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

REVA Academy for Corporate Excellence (RACE)

Prediction of Customer Lifetime Value in E-Commerce Business

MAHAPARA GAYASUDDIN

SRN: R19DM004

Date: 27-08-2022

MBA in Business Analytics

Capstone Project Presentation
Year: II

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

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.

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,2021	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,2020	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	This paper shows to determines 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.

Problem Statement

Business Problem | Analytics Solution

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 customers and end up attempting to solve the following problems:

- How much should they invest in bringing in fresh business to sell their product?
- How do various customer groups interact with a product?
- How successful is their product in the real world?
- How can they increase demand for their goods while maximizing profits from each customer?

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 have been applied to the case study in the industry to make this estimation.

Problem Statement

Business Problem | Analytics Solution

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. 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 behaviour of its users, who will bring in the most money over the course of a month using the statistics at hand. 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.

According to the 20:80 rule of marketing, 20 percent of a company's customers account for 80 percent of its earnings, and keeping an existing customer is far less expensive than finding new ones. The true problem is to find these customers and turn them into partners.

Project Objectives

Primary & Secondary Objectives | Expected Outcome

- The importance of CLV in the Retail and E-commerce Industry has grown in recent years, and it has become a requirement for an organization to keep track of their customers and calculate their lifetime value in order to understand their business prospects.

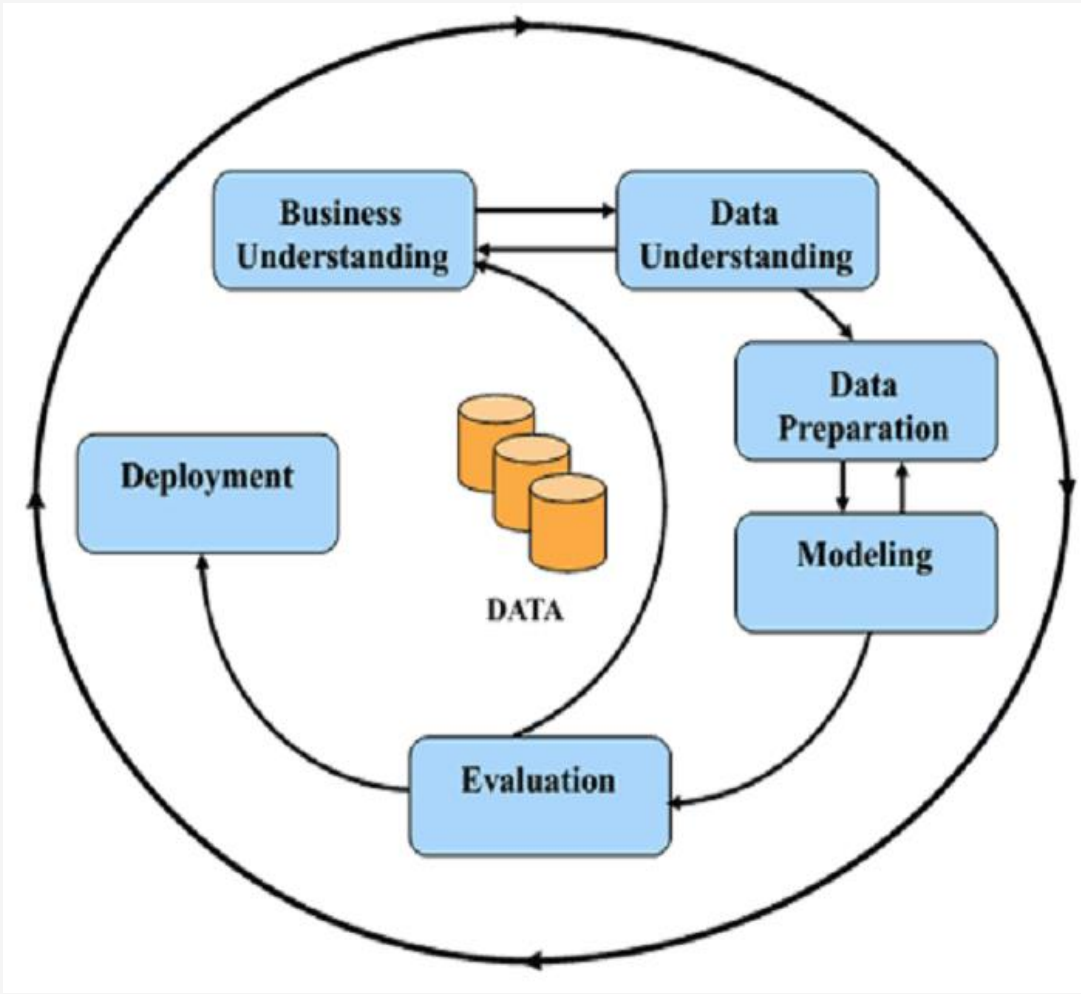
The study's objectives would be as follows:

- 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.
- To understand how a cohort behaves across time and compare it to 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.
- 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.
- To predict item-wise sales of the retail store using machine-learning algorithms.

Project Methodology

Conceptual Framework | Research Design

- **Understanding Business Requirements** — What is the specific ask from the business? Convert the same into a measurable and specific goal and formalize it as a problem statement.
- **Data Understanding** — This phase revolves around data gathering, exploration and comprehension.
- **Data Preparation** - This phase includes final data set selection, Cleaning, Transforming data etc.
- **Data Modeling** - Modelling is the heart of data analytics. One can think of a model as a black box which takes relevant data as input and gives an ML model as an output.
- **Model Evaluation** - The essence of model evaluation is dependent on the work that has happened in the previous 4 steps. If the results obtained from the model evaluation are not satisfactory, we reiterate the whole process, otherwise, we can move on to the implementation of the model. Evaluation is necessary to ensure that your model is robust and effective.
- **Deployment** - After model evaluation, the final model should go through thorough testing and then it should be deployed



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 predicted CLV 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.

Data Understanding

Data Collection | Variables

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

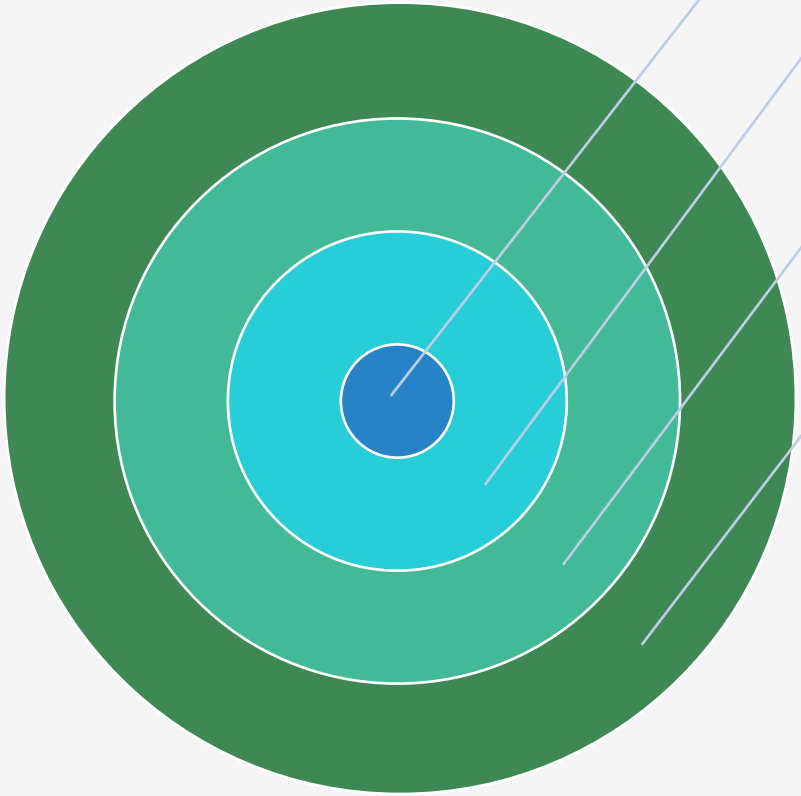
- Customer data
- Invoice data
- Customer transaction data
- Product purchased by the customers and the respective Quantity and amount

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



The columns "Description" and "CustomerID" contain null values. A unique CustomerID should be associated with each Invoice Number . Here the null values have been discarded, since it was difficult to discover the linkage.

Records with negative order quantities and monetary values were eliminated. These were cancelled orders.

For different transactions, the same item's unit price varies.

