Game Classification and Analysis Based on Machine Learning-Based Methods

Zhen Li[1], Dingzhuoya Wang[2,*] and Yanzhao Zou[3]
[1]Moonshot academy, Beijing, 100098, China
[2]Shenzhen Zhengde High School, Shenzhen, 518000, China
[3]Wuhan Britain China School, Wuhan, 049452, China
*

Keywords: Machine Learning, Game Rating, AI.

Abstract: It is clear that we should pay high attention to video games on a variety of degrees. This includes the rating of them (Game rating). The efficiency for labor to do this work is highly limited and the reliability is unstable, so try to let Artificial Intelligence (AI) do this work or assist related personnel. The goal is to construct an AI that can help evaluate the ratings of each video game with basic details of the contents that the game contains. We try different AI models and use datasets on Kaggle for AI training. We also take multiple indicators such as accuracy to compare the performance of those models. This study is conducted on those datasets on Kaggle, the result shows that Extreme gradient boosting (XGboost) has advantages over others to some degree. XGboost improves data fitting and inference due to its powerful representation ability. The proposed plan can help staff improve labor efficiency and reliability in-game rating.

1 INTRODUCTION

With the development of technology and the strength of the Internet, the electronic game has gradually evolved and has advanced into every aspect of daily life, offering new entertainment forms to the world. However, as the gaming market continues to expand and game content diversifies, the issue of game rating has become increasingly prominent.

Numerous scholars have conducted in-depth research on game ratings in recent years. For instance, Smith (2018) emphasized the vital role of game rating systems in protecting underage players and guiding consumers toward suitable games. Johnson (2020) highlighted the complexity and challenges associated with game rating in his research. Furthermore, Brown (2019) emphasized that due to cultural and regional disparities.

In addition to the aforementioned studies, scholars have also analyzed game ratings from a policy-making and market perspective. For example, Lopez (2021) noted that in some countries, governments have begun implementing game rating systems to protect minors from harmful games. These rating systems also provide valuable insights for businesses in terms of market positioning and product promotion. Furthermore, Miller (2022) emphasized that the expansion and growth of games is becoming more and more important to businesses. This suggests that when establishing a global game rating system, it is imperative to take into account the cultural backgrounds and market demands of different countries and regions.

In summary, game rating is a complex and challenging issue. Despite numerous difficulties and challenges, the need for establishing more comprehensive and accurate game rating systems will become increasingly pressing with technological advancements and heightened awareness of minors' protection.

With the development of recent years, machine learning (ML) has achieved a series of significant results. ML may have seemed to be a new topic. However, the fact is that machine learning can be traced back to the 20th century when people started to explore artificial neural networks. Warren McCulloch and Walter Pitts in 1943 proposed a hierarchical model of neural networks and created the theory of computational models of neural networks, laying the foundation for the development of machine learning (McCulloch & Pitts 1943). Alan Matheson Turing proposed the famous "Turing Test" in 1950 and conducted experiments with artificial intelligence as an important research topic. By the 2000s, several

Li, Z., Wang, D. and Zou, Y.
Game Classification and Analysis Based on Machine Learning-Based Methods.
DOI: 10.5220/0012866800004547
In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 382-386
ISBN: 978-989-758-690-3

different machine learning models including random forest had been published by loads of professionals.

However, no one has yet explored Artificial Intelligence (AI) and game groupings. Since this is essentially a classification task, performance can be improved by introducing a classification model such as extreme gradient boosting (XGBoost). By introducing a machine learning model to learn features of the dataset(s) of game contents and their ratings, the association between features can be effectively captured to improve the model's representation of the input.

The main objective of this study is to analyze the machine learning-based methods for game classification. Specifically, first, logistic regression (LR), XGBoost, random tree (RF), decision tree (DT), and gradient boosting decision tree (GBDT) are used in the research. XGBoost supports various objective functions with an efficient linear model solver and tree learning. LR allows for the use of continuous or categorical predictors and it's useful for analyzing observational data. DT is a data mining technique and forecasting algorithm commonly used to create classification systems and develop target variables based on multiple covariates. RF combines multiple decision tree stacks. In environments where the number of variables exceeds the number of observations, the RF signal behaves well. For GBDT, it is a predictive analysis used to explain the relationship between binary differential variables. Secondly, the predictive performance of the different models is analyzed and compared. In addition, we process the dataset, such as removing duplicate values. Meanwhile, we use five indicators to compare the advantages and disadvantages of several machine learning methods. The experimental results demonstrate that XGBoost performs excellently. The results obtained from this study provide a good choice for game grading methods.

