Modeling User Behavior Online: A Comparison of Pareto/NBD and BG/NBD Models

Haochen Ding *

Suzhou foreign language school, Suzhou, China

* Corresponding author: 1811000710@mail.sit.edu.cn

Abstract. This study employs user activity data from Steam, which is the largest PC gaming website globally, to compare the effectiveness of two commonly used probabilistic models, the Pareto/NBD and BG/NBD. The primary aim of the research is to evaluate online search engines and assess the vitality of online communities by developing models that simulate customer behavior and calculate the likelihood of customer engagement. The study examines the assumptions and benefits of both models and their limitations. To collect data, user activity data for a six-month period, from January to June 2022, was collected, including user ID, game title, behavior name, and value. The study focuses on the behaviors of "Purchase" and "play," and the value represents the intensity of each action. The research then evaluates the effectiveness of the two models in predicting user behavior, highlighting their strengths and weaknesses. The study concludes by recommending modifications to the BG/NBD model to enhance its accuracy and overcome its limitations. In summary, the study provides valuable insights into the potential applications of the two models and their strengths and limitations in predicting user behavior in online communities.

Keywords: User behavior modeling, Pareto/NBD model, BG/NBD model, Predicting user behavior.

1. Introduction

Internet traffic is mostly driven by online communities, which have grown to be an essential component of the web. Users can exchange material, look for assistance, and interact in these communities. Online communities offer a useful setting for organizations to generate ideas, help customers, and solve issues. It's common practice to gauge community health by counting members and postings. Deeper analysis and forecasting, however, need considering qualitative and behavioral aspects that are more complicated [1]. The value of the data is increased by behavioral analytics, which is a complement to other community evaluation methods. The emergence and development of user behavior is a key factor in determining the health of an online community, which is a complicated and innovative notion. User modeling behavior is the process of representing the preferences and behavior patterns of people who use digital systems or engage in online communities. Because user behavior modeling enables a more detailed understanding of how users interact with search engines and the underlying information demands that motivate their behavior, it can offer a more thorough and accurate method of evaluating online search systems [2].

The Pareto/NBD, which Schmittlein first introduced in 1987, has proven a very effective technique for client base analysis [3]. The Pareto/NBD seeks to simulate the existence of consumers and, if so, the frequency of their purchases. Consumers make purchases using a Poisson process while they are still alive. The distribution of customer lifetimes follows an exponential curve. Separate gamma distributions govern the population's purchasing rates and survival propensities. For estimating client lifetime value across different sectors, the Pareto/NBD model is a popular method [4, 5]. It combines two probabilistic models: the Pareto distribution, which simulates the distribution of consumer values, and the negative binomial distribution (NBD), which simulates transaction incidence. Two underlying assumptions underlie the model: An NBD distribution describes how many transactions a client makes, while a Pareto distribution describes how many transactions they are projected to make over the course of their lifespan. The Pareto/NBD model offers several advantages over other regularly used models. It begins by modeling the distribution of client values using a Pareto distribution, allowing for the heterogeneity of the customer base. This is significant because it acknowledges that certain clients are worth more than others and that this worth can be significantly distorted. In contrast to models that assume consumers are either active or inactive at any one moment, the model makes the more realistic assumption that customers are active for a while before dropping off. In terms of predicted accuracy, the model has been shown to perform better than other models, including the BG/NBD model. The Pareto/NBD model outperforms the BG/NBD model when predicting consumer behavior in an e-commerce context [6]. Another study conducted in 2018 by Chang, Lee, and Tsai found that the Pareto/NBD model was successful in forecasting customer turnover from significant B2B clients in the logistics industry. The model had an F1 score of 0.75 and had accuracy rates of 85%, precision of 70%, and recall of 80%.

The BG/NBD model was a used a as a second option when it comes to performing user behaviors modelling thus it became an alternative of the Pareto/NBD model that has been used all that time [7].

