# Prediction of Customer Lifetime Value and Sales in E-Commerce Business

Mahapara Gayasuddin
Research Scholar
RACE, Reva University
mahapara.ba05@reva.edu.in

Krishna Kumar Tiwari
Mentor
Jio, General Manager
Krishna.Tiwari@ril.com

Mithun Dolthody Jayaprakash
Mentor
RACE, Reva University
mithun.dj@reva.edu.in

***Abstract*—**Customer lifetime value has emerged as an important metric for identifying and reaching out to Customers who make larger and more frequent contributions. As a result, this parameter is dependent on the marketing industry. It is critical to understand the value of a customer's purchases and to recurrently monitor their transaction frequency and value to accurately determine their Customer Lifetime Value (CLV). Also, for any retail company, predicting sales is one of the most crucial business challenges. A company can better manage its inventory if it can forecast how much of each item it will sell each month. Additionally, sales forecasts assist in focusing marketing efforts to improve sales prospects. In order to implement the concept, a two-step strategy was taken. It is begun by estimating the frequency of future transactions from clients. The rate at which users will eventually leave the system has also been anticipated by us. Pareto/NBD or BG/NBD have been utilized to find them. These findings were utilized to determine the monetary value of our consumers. Additionally, the customers have been segmented based on RFM values, and then each group is examined separately in terms of Revenue with Frequency, Revenue with Recency, and Recency with Frequency. Furthermore, in order to forecast the monthly sales volume of each item, a machine learning algorithm has been developed. Quantity is therefore the target variable. Since it is a continuous variable, a regression algorithm will be used.

***Keywords—Probabilistic Models, BG/NBD, Pareto/NBD, CLV Modelling techniques, RFM, Customer Segmentation.***

## *1. Introduction*

Customer lifetime value (CLV) is a concept that has generated interest due to the shift toward a consumer-centric strategy in marketing and the growing accessibility of customer transaction data.

Although customer equity is viewed as an intangible asset that is challenging to quantify, it is possible to estimate its value accurately as technology and science advance. Companies require techniques and benchmarks to manage their customers, and one of them is determining the Client Lifetime Value of each customer to determine how valuable each customer is (CLV). The present value of anticipated future cash flows from consumers is known as CLV [1].

Companies can better understand their customers by using CLV. Once the customers have been classified, the company can tailor its offerings to the demands and behaviour of each group. CLV can assist businesses in understanding the potential worth of their consumers [2]. The majority of empirical research on "lifetime value" has actually computed customer profitability based only on customers' previous behaviour because projecting future revenue streams is difficult. However, in order for our measurements to be true to the concept of CLV, they must look to the future rather than the past. Our capacity to accurately predict future revenues has been a substantial impediment, especially in the event of a "non-contractual" scenario (i.e., when the moment when clients become "inactive" is unseen) [3]. It will be possible to clearly determine the value of each type of customer by determining the lifetime value of the customer segment [4].

According to OnurDogan in his journal, clustering, one of the data mining tasks, has been used to group individuals and objects. In the research, it was also mentioned how important it is to categorise customers so that businesses can tailor their products to the specific wants and needs of their customers [5].

Although CLV can be broken down into many other categories, this study follows [6] theory that it can be broken down into three key management processes: customer acquisition, customer retention, and customer development.

### *A. CUSTOMER ACQUISITION*

Some businesses' methods are ineffective at accurately identifying their profitable consumers. Choosing the ideal clients to target and acquiring them requires careful consideration of factors such as future profitability, firm products, and overall business risk. Those fresh product startups and new business ventures who wish to draw in more clients can benefit from customer acquisition.

Customer acquisition, according to [7] refers to a new or lapsed customer's first purchase. Particularly with new clients who might not be a good fit for the company's value proposition, this kind of process could be dangerous and expensive [8].

### *B. CUSTOMER RETENTION*

Customer retention is typically more affordable and simpler than customer acquisition, particularly in consistent markets with slow growth rates. Maintaining successful clients boosts a business's overall profitability [9].

Customer retention is also the likelihood that a customer will remain "alive" or continue doing business with a company. Customers must notify the company when they end their relationship under contractual arrangements (such as cell phones and magazine subscriptions). However, a company must determine whether a consumer is still active in non-contractual contexts (such as when purchasing books from Amazon) [10].

