National College of Ireland

Predicting Sales and Analysing Customer Lifetime Value (CLV) in the E-Commerce Industry Using Machine Learning Methods

MSc Research Project
MSc in FinTech

SAFA MASOOD
Student ID: x22186506

School of Computing
National College of Ireland

Supervisor: Faithful Onwuegbuche

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	SAFA MASOOD
Student ID:	x22186506
Programme:	MSc in FinTech
Year:	2024
Module:	MSc Research Project
Supervisor:	Faithful Onwuegbuche
Submission Due Date:	16/09/2024
Project Title:	Predicting Sales and Analysing Customer Lifetime Value (CLV) in the E-Commerce Industry Using Machine Learning Methods
Word Count:	6843
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Safa Masood
Date:	15th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Sales and Analysing Customer Lifetime Value (CLV) in the E-Commerce Industry Using Machine Learning Methods

SAFA MASOOD
x22186506

Abstract

The advancement of e-commerce has brought a major change in the business sector by allowing companies to connect with customers globally and offer different products and services. To manage an e-commerce business prediction of sales and calculating the Customer Lifetime Values are the pivotal components in formulating the strategy, customer differentiation, and utilization of resources. However, the data produced by e-commerce such as customer interactions, buying and selling transactions, and market forces of demand and supply form a real challenge to extract meaningful information and intelligence. Another challenge is the heterogeneity of the customers as e-commerce firms are dealing with customers from different backgrounds. It is important to anticipate such conduct for customizing the marketing, maintaining the clients, and optimization of supplies. However, it is crucial to find such customers who are worthy of constant revenue generation called Customer Lifetime Values (CLV). To address these challenges, the research focuses on analyzing the suitability of various regression algorithms in predicting sales and analyzing the customer lifetime value (CLV) in the context of e-commerce. The objective of this research includes the implementation of various regression-based machine algorithms for forecasting sales based on the Brazilian e-commerce dataset. After our analysis, we have identified Random forest as the most accurate model for predicting sales with a minimum error score. The accurate prediction of Sales and analysis of CLV in our research allows optimization of the acquisition of customers, retention, and overall profitability for the business.

1 Introduction

The E-Commerce landscape has dramatically altered the world of online shopping by providing a global marketplace with a range of products and services. Accurate predictions of sales and Customer Lifetime Value are important for a business to succeed in the competitive landscape. Accurate sales predictions help managers understand and make inventory management decisions, pricing strategies, and marketing campaigns, and help ensure companies are meeting consumer needs. Customer Lifetime Value reveals which customers are most valuable to an organization, and can guide companies toward targeted marketing strategies and provide personalized experiences that can help a company become more profitable. The analysis of sales predictions, and Customer Lifetime Value allow companies to use data to help make decisions that drive customer satisfaction and

organizational profitability. "Customer Lifetime Value" (CLV/CLTV) is the measure of approximate value, which is expected to be put forward by the customer to a business throughout a customer-provider relationship. In this regard, the accountability for factors continuously impacting this relationship varies indifferently where the money spent by a customer over the period, customer purchase behavior, likelihood of purchasing again, and rate of recommendations to others are critically contributing to CLV. It is evident from industrial experts that customer lifetime value is important in the global business landscape as it enables companies to identify and prioritize valuable and potential customers while allocating resources in the process of retaining and maximizing customers' values Paul (2023). While understanding this, an accurate prediction of customer lifetime value enables firms to make justifiable and informed decisions regarding the marketing process, sale improvement, and enhancing customer service strategies. The prediction of customers' future purchases and lifetime values are key metrics that help in managing different marketing campaigns as well as optimizing marketing trends.

Indeed, CLV prediction and evaluating sales are challenging tasks in terms of future purchases of customers, and critical improvement in the prediction process has been bestowed with the consideration of key aspects Pollak (2021). The prediction of CLV based on customer behavior on future purchases when evaluating the relationship between businesses and customers is non-contractual. In recent years, the e-commerce industry has received generous attention considering the revolutionary trend in business amid the situation of globalization and technological enhancement Laksono and Wulansari (2022). In the modern economy, e-commerce has been playing an essential role, which therefore attracts the attention of researchers to understand their contribution to business growth. In this competitive era, building a long-term customer relationship requires firms to maintain integrated service infrastructure and marketing tactics, thus summing up the "customer lifetime value" (CLV) and "corporate customer value" (CCV). Hence, the significant estimation of the CLV and CCP rate enables firms to make business decisions more accurately.

Indeed, the intensity of e-commerce growth and sales performance maintenance in the new platform are enhanced by CLV value. However, as stated earlier, the future prediction of customer purchases has been increasingly challenging for a wide range of factors among which continuous dependence on massive customer data and accuracy in the prediction of CLV are critical considerations. In this regard, extensive research has determined the contribution of various models that are often used including "recency monetary frequency" (RFM), probabilistic models, econometric methods, simple diffusion methods, and many more. However, not all models are efficient in accurately predicting the CLV value because of which dilemmas in customer retention and relationship establishment are identified. While understanding the limitations of existing models, recent approaches to machine learning models have shown improvement in the prediction process. ML-based prediction of CLV value largely depends on the feature selection process from databases where data complexity is a factor enhancing the prediction rate.

One such ML-based algorithm is the linear regression that enables CLV prediction based on considering small-scale features while the application of neural architecture needs more complex datasets. Upon highlighting the context, the current study utilized "Brazilian Public E-commerce Dataset" . Which consists of nearly 100k orders available at the Olist

Store where the range is identified from 2016-2018 Olist and Magioli (2018). It has been developed across multiple marketplaces available in Brazil. The features obtained from the dataset will enable us to view the order based on multiple dimensions, thus providing significant information that is useful in the prediction of sales performance and customer retention aspects in the e-commerce industry. Thus, the current study utilized 3 different machine learning models Linear regression, random forest, and XGBoost to predict the sales of e-commerce businesses. Each of these machine learning model performances will be evaluated by calculating Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). For years companies have been using extensive data that can support business decision-making. Hence, the application of this extensive dataset in the study provides a suitable consideration for its use to extract features for training machine learning models for prediction purposes.

1.1 Research Objectives

Our research will revolve around the following research objectives.

- To perform extensive Exploratory Data Analysis (EDA) to uncover patterns, trends, and insights within the dataset that can inform the development of predictive models.
- To develop predictive models for accurately forecasting sales in the eCommerce sector using machine learning algorithms, informed by insights gained from EDA.
- To analyze and compare the effectiveness of different machine learning models—XGBoost, Random Forest, and Linear Regression in predicting Customer Lifetime Value (CLV) and sales outcomes.
- To evaluate the performance of the predictive models using key metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to identify the most effective algorithm.

1.2 Research Question

We will address the following research question in our study.

- How can machine learning algorithms, informed by extensive Exploratory Data Analysis (EDA), be effectively utilized to predict sales and analyze Customer Lifetime Value (CLV) in the e-commerce industry?

2 Related Work

In the current chapter, evidence-based information from the literature has been established that provides insights into a similar prediction process in the e-commerce sector. The information established from studies has presented evidence on different models that have been used by experts in predicting customer lifetime value (CLV) within e-commerce business segments and how it has enabled companies to retain maximum profit. At the same time, the importance of datasets has been determined with the feature extraction process used by many studies for the predictive analysis process.

