REVA UNIVERSITY
Bengaluru, India

A Project Report on

Prediction of Customer Lifetime Value in E-Commerce Business

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration

In Business Analytics

Submitted By

Mahapara Gayasuddin

R19DM004

Under the Guidance of

Krishna Kumar Tiwari

General Manager ML

Platform, Jio

REVA Academy for Corporate Excellence - RACE

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

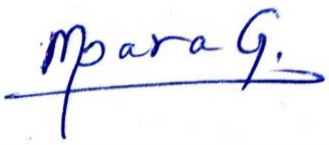
race.reva.edu.in

August, 2022

REVA UNIVERSITY
Bengaluru, India

Candidate's Declaration

I, **Mahapara Gayasuddin** hereby declare that I have completed the project work towards the second year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Prediction of Customer Lifetime Value in E-Commerce Business** under the supervision of Mr. Krishna Tiwari. This report embodies the original work done by me in partial fulfilment of the requirements for the award of the degree for the academic year **2022.**



Place: Bengaluru
Date:27-08-2022

MAHAPARA GAYASUDDIN
Signature of Student

REVA UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the project work entitled **Prediction of Customer Lifetime Value in E-Commerce Business** carried out by **Mahapara Gayasuddin** with **R19DM004** is a bonafide student of REVA University, is submitting the second year project report in fulfilment for the award of **MBA** in Business Analytics during the academic year 2022.The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said degree.

Signature of the Guide
Mr. Krishna Kumar Tiwari

Signature of the Director
Dr. Shinu Abhi

External Viva
Names of the Examiners

1. Dr. Sai Hareesh, Research Expert, SAP Labs India
2. Pradeepta Mishra, Director – AI, L&T InfoTech

Place: Bengaluru
Date: 27-08-2022

REVA UNIVERSITY
Bengaluru, India

Acknowledgement

During this hard time even when it wasn't possible to connect physically, we all have come together through the virtual platform with the support of Dr. Shinu and under the guidance of my mentor Mr. Krishna Kumar Tiwari Sir, with the help of my classmates and Reva technical staff, I have done my project successfully. So, I would like to thank each and everyone for their support.

I would like to thank Hon'ble Chancellor, Hon'ble Chancellor, Dr. P Shayma Raju, Vice Chancellor, Dr. M. Dhananjaya and Registrar, Dr. N Ramesh for providing us with great Infrastructure and quality education at Reva University.

Place: Bengaluru
Date: 27-08-2022

REVA UNIVERSITY
Bengaluru, India

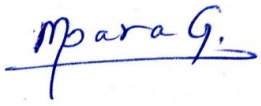
Similarity Index Report

This is to certify that this project report titled **Prediction of Customer Lifetime Value in E-Commerce Business** was scanned for similarity detection. Process and outcome is given below.

Software Used: Turnitin
Date of Report Generation: 26-08-2022
Similarity Index in %: 6%
Total word count: 11786
Name of the Guide: Mr. Krishna Kumar Tiwari

Place: Bengaluru

Name of the Student: Mahapara Gayasuddin



Date: 27-08-2022

Signature of Student

Verified by: M N Dincy Dechamma

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	CLV	Customer Lifetime Value
2	BG/NBD	Beta Geometric / Negative Binomial Distribution Model
3	Pareto/NBD	Pareto/ Negative Binomial Distribution Model
4	MBG/NBD	Modified Beta Geometric/ Negative Binomial Distribution Model
5	RFM	Recency,Frequency,Monetary
6	MBA	Market Basket Analysis
7	RS	Recommender System
8	CPA	Cost Per Acquisition

List of Figures

List of Tables

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).

To give this information to the marketing team for the campaign & Cost Per Acquisition (CPA) optimization, it is required to forecast the customer's lifetime value and segment the customers depending on their Life Time Value. Recognizing, segmenting, and ranking consumers is one of the key problems in customer-oriented firms.

The goal of the project is to look into different methods for calculating prospective income produced by a certain set of active consumers in the framework of non-contractual, ongoing business. “The probabilistic models (Pareto-NBD, BG-NBD, MBG-NBD, & Gamma-Gamma) have been applied to make this estimation”. Unsupervised machine learning was also used to undertake customer segmentation to demonstrate an effective tool for strategy development and improve customer-centric marketing.

To further get more Insights the market basket analysis is done using association analysis and a collaborative filtering recommender system based on items is developed. The co-occurrence matrix technique was also used to identify products that were frequently purchased in tandem. In the project, several aspects have been looked at that enabled to conclude the store. 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. A Dashboard has been prepared using Google Data Studio to Predict the CLV and to provide a better understanding to the marketing team and the CLV prediction model has been deployed in *Streamlit.*

Keywords: ***Probabilistic Models, BG/NBD, Pareto/NBD, MBG/NBD, CLV Modelling techniques, Customer Segmentation, Market Basket Analysis.***

Contents

Chapter 1: Introduction

One of the most essential e-commerce Key Performing Indicators (KPI) is customer lifetime value. They have disregarded CLV, one of the most important indicators for any business. Only 34% of the marketers assessed in a UK study were found to be "fully aware of the phrase and its meanings." Additionally, only 24% of respondents said their business successfully monitored CLV (Palalı, 2021). Because the cost of gaining clients is more than anticipated, many start-ups also struggle to make a name for themselves in the marketplace. When compared to new consumers, existing customers spend 67% more. Therefore, promoting repeat business is a wise approach to include in your marketing mix (SaasOptics, 2022).

> "CLV is the net profit a repeat customer generates for your company over their lifetime. Companies calculate the cumulative value and frequency of all a customer's transactions to predict CLV. Companies can predict and improve their long-term financial health by shifting their focus from short-term value to lifetime value" (*Surveymonkey*, 2022).

"According to the Harvard Business Review, acquiring a new customer costs seven times more than retaining an existing one. Companies that increase their customer retention rate by 5% see an increase in profit of 25% to 95%. In the long run, a business strategy that attracts a small number of repeat customers outperforms one that attracts a large number of one-time customers" (*Surveymonkey*, 2022).

1.1. Advantages of CLV:

1. The financial impact of marketing campaigns, initiatives, and other activities can be evaluated using the metric known as CLV.
2. As a result, your company will be able to align and ladder up to larger financial targets within the organization—or, if you're a smaller operation, start developing them.
3. By enabling you to create loyalty targets and devote resources to underserved areas, CLV can also alter the way you think about marketing.

4. CLV will assist you in finding the right balance between short- and long-term marketing objectives and show that you have a better grasp of financial return on investment (Library, 2022).

1.2. Importance of CLV:

One of the key purposes for monitoring CLV is customer retention. The likelihood of selling to a new prospective client is 5%-20%, while the likelihood of selling to an existing customer is 60%-70%, according to Marketing Metrics (Karolina Matuszewska, 2021). Therefore, increasing sales to loyal clients will provide much bigger earnings. This highlights the need of encouraging client loyalty as well. Regular customers tend to spend more money on your products, which helps you grow and promote your company. 81% of marketers feel that tracking CLV helps sales, according to a Criteo report (Karolina Matuszewska, 2021).

“CLV is essential to the existence of financially stable e-commerce businesses that can grow naturally and sustainably. This is so because CLV is an investment that pays off in the long run with a greater ROI and better unit economics. It's a fundamentally different strategy than focusing on immediate sales. The problem is that acquisition-based growth demands continual marketing expenditures, and you can only grow to the extent that you can afford to spend (for example, consider Facebook adverts and Google AdWords” (Strategy, 2022).

"CLV is without a doubt the most comprehensive metric for marketing analytics. Future income is affected by several factors, including the number of new customers, the cost per order, customer retention rates, conversion rates, and more, but CLV collects all pertinent information for each client. Simply said, it's the estimated profit you'll make from each consumer in your company. With the right calculations, this measure can help you develop your e-commerce firm quickly. You won't lose money since you'll know exactly how much money you're making " (Library, 2022).

1.3. Maximization of CLTV:

The most important variables that you can control are the proportion of repeat purchases, the average transaction value, and the average customer retention rate. So, the following are some methods to increase client lifetime value:

Surveys

By creating efficient strategies for boosting customer loyalty, long-term relationships can be built. Due to their need for and familiarity with a given brand, consumers are more likely to continue with such brands.

A customer feedback survey is an innovative strategy to interact with customers and foster loyalty. Experts claim that website-based surveys will unquestionably assist us in gaining insight into the perspective of a customer when evaluating any goods or service as well as a clear understanding of their intentions and expectations or even the modifications that must be made to each specific product. For instance, as shown in Figure No 1.1:

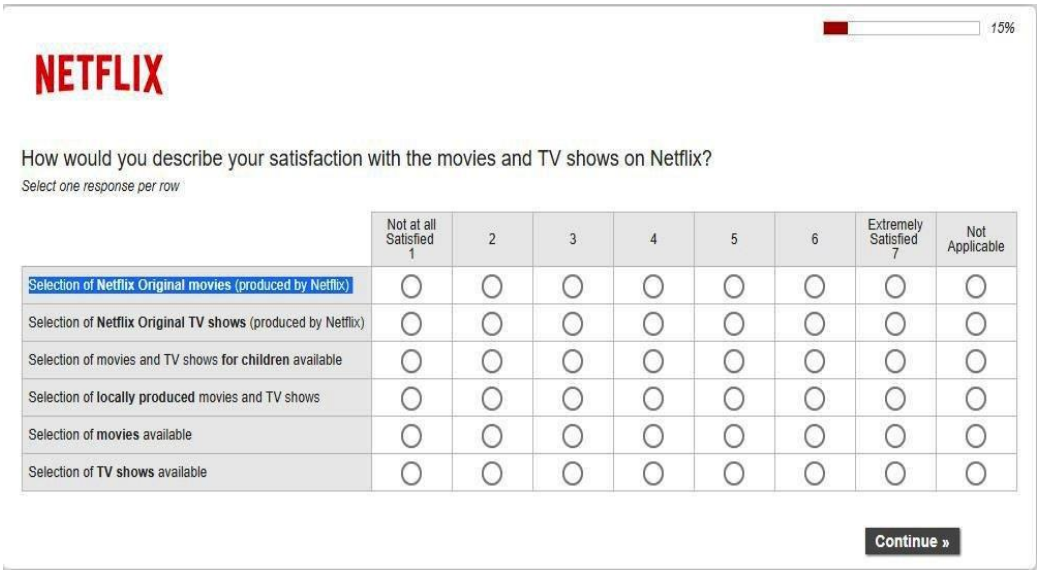


Figure No. 1.1: Netflix Customer Satisfaction Survey

Staying Relevant

As shown in Figure No. 1.2 staying at the forefront of your customer's minds is another successful tactic for increasing the number of repeated business you receive. If your brand is tagged as relevant and is at the top of the suggestion list, clients will undoubtedly consider it when they have a need.

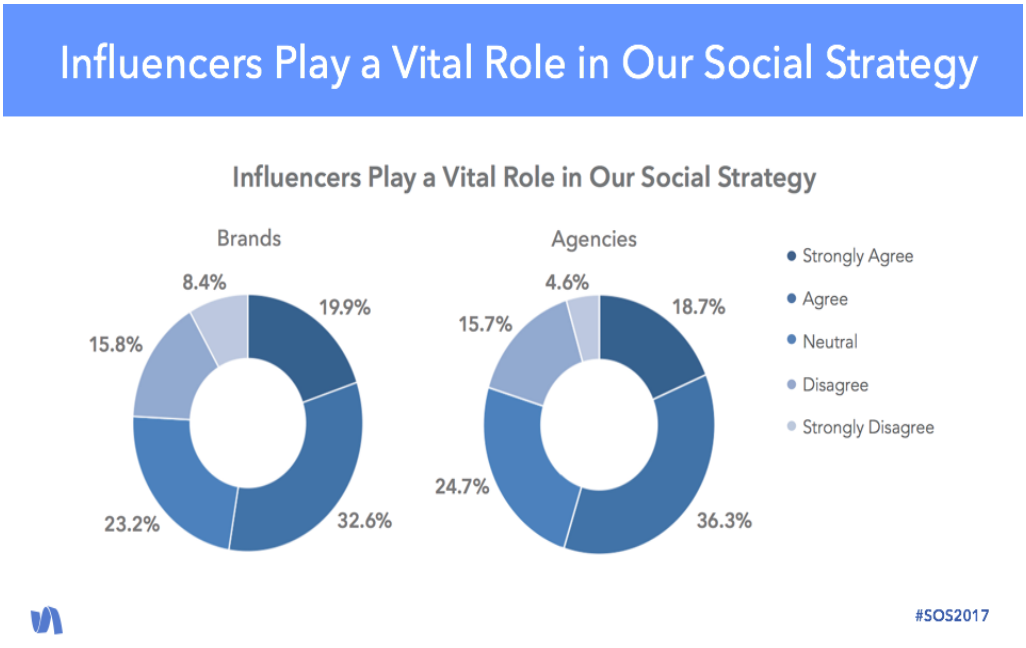


Figure No. 1.2: Role of Influencers in Social Strategy

Cross-Selling and Up-Selling

A great approach to increase sales and customer loyalty is to fill the cart to the brim with fictional and irrelevant goods. Additionally, the rate of basket abandonment suffers. In this sense, Amazon has demonstrated exceptional foresight by providing superior shipping services, unique pricing, and access to a sizable movie collection via its Prime membership. In essence, by providing their clients with reduced extras, they maintain their brand's awareness as shown in Figure No. 1.3.

Figure No. 1.3: Up-Selling and Cross-Selling Example

Loyalty Programs

Loyalty programs can increase customers' lifetime values. This tactic compensates current customers who do business with the business for a longer period of time. The process of determining your most valuable clients was ongoing. Each of these tactics will assist you in raising both your long-term success rate and the lifetime value of your clients. Perhaps the most beneficial aspects of your business are outstanding buyer feedback or excellent customer service. Some businesses may get the most bang for their buck by launching a profitable loyalty program. What works best for you will completely determine the final plan. Continue to experiment and keep track until you discover your optimal CLTV recipe, proceed.

An outstanding illustration of a successful e-Commerce program is Amazon Prime. Compared to non-Prime members, who spend $400–$500 a year on Amazon, an average Prime member spends around $1.4K annually (Jawad Khan, 2021).

Provide A Special Customer Experience

By consistently enhancing the customer experience your company provides, you may raise your client's lifetime value in one of the quickest possible methods as shown in Figure No. 1.4. 43% of consumers are willing to spend more for a product or service if it provides a better experience, according to a PWC survey. Customer experience is cited by 75% of US consumers

and 65% of UK consumers as being important when making a purchase decision (Jawad Khan, 2021).

Figure No. 1.4: Customers Experience helps people decide between buying options

Customer Experience in Online Retail:

It is the general opinion a consumer has of a business after doing business with it. It covers every aspect, such as website UX, payment convenience, and product quality. Here are the aspects of customer experience that customers appreciate most, per the same PWC study that that previously stated as shown in Figure No.1.4 (Jawad Khan, 2021).

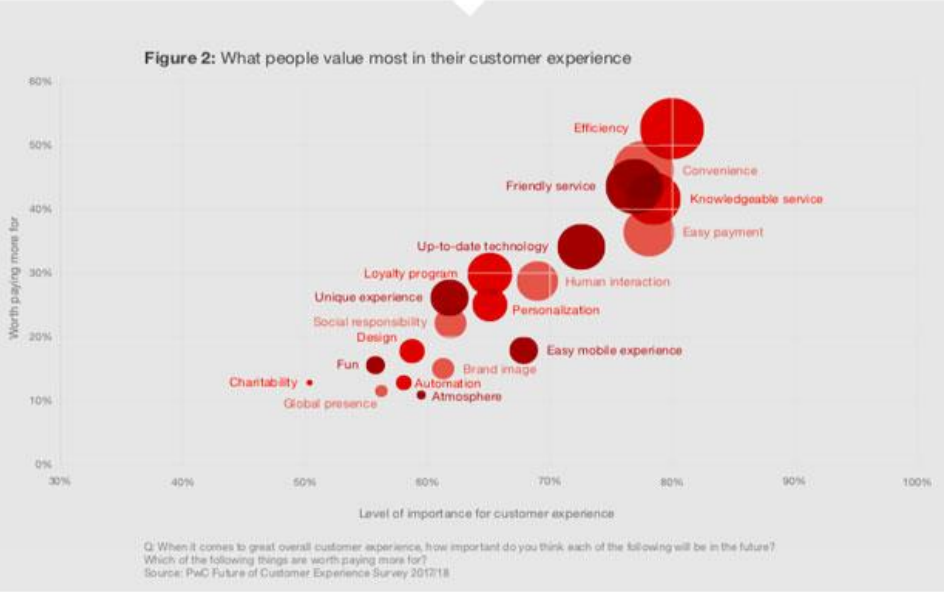


Figure No. 1.5: What people value the most in their customer experience

To deliver a memorable customer experience that immediately increases your customer's lifetime value, e-commerce business owners can enhance their website's product navigation, payment process, customer care, return handling process, and overall customer interactions as shown in Figure No. 1.5.

By providing individualised product recommendations and promoting similar products during the checkout process, you may also raise the average order value of your clients (Jawad Khan, 2021).

Chapter 2: Literature Review

The purpose of this study is to use big data analytics, to investigate the behaviour of bank clients through value segmentation as Customers' demands and expectations shift frequently, causing customer lifetime value to fluctuate. Client segment dynamics can be used as a predictor of customer lifetime value (Mosaddegh et al., 2021)**.** This study uses the K-means clustering approach to examine Customer Lifetime Value (CLV) and cluster it into customer segments (Najib et al., 2019).

The purpose of this study was to use association rules to create a market basket analysis. As a result, these rules can be used to place products in the supermarket. As a result, sales of these products will rise, and supermarket revenue will rise as well (Ünvan, 2021)**.**

This proposes a framework for consumer segmentation that encodes each customer's behaviour as a time-series sequence of the Recency, frequency, and monetary variables, and then uses time-series clustering methods to do so. Customer segmentation is a frequently used analytical technique for identifying separate customer segments (Abbasimehr & Bahrini, 2022)**.**

This discusses a statistical examination of the prediction capacities of various customer lifetime value (CLV) models that could be employed in online shopping in e-commerce company settings (Jasek et al., 2018)**.**

The purpose of this article is to examine the utility of fundamental approaches for calculating client lifetime value in businesses. The fundamental economic component influencing the formation and maintenance of long-term lucrative client relationships is customer lifetime value. It's also important in making judgments about acquiring new customers and keeping existing ones. As a result, it has an impact on the company's capacity to continue operating (Lew, 2017)**.**

Recommender systems, often known as recommendation systems (RSs), are software tools and approaches that provide users with recommendations based on their requirements. This document seeks to describe the limitations and benefits of various recommendation algorithms (Shah et al., 2018)**.**

By identifying connections between various things in customers' shopping baskets, this approach looks into their purchasing behaviours. By acquiring information on which things are commonly purchased by customers, merchants can improve their marketing campaigns by identifying such relationships. This Provides a Survey of market basket analysis data mining algorithms (Enabled et al., 2019).