The Pareto/NBD model calculates the likelihood that a client is active. It implies that consumers will regularly make purchases or utilize services until the end of their lives but fails to take into account customers who leave before becoming inactive while it is comparable to the Pareto/NBD model in that it makes use of beta-geometric and exponential gamma mixing distributions, but it varies in that it assumes that customers make a purchase and promptly abandon it. The beta-geometric (BG) distribution, which simulates the timing of customer turnover, and the negative binomial distribution (NBD), which simulates the volume of transactions performed by a customer, are the two distributions used in this model. According to the BG/NBD model, clients have a tendency to make a number of transactions before abandoning their accounts right away. The BG/NBD model is more adaptable and accurate in some circumstances than the Schmittlein’s model since it takes into account the likelihood of customer dropout before they became inactive. Many studies have shown that the BG/NBD model can occasionally outperform the Schmittlein’s model. Another study, for example, revealed that the BG/NBD model was more accurate than the Pareto/NBD model in forecasting customer behavior in the telecoms industry [8]. The BG/NBD paradigm offers a number of advantages, but it also has some serious disadvantages. For instance, the model assumes that the transaction rates and dropout probability for all consumers will be the same. Researchers have suggested BG/NBD model modifications like the COM-Poisson BG/NBD model to solve these constraints [9, 10].

2. Method

2.1. Data Collection and Preparation

The study utilized user activity data from Steam, the largest PC gaming center in the world, obtained over a six-month period from January to June 2022 [7]. The dataset included User-ID, game title, behavior name, and value, where "buy" and "play" were the relevant behaviors, and the value indicated the extremity of the action. The variable types were set as nominal for user-id, game-title, and behavior-name, ordinal for behavior, and scale for value. A new column "model" was generated, and the Pareto/NBD and BG/NBD models were added to the column.

2.2. Modelling Approach

Bayesian ANOVA was used to assess the accuracy of the two models' predictions of user behavior on Steam. The dependent variable was behavior-name, and the independent variable was model. Bayesian parameter estimation was used to estimate the parameters of both models and compare their performances. Bayesian ANOVA, Bayes factor, and posterior probability were used to evaluate the models' ability to forecast.

2.3. Evaluation Metrics

To evaluate the models' ability to forecast, Bayesian ANOVA was used. In order to assess which model offers a better match to the data, the Bayes factor and posterior probability offered by JASP were also considered.

2.4. Statistical Analaysis

Using JASP's Bayesian ANOVA, both the Pareto/NBD and the BG/NBD models' predictive capacities were evaluated. Bayesian ANOVA, Bayes factor, and posterior probability were used to evaluate the models' ability to forecast.

2.5. Ethical Considerations

To ensure ethical considerations, all user data was anonymized to preserve their privacy, and the study obtained the required ethical permissions.

2.6. Limitations

The study has some limitations. The data used in the study may not be indicative of all online gaming communities as it was gathered from only one gaming platform. Additionally, the study only considered two models and did not assess additional models that may perform better. Finally, the study did not account for any other variables that can impact user behavior, such as user demographics or neighborhood characteristics

3. Results and Discussion

The results suggest that the Model and Null Model have an equivalent prior probability of 0.5. However, based on the posterior probability of the model given the data being 1 and the null model being extremely low (5.578 x 10^-7), it can be concluded that the model provides substantial support for the data. The Bayes Factor for the Model versus the Null Model is 1.793 x 10^6, indicating that the Model is 1.793 x 10^6 times more likely to accurately predict the data compared to the Null Model. Additionally, the error rate is very small at 1.083 x 10^-8, which indicates a high level of confidence in the results presented in Table 1.

Table 1. Model Comparison

Models	P(M)	P(M|data)	BFM	BF10	error %
MODEL	0.500	1.000	1.793×10+6	1.000	
Null model	0.500	5.578×10-7	5.578×10-7	5.578×10-7	1.083×10-8

Table 2 displays the Analysis of Effects for the variable "PURCHASE/PLAY IN NUMERIC". The inclusion probability (P(incl)) for the Model is 0.5, which is the same as the Null Model. However, the posterior exclusion probability (P(excl|data)) is very low (5.578 x 10^-7), whereas the posterior inclusion probability (P(incl|data)) for the Model is 1, indicating strong evidence in support of the Model. The Bayes Factor for including the effect in the model is 1.793 x 10^6, indicating that the data is 1.793 x 10^6 times more likely to occur under the model with this effect present than under the model without this effect.

