National College of Ireland

# A Data-Driven Approach to Customer Segmentation and Customer Lifetime Value Prediction in Retail and E-Commerce

MSc Research Project
Data Analytics


Prachi Pradeep Patil
Student ID: X23109980

School of Computing
National College of Ireland

Supervisor: Prof. Teerath Kumar Menghwar

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | Prachi Pradeep Patil |
| **Student ID:** | X23109980 |
| **Programme:** | Data Analytics |
| **Year:** | 2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Prof. Teerath Kumar Menghwar |
| **Submission Due Date:** | 12/08/2024 |
| **Project Title:** | A Data-Driven Approach to Customer Segmentation and Customer Lifetime Value Prediction in Retail and E-Commerce |
| **Word Count:** | 6662 |
| **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:** | Prachi Pradeep Patil |
| **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): | |

# A Data-Driven Approach to Customer Segmentation and Customer Lifetime Value Prediction in Retail and E-Commerce

Prachi Pradeep Patil
X23109980

**Abstract**

The rapid advancement of digital technologies has transformed the retail and e-commerce landscape, boosting market competition and reshaping traditional business practices, making it essential for businesses to understand customer needs and behaviors. This proposed research focuses on improving Customer Segmentation and Customer Lifetime Value (CLTV) prediction in the Retail and E-Commerce sectors by integrating RFM (Recency, Frequency, Monetary) analysis with advanced clustering and regression techniques. The primary objective was to assess the effectiveness of these combined methods in enhancing marketing strategies through more accurate customer segmentation and value prediction. The methodology followed the Cross-Industry Standard Process for Data Minin (CRISP-DM) framework, starting with data collection and preprocessing, followed by the implementation of RFM analysis Hierarchical Clustering, Gaussian Mixture Models for segmentation, and Linear Regression, Random Forest, and Support Vector Regression models for CLTV prediction. The findings showed that Linear Regression using Scikit-learn, particularly when applied to log-transformed data, outperformed other models, with a significant improvement in accuracy and error reduction. LazyPredict further confirmed these results, with ElasticNetCV and LassoCV models showing superior performance. Despite limitations related to data quality and industry-specific findings, the research successfully proved the potential of these techniques in optimizing customer relationship management and targeted marketing efforts, contributing valuable insights for the retail and e-commerce industries.

# 1 Introduction

## 1.1 Background

In today's highly competitive retail and e-commerce landscape, understanding the customer behaviour is very crucial. It allows you to formulate several effective marketing strategies (Hilmy et al.; 2023),ASLANTAŞ et al. (2023) in the field. Businesses always aim to identify and retain high value customers, while optimizing their marketing budgets. Two major concepts of business industry - Customer Segmentation and Customer Lifetime Value (CLTV) are used in understanding customers and are considered as crucial techniques for prediction in this domain. By applying data analytics, retailer's can classify their customers in different segments and also predict the future value they contribute

to the companies or businesses. This insight empowers businesses to tailor their marketing efforts in order to maximize customer retention and profitability.
Customer Segmentation refers to the process of dividing a customers into different groups with similar characteristics, allowing for more clear targeted marketing strategies. Customer's are segmented on characteristics such as demographics, geographic location, pyschographics or behavioural patterns. RFM analysis is widely and commonly used behavioural segmentation technique ASLANTAŞ et al. (2023),Rathi and Karwande (2022). It measure how recently customers made a purchase, how often they purchase and how much they spend. On the other hand, Customer Lifetime Value also known as CLTV or CLV is a measure that projects the total value (revenue) a customer will bring to business over the entire period of relation. First purchase, repeat purchase and the average length of engagement is taken into consideration of CLTV. Accurately predicting CLTV allows businesses to focus their marketing efforts on high-value customers, thereby enhancing profitability and customer loyalty Fernandes et al. (2023), Gadgil et al. (2023), Sun et al. (2023). As this research aims to improve the predictive power of RFM analysis by implementing clustering techniques and supervised machine learning models to predict the Customer Lifetime Value.

## 1.2 Aim

The goal of this research is to explore the success of RFM analysis combined with various clustering techniques and supervised machine learning models, for predicting the CLTV and informing segmentation based marketing strategies in retail and e-commerce sector. Integrating advanced analytical methods, the research believes to improve the accuracy of customer value prediction and customer segmentation. Furthermore, the aim is to deliver clear insights that can be practically applied to improve marketing strategies, drive customer engagement, and ultimately increase business profitability. To provide a robust structure that can be adopted by retail and e-commerce businesses to optimize their customer relationship management practices and strategic marketing initiatives is a primary objective of this research.

## 1.3 Research Question

To what extent can RFM analysis and clustering techniques be used to predict Customer Lifetime Value (CLTV) and inform segmentation-based marketing strategies in the retail or e-commerce sector ?
**Sub Question:**

- How can customer segments derived from clustering be used to inform marketing strategies in the Retail or E-commerce sector?
- To compare the performance of different supervised machine learning models based on RFM metrics in predicting CLTV.

## 1.4 Research Objective

Research Objective to be addressed for the above mentioned research question are derived as -

- To evaluate the effectiveness of RFM analysis in segmenting customers in the retail and e-commerce sector.
- To apply various suitable clustering techniques such as Hierarchical Clustering and Gaussian Mixture Model to identify different customer segments.
- To compare the performance of different supervised machine learning models in predicting CLTV.
- To provide recommendations on how segmented customer data can be used to inform targeted marketing strategies.

## 1.5 Limitations

Certain research gaps and limitations have been acknowledged in this research as they can impact the results and applicability of conclusions. One major gap is the quality of the data available for data analysis, which can have impact on accuracy and reliability of the segmentation and predictive models. Moreover, the key findings and learning's may be specific for particular datasets and industry, limiting their applications to other industry or wider retail stores. The research study focuses on RFM analysis, potentially missing other important factors in customer behavior, such as social influence, market trends, or psychological aspects. Complex nature of implementing different clustering and machine learning models may lead to challenges like overfitting or computational errors. While the study aims to provide actionable ideas, the practical implementation of these approach in real word scenarios might face operational and organizational limitations too.
The research is divided into several sections to thoroughly discuss the research topic. Section 2 discussed related work, providing an overview of existing literature on customer segmentation, RFM analysis, clustering techniques, and CLTV prediction. Section 3 explains the Research Methodology, including data collection, data cleaning, exploratory data analysis, feature engineering, and selected machine learning models. Section 4 presents the design specifications, including the fundamental architecture and framework used in the study. Section 5 discusses the implementation steps and model specifications. Section 6 evaluates the findings from the analysis, while Section 7 concludes the report, summarizing key findings, discussing limitations, and suggesting directions for future research.

# 2 Related Work

## 2.1 Customer Segmentation

(Shirole et al.; 2021) and (Hilmy et al.; 2023) analyzed transactional data from a UK-based online retail store. They covered the period from December 2010 to December 2011. On the other hand ASLANTAŞ et al. (2023) used data from connected home appliances with a dataset of approximately 41,000 electronic home appliances. All studies involved data preprocessing, feature selection, and normalization before applying clustering algorithms. Shirole et al. (2021) and ASLANTAŞ et al. (2023) implemented the K-means algorithm, whereas Hilmy et al. (2023) compared K-means with K-Medoids. The evaluation metric used in all above research is Silhouette Index. Shirole et al. (2021)

achieved a score of 0.442, ASLANTAŞ et al. (2023) got scores ranging from 0.37 to 0.60 for different appliances and paper 3 achieved 0.74 which showcased that K-means outperformed K-Medoids. They also share common limitations such as, oversimplification by focusing solely on three factors (RFM) and the need for further refinement in feature selection and clustering approaches.