Author-Year	Key Findings/ Results	Advantages	Limitations
Laksono and Wulansari (2022)	A multivariate analysis has been performed to estimate the CLV where the RFM method has been used. As per the experimental outcome, the model provides highly accurate insights into the CLV values that are required to maintain customers in e-commerce companies.	The focused study has contributed knowledge to the utilization of big data by companies to make business decisions. It highlights the case of the e-commerce industry where customer purchasing behavior and their lifetime value are highly perceived to determine the size as well as the rate of higher transactions.	Limited evidence on the model efficiency in CLV prediction and typically focused on statistical analysis, which is complex and has estimation issues.
Platzer and Reutterer (2016)	As per the study implications, it is imperative to incorporate regularity within forecasting accuracy through extensive simulation and wide empirical applications.	The study provides information on customers' regularity and inter-purchase timing through a recency-frequency paradigm through which customer-based analysis has been performed.	The study provides a comprehensive review of the importance of statistical models such as the Pareto/NIBD model with necessary information on the simulation process. However, there is a lack of focus on a primary approach to determine the model accuracy.
Sun et al. (2023)	The study discloses knowledge of CLV and its importance to enterprises in managing customer relationships. As per the findings of experimentation, the ML model provides improved results in terms of CLV value under a non-contractual relationship.	Upon underlying the condition, the study provides useful insights into the feature engineering techniques, for example, data selection, pre-processing, and transformation using ML models.	It has been identified that limitations within a single data-mining process have provided ineffective CLV values for non-contractual relationships, thus posing research difficulties.
Chen (2018)	As per the findings obtained from the study, it has been observed that the model scores better than many existing techniques, thus helping airlines in finding potential customers.	Findings obtained from the study show the need to maintain high-quality customer retention resources through which competitiveness can be measured. In this approach, the study uses machine-learning models to prevent controversial dilemmas and ensure higher accuracy in CLV prediction.	The study's main focus is on the airline industry rather than e-commerce, thus, a need for further research to evaluate the model's accuracy in CLV prediction in e-commerce is mandatory.

Table 1: Summary of Literature Review

2.1 Sales and CLV Prediction Using Traditional Methods

The focus on the prediction of customer lifetime value (CLV) has been considered an essential task that holds significant value in estimating current as well as future purchasing intentions of customers. Across multiple industries, customer lifetime value (CLV) plays an important role in building and managing distinct customer relationships that can ensure business profit. Upon identifying the suitability, Platzer and Reutterer (2016) explained that it is imperative to utilize enhanced prediction methods that can positively emphasize an accurate estimation of the inter-purchase aspects of customers. In this regard, the "recency-frequency" paradigm focuses on customer-based analysis through the integration of regularity within the inter-purchase timing process with the modeling framework. In this regard, the above study introduces a renowned Pareto/NBD model that has been accounted for with a distinct level of regularity prediction across customers while replacing NBD components containing a "gamma distribution mixture". As Platzer and Reutterer (2016) explained, the technique has been used by industrial experts for many years, and seems to be effective; however, data requirements are a factor that deviates the outcome. Compared to this study, another study presented by Qismat and Feng (2023) has explained the relevance of RFM models where evidence-based information has provided comprehensive insights into the prediction accuracy of the Pareto/NBD model through which "recency-frequency" has been evaluated. As per the findings presented, it has been identified that the NBD model provides greater prediction accuracy compared to other techniques thus serving the effectiveness and efficiency levels of the method in estimating customer lifetime value (CLV).

In another study presented by Vanderveld et al. (2016), the information presented comprehensive insights into the importance of customer lifetime value (CLV) as a crucial part of business-customer engagement. The relationship that exists between companies and customers in the e-commerce industry has attempted to explore the long-term returns that can be achieved from this engagement. Upon understanding the value of customers in an e-commerce company, it is therefore imperative to perform an accurate prediction of CLV by leveraging suitable prediction models. In this regard, the information provided by Vanderveld et al. (2016) has presented the deployment of a novel framework - the "economic scoring model", whose effectiveness has been evaluated using a re-scoring process. This re-scoring process is based on various triggering events that often occur thus enabling a simultaneous scoring of the large customer base using a complex model. Upon understanding the suitability of the model, it can be stated that the deployment of the system has efficiently predicted the CLV of millions of customers, thus providing comprehensive insights into creating extensive business values as well as productive initiatives. In another study presented by Chamberlain et al. (2017), the author has described the importance of CLV prediction for companies. In this regard, the study introduces the contribution of this value prediction for a leading global fashion retail company. Understandably, the study uses state-of-the-art techniques that have applied an extensive domain of handcrafted features. Apart from this, an ensemble regressor has also been used to forecast CLV values, predict churn, and further evaluate customer loyalty. Although the process was extensive, it provides comprehensive insights based on feature representations as an extension of the CLV modeling. The findings established show that the method provided an improved result based on an exhaustive set comprising handcrafted features.

2.2 Sales and CLV Prediction Using Machine Learning Algorithms

In the above section, it has been identified that certain traditional methods have been used for years to predict the customer lifetime value (CLV), which is a key instrument for evaluating customer relationships for a company. Despite the methods being reliable and effective in the prediction process, most of them are time-consuming and require handcrafted features to efficiently manage the prediction accuracy. Therefore, the contemporary prediction process has identified the importance of sophisticated methods such as machine learning models. In this regard, the study presented by Norouzi (2024) informed that the accurate prediction of CLV has received paramount importance in optimizing customer relationships. The study introduces several "key performance indicators" (KPIs) such as NPS, ATV, and CES that promote a collective impact on customer lifetime value (CLV). Understandably, a neural architecture has been used that consists of dense layers that help in tuning and regularizing features. The model's prediction accuracy is estimated based on certain parameters, which shows that the prediction is highly efficient when optimized with NPS & CES values.

Curiskis et al. (2023) in their study has introduced a flexible method indicating a machine learning (ML) model that enables the prediction of CLV value for business-to-business (B2B) platforms. Notably, the supervised ML model introduced provides extensive flexibility in the prediction process with rich features as well as improving the forecasting of values As per the experimental result determined from the study, it has been observed that the model provides improved accuracy with greater benefits in overcoming the issues with CLV prediction through traditional heuristic models. At the same time, the supervised model introduced in the study has optimized the cost of marketing and potentially driven managerial insights within the context.

Amid the fierce competition identified in many sectors including telecom and e-commerce, companies are adopting specific marketing options to acquire and retain potential customers. In this approach, a focus, however, is given the significance of CLV prediction to analyze customer relationships with firms. Understandably, the information presented by Venkatakrishna et al. (2020) explained that the integration of market decision-making models to optimize customer-centric product values is an important approach. On the other hand, Fahim et al. (2020) explained that predicting customer lifetime value (CLV) using regression analysis evaluates customers' data through virtual retailers. This process encompasses a highly suitable prediction process that has necessarily discerned knowledge of customer loyalty and maximizes the count of high-value customers. Information presented in the study by Bauer and Jannach (2021) has acknowledged the limited predictability of revenue statistical methods that have enhanced research focus toward machine learning approaches. It has been observed that the study introduces an advanced ML technique tailored with "encoder-decoder" sequence-to-sequence "recurrent neural network" (RNN) fused with "augmented temporal convolutions". As per the novel features used by the model, it has been observed that the outcome provides a highly efficient result that can enhance competitive performance. Furthermore, information has provided insights into the conceptual importance of customer lifetime value as an essential metric in the identification of profitability. In the e-commerce and retail industry, the prediction of CLV provides distinct information on the way customers can be handled and managed efficiently in long-term relationships. Based on the empirical analysis, it

has been determined that real-time customer transactions in online platforms has verified the validity as well as the applicability of various customer segmentation methods. Therefore, it can be stated that the evolution in the prediction process over the years has extensively studied the importance of customer lifetime value (CLV) and its relevance to business practices.

With the rapidity identified in technological transformation in business, artificial intelligence (AI) and machine learning have served greater contributions in managing business processes efficiently amid global competition. Chen (2018) explained that the increased competition in some sectors such as airline, e-commerce, and telecom, customer acquisition, and long-term retention is a valuable task. With the sophisticacy and reliability identified with machine learning-based applications, researchers have extended their focus on such advancement. Considerably, Laksono and Wulansari (2022) explained that customer segmentation as well as resource allocation through the recency-frequency (RFM) analysis while applying k-Means clustering analysis provide extensive results. Alternatively, Sun et al. (2021) explained that using two enhanced ML models - "gradient boosting decision trees" (GBDT) and random forest (RF) have provided higher prediction performance compared to two classical methods - NBD and GGG models. Therefore, it can be stated that ML models are highly efficient in predicting sales and CLV for businesses. Based on the information presented by Sun et al. (2023), it has been determined that customer lifetime value prediction is essential for enterprises irrespective of their size and business agenda. Despite this fact, previously approached "single data mining" techniques show limitations in the efficient prediction of CLV. Understandably, the study has introduced a segmentation model using machine learning through which customer relationships and segmentation have been determined even with non-contractual relationship conditions. Yang et al. (2023) on the other hand, have introduced a feature extraction and prediction model "feature missing-aware routing-and-fusion network" (MarfNet) that has efficiently reduced missing features impact and increased sample interactions. Indicating the efficiency of the model, it can be stated that the result provides higher efficiency in overcoming the issues with missing features and increases the prediction performance with superiority.

2.3 Gap in the Literature

In our literature analysis, we noted several key research gaps. The majority of studies use traditional techniques and classic machine learning models to forecast customer lifetime value (CLV) and sales, and overlook the potential of more sophisticated algorithms and extensive comparative analysis. Moreover, we found no research focused on regional and product-specific insights which are critical to developing products for effective targeted marketing strategies. Additionally, there is a common failure to integrate a thorough exploratory data analysis (EDA) with predictive modeling, which undermines the depth and usability of the available insights. By using advanced machine learning algorithms, conducting comprehensive EDA and giving a detailed analysis of CLV across various regions and product categories, our analysis addresses these research gaps and provides a more robust and nuanced framework for optimizing customer value and productivity in the e-commerce industry

3 Methodology

The competition in the e-commerce industry is increasing day by day, so understanding customer behavior and predicting future sales is very important for making strategic decisions and long-term success. By leveraging advanced machine learning methods, valuable insights can be gained by businesses to understand customer purchasing patterns, preferences, and lifetime values. So, it is important to develop robust methodologies to predict sales and analyze the customer lifetime value (CLV) by leveraging machine learning methods with which e-commerce companies optimize their marketing strategies, improve the retention of customers, and generate the maximum revenue. Therefore, in this research, we have developed an effective methodology which consists of several sets of steps, to predict the sales from the Brazilian e-commerce dataset. The detailed description of the methodology for each step is discussed below and the flow diagram of methodology is shown in Figure 1.

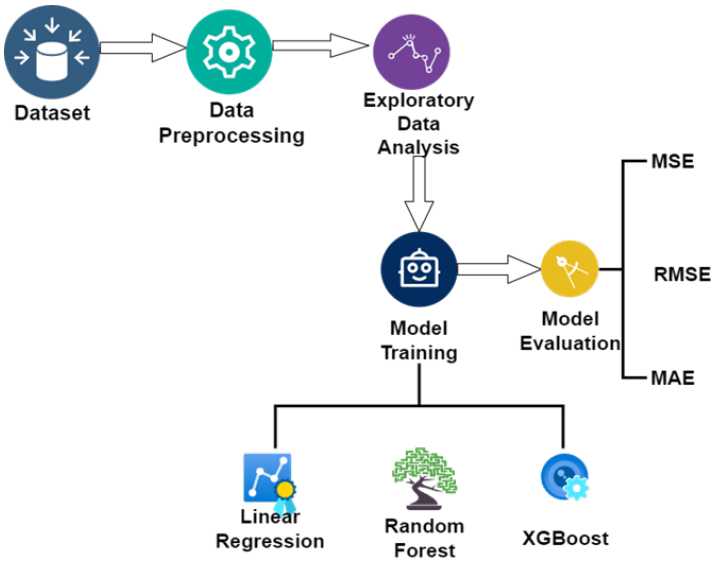

Figure 1: Methodology for predicting sales and CLV using Machine Learning Approach

3.1 Data Description

In this research, the dataset is taken from the Brazilian public datasets of the orders that are made at the Olist Store Olist and Magioli (2018). The dataset contains the records of 100k orders from the year 2016 to 2018 which are places at many marketplaces in Brazil. There are separate datasets for each entity such as customers data which contains 99441 rows and 5 columns, sellers dataset which contains 3095 rows and 4 columns, and products dataset contains 32951 rows and 9 columns. The features in the dataset allow viewing the order from various perspectives such as order status, price, payment, performance of the shipping, customer location, product attributes, and reviews of customers. Additionally, the Brazilian dataset is linked with zip codes to latitude and longitude coordinates which is added by the geolocation dataset. In the dataset, there might be multiple items in the order. Each item in the dataset might be filled by a unique seller. The data is grouped into various datasets for better understanding. The various datasets are the order payments dataset, products dataset, order review dataset, and others given the

data schema given in Figure 2.

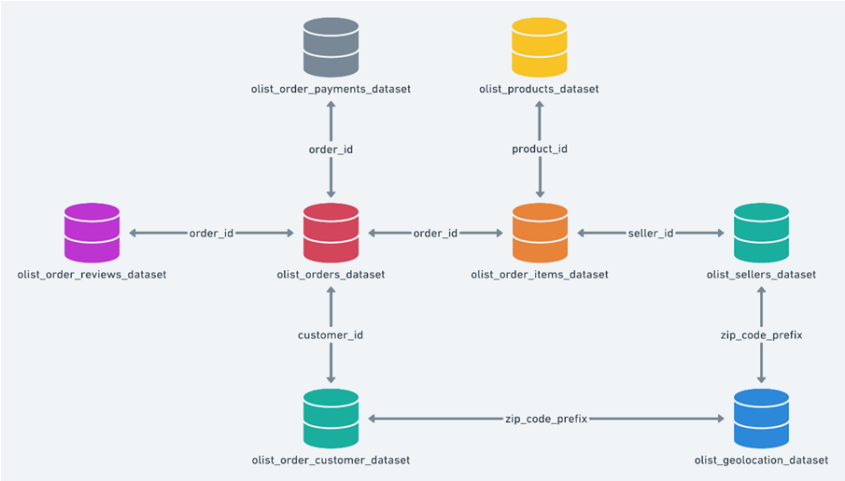

Figure 2: Schema of the data

3.2 Data Preprocessing

Data preprocessing is an important step in the data analysis and machine learning pipeline. It consists of cleaning, transforming, and organizing the raw data into a suitable format for analyzing the data and building the machine learning models. The data preprocessing improves the data quality, helps in enhancing the performance of the model, reduces the complexity of the model, and handles missing data. In this research, the data preprocessing is carried out by removing the duplicate values from the data as duplicate data increases the data size and by dropping it has no change on the dataset. All the datasets are summarized which gives the information from the data such as several columns and number of rows in the data. The summary of all the datasets is given in Figure 3.

	all_datasets	Columns	Column_name	Rows
0	customers	5	customer_id, customer_unique_id, customer_zip_code_prefix, customer_city, customer_state	99441
1	sellers	4	seller_id, seller_zip_code_prefix, seller_city, seller_state	3095
2	products	9	product_id, product_category_name, product_name_lenght, product_description_lenght, product_photos_qty, product_weight_g, product_length_cm, product_height_cm, product_width_cm	32951
3	orders	8	order_id, customer_id, order_status, order_purchase_timestamp, order_approved_at, order_delivered_carrier_date, order_delivered_customer_date, order_estimated_delivery_date	99441
4	order_items	7	order_id, order_item_id, product_id, seller_id, shipping_limit_date, price, freight_value	112650
5	order_payments	5	order_id, payment_sequential, payment_type, payment_installments, payment_value	103886
6	order_reviews	7	review_id, order_id, review_score, review_comment_title, review_comment_message, review_creation_date, review_answer_timestamp	99224
7	geolocation	5	geolocation_zip_code_prefix, geolocation_lat, geolocation_lng, geolocation_city, geolocation_state	738332
8	product_translation	2	product_category_name, product_category_name_english	71

Figure 3: Summary of the data

The dataset with the maximum number of columns is products and the dataset with the maximum number of rows is geolocations. Missing values are the values that contain no information and are filled with null values which can lead to poor performance of the model. In the research, the missing values are handled by filling with the median values and some are replaced by the other dataset. When the dataset is merged can lead to duplication of some rows, and these rows are de-duplicated as these rows will add nothing crucial to the data but lead to redundancy.

