REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

IIMB
तेजस्वि नावधीतमस्तु
भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

REVA Academy for Corporate Excellence (RACE)

DCAL
DATA TO DECISIONS

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

Presented by:

Mahapara Gayasuddin
Research Scholar, RACE, Reva University

Krishna Kumar Tiwari
Mentor, Jio, General Manager

Mithun Dolthody Jayaprakash
Mentor, RACE, Reva University

race.reva.edu.in

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Agenda

DCAL
DATA TO DECISIONS

01 Introduction
Back Ground | Current status | Why this study

02 Literature Review
Seminal works | Summary | Research Gap

03 Problem Statement
Business Problem | Analytics Solution

04 Project Objectives
Primary & Secondary Objectives | Expected Outcome

05 Project Methodology
Conceptual Framework | Research Design

06 Business Understanding
Business Context | Monetary Impact

07 Data Understanding
Data Collection | Variables

08 Data Preparation
Pre-processing | Process | Techniques

09 Descriptive Analytics
Univariate | Bivariate | Hypothesis

10 Modeling
Machine Learning | Model Evaluation | Insights

11 Model Deployment
Applications | Demo

12 Suggestions and Conclusions
Insights | Next Step | Future Scope

13 Annexure
References | Publications | Plagiarism Score

Introduction

Background | Current status | Why this study

- ✓ Customer Lifetime Value (CLV) is one of the key stats likely to be tracked as part of a customer experience program. CLV is a measurement of how valuable a customer is to your company for an unlimited period as opposed to just the first purchase. This metric helps you understand a reasonable cost per acquisition. CLV is the total worth to a business of a customer over the whole period of their relationship. It's an important metric as it costs less to keep existing customers than it does to acquire new ones, so increasing the value of your existing customers is a great way to drive growth.

- ✓ CLTV tells marketers how much revenue they can expect from one customer over the business relationship. The longer a customer continues to purchase from a company, the greater their lifetime value becomes.

- ✓ CLV helps marketers make smarter decisions by encouraging them to spend less time obtaining low-value clients.

- ✓ Customer retention is one of the primary reasons for measuring CLV. According to Marketing Metrics, the probability of selling to a new prospective customer is 5%–20%, while the probability of selling to an existing customer is 60%–70%. It follows that selling more to repeat customers will result in significantly higher profits. Regular customers tend to spend more money on your products, which helps you grow and promote your company. According to a Criteo survey, 81 % of marketers believe that tracking CLV increases sales.

- ✓ CLV is all about knowing your customers in terms of the value they provide for your company. Therefore, CLV can be used as a segmentation approach. Additionally, when we can segment our customer base, we may determine the most and least lucrative clients, cater to their needs, and make the most use of our resources. Customer profitability can increase the efficacy of marketing initiatives. Segmentation is a basic component of marketing.

- ✓ 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.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Literature Review

Seminal works | Summary | Research Gap

Title of the Paper	Author & Year	Journal/Source	Major Insights	Research Gap
Dynamics Of Customer Segments: A Predictor Of Customer Lifetime Value	Mosaddegh, Abdolreza Albadvi, Amir Sepehri, Mohammad Mehdi Teimourpour, Babak	Expert Systems with Applications	They have studied the dynamics of bank customers through value segments using big data analytics.	The study has been conducted on bank customers.
Buy- ' Till-you-die Models For Large Data Sets Via Variable Selection	Dimaano, Rafael Fader, Advisor Peter,2018	SEMANTIC SCHOLAR	The theory of the BTYD models is laid out and frameworks for incorporating regression elements to the model class are developed.	Only the BTYD Model has been used.
Comparative Analysis Of Selected Probabilistic Customer Lifetime Value Models In Online Shopping	Jasek, Pavel Vrana, Lenka Sperkova, Lucie Smutny, Zdenek Kobulsky, Marek,2019	Journal of Business Economics and Management	Eleven CLV models were used for comparision.	Only the model comparisions have been made.
Forecasting with Ensemble Methods: An Application Using Fashion Retail Sales Data	Orkun Berk Yüzbaşıoğlu, Hande Küçükaydin	ResearchGate	They have studied the ensemble methods of machine learning to predict short term store sales of a fashion retailer	The study has been conducted on fashion industry.
Big Data Analytics for Customer Lifetime Value Prediction	Avinash, Aslekar Sahu, Piyali Pahari, Arunima,2019	Telecom Business Review	To determine the dynamic view of customer behavior, future marketing strategies and to foster brand loyalty, prediction of a proper CLV model is much needed.	The paper is only talking about using Pareto-NBD & Gamma-Gamma Model for the prediction of CLV.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Problem Statement

Business Problem | Analytics Solution

- ✓ The goal of the project is to model CLTV in order to identify customers who are more likely to provide the company with high income in the future. It is attempted to predict this over a 90-day period (3 months). To date, preparing the data was the first stage. Age on Network and RFM (Recency, Frequency, and Monetary) features were employed in the model. These attributes have made it easier to estimate CLTV value.

- ✓ 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 (Pareto Negative Binomial Distribution or Beta Geometric Negative Binomial Distribution) 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.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Business Understanding

Business Impact | Challenges | Monetary Impact

- ✓ Customer lifetime value is a key metric for gaining a better understanding of the customers. It's a forecast of the value your customer relationship can bring to your company. This approach enables organizations to demonstrate the future value that their marketing initiatives can generate. Concentrating on CLV allows for creating an efficient strategy with concise budget planning. However, some customers are more valuable to the company than others. That is why it is critical to understand which ones you should prioritize and invest in first.

- ✓ 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 customer lifetime value.

- ✓ One of the most significant constraints and deficiencies of sales and trade marketing departments in terms of sales development in the E-Commerce industry is a lack of knowledge about which customer segments to target and how to deal with each one. Customer segmentation using the RFM method, as well as customer lifetime value (CLV) would be useful for sales, trade marketing, and marketing decisions in all industries, particularly active companies in the E-Commerce industry.

- ✓ The customer segmentation model assigns groups of customers to corresponding marketing strategies, allowing businesses to maximize profits. We can identify precise segments and strategize accordingly. This helps to create targeted campaigns than using the traditional blanket campaign approach.

Data Understanding

Data Collection | Variables

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 contains all the transactions occurring between 01/12/2019 and 09/12/2021 for a UK-based and registered Online Retail store. The company primarily offers distinctive 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.

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, a five-digit integral number assigned to every customer separately.
Country	Nominal	The name of the country where each customer resides.

Data Preparation

Pre-processing | Techniques

Feature Engineering

For the probabilistic approach followed, some additional data processing steps have been taken:

- ✓ The orders were grouped by day instead of Invoice because the minimum time unit used by the probabilistic model is a day.
- ✓ Only customers who bought something in the past 90 days are considered.
- ✓ Only the fields that were useful for the probabilistic model are retained.

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:

- ✓ Orders received prior to the threshold date are used to train the model.
- ✓ Orders received after the threshold date are used to determine the target value. The date chosen for our analysis is 2021-06-08.

Target and Feature Variables for Sales Prediction

A Machine learning algorithm is built to predict sale quantity of each item for a month. So, the target variable is Quantity. As it is a continuous variable, we will be using regression algorithms.

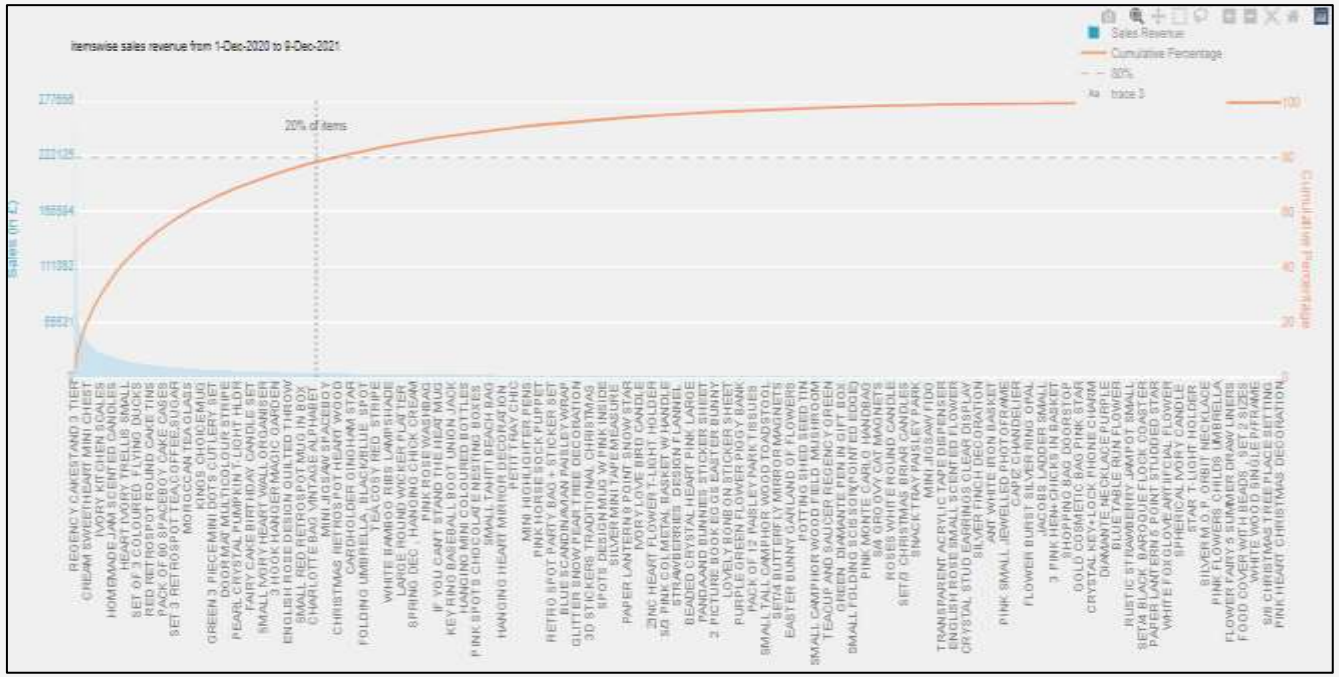
Train Test Split

For this purpose, we have hold out the data for last month from Nov-01-2021 to Dec-09-2021 as our test set, and the remaining data will be used to train our model.

Descriptive Analytics

Multivariate Analysis | Hypothesis

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.

CLV Calculation and Model

In the project, the CLV is determined in two steps:

1. To Determine the rate at with customers will make future transactions and the rate at which customers will drop out of the system in the future using Pareto/NBD or BG/NBD.

2. Calculate the monetary value of each customer.

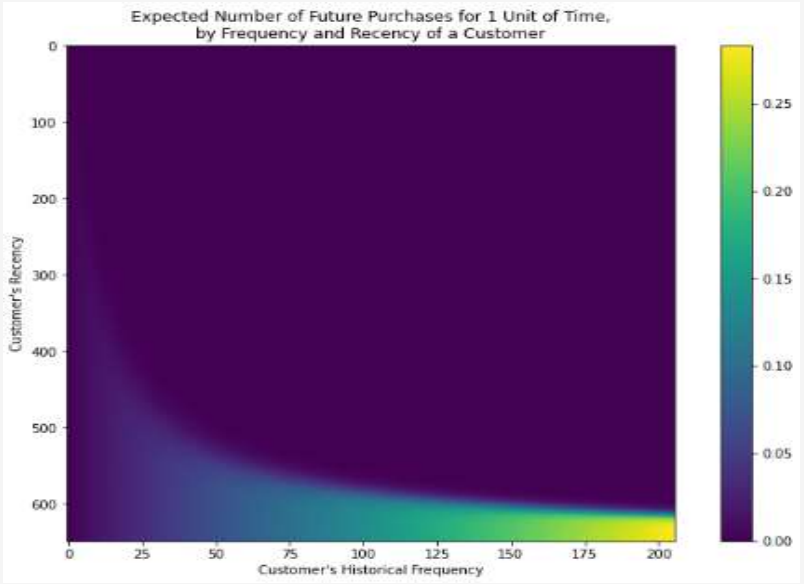
BTYD models (Pareto/NBD or BG/NBD) give us the following three outputs:

- $P(X(t) = x \mid \lambda, p)$ - probability of observing x transactions in given time t
- $E(X(t) \mid \lambda, p)$ - expected number of transactions in given time t
- $P(\tau > t)$ - probability of the customer being inactive at time t

- ✓ The outputs of the above-mentioned probabilistic model are used to model the future monetary value of the customers. The probabilistic method assumes that monetary value follows a gamma-gamma distribution.
- ✓ 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.

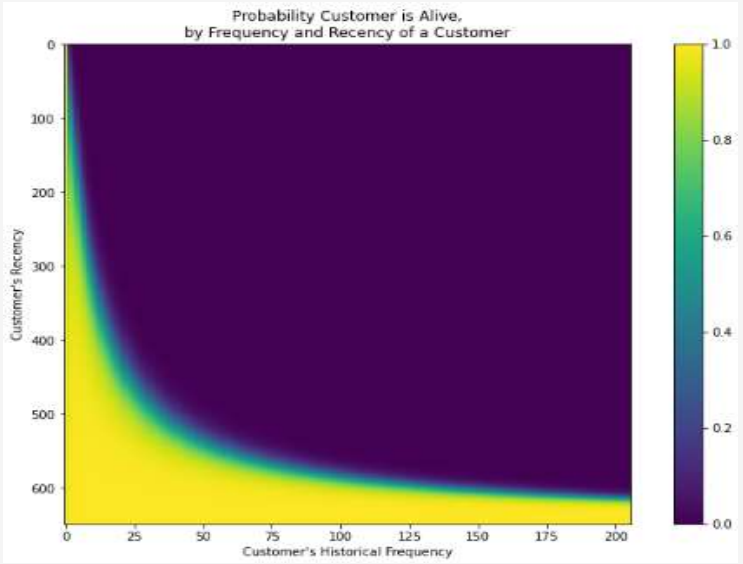
Modeling

Modeling Techniques | Modeling Process | Model Building



Frequency & Recency Matrix Using BG-NBD Model

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.



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 Figure, it can be seen that our 80% of customers have already churned or it can be said that they dropped.

Modelling

Customer Segmentation

One more analysis we can gain from our dataset is implementing the RFM value and segmenting our customers accordingly.

To achieve segmentation, we perform clustering on all 3 metrics individually - Recency, Frequency, and Monetary value. Our model is k-means and the optimum number of clusters obtained through elbow plot is 4.
On these different clusters, we perform weighted sum and achieve an overall score.

t6_cust['OverallScore'] = t6_cust['RecencyCluster'] + t6_cust['FrequencyCluster'] + t6_cust['RevenueCluster']

t6_cust.groupby('OverallScore')['recency','frequency_btyd','target_monetary'] .mean()

Through this process, a total of 8 overall score clusters are obtained.

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

Overall Score based on RFM

Modelling

After analyzing the mean Recency, Frequency, and monetary values of these clusters, we can observe three major groups being formed. We will label then Low Value, Mid Value, and High-Value customers.

Customer	Overall Score
Low-Value	0-3
Mid-Value	4-5
High-Value	5+

Since we have three groups, we want to strategize customer retention on their individual attributes as that would provide a higher retention value. Therefore, we want to understand where these customers lag i.e. whether they buy less frequently, or they buy low-value or less number of items, or they buy sporadically or have they not bought recently, and then plan to mitigate these issues.

We plot our customers against
- Revenue with Frequency
- Revenue with Recency
- Recency with Frequency

Modeling

Modeling Techniques | Modeling Process | Model Building

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

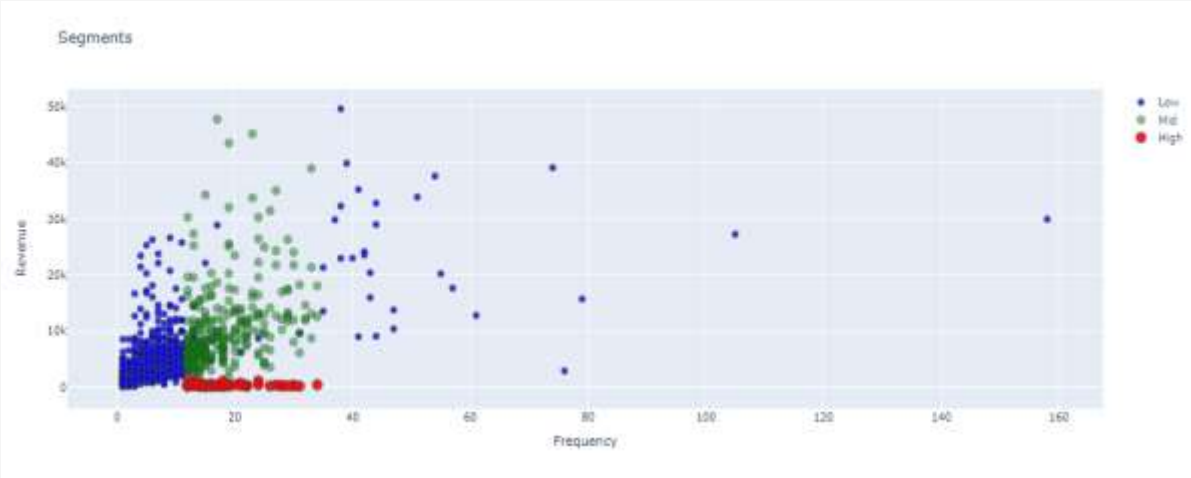
```
Segment
High-Value     2.821960
Low-Value     82.298615
Mid-Value     14.879425
Name: customer_id, dtype: float64
```

Labels	CostumerID Count
High-Value	2.82
Low	82.29
Medium	14.87

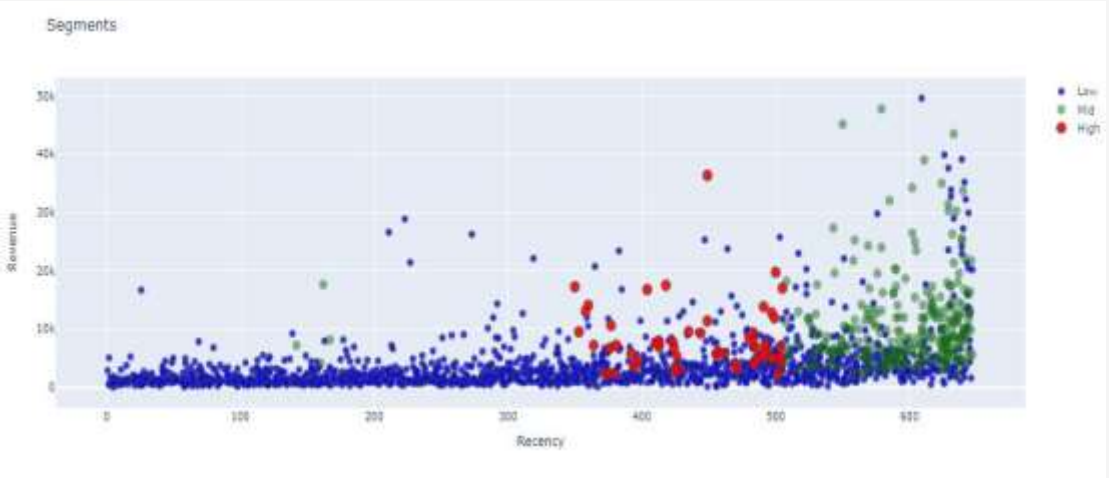
Clustering with K-Means

- ✓ For high-value customers, we know that these customers buy less frequency but are high revenue purchasers. They also haven't bought any products lately. Therefore we need to inquire if they are sleeping or dead customers. They could also be groups that purchase products based on certain period gaps, for instance, customers who shop based on season or customers who shop based on quarterly bonuses.
- ✓ For our mid-value customers, they haven't bought recently but have a considerable range with respect to frequency and revenue. They could be potential brand loyalists. They could also be high revenue generators that either buy large quantities or have an affinity towards expensive items. We need to dig a little deeper into this but overall need to increase their recency.
- ✓ For our low value, we know that they buy less frequency and have low revenue purchases but they have sporadic buying patterns, which is an attribute the business could leverage and improve.

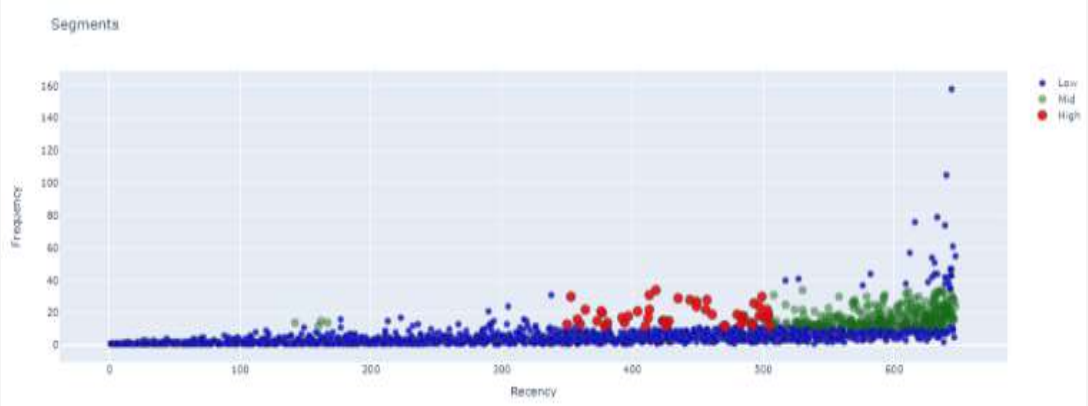
Modelling



Frequency vs Monetary



Recency vs Monetary



Frequency vs Recency

Model Evaluation

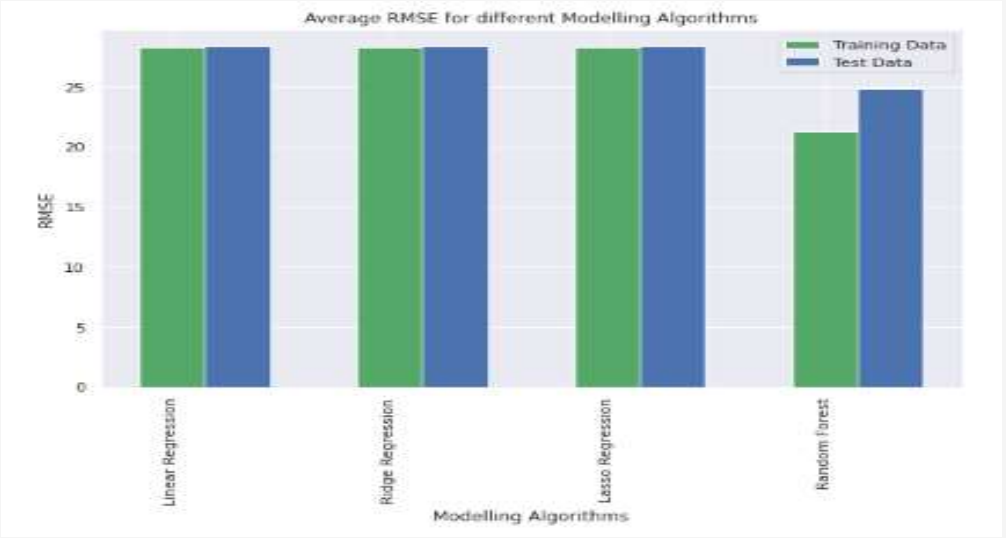
Results | Interpretation | Insights

Sales Forecasting using Machine Learning Algorithm

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

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

Model Comparison



Average RMSE for different Modelling Algorithms

On the test dataset, Random Forest exhibits the best performance. In the test data, Random Forest yields an RMSE of 24.84. Therefore, we have settled on Random Forest as our chosen algorithm.

Results and Insights

Key Findings | Suggestions

✓ 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
- Recency with Frequency

✓ Now we create individual strategies for these brackets and sometimes different strategies within one major bracket. Our strategies based on our analyzed hypothesis is as follows:

- For high-value customers, we know that these customers buy less frequency but are high revenue purchasers. They also haven't bought any products lately. Therefore we need to inquire if they are sleeping or dead customers. They could also be groups that purchase products based on certain period gaps, for instance, customers who shop based on season or customers who shop based on quarterly bonuses.
- For our mid-value customers, they haven't bought recently but have a considerable range with respect to frequency and revenue. They could be potential brand loyalists. They could also be high revenue generators that either buy large quantities or have an affinity towards expensive items. We need to dig a little deeper into this but overall need to increase their recency.
- For our low value, we know that they buy less frequency and have low revenue purchases but they have sporadic buying patterns, which is an attribute the business could leverage and improve.

✓ The objective of this paper was to predict sales for each item in a month. The model we came up with gives us decent results with RMSE of 24.84 on test data.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Conclusion and Future Work

Proposed solutions | Scope for future work

- ✓ 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.

- ✓ In the future, we can combine the customer segmentation model with the customer lifetime value model to predict the value of customers in each segment. This approach can also help us with segmented targeting i.e. identifying which of the inactive or sleeping customers should be targeted with a reactivation trigger, or which of the average customers can be pushed into the loyal customer category, etc. We can further 21 perform a/b testing, surveying, etc. to understand the feasibility of our recommendations through quantitative metrics.

REVA UNIVERSITY

Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

IIMB
तेजस्वि नावधीतमस्तु

भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

DCAL
DATA TO DECISIONS

Thank You!