Priyadarshni et al. (2023) analyzed credit card customers using a dataset of 8,950 entries. By implementing PCA for dimensionality reduction and achieving a silhouette score of 0.867026 with three clusters. Rathi and Karwande (2022) targets credit customers from a Nine Reload Credit Server. Dataset was reduced to 82,648 transactions from 368,829 entries, using the RFM model to identify profitable customers with different segments. Alzami et al. (2023) improved customer segmentation for an e-commerce company in Brazil with a dataset from Kaggle. By conducting EDA and determining four clusters, a silhouette score of 0.470 was achieved. Khumaidi et al. (2023) segments customers of dataset PT. Literata Lintas Media using sales transaction data, following the CRISP-DM framework. Three grades of customers were identified. The best possible numbers of clusters to consider are two for effective performance. K-means algorithm is commonly used algorithm with the goal of improving customer segmentation for better marketing strategies. Differences lie in the datasets used for analysis and specific methodologies: Priyadarshni et al. (2023) uses PCA for dimensionality reduction, Rathi and Karwande (2022) focuses on RFM variables for credit customers, Alzami et al. (2023) merges multiple data sources from Kaggle for e-commerce analysis, and Khumaidi et al. (2023) applies the CRISP-DM framework to a retail dataset of books and stationary. Research gap includes clustering accuracy and suggests the addition of features and using other advanced algorithm for better model performance.

Lewaa (2023), Wang et al. (2024), Aliyev et al. (2020) and Raj et al. (2023) had a common goal of research, to improve the CRM and targeted marketing startegies by identifying various customer segments. Each research used different datasets: an online retail dataset, auto parts sales transactions, bank customer data, and credit card user data. The methodologies involved preprocessing steps like handling missing values and outliers. Further followed by the application of clustering algorithms such as K-means, DBSCAN, and Hierarchical Clustering. The accurate number of clusters were figured out using methods like the elbow method and silhouette score. Results across these research showed effective segmentation, helped in personalized marketing strategies. Limitations include clustering accuracy, computational complexity, and the need for incorporating additional variables. As Lewaa (2023),Wang et al. (2024),Aliyev et al. (2020) and Raj et al. (2023) have focused on using RFM based clustering, specific improvements like including satisfaction variable in LRFMS model and separating of behavioral and demographic data for segmentation.

## 2.2 Customer Lifetime Value Prediction

The Fernandes et al. (2023), Gadgil et al. (2023) and Sun et al. (2023) shared a common objective of enhancing customer segmentation and lifetime value prediction using machine learning techniques. All study used the "Online Retail II" dataset from a UK-based online retailer. Fernandes et al. (2023), Gadgil et al. (2023) and Sun et al. (2023) implemented

the RFM model for customer segmentation but with different methodological approach. The Fernandes et al. (2023) implemented Customer Lifetime Value (CLV) analysis with RFM and K-Means clustering, 5 different customer segments were achieved, though it acknowledged the need for refinement due to potential biases. Gadgil et al. (2023) proposed a stacking-based ensemble model combining RandomForest, XGBoost, and ElasticNet, which outperformed traditional models, with the lowest Mean Absolute Error (MAE). It needs further exploration of model generalizability and efficiency. Sun et al. (2023) used a variety of machine learning algorithms, with the gradient boosting decision tree performing best, achieving a 95.12% AUC and 93.80% accuracy. Sun et al. (2023) suggested that further analysis of customer heterogeneity could enhance model applicability.

Surti et al. (2023) and Alsharafa et al. (2024) analysed larger datasets, including diverse features like demographic data and insurance policies, Carneiro and Miguéis (2021) focused on a smaller B2B dataset. Surti et al. (2023), Alsharafa et al. (2024) and Carneiro and Miguéis (2021) aimed to improve the accuracy of CLV predictions and customer segmentation, using methodologies that involved data preprocessing, feature selection, and the application of models like RFM, K-Means clustering, and advanced regression techniques. Surti et al. (2023) and Alsharafa et al. (2024) incorporated sophisticated models like AutoML Regression and ARIMA with Deep Neural Networks, achieving high prediction accuracy. While Carneiro and Miguéis (2021) used a simpler approach with the RFM model and K-Means clustering, achieving an 83.6% SSE in segmentation. Common limitations across Surti et al. (2023), Alsharafa et al. (2024) and Carneiro and Miguéis (2021) included the need for more data, the potential for integrating deep learning techniques, and the challenge of handling complex data processing tasks. The primary difference lied in the scope and complexity of the models used, Surti et al. (2023) and Alsharafa et al. (2024) focused more on predictive accuracy, while Carneiro and Miguéis (2021) focused on practical segmentation for decision-making.

# 3 Methodology

## 3.1 Research Methodology

This research uses the Cross-Industry Standard Process for Data Mining (CRISP - DM) methodology to segment customers and predict Customer Lifetime Value (CLV). CRISP-DM is widely used framework which guides us through a structured and iterative method that ensures each phase of data mining process is carefully addressed, in line with research or project goals. CRISP-DM is a six-phase process as shown in Figure 1. This methodology suits for Customer Segmentation and Customer Lifetime Value prediction as it guarantees a thorough and iterative approach to data mining, allowing for clear identification of business objectives. The advantages of CRISP-DM include its flexibility, applicability across various industries, and its ability to handle complex data mining tasks by facilitating continuous refinement and improvement of the models based on business goals.

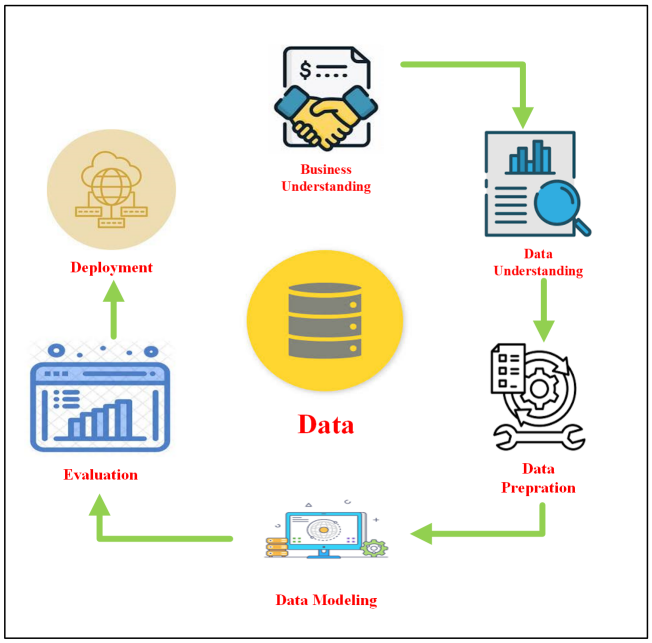


Figure 1: CRSIP-DM

### 3.1.1 Business Understanding

The research aims to better understand the customer behaviour and predict the customer lifetime value (CLTV) in order to implement the marketing strategies, and improve overall business performance. By categorizing the customers based on their purchasing behaviours, demographics, geographic and engagement parameters, objective is to identify the distinct clusters with similar behaviour. Predicting the CLTV for these clusters(customer's) will help in prioritizing marketing efforts, optimizing resource allocation and improving the retention strategies. The insights achieved from the research will empower the business to personalise deals & offers for each customer, improve customer satisfaction and eventually boost the revenue and profitability.