### *C. CUSTOMER DEVELOPMENT*

Since not all customers have the potential to develop, customer development focuses on a select few. The major objective of this method is to boost the growing value of retained consumers by increasing the value of retained customers to the business. In addition to this, [11] thought that the word "customer development" typically referred to two important areas of activity:

- Upselling: Increasing "share of wallet" by offering more to existing customers.
- Cross-selling is the practise of offering additional products to current customers

By upselling and cross-selling, the revenues generated by customers at any one time will alter.

## *2. Methodology and Data Collection*

The six phases of the research technique are depicted in Fig. 1. The research's initial phase consisted of developing the research

question and establishing its goals. Additionally, it supported the necessity and suitability of the suggested research in Section 1. The identification and justification of CLV models appropriate for use by e-commerce businesses involved in online purchasing were part of the second phase. At this point, the chosen models were also implemented in accordance with the models specified in Section 2.2. Based on the chosen models, the third phase of the process determined the data requirements. Based on this, it was possible to ascertain what information, in what format, and for how long will be required to carry out the research. Data was gathered from numerous e-commerce businesses in the needed structure during the fourth phase. Additionally, the acquired datasets that satisfied the requirements were pre-processed to meet the requirements of the various models. In Section 2.1, the data pre-processing is explained. The datasets from various e-commerce companies are described in Section 2.1.1.

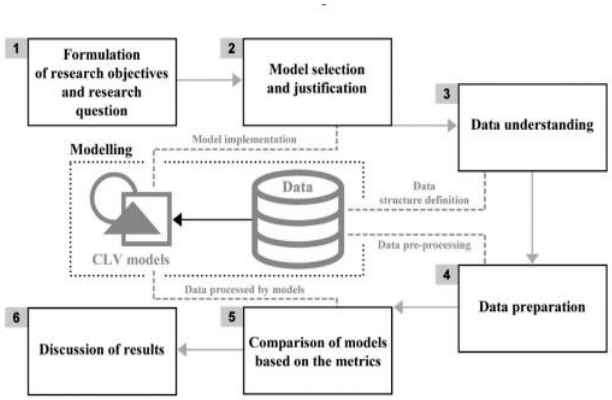


Fig. 1: Methodology of the Project

Based on the statistical indicators given in Section 3.4, the chosen CLV models were compared in the research's fifth phase. Additionally, this section explains how to conduct the comparison and defines a training and testing time. The research issue is addressed in the sixth and last phase, Section 4, and the findings are then the subject of a wider discussion, including pertinent managerial implications, in Section 5.

## *2.1 Data Collection and Pre-Processing*

This paper uses information about non-store online retail. Customers are not obligated and are free to sever their contact with the shop at any time without incurring any costs in such a non-contractual business environment. Due to this, determining whether a consumer is "alive," meaning they will make future transactions, or "dead," meaning they will never make a purchase in the future, is challenging. However, in this project, we have some information on them based on the transactions they have with the shop.

"Buy Till You Die" statistical models help to quantify the behavioral characteristics of the customers and calculate their lifetime value by predicting the number of future transactions that the customer will do and assigning a probability to the customer being "alive".

### *2.1.1 Description of Datasets*

The dataset used for the study is of the transactional data type and includes details on every transaction carried out in a UK-based, registered retail store between January 1, 2019, and September 9, 2021. The list of properties in the dataset includes "Customer ID, Invoice Number, Product Code, Product Description (name), Purchase Quantity, Invoice Date, Unit Price, and Country Name" as shown in Table 1.

| Attribute Name | Type | Description |
|---|---|---|
| Invoice | Nominal | Invoice number of the transaction. Nominal, is an intrinsic 6-digit number assigned specifically to each transaction. If this code starts with the letter 'c', it indicates a cancellation. |
| StockCode | Nominal | A 5-digit integral number known as the nominal is assigned to each unique product. |
| Description | Nominal | Product (item) name. |
| Quantity | Numeric | The quantities of each product (item) per transaction. |
| InvoiceDate | Numeric | Invoice Date and time. |
| Price | Numeric | Product price per unit in sterling. |
| CustomerID | Nominal | Customer number. Nominal, is a five-digit integral number assigned to every customer separately. |
| Country | Nominal | The name of the country where each customer resides. |

Table 1: Attributes of the Dataset

### *2.1.2* Data Preparation

Data cleaning is necessary since this data contains a large number of records without Customer IDs or with negative order quantities. To clean the data, the following steps were taken:

- Records that had negative order quantities and monetary values were filtered out.
- Only records with a Customer ID were kept.

