Classification of Instagram Users and Prediction of Engagement Rates Using Machine Learning

Vaishnavi Padiya[1], Dev Shah[2], Tanay Dave[3], Ruhina Karani[4]
Dept of Computer Engineering, Dwarkadas J Sanghvi College of Engineering, Mumbai, Maharashtra, India [2]

Abstract. In the rapidly evolving landscape of social media, understanding user behaviour and predicting engagement rates are crucial for optimizing content strategies and enhancing user experience. The paper represents a comprehensive approach toward classification of Instagram users and the prediction of engagement rates by state-of-the-art machine learning techniques. High variance in engagement rates from fake and inactive users mostly acts as a barrier to the learning process of the model. In this regard, we proposed a new approach to reduce these variations in order to enhance the performance of the model. Two-class classification, followed by a regression analysis. First of all, classification into real and fake accounts; another one is into active and inactive accounts. Further, the prediction of engagement rate was done only on the shortlisted real and active users; and the best performing model—MSE, that turned out to be 148.6313—was obtained from XGBoost. To add to this, a performance in the classifier—84.88% accuracy achieved by CatBoost in the real and fake account classification—and a Voting Classifier turning in an accuracy of 96.66% in active and inactive account classification. These results were useful for the multi-stage approach by using the strengths of various machine learning models to obtain strong results in classification and prediction. This study contributed to the growing field of social media analytics by providing insights into user classification and engagement prediction, with potential applications in marketing, content curation, and user retention strategies.

Keywords: Classification; Engagement Rate; Prediction ; Instagram ; Machine Learning.

1 Introduction

The proliferation of social media has had huge effects on how the world, especially individuals and organizations, interacts with the digital world. Among all these social media platforms, Instagram stands out as one of the largest and most influential, with millions of users worldwide. However, there lies a challenge of differentiating between fake and real Instagram users. Fake accounts alter the engagement metrics, misguide users, and thematize the whole integrity of the platform. To overcome these challenges our research aims to classify Instagram users and predict their engagement rate using advanced machine-learning techniques. Datasets from the work of Kristo Radion Purba, David Asirvatham, and Raja Kumar Murugesan [1] were used in this research and evaluation of models that can identify a real user apart from a fake one and alternatively an active account from an inactive one. Moreover, we investigate predictions of different engagement rates for different types of users to provide a detailed study of user behavior on Instagram and also a prediction of their engagement rates. The approach involved a multi-step methodology: first, the classification of the user on Instagram as real or fake; then, active fake users are differentiated from the inactive ones; finally, the engagement rates are predicted for users classified as real and active. The models were trained and cross-validated over such datasets, which differed with respect to the set of features used as input: post frequency, follower count, bio length, and whether the profile picture is available or not, among others. The accuracy of each model was carried out with rigorous checking with such metrics of effectiveness Mean Squared Error(MSE). The results of this research augment the theoretical body of social media analytics related to the development of models for user classification or any such engagement prediction on Instagram. Enhanced models better our nuanced understanding of user behavior and offer meaningful guidelines for keeping the integrity of the platform and optimization strategies for better user engagement.

2 Literature Review

Deep analysis of user behaviour and post performance on Instagram was done by Chang and Steve Gemine, pointing out the main elements that enhance user engagement [2]. Their work built the foundation for Purba, Asirvatham, and Murugesan, who further expounded this by predicting post popularity using machine learning, incorporating features like hashtags, image assessment, and user history with high accuracy [3]. Qian et al. conducted research on the effects of image aesthetics and dual-attention models on post popularity [4]; Zhang et al. considered factors related to the content that create variance in the visibility of a post [5]. This stream was further developed by Carta et al., who used Gradient Boosting for the prediction of post visibility, already very close to our approach of using sophisticated algorithms for classification and prediction tasks [6]. Abinowi and Aminudin used the Apriori method to optimize Instagram advertisement strategy respecting peak times of users' interest and top categories [7]. Rajeswari et al. tried to maximize visibility on Instagram with the help of machine learning [8]. Purba et al. proposed a new popularity measure "outsiders percentage" based on Random Forest regression [9]. Chaitanya et al. proposed further improvement by using dynamic feature weighting to update the models according to changes in user behaviour [10]. In fake account detection, Purba et al.[1] employed SVM and logistic regression with random forests to acquire high rates of precision and recall. Akyon and Kalfaoglu, and Bhat and Jain, have demonstrated that the use of newer datasets and advanced techniques significantly enhances the fake account detection [11][12]. Although extremely popular in the area of influencer marketing, there is still scant research available on this theme, hence allowing room for our research that aims to improve user engagement strategies with engagement rate prediction for real and active users.

