National College of Ireland

Utilizing Predictive Analytics to Enhance Retail Business Performance

MSc Research Project
MSc in Data Analytics

Aravind Hallimysore Kalegowda
Student ID: X22104275

School of Computing
National College of Ireland

Supervisor: Mayank Jain

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: …….Aravind Hallimsyore Kalegowda……

Student ID: ……x22104275…..……

Programme: …… MSc in Data Analytics …… **Year:** ……2023……..

Module: …… MSc Research Project ……

Supervisor: ……Mayank Jain……

Submission Due Date: ……31/1/2024……

Project Title: Utilizing Predictive Analytics to Enhance Retail Business Performance

Word Count: ……9700…… **Page Count**……34……..

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: ……Aravind Hallimysore Kalegowda……

Date: ……28/1/2024……

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):	

Utilizing Predictive Analytics to Enhance Retail Business Performance

Aravind Hallimysore Kalegowda
X22104275

Abstract

The study focuses on the changing world of e-commerce, specifically looking at how important it is to group customers and predict sales to improve business strategies, with a special interest in customer behavior. The goal is to use predictive analytics in retail to make better business decisions and improve how businesses operate.

The research uses predictive models and K-means clustering to make better decisions and work more efficiently in retail. The findings show that machine learning models like Random Forest and CatBoost are very good at predicting retail sales, with an R^2 score of 0.98. K-means clustering effectively groups customers based on how recently and frequently they buy, and how much they spend, leading to a Silhouette score of 0.93. This helps in creating focused marketing strategies. Overall, the study shows that advanced machine learning models are great for predicting retail sales and that K-Means clustering is useful for grouping customers and planning sales strategies, which helps in making smarter business decisions.

1 Introduction

The rapid evolution of the retail industry necessitates innovative solutions to increase decision-making and performance. Leveraging predictive analytics addresses the requirement by offering insights into customer behaviour, pricing strategies, and males' prediction, and market dynamics.

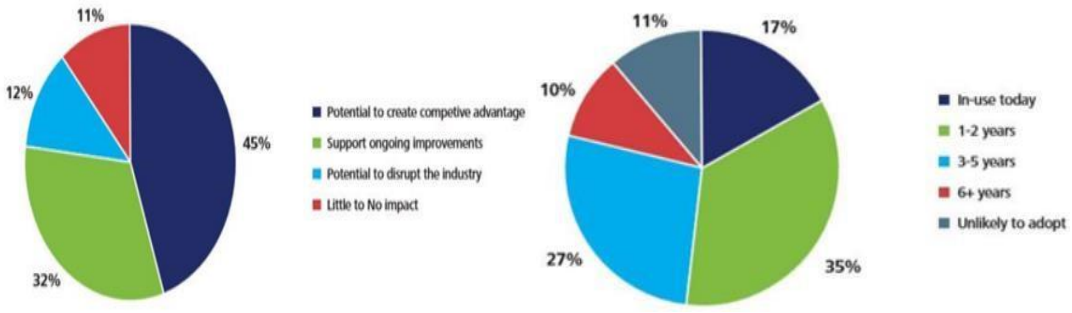

Figure 1: Trend of predictive analytics for the retail industry in 2019 (Vyas, 2019)

Figure 1 illustrates the historical and projected trajectory of predictive analytics within the retail sector, portraying market size dynamics from 2016-2022. The forecast indicates a market size of 6.2 billion U.S. dollars in 2018. Additionally, Wassouf et al., (2020) have

emphasised the advantages and challenges of applying predictive analytics in retail, underscoring the potential for more profitable pricing decisions. In addition, Bousdekis et al., (2021) have highlighted the inadequacy of conventional decision-making methods to necessitate data-driven approaches. This background underscores the significance of the proposed project in addressing the unrealised potential of predictive analytics in the retail sector, aligning with the evolving demands of the retail industry.

Research Question and Objectives

R^2 The research mainly aims to explore the full potential of predictive analytics in retail by handling the challenges businesses face in efficiently leveraging data.

The research study aims to harness the power of predictive analytics in the retail sector for enhanced decision-making and improved operational performance.
The main objective of the research to understand the implementation of predictive analytics in retail sector and the role it plays in improving business decisions and operational performance.
The question that the research tends to answer is that How implementation of Predictive Analytics can be applied in the retail sector for improving business decisions and improve operational performance?
The research introduces novelty by comprehensively addressing the unrealised potential of predictive analytics in retail decision-making. The research pioneers the development of intricate predictive models by integrating diverse data resources, involving customer interactions, market trends, and historical sales data. The emphasis on optimizing inventory levels, targeted marketing, and sales predicting aligns with contemporary challenges in the retail sector, offering a unique and impactful contribution to advancing the significant application of predictive analytics for heightened operational performance.

Outline the Structure of the Project

The introduction chapter helps to understand the principal aim of this project along with the significance of predictive analytics in retail. Related Work reviews existing literature and studies to provide a comprehensive understanding. Moreover, the Research Methodology contains a detailed approach taken in developing predictive models, involving data sources, and techniques. Additionally, Design Specification defines the structure and components of the predictive models, aligning with research objectives. The implementation then defines the practical execution of the predictive models, incorporating customer segmentation and sales predictions. The evaluation chapter assesses the effectiveness and accuracy of the implemented models. Lastly, Conclusion and Future Work summarizes key findings and outlines avenues for further research in predictive analytics for retail.

2 Related Work

Within the context of the retail industry, sales prediction plays a significant role in managing supply chain networks and operations between manufacturers and retailers. Consequently, external factors such as market volatility may also influence the accuracy of predictions. These limitations underscore the requirement for a nuanced approach and strategic considerations to deploy predictive analytics within the retail sector. As per the viewpoint of Sajawal et al. (2023), sales forecasting is a challenging task for the maintenance of effective

inventory management, customer services and financial planning in the retail sector. The study by Sajawal et al. (2023) demonstrated that changing consumer behaviour, economic conditions and supply-demand equilibrium plays a significant role in the prediction of retail sales.

According to Seyedan and Mafakheri (2020), customer behaviour and market trends fundamentally create an impact on retail sales prediction. The research by Seyedan and Mafakheri (2020) has applied clustering and neural network algorithms for predicting retail sales and demand. Findings stated that shifts in consumer demand due to external conditions like economic downtime, social and political issues and more significantly impacted retail sales. Main strength of the review is the use of regression and neural networks for predicting retail sales. On the other hand, main drawback of the review was the lack of association between sales forecasting and closed-loop supply chain (CLSC) operations.

According to Nouf Fahad Al-Mufadi et al. (2023) data mining can be considered as a fundamental component of modern businesses, providing elements to diverse businesses in enhancing customer affinities along with improving sales performance through the prediction of future sales. Nouf Fahad Al-Mufadi et al. (2023) highlighting the low accuracy and low explanatory power of the predictive model indicating that proper sales cannot be predicted. Yet it aims to develop machine learning models that can accurately predict retail sales with high explanatory power.

The study by Saraswathi et al. (2021) proposed a retail sales prediction system based on a machine learning algorithm utilising sales data of Rossamann stores. As per the viewpoint of Saraswathi et al. (2021) sales forecasting in retail landscape is highly influenced by the purchasing behaviour of customers. The showed that obtained from the predictive models have shown a low R^2 which is coefficient of determination, questioning the explanatory capability of the model. This can be treated as the main drawback of this study. Therefore, the identified research scope is to develop a sales prediction system that can help retail companies project sales, leading to boost demand growth and high sales generation.

Customer segmentation is an important aspect of marketing activities in the retail sector, helping retail companies to deploy customer-centric targeting strategies. The study by Miguel Alves Gomes and Meisen (2023) applied RFM methodology for developing clusters of retail customers by developing the KMeans clustering algorithm (Wong, 1979). Main strength of the review is the high silhouette score, denoting the compactness of the samples with the defined clusters. Similarly, Anitha and Patil (2019) have also applied the RFM approach for the development of the segmentation of retail customers based on recency, frequency, and monetary value of purchase. The obtained silhouette score was close to 1, representing optimal comparatively to other clusters. On the other hand, limitations are included in terms of the scope of further investigation of the Sales Recency, frequency and Monetary for the development of optimal clusters. The study by Tripathi, Bhardwaj and Poovammal (2018) utilised the clustering method for segmenting the customers based on monetary value. Due to the exclusion of recency and frequency of purchasing, developed clusters imposed a lack of feasibility and applicability in the real-world retail landscape, demanding further optimisation. On the other hand, Tabianan, Velu and Ravi (2022) included recency, monetary and frequency aspects of customer segmentation to develop an optimal number of clusters based on geographic, demographic, and psychographic profiles of the customers. The strengths of the research study underline in the practical applicability of the solution for the segmentation of retail customers. But there is a noticeable research gap in investigating the optimal combination of sales recent, frequency and monetary value to evaluate clusters that align with the diverse dynamics of real-world retail landscapes.

Bradlow et al. (2017) have delved into the transformative role of big data and predictive analytics in the retail section. It comprehensively explores five key dimensions of data in

retail such as customer, product, time, location, and channel. The research paper also underscores the evolving landscape of retail through the integration of new data resources, statistical tools, and domain knowledge, encompassing the continued relevance of theoretical insights. Methodologically, the endeavour employs Bayesian analysis techniques and predictive analytics on big data, supported by a significant field experiment. The authors highlight the ethical and privacy considerations associated with big data in retail. A notable strength lies in its multidimensional approach, contributing to the nuanced understanding of the field. Consequently, potential limitations involve the specificity of the retail context and the requirement for cautious data handling. The research helps in exploring the applications and challenges of primitive analytics in retail. This paper inspires the research study by emphasising the multi-dimensional impact of big data and predictive analytics along with specificity in retail.

The research by Alqhatani et al., (2022) represents a comprehensive framework for retail business analytics, integrating hybrid machine learning and business intelligence. The approach combines descriptive and diagnostic analysis through business intelligence modelling, enabling executives and managers to conduct adequate analyses. Machine learning modelling and clustering are employed for predictive analytics, highlighting potential customers, products, and time intervals. The author uses a resource-efficient and automated framework, integrating seamlessly with operational data pipelines. Strengths include the holistic 360-degree view, efficient data analysis, and a dynamically filtered dashboard.

Multiple organisations in the retail industry utilise the latest technologies in order to improve their business performance as well as customer satisfaction. Malik et al. (2023) focused on small suppliers and merchants, using machine learning algorithms such as XGBoost, Random Forest in order to estimate product demand, thereby assisting improved inventory control. The findings by Malik et al. (2023) suggest strategies that use ample data to improve forecasting accuracy and therefore also maximises profit gains. Despite significant focus on machine learning model development, the research fails to appropriately address the challenges being faced by the retail firms for which the models are developed.

Pavlyuchenko and Panfilov (2021), have focused on the FMCG market which uses broad scale predictive analytics to improve sales planning and forecasting within the existing business model. The findings by Pavlyuchenko and Panfilov (2021) has resulted in the development of a predictive model utilising machine learning models which can increase efficiency of company stock as well as shipment planning. However, the accuracy of the models is dependent on the data input which can affect different companies based on their sales and stock data.

Big data technology and predictive analysis demonstrate advanced potential for business intelligence, which significantly help organisations in dynamic decision making. According to Panda (2020), in the competitive retail industry, knowing customer behaviour using analytics is significant for sustained growth and survival. The research emphasised multinational retail operators who are implementing customer analytics systems and Big Data Analytics, especially in areas such as market research, sales forecasting, product optimization and targeted marketing. Meanwhile, the research by Chen, Li and Wang (2022) highlights the over-reliance on big data and predictive analytics for business intelligence through a systematic analysis. The findings by Chen, Li and Wang (2022) emphasised the increasing importance of the latest technology, especially in handling retail firms after the pandemic and provide significant operation management approaches for future operations of the firms. This research is significant in understanding the dependence of retail businesses on big data and how this can affect the business performance.