### 3.1.2 Data Understanding

Data Understanding is the second and critical step in CRISP-DM methodology in data mining process, where the analyst examines and becomes familiar with the available data. This phase involves several steps such as checking the structure, content and quality of the data, as well as identifying missing values, understanding the relation between features and determining appropriate techniques for cleaning and preprocessing the data. This helps in ensuring that data is suitable for the analysis and fits with business objectives. The data gathered should be in a suitable form as it should get imported easily using the programming language used in the research. The data gathered can in any form like .csv, .xlsx, XML, JSON etc. In this research, two datasets are used -'Online Retail Dataset' [1] from a UK based company, available on UCI machine repository, which includes 8 features, and 'Retail Insights: A Comprehensive Sales Dataset' [2] from Kaggle, containing 24 features. These datasets provide detailed information on customer transactions,

[1] https://archive.ics.uci.edu/dataset/352/online+retail
[2] https://www.kaggle.com/datasets/rajneesh231/retail-insights-a-comprehensive-sales-dataset

demographics, and sales patterns, setting the foundation for Customer Segmentation and CLTV prediction models as shown in Table 1 and 2.

Table 1: Dataset 1 features

| **Dataset Features** | **Description** |
|---|---|
| **InvoiceNo** | Unique number assigned to each transaction |
| **StockCode** | Unique number assigned to each distinct product |
| **Description** | Product Name |
| **Quantity** | Number of items per transaction |
| **InvoiceDate** | Day and time on which transaction was generated |
| **UnitPrice** | Product per unit price |
| **CustomerID** | Unique number assigned to each customer |
| **Country** | Country name where customer resides |

Table 2: Dataset 2 Features

| **Dataset 2 Features** | |
|---|---|
| Order No | Product Name |
| Order Date | Product Category |
| Customer Name | Product Container |
| Address | Ship Mode |
| City | Ship Date |
| State | Cost Price |
| Customer Type | Retail Price |
| Account Manager | Profit Margin |
| Order Priority | Order Quantity |
| Discount | Sub Total |
| Total | Shipping Cost |
| Discount $ | |

### 3.1.3 Data Preparation

Sometimes the performance of the machine learning model can be significantly diminished by the noise and abnormalities present in data. So at this point, data cleaning and preprocessing of the dataset is necessary. Data preparation is the process of making raw data ready for after processing and analysis. The key methods are to collect, clean, and label raw data in a format suitable for machine learning (ML) algorithms, followed by data exploration and visualization.

In this research, after loading the data in dataframe, the data was further analyzed to understand the data structure, numerical columns, categorical columns present in the dataset's. The statistical values of each column were studied and dataset's summary was understood for better analysis. Null values and duplicates were checked in both datasets, and rows with null values were removed from both dataset's to ensure the data was clean. In Dataset 2, columns containing special characters such as '$' and ',' were cleaned by removing these characters and converting the values into numeric format, ensuring consistent and clean data for processing. In both datasets, columns like Invoice Date and Order Date were converted to a proper date format using the 'pd.to_datetime()' function.

This function makes sure that the dates are accurately formatted, which is necessary for tasks like Customer Segmentation and CLV prediction, as it allows for precise analysis of time-based data. Different functions were used to extract specific components from 'Invoice Date' for better understanding and visualization of year, month, day, hour and day of week on granular manner. This breakdowns allows for identifying the trends, patterns and customer behaviour in different time frames. Some redundant features from dataset 2 such as Account Manager, Product Container and many more were eliminated to reduce the complexity. Key step of finding outliers and removing it was done using the IQR method, particularly for the Recency, Frequency, Monetary metrics to prevent the skewed analysis. Additionally, data transformation techniques such as log transformation were applied to normalize the features, stabilizing variance and reducing the skewness. Data scaling was performed to make sure that all features contribute equally for model's performance. At last for Customer Lifetime Value Prediction, the dataset was splitted in training (80%) and testing (20%) data. For proper machine learning model deployment and to achieve the reliable results, all necessary data preparation steps are performed for targeted marketing and customer retention strategies.

### 3.1.4 Modelling

Data modelling involves selecting and applying various data mining techniques on pre-processed data in order to achieve the actionable insights. Modelling entails selecting the suitable algorithms based on thorough understanding of datasets, its features and research question. In this research of Customer Segmentation and Customer Lifetime Value Prediction, several machine learning models are used to gain meaningful insights. RFM is a marketing technique used to evaluate and segment customers based on their purchasing behavior. RFM analysis stands for Recency, Frequency, and Monetary analysis. For Customer Segmentation, RFM analysis was followed by the applying Hierarchial Clustering and Gaussian Mixture Model (GMM). And for CLTV, regression models like Linear, Random Forest and Support Vector Regression were used. These models helps in accurately segmenting customers and predicting their future values, enabling targeted marketing and strategic decision-making.

### 3.1.5 Evaluation

Evaluation is next step followed after Data Modelling, where the developed models are assessed to determine effectiveness and relevance to the business goals. In this phase, the very first step is to select the correct evaluation method depending on the nature of dependent variable. Selecting the right evaluation metrics is very important because entirely depending on the model accuracy can result in inaccurate results.
This research involving both clustering and regression tasks, specific evaluation metrics are used for each. For Customer Segmentation, models like Hierarchical Clustering and Gaussian Mixture Model (GMM) are classified using metrics such as Silhouette Score, Calinski-Harabasz Index, Davies-Bouldin Index, Akaike information criterion (AIC) and the Bayes information criterion (BIC). These metrics analyses the quality of the clusters by measuring their cohesion and separation, Well defined clusters are indicated with the higher Silhouette and Calinski-Harabasz scores, and lower Davies-Bouldin values suggests compact and distinct clusters. AIC and BIC help in selecting the best model by balancing fit and complexity, with lower values indicating a more ideal model. These evaluations ensures accurate identification of different customer segments for effective targeting and

personalized strategies. For Customer Lifetime Value Prediction, models like Linear Regression, Random Forest, and Support Vector Machine are examined using various metrics. For Regression, key metrics include Mean Squared Error (MSE), $R^2$ Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). MSE and RMSE calculates the average squared differences between predicted and actual values, with lower values indicating better model accuracy. The $R^2$ Score shows the proportion of variance explained by the model, with values closer to 1 indicating better fit. MAE and MAPE provide insights into the average magnitude of errors, with lower values suggesting more accurate predictions. These metrics are used to measure the model's predictive performance, leading to accurate estimation of customer value for targeted marketing and strategic planning.
Comparing these various evaluation parameters among different models, business can select the most suitable model for their needs, leading to precise and useful insights.This phase also includes the visualization of model results, such as plotting predicted vs actual values, which provides a clear understanding of model performance and potential areas for improvement.

### 3.1.6 Deployment

Deployment is the final step in CRISP-DM methodology, where the findings of data mining process are put forward in practical use. Integrating the chosen models into business processes or systems allows for data driven decision making. However, this research is more focused on the research side, the deployment stage will not be included. As a result, the focus will be on discussing the future scope and steps ahead for advancing this research. This includes exploring additional data sources, refining models, and potentially shifting from research to practical applications in real-world cases. By doing so, the research can continue to evolve and provide valuable insights for future implementation.

## 3.2 Data Visualization

Exploratory Data Analysis (EDA) is the process of exploring and visualizing data to discover underlying patterns, relationships, and anomalies of data before applying any modeling techniques. Considering Customer Segmentation and Customer Lifetime Value Prediction, EDA plays a important role in gaining deeper insights into customer behaviors and product interactions, helping to identify key factors that drive customer value, allowing for more accurate segmentation and prediction models.

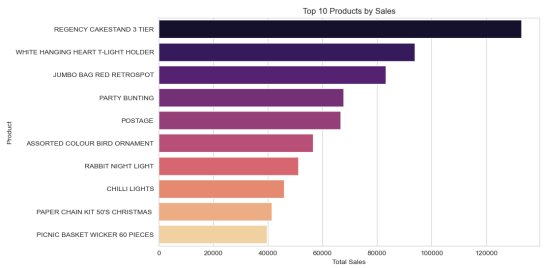


Figure 2: Dataset 1-Top 10 Products by Sales

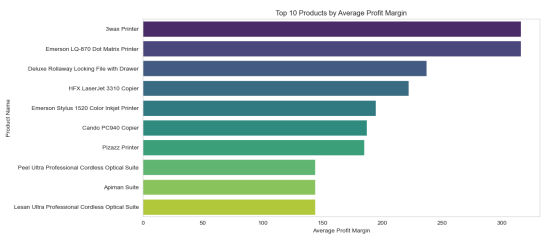


Figure 3: Dataset 2-Top 10 Products by Profit Margin

The bar graph shown above in Figure 2 and Figure 3 visually represents the Top 10 products of Dataset 1 and 2. Figure 2 shows the top 10 products by total sales, with

"REGENCY CAKESTAND 3 TIER" as the top-selling item, while Figure 3 highlights the top 10 products by average profit margin, showing "3wax Printer" as the most profitable product.

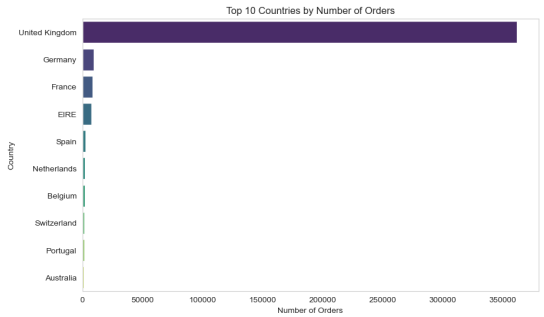


Figure 4: Dataset 1-Top 10 Countries by Orders

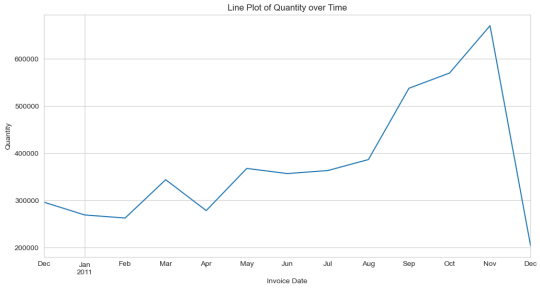


Figure 5: Dataset 1-Quantity over Time

Figure 4 shows the top 10 countries by the number of orders, with the United Kingdom having most orders, highlighting it as main market. Figure 5 line chart shows the quantity of items sold over time, displaying seasonal trend where sales peak towards the end of the year and then drop sharply in December. These visuals helps in understanding where most sales are coming from and how sales change with the seasons, which is important for planning inventory and marketing efforts.

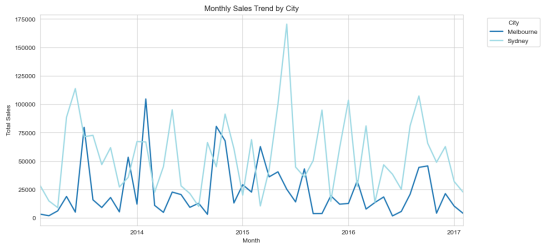


Figure 6: Dataset 2-Monthly sales trend by city

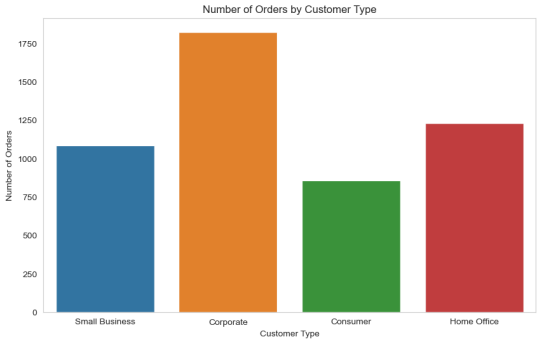


Figure 7: Dataset 2-Orders by Customer Type

Line chart as shown in Figure 6 shows the monthly sales trend by city, comparing Melbourne and Sydney over time. It indicates significant fluctuations over time with seasonal peaks. Figure 7 Bar chart highlights the number of orders by customer type, Corporate type customers placing maximum number of orders. These visuals helps in understanding geographic sales patterns and customer segment behaviours.

# 4 Design Specification

This section discusses the design structure of the process followed in this research. The design specification describes the structured flow and used algorithms, ensuring that

research's implementation and objectives are clear and complete. Combination of unsupervised and supervised models provides a thorough approach to understand customer behaviour and predicting their lifetime values.

The research requires a structured plan were unsupervised and supervised models are used. The implementation has multiple steps, starting from data collection to model deployment, making sure a comprehensive and effective analysis process is followed. Structured methodology is important for achieving the accurate findings and pedictions as it showcases a thorough understanding of advanced analytical methods or models.

**Framework**

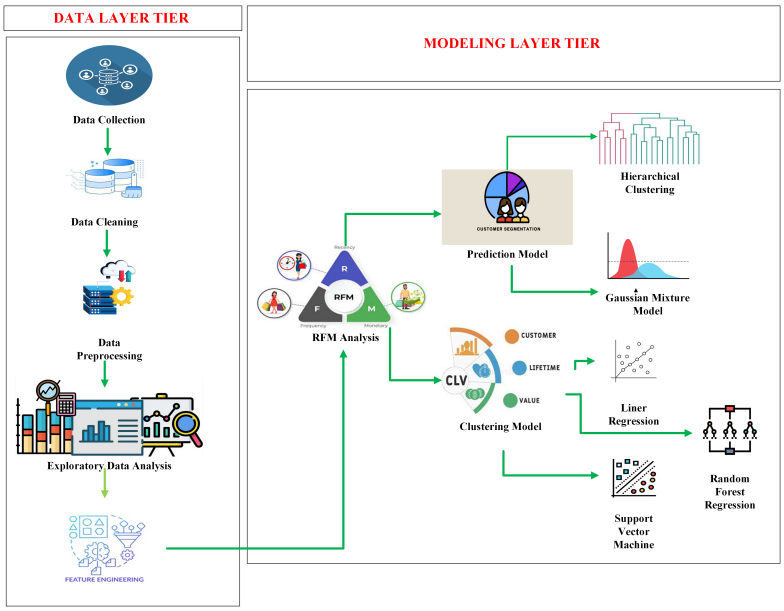


Figure 8: Project Flow

- **Data Collection and Cleaning:** The very first step includes collecting the data. In this research the data is collected from Kaggle and UCI repository, followed by data cleaning - checking missing values, duplicates, outliers. This steps ensures the quality and accuracy of the data for future analysis.

