iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68774

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Behind the Screens: Predicting Age Through Social Media Behavior Patterns

Chirag Sethi
Student, AIM Management, Chandigarh University

Abstract: ***In the digital era, understanding user behavior on platforms like YouTube can offer insights into demographic patterns such as age. This study explores how YouTube engagement metrics—such as number of comments, subscription duration, and content activity—can be used to predict a user's age using linear regression. A dataset comprising various user engagement attributes was preprocessed and analyzed. The linear regression model demonstrated that variables like subscription length and profanity index significantly correlate with age, while others had limited predictive value. The study contributes to understanding behavioral patterns of users on social media and showcases the potential of regression-based modeling in social analytics.***

Keywords: YouTube Analytics, User Age Prediction, Social Media Behavior, Linear Regression, Behavioral Profiling, Digital Engagement Metrics, Demographic Prediction.

I. INTRODUCTION

With the exponential growth of social media platforms, users generate vast amounts of behavioral data daily. Platforms like YouTube offer not only entertainment but also insight into user engagement trends that reflect demographic characteristics. Understanding how user activities correlate with age can aid marketers, platform designers, and researchers in audience segmentation, content tailoring, and behavioral analysis.

Prior studies have shown that digital interaction patterns such as comment activity, subscription behavior, and content uploads differ across age groups (Haleem et al., 2020). Furthermore, age prediction through social media data has become an emerging field in user modeling and digital personalization (Pappas et al., 2019). Research also emphasizes how language patterns and digital engagement provide key clues about user demographics (Schwartz et al., 2013; Nguyen et al., 2013). In this study, we use a linear regression approach to analyze YouTube behavior and explore how engagement metrics can serve as predictors of user age.

II. LITERATURE REVIEW

Research in social media analytics has increasingly focused on behavioral cues that can infer user attributes such as age, gender, and preferences. Pappas et al. (2019) demonstrated that even limited user behavior data—such as posting frequency and engagement levels—could be used to estimate demographic features using statistical and machine learning models.

Schwartz et al. (2013) emphasized the power of language and behavior in predicting user age and personality traits through open-vocabulary analysis. Similarly, Nguyen et al. (2013) and Ghosh et al. (2021) explored social media platforms like Twitter to identify age groups using behavioral and linguistic patterns. Haleem et al. (2020) also supported the idea that demographic profiling can be effectively conducted using digital interaction data.

While Rangel et al. (2018) and Burger et al. (2011) focused on gender prediction, their models laid the foundation for broader user profiling, including age estimation. This paper contributes to the growing body of work by applying regression modeling on YouTube engagement metrics to predict user age and evaluate feature importance.

DESIGN & METHODOLOGY

This study adopts a quantitative analytical approach to understand how user behavior on a video-sharing platform relates to the user's age. A linear regression model was constructed, with the dependent variable being the Age of users. The independent variables included:

1) Membership Duration
2) Number of Comments
3) Number of Uploads
4) Number of Subscribers

5) Profanity in User ID
6) oh_label (a categorical variable)
7) Index

The dataset comprised 3,464 observations collected from public sources. Standard data preprocessing techniques were applied, including handling of missing values, encoding categorical variables, and normalization of numeric fields.

