Title 3:

Enhanced Prediction of Sales in Video Games Using Random Forest Algorithm in Comparison with Generalized Linear Model Algorithm to Improve Accuracy.

J Sai Chandu[1], K. V. Kanimozhi[2]

J Sai Chandu
research scholar
BE computer science
SIMATS engineering
SIMATS institute of medical and technical sciences
SIMATS university, chennai, tamil nadu, india, pincode: 602105
saichanduj1216.sse@saveetha.com

K. V. Kanimozhi
research guide, corresponding author
department of computer science and engineering
SIMATS engineering,
SIMATS Institute of Medical and Technical Sciences,
SIMATS University, Chennai, Tamil Nadu, India, Pincode: 602105,
kanimozhikv.sse@saveetha.com

Keywords: Comparison, sales, video games, prediction, Generalized Linear Model algorithm, Random Forest.

Abstract:

Aim: To enhance forecasting, this research contrasts the Random Forest with the Generalized Linear Model. In order to identify the algorithm that most reliably predicts purchases, this study will compare the working of these methodologies. **Materials and methods:** In order to precisely estimate how many gaming copies will sell, I used 2 different techniques in this study: the Random Forest and Generalized Linear Model. I evaluated these two methods to discover which one worked better by observing the game industry sales information. In this study, the number of iterations for each method is 10, and the sample size was found to be ten per group. Consequently, p=0.12. **Results:** The accuracy of the Generalized linear model and Random Forest Model varies significantly, as this analysis demonstrates. 86.40% was the precision value of the Random Forest which was far greater than 74.89% the Generalized Linear model. **Conclusion:** It says that the method known as Random Forest outperforms the Generalized Linear Model approach. Random Forest got 86.40% precision value, whereas the Generalized Linear model managed 74.89%. This tells that it has a superior ability to analyze video game sales. Hence, Random Forest is a preferable one for individuals in the gaming business for accurate sales forecasts.

Keywords: Sales, Machine Learning, Random Forest Algorithm,Generalized Linear Model Algorithm, accuracy improvement, Comparison, Video Games, Prediction.

Introduction:

To improve forecasting accuracy, these methods were utilized for assuming the sales in gaming, random forest and the Generalized Linear model method(Kowert and Quandt 2020). In the current situations, the Games industry is growing in a competitive way, accurate precision is crucial for creators in making choices(Carton 2022). This study explores the applications of past purchases for guessing developments in sales(Ruffino 2018).

Overall, 252 research papers in Google Scholar & 1291 research papers in IEEE Xplore on video game sales prediction(McLinden et al. 2024). Based on numerous research conducted on the topic, there is a considerable consensus that Random Forest outperforms alternative algorithms in video game sales forecasting, as these articles describe(Park et al. 2024). Studies have shown time and time again that the algorithm known as Random Forest predicts more accurately and robustly than other methods(Khaleghi et al. 2024)(Wang et al. 2024). Random Forest is a good option for managing large sets of data for a constantly shifting industry because of its scalability and agility(Labrador et al. 2023).

Lack of consideration when evaluating Random Forest using Generalized Linear Model Algorithm techniques, particularly whenever it comes to video game sales forecasting, is a significant issue in many of the publications published today(Doran 2020). In order to anticipate video game sales, my research will contrast the Random Forest and Generalized Linear Model (Wong et al. 2024).

Materials and Methods:

The task that suggested was finished within this SIMATS university. For this study, a total of two sets were chosen. Set 1 employed the intervention technique, while Set 2 employed the generalized linear model (Mahmud et al. 2023). The Random Forest and Generalized Linear Model algorithms were run at various intervals on a dataset with 20 items. The calculation employed with alpha = 0.05 and beta = 0.2.

The game name and sales statistics from different nations are the two most important components of the dataset used for this investigation. These characteristics are essential to raising the sales prediction accuracy (%) since they form the basis of the dataset's description. The dataset will be of higher quality and yield more reliable analysis if preprocessing is done regularly. Feature extraction and additional cleaning methods come next after pre-processing(Marionneau et al. 2024). Using key features such as game name and geographical sales statistics, our goal is to find out the main components that will affect the sales.

Random Forest Algorithm:

This method is a strong & flexible ensemble learning method that can be applied to both regression and classification problems. The basic method is to build several decision trees during the training stage, each one from a different random subset of the dataset. This technique, called bootstrapped sampling, involves selecting random samples with replacement so that the program can produce a variety of data subsets. Additionally, to determine the optimal split at each decision tree node, a random subset of characteristics is used, which lowers correlation between the trees and improves overall model performance.