One of the most extensively used and successful technologies in recommender systems is collaborative filtering recommendation. This paper summarises the recommender system and traditional recommendation algorithm, then moves on to the collaborative filtering recommendation algorithm's associated work (Zhao et al., 2021)**.**

Collaborative filtering is a technique that makes predictions based on the relationships between people and items. The purpose of the recommendation system is to create a score function that combines the results of computing user and item similarity. This evaluates many evaluation indicators that represent the recommender system's efficacy (Shakirova, 2017)**.**

This article is a review that examines existing research on CLV and covers studies that are related to it that have already been published by others. It demonstrates the value of customers to a business throughout the business. Additionally, CLV is a statistic for determining whether or not a group of consumers has sufficient value to be acquired, retained, and/or continued in the connection. CLV continued three key components cutomer acquisition, customer retention, and customer development—are examined through this study (Feiz et al., 2016).

The “Buy-Til-You-Die (BTYD) models”, a class of probability models, are examined in this work, along with a proposal for employing variable selection techniques to implement these methods to enormous sets of data. The BTYD models' theory is presented, and models that use frameworks to include regression aspects classes are created (Dimaano & Fader, 2018).

The Pareto/NBD model, which is frequently employed in customer base analysis, will be extended and examined in this work. The Pareto/NBD will be illustrated and estimated in this study using a "recency-only" approach. The "recency/frequency" model will be contrasted with the model after it has been fitted to both real-world and simulated data (Scholars & Rajagopalan, 2018).

The authors of this study draw the following conclusions: when Apriori algorithms are used, information can be obtained in the form of a combination of consumer purchase patterns based on consumer transaction data, and sales estimates can be made using the information obtained from the analysis of information systems in identifying consumer purchasing patterns. (Series, 2020).

Aslekar Avinash and his coauthors think that using historical customer lifetime value, which ignores time, is the most straightforward RFM method to determine CLV. This strategy uses historical transactions to identify clients who have exhibited the same behavioural pattern over a similar time period. Therefore, the CLV projection period is when the most recent transaction was made. However, it is of no use in anticipating the customers' upcoming actions. To identify client habits and therefore anticipate their future behavior is the idea behind CLV prediction utilising RFM models (Avinash et al., 2019).

In this study, a fashion retailer's short-term shop sales are predicted using ensemble machine learning techniques. With the aid of the “bagging tree regressor, random forest regressor, and gradient boosting regressor algorithms”, sales projections for a range of goods at various retailers are created over a period of three months. With the help of sales data from a Turkish fashion retailer, algorithms are trained and evaluated (Olkhov, 2019).

Chapter 3: Problem Statement

In a normal non-contractual situation, B2C businesses struggle to identify and forecast client buying trends. As a result, they find it difficult to comprehend the true value of their clients and end up attempting to solve the following problems:

1. How much should they invest in bringing in fresh business to sell their product?

2. How do various client groups interact with a product?

3. How successful is their product in the real world?

4. How can they increase demand for their goods while maximizing profits from each client?

The company selected for this project is one of many retail stores that put in a lot of effort to keep the UK's retail industry running smoothly. It is a business that started in London, England, and now has stores all over the world. In the UK, it manages close to 100 retail locations. One of the businesses included in the FTSE 250 Index. The business generates 2,277 million in sales annually (GBP). It has a customer base both inside as well as the outside UK. The market serving online retail has experienced intense competition over the past year, and its sales revenues have been trending somewhat lower.

To solve this issue, the business is currently intending to modify its marketing approach. According to their customer's CLTV values, they aim to project the future behavior of its users, who will bring in the most money over the course of a month using the statistics at hand. Also, to determine the association between various products in the basket by analyzing the customer purchase pattern of multiple items and recommending them the products. Furthermore, 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.

Chapter 4: Objectives of the Study

The importance of CLV in the Retail and E-commerce Industry has grown in recent years, in order to assess their commercial possibilities, it has recently become necessary for an organization to follow their customers and determine their lifetime worth. The study's objectives would be as follows:

1. Understand business requirements to formulate an effective customer analytics strategy for the Retail & E-commerce industry. Perform descriptive analysis & data visualization to derive intuitions about customer behavior and hypothesis from data.

2. To understand how a cohort performs over time and how it stacks up against other cohorts. Develop actionable customer segments & predictive models for activation, cross-sell and retention. Develop & implement customer lifetime value framework. Provide prescriptive analytics on derived segments and measure engagement success of various marketing strategies and finally deploy the model to predict the CLV.

3. To use the Market basket analysis method for locating products grouped together and subsequently identifying client buying habits. To provide a forecast or “Top-N recommendation” for the active user using Collaborative Filtering.

4. To predict item-wise sales of the retail store using machine-learning algorithms.

It is intended to provide managers with a framework for measuring their valuable intangible assets known as clients. The model has been developed after conducting an exploratory study on data availability, and then employing the most acceptable method to forecast the CLV for the dataset available.

Chapter 5: Project Methodology

For model development, I attempted to adhere to the CRISP-DM standard as shown in Figure No. 5.1.

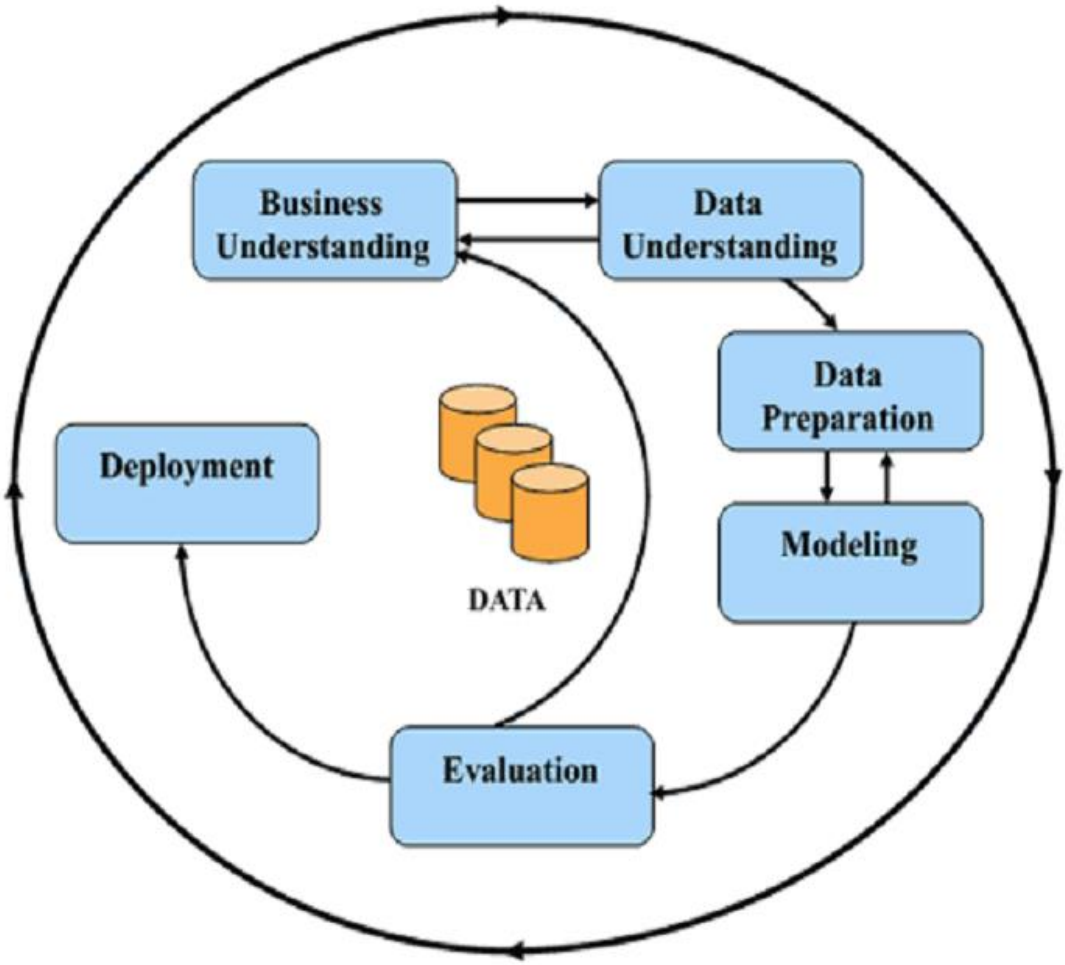


Figure No. 5.1: “The cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts” (*Prakhar Gurawa*, 2021).

As previously discussed in this section, the entire project was completed in the order of steps that is used. The diagram below only represents the framework of the entire project.

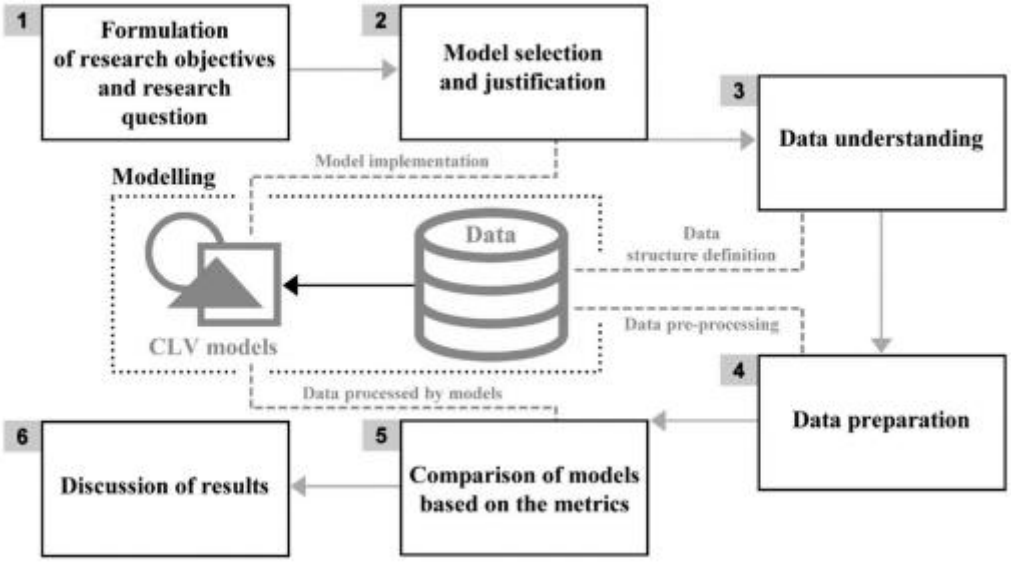


Figure No. 5.2: Methodology of the Project

Figure No. 5.2 shows the Methodology of the Project which is explained below:

Step 1: Identifying Suitable Dataset

After review, it was critical for me to select a dataset that met all of the requirements for using CLV modeling techniques. This was our main restriction while choosing the right dataset because some variables are crucial for forecasting the CLV.

Step 2: Pre-Processing of the Dataset

Because datasets may contain missing values, outliers, and duplicate entries, which will surely affect the predictions, it is crucial to utilize the appropriate pre-processing procedures. Additionally, the proper pre-processing methods has been applied before using it for prediction.

Step 3: Deciding on Suitable Modelling Techniques

Following pre-processing, it was necessary to choose the modelling approaches that would yield the best results for this particular dataset. After careful research and study, it was chosen to use probabilistic modelling technique as well as other ML modelling technique.

Step 4: Evaluation of the chosen Models

The dataset was then applied to the selected models to calculate the customer lifetime value, and the models were assessed based on a range of indicators.

Step 5: Results and Findings

The outcomes for these models were recorded, examined, and analyzed in terms of their capacity to more precisely and completely forecast CLV.

Step 6: Conclusion and Future work

The work has been finished by talking about the findings and how they were interpreted, as well as by identifying potential insights that could aid the marketing teams in achieving their objectives.

Chapter 6: Business Understanding

CLV is a key metric for gaining a better understanding of the customers as shown in Figure No. 6.1. It's a projection of the potential value your customer relationship will have for your business. Organizations can use this strategy to show the potential worth of their marketing initiatives in the future. Focusing on CLV enables the development of an effective strategy with clear budgeting. Some clients, however, are more important to the business than others. Knowing which ones to focus and invest in initially is crucial because of this (Karolina Matuszewska, 2021).

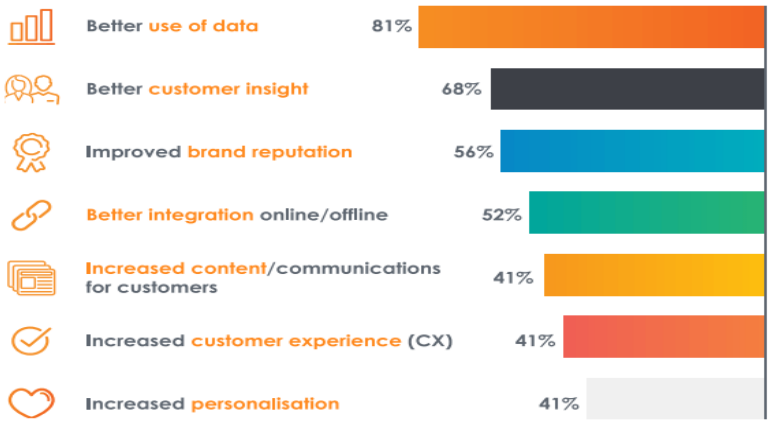


Figure No. 6.1: What, according to a Criteo survey, boosts CLV in business?

It is also crucial in decisions about acquiring new customers and retention of the current ones. To calculate the customer's cumulative profitability, it is also necessary to estimate the time of collaboration with him, which introduces some subjectivity into the estimation of CLV (Lew, 2017).

Marketing relies heavily on segmentation, and include client profitability could enhance the efficiency of marketing initiatives. Additionally, firms are under pressure to implement marketing plans that are specific to each customer. (Ronan Martin, 2019).

“Customers' expectations are changing; thus businesses must comprehend the current customer requirements. Using CLV segmentation, clients are divided into groups according to the elements that affect their purchasing decisions”. (Ronan Martin, 2019).

“Lack of understanding about which consumer groups to target and how to interact with each one is one of the most important limitations and weaknesses of sales and trade marketing departments in terms of sales development in the E-Commerce business. In all industries, customer segmentation using the RFM approach and CLV would be helpful for sales, trade marketing, and marketing decisions, especially for enterprises involved in the E-Commerce sector.” (Mohammadian & Makhani, 2019).

Let's go over this in greater depth. As previously stated, Customer Lifetime Value offers numerous benefits. Here is a handful of them:

1. “Finding those Customers is essential because the Pareto principle states that 20% of a company's current customers generate 80% of its future income”. This also contributes to a company's long-term goal of achieving loyalty.

2. Customer Satisfaction Analysis can be performed by examining changes in their CLV Scores. This will increase investment in automating internal processes, offering coupons, and so on.

3. Recognize early warning signs of customer defection and identify the customer segments responsible for the problem.

4. It informs businesses about the estimated time it will take to recoup the Customer Acquisition Cost (CAC) and helps them decide whether or not to proceed with the investment.

Let's talk about the business outcomes and the areas where CLTV will be useful.

1. Because the percentage of customers that were focused on retaining is lower, analysis can be performed and additional schemes can be developed aiming at increasing customer retention.

2. Specific segments can be targeted and strategize accordingly. This aids in the creation of targeted campaigns rather than the traditional blanket campaign approach.

Chapter 7: Data Understanding

The following information is examined and combined to produce sales data over a 24-month period:

Customer data

Invoice data

Customer transaction data

Product purchased by the customers and the respective Quantity and amount

This transactional data set includes every transaction made by a UK-based, registered online retail store between December 1, 2019, and December 9, 2021. The company's main product line is unique gifts for all occasions. The company has a large number of wholesalers as clients. The dataset contains transaction-level data of customers with 1067371 rows and 8 columns as shown in Table No. 7.1 **("*UCI Machine Learning Repository*, n.d.").**

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 No. 7.1: Attributes of the Dataset

Exploratory Data Analysis (EDA)

The first stage is understanding the primary metrics that the business wants to follow. These metrics rely on the company's product, position, targets, and more. The majority of companies now monitor their key performance indicators (KPIs). The primary KPIs in this illustration could be those that pertain to revenue, such as monthly revenue, monthly active customers, the

average order value, order frequency, new customer vs. existing customer revenue, and cohort analysis to gauge customer retention.

Worldwide Distributions

Customers of the retailer come from all around the world. On a globe map, the distributions will be examined as shown below in Figure No. 7.1 and Figure No. 7.2.

- Orders

- Customers

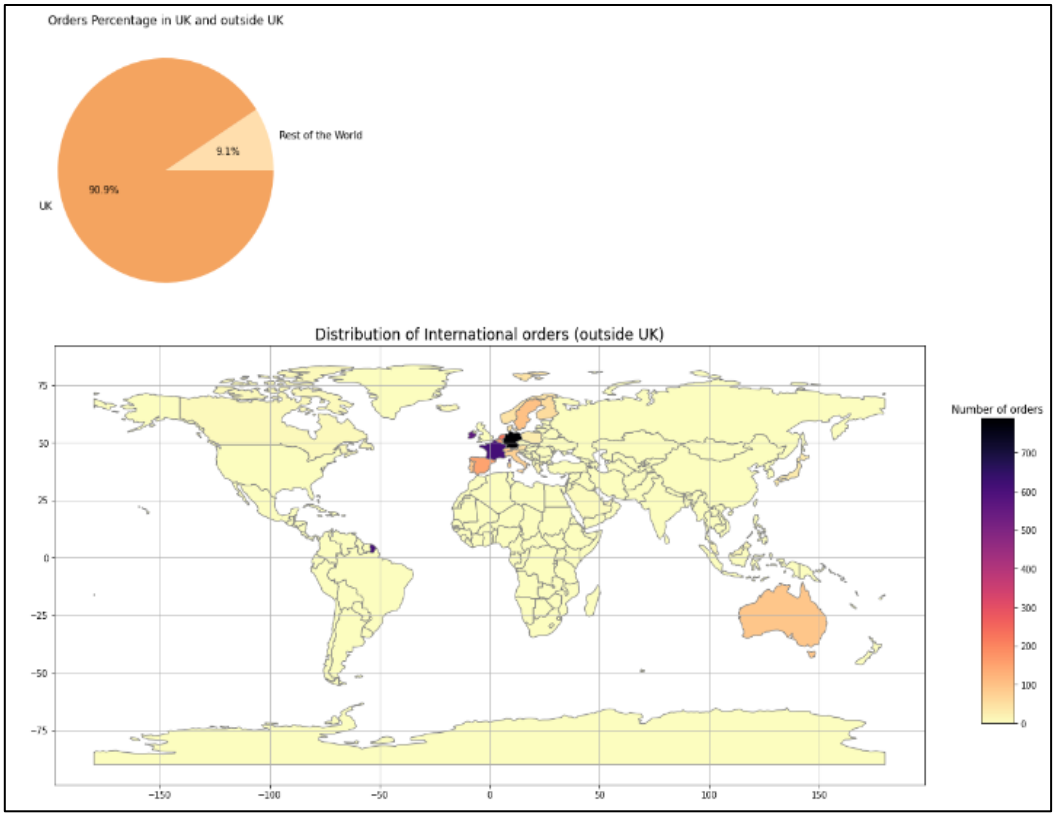


Figure No. 7.1: Orders Percentage in the UK and the Outside UK

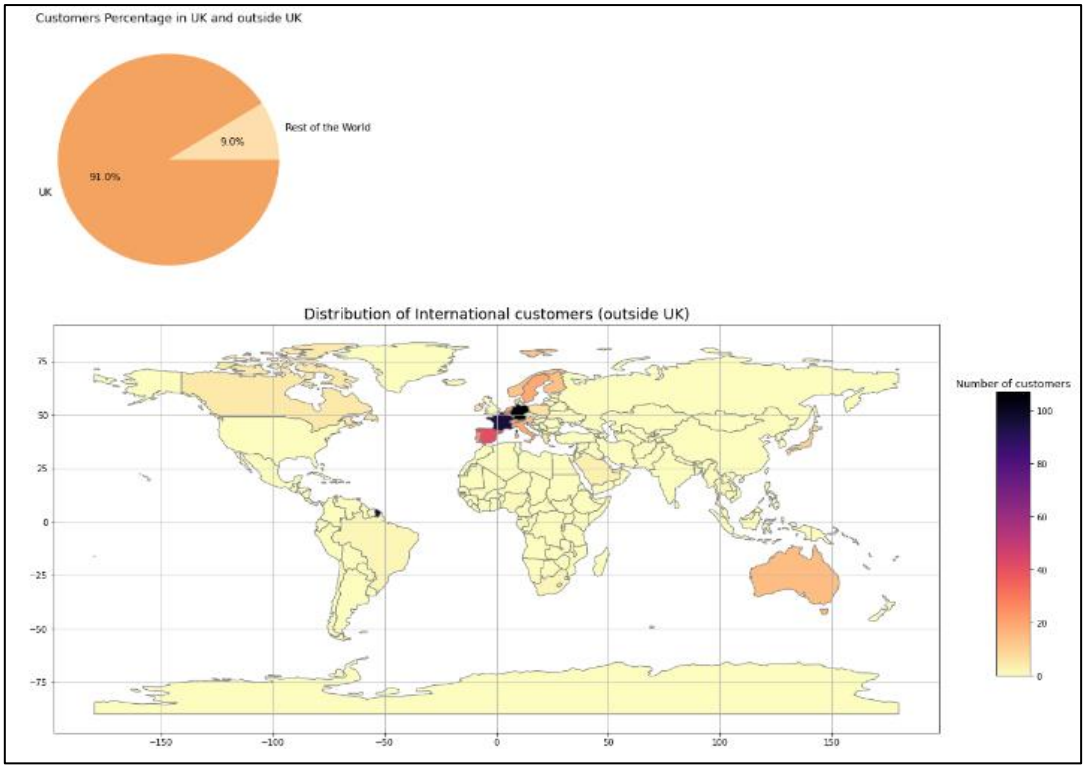


Figure No. 7.2: Customers Percentage in the UK and the Outside UK

Sales by Country

The region where the company makes the most money is the United Kingdom as shown in Figure No. 7.3. In order to conduct this research, it is required to concentrate on some more EDA.

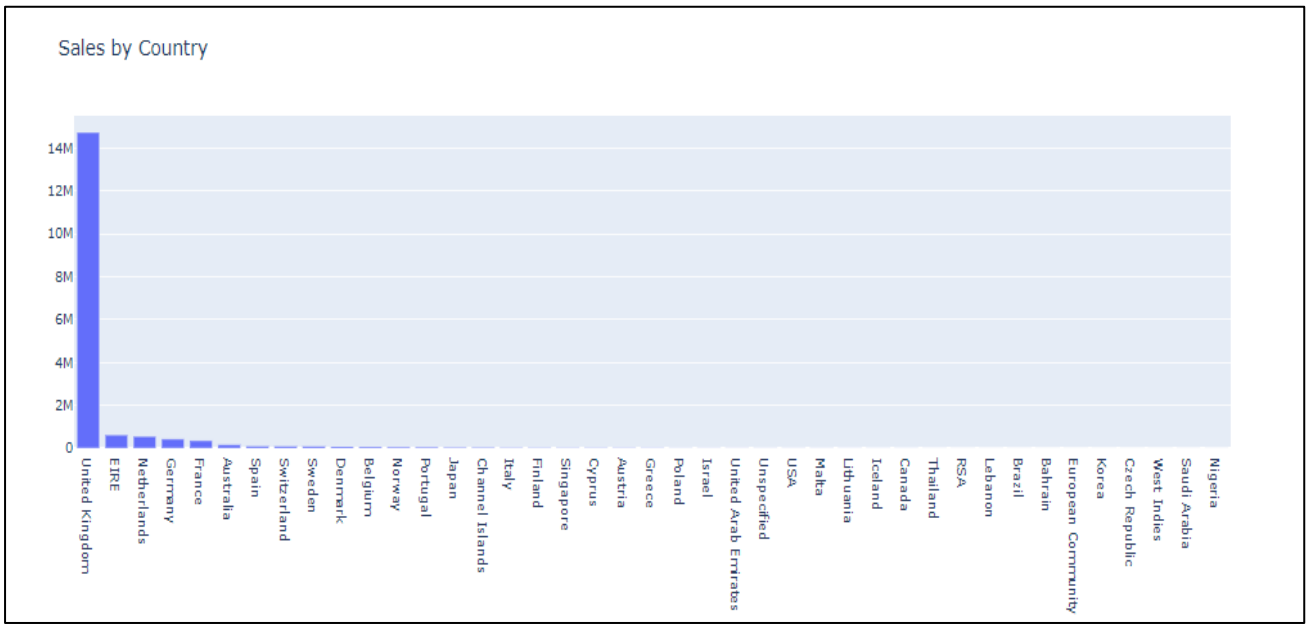


Figure No. 7.3: Sales by Country

Most Bought products in the UK

Figure No. 7.4 shows the most bought products in the country UK.

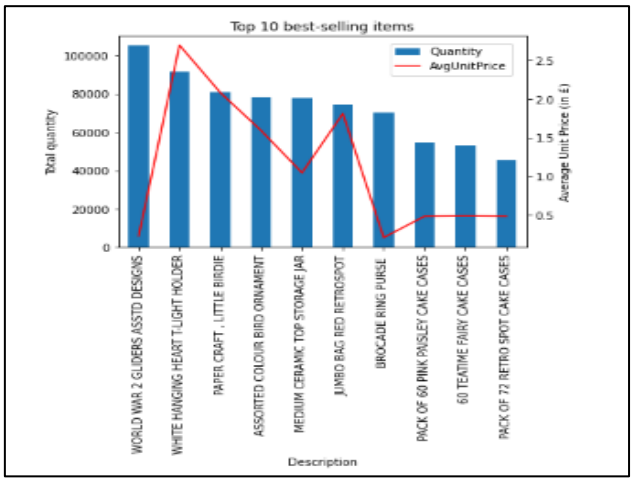


Figure No. 7.4: Most Bought products in the UK

Top Revenue Grossing Items

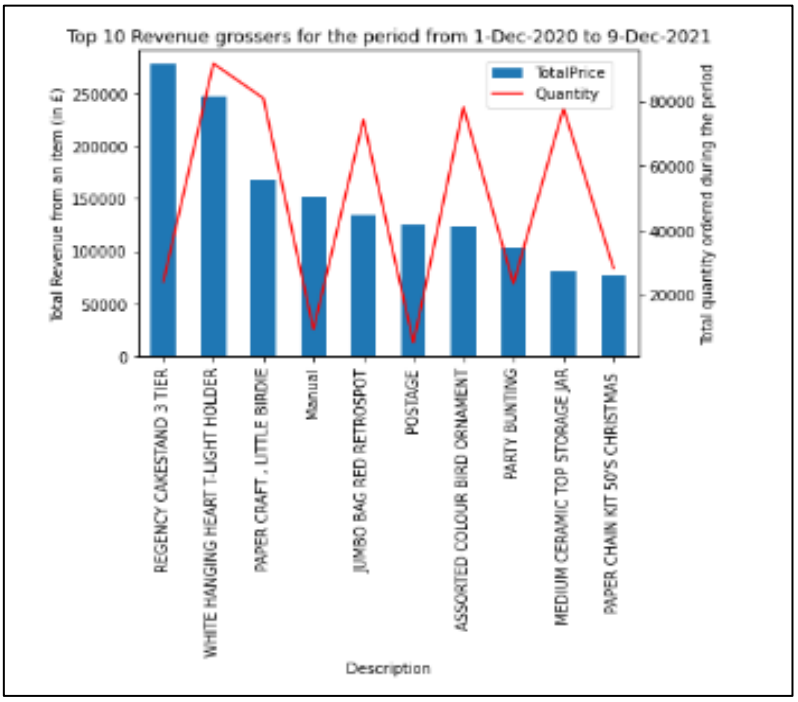


Figure No. 7.5: Top 10 Revenue Grosser Items

It can be concluded from the Figure No. 7.5 that,

1. We can see that 90% of customers the majority of the products are sold by the retailer in the UK, then several other European nations. The average unit price of the top-selling goods varies. Therefore, it doesn't seem to be related to the price.

2. The top-selling item i.e. Regency cakestand 3 Tier sells more than twice the Paper chain kit Christmas 50’s Christmas.

3. The top revenue earner i.e. Regency cakestand 3 Tier is almost 13% of its closest rival i.e. White Hanging Heart T-Light Holder.

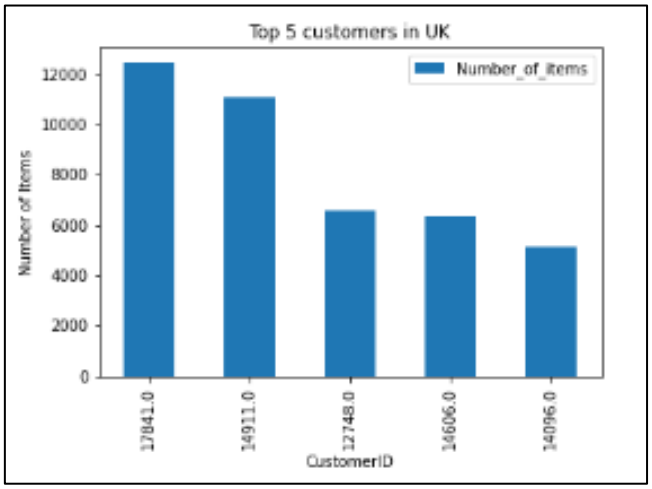


Figure No. 7.6: Top 5 Customers in the UK

From Figure No. 7.6, it can be seen that the Customer having CustomerID 17841 is the top Customer of the Retail store.

Date Time Analysis

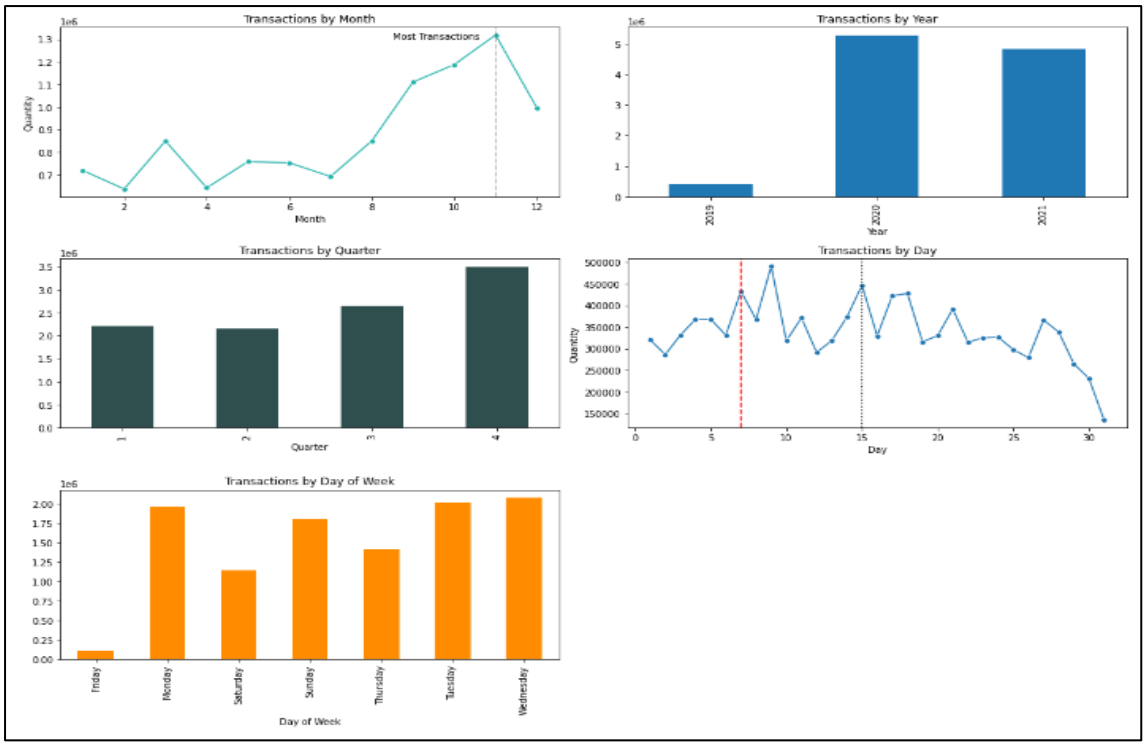


Figure No. 7.7: Date Time Analysis of the Sales

From Figure No. 7.7 it can be concluded that:

1. Due to the holiday seasons, November was the month with the most transactions.

2. 2020 and 2021 are the two years with the greatest trades, respectively.

3. The biggest number of transactions occurred in Q4.

4. It was also noted that consumers tend to make more purchases around the end of the first week and the beginning of the third week.

5. Wednesday is the day that people love to shop, followed by Tuesday and Thursday.

Monthly Sales

The graph in Figure No. 7.8 depicts an upward trend for the income earned through November 2021. (as the December data is incomplete). The company's monthly revenue was from

568.101K to 950.69K up until September 2021. Since then, it has had tremendous growth in sales, hitting 1.156M in November 2021.

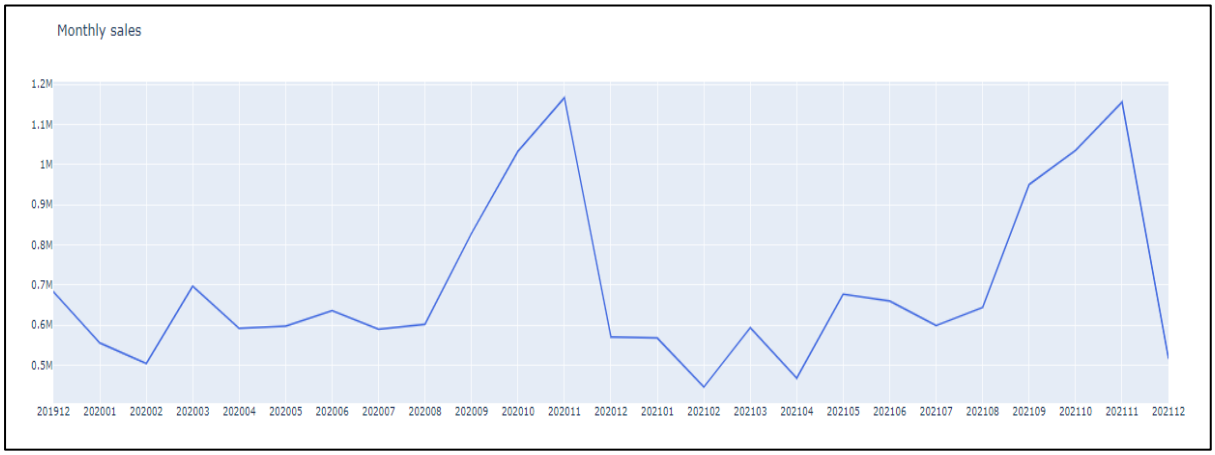


Figure No. 7.8: Monthly Sales

Monthly Growth Rate

Figure No. 7.9 shows that with almost 84.18% growth from the month before, September 2021 was an exceptional month. With a growth rate of 44.62%, May 2021 was likewise a very strong month. Both March and May 2021 saw increases of over more than 30%, but the underwhelming results of the preceding months may be to blame for 2021's January, February and April both had poor performances will less than -21%.

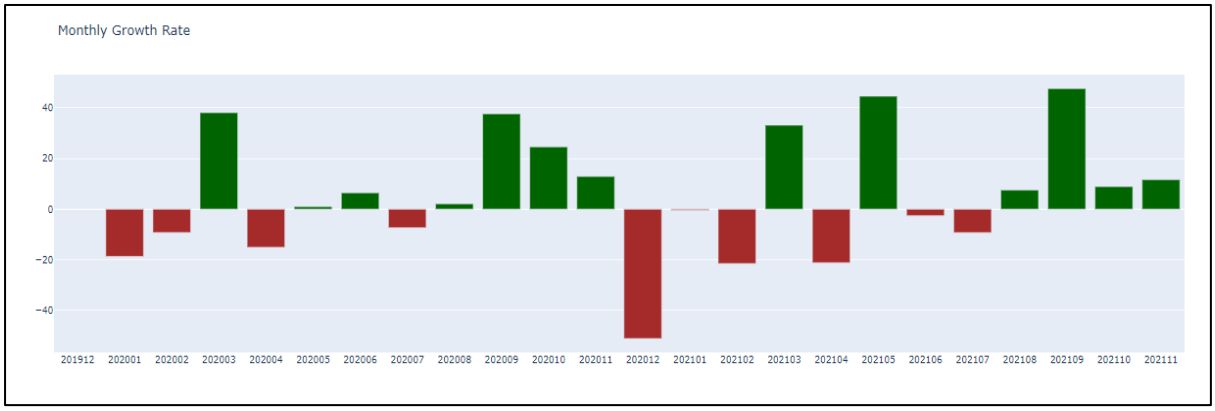


Figure No.7.9: Monthly Growth Rate

Monthly Active Customers

The company lost over 144 customers in January 2021, dropping from 885 in December 2010 to 741 in January 2021, a -18.27% fall. Similar to this, in April 2021 the company lost -12.11% of its customers, going from 974 to 856 as shown in Figure No. 7.10.

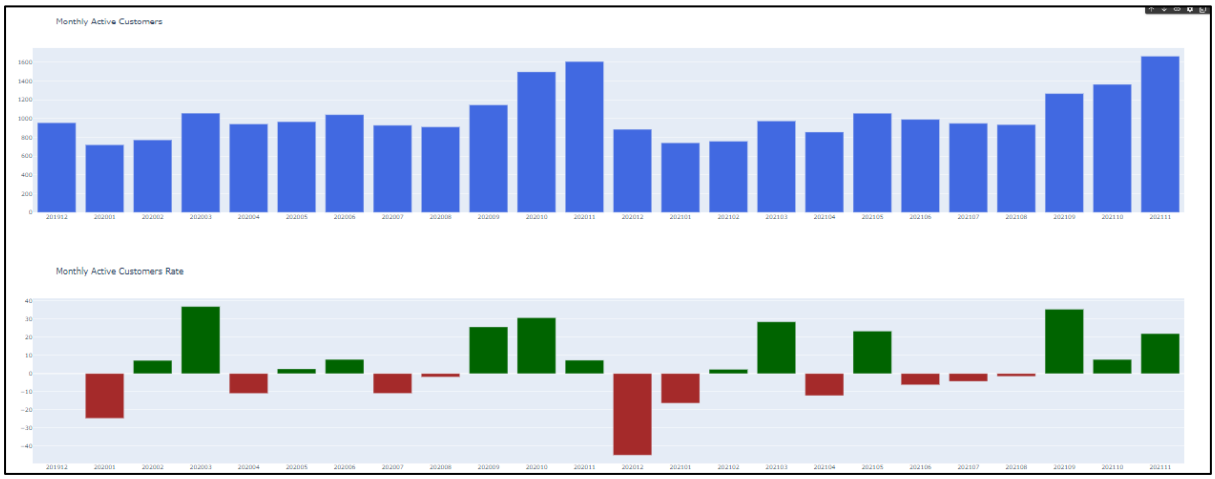


Figure No. 7.10: Monthly Active Customers

Monthly Order Count

Figure No. 7. 11 shows that Between December and January 2021, there were 413 fewer orders than there were at that time, a fall of -29.5%. Up until May 2021, orders increased by 35.34%. Orders decreased once more until August 2021 by -4.45 % before ultimately increasing by 37.22% in the month of September 2021 up until November 2021.

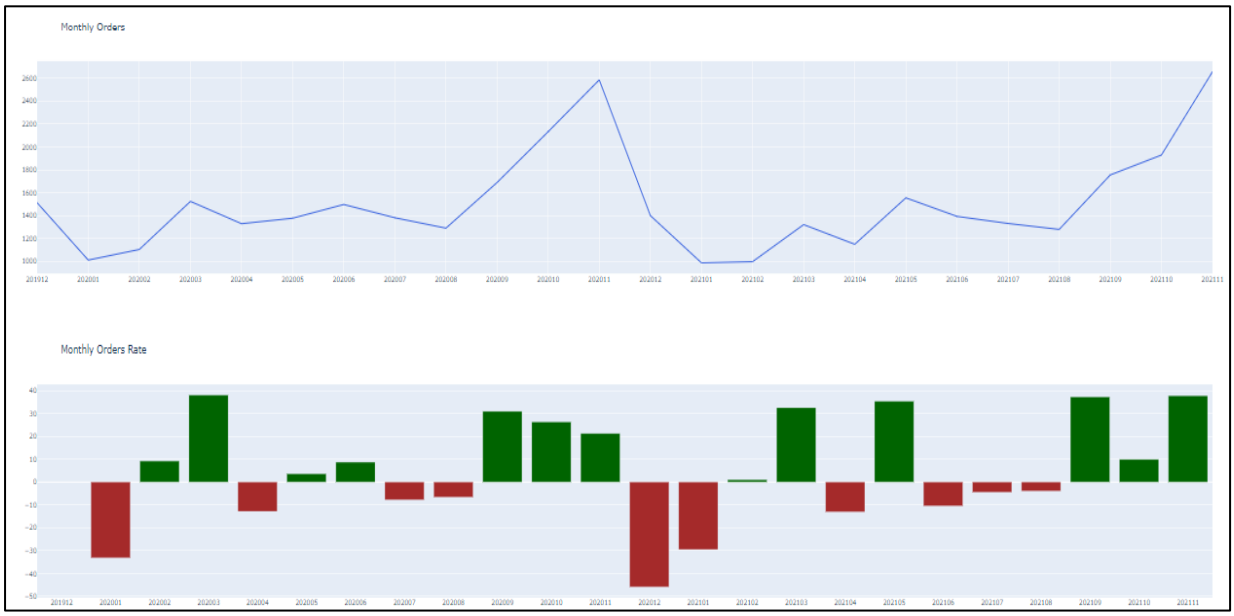


Figure No. 7.11: Monthly Order Count

Average Order Value

The company's average decreased by -21.16% from March to April 2021, but it then increased until September 2021, rising to 47.61% as shown in Figure No. 7.12.

Figure No. 7.12: Average Order Value

New Customer vs. Existing Customer Revenue

Figure No. 7.13 shows the revenue generated by the Customers declines over time. However, the existing customer base exhibits a promising tendency, suggesting that the company keeps the majority of its clients.

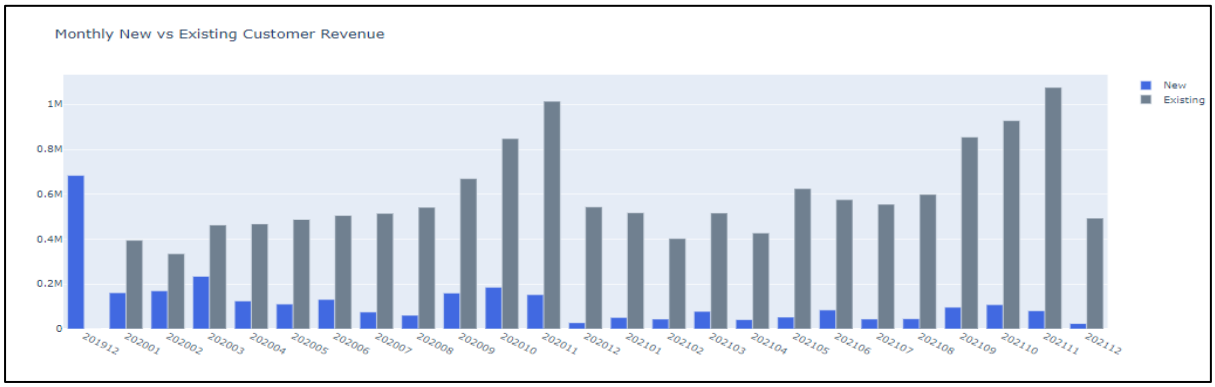


Figure No. 7.13: New Customer vs. Existing Customer Revenue

Monthly Retention Rate

The monthly retention rate is computed in order to do the analysis.
Retention rate is crucial since it shows how well your company is doing. A high retention rate indicates active and engaged users, which may indicate more lucrative opportunities.
The following could be learnt from the retention rate (AppLovin, 2021).

1. How likely it is that you can keep each new client you get?
2. How long each customer will be retained if you stick to your current tactics?
3. How much the business might expand in the future (AppLovin, 2021).

As can be seen from Figure No. 7.14 that with the greatest rate at 47% in January 2020 and the lowest at 3% in March 2020, the retention rate is generally favourable. In terms of customer retention, January, April, July and December are the best months.

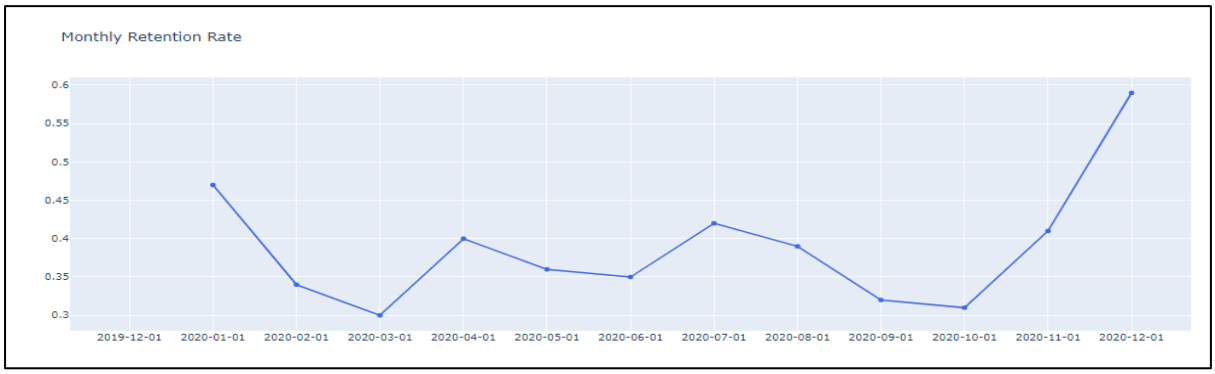


Figure No. 7.14: Monthly Retention Rate

Cohort Analysis

In this case, there is 25 cohorts total, with 25 cohort indices. Higher values correspond to blue colours that are lighter as shown in Figure No. 7.15.
As a result, the light blue shade with 34% in the 2021-07 cohort month of the 5th cohort index indicates that 35% of the cohorts that signed in July 2021 were still active at that time.
Likewise, it can be seen that just 32% of the group that signed up in October 2021 was still active a month later.

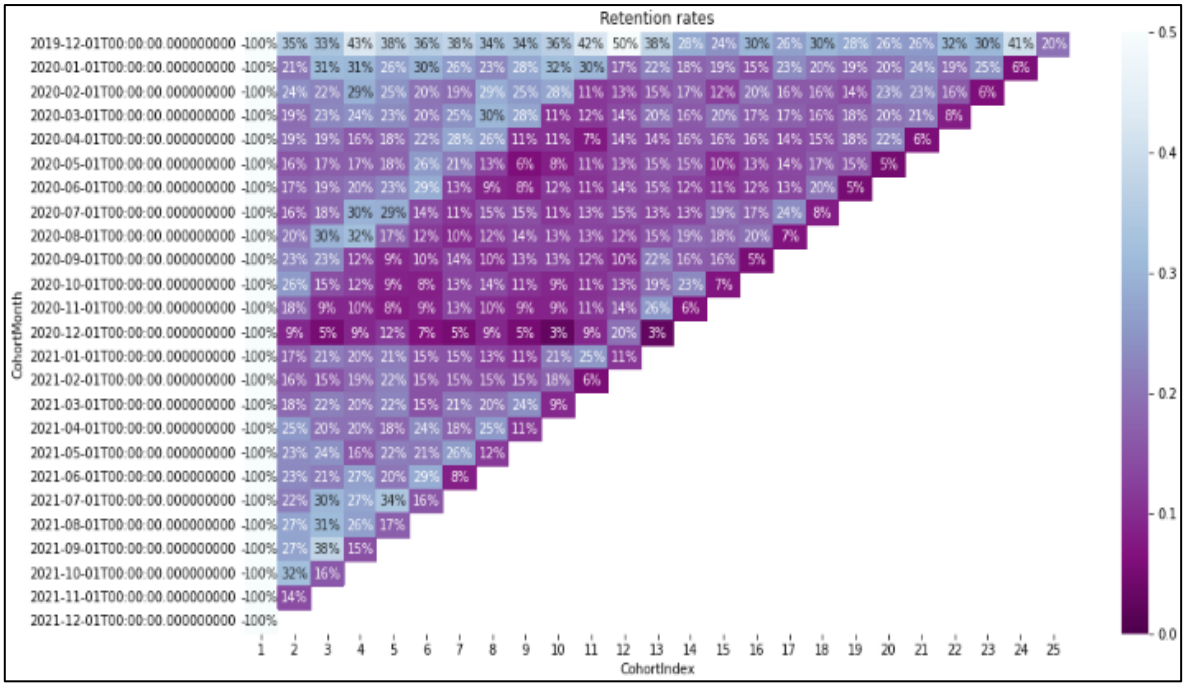


Figure No. 7.15: Retention by Cohort

Pareto Principal (80:20 Rule)

Our online shop has 5878 different Customers and 5283 different products for sale from December 1, 2020, until December 9, 2021. The Pareto principle is useful in this situation because with so many products and clients to concentrate on, the business can only concentrate on 20% of them.80 % of its sales must come from these.

The Pareto principle applies to the products in our dataset because just **22% of all the items account for 80% of the sales revenue.** 1162 items, or 22% of all goods, are included. 80 % of sales revenue is generated by just 23% of all customers. 1352 items, or 23% of all customers, are included. It can be observed how the Pareto principle applies to our dataset as shown in Figure No. 7.16.

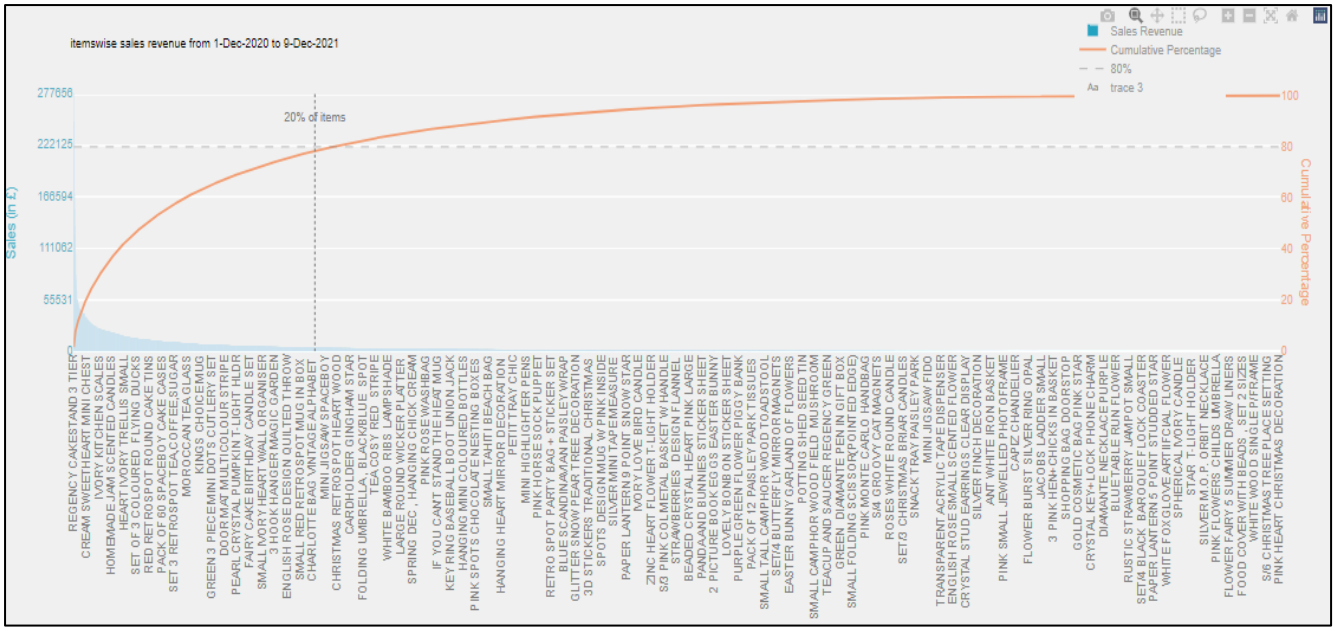


Figure No. 7.16: Items-Wise Sales Revenue

Chapter 8: Data Preparation

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”.

The sample dataset view mentioned above is displayed in Figure No. 8.1.

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	CustomerID	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	01-12-2019 07:45	6.95	13085	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	01-12-2019 07:45	6.75	13085	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	01-12-2019 07:45	6.75	13085	United Kingdom
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	01-12-2019 07:45	2.1	13085	United Kingdom
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	01-12-2019 07:45	1.25	13085	United Kingdom
489434	22064	PINK DOUGHNUT TRINKET POT	24	01-12-2019 07:45	1.65	13085	United Kingdom
489434	21871	SAVE THE PLANET MUG	24	01-12-2019 07:45	1.25	13085	United Kingdom
489434	21523	FANCY FONT HOME SWEET HOME DOORMAT	10	01-12-2019 07:45	5.95	13085	United Kingdom
489435	22350	CAT BOWL	12	01-12-2019 07:46	2.55	13085	United Kingdom
489435	22349	DOG BOWL , CHASING BALL DESIGN	12	01-12-2019 07:46	3.75	13085	United Kingdom
489435	22195	HEART MEASURING SPOONS LARGE	24	01-12-2019 07:46	1.65	13085	United Kingdom
489435	22353	LUNCHBOX WITH CUTLERY FAIRY CAKES	12	01-12-2019 07:46	2.55	13085	United Kingdom
489436	48173C	DOOR MAT BLACK FLOCK	10	01-12-2019 09:06	5.95	13078	United Kingdom
489436	21755	LOVE BUILDING BLOCK WORD	18	01-12-2019 09:06	5.45	13078	United Kingdom

Figure No. 8.1: Sample of the Dataset

8.1 Data cleaning

This dataset contained a large number of records that lacked Customer IDs or had negative order quantities, necessitating data cleaning.

To clean the data, the following steps were taken:

1. The columns "Description" and "CustomerID" contain null values. A unique CustomerID should be associated with each Invoice Number in order to fill in the gaps, it is attempted using the Invoice and CustomerID linkage. Here the null values have been discarded, since it was difficult to discover the linkage.
2. Orders that have been cancelled will be eliminated because, according to the information about the attributes and columns in our dataset, entries with a "C" in the "Invoice" field are orders that have been cancelled.

3. These were cancelled orders. Only one product description should be assigned to each stock code; nevertheless, the dataset disregards this constraint.

4. If the dataset were constrained from many tables, the cause can once again be a lack of constraints or incorrect database joins.

5. Customers who purchased something within the last 30 days are only considered.

8.2 Feature Engineering

1. Some attributes in this dataset were not necessary to forecast CLV. Therefore, a key component of our preprocessing and, ultimately, our dataset, was the feature extraction. "All other attributes but the customer ID, invoice number, date, quantity, and unit price" have been removed, and other features have been eliminated.
2. Following the division of the data into training and target intervals, aggregated data is used to generate features and targets for each customer. Aggregation for the probabilistic model is limited to Recency, frequency, and monetary (RFM) fields.

The following are the new features:

2.1 Frequency is a measure of how frequently a client has made a repeat purchase. This indicates that there have been less purchases overall by one.

2.2 T denotes the customer's age in the selected time units (daily, in our dataset). This is the time span between a customer's initial purchase and the end of the investigational period.

2.3 Recency is the customer's age whenever they made their most recent purchases. The interval between a customer's earliest and most recent purchases is represented by this number.

3. A product's unit price fluctuates between different transactions. This creates issues when averaging the data. For the final column, here the TotalPrice (which is Quantity multiplied by Price) is being included. The quantity and total cost were added during the aggregation process, and the unit price was calculated using the combined numbers. Because the probabilistic model's minimum time unit is a day, the orders were grouped by day rather than the invoice number.

8.3 Train-test 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.

Chapter 9: Modeling

“The CLV is defined as the total profit” that can be anticipated from a customer over the course of their relationship with the firm.

Modelling through historical approach:

“To calculate CLV, these models use historical data. In this category, there are two major models: aggregate models and cohort models. The aggregate model computes the CLV using the average revenue per customer from previous data. Cohort models divide data into cohorts based on characteristics such as transaction date, and then calculate average revenue per cohort. This is used to calculate the CLV of each cohort” (Hariharan S, 2020).

Probabilistic models:

Figure No. 9.1” shows that by applying a probability distribution to the data, these models predict how many transactions and how much money a client will spend in the future. Numerous probabilistic models can be used to determine CLV. Remember that not all variables are influenced by a single model.”.

“Independent of monetary factors like the customer's expected spending, factors like purchase frequency and churn rate are calculated. The figure demonstrates this. The Pareto/NBD (Pareto Negative Binomial Distribution) model and the BG/NBD (Beta Geometric Negative Binomial Distribution) model are two widely used models that have been used to forecast future transactions. Both models can be used according to the same steps. Because it is necessary to account for the discrete-time analogue, the BG/NBD model has been used in this work as one of our models for the comparison analysis. Gamma-Gamma and the BG/NBD model have been employed to forecast the expected amount of cash a consumer will part with. These models were created using the lifetimes library and written in Python.” (Hariharan S, 2020).

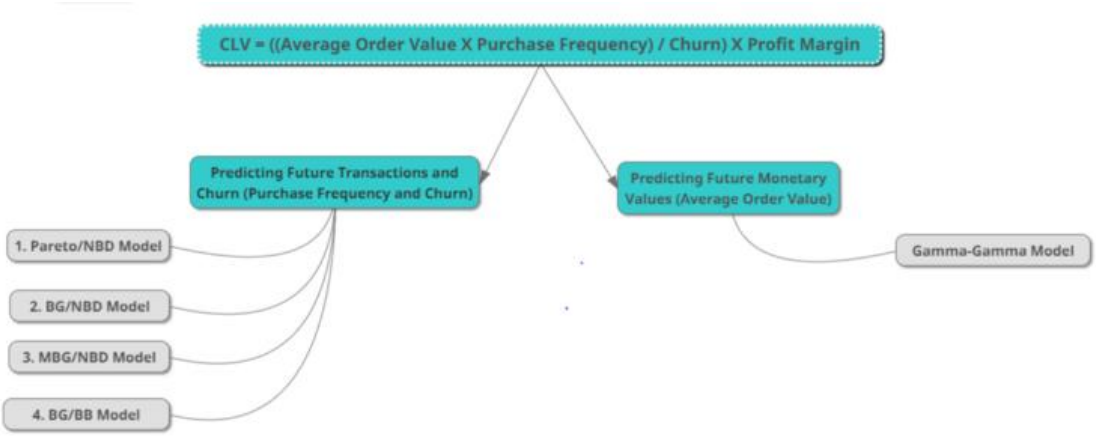


Figure No. 9.1: "The Pareto/NBD Model: A Good Starting Point for CLV Modeling"

"Pareto/NBD only considers the number of purchases and lifetime. It makes no mention of the monetary value. There are several models that deal with monetary value; I chose the Gamma-Gamma extension to the Pareto/NBD model" (Gauthier, 2017).

Regarding the customer population, the Pareto/NBD model makes the following assumptions:

1. **"Purchasing activity exhibits a Poisson distribution with rate:** In other words, these purchases happen at different times, but the pace (measured in counts/unit time) is always the same. The consumer level inter-purchase time should therefore have an exponential distribution".
2. **"The lifetime distribution exhibits an exponential distribution with slope:** Such a distribution has an expected value of 1/μ, which is equal to the user's lifespan".
3. **"Two prior gamma distributions that reflect our perception of how these latent parameters are distributed among the customer population restrict the latent parameters:** The purchase count and lifetime parameters for these two gamma distributions are (r,) and (s,), respectively. Finding these four factors is the goal. These are the only actionable metrics that can be derived". (Gauthier, 2017).

“The Gamma-Gamma Extension to the Pareto/NBD Model”:

The Pareto/NBD model, as previously stated, focuses on modeling lifetime and purchase count. Gamma-Gamma, the monetary value extension to the Pareto/NBD model shown on the right side of the chart, makes a few assumptions:

1. At the customer level, the transaction/order value fluctuates at random around the average transaction value for each client.
2. “The observed mean value is a poor approximation of the latent mean transaction value E(M), where M is the monetary value”.
3. Although the Customers' average transaction amount varies, these values remain constant.
4. The transaction process has no impact on the average value distribution among clients. In other words, the lifetime and purchase count components of the model are not necessary for the modeling of monetary worth. It's possible that this is false in real-world commercial circumstances (Gauthier, 2017).

In the project, it is attempted to model the Customer Lifetime Value in order to identify customers who are more likely to provide high revenue to the business in the future. It is attempted to forecast this over a 30-day period (1 month). For this, the first step was to prepare the data. In the model, RFM (Recency, Frequency, and Monetary) features have been used. Instead of using an aggregate or cohort level for this study, the customer level is being used to forecast the Customers lifetime value.

“To estimate the predicted monetary value as part of the customer lifetime value forecast, the non-contractual with continuous purchase opportunity is being used and the best fitting models will be the BG-NBD Model, Pareto-NBD Model, Modified BG Model, and Gamma-Gamma Distribution Model”.

Then it was intended to learn about the most lucrative customer segment that is most profitable and to whom the marketing team can target to deliver the best-optimized campaigns. Delivering the needed data to the marketing team is our responsibility, and in order to do that, it shall cluster the expected metrics.

Clustering for Customer Segmentation

K-means clustering is a prominent vector quantization technique for cluster analysis in data mining. It was originally developed for signal processing. The goal of the k-means clustering is to divide n observations into k clusters, each of which is made up of the observations that belong to the cluster that has the closest mean to it and acts as the prototype of the cluster. As a result, the data space is divided into Voronoi cells (RPubs, 2019).

When you have unlabeled data, you can utilise a type of unsupervised learning called K-means clustering. The K variable acts as a stand-in for the number of groups, and the goal of this algorithm is to find groups in the data. Each data point is iteratively assigned to one of the K groups based on the provided attributes. Based on feature similarity, data points are clustered. The K-means clustering algorithm's findings are as follows:

1. It is possible to annotate new data using the K cluster centroids.
2. Labels of the training data (each data point is assigned to a single cluster)

Clustering enables you to identify and examine the groups that have emerged naturally rather than creating groupings before looking at the data. The following example shows how to determine the number of groups in the "Choosing K" stage (RPubs, 2019).Each cluster's centroid consists of a set of feature values that characterize the resulting groups. The type of group that each cluster represents can be qualitatively interpreted by looking at the centroid feature weights (RPubs, 2019). The clustering results is mentioned in Chapter 11.

“Market Basket Analysis: Using Apriori Algorithm to define the Changing Trends in Market Data”

The primary goal of a Market Basket Analysis in marketing is to arm the merchant with knowledge about consumer purchasing patterns so that they may make informed decisions (Kaur & Kang, 2016). By analyzing purchasing patterns and carefully anticipating consumer choices, MBA helps retailers to profit from consumer behavior. The MBA algorithms explore "what-if" scenarios involving assortment changes to identify cross-selling opportunities and

innovative planogram concepts for surgically accurate item marketing for stores (applexus, 2021).

It is frequently linked to expressions from online sellers like "customers who bought this item also bought" or Netflix's "See what's next" as shown in Figure No. 9.2. Market basket analysis is quickly becoming the primary force behind e-commerce upselling and cross-selling as consumers and retailers become more and more dependent on recommendation engines for additional purchases (applexus, 2021).

Figure No. 9.2: Netflix's "See what's next."

In this project MBA method has been used for locating products grouped together and subsequently identifying client buying habits. Apriori algorithms produce association rules for better prediction of customer behaviour. It recognizes the things in a data collection and expands them to ever-larger groupings of items. The association rules are a very helpful technique for analyzing behavioural patterns and are not just applicable for "market basket analysis".

The "Apriori algorithm" results used in this study are simple to comprehend and interpret. Another benefit is that the method performs well with huge data sets, making it possible to extract important information that is typically challenging when there are many dimensions (Sylwia Wrona, 2022). Association rule mining can provide useful insights for several of a retailer's most crucial tactics, including Customer analytics, Market Basket Analysis, and Product Clustering (Lim, 2022).

Chapter 10: Model Evaluation

Customer Lifetime Value using the Probabilistic method:

In the project, it is attempted to model the Customer Lifetime Value using the probabilistic models "BG-NBD Model, Pareto-NBD Model, Modified BG Model, and Gamma-Gamma Distribution Model" for estimating the expected monetary value as a part of the customer lifetime value prediction, and segmenting the customers based on their LTV so that this information could be provided to the marketing team to enhance the business profit margin. Unsupervised machine learning was also used to undertake customer segmentation in order to demonstrate an effective tool for strategy development.

The technique of clustering involves organizing all the data into clusters based on the patterns present. Here, the data is divided into three categories so that the retailer can easily target customers with shared interests rather than having to develop unique marketing plans for each individual client.

A cluster of data should contain only comparable data points. There must be substantial differences between every data point in every cluster. Customers in a given cluster may have different needs if they are not similar to one another. It is easier for the merchant to employ targeted marketing when similar data points are contained within the same cluster.

Transactions for an online retail store have been predicted over the course of the upcoming month utilizing RFM features as shown in Figure No. 9.3 and Figure No. 9.4. Recency, Frequency, and Monetary, sometimes known as RFM where:

"Recency" measures the length of time since a user last made a transaction. In most circumstances, a customer is more most likely to react to communications from a brand the more recently they have connected with it.

"Frequency" means the quantity of purchases done by a user. Customers who frequently participate in activities are more involved and likely to be more loyal than those who rarely do so.

“Monetary” indicates the total amount of money a user has spent on a service or a brand over a certain period of time. Customers that are loyal and spend more money than other customers typically deserve special treatment (Sendpulse, n.d.).

	CustomerID	frequency	recency	T	monetary value
0	12346.0	10.0	400.0	725.0	7746.646000
1	12347.0	7.0	402.0	404.0	615.714286
2	12348.0	4.0	363.0	438.0	449.310000
3	12349.0	4.0	717.0	735.0	1107.172500
4	12350.0	0.0	0.0	310.0	0.000000

Figure No. 9.3: RFM Estimation

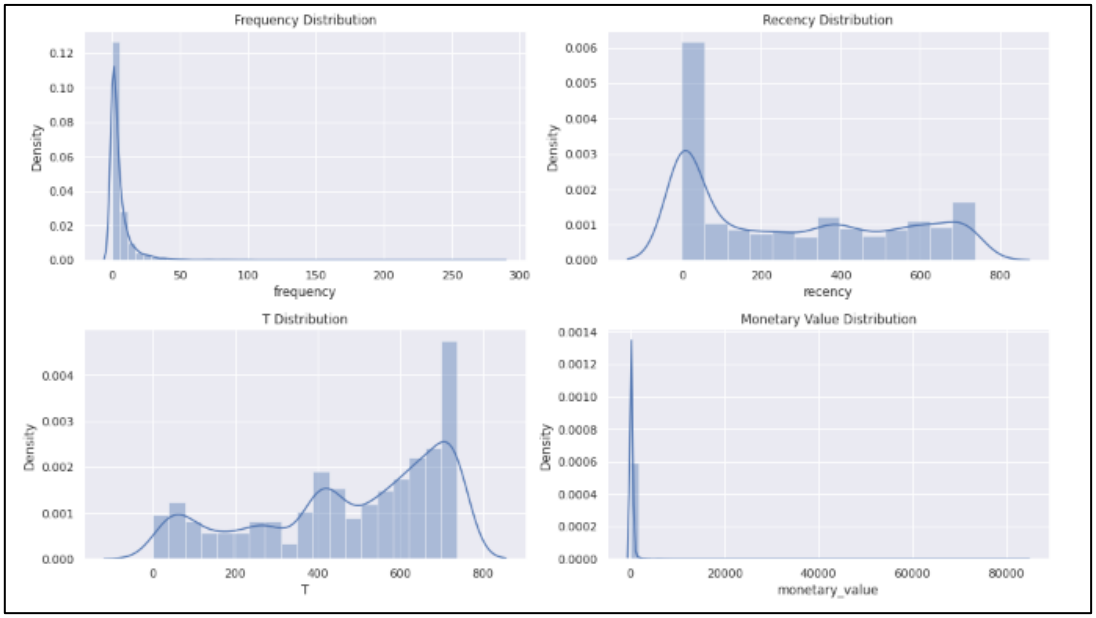


Figure No. 9.4: Recency/Frequency/Monetary Distribution

BG/NBD Model

The "Buy till you die" (abbreviated as BTYD) model, which is based on transaction frequency and when the previous transaction was made, captures the non-contractual customer's purchasing behaviour (also known as Recency). The CLV is calculated using Bayesian probability theories the formula.” CLV is the estimated present value of the customer's expected future cash flows. It is a notion that considers the future and should not be confused

with past customer profitability". The justification for employing the probability model is that, even if we are aware of the variables, such as those relating to marketing, people, and situations, it is nevertheless accepted that the result will be stochastic (Mishra, 2020).

Frequency/Recency Matrix

Example: Think about a customer who made purchases from us every day for three weeks in a row but who hasn't contacted us in months. What are the possibilities that they are "living" right now? very little On the other hand, a client who typically makes a purchase from us once every quarter and did so last quarter is probably still around. It can be seen from this relationship through the Frequency/Recency matrix, which determines the estimated number of transactions a fictitious consumer will make in the forthcoming time frame given his or her Recency (age at last purchase) (the number of repeat transactions he or she has made).

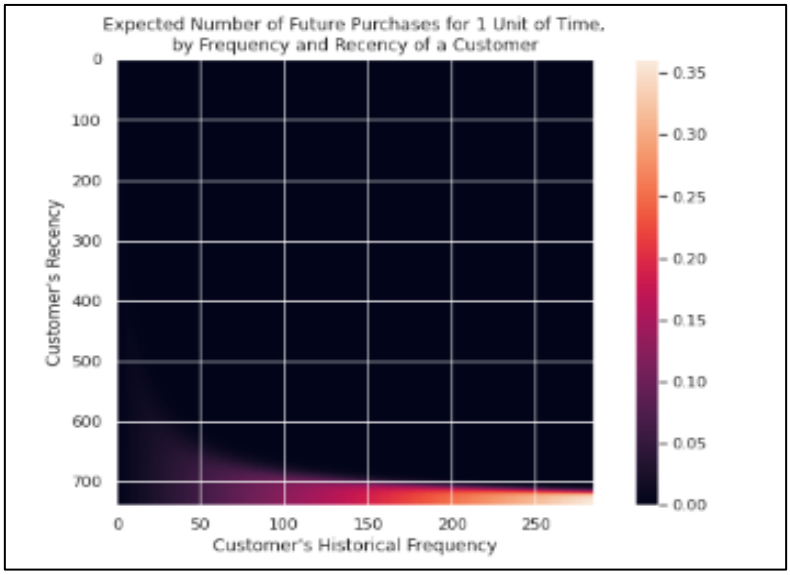


Figure No. 9.5:" Frequency & Recency Matrix Using BG-NBD Model"

From the above Figure No. 9.5, it can be seen that our best customers are where the frequency is 250 and Recency is 700 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:

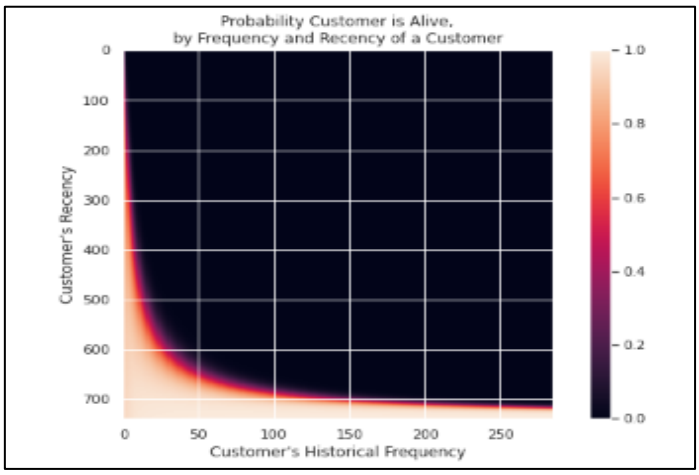


Figure No. 9.6: Probability Customer is Alive Using BG/NBD Model

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 the above Figure No. 9.6, it can be seen that our 80% of customers have already churned or it can be said that they dropped.

	index	CustomerID	frequency	recency	T	monetary_value	probability_alive	predicted_purchases
0	2565	14911.0	284.0	737.0	738.0	1023.546408	0.999268	10.800953
1	402	12748.0	206.0	735.0	735.0	258.904320	0.999298	7.871598
2	2965	15311.0	201.0	738.0	738.0	567.877612	0.999281	7.651860
3	5495	17841.0	193.0	736.0	737.0	352.432850	0.999042	7.355876
4	2260	14606.0	181.0	735.0	736.0	164.289337	0.998994	6.908522
5	743	13089.0	161.0	735.0	737.0	702.166522	0.998648	6.138042
6	1810	14156.0	147.0	729.0	738.0	2124.971565	0.994671	5.577115
7	2181	14527.0	135.0	735.0	737.0	194.985852	0.998493	5.150112
8	4076	16422.0	118.0	721.0	738.0	524.753136	0.984005	4.433800
9	1452	13798.0	101.0	730.0	731.0	742.028416	0.998380	3.888581

Figure No. 9.7: Ranking of top Customers

The top 10 consumers who the model predicts will make purchases the following day are listed above in Figure No. 9.7. It can be observed that the consumer who has bought from us 284 times and just recently will likely make another purchase during the upcoming period. While

the other three columns show their current RFM stats, the predicted purchases column shows how many purchases they anticipate making. The Pareto/NBD model predicts that since these people are currently our top customers, they will soon make further purchases.

“Gamma-Gamma Model”

The “Gamma-Gamma Submodel” will forecast the expected average profit for each customer as well as model the expected average profit distribution.

1. The total amount of a customer's transactions will be randomly dispersed around the average of that customer's monetary value.

2. While the average transaction value may fluctuate over time for different consumers, it does not alter for any particular client.

3. Gamma distribution will be used to spread the average transaction value among all consumers (Palalı, 2021).

From the below Figure No. 9.8, three things are very crucial to note:

1. Time: Time is measured in months for this parameter in the customer lifetime value() method. For example, t=1 denotes a month, and so forth.

2. Frequency: It is required to define the time unit the data is in using the frequency parameter. If our data is on a daily basis, use the letters "D," "M," and so forth.

3. Discount rate: This parameter is based on the idea of “DCF (discounted cash flow)”, in which the yields its present value by applying a discount rate to the monetary worth of a future cash flow. It is stated in the documentation that it is 0.01 on a monthly basis (or 12.7% annually).

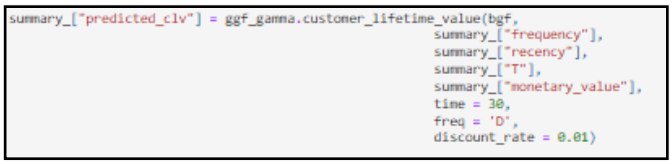


Figure No. 9.8: Predicting Customer Lifetime Value for the next 30 days.

“Pareto-NBD Model”

The “Pareto/negative binomial distribution (NBD) Model”, which simply considers the total number of purchases made over the course of a lifetime, is one of the most well-known and frequently applied RFM models for calculating CLV (Schmittlein, Morrison, and Colombo, 1987). This model makes the assumption that before a consumer "dies," they are initially "alive" (actively making purchases) for an unobserved amount of time (permanently inactive). g The Pareto distribution, also known as an exponential distribution with a gamma-distributed dropout rate, captures customers while they are still alive. The most recent data, the most frequent data, and the duration of the observation period are utilized to forecast a number of future consumer transactions (Fader et al., 2005).

The “Pareto/NBD” seeks to model the existence of customers and, if so, the frequency of their purchases. Customers make purchases using a Poisson process while they are still alive. The distribution of customer lifetimes follows an exponential curve.
Separate gamma distributions describe the population's purchasing rates and survival propensities.

A number of managerial issues can be resolved using the Pareto/NBD and other comparable models, such as calculating the number of "active" consumers, ranking customers according to how likely they are to still be "alive," and forecasting future transaction levels as shown in Figure No. 9.9 and Figure No. 9.10 (Scholars & Rajagopalan, 2018).

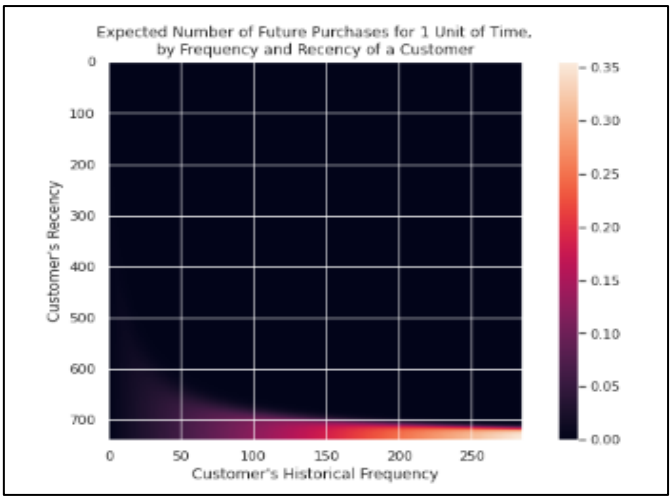


Figure No. 9.9: Frequency & Recency Matrix Using Pareto-NBD Model

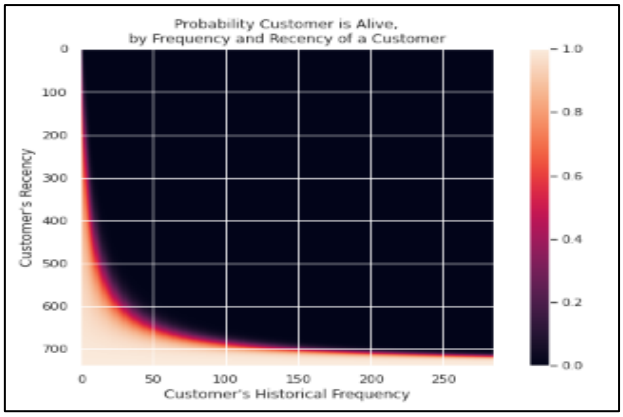


Figure No. 9.10: Probability Customer is Alive Using Pareto/NBD Model

MBG Model

The following list of assumptions applies to this model:

1. The transaction rate (λ) of an active client follows a Poisson pattern.
2. A Gamma distribution with shape (Γ) and scale (α) parameters controls heterogeneity across the transition rate (λ).
3. At time zero and with a consistent likelihood following each purchase, customers become inactive (p)
4. Based on the Gamma distribution with parameters a and b, probability (p) is diverse between consumers.
5. The "transaction rate (λ)" and "drop out probability (p)" differ separately across customers and are not correlated (*Analyticsindiamag*, 2021).

Figure No. 9.11 & Figure No. 9.12 shows the Frequency & Recency Matrix and Probability of Customers who are still alive.

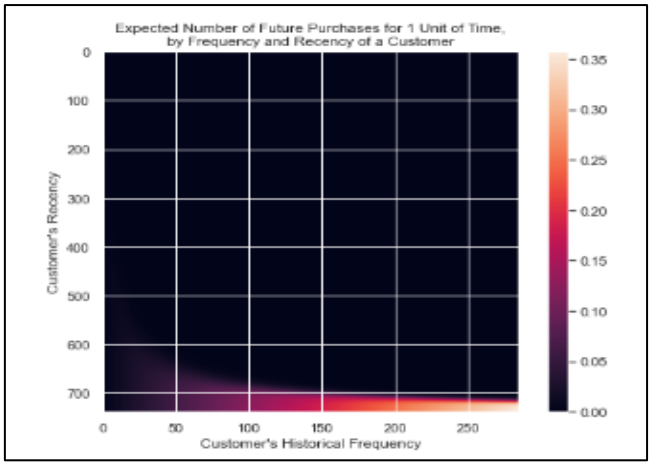


Figure No. 9.11: Frequency & Recency Matrix Using MBG Model

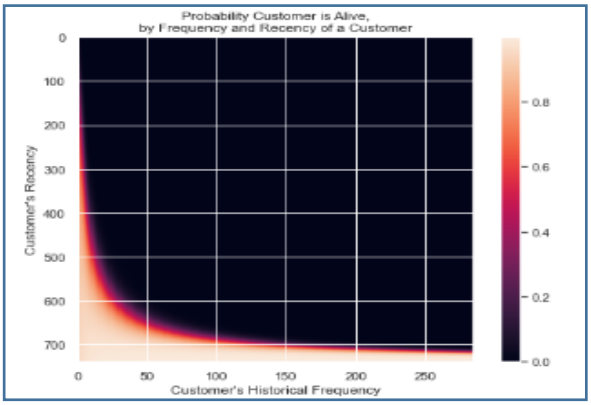


Figure No. 9.12: Probability Customer is Alive Using MBG Model

Finally, the comparison of the above described three models i.e. BG-NBD, Pareto-NBD & MBG Model has been made as shown in Figure No. 9.13.

	BG-NBD	Pareto-NBD	MBG-NBD
MSE Purchase Error	4.337883	4.335935	4.346083
RMSE Purchase Error	2.082758	2.082291	2.084726
Avg Purchase Error	0.411798	0.412367	0.417090

Figure No.9.13: Comparison between the Models

As can be seen, there isn't much of a difference between both models' performances, however, the Pareto NBD model performs a little bit better when it comes to minimizing the MSE & RMSE Errors.

CLV Calculation and Model

In the project, the CLV is determined in two steps:
1. Using Pareto/NBD, estimate the rate at which customers will make future transactions and the rate at which they will exit it.
2. Determine the financial value of each client using Gamma-Gamma Model.

Market Basket Analysis Using Apriori Algorithm

A type of affinity analysis known as market basket analysis, or MBA, has long been employed in the retail industry. It offers a computational tool for discovering typical relationships between objects—typically products—from which strategy can be developed.

In order to locate "often recurring item sets," the Apriori method is used. A user-defined "support" threshold determines how frequently an item set of related items will appear together, such as things in a basket of goods. It aids in the discovery of frequent item sets in transactions and pinpoints the laws of association between these items (*Practical Data Science*, n.d.).

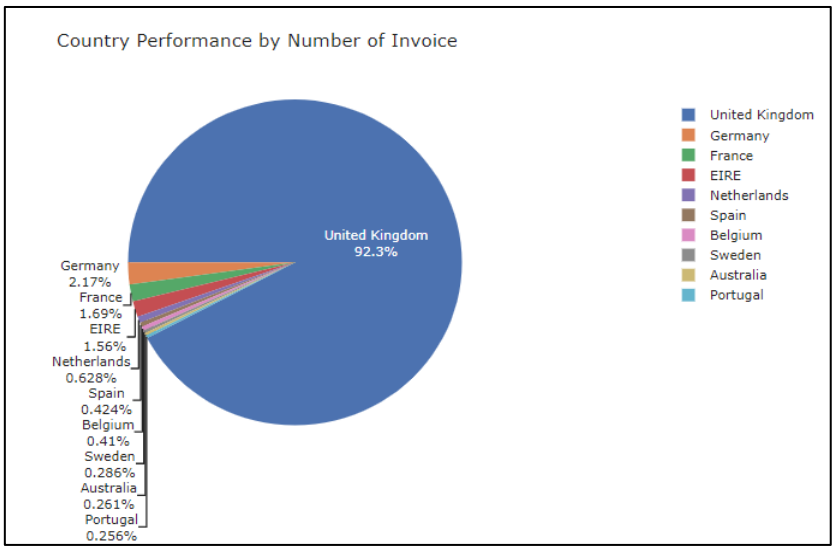


Figure No. 9.14: Top 10 countries performance by the number of Invoices

The visualization in Figure No. 9.14 shows that 92.3% of the sales/transactions are from the UK. Therefore, the transactions that come from the UK will only be utilized in order to make this project easier to complete and more individualized. The basket data will be created once a better understanding is developed of why to only use UK transactions. The quantity of each item purchased per transaction (Invoice) will be included in this basket data (Palalı, 2021).

Description	DOORMAT UNION JACK GUNS AND ROSES	3 STRIPEY MICE FELTCRAFT	4 PURPLE FLOCK DINNER CANDLES	50'S CHRISTMAS GIFT BAG LARGE	ANIMAL STICKERS	BLACK PIRATE TREASURE CHEST	BROWN PIRATE TREASURE CHEST	Bank Charges	CAMPHOR WOOD PORTOBELLO MUSHROOM	CHERRY BLOSSOM DECORATIVE FLASK	...	ZINC STAR T-LIGHT HOLDER	ZINC SWEETHEART SOAP DISH	ZINC SWEETHEART WIRE LETTER RACK
Invoice														
489434	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
489435	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
489436	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
489437	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
489438	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
581582	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
581583	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
581584	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
581585	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
581586	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

33541 rows × 5249 columns

Figure No. 9.15: UK Market Basket Model

The data has been grouped by the transaction (Invoice) & the products (Description) and displayed the values of the quantity of each item purchased using the positive quantity and transaction from UK-only data. The value is then added up and unstacked as shown in Figure No. 9.15.

The data frame's index has been modified to the invoice number so that it could see how many of each item was purchased per invoice number. Essentially, this dataframe represents the "basket" that consumers "carry on" to the register at our store. It reveals the price at which this consumer / transaction (Invoice) purchased a certain good. The value 0 shows that the client didn't purchase the specific item. Providing the support value of 0.03, and while using the Apriori method, it may define the frequent data that is desired.

After Installing the mlxtend package, now It's time to use the association rules after using the apriori algorithm to identify the commonly purchased item. We could extract information and even learn which goods are more profitable to be sold together via association rules.

From the results shown below in Figure No. 9.16, it could find that that there were 33541 transactions that purchased multiple goods. This suggests that 91.67% of the basket data represents a transaction in which more than one item was purchased.

	support	itemsets	length
0	0.151036	(WHITE HANGING HEART T-LIGHT HOLDER)	1
1	0.091219	(REGENCY CAKESTAND 3 TIER)	1
2	0.080957	(ASSORTED COLOUR BIRD ORNAMENT)	1
3	0.076833	(JUMBO BAG RED RETROSPOT)	1
4	0.062252	(PARTY BUNTING)	1
...	...	...	...
68	0.031110	(JUMBO STORAGE BAG SKULLS)	1
69	0.030753	(LUNCH BAG RED SPOTTY)	1
70	0.030428	(NO SINGING METAL SIGN)	1
71	0.030396	(PINK BLUE FELT CRAFT TRINKET BOX)	1
72	0.030363	(VINTAGE HEADS AND TAILS CARD GAME)	1

Figure No. 9.16: Frequently Bought Items

By providing the support value while using the Apriori method, it may define the frequent data that is desired. In this instance, a commonly purchased item has been defined as one that represents up to 3% of the whole transaction, therefore I'll provide a support value of 0.03. The amount of items purchased is then contained in a new column established called length.

As can be seen from the above Figure No. 9.16, there were 73 transactions for what are thought to be frequently purchased items. The **WHITE HANGING HEART T-LIGHT HOLDER**, which has a support value of 0.151028, is the item that customers most usually purchase, as seen in the image. It indicates that out of the total transaction, the item is purchased 5065 times. Information could be extracted and even can be learned which products are more profitable to sell together via association rules.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.055702	0.052129	0.031570	0.566764	10.872332	0.028666	2.187885
1	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.052129	0.055702	0.031570	0.605607	10.872332	0.028666	2.394311
2	(WHITE HANGING HEART T-LIGHT HOLDER)	(RED HANGING HEART T-LIGHT HOLDER)	0.151028	0.050213	0.035532	0.235269	4.685441	0.027949	1.241988
3	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.050213	0.151028	0.035532	0.707633	4.685441	0.027949	2.903785

Figure No. 9.17: MBA Using Apriori Algorithm

Since these two items have the highest "lift" values, it is clear from the association rules results that “WODEN FRAME ANTIQUE WHITE” and “WOODEN PICTURE FRAME WHITE FINISH” are the things that have the highest association with one another. The relationship between the things will be stronger the greater the lift value. It can be stated that two items are related if the lift value is greater than 1, which suffices. The greatest value in this instance is 10.872332, which is an extremely high number. This indicates that selling these two things together would be a great idea.

Additionally, it can be observed from Figure No. 9.17 that the support values of “WODEN FRAME ANTIQUE WHITE” and “WOODEN PICTURE FRAME WHITE FINISH” are 0.031570 and 3.15%, respectively, of the whole transaction, respectively. It occurs 1059 times in total.

Even more, information could be gained from confidence. The antecedent value i.e. 0.055702 in this instance is greater than the consequent value i.e. 0.052129. This indicates that it shall follow Rule No. 1, which is WOODEN PICTURE FRAME WHITE FINISH → WODEN FRAME ANTIQUE WHITE. This indicates that customers are more likely to purchase “WODEN FRAME ANTIQUE WHITE AFTER” purchasing “WOODEN PICTURE FRAME WHITE FINISH”, not in the opposite direction. This could be extremely useful information because it lets us know which things to provide discounts on. If a buyer purchases a WOODEN PICTURE FRAME WHITE FINISH, it may offer a discount on the WODEN FRAME ANTIQUE WHITE (Palalı, 2021).

Collaborative Filtering based systems:

These systems suggest products to users based on the user’s similarity to other users in the system or products that are similar to previous products the user has shown interest in (Hariharan S, 2020).It is further divided into two categories as shown in Figure No. 9.18:

User-based Collaborative filtering: The recommender system uses several similarity metrics to identify users who are similar to the target user in an effort to make suggestions for products based on those users' shared tastes. Here, calculating similarity is a crucial duty.

Item-based Collaborative Filtering: The recommender system looks for products based on the user's past choices and then suggests comparable items to the user. The user may be interested in these things.

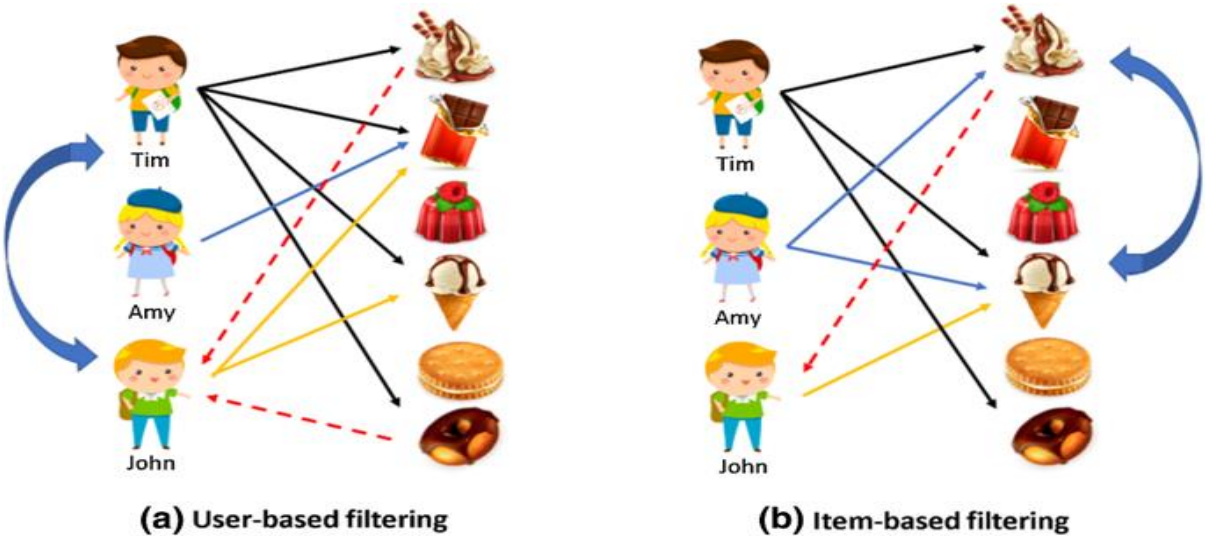


Figure No. 9.18: Collaborative filtering methods

The proposed system makes use of cosine-based similarity to implement User-based and Item-based collaborative filtering to generate recommendations for the active user, as both in practice and research, as well as in information filtering applications and E-commerce applications, collaborative filtering has proved quite effective.

Memory Based collaborative filtering algorithm: Memory-based algorithms produce predictions by using the complete user-item database. These algorithms identify a set of users known as neighbors who have historically sided with the target user using statistical techniques. These systems combine the preferences of neighbors after a neighborhood of users has been established to provide a forecast or "Top-N recommendation for the active user". The methods—"also referred to as nearest-neighbor or user-based collaborative filtering"—are more well-liked and frequently employed in daily life (Sarwar et al., 2001).

Figure No. 9.19 shows items that can be recommended to B based on the preferences of A are:

Items to Recommend to B
{'84086C', 20615, 21832, 21864, 20652, 22348, 22412, 21171, 21908, '79066K', '79191C', 21915, 22620}

StockCode	Description
21864	UNION JACK FLAG PASSPORT COVER
21908	CHOCOLATE THIS WAY METAL SIGN
21832	CHOCOLATE CALCULATOR
22348	TEA BAG PLATE RED SPOTTY
79191C	RETRO PLASTIC ELEPHANT TRAY
21171	BATHROOM METAL SIGN
21915	RED HARMONICA IN BOX
20652	BLUE SPOTTY LUGGAGE TAG
20615	BLUE SPOTTY PASSPORT COVER
79066K	RETRO MOD TRAY
84086C	PINK/PURPLE RETRO RADIO
22412	METAL SIGN NEIGHBOURHOOD WITCH
22620	4 TRADITIONAL SPINNING TOPS
20652	BLUE POLKADOT LUGGAGE TAG
22348	TEA BAG PLATE RED RETROSPOT
20615	BLUE POLKADOT PASSPORT COVER

Figure No. 9.19: Items to recommend to B

The top 10 Similar items to recommend are shown in Figure No. 9.20:

[23166, 23165, 23167, 22993, 23307, 22720, 22722, 23243, 23306, 22961]

StockCode	Description
23166	MEDIUM CERAMIC TOP STORAGE JAR
23165	LARGE CERAMIC TOP STORAGE JAR
23167	SMALL CERAMIC TOP STORAGE JAR
22993	SET OF 4 PANTRY JELLY MOULDS
23307	SET OF 60 PANTRY DESIGN CAKE CASES
22720	SET OF 3 CAKE TINS PANTRY DESIGN
22722	SET OF 6 SPICE TINS PANTRY DESIGN
23243	SET OF TEA COFFEE SUGAR TINS PANTRY
23306	SET OF 36 DOILIES PANTRY DESIGN
23306	SET OF 36 PANTRY PAPER DOILIES
22961	JAM MAKING SET PRINTED

Figure No. 9.20: Top-10 Similar Items

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 re

gression 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 has been trained on many algorithms, evaluate them, an d choose the one that performs the best based on our assessment criteria to determine the mod el that will serve our needs best.

The following mentioned algorithms has been used:

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

Random Forest algorithm

Another ensemble machine learning algorithm that uses the bagging method is Random Forest. This approach is an expansion of the bagging estimator. Decision trees serve as the foundation estimators in random forests. In contrast to the bagging meta estimator, random forest chooses a set of features at random, using those characteristics to determine the optimum split at each decision tree node (Hariharan S, 2020).

Chapter 11: Deployment

Figure No. 10.1 and Figure No. 10.2 shows the model for the prediction of CLV which has been deployed using Streamlit.

Figure No. 10.1: CLV Prediction App using Streamlit

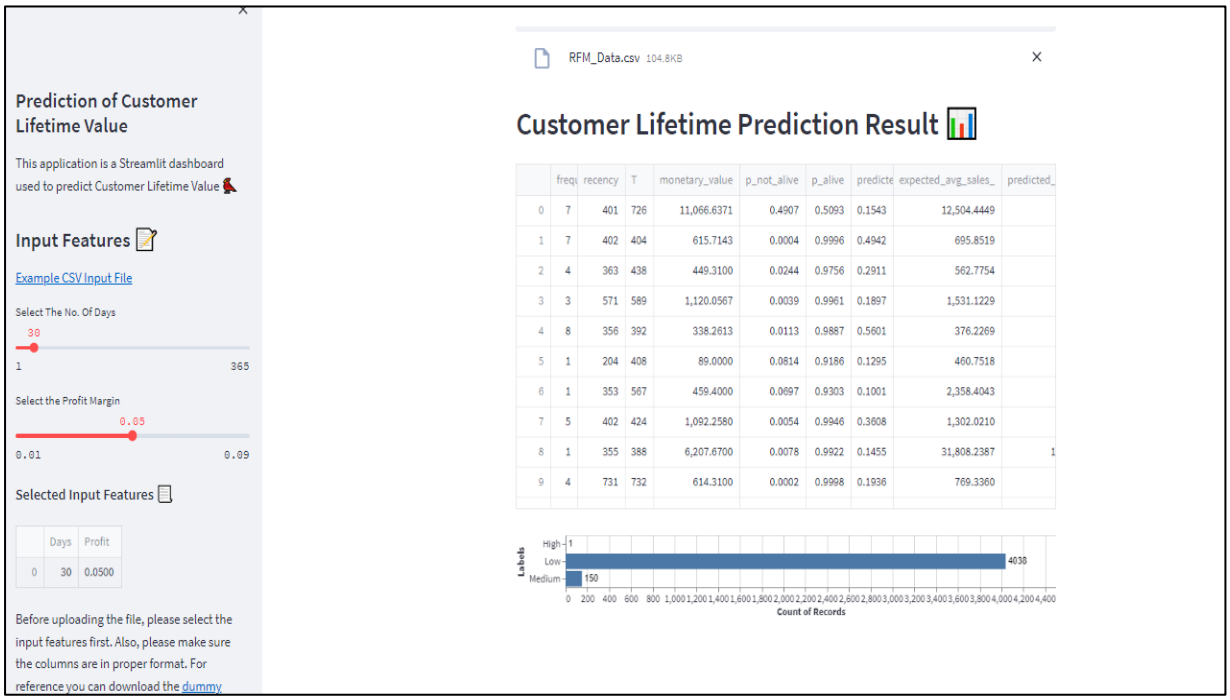


Figure No. 10.2: Customer Lifetime Prediction Result

Chapter 12: Analysis and Results

“The Pareto/NBD Model”

The Pareto/NBD has been a very effective technique for analyzing consumer bases. The Pareto/NBD seeks to model the existence of customers and, if so, the frequency of their purchases. Customers make purchases using a Poisson process while they are still alive. The distribution of customer lifetimes follows an exponential curve. Separate gamma distributions describe the population's purchasing rates and survival propensities (Scholars & Rajagopalan, 2018).

The Pareto/NBD model is expanded upon by the “Gamma-Gamma model”. The monetary value component is not the main focus of Pareto/NBD. “The Gamma-Gamma model”, however, gives each of these future purchases a monetary value. “The Gamma-Gamma model” is a solid strategy because it analyses the financial aspect of each transaction before estimating the likelihood that the customer will remain a customer (Avinash et al., 2019).

The goal of the project is to look for estimate potential revenue (CLV) created in the context of non-contractual-continuous business by a certain group of active consumers using the “Pareto-NBD Model” as the MSE & RMSE of the “Pareto-NBD Model” is less compared to “BG-NBD AND MBG-NBD Model”. As part of the forecast of Customer lifetime value, the expected monetary value is estimated using the Gamma-Gamma Distribution model. Unsupervised machine learning was also used to undertake customer segmentation in order to demonstrate an effective tool for strategy development.

Transforming for Pareto/NBD Model

For the Pareto/NBD, first, the data has to be prepared. Fortunately, the Pareto/NBD model only needs three variables for each customer: “Recency, frequency, and monetary value”. This is taken from the dataset's variables. The amount of days that passed between the customer's most recent purchase and their first one is known as Recency. The total number of repeat purchases made throughout the observation period is known as frequency, and the mean purchase value in dollars is known as monetary value. These variables were extracted with the use of the

lifetimes Python programme. This programme performs an automatic calculation of the three variables as well as the overall observation time T.

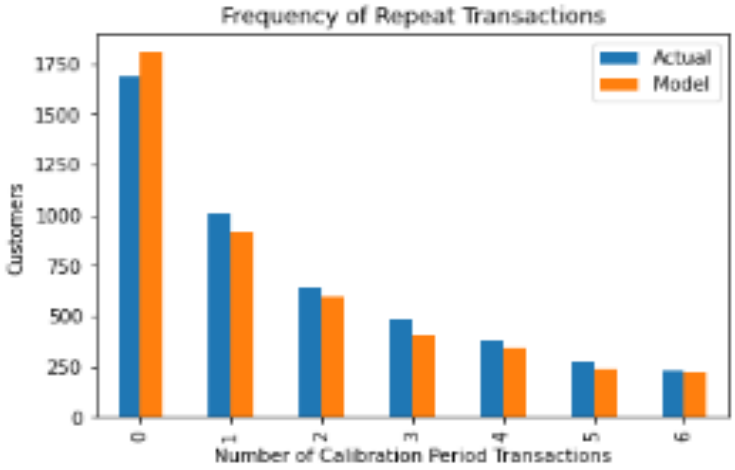


Figure No. 11.1: Frequency of repeat Transactions

The comparison of repeat purchases between the actual and the model's predictions throughout the calibration phase is shown in the plot above in Figure No. 11.1. The better, the closer the actual values and predictions are. The actual numbers are shown in the blue area, while the projections are shown in the orange area. This plot shows how many purchases each customer made throughout our observation period for all of the customers in our data, as well as what the model predicted. More significantly, it demonstrates that the model is not flawed. Consequently, the analysis can be carried out. It is clear that our simulated data and real data closely match each other.

The dataset can be divided into a holdout dataset and a dataset for the calibration period. This is crucial because it is required to test how well our model works with hypothetical. “Lifetimes has a function that divides the dataset”. For the analysis the following parameters have been selected:

Calibration_period_end = 2021-06-08

Observation_period_end = 2021-12-09

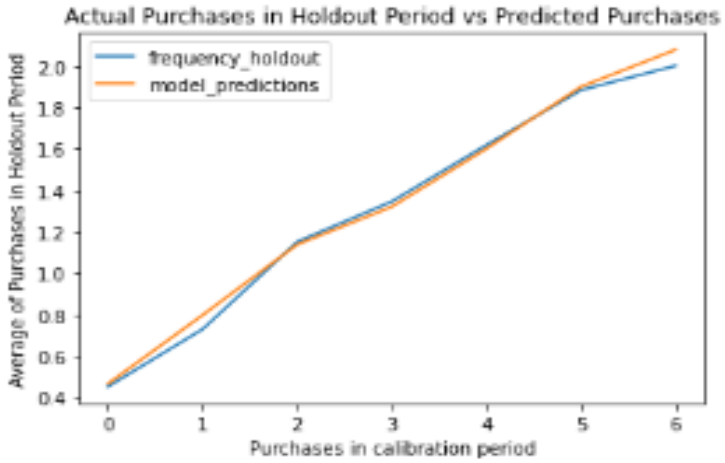


Figure No. 11.2: “Actual Purchases in Holdout Period Vs Predicted Purchases”

The data has been divided into an in-sample (calibration) and validation (holdout) period for this graphic. Beginning on 2021-06-08, the sampling period runs; the validation period runs from 2021-06-09 to 2021-12-09. The graphic classifies all customers according to the number of repeat purchases they made during the calibration period (x-axis) and then takes the average of those repeat purchases during the holdout period (y-axis). The y-model axis's prediction and the actual result are represented by the green and orange lines, respectively. As can be seen from Figure No. 11.2, the model is very good at predicting the behavior of the consumer base outside of the sample (*Towardsdatascience*, 2021).

Data points are organised using K means into discrete, non-overlapping groupings. K-Means clustering is unsupervised: no labelled data for this clustering; It divides objects into clusters that have things in common and are different from things in another cluster. The Customer segmentation count is shown below in Table No. 11.1.

Labels	**CostumerID Count**
High-Value	0.023872
Low	96.228217
Medium	3.747911

Table No. 11.1: Customer Segmentation Using K-Means

Segment K-means	predicted_purchases	expected_avg_sales_	predicted_clv_1month	profit_margin
Low	0.283328	820.983067	4.073496e+03	203.674777
High	0.187433	863116.959561	3.648665e+06	182433.244475
Medium	1.905651	1066.185421	3.971256e+04	1985.628075

Figure No. 11.3: Overview of the Customer Segment

From Figure No. 11.3 it can be concluded that:

1. The High-Value Customers purchase less frequently but with a higher monetary amount of 168469.6£. Additionally, they haven't recently made any purchases. It must thus be asked 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 business should devote the majority of its marketing budget to this segmentation so that the online store may begin to produce and send customers marketing materials in an effort to grab their attention. The business may use email marketing to promote new product information, offer VIP service to boost customer happiness, send customer surveys to determine the demand for particular products, and stock up on enough inventory.
2. Our Mid-Value Customers haven't purchased recently, but they fall within a wide range in terms of frequency and revenue. 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. It has to be looked at a little further, but it also need to make them more recent generally.
3. 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.

Sales Forecasting using Machine Learning Algorithm

The project's goal was to forecast monthly sales for each item using different Machine Learni ng Models.

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 Table No. 11.2 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

Table No. 11.2: Model Comparision

The training and test RMSE for each method is shown graphically below in Figure No. 11.4. 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.

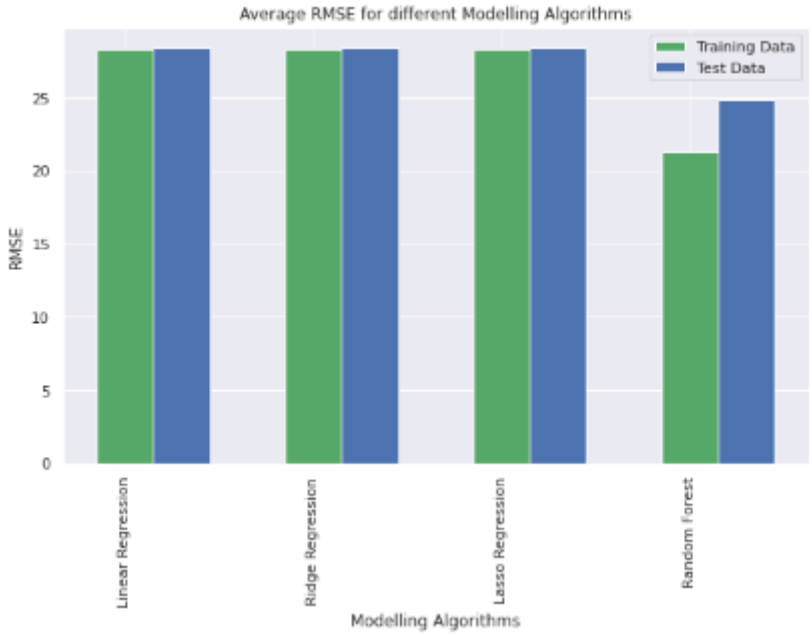


Figure No.11.4: Average RMSE for different Modelling Algorithms

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.

A Dashboard for the prediction of Customer Lifetime value has been prepared using Google Data Studio as shown below in Figure No. 11.5. It is prepared to explain the most crucial key performance indicators (KPIs) for the company. The marketing team can use it to track and ke ep track of the marketing performance so can decide which area to concentrate on using data.

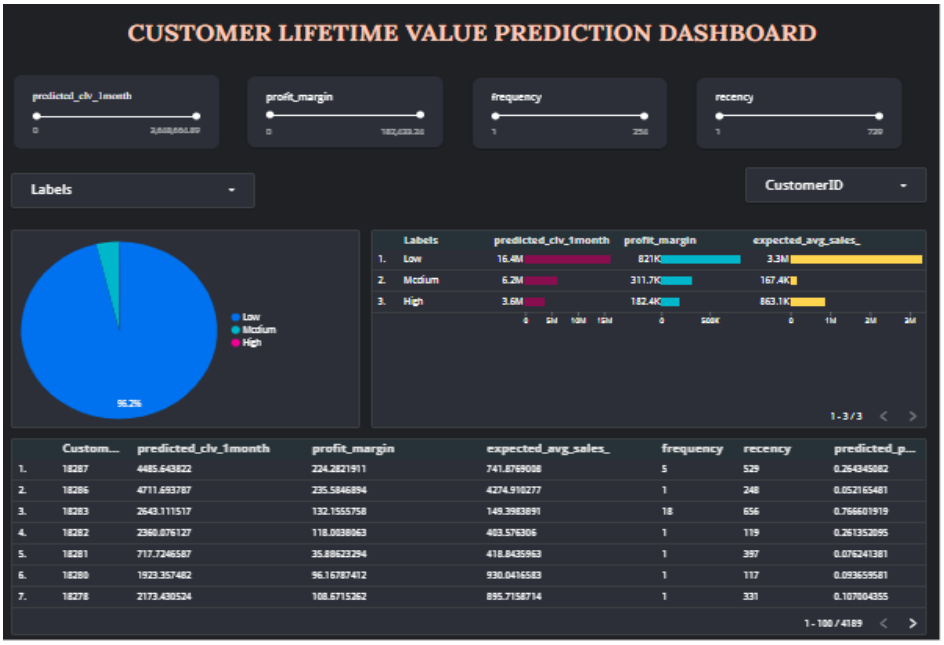


Figure No.11.5: Dashboard for the Prediction of CLV

It can be seen from the below Figure No. 11.6 of the Dashboard how the Parameters change when the frequency is changing. This helps us to give a better understanding of each Customer.

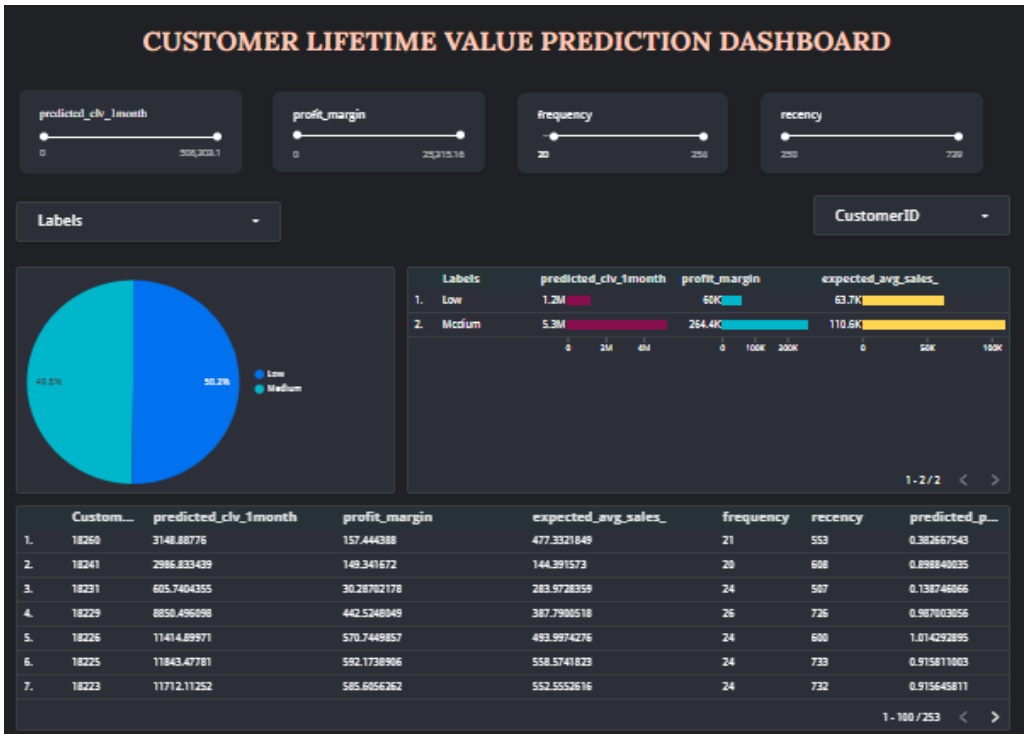


Figure No. 11.6: Measurement of the Key Parameters

Chapter 13: Conclusions and Recommendations for future work

The goal of the study is to look into different approaches for estimating potential revenue CLV produced by a certain set of active consumers in the context of non-contractual-continuous business. “The probabilistic models (Pareto-NBD, BG-NBD, MBG-NBD, & Gamma-Gamma) have been applied to the case study in the industry to make this estimation”. Unsupervised machine learning was also used to undertake customer segmentation in order to demonstrate an effective tool for strategy development. Customers have been successfully divided into the Low, Medium, and High groups in this work. Customers in the group High produce the most revenue for the business, whilst those in the group Low produce the least revenue.

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 (Najib et al., 2019). Finally, given that loyal customers contributed the majority of sales, it can be concluded that businesses should prioritize customer retention.

This study served as a starting point for more research because of its limitations and the wide range of CLV-related research prospects (Tavakolijou, 2009). 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 (Tavakolijou, 2009).

It has been proposed that account additional variables should be taken into account which is not covered in this study as well as additional qualitative elements affecting the industry's CLV (Tavakolijou, 2009). To carry out the identical investigation using the AHP technique and evaluate the outcomes (Tavakolijou, 2009).

Bibliography

Abbasimehr, H., & Bahrini, A. (2022). An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation. *Expert Systems with Applications*, *192*, 116373. https://doi.org/10.1016/J.ESWA.2021.116373

analyticsindiamag. (2021). https://analyticsindiamag.com/

applexus. (2021). *Market Basket Analysis in Retail & CPG Analytics to Increase Sales Revenue and Market Share*. https://www.applexus.com/blogs/market-basket-analysis-in-retail-and-cpg-analytics-to-increase-revenue

AppLovin. (2021). *What is a Good Retention Rate and Why Does it Matter?* https://www.applovin.com/blog/what-is-retention-rate/

Avinash, A., Sahu, P., & Pahari, A. (2019). Big Data Analytics for Customer Lifetime Value Prediction. *Telecom Business Review*, *12*(1), 46–49. https://academica.edu.pl/reading/readMeta?cid=33117012&uid=45988175

Dimaano, R., & Fader, A. P. (2018). *Buy- ' Til-You-Die Models for Large Data Sets via Variable Selection*. 1–22.