2 METHODOLOGY

2.1 Dataset Description and Preprocessing

We chose the dataset of the Entertainment Software Rating Board (ESRB) as the benchmark of this study (Kaggle 2023). This dataset contains the names for 1883/1895 (before/after data cleaning) games with 32/34 of ESRB rating content with the name as features for each game (after cleaning dataset). Some single binary vectors are used for representing the features of ESRB content, and short strings are used to represent the rating from ESRB of each game (recorded in the "esrb rating" column). Also, we remove columns "console" and "no descriptors" for they have exactly no effect on the ESRB rating of games. Some records in the dataset have repeated game names and the same binary vectors (columns removed do not count), so they are also removed.

2.2 Proposed Approach

The main purpose of this study is to investigate better ways to classify games. In this study, we first employ XGBoost, LR, DT, RF, and GBDT. Next, we preprocess the data, filter for its missing and duplicate values, and remove the duplicate values. Thirdly, we analyze and compare the prediction performance of different models using five indicators: accuracy, Area Under Curve (AUC), precision, recall, and FL-score. The process is shown in Figure 1.

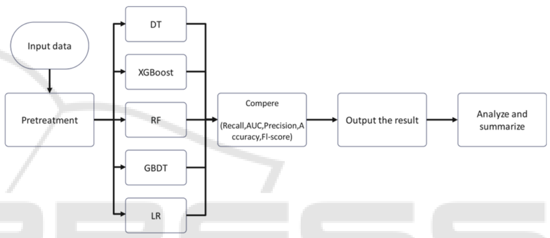

Figure 1. The pipeline of this study (Picture credit: Original).

2.2.1 XGBoost

XGBoost is a scalable machine-learning system that is primarily used for tree climbing. Its influence is widely recognized in several machine learning and data mining operations (Chen & Guestrin 2016). XGBoost provides parallel wood reinforcement (GBDT, also known as GBM) to solve many data science problems quickly and accurately. It expands with the second-order Taylor formula to optimize loss function to guarantee calculating accuracy, constant terms are removed, and there are also loss function terms optimizations; meanwhile, regular items are used to avoid overfitting, they are expanded, the constant term is removed again, and the regular item is optimized; At last, it combines the coefficient of the primary and quadratic terms above to get the final objective function. Blocks storage structure also allows it for parallel calculating.

2.2.2 LR

LR is a logistic function-based statistical method that is used for binary classification problems. The logistic function transforms linear combinations of input variables into probabilities.

In LR, the input variables are multiplied by regression coefficients (weights) and added to a constant term (intercept) to create a linear predictor. This linear prediction is then converted using a Boolean function to generate a probability value between 0 and 1. The output of the logistic regression model is the predicted probability of an event occurring for a given input.

One of the key advantages of LR is its simplicity and interpretability. The model can be expressed as a set of linear equations, making it easy to understand and visualize. Additionally, LR is robust to outliers and can handle both binary and continuous response variables.

In summary, LR is a simple and interpretable statistical method for binary classification problems. It transforms linear combinations of input variables into probabilities using the logistic function and can handle various types of response variables.

2.2.3 DT

DT is a flowchart-like tree structure, where rectangles represent each inner node and ellipses represent individual end nodes. DT can be applied sequentially or in parallel, depending on the amount of information, the position of the available memory in computing resources, and the measurement of the algorithm (Priyam et al. 2013).

In DT training, the DTs are retrieved from identified learning cases, represented by pipes with the value of attributes and layers markers. Destination tree training usually starts with an empty tree with all the learning information. This is a recursive process from top to bottom. Select attributes, find the data attributes that can best format the part, use them as the root attribute, and then divide the training data into non-overlapping subsets corresponding to the values of split attributes.

DT has many interesting features, such as simplicity, ease of knowledge, and the ability to handle mixed data types with some freedom. This makes DT-Learning one of the most successful learning algorithms among machine learning algorithms today (Song & Ying 2015). Compared with other classification methods, the construction speed of decision trees is relatively fast. Trees can be easily converted into SQL statements for efficient access to databases. Compared with other classification methods, decision tree classifiers achieve similar, sometimes even better accuracy (LaValley 2018).

2.2.4 RF

RF technology is a regression tree technology that allows users to control aggressive and predictable plantings with high predictive accuracy. Biermann's RF uses randomization to generate many DTs. Combine the production of these plants into a single output, either by voting on classification problems or mean regression problems.

