ISSN 2959-6157

Predicting Online Gaming Behavior: A Comparative Analysis of Random Forest and Gaussian Naive Bayes Models

Tianchen Xie

Department of Computer Science, New York University (Shanghai), Shanghai, China
tx813@nyu.edu

Abstract:

With the rising impact and economic benefits of the online gaming industry, online gaming companies need to understand players' behavior to evaluate their products better and adjust their developing strategy. This paper uses machine learning models to predict players' online gaming behaviors. The study utilizes the dataset from Kaggle, which contains 40,034 samples with 13 features related to player demographics, game preferences, and gameplay behaviors. The primary objective is to classify two target variables: 'EngagementLevel' and 'InGamePurchases.' Data preprocessing steps included handling missing values, encoding categorical features, normalizing numerical data, and addressing class imbalances using the Synthetic Minority Over-sampling Technique (SMOTE). The Random Forests (RF) and Gaussian Naïve Bayes (GNB) models were employed to predict the target variables. The RF model significantly outperformed the GNB model in both classification tasks, especially when predicting the in-game purchases. The RF model's robustness is attributed to its ability to handle complex, non-linear relationships and interactions between features, while the GNB model's performance was limited by its assumptions of feature independence and normally distributed data. The findings suggest that RF is a more effective tool for predicting online gaming behaviors, particularly in scenarios with complex feature interactions. Future research could incorporate additional machine learning models and more diverse datasets to further enhance predictive accuracy and offer a more comprehensive understanding of online gaming behaviors.

Keywords: Online gaming; machine learning; artificial intelligence.

1. Introduction

Video games usually refer to electronic games where players interact with a visual display using controls to achieve specific objectives or explore virtual worlds. Generally, video games can be divided into two basic forms: online and offline. The latter often have fixed start and finish points, enabling players to achieve by themselves. On the other hand, online games have constantly evolving tasks added by game developers and thus usually need players to cooperate in real-time [1]. Since the late 1990s and 2000s, the integration of online multiplayer features and expansion with consoles has transformed gaming into a more social experience. The rise of broadband internet further accelerated this trend, making online gaming a more popular form than offline gaming. After nearly 20 years of development, the online gaming market has already become a neglectable and profitable industry. The global online gaming market size was valued at USD 87.22 billion in 2023 and is expected to reach USD 229.85 billion by 2033, according to a research report published by Spherical Insights & Consulting [2]. To stand out from the competitive market, online game companies need to predict users' behavior so that they can adjust their development plan and improve their game quality accordingly, especially players' engagement level and whether they will make in-game purchases. Since machine learning models have strong feature extraction and prediction capability, it is ideal to apply them in the scenario of online gaming behavior predictions.

Many former studies have already applied machine learning models to predict multiple aspects related to online gaming. For instance, a study carried out by Sanghvi has applied the cooperation between artificial intelligence and blockchain technology to establish a robust system for detecting and moderating hate speech in online gaming. The study showcases the effectiveness of machine learning models, particularly gradient boosting, in accurately analyzing and categorizing hate speech, achieving an accuracy rate of 86.01% [3]. Faraz's study combines fastText and machine learning classifiers such as support vector machine, k-nearest neighbors, random forests, and

XGBoostmto to build a comprehensive framework that mitigates predatory behavior on various gaming platforms. The classifier achieves metrics of accuracy, precision, recall, F1-score, and F0.5-score at 0.99 for all when evaluated on the PAN12 dataset [4]. Ribeiro and Bao's study applied principal component analysis, least absolute shrinkage, and selection operator, and a novel continual learning modeling approach based on the combination of random forests with ordinary least squares to validate the relationship between the revenue change of the influential online gamer and increasing intensity of the indirect network effect exerted on viewers [5]. The studies show that machine learning models have already proved their strength in making predictions and extracting features. Therefore, this paper aims to apply machine learning models to predict players' engagement levels and purchasing tendencies.

In this regard, this paper used the source from the Kaggle website as the dataset. The random forests (RF) and Gaussian naïve Bayes (GNB) models were applied to predict the target feature 'engagement level' and 'in-game purchases' in the dataset and analyze the feature importance when predicting. The result of the prediction validates the availability of the research method.

