App User Segmentation

Anusha Manjunath Saboji [1] , Dr. Geetha M [2]
[1] *Student, Department of MCA, BIET, Davanagere*
[2] *Associate Professor, Department of MCA, BIET, Davanagere*

ABSTRACT

In today's highly competitive digital landscape, retaining users is as crucial as acquiring them, prompting the need for effective strategies to identify and engage users at risk of disengagement. This paper presents a machine learning–based approach utilizing KMeans clustering to segment mobile app users into three categories—Retained, Churn, and Needs Attention—based on key behavioral metrics such as average screen time, spending patterns, ratings, review activity, password reset frequency, and last activity timestamp. By analyzing these multidimensional indicators, the system provides actionable insights that enable businesses to proactively tailor retention strategies and enhance user experience. The proposed solution is implemented as a Flask-based web application featuring role-based access control, where administrators can manage users and maintain a Frequently Asked Questions (FAQ) section, while users can register, log in, and submit behavioral data to receive segmentation results.

Keywords: KMeans, App User Segmentation, Churn, Needs Attention

I. INTRODUCTION

In today's highly competitive digital environment, understanding user behaviour has become a critical factor for the success and sustainability of mobile applications.

With the increasing availability of user activity data, businesses are shifting their focus from mere acquisition to long-term user retention. The primary challenge lies not only in attracting users but in retaining them effectively, as disengaged users often churn—uninstalling or abandoning the app entirely. Consequently, identifying users who are likely to churn or require engagement interventions can significantly enhance app performance, user satisfaction, and overall business growth.

This study proposes a machine learning-based user segmentation system designed to categorize users into three behavioural segments: *Retained*, *Churn*, and *Needs Attention*. The classification is based on key behavioral metrics, including average screen time, in-app spending, app ratings, review activity, password reset frequency, and time since last activity. By employing the K-Means clustering algorithm, the system learns patterns within user data to accurately group users into the aforementioned categories, enabling targeted retention strategies.

A Flask-based web application has been developed as part of this research to operationalize the segmentation model. This application allows users to input relevant behavioral data and receive real-time feedback on their predicted segment. In addition, interactive data visualizations, such as scatter plots, are incorporated to illustrate relationships between user behavior and engagement segments. Users also have the option to download the segmented dataset in CSV format for further analysis or integration with external systems.

The application supports two user roles: **Admin** and **User**. Admins are provided with tools to manage registered users and maintain an FAQ repository, while general users can register, access personalized prediction services, and consult FAQs for guidance. This dual-role system enhances the usability and manageability of the platform, ensuring a comprehensive and user-centric experience.

The potential impact of this project is substantial, offering actionable insights into user engagement and aiding mobile app developers and businesses in designing data-driven strategies to reduce churn and foster sustained user interaction. By leveraging machine learning for behavior-based segmentation, the project underscores the significance of predictive analytics in optimizing app retention and overall user experience.

II. RELATED WORK

Interpretable New User Clustering and Churn Prediction on a Mobile Social Application

Authors: Yang, C., Shi, X., Luo, J., & Han, J. (2019)

This paper introduces ClusChurn, a two-step framework combining clustering and sequence modeling. First, it uses interpretable clustering on initial user behavior and network features. Then, it employs an LSTM with attention to predict churn, even with limited data for new users. The model achieves state-of-the-art performance and is deployed in a live Snapchat environment for real-time churn intervention. This paper demonstrate the power of integrating unsupervised clustering with supervised sequence modeling for churn. They group users based on initial engagement and social graph features, then feed those patterns into an LSTM to predict churn. Their real-world application proves the approach's effectiveness and deployability.[1]

An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry

Authors: Bilal, S. F., Almazroi, A. A., Bashir, S., Khan, F. H.,

This study evaluates combinations of clustering (K-means, K-medoids, X-means) with multiple classifiers, including ensemble methods. The hybrid model achieves up to 94.7% accuracy on telecom datasets by pre-segmenting users before classification. This paper underscore the value of a two-stage pipeline: first clustering, then applying classification. Their hybrid system outperforms standalone models, highlighting how behavioral segmentation via clustering can improve churn prediction accuracy.[2]

Behavioural Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Customer Defection and Loyalty

Authors: Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J. (2015)

