National College of Ireland

Predicting Customer Lifetime Value (CLV) in UK and Brazil using Machine Learning and Deep Learning: A Comparative Analysis

MSc Research Project
Data Analytics

Christy Davis Maliyekkal
Student ID: x22151648

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Christy Davis Maliyekkal		
Student ID:	x22151648		
Programme:	Data Analytics	**Year:**	2024
Module:	MSc Research Project		
Supervisor:	Teerath Kumar Menghwar		
Submission Due Date:	31/01/2024		
Project Title:	Predicting Customer Lifetime Value (CLV) in UK and Brazil using Machine Learning and Deep Learning: A Comparative Analysis		
Word Count:	7753		
Page Count	22		

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:	Christy Davis Maliyekkal
Date:	31/01/2024

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 Customer Lifetime Value (CLV) in UK and Brazil using Machine Learning and Deep Learning: A Comparative Analysis

Christy Davis Maliyekkal
x22151648

Abstract

Understanding and efficiently deciding Customer Lifetime Value (CLV) is essential for keeping competitive advantage in the modern era of e-commerce. This study aims to taking into account the differences between two very different markets: Brazil and the United Kingdom, to clarify the complex effectiveness of CLV. The primary objective is to provide insightful information that helps companies operating in these various e-commerce environments make more informed decisions. The project is organized, starting with a careful data collection procedure to guarantee a thorough portrayal of customer behavior and transactions. The datasets from Brazil and the UK form the basis for the studies that follow. The XGB Regressor, Support Vector Machine (SVM), and Random Forest algorithms are used to model CLV trends for each market under the machine learning framework. Moreover, the deep learning technique makes use of the Multilayer Perceptron (MLP) Regressor to identify complexities and connections in the data. The Brazilian e-commerce market performed better than the UK e-commerce by showing better performance in accuracy and error pattern. Random Forest and MLP Regressor are the better performed algorithms. A comparison of the models that clarifies the advantages and disadvantages of each strategy. The results not only further our understanding of CLV but also provide useful information for companies looking to adjust their strategy to the particularities of the Brazilian and UK e-commerce markets. To put it simply, our research acts as a role model for companies, helping them navigate the complex world of consumer dynamics and provide a path forward for utilizing advanced analytics to achieve long-term profitability and success.

Keywords: Customer Lifetime Value, CLV, Machine Learning, Deep Learning, XGB Regressor, SVM, Random Forest, Multi-Layer Perceptron (MLP) Regressor.

1 Introduction

The capacity to not only comprehend but also anticipate client behaviour stands as a vital pillar of success for businesses spanning a variety of industries in today's ever-changing and fiercely competitive business environment. The measurement of Customer Lifetime Value (CLV), which acts as a crucial indicator of the total value a client delivers to a business throughout their contact with the brand, is at the core of this venture. CLV encompasses a variety of elements, including the money a customer generates over time as well as more subtle facets like their purchasing patterns, loyalty levels, and the length of their engagement with the business. It may be difficult to fully understand the complex nuances of consumer behavior using typical analytical techniques and statistical methodologies. So, the thesis of this study is based on this hypothesis. In order to uncover deeper understandings about CLV patterns, promotes the integration of advanced machine learning and deep learning approaches, empowering organizations to make data-driven decisions that go beyond the

constraints of traditional methodologies. This research study's primary issue, which has many different aspects, relates to the need for developing a thorough understanding of Customer Lifetime Value (CLV) patterns in two particular geographic markets, namely the United Kingdom and Brazil, as well as the ability to predict and compare them. Despite the fact that both regions participate in the larger e-commerce market, they each have unique difficulties, possibilities, and customer dynamics that call for a specialized analytical approach. The main difficulty is in accurately predicting CLV for clients in both markets and in identifying the numerous variables that affect this important indicator in each market's environment. In order to better understand this issue, it is critical to understand how cultural quirks, economic inequalities, and different market dynamics all interact to influence consumer behavior in the UK and Brazil. The identification of these characteristics is crucial for companies working in these areas since it forms the basis for strategic marketing activities and resource allocation plans.

Given the complexity of contemporary consumer behavior, conventional statistical techniques that have long been used in the field of CLV analysis may no longer be adequate. The need of utilizing cutting-edge machine learning and deep learning techniques is thus highlighted by this research. With the initiation of these approaches, businesses may be able to obtain deeper, more complex perception from their data, encouraging them to take more strategic, well-targeted actions. By solving this issue, the research aims to bring down the gap between traditional CLV analysis and modern data-driven methodologies that are becoming highly important in today's powerful corporate climate. The main motive behind this study is that the existing research completely ignored the CLV analysis. Here have two e-commerce markets and it is possible to compare the Customer Lifetime Value structure between the two different markets. It helps them to understand, how to improve their way of marketing and how the clv trend is going. In previous papers, they undergone CLV analysis but a comparison between two countries or e-commerce is done by none of them. So, the current research is trying to compare two markets and the distribution of their customer's CLV.

Overall, this project aims to evaluate, understand and predict the CLV pattern of Brazil and UK e-commerce market by implementing both machine learning and deep learning techniques. It consists of two e-commerce datasets Brazilian e-commerce dataset and UK e-commerce dataset. The dataset then undergone preprocessing, feature engineering, model training, predicting, and comparison to determine the customer behavior and pattern. The research question and related aims and objectives is as follows:

Research Question:

How can a comparative analysis of machine learning and deep learning models be utilized to accurately predict and understand the Customer Lifetime Value (CLV) trends, considering the unique market dynamics and influential variables in both the UK and Brazil?

Aims and Objectives:

This study's main aim is to forecast and contrast CLV patterns in the UK and Brazil with the following objectives in mind:

- To create CLV predictive models through the use of machine learning and deep learning methods.
- To examine and contrast CLV trends between the UK and Brazil.
- To determine the key variables that have the greatest impact on CLV in each market.

Structure of the Document: the following section of the research includes, section 2. Related works which critically evaluates the previous works by reviewing and studying and how they contribute to our study. Section 3 is the methodology part which explains in detail about the data, preprocessing, feature engineering, model training and prediction. Section 4 is the design specification part and it specifies how the plan and methodology been outlined. Section 5 is the implementation. Here implement the methods and plan and will explain in detail about the process. Result and Discussion is the next part, section 6, where there explain the results in detail and discuss how it efficiently answers our research problem, aim and objectives. And in the final part 7, makes a conclusion about all findings that derived and what all future works can be done to the research that here initiated.

2 Related Work

This section of our study deeply evaluates the previous studies and explains how it relevant to our current study.

2.1 Global E-Commerce Market

The marketing pattern has widely changed globally with the introduction of electronic commerce. (Xiao et al.; 2017) in their research investigated about the growth of the e-commerce in European countries like UK, Germany and France through a professional perspective. European e-commerce started lately but exposed a huge potential growth, globally leading, high infrastructure and security measures. But the success of their e-commerce is sometimes varied because of the lifestyle and cultural variations across the countries. They also faced with the online payment complexities. (Xiao et al.; 2017) contributes realization of the growth of e-commerce in Europe. They expose the challenges with respect to cultural differences and diverse success models, which sheds light to future research. As the current research comparing the e-commerce pattern of two markets (Xiao et.al; 2017) made an overview of the European e-commerce markets by highlighting its strength and weakness. The future development in the field of e-commerce is focused by (Abid et al.;2020) and they researched by highlighting the advancements like cryptocurrency, which is a digital currency system and other innovations. UK, Norway, Finland, Korea and China have the largest dealers and buyers in the world. The study contributed an advanced knowledge in the field of e-commerce by forecasting the future development and (Abid et al.;2020) also suggests an in-depth exploration in the e-commerce platforms and investigate the advancements on online websites like c2c and their effect on the e-commerce markets. Overall, the studies show the need of advancement in the e-commerce market as it growing fast. Their payment preferences are also changed. Not only Europe, the global market is

emerging faster, customers preferring the online marketing than the offline shopping because of the ease of use, comfort and inevitable options and choices.