2. Method

2.1 Dataset Preparation

The dataset utilized in the study was obtained from the Kaggle website [6]. The dataset has 40, 034 data items and 13 features, of which 8 are numerical and 5 are categorical. The features include player demographics ("PlayerID", "Age", "Gender", "Location"), game characteristics ("GameGenre", "GameDifficulty"), and gameplay behaviors ("PlayTimeHours", "InGamePurchases", "SessionsPerWeek", "AvgSessionDurationMinutes", "PlayerLevel", "AchievementsUnlocked", "EngagementLevel"). The primary objective of this research was to classify two target variables: 'InGamePurchases' and 'EngagementLevel.'The former is a numerical feature, with 0 representing not having made in-game purchases and 1 representing having made in-game purchases. The latter is a categorial feature with three levels: low, medium, and high, showing to what extent have players engaged in the game.

However, two significant issues were identified within the raw dataset. First, most of the numerical features exhibit non-normal distributions. As GNB assumes the features follow a normal distribution, non-normal features may lead to biased predictions and lower accuracy. Second, the 'InGamePurchases' target variable is highly imbalanced, with nearly 80% of instances falling into Category 0 and only 20% into Category 1, as illustrated in Fig. 1. Imbalanced data is a challenge for standard algorithms, affecting their performance since such algorithms were designed to maximize accuracy, and this measure is biased towards the majority class [7]. To address these challenges, comprehensive data preprocessing was undertaken.

The preprocessing phase in this study involved five key steps. Initially, the dataset was inspected for missing and duplicate values using the functions from the Pandas library; no such values were detected. Next, the 'PlayerID' feature was excluded from the dataset since it was irrelevant to the prediction. To enhance the classification algorithms' performance and standardize the feature space, functions from the Scikit-learn library were applied to encode the categorical variables and normalize the numerical features.

The final step was splitting the dataset into training and testing subsets. In this study, 30% of the data was allocated for testing and the random state was set to 42 to ensure consistency across analyses. To address the issue of class imbalance in the 'InGamePurchases' variable, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE is an over-sampling approach in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement [8]. It helps balance the class distribution by taking each minority class sample and introducing synthetic examples along the line segments connecting that sample to its nearest neighbors in the feature space.

Fig. 1 Distribution of in-game purchases (Photo/Picture credit: Original)

2.2 Machine Learning-based Prediction

In this study, RF and GNB from Scikit-learn were used. The performance of both models was evaluated using the confusion matrix result, including their precision, recall, F1-score, and overall accuracy.

2.2.1 Random forests

The random forests (RF) model is a supervised classification method based on the combination of Breiman's "bagging" and random selection of features to construct a collection of decision trees with controlled variance [9]. Each decision tree is trained on different parts of the same training data and then combined to produce a more accurate and stable outcome than any single tree would produce. Because of the Law of Large Numbers, they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors [10].

One of the most significant hyperparameters of an RF model is the number of decision trees. Increasing the number of decision trees generally improves the model's performance. However, it will also lead to drawbacks such as increased computational cost and complexity. To optimize the model's performance, the grid search was applied to find the best tree number from 200 to 1000. Grid Search finds the optimal number of trees by traversing different values of n_estimators in the parameter grid and using cross-validation to evaluate the performance of each model. This process ensures that the number of trees chosen gives good generalization over unseen data. The intrinsic feature importances were also shown to help interpret the model and compare each feature's contribution.

2.2.2 Gaussian naïve bayes

Naïve Bayes is a set of supervised learning algorithms that apply Bayes' theorem with the "naive" assumption of independence between every pair of features [11]. Gaussian Naïve Bayes (GNB) is a variant of the Naïve Bayes classification method specifically designed for continuous features that are assumed to follow a Gaussian (normal) distribution. This approach allows for efficient computation of probabilities for continuous data by estimating the mean and variance of the features for each class. During classification, GNB calculates the likelihood of a sample belonging to each class based on these Gaussian distributions and assigns the class with the highest posterior probability. In this paper, the hyperparameters of GNB were set to be the default value.

3. Results and Discussion

3.1 Engagement Level Prediction

Grid search was applied to the RF model to find the best number of decision trees before making the prediction. As Fig. 2 shows, when the number of decision trees was equal to 500, the mean accuracy was the highest and had the least time consumption. Therefore, the decision tree number was set to 500 when predicting the engagement level.