There are two ways to randomize. Firstly, perform replacement sampling (bootstrap sampling) on the dataset. The process of aggregating new samples in this way is called "guided aggregation" or "bagging". It cannot be guaranteed that every subject will appear in the new sample, and some subjects may appear more than once. In a big dataset with n subjects, the probability of being excluded from bootstrap samples of size n converges to 1/e or approximately 37%. These omitted or "out of the box" subjects constitute a useful set of data for testing decision trees developed from sample subjects.

The second randomization occurs at the decision node. At each node, a certain number of predictors will be selected. For a set with p predictive factors, a typical number is the rounded square root of p - although this parameter can be chosen by analysts. Then, the algorithm tests all possible thresholds for all selected variables and selects the combination of variable thresholds that produces the best segmentation - for example, the segmentation that most effectively separates cases from controls. This random selection of variables and threshold testing will continue until reaching a "pure" node (containing only cases or controls) or some other predefined endpoint. Repeat the entire tree growth process (usually 100 to 1000 times) to grow an RF.

The biggest advantage of RF is that their interactions or nonlinear properties do not need to be specified in advance as required by other parameter survival models (Biau & Scornet 2016).

2.2.5 GBDT

GBDT is a combination of a gradient enhancement algorithm and a decision tree algorithm. Weak students who choose GBDT are an important tree for optimizing the loss function. Boosting, a technique that combines and creates weak learners through an iterative approach to strong learners, was chosen as the primary blended learning method of GBDT. The gradient pulse algorithm differs from other pulse methods in that it updates the loss and gradient functions to complete the learning process. The decision tree algorithm has a tree structure that displays the test results for each field and is a basic classification and regression technique. Represents a characteristic test for each category of each internal node and each leaf node. In short, as an integrated machine learning algorithm, GBDT is superior to some traditional machine learning methods for better prediction accuracy (Wang et al. 2018).

GBDT calculations are very expensive for applications with high dimensional sparse output. Each time iteration, GBDT builds a regression tree to fit the previous tree residuals. Only one iteration increases the density of the residue rapidly, while N is the meaning of samples, and L is the meaning of L, the number of labels (the size of the output space). Therefore, there is at least O (NL) time and memory required in the construction of the GBDT tree. This allows GBDT to run on large applications (e.g., millions) for both N and L (Si et al. 2017).

3 RESULT AND DISCUSSION

3.1 Performance Analysis in Accuracy, AUC

As the result shown in Figure 2, in terms of accuracy, there was no significant difference between DT, RF, GBDT, and XGBoost, with data values ranging from 0.825 to 0.865. RF and XGBoost show slight advantages, while LR is weaker than the other four algorithm models. XGBoost performed the best, with an accurate value of 0.865, followed by RF at 0.860. LR's value of the data is 0.830, which is about 0.025 weaker than the first four. Both DT and GBDT are around 0.850. This is due to the low adaptability of the LR algorithm to this data. LR is more suitable for building models with linear correlations, so it performs poorly in all aspects when analyzing the nonlinear data in this dataset (Luis & Augusto 2019).

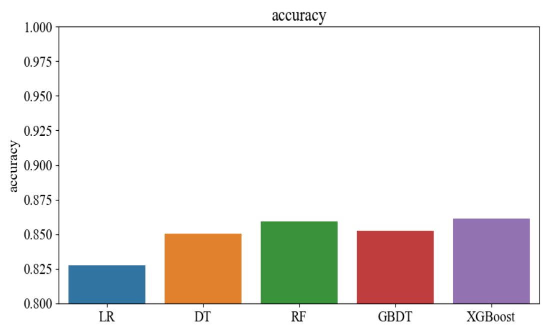

Figure 2: The analysis results for accuracy (Picture credit: Original).

As the result shown in Figure 3, In terms of AUC, there is not much difference between LR, RF, GBDT, and XGBoost, the value of the data ranges from 0.965 to 0.975. The RF is slightly higher, and the DT is significantly lower than the other four learning methods, with a difference of about 0.020. The gap between RF, GBDT, and XGBoost is small, the value of the data is about 0.05. It can be seen that the use of DT can play a role in cost savings. DT is easier to implement and easier to understand than other classification algorithms. Its easy-to-build nature saves it a certain amount of AUC.

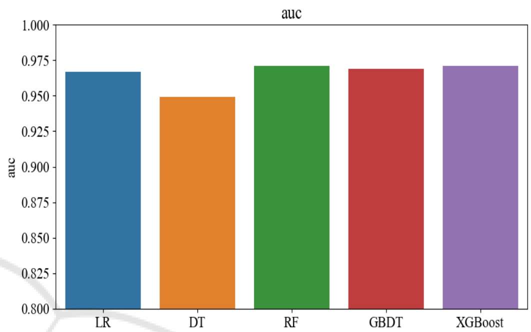

Figure 3: The analysis results for AUC (Picture credit: Original).