Retailing industry today is changing at a faster pace with the help of the latest technologies such as big data and analytics. Verma, Malhotra and Singh (2020) have addressed the challenges experienced by retailers in the present era of ecommerce shopping through a MapReduce framework in big data analytics. The findings have focused on the transition from conventional to modern retailing, focusing on the need of merchants to understand and meet customer expectations. The MapReduce-based Apriori method, implemented as the intelligent retail mining tool (IRM tool), helps to effectively discover shopping trends while also resolving scalability issues and fault tolerance constraints. On the other hand, Saurav (2018) highlighted the transition of the retail industry in which the path of consumer purchasing choices is digitised. The emphasis has changed from data collecting to insight extraction for differentiation, competitive advantage and improved buyer experiences. However, the persistent challenge is dealing with a massive amount of unstructured big data. Saurav (2018) identified the potential of predictive analysis through A/B testing on online activity data of customer demands, preferences. However, there remains a challenge in incorporating incremental revenue generation systems in the model which can further boost business performance.

The retailing industry has been a major contributor to the global economy, which is significantly influenced by predictive dynamics. Rooderkerk, DeHoratius and Musalem (2022) conducted an extensive study on the evolution of retail Analytics, evaluating different academic studies and interviewing retailing practices. The bibliometric research found a fast-increasing dynamic sector with major decision areas, a heavy reliance on perspective Analytics and a growing emphasis on big. Rooderkerk, DeHoratius and Musalem (2022) revealed that retail analytics leads to adoption challenges for many retailers who are not introduced to the latest technologies. The authors highlighted the "think big, start small, scale fast" strategy in order to cope up with the retaining analysis in daily retail operations. In contrast, Smith and Côté (2022) investigated the influence of predictive analytics in pop-up retail stores. Their findings highlighted those predictive analytics have reduced unsold products by 40%, allowing supply chain management to be optimised for improved efficiency. However, there is a gap in qualitative explanation of the challenges faced by retailers across the world.

The integration of real time video analysis and consumer transaction data demonstrates an innovative technique for offline purchasing decision making with current advances in predictive modelling. Li et al. (2023) showed the significance of connecting emotional responses in customer conduct analysis by carrying out computer vision strategies, facial recognition and AI models. The findings connected the retailing elements by demonstrating the usefulness of video-based content in distinguishing complex buying decision determinants, as well as furnishing advertisers with a significant video-based content arrangement. Conversely, Ali and Essien (2023) examined the meaning of Big Data Analytics in advancing outbound logistical factors in the retailing business. The non-transferable nature of perceived Big Data Analytics (BDA) benefits was highlighted in their findings, highlighting the significance of top-level management support in a variety of organisational contexts. Ali and Essien (2023) recommended a structure that portrays what changes in natural situations can mean for BDA direction, stressing the variables, for example, supply chain development levels, connectivity and outsourcing services which impact business execution.

The retail industry is going through an enormous change because of the consolidation of most recent innovations like big data which assists in better comprehending client requests and buying expectations. Prasad and Venkatesham (2021) featured the impact of big data on

retail, showing strategies, for example, in store advertising approaches, customised coordinated marketing, perceiving significant shoppers and noticing client propensities for buying. Simultaneously, Ibrahim and Wang (2019) utilised Twitter data to recognise the difficulties of online retail and its clients, focusing on delivery, product and client care as huge subjects for the sentiment analysis. The finding emphasised the significance of user generated content by highlighting its impact on improvement of online retail services and addressing growing concerns such as online engagement and in-store experience. Using user generated content has been helpful for in-depth analysis of customer loyalty, brand recognition and customer satisfaction.

Aim of the research studies	**Method**	**Findings**	**Limitations**
Predicting retail sales of Citadel POS using machine learning models (Sajawal et al. 2023)	Dataset used (Citadel Point of Sales) Models – Linear Regression, Random Forest Regression, Gradient Boosting Regression, Time Series models like ARIMA, LSTM.	XGBoost Regression outperformed Time Series achieving the best performance of MAE 0.516 and RMSE 0.63	Limitations involve potential overfitting with a small dataset and sensitivity to hyperparameter tuning with definite challenges to capture complicated temporal patterns.
Investigating predictive Big Data Analytics in demand for supply chain forecasting (Seyedan and Mafakheri 2020)	Time-series forecasting, clustering, KNN, neural networks, regression, SVM	Lacking in the proper application of BDAs for forecasting demand for closed looped supply chains.	This includes a scarcity of studies focusing on the applications of BDA in Closed-Loop Supply Chains with reverse logistics. it needs more investigation into datadriven approaches in these specific domains.
To develop a linear regression model for analysing superstore sales and propose the best method (Nouf Fahad Al-Mufadi et al. 2023)	Linear Regression	The study achieved an R2-score of 0.016 and MAPE of 27.8%.	The study indicates limited exploratory power in the linear regression model, suggesting moderate predictive accuracy.

To forecast the sales of Rossamann stores using machine learning. (Saraswathi et al. 2021)	Simple and Multiple Linear Regression	The results helped in boosting demand growth and sales.	The study involves potential over-reliance on linear regression which oversimplifies complicated sales dynamics.
To provide a structured overview of segmentation processes (Miguel Alves Gomes and Meisen 2023)	105 publications from 2000 to 2022 and conduct systematic literature	K-means clustering is the most commonly used segmentation.	Limitations involve potential bias in the selection of publications as the study focuses on the 2000 to 2022 timeframe.
To analyse historic sales data and purchasing pattern of customers (Anitha and Patil 2019)	K-means clustering has been implemented based on transactional and retail dataset	The results have shown that k-means clustering has helped in segregating sales based on Recency, Frequency and Monetary.	The analysis relies on historical data, and other factors that are affecting purchasing behaviour may not be fully captured in this study.
Explore importance of customer segmentation for achieving competitive advantage (Tripathi, Bhardwaj and Poovammal 2018)	Implementation of KMeans and Hierarchical Clustering	Clustering helps in understanding commercialisation strategies for making strategic decisions.	Limitations involve thorough clustering, potential sensitivity to initial conditions in K-Means along with the challenge of capturing significant relationships among variables in hierarchical clustering.
The aim is to optimise the experimental similarities within clusters of E-commerce customers. (Tabianan, Velu and Ravi 2022)	E-commerce dataset of Malaysia from MDEC repository for machine learning development based on customer behaviour and implement K-Means Clustering.	The segmentation of customer has been done properly based on Kmeans clustering.	It encompasses potential bias in the selected E-commerce dataset and generalisability challenges to broader customer behaviour beyond the scope of the dataset.
The paper aims to examine the opportunities and possibilities of big data retailing related to customer, product, and time (Bradlow et al. 2017).	Statistical methods for predictive analysis.	A proper understanding of privacy issues can be achieved.	Challenges occurred during the representation of realworld complexities in statistical models with evolving privacy concerns.

To manage inventory of small stores Malik et al. (2023)	Predictive analytics system through machine learning	Maximising profits for small stores along with sales forecasting	Lack of qualitative instances and current circumstances of retail industry
To improve sales planning and forecasting within the existing business model Pavlyuchenko and Panfilov (2021)	Predictive model utilising machine learning models	Increase efficiency of company stock as well as shipment planning	Accuracy of the models are dependent on the data input
Analyse customer behaviour, buying patterns Panda (2020)	Systematic review	Omni channel marketing strategies, implementing customer analytics systems and Big Data Analytics	Lack of focus on predictive modelling implementation
To explore current research studies, historic developing trends, and the future direction Chen, Li and Wang (2022)	681 non-duplicate publications	Increasing importance of the latest technology, especially in handling retail firms	More focus on literature study, rather than quantitative analysis of the practical problem
To address the challenges experienced by retailers in the present era of ecommerce shopping Verma, Malhotra and Singh (2020)	MapReduce based Apriori (MR-Apriori) algorithm in the form of Intelligent Retail Mining Tool	Justify the effectiveness of the proposed algorithm through speed-up, size-up, and scale-up evaluation parameters	Lack of real time analysis

To emphasise path of consumer purchasing choices Saurav (2018)	Predictive analysis through A/B testing on online activity data	Identified the potential of predictive analysis	Limited evaluation metrics
To analyse the past, present, future of the	Bibliometric research of 123 retail analytics from	Found a fast increasing dynamic sector with	Over-dependence on
retail analytics Rooderkerk, DeHoratius and Musalem (2022)	2000 to 2020	major decision areas, a heavy reliance on perspective Analytics and a growing emphasis on big	historical data

Table 2: Understanding the related work.

3 Research Methodology

3.1 Detailed procedures of this research

In order to the research for proper prediction of sales and understanding of customer segmentation the following steps have been implemented:

Step 1: Collecting Data from Retailers and Online Sources

Historical sales data including consumer interactions and quantitative retail market data were gathered from participating retailers along with reliable online resources. This has ensured compliance with data protection legislation and ethical considerations. The store sales data involves the details of the store with the specific amount of store sales with ID. And also the online retailer dataset includes the invoice no, stock code and customer ID with invoice date helping to predict the customer segmentation with the utmost accuracy.

Step 2: Feature Engineering and Data Preprocessing

Data preprocessing (checking null values, handling outliers) and label encoding (converting categorical values into numeric and handling the missing values) have been performed in order to enhance the structural integrity and reduce data complexity in retail sales data. The data were well defined and refined through the application of feature engineering (data validations and data splitting) to enhance data reliability leading to the enhancement of model performance.

Step 3: Development and Training of predictive models

Predictive models like RandomForestRegressor (Breiman, 2001), XGBRegressor (dairu & Shilong, 2021),
LGBMRegressor (Harsha Chamara Hewage, n.d. 2021), CatBoostRegressor (Ding, et al., 2020), DecisionTreeRegressor (Arno De Caigny, 2018), and LinearRegression (Dastan

Maulud, 2020) have then been employed for retail sales predicting, along with pricing optimising using pre-processed data (Sajawal et al. 2023). On the other hand, K-Means Clustering was used for customer segmentation. At first, models are trained to identify patterns and correlations.

Step 4: Model validation

Predictive models are evaluated by using metrics such as Mean Absolute Error, Mean Squared Error, Root Mean Squared Error and R2 score (Coefficient of Determination). Cross-validation methods have also been employed to test models with data and assess their applicability in real retail environments. Silhouette score was considered as the performance metric for the K-Means clustering model (Nicholas, 2020).

3.2 Equipment and materials used in the research.

The entire study adheres to the CRISP-DM technique, contributing a systematic framework to guide the investigation of data and to maintain methodical clarity throughout the research process (Schröer, Kruse and Gómez, 2021). This has helped in understanding the process of using the dataset for predicting sales and doing customer segmentation analysis. The proper implementation of the secondary quantitative data has been used which has been collected from Kaggle related to online retail and Supermarket branch store sales dataset (Kaggle, 2023). This has helped in developing proper predictive models for predicting sales and also incorporating the suitable unsupervised machine learning algorithm such as clustering analysis to predict customer segmentation with utmost accuracy.

3.3 Gathering of the samples.

The description of the datasets is given below. The samples that have been selected for doing the research is based on purposive sampling where the Online Retail data is used for customer segmentation and Supermarket store branch sales are used for predictive analysis. Both the dataset of store sales and customer segmentation have some missing values and outliers, which is why the datasets were pre-processed and cleaned, and validated through some significant steps to ensure data integrity and reliability.

Online retail: The dataset is an transnational data containing all the transactions that has happened between 01/12/2010 and 09/02/2011. The dataset contains information for UK-based registered as well as nonregistered retail stores online. The dataset is related to the company that mainly deals with gifts for every occasion. The dataset also reveals that most of its customers are wholesalers (Kaggle, 2023a). The dataset includes variables like ‘invoice number’, ‘stock or product code’, ‘description of each product’, ‘quantity of products for every transaction’, ‘date and time of the invoice generated’, ‘unit price’, ‘ID of each customer and country’. The challenges associated with the dataset involve managing store-related data and handling detailed transaction records with relevant information such as invoice numbers, stock code and many more.

Supermarket store branch sales: The dataset contains information related to the sales of grocery stores in the US which are larger and provides wider options for customers (Kaggle, 2023b). The dataset includes information related to Store ID, Physical area of each store measured in Yard square. Number of different products, Average number of customers visiting the store and also the sales the store made in $s. This dataset has helped in implementing proper predictive methods for understanding sales of the company in the future.

The features that have been implemented in doing the calculation on raw data include data cleaning measurements like removing null, and duplicate values checking for outliers in the dataset, and removing the variable. The measurement of data cleaning in the raw data has

helped in improving the accuracy of the models developed as well as helping in summarizing the nature of the data. After cleaning the data, exploratory data analysis has been conducted.

3.4 Describing the statistical analysis used in the research.

Statistical analysis that has been used in the research includes descriptive statistics for showing the distribution and nature of the data. Correlation analysis has also been implemented for understanding relationship between the variables. Supervised machine learning models like Random Forest Regressor, XGB Regressor, LGBM Regressor, Cat Boost Regressor, Decision Tree Regressor, and Linear Regression have been implemented for predicting sales after splitting the data into training and testing which has helped in understanding the best model based on R-square, MSE, RMSE and MAE. Moreover, K-Means Clustering has also been used for customer segmentation based on optimal clusters.

4 Design Specification

The software analysis of the Store sales dataset mainly focuses on a comprehensive analysis of retail store data, leveraging diverse Python Libraries and machine learning models. The entire workflow can be broadly categorised into diverse specific stages such as importing necessary libraries, The models used involve Random Forest Regressor, XGB Regressor, LGBM Regressor, Cat Boost Regressor, Decision Tree Regressor, and Linear Regression. These models were selected for their diverse algorithms and abilities to capture complicated relationships within the data. The data preprocessing included removing outliers from the features to enhance model performance. Standard Scaler was then applied to scale all the features, and the dataset was split into training and testing sets (80/20 split) by widely using a random state of 42 for reproducibility. The models were eventually trained on the training set and evaluated on the test set by utilising metrics such as Mean Square Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) along with R-squared score. A data mining step involves creating additional features to enhance the predictive power and understand the purchasing power of a store, considering ratios of items to customers, store size to customers, and store size to items. K-Means Clustering was also applied to group the data into 10 different clusters, contributing to valuable insights into customer purchasing behaviour. The optimal number of the cluster is 4 which is considered for the entire development of the clustering algorithm.

The model comparison phase used 10-fold cross-validation to assess the performance of the models robustly. Cat Boost Regressor emerged as the top-performing model with the highest R-squared score. This comprehensive approach assures a thorough understanding of the data, effective preprocessing, and a model selected process that considers diverse algorithms along with features, contributing to a robust as well as accurate prediction of store sales in the retail domain.

For the development of the K Means Clustering for customer segmentation, the RFM approach has been taken into consideration. Along with recency, frequency, and monetary, a fourth feature is ‘interpurchase time’. For the initial development of the K Means model, an elbow method has been considered. In the elbow method, the elbow point indicates the instances where the rate of decrease slows down. The optimal number of clusters considered for the development of the K Means clustering algorithm is 4. In addition to that, the initialised K Means Model is developed with K= 4 and maximum iterations = 50. The centroid of each of the clusters is computed and the RFM data have been fitted to the Clustering model. For the evaluation of the performance of the K Means Clustering model, the Silhouette score is calculated which has helped in measuring how well-separated the clusters are and ranges from -1 to 1, where a higher score indicates better-defined clusters.

The development of the K Means model is based on a systematic approach, in which a systematic step-by step approach has been followed starting from selecting the initial K value to assessing the quality of clusters through the use of performance metrics like Silhouette score.

5 Implementation

The study employs a systematic approach, combining quantitative and qualitative methods to implement predictive analytics in the retail sector.

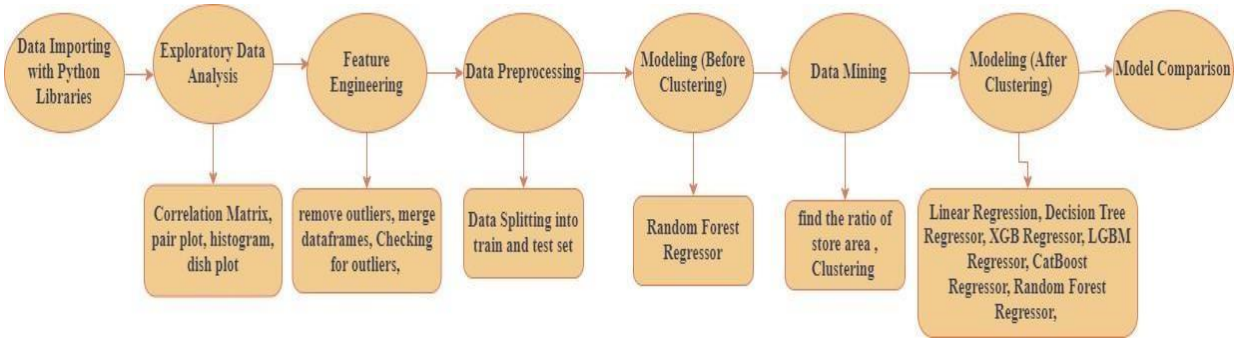

Figure 2: Flow Diagram for the proposed methodology of store sales prediction

In this comprehensive retail sales prediction, the implementation starts with importing necessary libraries for data manipulation, visualisation, and modeling [refer to figure 2]. The dataset "Stores.csv" is loaded using pandas, and initial exploratory data analysis (EDA) is performed. Key statistics and data visualisations such as correlation heatmaps, pair plots, and distribution plots contribute to valuable insights into the relationships between features such as store area, items available, daily customer count, and store sales. Feature engineering helps to create ratios for understanding purchasing power. Data pre-processing includes scaling and splitting the dataset into train and test sets. The modelling phase begins with radon forest regression (before clustering). Afterward, data mining is performed to determine the optimal number of clusters for target sales prediction using K-means clustering. The dataset is then joined with the target variable for clustering-based modeling.

After Clustering, the six regression models (Linear Regression, Decision Tree Regressor, XGB Regressor, LGBM Regressor, CatBoost Regressor, and Random Forest Regressor) are implemented with crossvalidation for robust evaluation. Rohaan, Topan, and Groothuis-Oudshoorn (2022), have stated that the kfold cross-validation helps in preventing overfitting. The performance metrics of these models involve MAE, MSE, RMSE, and R-squared scores, which are compared to identify the most effective model.

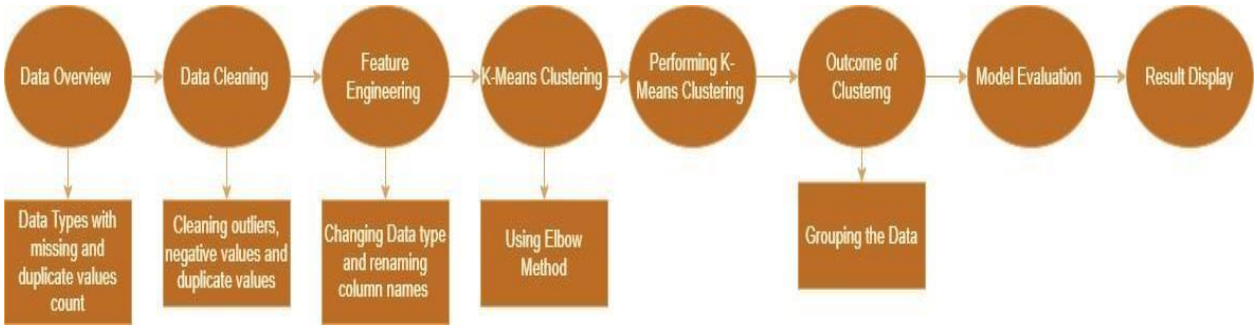

Figure 3: Flow Diagram for Customer Segmentation Analysis

In the implementation of the RFMT (Recency, Frequency, Monetary, Interpurchase Time) model, several crucial steps were undertaken to analyse and segment customer behaviour. The dataset underwent a comprehensive overview, involving shape, data types, missing values, duplicates, and statistics. Data cleaning included removing NaN values from the CustomerID column, catering products from InvoiceNo containing "C" and eliminating duplicates. Outliers in Quantity and Unit Price were also addressed using a threshold-based approach, enhancing the robustness of the dataset. Feature engineering played an essential role, introducing the RFM variables, and a novel feature, Interpurchase_Time, reflecting the average time gap between customer shopping trips. This is calculated based on the shopping cycle, offering a nuanced understanding of customer behavior.

The K-Means algorithm was employed for clustering, determining the optimal number of clusters (K= 4) using the Elbow method. The clustering results were evaluated using centroid visualisation and a scatter plot, depicting distinct clusters. Evaluation metrics, such as the Silhouette Score were used to measure the quality of clustering. The Silhouette Score of approximately 0.938 indicates a well-distinguished clustering structure. The robustness of the RFMT model is underscored by its capability to capture nuanced customer behaviour through thoughtful feature engineering along with the application of advanced clustering techniques.

6 Evaluation

This section mainly involves the EDA, and machine learning models implementation for sales prediction and customer segmentation. The study helps to explore the most effective model for store sales prediction in the retail sector by comparing all these advanced machine learning models.

6.1 Exploratory Data Analysis for Store Sales Prediction

The heatmap of **Figure 4** illustrates the correlation matrix of the variables in the Store sales dataset. Store areas and items available have a linear relationship, having more influence on store sales compared to daily customer count and item availability.

The seaborn pairplot visualises relationships between variables in **Figure 5** In this plot, the availability of the items and store area exhibit a strong positive correlation, suggesting that as the store area increases, the number of items available also tends to rise. The daily customer count also affects the frequency of the store sales.

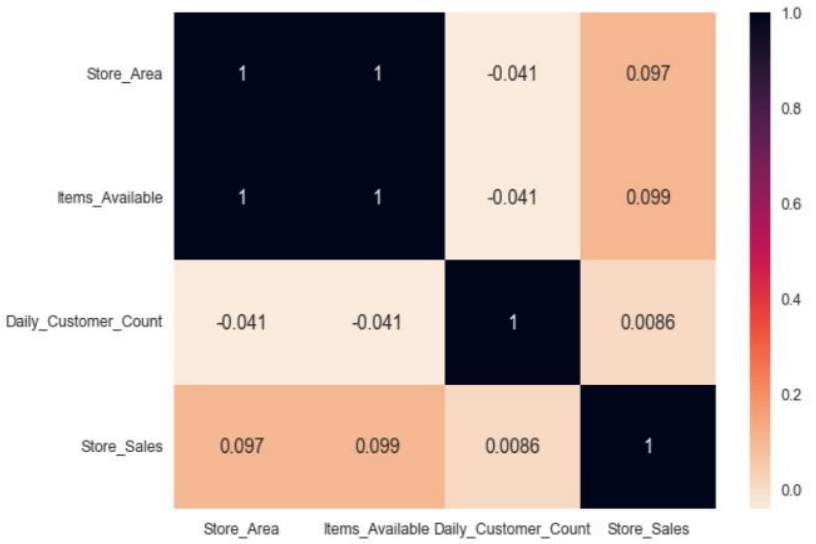

Figure 4: Correlation Matrix for Sales Prediction

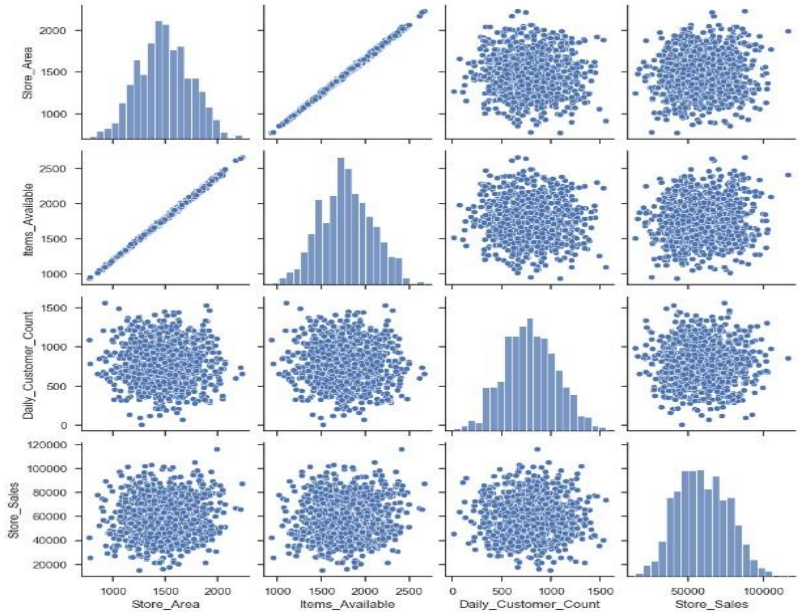

Figure 5: Pair Plot for Sales Prediction

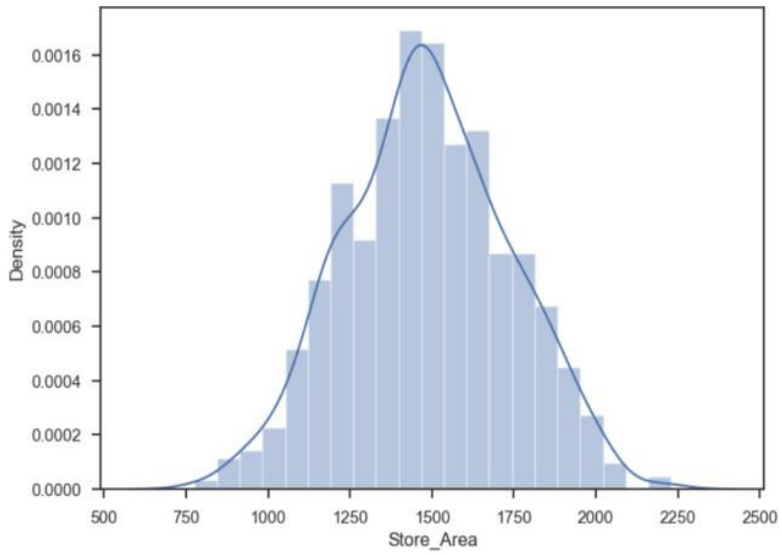

Figure 6: Distribution of Store Area

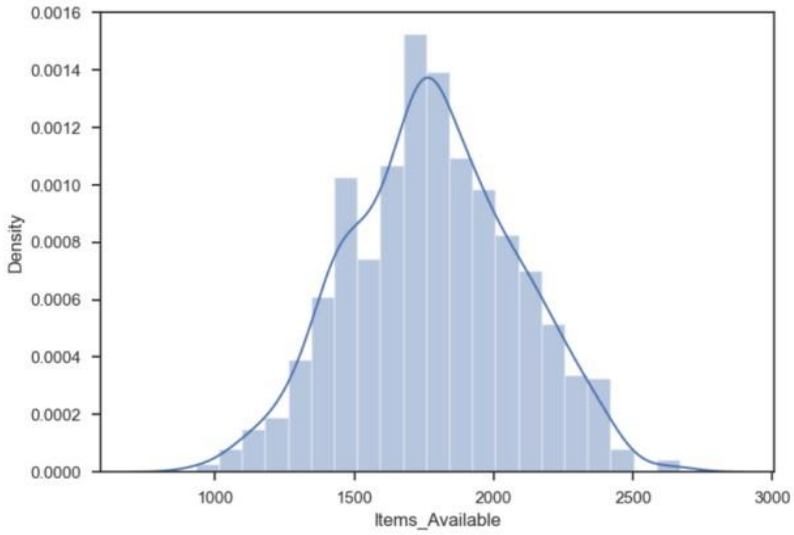

Figure 7: Distribution of Available items

The histogram in **Figure 6** shows the unimodal distribution of store area, suggesting that the majority of stores have a concentrated range of area sizes. This unimodal distribution indicates a concentrated range of sizes among the majority of stores, offering insights for effective resource allocation with strategic planning in the longer run.

Figure 7 exhibits the right-skewed distribution, indicating that more stores offer a limited range of available items. There are few stores with an extensive inventory.

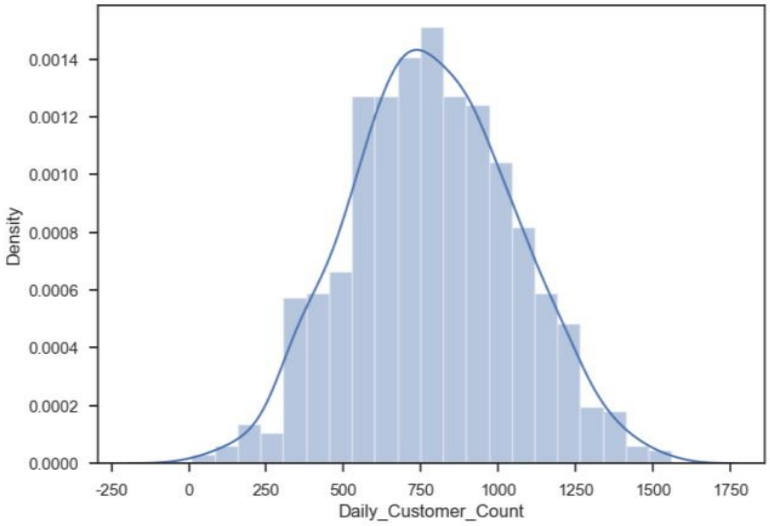

Figure 8: Density of Daily Customer Count

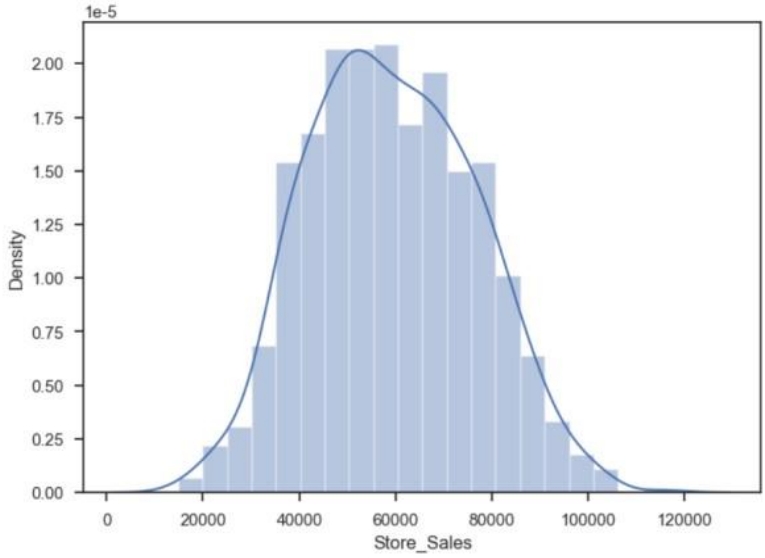

Figure 9: Density of Store Sales

The histogram of "Daily_Custmer_Count" in **Figure 8** represents a roughly normal distribution, implying that more stores attract a moderate to high number of daily customers, with a few outliers experiencing exceptionally high footfall.
The density of store sales reveals a right-skewed distribution, suggesting that a majority of stores generate a moderate to high range of sales, while a few outliers accomplish significantly higher sales volumes **[refer to figure 9]**.

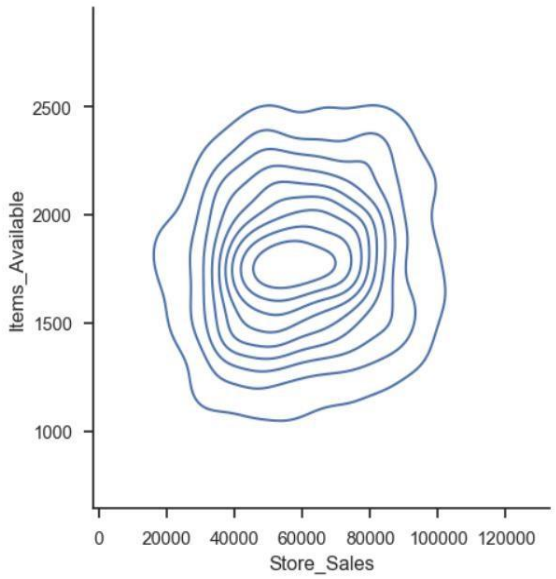

Figure 10: Items available vs. store sales

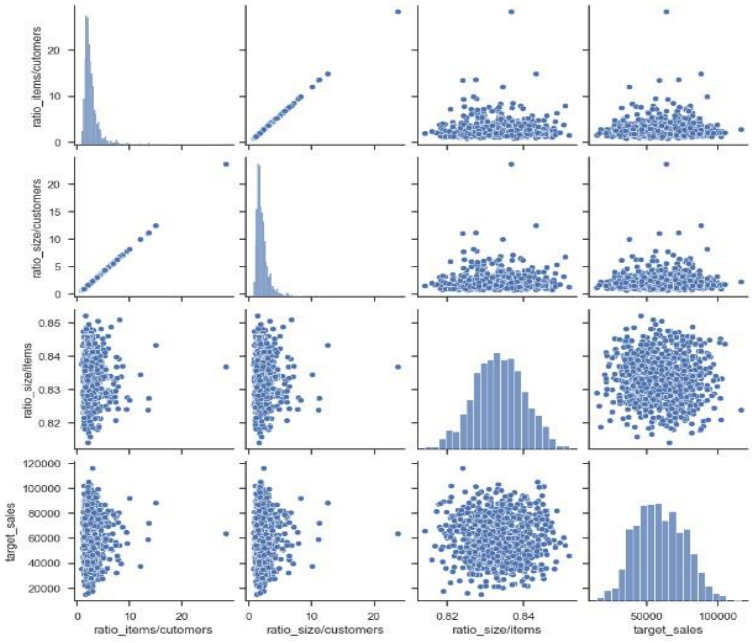

Figure 11: Pair Plot for average sales data

The seaborn displot in **Figure 10** shows a concentration of stores with moderate sales and a varying number of items available. **Figure 10** suggests that certain stores achieve high sales with diverse item counts.

The pairplot in **figure 11** visualises relationships among average target sales, items available, and daily consumers. This also indicates a positive correlation between sales and items available, implying that more items might contribute to higher sales.

6.2 Exploratory Data Analysis for Customer Segmentation Analysis

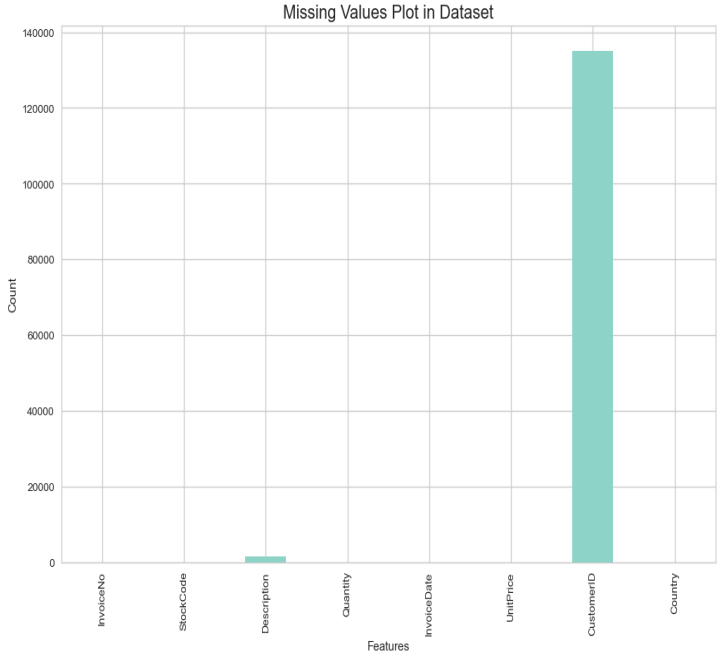

Figure 12: Count of missing values

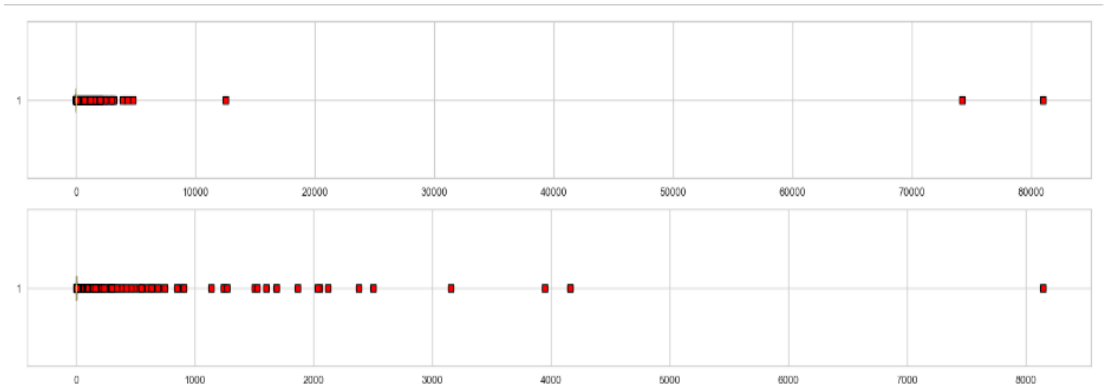

Figure 13: Checking Outliers

The missing values analysis of **Figure 12** reveals that the "Description" column has 0.27% missing values, while the "CustomerID" column has a substantial 24.93% missing values. This analysis helps to understand the requirement of data handling strategies for customer segmentation such as dropping duplicate and null values, checking negative values, and leaning outliers.

The boxplots of **Figure 13** illustrate the presence of outliers in "Quantity" and "UnitPrice". Outliers are identified as individual points beyond the whiskers, encompassing the need for outlier removal. Checking outliers is crucial in identifying and addressing data anomalies, assuring statistical robustness along with the reliability of analytical insights.

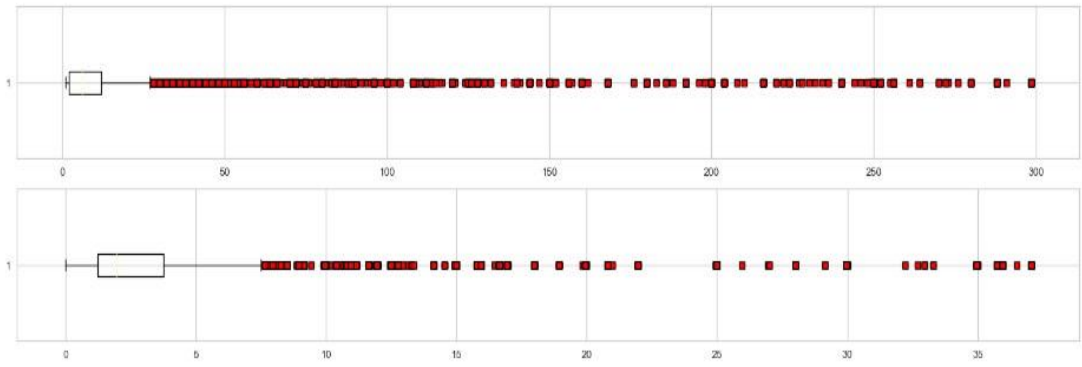

Figure 14: After removing outliers, checking them by box plot.

After removing the outliers based on the threshold values, the box plot ensures a more robust dataset. Postoutlier removal box plots show a comparatively cleaner distribution, confirming the successful data processing of the customer segment dataset.

6.3 Evaluation of Different Machine Learning Models for Store Sale Prediction

6.3.1 Random Forest Regressor

Model Name	**MSE**	**RMSE**	**MAE**	**MAPE**	**Training Score**	**Mean Crossvalidation Score**	**k-fold CV average score**
Random Forest Regressor	333971169.55	18274.88	15211.55	29.61%	0.41866597 56432295	-0.05	-0.07

Table 3: Random Forest Regression Output

Table 3 denotes the satisfactory performance of the Random Forest Regression model with MSE (333,123,879.02). RMSE (18,251.68), and MAE value (15,314.47). The MSE exceeds MAE, and RMSE values signifying sensitivity to outliers or skewed residuals. The positive training score (0.42) indicates the model captures underlying patterns. However, the negative mean cross-validation score and K-fold CV average scores (-0.04, -0.05) raise concerns about overfitting or inadequacies in the generalisation of the model. These results help to assess the model before clustering, providing insights into the intrinsic limitations and strengths especially for informed decision-making.

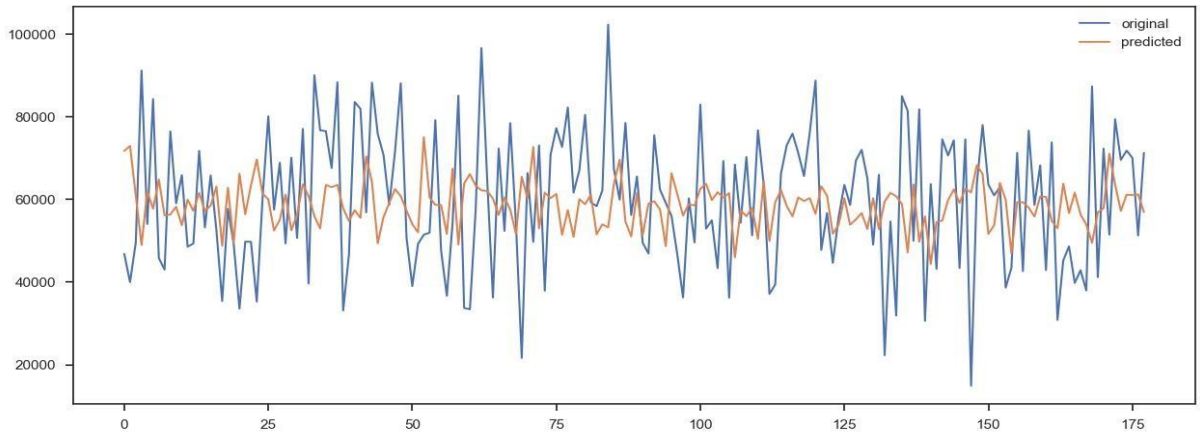

Figure 16: Original vs. predicted sales data

The plot compares the original sales values with the predicted values from the Random Forest model. **Figure 16** illustrates minimal differences between the actual and predicted sales values. The small divergences indicate that the model effectively approximates the true sales data, indicating its ability to make accurate predictions with slight variations.

6.3.2 K-Means Clustering

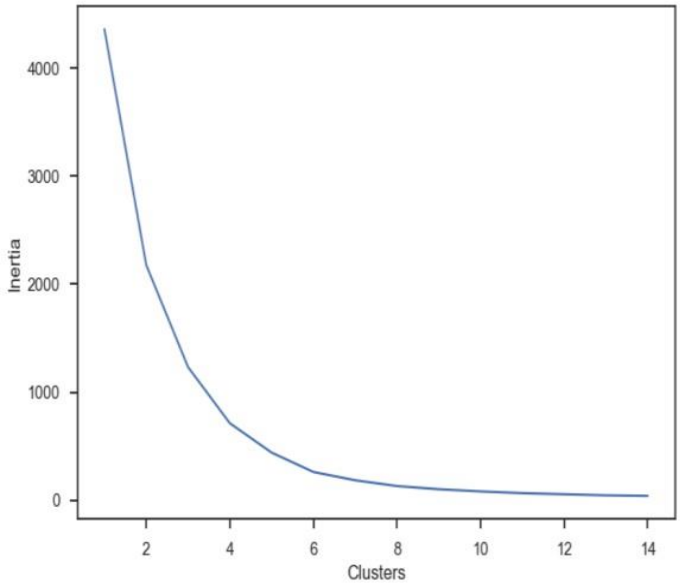

Figure 17: Elbow Method for Optimal K

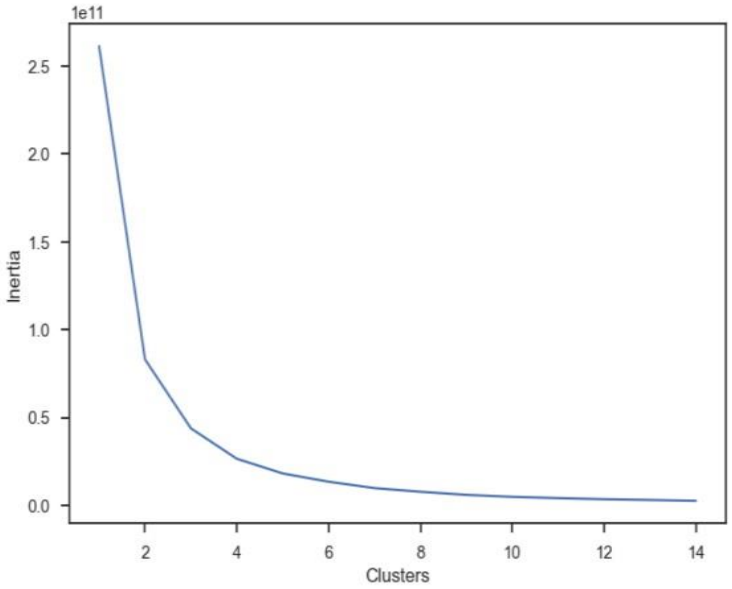

Figure 18: Elbow Method for Optimal K after creating a group of target sales

The graph displays the relationship between the number of clusters and the inertia, indicating the sum of square distances within each cluster. The Elbow in **Figure 17** occurs around 6 clusters, where the rate of inertia reduction slows down (parallel to the x-axis). This indicates that grouping the data into this number of clusters effectively captures the underlying patterns and variability in customer purchasing power.

Sl. No.	Item ratios for Customers	Size ratios for customers	Size ratio for items	Labels
0	3.700000	3.130189	0.845997	3
1	8.342857	6.957143	0.833904	3
2	2.234722	1.861111	0.832815	6
3	2.819355	2.340323	0.830092	6
4	4.691111	3.933333	0.838465	9

Table 4: Evaluating cluster labels to the original data.

The data has been clustered into 10 groups based on the ratios of items, store size, and their combinations. Each record shown in **Table 4** is assigned a label indicating its cluster. These clusters provide insights into different customer segments, potentially aiding store management strategies for enhanced business understanding and decision-making.

The graph in **Figure 18** shows an Elbow point at K= 6, indicating that the data can be grouped into 10 clusters based on the labels assigned in the previous step. This indicates a suitable segmentation for target sales groups. The inertia measures the sum of squared distance within clusters, and the elbow point signifies the optimal balance between

maximising inter-cluster variance and reducing intra-cluster variance (Nainggolan et al., 2019). This implies a meaningful grouping of target sales for improving the analysis and strategy formulation.

ratio_items/cutomers	ratio_size/customers	ratio_size/items	Labels	Target_Groups
3.7000	3.1301	0.84599	3	3
8.3428	6.9571	0.83390	3	5
2.2347	1.8611	0.83281	6	6
2.8193	2.3403	0.83009	6	6
4.6911	3.9333	0.83846	9	0

Table 5: Assigning cluster labels to the original data after creating a group of targets sales.

Table 5 displays ratios of items per customer, size per customer, and size per item, along with labels assignment during clustering (Labels) and subsequent grouping for target sales (Target_Groups). For instance, the item-to-customer ratio is 3.7, the size-to-customer ratio is 3.13, and the size-to-item ratio is 0.85, corresponding to Label 3 and target Group 3. Thus, these labels and target groups help categorise customers depending on their purchasing behavior.