3.3 Exploratory Data Analysis

Data analysis is the method of describing, summarizing, and evaluating the data which involves inspecting, cleaning, transforming, and modeling the data which helps in discovering meaningful information, concluding, and supporting decision-making. It is the key component in the machine learning pipeline which enables extraction of meaningful information from the data. In the research, the data analysis is carried out by analyzing the different datasets such as payment type analysis as shown in Figure 4.

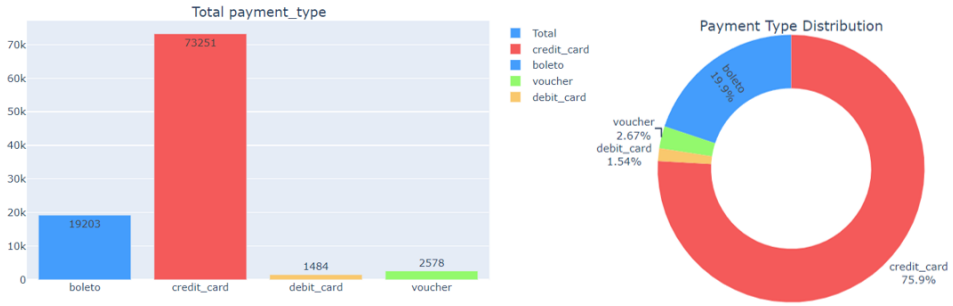

Figure 4: Count of payment type and distribution of payment

Based on the analysis of the count of payment type, Figure 4 represents a bar chart and pie chart describing the count of payment type and the distribution of the payment type. Most of the payments are made by customers having credit cards and the smaller number of payments are done by debit card customers. The percentage of each mode of payment is shown in a pie chart which shows amongst all payments made by the user credit card is used by 75.9% of the users, boleto is used by 19.9% of the users and 3.2% of the user used vouchers and debit card.

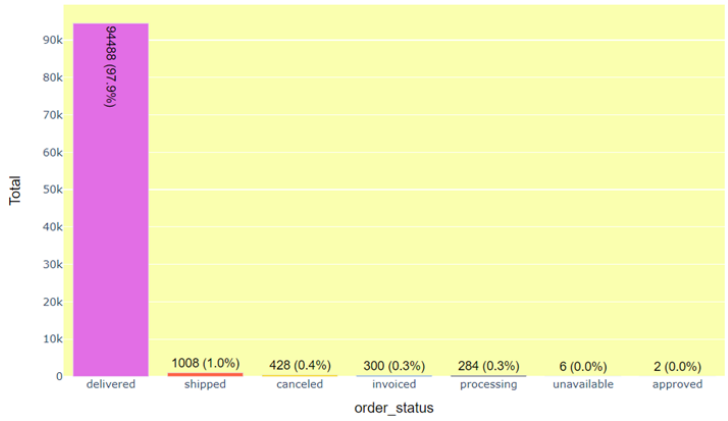

Figure 5: Total Order Status

Another analysis for total order status, Figure 5 represents a bar chart which shows that 97.8% of the orders of status delivered and the remaining percentage are with status shipped, canceled, invoiced, processing, unavailable, and approved. The highest number of orders went from delivered ones. Only 3% of all orders came from the other status.

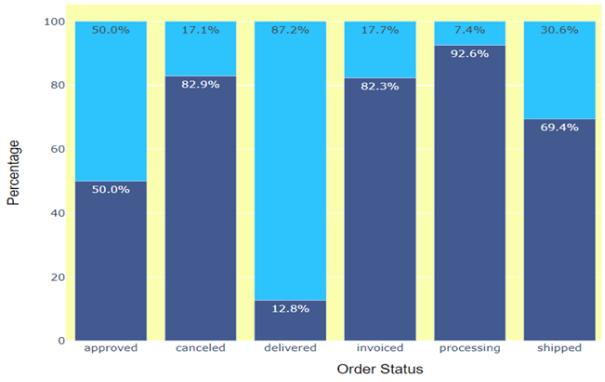

Figure 6: Order status with percentage of reviews

Furthermore, the analysis for the order status with the percentage of reviews is given in Figure 6 which shows a bar chart for the order status delivered most of the orders are with positive reviews i.e., 85% and only 12.8% are negative reviews.

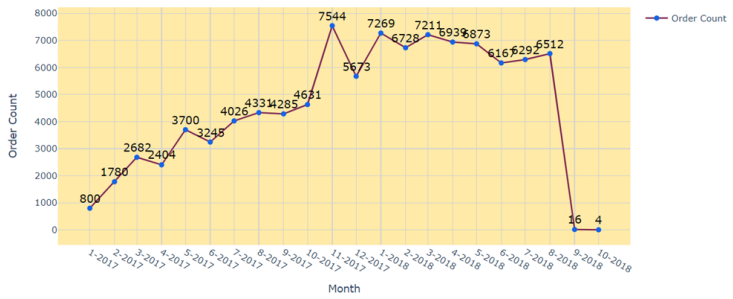

Figure 7: Order trend by month in 2017-2018

Based on another analysis of the dataset, we have plotted a line chart to understand the order trend by month as shown in Figure 7 which shows the trend of order by month in the year 2017-2018. From the figure, it can be concluded that the trend in orders increased from 800 to 7544 from January to November 2017, and the trend gradually decreased from December 2017 to August 2018 and there is a sudden decrease of orders from 6512 to 16 in month July to August 2018.

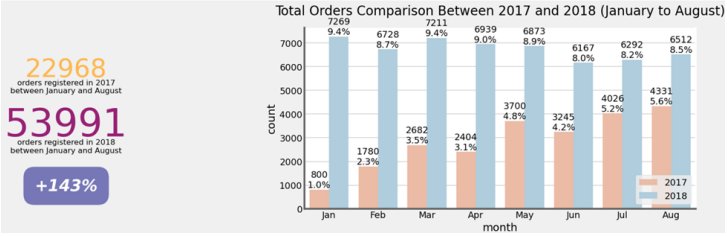

Figure 8: Total order comparison between 2017 and 2018 (January to August)

The analysis of the total order comparison between 2017 and 2018 shown in Figure 8 shows a bar graph that represents a clear increase in the number of orders from January to August in the years 2017 and 2018. From the figure, it can be concluded that there is an increase of 143% in January from 22968 orders in 2017 to 53991 orders in 2018. It can be seen that there is an increase in orders from January to August which indicates a consistent upward trend in the order volume. The trend of growth can be important in the analysis of business and strategic planning.

