Journal of Current Science and Technology, April - June 2025 Vol. 15 No. 2, Article 103

ISSN 2630-0656 (Online)

Cite this article: Mulla, R., Potharaju, S., Tambe, S. N., Joshi, S., Kale, K., Bandishti, P, & Patre, R. (2025). Predicting player churn in the gaming industry: a machine learning framework for enhanced retention strategies. *Journal of Current Science and Technology*, *15*(2), Article 103. https://doi.org/10.59796/jcst.V15N2.2025.103

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Predicting Player Churn in the Gaming Industry: A Machine Learning Framework for Enhanced Retention Strategies

Rahesha Mulla[1], Saiprasad Potharaju[2,*], Swapnali N Tambe[3], Suvarna Joshi[4], Kalpana Kale[4], Pancham Bandishti[4], and Ruchita Patre[4]

[1]Department of Artificial Intelligence & Machine Learning, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[2]Department of Computer Science and Engineering, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[3]Department of Information Technology, K. K.Wagh Institute of Engineering Education & Research, Nashik, MH, India
[4]Department of Computer Science and Engineering, MIT ADT University, Pune, India

[*]Corresponding Author; E-mail: psaiprasadcse@gmail.com

Received 27 November 2024; Revised 9 December 2024; Accepted 21 December 2024; Published online 25 March 2025

Abstract

This research presents a prediction of gaming player churn along with a thorough analysis. It employs predictive modeling techniques utilizing machine learning approaches to predict player churn (customer attrition) on gaming platforms. Using real-world gaming data from player demographics, in-game purchases, social interactions, and historical gaming behavior, this study proposes a new framework that integrates data preprocessing, segmentation, and predictive modeling to determine which players will churn. Additionally, it uses Logistic Regression and Random Forest, a powerful ensemble learning algorithm, to estimate player churn within a limited time horizon. We found that this approach accurately identified potential churners through a thorough exploration and understanding of the dataset. This predictive model provides insight into the key factors influencing player attrition, allowing game developers to take countermeasures to prevent churn risks and improve player retention strategies. In addition, Power BI insights highlight the key factors influencing player churn. These findings provide actionable recommendations for game developers to mitigate churn risks and enhance player retention strategies. This study contributes to predicting player turnover in the gaming industry, providing a valuable tool for fostering sustainable growth and profitability.

Keywords: *predictive modeling; machine-learning model; Power-BI for prediction; churn analysis; classification*

1. Introduction

In the gaming industry, one of the most significant challenges is players churn—the phenomenon of players leaving a game. This issue directly impacts revenue, disrupts community dynamics, and threatens the long-term sustainability of gaming platforms. Understanding the mechanics of player churn and employing predictive techniques are critical for designing effective retention strategies that maintain a healthy player base (Panimalar et al., 2023). Specifically, this study investigates a case example of Candy Crush to analyze player behavior and suggest retention strategies that optimize satisfaction and performance in the gaming market. This research focuses on churn prediction because it is crucial for industry stakeholders and essential for gaming sector growth.

The gaming industry relies on understanding and predicting player attrition to create tailored player experiences, maximize engagement, attract new players, and expand business opportunities. The goal of this research is to offer academic insights and real-world applications that enable game developers to leverage player behavior data for churn prediction. By enhancing reliability and sustainability in the industry, this study aims to improve player satisfaction through better prediction techniques and retention strategies.

1.1 Related Work

As gaming business models evolve, so does the analysis of player behavior and its implications for business intelligence. The gaming industry has seen a rise in the use of predictive analytics and machine learning to understand and reduce churn, driven by growing interest and investment in advanced analytics. Extensive literature on churn prediction presents various frameworks and models aimed at understanding and mitigating churn. This research builds on studies by Li et al., (2021) and Karmakar et al., (2022), who developed personalized and statistical models to analyze player behavior, motivations, and retention strategies. Our study extends their findings by incorporating more sophisticated algorithms and real-time data integration for more precise and adaptive churn prediction. Based on Personalized Perceived Difficulty (PPD) and Dynamic Difficulty Influence (DDI), Li et al., (2021) presented a Difficulty-Aware Framework (DAF) for predicting player churn and intervention. Utilizing Random Forest algorithms and Power BI, this study advances previous research by improving player retention predictions with more accurate churn predictions and actionable retention strategies. This method enhances dropout rate prediction by modeling engagement, collaboration, and achievement at each level. Our research builds on these efforts by incorporating Random Forest algorithms and Power BI, offering more precise churn predictions and practical strategies to enhance player retention. Karmakar et al., (2022) developed a statistical framework for retention analysis in freemium RPGs, using a generalized linear mixed model to capture player motivations, progression, and churn. Their method improved dropout rate prediction by analyzing engagement, collaboration, and achievement at each level. In contrast, our research utilizes dynamic feature selection and real-time data integration for adaptive churn prediction and targeted retention strategies, resulting in a more robust solution to changes in the gaming environment. Flunger et al., (2022) examined analytics related to churn prediction and customer lifetime value (CLV) prediction in the F2P business model. Their study discussed methods, metrics, and techniques used in game analytics.

However, our research integrates real-time data analysis with personalized intervention strategies to significantly reduce churn and enhance player retention. Perišić et al., (2022) addressed the challenge of defining churn in a non-contractual setting by proposing four definitions based on user activity and engagement, which were evaluated using the Jaccard similarity coefficient. In correlation, our research adopts a dynamic and context-dependent churn definition methodology that enables more accurate predictions and tailors' interventions to improve player retention. Perišić et al., (2021) proposed an extended RFM (Recency, Frequency, and Monetary)-based churn prediction model for mobile games, considering user lifetime, intensity, and reward. Our research builds on this effort by incorporating comprehensive player behavioral analytics and enhancing churn analysis and retention strategies through the use of advanced visualization tools. This approach is complemented by the application of machine learning and advanced analytics in churn prediction, following the work of Wang et al., (2022) and Singh et al., (2023), who analyzed player data and improved retention using various predictive models such as decision trees, random forests, and unsupervised learning approaches.

