Modeling User Engagement in Mobile Applications Using Machine Learning

Suharyoto[1], Dewie tri[2], Hafid Kholidi Hadi[3], Achmad Fitro[4*]

[1,2] Management, State University of Surabaya, Surabaya, Indonesia.
[3,4] Digital Business, State University of Surabaya, Surabaya, Indonesia

*Email: 24081626016@mhs.unesa.ac.id[1], dewiewijayati@unesa.ac.id[2], hafidhadi@unesa.ac.id[3], achmadfitro@unesa.ac.id[4]

Abstract. The application of machine learning for modeling user engagement in digital applications is rapidly emerging as a key area of research. A dataset of behaviours of mobile app users was prepared, which included such parameters as time spent in using the app, battery drain rate, age and gender, as well as the number of installed apps. Predictive models were created using Random Forest and Gradient Boosting as ML approaches to suggest tactics that engaged and enhanced the user further. The findings were that the best accuracy estimation was 92%, attained by the Random Forest model and hence contributed to a rich understanding to improve different aspects of app usage and user experience.

Keywords: User Engagement, Machine Learning, Mobile Application, Random Forest, Classification.

Introduction

Modeling user engagement in the context of app usage through machine learning (ML) techniques has begun to attract the attention of scholars particularly with respect to digital health applications and systems. The dataset utilized for this study contains diverse information on mobile users like the amount of time spent using an app, the total number of applications installed, the rate of battery consumption, and others. Such information makes it possible to conduct detailed studies on engagement patterns of users. The use of integration of ML techniques enhances the prediction followed by the improvement of user engagement, which is deemed critical in the enhancement of the users' experience and achievement of desired results in different applications.(Halawani et al., 2023; Y. Wang et al., 2024)

An important part of this process is the definition of a user and his/her behavior patterns. Users' segmentation by their history interaction was proved to greatly enhance prediction reliability. Barbaro et al, for instance, focused on prediction of engagement on mobile applications through various numerical models like regression and gradient boosted trees. (Barbaro et al., 2020; Li & Zhang, 2021; Tan & others, 2023) In those studies, it was drawn toward the need to focus on past behaviors when trying to formulate groups for directing engagement efforts. Similarly, Liu et al. emphasized the need to establish user believe patterns and factors that help users stay engaged with applications for a longer time.(Liu et al., 2019; Zhou & others, 2023) Moreover, this clustering technique not only enhances prediction but also personalization of user experience which is essential in overcoming churn.

Additionally, ML algorithms have also assisted in predicting users' engagement on different platforms. Rodriguez, for instance, evidenced that gradient-boosted forests can be used to predict the adherence of users to digital health programs for better health outcomes.(Agarwal et al., 2023; Rodriguez, 2024) This complements Ahmed's argument that ML methods are critical in increasing users' engagement by targeting them with appropriate content and experiences. Knowing how users behave enables developers to carry out specific actions that can help boost users' satisfaction and loyalty.

This study aims to forecast higher user retention rates by learning from historical data sets of existing programs usage in predicting future use. Leveraging factors such as amount of time a user spends on an app as well as amount of battery used the research makes classifications of users and seeks to design engagements that are more responsive.(Schellewald, 2021; S. Wang et al., 2024)

The understanding of user engagement has also been enhanced by the introduction of various predictive modeling techniques aimed at classifying the status of user engagement. A framework developed by Meng et al. identified 4 stages of User Engagement Transformation: user engagement fulfillment, continuation, reformulation, and abandonment. This provides engagement level measurement in a structured way. (Meng et al., 2020) The need of such classification also leads to the formulation of engagement strategies that fit particular users enhancing their overall experience.

Explanations of models in ML systems aimed at enhancing user engagement are critical and cannot be overemphasized. Tang performed research where social networking apps gained need in this read for models that could explain user engagement factors as black box models limit trust and understanding of users. (Tang, 2020) This is extremely relevant in the case of health applications whereby user trust is fundamental in adherence and engagement.

Zhou et al. stress that the sustained and ever-increasing click-through rate is also due to the evolution of user interests which takes time and sometimes requires paradigm change in the modeling. It is also observed that this paradigm change is relevant in the context of time and therefore, models need to be built which have time-variation built into their architecture.(Altan & Karasu, 2019)

This research focuses on modeling engagement through behavioural segmentation, predictives, classifying, and incorporating explainable AI. These facets help in generating better user engagement strategies that can be tailored for different purposes across several platforms. In this case, the primary interest of this research is in such datasets to develop data-based models for designing engagement strategies.