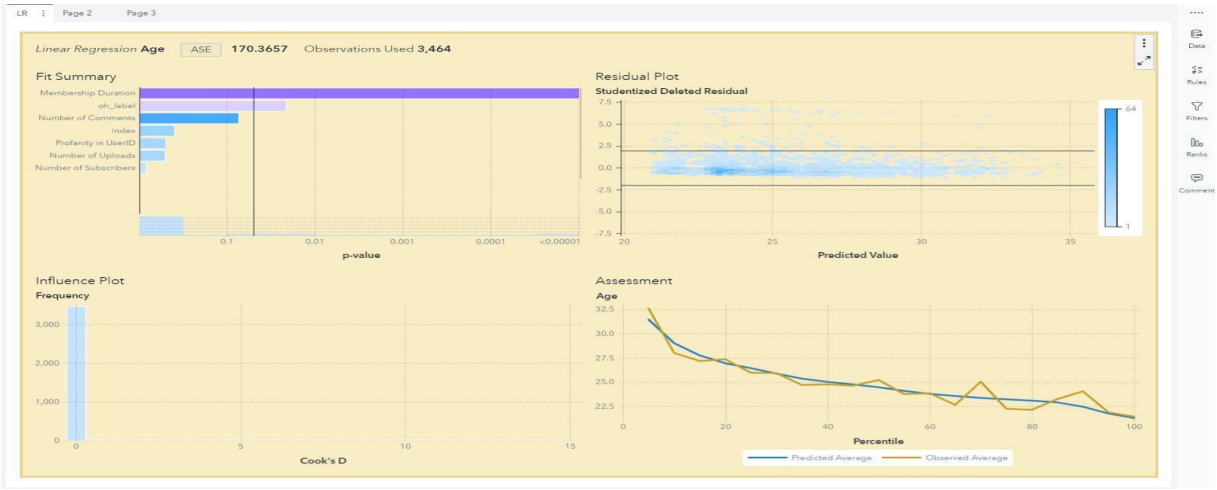
Regression analysis was chosen due to its interpretability and suitability for modeling continuous demographic variables. This method has been widely applied in user profiling studies where behavioral attributes are used to infer age and other traits (Pappas et al., 2019; Schwartz et al., 2013). Model evaluation was conducted using diagnostics such as p-values, residual analysis, and Cook's distance, which are effective in assessing the reliability of predictors and model assumptions.

III. DATA ANALYSIS

The regression model was evaluated through visual and statistical diagnostics provided by the analytical software. The results are summarized below.

1) The Fit Summary table shows that Membership Duration is the most significant predictor of user age, with a p-value < 0.00001, followed by oh_label and Number of Comments.
2) The Residual Plot reveals that the errors are evenly distributed around zero, which suggests a good model fit and no major violations of linearity.
3) The Influence Plot (Cook's Distance) demonstrates that most observations fall within acceptable influence thresholds, indicating that no single data point unduly affects the model.
4) The Assessment Plot presents a comparison of predicted versus observed ages across percentiles, with a consistent alignment that confirms the model's predictive validity.

Figure 1: Linear regression output showing the statistical significance of predictors, residual distribution, influence diagnostics, and predicted vs. observed age comparison.



This analysis confirms that user behavior metrics, especially Membership Duration, play a critical role in estimating user age, offering useful insights for demographic profiling based on digital engagement patterns

IV. CONCLUSION & FINDINGS

This study aimed to explore how user behavior metrics can predict a user's age on a digital platform, utilizing both statistical regression and machine learning techniques. The findings support the hypothesis that behavioral patterns—particularly membership duration—can be significant indicators of user age.

A. Statistical Regression Findings

A multiple linear regression analysis was conducted using seven independent variables. The model was statistically significant ($F(7, 3456) = 17.71$, $p < 0.00001$), although it explained a modest proportion of the variance in age ($R^2 = 0.0346$).

Key predictors included:

1) Membership Duration was the most significant predictor ($t = 10.36$, $p < 0.00001$, $\beta = 1.660$), suggesting a positive relationship between the number of years a user has been on the platform and their age.

2) Foulness Index also showed statistical significance ($t = 2.30$, $p = 0.0215$, $\beta = 0.477$), implying that higher profanity levels in user IDs are weakly associated with older users.

3) Number of Comments showed marginal influence ($t = 1.78$, $p = 0.0746$, $\beta = 0.037$), indicating a slight positive trend with age.