We further extend these methods by incorporating dynamic features and leveraging real-time analytics to enhance utility across different player segments. For mobile games, Óskarsdóttir et al., (2022) demonstrated that player churn can be influenced by social network features and first-day behavior. To this end, our study introduces a broader set of behavioral indicators and more sophisticated machine learning (ML) algorithms to provide additional insights and improve predictive accuracy for enhancing engagement and retention mechanisms. Additionally, Wang et al., (2022) emphasized the need for innovative churn analysis that incorporates unique perspectives to adapt to diverse gaming environments and maintain a stable paying user base. Furthermore, our research extends feature selection and model optimization techniques for player segmentation and churn definition, ensuring that retention strategies remain adaptive to various gaming environments. These enhancements help maintain consistent performance and adaptability in a rapidly evolving gaming landscape.

To combat high churn rates in digital game-based learning (DGBL), this study analyzes churn and develops a machine learning-based predictive model using logistic regression, decision trees, and random forests (Kiguchi et al., 2022). Kiguchi et al., (2022) employed this approach and, using logistic regression, achieved high effectiveness, obtaining an AUC of 0.9225 and an F1-score of 0.9194. Singh et al., (2023) utilized predictive analytics to develop models for forecasting and preventing player churn in online multiplayer games. Data from gaming platforms was preprocessed and used to train five machine learning models, namely decision trees, logistic regression, random forests, and AdaBoost. These models achieved predictive accuracies between 90% and 99.1% on out-of-sample data. The focus of this research is on proactive churn management as a strategy for ensuring revenue stability and customer loyalty. According to Weiss et al., (2023), an unsupervised learning model was used to predict user churn in online recreational games through three techniques: Principal component analysis (PCA) to quantify engagement diversity, Decision trees, and Random forests, as explored in the study by Kiguchi et al., (2022). Their approach demonstrated high effectiveness, achieving an AUC of 0.9225 and an F1-score of 0.9194 with logistic regression. In the context of online multiplayer games, Singh et al., (2023) applied predictive analytics to design models for player churn prediction and prevention. By preprocessing gaming platform data, various machine learning models including decision trees, logistic regression, random forests, and AdaBoost achieved predictive accuracies between 90% and 99.1% on out-of-sample data. This research highlights the importance of proactive churn management in maintaining revenue stability and player retention. Weiss et al., (2023) proposed an unsupervised learning approach to predict user churn in online recreational games by quantifying engagement diversity through principal component analysis (PCA). Geometric variability in user data trajectories allowed for tracking engagement trends and enabled competitive performance compared to black-box machine learning models. Through this study, we recommend intuitive and explainable decision-rule algorithms for effective churn prediction in online gaming platforms.

Xiong et al., (2023) proposed a data-driven framework for analyzing player churn in online games, leveraging explainable AI methods to predict and understand churn causes. Their approach, validated on Justice PC, aids game publishers in making informed decisions to improve player retention and game quality. The study by Zhao et al., (2023) introduced perCLTV, a system for personalized Customer Lifetime Value (CLTV) prediction in online games. It addresses personalization across churn and payment tasks using sequential gated multi-task learning. Validated on real-world datasets, perCLTV outperformed baseline models, enhancing precision marketing in online games.

With diverse data sources post-installation, Yun et al., (2023) introduced MDLUR for LTV prediction in Club Vegas Slots. Specific studies and case analyses, including those by Kawale et al., (2019), have focused on player behavior clustering and analyzing different game types, such as MMORPGs and online mini-games. These studies emphasize the importance of understanding social dynamics and player engagement metrics. Using advanced visualization tools, we extend these findings by applying a data-driven approach to refine retention strategies and address issues related to unbalanced matchmaking and disparate player experiences.

In the study by Perišić et al., (2023), a hybrid dissimilarity measure was developed for clustering player behavior in mobile games, leading to improved churn prediction. They combined categorical and quantitative data with normalized linear distances and demonstrated how this approach enhanced performance when using PAM clustering on real-world datasets.

The challenges in online minigames were analyzed in World by Yang et al., (2022). By applying exponential smoothing and BP neural networks to player data, the study examined game popularity and player score predictions from January 2022 to January 2023. The findings demonstrated high feasibility and low fitting error in the proposed approach.

Kawale et al., (2009) proposed a churn prediction model for MMORPGs like EverQuest II, integrating social influence and player engagement metrics. Their modified diffusion model enhanced churn prediction accuracy by leveraging social dynamics and activity-based engagement metrics. Kang et al., (2024) presented a churn prediction model for free-to-play games, defining churn formally and assessing generic game features such as playtime. Their universal churn model improved player retention strategies through data-driven insights in a competitive gaming market. Kim et al., (2017) focused on churn prediction in mobile and online casual games, using play log data to analyze churn among new players. Their research employed logistic regression, gradient boosting, random forests, CNN, and LSTM models, emphasizing feature selection and churn period definitions (OP and CP) for effective churn management.

Yang et al., (2022) introduced novel churn prediction features based on playtime regularity metrics derived from entropy measures. Their research validated these features across datasets from six online games, demonstrating superior churn prediction accuracy and highlighting their utility in refining retention strategies. Kang et al., (2024) investigated the impact of match experiences on churn in competitive games, shedding light on factors influencing player retention. Their study found that unbalanced matchmaking and match results influenced player churn, suggesting improvements in matchmaking to reduce skill gap variance. However, our research addresses these issues by utilizing Random Forest to predict churn based on diverse variables, visualized through Power BI.