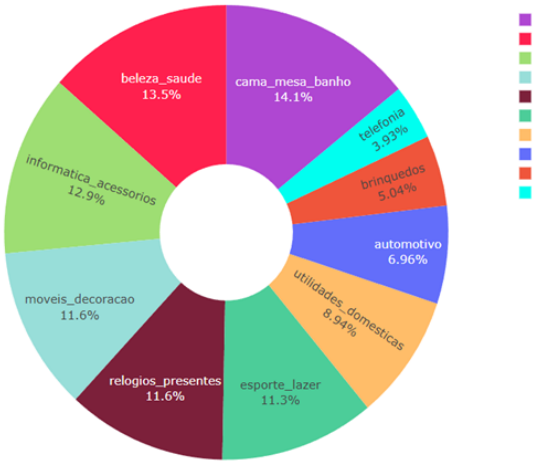

Figure 9: Top 10 products based on revenue

The analysis of the top 10 products based on revenue is shown in Figure 9 which represents a pie chart that provides the top 10 products based on revenue. From the pie chart, the top revenue generators are Cama Mesa Banho with 14.1% of total revenue, Beleza saude contributing 13.5% of total revenue and Informatica accessories for 12.9% of the revenue. Utilidades Domesticas, Automotivo, Brinquedos, and Telefonia generated low revenue.

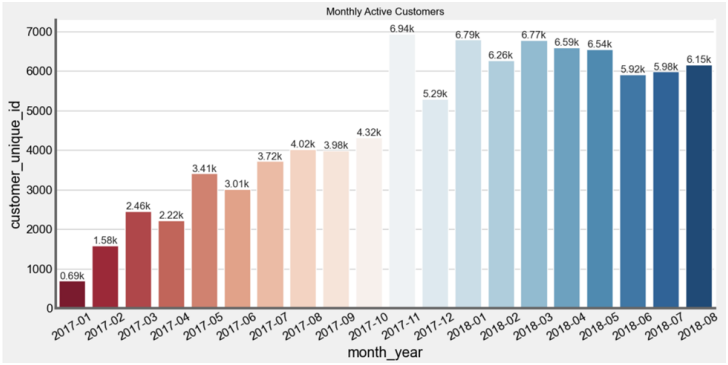

Figure 10: Monthly Active Customers

Another analysis for the monthly active customers in Figure 10 represents a bar chart for the number of unique customer visits per month from January 2017 to August 2018. From the figure, it can be seen that there is a clear upward trend in the number of customers. Also, it can be noted that there was a jump in customer numbers around mid-2017 and early 2018 which indicates significant growth. A seasonal pattern can be noted where customer activity spikes, which can help make marketing strategies or know the customer behavior.

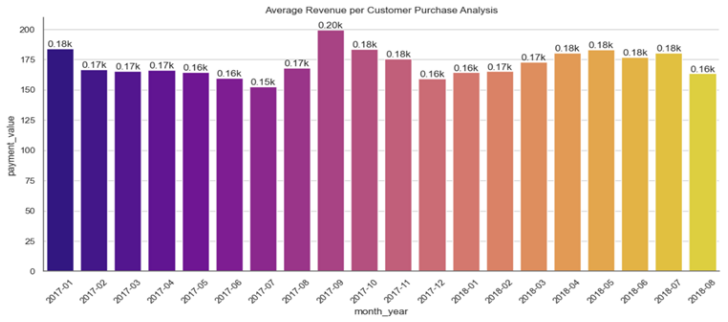

Figure 11: Average revenue per customer purchase analysis

A comparison of average revenue per customer purchase is plotted in Figure 11 which shows a bar chart that gives insights into the average revenue per customer purchase from January 2017 to August 2018. There is a slight upward trend can be shown by the average revenue per customer purchase over the period which indicates an increase in the value of each purchase. Also, it can be noted that there is a peak around July each year, suggesting seasonal or promotional impacts on customer spending.

Figure 12: Monthly Retention Rate

A visualization of the monthly retention rate is shown in Figure 12 which shows a bar chart for the monthly retention rate and gives insights into retention rates from January 2017 to December 2018. It can be observed, that in November 2017 the retention rate was high reaching 0.74. A lower retention rate can be seen in April 2018 at 0.21.

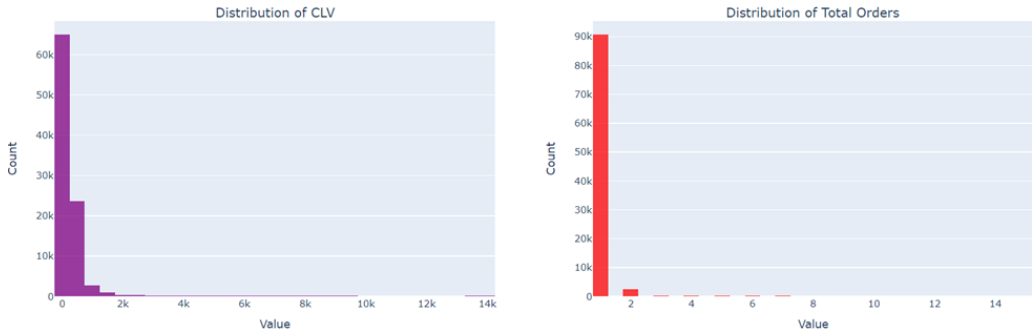

Figure 13: Combined distributions of CLV and Total Orders

A bar chart for the distribution of CLV and the distribution of Total orders is plotted in Figure 13. From the figure it can be observed that the majority of the customers have low customer lifetime value, having the highest frequency in the range (0-2k). It can also be observed that there is a decrease in the frequency of the customers significantly as the CLV increases which indicates few customers with high lifetime value. For the distribution of total orders, it can be seen that most of the customers have placed a low number of orders, with the highest frequency. As the total order increases the frequency of customers decreases which suggests that only a few customers place many orders.

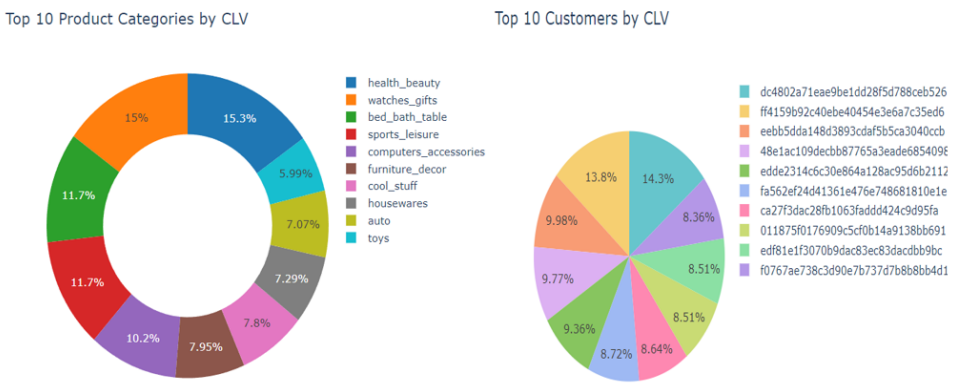

Figure 14: Distribution of top 10 product categories by CLV and distribution of top 10 customers by CLV

A donut chart is visualized for the analysis of the distribution of the top 10 categories by CLV and a pie for the top 10 customers by CLV in Figure 14. The donut chart for top product by CLV provides the information about which product categories by CLV. The health & beauty category have the highest share at 15.3% and Toys among all the categories have the smallest share of 4%. The pie chart for the top 10 customers by CLV gives the insights into distribution of CLV among the top customers.

Figure 15: Distribution of CLV by Customer State

An analysis of the distribution of CLV by customer status is visualized in Figure 15 which represents a box plot for the distribution of CLV by customer state which provides information about the CLV across different states. State SP with the highest data point and some highest CLV that represent a significant customer base having high lifetime value.

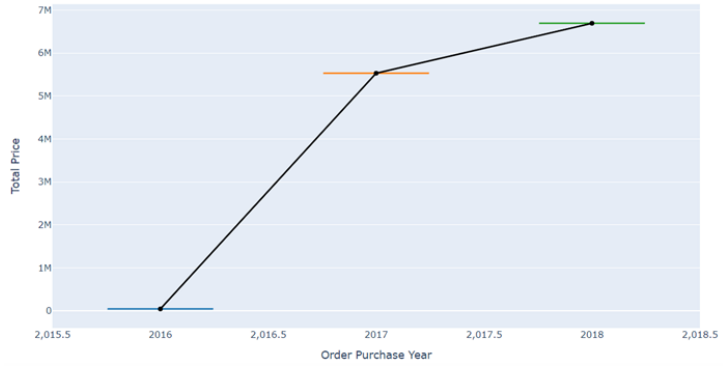

Figure 16: CLV distribution by year

A line chart is plotted in Figure 16 for the distribution of CLV by year which gives information about the growth of CLV over time. From the graph, it can be concluded that there is a steady increase in the total price from mid-2015 to mid-2018. There is a notable steep growth in Customer Lifetime Value from mid-2015 to the end of 2106. The is a continuous slow rate of growth after 2016 and just reached 6 million by mid-2018.

3.4 Feature Engineering

Feature engineering is the model of selecting, and transforming of variables to improve the performance of the machine learning algorithm. Feature engineering helps in selecting features that are more relevant to the model training, and also it helps in converting raw data into a suitable format for training the algorithm such as normalization, scaling, and encoding of categorical variables. Necessary information is provided by the feature engineering that is crucial for making accurate predictions. In this study, a systematic approach is employed to train and evaluate machine learning models for making predictions. The original dataset contains details of various customers and products with order, and delivery timestamps, which are pre-processed for model training. The columns that contain datetime information like order purchase timestamp, and order delivered customer data are converted to datetime format. Features such as year, month, day, day of the week, etc. are extracted from the order purchase time stamp, the daytime was categorized into dawn, morning, afternoon, and night. The target variables and features variables are separated from the data and categorical variables are encoded using Label Encoder which makes it suitable for machine learning algorithms. The features are normalized to mitigate the impact of varying scales. Using MinMaxScaler, the features are scaled in the range [0,1]. To reduce the high dimensionality of data from the dataset to mitigate potential multicollinearity, Principal Component Analysis (PCA) is employed on the data. The original feature space is transformed into a set of linearly uncorrelated components using PCA. A cumulative explained variance graph is plotted against the number of components to determine the number of components that are optimal for our analysis which is shown in Figure 17.

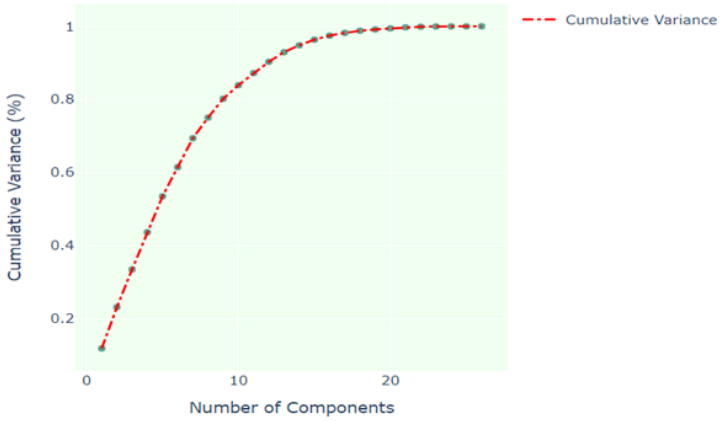

Figure 17: Principal Component Analysis (PCA)

3.5 Model Training

Model training is a crucial part of making accurate predictions. In this study, the model training stage consists of preparing the data, relevant features, and selection of appropriate machine learning models Three machine learning algorithms are employed on the data for training and evaluation of the result. The three machine learning models are linear regression which is a statistical approach and a base model for predicting the target values using a linear approach. Random forest regressor which is an ensemble machine learning algorithm for predicting numerical values using multiple decision trees and lastly XGBoost which is a gradient boosting machine learning algorithm for predicting numerical values efficiently and accurately. Each machine learning algorithm is trained on the training set by splitting the dataset into an 80% training set and for testing 20% data set is used.

3.6 Model Evaluation

After training the models on the training data, the final step is to evaluate the performance of the machine learning algorithms on the test data. This evaluation is important to find out how well the models generalize on the unseen data. Three performance metrics are evaluated on the data. The three performance matrices are Mean Square Error (MSE) which is the average of the squared differences between the predicted values and actual values to penalize the large error more heavily because of the squaring difference which makes it sensitive to outlier and MSE also helps in knowing the variance in the errors. Root Mean Squared Error (RMSE) which is the square root of the mean squared error. Mean Absolute Error (MAE) which is the average of the absolute difference between the actual output and predicted output. These metrics are evaluated on the test dataset to assess the accuracy of the models, and model comparisons, these metrics give an interpretable measure of error, which helps in understanding the performance of the model in real-world terms.

4 Design Specification

The research focuses on the regression task which is carried out by implementing three machine learning algorithms: Linear Regression, Random Forest, and XGBoost. A detailed explanation of these algorithms is given below:

4.1 Linear Regression

Linear regression is a statistical approach that is used to set up a linear relationship between a dependent variable and one or more independent variables referred to as features. A line is fitted called the best-fit line or regression line that predicts the output values from the features. The simplicity of linear regression makes it a good start to predict the values. A clear insight into the relationships between the dependent and independent variables is provided by the algorithm. Linear regression provides an important benchmark for evaluating model performance on complex data. The coefficient of linear regression represents the effect on the target variable for each variable that provides insights into the factors that influence sales and CLV. An architectural diagram of the linear regression is shown in Figure 18.

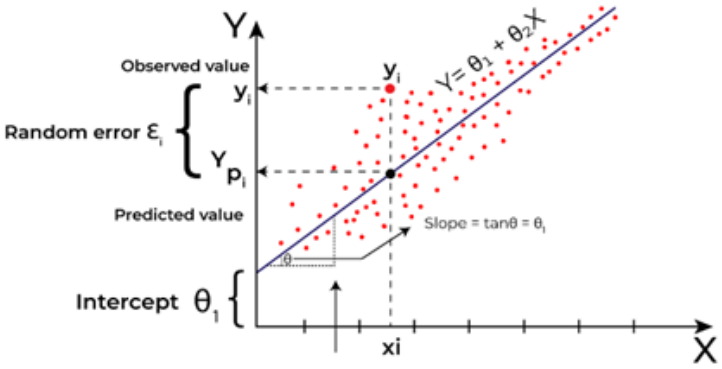

Figure 18: Linear Regression Model Architecture Sarmento and Costa (2017)

4.2 Random Forest

Random Forest is an ensemble learning technique that is used for solving classification and regression tasks. Random Forest constructs multiple decision trees during training and evaluates the average prediction (in case of regression). The risk of overfitting is reduced in Random Forest due to the ensemble method which generalizes better on new data. High predictive accuracy is provided by random forest that leverages the strength of several trees that capture complex patterns from the data. The e-commerce data consists of complex interactions and non-linear relationships between input variables and output variables. The ability to capture complex patterns makes it an efficient choice for predicting sales and CLV. The robustness and versatility of random forest made a better model to predict the sales and analyze CLV. The diagram for the random forest is given in Figure 19.

