Investigating the Impact of Gameplay Hours on Player Recommendations in Steam Games: A Comparative Analysis Using Logistic Regression and Random Forest Classifiers

Yusuf Durachman[1,*], Abdul Wahab Abdul Rahman[2]

[1]Department of Information System, State Islamic University Syarif, Hidayatullah Jakarta, Indonesia

[2]Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia

ABSTRACT

The study delves into the complex relationship between gameplay hours and player recommendations on the Steam platform, leveraging both Logistic Regression and Random Forest classifiers to analyze the data. The findings underscore a strong correlation between hours played and the likelihood of recommending a game. Specifically, longer gameplay hours generally indicate higher engagement levels, which often translate into a greater propensity for players to recommend the game. However, this trend is not universally applicable; a subset of users with high playtime did not recommend their games, highlighting that engagement alone does not guarantee satisfaction. Factors such as game quality, unmet player expectations, and individual preferences may influence these outcomes. The Logistic Regression model provided a clear linear understanding of the data, demonstrating that hours played significantly affect recommendation likelihood. Its coefficients suggested a positive relationship, making it a useful tool for interpreting the odds of recommendation changes based on gameplay hours. Nonetheless, the model's limitations became evident in its inability to capture intricate, non-linear patterns within the data. In contrast, the Random Forest classifier excelled by capturing complex interactions and offering robust predictive accuracy. This model utilized ensemble learning to analyze various decision trees, thereby revealing more nuanced insights into player behaviors. Feature importance scores derived from Random Forest confirmed that hours played was a critical variable, but also highlighted the potential significance of other factors contributing to player recommendations. Model performance metrics further reinforced these observations. The Random Forest classifier outperformed Logistic Regression in terms of accuracy (82.65% compared to 81.26%), precision, recall, and the F1-score, while also delivering a higher Area Under the Curve (AUC-ROC), indicating superior discriminative power. These results suggest that Random Forest is more suitable for capturing the multifaceted dynamics of player engagement and recommendations. This comprehensive comparison illustrates how different modeling approaches can yield valuable, yet varying, insights into gaming data.

Keywords Gameplay hours, Player engagement, Player recommendations, Logistic Regression analysis, Random Forest model

Submitted 30 December 2024
Accepted 23 January 2025
Published 28 February 2025

Corresponding author
Yusuf Durachman,
yusuf_durachman@uinjkt.ac.id

Additional Information and Declarations can be found on page 73

DOI: 10.47738/ijrm.v2i1.21

Introduction

The gaming industry has undergone substantial growth and transformation over

How to cite this article: Y. Durachman, and A. W. A. Rahman, "Investigating the Impact of Gameplay Hours on Player Recommendations in Steam Games: A Comparative Analysis Using Logistic Regression and Random Forest Classifiers," *Int. J. Res. Metav.*, vol. 2, no. 1, pp. 52-77, 2025.

the past few decades, significantly influenced by the emergence of digital distribution platforms such as Steam. In 2020 alone, Steam reported the release of 10,263 new games, a dramatic increase from the 276 games available in 2010, underscoring the rapid expansion of game offerings accessible to consumers. This surge reflects a broader trend where digital platforms have become indispensable for both game distribution and community engagement, enabling developers to reach a global audience more efficiently than traditional retail methods [1]. Steam has solidified its position as a dominant force in the digital game distribution market, boasting approximately 555 million active users and housing over 12,000 games in its extensive database. Beyond facilitating game purchases and downloads, Steam integrates social networking features that allow users to discuss and share their gaming experiences, thereby fostering a vibrant community that enhances user engagement and retention [2].

Player engagement metrics, such as hours played and recommendations, are pivotal in evaluating game performance and gauging player satisfaction. These metrics offer critical insights into player behavior, preferences, and the overall effectiveness of a game. Research has demonstrated that increased hours spent playing games are often associated with improved cognitive skills and higher engagement levels [3], highlighting the importance of playtime in assessing how well a game retains its audience. Additionally, player recommendations serve as a significant indicator of satisfaction and perceived game quality, with studies indicating that players who invest more time in a game are more likely to endorse it to others [4]. This correlation suggests that fostering longer play sessions can enhance player satisfaction and promote organic growth through positive word-of-mouth.

Furthermore, data mining techniques are instrumental in extracting actionable insights from these engagement metrics. By analyzing vast amounts of gameplay data, developers can identify patterns in player behavior, predict player preferences, and inform strategic decisions in game design and marketing [5], [6]. Consequently, leveraging data mining not only aids in understanding current player dynamics but also in anticipating future trends, thereby enabling the creation of more engaging and satisfying gaming experiences.

Understanding the relationship between the amount of time players spend on a game and their likelihood to recommend it represents a core issue in assessing game performance and player satisfaction. Player engagement metrics, such as hours played and recommendations, offer critical insights into player behavior and preferences, which are essential for evaluating how effectively a game retains its audience [3]. The significance of this relationship extends to game developers and marketers, who rely on these metrics to enhance player satisfaction and drive game success. By comprehensively analyzing how gameplay duration influences recommendation rates, stakeholders can identify key factors that contribute to positive player experiences and address areas that may detract from overall satisfaction [4].

The primary objective of this research is to investigate how hours played influence recommendations in Steam games. Additionally, the study aims to compare the effectiveness of Logistic Regression and Random Forest classifiers in modeling this relationship. By evaluating these two algorithms, the

research seeks to determine which model offers superior predictive performance, thereby providing a methodological framework for future analyses. Furthermore, the findings are intended to offer actionable insights that inform game development and marketing strategies, enabling developers to optimize game design for enhanced player engagement and satisfaction. The research addresses the following questions: Is there a significant correlation between gameplay hours and the likelihood of recommending a game? Which algorithm, Logistic Regression or Random Forest, provides a better predictive performance for this relationship?

This research contributes to a growing body of literature exploring data-driven modeling techniques and player engagement dynamics in digital platforms. In sentiment analysis, [7] and [8] demonstrate how data sampling and consumer interactions can refine prediction models. Similarly, targeted applications of ensemble learning approaches have been highlighted in [9] and [10] underscoring the versatility of Random Forest classifiers in predicting user behaviors within digital marketing domains. Further, the intricate relationships between user engagement metrics and digital assets have been explored through [11] and [12] emphasizing the predictive capabilities of correlation analysis in complex markets. In virtual asset management and valuation, [13] and [14] illustrate how social and market dynamics within emerging digital ecosystems can be decoded for practical applications.

Literature Review

Player Engagement Metrics in Gaming Research

The concept of "hours played" is a fundamental metric in understanding player retention and satisfaction within the gaming industry. Research has consistently demonstrated that the amount of time players spends engaging with a game can significantly influence their overall experience and level of satisfaction. Research [15] highlight that player satisfaction is often driven by positive reinforcement mechanisms, such as rewards and a sense of immersion, which are closely tied to time spent in-game. This aligns with findings by [16], who suggest that player satisfaction derived from meaningful choices and autonomy is particularly influential for less experienced players. As players become more familiar with the game mechanics, the hours spent playing enhance their experience, contributing to increased satisfaction. Furthermore, [17] identify a positive correlation between gameplay duration and behavioral patterns, suggesting that longer playtimes may lead to heightened engagement levels.

However, the implications of extended gaming hours extend beyond simple satisfaction metrics, as excessive playtime can also lead to negative psychological impacts. Study [18] notes a high prevalence of video game addiction among players, particularly those engaging in five or more hours of gameplay daily. Such extensive gaming habits have been linked to issues like anxiety and depression, as observed in [19] research, which explores the differential impacts of gaming on male and female players. Moreover, the concept of "tolerance" in gaming, discussed by [20], suggests that players may require increasingly longer sessions to achieve the same level of satisfaction over time, potentially leading to addictive behaviors. While hours played serve as a valuable indicator of player retention and satisfaction, understanding its dual impact is crucial for promoting healthier gaming environments and

identifying potential risks associated with excessive engagement.

The significance of recommendations in video gaming is increasingly recognized as a powerful proxy for player satisfaction, with substantial implications for game popularity and sales. Recommendations often reflect players' personal experiences and satisfaction levels, influencing their likelihood of endorsing the game to others. Autonomy, for example, plays a central role in shaping player satisfaction and, subsequently, their willingness to recommend games. Autonomy satisfaction is particularly impactful for new players, who are more likely to recommend a game if they feel their choices are meaningful and tailored to their preferences. This notion aligns with [21] findings, which suggest that social interactions in gaming enhance engagement and drive positive recommendations. Games that support player agency and community involvement are more likely to receive favorable recommendations, thus boosting their visibility and attractiveness to potential players.