Generalized linear Model:

This method is a versatile framework for representing the connection between a response variable and predictor factors. In GLM, the response variable has Gaussian & Poisson probabilities. Initially, the linear predictor is built as a linear combination of predictor variables, frequently supplemented by an intercept term. The link function translates the linear predictor to the response variable's scale, ensuring that the predicted values are within a reasonable range. The link function is chosen based on the nature of the response variable and the intended model assumptions. Finally, the probability distribution describes the variability in the response variable and enables the estimate of model parameters using maximum likelihood estimation or other optimization techniques. GLM is a strong and interpretable technique to regression analysis, classification, and count data modeling, with applications in domains as diverse as statistics, machine learning, and econometrics.

Statistical Analysis:

Quantitative analysis is conducted using IBM SPSS. I will assess the significance of differences in prediction of accuracy using techniques like Hypothesis Testing. Additionally, analysis was conducted of the variables affecting each algorithm's sales estimate. My objective is to conduct a Statistical Analysis and identify the optimal strategy for improving the accuracy of algorithms for video game sales.

Result:

The RF method produced a precision value of 86.40% in the video game sales forecast, was higher than the precision value of 74.89% produced by the Generalized Linear Model approach. Random Forest continuously beats the Generalized Linear Model in 10 iterations shown in 1st table. With a balance value of 82.84, a sd of 2.25, and a se of the balance of 0.90, Table 4 presents the results of the RF method. In contrast, the results of the Generalized Linear Model show a balance of 74.52, along ssd of 2.25 & a se of the balance of 0.90. Compared to the Generalized Linear Model approach, these outcomes imply that the RF technique, on average, produces greater precision & shows fewer differences. As demonstrated in fig.1, the RF predicts more correctly than the generalized linear model.

Discussion:

According to this study, the Random Forest works good than the generalized linear model when it comes to predicting video game sales, with a precision of 86.50% as opposed to 74.49%(Zhang and Bi 2024)(Ruffino 2018). This accuracy value shows random forest is better than than generalized linear model(Anderson, Allen, and Groves 2019).

The outcomes of this research, which indicate that the RF Algorithm is more accurate, are consistent with previous findings by Vince A(McLinden et al. 2024). In comparable prediction tasks, Random Forest works better than other algorithms, according to Vince A. However, research like that of Peter L. has shown conflicting results, supporting the superior performance of generalized linear models on certain circumstances(Pavlov 2019). With this discrepancy, highly cited works exhibit a consistent trend in indicating Random Forest's betterness in every way.(Cook and Draycott 2022).

This study backs the literature's overall conclusion that the method known as Random Forest has the greatest predictive value when forecasting video game sales(Rollinger 2020). These outcomes with past research display Random Forest's reliability in this field(Hansch 2024). Investigating these nuances is critical for developing more reliable and universally applicable forecasting techniques that can accurately depict the problems in the gaming business. Sustained working in this area improves our recognition of forecasting and leads to more precise and helpful insights for those in the industry(Hansch 2024; McCullagh 2019).

Our study's conclusion emphasizes how important it is to select the best prediction model based on the features of the dataset and the particular goals of the investigation. Although Random Forest is superior at identifying complex patterns and providing accurate revenue forecasts for games, Generalized Linear Models (GLMs) are more comprehensible and flexible when it comes to handling different distributional assumptions. The choice between these models is influenced by variables like processing resources, interpretability requirements, and dataset complexity. But as machine learning is a dynamic field, more research may be conducted on ensemble methods that combine the best aspects of GLMs and Random Forests. Such ensemble approaches have the potential to improve the interpretability and accuracy of video game sales projections by combining the strong predictive capability of Random Forest with the interpretability of GLMs. Furthermore, investigating developments in machine learning techniques and algorithms may enhance the prediction powers in this field, enabling publishers and game creators to make better-informed decisions.

Conclusion:

This study discovered that the Random Forest outperforms the generalized linear model technique for predicting video game sales. Random Forest forecasted sales with 86.40% accuracy, but generalized linear model only got 74.49%. This suggests that Random Forest is good at forecasting. For those in the gaming business who require precise sales estimations, Random Forest is the best option. Its capacity to provide precise estimates is critical for taking informed judgments and planning in the ever evolving gaming business.

Declaration:

Conflicts of interest are absent from this paper.

Authors Contribution:

Author JSC was in charge of writing the manuscript as well as managing data collection and analysis. The text's critical evaluation, data validation, and conceptualization were all handled by author KVK.

Acknowledgements:

The authors recognize the support of SIMATS Engineering in order to effectively complete this study.

Funding:

We are appreciative of the following organizations' financial support, which enabled us to complete this study.

1. INautix Technologies, India.
2. SIMATS engineering.
3. SIMATS university.
4. SIMATS institute of medical and technical sciences.

References:

Anderson, Craig A., Johnie J. Allen, and Christopher L. Groves. 2019. *Game On!: Sensible Answers about Video Games and Media Violence*.

Carton, Christopher. 2022. *A Guide to Video Game Movies*. White Owl.