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

Data Preparation

Feature Engineering

- Only the fields required by the probabilistic model are kept i.e. Customer ID, Invoice Number, Invoice Date, Quantity, and Unit Price.
- Frequency is a measure of how frequently a client has made a repeat purchase. This indicates that there have been fewer purchases overall by one.
- 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.
- Recency is the customer's age at the time of their most recent purchases. This is the amount of time between a customer's first and most recent purchases. (Therefore, the Recency is zero if they have only made one purchase.
- For the final column, here the Totalprice (which is Quantity multiplied by Price) is included.

Train-test split

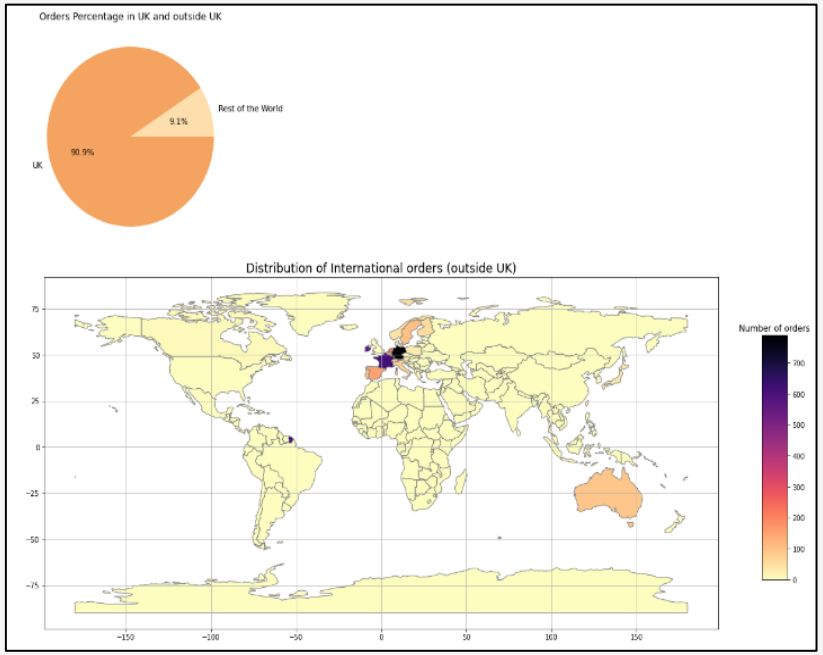
In order to prepare the data for training the model, a threshold date had to be chosen. That date divides the orders into two parts:

- 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.

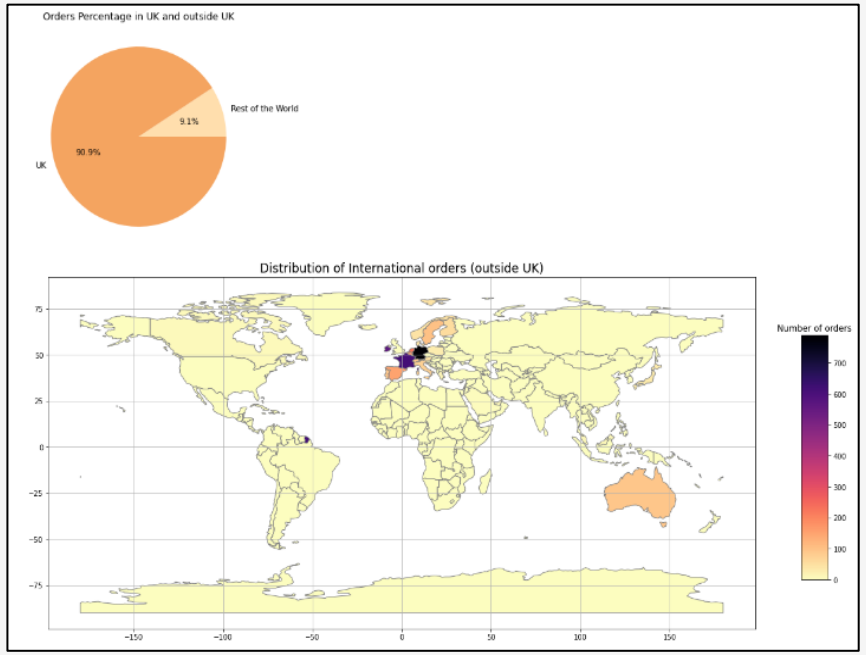
Descriptive Analytics

Multivariate Analysis | Hypothesis

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.



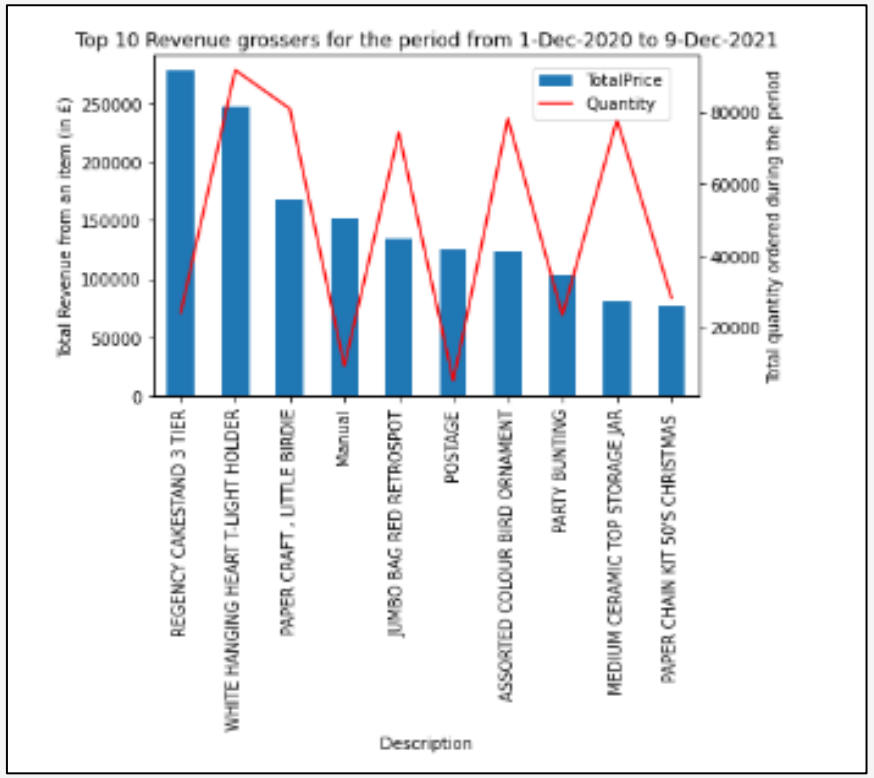
Orders Percentage in the UK and the Outside UK



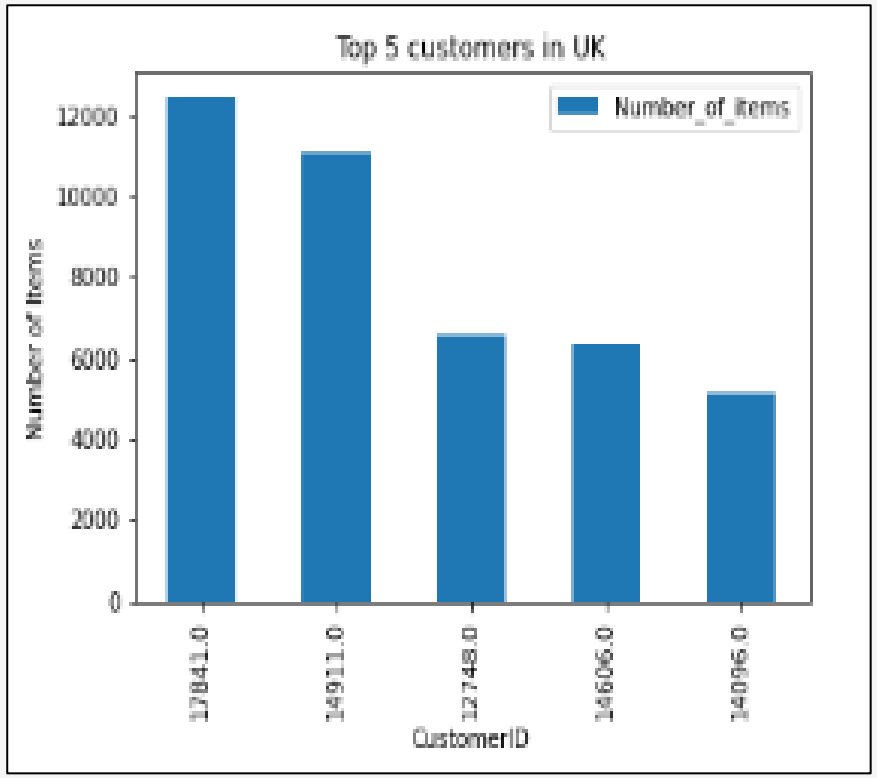
Customers Percentage in the UK and the Outside UK