Furthermore, the concept of perceived justice in gaming environments influences both satisfaction and recommendations, particularly in games that utilize the freemium model. Research [22] argue that dissatisfaction with monetization practices can lead to a reluctance among players to make in-game purchases, which, in turn, reduces their likelihood of recommending the game to others. Ensuring a fair and satisfying gaming experience is essential for fostering positive word-of-mouth and maximizing game sales. Additionally, enjoyment plays a critical role in players' intentions to continue playing and to recommend games. Study [23] illustrate that enjoyment directly impacts players' desire to keep engaging with a game, which is closely tied to their likelihood of endorsing it. Social dynamics further underscore the importance of recommendations as [24] find that social facilitation enhances enjoyment, thereby increasing players' propensity to recommend the game. Collectively, these factors illustrate that recommendations are a key measure of player satisfaction, influenced by autonomy, perceived justice, and enjoyment, all of which contribute to the game's popularity and commercial success.

Relationship Between Gameplay Hours and Recommendations

The relationship between time spent on video games and player recommendations has been a central focus of research, particularly in understanding how gaming duration impacts well-being and social interactions. Studies indicate that the correlation between gaming time and recommendations is multifaceted, influenced by individual motivations, the nature of the game, and the context in which gaming occurs. Study [25] highlight the potential for gaming companies to address problematic gaming behaviors by using player data to identify those who may require support services. This proactive approach reflects the industry’s growing recognition of its role in promoting healthy gaming habits, particularly for players who engage in excessive gaming. Moreover, [26] demonstrate that the context of gaming plays a crucial role, as gaming driven by social or recreational motives can mitigate some negative effects, such as depression and social withdrawal. In contrast, compulsive or excessive gaming often results in adverse outcomes, reinforcing the importance of understanding players' motivations behind extended playtime.

Additional research underscores how player motivations influence the relationship between gaming time and well-being, which in turn affects player

recommendations. Research [27] found that players who engage in gaming for intrinsic reasons, such as enjoyment, experience positive effects on well-being and are more likely to recommend the game to others. This positive correlation between intrinsic motivation and well-being suggests that players' underlying motivations are critical in shaping their overall gaming experience and willingness to endorse the game. Conversely, the research indicates that extrinsic motivations can lead to negative gaming experiences and reduce the likelihood of recommendations, highlighting the dual impact of gaming time based on different motivational factors [27]. Further, [28] suggest that players' psychological characteristics significantly impact their satisfaction with gaming, illustrating that the quality of gaming experiences is as crucial as the time spent playing. Study [29] support this complexity, indicating that while the effects of gaming on well-being may be minimal on average, they are heavily contingent on social contexts and interactive elements within the game.

The relationship between increased engagement in video games and player recommendations is grounded in several theoretical frameworks, notably Self-Determination Theory (SDT) and Flow Theory, which provide insight into how engagement influences players' likelihood to recommend games. According to SDT, three fundamental psychological needs—autonomy, competence, and relatedness—drive human motivation. Research [30] demonstrate that games fulfilling these needs enhance player enjoyment and motivation, increasing the likelihood of positive recommendations. Similarly, [31] propose that meeting these intrinsic needs fosters a deeper connection to the game, leading to greater satisfaction and more favorable word-of-mouth recommendations. When players feel empowered by meaningful choices, experience competence through challenging gameplay, and engage with others, they are more likely to perceive the game positively and advocate for it within their social circles.

Flow Theory further elucidates how the state of complete immersion, or "flow," contributes to player satisfaction and recommendations. [32] describe flow as a state where players are fully absorbed in the gaming experience, leading to a heightened sense of achievement and enjoyment. This immersive experience is often associated with positive recommendations, as players who achieve a state of flow are more likely to recall the game favorably and share it with others. [33] expand on this by integrating goal-setting theory with flow, explaining how achievement-oriented features in games can enhance user loyalty and ultimately increase recommendations. Effective feedback and reward systems are also instrumental in sustaining engagement as [34] emphasize that these mechanisms promote social interaction and reinforce player competence, making players more likely to recommend the game. Conversely, [35] discuss how unmet needs, such as frustration with a lack of autonomy or competence, can result in negative experiences and reduce the likelihood of recommendations. This dual nature of engagement highlights that while high engagement may lead to positive recommendations, negative experiences can have the opposite effect, underscoring the importance of understanding player motivations and experiences in game design.

Data Mining Techniques in Gaming Analytics

Supervised learning has become a crucial approach in gaming analytics, with classification algorithms like Logistic Regression, Random Forest, and Support Vector Machines commonly employed to predict player behaviors. These

algorithms leverage historical data to classify players based on patterns in their actions, preferences, and potential future behaviors, offering valuable insights for game developers and researchers. In free-to-play (F2P) mobile games, for instance, predictive modeling is frequently utilized to analyze player engagement and retention, as it helps in understanding how players interact with various game elements. Study [36] emphasize the importance of machine learning models in analyzing large-scale player data, noting that algorithms such as Random Forest and SVM have shown robust performance in predicting player behaviors and retention outcomes. Their study demonstrates that these models are particularly effective in handling vast amounts of data, providing actionable insights for enhancing player engagement strategies.

Beyond digital gaming, classification algorithms have also found applications in traditional sports analytics. For example, [37] explored the use of pattern-mining algorithms to classify player movements in rugby league, which offers insights into coaching strategies and player development. Similarly, [38] applied supervised machine learning techniques to predict college basketball players' performance efficiency, highlighting the versatility of these algorithms in various contexts. This body of research underscores the adaptability of classification models across gaming environments. Additionally, advanced techniques like deep learning and Hidden Markov Models (HMM) have enhanced predictive capabilities, enabling more sophisticated analyses of player behaviors. These studies collectively illustrate how supervised learning is essential for understanding player behavior and optimizing game design and retention strategies.

The comparison of Logistic Regression (LR) and Random Forest (RF) has been extensively examined across diverse fields, providing insights into the strengths and limitations of these models. In a large-scale benchmark study, [39] evaluated the performance of LR and RF across multiple datasets, revealing that RF consistently outperforms LR, particularly in terms of c-statistics, which measure the model’s discriminatory power. The study found a mean difference in c-statistics favoring RF, suggesting its superior predictive accuracy in handling complex datasets. Similarly, [40] observed higher average c-statistics for RF compared to LR across 243 real datasets, indicating that RF generally offers more reliable performance, especially in datasets where non-linear relationships and variable interactions are present.

In clinical settings, the predictive capabilities of RF and LR have been compared to assess their effectiveness in high-stakes scenarios. Research [41] examined the use of these algorithms for predicting clinical deterioration, reporting that RF achieved an AUC of 0.80, surpassing LR's AUC of 0.77. The study attributes RF's improved accuracy to its ability to account for non-linear interactions without extensive data preprocessing, which is often required for LR. [42] similarly reported superior performance of RF in predicting mortality in sepsis patients, with RF achieving an AUC of 0.86 compared to LR’s 0.76. However, the potential for overfitting remains a consideration for RF, particularly with smaller datasets, as highlighted by [41]. Despite this limitation, the comparative studies consistently suggest that RF is more versatile and accurate in complex scenarios. However, LR may be preferable for simpler datasets due to its straightforward implementation and lower risk of overfitting.

Logistic Regression

Logistic Regression is a widely used statistical method for binary classification, where the objective is to classify instances into one of two distinct categories. This algorithm is based on the logistic function, also known as the sigmoid function, which maps predicted values to a probability between 0 and 1. In binary classification problems, this probability is then used to determine the class label based on a threshold, typically set at 0.5. The core of Logistic Regression lies in its ability to model the relationship between a dependent variable and one or more independent variables, which makes it applicable to a variety of fields such as medical diagnosis, fraud detection, and marketing analytics [43]. Logistic Regression is particularly valued for its interpretability, as the model coefficients can indicate the direction and strength of the relationship between predictor variables and the outcome, making it a useful tool for understanding underlying data patterns [44].

The binary classification framework, as used in Logistic Regression, evaluates model performance through key metrics, including True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These metrics allow researchers to assess a model's accuracy and reliability, particularly in distinguishing between two classes. Logistic Regression is well-suited for binary classification scenarios where data is relatively balanced. However, it can be adapted for imbalanced datasets by incorporating techniques such as class weighting and sampling adjustments. Additionally, binary classification techniques like Logistic Regression can be extended to handle multi-class problems through methods such as Error Correcting Output Coding (ECOC), which transforms a multi-class problem into several binary classification tasks, thereby leveraging the binary classification framework to improve overall performance [45], [46]. This adaptability underscores the versatility of Logistic Regression, reinforcing its position as a fundamental method in supervised learning for binary and multi-class classification tasks.