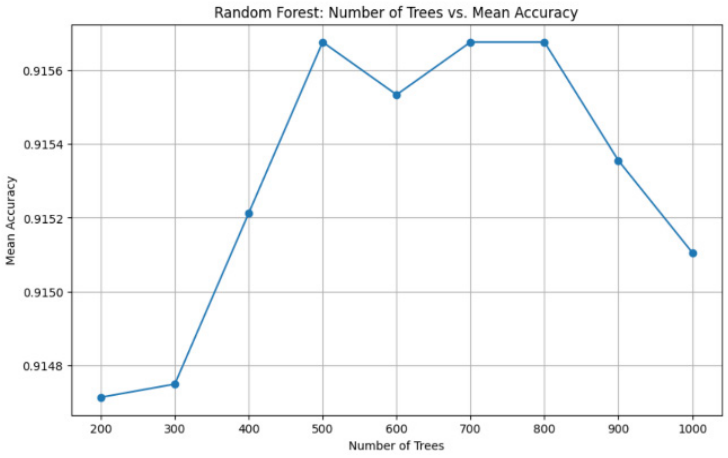

Fig. 2 Grid search result for RF model in engagement level prediction (Photo/ Picture credit: Original)

Then the two models were applied to predict the engagement level, the result was shown in Table 1. The results proved the superiority of RF in most metrics. For precision, the two models performed only a slight difference in Class Low and Medium. However, RF showed a significantly higher precision (0.91) compared to GNB (0.79) for class high, which meant it was better in correctly identifying instances of high engagement level. For recall, RF consistently outperforms GNB for Classes low and medium with values of 0.88 for both classes, compared to 0.77 and 0.69 respectively for GNB. This indicates that RF is more effective at capturing the true instances of these classes. Combining the precision and recall, the F1-score further highlights RF's superior performance. RF achieves higher F1-scores across all classes: 0.90 for Class Low, 0.89 for Class Medium, and 0.93 for Class High, while GNB only achieved 0.84, 0.77, and 0.86 respectively for each class. This suggests that RF maintains a better balance between precision and recall, leading to more reliable overall performance. The overall accuracy of the RF model is 0.91, which is notably higher than the 0.84 achieved by GNB.

The superior performance of the RF model can be attributed to its ability to handle complex, non-linear relationships between features and its robustness against noise and irrelevant features. In contrast, GNB, which assumes feature independence and normally distributed data, may struggle with the complexities of the dataset, leading to lower performance metrics. These findings suggest that RF is a more appropriate choice for this classification problem, especially when high recall and balanced performance across classes are crucial.

Table 1. Prediction results for engagement level

Metric	Class	RF	GNB
Precision	Low	0.92	0.93
	Medium	0.91	0.89
	High	0.91	0.79
Recall	Low	0.88	0.77
	Medium	0.88	0.69
	High	0.95	0.95
F1-Score	Low	0.90	0.84
	Medium	0.89	0.77
	High	0.93	0.86
Accuracy		0.91	0.84

Fig. 3 displays each feature's importance of RF when predicting the engagement level. Features related to playing time were found to contribute to the model's prediction significantly. "SessionsPerWeek", "AvgSessionDura-

tionMinutes", and "PlayTimeHours" respectively ranked first, second, and fourth. Meanwhile, "PlayerLevel" also showed a high importance and ranked third. Demographic and game-related features like "Age," "GameGenre," "Location," "GameDifficulty," and "Gender" have relatively lower importance scores. These results suggest that behavioral metrics, especially how long players spend playing games, are generally more crucial for understanding engagement than demographic or game-specific factors.