Additional data processing procedures have been taken for the probabilistic technique used:

- The orders were organized by day rather than invoice because the probabilistic model uses a day as the least time unit.
- Only customers who have made a purchase in the last 90 days are taken into account.
- Only the fields necessary for the probabilistic model are kept.

### *2.1.3 Test-Train Split*

In order to prepare the data for training the model, a threshold date had to be chosen. That date divides the orders into two parts:

1. Prior to the threshold date, orders are used to train the model.

2. The goal figure is established using orders that arrive after the threshold date. Our analysis will be conducted during 2021-06-08.

Aggregated data is utilized to build features and targets for each client after the data has been divided into training and target

intervals. The aggregate for the probabilistic model is restricted to the Recency, Frequency, and Monetary (RFM) fields.

The new features are defined as follows:

- monetary_btyd: The average of all orders' monetary values for each customer during the features period. The probabilistic model assumes that the value of the first order is 0. This has been manually enforced.
- Recency: The time between the first and last orders that were placed by a customer during the features period.
- frequency_btyd: The number of orders placed by a customer during the features period minus the first one.
- frequency_btyd_clipped: Same as frequency_btyd, but clipped by cap outliers.
- monetary_btyd_clipped: Same as monetary_btyd, but clipped by cap outliers.
- target_monetary_clipped: Same as target_monetary, but clipped by cap outliers
- Target_monetary: The total amount spent by a customer

*2.2 CLV Calculation and models*

In our paper, the Customer Lifetime Value is calculated in two steps:

1. Using Pareto/NBD or BG/NBD, calculate the rate at which customers will make future purchases and the rate at which they will leave the system in the future.
2. Calculate the monetary value of each customer.

The Following Assumptions have been made:

- Number of transactions made by an active customer follows a Poisson Process given transaction rate of λ, which is E [# transactions in a given period of time]
- Heterogeneity in λ among customers follows a Gamma Distribution
- Probability of customer becoming inactive after every transaction is p
- Heterogeneity in p among customers follows a Beta Distribution
- λ and p is independent among different customers

BTYD models (Pareto/NBD or BG/NBD) give us the following three outputs:

- P(X(t) = x | λ, p) - probability of observing x transactions in given time t
- E(X(t)| λ, p) - expected number of transactions in given time t
- P(τ > t) – the probability of the customer being inactive at time t

The expected number of transactions for a client with prior observed behaviour specified by x, tx, and T is then determined using these fitted distribution parameters, where x is the number of historical transactions and tx is the date of the most recent purchase and T = The customer's age [12].

$$E\big(Y(t) \mid X = x, t_x, T, r, \alpha, a, b\big) = \frac{\frac{a+b+x-1}{a-1}\left[1-\left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x} {}_2F_1\left(r+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t}\right)\right]}{1+\delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}} \quad (1)$$

The expected number of transactions in a future period of length t for an individual with past observed behavior ($X = x, t_x, T$; where x = n. historical transactions, $t_x$ = time of last purchase, and T = Age of a customer) given the fitted model parameters r, α, a, b.

The outputs of the probabilistic model outlined above are utilized to project the customers' future financial worth. The probabilistic approach presupposes that the distribution of monetary value is gamma-gamma. For both the models, the python package called Lifetimes is used.

*2.3 Predicted RFM Analysis*

To assess the precision of our CLTV model prediction, RMSE is employed. The RMSE for the Pareto/NBD model is $3166.96, whereas the RMSE for the BG/NBD model is $3150.40. so we have gone ahead with BG/NBD Model to calculate the CLV.