The study by Xiong et al., (2023) applied survival analysis to predict customer churn in automobile insurance, using demographic and policy data collected over six years. Significant factors such as age, marital status, and claim indicators were identified, with K-Prototypes clustering revealing distinct customer segments that require targeted retention strategies.

The gaming industry is highly competitive, with numerous gaming platforms competing for users' attention. Player retention is a critical metric that determines a gaming platform's success by driving revenue generation and enhancing user engagement. By understanding the factors contributing to churn and predicting churn events in advance, gaming companies can proactively retarget players, ultimately improving the overall gaming experience. Churn, or gameplay abandonment, remains a major challenge in the gaming industry. It involves a decline in player activity, voluntary discontinuation of gameplay, losses in revenue, disruptions in community dynamics, and threats to platform sustainability.

Churn typically begins with game installation and high initial engagement. Over time, player activity declines for various reasons, putting them at risk of churn. This process is illustrated in Figure 1. During periods of high player engagement, activity levels are elevated; however, over time, plateaus and subsequent declines indicate an increased risk of churn. Understanding this process is crucial for gaming companies as it allows them to optimize targeted retention strategies and strengthen player loyalty. The core challenge is to analyze why and how player churn occurs and to develop predictive techniques that can help mitigate it.

The goal of this task is to predict whether a player is likely to churn based on a dataset containing various gameplay and engagement metrics. The dataset includes features such as purchase frequency, total booster usage, level progression rate, time since last purchase, average daily playtime, overall engagement score, and scaled social engagement ratio. We propose an approach to analyze these features and develop a strong, robust predictive model to identify players at risk of churn. To achieve this, we build a churn prediction model using machine learning methods, specifically Random Forest and Logistic Regression algorithms, to enhance accuracy and provide actionable insights for player retention strategies.

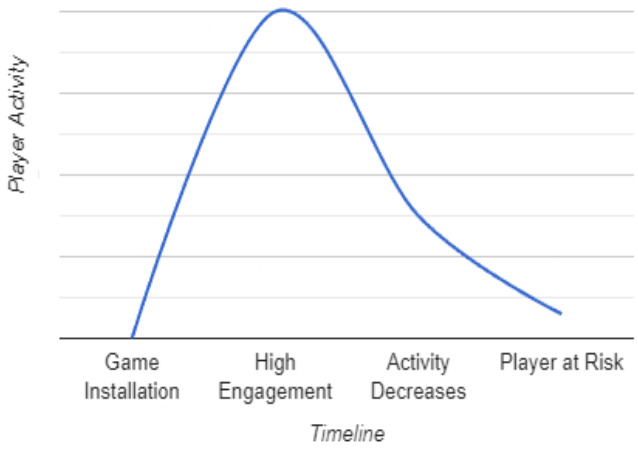


Figure 1 Player Activity Analysis (Singh et al., 2023)

2. Objectives

The goal of the research is to create a predictive tool that can identify gamers who are in danger of quitting by examining comprehensive data on player interactions and activities. Enabling game producers to use tailored tactics to maintain player engagement and lower churn rates is the goal.

3. Methodology

The research methodology as depicted in Figure 2 consists of data acquisition using a synthesized dataset, followed by preprocessing to address inconsistency issues. Exploratory data analysis (EDA) provided insights, and feature engineering generated prediction conditions. Multiple machine learning algorithms were evaluated, and the best-performing model was selected. Principal Component Analysis (PCA) has been used to enhance efficiency. The chosen model underwent training and evaluation, with further research exploring optimal feature extraction and churn period durations for improved performance.

3.1 Data Collection

The data collection process involves accessing and extracting data from the *Candy Crush* game platform. We capture various aspects of player interactions, including levels played, number of attempts, time spent on different features, use of boosters, and social interactions such as sending friend invites and participating in guild chat messages. Additionally, data on in-game purchases is collected, including the amount spent, purchase timing, and purchase type. This data is obtained through the game's internal analytics system, which continuously monitors and records player activities in real-time as they engage with the game.

3.2 Data Preprocessing

Before training any predictive models, we preprocess the collected data to ensure high quality and usability. This process includes several key steps, such as data cleaning, handling missing values, and normalization. Data cleaning involves identifying and addressing missing, incorrect, or inconsistent data to ensure that the dataset remains complete, valid, and meaningful. To handle missing values, techniques such as imputation and deletion are applied to maintain data integrity. Additionally, numerical features are adjusted and normalized to ensure a consistent scale across all variables, preventing features with larger magnitudes from dominating the modeling process. Without these preprocessing steps, the data would not be suitable for analysis, and the predictive models would likely yield poor results.

3.3 Exploratory Data Analysis

For forecasting player churn, exploratory data analysis (EDA) involves a thorough examination of the dataset to understand player behavior patterns and identify those at risk of churning. To begin, we analyze descriptive statistics such as the mean, median, and standard deviation to assess central tendencies and variability within the data. We then visualize key feature distributions using histograms, box plots, and density plots to detect outliers and assess any skewness. To explore relationships between variables, we utilize scatter plots and correlation matrices to identify potential predictors of churn. Additionally, segmentation analysis is performed to group players based on demographic and behavioral attributes, helping to evaluate their likelihood of churning. Another crucial aspect of EDA is analyzing temporal trends in player activity and churn rates over time to uncover any seasonal patterns or long-term trends. The goal of EDA is to gain a high-level understanding of the dataset's structure and patterns, ultimately guiding the next steps in the predictive modeling process.

3.4 Feature Engineering

The conversion of raw data into attributes to provide insights on player behavior and engagement trends is an important aspect of model development and is called feature engineering. For example, we have generated new variables, such as the average daily login duration required to complete any given task. We also extract features related to player activity, including the level and frequency of specific actions performed within the game. Additionally, we gather attributes such as the rate of advancement in levels, total booster utilization, and the proportion of social interactions, aiming to capture how players engage with game assets and participate in the gaming community.