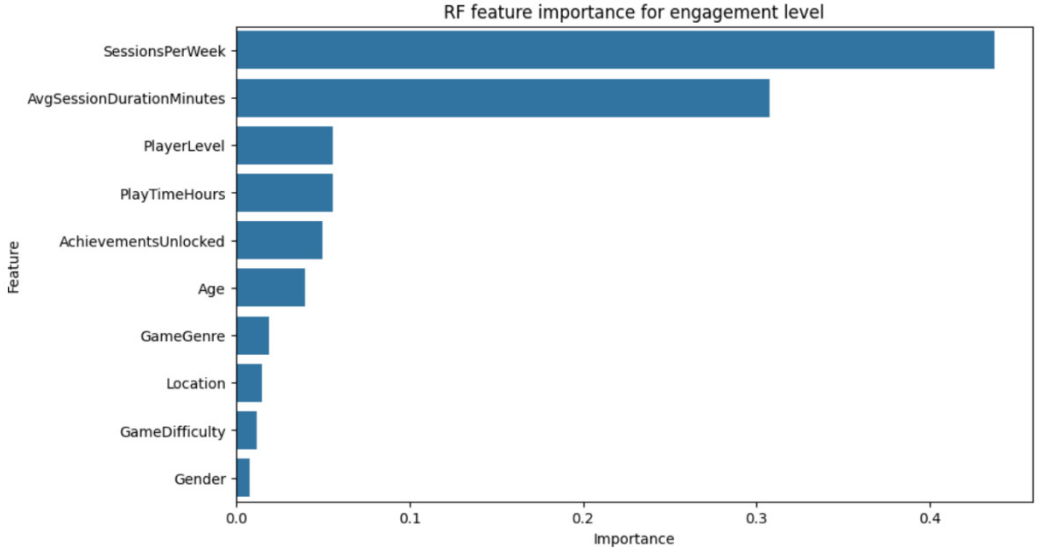

Fig. 3 RF feature importance for engagement level (Photo/Picture credit: Original).

3.2

In-game Purchases Prediction

Fig. 4 shows that when the number of decision trees was 900, the mean accuracy was the highest, slightly higher than the accuracy when the decision tree number was 700. Therefore, the hyperparameter was set to 900 and repeated the aforesaid prediction process. The results are shown in Table 2.

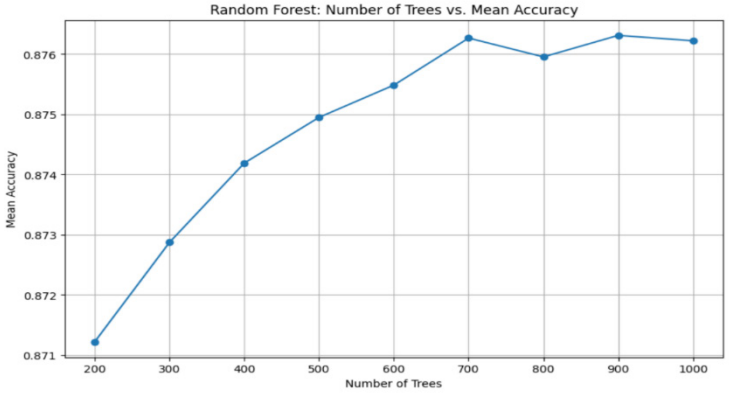

Fig. 4 Grid search result for RF model in in-game purchases prediction (Photo/Picture credit: Original)

Table 2 shows that RF is significantly better suited for this binary classification task. Among the precision, recall, and F1-scores of both classes, GNB all performed very nonideally with results only around 0.54. As a comparison, the three metrics of RF all exceeded 0.8 and RF performed especially ideally in the precision of Class Not purchased and the recall of Class Purchased. This indicates RF's reliability in identifying true non-purchasers and capturing the majority of true purchasers. The overall accuracy of RF is 0.88, which is considerably higher than the 0.54 accuracy achieved by GNB.

Table 2. Prediction results for in-game purchases

Metric	Class	RF	GNB
Precision	Purchased	0.83	0.54
	Not purchased	0.95	0.54
Recall	Purchased	0.96	0.53
	Not purchased	0.80	0.55
F1-Score	Purchased	0.89	0.54
	Not purchased	0.87	0.55
Accuracy		0.88	0.54

When predicting the in-game purchase, RF still showed its reliability due to its nature as a non-linear ensemble method. However, GNB performed extremely nonideally because several of its assumptions based on the Bayes theory were not satisfied. First, GNB assumes that the numerical features follow a Gaussian distribution within each class. It expects the data to have a bell-shaped curve with most values concentrated around the mean. Such an assumption did not hold in this dataset as shown in Fig. 1: the distribution of the in-game purchase is significantly left-skewed. Also, contrary to RF, GNB is a linear classifier, meaning it cannot capture complex, non-linear relationships between features and the target variable.

Fig. 5 shows the feature importance of RF when predicting the in-game purchase. Though "SessionsPerWeek" again emerges as the most important feature, the demographic feature "Age" and game-specific feature "AchievementsUnlocked" appear as the second and third most important features. The rising importance may be attributed to the different levels of financial independence or spending habits affected by age and psychological stages that have evolved with the increasing number of achievements unlocked. This may indicate that, unlike engagement level prediction, the in-game purchase prediction needs to consider both the players' time consumption as well as the players' characteristics.