This work presents an end-to-end churn prediction pipeline: exhaustive feature engineering, selection, and feeding into supervised models on massive telco data. Their models achieve 89.4% predictive accuracy. This research highlights the importance of comprehensive feature engineering in churn modeling. Although not clustering-based, their robust feature-driven supervised pipeline demonstrates effective churn detection in large-scale mobile networks.[3]

A Hybrid Data Mining Method for Customer Churn Prediction

Authors: E. Jamalian, R. Foukerdi. (2018)

This paper presents a hybrid approach for predicting customer churn to improve retention in competitive markets. By combining data fusion and feature extraction techniques, the study trains two classifiers—LOLIMOT and C5.0—on selected features and integrates their outputs using weighted voting. Tested on real telecommunications data, the method demonstrates enhanced accuracy over individual models, offering an effective solution for churn management and helping businesses reduce customer attrition.[4]

The use of knowledge extraction in predicting customer churn in B2B

Authors: Jamjoom, A. A. (2021)

This study explores churn prediction in B2B insurance using clustering (K-means) and classification. Results show clustering aids in grouping clients into distinct segments, improving interpretability and prediction performance. Jamjoom's research confirms K-means as a valuable exploratory tool for segmenting B2B clients prior to prediction. The study highlights interpretability as an important advantage of the clustering step in churn pipelines.[5]

Early Churn Prediction from Large Scale User-Product Interaction Time Series

Authors: Shamik Bhattacharjee, Utkarsh Thukral, Nilesh Patil,

This paper presents a comprehensive study on early prediction of user churn in large-scale, business-to-customer environments, with a focus on the fantasy sports industry. The authors propose treating churn prediction as a multivariate time series classification problem, leveraging user activity data and deep neural networks to improve accuracy. Their approach reduces the need for extensive feature engineering and outperforms traditional machine learning methods, enabling businesses to better understand attrition trends and implement more effective retention strategies in complex and volatile user environments.[6]

Customer Churn Analysis Using Machine Learning

Authors: Ritika Tyagi & K. Sindhu.

This paper tackles the challenge of understanding and predicting customer churn by developing a machine learning-based model that streamlines both data analysis and prediction tasks. The authors first conduct exploratory data analysis using graphical methods to identify which features are most correlated with churn. They then build and evaluate multiple classification algorithms, applying cross-validation and hyperparameter tuning to optimize performance. An ensemble approach is also used to further boost accuracy. Ultimately, the XGBoost classifier outperforms other models, making it the preferred choice for end-to-end customer churn prediction, as it effectively handles complex feature interactions and delivers high predictive accuracy.[7]

Modeling and Customer Churn Prediction Using Deep Learning

Authors: Gayathri Sundaram, Venkateswar Reddy, Thirupathi Reddy & Rakesh Reddy (2024),

This study introduces a deep learning–based model, incorporating optimized LSTM networks, to predict telecom customer churn using sequence learning, combined with clustering-guided marketing interventions. Conducted in Chennai, their hybrid LSTM+clustering framework processes telecom usage sequences to detect churn patterns. The integration of clustering-driven marketing scenarios adds practical interpretability, reinforcing the relevancy of combining unsupervised segmentation with temporal behavior modeling.[8]

Modeling and customer churn prediction using deep learning,

Authors: Gayathri Sundaram, Venkateswar Reddy, Thirupathi Reddy, Rakesh Reddy.

This study investigates customer churn prediction in the telecom sector using deep learning techniques. The authors propose a feature set optimized Long Short-Term Memory (LSTM) model to predict churn based on a real-world telecom dataset. To enhance predictive performance, boosting methods are applied alongside the LSTM model. The research also explores marketing strategies informed by clustering results and simulates targeted marketing activities to maximize profit. The findings highlight the effectiveness of deep learning and advanced feature engineering in improving churn prediction accuracy and supporting strategic customer retention efforts.[9]

Customer Churn Prediction Using ML Algorithms.

Authors: Prajwal Waghole, Kalpna Saharan, Varad Wanwase, Shantanu Wasnik, Rajratan Gokhale.