2.2 Customer Lifetime Value Analysis

Customer Lifetime Value Analysis is essential for an organisation for the growth of their profit, by analysing their valuable customers. (Kailash et al.; 2023) performed a CLV analysis on an IBM Watson dataset, modelling performed using the most prominent machine learning algorithms Linear Regression, Decision Tree, SVM and Random Forest. By analysing CLV (Kailash et al.; 2023) were able to increase their sales, improved product recommendation and improved the customer relationship. A further comparison between other markets will help them to improve their performance buy understanding their key points. (Kumar et al.; 2023) focused on the importance of CLV, improving Customer Relationship Management with the help of machine learning, found out that the machine learning performs better than traditional approaches, suggests advanced machine learning methods for more accurate CLV predictions. (Myburg et al.; 2022) introduces mainly XGBoost and K-means clustering and outperforming the traditional RFM models with an accuracy of 78% for CLV segmentation, usage of additional purchase data showed no improvement and should explore more comparative techniques. Another research by (Yean et al.; 2010) addresses the lack of integrated guidelines for CLV, highlights the requirement of CLV prediction and for further refinement outlines a research roadmap. (Gumber et al.;2021) for predicting customer behaviour used XGBoost in clickstream analysis, obtained an accuracy of 91.04%, emphasises the importance of understanding and predicting the customer behaviour, tried to achieve a balance between accuracy and model performance. A study by (Upadhyay, A., 2023) explains about the customer behaviour in UK and Brazil using RFM analysis. This study formed a base for our current research, while providing valuable insights, they hold some limitations like not exploring the CLV and advanced machine learning models. All these studies contributed as a baseline for the current study by analysing different modes of predicting CLV.

2.3 Machine Learning in E-Commerce

Machine Learning techniques helps in analyzing the trend happening in the e-commerce platform. (Alquhtani et al.; 2022) concentrated on the e-commerce reviews by analyzing the customer sentiments with the help of machine learning classifiers such as Support Vector Machine (SVM), Naïve Bayes, Logistic Regression and Neural Network and the objective is to identify the most effective classifier. The study provided a valuable insight to enhance the customer relationship and they lacked exploring factors influencing classifier performance and comparison. Similar study by (Zulfiker et al.; 2022) used six machine learning algorithms, Multinomial Naïve Bayes, Linear Regression, SVM, Decision Tree and Random Forest, obtained a highest accuracy of 90.68% and they failed to provide the impact of the cultural modulation on sentimental analysis. (Panwar et al.; 2021) did an Exploratory Data Analysis for the better implementation of machine learning techniques, so that many e-commerce datasets can be able to undergo through the initial implementation done by them.

Research investigated by (Hamsagayathri et al.; 2020) investigates about the women's clothing e-commerce using classifier algorithms using SVM and REPTree classifiers and they recommend the products based on the feedback of the customers. They obtained a classification accuracy of 91.43%, which represents a high precision. Validated the model using K-fold cross-validation and using confusion matrices, they were able to provide valuable information about the e-commerce market and there is a notable limitation on exploration of future directions and cross validation of e-commerce trends. Similar study conducted by (Win et al.;2023) is an e-commerce model conducted for Myanmar hotel industry, helped to understand the customer behaviour and improved the industrial quality, used machine learning algorithms and it supported the creation of better e-commerce model.

2.4 Deep Learning in E-commerce

(Subroto et al.; 2022) examines the effect of e-commerce demand prediction using Multi-Layer Perceptron (MLP) and XGBoost, finds that improved stacked generalization with MLP, provides reasonable insights for optimizing e-commerce prediction models. (Chen et al.;2018) utilized Convolutional Neural Networks (CNNs) to predict the customer lifetime value mainly in videogames, pointing out its accuracy and efficiency. Limited comparisons done by them. Overall, they contributed significant insights on deep learning techniques in the area of CLV prediction. (Xiao, 2020) explored about the information management in e-commerce using neural networks, and they enhanced the model performance, improved user satisfaction and sales hike. All these studies resulted as an overview of the relevance of deep learning in the e-commerce system and highlighted how deep learning can accurately predict the customer lifetime value.

3 Research Methodology

The research methodology part is explained here using the following figure.1 and it consist of data gathering, preprocessing, feature engineering, Exploratory Data Analysis (EDA), data modelling, prediction, evaluation and comparison.

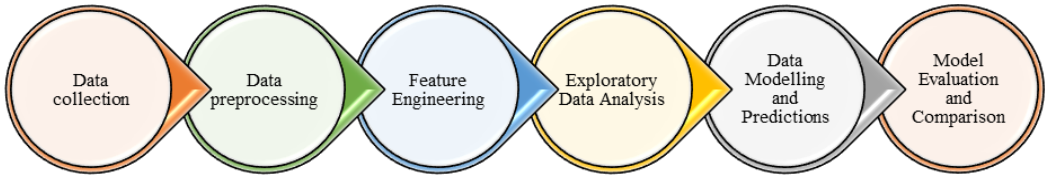


Figure.1. Methodology steps

Data Collection: Two e-commerce dataset is collected, the Brazilian e-commerce dataset and UK e-commerce dataset. This step is necessary and inevitable as it gathering and preprocessing the relevant datasets. The Brazilian dataset consist of 9 separate datasets consist of consumer details, specifically location, payments, order details, seller information and transactions. And the UK dataset consist of 8 features with 541910 rows.

Data Preprocessing: In the data analysis pipeline data preprocessing is a critical step. The selected data sets are cleaned, removed the duplicates and handled the missing values. Here in the case of Brazilian datasets the preprocessing is performed on the customer orders related datasets. It includes handling the duplicates by identifying the duplicate rows. Handled the missing values by dropping method. The data also merged for a comprehensive dataset for further analysis. Removed the outliers. Features like year, month, day and hour are extracted from the date columns to enable better evaluation. The one hot encoding is implemented to format the categorical features like order status and payment type that is suitable for the machine learning models. For the UK dataset, the data is filtered and only included the United Kingdom records. Also, the rows with null values dropped.

Feature Engineering: Feature engineering involves adding new characteristics or altering preexisting ones to improve the dataset's quality and suitability for modelling. Let's go over the feature engineering procedures for the datasets from Brazil and the UK. In the Brazilian dataset, Time Stamp Conversion is performed to extract additional temporal information, date columns (like order_purchase_timestamp) are turned to datetime objects. By combining the dataset according to customer_id and adding the payment_value for every customer, the Customer Lifetime Value (CLV) is determined. To evaluate the distribution of CLV over several categorical variables, including payment_type and order_status, boxplots are made.

In the case of UK dataset, one new characteristic, TotalPurchase, is generated by multiplying the unit price (UnitPrice) by the quantity (Quantity). The dataset is arranged by CustomerID, and each customer's CLV is estimated by adding their TotalPurchase amounts. To see the distribution of CLV over time, categorized by invoice year and invoice month, boxplots have been generated.

Feature engineering in both datasets entails generating extra temporal features, computing CLV, and displaying the distribution of CLV along several dimensions. The information provided by these well-designed elements is crucial for comprehending consumer behaviour and enhancing the precision of CLV forecasts.

Exploratory Data Analysis: Understanding a dataset's characteristics and trends requires implementing exploratory data analysis (EDA). EDA is performed for the Brazil and UK datasets. Initially in the Brazilian dataset, basic dataset information, such as data types and non-null counts, can be shown using the info () function. To visualize the distribution of order payment values a histogram is created. To depict the distribution of order status, a bar plot is created. Visualized a bar plot for the distribution of payment methods. To understand the distribution of customer states in Brazil, another bar plot is created. To show how Customer Lifetime Value (CLV) varies according on order status and payment type, boxplots are made. The correlation matrix between numerical features is shown via a heatmap.

The UK dataset is examined for duplicate rows and missing variables. A heatmap is used to build and display the correlation matrix. To see the Customer Lifetime Value (CLV) distribution, a histogram is made. A greater understanding is made possible by these EDA procedures, which offer insights into the distribution of important variables, trends over time, and correlations between characteristics.

Data Modelling and Predictions: Both the machine learning and deep learning techniques are implemented for predicting the best model. For evaluating the Brazilian e-commerce trained some of the machine learning models using XGB regressor, Support Vector Machine

(SVM) and Random Forest. And for deep learning, implemented the MLP regressor. For the UK e-commerce trained the same algorithms in machine learning and for the deep learning part. And examined these models using the R-squared value, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE).

Model Evaluation and Comparison: Customer Lifetime Value (CLV) prediction accuracy of the Random Forest model is the basis for evaluation, because it performs better than other algorithms. The ability of the MLP Regressor to forecast CLV is evaluated. Plotting the expected and actual CLV trends over time is done using both the Random Forest and Deep Learning model. For subsequent use, the trained Random Forest model is stored. Based on its capacity to forecast Customer Lifetime Value (CLV) for the UK dataset, the Random Forest model is assessed. The effectiveness of the MLP Regressor in forecasting CLV for the UK dataset is evaluated. Plotting real CLV trends against anticipated CLV trends over time is done using both the Random Forest and Deep Learning models. In general, the trend analysis and assessment metrics offer a thorough understanding of how well the models forecast CLV for the datasets from Brazil and the UK. By comparing the models, one can determine which one most efficiently fits the features of the dataset by understanding about the advantages and disadvantages of each.