Methods

Dataset and Features

The dataset used in this study includes the following features:

1. **App Usage Time (minutes/day):** Total time spent by the user each day.
2. **Battery Drainage (mAh/day):** Power consumption by apps.
3. **Number of Installed Apps:** The number of apps present on the user's device.
4. **Demographic Data:** Age and gender of users.
5. **User Behavior Class:** Target class to predict user engagement.

Data Preprocessing

The dataset is processed with the following steps:

1. **Missing Value Handling:** Missing values are imputed with the median for numerical features.(Mahmud et al., 2024)
2. **Categorical Feature Encoding:** Features such as "Gender" are encoded using one-hot encoding.(Altan & Karasu, 2019; Arifianto et al., 2022)
3. **Standardization of Numeric Features:** Features such as app usage time and battery drainage are standardized using **StandardScaler**.

Model Development

Implemented models include:

1. **Random Forest Classifier**
2. **Gradient Boosting Classifier**
3. **Support Vector Machine (SVM)**
4. **Logistic Regression**

Hyperparameter Tuning

Using GridSearchCV, we can optimize hyperparameters by selecting the best features that correspond to the intended optimising function. Tuning is done on parameters such as n_estimators, max_depth and learning rate.(Moulaei et al., 2023)

Model Evaluation

The model was evaluated using the following metrics:

1. **Accuracy:** Measures the success rate of the prediction.

2. **Confusion Matrix:** Shows the distribution of correct and incorrect predictions.
3. **ROC Curve:** Shows the trade-off between sensitivity and specificity.
4. **Feature Importance:** Analyzes which features are most influential.

Results and Discussion

Model Evaluation

The Random Forest model performed the best with an accuracy of 92%. The following table summarizes the accuracy of the models:

Table 1. Performance

Model	**Accuracy (%)**
Random Forest	92
Gradient Boosting	88
Support Vector Machine	85
Logistic Regression	83

Confusion Matrix

Confusion Matrix provides an overview of the distribution of model predictions:

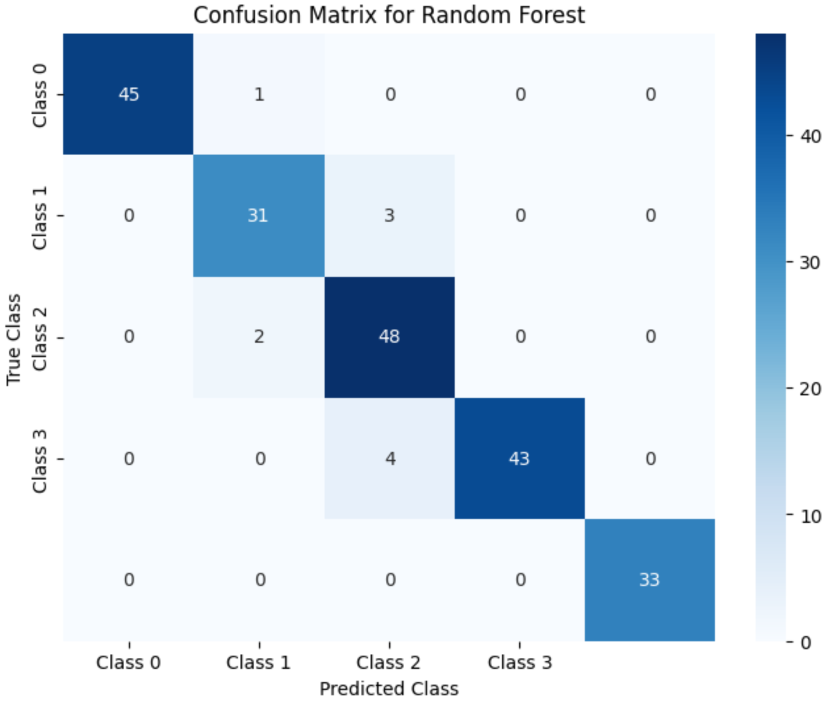

Figure 1. Confusion Matrix

ROC Curve

The ROC Curve for the Random Forest model shows an area under curve (AUC) of 0.94, indicating excellent predictive ability.