Random Forest Classifier

Ensemble learning methods, particularly Random Forest, have gained significant attention in machine learning due to their ability to enhance predictive accuracy by combining multiple models. Ensemble methods leverage the collective strengths of various algorithms, resulting in more robust predictions and improved generalization compared to single models. Random Forest, an ensemble of decision trees, operates by constructing multiple decision trees during training and aggregating their predictions to achieve a final result. This approach mitigates issues commonly associated with individual decision trees, such as high variance and susceptibility to overfitting, thereby producing a model that is more stable and accurate across different datasets. Studies have shown that ensemble methods like Random Forest outperform traditional single models in various domains, including medical diagnosis and genetics, due to their ability to handle complex, high-dimensional data [47].

One of the primary advantages of ensemble learning methods is their capacity to reduce overfitting, which is particularly beneficial in scenarios with limited data. By averaging predictions from multiple decision trees, Random Forest minimizes the likelihood that the model will adapt too closely to training data noise, thus enhancing generalization. This capability is especially relevant in fields such as bioinformatics, where data scarcity can pose challenges for model reliability. Research [48] emphasize that ensemble techniques, including

Random Forest, can alleviate small sample size issues by incorporating multiple classification models, thus yielding more accurate and reliable predictions. Additionally, the diversity of base classifiers in an ensemble enables Random Forest to handle noise and uncertainty more effectively, making it a versatile tool in complex data-driven applications [49].

At the core of Random Forest lies the decision tree, a supervised learning model used for classification and regression tasks. Random Forest constructs numerous decision trees during training, each based on a random subset of the training data, a technique known as bootstrapping. This sampling method ensures that each tree learns from different data points, thereby enhancing the ensemble's diversity [50]. Each tree operates independently, splitting the data based on feature values to make predictions. The final output of a Random Forest model is obtained by aggregating the predictions from all individual trees, typically using majority voting for classification or averaging for regression. This aggregation process significantly reduces the variance of the model, leading to more reliable predictions by smoothing out individual tree errors.

The decision trees within a Random Forest can be constructed using various algorithms, such as CART (Classification and Regression Trees) or C4.5. The choice of algorithm can affect the structure and performance of each tree, but the ensemble nature of Random Forest allows it to capitalize on the strengths of these different tree-building techniques [51]. Additionally, Random Forest introduces randomness at each split by selecting a subset of features from which to choose the best split. This technique reduces correlation among trees and further minimizes generalization error. This random feature selection is crucial for improving the robustness of the model and enhancing its ability to capture complex patterns in the data, making Random Forest particularly effective for high-dimensional classification tasks [50], [52].

Method

The research method for this study consists of several steps to ensure a comprehensive and accurate analysis. The flowchart in Figure 1 outlines the detailed steps of the research method.

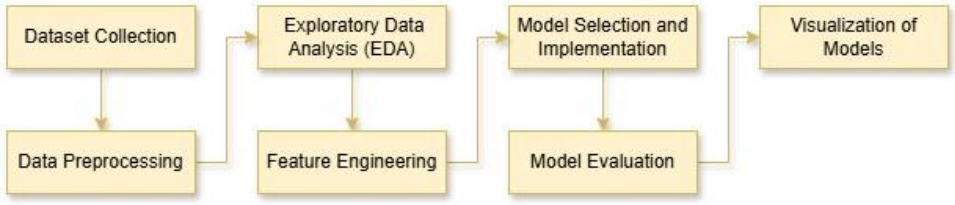

Figure 1 Research Method Flowchart

Data Description

The dataset used for this research, `steam_game_reviews.csv`, contains player reviews of various games on the Steam platform, providing a rich source of data for analyzing player behaviors and engagement metrics. This dataset consists of 992,153 entries, each capturing detailed information about individual player experiences. The primary columns include `review`, `hours_played`, `helpful`, `funny`, `recommendation`, `date`, `game_name`, and `username`. The data is diverse, encompassing reviews from a broad range of games and a large

number of unique players, reflecting the diversity of gaming experiences on the Steam platform. The column `review` contains the textual content of player reviews, providing qualitative insights into player opinions and sentiments. Meanwhile, `username` and `game_name` identify the player and the game being reviewed, respectively.

A key focus of this study is the `hours_played` column, which records the total hours spent by each player on the game. This numeric feature serves as a crucial predictor in understanding player engagement and its potential influence on recommendations. The `recommendation` column, another critical variable, captures whether a player recommends the game (`Recommended` or `Not Recommended`). This binary classification is the target variable for our predictive modeling efforts. The `helpful` and `funny` columns capture the number of users who found the review helpful or funny, respectively, providing additional context regarding the perceived quality and impact of each review. Furthermore, the `date` column indicates when the review was posted, offering potential insights into temporal trends in player engagement and recommendations. Collectively, these features enable a comprehensive analysis of player behaviors, with a particular emphasis on the relationship between gameplay hours and the likelihood of recommending a game.

Preliminary analysis of the dataset revealed several important characteristics and considerations. The `hours_played` column exhibited a wide range of values, with some players reporting minimal gameplay and others recording extensive hours, reaching over thousands of hours in certain cases. This variability highlights the diverse engagement levels among players and necessitates data preprocessing to manage potential outliers. The `recommendation` column displayed a class imbalance, with 805,782 entries marked as `Recommended` and 186,371 entries marked as `Not Recommended`. This imbalance underscores the importance of considering appropriate evaluation metrics during model assessment to ensure a fair representation of both classes. Additionally, the `helpful` and `funny` columns contained mixed data types due to formatting inconsistencies, requiring conversion to numeric data types for effective analysis.

Missing values were present in the `review` and `username` columns, with 503 missing entries in `review` and 81 in `username`. While these columns provide contextual information, they were not critical to the primary analysis and could be addressed through appropriate data handling techniques. Overall, the dataset offered a rich blend of quantitative and qualitative information, enabling a multifaceted exploration of player behaviors and engagement metrics on the Steam platform. Through careful preprocessing and feature selection, this study aimed to extract meaningful insights into how gameplay hours influence player recommendations, leveraging both logistic regression and random forest classifiers to model this relationship.

Data Preprocessing

Data preprocessing began with an assessment of missing values in key columns, specifically `hours_played` and `recommendation`. The initial inspection indicated that neither of these columns contained missing entries, ensuring a complete dataset for the critical variables of interest. Despite the absence of missing values for these fields, general strategies for handling

potential missing data were incorporated to maintain data integrity and adaptability for future analyses. Any rows with missing values in `hours_played` or `recommendation` would have been dropped to avoid bias in predictive modeling. Additionally, other columns, such as `review` and `username`, exhibited missing values. While these fields provided contextual information, their absence did not directly impact the primary analysis, thus they were addressed through data imputation or removal based on the study's objectives.

The data cleaning process focused on ensuring the correctness of data types and preparing the dataset for analysis. Columns such as `hours_played`, `helpful`, and `funny` contained numeric data stored as strings with embedded commas, requiring conversion to numeric types. This conversion was achieved by removing commas and coercing values to numeric formats. Any non-numeric values that emerged during this conversion were treated as missing and subsequently removed to maintain data consistency. Furthermore, the `recommendation` column, which contained categorical values (`Recommended` and `Not Recommended`), was encoded into a binary format, with `1` representing `Recommended` and `0` representing `Not Recommended`. This transformation facilitated efficient modeling with binary classification algorithms, such as logistic regression and random forest classifiers, by providing a standardized target variable for prediction.

To ensure the robustness of predictive models, outlier detection and treatment were conducted on the `hours_played` feature. Using the interquartile range (IQR) method, the lower and upper bounds for acceptable values were defined as Q1 - 1.5 IQR and Q3 + 1.5 IQR, respectively. The analysis identified a substantial number of outliers, with `hours_played` values exceeding the upper bound or falling below the lower bound. These outliers could skew the model's performance and introduce bias, necessitating careful handling. Two approaches were considered: removing the outliers or capping their values at the established bounds. For this study, outliers were removed to ensure a clean and representative dataset, resulting in a final dataset size of 858,081 entries. This decision aimed to reduce noise while maintaining the validity of the `hours_played` metric for predicting player recommendations.

Following data cleaning and outlier management, the dataset was reviewed to ensure all transformations were correctly applied and that no new inconsistencies emerged. The `hours_played` column was confirmed as a continuous numeric variable, and the `recommendation_binary` column provided a binary target for classification. Remaining columns, such as `helpful` and `funny`, were retained as additional features to provide context on user engagement and review impact. The final dataset was then ready for modeling, with a clear structure and consistent data types across all fields. This comprehensive preprocessing pipeline laid the groundwork for effective modeling using logistic regression and random forest classifiers, enabling a thorough analysis of the relationship between gameplay hours and player recommendations on the Steam platform.