This paper presents a machine learning-based approach to predicting customer churn using demographic and behavioral data from the telecom industry. The authors developed and compared predictive models using support vector machines (SVM), random forests, and decision trees, finding that random forests consistently delivered the best performance across accuracy, precision, recall, and F1-score metrics. The study highlights the effectiveness of machine learning in identifying at-risk customers, enabling businesses to implement proactive retention strategies. The authors also suggest future research directions, such as incorporating additional features or ensemble methods, to further improve churn prediction. Overall, the research provides actionable insights for organizations seeking to reduce churn and foster long-term customer loyalty through data-driven decision-making.[10]

III. METHODOLOGY

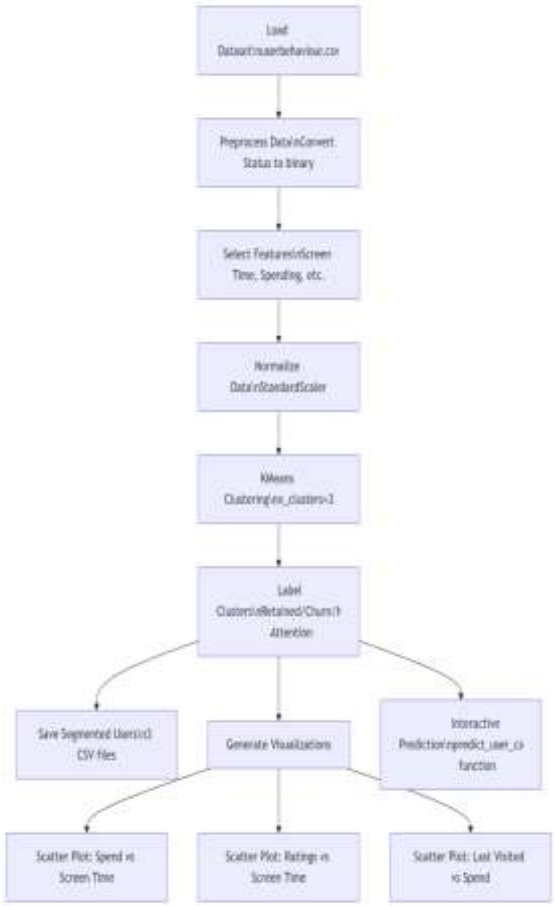


Fig 3.1.1: Architecture Diagram

3.1 Data collection

The dataset named userbehaviour.csv is used for analysis. It contains multiple features reflecting user interactions with the app such as average screen time, spending behavior, review activity, app ratings, password reset count, last activity time, and installation status. This data forms the foundation for clustering and behavior prediction.

3.2 Data preprocessing

The preprocessing step includes cleaning and formatting the data. The 'Status' column is converted into numeric form where 'Installed' is mapped to 0 and 'Uninstalled' to 1. Only the essential behavior-related features are selected for clustering. The selected features are scaled using StandardScaler so that all features contribute equally to the distance measurement during clustering.

3.3 Clustering using KMeans

The KMeans clustering algorithm is applied with three clusters to segment users into behavior groups. The model is trained on the normalized data, and each user is assigned to one of the three clusters. These clusters are then labeled meaningfully as Retained, Churn, and Needs Attention based on behavioral patterns observed.

3.4 Segment mapping and storage

After clustering, the predicted cluster labels are added to the dataset, and each label is mapped to a user-friendly segment name. The data is then filtered based on these segments and saved into three separate CSV files for further analysis or integration with other systems. Each file corresponds to a specific user group.

3.5 Data analysis

Basic statistical insights are generated to understand user behavior trends. This includes calculating the average, maximum, and minimum screen time and spending values. Additionally, the distribution of users across the three segments is printed to observe how users are spread among different behavior groups.

3.6 Visualization

Scatter plots are created to visually explore relationships between different features and how users are grouped. Plots are generated for screen time versus spending, screen time versus ratings, and last visited minutes versus spending. Each point is colored based on the user segment to make clusters easy to interpret.

3.7 Interactive prediction

A command-line input function is implemented to allow users to enter their app usage metrics. These inputs are scaled using the same scaler used during training. The trained KMeans model predicts the segment to which the user belongs, providing an instant behavioral category prediction as either Retained, Churn, or Needs Attention.

IV. TECHNOLOGIES USED

- Flask is used as the web framework to build the backend of the application. It allows handling routes, displaying forms for user input, and connecting machine learning results to the web interface.
- Python is the main programming language used in this project. It is used for data loading, cleaning, analysis, clustering, and making predictions based on user input.
- Pandas is used for loading and manipulating the dataset. It helps in filtering records, modifying columns, and saving segmented data into separate CSV files.
- NumPy is used for handling numerical arrays and preparing user input data for prediction. It simplifies array operations required before scaling and clustering.
- Matplotlib and Seaborn are used for data visualization. These libraries help create scatter plots that represent the relationship between screen time, app spending, ratings, and other features.
- Scikit-learn is used for implementing machine learning. StandardScaler is used for feature normalization and KMeans is used to perform clustering on user behavior data.
- CSV file format is used for reading input data and exporting the clustered user segments into individual files like retained users, churn users, and needs attention users. This makes it easy to store and share results.

