**PROJECT REPORT**

**PREDICTING CUSTOMER LIFETIME VALUE**

**NAME: -** NISHANT MUDGAL

**UTA ID** - 1002117213

**BACKGROUND OF STUDY**:

Customer Lifetime Value (or CLV) means the monetary value a customer would bring to a business from their first purchase throughout their lifespan, considering they remain associated with it. This metric is predominantly being used by E-commerce businesses to gauge and predict an organization’s future profit. CLV assists businesses in optimizing their marketing techniques along with the costs incurred in retaining high-value customers and customer acquisition, identifying the most profitable products, improving marketing techniques, etc. In recent years, advancements in machine learning and data analysis have made it possible to predict CLV with greater accuracy and precision, enabling companies to develop more targeted and effective marketing strategies.

**RESEARCH PROBLEM:**

This study examines the importance of predicting a customer’s lifetime value from a business perspective ranging from their first purchase to the last. This study would also highlight the type of customers that need to be retained.

**RESEARCH PROBLEM:**

* Why do businesses need to calculate the Customer’s Lifetime Value?
* Which type of customer’s lifetime value should be predicted?
* How much profit would a customer bring to the business?
* Which type of model should be used for predicting the Customer Lifetime Value?

**LITERATURE REVIEW**

For this project, I identified several reputable journals in the field of Customer Relationship Management, including ‘The Journal of Business Research’, ‘Journal of Service Research’, ‘2nd International Conference on Computer Research and Development, ICCRD 2010’, ‘Journal of Interactive Marketing’, ‘TQM Journal’ and ‘Journal of International Marketing’.

After identifying the journals, I conducted a search for peer-reviewed articles related to the

topic of interest. I filtered out at least 30 relevant articles based on the title and abstract.

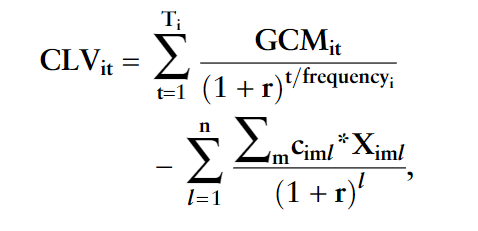
**CLV (CUSTOMER LIFETIME VALUE)**

CLV refers to “the present values of the future cash flow attributed to the customer relationship” (***Zhang, et al,2016, pt.2***). In other words, how much profit or monetary value a customer would bring throughout his lifetime to a company. In most cases, the lifetime of a customer is captured within the first three years (***Gupta and Lehman 2005***). CLV is one of the most critical factors to consider when making marketing decisions and is an important aspect of Customer Relationship Management (CRM). While calculating CLV it is important to identify the customers who will bring the maximum profit by segregating them into various categories (high, medium, low) based on profitability and then focusing the marketing efforts on the most beneficial ones.

CLV is not limited to just the e-commerce market but is applicable to almost every other industry. It has gained well acceptance from the company’s management to assist their business management process. According to the Predictive Analytics World Report (Feb 2009), 51.7% of survey respondents among the vendors and consultants also claimed that CLV modeling was best to improve the relationship with their customers.

The CLV metric is a function of the purchase frequency, the predicted customer margin, and the marketing resources allocated to the customer (***Kumar and Pansari 2016***).

Earlier research (***Venkatesan and Kumar 2004***) computed the CLV metric using the below formula: -



where,

CLVit = lifetime value of customer ‘i’ in time ‘t’;

GCMit = predicted gross contribution margin from customer i in period t, measured in dollars;

r = discount rate for money;

ciml = unit marketing cost for customer ‘i’ in channel ‘m’ in year ‘l’;

ximl = number of contacts to customer i in channel m in year l;

frequencyi = predicted purchase frequency of customer i in each year;

n = number of years to forecast; and

Ti = predicted number of purchases made by customer i until the end of the planning period.

Although there are various other versions of the formula used to calculate CLV, all of them are derived from the most basic one in ‘**Customer Lifetime Value’** by **Torcy, Guy de:-**

**CLV = (total customer revenue)\*(number of loyal years)\*(company profit margin)**

Most existing CLV models have 3 basic elements: revenue from the customer, the costs of serving a customer, and the customer retention rate. At the same time, there are several drivers that affect these elements which will be discussed later in this review.

**KEY METRICS OF CLV**

1. **Revenue from the customer (Purchase History):**

This variable is computed by counting the number of purchases made by a customer during a time period. A retailer would like this count to increase as much as possible, hence, it is their objective to ensure repeated purchase activity in the future which generates revenue. ***Morgan and Hunt (1994)*** argue that satisfactory interactions lead to greater trust which in turn leads to a long-lasting relationship and eventually higher profits.

However, ***Reinartz and Kumar (2003)*** show that a customer with moderate purchasing frequency in a noncontractual setting tends to stay longer with a business. A noncontractual setting is one in which the firm does not observe a customer defection and a relationship between customer purchase and customer lifetime is not certain (***Fader et al. 2005***; ***Schmittlein and Peterson 1994***). Whereas, a contractual context is one in which customer defection is observed and a longer customer lifetime implies higher customer lifetime value (***Thomas 2001***, ***Bolton 1998***, ***Bhattacharya 1998***). Nonetheless, in the future, it is imperative to increase a customer’s purchasing frequency.

1. **Gross Contribution Margin:**

Gross Contribution Margin per purchase is calculated by deducting the cost of goods sold from the revenue received from a customer after each purchase. This component is affected by the factors such as the number of times a business contacts a customer, and different promotions or discounts offered alongside other marketing activities focused on the customer.

1. **Cost of serving a customer (Marketing costs):**

Marketing cost is the cost of various activities that a firm incurs in maintaining the relationship with customers. These costs include both development and retention costs. A major portion of these costs comes from various marketing strategies and promotional activities that a business indulges in using emails, social media, flyers, advertisements, telephone calls, face-to-face interactions, etc. These costs are variable in nature and are entirely at a business’s discretion.

**VARIOUS MODELS DEVELOPED SO FAR: -**

Different models for measuring CLV arrive differently at estimates of future customer purchase behavior. Some models consider discrete time intervals and assume that each customer spends a given amount (an average amount spent in the data) during each interval of time. This information along with some other assumptions about the customer’s lifetime health is used to estimate the lifetime value of each customer by a discounted cash-flow method (***Berger and Nasr 1998***).

In another model, ***Rust et al (2004)*** combined the frequency of category purchases, brand-switching patterns, and the firm’s contribution margin to estimate the lifetime of each customer. Because customer purchase behavior may change over their lifetime with the firm, hence, methods that consider the customer’s past purchase behavior to predict their lifetime value have an edge over the other ones.

A popular method that follows such an approach in the noncontractual context is the negative binomial distribution (NBD) – Pareto model by ***Schmittlein et al (1987)***.

Another approach that can naturally incorporate past purchase behavior outcomes into future expectations is a Bayesian approach ***(Rossi and Allenby 2003).*** Bayesian methods can incorporate such prior information in the structure of the model easily. Furthermore, this approach can be used in any context. Therefore, we use such an approach to measure customer lifetime value, leveraging this extra information available to the firm.

**DATA DESCRIPTION**:

For this study, I will be using a public dataset available on the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/online+retail).

The dataset used in this research project contains information about transactions that took place between 01/12/2010 and 09/12/2011 for the UK- based and registered non-store online retail. The company sells unique all-occasion gifts with the majority of the customer base being wholesalers.

Attribute Information:

* Invoice No: Invoice number. Nominal, is a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation
* Stock Code: Product (item) code. A 5-digit integral number is uniquely assigned to each distinct product.
* Description: Product (item) name.
* Quantity: The quantities of each product (item) per transaction. Numeric
* Invoice Date: Invoice Date and time. Numeric, the day and time when each transaction was generated
* Unit Price: Unit price. Numeric, Product price per unit in sterling
* Customer ID: A 5-digit integral number uniquely assigned to each customer
* Country: The name of the country where each customer resides.

The dependent variable (DV) in this dataset is the Customer Lifetime Value while the independent variables (IVs) are frequency of purchase, recency of purchase, and the total amount of purchases marketing and retention efforts made by the company. The size of the dataset is 22.6 MB with 541,910 observations.

The data was then cleaned and preprocessed using Python libraries. The final dataset is in .xls format and will be analyzed using the Beta Geometric/Negative Binomial Distribution (BG/NBD) and Gamma- Gamma model.

**ANALYSIS:**

The code performs several data cleaning and analysis tasks to predict customer lifetime value (CLV). Firstly, the code extracts relevant columns such as ‘CustomerID,’ ‘Invoice Date’, ‘Quantity’, ‘Invoice No’, and ‘Total Sales’ (computed by multiplying the column ‘Quantity’ with ‘Unit Price’). Next, it performs descriptive statistics of the data to gain insights into the data distribution. The code identifies negative values in the ‘Quantity,’ ‘Unit Price’, and ‘Total Sales’ columns, which may indicate returned orders or discounts offered to customers. The negative values are removed from the data. Further, the code checks for null values and removes records without a ‘CustomerID’ since they are not needed for the analysis. The code then computes the most recent and first transaction dates, total revenue generated, total quantities sold, and the total number of unique customers. Finally, the program prints a summary of the cleaned data. These steps are crucial for predicting customer lifetime value accurately.

In this research, the Beta Geometric/Negative Binomial Distribution model (BG/NBD) is used which is one of the most used statistical methods to predict the customer lifetime value which also acts as an alternative to the Pareto/NBD model. It predicts future transactions for a customer. For this research, the BG/NBD model is combined with the Gamma-Gamma model adds the monetary aspect of the customer transaction, and eventually helps in calculating and predicting the Customer’s lifetime value. The BG/NBD has a few assumptions:

* When a user is active, the number of transactions in a time t is described by Poisson distribution with rate lambda.
* Heterogeneity in transactions across users (difference in purchasing behaviour across users) has Gamma distribution with shape parameter r and scale parameter a.
* Users may become inactive after any transaction with probability p and their dropout point is distributed between purchases with Geometric distribution.
* Heterogeneity in dropout probability has Beta distribution with the two shape parameters alpha and beta.
* Transaction rate and dropout probability vary independently across users.

Post-cleaning of the data, the program creates an RFM (Recency, Frequency, Monetary value) summary from the transactional data using the lifetimes’ package. Then a graph depicting the distribution of frequency to understand customer frequency was plotted and we found out how many one-time buyers are there. The is then fitted to the summary data to compute the probability of the customer being alive. The relationship between frequency and recency is then plotted using the ‘plot\_probability\_alive\_matrix’ function. The Gamma-Gamma model is used to estimate the average sales for each customer. The predicted value of the average sale is compared to the actual value to check for accuracy. Finally, the CLV is predicted for each customer for the next 2 months using the ‘customer\_lifetime\_value’ function. Overall, the analysis involves fitting different models to the data, exploring the relationship between various features, and predicting the CLV for each customer using the estimated models.

**FINDINGS AND DISCUSSION:**

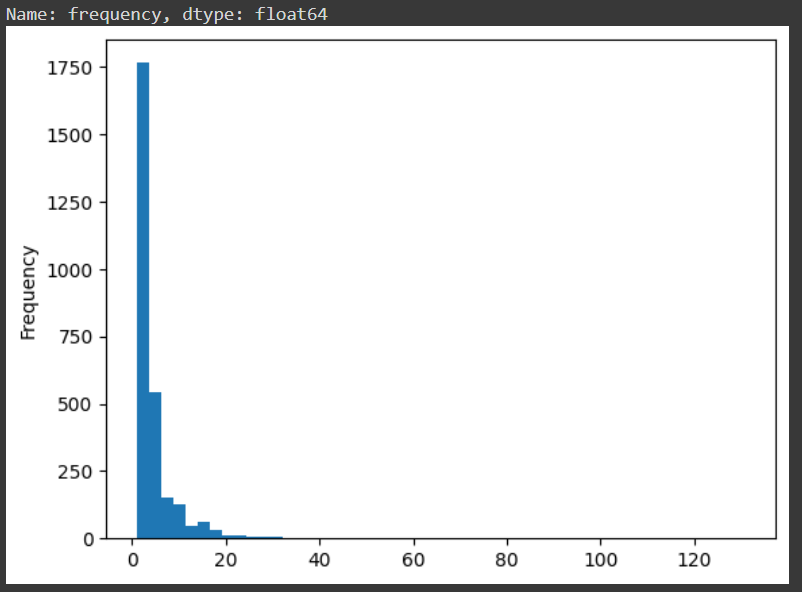
The project aims to predict customer lifetime value using the Beta Geometric/Negative Binomial Distribution (BG/NBD) model.

The dataset used is "Online Retail" and it has a shape of (541,909, 8), meaning it contains 541,909 records with 8 columns. The columns containing irrelevant information are dropped. The cleaned dataset has 531,283 rows and 5 columns containing the relevant data, which includes "CustomerID," "InvoiceNo," "InvoiceDate," "Quantity," and "UnitPrice." A new column, "Total Sales," is created by multiplying the "Quantity" and "UnitPrice."

Further analysis of the dataset revealed that some values in "Quantity," "UnitPrice," and "Total Sales" were negative. These negative values could be due to returned orders or discounts offered to customers. Therefore, they are removed from the dataset since only positive values are relevant for this analysis.

Out of the total records, 25% of the rows do not have a "CustomerID," which is not needed for this analysis. Therefore, those rows are removed, leaving 397,924 rows in the final dataset.

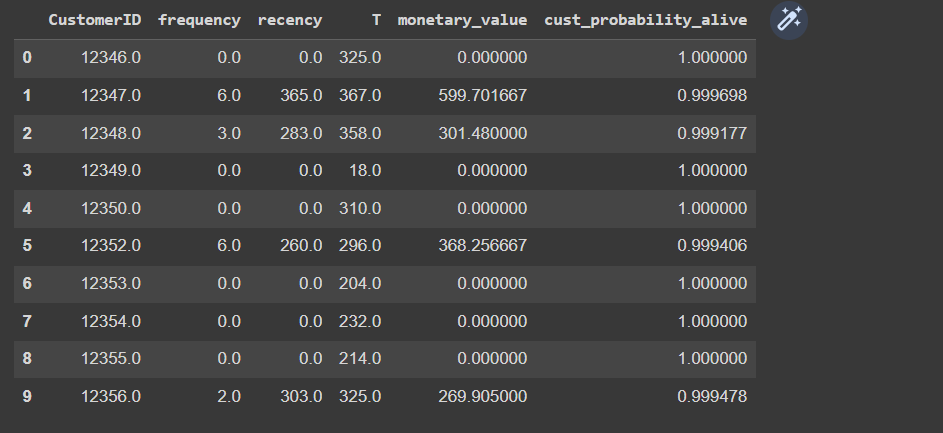
The cleaned dataset is then used to create a summary data frame, where the Recency, Frequency, and Monetary value (RFM) of each customer are calculated. The summary data frame is then used to create a frequency histogram and compute the percentage of one-time buyers.



From the histogram, we can see that most customers have made transactions with a revenue of less than 200 units, with a small number of customers having made transactions with revenue above 1000 units. The distribution is skewed to the right, which means that a few customers have contributed significantly to the total revenue. This information is useful in understanding the revenue pattern and identifying high-value customers.

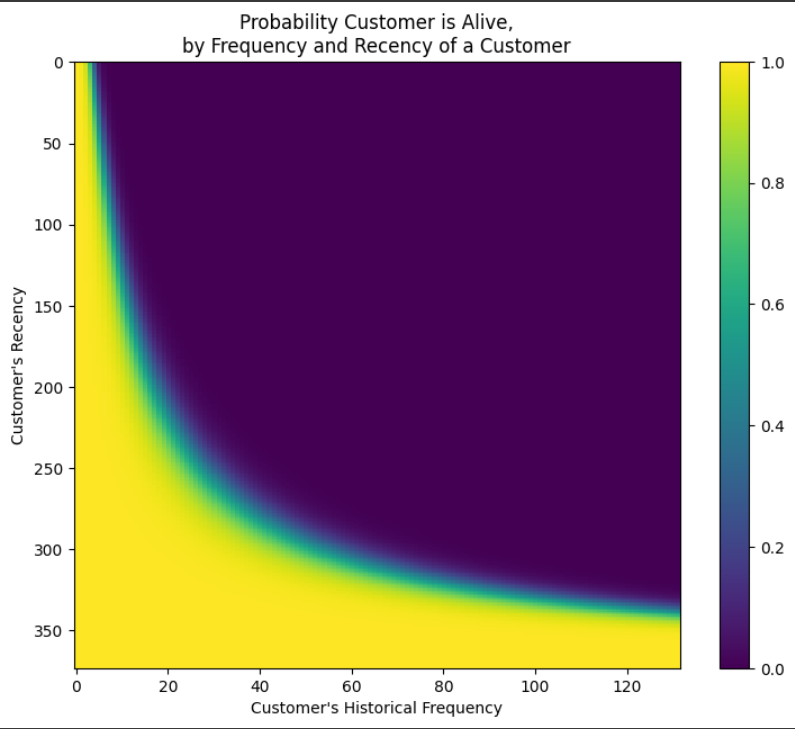
Next, the cleaned dataset is fitted to the Beta Geometric/Negative Binomial Distribution (BG/NBD) model using the "BetaGeoFitter" class in the Lifetimes package. The probability of the customer being alive is computed using the model.

The below output cell shows the probability of the first 10 customers:

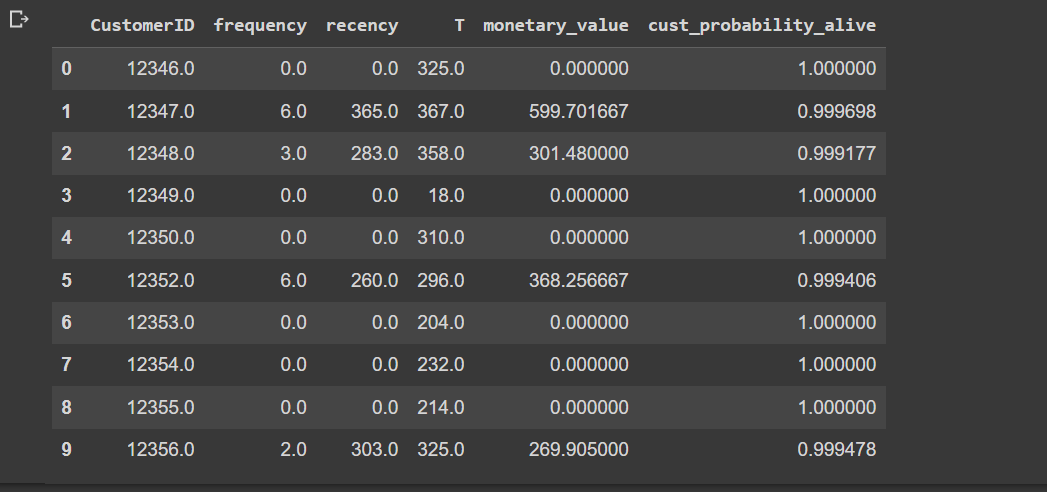


The probability of a customer being alive is calculated using the frequency (how frequently a user made a transaction) and recency (time between the user’s first and last transaction). If a customer made a transaction multiple times, and the time between their first and last transaction is high, then the probability of that user being alive is low and likewise, if the frequency and recency are high then, the probability of that user being alive is high.

The relationship between frequency and recency is plotted, and the predicted number of transactions is computed using the model. A graph of the probability of a customer being alive, by frequency and recency is plotted using the matplotlib library of Python.

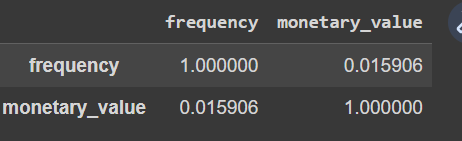


Using the Beta Geo Fitter class in Python’s lifetimes’ package, the probability of a customer being alive was calculated and stored under the column heading ‘cust\_probability\_alive’. The top 10 rows are displayed below:

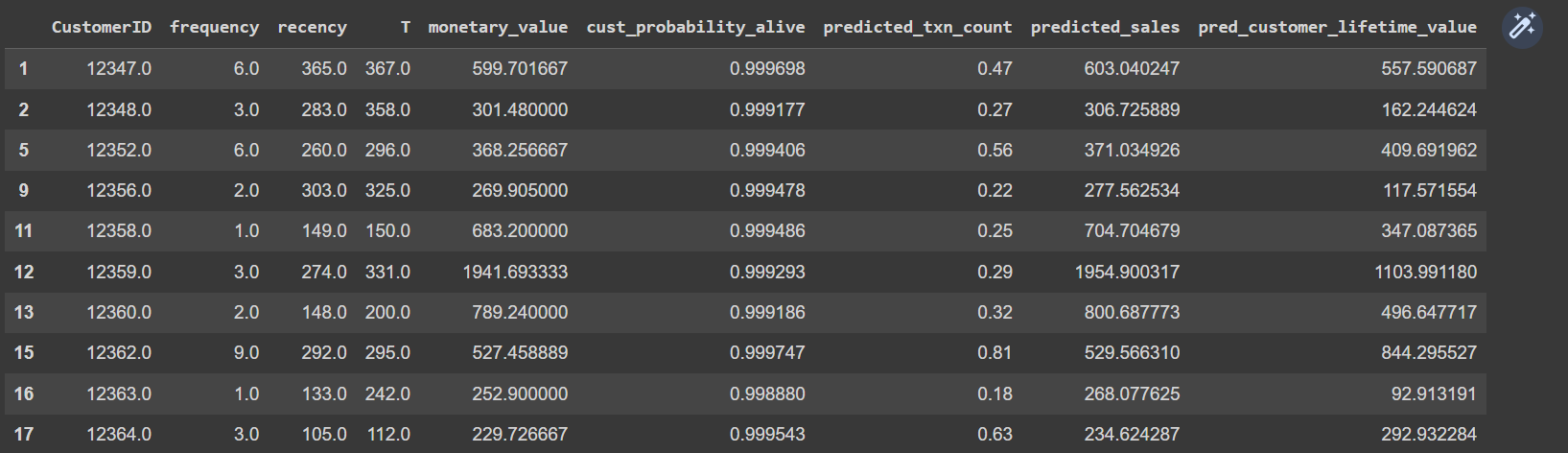


Next using the combination of the BG/NBD and Gamma-Gamma model, future transactions for each customer were predicted and the values are stored under the column ‘predicted\_txn\_count’. The value is calculated by dividing the frequency of a customer’s transaction by the recency and multiplying the result by the number of days for which we are trying to predict.

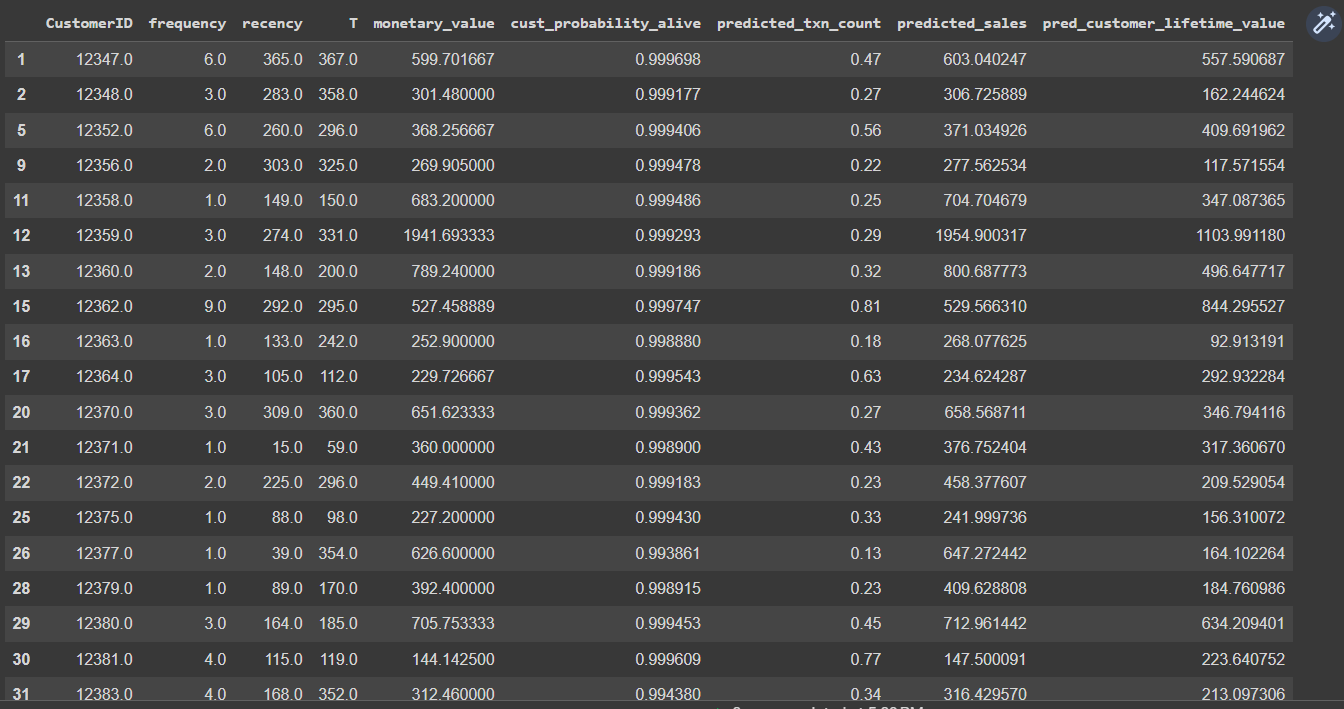
Next, the correlation between frequency and monetary value is evaluated using the corr() function which comes out to be weak (The correlation coefficient between "frequency" and "monetary\_value" is 0.015906, which is close to zero. This indicates that there is little or no linear relationship between the two variables.), indicating that the model can be fitted to the data.