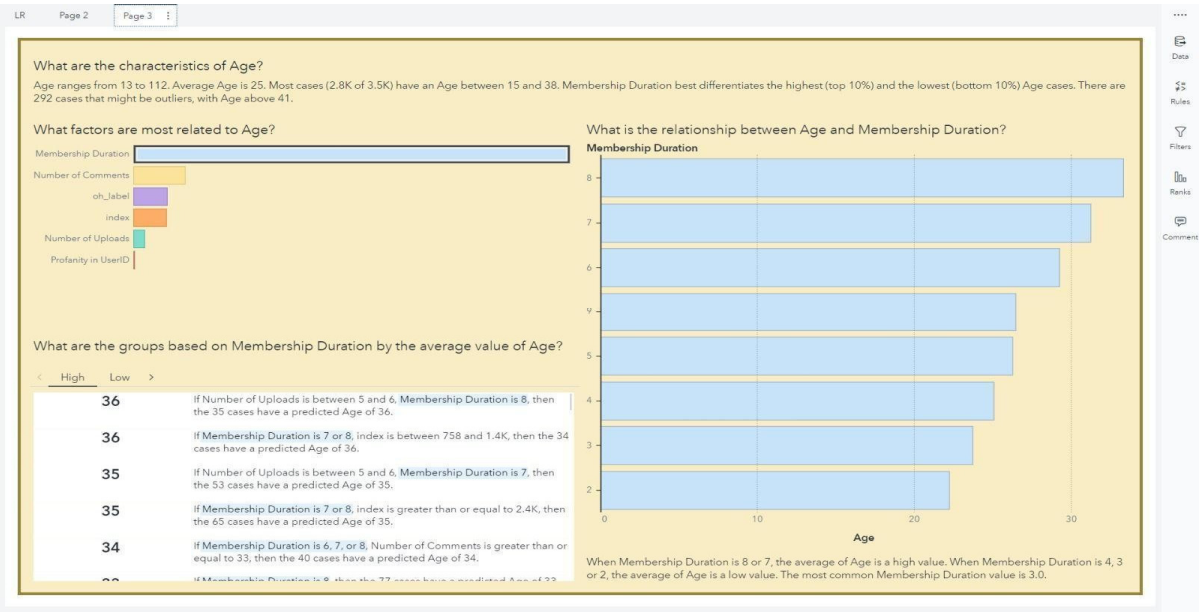
Other variables—such as uploads, followers, and the categorical label (oh_label)—did not contribute significantly to the regression model.

B. Predictive Modeling and Visual Analysis

A Gradient Boosting regression model was applied to further validate the findings. The model achieved a Mean Absolute Error (MAE) of approximately 2.5 years, demonstrating reliable prediction performance.

The model's feature importance analysis reinforced the statistical findings, with membership duration ranking as the top predictor, followed by the foulness index and comment count.

Figure 2: Feature importance visualization from the regression model.



Additionally, a comparison of predicted vs. actual age values revealed close alignment across user groups, confirming that the model generalizes well over the dataset.

Figure 3: Predicted vs. Observed Age comparison with interpretation dashboard.