- **Data Preprocessing and EDA:** Includes data normalization, data transformation, feature engineering and encoding the categorical variables if necessary. Exploratory Data Analysis is performed to visualize the data distributions, correlations between features and trends, offering insights for future analysis,

- **Feature Engineering and Data Preparation:** Key features are transformed from the datasets focusing on Recency, Frequency, Monetary parameters. Transaction Data features are standardized using appropriate date format, verifying consistency and simplified analysis. Engineered features are important for Customer Segmentation and Customer Lifetime Value Prediction, improving the accuracy and interpretability of model.

- **RFM Analysis:** RFM analysis is strong marketing technique that categorizes the customers based on their purchasing behaviour and past transactions, acting as a starting point for both Customer Segmentation and CLTV Prediction.
- **Customer Segmentation:** Segmentation involves the use of unsupervised models - Hierarchical Clustering and Gaussian Mixture Model (GMM). Hierarchical Clustering divides the unknown data into clusters, forming a hierarchy that can be visualized using a tree-like structure called dendogram. GMM is a probabilistic model that offers the flexibility in identifying the overlapping segments. Both methods are unsupervised, allowing the data to reveal inherent groups.
- **CLV Prediction:** Supervised machine learning models such as Linear Regression, Random Forest Regression and Support Vector Regression are implemented to predict the Customer Lifetime Value. These models are trained on historical purchase data and RFM features, offering robust CLTV estimates.

The design specification explains each model used in this research, each have specific purpose in the analysis and prediction of customer behavior. These models range from unsupervised clustering techniques to supervised regression models, providing a thorough approach to customer segmentation and their lifetime value prediction.

## 4.1 RFM Analysis

RFM, stands for Recency, Frequency, and Monetary value, is a widely used marketing technique for evaluating and dividing customers based on their buying habits and transaction history. The main objective of the model is to determine customers based on, how recently they made a purchase, how frequently they buy, and how much they spend. Such segmentation helps businesses to customize their marketing strategies to different customer groups, such as 'champions', 'loyal customers', 'potential loyalist' or 'at-risk' customers. The RFM process involves data preparation, calculating RFM metrics, scoring, segmenting, analysis, and implementing targeted actions to improve customer engagement and retention.

## 4.2 Customer Segmentation Models

### 4.2.1 Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning algorithm that forms a hierarchy of clusters, visualized using a dendrogram plot. Hierarchical clustering has two types - agglomerative - where each data point starts as its own cluster and merges with others, and divisive - where all points start in one cluster and split. This method does not require prior knowledge of the number of clusters, making it flexible for exploring data structures. It is particularly useful in market segmentation, anomaly detection, and social network analysis, as gives clear picture of data relationships and similarities.

### 4.2.2 Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a probabilistic, unsupervised learning model that displays data as a mixture of several Gaussian distributions. GMM handles overlap in clusters using the Expectation-Maximization algorithm, iterating the improvement of its

parameters to maximize the likelihood of the observed data. Due to flexibility model allows to identify clusters of various shapes and sizes, making it suitable for applications like customer segmentation, image processing, and anomaly detection, where understanding complex data distributions is crucial.

## 4.3 Customer Lifetime Value Prediction Models

### 4.3.1 Linear Regression

Linear Regression is a supervised machine learning model that aims to predict continuous results based on input features. Model operates by fitting a linear equation, reducing the discrepancy between predicted and actual values. Ordinary Least Squares is most commonly used method in Linear Regression. The simplicity and interpret-ability of Linear Regression makes it as widely used model for tasks like predicting housing prices or sales forecasts. Key evaluation metrics include $R^2$, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which helps in predicting the model's accuracy and explanatory power.

### 4.3.2 Random Forest Regression

Random Forest is a effective supervised machine learning algorithm, that uses the collection of decision trees to evaluate continuous variables. Risk of model over fitting and improved accuracy is achieved by combining the predictions of created multiple decision trees. Random forest has the ability to handle the high dimensions datasets and maintain a good balance between bias and variance. Random Forest is widely used in applications such as predicting housing prices, fraud detection, and customer behavior analysis, offering reliable and precise predictions.

### 4.3.3 Support Vector Regression

Support Vector Regression (SVR) applies the idea of Support Vector Machines (SVM) to regression tasks, predicting continuous results by finding a hyperplane that best fits the data within a specified range of tolerance. By using suitable kernel function, SVR can handle both linear and non linear, thus being versatile for most data scenarios. Three main parameters to be optimized in the model are - Choice of kernel, regularization parameter (C) and epsilon.This model is particularly useful in financial forecasting, real estate price prediction, and time series analysis.

# 5 Implementation

Implementation section represents practical fulfillment of Customer Segmentation and Customer Lifetime Value Prediction. This section will discuss implementation process from the start to end that was followed for the purpose of this research. The implementation of models is executed by Jupyter notebook itself using programming language Python from the open source platform Anaconda. Various libraries such as 'Pandas', 'Numpy', 'Matplotlib', 'Seaborn', 'Sci-kit learn' are used for data manipulation, visualization and model deployment. This research aims to understand the customer behaviour and predict their lifetime value to improve the marketing strategies and business performance. The data preparation was carried out using two datasets: 'Online Retail Data'

and 'Retail Insights: A Comprehensive Sales Dataset'. Data preparation steps ensures consistent and clean data for analysis.

## 5.1 RFM Analysis

In this research the RFM models were applied on both the datasets, to classify the customer's based on their purchasing behaviour. The very first step involved cleaning, preprocessing the data, converting the date in standard format, filtering out the returns, removing the invalid inputs from numerical variables and dropping the unnecessary columns for better understanding. The RFM model measured three key parameters - Recency (Days since last purchase), Frequency (Number of Transactions) and Monetary (Total spending). In Dataset 1, the 'Customer ID', while in Dataset 2, the 'Customer Name' was considered as the common factor for allocating RFM metrics. Customer's were assigned scores from 1 to 5 for each metric, which were then further combined into an overall RFM score.

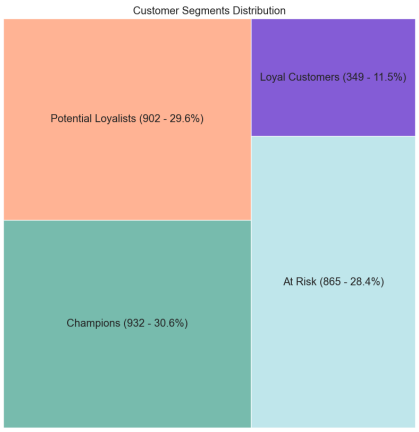


Figure 9: RFM Segments for Dataset 1

As shown in Figure 9, customer's are divided in different categories such as 'Champions,' 'Loyal Customers,' 'Potential Loyalists,' 'Needs Attention,' and 'At Risk'. These segments helps to gain clear insights on customer behaviour and their preferences. Visualisations of the RFM scores and customer segments were also created using the seaborn, providing a depth view of customer base. This segmentation of customers helps to understand the customer demographics, but it acts as framework for predicting CLTV.