4 Design Specification

In this section deeply explains the research methodology. In order to enable other researchers to perform the study and guarantee the validity and rigor of the research findings, the design specification offers a thorough and clear overview of the research technique.

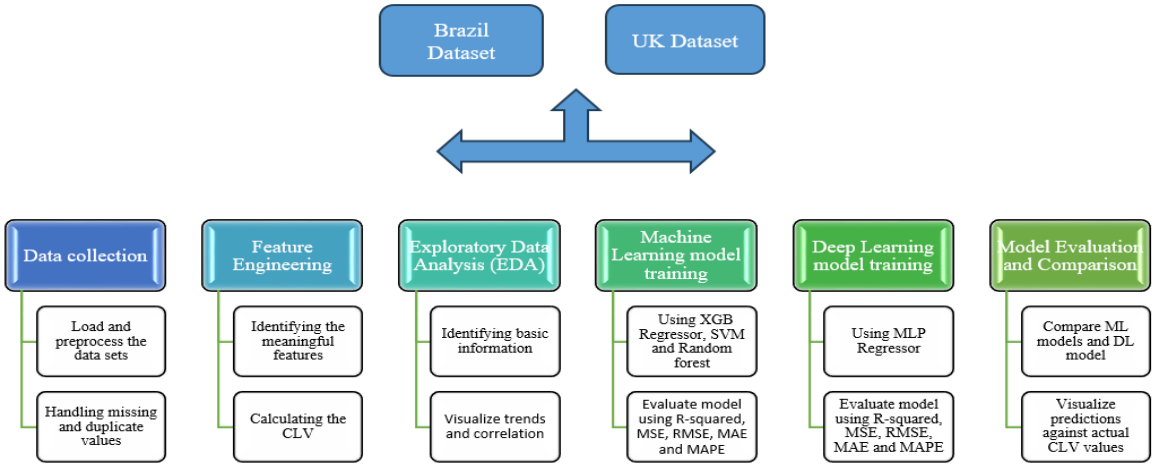


Figure.2. Design of the Process

The above figure 2 depicts the overall implementations from the data collection to the model evaluation, comparisons and results. Firstly, the Brazilian e-commerce data set, obtained from Kaggle, which is a csv file with 9 separate csv files, contains details on 100,000 orders placed at several Brazilian marketplaces between 2016 and 2018. According to its variables, one can examine an order from a number of perspectives, including order status, pricing, payment, and freight performance, as well as customer location, product

qualities, and customer reviews. A geolocation dataset that connects Brazilian zip codes to lat/lng coordinates was also made public. The UK dataset obtained from UCI Machine Learning Repository website, is an international data collection that includes every transaction made by a UK-based, registered online retailer without a physical store between December 1, 2010, and September 12, 2011.The company primarily offers special gifts for every occasion. A huge part of this company's clients are wholesalers. The collected data loaded and then pre-processed, handled the missing and duplicate values.

As part of the feature engineering, in both datasets requires developing additional temporal features, computing CLV, and displaying the distribution of CLV along different dimensions. The information suggested by these elements is pivotal for realise the consumer behaviour and improving the accuracy of CLV forecasts. The validation of Customer Lifetime Value (CLV) may vary based on the particular objectives and business scenario. In general, CLV represents the approximate total income that a business can predict, receiving from a customer over the period of their partnership. The CLV for Brazil is obtained by adding the 'payment_value' for every customer. In this case, the payment values for each client are added together to obtain the CLV. The CLV is the total of these transactional values for a particular client, and it is considered that the 'payment_value' indicates the revenue earned from each transaction. CLV for UK is examined using the total amount of purchases done by each consumer. In this instance, the quantity multiplied by the unit price, or "TotalPurchase," is used to calculate the CLV. The CLV is the sum of these transactional values for each customer, and it is considered that the 'TotalPurchase' represents the revenue made from each transaction. A relevant exploratory data analysis is performed to evaluate the basic information and visualized the trends and correlation.

After the EDA move on to predicting the model using the machine learning and deep learning model. As figure 2 represents, both datasets are subjected to machine learning and deep learning techniques. Initially, three algorithms were implemented and selected the best algorithm among them to find the best model, as part of the machine learning model training. Here implemented the XGB Regressor, SVM and Random Forest. Also implemented the Multi-Layer Perceptron (MLP) Regressor for the deep learning model training. Regression can be examined by this neural network called MLP Regressor. Regression in the context of machine learning can be approached as a mapping between two spaces, each of it may have arbitrary dimensions. After that they subjected to evaluate using the R-squared value, MSE, RMSE, MAE and MAPE. A relevant comparison between them is gathered using these matrices and visualized these trends for the better comparison. By these it is possible to come to know how can a comparative analysis of machine learning and deep learning models be utilized to accurately predict and understand the Customer Lifetime Value (CLV) trends, considering the unique market dynamics and influential variables in both the UK and Brazil.

5 Implementation

The data collection, experimentation, survey administration, and execution of any other planned procedures take place within the implementation phase.

We can start with the **Brazil dataset**. As already mentioned, the dataset consists of 9 separate datasets. The data is a publicly available dataset and it is collected from Kaggle. We implemented all the analysis in python using Jupiter notebook and imported the necessary and proper libraries such as Pandas, numpy, matplotlib.pyplot, seaborn, xgboost, sklearn.metrics,sklearn.model_selection,sklearn.preprocessing,sklearn.compose,sklearn.pipeline,sklearn.ensemble.RandomForestRegressor, sklearn.neural_network.MLPRegressor, joblib, tensorflow.keras.models.sequential, and tensorflow.keras.layers.Dense. Initially we load the datasets that is necessary and useful for our analysis. Dropped duplicate rows using drop_duplicates and handled the missing values using dropna.

```
# Load the datasets
customers = pd.read_csv('olist_customers_dataset.csv')
geolocation = pd.read_csv('olist_geolocation_dataset.csv')
order_items = pd.read_csv('olist_order_items_dataset.csv')
order_payments = pd.read_csv('olist_order_payments_dataset.csv')
order_reviews = pd.read_csv('olist_order_reviews_dataset.csv')
orders = pd.read_csv('olist_orders_dataset.csv')
```

Figure.3. Loaded Datasets

The above figure 3 shows the datasets we loaded out of the 9 datasets. We have selected only 6 out of the 9 datasets, that are relevant to our research. After cleaning the data, we merged the above datasets in to a single dataset for more comprehensive analysis. We removed the outliers for the better model performance. As part of the feature engineering, we selected the relevant features. They are, 'customer_id', 'order_purchase_timestamp', 'price', 'payment_value' and we calculated the CLV.

```
# Select relevant features and target variable (CLV)
features = merged_data[['customer_id', 'order_purchase_timestamp', 'price', 'payment_value']]
target = merged_data[['customer_id', 'payment_value']]

# Calculate CLV for each customer
clv = target.groupby('customer_id')['payment_value'].sum().reset_index()
clv.rename(columns={'payment_value': 'CLV'}, inplace=True)
```

Figure.4. CLV Calculation

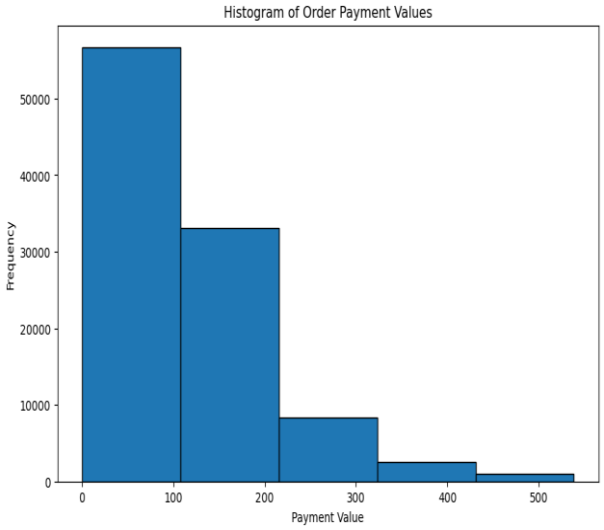
Figure 4 shows how the CLV is calculated. As part of this, calculated the sum of the payment values for each customer and it substantially aggregates total payment value for each customer. It also groups the target variable by ‘customer_id’. After calculating the CLV, merged the CLV back to the data. As a result, we can view the CLV of each customer along with the data. Exploratory Data Analysis (EDA) is the next inevitable part of the analysis. Started by the basic statistical information analyzed by the info().

```
Int64Index: 101783 entries, 0 to 115017
Data columns (total 23 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_id                       101783 non-null  object
 1   customer_id                    101783 non-null  object
 2   order_status                   101783 non-null  object
 3   order_purchase_timestamp       101783 non-null  object
 4   order_approved_at              101783 non-null  object
 5   order_delivered_carrier_date   101783 non-null  object
 6   order_delivered_customer_date  101783 non-null  object
 7   order_estimated_delivery_date  101783 non-null  object
 8   customer_unique_id             101783 non-null  object
 9   customer_zip_code_prefix       101783 non-null  int64
 10  customer_city                  101783 non-null  object
 11  customer_state                 101783 non-null  object
 12  order_item_id                  101783 non-null  int64
 13  product_id                     101783 non-null  object
 14  seller_id                      101783 non-null  object
 15  shipping_limit_date            101783 non-null  object
 16  price                          101783 non-null  float64
 17  freight_value                  101783 non-null  float64
 18  payment_sequential             101783 non-null  int64
 19  payment_type                   101783 non-null  object
 20  payment_installments           101783 non-null  int64
 21  payment_value                  101783 non-null  float64
 22  CLV                            101783 non-null  float64
dtypes: float64(4), int64(4), object(15)
memory usage: 18.6+ MB
```