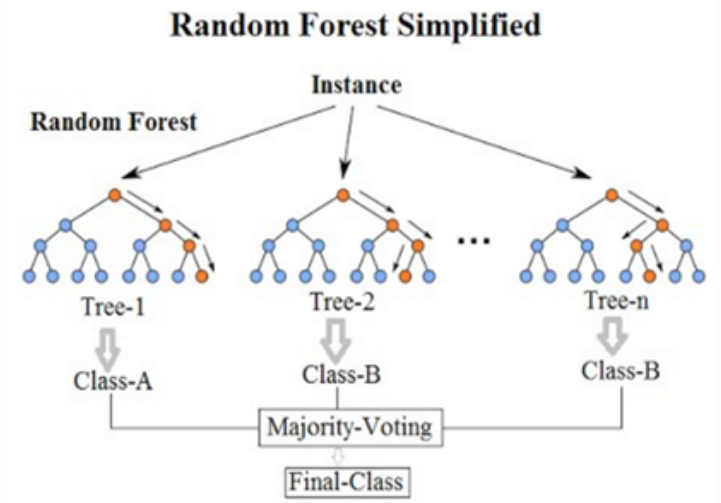

Figure 19: Random Forest Algorithm Architecture Cutler et al. (2011)

4.3 XGBoost Algorithm

XGBoost is a gradient-boosting technique which is an ensemble learning method for evaluating the regression task. Multiple decision trees are made in sequence where each tree tries to correct the errors which was made by the previous tree. Due to the superior performance, the robustness of XGBoost the ability to prevent overfitting, and the efficiency in handling large data make a better model to predict the sales and Customer Lifetime Value (CLV) in the e-commerce industry. An architectural diagram of the model is given in Figure 20.

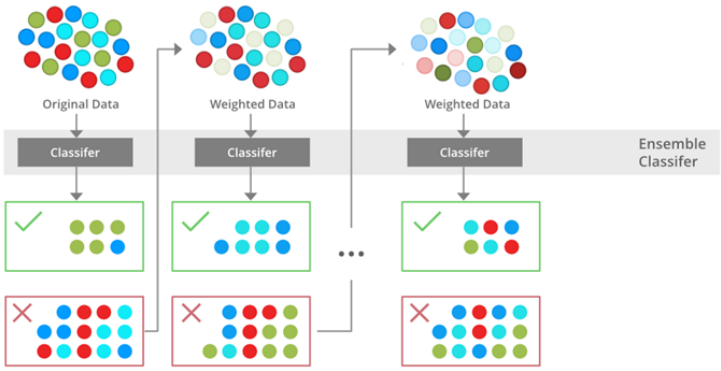

Figure 20: XG Boost Algorithm Architecture Chen and Guestrin (2016)

5 Implementation

This research is performed in the Windows operating system, with 8GB RAM, in Python language using Jupyter Notebook IDE, focusing on predicting sales value and analyzing Customer Lifetime Value (CLV) using machine learning algorithms. Key libraries used in the research include NumPy for numerical operations and array handling, Pandas for data manipulation and transformation, Matplotlib and Seaborn for visualizing data

distributions and patterns, and Plotly for creating interactive plots and visualizations. There is a need to predict sales and CLV to optimize business operations, increase customer experience, and drive growth. The research is carried out by collecting the data from Kaggle, a repository of datasets. The implementation involves converting various columns to datetime using different formats such as to_datetime from pandas to convert strings into dates, times, and months, which are then separated. Features that are more relevant to predicting sales values are selected. Following feature selection, the data is separated into target values and feature values, and the categorical features are labeled and encoded as 0 and 1 using LabelEncoder from Scikit-learn to convert categorical features into numerical features. The features are then scaled within the range of 0 and 1 using MinMaxScaler from Scikit-learn to improve the results. After label encoding and scaling of feature variables, the features are further selected using PCA by explaining cumulative variance. With the relevant features selected, the dataset is then split into training and testing sets using the train_test_split method from Scikit-learn to ensure the model's robustness and generalizability on unseen data. Three machine learning algorithms—Linear Regression, Random Forest, and XGBoost—are fitted on the data to capture complex patterns and improve performance on unseen data. The models are evaluated using three performance metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to assess the models' effectiveness.

Component	**Specification**
CPU	Intel Core i7 (octa core)
RAM	8 GB
Storage	512 SSD
Operating System	Windows 10
Python Version	3.10
ML libraries	Pandas, NumPy, Scikit-learn, LinearRegression, RandomForestRegressor, XGBoost, Seaborn, Plotly, Matplotlib
IDE	Jupyter Notebook
Performance Metrics	MSE, RMSE, MAE

Table 2: System Requirement and Resource Details

6 Evaluation

This research is carried out by implementing three machine learning algorithms: Linear regression, random forest regression, and XGBoost. The problem given is the regression problem to evaluate the sales value of the product based on the features that are selected using Principal Component Analysis. Three performance metrics are used to evaluate the performance of the algorithms. The three metrics are Mean Squared Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to evaluate the model's ability to generalize on the test data. The comparison and evaluation of these metrics are discussed.

6.1 Evaluation Based on Mean Squared Error (MSE)

Mean Squared Error gives the average of the squared difference between the actual output and the predicted output by penalizing the large error more heavily because of the squaring difference which makes it sensitive to the outliers. In the research, the MSE value achieved by the Linear Regression model is 32900.08, while the Random Forest models attain an MSE value significantly lower at 26181.65and the MSE value attained by XGBoost is 27196.53 which is lower than linear regression but higher than random forest. The lower MSE values tell that the model is a good fit for the prediction. and in our research, the MSE value of the Random Forest algorithm exhibits the best performance among the three algorithms. This comparison addresses the effectiveness of model selection in minimizing prediction errors. Figure 21 presents a horizontal bar chart to provide a comparative study of the models based on MSE values.

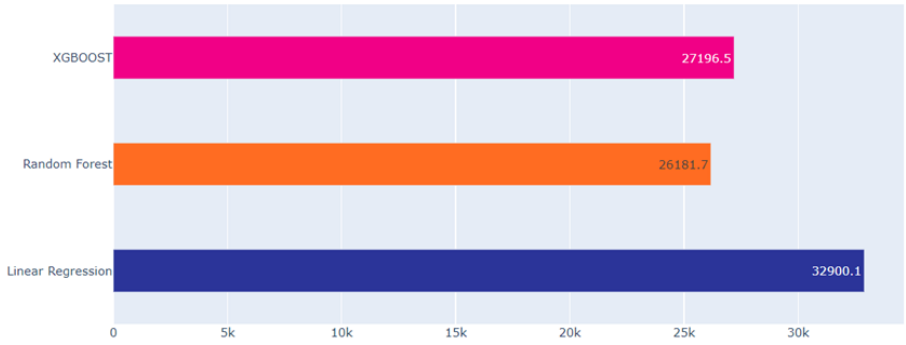

Figure 21: Mean Squared Error Comparison

6.2 Evaluation Based on Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) gives the squared root the mean squared error. RMSE gives an error metric on the same scale which makes it easier to interpret the result. A larger error is penalized by RMSE but is more interpretable because of the same units as the target variable. In the research, the RMSE value achieved by the linear regression model is 181.38, while the accuracy attained by Random Forest regression is lower than liner regression which is 161.81 and the XGBoost attains the RMSE value of 164.91 which is lower than linear regression but higher than Random Forest. Thus, lower the RMSE value of the model, the model is better for predicting the values, and based on RMSE, random forest models give superior performance. Figure 22 shows a sunburst chart for the comparative analysis of three models based on RMSE.