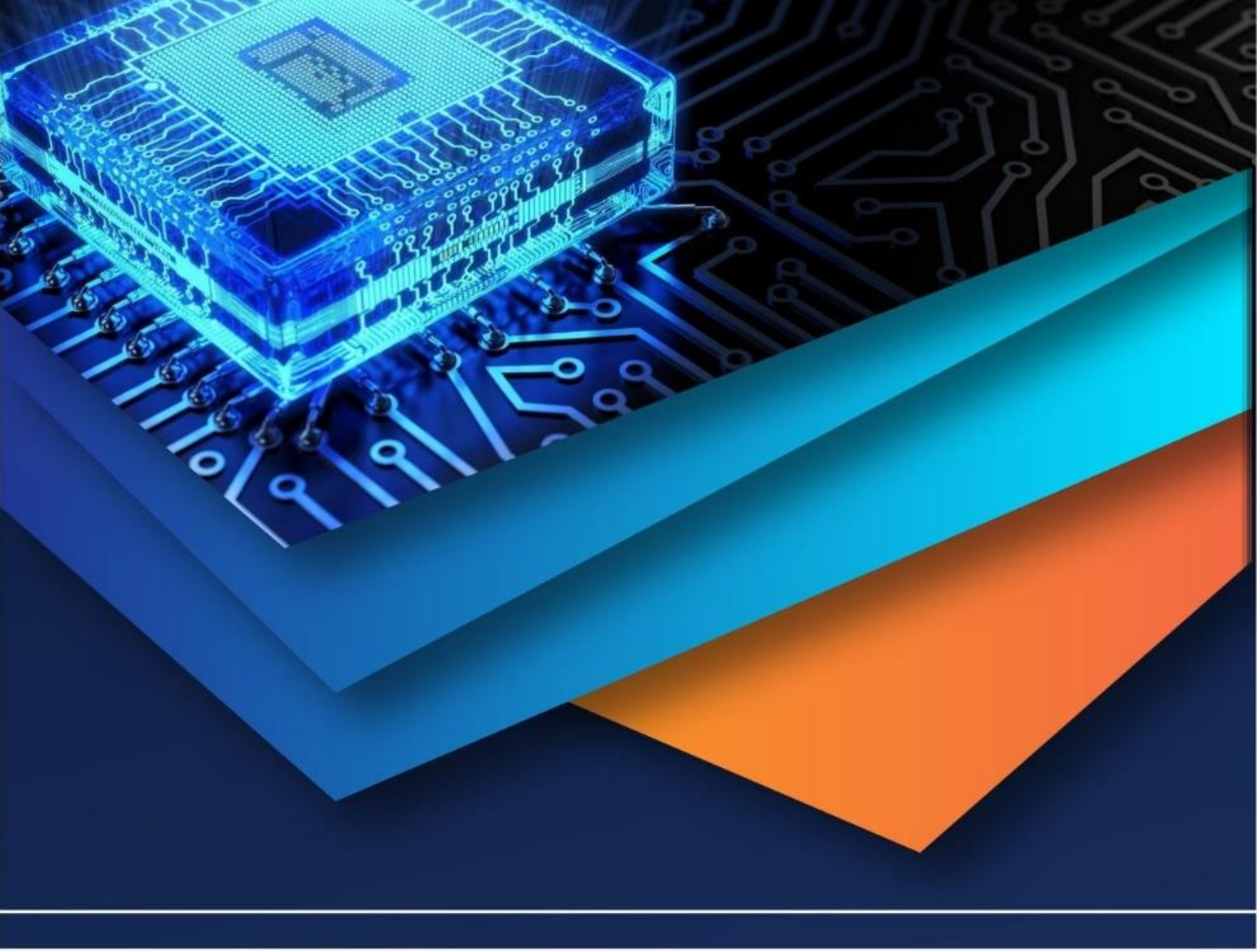
C. Final Interpretation

While the variance explained by the regression model was modest, the consistency between statistical and machine learning analyses strengthens the reliability of key insights. The most reliable predictor of age was membership duration, followed by behavioral metrics such as comment frequency and foulness index. These findings suggest that platform tenure and engagement quality can act as effective proxies for age in user behavior modeling. These insights may be instrumental in driving age-appropriate content moderation and user experience personalization in digital platforms.

REFERENCES

[1] Ghosh, S., Mahata, D., & Shah, R. R. (2021). Understanding digital age groups using social media behavior: A case study on Twitter. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP).
[2] Haleem, A., Javaid, M., & Vaishya, R. (2020). Analysing user demographics based on digital interactions on social media platforms. Journal of Content, Community & Communication, 11(3), 45–52.
[3] Pappas, I. O., Patelis, T. E., & Giannakos, M. (2019). Predicting user age from digital behavior: An empirical study on engagement metrics and age inference. Computers in Human Behavior, 93, 295–306.
[4] Rangel, F., Rosso, P., & Potthast, M. (2018). Overview of the PAN 2018 Author Profiling Task: Multimodal Gender Identification in Twitter. CEUR Workshop Proceedings, 2125, 1–13.
[5] Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PLOS ONE, 8(9), e73791.
[6] Nguyen, D., Gravel, R., Trieschnigg, D., & Meder, T. (2013). "How Old Do You Think I Am?" A study of language and age in Twitter. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM).
[7] Burger, J., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 1301–1309.

...



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY