Machine Learning Models in Predicting Gaming Popularity: A Comparative Analysis

Vidath Kuna

Suchitra Academy International School, Suchitra junction, Quthubullapur, Hyderabad - 500067

Abstract

This paper provides an insightful exploration into the application of machine learning (ML) in predicting the player base of video games, shedding light on the pivotal role ML plays in the gaming industry. The introductory section succinctly defines machine learning and its relevance, emphasing its ability to analyse patterns and make data-driven predictions. Focusing specifically on the gaming sector, the paper delves into the significant impact of ML on understanding player behaviour, optimizing user experiences, and enhancing overall game performance. Three prominent ML models — Logistic Regression, Decision Tree and Random Forest — are comprehensively examined for their efficacy in forecasting the base number of players owning a game. To validate the models, a publicly available dataset is employed and the study aims to unravel the strengths and weaknesses of each model, offering valuable insights for developers and stakeholders in the gaming industry. Random forest model emerged as the one with maximum accuracy of 89.38%. The paper contributes to the growing body of knowledge on ML applications in gaming, showcasing the potential of predictive analytics in anticipating and meeting player demands.

Keywords: Machine Learning, gaming industry, predictive analytics, player engagement

1. Introduction

Machine learning (ML) is a type of Artificial Intelligence (AI) which allows us to train a machine some specific things which it will perform further without a lot of supervision. Most of us misinterpret machine learning and artificial intelligence as being related to robots and advanced computer science but they are present in simple instances which we come across like face detection, friend recommendations on social media and suggestions on OTT platforms. The term "Artificial Intelligence" (AI) has long been prevalent in the video games industry, yet, despite decades of use and recent noteworthy advancements in academic AI research, particularly in different methods and real-world applications, the adoption of contemporary AI techniques like machine learning and deep learning remains surprisingly limited in commercial video games [1].

ML can be classified into supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, for example, the machine learning model has to be fed with images, and what the images depict — this data is known as labelled data. Further, when the machine is given an input to identify, it will recognise the image. In unsupervised learning, the teachable machine is given multiple samples of data, and the machine model will recognise the pattern and sort the given data in an order. For instance, if you feed a machine with sample images of mangoes, apples, and bananas, the machine learning model will categorise the images. Hence, unsupervised learning is used to spot patterns, and supervised learning is used to identify specific things. In reinforcement learning when data, say, an image sample, is given, the machine observes and recognises the sample and gives its output. If the output is correct the machine saves the data and becomes more accurate. When the output is wrong, the user corrects it by giving it a necessary command. In simple terms the machine learns from its mistakes and experiences.

The right machine learning solution can be chosen depending upon the problem statement, the size or quality and nature of the data and complexity of the algorithm. Algorithms or methods of solving a particular problem. Algorithms are of three types, classification, regression and clustering. Classification is used when the output is a simple yes or no. Regression is used when a value needs to be predicted like the price of houses in a locality, or when the data needs to be organised to find patterns, like in the case of different types of citrus fruits.

Machine learning and artificial intelligence have profoundly transformed the landscape of the gaming industry, ushering in a new era of innovation and immersive experiences. The primary objective of this paper is to conduct an investigation into the effectiveness of three prominent machine learning models—Logistic Regression, Decision Tree, and Random Forest—in predicting the estimated number of players owning the game. To validate these models, a publicly available dataset is utilised. The study is specifically focused on identifying the model that yields the highest accuracy in forecasting the base number of players for a given game. Through this analysis, the paper aims to contribute valuable insights that can guide decision-makers and developers in the gaming industry towards selecting the most accurate and reliable predictive model for player engagement.

1.1. AI and ML in the gaming Industry

Games provide a huge platform for the machine learning models to learn from. Gaming industry is one of the largest growing industries. There is a huge demand for gaming nowadays, and machine learning plays a major role in shaping the gaming algorithms. Applications of machine learning in games include learning how to play the game well, player modelling, adaptability, model interpretation, and of course, performance.

Another major use is in non-player characters (NPCs) with advanced AI capabilities that can exhibit a bit more realistic and natural behaviours, making the game feel more alive. The machine learns to play the game as an NPC and grasps the way the game is played, this helps the machine learning model to reduce the small inaccuracies in the game. The machine learns about the players, when the machine plays the game. It starts to learn about the opponents, the tactics used by them to play the game and observe glitches from the players point of view.

The characters learn from other player interactions and change according to the circumstances. For example, in the game ‘last of us part II ’an AI-driven NPC invites the enemies, showcasing advanced behaviours, such as teamwork, communication and planned decision making, resulting in a complex gameplay experience. ML plays a major role in bug detection in games. Some machine learning algorithms can be trained to identify common bugs, glitches in the game code. ML models also play a major role in leaderboards in most of the competitive games. In a competitive online game called ‘Fortnite’, the machine learning algorithms arrange the leaderboard by analysing the statistics of the players in a match of hundred players [2, 3].

1.2. Machine learning models

a. **Logistic Regression:** Logistic Regression is a handy tool for predicting the likelihood of an event occurring. Picture trying to forecast whether your favourite sports team will win a game based on factors like player performance and past results. Logistic Regression takes these factors and transforms them into a probability score between 0 and 1, where 0 means very unlikely and 1 means very likely. At its core is the logistic function, resembling an "S" curve, which converts any input value into a value between 0 and 1. In our sports prediction example, if the sigmoid output is close to 0, it implies a low chance of your team winning, and if it's close to 1, a high chance. Logistic Regression fine-tunes the weights assigned to each input (like player performance) to minimize the difference between its predictions and the actual outcomes in the training data. It's widely used for binary choices, making it valuable in situations where you want to predict whether something will or won't happen.

b. **Decision Tree:** A Decision Tree is a powerful tool in machine learning that helps make decisions based on input features. Imagine you're deciding what to wear each day. A decision tree for this might start with a question like "Is it sunny?" If yes, it branches into further questions like "Is it hot?" or "Is there a breeze?" These questions form a tree-like structure, leading to different clothing decisions. In machine learning, a Decision Tree does something similar. It asks a series of questions about input data, leading to a final decision or prediction. Each question represents a "split" in the data, where the goal is to create groups that are as pure as possible regarding the predicted outcome. The principle is to find the most informative questions that best separate the data, making the tree efficient in predicting outcomes for new, unseen data. Decision Trees are interpretable and useful for both classification and regression tasks, making them a widely-used model.

c. **Random Forest:** The Random Forest model is like a smart and diverse group of decision trees working together to make accurate predictions. Picture trying to make a decision with advice from many different people rather than relying on just one person's opinion. Similarly, a Random Forest combines multiple decision trees, each trained on different parts of the data. Here's how it works: Instead of relying on a single decision tree, which might be influenced by peculiarities in the data, the Random Forest builds many trees with different subsets of the data and features. Each tree "votes" on the outcome, and the most popular decision becomes the final prediction. This approach helps overcome overfitting (making decisions based on noise in the data) and generally improves accuracy and robustness. Random Forests are versatile and effective, making them popular for tasks: classification and regression.

2. Methods

For this study, a publicly available dataset was obtained from Kaggle.com [4], which combines data from Steam and SteamSpy APIs, encompassing information on 27,000 games. The dataset includes crucial details such as game genre and the estimated number of owners. The dataset comprises of 27075 rows and 18 columns; the columns are defined as follows:

- Appid (Unique identifier)
- Name (Title of app/game)
- Release_date (Release date)
- English (Language support),
- Developer (Developer name(s))
- Publisher (Publisher name(s))
- Platforms (Supported platforms)
- Required_age (Minimum age requirement)
- Categories (Game categories)
- Genres (Game genres)

- Owners (Estimated number of owners, including lower and upper bounds)
- Steamspy_tags (Top SteamSpy game tags)
- Achievements (Number of in-game achievements)
- Positive_ratings (Number of positive ratings)
- Negative_ratings (Number of negative ratings)
- Average_playtime (Average user playtime)
- Median_playtime (Median user playtime)
- Price (Current full price in GBP)

For the analysis, Google Colaboratory was utilised for coding in Python. Various machine learning models were employed to predict the estimated number of owners, a crucial metric indicating the game's popularity and user reach through purchases or other acquisition means. This predictive analysis aimed to uncover patterns and relationships within the dataset, facilitating insights into the factors influencing a game's ownership metrics. Figure 1 shows the schematic representation of the steps followed for the analysis approach.

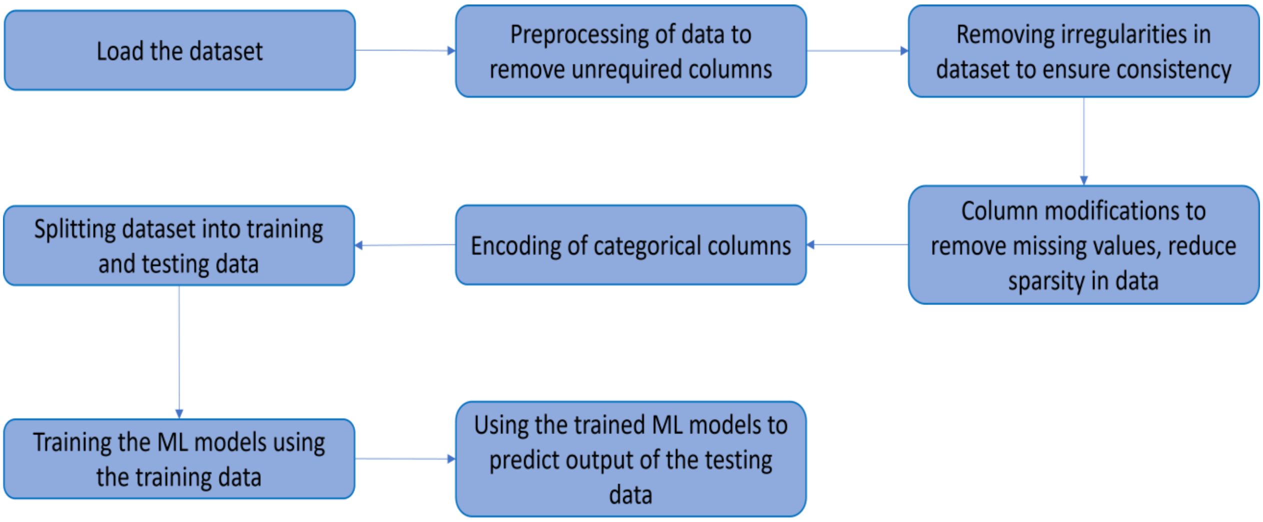
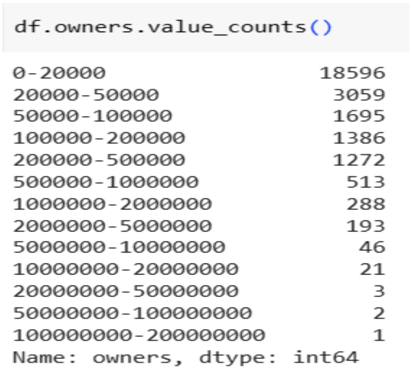

Load the dataset
Preprocessing of data to remove unrequired columns
Removing irregularities in dataset to ensure consistency
Column modifications to remove missing values, reduce sparsity in data
Encoding of categorical columns
Splitting dataset into training and testing data
Training the ML models using the training data
Using the trained ML models to predict output of the testing data

Figure 1. Schematic representation of steps followed for Machine Learning execution

Initially, we sought to understand the distribution of values in the ‘Owners ’column. Our analysis revealed that a substantial majority of values fell within the 0-20000 range, as illustrated in Figure 2. Consequently, we categorised the ownership values into three groups: 0-20000, 20000-500000, and >500000.

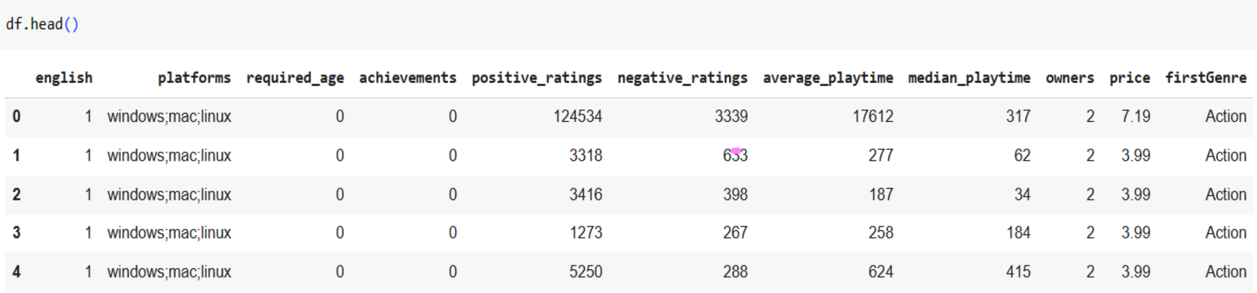


```
df.owners.value_counts()

0-20000                  18596
20000-50000               3059
50000-100000              1695
100000-200000             1386
200000-500000             1272
500000-1000000             513
1000000-2000000            288
2000000-5000000            193
5000000-10000000            46
10000000-20000000           21
20000000-50000000            3
50000000-100000000           2
100000000-200000000          1
Name: owners, dtype: int64
```

Figure 2. Distribution of values in the ‘owner ’number categories in the original dataset.

To streamline the dataset for analysis, columns such as ‘appid,‘ ’name,‘ ’release_date,‘ ’developer, ’ ‘publisher,‘ ’categories, ’and ‘steamspy_tags ’were excluded due to their limited relevance. Additionally, the ‘genres ’column contained multiple comma-separated values in different rows. To address this, we retained only the first genre from each row, creating a new column named ‘firstGenre. ’The pre-encoded dataset is depicted in Figure 3.



```
df.head()
```

	english	platforms	required_age	achievements	positive_ratings	negative_ratings	average_playtime	median_playtime	owners	price	firstGenre
0	1	windows;mac;linux	0	0	124534	3339	17612	317	2	7.19	Action
1	1	windows;mac;linux	0	0	3318	633	277	62	2	3.99	Action
2	1	windows;mac;linux	0	0	3416	398	187	34	2	3.99	Action
3	1	windows;mac;linux	0	0	1273	267	258	184	2	3.99	Action
4	1	windows;mac;linux	0	0	5250	288	624	415	2	3.99	Action

Figure 3. The dataset used for ML models after encoding.

Subsequently, one-hot encoding was applied to only three categorical columns: ‘platforms.‘ ’owners, ’ ‘firstGenre. ’As discussed earlier, the target output column is. As discussed earlier, the target output column is ‘owners. ’The data was then divided into training and testing sets, with 20% allocated for testing. A random state of 2 was selected to ensure consistent data splitting. The dataset was then fed into three machine learning models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier.

3. Results and Discussions

The ‘owners ’column has 3 categories and hence this is a classification problem, so the accuracy score was calculated using the confusion matrix. Table 1. presents the accuracy scores obtained for each of the three models.

Table 1: Accuracy scores for training and testing data prediction using various machine learning models.

