RFM Analysis for Customer Lifetime Value with PARETO/NBD Model in Online Retail Dataset

Rama Aria Megantara1*, Farrikh Alzami1, Ahmad Akrom1, Ricardus Anggi Pramunendar1, Dwi Puji Prabowo1, Sasono Wibowo1, Ritzkal1,2
1Faculty of Computer Science, Universitas Dian Nuswanto, Semarang, Indonesia
2Department of Informatics Engineering, Universitas Ibn Khaldun, Bogor, Indonesia
E-mail: aria@dsn.dinus.ac.id

ABSTRAK
In recent years, there has been a growing interest in analyzing Customer Lifetime Value (CLV) due to its ability to provide valuable insights into customer profitability and worth. CLV analysis predicts the net profit attributed to the entire future relationship with a customer. This analysis involves calculating the present value of a customer's expected future spending with the company, facilitating an understanding of the economic value of long-term customer relationships. CLV analysis empowers businesses to identify their most profitable customers and develop strategies for retaining them, ultimately maximizing long-term profitability. CLV analysis relies on various models and techniques, including the RFM analysis categorizes customers based on recency, frequency, and monetary value, helping to segment customers and predict future behavior. Then, The Pareto/NBD model combines probability distributions to estimate CLV and is commonly used for customer base analysis. This research article explores the application of RFM analysis for estimating customer lifetime value using the Pareto/NBD model in an online retail dataset. This metric is crucial for businesses as it assists in identifying valuable customers and formulating retention strategies to maximize long-term profitability.

Introduction
In recent years, there has been a growing interest in customer lifetime value analysis as it provides valuable insights into the profitability and worth of customers (Alda’ Omar Dandis et al., 2021; De Marco et al., 2021; Mandal, 2023). Customer Lifetime Value (CLV) analysis is a method that predicts the net profit attributed to the entire future relationship with a customer. This analysis involves calculating the present value of the total amount of money a customer is likely to spend in the future with the company. This helps in understanding the economic value of the customer relationship in the long run (Mensouri & Azmani, 2022). CLV analysis allows businesses to identify the most profitable customers and develop strategies to retain them, thus maximizing their long-term profitability (Bolané et al., 2018). Firms can use this analysis to allocate marketing resources more efficiently, target customers who are likely to yield the highest return on investment, and devise personalized customer retention strategies (Sliz & Delitsis, 2021).

Customer lifetime analysis is often conducted using various models and techniques, such as the RFM (Recency, Frequency, Monetary) model (Sabuncu et al., 2020), the Pareto/NBD (Negative Binomial Distribution) model (Mzoughia et al., 2018), and machine learning algorithms (Sun et al., 2023). These models help in segmenting customers, predicting future behavior, and estimating customer lifetime value (Akhmetbek, 2022).

Thus, estimates of future behavior can be derived from extant marketing models that allow for a customer's latent transition to inactivity, here the research about combining Customer lifetime analysis with other methods is considered.

Various models and methods have been developed to estimate customer lifetime value, with RFM analysis being one of the most commonly used approaches (Jasek et al., 2018). RFM analysis is a technique that categorizes customers based on their purchasing behavior across three dimensions: recency, frequency, and monetary value (Carrasco et al., 2019). Recency refers to the time elapsed since a customer's last purchase, indicating the likelihood of future purchases. Frequency measures the number of
times a customer makes purchases within a specific timeframe, reflecting their loyalty and engagement with the business. Monetary value represents the total amount of money spent by customers, indicating their level of customer demand.

The Pareto/NBD model, also known as the Pareto/Negative Binomial Distribution model, is a probabilistic model used in customer base analysis and customer lifetime value (CLV) estimation (Jasek et al., 2019; Simon & Adler, 2022). It combines the Pareto distribution, which models the probability of a customer making a purchase, and the Negative Binomial distribution, which models the number of purchases made by a customer. It model is based on the assumption that customer purchase behavior follows a double Pareto distribution with two parameters: the transaction rate and the dropout rate (Xie, 2022).

By using the PARETO/NBD model in online retail datasets, businesses can gain valuable insights into customer behavior and make more informed decisions regarding marketing strategies and customer relationship management (Chen et al., 2022). There are several reason why the Pareto/NBD model is commonly used: 1) Accurate prediction of customer behavior: The Pareto/NBD model has been shown to provide accurate predictions of customer lifetime value and repeat purchase behavior(Xie, 2022). It takes into account both the probability of a customer making a purchase and the number of purchases they are likely to make. 2) Flexibility and applicability: The Pareto/NBD model can be applied to various industries and business settings, including online retail, e-commerce, and non-contractual relationships (Mzoughia et al., 2018). 3) Integration with other models: The Pareto/NBD model can be integrated with other models, such as the BG/NBD (Beta Geometric/NBD) model, to enhance its predictive capabilities and improve parameter estimation. This allows businesses to leverage the strengths of multiple models and obtain more accurate insights into customer behavior (Jasek et al., 2019).

