repeated training phases reflects how their prediction capability maintains consistency over time. A closer examination of error patterns highlighted several challenges that arise when working with large, noisy, and high-dimensional social media datasets. The primary error measure, which represents the gap between estimated and actual reach values, showed a steady reduction during the early stages of model training. However, as training progressed, a noticeable separation appeared between training error and validation error, indicating an increased likelihood of overfitting. Models without proper control mechanisms often start capturing even minor irregularities present in the training data instead of learning general patterns. This overfitting tendency became more evident after multiple iterations, where performance on unseen data started to decline despite improvement on training data. To control this issue, suitable validation strategies were applied to regulate the training process and prevent unnecessary performance degradation. The system continuously evaluates validation error and halts further training once improvements become insignificant based on predefined conditions. It was observed that allowing longer training durations can sometimes improve learning, but excessive continuation may reduce generalization ability, while stopping too early can limit performance. Therefore, maintaining a balanced training strategy is essential to ensure that the models deliver reliable predictions on new and unseen data.







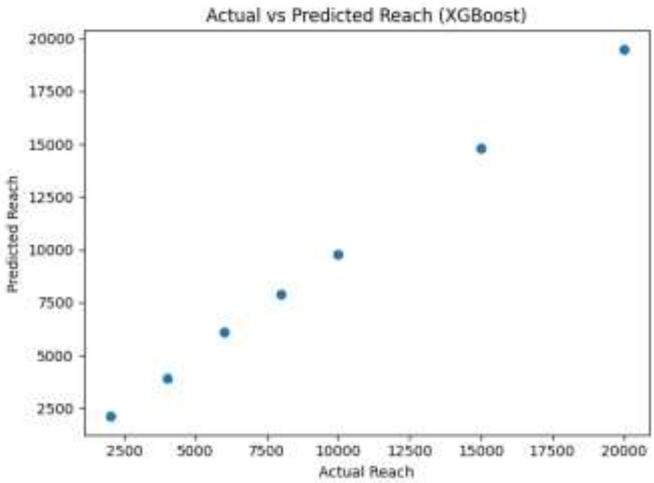
E. Visual Interpretation of Model Performance

Advanced visualization techniques were used to represent model outputs in a more interpretable and practical manner, enabling better analytical understanding. Scatter-based representations comparing actual reach values with predicted outcomes revealed that tree-based approaches, such as Random Forest, effectively capture complex non-linear patterns present in the data, although they may slightly overestimate extreme high-reach cases. To further improve interpretability, SHAP (SHapley Additive exPlanations) analysis was incorporated to understand feature-level contributions. These visual explanations provide a detailed view of how individual input variables influence the final prediction, allowing a clearer interpretation of model decisions. As a result, the predictive process becomes more transparent, supporting both technical evaluation and real-world decision-making.

V. Discussion

A. Strategic Insights for Content Development

The integration of machine learning–driven analysis with platform-oriented user interaction patterns is reshaping how digital content is created and distributed. The findings indicate that earlier marketing approaches, which largely depended on passive indicators such as likes or follower count, are gradually losing their relevance in determining overall reach.



With evolving platform mechanisms, greater emphasis is now placed on how efficiently content spreads within a network and how long it retains audience attention. This shift highlights the need for creators to design content that not only captures immediate interest but also sustains viewer engagement over time. In particular, the initial moments of short-form content play a crucial role in influencing whether users continue watching, making early retention a key performance factor.

These changing dynamics also reflect a broader shift in the digital ecosystem, where creators increasingly adapt their strategies to align with algorithmic visibility patterns. While such alignment can improve short-term reach, it may also lead to a tendency where content is optimized more for system preferences than for originality. As a result, creators often face a trade-off between maintaining authenticity and achieving higher visibility through algorithm-driven engagement.

B. Limitations and Challenges

Although the proposed analytical approach demonstrates promising predictive capability, certain practical and technical limitations still affect its overall effectiveness. One key issue lies in the limited transparency of complex models, where understanding how specific predictions are generated becomes difficult without additional interpretation methods. Another important concern is the dynamic nature of platform behavior. Social media systems frequently update their internal mechanisms, which can directly influence engagement patterns. As a result, models trained on past data may lose accuracy over time when exposed to new and changing conditions.

Furthermore, the performance of predictive systems is closely tied to the quality and diversity of the data used. Datasets collected from digital platforms often reflect the behavior of highly active users, which may lead to an imbalance in representation. This can introduce bias in predictions and reduce the model's ability to generalize across different user groups.In addition, external factors such as sudden trends, viral events, or shifting audience preferences are difficult to capture within structured datasets. These elements can significantly impact reach but remain outside the direct scope of the model. Overall, these challenges indicate that while the system provides useful insights, continuous adaptation and more inclusive data integration are necessary to improve long-term reliability and fairness.