| | CustomerID | actual_total | predicted_num_purchases | predicted_value | predicted_total | error | predicted_purchases | predicted_CLV | CLV |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 12347 | 4921.53 | 0.0 | 0.0 | 4420.486919 | 501.043081 | 1.478590 | 1018.096919 | 50.904846 |
| 1 | 12348 | 2019.40 | 0.0 | 0.0 | 2269.090328 | -249.690328 | 0.910700 | 559.690328 | 27.984516 |
| 2 | 12349 | 4428.69 | 0.0 | 0.0 | 3087.541811 | 1341.148189 | 0.461909 | 416.401811 | 20.820091 |
| 3 | 12352 | 2849.84 | 0.0 | 0.0 | 2550.295710 | 299.544290 | 1.370917 | 644.685710 | 32.234285 |
| 4 | 12356 | 6371.73 | 0.0 | 0.0 | 8018.299823 | -1646.569823 | 1.146721 | 1704.919823 | 85.245991 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1944 | 18273 | 357.00 | 0.0 | 0.0 | 443.109999 | -86.109999 | 0.535323 | 137.109999 | 6.855500 |
| 1945 | 18276 | 1656.52 | 0.0 | 0.0 | 1576.457402 | 80.062598 | 0.646512 | 255.797402 | 12.789870 |
| 1946 | 18277 | 1180.05 | 0.0 | 0.0 | 1327.462743 | -147.412743 | 0.516341 | 257.792743 | 12.889637 |
| 1947 | 18283 | 2664.90 | 0.0 | 0.0 | 2086.504437 | 578.395563 | 2.021399 | 360.304437 | 19.015222 |
| 1948 | 18287 | 4182.99 | 0.0 | 0.0 | 3672.257890 | 510.732110 | 0.711216 | 561.267890 | 28.063395 |

Fig. 2: CLV Calculation using BG/NBD Model

Consider for instance a customer that has made a purchase every day for four weeks straight, and then is inactive for months. What are the chances he/she is still “alive”? The chances are pretty slim. On the other hand, a customer that historically made a purchase once a quarter, and again last quarter, is likely still alive. This can be visualized using the frequency/recency matrix, which computes the expected number of transactions an artificial customer is to make in the next time period, given his recency (age at last purchase) and frequency (the number of repeat transactions he has made).

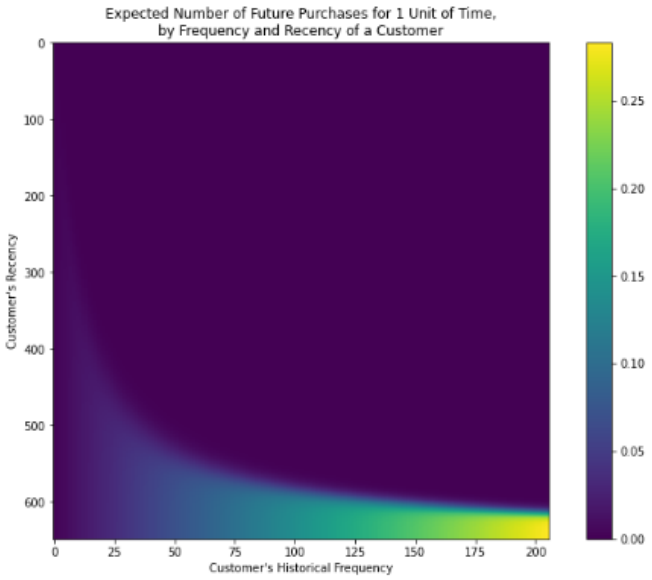


Fig. 3: Frequency & Recency Matrix Using BG-NBD Model

From the above Fig. 3, it can be seen that our best customers are where the frequency is 200 and Recency is 600 plus. Future best customers will probably be those who have lately made a lot of purchases. Customers who have made numerous purchases but not recently (top-right corner) have likely stopped shopping there.
Additionally, there is that tail that represents the consumer who spends infrequently. Since they haven't been seen recently, it can't be assured if they dropped out or were simply in between transactions, but they may buy again. It can be predicted which customers are still alive:

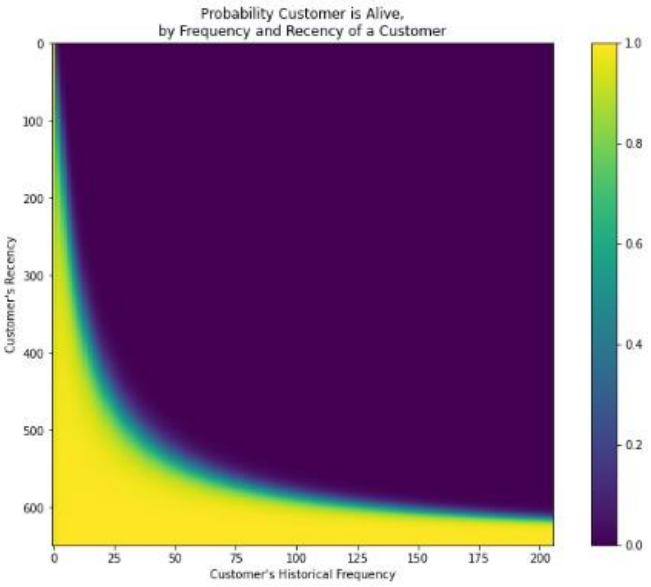


Fig. 4: Probability Customer is Alive Using BG/NBD Model

Customers who have recently made a purchase are nearly certainly still "alive". Customers that frequently made purchases in the past but not recently are likely no longer present. And the more they had previously purchased, the more probable it was that they would stop. They are shown in the upper-right corner. From above Fig. 4, it can be seen that our 80% of customers have already churned or it can be said that they dropped.

### *2.4 Customer Segmentation*

Clustering is used on each of the three criteria — recency, frequency, and monetary value — separately to achieve segmentation. Our model uses k-means, and the elbow plot's recommended number of clusters is 4. We apply a weighted sum to these various clusters to produce an overall score as shown below in Fig. 5.

| OverallScore | recency | frequency_btyd | target_monetary |
|---|---|---|---|
| 0 | 262.327411 | 3.474619 | 2341.824873 |
| 1 | 565.346505 | 7.057751 | 4461.122918 |
| 2 | 122.346341 | 4.895122 | 2966.897902 |
| 3 | 426.146497 | 7.388535 | 6039.409703 |
| 4 | 597.916058 | 20.485401 | 12913.294380 |
| 5 | 484.312500 | 42.750000 | 156164.135625 |
| 6 | 441.407407 | 18.592593 | 8400.639815 |
| 7 | 428.000000 | 18.000000 | 144458.370000 |

Fig. 5: Overall Score based on RFM

It can be seen that three main groups formed after examining the mean Recency, Frequency, and monetary values of these clusters. Following that, we will assign labels to Low Value, Mid Value, and High Value consumers. These are binned into three segments

- 0 to 3: Low Value
- 4 to 5: Mid Value
- 5+: High Value

The Fig. 6 shows the Customer Segmentation using K-means. Segment 0 is with 82.29% customers are the low value customers and Segment 2 with close to 15% customers. The marketing team should target this group to retain them and provide them with offers.

```
t6_cust.groupby('Segment').customer_id.count()/t6_cust.customer_id.count()*100
```

```
Segment
High-Value     2.821960
Low-Value     82.298615
Mid-Value     14.879425
Name: customer_id, dtype: float64
```

Fig. 6: Clustering with K-Means

We want to strategize customer retention based on each group's unique qualities because this will increase retention value as we have three groups. Therefore, in order to address these concerns, we need to determine where these customers are falling behind, such as whether they purchase things of lower value or in smaller quantities, less frequently, inconsistently, or not at all.