## 5.2 Hierarchical Clustering

While implementing the Hierarchical Clustering for Customer Segmentation, outliers were checked and removed using IQR method from RFM obtained metrics dataframe. This step guarantees that the values do not skew the clustering performance. Using StandardScaler, the data was standardized to normalize the features, making them equivalent to scale. The clustering was followed by agglomerative strategy, starting with each data point as its cluster and grouping the most similar clusters iteratively. To link the cluster's with each other with minimum variance 'ward' method was used within each cluster. The cluster's were identified by analyzing the dendogram and setting threshold distance as

shown in Figure 10. Threshold helped in defining the cluster's, which are then visualised using PCA for 2D representation of the data.

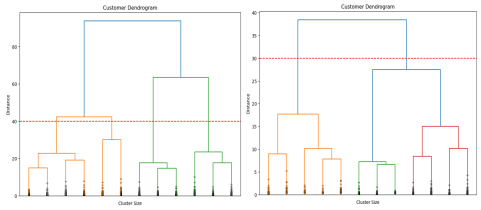


Figure 10: Dendogram of Dataset 1 and 2

The significance of these performed steps lies in properly grouping the customer's as shown in Figure 11 based on transactional behaviour, which results in precise segmentation. This method helps to identify different customer cluster's enabling targeted marketing strategies for business and better resource allocation.

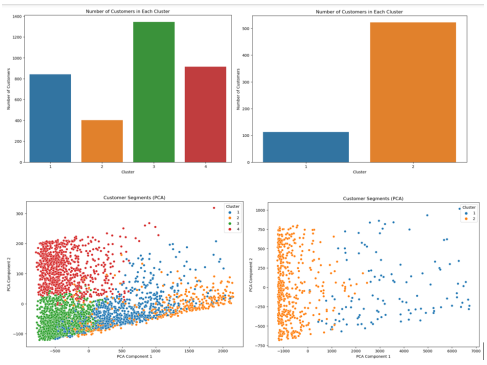


Figure 11: Customer Segmentation using Hierarchical Clustering for both datasets

Dendogram and PCA scatter plot helps us to gain insights on the original structure of data, which is important for understanding customer dynamics and boosting business strategies in Customer Segmentation.

## 5.3 Gaussian Mixture Model

Several steps are followed for segmentation by using GMM model, starting with calculating the required number of clusters by measuring the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) for different models. These parameters helps to select the model that best fits the data by balancing model complexity, with lower values indicating better models. As shown in Figure 12, the elbow points on the BIC and AIC plots suggests the ideal number of cluster's to be considered.

After calculating the ideal number of clusters Figure13, the model is fitted to the data by chosen number of components and data points are allocated to clusters based on probabilistic distribution, allowing indirect clustering where the data points may belong to various clusters with varying probabilities. This models flexibility is useful in segmentation of customers as it gives us the complex understanding of customers behaviour and choices. The clusters are then visualized using PCA scatter plot for better understanding of the data.

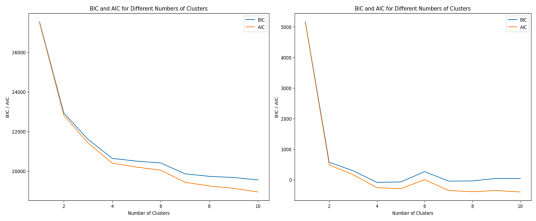


Figure 12: BIC and AIC plots for both Datasets

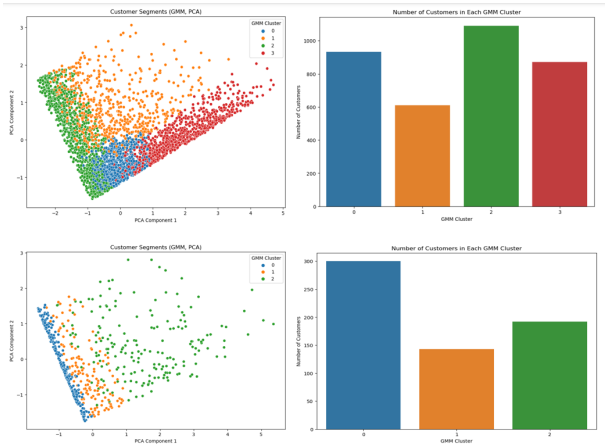


Figure 13: Customer Segmentation using GMM for both datasets

## 5.4 Regression Models

In this research, Linear, Random Forest and Support Vector Regression models are implemented for Customer Lifetime Value prediction. RFM metrics dataframe is considered as input (Recency, Frequency) and output (Monetary) features. The process starts with splitting the data in training (80%) and testing (20%) sets, and then standardizing the features to ensure uniform impact on model's performance. Then model is trained on the training data and evaluated on the test data using metrics such as Mean Squared Error (MSE), $R^2$ Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Further log transformation was performed on both input and output features. Log transformation helps to stabilize the variance, lower skewness, and to handle the outliers if present more efficiently, resulting in a more distribution of the data. The implementation of log transformation improves the model's ability to manage heteroscedasticity and improve overall accuracy. This research focuses on using RFM features. Through feature engineering, additional features can be included into the data, which can improve the prediction of Customer Lifetime Value (CLTV) and the model's performance.This integrated approach allows for a clear understanding of the relation between customer metrics and their lifetime value.

# 6 Evaluation

## 6.1 Hierarchical clustering

The evaluation parameters used for Hierarchical Clustering are Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index, as they shows different characteristics of cluster. As shown in Table3 for Dataset 1, the Hierarchical Clustering model has a

Table 3: Hierarchical Clustering Evaluation Metrics

| **Hierarchical Clustering** | **Ideal Values** | **Dataset 1** | **Dataset 2** |
|---|---|---|---|
| **Silhouette Score** | Closer to 1 is better | 0.36 | 0.50 |
| **Calinski–Harabasz Index** | Higher is better | 2652.20 | 399.79 |
| **Davies–Bouldin Index** | Closer to 0 is better | 1.07 | 1.02 |

Silhouette Score of 0.368, a Calinski-Harabasz Index of 2652.20, and a Davies-Bouldin Index of 1.073. Whereas, Dataset 2 shows a higher Silhouette Score of 0.501, a much lower Calinski-Harabasz Index of 399.79, and a slightly better Davies-Bouldin Index of 1.026. The higher Silhouette Score in Dataset 2 suggests well-defined clusters, while the lower Calinski-Harabasz Index suggests the separation between clusters is totally different than in Dataset 1. The slightly lower Davies-Bouldin Index for Dataset 2 says better intra-cluster connection.
Overall, Dataset 2 has more cohesive clusters, while Dataset 1 shows clearer cluster separation, reflecting different feature of clustering quality over the datasets.