Cook, Kate, and Jane Draycott. 2022. *Women in Classical Video Games*.

Doran, John P. 2020. *Unity 2020 Mobile Game Development: Discover Practical Techniques and Examples to Create and Deliver Engaging Games for Android and iOS, 2nd Edition*. Packt Publishing Ltd.

Hansch, Ronny. 2024. *Handbook of Random Forests: Theory and Applications for Remote Sensing*. World Scientific Publishing Company.

Khaleghi, Ali, Abbas Narimani, Zahra Aghaei, Anahita Khorrami Banaraki, and Peyman Hassani-Abharian. 2024. "A Smartphone-Gamified Virtual Reality Exposure Therapy Augmented With Biofeedback for Ailurophobia: Development and Evaluation Study." *JMIR Serious Games* 12 (March): e34535.

Kowert, Rachel, and Thorsten Quandt. 2020. *The Video Game Debate 2: Revisiting the Physical, Social, and Psychological Effects of Video Games*. Routledge.

Labrador, Marta, Iván Sánchez-Iglesias, Mónica Bernaldo-de-Quirós, Francisco J. Estupiñá, Ignacio Fernandez-Arias, Marina Vallejo-Achón, and Francisco J. Labrador. 2023. "Video Game Playing and Internet Gaming Disorder: A Profile of Young Adolescents." *International Journal of Environmental Research and Public Health* 20 (24). https://doi.org/10.3390/ijerph20247155.

Mahmud, Shohel, Md Abdullah A. Jobayer, Nahid Salma, Anis Mahmud, and Tanzila Tamanna. 2023. "Online Gaming and Its Effect on Academic Performance of Bangladeshi University Students: A Cross-Sectional Study." *Health Science Reports* 6 (12): e1774.

Marionneau, Virve, Jani Selin, Antti Impinen, and Tomi Roukka. 2024. "Availability Restrictions and Mandatory Precommitment in Land-Based Gambling: Effects on Online Substitutes and Total Consumption in Longitudinal Sales Data." *BMC Public Health* 24 (1): 809.

McCullagh, P. 2019. *Generalized Linear Models*. Routledge.

McLinden, Shea, Peter Smith, Matt Dombrowski, Calvin MacDonald, Devon Lynn, Katherine Tran, Kelsey Robinson, Dominique Courbin, John Sparkman, and Albert Manero. 2024. "Correction to: Utilizing Electromyographic Video Games Controllers to Improve Outcomes for Prosthesis Users." *Applied Psychophysiology and Biofeedback*, March. https://doi.org/10.1007/s10484-024-09636-3.

Park, Kyeongwoo, Hyein Chang, Jin Pyo Hong, Myung Hyun Kim, Sohee Park, Jin Young Jung, Dahae Kim, Bong-Jin Hahm, and Ji Hyun An. 2024. "The Effect of Time Spent on Online Gaming on Problematic Game Use in Male: Moderating Effects of Loneliness, Living Alone, and Household Size." *Psychiatry Investigation* 21 (2): 181–90.

Pavlov, Yu L. 2019. *Random Forests*. Walter de Gruyter GmbH & Co KG.

Rollinger, Christian. 2020. *Classical Antiquity in Video Games: Playing with the Ancient World*. Bloomsbury Publishing.

Ruffino, Paolo. 2018. *Future Gaming: Creative Interventions in Video Game Culture*. MIT Press.

Wang, Jiadong, Yu Wang, Qian Ou, Sengze Yang, Jiajie Jing, and Jiaqi Fang. 2024. "Computer Gaming Alters Resting-State Brain Networks, Enhancing Cognitive and Fluid Intelligence in Players: Evidence from Brain Imaging-Derived Phenotypes-Wide Mendelian Randomization." *Cerebral Cortex* 34 (3). https://doi.org/10.1093/cercor/bhae061.

Wong, Rosa S., Keith T. S. Tung, Frederick K. W. Ho, Wilfred H. S. Wong, Chun Bong Chow, Ko Ling Chan, King Wa Fu, and Patrick Ip. 2024. “Effect of a Mobile Game-Based Intervention to Enhance Child Safety: Randomized Controlled Trial.” *Journal of Medical Internet Research* 26 (February): e51908.

Zhang, Gong, and Shulei Bi. 2024. “Evolutionary Game Analysis of Online Game Studios and Online Game Companies Participating in the Virtual Economy of Online Games.” *PloS One* 19 (1): e0296374.