To improve the accuracy of predictions regarding when players stop playing, these features are incorporated into the dataset, augmenting it with relevant information.

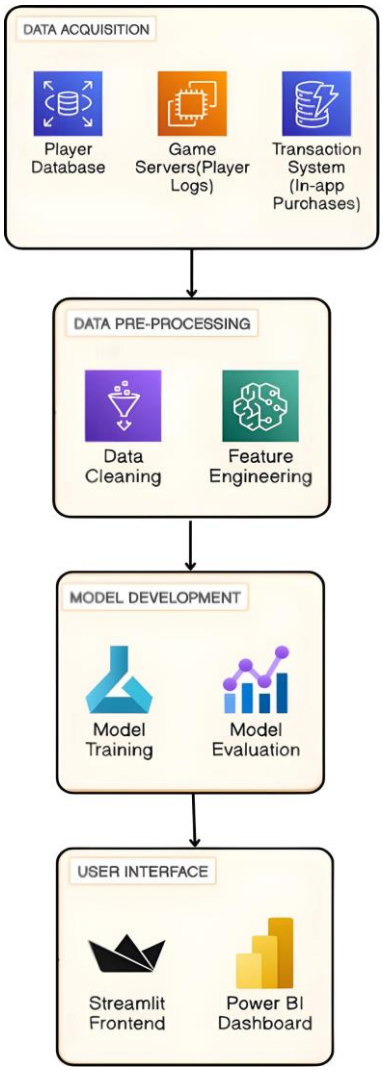


Figure 2 Research Methodology

3.5 Model Selection

We utilize both Random Forest and Logistic Regression models. Given the complexity of player behavior data and the critical importance of accurate predictions for informing retention strategies, we adopt a multi-modal ensemble approach to enhance predictive accuracy and robustness.

1) Random Forest: This model is selected for its ability to handle complex, nonlinear relationships within the data and its strong performance in predictive modeling tasks. As an ensemble of decision trees, Random Forest aggregates individual predictions, making it particularly suitable for analyzing diverse gaming data.
2) Logistic Regression: In contrast, Logistic Regression is a simple yet effective algorithm for modeling the probability of a binary outcome, such as player churn prediction. Its interpretability and efficiency make it a valuable baseline for comparison.

The performance of both models is evaluated using key metrics, including accuracy, precision, recall, and F1-score, to determine the most effective approach for our research.

3.6 Principal Component Analysis

Karl Pearson originally described the search for "lines and planes of closest fit to systems of points in space" as Principal Component Analysis (PCA)—a method for reducing the dimensionality of data.

Later, Fisher and MacKenzie suggested that PCA was a more appropriate technique than analysis of variance for modeling response data. PCA is a statistical technique that transforms a high-dimensional dataset into a lower-dimensional one while preserving as much variance as possible. By identifying the most significant patterns in the data (known as principal components), PCA enhances interpretability while minimizing information loss. This makes PCA particularly useful in various domains, including data mining, image processing, and predictive modeling (Pechprasarn et al., 2023).

In our player churn forecasting project, PCA serves as a powerful tool to simplify the analysis of complex player attributes and behaviors. By extracting orthogonal variables (principal components) from the dataset, PCA enables compact feature selection and dimensionality reduction, helping to uncover underlying patterns and correlations in the data. This improves the accuracy of predictive modeling while maintaining essential player behavior information.

By projecting the high-dimensional feature space into a lower-dimensional representation, PCA not only simplifies the modeling process but also enhances model efficiency and generalization. When integrated with the Random Forest algorithm, PCA optimizes model performance by reducing computational complexity while preserving key predictive determinants of player churn. Ultimately, this approach equips gaming companies with actionable insights to enhance player retention strategies, drive engagement, and promote growth and profitability in the gaming industry.

By employing PCA in our analysis, we effectively reduce dataset dimensionality without compromising critical player behavior and engagement data. This streamlined representation enhances the performance of predictive models, such as Random Forest, by reducing overfitting, improving computational efficiency, and increasing model interpretability.

3.7 Model Training

In this study, we developed predictive models to estimate when players might stop playing the game. To ensure the highest possible accuracy, we followed a structured machine learning workflow.

First, we split the dataset into two parts:

1) Training data: A large portion of the dataset used to train the model.
2) Testing data: A separate portion used to evaluate model performance.

Using the training data, we provided the model with historical player interaction data, including information on when players disengaged from the game. During this process, we performed data preprocessing, which involved cleaning errors, ensuring numerical values were within valid ranges, and formatting different data types for consistency and proper model interpretation. Next, we trained the model using advanced machine learning algorithms, specifically Random Forest and Logistic Regression, to make accurate churn predictions. These models were fine-tuned by adjusting hyperparameters to optimize performance. Once training was complete, we evaluated the model using the testing data to assess its predictive accuracy.

If the model demonstrated strong performance, it was leveraged as a valuable tool for game companies to analyze player behavior, understand churn patterns, and develop strategies to improve player retention. Throughout this process, the model continuously improved through iterative evaluations and refinements. Ultimately, these trained models provided actionable insights to game designers, helping them implement effective strategies to enhance player engagement and retention while optimizing the overall gaming experience.

3.8 Model Evaluation

Model evaluation is a critical step in assessing how well the predictive model can forecast player churn and the accuracy of its results. This phase ensures that the trained models effectively generalize new, unseen data by testing their performance on a separate testing subset that was set aside during data preparation. To measure model performance, we employ key evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive understanding of the model's predictive capabilities and its ability to distinguish between churners and non-churners. A confusion matrix is used to visualize the classification results, comparing predicted vs. actual churn outcomes. This helps in identifying the model's strengths and weaknesses in detecting true churners (players likely to stop playing) and non-churners (players who remain engaged). By analyzing misclassifications, we can refine the model and improve its predictive power. Through this rigorous evaluation process, we aim to assess the efficiency and reliability of the predictive models in capturing player churn dynamics. The insights derived from these evaluations assist gaming platform leaders in making data-driven decisions to enhance player retention strategies and improve overall engagement.

