Unlocking Business Potential in FMCG with Predictive Analytics: A Machine Learning Approach

MSc Research Project

MSc in Data Analytics

Utkarsh Sharma

Student ID: x23170450

School of Computing

National College of Ireland

Supervisor: Bharat Agarwal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Utkarsh Sharma		
Student ID:	x23170450		
Programme:	MSc in Data Analytics	**Year:**	2024-25
Module:	MSc Research Project		
Supervisor:	Bharat Agarwal		
Submission Due Date:	12th December, 2024		
Project Title:	Unlocking Business Potential in FMCG with Predictive Analytics: A Machine Learning Approach		
Word Count:	8706 **Page Count :**19		

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: Utkarsh Sharma

Date: 11/12/2024

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):	

Unlocking Business Potential in FMCG with Predictive Analytics: A Machine Learning Approach

Utkarsh Sharma

x23170450

Abstract

The Fast-Moving Consumer Goods (FMCG) segment is one of the most important segments in the world economy comprising goods that can be sold quickly at rather low margin – foods & beverages, non food personal healthcare products, home care etc. Traditional models of the FMCG segment have burdened organizations with the problems of low sales forecast, customer actions examination, and stock manage. Many of these models are based upon historical sales data and last tendencies plus simple forecasting methods which do not reflect current market or consumer demands, seasonality, and variability of many aspects of supply chain. These challenges are addressed in this study by using advanced machine learning (ML) models and algorithms for decision-making process in the FMCG operations. Using the tools of predictive sales, customer behavior analysis using K-Means clustering and optimisation of an inventory through reinforcement learning, the paper gives a more comprehensive approach. In this paper, the authors highlight an integrated system for supply chain management and sales forecasting with the assistance of ML methods. The study focuses on three primary domains: forecasting of sales, understanding of the customer behaviour and the management of supply chain efficiently. In the first domain, five models of ML Linear Regression, Random Forest, Decision Tree, SVR, KNN was used to forecast the retail sales to which the Random Forest was exclusively identified as the most suitable model within the three performance indicators of RMSE as well as R². The second domain used K-Means clustering to assess customers' behavior; this broke down the customers into three groups based on their retail, transfer and warehouse sales. These segments offered valuable information when it comes to the management of stock and sales outlets. The third domain was concerned with stock-up or stock-down decisions based on seasonal sales forecasts of the organisation's products. Moreover, for the formation of an inventory control system, Reinforcement Learning was used to find the optimal between stock surplus and stockout, holding cost, and profit.

Keywords: Fast-Moving Consumer Goods (FMCG), Sales Prediction, Machine Learning (ML) Models, Random Forest, K-Means Clustering, Customer Behavior Analysis

1 Introduction

1.1 Background

The possibility of forecasting in the FMCG sector is a strong factor in using predictive analytics where the results depend on the analysis of trends and experience Juan (2024). Also referred to as using of various statistical formulas and machine learning procedures, predictive analytics helps to make appropriate decisions that may improve business operations and sales effectiveness. In the fast-moving FMCG market, low product differentiation, but high heterogeneity, coupled with fluctuating consumer preferences, it becomes a significant challenge for organizations to manage their inventory as well as their marketing and customer relationship. Purchasing data patterns, seasonal buying and other relevant external factors can be used effectively to predict buying patterns and, therefore, better manage the supply chain. This proactive management extends not just beyond creating savings towards waste and inventory turnover, but also marketing messages that hit home with the intended entrants. Therefore, predictive analytics can act as a scope for FMCG corporations aspiring for better customer satisfaction and enhanced sales performance Kukard (2023), and identify organizational effectiveness throughout the rivalry thus contributing to the enhancement of competitiveness in today’s emerging economy.

1.2 Aim of the Study

This study aims to enhance decision-making and operational efficiency in the FMCG sector by focusing on three key domains: prediction of sales, further analyzing customer behaviour by using a K-means clustering technique, and optimization of supply chain management. In the first domain of the study, we shall consider the following machine learning models; Linear regression, Random Forest, Decision tree, SVR, and KNN in its attempt to approximate sales performance. These models will be assessed and judged by their performances given by RMSE, R^2 to identify best predictor for future sales trends. The second domain is about studying the customers’ behaviour to perform the K-means clustering to segment them according to their purchasing behaviour to better understand the customers needs and to effectively target marketing campaigns. The third domain is about improving supply chain by using predictive analytics in order to predict demand and consequently eliminate unnecessary and costly supply chain activities. The findings of this study hope to make recommendations which are to be derived from the best practices of using sales prediction, customer behavior analysis and supply chain flow in the FMCG Industry should contain recommendations on how to optimize the use of scarce resources, achieve more of the set sales targets and increase efficiency in the supply chain flow of products in the fmcg industry. While the integration of the three domains – prediction of sales, customer engagement, and operation flows – with the help of more sophisticated machine learning models will lead to improvement in the accuracy of forecast and effectiveness of sales and operations planning.

1.3 Research Objectives

There are some research objectives in this study are as follows:

1. To evaluate and compare the performance of machine learning models (Linear Regression, Random Forest, Decision Tree, SVR, KNN) in predicting future sales and identifying the best-performing model using RMSE and R^2 metrics.

2. To analyze customer behaviour using K-means clustering to identify distinct segments based on purchasing patterns for targeted marketing and product offerings.
3. To leverage predictive analytics to forecast product demand, optimize inventory management, and improve the efficiency of supply chain operations in the FMCG sector.

1.4 Research Questions

There are some research objectives in this study are as follows:

1. Which machine learning model (Linear Regression, Random Forest, Decision Tree, SVR, KNN) provides the most accurate sales predictions, based on RMSE and R^2 metrics?
2. How can K-means clustering be used to segment customers based on their purchasing behaviour, and what insights can these segments provide for targeted marketing strategies?
3. How can predictive analytics improve supply chain efficiency by accurately forecasting product demand and optimizing inventory management?

2 Literature Review

2.1 Introduction to Predictive Analytics in FMCG

Predictive analytics in the Fast-Moving Consumer Goods (FMCG) using statistical tools Sazu et al. (2022) as well as other computational methodologies that consider certain aspects of large data sets and using these aspects to reach probable conclusions. Through the efficiency and effectiveness of using a vast amount of data generated from sales, customers' behavior, and the market in general, FMCG firms are put in a good position to take the appropriate decisions that are likely to yield the right outcomes in the management of stocks Wikensten (2022), marketing of their products, and giving the customers the value they seek in the appropriate time that is most convenient for them. The incorporation of predictive analytic leads to accurate forecasting of the consumer demand Punia and Shankar (2022), this ensures that organizations avoid stocking up on products as well as avoiding to have excess unsold stock, this leads to efficiency in the organizations' operations. In addition, customer segmentation and the use of predictive modeling allow for targeted promotional activities so that marketing messages are appropriate for particular targeted audiences. This capability has not only a direct impact in increasing the overall sales, but also measurably increase the level of satisfaction and consequently customer loyalty. With the ever distancing FMCG competition and skyrocketing stakes on analytics, predictive analytics is very important to organizations with the intention of helping them in using explicit and intensive data in strategic planning and also in decision making.

2.2 Predictive Analytics in FMCG Sales Optimization

This section delves into the role of predictive analytics in optimizing sales within the fast-moving consumer goods (FMCG) sector, highlighting two critical areas: sales forecasting models as well as market basket analysis. SARIMA and other machine learning methods like

LSTM are useful in refining the common sales forecasting models, and thus raising the level of inventory management to avoid high proportion of stock out conditions. Also, it reveals the pattern of purchases to target customers including implementing techniques such as cross selling and up selling. In this way, the FMCG companies will be able to utilize the preferable analytics and make more precise decisions on how to increase customers' satisfaction, and as a result sales performance as well as on how to regulate the inventory in the most effective way.

This approach is stated by the paper of Bartwal et al. (2024) for which an integrated model that integrates Seasonal Autoregressive Integrated Moving Average (SARIMA) for the non seasonal demand and regression for the seasonal demand, has been developed for the improving of forecast accuracy of the periodically seasonal products. One of the major issues which can create problem in achieving consistent extent of accuracy is multifaceted nature of causal variables, which define demand. The performance of the proposed model is compared with SARIMA, Artificial Neural Network (ANN), and SARIMAX model in terms of different error measure of forecast. Overall, the proposed integrated model yields the highest accuracy in terms of all the performance measures, generating useful managerial implications while also demonstrating suitably for a wide range of industries to ensure this approach's versatility.

The research paper presented by Regina and Jodiawan (2021) is based on a case study of a national distributor firm dealing with FMCG product that will focus on the forecasting techniques used for the PPIC division of the company which at present utilizes moving average techniques in combination with managerial judgement while not quantifying the accuracy of the forecast. Based on four time series techniques, namely ARIMA, MA, DES, and RL techniques, the study assesses the forecast accuracy of overall demand of 12 classes of SKUs. A major research question answered in this research involves, how to make data stationary for accurate ARIMA forecasting; for DES, it is done by incorporating alpha and gamma values obtained from the ARIMA method. In other words, the overall performance of the forecasting methods studied highly depends on the specific demand characteristics of each class A SKU. Therefore, the paper recommends that the company use time series forecasting procedures that have lower error rates for each of the four SKU classes currently in use in an effort to increase the general accuracy of demand forecasting.

According to a study by Godoy et al. (2022) who underscores the importance of demand forecasting in enabling organizations to meet future customer needs; many Colombian firms rely on traditional regression techniques whose outcome is often imprecise demand forecasts. However, to enhance the forecast accuracy, this study examines Stochastic Time Series Models: Seasonal Autoregressive Integrated Moving Average Multiple Linear Regression (SARIMA-MLR) in contrast to Recurrent Neural Network: Long Short-Term Memory (LSTM) networks for demand in the beauty and make-up sector particularly nail varnish. In this type of research the greatest challenge remains in how to provide better estimates to combat limitations of traditional models. The results show that the acknowledgement of the previous steps using the RNN architecture with two LSTM layers has a higher predictive capability than SARIMA-MLR method for four chosen FMCG products. This implies that an industry that embraces new methods of machine learning can be able to improve on the accuracy of demand forecasting by a large percentage.

Another research proposed by Chaowai and Chtima (2024) seeks to improve the inventory control system of FMCGs in a small trading company which is highly specialized on high volatility products caused by promotions. Analysis of the overall promotional campaigns showed that there was a highly significant lift in demand where promotional factors contributed to an increase of 60.44% in case of WM 85ML and 161.76% for SW 85ML; seen in Figure 28 below; it is clear why these factors have to be incorporated into demand forecasting. The approach consists in the identification of the existing inventory system, problem definition, solution, and the final evaluation of the results. To enhance the accuracy of the forecasts, the authors resorted to quantitative time-series techniques including Exponential Smoothing and both Additive and Multiplicative versions of Holt Winters' Method; these enhanced the forecast precision by about 90%. A major contribution of the research relates to the formulation of a new order policy known as the Periodic Review with Safety Stocks and Reorder Point which is a variant of both the fixed-order quantity and fixed-time period models. It has immediate realistic organisational implications for maintaining ideal stock holding, supporting constant promotions and is estimated to cut down stock cost by around 30% as per the fluctuating demand pattern of FMCG sector.

The study which is been given by Gijare and Kumar (2023) who has proposed a market basket analysis which intends to establish the following relationship patterns of the consumers of a lifestyle store in Mumbai. The research uses data from sales of one year and is acquired from the retail software industry; prediction of the sales of products and, co- sales of items (the probability that two or more items are bought together) is done through the use of Microsoft Associative algorithm in the Business Intelligence Development Studio application. One of the major issues that will face this exploratory research is how best to describe and explain consumer purchasing behaviours in a way that is operationally useful. The outcomes of the market basket analysis are useful to retailers as a base for orchestrating sales campaigns, as well as for configuring the strategies of upselling and cross-selling; similarly, the positioning of the products in the store layout is informed by the market basket analysis. Having the information from the analysis, business can make necessary changes to the product offering, place products that sell well, and develop special offers on the most selling categories to improve customers' satisfaction. The observations made are also instrumental in specific marketing and advertising approach, which in turn are beneficial to businesses and increased consumers' participation.