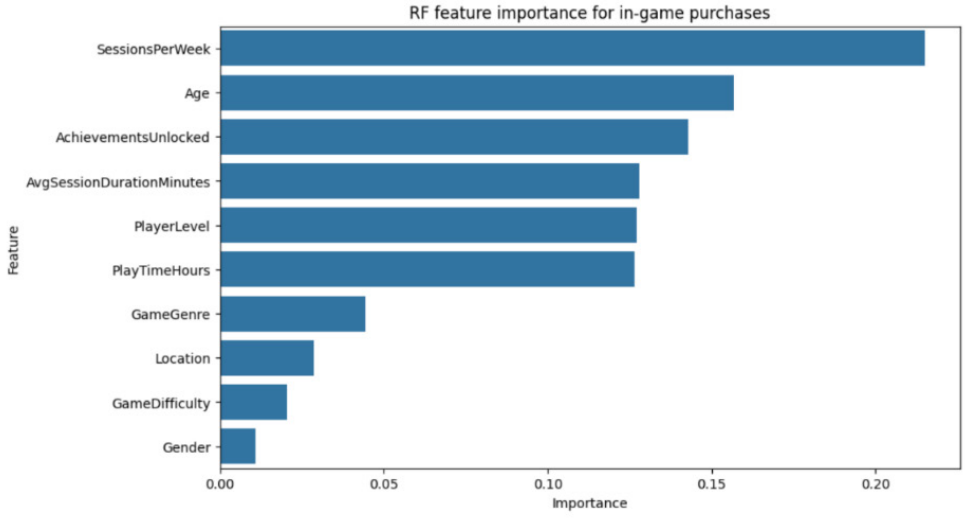

Fig. 5 RF feature importance for in-game purchases (Photo/Picture credit: Original)

4. Conclusion

In this work, machine learning algorithms were applied to predict online gaming behaviors including players' engagement level and whether they would make in-game purchases. The random forests and Gaussian Naïve Bayes algorithms were used to classify the two target features in the dataset. The results demonstrated that the RF model outperformed GNB in both classification tasks. This superiority can be attributed to RF's ability to handle complex, non-linear relationships in the data, while GNB's performance was limited by its assumptions of feature independence and normal distribution, which did not hold for the dataset used. Despite these findings, the study's limitations include reliance on only two machine learning models and a single dataset. Future research could explore

additional models, such as XGBoost or deep learning approaches, to further improve prediction accuracy. Additionally, expanding the dataset or incorporating more diverse features could offer a more comprehensive analysis of online gaming behavior.

References

[1] Griffiths M. Problematic online gaming: Issues, debates and controversies. Медицинская психология в России. 2015(4 (33)):5.

[2] Yahoo Finance, Global Online Gaming Market Size [2024-2032]. 2023 Aug 8. Available from: https://finance.yahoo.com/news/global-online-gaming-market-size-050000553.html. Accessed 2024 Aug 10.

[3] Sanghvi H, Bhavsar R, Hundlani V, Gohil L, Vyas T, Nair A, et al. MetaHate: AI-based hate speech detection for secured online gaming in metaverse using blockchain. Security and Privacy. 2023 Sep 13;7(2).

[4] Faraz A, Ahsan F, Mounsef J, Karamitsos I, Kanavos A. Enhancing Child Safety in Online Gaming: The Development and Application of Protectbot, an AI-Powered Chatbot Framework. Information. 2024 Apr 19;15(4):233.

[5] Ribeiro VM, Bao L. Professionalization of Online Gaming? Theoretical and Empirical Analysis for a Monopoly-Holding Platform. Journal of Theoretical and Applied Electronic Commerce Research. 2021 Jan 11;16(4):682–708.

[6] Kaggle "Predict Online Gaming Behavior Dataset" 21 June 2024, www.kaggle.com/datasets/rabieelkharoua/predict-online-gaming-behavior-dataset.

[7] Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences. 2018 Oct 1;465:1–20.

[8] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 Jun 1;16:321–57.

[9] Mantas CJ, Castellano JG, Moral-García S, Abellán J. A comparison of random forest based algorithms: random credal random forest versus oblique random forest. Soft Computing. 2018 Nov 17;23(21):10739–54.

[10] Breiman L. No Title. Machine Learning. 2001 Jan 1;45(1):5–32.

[11] Lou W, Wang X, Chen F, Chen Y, Jiang B, Zhang H. Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes. PLoS ONE. 2014 Jan 24;9(1):e86703. Available from: https://doi.org/10.1371/journal.pone.0086703