3 Dataset

The datasets used in this research work form the basis of the work by Kristo Radion

Purba, David Asirvatham, and Raja Kumar Murugesan, titled "Classification of Instagram Fake Users Using Supervised Machine Learning Algorithms" 2020 [1]. There have been majorly two datasets used, serving two different purposes:

Dataset 1: Real and Fake Users It had two classes of users: real and fake. As such, this dataset was particularly created to train a classification model that can efficiently differentiate between the classes of real and fake Instagram accounts. The dataset had one primary goal - to create a baseline that efficiently separates real and fake classes of users by implementing a series of machine learning algorithms to attain accurate and reliable classification results.

Dataset 2: Mixed User Types The second dataset consisted of four different classes of users: real, active fake, inactive fake, and spam fake. This dataset is further divided into two subsets, both targeting different problems related to classification and prediction.

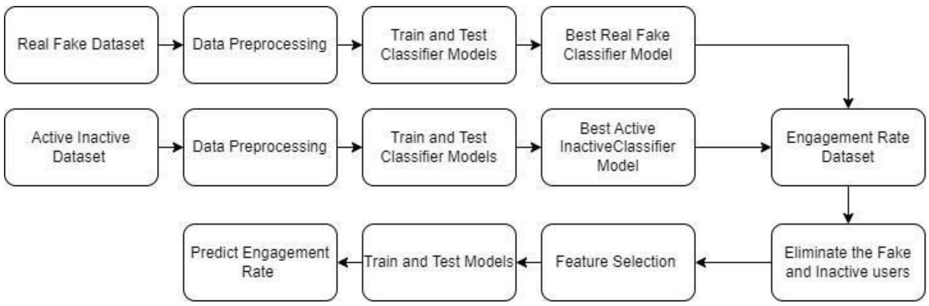

Fig.1 Methodology

Subset 1: Active and Inactive Fake Users This was a subset focused on fake users alone, whose purpose was to segregate active and inactive fake accounts. It is used for training a classification model to classify the various activity. levels among fake accounts. User patterns of engagement and behavior formed a very vital part of this subset.
Subset 2: Spam and Real Users This subgroup includes spams and real users that constituted the test data for evaluating the overall performance of the classification models.
This would therefore establish the efficacy of the models with respect to differentiating regular users from a myriad of classes of spam accounts, thus validating the strength of the classification approach. These datasets could give a deep, full insight into the behaviour of Instagram users. They have made possible the further application of machine learning techniques to improve the accuracy in classifying users and make predictions, hence providing a granular understanding of social media dynamics and user engagement. Variables from the datasets included a number of features, including the total number of posts, the number of accounts followed, and the total number of followers, the length of the bio, the existence of a profile picture represented by a binary value, whether the account had an external URL in the profile, the average caption

length of all posts, and the percentage of captions containing three or fewer characters. The dataset also captured the percentage of non-image media posts (ni), the percentage of posts tagged with a location (lt), the average number of hashtags used in posts (hc), the average usage of promotional keywords (pr) and follower-hunting keywords (fo) in hashtags, the average cosine similarity between pairs of posts (cs), and the average interval between posts in hours (pi). Engagement rates were calculated separately for likes (erl) and comments (erc), and these engagement rates were summed to derive a combined engagement rate for prediction purposes. The dataset also included the following labels: 2-Class User Labels: r (real/authentic user), f (fake user/bought followers) 4-Class User Labels: r (real/authentic user), a (active fake user), i (inactive fake user), s (spammer fake user) These features and labels were crucial in enabling a detailed and accurate classification of Instagram users, as well as in predicting engagement rates with higher precision.