Exploratory Data Analysis (EDA)

EDA began with a summary of descriptive statistics for the `hours_played` feature, a key variable in understanding player engagement. The mean `hours_played` was found to be 88.62 hours, indicating that, on average,

players spent a significant amount of time engaging with the games in the dataset. The median value of 44.50 hours further suggested a distribution skewed by higher values, as the median was lower than the mean. The mode, recorded at 0.20 hours, highlighted that a notable portion of players had minimal engagement, reflecting a pattern of quick interactions or potentially early exits from games. This distribution underscored the need to explore whether prolonged gameplay correlated with positive player experiences and recommendations.

The distribution of recommendations was analyzed next, focusing on the `recommendation_binary` column. Approximately 81.26% of reviews in the dataset were marked as `Recommended` (encoded as 1), while 18.74% were `Not Recommended` (encoded as 0). This imbalance indicated that most players had a favorable experience with the games they reviewed. Such an imbalance could influence predictive modeling performance and warranted careful consideration in the subsequent analysis to ensure balanced evaluation metrics.

To further understand the distribution and potential patterns in the `hours_played` data, a histogram was plotted. The histogram showed a right-skewed distribution, with many players spending relatively few hours on a game, while a smaller portion recorded substantially longer playtimes. The presence of long-tail values emphasized the diversity of player engagement, from casual users to dedicated players with extensive gameplay hours. Additionally, a kernel density estimate (KDE) line overlaid on the histogram provided a smoothed representation of the density, offering further insights into the data's central tendency and dispersion.

A box plot comparing `hours_played` across recommendation statuses provided a clearer understanding of how gameplay hours differed between `Recommended` and `Not Recommended` reviews. Players who recommended games tended to have higher median playtimes, suggesting a potential positive correlation between longer engagement and favorable reviews. However, the box plot also highlighted significant variability within each recommendation group, indicating that factors beyond mere playtime might influence player satisfaction and recommendations. This finding motivated the exploration of other predictors and interactions in modeling.

Finally, a scatter plot depicting the relationship between `hours_played` and `recommendation_binary` was generated to visualize potential trends or clusters. Slight jitter was added to the `recommendation_binary` values to enhance interpretability and reduce overlap. The plot revealed a general trend where higher `hours_played` values appeared more frequently associated with `Recommended` outcomes. However, there were exceptions, with some high-playtime users not recommending their games, suggesting that prolonged engagement does not universally guarantee satisfaction. This nuanced relationship between playtime and recommendations highlighted the complexity of player behavior and the need for comprehensive modeling to capture these dynamics effectively.

Feature Engineering

Feature engineering involved the transformation of key numeric variables to ensure consistent scaling and improved performance during model training. The

`hours_played` feature exhibited a wide range of values, from minimal playtime to extensive hours, which could potentially introduce bias and affect model convergence. Two standard normalization techniques were applied: Min-Max Scaling and Z-score normalization. Min-Max Scaling transformed the `hours_played` values into a range between 0 and 1, preserving relative differences while eliminating large disparities in magnitude. This approach helped reduce sensitivity to outliers while maintaining interpretability within a fixed scale. Additionally, Z-score normalization was used to standardize `hours_played` by centering the values around a mean of 0 with a standard deviation of 1. This method proved beneficial in cases where normally distributed input data was desirable for certain machine learning algorithms. The resulting features, `hours_played_minmax` and `hours_played_zscore`, provided flexible options for modeling and improved the robustness of downstream predictive analyses.

The decision to apply both scaling techniques offered flexibility for model selection, as different algorithms may benefit from varying data distributions. Algorithms like logistic regression, which are sensitive to feature magnitudes, often perform better with normalized input, while tree-based models, such as random forest, are generally less affected by feature scaling. However, preprocessing ensured that potential data-related biases or scaling issues were mitigated across all modeling approaches.

To facilitate binary classification tasks, the `recommendation` feature, originally represented as categorical labels (`Recommended` and `Not Recommended`), was encoded into a binary numeric format. The transformation mapped `Recommended` to 1 and `Not Recommended` to 0, resulting in a `recommendation_binary` feature. This encoding was critical for enabling the application of logistic regression and random forest classifiers, which require a binary target variable. A review of the dataset confirmed that all entries in the `recommendation_binary` column were correctly encoded, ensuring consistency and eliminating potential errors during model training.

The robustness of the binary encoding process was further validated through a thorough review of unique values in the `recommendation_binary` column. This step was necessary to confirm that no unexpected or inconsistent values were present. The encoded feature provided a clear and interpretable target for modeling, streamlining the prediction of player recommendations based on `hours_played` and other relevant features. This transformation not only facilitated the development of predictive models but also aligned with common practices in binary classification tasks, ensuring compatibility with a range of algorithms used in the study.

Model Selection and Implementation

Logistic Regression was selected for its simplicity, interpretability, and effectiveness in binary classification tasks. This model maps the relationship between the predictor variable `hours_played` and the binary outcome `recommendation_binary` using a logistic function to estimate the probability of a positive recommendation. The implementation involved splitting the dataset into training (80%) and testing (20%) sets to ensure unbiased model evaluation. The training set was used to fit the Logistic Regression model, while the testing set evaluated its performance on unseen data. During model fitting, the `solver`

parameter was set to `liblinear` due to its suitability for small datasets and binary classification. The primary goal was to estimate the relationship between gameplay hours and player recommendations, leveraging the linear regression coefficient to interpret the impact of changes in gameplay time on the likelihood of a positive recommendation.

Predictions on the testing data yielded classification outputs and probability scores for each observation. These predictions enabled the computation of key performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC scores. Logistic Regression offered a transparent view of how gameplay hours influenced recommendations, providing baseline insights for comparison with more complex models. The evaluation results highlighted that while the model performed well overall, there were limitations in capturing potential non-linear relationships in the data.

In contrast, the Random Forest Classifier was chosen for its ability to capture non-linear relationships and interactions between features. This ensemble method constructs multiple decision trees, each trained on a bootstrap sample of the training data, and aggregates their predictions to improve accuracy and robustness. Hyperparameter tuning was conducted using grid search to identify the optimal combination of parameters, such as the number of trees (`n_estimators`), maximum tree depth (`max_depth`), minimum samples per leaf (`min_samples_leaf`), and minimum samples required to split a node (`min_samples_split`). The best-performing model was determined through cross-validation, ensuring that the chosen parameters generalized well across different data subsets. The selected Random Forest model was then applied to the testing set to generate predictions.

The model's evaluation demonstrated higher predictive accuracy and recall compared to Logistic Regression, indicating its strength in capturing complex patterns within the data. Feature importance assessment further emphasized the significance of `hours_played` in predicting recommendations, providing valuable insights for understanding player engagement dynamics. Random Forest's ability to handle feature interactions and reduce overfitting through bagging techniques offered a robust approach to modeling, making it a suitable complement to the simpler Logistic Regression model.

Model Evaluation

Model evaluation focused on several performance metrics to comprehensively assess predictive accuracy and reliability. Accuracy, representing the proportion of correct predictions, was a primary measure used to compare model performance. However, accuracy alone was insufficient due to the imbalanced nature of the dataset, where the majority of players recommended their games. Precision and recall were calculated to evaluate the balance between true positives and false positives, with the F1-score serving as the harmonic mean of these two metrics, offering a balanced view of model performance. The AUC-ROC score quantified the model's ability to distinguish between the two recommendation classes, providing insights into its overall discriminative power.

Cross-validation using Stratified K-Fold further ensured model robustness and generalizability by evaluating performance across multiple folds. Logistic Regression achieved a mean AUC of 0.6109, while Random Forest

demonstrated a higher mean AUC of 0.6253, reflecting its superior capacity to capture complex relationships within the data. These findings underscored the need for both simple and complex modeling approaches, depending on the complexity and distribution of the data.

Visualization of Models

Visualization played a crucial role in interpreting model performance and results. Receiver Operating Characteristic (ROC) curves were generated for both Logistic Regression and Random Forest models, illustrating the trade-off between true positive and false positive rates. The ROC curves provided a visual representation of each model's discriminative ability, with the Random Forest curve demonstrating a slightly higher AUC, indicative of better overall performance. Confusion matrices were also constructed to depict the distribution of predicted and actual values, highlighting the number of true positives, true negatives, false positives, and false negatives for each model. This visualization offered a clear perspective on each model's predictive strengths and areas for improvement.