Descriptive Analytics

Multivariate Analysis | Hypothesis



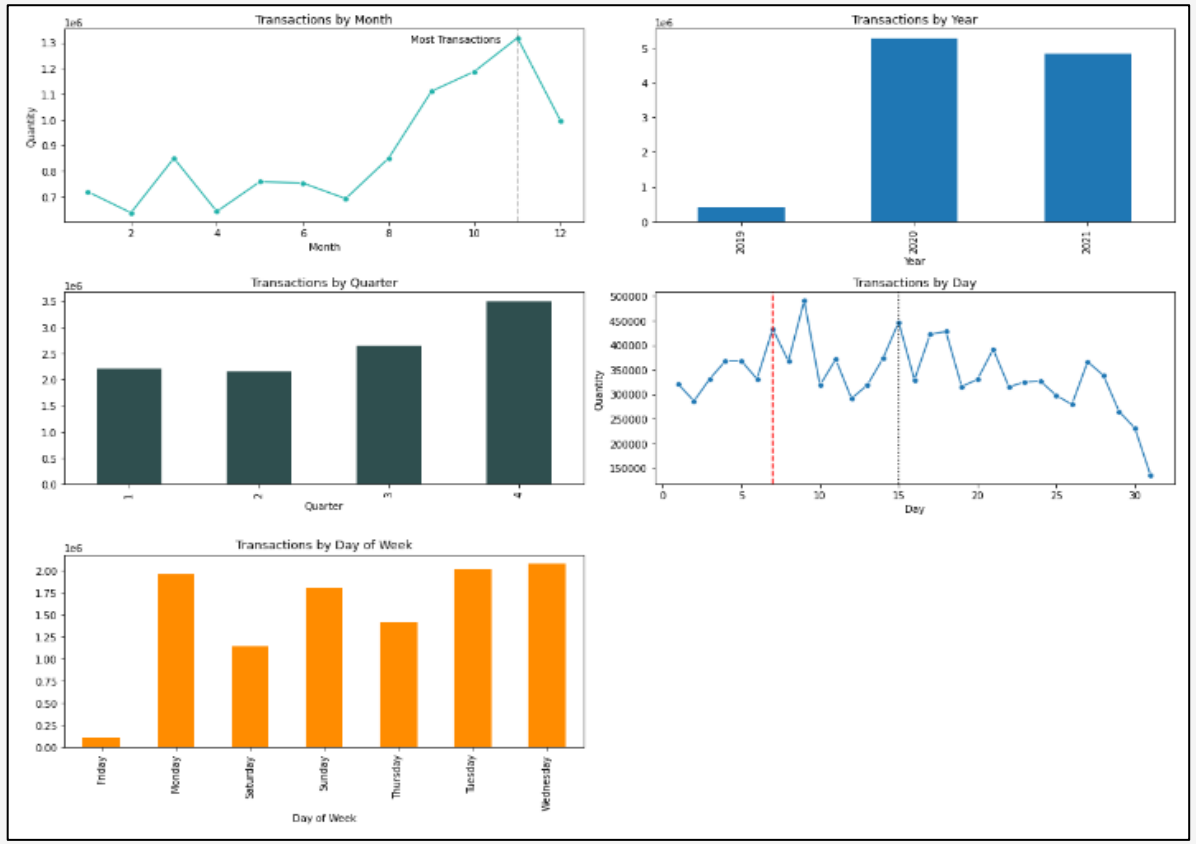
Top 10 Revenue Grosser Items



Top 5 Customers in the UK

Descriptive Analytics

Multivariate Analysis | Hypothesis

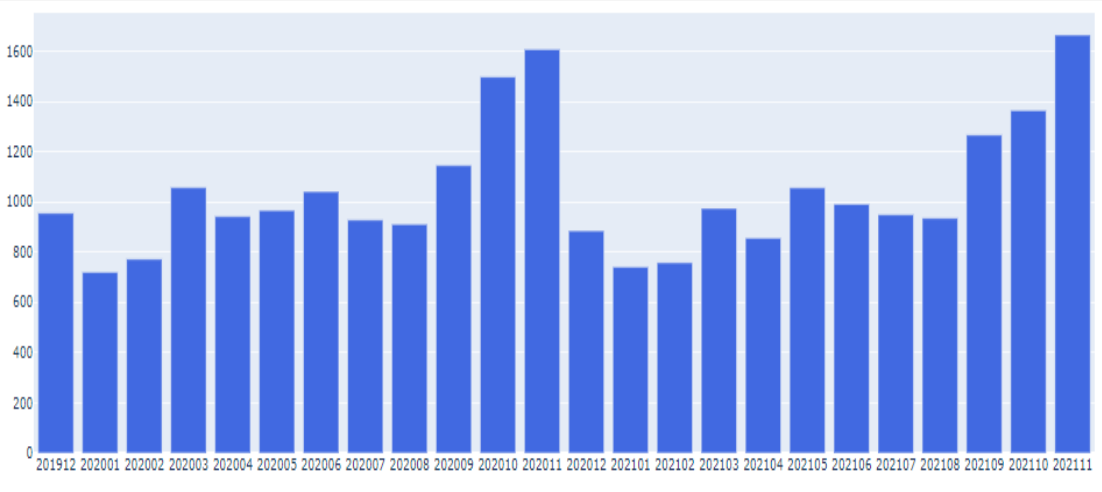


Date Time Analysis

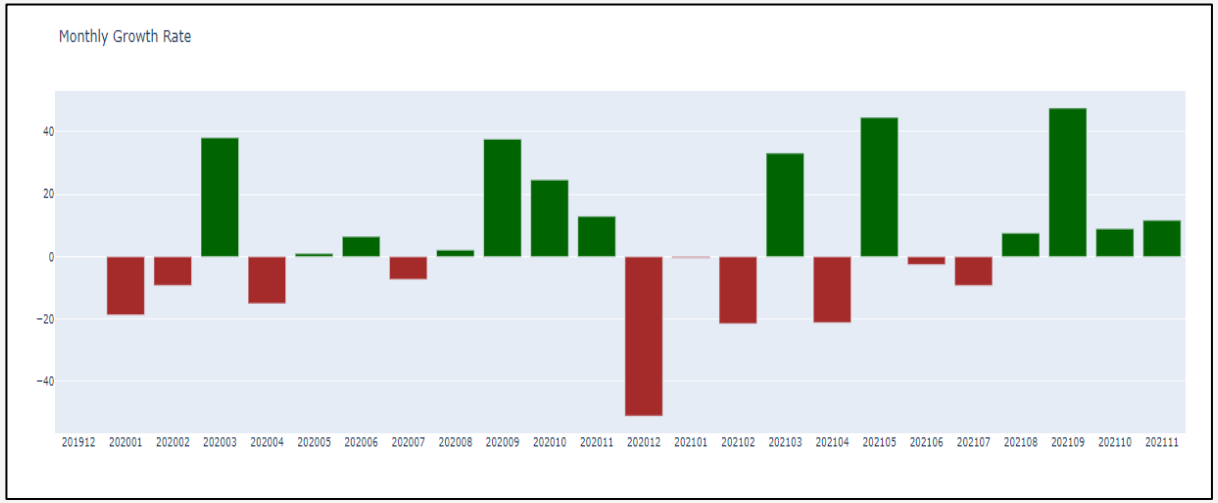
1. Due to the holiday season, November was the month with the most transactions.
2. 2020 and 2021 are the two years with the greatest trades, respectively.
3. The biggest number of transactions occurred in Q4.
4. It was also noted that consumers tend to make more purchases around the end of the first week and the beginning of the third week.
5. Wednesday is the day that people love to shop, followed by Tuesday and Thursday.

Descriptive Analytics

Multivariate Analysis | Hypothesis



Monthly Active Customers



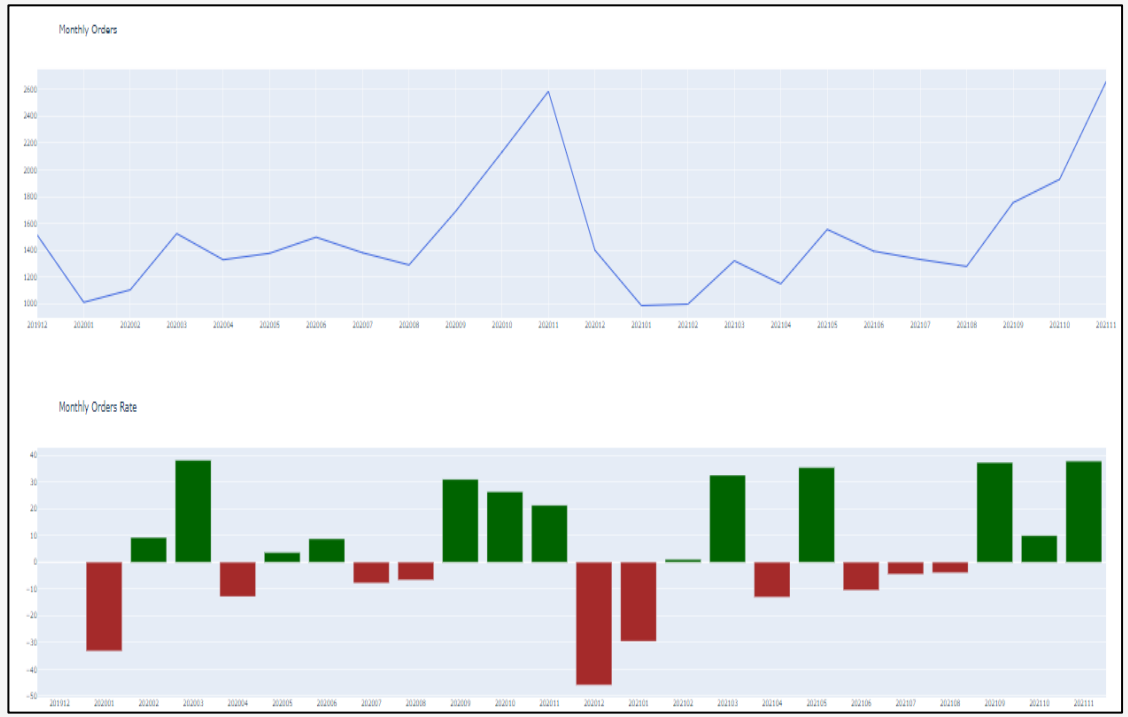
Monthly Growth Rate

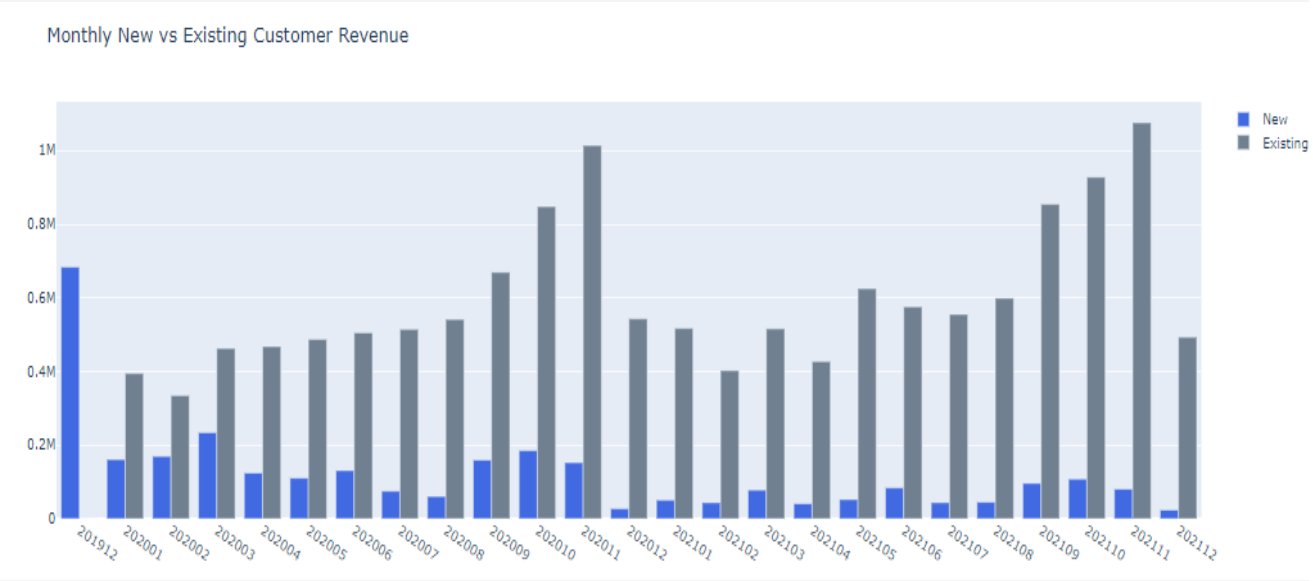
The company lost over 144 customers in January 2021, dropping from 885 in December 2010 to 741 in January 2021, a -18.27% fall. Similar to this, in April 2021 the company lost -12.11% of its customers, going from 974 to 856. We can see the maximum number of active customers in Nov'21 i.e. 1664.

The figure 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%.

Descriptive Analytics

Multivariate Analysis | Hypothesis





Monthly Order Count

The figure 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.

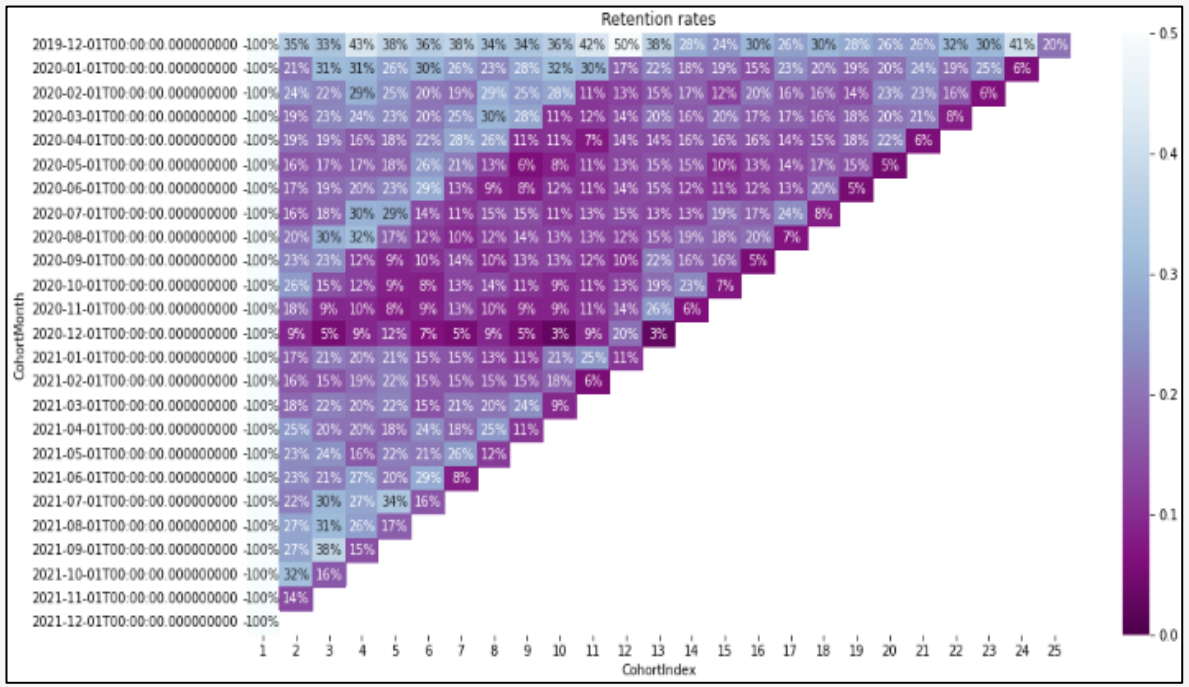
Monthly New vs Existing Customer Revenue

We can see that the number of existing customers is more compared to the new customers i.e. 1.07M in Nov'21 compared to the new customers i.e. 107.95k.

Descriptive Analytics

Multivariate Analysis | Hypothesis

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.

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.

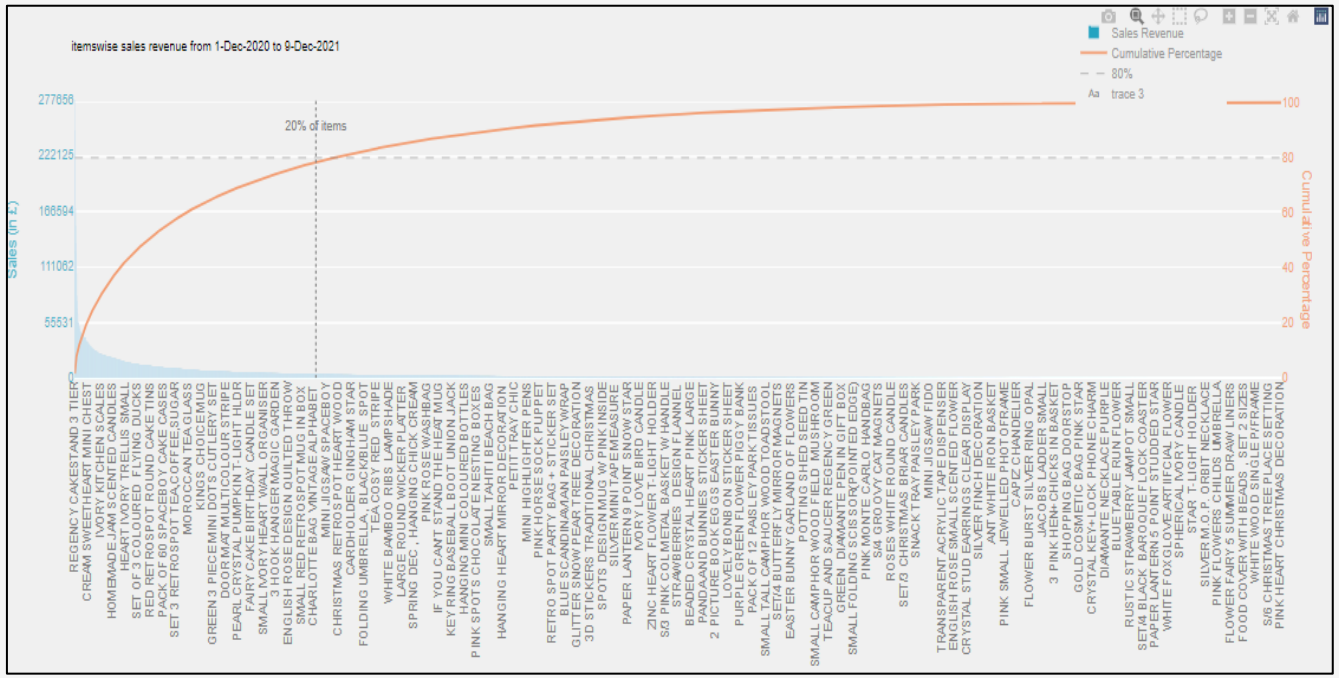
23 % of customers contribute to 80% of sales 23 % means 1352 customers
23

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

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.

Modeling

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 leave the system in the future.

2. Determine the financial value of each client using Gamma-Gamma Model.

- 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.

	CustomerID	frequency	recency	T	monetary_value
0	12346	7.0	401.0	726.0	11066.637143
1	12347	7.0	402.0	404.0	615.714286
2	12348	4.0	363.0	438.0	449.310000
3	12349	3.0	571.0	589.0	1120.056667
4	12350	0.0	0.0	310.0	0.000000

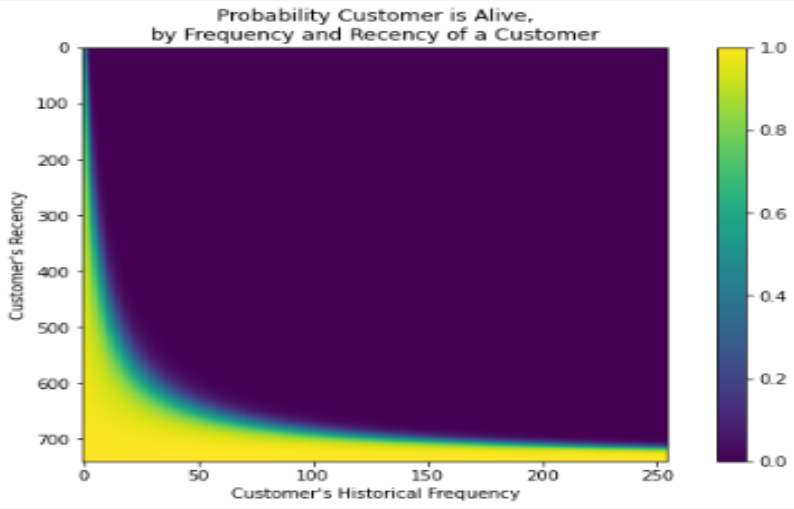
RFM Summary

RFM analysis is used to quantitatively determine the recency of purchase done by the customer, purchasing frequency of a customer and the expense done by the customer (monetary) from this, probability of customers who are surely alive, can be predicted. Customers having very high frequency and very high recency are likely to be the best customers in future.

Modeling

Modeling Techniques | Modeling Process | Model Building

From the figure, 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.



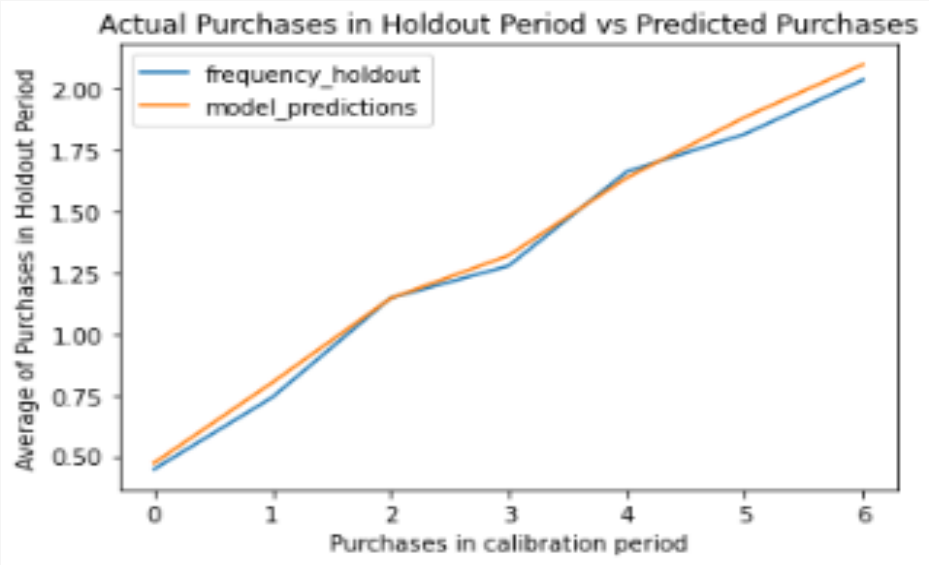
Customers who have recently made a purchase are nearly certainly still "alive". Customers who have made numerous purchases in the past but not recently are probably no longer there. And the more they had previously purchased, the more probable it was that they would stop. They are shown in the upper-right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

Modeling

Modeling Techniques | Modeling Process | Model Building

	CustomerID	frequency	recency	T	monetary_value	p_not_alive	p_alive	predicted_purchases
0	12346	7.0	401.0	726.0	11066.637143	0.465187	0.534813	0.155209
1	12347	7.0	402.0	404.0	615.714286	0.000467	0.999533	0.485818
2	12348	4.0	363.0	438.0	449.310000	0.025307	0.974693	0.270791
3	12349	3.0	571.0	589.0	1120.056667	0.003551	0.996449	0.168606
4	12350	0.0	0.0	310.0	0.000000	0.167712	0.832288	0.048093
...	...	...	...	...	...	...	...	...
5873	18283	18.0	656.0	659.0	142.666667	0.000541	0.999459	0.766602
5874	18284	0.0	0.0	431.0	0.000000	0.222904	0.777096	0.034122
5875	18285	0.0	0.0	661.0	0.000000	0.309941	0.690059	0.020808
5876	18286	1.0	248.0	724.0	833.480000	0.201957	0.798043	0.052165
5877	18287	5.0	529.0	571.0	622.276000	0.009873	0.990127	0.264345

Pareto-NBD Model Result




```
#dividing our dataset into training & holdout
pareto_summary_cal_holdout = calibration_and_holdout_data(data, "CustomerID", "InvoiceDate",
                                                          calibration_period_end = '2021-06-08',
                                                          observation_period_end = '2021-12-09')
```

Dividing the Dataset into Train-Test Split

In this plot, we separate the data into both an in-sample (calibration) and a validation (holdout) period. The sample period consists of the beginning of 2021–06–08; the validation period spans from 2021–06–09 to 2021–12–09. The plot groups all customers in the calibration period by their number of repeat purchases (x-axis) and then averages over their repeat purchases in the holdout period (y-axis). The orange and blue lines present the model prediction and actual result of the y-axis respectively. As we can see, our model is able to very accurately predict the customer base's behavior out of the sample.

Modeling

Modeling Techniques | Modeling Process | Model Building

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.

```
ggf_filter["predicted_clv_1month"] = ggf_model.customer_lifetime_value(pareto_model,
                                                                       ggf_filter["frequency"],
                                                                       ggf_filter["recency"],
                                                                       ggf_filter["T"],
                                                                       ggf_filter["monetary_value"],
                                                                       time = 30,
                                                                       freq = 'D',
                                                                       discount_rate = 0.01)
```

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 present value of a future cash flow is obtained by applying a discount rate to the future cash flow's monetary value. It is stated in the documentation that it is 0.01 on a monthly basis (or 12.7% annually).

Modeling

Modeling Techniques | Modeling Process | Model Building

Top 5 customers with high CLV

	CustomerID	predicted_clv_12months
4061	16446	9.733925e+06
5692	18102	1.408517e+06
2277	14646	1.265768e+06
189	12536	9.602072e+05
5050	17450	7.856863e+05

Finally, the Customer Lifetime Value for Non-contractual business has been predicted using the Pareto-NBD Model.

CustomerID	frequency	recency	T	monetary_value	p_not_alive	p_alive	predicted_purchases	actual30	error	expected_avg_sales_	predicted_clv_1month	predicted_clv_12months	profit_margin
12346	7	401	726	11066.63714	0.465187415	0.534813	0.15520923	0.523691	0.368482	12504.44487	35772.18377	99428.87241	1788.609189
12347	7	402	404	615.7142857	0.00046689	0.999533	0.485818254	0.522388	0.03657	695.8518605	7706.961492	20914.65085	385.3480746
12348	4	363	438	449.31	0.025306736	0.974693	0.27079086	0.330579	0.059788	562.7753882	3444.569575	9373.182575	172.2284788
12349	3	571	589	1120.056667	0.003550744	0.996449	0.168606403	0.157618	-0.01099	1531.122915	5935.296147	16338.01071	296.7648073
12352	8	356	392	338.26125	0.012356667	0.987643	0.556059764	0.674157	0.118098	376.2268987	4742.312084	12856.83462	237.1156042
12353	1	204	408	89	0.083804287	0.916196	0.099218143	0.147059	0.047841	460.7517958	1018.326606	2764.36292	50.91633029
12355	1	353	567	459.4	0.063870069	0.93613	0.076195436	0.084986	0.00879	2358.404253	4012.790137	11028.05502	200.6395069
12356	5	402	424	1092.258	0.005680928	0.994319	0.343961622	0.373134	0.029173	1302.020976	10199.72989	27723.98635	509.9864945
12357	1	355	388	6207.67	0.008391406	0.991609	0.112043796	0.084507	-0.02754	31808.23873	81126.55278	219869.5806	4056.327639
12358	4	731	732	614.31	0.000162635	0.999837	0.176355242	0.164159	-0.0122	769.3359693	3144.781603	8744.507148	157.2390802
12359	9	678	735	901.4111111	0.01366539	0.986335	0.356214277	0.39823	0.042016	990.048421	8087.800417	22493.85886	404.3900209
12360	7	604	656	590.6985714	0.012422337	0.987578	0.314113805	0.347682	0.033568	667.5863769	4802.019205	13282.40108	240.1009603
12361	3	396	683	134.0166667	0.137745572	0.862254	0.127766156	0.227273	0.099507	183.5757891	502.9719063	1393.86546	25.14859531
12362	10	736	739	522.623	0.000493543	0.999506	0.396077251	0.407609	0.011531	568.4672422	5215.246975	14508.6322	260.7623488

Modeling

Modeling Techniques | Modeling Process | Model Building

Customer Segmentation Using Predicted CLV

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

Overview of the Customer Segment

Labels	CostumerID Count
High-Value	0.023872
Low	96.228217
Medium	3.747911

Customer Segmentation Using K-Means

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 needs 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 behavior, which the company may capitalize on and enhance.

Modeling

Market Basket Analysis Using Apriori Algorithm

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
489435	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
489437	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
581582	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
581584	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
581586	0	0	0	0	0	0	0	0	0	0	...	0	0	0

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.

Modeling

Modeling Techniques | Modeling Process | Model Building

	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

MBA Using Apriori Algorithm

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.

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.

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.

Modeling

Collaborative Filtering based systems

User-based Collaborative Filtering

```
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

The figure shows items that can be recommended to B based on the preferences of A are:

Item-based Collaborative Filtering

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

                                Description
StockCode
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
```

The top 10 Similar items to recommend

Model Evaluation

Results | Interpretation | Insights

Customer Lifetime Value For Non-Contractual Business using the Probabilistic method

	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

Model Comparisons of different Probabilistic Methods

The goal is to focus on empirical statistical analysis and predictive abilities of selected probabilistic CLV models that show very good results in an online retail environment compared to different model families. For comparison, eleven CLV models were selected. Probabilistic models have achieved overall good and consistent results on the majority of the studied transactional datasets, with BG/NBD, Pareto/NBD, and MBG-NBD 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. So, will go ahead with the Pareto-NBD Model for the prediction of Customer Lifetime Value for non-contractual relations in online retail and its segmentation.