4. Results

Data was collected from Candy Crush, encompassing player interactions such as gameplay, social engagement, and in-game purchases. This synthetically generated dataset provided real-time-like insights into player behavior.

4.1 Exploratory Data Analysis

Exploratory Data Analysis (EDA) involves statistical techniques and visualizations to uncover patterns, such as monthly in-app purchases (as depicted in Figure 3) and potential indicators of churn. This process includes descriptive statistics, histograms, scatter plots, and segmentation analysis to understand data distribution, identify outliers, explore variable relationships, and discern temporal trends. EDA provides critical insights that guide subsequent steps in predictive modeling, ensuring a more data-driven approach to churn prediction and player retention strategies.

4.2 Model Results

We use specific benchmarks to identify players who may discontinue using the gaming service. By comparing individual player data against these benchmarks, the system can determine which players are at risk of churning. Players meeting the risk criteria are assigned a churn label. This information is visualized using a count plot, which displays the number of players likely to churn versus those expected to stay. This visualization provides valuable insights into player retention trends, enabling data-driven decisions to enhance player engagement and maintain a healthy gaming platform.

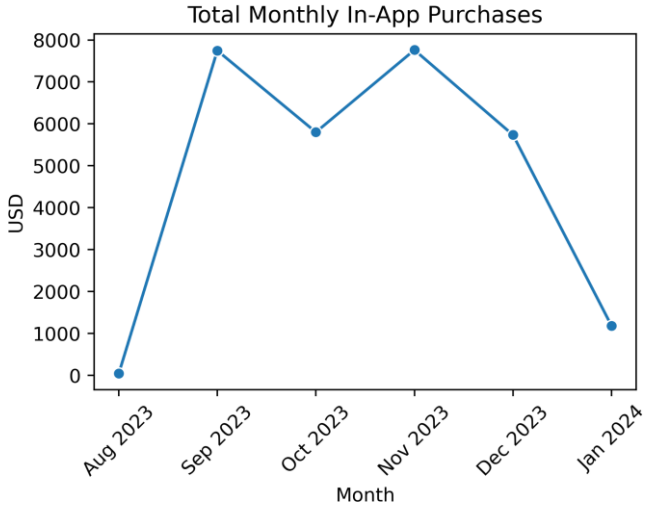


Figure 3 Monthly In-App Purchases

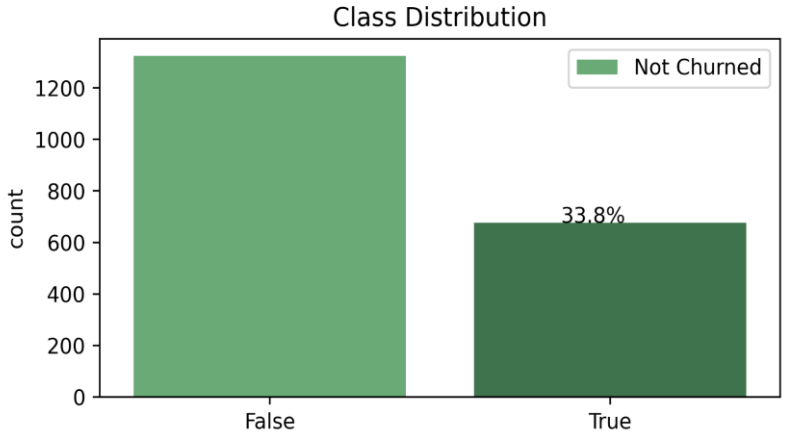


Figure 4 Churn Distribution

Table 1 Classification Report and Accuracy of Logistic Regression

	Precision	Recall	F1-score	Support
False	0.89	0.85	0.87	261
True	0.75	0.81	0.78	139
Accuracy	0.84			
Macro avg	0.82	0.83	0.82	400
Weighted avg	0.84	0.84	0.84	400

We used a Random Forest model for this prediction, which is a subtype of machine learning that constructs multiple decision trees to analyze different parts of the data. This ensemble approach allows for more accurate predictions while preventing the model from overfitting the sample data. After training the model on the dataset, we evaluated its performance, achieving an accuracy of 97%. This means the model correctly predicts whether players will change teams or remain in their current team 97% of the time. The classification report, shown in Tables 1 and 2, provides detailed metrics such as precision, recall, and F1-score for different player groups (those who left and those who stayed). This helps assess how well the model performs for each category.

A confusion matrix is used to illustrate the model's accuracy in determining whether a player will stay or leave. It provides a breakdown of true positives, false positives, true negatives, and false negatives, offering insights into where the model may have misclassified instances. Additionally, Principal Component Analysis (PCA) enhances the model by reducing the data's complexity without losing critical information. This improves processing speed and can potentially enhance predictive performance.

The ROC curve visualizes the model's ability to differentiate between players who will leave and those who will stay by comparing true positive rates against false positive rates across various thresholds. The AUC score quantifies overall model performance, with higher values indicating better differentiation. The Random Forest model achieved an AUC score of 0.93, demonstrating strong predictive capabilities.

Overall, the Random Forest model is well-suited for predicting player behavior accurately. The incorporation of PCA further enhances its efficiency, and additional analyses provide deeper insights into its performance.

4.3 Logistic Regression

For Logistic Regression, we trained a model on the preprocessed data and achieved an accuracy of 83% on the test set. The classification report, shown in Table 1, provides a detailed evaluation of the model's performance across different metrics.