The values under the column ‘predicted\_sales’ represent the conditional expectation of the average profit per transaction made by the company and are calculated using the ‘conditional\_expected\_average\_profit()’ function.



Finally, we calculate the Customer Lifetime Value for the next 2 months by using the customer\_lifetime\_value() function that takes in the BG/NBD model as one of the parameters alongside the other predictor variables. The output of the first 20 rows is shown below:

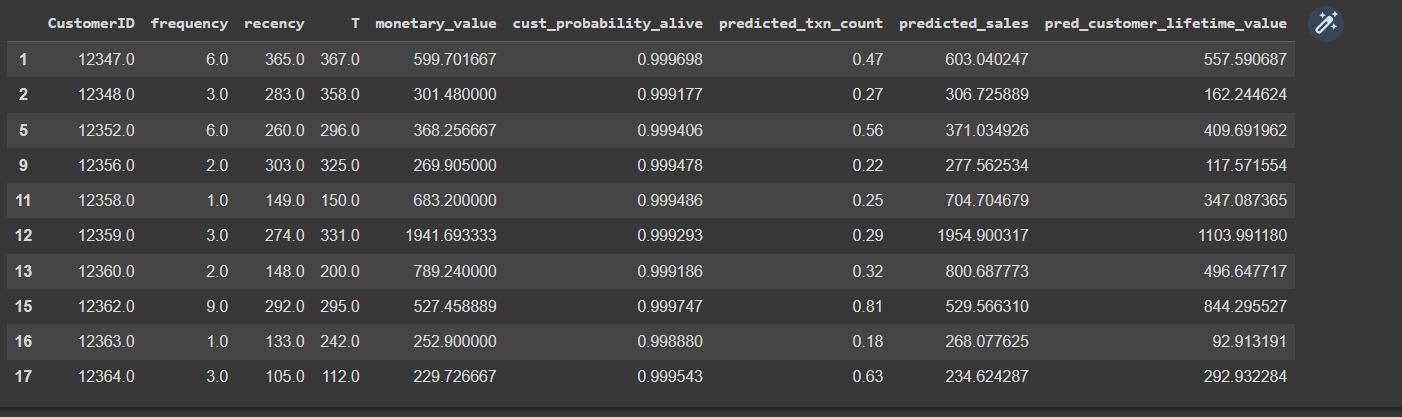


Overall, this analysis provides insights into customer lifetime value using the BG/NBD and the Gamma Gamma model.

This research provides one of the ways to predict the Customer Lifetime Value, that can help a business identify areas of improvement in their products or services by highlighting the factors that drive customer loyalty and value.

This project will aid a company to prioritize its marketing efforts toward high-value customers which would be those who have a high probability (cust\_probability\_alive) of being alive and a high predicted transaction count because these types of customers would be profitable in the long run.

This research also shows the predicted profit (‘predicted\_sales’) a customer will bring to an organization while being associated with the business.



Moreover, using BG/NBD model alongside the Gamma Gamma model for predicting CLV has an edge over other models, as it provides two different dimensions for an analyst to interpret the CLV. BG/NBD tries to predict future transactions and the Gamma Gamma model adds the monetary aspect to the overall model.

**LIMITATIONS: -**

One of the limitations of this research is that this model may not be generalizable to all customer segments, as it may be biased compared to the other datasets out there.

Secondly, this research did not incorporate the dataset that contains customer demography and external factors like economic conditions, where segments could have been created and then the value of CLV be predicted. The segment-level information is vital and can be used for personalized targeting of customers.