6.3.3 Linear Regression

MSE	275247518.23
RMSE	16590.59
MAE	14121.19
Training Score	0.02002072
Mean cross-Validation score	-0.00
K-fold CV average score	-0.01
R_score	-0.01

Table 6: Outcome of Linear Regression

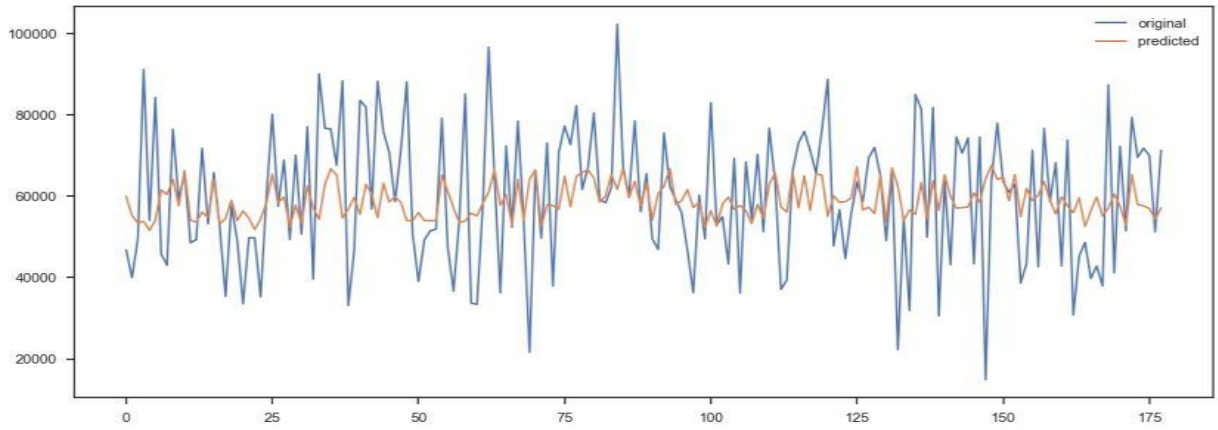

Figure 20: Original vs predicted sales data.

The Linear Regression model accomplished in **Table 6,** MSE of 275,247,518.23, RMSE of 16,590.59, and MAE of 14,121.19. The training score is 0.02, suggesting limited predictive power. The negative R-squared value (0.01) indicates the model performs poorly, possibly due to a weak relationship between the features and target variable. This emphasises the requirement for more sophisticated models or feature engineering. The results underscore the challenges in predicting the target variable utilising a linear regression approach in the context.

The predicted values shown in **Figure 20** do not precisely match the actual data, suggesting limitations in the accuracy score of the Linear Regression model. However, both the original and predicted data maintain a consistent trend, signifying that the model captures some aspects of the underlying patterns but lacks the precision to reproduce the exact values.

6.3.4 Decision Tree Regressor

MSE	9970390.45
RMSE	3157.59
MAE	2431.52
Training Score	1.0
Mean cross-Validation score	0.96
K-fold CV average score	0.96
R_score	0.96

Table 7: Decision Tree Regression

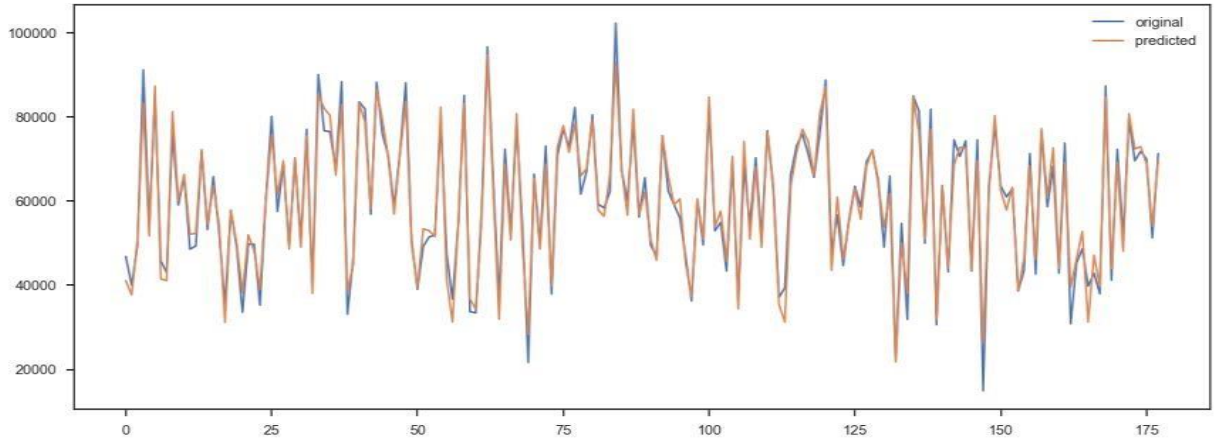

Figure 22: Plot for decision tree regression

The Decision Tree Regression exhibits superior performance compared to the Linear Regression model. As shown in **Table 7** it accomplished a comparatively lower MSE (9,970,390.45), suggesting better accuracy in predicting sales. The perfect training score of 1.0 along with high ross-validation scores of 0.96 highlight its robustness. This model outperforms Linear regression, showing its effectiveness in capturing non-linear relationships within the data.

The Decision Tree Regressor demonstrates remarkable accuracy in predicting sales, aligns closely with the actual data. The model captures intricate patterns, resulting in a visually impressive match between the original and predicted sales trends **[refer to figure 22]**. Its

capability to adapt to non-linear relationships contributes to this effective predictive performance.

6.3.5 XGB Regressor

MSE	7544231.76
RMSE	2746.68
MAE	2066.74
Training Score	0.999644
Mean cross-Validation score	0.97
K-fold CV average score	0.97
R_score	0.97

Table 8: XGBoost Regression model

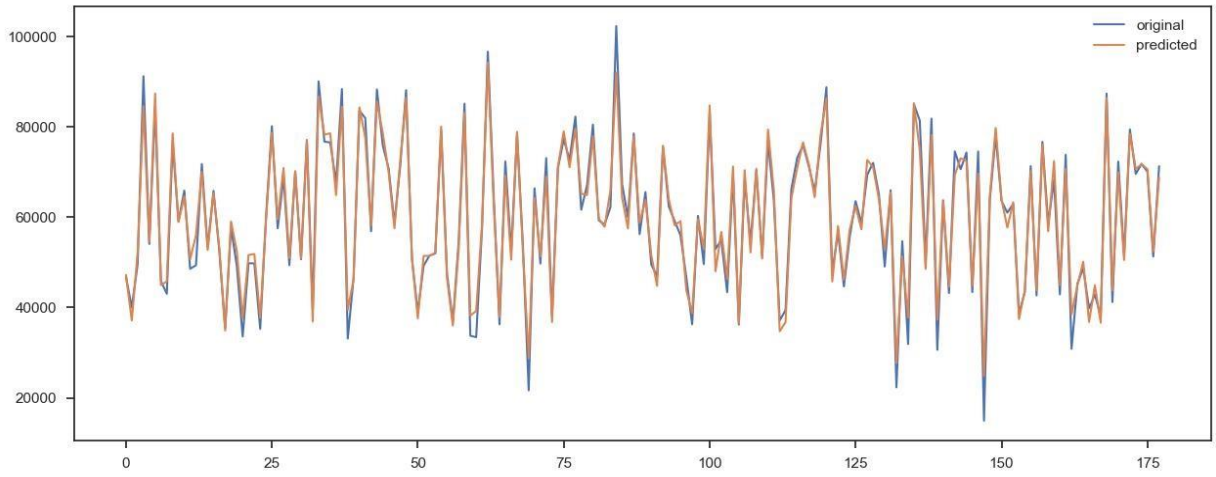

Figure 24: Plot for XGBoost regression model

The XGBoost regressor in **Table 8** illustrates superior predictive accuracy with comparatively low MSE (7544231.76), RMSE (2746.68), and MAE (2066.74). It accomplishes near-perfect training (0.99) and crossvalidation scores (0.97), outperforming both Decision Tree and Linear regression models. The high Rsquared value of 0.97 indicates an excellent fit to the data, encompassing its robustness and efficacy in capturing complicated relationships.

Figure 24 illustrates a close alignment between the original sales data and the predicted values generated by the XGBoost Regressor model. The smooth overlap indicates that the model effectively captures the underlying patterns in the test data, reinforcing its predictive ability and reliability.

6.3.6 LGBM Regressor

MSE	6213748.70
RMSE	2492.74
MAE	2006.19
Training Score	0.99210918121
Mean cross-Validation score	0.94
K-fold CV average score	0.94
R_score	0.94

Table 9: LGBM regression model

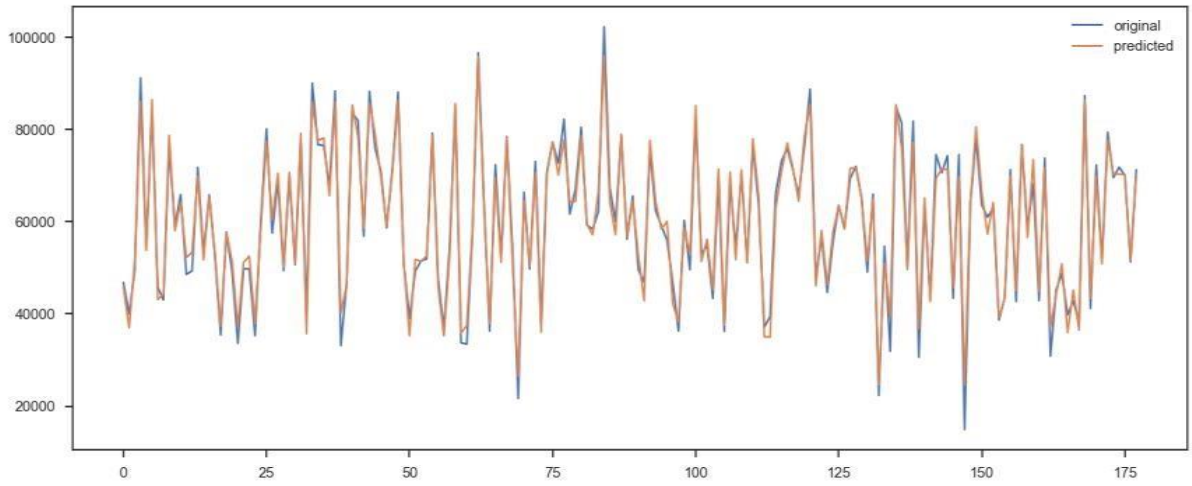

Figure 26: time series plot for LGBM model

The LGBM regressor illustrates shown in **Table** 9 strong predictive performance with an MSE of 6,213,748.70, RMSE of 2,492.74, and MAE of 2,006.19. Its high training score of 0.99 and consistent cross-validation score of 0.94 with an R-squared value (0.98) affirm its robustness, making it a competitive choice compared to the Decision Tree, Linear Regression, and XGBoost models.

Figure 26 displays a compelling alignment between the actual and predicted sales data, suggesting the accuracy of the LGBM model in capturing the underlying trends. The model effectively predicts sales, illustrating its ability for precise predictions and suggesting a reliable fit to the test data.

6.3.7 CatBoost Regressor

MSE	6736665.31
RMSE	2595.51
MAE	2001.20
Training Score	0.99455268
Mean cross-Validation score	0.98
K-fold CV average score	0.98
R_score	0.98

Table 10: CatBoost Regression

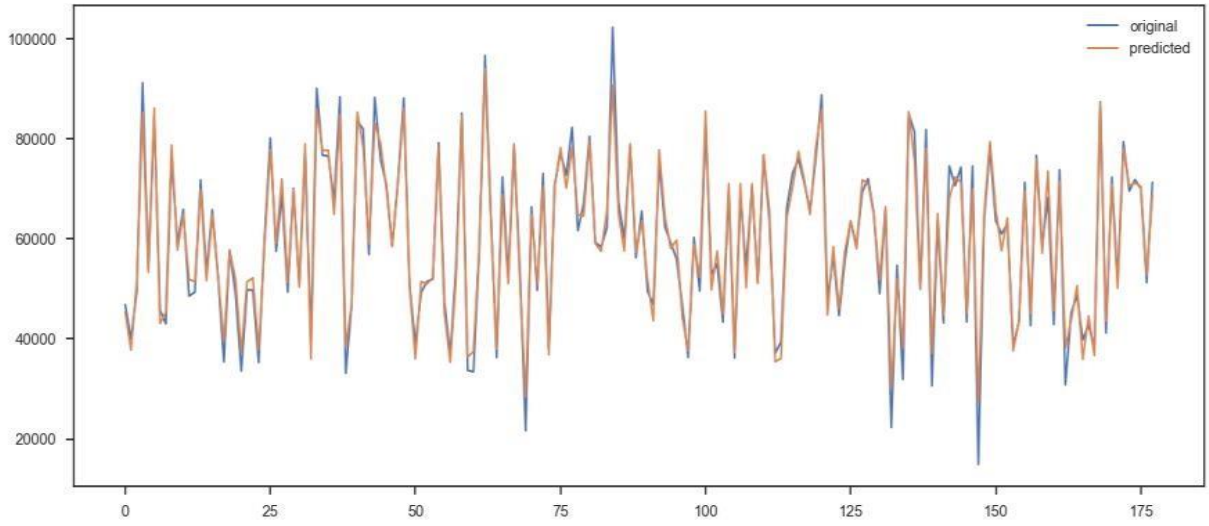

Figure 28: Sales test and predicted data plot.