V. Result

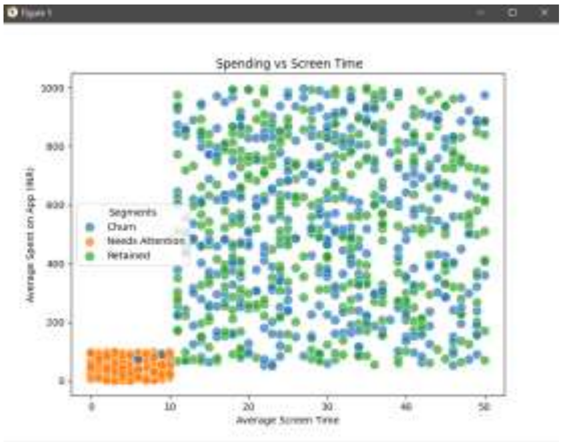


Fig 5.1: Spending vs Screen Time

This screen visualizes the relationship between users' average screen time and their average spending on the app, segmented by user type. Users who spend more time on the app generally tend to spend more money, with distinct behavioral clusters (Retained, Churn, Needs Attention) clearly visible in the scatter plot.

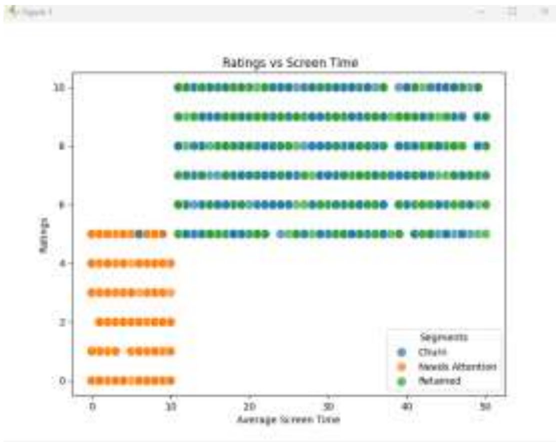


Fig 5.2: Rating Vs Screen Time

This chart shows how user ratings relate to their average screen time, separated by user segments.

Higher ratings are generally associated with users who spend more time on the app, while lower ratings and minimal screen time are mostly observed among users likely to churn.

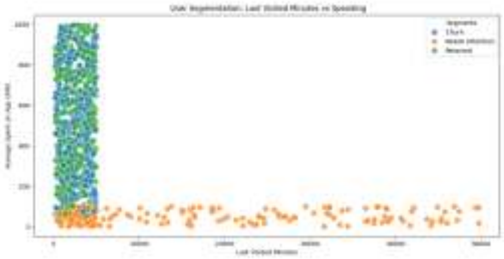
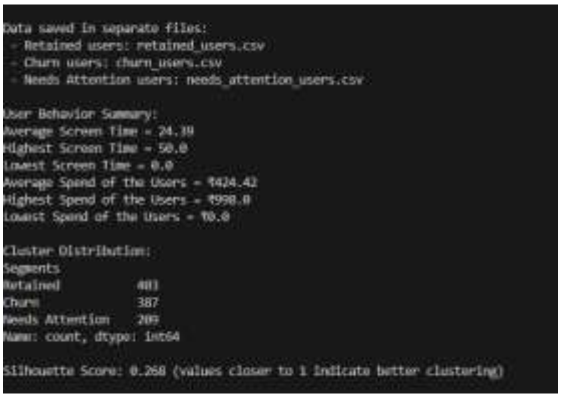