4 Methodology

The work had two developments: the first one developed two various classifiers that could separate a fake from a real Instagram account and an active from an inactive one, and the second one used both, relying on a partial set of models to predict engagement rates for real and active accounts. This approach is designed to consider that fake and inactive accounts often exhibit high variability in engagement rates, which will, therefore, impact the model's performance. On the other hand, a real and active account usually has a more stable rate of engagement that gives a firmer basis for the right prediction. The methodology is shown in Fig. 1 The engagement rate variations for both real and active accounts, as well as fake and inactive accounts, were analyzed and are depicted in Fig 2 and Fig 3 . As illustrated in Fig 3, the engagement rates of inactive and fake users exhibited considerable variability, which posed challenges for accurate prediction. Consequently, the methodology adopted in this study focused exclusively on real and active users for engagement rate prediction.

The engagement rate variations for both real and active accounts, as well as fake and inactive accounts, were analysed and are depicted in Fig 2 and Fig 3 . As illustrated in Fig 3, the engagement rates of inactive and fake users exhibited considerable variability, which posed challenges for accurate prediction. Consequently, the methodology adopted in this study focused exclusively on real and active users for engagement rate prediction.

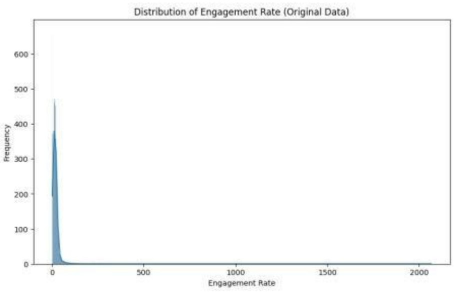

Fig.2 Real and Active Users.

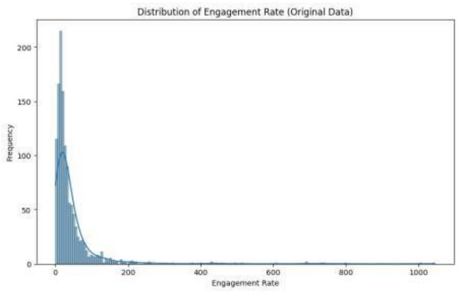

Fig.3 Fake and Inactive Users.

4.1 Real-Fake Instagram Account Classification

The first step in our methodology involved training a classification model to distinguish between real and fake users. For this, we utilized Dataset 1, which contained labels for real and fake users. During preprocessing, we converted the labels: 'r' for real users was transformed to 0, and 'f' for fake users was converted to 1. Various features such as the number of posts, number of followers, biography length, and picture availability were selected for training the model. Several classifiers were employed to enhance the detection of real versus fake accounts. Gradient Boosting, Random Forest, and Extra Trees are chosen for their ability to capture really complex, nonlinear relationships among features. This will make them very efficient with regard to the recognition of tiny patterns. Gradient Boosting is used due to its iterative approach toward leveraging accuracy, especially for class-imbalanced datasets. Other advanced boosting methods applied were XGBoost, LightGBM, and CatBoost. Each of them brings their own different strengths to bear: regularization, efficiency on large data, and robust management of categorical features. All models were benchmarked against metrics like Training and Testing accuracynly two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

4.2 Active-Inactive Instagram Account Classification

The second step was to further distinguish between the active and inactive fake users with another subset of Dataset 2. In the subset, there are only active fake and inactive fake users. In the preprocessing phase, we converted the labels where 'i' for inactive users was converted into 1 and 'a' for the active user was converted to 0. Features that measure the activity of every user, such as posting frequency, interaction with their followers, and recent activities, were used while training the model. For robust classification, various supervised machine-learning algorithms have been employed to model the binary outcome of user activity status based on their strength in linear classification. The methods chosen are Stochastic Gradient Descent and Passive Aggressive; these were selected because they are efficient to deal with large datasets and robust to changes in real-time user behaviour. Quadratic Discriminant Analysis was also able to capture probable nonlinear trends in user activity. The Voting Classifier aggregated predictions from different models to improve accuracy. Naive Bayes was also employed for being simple and having efficiency in treating categorical data commonly existing in user behavioural analyses. All of them combined helped distinguish active from inactive users by learning from past interaction data. These developed models were then compared based on different metrics, such as accuracy during training and testing.