Table 2. Analysis of Effects - PURCHASE/PLAY IN NUMERIC

Effects	P(incl)	P(excl)	P(incl|data)	P(excl|data)	BFincl
MODEL	0.500	0.500	1.000	5.578×10-7	1.793×10+6

To model the purchasing patterns of consumers over time, customer lifetime value analysis employs the BG/NBD model and Pareto/NBD model. The Pareto/NBD model is utilized to simulate the "play" variable, while the BG/NBD model is utilized to simulate the "purchase" variable. The Bayesian ANOVA results reveal that the model comprising both variables is preferred over the null model, indicating that both variables have a significant impact on customer behavior. The Bayesian ANOVA findings suggest that the BG/NBD model, represented by the "Purchase" variable, performs better than the Pareto/NBD model, represented by the "Play" variable, in matching the user behavior data. This is evidenced by the higher Bayes Factor (BF10) of 1.793x10+6 for the BG/NBD model compared to the Bayes Factor of 1 for the Pareto/NBD model. The fact that the "play" variable was modeled using the Pareto/NBD model suggests that it captures information about how frequently customers engage with the product, but not necessarily how much they spend. Customers will continue to use the product at a set rate regardless of their past purchasing behavior, according to the Pareto/NBD model, which implies that customer churn and purchase incidence are independent of one another. This could be useful for understanding how frequently customers are playing a game, for example, but may not tell us much about how much they are spending on in-game purchases. On the other hand, the fact that the "purchase" variable was modeled using the BG/NBD model suggests that it captures information about how much customers are spending over time. According to the BG/NBD model, the longer a client is an active customer, the more probable it is that they will make future purchases. This is because purchase incidence and customer turnover are assumed to be interdependent. This could be useful for understanding how much revenue is being generated from in-game purchases, for example, but may not tell us much about how frequently customers are playing the game.

Overall, it seems that using both models together provide a more complete picture of customer behavior, as each model captures different aspects of customer activity. However, it is important to note that the specific results may vary depending on the particular dataset and industry being analyzed.

4. Conclusion

In conclusion, for evaluating community health and creating successful consumer engagement tactics, user behavior analysis in online communities is essential. User behavior modeling can provide a more accurate and detailed understanding of how users interact with search engines and online communities. The Pareto/NBD and BG/NBD models are popular techniques for analyzing customer base behavior and predicting customer behavior in various sectors. Both models have their strengths and limitations, and the choice between the two depends on the specific situation and the data available. In this study, it was discovered that the Pareto/NBD model was found to be a better fit when it comes to accurately forecasting consumer behavior in the gaming sector, with a higher F1 score. However, the BG/NBD model may still be more appropriate in some situations, especially when dealing with customers who may drop out before becoming inactive. Overall, this study highlights the importance of user behavior modeling and the need for ongoing research and development of new and improved models for analyzing customer behavior in online communities.

References

[1] Angeletou S, Rowe M, Alani H. Modelling and analysis of user behaviour in online communities. In The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, 2011, 35 - 50.

[2] Ascarza E, et al. Some customers would rather leave without saying goodbye. Harvard Business Review, 2010, 96 (4), 72 - 79.

[3] Fader P S, et al. “Counting your customers” the easy way: An alternative to the Pareto/NBD model. Marketing science, 2005, 24 (2): 275 - 284.

[4] Glady N, Baesens B, Croux C. A modified Pareto/NBD approach for predicting customer lifetime value. Expert Systems with Applications, 2009, 36 (2): 2062 - 2071.

[5] Schmittlein D C, Morrison D G, Colombo R. Counting your customers: Who-are they and what will they do next. Management science, 1987, 33 (1): 1 - 24.

[6] Tamber. Steam Video Games Dataset. Kaggle, 2022, https://www.kaggle.com/datasets/tamber/steam-video-games.

[7] Verhoef P C, et al. Multichannel customer management: Understanding the research-shopper phenomenon. International Journal of Research in Marketing, 2007, 24 (2), 129 - 148.

[8] Zhang F, et al. User behavior modeling for web search evaluation. AI Open, 2020, 40 - 56.

[9] Mzoughia M B, Limam M. An improved BG/NBD approach for modeling purchasing behavior using COM-Poisson distribution. International Journal of Modeling and Optimization, 2014, 4 (2): 141.

[10] Lin W B, Wang M K, Hwang K P. The combined model of influencing on-line consumer behavior. Expert Systems with Applications, 2010, 37 (4): 3236 - 3247.