Research Article

Predicting Instagram Post User Engagement With Machine Learning Models

Aarav Kolhe

Allen High School, 300 Rivercrest Blvd, Allen, TX 75002, USA

ABSTRACT

This study investigates the practical application of statistical machine learning techniques to predict average Instagram post user engagement based on an account's follower count. It explores three distinct models—linear Regression, Random Forest Regression, and Neural Networks—for their effectiveness in modeling engagement patterns. Using recent Instagram data, each model is trained on the average user engagement for a profile. The predicted data is then compared to the actual data to determine the most accurate and viable model. The Neural Network excelled in capturing variance, having the highest R-squared value of the three models tested, but struggled with overfitting. Random Forest handled non-linear patterns well, having the lowest mean squared error out of the three models, but tended to overestimate. LASSO Regression was a balance between both the Neural Network and Random Forest Model, maintaining variance capture while reducing overestimation. Future research could refine models or explore hybrid approaches for better scalability. Machine learning shows promise in predicting post popularity, but further improvements are needed to aid social media creators and developers.

Keywords: Engagement; AI; Followers; Posts; Performance

INTRODUCTION

Post user engagement on social media platforms refers to the interactions a post receives from its audience, serving as a key metric for measuring content effectiveness and audience interest. Engagement encompasses a variety of elements, including likes, comments, shares, saves, and click-throughs, each reflecting a different type of user interaction (1). These elements provide insights into user preferences and behavior, allowing content creators, brands, and marketers to evaluate the success of their strategies. Understanding post engagement is essential for optimizing content to foster stronger connections with the target audience and achieve specific goals, such as increasing visibility, driving traffic, or enhancing brand loyalty. This paper examines post engagement as a multifaceted metric and explores predictive models to better understand its relationship with account follower counts. However, predicting engagement solely based on follower count can be challenging due to varying audience behaviors and platform algorithms. To address this, machine learning (ML) techniques provide a powerful toolkit for developing predictive models. This research

Corresponding author: Aarav Kolhe, E-mail: aarav.kolhe29@gmail.com.

Received March 13, 2025**; Accepted** April 23, 2025
https://doi.org/10.70251/HYJR2348.328387

explores the use of Linear Regression, Random Forest Regression, and Neural Networks to model the relationship between follower count and engagement. By leveraging these methods, the study aims to uncover patterns and improve the accuracy of engagement predictions, offering actionable insights for social media strategy optimization.

MATERIALS AND METHODS

Web scraping

To collect Instagram data for this research, we employ *PhantomBuster*, an automation tool tailored to interact with platforms like Instagram. The process involves the following steps:

Using PhantomBuster. PhantomBuster focuses on automating actions in web browsers. It is designed to mimic user interactions with Instagram's online platform. The tool automates the actions of scrolling and clicking to retrieve engagement statistics for posts. This method only collected the posts' like and comment counts. The number of followers for the selected accounts was found using third-party analytics tools. The collected data is incorporated into a CSV for the analytical process.

Pipeline Overview. A list of Instagram accounts was imported for targeted scraping. The tool extracts the number of likes and comments from each post and the number of followers from the corresponding profile.

This data is then placed in a data sheet where user engagement for each post is calculated. The collected data is then refined by removing outliers and duplicates to ensure. This methodology provides a scalable, efficient, and ethical approach to collecting Instagram data, enabling the subsequent application of machine learning models to analyze post-engagement patterns.

Dataset

Over 1000 posts were scraped. Private accounts and their data were not used. Bot accounts were excluded from the dataset because they are outside this study's focus. Follower count was split into five categories: 1,000,000 followers - 500,000 followers, 499,999 followers - 100,000 followers, 99,999 followers - 10,000 followers, 10,000 followers - 5,000 followers. This range of 1,000,000 to 5,000 followers was selected to avoid the potential outliers that come from profiles outside of this range. These outliers are caused by either an insuffucient number of followers to have any post interaction or having too many followers, with many of them possibly being dead or inactive accounts that don't engage with the platform. Twenty random posts were selected from each profile, with only the necessary observed numerical data being used. No content was collected, including posts, videos, photos, descriptions, and individual comments.

LASSO Regression Model

LASSO Regression (Least Absolute Shrinkage and Selection Operator) is a linear regression technique incorporating L1 regularization to enhance model performance and interpretability (2). The model minimizes the sum of squared residuals while adding a penalty proportional to the absolute values of the regression coefficients. This regularization forces some coefficients to shrink to zero, effectively selecting a subset of relevant features. Lasso is instrumental when working with high-dimensional data, as it helps mitigate overfitting by simplifying the model. LASSO Regression explains the relationships between predictors and the target variable by focusing only on the most impactful variables.