The Random Forest model's feature importance plot highlighted the significance of `hours_played` in influencing recommendations, confirming its critical role in player engagement analysis. Although only one feature was used in this study, the feature importance score emphasized how varying gameplay hours impacted player behavior. This visualization validated the model's interpretive capabilities and reinforced the utility of Random Forest in capturing complex, non-linear relationships in player data.

Result and Discussion

Descriptive Statistics and EDA Findings

Descriptive statistics provided critical insights into the central tendencies and variability within the hours_played feature. The mean hours_played across all entries was 88.62 hours, reflecting a substantial level of engagement among players, while the median was 44.50 hours, suggesting that half of the players spent fewer hours than this threshold. This difference between the mean and median, alongside a mode of 0.20 hours, highlighted a right-skewed distribution where a small subset of players exhibited disproportionately high playtimes. The standard deviation further confirmed significant variability in playtime, underscoring the diverse engagement levels present within the dataset. Such variations necessitated further exploration to determine whether prolonged gameplay correlated positively with player satisfaction, as reflected in their recommendations.

The distribution of player recommendations showed a marked imbalance, with 697,276 entries categorized as Recommended (encoded as 1) and 160,805 as Not Recommended (encoded as 0). This distribution indicated that approximately 81.26% of players had a positive experience with the games they reviewed, while 18.74% expressed dissatisfaction. This imbalance presented challenges in modeling, as it increased the likelihood of biased predictions. Therefore, careful consideration was given to evaluation metrics to ensure a balanced assessment of model performance across both classes.

Visual exploration of the data provided additional context to the descriptive statistics. A histogram depicting the distribution of hours_played (Figure 2)

revealed a right-skewed pattern, consistent with the statistical summary. The majority of players spent relatively few hours engaging with games, while a smaller subset exhibited extensive gameplay. This distribution emphasized the importance of understanding the factors driving both casual and intense engagement.

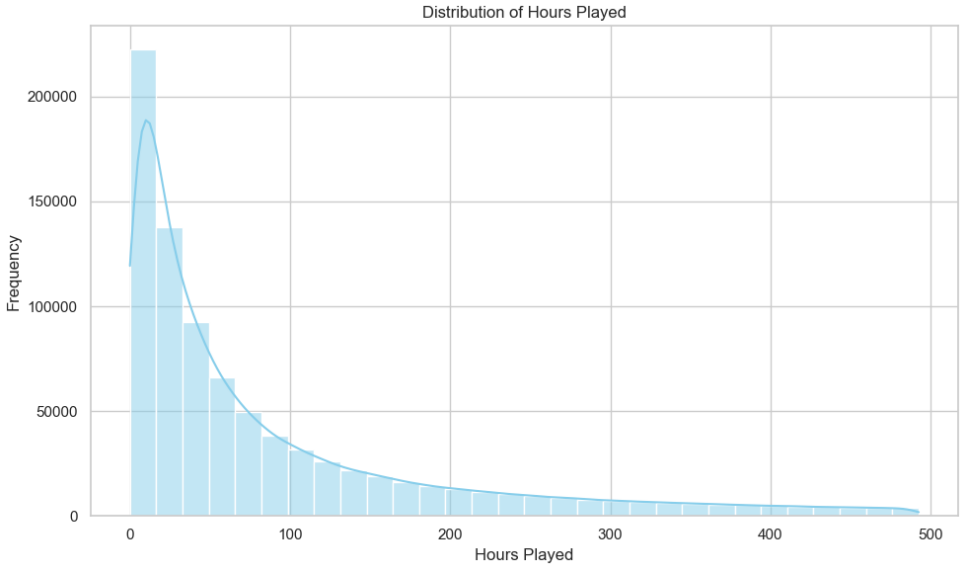

Figure 2 Distribution of Hours Played

Additionally, a box plot comparing hours_played across the two recommendation categories (Figure 3) illustrated clear differences. Players who recommended games typically exhibited higher median playtimes, suggesting that longer gameplay hours were generally associated with more favorable experiences. However, the presence of outliers within both groups indicated that extended playtime alone did not guarantee a positive recommendation, pointing to the complexity of player satisfaction dynamics.

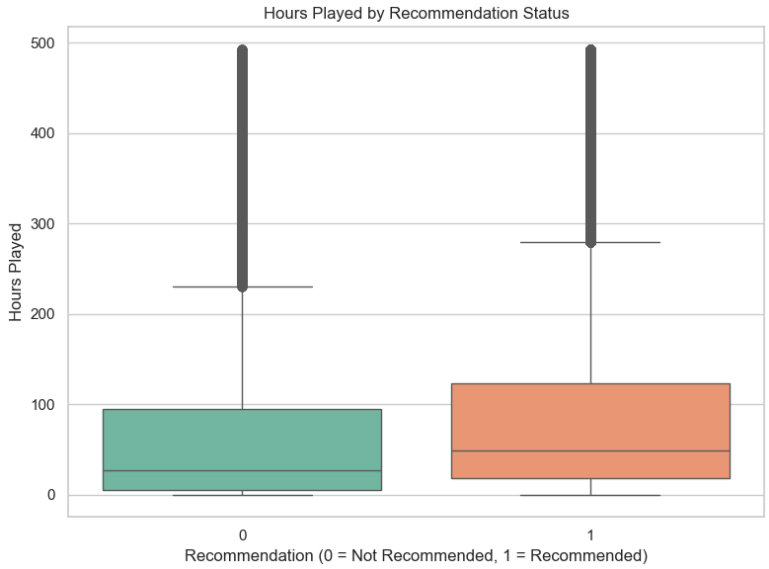

Figure 3 Boxplot of Hours Played by Recommendation Status

Model Performance Comparison

The evaluation of model performance focused on five key metrics: accuracy, precision, recall, F1-score, and the AUC-ROC (Area Under the Receiver Operating Characteristic curve). Logistic Regression achieved an accuracy of 81.26%, indicating that it correctly predicted player recommendations in approximately 81% of cases. Precision, measuring the proportion of positive identifications that were actually correct, was also 81.26%, suggesting that the model was accurate when it predicted a positive recommendation. The recall metric, which assesses the model's ability to identify all positive recommendations, was a perfect 100%. This high recall indicates that Logistic Regression successfully captured all instances of positive recommendations, albeit at the potential cost of misclassifying some negative cases as positive. The F1-score, which balances precision and recall, was 0.8966, reflecting a strong performance in terms of predictive accuracy. However, the AUC-ROC score was 0.6124, highlighting that the model's ability to discriminate between recommended and not recommended games was somewhat limited.

The Random Forest Classifier demonstrated superior performance across most metrics compared to Logistic Regression. It achieved an accuracy of 82.65%, reflecting a slight improvement in overall prediction accuracy. Precision for Random Forest was 83.59%, indicating a higher proportion of correctly identified positive recommendations compared to Logistic Regression. The recall score of 97.85% showed a slight decrease relative to Logistic Regression but still represented a strong ability to capture true positives. The F1-score, a balance between precision and recall, was 0.9016, marking an improvement over the logistic model and reflecting a better balance between identifying positive recommendations and minimizing false positives. The Random Forest model also outperformed Logistic Regression in terms of AUC-ROC, achieving a score of 0.6268, demonstrating a stronger capacity to differentiate between the two classes.

To further illustrate model performance, Receiver Operating Characteristic (ROC) curves were plotted for both Logistic Regression and Random Forest classifiers (Figure 4). The ROC curve for Random Forest displayed a more favorable trajectory, with a higher true positive rate (TPR) at various false positive rate (FPR) thresholds compared to Logistic Regression. This result reinforced the conclusion drawn from the AUC-ROC scores, with Random Forest providing superior discriminatory power between recommended and not recommended classes. The visual representation underscored the Random Forest's ability to capture nuanced patterns in the data that the linear logistic model may have overlooked.