The CatBoost Regressor exhibits impressive performance as shown in **Table 10** with an MSE of 6736665.31, RMSE of 2595.51, and MAE of 2001.20. It outperforms the LGBM Regressor, showing superior accuracy and predictive abilities. The high training (0.99) and cross-validation scores (0.98) emphasise its robust fit to the data.
The above plot represents a close alignment between the actual sales and the predicted sales data made by the CatBoost Regressor. In **Figure 28,** the predicted values significantly follow the trend of the original data, suggesting the accuracy of the model to capture the underlying patterns in the sales dataset.

6.3.8 Random Forest Regressor

MSE	6085161.78
RMSE	2466.81
MAE	1928.23
Training Score	0.9965419
Mean cross-Validation score	0.98
K-fold CV average score	0.98
R_score	0.98

Table 11: Random Forest Regression

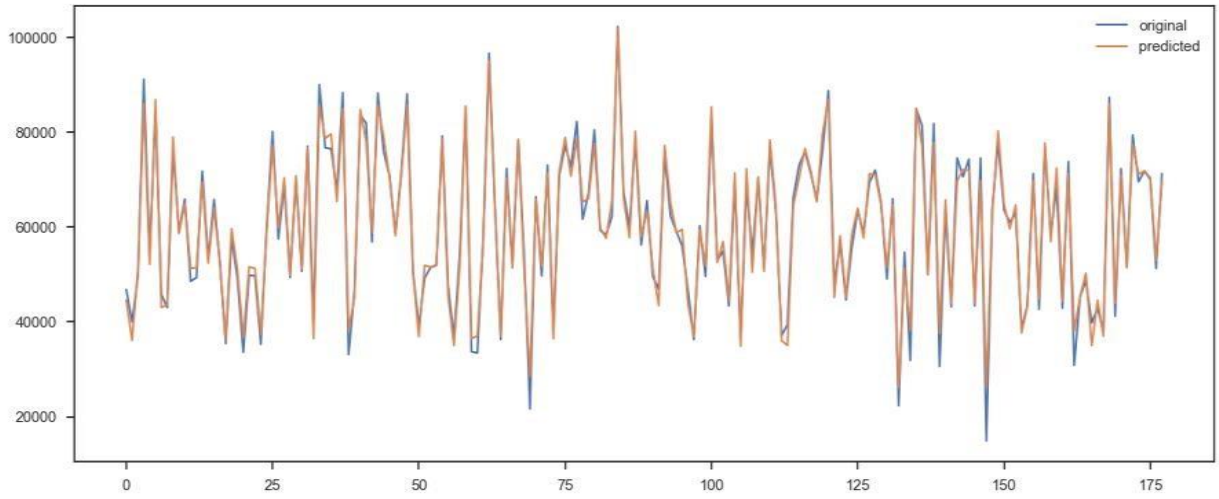

Figure 30: Sales Test and Prediction data for RF model

The current Random Forest Regressor outperforms the previous version as shown in **Table 11** (before clustering) showing significantly improved performance metrics. The MSE (6085161.78), RMSE (2466.81), MAE (1928.23), and R-squared score (0.98) indicate enhanced accuracy and predictive capability **[refer to Figure 29]**. This mainly underscores the significance of refining the model after clustering to accomplish superior results in sales prediction.

The graph in **Figure 30** illustrates the comparison between actual sales data and predicted values utilising a refined Random Forest Regressor. The close alignment between the two lines also implies the accuracy of the model reinforcing its effectiveness in predicting sales with improved precision.

Model Name	R-Squared Value	MSE	RMSE	MAE
XGB Regressor	0.97	8651550.52	2941.35	2181.04
LGBM Regressor	0.97	6867481.82	2620.59	2113.72
Cat Boost Regressor	**0.98**	7485044.54	2735.88	2116.07
Random Forest Regressor	**0.98**	**6283989.01**	**2506.79**	**2001.23**
Decision Tree Regressor	0.96	11514901.69	3393.36	2721.52
Linear Regression	0.00	270688558.55	16452.62	13125.63

Table 12: Comparing outcomes of Machine Learning Models

Depicts the entire R-squared value comparison for each one of the used machine learning models for sales prediction. The CatBoost Regressor and Random Forest model achieved the highest R-squared score of 0.98, surpassing XGBoost (0.97) and LGBM (0.97) as the CatBoost Regression model has comparatively lower MSE, RMSE, and MAE values. This suggests that CatBoost is the best-fitted model among all the regressor models, showing its superior performance in predicting retail sales.

6.4 Machine Learning Models for Customer Segmentation

6.4.1 K Means Algorithm

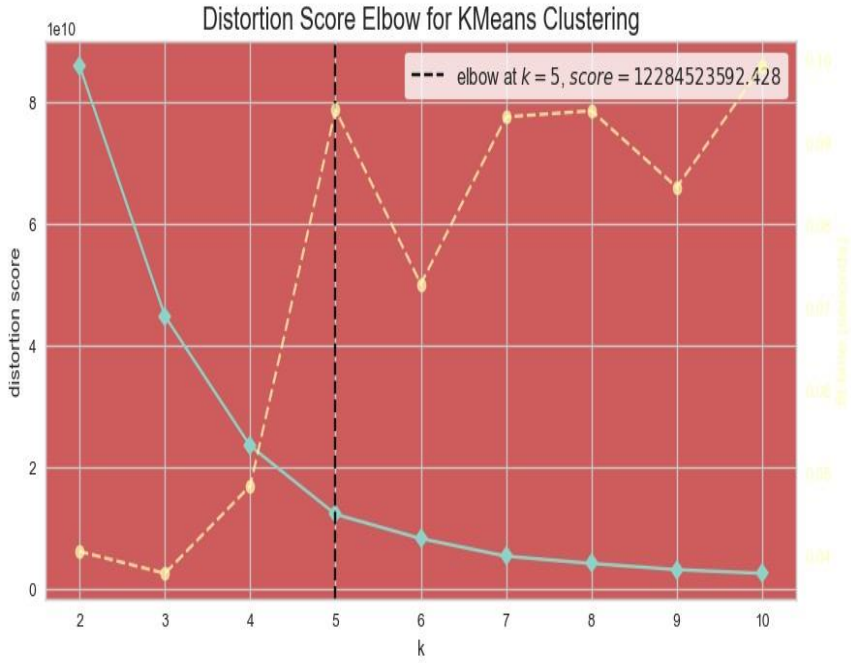

Figure 32: K-Means Clustering K-Means clustering for customer segmentation.

CustomerID	Recency	Frequency	Monetary	Interpurchase_ Time	Clusters
12347.000000	2	7	4310.000000	52	0
12348.000000	75	4	1770.780000	70	0
12352.000000	36	8	1756.340000	32	0
12356.000000	22	3	2811.430000	100	0
12358.000000	1	2	1150.420000	74	0

Table 13: Cluster labels to the original data

The Elbow Method is employed to identify the optimal number of clusters (k) for K-Means clustering. **Figure 32** displays the relationship between the number of clusters and the within-cluster sum of squares. In this case, the graph indicates diminishing returns in reducing the sum of squares after K= 5, Thus, K= 5 is identified as the optimal number of clusters for a balanced and effective grouping. The yellow line in the graph tracks the computation time

required to fit the KMeans model for each cluster count showing minimal variation across the range of k values tested.

Table 13 displays customer segments obtained from K-Means clustering based on recency, frequency, monetary value, and interpurchase time. For instance, customers in Cluster 0, characterised by moderate recency, high frequency, substantial monetary value, and shorter inter purchase time, illustrate a potentially valuable segment for targeted marketing strategies.

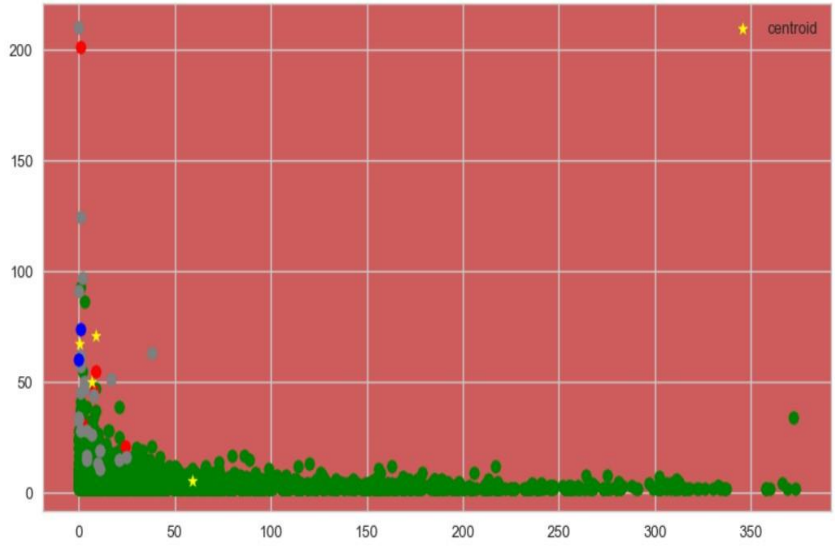

Figure 33: Quality of clustering in the data set

The scatter plot in **Figure 33** represents the quality of clustering in the dataset, in which each color represents a distinct cluster. Centroids (yellow stars) illustrate cluster centers. Notably, the clusters exhibit clear separation depending on recency, Monetary and frequency, suggesting effective segmentation by the K-Means algorithms for targeted marketing strategies.

```
Silhouette score : 0.9386494916906333
```

Figure 34: Evaluating Silhouette score.

The Silhouette score of 0.94 indicates excellent clustering quality, as it approaches the ideal value of 1. This metric evaluates how well-defined and separated the clusters are. The high score in ***Figure 34*** affirms the effectiveness of K-means in creating distinct and cohesive clusters based on Recency, Frequency, and Interpurchase Time.

6.5 Discussion

The study illustrates a unique and comprehensive approach to sales prediction and customer segmentation. This section helps to understand the effectiveness of the entire analysis of this study.

6.5.1 Sales Prediction Analysis

The study surpasses **"Sales Predicted Based on Data Mining Techniques"** in several aspects. This methodological diversity allows for a more nuanced analysis of sales prediction. Moreover, the incorporation of K-Means clustering for sales prediction enhanced the depth of this study, contributing to meaningful insights into diverse customer segments for targeted marketing. Moreover, the thorough evaluation using metrics like MSE, RMSE, MAE, and R-squared values (up to 0.98) along with diverse visualisations, confirming the effective performance of the model and addressing key research questions to contribute valuable insights in filling the gaps in the existing literatures. Additionally, while Zhang et al. (2023) focused on enhancing e-commerce management efficiency through a novel online sales predicting model (SFOR-ELM), the interactive refinement of the Random Forest Regressor in this study highlights a unique and effective strategy, resulting in superior predictive capabilities. Thus, the methodological richness and performance refinement of this study set it apart from Nouf Fahad Al-Mufadi et al. (2023) and Zhang et al., (2023), offering a more accurate approach to sales prediction in the retail domain.