## 6.2 Gaussian Mixture Model

Table 4: Gaussian Mixture Model Evaluation Metrics

| **Gaussian Mixture Model** | **Dataset 1** | **Dataset 2** |
|---|---|---|
| **Silhouette Score** | 0.15 | 0.11 |
| **Calinski-Harabasz Index** | 1323.42 | 205.70 |
| **Davies-Bouldin Index** | 1.46 | 1.93 |
| **BIC** | 20633.86 | 292.70 |
| **AIC** | 20393.56 | 163.55 |

The evaluation metrices of the Gaussian Mixture Model (GMM) as shown in Table 4 of both datasets tells about clustering quality and model fit. Dataset 1, metrics shows slightly improved clustering with a Silhouette Score of 0.1585, a Calinski-Harabasz Index of 1323.42, and a Davies-Bouldin Index of 1.4679, showing more distinct and cohesive clusters as compared to Dataset 2. Yet, both datasets has relatively low Silhouette Scores, which tells about lack of clear clustering structure. Dataset 2, on the other hand, shows a much lower Bayesian Information Criterion (BIC) of 292.70 and Akaike Information Criterion (AIC) of 163.55, showcasing a simpler model fit with few parameters, even with lower cluster quality.

In all, Dataset 1 shows better-defined clusters and Dataset 2 offers a more compact model, balancing model complexity and cluster quality.

## 6.3 Regression Models

The evaluation of regression models included two methods - standard evaluation and log-transformed data evaluation. The metrics used include Mean Squared Error (MSE), $R^2$ Score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The following Table 5 and 6 showcases summary of the results for all three models for both datasets.

Table 5: Evaluation Metrics using Sci-kit Learn Library for Dataset 1

| Dataset 1 **Models** | **Linear** | | **Random Forest** | | **SVR** | |
|---|---|---|---|---|---|---|
| | **Standard** | **Log Transformed** | **Standard** | **Log Transformed** | **Standard** | **Log Transformed** |
| **MSE** | 243300.79 | 0.4237 | 311527.22 | 0.5525 | 340660.98 | 0.4418 |
| **$R^2$ Score** | 0.4331 | 0.5557 | 0.2741 | 0.4206 | 0.2062 | 0.5367 |
| **MAE** | 361.84 | 0.5132 | 391.54 | 0.5732 | 400.37 | 0.4845 |
| **RMSE** | 493.26 | 0.6509 | 558.15 | 0.7433 | 583.66 | 0.6647 |
| **MAPE (%)** | 99.19 | 8.87 | 98.16 | 9.87 | 112.25 | 8.47 |

Table 6: Evaluation Metrics using Sci-kit Learn Library for Dataset 2

| Dataset 2 **Models** | **Linear** | | **Random Forest** | | **SVR** | |
|---|---|---|---|---|---|---|
| | **Standard** | **Log Transformed** | **Standard** | **Log Transformed** | **Standard** | **Log Transformed** |
| **MSE** | 2993633.51 | 1.5848 | 3466088.07 | 1.8903 | 5037631.15 | 1.6302 |
| **$R^2$ Score** | 0.2580 | 0.3411 | 0.1409 | 0.2141 | -0.2486 | 0.3222 |
| **MAE** | 1028.06 | 0.9279 | 1082.01 | 1.0388 | 1226.51 | 0.9305 |
| **RMSE** | 1730.21 | 1.2589 | 1861.74 | 1.3749 | 2244.47 | 1.2768 |
| **MAPE (%)** | 290.87 | 15.50 | 162.86 | 17.49 | 180.18 | 15.26 |

In Dataset 1, Linear Regression outperformed Random Forest and SVR with higher $R^2$ scores (0.433 for standard evaluation, 0.556 for log-transformed evaluation) and lower Mean Squared Errors (243,300.79 for standard evaluation, 0.424 for log-transformed evaluation). Likewise for Dataset 2, Linear Regression consistently outperformed the other models with better $R^2$ scores (0.258 for standard evaluation, 0.341 for log-transformed evaluation) and lower MSEs (2,993,633.51 for standard evaluation, 1.585 for log-transformed evaluation). Models showed improved performance accross all metrics, reducing the MAPE on applying log transformation. While Linear Regression continued to remain the most effective model. Slight improvements were seen in Support Vector model on applying log transformation. Results prove that data normalization improves the predictive accuracy of regression models.

After evaluating the regression models using sci-kit learn library, further the models are evaluated using Lazy predict library. The evaluation metrics of both datasets using Lazy Predict provides valuable insights for the performance of regression models. For Dataset 1, the Gradient Boosting Regressor performed best under standard evaluation with an R-Squared of 0.46. Whereas models like HuberRegressor and NuSVR showed superior performance with log-transformed data, achieving an R-Squared of 0.56. On the other hand Dataset 2 shows weaker performance accross all the models with OrthogonalMatchingPursuit leading in the standard evaluation with an R-Squared of 0.26. But using log transformation improved the models accuracy with ElasticNetCV and LassoCV

achieving the highest R-Squared of 0.35.

Table 7: Evaluation Metrics using Lazy Predict Library for Dataset 1

| Model | Adjusted R-Squared | R-Squared | RMSE | Time Taken (s) |
|---|---|---|---|---|
| **Standard Evaluation** | | | | |
| GradientBoostingRegressor | 0.46 | 0.46 | 479.31 | 0.55 |
| SGDRegressor | 0.43 | 0.43 | 493.14 | 0.02 |
| TransformedTargetRegressor | 0.43 | 0.43 | 493.26 | 0.02 |
| OrthogonalMatchingPursuitCV | 0.43 | 0.43 | 493.26 | 0.03 |
| LinearRegression | 0.43 | 0.43 | 493.26 | 0.02 |
| **Log-Transformed Evaluation** | | | | |
| HuberRegressor | 0.56 | 0.56 | 0.65 | 0.05 |
| NuSVR | 0.55 | 0.56 | 0.65 | 1.23 |
| LassoLarsIC | 0.55 | 0.56 | 0.65 | 0.02 |
| OrthogonalMatchingPursuitCV | 0.55 | 0.56 | 0.65 | 0.04 |
| TransformedTargetRegressor | 0.55 | 0.56 | 0.65 | 0.02 |

Table 8: Evaluation Metrics using Lazy Predict Library for Dataset 2

| Model | Adjusted R-Squared | R-Squared | RMSE | Time Taken (s) |
|---|---|---|---|---|
| **Standard Evaluation** | | | | |
| OrthogonalMatchingPursuit | 0.25 | 0.26 | 1725.05 | 0.02 |
| OrthogonalMatchingPursuitCV | 0.25 | 0.26 | 1725.05 | 0.00 |
| SGDRegressor | 0.25 | 0.26 | 1729.55 | 0.01 |
| LassoCV | 0.25 | 0.26 | 1730.11 | 0.10 |
| LassoLars | 0.25 | 0.26 | 1730.12 | 0.01 |
| **Log-Transformed Evaluation** | | | | |
| ElasticNetCV | 0.34 | 0.35 | 1.25 | 0.10 |
| LassoCV | 0.34 | 0.35 | 1.25 | 0.24 |
| PoissonRegressor | 0.34 | 0.35 | 1.25 | 0.03 |
| LarsCV | 0.34 | 0.35 | 1.25 | 0.00 |
| LassoLarsCV | 0.34 | 0.35 | 1.25 | 0.02 |

The results obtained as shown in Table 7 and 8 suggests that using log transformation improved the models ability to identify underlying patterns in both datasets. For Dataset 2, the improvement in R-Squared suggests better handling of skewed data. The RMSE values are consistently lower in log transformed evaluations. As a result, it highlights the the importance of data normalization for improving the models accuracy. Ultimately the

results suggests Dataset 1 models are relatively robust. But Dataset 2 achieves better results using log transformation, proving the importance of this approach in fine tuning the predictive accuracy on different datasets.