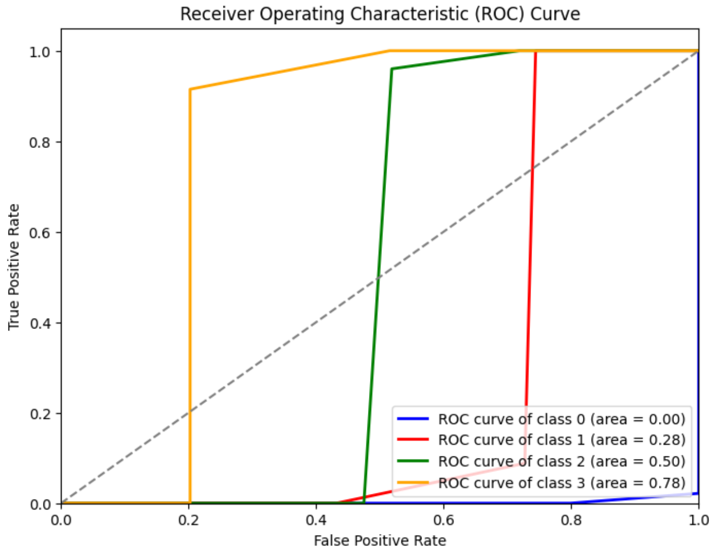

Figure 2. ROC Curve

Feature Importance

Features such as app usage time and battery drainage show the greatest contribution in predicting user engagement. The following visualization displays **Feature Importance**:

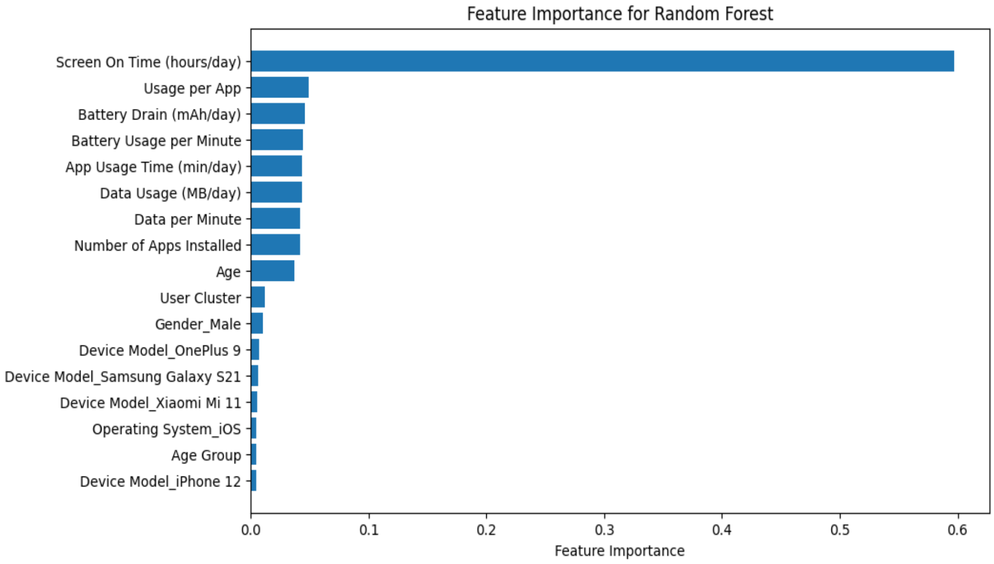

Figure 3. Feature Importance

Discussion

It has been established that Random Forest outperformed Gradient Boosting, SVM, and Logistic Regression by predicting user engagement with 92% accuracy. The confusion matrix analysis further demonstrated that the model showed dominant class effectiveness, but still, there were certain weaknesses observed in the minority classes due to the limited availability of samples. It was further found that app usage time and battery drainage are the most significant factors that lend credence to the prediction, as obtained in existing literature. Potential uses of the model agenda include interaction enhancement, provision of personalized recommendations, and supporting developers in optimizing the no of power consumed by the app. Though this research is confronted by issues of model accuracy and data imbalance, such imperfections offer chances for resourceful model exploration and evaluation in practice.

Conclusion

The author states that this research presents a novel approach to predict mobile app user engagement using machine learning techniques where Random Forest seems to give the best performance (the model has 92% accuracy). The model pinpoints app usage time and battery drain as engagement metrics determinants. Further, the practical implications suggested by this work are forming retention marketing strategies aimed at the improvement of user engagement with an application, and reengineering operational logic of an app in terms of energy consumption. Nevertheless remarkable results are achieved, observations should be further augmented to handle the issues of data imbalance and to foresee people's changes in the behavior over time..