Figure.5. Basic Statistical Information

In figure 5 we can see 22 features including the customer CLV. There are 101783 entries for each, and all of them are non-null values. We removed outliers from our data for better performance.

As part of the EDA, we performed some basic visualizations as well to analyze the data. Firstly, a histogram depicting the order payment values.



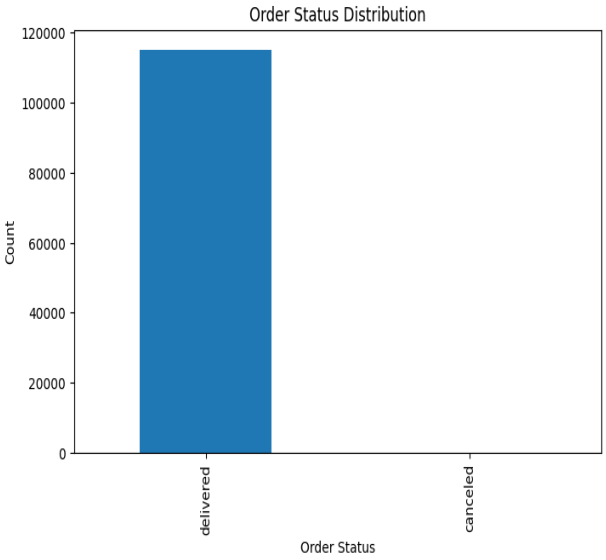


Figure.6. Distribution of Order payment values

Figure.7. Order status distribution

The above figure 6 representing a histogram indicating the distribution of payment values. Here the X axis represents the payment value and Y axis representing its corresponding frequency. The figure itself shows like skewed distribution with the customer's payment is highly within the range from 0 to 100. And the frequency is about more than 50000. More than 30000 customers making payment between the range of 100 to 200. Below 10,000 making payments between 200 to 300. Only a small portion is within 400 to 500 range. This histogram represents the visualization of payment values that makes us understand where all the majority of transactions fall within a certain range. The payment is distributed mainly on smaller amounts.

The figure 7 is a bar plot representing the order status. The X axis indicating whether it is delivered or cancelled. Y axis represents the count of the order. It's simply visible that there is nothing of the order cancelled. The data itself consist of the orders that is delivered. There are more than 100000 orders that is delivered. This bar plot helps us to understand that is there any orders cancelled by the customers, and it showing no cancellation.

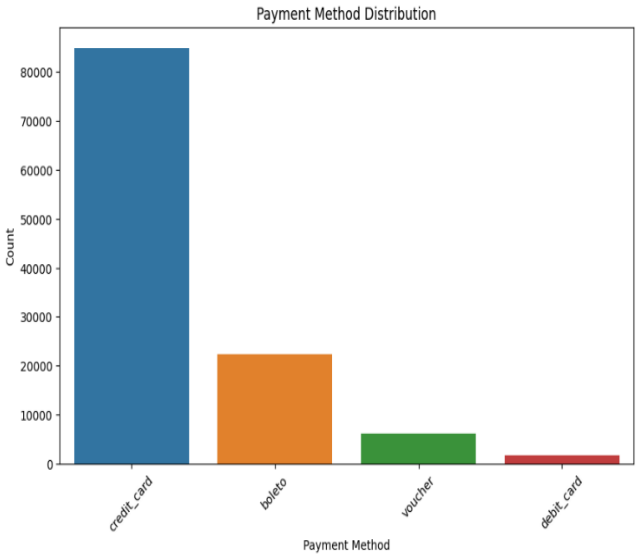


Figure.8. Payment Method Distribution

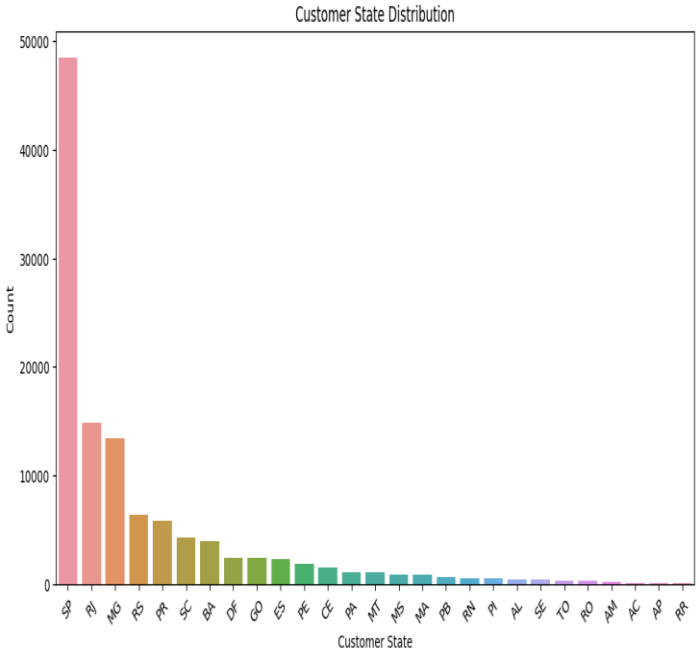


Figure.9. Customer State Distribution

Figure .8 showing the distribution of payment methods that the customer made. Most of the customers used credit card for their transaction. More than 80,000 transactions are made through credit card. Second largest is through boleto. Boleto is a form of payment method in Brazil regulated by the Central Bank of Brazil. More than 20,000 transactions are through the boleto payments. Less than 10,000 transactions made through the voucher and only a few percent done through the debit card. Credit card is the mostly used payment option and debit card is the least preferred payment option.

Figure.9. is a bar graph showing the customer state distribution. This Brazilian E-commerce dataset consists of 27 states in Brazil. The X axis consists of customer states and Y axis represents the count. The most of the customers is from the state SP, indicating Sao Paulo, a state in southeastern Brazil. this state representing a count more than 40,000. The second largest is RJ, indicating Rio de Janeiro is about 15,000. And the third largest is MG (Minas Gerais). Rest of the states representing a smaller portion. In general Sao Paulo is the state representing majority of the customers in the Brazilian e-commerce dataset.

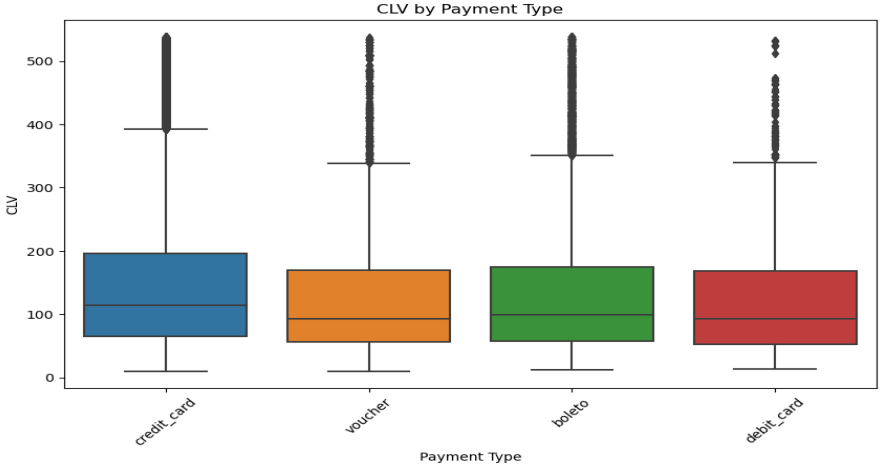


Figure. 10. Boxplot of CLV by Payment Type

By analysing this figure.10, boxplot depicting the CLV by payment type. Here The CLV range for credit card payments is greater. With no outliers and a smaller CLV range, voucher payments show a more uniform CLV distribution. Boleto payments exhibit almost same kind of CLV distribution like voucher payments. Like vouchers, debit card payments have a limited CLV range with likely smaller outliers. Here it shows that the payment method also be a part of CLV distribution like, for example, customers using credit card might make more purchases or make frequent transactions which will show a hike in their CLV.

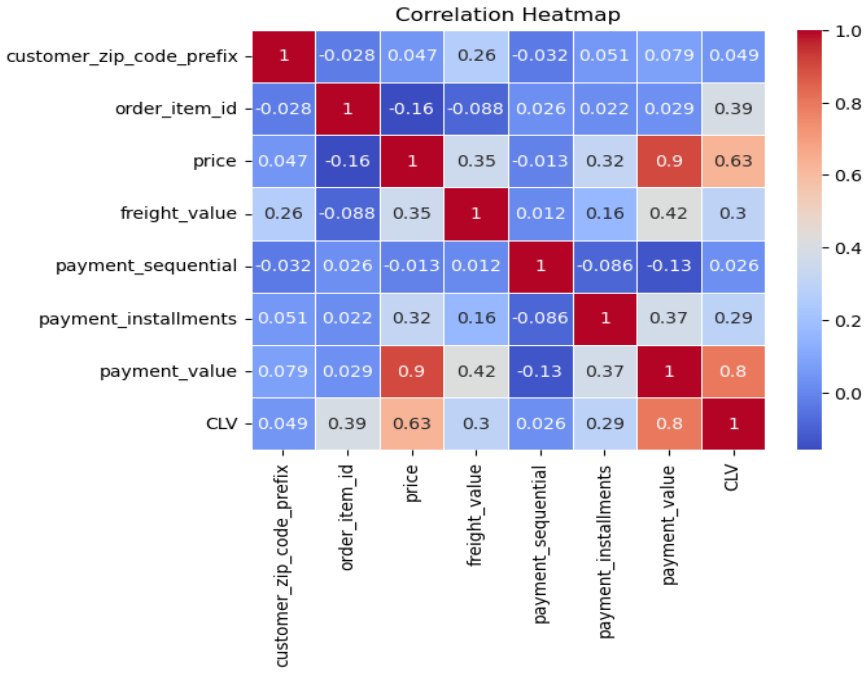


Figure.11. Heat Map

A heatmap is a form of visualization which use colors to represent the values in a correlation matrix and it is showed in figure.11. The color dark red denotes a highly positive association. The hue turns a deeper shade of red as the correlation gets closer to +1. Dark blue denotes a highly inverse relationship. The color becomes a deeper shade of blue as the correlation gets closer to -1. White denotes no correlation (or a correlation that is almost zero). Features like payment value and price positively correlated to the CLV. Payment sequential and order item id indicating a highly inverse relationship, freight value and customer zip code are negatively correlated. Here it is visible that the payment value and price contributing highly to the CLV distribution of each customers making orders from this e-commerce organization.

For analyzing the **UK dataset**, as already mentioned it consist of 8 features with 541910 rows. Implemented all the analysis in the python using Jupiter notebook and imported the necessary libraries such as pandas, numpy, matplotlib.pyplot, seaborn, xgboost. Initially load the datasets that is necessary and useful for our analysis. Dropped duplicate rows using drop_duplicates and handled the missing values using dropna. After the preprocessing calculated the CLV. Also tried to remove the outliers from our data.

```
# Group by 'CustomerID' and calculate the sum of 'TotalPurchase' as CLV
features = data_uk.groupby('CustomerID')['TotalPurchase'].sum().reset_index()
features.rename(columns={'TotalPurchase': 'CLV'}, inplace=True)

# Merge CLV back into the original dataset
data_uk = pd.merge(data_uk, features, on='CustomerID')
```

Figure.12. Calculation of CLV

Above shows how calculated the CLV for the UK e-commerce. The CLV is calculated by adding up the 'Total Purchase' for each of the customers. After merging back, the CLV to the dataset, saved the preprocessed data as a csv file. This generated 'data_uk' The CLV values for UK customers are contained in the Data Frame. This CLV provides an estimate of the customer's worth to the firm over the specified period and is based on the total purchase amounts per customer.

The next most important step is the EDA, for that like the Brazilian dataset analysed basic statistics using info() and performed some important visualization that can interpret and study about the data.

```
RangeIndex: 361807 entries, 0 to 361806
Data columns (total 10 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   InvoiceNo      361807 non-null  object
 1   StockCode      361807 non-null  object
 2   Description    361807 non-null  object
 3   Quantity       361807 non-null  int64
 4   InvoiceDate    361807 non-null  object
 5   UnitPrice      361807 non-null  float64
 6   CustomerID     361807 non-null  float64
 7   Country        361807 non-null  object
 8   TotalPurchase  361807 non-null  float64
 9   CLV            361807 non-null  float64
dtypes: float64(4), int64(1), object(5)
memory usage: 27.6+ MB
```

Figure.13. Basic Statistical Info

Figure 13 shows that there are 9 variables that are non-null and each variable consist of 36,1807 entries. The 9th variable is the CLV that calculated for each customer. Below can view the visualizations as part of the EDA.

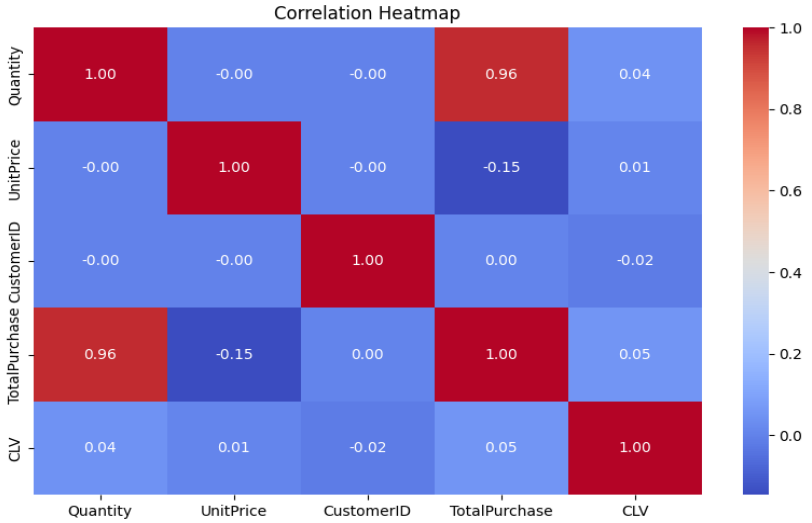


Figure .14. Correlation Heat Map

Figure 14 shows that the total purchase and quantity have a strong positive correlation and strong negative correlation between unit price and total purchase. Here considered the features like quantity, unit price, customer id, total purchase and CLV. A heatmap is a data visualization where colors can represent the values in a correlation matrix. The color dark red

denotes a highly positive association. The hue turns a deeper shade of red as the correlation gets closer to +1. Dark blue denotes a highly inverse relationship. The color becomes a deeper shade of blue as the correlation gets closer to -1.

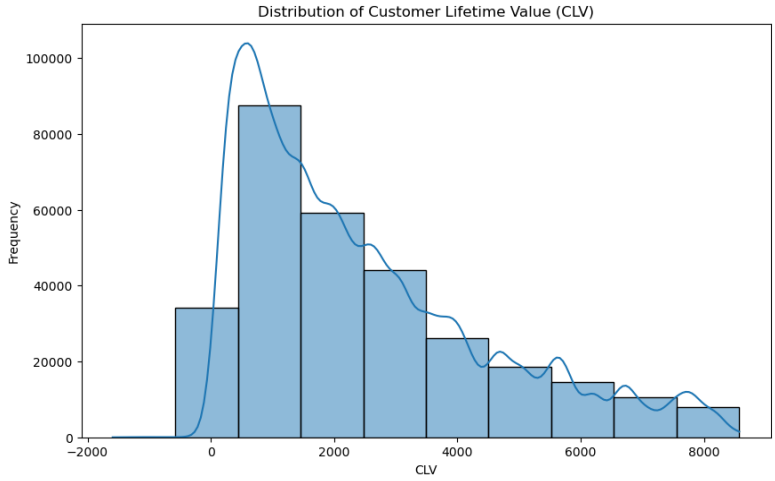


Figure.15. Distribution of CLV (UK)

Here the figure.15 depicting the CLV distribution of UK customers. The X axis represents the Customer Lifetime Value (CLV) and Y axis representing the corresponding frequency. The diagram itself shows a skewness as most of the value is representing below 2000. Second largely, between 2000 is distributed. Thirdly, distributed between 2000 to 4000. After that a smaller portion of the value displayed between the ranges 4000 to 6000 and 6000 to 8000. In general, most of the customer's CLV is distributed between 2000.

As the plot is reasonably skewed, tried to perform log transformed target plot.

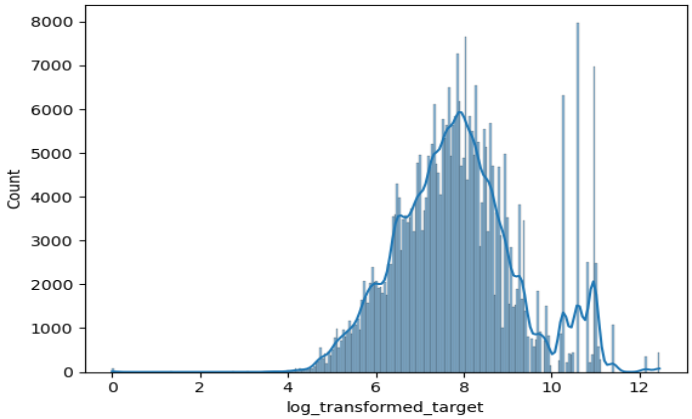


Figure.16. Log transformed target plot for CLV

Here in figure 16 tried to normalize our target variable CLV using the log transformed target variable plot. Now the variable is transformed from skewed distribution to a normally distributed variable.

Predicting the Model using Machine Learning and Deep Learning (UK and Brazil):

In the research, have used three different algorithms xgb Regressor, SVM and Random Forest in the machine learning model training. Extreme Gradient Boosting, or XGBoost, is an esteemed and potential machine learning technique. SVM is a flexible approach that can be

applied to assignments involving regression and classification. Finding a hyperplane that estimates the relationship between input features and target values is the goal of SVM in regression. A technique known as ensemble learning called Random Forest creates a large number of decision trees during training and outputs the mean prediction (regression) or mode of the classes (classification) of each individual tree. When compared to individual trees, it increases accuracy and decreases overfitting.

In the deep learning section, implemented a kind of feedforward neural network that comprises of an input layer, one or more hidden layers, an output layer known as Multi-Layer Perceptron. MLP is trained to realize the relationship between input features and continuous target variables in terms of regression. Below shows the R-Squared value, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) value accured for the indicated algorithms.

Table.1. Performance of Algorithms (Brazil)

Model	R-Squared	MSE	RMSE	MAE	MAPE
XGB Regressor	0.97	263.89	16.24	4.27	0.03
Support vector Machine	0.82	2162.55	46.50	17.06	0.03
Random Forest	0.98	252.11	15.87	2.60	0.01
MLP Regressor (DL)	0.97	283.15	16.82	5.14	0.04

Table.2. Performance of Algorithms (UK)

Model	R-Squared	MSE	RMSE	MAE
XGB Regressor	0.39	165732831.23	12873	6725
Random Forest	0.60	109761492.35	10476	4875.72
MLP Regressor (DL)	0.59	111248700.45	10547	4884

Above table 1 and 2 shows the values obtained for each algorithm to predict the best CLV model for both UK and Brazil. Further the evaluation of the model can be done in the following section.

6 Evaluation

The evaluation part is as important as the implementation part. We can start with evaluating the performance of Brazil.

6.1 Evaluation of Brazil:

From Table.1 it is visible that the Random Forest algorithm is performing efficiently when comparing to other algorithms. It showed a higher accuracy and least number of errors, like lower MSE, RMSE, MAE and MAPE. Good fit to the data is evident by a high R-squared. Accurate predictions are indicated by the other performance metrices. XGB Regressor and SVM also performing well. SVM showed a lower accuracy compared to other ones. Still by considering the accuracy and error rates, Random Forest is highlighted by its performance and it can have good explanatory power than the other models.

The Deep Learning algorithm MLP Regressor also performing well for the data with high accurate value and indicates best fit and good performance. The MLP Regressor (DL) has the highest R-squared (0.97), meaning that the model accounts for around 97% of the variance in CLV. In general, all the algorithms performed accurately for Brazil with lower error rates. Both the Machine learning and Deep Learning algorithms is able to perform efficiently to predict better CLV.

6.2 Evaluation of UK:

Table.2 displays the performance of algorithms for UK. Here, when comparing to Brazil, algorithms showed an unsatisfactory performance. The accuracy is lower and error rates are high. The algorithms were unable to predict the model accurately. Here also the Random Forest model is performing better than the other algorithms. There shows a moderate performance, suggesting that these models account for a considerable amount of the variance in CLV. the XGB Regressor appears to explain less of the variance in CLV in the UK with a lower R-Squared.

In the case of Deep Learning, the MLP regressor also showed a moderate performance, suggesting that it also able to represent the model moderately.

General Findings: In terms of R-squared, MSE, RMSE, MAE and MAPE the Random Forest (Machine Learning) and MLP Regressor (Deep Learning) continuously beats other models for Brazil. R-squared values are greater in Brazil than in the UK, which suggests that the models there fit the data better. Although Random Forest does well in both nations, it outperforms XGB in the UK. UK is where SVM performs the worst, as seen by its high error rates and poor R-squared.

Now move on to examine how the CLV trends is distributed over time by comparing both the e-commerce department. The following figures shows how the CLV trends is distributed using the actual CLV and predictions. Evaluate CLV trends using ML algorithm Random Forest, because it showed greater accuracy and lower errors. Also evaluate using the deep learning algorithm MLP Regressor. This will help us to identify which technique (Machine Learning and Deep Learning) performs better.

6.3 CLV trends over time (Brazil)

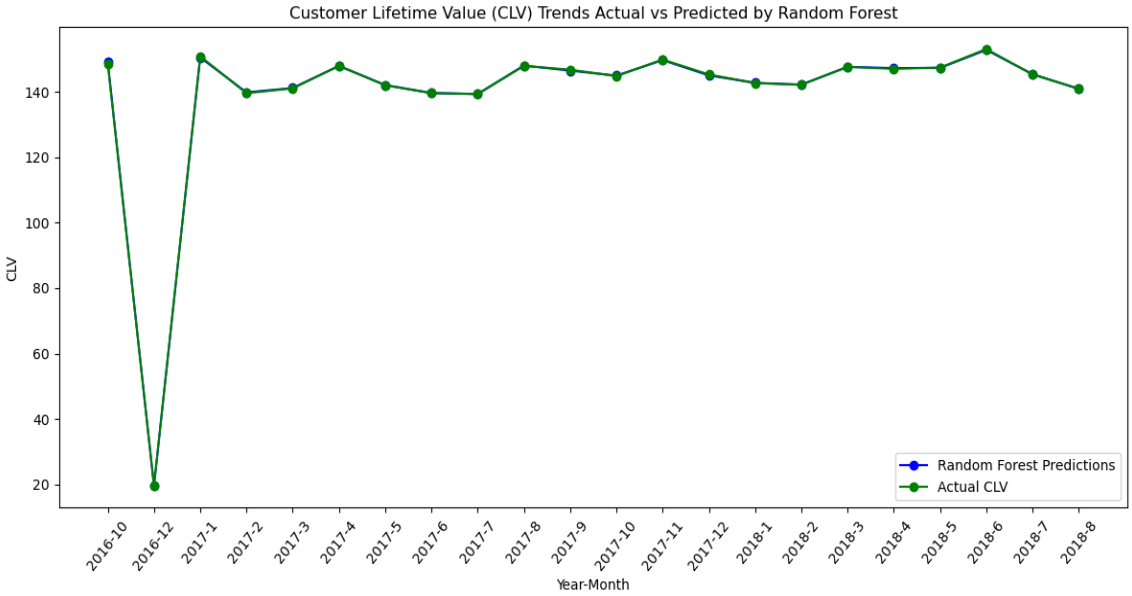


Figure.17. CLV trends Actual and Predicted Values (Random Forest)

The figure.17 depicted the CLV trends using the actual and predicted values by Random Forest. The X axis represents the year along with month and Y axis represents the corresponding CLV. There we can see up and down of CLV across time periods. Initially CLV is at its peak in 2016, October and after that a huge decline happened in December, then inclined in 2017, January. After that the CLV is not declining too much or inclining. Both the lines of actual CLV and predicted CLV is going on the same pattern and their similarity depicts that the model is performing well and it reminds that the factors influencing CLV is trained well. The trend of the line representing Random Forest forecasts and the real CLV are comparable. The general trends in CLV variations appear to be captured by the model.