Thus, this research article aims to explore the application of RFM analysis for customer lifetime value using the Pareto/NBD model in an online retail dataset. Customer lifetime value is a crucial metric for businesses, as it helps in identifying the most valuable customers and devising strategies to retain them and maximize their long-term profitability.

By any means, our research is essentially using the insights gained from RFM (Recency, Frequency, Monetary) data to inform the parameters of the Pareto/NBD model. This combination allows businesses to make more accurate predictions about customer behavior, including customer lifetime value, and helps in the development of targeted marketing and retention strategies.

Method

The methodology of this research article involves using RFM analysis with the Pareto/NBD model to estimate customer lifetime value. To fulfill this task, a dataset containing information such as Invoice, Stock Code, Description, Quantity, Invoice Date, Price, Customer ID, and Country was derived from an online retail dataset. The dataset was then used to calculate the RFM scores for each customer. The RFM scores were calculated based on three main factors: Recency, Frequency, and Monetary value. Recency refers to the number of days since the customer's last purchase, frequency represents the number of purchases made by the customer within a specified time period, and monetary value indicates the average amount of money spent by the customer on each purchase. The RFM scores were then used in conjunction with the Pareto/NBD model to estimate the customer lifetime value.
Thus, from figure 1, the detailed procedures can be described as follows:

**First.** Load dataset. Here we obtained the dataset from UCI dataset with the link: [https://archive.ics.uci.edu/ml/machine-learning-databases/00502/online_retail_II.xlsx](https://archive.ics.uci.edu/ml/machine-learning-databases/00502/online_retail_II.xlsx). Here, we are using sheet 2009-2010 (525461 rows) and sheet 2010-2011 (541910 rows), then we append the two sheet become one dataset. Resulted in 1067371 rows.

**Second.** Drop rows which have missing values (where description is empty or NA). Resulted in 1062989 rows.

**Third.** Drop rows which does not have Customer ID (due to we want to utilized customer lifetime value, we need the customer ID to obtain the characteristics). Resulted in 824364 rows.

**Fourth.** Conduct date time analysis. Here we are using Month, Time, Year, Day, Quarter, Day of Week.

![Figure 2. time analysis](https://example.com/time_analysis.png)
From Figure 2, we can see transaction year 2009 is less than 2010 and 2011. This stated that the company begin their operations at 2009 in other countries.

First. Conduct RFM analysis. Here we are using this methods:
First). Create total amount features which comes from quantity * price
Second). Using lifetimes library by using summary_data_from_transaction_data function with features that considered: customer ID, invoice date, total amount.
Third). RFM is generated as seen at figure

![RFM results](image)

Second. Conduct PARETO/NBD model
First). Create pareto model by utilizing frequency, recency, and time
Second). Create feature predicted purchases by considering time, frequency paretp, recency pareto, and time pareto
Third). Create feature actual purchases by considering:
Fifth). Calculate probability alive and not alive (whether the customer will come again or not).
Sixth). Calculate how much probabilty the customer will buy again in next 30 days by using.

Results

From the calculation using lifetimes library, we obtained results as figure 4 as follows:

![Prediction v/s Actual](image)
From figure 4, by using penalizer coefficient range from (0.001 to 30), we can see that 0.1 is suitable as coefficient range as it able to minimizing the error.

Thus, after obtaining the error, we put together the coefficient range into lifetimes library, and calculate probability alive and not alive which resulted in figure 5 as follows:

![Figure 5. snippets of pareto CLV](image)

From figure 5, we can see that customer ID 12346 probability not alive (will not comeback) is higher than probability alive. Then, customer ID 12349 have probability alive is higher than probability not alive. Thus, we need consider customer ID 12349 so it will comeback again to purchase.

Next part, is to know, how much probability the customer will back in 30 days duration which can see at figure 6

![Figure 6. snippet of probability customer will buy in next 30 days](image)

From figure 6, we can see that customer ID 12347 predicted that it will buy in next 30 day is 0.49 and the actual it bought in next 30 days is 0.52. It also confirmed that when RFM is higher, the probability in next buying is also higher. Thus, we can see that the PARETO/NBD can be used for next buy prediction.

### Conclusion

The results of this study showed that the combination of RFM analysis and the Pareto/NBD model is an effective method for estimating customer lifetime value. The analysis revealed that recency, frequency, and monetary value are significant factors in predicting customer lifetime value. Customers who have high scores in any of the three RFM components are more likely to have a higher lifetime value. Furthermore, the research found that the Pareto/NBD model accurately captures the probability of a customer being active at a particular point in time. This study fills a gap in the literature by showcasing the effectiveness of the Pareto/NBD model in combination with RFM analysis for estimating customer lifetime value using an online retail dataset. The findings from this research have important implications for businesses operating in the online retail industry. Firstly, these findings highlight the importance of focusing on recency, frequency, and monetary value when trying to predict customer lifetime value. Secondly, the use of the Pareto/NBD model in conjunction with RFM analysis offers a straightforward and efficient method for estimating customer lifetime value.

### Bibliography


Popa, A.-L., Sasu, D. V., Department of Management Marketing, Faculty of Economic Sciences, University of Oradea, Oradea, Romania, Tarcza, T. M., & Department of Management