4.3 Engagement Rate Prediction

In the final stage of the methodology, the prediction of the engagement rate was done for users who were classified as real and active in the previous classification stages. Feed the classification models with the test dataset consisting of both spam and real users that were appropriately scaled and transformed during preprocessing. From the results, select only the users identified as real and active for analysis of the engagement rate. Various feature selection methodologies were applied, including the correlation matrix and Correlation Coefficient, for their strength in identifying linear relationships and dealing with multi-collinearity. While that was not enough, the Variance Inflation Factor also ensured the independence of features, and neural networks with permutation importance captured complex, nonlinear relationships key to social media data. LassoCV provided feature selection and regularization to prevent overfitting, while Recursive Feature Elimination further fine-tuned the feature set down to just the most predictive variables. Mutual information was used since it can capture both linear and nonlinear dependencies, hence ideal in the intricate interactions of social media data. Table 1 shows the various features that each method selected. There was a collection of regression models to enhance robustness in prediction. ElasticNet was included; regularization is for avoiding overfitting, especially in high dimensions. Support Vector Regression (SVR) effectively captured complex relationships, Gradient Boosting enhanced accuracy through iterative learning, CatBoost was particularly used because of its effectiveness in handling categorical data and reducing overfitting, while LightGBM was valued for its speed and scalability, making it ideal for large datasets. XGBoost was employed for its robustness and regularization techniques, which improved model generalization. These models efficiently captured patterns in the data to predict engagement rates by leveraging their gradient-boosting capabilities, making them well-suited for regression tasks in social media analytics, making each of these models well-suited for predicting Instagram engagement rates in a complex, multi-feature environment. The model performance was assessed using regression metrics such as mean squared error(MSE).

Table 1. . Selected Features by Various Feature Selection Techniques.

Feature Selection Method	Selected Features
Correlation Matrix	pi, flg, pos, cs, hc, pic, fo, pr, ni, flw, cz, cl, bl
VIF	flw, pic, pos, ni, pi, pr, fo, lt, hc, flg, lin, bl, cl
Random Forest	pi, fo, flw, lin, pos, flg, cl, hc, bl, cs, ni, lt, cz
Neural Networks	pos, flw, pi, bl, lin, cs, fo, ni, hc, lt, pic, pr, cz
LassoCV	pos, flw, flg, bl, pic, lin, cl, cz, ni, lt, hc, pr, fo
Correlation Coefficient	pos, flg, lin, cs, pi
Recursive Feature Elimination	pic, lin, cz, ni, lt, hc, pr, fo, cs, pi
Mutual Information	lin, pos, flg, flw, pi, cs, bl, cl, lt, ni, cz, hc, fo

5 Results

5.1 Real-Fake Instagram Account Classifier performance

The Training and Testing Accuracies for different classifiers were obtained for the Real Fake Account Detection as shown in Table 2. The results for the real versus fake Instagram account classification showed that while most models exhibited high accuracy on the training data, their performance on the test data reveals varying degrees of generalization. The Random Forest and Extra Trees models achieved near-perfect accuracy on the training set (99.99%), but their test accuracies dropped significantly to around 83%, indicating potential overfitting due to model's complexity.To address, Gradient Boosting, XGBoost, LightGBM, and CatBoost models were used which = showed a more balanced performance with better generalization across unseen data, with training accuracies ranging from 84.20% to 89.46% and test accuracies between 83.47% and 84.88%. Among these, CatBoost achieved the highest test accuracy of 84.88%, suggesting it may offer the best balance between learning from the training data and generalizing to unseen data in this classification task.

5.2 Active-Inactive Instagram Account Classifier performance

The classification models chosen to distinguish between active and inactive fake Instagram users were evaluated as shown in Table 3 The results for the active versus inactive Instagram user classification indicate that most models performed well, with high accuracy on both the training and test datasets. The Voting Classifier emerged as the top performer, achieving the highest test accuracy of 96.66%, closely aligning with its training accuracy of 96.98%, suggesting strong generalization. Models like the SGD Classifier, Passive Aggressive, and Quadratic Discriminant.

Table 2. Results for Real vs. Fake Classifier.

Model	Train Accuracy	Test Accuracy
Random Forest	99.99	83.98
Gradient Boosting	84.20	83.47
XGBoost	89.46	84.23
LightGBM	86.17	84.49
CatBoost	88.03	84.88
Extra Trees	99.99	83.16

Analysis also demonstrated balanced performance, with test accuracies around 9495%, indicating reliable classification. Logistic Regression, SVC, and Na¨ıve Bayes performed slightly lower, with test accuracies in the range of 92-93%, still reflecting good, consistent performance across training and testing data.