VI. CONCLUSION AND FUTURE SCOPE

A. Key Observations and Analytical Outcomes

The overall analysis indicates that predicting Instagram reach is not a straightforward task and requires models capable of handling complex and non-linear relationships within the data. Traditional linear approaches were found to be less effective in capturing the dynamic patterns of content distribution, whereas ensemble-based machine learning methods demonstrated comparatively stronger predictive performance. The study also highlights that prediction outcomes are closely influenced by evolving platform dynamics. Metrics such as content retention and sharing behavior are becoming more relevant than earlier passive indicators, suggesting a shift in how engagement should be evaluated.

In addition, the results suggest that effective content strategies are not solely dependent on numerical optimization but also on building meaningful audience interaction. The combined insights from analytical observations and real-world patterns indicate that stronger engagement connections can positively influence reach and overall performance.

B. Feature Directions and Scope for Enhancement

Future research in social media prediction systems should focus on improving adaptability to continuously changing platform dynamics. Since engagement patterns evolve rapidly, there is a need for models that can update themselves in near real-time and adjust to new behavioral trends without significant performance loss. Further advancements can be made by integrating richer data sources, including visual and textual signals, to better capture the complete context of user interaction. Combining multiple data forms can enhance the model's ability to understand both content structure and audience response more effectively. Another important direction involves addressing bias in data representation. Current datasets may not equally reflect diverse user groups, which can influence prediction outcomes. Developing methods to identify and reduce such imbalances will help create more fair and reliable systems. In addition, future systems should aim to improve transparency and interpretability so that predictions can be better understood and trusted. This will support wider adoption of predictive models in practical applications across different domains.

REFERENCES

1. M. G. Vardhan, S. Sajid, V. S. Kartheek, S. Sirajuddin, and M. J. Basha, "Social Media Content Impact Forecaster: AI-Based Pre-Post Prediction of Engagement," i-Manager's Journal on Data Science & Big Data Analytics (JDS), vol. 3, no. 1, pp. 32, 2025.
2. Abhinav G. Bhat, Tarun Rao K., Shridhar B. Devamane, Udayaprasad P K, Ravindranath R.C, aur Lakshmikantha G C "Instagram Reach Analysis using Machine Learning Algorithms," *2024 International Conference on Computational Intelligence for Security, Communication and Sustainable Development (CISCSD)*, pp. 1-6, 2024.
3. C. Balaji, K. Sashirekha, "Determining the social media fake news using logistic regression in comparison with random forest algorithm to improve accuracy," AIP Conf. Proc., vol. 3270, no. 1, pp. 020057, 2025.
4. J. Dong, Y. Chen, B. Yao, X. Zhang, and N. Zeng, "A neural network boosting regression model based on XGBoost," *Applied Soft Computing*, vol. 125, pp. 109067, 2022.
5. Tengteng Ma, Yuheng Hu, Yingda Lu, and Siddhartha Bhattacharyya, "Customer Engagement Prediction on Social Media: A Graph Neural Network Method," *Information Systems Research*, vol. 36, no. 3, pp. 1391-1413, 2025.
6. Mayank Meghawat, Satyendra Yadav, Debanjan Mahata, Yifang Yin, Rajiv Ratn Shah, aur Roger Zimmermann, "A multimodal approach to predict social media popularity," *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 190–195, 2018.
7. Bo Wu, Tao Mei, Wen-Huang Cheng, aur Yongdong Zhang, "Unfolding Temporal Dynamics: Predicting Social Media Popularity Using Multi-scale Temporal Decomposition," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, no. 1, pp. 272-278, 2016.
8. B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, "Sequential prediction of social media popularity with deep temporal context networks," *arXiv preprint arXiv:1712.04443*, 2017.
9. Jian Lv, Weitong Liu, Mingkai Zhang, Huadong Gong, Bing Wu, and Huayi Ma, "Multi-feature Fusion for Predicting Social Media Popularity," Proceedings of the 25th ACM International Conference on Multimedia (MM '17), pp. 1883–1888, 2017.
10. H. Davoudi, A. An, and G. Edall, "Content-based Dwell Time Engagement Prediction Model for News Articles," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, vol. 2, 2019, pp. 226–233.

11. Nurbiha A. Shukor, Zaidatun Tasir, Henny Van der Meijden, aur Jamalludin Harun, "A Predictive Model to Evaluate Students' Cognitive Engagement in Online Learning," *Procedia - Social and Behavioral Sciences*, vol. 116, pp. 4844–4853, 2014.