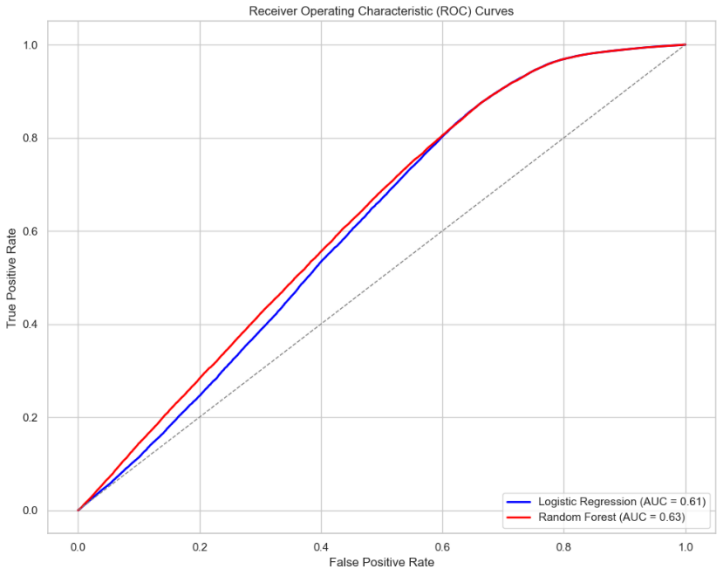

Figure 4 ROC Curves

Confusion matrices for both models provided additional clarity regarding their predictions, shown in Figure 5. The Logistic Regression confusion matrix showed a high number of true positives but also revealed a non-negligible number of false positives, aligning with its perfect recall but limited precision. The Random Forest confusion matrix, on the other hand, demonstrated a better balance between true positives and true negatives, with fewer false positives. This difference suggested that Random Forest was more effective at correctly identifying both positive and negative recommendations while maintaining high accuracy. Collectively, these visualizations offered a clear and comprehensive picture of each model's strengths and limitations in predicting player recommendations based on gameplay hours.

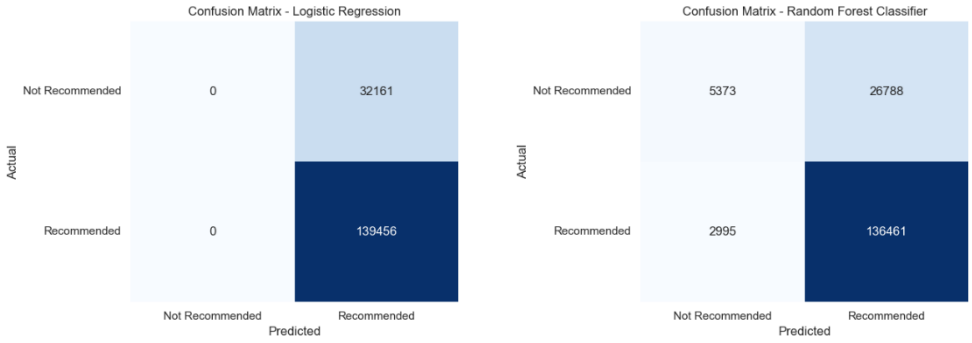

Figure 5 Confusion Matrix of LR and RF Classifier

Interpretation of Results

The relationship between `hours_played` and `recommendation` exhibited a positive correlation, indicating that players who invested more time in games were generally more likely to provide a positive recommendation. This relationship, while suggestive, was not entirely linear, as evidenced by the

variability within the data. Some players with high playtime still opted not to recommend the games they played, highlighting that gameplay hours alone are not the sole determinant of player satisfaction. The analysis demonstrated that while longer engagement often led to positive experiences, other factors such as game quality, enjoyment, and individual expectations played a significant role.

The Logistic Regression model provided a straightforward interpretation of the relationship between `hours_played` and recommendations. The model's coefficients indicated a statistically significant positive association between gameplay hours and the likelihood of a positive recommendation, with an odds ratio greater than one. This finding suggested that an increase in hours played was associated with higher odds of recommending a game. However, the linear nature of the model limited its ability to capture complex patterns and interactions, making it less adaptable to nuanced variations in the data.

In contrast, the Random Forest classifier provided deeper insights through its handling of non-linearity and interactions within the dataset. The feature importance scores highlighted `hours_played` as a critical predictor of recommendations, underscoring its influence on player satisfaction. Unlike the linear coefficients in Logistic Regression, the Random Forest model captured more intricate relationships and non-linear trends, offering a nuanced understanding of how variations in gameplay time impacted recommendations. This flexibility allowed the model to account for complex interactions that may have been overlooked by simpler approaches.

Between the two models, the Random Forest classifier demonstrated superior performance across key evaluation metrics, including accuracy, precision, and F1-score. The model's ability to capture non-linear relationships contributed to its enhanced predictive capacity, as reflected by a higher AUC-ROC score compared to Logistic Regression. The improved performance of Random Forest was attributed to its ensemble learning approach, which reduced variance and provided more robust predictions by aggregating the outputs of multiple decision trees. Additionally, Random Forest's feature importance analysis offered practical insights into the impact of gameplay hours, reinforcing its value in complex predictive modeling.

The relative underperformance of Logistic Regression could be linked to its reliance on a linear relationship between `hours_played` and recommendations, limiting its capacity to model more intricate patterns present in the data. While Logistic Regression provided a transparent and interpretable model with significant coefficients, its lack of flexibility rendered it less effective in capturing the full spectrum of relationships within the dataset. This comparison highlighted the importance of model selection based on data complexity, with Random Forest emerging as a more suitable choice for capturing the nuances of player behavior in this context.

Discussion

The findings of this study offer important implications for game developers seeking to enhance player engagement and satisfaction. The positive correlation between `hours_played` and recommendations suggests that fostering prolonged player engagement may lead to more favorable perceptions of the game. This underscores the need for game designs that promote

sustained interest, such as well-paced content updates, compelling narratives, or community-driven features that encourage longer play sessions. By strategically designing experiences that enhance immersion and replayability, developers can potentially increase the likelihood of positive recommendations, thereby leveraging word-of-mouth promotion within the gaming community.

Moreover, the variability observed among players with extensive playtime who did not recommend the games points to the importance of individualized player experiences. Developers could benefit from more granular insights into player behaviors, enabling them to tailor in-game rewards, difficulty levels, or social interaction opportunities to better align with diverse player expectations. This tailored approach could help bridge the gap between gameplay hours and satisfaction, fostering more consistent positive recommendations across a broader player base.

The results align with existing theories on player engagement and satisfaction but also reveal nuances that extend beyond simplistic linear associations. The observed positive association between `hours_played` and recommendations supports Self-Determination Theory (SDT), which emphasizes the role of autonomy, competence, and relatedness in fostering intrinsic motivation and enjoyment during gameplay. Players who engage for extended periods may experience a stronger sense of mastery and connection with game mechanics, leading to increased satisfaction and recommendations. However, the divergence observed in cases where high playtime did not lead to recommendations highlights potential areas where intrinsic motivators may be lacking or negative experiences, such as frustration or imbalance, may deter positive sentiments.

This study also builds upon prior research that links player engagement metrics to game success by providing empirical evidence through predictive modeling. The superior performance of the Random Forest classifier in capturing non-linear dynamics suggests that traditional linear models may oversimplify complex relationships between playtime and satisfaction. This finding emphasizes the importance of exploring advanced modeling techniques to better understand the multifaceted nature of player behavior, reinforcing the value of machine learning approaches in behavioral analytics.

From a practical perspective, the study’s findings can be leveraged by game developers and marketers to enhance player retention strategies and targeted marketing campaigns. Identifying players who engage deeply with a game but do not provide positive recommendations can help inform personalized interventions, such as targeted in-game incentives or feedback requests, aimed at improving satisfaction levels. This approach could foster a more engaged and loyal player base, ultimately driving higher recommendation rates and improved player retention metrics.

Additionally, the insights gained from the Random Forest model’s feature importance scores highlight the potential to refine recommendation algorithms on platforms like Steam. By incorporating nuanced metrics such as gameplay hours and their impact on player satisfaction, recommendation systems can be optimized to prioritize games likely to resonate with specific player segments. This data-driven approach can enhance user experiences and bolster sales by aligning game recommendations more closely with player preferences and

behaviors. In summary, the application of predictive modeling techniques offers actionable pathways for game developers and platforms to maximize player satisfaction, engagement, and long-term success.

Limitations

This study faced several data-related constraints that limited the scope of its findings. The analysis relied solely on the columns available in the `steam_game_reviews.csv` dataset, with key features like `hours_played` and `recommendation`. However, important contextual factors such as game genre, user demographics, and detailed sentiment analysis of reviews were absent. This omission may have led to a partial understanding of the factors influencing player recommendations, as the inclusion of genre-specific features or demographic data (e.g., age, region, or play preferences) could have provided more nuanced insights into player engagement and satisfaction patterns.

Additionally, the dataset only represented the snapshot of player behavior reflected in the available records, limiting the potential to capture evolving engagement trends over time. The absence of longitudinal data further restricted the ability to understand how player engagement and satisfaction may change with subsequent updates or content releases for games. This constraint emphasized the need for broader and more diverse datasets in future research to develop more comprehensive predictive models.

The two models used, Logistic Regression and Random Forest, each presented inherent limitations. Logistic Regression, being a linear model, assumes a linear relationship between `hours_played` and player recommendations. This assumption likely oversimplified the complex interactions present within the data, potentially leading to reduced predictive accuracy and an inability to capture non-linear relationships effectively. Conversely, the Random Forest classifier, while more flexible and capable of capturing non-linearities, posed a risk of overfitting, particularly when handling datasets with many correlated features or complex structures. The hyperparameter tuning process and cross-validation techniques mitigated some of these risks, but overfitting remains a potential concern that could limit the model's generalizability to unseen data.