Similarly, the confusion matrix in Figure 5 provides insights into the model's ability to correctly classify churned and non-churned instances, while the ROC curve in Figure 6 illustrates its discrimination capability, achieving an AUC score of 0.93.

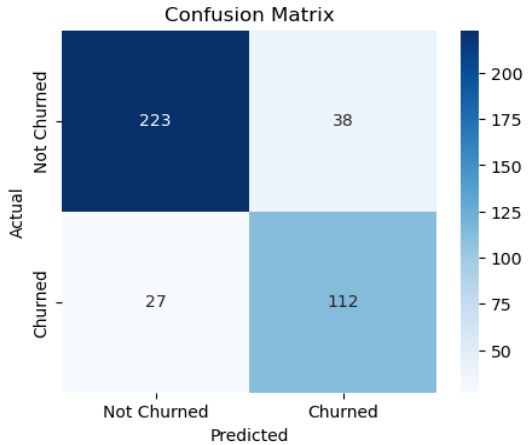
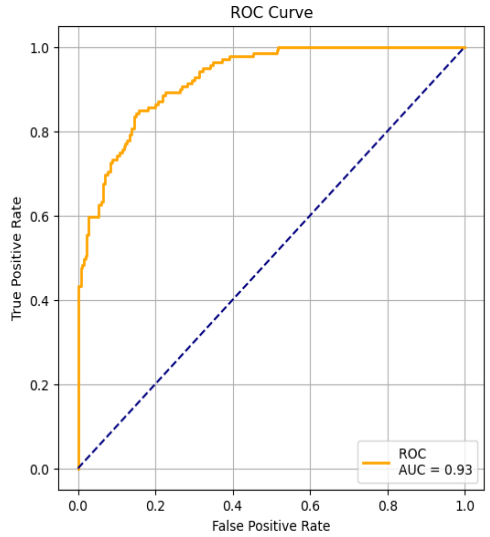


Figure 5 Confusion Matrix of Logistic Regression Model



AUC Score: 0.925246010088481

Figure 6 ROC Curve of Logistic Regression Model

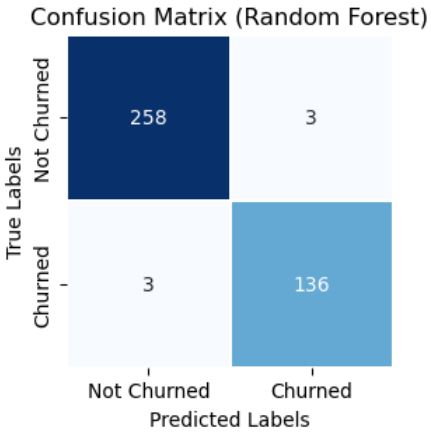


Figure 7 Confusion Matrix of Random Forest Model

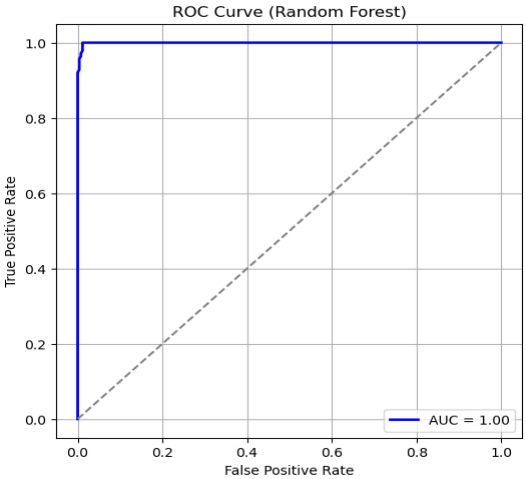


Figure 8 ROC Curve of Random Forest Model

Table 2 Classification report and accuracy of random forest

	Precision	Recall	F1-score	Support
False	0.99	0.99	0.99	261
True	0.98	0.98	0.98	139
Accuracy	0.98			
Macro avg	0.98	0.98	0.98	400
Weighted avg	0.98	0.98	0.98	400

The output includes key performance metrics such as accuracy, the classification report, and the AUC score. Additionally, visualizations of the confusion matrix, churn probability distribution, and ROC curve are generated to offer a comprehensive understanding of the model's performance.

4.4 Random Forest

A Random Forest model was trained on the preprocessed data. Random Forest is an ensemble learning method that constructs multiple decision tree classifiers on different subsets of the dataset and aggregates their outputs to enhance predictive accuracy while reducing the risk of overfitting. The model's performance was evaluated on the test set, achieving an accuracy of approximately 94%. However, the statement that "98% of the model's predictions were correct" contradicts the reported accuracy. If the accuracy is 94%, it means 94% of the predictions were correct, not 98%.

The classification report in Table 2 shows insights into the model's performance across distinct evaluation metrics such as precision, recall, and F1-score for each class (churned and non-churned), helping in understanding its effectiveness for each category.

The confusion matrix in Figure 7 visualizes the model's proficiency in correctly categorizing churned and non-churned instances, providing a detailed breakdown of true positives, false positives, true negatives, and false negatives. This analysis aids in identifying the types of errors made by the model. Additionally, the implementation of Principal Component Analysis (PCA) enhances the model's performance by reducing dimensionality while retaining essential information, thereby improving computational efficiency and potentially enhancing predictive accuracy. The output includes key printed metrics such as accuracy, the classification report, and the AUC score. Furthermore, visualizations such as the confusion matrix, churn probability distribution, and ROC curve in Figure 8 are generated to offer a comprehensive understanding of the model's performance. After experimenting with various models for binary classification to predict whether a player is likely to churn, we found that logistic regression provides a straightforward and effective approach for making predictions. However, for a more in-depth analysis of the factors influencing churn and for more nuanced predictions, random forest proved to be more robust. It effectively handles complex interactions between variables while also providing insights into the importance of different features, making it an invaluable tool for developing targeted retention strategies.