Table 3. . Results for Active vs. Inactive Classifier.

Model	Train Accuracy	Test Accuracy
Logistic Regression	93.60	93.50
SVC (Support Vector Classifier)	93.24	92.57
SGD Classifier	95.24	94.49
Passive Aggressive	94.48	94.40
Quadratic Discriminant Analysis	94.54	94.56
Voting Classifier	96.98	96.66
Naive Bayes	92.93	92.66

5.3 Engagement Rate Prediction performance

The proliferation of social media has had huge effects on how the world, especially individuals and organizations, interacts with the digital world. Among all these social media platforms, The evaluation of various regression models for predicting Instagram engagement rates provided insights into their performance are shown in Table 4. It showed varying levels of performance, with different regression techniques achieving different Mean Squared Error (MSE) values. ElasticNet, which combines the regularization techniques of Ridge and Lasso, resulted in an MSE of 200.9306, providing balanced regularization but was outperformed by other models. Support Vector Regression (SVR) had a slightly better MSE of 199.8865 by using a margin to focus on key data points. Gradient Boosting performed significantly better with an MSE of 142.2369, as it built sequential decision trees to correct errors iteratively. XGBoost, a refined boosting algorithm with additional regularization, achieved a competitive MSE of 148.6313, ensuring better model performance and generalization. CatBoost, which is particularly effective with categorical features and overfitting prevention, had an MSE of 177.4990, while LightGBM, known for its efficiency and speed in boosting, had a slightly lower MSE of 174.4335. Overall, Gradient Boosting demonstrated superior predictive performance in capturing engagement rate patterns

5.4 Model Testing

The Engagement rate Prediction model was then tested for the input features of an unseen user, the input features are shown in Table 5. These features were first scaled and tested on the Real Fake Classifier Model and Scaler and then separately on the Active Inactive Classifier Model and scaler, the results for the classification were Real and Active respectively and hence it was preceded by the Engagement Rate prediction which resulted in an Engagement Rate of 15.60.

Table 4. . Comparison of Models with their MSE, Descriptions, and Formulas.

Model	MSE	Description	Formula
ElasticNet	200.9306	Combines penalties of Ridge and Lasso for balanced regularization.	$\min \sum_{i=1}^{n} y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}^2 + \lambda_1 \sum_{j=1}^{p} \lvert\beta_j\rvert + \lambda_2 \sum_{j=1}^{p} \beta_j^2$
SVR	199.8865	Uses a margin to fit data, focusing on key points.	$\min \frac{1}{n} \sum_{i=1}^{n} \max(0, \lvert y_i - \hat{y}_i\rvert - \epsilon) + \lambda \lVert\beta\rVert^2$
Gradient Boosting	142.2369	Sequentially builds trees to correct errors.	$\min \sum_{i=1}^{n} y_i - \sum_{m=1}^{M} \nu f_m(x_i)^2$
XGBoost	148.6313	Boosting algorithm with regularization for optimal performance.	$\min \sum_{i=1}^{n} y_i - \sum_{m=1}^{M} f_m(x_i)^2 + \sum_{m=1}^{M} \Omega(f_m)$ $\Omega(f) = \gamma T + \frac{1}{2}\lambda \lVert w\rVert^2$
CatBoost	177.4990	Efficiently handles categorical features, reduces overfitting.	$\min \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \lambda\Omega(w)$ where L is a loss function and Ω is regularization.
LightGBM	174.4335	Known for efficiency and speed in gradient boosting.	$\min \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \lambda\Omega(w)$

Table 5. Engagement Rate prediction on unseen user

pos	flw	flg	bl	pic	lin	cl	cz	ni	lt	hc	pr	fo	cs	pi
125	2800	319	141	1	0	120	0	0.167	0.77	0	0	0	0.057	1349

Result: Predicted Engagement Rate: 15.60 (Real and Active)