Enabled, I., Location, P., & View, M. (2019). A Study on Market Basket Analysis and Association Mining. *In Proceedings of National Conference on Machine Learning*.

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, *42*(4), 415–430. https://doi.org/10.1509/jmkr.2005.42.4.415

Feiz, S., Ghotbabadi, A. R., & Khalifah, Z. B. (2016). Customer Lifetime Value in Organizations. *Asian Journal of Research in Social Sciences and Humanities*, *6*(5), 53. https://doi.org/10.5958/2249-7315.2016.00103.9

Gauthier, J.-R. (2017). *An Introduction to Predictive Customer Lifetime Value Modeling*. https://blogs.oracle.com/

Hariharan S. (2020). *AnalyticsVidya*. https://www.analyticsvidhya.com/

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

Jawad Khan. (2021). How To Improve Customer Lifetime Value in e-Commerce. *B2B SaaS Content Marketing Strategist*. https://www.indellient.com/blog/how-to-improve-customer-lifetime-value-in-e-commerce/

Karolina Matuszewska. (2021). *Customer lifetime value: what it is and why it is important for your business*. https://piwik.pro/blog/customer-lifetime-value-important-for-your-business/

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. *Procedia Computer Science*, *85*(Cms), 78–85. https://doi.org/10.1016/j.procs.2016.05.180

Lew, G. (2017). The importance of customer lifetime value in determining their profitability. *The Business and Management Review*.

Library, M. (2022). *Customer Lifetime Value and Why it Matters*. https://mailchimp.com/clv/

Lim, Y. (2022). *Data Mining: Market Basket Analysis with Apriori Algorithm*. https://towardsdatascience.com/data-mining-market-basket-analysis-with-apriori-algorithm-970ff256a92c

Mishra, R. (2020). *Predicting Customer Lifetime Value through Buy Till You Die model and the importance of segmentation*. https://medium.com/@richa.mishr01/predicting-customer-lifetime-value-through-buy-till-you-die-model-and-the-importance-of-f380af435dca

Mohammadian, M., & Makhani, I. (2019). RFM-Based customer segmentation as an elaborative analytical tool for enriching the creation of sales and trade marketing strategies. *International Academic Journal of Accounting and Financial Management*. https://doi.org/10.9756/iajafm/v6i1/1910009

Mosaddegh, A., Albadvi, A., Sepehri, M. M., & Teimourpour, B. (2021). Dynamics of customer segments: A predictor of customer lifetime value. *Expert Systems with Applications*, *172*. https://doi.org/10.1016/j.eswa.2021.114606

Olkhov, E. (2019). *Current and Future State of Recommender Systems*. Medium. https://medium.com/compassred-data-blog/current-and-future-state-of-recommender-systems-ea3c4669c6ba

Palalı, A. S. (2021). *Medium*. https://medium.com/

Practical Data Science. (n.d.). https://practicaldatascience.co.uk/

Prakhar Gurawa. (2021). https://prakhargurawa.medium.com/

pratomo, edwin agung, Najib, M., & Mulyati, H. (2019). Customer Segmentation Analysis Based on the Customer Lifetime Value Method. *Jurnal Aplikasi Manajemen*, *17*(3), 408–415. https://doi.org/10.21776/ub.jam.2019.017.03.04

Ronan Martin. (2019). Using Customer Lifetime Value As A Segmentation Strategy. *Digital*

Growth Stratergy. https://www.ronanmart.in/blog/customer-lifetime-value-segmentation-strategy/

RPubs. (2019). *RPubs*. https://rpubs.com/

SaasOptics. (2022). *B2B SaaS Financial Operations, Supercharged*. https://saasoptics.com/

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web, WWW 2001*, 285–295. https://doi.org/10.1145/371920.372071

Scholars, J. W., & Rajagopalan, R. (2018). *ScholarlyCommons ScholarlyCommons A Recency-Only Pareto/NBD A Recency-Only Pareto/NBD*. https://repository.upenn.edu/joseph_wharton_scholars/58

Sendpulse. (n.d.). *Accelerate sales and grow your business with SendPulse*. https://sendpulse.com/

Series, C. (2020). *Consumer purchase patterns based on market basket analysis using apriori algorithms Consumer purchase patterns based on market basket analysis using apriori algorithms*. https://doi.org/10.1088/1742-6596/1524/1/012109

Shah, K., Salunke, A., Dongare, S., & Antala, K. (2018). Recommender systems: An overview of different approaches to recommendations. *Proceedings of 2017 International Conference on Innovations in Information, Embedded and Communication Systems, ICIIECS 2017*, *2018-Janua*, 1–4. https://doi.org/10.1109/ICIIECS.2017.8276172

Shakirova, E. (2017). Collaborative filtering for music recommender system. *Proceedings of the 2017 IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017*, 548–550. https://doi.org/10.1109/EIConRus.2017.7910613

Strategy, E. M. and. (2022). *What Is Customer Lifetime Value And Why Is It Very Important*. https://www.bigcommerce.com/ecommerce-answers/what-is-customer-lifetime-value-and-why-is-it-very-important/

surveymonkey. (2022). https://www.surveymonkey.com/

Sylwia Wrona. (2022). *Association rules - market basket analysis*. https://rpubs.com/smwrona/871575

Tavakolijou, M. (2009). *A Model to Determine Customer Lifetime Value in Iranian Banking Industry*.

Towardsdatascience. (2021). https://towardsdatascience.com/

UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/

Ünvan, Y. A. (2021). Market basket analysis with association rules. *Communications in Statistics - Theory and Methods*. https://doi.org/10.1080/03610926.2020.1716255

Zhao, J., Zhuang, F., Ao, X., He, Q., Jiang, H., & Ma, L. (2021). Survey of Collaborative Filtering Recommender Systems. In *Journal of Cyber Security*. https://doi.org/10.19363/J.cnki.cn10-1380/tn.2021.09.02

Plagiarism Report[1]

Prediction of Customer Lifetime Value in E-Commerce Business

by Mahapara Gayasuddin

Submission date: 26-Aug-2022 06:37PM (UTC+0530)
Submission ID: 1887435378
File name: of_Customer_Lifetime_Value_in_E-Commerce_Business-Mahapara.docx (2.45M)
Word count: 11786
Character count: 63330

[1] Turnitn report to be attached from the University.

Prediction of Customer Lifetime Value in E-Commerce Business

ORIGINALITY REPORT

6%	4%	1%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

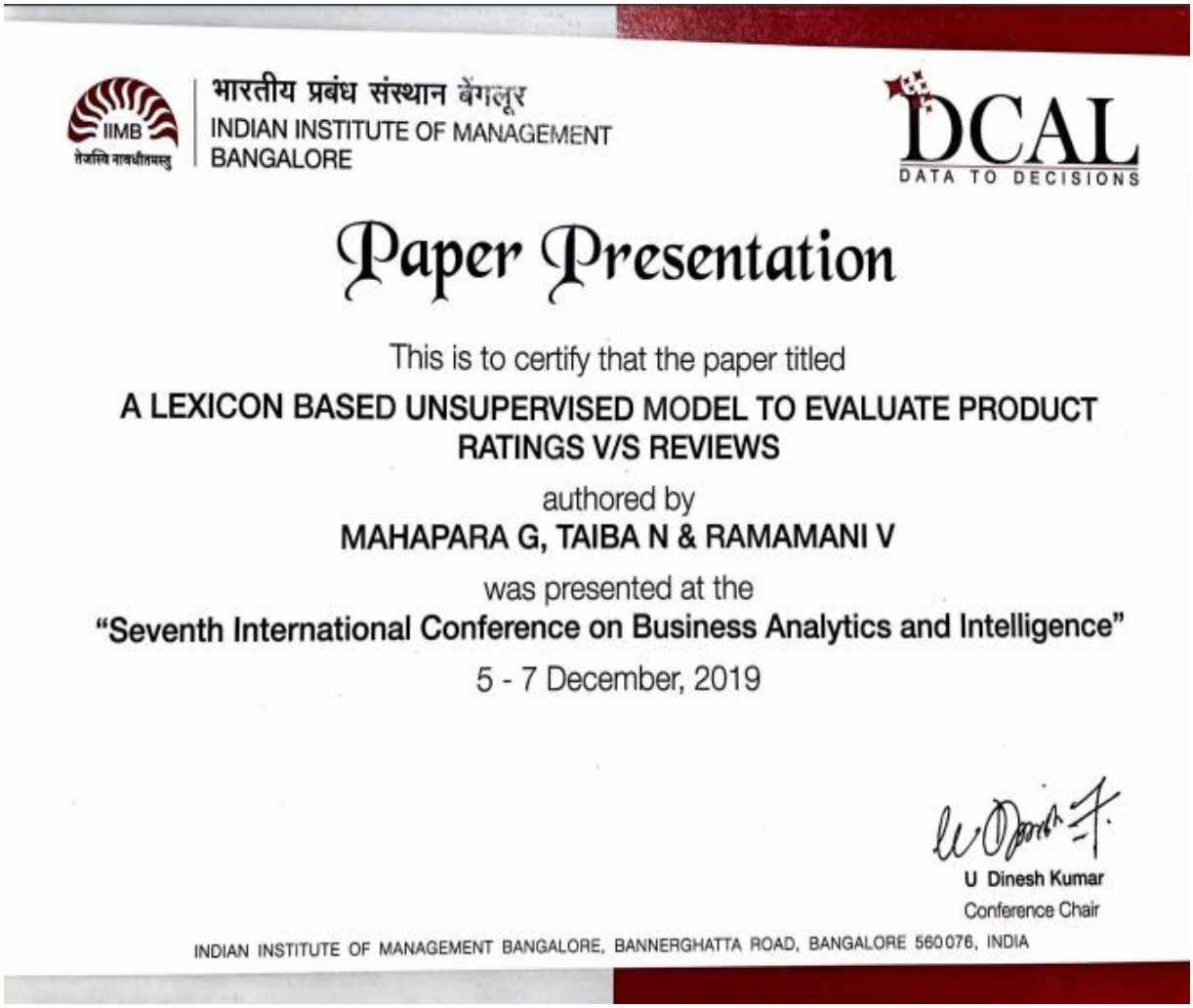
1	global.oup.com Internet Source	1%
2	www.dspace.dtu.ac.in:8080 Internet Source	1%
3	Submitted to Liverpool John Moores University Student Paper	<1%
4	Submitted to University of East London Student Paper	<1%
5	Submitted to Symbiosis International University Student Paper	<1%
6	Submitted to Postgraduate Institute of Management Student Paper	<1%
7	github.com Internet Source	<1%
8	Submitted to Cyryx College, Maldives Student Paper	<1%

9	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1%
10	Submitted to University of Exeter Student Paper	<1%
11	hdl.handle.net Internet Source	<1%
12	iopscience.iop.org Internet Source	<1%
13	www.nickventurella.com Internet Source	<1%
14	Submitted to National College of Ireland Student Paper	<1%
15	Submitted to Coventry University Student Paper	<1%
16	ukdiss.com Internet Source	<1%
17	Submitted to Queen Mary and Westfield College Student Paper	<1%
18	Submitted to University of North Texas Student Paper	<1%
19	www.coursehero.com Internet Source	<1%

20	Patrick Bachmann, Markus Meierer, Jeffrey Näf. "The Role of Time-Varying Contextual Factors in Latent Attrition Models for Customer Base Analysis", Marketing Science, 2021 Publication	<1 %
21	www.ir.dnb.no Internet Source	<1 %

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Publications in a Journal/Conference Presented/White Paper[2]



भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

IIMB
तेजस्वि नावधीतमस्तु

DCAL
DATA TO DECISIONS

Paper Presentation

This is to certify that the paper titled

A LEXICON BASED UNSUPERVISED MODEL TO EVALUATE PRODUCT RATINGS V/S REVIEWS

authored by

MAHAPARA G, TAIBA N & RAMAMANI V

was presented at the

"Seventh International Conference on Business Analytics and Intelligence"

5 - 7 December, 2019

U Dinesh Kumar
Conference Chair

INDIAN INSTITUTE OF MANAGEMENT BANGALORE, BANNERGHATTA ROAD, BANGALORE 560076, INDIA

[2] URL of the white paper/Paper published in a Journal/Paper presented in a Conference/Certificates to be provided.

Paper submitted:

Mahapara Gayasuddin, Krishna Kumar Tiwari, Mithun, “Prediction of Customer Lifetime Value in E-Commerce Business for Growth” 9th International Conference on Business Analytics and Intelligence, IIMB.

Submission Date: 20th October 2022.

Github Link

https://github.com/mahapara2411/Capstone-2

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.

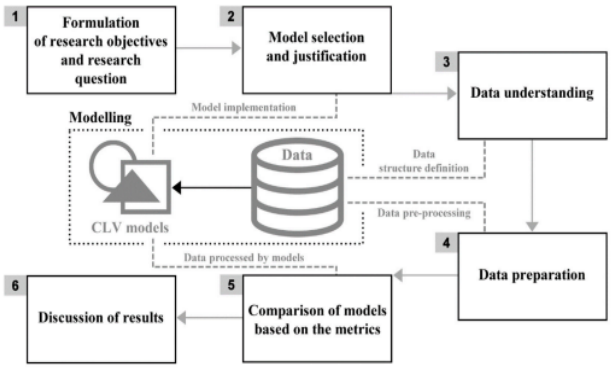


Fig. 1: Methodology of the Project

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].

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

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	0.481909	416.401811	20.820091
3	12352	2848.84	0.0	0.0	2550.295710	298.544290	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	380.304437	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).

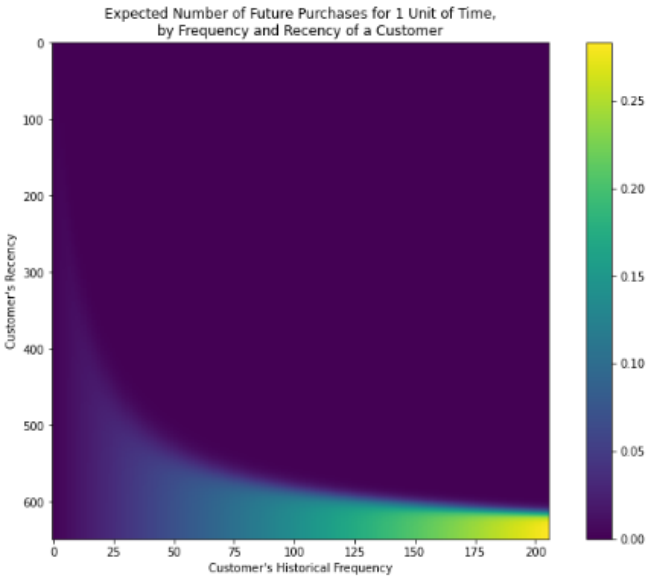


Fig. 3: Frequency & Recency Matrix Using BG-NBD Model

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:

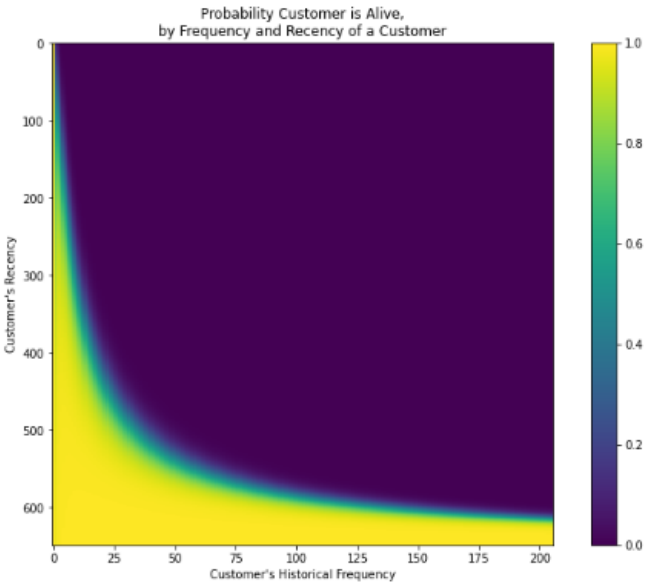


Fig. 4: Probability Customer is Alive Using BG/NBD Model

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

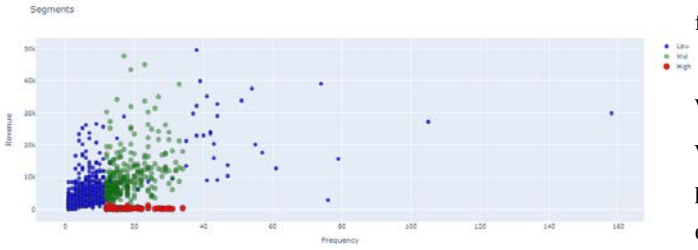


Fig. 7: Frequency Vs Monetary

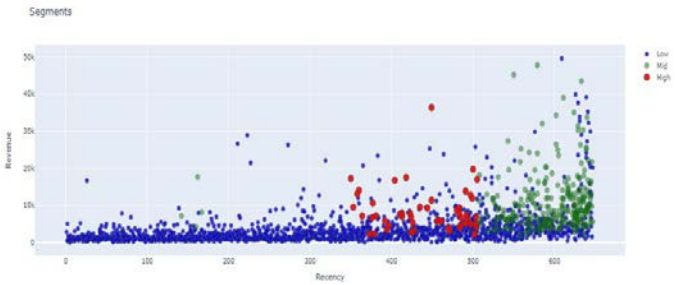


Fig. 8: Recency Vs Monetary

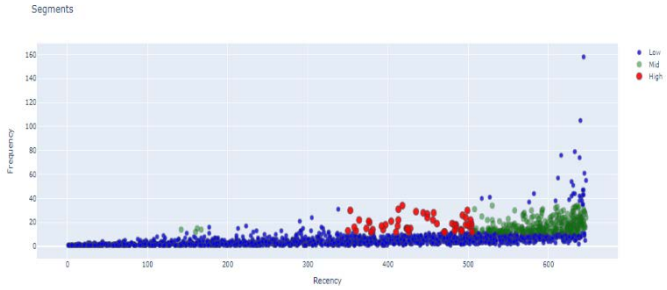


Fig 9: Frequency Vs Recency

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.

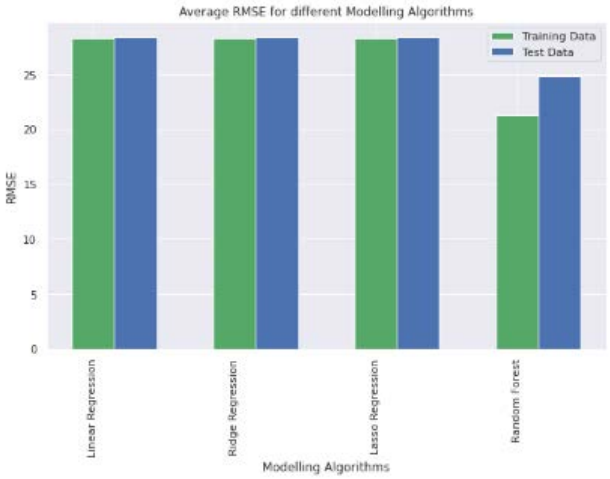


Fig. 11: Average RMSE for different Modelling Algorithms

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.