4.5 Power-BI Dashboard

Utilizing Power BI, an interactive dashboard was created to visualize the performance of the churn prediction model and provide key data insights.

- Churn Overview: Displays a summary of churn statistics, including the churn rate, total customers, and the number of churned customers.
- Feature Importance: Analyzes the most influential factors in predicting churn.
- Player Segmentation: Categorizes players based on churn probability or other relevant features.

In the first dashboard (Figure 9), the behavior of players who have churned, or stopped playing, is analyzed. Among the 676 churned players, the total amount spent was 7.00K. Regarding purchase frequency, the highest number of purchases was six, made by two players, while most players made between three and four purchases. When examining average daily playtime, these players engaged for approximately two to six hours per day.

Figure 9 Dashboard Outputs for Potential Churn Players

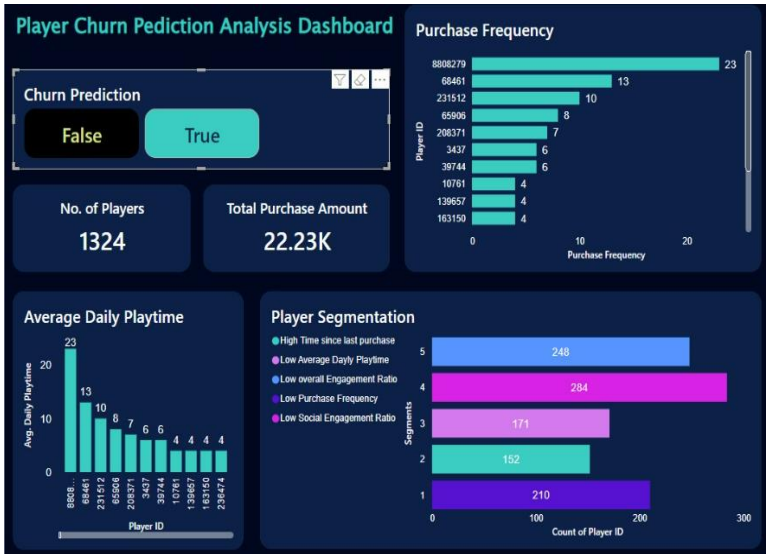


Figure 10 Dashboard Outputs for Non-Potential Churn Players

A significant portion of churned players, specifically 594 out of 676, exhibited low average daily playtime. All 676 churned players had a low overall engagement ratio, with 397 showing a low purchase frequency. Additionally, 427 players had low social engagement, indicating minimal interaction with other players. These figures suggest that churned players generally exhibit low engagement in terms of playtime, purchases, and social interaction. In contrast, Figure 10 focuses on players predicted not to churn, revealing a different pattern. This group consists of 1,324 players, with a total purchase amount of 22.23K significantly higher than that of the churned players. The highest purchase frequency in this group is an impressive 23 purchases by a single player, with several others making between 4 and 13 purchases. The average daily playtime for these non-churning players ranges from 4 to 23 hours, indicating a much higher level of engagement. Specifically, 248 players exhibited a high time since their last purchase, demonstrating sustained engagement despite not making recent purchases.

Although 284 players in this group had low average daily playtime, this number is considerably lower than that of the churned group. The number of players with a low overall engagement ratio is 171, significantly fewer than in the churned group.

Similarly, only 152 players had a low purchase frequency, and 210 had a low social engagement ratio. Comparing the two dashboards, it is evident that players predicted not to churn display higher overall engagement and make purchases more frequently. Non-churning players have a total purchase amount more than three times higher than churned players (22.23K vs. 7.00K). Their purchase frequencies and daily playtimes are significantly higher, with some players engaging for up to 23 hours daily and making as many as 23 purchases.

Conversely, Churned players show lower engagement through reduced playtime, fewer purchases, and limited social interaction. These patterns emphasize the need to boost engagement and spending to reduce churn. By targeting these factors, game companies can improve retention and sustain revenue. The interactive dashboard further supports this by enabling users to explore data, assess model performance, and generate actionable insights for customer retention strategies.

5. Discussion

Random Forest demonstrated a high accuracy of 97% on the test set, compared to an accuracy of 83% achieved by Logistic Regression on the same dataset. Random Forest is an advanced version of the decision tree, constructing multiple decision trees and aggregating their outputs to enhance predictive performance. This ensemble approach allows Random Forest to identify nonlinear patterns in the data, leading to improved management of player behavioral variability, feature interactions, and noise-like factors in the dataset.

However, compared to the proposed decision tree model, Logistic Regression exhibits lower accuracy due to its linear nature, which limits its ability to capture crucial and complex aspects of player churn. Nevertheless, its simplicity and interpretability make it a suitable starting point for further model development. Logistic Regression performs best when feature relationships are linear and when the data is well-separated without significant overlap. The shortcomings of Logistic Regression, particularly its lower accuracy, highlight the importance of selecting the appropriate model for churn prediction problems. Given the complexity of gaming datasets, Logistic Regression is often insufficient for capturing intricate player behavior. Instead, models such as Random Forest or neural networks are better suited for providing a more detailed analysis of player activity. These differences underscore the trade-off between model interpretability and predictive power-a balance that game developers must consider when implementing retention strategies.

6. Conclusion

In conclusion, this research provides meaningful insights into player churn prediction within the Candy Crush gaming platform. By applying machine learning techniques such as Random Forest and Logistic Regression, the team built effective models to predict when players are likely to stop playing. Through analysis of extensive gameplay data, key patterns were identified, including player behavior, in-game purchases, and social interactions. Careful data preparation and feature selection were essential in enhancing model accuracy. Among the tested algorithms, Random Forest demonstrated the best performance, thanks to its ability to model complex relationships and produce interpretable results. Additionally, Principal Component Analysis (PCA) was used to reduce dimensionality, streamline the dataset, and improve model efficiency. The findings of this study not only deepen our understanding of player churn but also offer practical tools for developers to improve player retention. By identifying when and why players disengage, gaming companies can design targeted strategies to maintain engagement and extend player lifespans. Looking forward, future research could incorporate real-time data and explore additional behavioral and psychological factors to further refine churn predictions. Ongoing innovation in data analytics will be key to adapting to player needs and maintaining competitive advantage in the dynamic gaming industry.