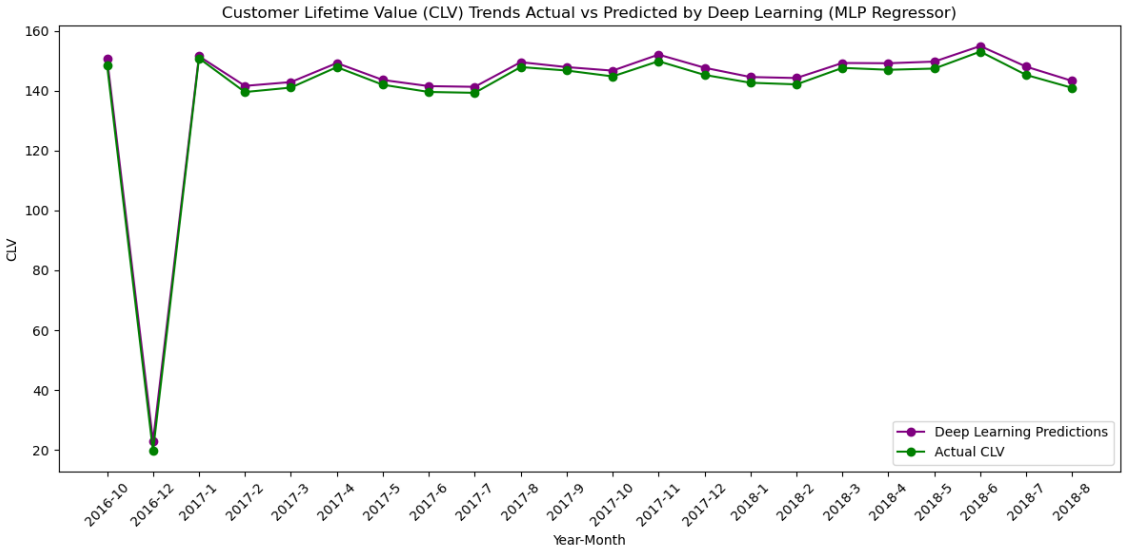


Figure.18. CLV trends Actual and Predicted Values (MLP Regressor)

Figure 18 is all about the CLV trends corresponding to actual to predicted values using MLP regressor. The X axis represents the year along with month and Y axis represents the corresponding CLV. When compared to the actual CLV, the MLP Regressor model frequently undervalues or predicts lower CLV values. Here the deep learning predictions outperforms than the actual CLV. Like the Random Forest predictions, here also the CLV is higher initially and then a huge decline happened in 2016, December, inclined after the next year. Then maintained a pattern of not having a huge decline. Showed continuously inclining and declining. The deep learning model is able to perform well than the actual CLV.

6.4 CLV trends over time (United Kingdom)

Now here able to see how the CLV trend pattern in the UK e-commerce, the figures are shown below.

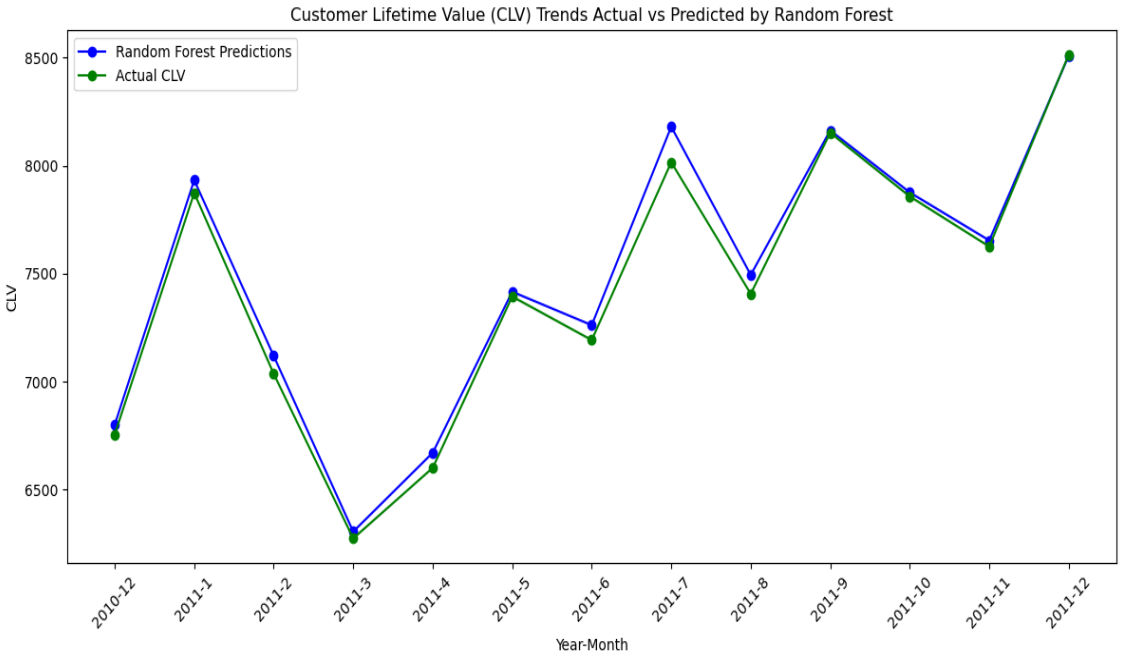


Figure.19. CLV trends Actual and Predicted Values (Random Forest)

Figure 19 displays the CLV trends by actual and predicted values using random forest algorithm. The X axis represents the year along with month and Y axis represents the corresponding CLV. Initially in 2010 showing lower CLV, later in 2011, January showing a hike in CLV. Then there is a huge decline happened in march of 2011 then simultaneously increased the and decreased. At the end showing a hike. Basically, this figure showing a growth in their CLV of the customers. Here the Random Forest is able to predict better than the actual CLV as it showing an increased value, which means random forest performing well in predicting the CLV.

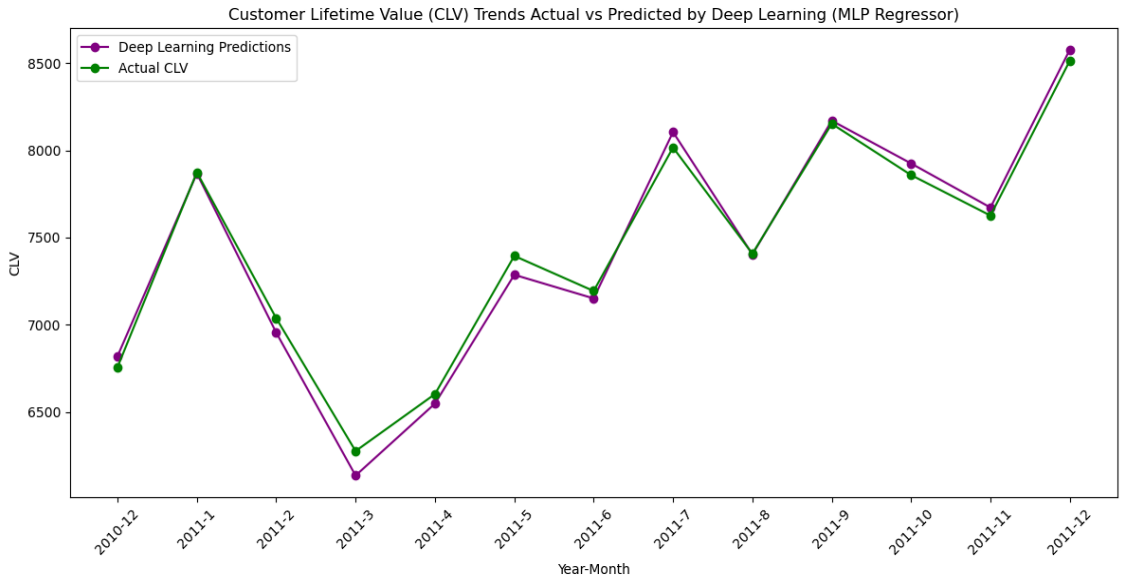


Figure.20. CLV trends Actual and Predicted Values (MLP Regressor)

In the figure.20, the X axis represents the year and Y axis represents the CLV values. The X axis represents the year along with month and Y axis represents the corresponding CLV. The above figure shows that the actual CLV performing well than the MLP Regressor at certain point, and after that the predicted CLV outperforms to predict the CLV trends. The deep learning predictions underperforms compared to the actual CLV. But both the lines going on the same pattern. The CLV is fluctuating throughout the years. The CLV is lower in the year 2011, march and higher in the year 2011, December. Overall showing a growth in the CLV of their customers as it is at its peak at last.