Model Evaluation

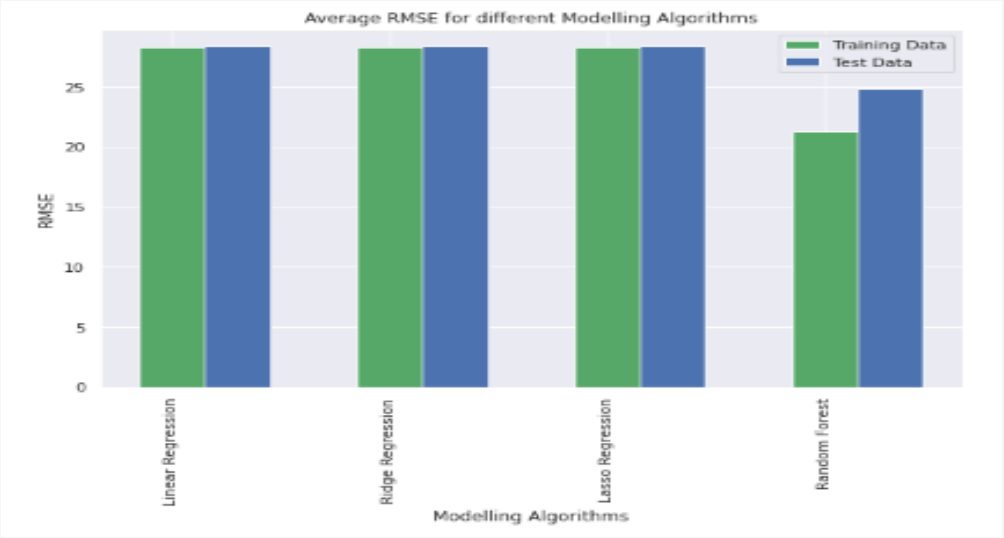
Results | Interpretation | Insights

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

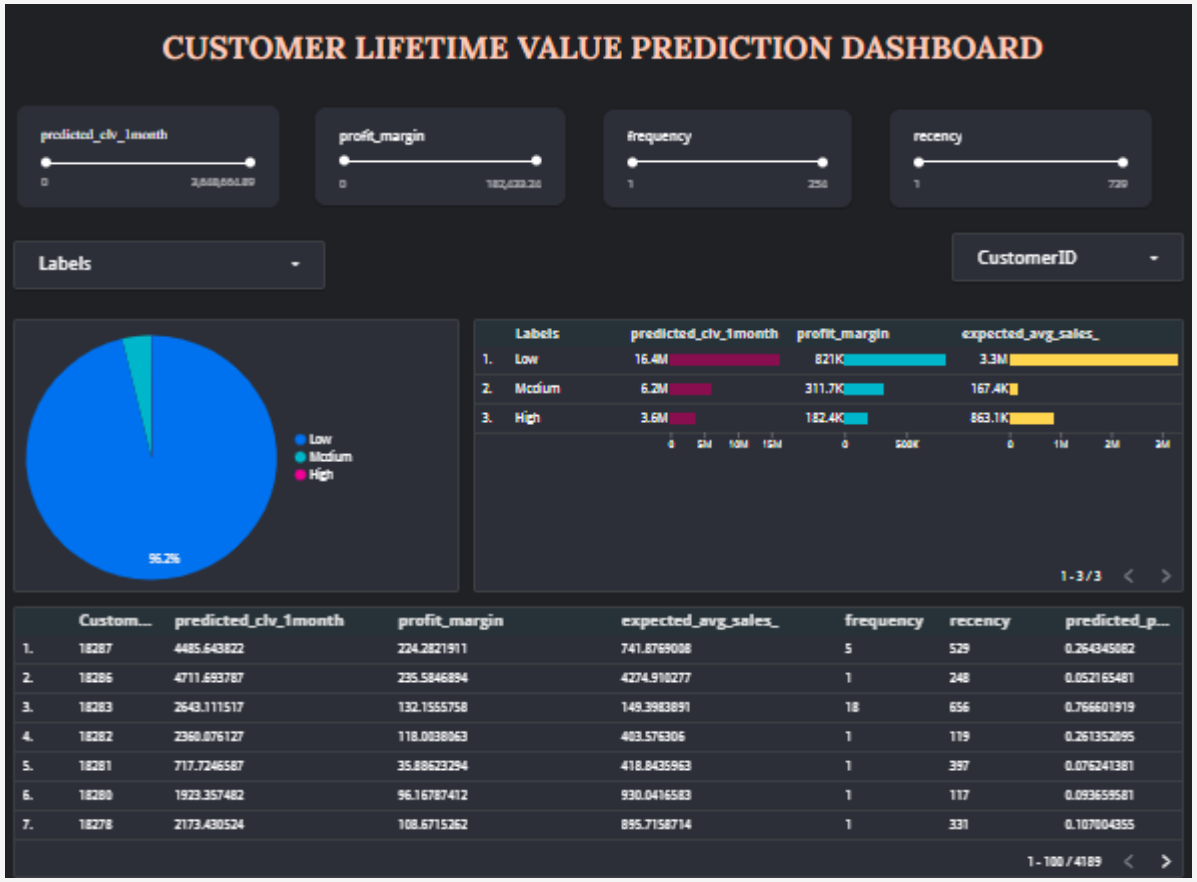


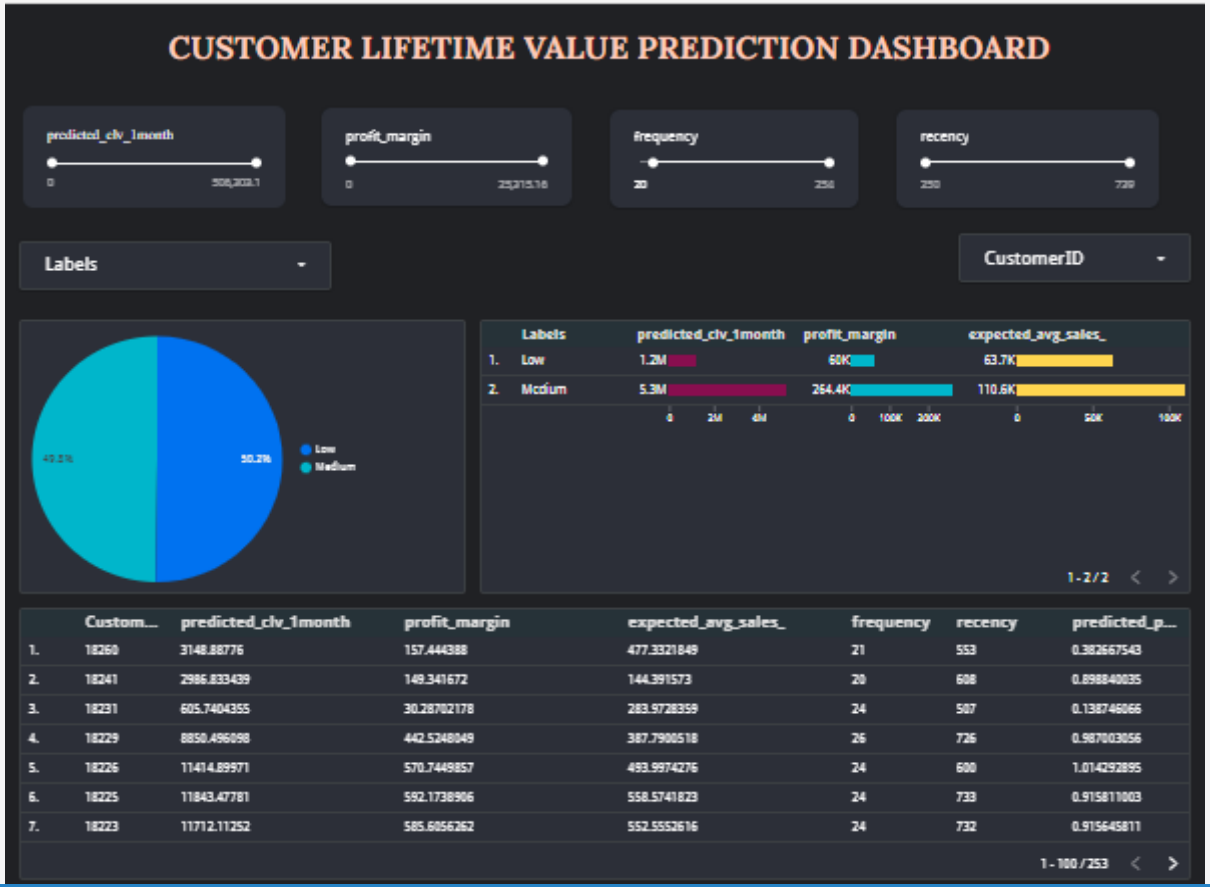
Average RMSE for different Modelling Algorithms

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.

Dashboard

A Dashboard for the prediction of Customer Lifetime value has been prepared using Google Data Studio as shown below in Figure. It is prepared to explain the most crucial key performance indicators (KPIs) for the company. The marketing team can use it to track and keep track of the marketing performance so can decide which area to concentrate on using data.