Table 1: pseudocode of RF algorithm

Input: video games sales prediction dataset
Output: improved accuracy for video game sales prediction
1. Dividing the dataset into two distinct subsets: one for model training and the other for performance evaluation to make sure the model performs well when applied to new data. 2. Create a Random Forest Classifier, which is a collection of decision trees with a preset number of trees (estimators) that may make collective predictions based on numerous tree outputs. 3. Each decision tree is generated by iteratively partitioning the training data into subsets depending on feature values, with the goal of maximizing the purity or homogeneity of each resulting node. 4. Following training, the Random Forest ensemble combines predictions from all trees. 5. Next, train the classifier with the training data. 6. After training, we make predictions for the testing set. 7. Finally, we find accuracy in the testing set. 8. Stop.

Table 2:pseudocode for generalized linear model

Input: video games sales prediction dataset

Output: improved accuracy for video game sales prediction

```
1. Load the dataset that includes both predictor variables (features) and the response variable (target).
2. Ensure that the dataset has been preprocessed.
3. Select an appropriate GLM family and link function based on the response variable's nature and distribution assumptions.
4. Fit the GLM model to the training data with a statistical library like statsmodels or a machine learning library like scikit-learn, specifying the desired family and link function.
5. Use the trained GLM model to predict the testing set's response variable.
6. Calculate the appropriate accuracy metric for your task.
7. For classification tasks, use true and predicted labels from the GLM model
8. Compare it with a random forest algorithm in accuracy.
```

Table 3: Accuracy values of 10 iterations:

Iteration number	Rf accuracy	GLM accuracy
1	85.2%	74.8%
2	84.21%	75.1%
3	84.3%	73.5%
4	86.4%	74.6%
5	85.6%	73.5%
6	82.4%	71.2%
7	85.9%	74.5%
8	86.9%	72.1%
9	83.2%	74.5%
10	80.2%	73.2%

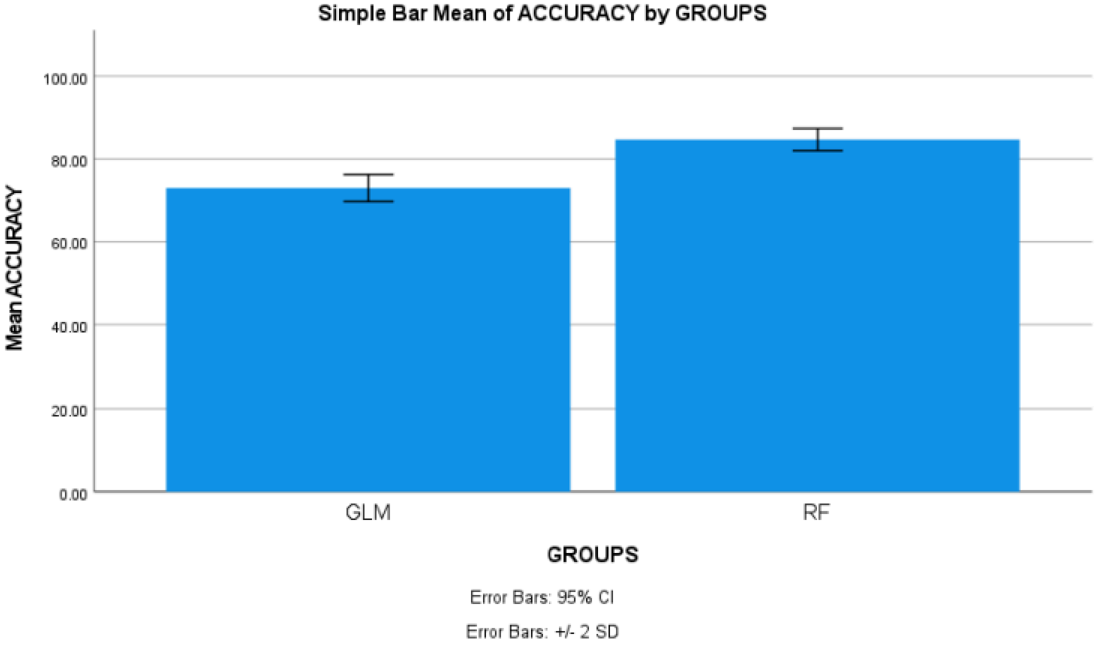
Table 4: group statistics

group	n	mean	standard deviation	standard error mean
RF	10	84.70	1.33749	0.42295
GBR	10	73.00	1.63299	0.51640

Table 5: Independent Sample Test

independent samples test										
		levene's test for equality of variances		t-test for equality of means						
									95% confidence interval of the difference	
		f	sig.	t	df	sig. (2-tailed)	mean difference	std. error difference	lower	upper
accuracy	equal variances assumed	0.6	0.12	17.528	18	0.00	11.70	0.66750	10.29	13.10236
	equal variances not assumed			17.428	17.327	0.00	11.700	0.66750	10.29	13.10628

Figure 1: Generalized Linear Model **vs** Random Forest


Simple Bar Mean of ACCURACY by GROUPS
Mean ACCURACY
GLM
RF
GROUPS
Error Bars: 95% CI
Error Bars: +/- 2 SD