## 6.4 Discussion

The research focused on addressing the research question, how effectively RFM analysis combined with clustering and regression models, helps to improve the customer segmentation and predict the accuracy customer lifetime value in Retail or E-commerce sector. Implementing clustering models like Hierarchical and GMM, helped to successfully discovered distinct customer segments, focusing on better understanding of customer behaviour. This segmentation of customer, enables the development of targeted and effective marketing strategies, for improved customer relationship management. However, segmenting research faced limitation, including dependency on RFM metrics alone, which may not capture the full aspects of customer behavior. Additionally, the quality of the available data could impact the potential of the findings.
For predicting Customer Lifetime Value, regression models like linear, random forest and support vector are used. Research compared the success rate of the models under standard and log transformed conditions using Scikit-learn and Lazypredict library. The findings proved that log transformation considerably improves the model performance. Results of Scikit-learn showcased that Linear Regression outperformed the other two models for both datasets, with log transformation improving accuracy and reducing error margins. LazyPredict backed this findings, where models like ElasticNetCV and LassoCV shows better performance on log transformed evaluations. This results higlights the importance of data normalization and selection of suitable models in accurately predicting CLTV.
In order to address the limitations faced in this research, future research can include additional features of customer's demographics over RFM metrics. This addition can enhance segmentation and predictions process. Likewise advanced models such as deep learning and ensemble methods can offer further improvements. Current research gaps can be experimented using different clustering and machine learning models, for more actionable insights on customer segmentation and CLTV prediction in Retail and E-commerce industry. Revisiting the research question—To what extent can RFM analysis and clustering techniques be used to predict Customer Lifetime Value (CLTV) and inform segmentation-based marketing strategies in the retail or e-commerce sector?—the research successfully met its objectives. Proving the implementation of applied techniques suggests actionable insights that can be incorporated directly for effective marketing efforts. Ultimately the research expands the field by providing a thorough structure for customer segmentation and CLTV prediction. It has practical application for businesses aiming to strengthen their customer relationship management and marketing strategies.

# 7 Conclusion and Future Work

**Conclusion**
The research aimed to explore the effectiveness of combining RFM analysis when combined with clustering and regression models for Customer Segmentation and Customer Lifetime Value (CLTV) prediction in the Retail and E-Commerce sectors. These methodology involved a structured way, starting with data collection and preprocessing, followed

by the implementation of RFM analysis, clustering techniques like Hierarchical Clustering and Gaussian Mixture Models, and regression models for CLTV prediction. Linear Regression, particularly with log-transformed data, outperformed other models, proving robust performance in predicting CLTV. The use of the LazyPredict library further highlighted the efficiency of different regression models. While the research showcased the potential of these techniques to improve customer segmentation and targeted marketing, it acknowledged limitations related to data quality and the specificity of findings to particular datasets and industry. Overall, the research findings proved the potential of combining these techniques to improve customer segmentation and CLTV prediction, thereby contributing valuable insights for marketing strategy optimization.

**Future Work**

Further research can explore the incorporation of additional features, social influence and market trends, beyond RFM metrics to improve the predictive accuracy of CLTV models. Advanced deep learning models can be implemented to refine the customer segmentation and prediction. Expanding the study across different datasets and industries can improve the generalizability of the findings. Key limitations that needs to be addressed, is variability in data quality, which has affected the performance of model. To overcome this limitation, more enhanced and advances data preprocessing models can be implemented, using high quality diverse datasets. Practical implementation of these models in real-world business enviornments, such as their application in dynamic marketing campaigns or personalized recommendation systems, can be important for promoting these insights to drive customer loyalty. To achieve practical success, it will be crucial to deal with operational and organizational constraints, like incorporating these models into current customer relationship management systems.

# References


Aliyev, M., Ahmadov, E., Gadirli, H., qizi Mammadova, A. I. and Alasgarov, E. (2020). Segmenting bank customers via rfm model and unsupervised machine learning, *ArXiv* **abs/2008.08662**.
**URL:** *https://api.semanticscholar.org/CorpusID:221186661*

Alsharafa, N. S., Madhubala, P., Moorthygari, S. L., Rajapraveen, K. N., Kumar, B., Sengan, S. and Dadheech, P. (2024). Deep learning techniques for predicting the customer lifetime value to improve customer relationship management, *Journal of Autonomous Intelligence* .

Alzami, F., Sambasri, F. D., Nabila, M., Megantara, R. A., Akrom, A., Pramunendar, R. A., Prabowo, D. P. and Sulistiyawati, P. (2023). Implementation of rfm method and k-means algorithm for customer segmentation in e-commerce with streamlit, *Ilkom Jurnal Ilmiah* .

ASLANTAŞ, G., GENÇGÜL, M., RUMELLİ, M., ÖZSARAÇ, M. and BAKIRLI, G. (2023). Customer segmentation using k-means clustering algorithm and rfm model, *Deu Muhendislik Fakultesi Fen ve Muhendislik* .

Carneiro, F. and Miguéis, V. L. (2021). Applying data mining techniques and analytic hierarchy process to the food industry: Estimating customer lifetime value, *Proceedings*

*of the International Conference on Industrial Engineering and Operations Management* .

Fernandes, A., Giri, B., Edison, B. and Tripathy, A. K. (2023). Consumer behavioral segmentation analysis, *2023 International Conference on Advanced Computing Technologies and Applications (ICACTA)* .

Gadgil, K., Gill, S. S. and Abdelmoniem, A. M. (2023). A meta-learning based stacked regression approach for customer lifetime value prediction, *Journal of Economy and Technology* .

Hilmy, F. M., Nurhaliza, R. A., Octava, M. Q. H. and Alfian, G. (2023). Web-based e-commerce customer segmentation system using rfm and k-means model, *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)* .

Khumaidi, A., Wahyono, H., Darmawan, R., Kartika, H. D., Chusna, N. L. and Fauzy, M. K. (2023). Rfm-ar model for customer segmentation using k-means algorithm, *E3S Web of Conferences* .

Lewaa, I. (2023). Customer segmentation using machine learning model: An application of rfm analysis, *Journal of Data Science and Intelligent Systems* .

Priyadarshni, S., Fathima, R., Urolagin, S., Bongale, A. and Dharrao, D. (2023). Unveiling customer segmentation patterns in credit card data using k-means clustering: A machine learning approach, *2023 International Conference on Modeling, Simulation Intelligent Computing (MoSICom)* .

Raj, S., Roy, S., Jana, S., Roy, S., Goto, T. and Sen, S. (2023). Customer segmentation using credit card data analysis, *International Conference on Software Engineering Research and Applications* .

Rathi, M. S. and Karwande, P. V. (2022). Review paper on customer segmentation approach using rfm and k-means clustering technique.

Shirole, R., Shirole, R., Salokhe, L., Salokhe, L., Jadhav, S. and Jadhav, S. (2021). Customer segmentation using rfm model and k-means clustering, *International Journal of Scientific Research in Science and Technology* .

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* .

Surti, M., Shah, V., Bharti, S. K. and Gupta, R. K. (2023). Customer lifetime value prediction of an insurance company using regression models, *2023 International Conference for Advancement in Technology (ICONAT)* .

Wang, S., Sun, L. and Yu, Y. (2024). A dynamic customer segmentation approach by combining lrfms and multivariate time series clustering, *Scientific Reports* .