Fig 5.3: Last Visited Minutes vs Spending Segment Explanation

This visualization highlights how user spending relates to the time since their last app visit, segmented by user group. Users in the "Churn" segment typically have not visited the app for a long time and show low spending, while "Retained" and "Needs Attention" users visit more frequently and tend to spend more on the app.



```
Data saved in separate files:
 - Retained users: retained_users.csv
 - Churn users: churn_users.csv
 - Needs Attention users: needs_attention_users.csv

User Behavior Summary:
Average Screen Time = 24.39
Highest Screen Time = 50.0
Lowest Screen Time = 0.0
Average Spend of the Users = ₹424.42
Highest Spend of the Users = ₹998.0
Lowest Spend of the Users = ₹0.0

Cluster Distribution:
Segments
Retained          403
Churn             387
Needs Attention   209
Name: count, dtype: int64

Silhouette Score: 0.268 (values closer to 1 indicate better clustering)
```

The KMeans clustering algorithm segmented the user base into three distinct groups: Retained (403 users), Churn (387 users), and Needs Attention (209 users). The average screen time across users was 24.39 minutes, with spending averaging ₹424.42, indicating varied engagement and monetization levels. Although the silhouette score of 0.268 suggests moderate cluster separation—common in complex behavioral datasets—the clusters provide meaningful insights into user behavior. The Retained group exhibited higher engagement and spending, while the Churn group reflected low activity and app uninstallation. The Needs Attention segment represents users with intermediate characteristics who may benefit from targeted retention strategies. These findings highlight the potential for personalized interventions to improve user retention and revenue.

VI. CONCLUSION

In conclusion, this study demonstrates how machine learning-based user segmentation enhances mobile app retention by categorizing users into Retained, Churn, and Needs Attention groups based on behavioral metrics. The Flask-based web application enables real-time predictions, interactive visualizations, and effective user management, providing actionable insights that help businesses design targeted retention strategies to reduce churn and foster sustained user engagement for long-term success.

VII. REFERENCES

[1]. Interpretable New User Clustering and Churn Prediction on a Mobile Social Application

Yang, C., Shi, X., Luo, J., & Han, J. (2019), DOI : arXiv:10.48550/arXiv.1910.01447

Publisher : arXiv.

[2] An ensemble based approach using a combination of clustering and classification algorithms to enhance customer churn prediction in telecom industry, Bilal, S. F., Almazroi, A. A.,

Bashir, S., Khan, F. H., & Almazroi, A. A. (2022), DOI: 10.7717/peerj-cs.854, Publisher: PeerJ Computer Science

[3]. Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Customer Defection and Loyalty, Khan, M. R., Manoj, J., Singh, A., & Blumenstock, J. (2015), DOI: 10.1109/BigDataCongress.2015.107, Publisher: IEEE Big Data Congress

[4]. A Hybrid Data Mining Method for Customer Churn Prediction, E. Jamalian, R. Foukerdi. (2018), Vol. 8, No. 3, 2018, 2991-2997, Publisher: Engineering, Technology & Applied Science Research.

[5]. The use of knowledge extraction in predicting customer churn in B2B, Jamjoom, A. A. (2021), DOI: 10.1186/s40537-021-00500-3, Publisher: Journal of Big Data

[6]. Early Churn Prediction from Large Scale User-Product Interaction Time Series, Shamik Bhattacharjee, Utkarsh Thukral, Nilesh Patil, DOI: /arXiv.2309.14390, Publisher: arXiv

[7]. Customer Churn Analysis Using Machine Learning, Ritika Tyagi & K. Sindhu, DOI:10.1007/978-981-19-0332-8_37, Publisher: Springer, Singapore (in: Proceedings of International Joint Conference on Advances in Computational Intelligence, Algorithms for Intelligent Systems series)

[8]. Modeling and Customer Churn Prediction Using Deep Learning, Gayathri Sundaram, Venkateswar Reddy, Thirupathi Reddy & Rakesh Reddy (2024), Publication: AIP Conference Proceedings, AIP Publishing, Chennai, DOI: 10.1063/5.0208736,

[9]. Modeling and customer churn prediction using deep learning, Authors: Gayathri Sundaram, Venkateswar Reddy, Thirupathi Reddy, Rakesh Reddy, DOI: 10.1063/5.0208736, Publisher: AIP Publishing (AIP Conference Proceedings, Vol. 3044, 2024)

[10]. Customer Churn Prediction Using ML Algorithms. Authors: Prajwal Waghole, Kalpna Saharan, Varad Wanwase, Shantanu Wasnik, Rajratan Gokhale, Publication Details: Year: 2024 | Volume: 15 | Issue: 02 | Pages: 70-75, Affiliation: RMD School College of Engineering, Pune, Maharashtra, India