The computational complexity and interpretability of Random Forest also present challenges. Unlike the simpler coefficient-based interpretation offered by Logistic Regression, understanding the nuanced decision-making of hundreds of trees can be difficult. While feature importance scores provided some insight, they lack the straightforward interpretability of traditional models, making practical application more challenging for developers and analysts unfamiliar with complex ensemble methods.

The findings of this study may be limited in their applicability beyond the Steam platform and the specific dataset used. Steam represents one of many digital distribution platforms, and while it holds a dominant position in the gaming market, player behaviors and recommendation patterns may differ on other platforms, such as Epic Games Store or console-based ecosystems. Additionally, the data used in this study reflected a particular point in time and may not fully capture broader trends across the gaming industry or in different cultural and regional contexts. Generalizability beyond the sample used is therefore limited, highlighting the need for broader validation across diverse datasets and platforms.

Future Research Directions

Future research could benefit from incorporating a wider range of variables to enhance the predictive power and interpretability of models. Including features such as game genre, user demographics (e.g., age, gender, geographic region), and detailed sentiment analysis of reviews could offer a more holistic view of player behavior and satisfaction. By accounting for genre-specific differences or demographic preferences, models may better capture the unique factors driving engagement and recommendations, leading to more tailored strategies for developers and marketers.

Sentiment analysis of player reviews could also provide deeper insights into the qualitative factors influencing recommendations, such as emotional attachment to gameplay elements, frustrations with game mechanics, or social interactions within multiplayer games. Combining these additional features with existing predictors would enable a richer and more nuanced understanding of player dynamics.

Exploring advanced modeling techniques could further enhance the predictive accuracy and robustness of models used in future studies. Gradient Boosting Machines (GBM), such as XGBoost or LightGBM, offer an alternative ensemble approach that often outperforms Random Forest by focusing on iterative improvement of weak learners. Neural Networks, particularly deep learning models, could also provide valuable insights by capturing complex patterns and interactions that simpler models might overlook. These advanced methods, while computationally intensive, offer potential for uncovering deeper relationships within the data, particularly when paired with large, diverse datasets.

Conducting longitudinal studies could offer a more comprehensive view of how player engagement and satisfaction evolve over time. Analyzing changes in `hours_played` and player recommendations as games receive updates, expansions, or community-driven content would shed light on the temporal dynamics of player behavior. Such studies would enable researchers to assess how factors like content freshness, community interaction, and evolving gameplay experiences impact long-term engagement and satisfaction. This approach would provide actionable insights for game developers seeking to foster sustained player loyalty and positive recommendations through ongoing content and engagement strategies.

Conclusion

The study provided valuable insights into the relationship between gameplay hours and player recommendations on the Steam platform. Results indicated a positive correlation between the number of hours played and the likelihood of recommending a game, suggesting that greater player engagement generally enhanced satisfaction levels. However, the data also revealed exceptions, as not all high-engagement players issued positive recommendations, pointing to other influential factors. The comparative analysis between Logistic Regression and Random Forest classifiers further demonstrated distinct strengths and weaknesses. While Logistic Regression offered straightforward interpretability with a statistically significant positive association, it struggled to capture non-linear relationships present in the data. On the other hand, the Random Forest model exhibited superior predictive performance, effectively accounting for

complex interactions but posing challenges in interpretability.

This study contributes to the understanding of player engagement metrics by empirically examining how gameplay hours influence player satisfaction and recommendations. The findings support the importance of engagement duration as a key factor in player experiences, aligning with theories of player retention and motivation. By comparing two different modeling approaches, the research highlights the strengths of machine learning techniques in capturing nuanced behavioral patterns. The study offers practical insights for game developers and marketers seeking to optimize player engagement and improve recommendation rates. The evidence underscores the potential of leveraging gameplay data to tailor game features and marketing strategies, thereby enhancing overall player satisfaction and fostering long-term player loyalty.

Based on the study's findings, game developers and marketers are encouraged to leverage gameplay hours data strategically to enhance player satisfaction. Features and updates that encourage extended play sessions, such as engaging storylines, regular content releases, and social interactions, can lead to increased positive recommendations. Furthermore, targeting players with personalized experiences based on their play patterns may enhance engagement and satisfaction, improving word-of-mouth recommendations. Marketers can utilize insights from predictive models to tailor promotional efforts, ensuring that campaigns resonate with players who demonstrate high engagement but have not yet converted to promoters of the game. Enhancing recommendation rates requires a comprehensive approach, integrating in-game content design with data-driven marketing strategies.

The study underscores the importance of data-driven decision-making in the gaming industry. Leveraging player engagement metrics, such as gameplay hours, offers a path to deeper understanding of player behavior and satisfaction, providing actionable insights for improving game design and marketing strategies. As gaming platforms and player preferences continue to evolve, ongoing research is essential to refine predictive models and identify additional factors influencing player engagement. Building upon this study's findings, future research can incorporate a broader range of variables, advanced modeling techniques, and longitudinal data to further enrich the understanding of player dynamics and maximize the impact of data-driven strategies in the gaming industry.

Declarations

Author Contributions

Conceptualization: Y.D.; Methodology: Y.D.; Software: A.W.A.R.; Validation: Y.D.; Formal Analysis: A.W.A.R.; Investigation: Y.D.; Resources: A.W.A.R.; Data Curation: A.W.A.R.; Writing—Original Draft Preparation: Y.D.; Writing—Review and Editing: A.W.A.R.; Visualization: A.W.A.R. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. D. Joseph, “Distributing Productive Play: A Materialist Analysis of Steam,” 2021.

[2] A. M. Thorhauge, “The Steam Platform Economy: From Retail to Player-Driven Economies,” New Media Soc., vol. 26, no. 4, pp. 1963–1983, 2022.

[3] J. D. Chisholm and A. Kingstone, “Action Video Games and Improved Attentional Control: Disentangling Selection- And Response-Based Processes,” Psychon. Bull. Rev., vol. 22, no. 5, pp. 1430–1436, 2015.

[4] R. Sifa, A. Drachen, and C. Bauckhage, “Large-Scale Cross-Game Player Behavior Analysis on Steam,” Proc. Aaai Conf. Artif. Intell. Interact. Digit. Entertain., vol. 11, no. 1, pp. 198–204, 2021.

[5] C. Bauckhage, K. Kersting, R. Sifa, C. Thurau, A. Drachen, and A. Canossa, “How Players Lose Interest in Playing a Game: An Empirical Study Based on Distributions of Total Playing Times,” 2012.

[6] E. A. E. Ahmed et al., “Comparison of Specific Segmentation Methods Used for Copy Move Detection,” Int. J. Electr. Comput. Eng. Ijece, vol. 13, no. 2, p. 2363, 2023.

[7] M. Liebenlito et al., “Active learning on Indonesian Twitter sentiment analysis using uncertainty sampling,” J. Appl. Data Sci., vol. 5, no. 1, Art. no. 1, 2024.

[8] B. Berlilana et al., “Exploring the Impact of Discount Strategies on Consumer Ratings,” J. Appl. Data Sci., vol. 5, no. 1, Art. no. 1, 2024.

[9] A. R. Hananto and B. Srinivasan, “Comparative Analysis of Ensemble Learning Techniques for Purchase Prediction,” J. Digit. Mark. Digit. Curr., vol. 1, no. 2, Art. no. 2, 2024.

[10] B. H. Hayadi and I. M. M. E. Emary, “Predicting Campaign ROI Using Decision Trees and Random Forests,” J. Digit. Mark. Digit. Curr., vol. 1, no. 1, Art. no. 1, 2024.

[11] A. R. Hananto and D. Sugianto, “Analysis of the Relationship Between Trading Volume and Bitcoin Price Movements Using Pearson and Spearman Correlation Methods,” J. Curr. Res. Blockchain, vol. 1, no. 1, Art. no. 1, Jun. 2024.

[12] Hery and A. E. Widjaja, “Predictive Modeling of Blockchain Stability Using Machine Learning to Enhance Network Resilience,” J. Curr. Res. Blockchain, vol. 1, no. 2, Art. no. 2, Sep. 2024.

[13] S. Yadav and A. R. Hananto, “Comprehensive Analysis of Twitter Conversations Provides Insights into Dynamic Metaverse Discourse Trends,” Int. J. Res. Metaverese, vol. 1, no. 1, Art. no. 1, Jun. 2024.