Random Forest Model

The Random Forest model is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting (4). Each tree in the forest is built using a random subset of features and training data, introducing diversity in the model. During prediction, the outputs of individual trees are averaged (for regression) or voted on (for classification) to generate the final result. Random Forests are highly robust, capable of capturing non-linear relationships, and effective at handling missing data and outliers. Additionally, they provide feature importance metrics, which help identify the most influential predictors in the dataset.

Neural Network Model

Neural Networks are a class of machine learning models inspired by the structure and functioning of the human brain. They consist of layers of interconnected nodes, or neurons, where each neuron processes input data and passes the output to subsequent layers (3). Neural networks excel at capturing complex, non-linear relationships in data through their ability to learn hierarchical representations. The model's architecture typically includes an input layer, one or more hidden layers, and an output layer. Using activation functions and backpropagation to adjust weights during training, neural networks iteratively minimize prediction errors (3). While computationally intensive, they are highly versatile and practical, particularly for large datasets and problems involving intricate patterns.

Average User Engagement

Each of the selected Machine Learning Models requires an x input and a y output to train on and predict. The number of followers is the x value, and the average user engagement value is the y value. In this study, user engagement equals the total interactions (total likes plus total comments) divided by the total number of followers from the post's user profile (5). User engagement will be kept in decimal form and not converted to a percentage.

Evaluation Metrics

To evaluate the performance of the machine learning models and the effectiveness of the adjusted engagement metric, several key evaluation metrics are employed, including:

- R^2 (R^2): The R^2 metric measures the proportion of variance that the models predict. A higher R^2 value indicates a better fit to the original data.
- Mean Squared Error (MSE): This metric measures the amount of error in statistical models' predictions. A higher value indicates more inaccurate predictions, whereas a lower value indicates more accurate predictions.
- Histogram: A histogram provides insights into the prediction's distribution and potential biases. It shows outliers

RESULTS

LASSO Regression

The histogram in Figure 1 illustrates the distribution of standardized user engagement predicted by the LASSO regression model. This visually represents the predictions generated by LASSO Regression, providing insight into its performance and revealing potential patterns. This histogram is skewed to the left, with most generated predictions concentrated above the mean. This reflects the typical linear relationship that LASSO Regression is known to capture in its data sets.

Random Forest

The histogram in Figure 2 shows the distribution of standardized user engagement predicted by the Random Forest model. It visually represents the predictions generated by Random Forest, providing insight into its performance and revealing potential patterns. This histogram is skewed to the right, with most of the model's predictions concentrated below the mean and a few outliers above.

Neural Network

The histogram in Figure 3 presents the distribution of standardized user engagement predicted by the neural network model. This visually represents the model's generated predictions, offering insight into its performance and revealing potential patterns. This bimodal histogram shows that the model's predictions are focused below the mean with a few outliers above the mean. This bimodal histogram means the neural network could capture possible nonlinear patterns in the dataset.

Mean Squared Error (MSE) and R-Squared (R^2)

Table 1 displays the Mean Squared Error and R-squared value for each model tested. MSE and R^2 provide insight

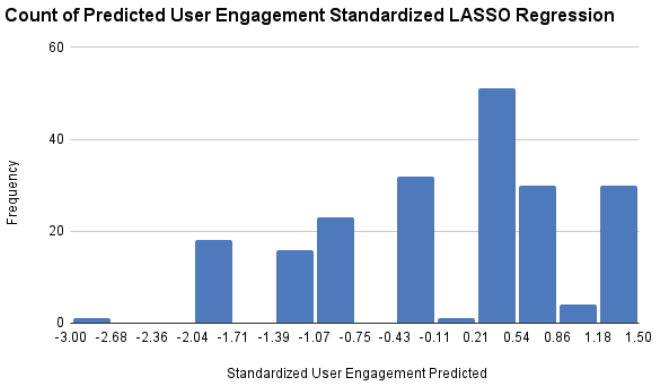

Figure 1. Histogram based on the count of User Engagement predictions standardized from a trained LASSO Regression model.

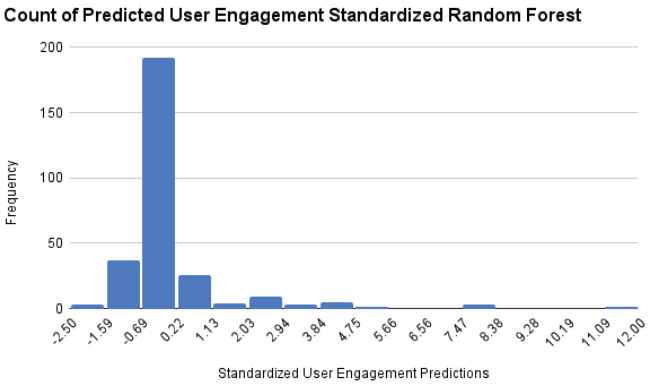

Figure 2. Histogram based on the count of User Engagement predictions standardized from a trained Random Forest model.