6 Future Scope

Future work can expand on this study by incorporating more diverse datasets from various social media platforms to generalize the classification and prediction models. Enhancing feature engineering techniques and exploring additional features like user interaction patterns and content analysis could improve model accuracy. Integrating advanced deep learning techniques and ensemble methods may provide further performance improvements. Additionally, developing real-time classification and prediction systems could offer practical applications for social media platforms and marketers. Ethical considerations, such as privacy and data security, should also be addressed in future research to ensure responsible use of these technologies. Expanding the scope of the study to include longitudinal analysis can provide insights into how user behaviour evolves over time. This would help in identifying long-term trends and patterns that are not immediately apparent in short-term data. Moreover, collaborating with industry stakeholders can lead to the development of tailored solutions that address specific business needs, enhancing the practical relevance and impact of the research.

7 Conclusion

This research is concluded by demonstrating a successful methodology for predicting the Instagram engagement rate through the integration of classification and regression techniques in a multi-stage approach. The classification of users at the very initial stage itself into real versus fake and active versus inactive handled effectively the challenges posed by high variability in engagement rates among inauthentic and inactive users. By implementing a multi-stage classification approach the research reduced the noise and variability in the dataset , thereby minimizing overfitting and improving robustness of engagement rate predictions. This system drastically improved the prediction of the rate of engagement filtering for real active users; XGBoost had the lowest Mean Squared Error of 148.6313. The classification models worked smoothly as well: CatBoost got 84.88% accuracy in classification between real and fake accounts, and a Voting Classifier reached 96.66% in the classification between active and inactive users. This methodology emphasized the importance of using sophisticated machine learning models to enhance predictions, which yielded valuable insights concerning user behavior on social media sites. Its findings bear implications for the optimization of marketing strategies, enhancing content delivery, and improving user engagement in the emerging field of social media analytics. In conclusion, this paper contributes valuable knowledge for the field of social media analytics and delivers a practical instrument for the improvement of online community management.

References

1. Purba, K. R., Asirvatham, D., & Murugesan, R. K. (2020). ”Classification of instagram fake users using supervised machine learning algorithms”. International Journal of Electrical and Computer Engineering, 10(3), 2763.
2. Chang, S. (2016). Instagram post data analysis. arXiv preprint arXiv:1610.02445.
3. Purba, K. R., Asirvatham, D., & Murugesan, R. K. (2021). Instagram post popularity trend analysis and prediction using hashtag, image assessment, and user history features. Int. Arab J. Inf. Technol., 18(1), 85-94.
4. Qian, C. J., Tang, J. D., Penza, M. A., & Ferri, C. M. (2017). Instagram popularity prediction via neural networks and regression analysis. IEEE Transactions on Multimedia, 19, 25612570.
5. Zhang, Z., Chen, T., Zhou, Z., Li, J., & Luo, J. (2018, December). How to become instagram famous: Post popularity prediction with dualattention. In 2018 IEEE international conference on big data (big data) (pp. 2383-2392). IEEE.
6. Carta, S., Podda, A. S., Recupero, D. R., Saia, R., & Usai, G. (2020). Popularity prediction of instagram posts. Information, 11(9), 453.
7. Abinowi, E. (2020). Analysis of Instagram posting for marketing using apriori method. PalArch’s Journal of Archaeology of Egypt/Egyptology, 17(10), 3094-3101.
8. Rajeswari, B., Sahithi, M. P., Harika, M. N. L., Alekhya, K. S. R., & Sathwika, M. (2024). Boosting Instagram Visibility with Machine Learning.
9. Purba, K. R., Asirvatham, D., & Murugesan, R. K. (2020). An analysis and prediction model of outsiders percentage as a new popularity metric on Instagram. ICT Express, 6(3), 243248.
10. Viswanatha Reddy, G., Chaitanya, B. S. N. V., Prathyush, P., Sumanth, M., Mrinalini, C., Dileep Kumar, P., & Mukherjee, S. (2024). DFW-PP: dynamic feature weighting-based popularity prediction for social media content. The Journal of Supercomputing, 80(4), 57085730.
11. Akyon, F. C., & Kalfaoglu, M. E. (2019, October). Instagram fake and automated account detection. In 2019 Innovations in intelligent systems and applications conference (ASYU) (pp. 1-7). IEEE.
12. Bhat, S. V., & Jain, N. N. Fake Account Detection on Instagram.