3.2 Performance Analysis in Precision, Recall and Fl-score

As the result shown in Figure 4, in terms of precision, XGBoost has the best performance, the data value is around 0.865. The next order is RF, DT, GBDT, and LR. For RF, its value of data is 0.860. For DT, it's 0.858. For GBDT, it's 0.852. For LR, it's 0.26. The algorithm model of LR has significantly lower precision in it. XGBoost's parallel computing capabilities give it an edge.

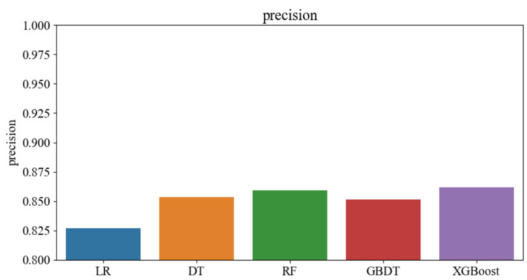

Figure 4: The analysis results for precision (Picture credit: Original).

As the result shown in Figure 5, in terms of recall, XGBoost still has an advantage, and RF also has an advantage, second only slightly to XGBoost. The GBDT is not significantly different from the DT, the value of the data ranges from 0.850 to 0.855. The disadvantage of LR is still quite obvious. It can be seen that XGBoost still has an advantage.

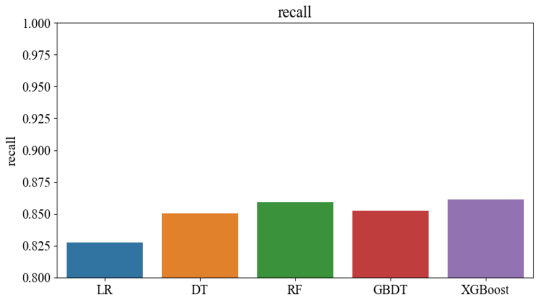

Figure 5: The analysis results for recall (Picture credit: Original).

As the result shown in Figure 6, on fl-score, XGBoost and RF have an advantage of approximately 0.865. The GBDT takes second place, and the difference between the DT and it is not significant, the value of the data is about 0.825. LR is significantly lower than the top four. Its value of data is 0.826.

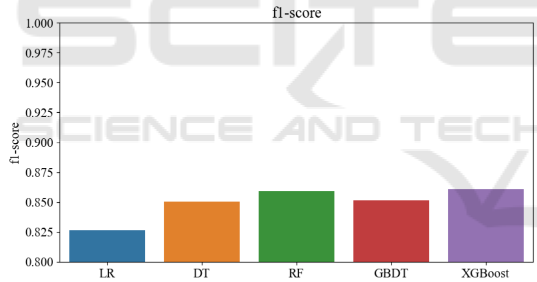

Figure 6: The analysis results for Fl-score (Picture credit: Original).

4 CONCLUSION

The main purpose of this project is to analyze and find a better game classification method based on machine learning. In this study, we used XGBoost, LR, DT, RF, and GBDT for purposes. In the beginning, we analyze the data. We also deduplicated preprocessing the dataset. Next, we input the data into XGBoost, LR, DT, RF, and GBDT for modeling. After that, we compared the advantages and disadvantages of several machine learning methods using five indicators: accuracy, AUC, precision, recall, and FL-score. Extensive experiments are conducted to evaluate the proposed method. Experimental results show that XGBoost has the best performance. The results of this study provide a good choice for game grading methods. In the future, A will be considered as the research objective for the next stage. The research will focus on experimenting with different, larger datasets and use more resources and time to refine the results.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

A. Priyam, G. R. Abhijeeta, A. Rathee, S. Srivastava, International Journal of current engineering and technology, pp. 334-337 (2013).

F. Luis, B. Augusto, "A game analytics model to identify player profiles in singleplayer games," In *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (2019).

G. Biau, and E. Scornet, Test, pp. 197-227 (2016).

J. Wang, P. Li, et. al, Applied Sciences p. 689 (2018).

Kaggle - video-games-rating-by-esrb, 2023, available at https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb.

M. P. LaValley, Circulation, pp. 2395-2399 (2018).

S. Si, H. Zhang, S. S. Keerthi, et. al, "Gradient boosted decision trees for high dimensional sparse output," In *International conference on machine learning* (2017) pp. 3182-3190

T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016) pp. 785-794.

W. S. McCulloch and W. Pitts, The bulletin of mathematical biophysics, pp. 115-133 (1943).

Y. Y. Song, and L. U. Ying, Shanghai archives of psychiatry, p. 130 (2015).