7. References

Flunger, R., Mladenow, A., & Strauss, C. (2022). Game analytics: Business impact, methods and tools. In *Developments in information & knowledge management for business applications* (Vol. 3, pp. 601–617). Springer. https://doi.org/10.1007/978-3-030-77916-0_19

Kang, H., Suh, C., & Kim, H. K. (2024). Match experiences affect interest: Impacts of matchmaking and performance on churn in a competitive game. *Heliyon, 10*(3), Article e24891. https://doi.org/10.1016/j.heliyon.2024.e24891

Karmakar, B., Liu, P., Mukherjee, G., Che, H., & Dutta, S. (2022). Improved retention analysis in freemium role-playing games by jointly

modelling players' motivation, progression and churn. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *185*(1), 102-133. https://doi.org/10.1111/rssa.12730

Kawale, J., Pal, A., & Srivastava, J. (2009). *Churn prediction in MMORPGs: A social influence based approach* [Conference presentation]. International Conference on Computational Science and Engineering. August 29-31, 2009, IEEE, Vancouver, BC, Canada. https://doi.org/10.1109/CSE.2009.80

Kiguchi, M., Saeed, W., & Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest. *Applied Soft Computing*, *118*, Article 108491. https://doi.org/10.1016/j.asoc.2022.108491

Kim, S., Choi, D., Lee, E., & Rhee, W. (2017). Churn prediction of mobile and online casual games using play log data. *PloS One*, *12*(7), Article e0180735. https://doi.org/10.1371/journal.pone.0180735

Li, J., Lu, H., Wang, C., Ma, W., Zhang, M., Zhao, X., ... & Ma, S. (2021). *A difficulty-aware framework for churn prediction and intervention in games* [Conference presentation]. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. August 14-18, 2021, Virtual Event Singapore. https://doi.org/10.1145/3447548.3467277

Óskarsdóttir, M., Gísladóttir, K. E., Stefánsson, R., Aleman, D., & Sarraute, C. (2022). Social networks for enhanced player churn prediction in mobile free-to-play games. *Applied Network Science*, *7*(1), Article 82. https://doi.org/10.1007/s41109-022-00524-5

Panimalar, S. A., & Krishnakumar, A. (2023). A review of churn prediction models using different machine learning and deep learning approaches in cloud environment. *Journal of Current Science and Technology*, *13*(1), 136-161. https://doi.org/10.14456/jcst.2023.12

Pechprasarn, S., Wattanapermpool, O., Warunlawan, M., Homsud, P., & Akarajarasroj, T. (2023). Identification of Important Factors in the Diagnosis of Breast Cancer Cells Using Machine Learning Models and Principal Component Analysis. *Journal of Current Science and Technology*, *13*(3), 642-656. https://ph04.tci-thaijo.org/index.php/JCST/article/view/700/415

Perišić, A., & Pahor, M. (2021). RFM-LIR feature framework for churn prediction in the mobile games market. *IEEE Transactions on Games*, *14*(2), 126–137. http://dx.doi.org/10.1109/TG.2021.3067114

Perišić, A., & Pahor, M. (2023). Clustering mixed-type player behavior data for churn prediction in mobile games. *Central European Journal of Operations Research*, *31*(1), 165-190. https://doi.org/10.1007/s10100-022-00802-8

Perišić, A., Jung, D. Š., & Pahor, M. (2022). Churn in the mobile gaming field: Establishing churn definitions and measuring classification similarities. *Expert Systems with Applications*, *191*, Article 116277. https://doi.org/10.1016/j.eswa.2021.116277

Singh, K. K., Rohatgi, S., & Singh, M. P. (2023). An empirical study to predict churn of online multiplayer games and its impact on revenue of the game developing company. *Journal of Statistics and Management Systems*, *26*(5), 1161–1174. https://doi.org/10.47974/JSMS-1169

Wang, G. Y. (2022). Churn prediction for high-value players in freemium mobile games: using random under-sampling. *Statistika: Statistics and Economy Journal*, *102*(4), 443-453. https://doi.org/10.54694/stat.2022.18

Weiss, I., & Vilenchik, D. (2023). Predicting churn in online games by quantifying diversity of engagement. *Big Data*, *11*(4), 282-295. https://doi.org/10.1089/big.2022.0109

Xiong, Y., Wu, R., Zhao, S., Tao, J., Shen, X., Lyu, T., ... & Cui, P. (2023). *A data-driven decision support framework for player churn analysis in online games* [Conference presentation]. August 6-10, 2023, https://doi.org/10.1145/3580305.3599759

Yang, W., Huang, T., Zeng, J., Chen, L., Mishra, S., & Liu, Y. E. (2022). Utilizing players' playtime records for churn prediction: Mining playtime regularity. *IEEE Transactions on Games, 14*(2), 153–160. https://doi.org/10.1109/TG.2020.3024829

Yun, J., Kwak, W., & Kim, J. (2023). *Multi datasource LTV user representation (MDLUR)* [Conference presentation]. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data. August 6-10, 2023, Long Beach CA USA. https://doi.org/10.1145/3580305.3599871

Zhao, S., Wu, R., Tao, J., Qu, M., Zhao, M., Fan, C., & Zhao, H. (2023). perCLTV: A general system for personalized customer lifetime value prediction in online games. *ACM Transactions on Information Systems, 41*(1), 1–29. https://doi.org/10.1145/3530012