6.5.2 Customer Segmentation Analysis

In this study on machine learning models for customer segmentation, the K-Means algorithm is applied, and the optimal number of clusters is determined using the Elbow method. The resulting clusters are analysed based on recency, frequency, monetary value, and interpurchase time, showing their effectiveness for targeted marketing strategies. This entire study and "RFM model for customer purchase behaviour using KMeans algorithm" by Anitha and Patil (2019), both use K-Means for customer segmentation, but the former additionally employs the Elbow Method for cluster optimisation. Anitha and Patil's method has accomplished a Silhouette score of 0.362 for 3 clusters and 0.349 for five clusters, suggesting moderate clustering effectiveness depending on Recency, Frequency, and Amount Logs. The proposed model illustrates superior performance, accomplishing a Silhouette score of 0.94 indicative of well-defined and separated clusters, enhancing its applicability for precise customer segmentation along with marketing strategy formation. In comparison with the research paper by Tabianan, Velu, and Ravi, (2022), both studies employ K-Means clustering techniques. However, the emphasis in the study by Tabianan, Velu, and Ravi, (2022) lies in optimising experimental similarity within clusters and maximising between clusters to analyse purchase behaviour. Unlike the broader focus on recency, frequency, and monetary values in the other study, this entire study specialises in understanding and leveraging customer behavioural factors, offering tailored strategies for sustainable customer satisfaction, and enhanced business profitability.

7 Conclusion and Future Work

7.1 Conclusion

In conclusion, this study represents a comprehensive and innovative approach to retail sales prediction and customer segmentation, surpassing existing models. Nouf Fahad Al-Mufadi et al. (2023) rely solely on linear regression, but this research study has explored a diverse range of advanced machine learning models, involving Random Forest, Decision Tree, XGBoost, LGBM, and CatBoost. Integrating diverse advanced machine learning models and refining predictions through interactive clustering contributes to a nuanced understanding of customer behavior. The emphasis on model diversity, unique refinements, and performance

metrics distinguishes this research, offering a robust foundation for enhancing decision-making in the retail sector.

7.2 Recommendation

This study recommends retail practitioners leverage advanced machine learning models, especially those encompassing the refined Random Forest regressor and CatBoost regressor for sales prediction. Implementing K-Means clustering with the Elbow method for customer segmentation is also advised to enhance targeted marketing strategies. Additionally, regular model updates and continuous exploration of feature engineering techniques are essential for sustained accuracy and business success.

7.3 Future Scope

Future research may explore the integration of emerging technologies like deep learning for enhanced predictive analytics in retail. Investigating dynamic clustering methods along with real-time data streams could further refine customer segmentation. The research has inherent limitations while it contributes predictive analytics, along with immense potential for retail. Challenges may arise in assuring data quality and managing the vast amount of retail data. In addition, the implementation of complicated predictive models requires a high level of expertise and resources, posing practical challenges for some businesses. Examining the impact of external factors such as economic trends on the sales prediction model would also contribute to a more comprehensive understanding.

8 References

Ali, M. and Essien, A. (2023). How can big data analytics improve outbound logistics in the UK retail sector? A qualitative study. doi:https://doi.org/10.1108/jeim-08-2022-0282.

Alqhatani, A., Ashraf, M.S., Ferzund, J., Shaf, A., Abosaq, H.A., Rahman, S., Irfan, M. and Alqhtani, S.M. (2022). 360° Retail Business Analytics by Adopting Hybrid Machine Learning and a Business Intelligence Approach. *Sustainability*, 14(19), p.11942. doi: https://doi.org/10.3390/su141911942.

Anitha, P. and Patil, M.M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, [online] 34(5), pp.1785–1792. doi: https://doi.org/10.1016/j.jksuci.2019.12.011.

Bousdekis, A., Lepenioti, K., Apostolou, D. and Mentzas, G. (2021). A Review of Data-Driven Decision-Making Methods for Industry 4.0 Maintenance Applications. *Electronics*, [online] 10(7), p.828. doi: https://doi.org/10.3390/electronics10070828.

Bradlow, E.T., Gangwar, M., Kopalle, P. and Voleti, S. (2017). The Role of Big Data and Predictive Analytics in Retailing. *Journal of Retailing*, 93(1), pp.79–95. doi: https://doi.org/10.1016/j.jretai.2016.12.004.

Chen, Y., Li, C. and Wang, H. (2022). Big Data and Predictive Analytics for Business Intelligence: A Bibliographic Study (2000–2021). *Forecasting*, [online] 4(4), pp.767–786. doi:https://doi.org/10.3390/forecast4040042.

Ibrahim, N.F. and Wang, X. (2019). A text analytics approach for online retailing service improvement: Evidence from Twitter. *Decision Support Systems*, 121, pp.37–50. doi:https://doi.org/10.1016/j.dss.2019.03.002.

Kaggle (2023). *Retail Data*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/abdurraziq01/retail-data.
Li, R., Ghose, A., Xu, K. and Li, B. (2023). Predicting Consumer In-Store Purchase Using Real-Time Retail Video Analytics. *Social Science Research Network*. doi:https://doi.org/10.2139/ssrn.4513385.

Malik, A., Dargar, G., Sharma, A. and Pandey, P. (2023). *Predictive Analysis for Retail Shops using Machine Learning for Maximizing Revenue*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICICCS56967.2023.10142634.

Miguel Alves Gomes and Meisen, T. (2023). A review on customer segmentation methods for personalized customer targeting in e-commerce use cases. *Information Systems and e-Business Management*, [online] 21, pp.527–570. doi: https://doi.org/10.1007/s10257-023-00640-4.

Nainggolan, R., Perangin-angin, R., Simarmata, E. and Tarigan, A.F. (2019). Improved the Performance of the KMeans Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series*, 1361, p.012015. doi:https://doi.org/10.1088/1742-6596/1361/1/012015.

Nouf Fahad Al-Mufadi, Nawadher Alblihed, Alhabeeb, S., Alhumud, S. and Selmi, A. (2023). Sales Prediction Based on Data Mining Techniques. *2023 3rd International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, [online] pp.1–6. doi: https://doi.org/10.1109/esmarta59349.2023.10293312.

Nouf Fahad Al-Mufadi, Nawadher Alblihed, Alhabeeb, S., Alhumud, S. and Selmi, A. (2023). Sales Prediction Based on Data Mining Techniques. doi:https://doi.org/10.1109/esmarta59349.2023.10293312.

Panda, B. (2020). Need of Business Analytics and Prediction Modeling in Retail Marketing in Indian Context. *Asian Journal of Managerial Science*, 9(1), pp.18–24. doi:https://doi.org/10.51983/ajms-2020.9.1.1635.

Pavlyuchenko, K. and Panfilov, P. (2021). Application of Predictive Analytics to Sales Planning Business Process of FMCG Company. *Proceedings of the 13th International Conference on Management of Digital EcoSystems*. doi:https://doi.org/10.1145/3444757.3485174.

Pitka, T. and Bucko, J. (2023). Segmenting Customers with Data Analytics Tools: Understanding and Engaging Target Audiences. *Acta Informatica Pragensia*, 12(2), pp.357–378. doi: https://doi.org/10.18267/j.aip.220.

Prasad, J.Phani. and Venkatesham, T. (2021). Big Data Analytics- In Retail Sector. *International Journal of Computer Science and Mobile Computing*, 10(7), pp.34–38. doi:https://doi.org/10.47760/ijcsmc.2021.v10i07.005.

Rohaan, D., Topan, E. and Groothuis-Oudshoorn, C.G.M. (2022). Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-

sales service provider. *Expert Systems with Applications*, 188, p.115925. doi:https://doi.org/10.1016/j.eswa.2021.115925.

Rooderkerk, R.P., DeHoratius, N. and Musalem, A. (2022). The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners. *Production and Operations Management*, [online] 31(10). doi:https://doi.org/10.1111/poms.13811.

Sajawal, M., Usman, S., Alshaikh, H.S. and Hayat, A. (2023). Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques. *Retail*. [online] doi: http://dx.doi.org/10.54692/lgurjcsit.2022.06004399.

Saraswathi, K., Renukadevi, N.T., Nandhinidevi, S., Gayathridevi, S. and Naveen, P. (2021). Sales prediction using machine learning approaches. *PROCEEDINGS OF THE 4TH NATIONAL CONFERENCE ON CURRENT AND EMERGING PROCESS TECHNOLOGIES E-CONCEPT-2021*, [online] 2387(1), p.140038. doi: https://doi.org/10.1063/5.0068655.

Saurav, S. (2018). Realizing the Potential of Retail Analytics. *Advances in logistics, operations, and management science book series*. doi:https://doi.org/10.4018/978-1-5225-3056-5.ch004.

Schröer, C., Kruse, F. and Gómez, J.M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model.
Procedia Computer Science, [online] 181, pp.526–534. Available at: https://www.sciencedirect.com/science/article/pii/S1877050921002416.
Seyedan, M. and Mafakheri, F. (2020). Predictive Big Data Analytics for Supply Chain Demand forecasting: methods, applications, and Research Opportunities. *Journal of Big Data*, [online] 7(1), pp.1–22. doi: https://doi.org/10.1186/s40537-020-00329-2.

Smith, M.A. and Côté, M.J. (2022). Predictive Analytics Improves Sales Forecasts for a Pop-up Retailer. *INFORMS Journal on Applied Analytics*. doi:https://doi.org/10.1287/inte.2022.1119.

Tabianan, K., Velu, S. and Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability*, 14(12), p.7243. doi:https://doi.org/10.3390/su14127243.

Tripathi, S., Bhardwaj, A. and Poovammal, E. (2018). Approaches to Clustering in Customer Segmentation. *International Journal of Engineering & Technology*, [online] 7(3.12), p.802. doi: https://doi.org/10.14419/ijet.v7i3.12.16505.

Verma, N., Malhotra, D. and Singh, J. (2020). Big data analytics for retail industry using MapReduce-Apriori framework. *Journal of Management Analytics*, 7(3), pp.1–19. doi:https://doi.org/10.1080/23270012.2020.1728403.

Vyas, M. (2019). *Predictive Analytics Trend for Retail Industry in 2019 | Sigma Data Systems*. [online] Sigma Data Systems. Available at: https://www.sigmadatasys.com/top-predictive-analytics-trends-for-retail-industry-in-2019and-beyond/ [Accessed 31 Oct. 2019].

Wassouf, W.N., Alkhatib, R., Salloum, K. and Balloul, S. (2020). Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study. *Journal of Big Data*, [online] 7(1). doi: https://doi.org/10.1186/s40537-020-00290-0.

Zhang, B., Tseng, M.-L., Qi, L., Guo, Y. and Wang, C.-H. (2023). A comparative online sales forecasting analysis: Data mining techniques. *Computers & Industrial Engineering*, [online] 176, p.108935.
doi:https://doi.org/10.1016/j.cie.2022.10893 Tripathi, Bhardwaj and Poovammal (2018)

Arno De Caigny, K. C. a. K. W. D. B., 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. s.l.:s.n.
Breiman, L., 2001. Random Forests. s.l.:s.n.
dairu, X. & Shilong, Z., 2021. Machine Learning Model for Sales Forecasting by Using XGBoost. s.l.:s.n.
Dastan Maulud, A. M. A., 2020. A Review on Linear Regression Comprehensive in Machine Learning. s.l.:s.n.
Ding, J., Chen, Z., Xiaolong, L. & Lai, B., 2020. Sales Forecasting Based on CatBoost. s.l.:s.n.
Harsha Chamara Hewage, H. N. P., n.d. Retail Sales Forecasting in the Presence of Promotional Periods. s.l.:s.n.
Wong, J. A. H. a. M. A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. s.l.:Oxford University Press. Nicholas, K. R. S. a. C., 2020. Cluster Quality Analysis Using Silhouette Score. Sydney: s.n.