[14] T. Wahyuningsih and S. C. Chen, “Determinants of Virtual Property Prices in Decentraland an Empirical Analysis of Market Dynamics and Cryptocurrency Influence,” Int. J. Res. Metaverese, vol. 1, no. 2, Art. no. 2, Sep. 2024.

[15] D. L. King, M. C. E. Herd, and P. Delfabbro, “Tolerance in Internet Gaming Disorder: A Need for Increasing Gaming Time or Something Else?,” J. Behav. Addict., vol. 6, no. 4, pp. 525–533, 2017.

[16] M. Kosa and A. Uysal, “The Role of Need Satisfaction in Explaining Intentions to Purchase and Play in Pokémon Go and the Moderating Role of Prior Experience.,” Psychol. Pop. Media, vol. 10, no. 2, pp. 187–200, 2021.

[17] F. D. Buono et al., “Delay Discounting of Video Game Players: Comparison of Time Duration Among Gamers,” Cyberpsychology Behav. Soc. Netw., vol. 20, no. 2, pp. 104–108, 2017.

[18] N. Alrahili, “The Prevalence of Video Game Addiction and Its Relation to Anxiety, Depression, and Attention Deficit Hyperactivity Disorder (ADHD) in Children and Adolescents in Saudi Arabia: A Cross-Sectional Study,” Cureus, 2023.

[19] C. M. Ohannessian, “Video Game Play and Anxiety During Late Adolescence: The Moderating Effects of Gender and Social Context,” J. Affect. Disord., vol. 226, pp. 216–219, 2018.

[20] B. Walia, J. Kim, I. Ijere, and S. Sanders, “Video Game Addictive Symptom Level, Use Intensity, and Hedonic Experience: Cross-Sectional Questionnaire Study,” Jmir Serious Games, vol. 10, no. 2, p. e33661, 2022.

[21] D. Ghuman and M. D. Griffiths, “A Cross-Genre Study of Online Gaming,” Int. J. Cyber Behav. Psychol. Learn., vol. 2, no. 1, pp. 13–29, 2012.

[22] H. Xiong and J. Yu, “Virtual Goods Purchase, Game Satisfaction and Perceived Justice: An Empirical Study of Players of PVP Mobile Games,” J. Virtual Worlds Res., vol. 13, no. 2–3, 2020.

[23] J. Merikivi, D. T. Nguyen, and V. K. Tuunainen, “Understanding Perceived Enjoyment in Mobile Game Context,” 2016.

[24] T. Cole, D. J. K. Barrett, and M. D. Griffiths, “Social Facilitation in Online and Offline Gambling: A Pilot Study,” Int. J. Ment. Health Addict., vol. 9, no. 3, pp. 240–247, 2010.

[25] O. Király et al., “Policy Responses to Problematic Video Game Use: A Systematic Review of Current Measures and Future Possibilities,” J. Behav. Addict., vol. 7, no. 3, pp. 503–517, 2017.

[26] C. Hellström, K. W. Nilsson, J. Leppert, and C. Åslund, “Effects of Adolescent Online Gaming Time and Motives on Depressive, Musculoskeletal, and Psychosomatic Symptoms,” Ups. J. Med. Sci., vol. 120, no. 4, pp. 263–275, 2015.

[27] N. Johannes, M. Vuorre, and A. K. Przybylski, “Video Game Play Is Positively Correlated With Well-Being,” R. Soc. Open Sci., vol. 8, no. 2, 2021.

[28] E. Petrovskaya and D. Zendle, “The Relationship Between Psycho-Environmental Characteristics and Wellbeing in Non-Spending Players of Certain Mobile Games,” R. Soc. Open Sci., vol. 10, no. 1, 2023.

[29] M. Vuorre, N. Johannes, K. Magnusson, and A. K. Przybylski, “Time Spent Playing Video Games Is Unlikely to Impact Well-Being,” R. Soc. Open Sci., vol. 9, no. 7, 2022.

[30] W. Peng, J.-H. Lin, K. A. Pfeiffer, and B. Winn, “Need Satisfaction Supportive Game Features as Motivational Determinants: An Experimental Study of a Self-Determination Theory Guided Exergame,” Media Psychol., vol. 15, no. 2, pp. 175–196, 2012.

[31] A. K. Przybylski, C. S. Rigby, and R. M. Ryan, “A Motivational Model of Video Game Engagement,” Rev. Gen. Psychol., vol. 14, no. 2, pp. 154–166, 2010.

[32] C. Harteveld and S. C. Sutherland, “Personalized Gaming for Motivating Social and Behavioral Science Participation,” 2017.

[33] C.-I. Teng, S.-K. Lo, and Y.-J. Li, “How Can Achievement Induce Loyalty? A Combination of the Goal-Setting Theory and Flow Theory Perspectives,” Serv. Sci., vol. 4, no. 3, pp. 183–194, 2012.

[34] S. Bigdeli et al., “Underpinning Learning Theories of Medical Educational Games: A Scoping Review,” Med. J. Islam. Repub. Iran, 2023.

[35] D. J. Mills, M. Milyavskaya, N. L. Heath, and J. L. Derevensky, “Gaming Motivation and Problematic Video Gaming: The Role of Needs Frustration,” Eur. J. Soc. Psychol., vol. 48, no. 4, pp. 551–559, 2017.

[36] A. Drachen et al., “Rapid Prediction of Player Retention in Free-to-Play Mobile Games,” 2016.

[37] V. E. Adeyemo, “Identification of Pattern Mining Algorithm for Rugby League Players Positional Groups Separation Based on Movement Patterns,” Plos One, vol. 19, no. 5, p. e0301608, 2024.

[38] N. Yahyasoltani, “Learning Performance Efficiency of College Basketball Players Using TVAE,” Ieee Access, vol. 11, pp. 130186–130196, 2023.

[39] R. Couronné, P. Probst, and A. Boulesteix, “Random Forest Versus Logistic Regression: A Large-Scale Benchmark Experiment,” BMC Bioinformatics, vol. 19, no. 1, 2018.

[40] P. C. Austin and F. E. Harrell, “Predictive Performance of Machine and Statistical Learning Methods: Impact of Data-Generating Processes on External Validity in the ‘Large N, Small P’ Setting,” Stat. Methods Med. Res., vol. 30, no. 6, pp. 1465–1483, 2021.

[41] M. M. Churpek, T. C. Yuen, C. Winslow, D. O. Meltzer, M. W. Kattan, and D. P. Edelson, “Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards,” Crit. Care Med., vol. 44, no. 2, pp. 368–374, 2016.

[42] C.-Y. Kuo, L.-C. Yu, H.-C. Chen, and C.-L. Chan, “Comparison of Models for the Prediction of Medical Costs of Spinal Fusion in Taiwan Diagnosis-Related Groups by Machine Learning Algorithms,” Healthc. Inform. Res., vol. 24, no. 1, p. 29, 2018.

[43] J. Holewik, G. Schaefer, and I. Korovin, “Imbalanced Ensemble Learning for Enhanced Pulsar Identification,” pp. 515–524, 2020.

[44] M. H. A. Hamid, M. Yusoff, and A. Mohamed, “Survey on Highly Imbalanced Multi-Class Data,” Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 6, 2022.

[45] B. Li and C. Vogel, “Improving Multiclass Text Classification With Error-Correcting Output Coding and Sub-Class Partitions,” pp. 4–15, 2010.

[46] H. Xue and S. Chen, “Orthogonality-based Label Correction in Multi-class Classification,” Electron. Lett., vol. 49, no. 12, pp. 754–756, 2013.

[47] V. Fuh-Ngwa et al., “Ensemble Machine Learning Identifies Genetic Loci Associated With Future Worsening of Disability in People With Multiple Sclerosis,” Sci. Rep., vol. 12, no. 1, 2022.

[48] P. Yang, Y. H. Yang, B. B. Zhou, and A. Y. Zomaya, “A Review of Ensemble Methods in Bioinformatics,” Curr. Bioinforma., vol. 5, no. 4, pp. 296–308, 2010.

[49] M. Pratama, W. Pedrycz, and E. Lughofer, “Evolving Ensemble Fuzzy Classifier,” Ieee Trans. Fuzzy Syst., vol. 26, no. 5, pp. 2552–2567, 2018.

[50] V. Kulkarni and P. K. Sinha, “Pruning of Random Forest Classifiers: A Survey and Future Directions,” 2012.

[51] B. Xu, J. Z. Huang, G. J. Williams, M. J. Li, and Y. Ye, “Hybrid Random Forests: Advantages of Mixed Trees in Classifying Text Data,” pp. 147–158, 2012.

[52] S. Bernard, L. Heutte, and S. Adam, “On the Selection of Decision Trees in Random Forests,” 2009.