The customers are plotted against

- Revenue with Frequency
- Revenue with Recency
- Recency with Frequency

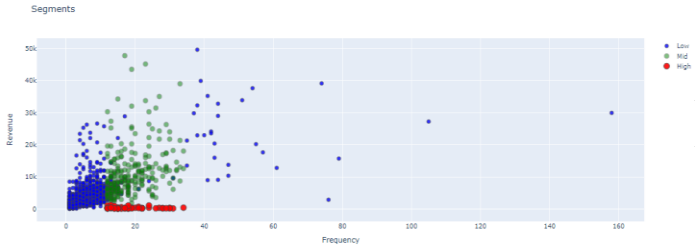


Fig. 7: Frequency Vs Monetary

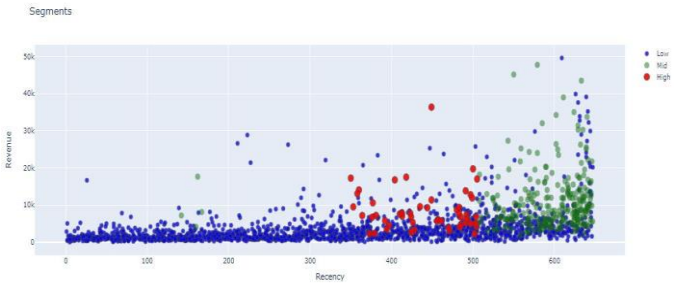


Fig. 8: Recency Vs Monetary

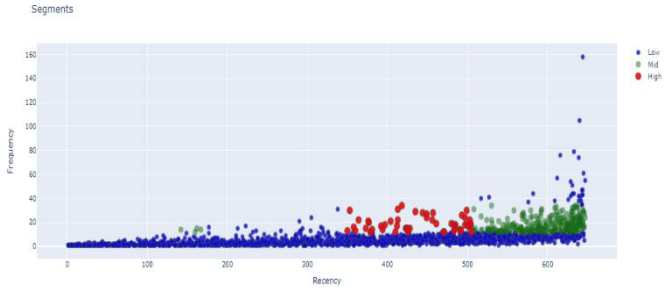


Fig 9: Frequency Vs Recency

For these brackets, we now develop unique strategies, sometimes even inside a single large bracket. Following are our strategies based on the analysis of our hypothesis.

We are aware that high-value clients purchase less frequently but generate higher revenue. Additionally, they haven't recently made any purchases. We must thus ask if they are sleeping or deceased consumers. They may also be groups of people who make purchases depending on time intervals, such as seasonal shoppers or people who make purchases based on quarterly bonuses.

The mid-value clients haven't purchased recently, but they do have a wide range in terms of frequency and income. They might develop into ardent brand supporters. They might also be high income producers who either make frequent major purchases or have a penchant for pricey goods. We need to look at this a little further, but we also need to make them more recent generally.

We are aware of the low revenue and low frequency of our low value customers' transactions, but they also exhibit erratic purchasing behaviour, which the company may capitalize on and enhance.

*2.5 Sales Prediction using Machine Learning Model*

In order to forecast the monthly sales volume of each item, a machine learning algorithm has been developed. Quantity is therefore the target variable. Since it is a continuous variable, a regression algorithm will be used. The data from November 2021 through December 2021 will serve as our test set for this project, and the remaining data will be used to train our model. 180,661 observations are in the train set after the split, while 36,092 observations are in the test set, for a ratio of 83:17. The data have been trained on many algorithms, evaluate them, and choose the one that performs the best based on our assessment criteria to determine the model that will serve our needs best.

The following mentioned algorithms have been used:

1. Linear Regression
2. Regularization Model – Ridge
3. Regularization Model - Lasso
4. Ensemble Model - Random Forest

**Performance Evaluation**

RMSE (Root Mean Square Error) of the prediction and time spent to fit/predict the model has been used to compare the performance of several methods and choose the best. It is preferable to use a model with the lowest RMSE and time taken.

**Model Comparision**

In the Fig. 10 below, the RMSE for train and test datasets for several algorithms has been compared

| Modelling Algo | Train RMSE | Test RMSE | Hyperparameters | Training+Test Time(sec) |
|---|---|---|---|---|
| Random Forest | 21.222628 | 24.849052 | {'n-jobs':-1,'n-estimators':1000,'min samp…} | 6706 |
| Linear Regression | 28.313427 | 28.364165 | | 0.51 |
| Ridge Regression | 28.313427 | 28.364170 | {'alpha': 145} | 6.91 |
| Lasso Regression | 28.313722 | 28.366796 | {'alpha': 0.24} | 31.32 |

Fig. 10: Model Comparision

The training and test RMSE for each method is shown graphically below in Fig. 10.

Even though the best methods are overfitted, still the model has obtained low RMSE for the test dataset, therefore, this does not cause us too much concern.