**REFERENCES:**

***THE DATASET USED FOR THIS PROJECT CAN BE FOUND*** [***HERE***](https://docs.google.com/spreadsheets/d/184SJh3s7s_OlV1sx28sA6eW5OAME3CmI/edit?usp=sharing&ouid=108339557363813137685&rtpof=true&sd=true)

***OR USE THE BELOW LINK***

**(https://docs.google.com/spreadsheets/d/184SJh3s7s\_OlV1sx28sA6eW5OAME3CmI/edit?usp=sharing&ouid=108339557363813137685&rtpof=true&sd=true)**

Chen, S. (2018). Estimating Customer Lifetime Value Using Machine Learning Techniques. In *Data Mining*. InTech. <https://doi.org/10.5772/intechopen.76990>

Yean, L. C., & Khoo, V. K. T. (2010). Customer relationship management: Lifecycle of predicting customer lifetime value. *2nd International Conference on Computer Research and Development, ICCRD 2010*, 88–92. <https://doi.org/10.1109/ICCRD.2010.24>

Malthouse, E. C., & Blattberg, R. C. (2005). Can we predict customer lifetime value? *Journal of Interactive Marketing*, *19*(1), 2–16. <https://doi.org/10.1002/dir.20027>

Yılmaz Benk, G., Badur, B., & Mardikyan, S. (2022). A New 360° Framework to Predict Customer Lifetime Value for Multi-Category E-Commerce Companies Using a Multi-Output Deep Neural Network and Explainable Artificial Intelligence. *Information (Switzerland)*, *13*(8). <https://doi.org/10.3390/info13080373>

Pollak, Z. (n.d.). *Predicting customer lifetime value-e-commerce use case*. [www.clearly.ca](http://www.clearly.ca)

Borle, S., Singh, S. S., & Jain, D. C. (2008). Customer lifetime value measurement. *Management Science*, *54*(1), 100–112. <https://doi.org/10.1287/mnsc.1070.0746>

Dandis, A. O., & Al Haj Eid, M. B. (2022). Customer lifetime value: investigating the factors affecting attitudinal and behavioural brand loyalty. *TQM Journal*, *34*(3), 476–493. <https://doi.org/10.1108/TQM-12-2020-0311>

Kumar, V., Venkatesan, R., Bohling, T., & Beckmann, D. (2008). The power of CLV: Managing customer lifetime value at IBM. *Marketing Science*, *27*(4), 585–599. <https://doi.org/10.1287/mksc.1070.0319>

Kumar, V., & Pansari, A. (2016). National culture, economy, and customer lifetime value: Assessing the relative impact of the drivers of customer lifetime value for a global retailer. *Journal of International Marketing*, *24*(1), 1–21. <https://doi.org/10.1509/jim.15.0112>

Zhang, H., Liang, X., & Wang, S. (2016). Customer value anticipation, product innovativeness, and customer lifetime value: The moderating role of advertising strategy. *Journal of Business Research*, *69*(9), 3725–3730. <https://doi.org/10.1016/j.jbusres.2015.09.018>

Berger, P. D., Eechambadi, N., George, M., Lehmann, D. R., Rizley, R., & Venkatesan, R. (2006). From customer lifetime value to shareholder value: Theory, empirical evidence, and issues for future research. In *Journal of Service Research* (Vol. 9, Issue 2, pp. 156–167). <https://doi.org/10.1177/1094670506293569>

Predictive Analytics World Survey Report, Feb 2009, Available at [http://www.predictiveanalyticsworld.com/Predictive-Analytics-World -Survey-Report-Feb-2009.pdf](http://www.predictiveanalyticsworld.com/Predictive-Analytics-World%20-Survey-Report-Feb-2009.pdf)

Berger, P. D., & Nasr, N. I. (1998). *CUSTOMER LIFETIME VALUE: MARKETING MODELS AND APPLICATIONS*.

Méndez-Suárez, M., & Crespo-Tejero, N. (2021). Why do banks retain unprofitable customers? A customer lifetime value real options approach. *Journal of Business Research*, *122*, 621–626. <https://doi.org/10.1016/j.jbusres.2020.10.008>

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. Informatics, 5(1). <https://doi.org/10.3390/informatics5010002>