Machine Learning Model	**Test Data Accuracy Score (%)**
Logistic Regression	82.12
Decision Tree Classifier	85.35
Random Forest Classifier	89.38

The accuracy scores obtained for the machine learning models reveal varying degrees of predictive performance in forecasting the number of owners for games in the gaming industry dataset. Logistic Regression, with an accuracy score of 82.12%, demonstrates a respectable level of predictive accuracy. This model is well-suited for binary classification tasks and performs reasonably well in discerning ownership categories. The Decision Tree Classifier, boasting an accuracy score of 85.36%, exhibits improved performance compared to Logistic Regression. Decision trees are adept at capturing complex relationships in the data, contributing to a more accurate prediction of ownership categories. The Random Forest Classifier outperforms both Logistic Regression and Decision Tree Classifier, attaining an accuracy score of 89.38%. By aggregating predictions from multiple decision trees, the Random Forest model enhances accuracy and robustness, showcasing its effectiveness in predicting the number of owners in this gaming dataset.

In summary, the escalating accuracy scores from Logistic Regression to Decision Tree Classifier and finally to Random Forest Classifier underscore the incremental improvement in predictive capabilities. The Random Forest model, by leveraging ensemble learning, emerges as the most accurate predictor for estimating the number of owners in this dataset.

4. Conclusion

In this paper, we have explored the applications of machine learning and artificial intelligence in the gaming industry. The aim was to investigate the effectiveness of various machine learning models in predicting the number of owners for video games using a publicly available dataset. Data preprocessing steps were undertaken, such as the removal of irrelevant columns and modification of the ‘Owners ’and ‘Genres ’columns, preparing the dataset for subsequent machine learning analysis. Three prominent machine learning models—Logistic Regression, Decision Tree Classifier, and Random Forest Classifier—were employed to predict the number of owners. The results showcased varying degrees of accuracy, with the Random Forest Classifier emerging as the most accurate predictor. The findings presented in this discussion highlight the significance of machine learning techniques in unraveling patterns within gaming datasets. The paper contributes to the growing body of knowledge regarding data-driven insights for game developers, emphasising the importance of accurate predictions for player engagement and overall game success. The integration of machine learning and artificial intelligence in the gaming landscape is a transformative force, opening new avenues for personalised gaming experiences and informed decision-making.

5. Conflict of Interest

The authors declare no conflict of interest in conducting this research and preparing the associated paper. The research was conducted with full transparency and integrity, and the authors did not receive any financial or non-financial benefits that could influence the objectivity or impartiality of the study. There are no affiliations with organisations or entities that might have a direct interest in the research outcomes. The findings and conclusions presented in the paper are solely based on the analysis of the gaming industry dataset and the application of machine learning models. The authors are committed to upholding ethical standards in research and ensuring the accuracy and reliability of the information presented in the paper.

6. Acknowledgement

The authors extend heartfelt gratitude to their parents for unwavering support, teachers for guidance, friends for encouragement, and mentors for invaluable insights. Their collective contributions have been instrumental in shaping and enhancing the research.

7. References

1. Pfau J., Smeddinck J.D., Malaka R., “The Case for Usable AI: What Industry Professionals Make of Academic AI in Video Games”, In Extended Abstracts of the 2020 Annual Symposium on Computer-

Human Interaction in Play (CHI PLAY '20), Association for Computing Machinery, New York, NY, USA, 2020, 330–334. https://doi.org/10.1145/3383668.3419905

2. T. Rath and N. Preethi, "Application of AI in Video Games to Improve Game Building," 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India, 2021, pp. 821-824, doi: 10.1109/CSNT51715.2021.9509685
3. Michael Bowling M., Furnkranz J., Graepel T., Musick R., "Machine Learning and Games", Machine Learning, 2006, 63:211–21, DOI: 10.1007/s10994-006-8919-x.
4. Steam Store Games (Clean dataset). (2019, June 12). Kaggle. https://www.kaggle.com/datasets/nik-davis/steam-store-games