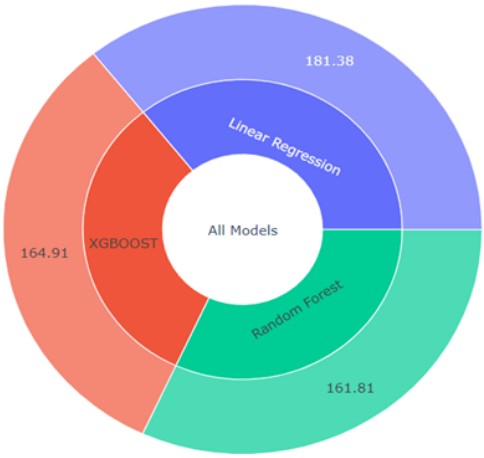

Figure 22: Comparison of Three models based on RMSE

6.3 Evaluation Based on Mean Absolute Error (MAE)

Mean absolute error is the mean of the absolute difference between the actual output and predicted output. MAE gives a measure of the accuracy of the model by taking the average of absolute error, providing the average magnitude of error without considering their direction. In our experiment, linear regression attains the MAE of 84.56, while the Random Forest models achieve the MAE of 70.15. The MAE achieved by XGBoost models is 75.269348 which is significantly less than linear regression but greater than Random Forest. Thus, a model is said to be more accurate for predicting the values if its MAE value is less compared to other models, and among the three algorithms, the random forest model demonstrates better performance based on MAE. Figure 23, shows a scatter plot that gives a comparative analysis of the three models.

Figure 23: Comparison of Mean Absolute Error (MAE)

6.4 Discussion

In the research for predicting the sale, three machine learning algorithms are employed: Linear Regression, Random Forest, and XGBoost, and these models are evaluated on three performance metrics MSE, RMSE, and MAE. Random Forest gives consistent performance across all metrics which represents the effectiveness of Random Forest in this

regression problem. Complex patterns are captured by the ensemble nature of the random forest which may not be captured by simple models such as Linear Regression. On the other hand, XGBoost also demonstrates strong performance but not better as compared to Random Forest. Thus, the results stress the significance of choosing accurate algorithms and scores when comparing the performance of machine learning models Random forest is found to be the most accurate model for predicting sales. According to the CLV analysis, most customers have relatively low lifetime values, while a small proportion of customers are driving significant revenue. The product categories that contribute more to the CLV are health and beauty, and watches and gifts. The region where customers have the highest lifetime values is São Paulo, followed by Rio de Janeiro. This contrast between low CLV customers and higher-value customers points to potential for targeted marketing in certain regions. Over time, the trend has been an increase in CLV, reflecting positive business growth. Insights from these analyses identify areas for strategic improvement to enhance customer value and business performance.

7 Conclusion and Future Work

In this study, we used the Brazilian e-Commerce Public Dataset by Olist to predict sales and analyze Customer Lifetime Value (CLV) using several machine learning (ML) algorithms. After conducting a wide range of exploratory data analysis (EDA), we detected critical patterns and trends that strengthened our predictive model. We selected three ML algorithms—Linear Regression, Random Forest, and XGBoost—and assessed them with the support of their Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). Random Forest had the best performance among all the other algorithms, followed by XGBoost and Linear Regression.

We found that most of the customers have low lifetime values, and a small percentage of customers have significantly higher lifetime values. We also identified significant product categories driving CLV such as health and beauty, watches and gifts, and bed, bath, and table. Geographically, we saw that states like São Paulo and Rio de Janeiro had considerably higher CLV. Over time, there was a significant growth in the CLV, indicating the business grew consistently as well.

The future work of this research is to enhance the predictive ability of the model by performing Grid Search or Bayesian optimization which could improve the performance of the model. Advance machine learning or deep learning model can be used to enhance the performance. Models such as Gradient Boosting Machine or neural networks. The model can be deployed to real-world application such as recommendation system, sales forecasting tools or CLV estimation platform which enable time to time evaluation of model and the refinement of the model base on the real-time data.

References

Bauer, J. and Jannach, D. (2021). Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models, *ACM Transactions on Knowledge Discovery from Data (TKDD)* **15**(5).
URL: *https://doi.org/10.1145/3441444*

Chamberlain, B. P., Cardoso, Â., Bryan Liu, C. H., Pagliari, R. and Deisenroth, M. P. (2017). Customer lifetime value prediction using embeddings, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1753–1762.
URL: *https://doi.org/10.1145/3097983.3098123*

Chen, S. (2018). Estimating customer lifetime value using machine learning techniques, *Data Mining*, IntechOpen.
URL: *https://doi.org/10.5772/INTECHOPEN.76990*

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system, pp. 785–794.

Curiskis, S., Dong, X., Jiang, F. and Scarr, M. (2023). A novel approach to predicting customer lifetime value in b2b saas companies, *Journal of Marketing Analytics* **11**(4): 587–601.
URL: *https://doi.org/10.1057/S41270-023-00234-6/FIGURES/5*

Cutler, A., Cutler, D. and Stevens, J. (2011). *Random Forests*, Vol. 45, pp. 157–176.

Fahim, S. F., Mohshiu, M. and Khan, I. (2020). Customer lifetime value prediction using the regression model, *International Research Journal of Modernization in Engineering* **3445**.
URL: *https://www.irjmets.com/papers/3445.pdf*

Laksono, B. C. and Wulansari, I. Y. (2022). Estimating customer lifetime value in the e-commerce industry using multivariate analysis, *Proceedings of The International Conference on Data Science and Official Statistics*, Vol. 2021, pp. 507–518.
URL: *https://doi.org/10.34123/ICDSOS.V2021I1.161*

Norouzi, V. (2024). Predicting e-commerce clv with neural networks: The role of nps, atv, and ces, *Journal of Economy and Technology* **2**: 174–189.
URL: *https://doi.org/10.1016/J.JECT.2024.04.004*

Olist, D. and Magioli, F. (2018). Brazilian e-commerce public dataset by olist.
URL: *https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce*

Paul, R. (2023). Predicting customer lifetime value with machine learning: A comprehensive guide.
URL: *https://rudrendupaul.medium.com/predicting-customer-lifetime-value-with-machine-learning-a-comprehensive-guide-f1b0ffd7d6f8*

Platzer, M. and Reutterer, T. (2016). Ticking away the moments: Timing regularity helps to better predict customer activity, *Marketing Science* **35**(5): 779–799.
URL: *https://doi.org/10.1287/MKSC.2015.0963*

Pollak, Z. (2021). Predicting customer lifetime values – ecommerce use case.
URL: *https://www.researchgate.net/publication/349234510*$Predicting_Customer_Lifetime_Values--e$
$commerce_use_case$

Qismat, T. and Feng, Y. (2023). Comparison of classical rfm models and machine learning models in clv prediction.

Sarmento, R. and Costa, V. (2017). *Introduction to Linear Regression*.

Sun, Y., Cheng, D., Bandyopadhyay, S. and Xue, W. (2021). Profitable retail customer identification based on a combined prediction strategy of customer lifetime value, *Midwest Social Sciences Journal* **24**(1): 104–127.
URL: *https://doi.org/10.22543/0796.241.1053*

Sun, Y., Liu, H. and Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model, *Heliyon* **9**(2): e13384.
URL: *https://doi.org/10.1016/J.HELIYON.2023.E13384*

Vanderveld, A., Pandey, A., Han, A. and Parekh, R. (2016). An engagement-based customer lifetime value system for e-commerce, *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 293–302.
URL: *https://doi.org/10.1145/2939672.2939693*

Venkatakrishna, M. R., Mishra, M. P., Sneha, M. and Tiwari, P. (2020). Customer lifetime value prediction and segmentation using machine learning, *International Journal of Research in Engineering and Science (IJRES)* **9**: 36–48.
URL: *www.ijres.org*

Yang, X., Jia, B., Wang, S. and Zhang, S. (2023). Feature missing-aware routing-and-fusion network for customer lifetime value prediction in advertising, *WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining*, pp. 1030–1038.
URL: *https://doi.org/10.1145/3539597.3570460*