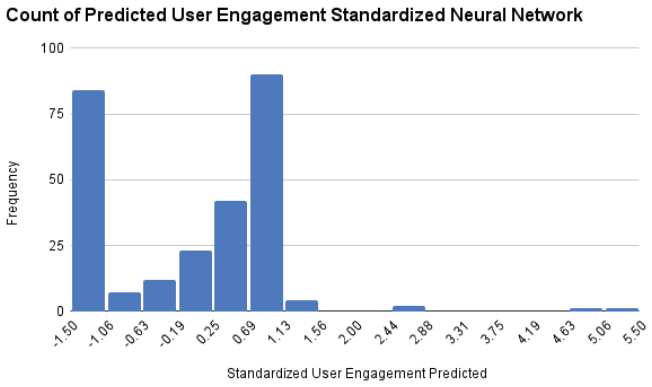

Figure 3. Histogram based on the count of User Engagement predictions standardized from a trained Neural Network model.

Table 1. MSE and R^2 value of each model's predictions for User Engagement compared to the actual User Engagement

	MSE	R^2
LASSO Regression	5.5460	0.0789
Random Forest	1.2340	0.0262
Neural Network	11.5461	0.1359

into each model's possible errors and ability to capture the data variance. These values will compare each model's overall performance and accuracy.

DISCUSSION

LASSO Regression Model

The LASSO Regression Model demonstrated its potential to improve User Engagement calculations. The histogram with a curve is lightly skewed to the left, suggesting that the model underpredicts actual values (Figure 1). However, an MSE of 5.5460 indicates that the model is relatively more accurate when compared to the Neural Network (Table 1). Also, the R-squared value of 0.0789 suggests that the LASSO model can better capture the data's variance than the Random Forest model (Table 1). The LASSO regression model seems to be the most accurate out of the three tested in the study.

Random Forest Model

The Random Forest Regression Model showcased its ability to minimize errors when predicting User Engagement values. The histogram has a skewed curve to the right, suggesting that the Random Forest model underestimates high engagement values (Figure 2). If the model rarely predicts high values, it might not correctly capture patterns associated with high engagement. Also, with an R^2 value of .0262, this model could not capture the variance in the given data set as effectively as the other two models tested (Table 1). However, the model is shown to have minimal errors compared to the Neural Network and LASSO Regression models, having an MSE of 1.2340 (Table 1). Discrepancies between predicted and actual values highlight areas for improvement, especially when the model overestimated user engagement.

Neural Network

The Neural Network model exhibited promising predictive capabilities for User Engagement. Observing the histogram, the standardized predicted user engagement values that form a bi-modal curve are concentrated on the left side of the histogram, suggesting the model predicts user engagement at multiple standard levels rather than following a simple, regular, or uniform pattern (Figure 3). This might explain why the Neural Network had the highest R^2 (0.1359) value out of all three models, capturing the data's variance best (Table 1). However, the Neural Networks MSE (11.5461) shows that the model predicts User Engagement values with more error than the LASSO or Random Forest Model (Table 1). Being the most complex of the three, there is still much improvement for potential future models and testing utilizing Neural Networks for predicting social media User Engagement.

CONCLUSION

This study showcased the comparative analysis of LASSO Regression, neural networks, and random forests for predicting user post engagement on Instagram, highlighting distinct strengths and limitations for each model. With its interpretability, the Neural Network identified key predictors, demonstrating superior performance when capturing variance in the data, but it substantially overfits or underfits when predicting actual values. The Random Forest captured non-linear patterns and interactions, more effectively predicting values, but still overestimating them. LASSO Regression balances the two models, offering power in capturing variance and more resilience to overestimating values.

Future research could explore hybrid approaches or fine-tune these models to optimize accuracy and scalability for dynamic social media environments. It

could also use larger datasets by primarily looking at archival statistics and third-party analytical software, which are more ethical. More research is needed to improve these models so social media creators and developers can utilize the proper tools to improve their posts or platforms.

FUNDING SOURCES

None.

DECLARATION OF CONFLICT OF INTEREST

None.

REFERENCES

1. Eckstein, M. (2024, January 29). Social Media Engagement: Why It's Important and How to Do It Well - the Buffer Blog. Buffer Library. https://buffer.com/library/social-media-engagement/(accessed on 2024-10-13)
2. Emmert-Streib, F., & Dehmer, M. (2019). High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Machine Learning and Knowledge Extraction*, *1*(1), 359–383. https://doi.org/10.3390/make1010021(accessed on 2024-10-15)
3. GeeksforGeeks. (2019, January 17). *Neural Networks | A beginners guide*. GeeksforGeeks. https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/(accessed on 2024-10-20)
4. Yiu, T. (2019, June 12). *Understanding Random Forest*. Medium; Towards Data Science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2(accessed on 2024-10-20)
5. Zote, J. (2024, September 26). Instagram engagement rate: How to calculate yours in 2024. Sprout Social. https://sproutsocial.com/insights/instagram-engagement-rate/(accessed on 2024-10-22)