References

Agarwal, H., Wang, X., Kulkarni, N. R., Tao, S., & Demers, C. (2023). Application of machine learning in ensuring viral safety of biotherapeutics: Case study demonstrating prediction and optimization of viral clearance performance of anion exchange chromatography. *Current Research in Biotechnology*, *6*(July), 100140. https://doi.org/10.1016/j.crbiot.2023.100140

Altan, A., & Karasu, S. (2019). The effect of kernel values in support vector machine to forecasting performance of financial time series and cognitive decision making. *The Journal of Cognitive Systems*, *4*(1).

Arifianto, T., Sunaryo, S., & Moonlight, L. S. (2022). Penggunaan Metode Support Vector Machine (SVM) Pada Teknologi Mobil Masa Depan Menggunakan Sidik Jari. *Jurnal Teknik Informatika Dan Teknologi Informasi*, *2*(2), 3–6.

Barbaro, E., Grua, E., Malavolta, I., Stercevic, M., Weusthof, E., & Hoven, J. (2020). Memodelkan dan memprediksi keterlibatan pengguna dalam aplikasi seluler. *Data Science*, *3*(2), 61–77. https://doi.org/10.3233/ds-190027

Halawani, H. T., Mashraqi, A. M., Badr, S. K., & Alkhalaf, S. (2023). Automated sentiment analysis in social media using Harris Hawks optimisation and deep learning techniques. *Alexandria Engineering Journal*, *80*(July), 433–443. https://doi.org/10.1016/j.aej.2023.08.062

Li, J., & Zhang, Y. (2021). Customer Data Classification Using SVM. *Journal of Financial Technology*.

Liu, Y., Shi, X., Pierce, L., & Ren, X. (2019). Mengkarakterisasi dan meramalkan keterlibatan pengguna dengan grafik tindakan dalam aplikasi. *Proceedings of the International Conference On ...* https://doi.org/10.1145/3292500.3330750

Mahmud, T., Karim, R., Chakma, R., Chowdhury, T., Hossain, M. S., & Andersson, K. (2024). A Benchmark Dataset for Cricket Sentiment Analysis in Bangla Social Media Text. *Procedia Computer Science*, *238*(2019), 377–384. https://doi.org/10.1016/j.procs.2024.06.038

Meng, R., Yue, Z., & Glass, A. (2020). Memprediksi status keterlibatan pengguna untuk evaluasi online asisten cerdas. *Arxiv Preprint*. https://doi.org/10.48550/arxiv.2010.00656

Moulaei, K., Sharifi, H., Bahaadinbeigy, K., Haghdoost, A. A., & Nasiri, N. (2023). Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis. *International Journal of Medical Informatics*, *179*(October), 105243. https://doi.org/10.1016/j.ijmedinf.2023.105243

Rodriguez, D. (2024). Memanfaatkan pembelajaran mesin untuk mengembangkan fenotipe keterlibatan digital pengguna dalam program pencegahan diabetes digital: studi evaluasi. *Jmir Ai*, *3*, e47122. https://doi.org/10.2196/47122

Schellewald, A. (2021). Communicative Forms on TikTok: Perspectives From Digital Ethnography. *International Journal of Communication*, *15*.

Tan, X., & others. (2023). Impact of Data Attributes on SVM Performance in Finance. *Journal of Data-Driven Finance*.

Tang, X. (2020). Mengetahui nasib Anda: persahabatan, tindakan, dan penjelasan temporal untuk prediksi keterlibatan pengguna di aplikasi sosial. *Arxiv Preprint*. https://doi.org/10.48550/arxiv.2006.06427

Wang, S., Xiao, X., & Ding, Q. (2024). A novel fractional system grey prediction model with dynamic delay effect for evaluating the state of health of lithium battery. *Energy*, *290*, 130057. https://doi.org/10.1016/j.energy.2023.130057

Wang, Y., Liu, Z. L., Yang, H., Li, R., Liao, S. J., Huang, Y., Peng, M. H., Liu, X., Si, G. Y., He, Q. Z., & Zhang, Y. (2024). Prediction of viral pneumonia based on machine learning models analyzing pulmonary inflammation index scores. *Computers in Biology and Medicine*, *169*(October 2023), 107905. https://doi.org/10.1016/j.compbiomed.2023.107905

Zhou, T., & others. (2023). Machine Learning in Credit Card Application Classification. *Journal of Data Science in Finance*.