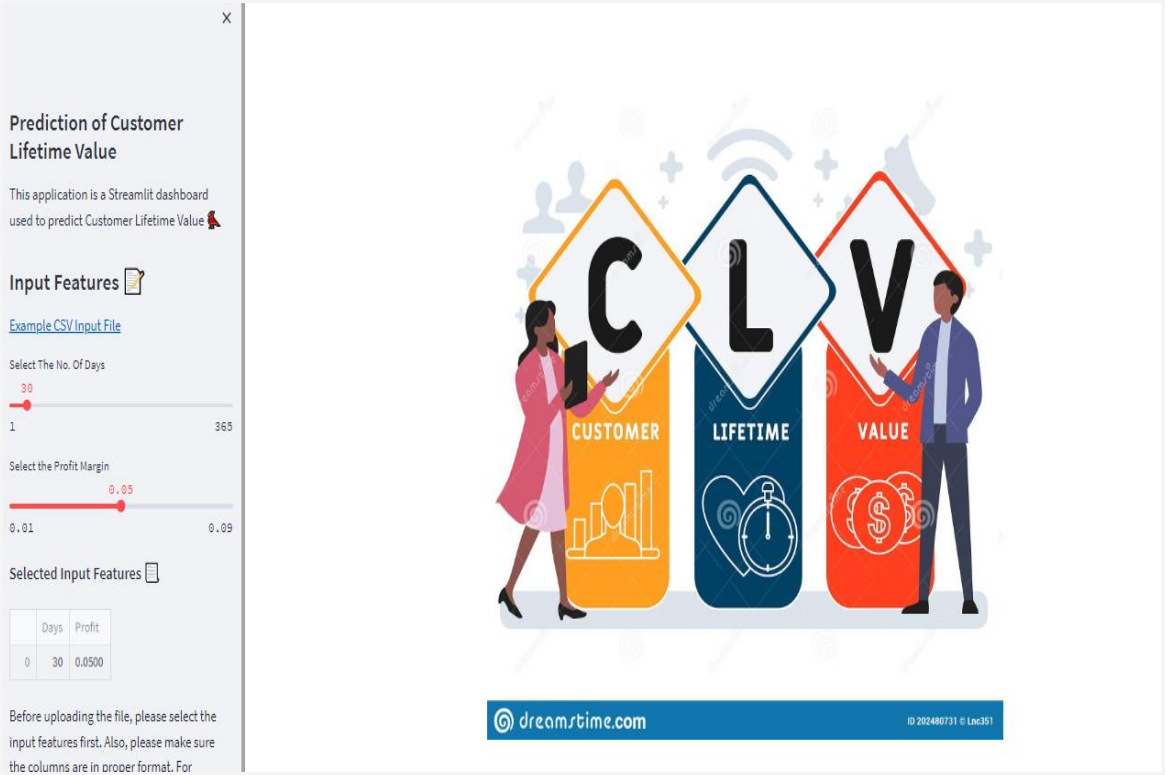


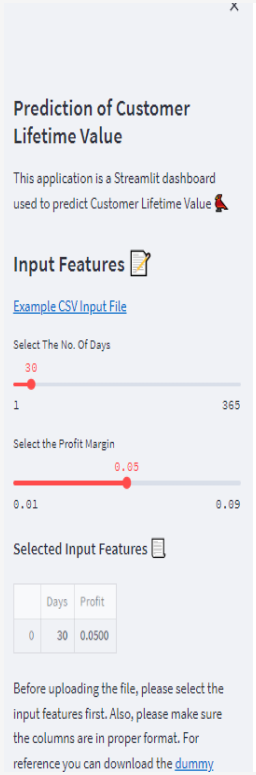


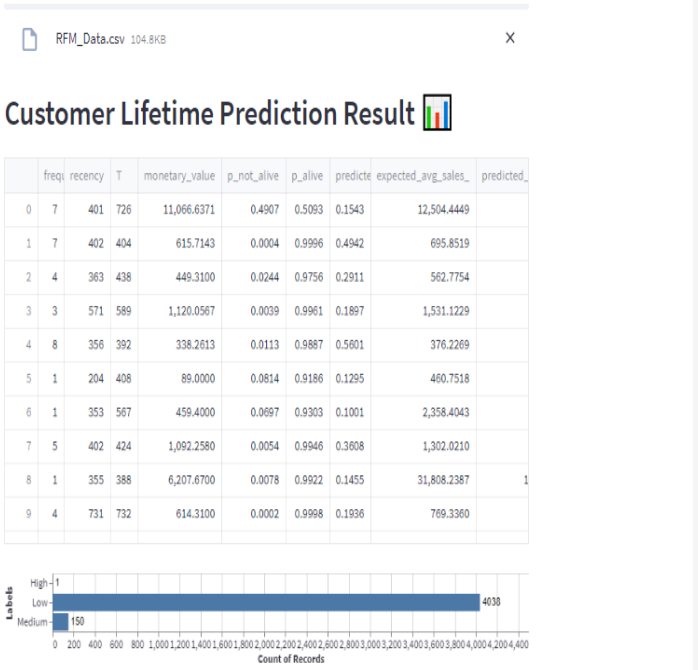
Model Deployment

Demonstration

The Customer Lifetime value prediction model has been deployed in ***Streamlit***. We just need to upload the RFM data to get the customer lifetime value.







Results and Insights

Key Findings | Suggestions

- We have made the comparison between three probabilistic models i.e. Pareto/NBD, BG/NBD, and MG/NBG models, and validated them through their MSE and RMSE. The Pareto NBD model is seen to perform a little bit better when it comes to minimizing the MSE & RMSE Errors. So, we have gone ahead with the Pareto-NBD Model for the prediction of Customer Lifetime Value for non-contractual relations in online retail and its Segmentation.

- This study aims to analyze the Customer Lifetime Value (CLV) of the customer and cluster it into customer segmentation using the K-means cluster method. The results showed the highest average CLV value is of CustomerID "**16446**" with a CLV of **£3648664.89**. The customer clustering produced in this study is three segments with the majority of customers at segment Low.

- The Consumer Purchase Pattern shows that the application of the Apriori algorithm is very good to be implemented because it can produce a purchase pattern based on the transaction data used in this study. It is expected that the results of consumer purchasing patterns can help online retail managers in making decisions to get even better profits.

- The Collaborative recommendation model helps in recommending the products to user B based on the interests of a similar user A.

- The item-wise sales for an online retail store have been predicted using different machine learning methods, among them the Random Forest yields an **RMSE** of **24.84**. Therefore, we have settled on Random Forest as our chosen algorithm.

Conclusion and Future Work

Proposed solutions | Scope 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.
- With Low segmentation, customers dominate the outcomes of customer CLV analysis. For customers in the low segment, the approach should be centred 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. 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, we can conduct the same study in other industries like insurance, Banking or telecommunication industry and be able to compare the results in various industries. I propose that we should take into account additional variables not covered in this study as well as additional qualitative elements affecting the industry's CLV. To carry out the identical investigation using the AHP technique and evaluate the outcomes.

References

Bibliography | Webliography

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

Annexure

Additional Information | Plagiarism score

Prediction of Customer Lifetime Value in E-Commerce Business

ORIGINALITY REPORT

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

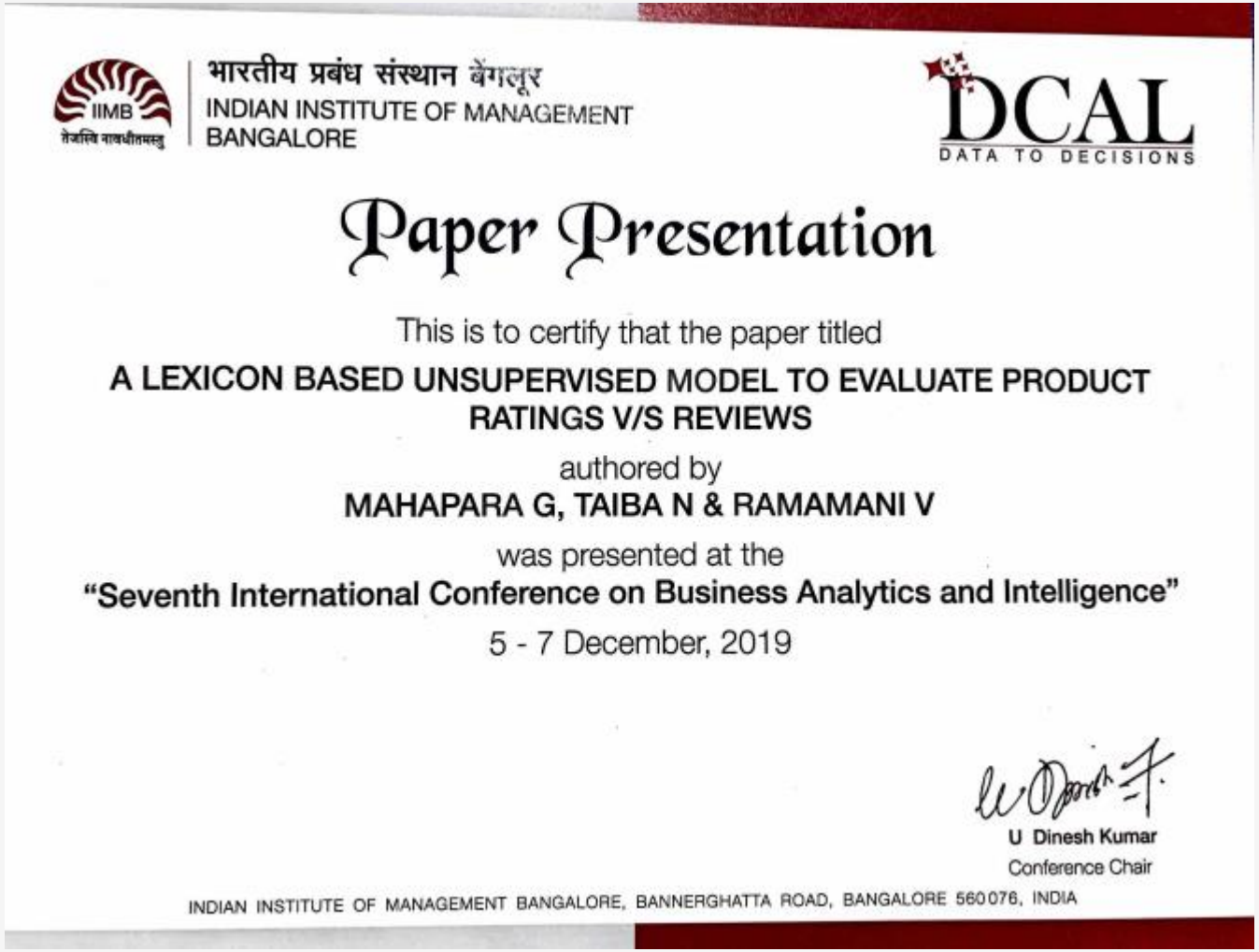
PRIMARY SOURCES

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 bibliography On
Exclude matches < 10 words

Annexure

Publications | Conferences



भारतीय प्रबंध संस्थान बेंगलूर
INDIAN INSTITUTE OF MANAGEMENT
BANGALORE

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

REVA
UNIVERSITY

Bengaluru, India

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

Thank You!