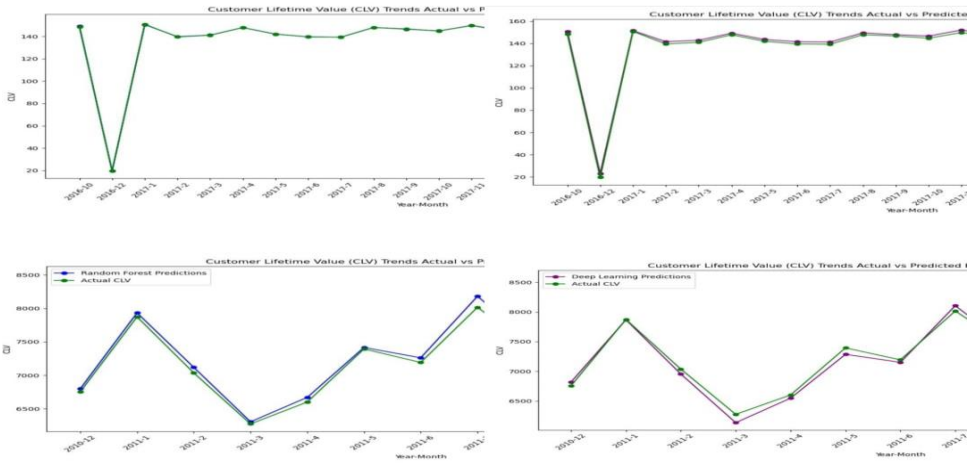


Figure.21. An overall combined view of the CLV trends between actual and predicted values of both Brazil and UK

The figure 21 displays and aims to receive an overall combined view of the trends. The above two plots in this figure.21 are CLV trends of Brazil by machine learning and deep learning,

and the below two plots in figure.21 are CLV trends of UK by machine learning and deep learning.

6.5 Discussion

Both the Machine learning and Deep Learning models performed well for predicting the CLV in Brazil and UK e-commerce and it contributed well for understanding the importance and effectiveness of these models. Random Forest performs well with high accuracy and lower error rates for Brazil and stands as a good fit for the data. The MLP Regressor in Deep Learning also showed higher accuracy and reduced error rates. Our finding is that, both the Machine Learning and deep Learning techniques can effectively predict the CLV of customers for the Brazilian e-commerce market.

When considering the UK e-commerce, the performance of the algorithms is unsatisfactory with low accuracy and high level of error rates. Here also the Random Forest performed better, but the accuracy level is moderate. Tried to improve the performance by an effective preprocessing, data analysis, feature engineering, removed outliers, hyperparametric tuning, still the result appears to be less effective. The difference in both market's performance is based on the consumer behavior, market dynamics and some other external factors. The CLV trends over time for both countries, revealed Random Forest having impressive capability to predict the growth of the CLV.

The major limitation of the study is the under performance of the UK, which needs extended evaluation. The UK dataset consist only limited features. So, it's better to include additional dataset or relevant features to improve model performance.

7 Conclusion and Future Work

Considering our research question and objectives, made some conclusions and recommending some future works. The research question is about how can a comparative analysis of machine learning and deep learning models be utilized to accurately predict and understand the Customer Lifetime Value (CLV) trends, considering the unique market dynamics and influential variables in both the UK and Brazil. And our objectives are to create CLV predictive models through the use of machine learning and deep learning methods, to examine and contrast CLV trends between the UK and Brazil and to determine the key variables that have the greatest impact on CLV in each market.

For that compared both the markets by calculating each customer CLV and predicted the model using both machine learning and deep learning techniques. Analyzed the trend by plotting the actual CLV and predicted CLV. From that made the following conclusions,

- Created predictive models for both markets using machine learning and deep learning techniques.
- For both Brazil and the UK e-commerce market stream, the Random Forest model gave better accuracy and lower errors than the Support Vector Machine (SVM) and XGBoost (XGB) models in terms of accuracy. The Customer Lifetime Value (CLV) for both nations were accurately predicted by Random Forest (Machine Learning) with good performance.

- All machine learning models (SVM, XGB, and Random Forest) in Brazil displayed quite high R-squared values, signifying a strong match with the data. In contrast to the other models, the SVM model's R-squared value was noticeably lower. R-squared values suggest a reasonable match to the data, and machine learning models Random Forest performed moderately well in the UK.
- When compared to machine learning models, the Deep Learning model (MLP Regressor) consistently showed good performance in Brazil with higher R-squared values and lower MSE and a moderate performance Showed in UK. This shows that the deep learning method could be able to better capture intricate patterns in CLV.
- The examination of order statuses, payment methods, and CLV trends gave insights into the variables affecting CLV in Brazil and the UK. Boxplots and trend analysis showed that the effects of payment types and order statuses on CLV seemed to differ. In Brazil, People using credit card showed high CLV rates than others. Price and payment value are the variables effecting CLV. In UK, quantity and total purchase effecting CLV.
- Brazil e-commerce showing better performance than UK. UK need to capture the marketing strategies of Brazil and needs improvements.

FUTURE WORK: Both the datasets are large enough and especially Brazil contains several features, suitable for sentimental analysis. The accuracy of CLV predictions may be improved by investigating more attributes and data sources. Model performance may be enhanced by adjusting hyperparameters and experimenting with various model techniques.

References

Xiao, Z., 2017. The development of e-commerce in Europe.

Abid, S., 2020. Future of e-commerce: An analysis of ecommerce in retail business. *International Research Journal of Electronics and Computer Engineering*, *5*(4), pp.10-14.

Kailash, H., Kanwar, K., Sonia, S. and Kant, R., 2023, May. Machine Learning Algorithms for Predicting Customers' Lifetime Value: A Systematic Evaluation. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 538-541). IEEE.

Kumar, A., Singh, K. U., Kumar, G., Choudhury, T., and Kotecha, K., 2023. Customer Lifetime Value Prediction: Using Machine Learning to Forecast CLV and Enhance Customer Relationship Management. In *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-7). IEEE.

Myburg, M. and Berman, S., 2022, November. Customer Lifetime Value Prediction with K-means Clustering and XGBoost. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 298-302). IEEE.

Yean, L.C. and Khoo, V.K., 2010, May. Customer relationship management: lifecycle of predicting customer lifetime value. In *2010 Second International Conference on Computer Research and Development* (pp. 88-92). IEEE.

Gumber, M., Jain, A. and Amutha, A.L., 2021, May. Predicting Customer Behavior by Analyzing Clickstream Data. In *2021 5th International Conference on Computer, Communication and Signal Processing (ICCCSP)* (pp. 1-6). IEEE.

Upadhyay, A., 2023. *An analysis of e-commerce purchase behaviour across the UK and Brazil* (Doctoral dissertation, Dublin, National College of Ireland).

Alquhtani, S.A. and Muniasamy, A., 2022, July. Analytics in Support of E-Commerce Systems Using Machine Learning. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)* (pp. 1-5). IEEE.

Zulfiker, S., Chowdhury, A., Roy, D., Datta, S. and Momen, S., 2022, December. Bangla E-Commerce Sentiment Analysis Using Machine Learning Approach. In *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)* (pp. 1-5). IEEE.

Panwar, M., Wadhwa, A. and Pippal, S., 2021, December. An Overview: Exploratory Data Analytics on E-Commerce Dataset. In *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 91-93). IEEE.

Hamsagayathri, P. and Rajakumari, K., 2020, January. Machine learning algorithms to empower Indian women entrepreneur in E-commerce clothing. In *2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-5). IEEE.

Win, T.N. and Lwin, N.K.Z., 2023, February. Machine Learning Based Ecommerce Model for Myanmar Hotel Industry. In *2023 IEEE Conference on Computer Applications (ICCA)* (pp. 83-85). IEEE.

Subroto, C.A.M. and Akbar, S., 2022, September. The Effect of Preprocessing Techniques on Stacked Generalization and Stand-Alone Method for E-commerce Demand Prediction. In *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)* (pp. 1-6). IEEE.

Chen, P.P., Guitart, A., del Río, A.F. and Periánez, A., 2018, December. Customer lifetime value in video games using deep learning and parametric models. In *2018 IEEE international conference on big data (big data)* (pp. 2134-2140). IEEE.

Xiao, P., 2020, September. Information management of e-commerce platform based on neural networks and fuzzy deep learning models. In *2020 International conference on smart electronics and communication (ICOSEC)* (pp. 532-535). IEEE.