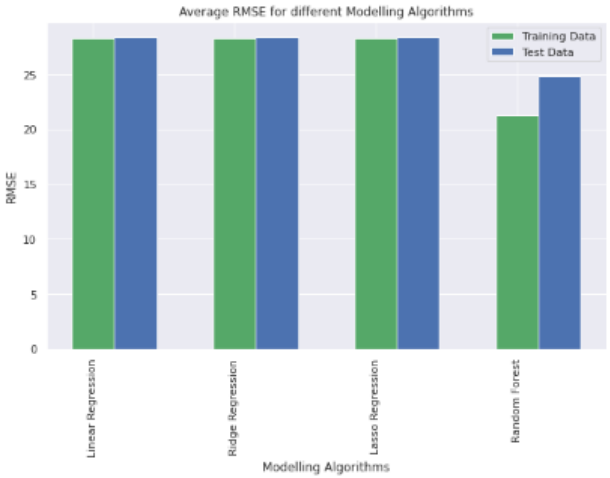


Fig. 11: Average RMSE for different Modelling Algorithms

On the test dataset, Random Forest exhibits the best performance. Need to consider how long it took to fit the model before choosing one of them. In each algorithm, the prediction time was significantly shorter than the fit time. In the test data. Random Forest yields an RMSE of 24.84. Therefore, have settled on Random Forest as our chosen algorithm.

## 3. Conclusion

We have used the Pareto/NBD and BG/NBD models to predict the Customer Lifetime Value. Furthermore, we have performed customer segmentation on RFM values to get 3 major groups as mentioned and have analyzed them individually with respect to Revenue with Frequency, Revenue with Recency and Recency with Frequency.

With Low segmentation, customers dominate the outcomes of customer CLV analysis. For customers in the low segment, the approach should be centered on upselling and cross-selling tactics, or on tactics to boost sales and increase revenue, which will raise the CLV of the customer. The tactical approach that can be taken is to increase efficiency, better the price clause when the work contract ends, or discontinue the partnership if the price adjustment cannot be agreed upon because the Low segment tends to produce negative CLV. Finally, given that loyal customers contributed the majority of sales, it can be concluded that businesses should prioritize customer retention.

For any retail company, predicting sales is one of the most crucial business challenges. A company can better manage its inventory if it can forecast how much of each item it will sell each month. Additionally, sales forecasts assist in focusing marketing efforts to improve sales prospects.

This study served as a starting point for more research because of its limitations and the wide range of CLV-related research prospects. In the future, the same study can be conducted in other industries like insurance, Banking, or telecommunication industry and be able to compare the results in various industries.

## REFERENCES


[1] Pfeifer, Phillip E., Mark E. Haskins, and Robert M. Conroy (2005), "Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending," Journal of Managerial Issues, forthcoming.

[2] Kim, E., and Lee, B. 2007. An Economic Analysis of Customer Selection and Leverage Strategies in A Mar- ket where Network Externalities Exist. Decision Support Systems. 44 (1) 124-134.

[3] Bell, David, John Deighton, Werner J. Reinartz, Roland T. Rust, and Gordon Swartz (2002), "Seven Barriers to Customer Equity Management," Journal of Service Research, 5 (August), pp.77–86.

[4] Buraera J, Kadir. Abd, Alam S. 2014. Customer Lifetime Value SegmenKonsumerdan Retail pada PT. Bank Negara Indonesia (Persero) Tbk. JurnalAnalisis ISSN, 3 (2).

[5] Onur, D., Ejder, A., and ZekiAtil, B. 2018. Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry. International of Contemporary Economics and Administrative Sciences, pp.1-20.

[6] Buttle, F. (2008). Customer Relationship Management: Concepts and Technologies. (2nd ed).Elsevier Butterworth-Heinemann.

[7] Gupta, S. and Zeithaml, V. (2006). Customer Metrics and Their Impact on Financial Performance. Marketing Science. 25 (6), pp.718- 739.

[8] Bolton, R.N. and Tarasi, C.O. (2007). Managing Customer Relationships. ed) Emerald Group Publishing Limited.

[9] Kumar, V. and Rajan, B. (2009). Profitable Customer Management: Measuring and Maximizing Customer Lifetime Value. Management Accounting Quarterly. 10 (3), pp.1-18.

[10] Gupta, S. and Zeithaml, V. (2006). Customer Metrics and Their Impact on Financial Performance. Marketing Science. 25 (6), pp.718- 739.

[11] Murphy, J.A. (2005). Converting Customer Value: From Retention to Profit. (1st ed).Wiley.

[12] Fader, Peter S. Hardie, Bruce G.S. Lee, Ka Lok (2005), "RFM and CLV: Using iso-value curves for customer base analysis" , pp. 415-430.


**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**