2.3 Predicting Customer Churn in FMCG

This section is specifically devoted to the methodologies and issues of customer churn forecasting in the context of the fast-moving consumer goods (FMCG) industry. It underlines the competitive non-catalogue character of the market, in which clients often 'jump' from one brand to another because the choice of similar goods is as diverse as possible. Some of the main approaches adequate with churn are the use of predictive analysis, historical data analysis, as well as machine learning algorithms which were already mentioned, including RFM analysis and several types of cluster models. The discussion will focus on the key issues such as the analysis of purchasing behavior of the customer, variations in customer demand, and the use of the relevant marketing strategies to increase customer loyalty and increase business performance.

In the study for Predicting Customer Churn in FMCG, there is a study which is given by Gunesen et al. (2021) who proposed a complete framework for analyzing churn behavior of the highest-spending customers of different FMCG brands, particularly for detergents, fabric conditioners, shampoos, and carbonated soft drinks in the Turkey market for 2018 and 2019. The plan entails exploratory data analysis and feature creation that ensures the creation of several predictive models depending with the algorithms to be used. One of the main problems that was met was the non-contract basis of the FMCG market which makes it difficult to retain customers due to a large number of brands and comparable products which means that buyers do not remain loyal to a particular brand and buy both the company's product and its competitors' ones. In this regard, the study used the RFM analysis model, and the K-Means clustering model to obtain practical data for operating a business and to develop and implement marketing strategies. Consequently, it was evident from the obtained results that the proposed churn prediction models, if applied, can help increase customers' retention and in turn, the improvement of revenue contributions as a result of carefully planned marketing and customers' retention efforts.

Mahdi and Jabbari (2024) has provided a method for anticipating customer churn in the FMCG business especially within the sales and distribution capillary food industry company. The proposed approach uses transactional data, which is collected from point of sale (POS) systems, to generate features that expose the customer purchasing behavior that is a sound starting point for churn prediction and prevention. One of the competitive threats faced was that the FMCG market is mainly non-contractual where the consumers are served with many brands, and substitutes, hence it was hard for the companies to know how to act when their customers are less committed. As a remedy for this, the research used both fundamental Machine Learning Approaches and the state of the art Deep Learning Models such as CNN and RBM to boost the model predictability. The findings illustrated the effectiveness of illustrating comparisons of the abovementioned methods for offering a better churn predictability in the FMCG consumer base, giving insights to improve customer retention.

In Predicting Customer Churn in FMCG another work was conducted by Ritesh (2024) who stated that there is need to incorporate predictive analytics in scrutinizing tendencies in the fast moving consumer goods market. As to the proposed approach it is crucial to note that much attention is paid to the application of historical data, machine learning, statistical approaches for developing valid models for predicting consumer behaviors and effective commercial promotion of delivering efficient, fast, and effective operations. One of the major issues in this regard is volatility which in certain ways exists as the rate at which consumers' tastes and preferences change and the volatility found in the market makes it difficult for firms to anticipate future demand and control supply chain. However, through attempting to identify these patterns and trends in the former data, the research demonstrates how predictive analytics can be highly beneficial from future results and enhancing digitally mediated interactions. The findings show that adopting big data predictive analytics in the FMCG industry enables not only anticipating customer attrition but also improving trading campaigns and optimizing stores' stock management, thus facilitating business strategy and operations.

3 Methodology

3.1 Libraries Imported

In this study data processing, data visualization, modeling and evaluation tasks were done using several libraries available in python. Pandas was used widely for data pre-processing and analysis; dataframes used in this work were created, cleaned and transformed in pandas library. For basic calculations and numerical data handling the software NumPy were used especially for arrays and matrices. In addition for data visualization Matplotlib and Seaborn libraries were used to create heatmaps, scatter and bar plots in order to analyze for specific trends or relations, if any, or to look for any outlying values. For the implementation of the predictive model scikit-learn was used, for the selection of features the function train_test_split was used and for using a machine learning model for predicting numeric target variable RandomForestRegressor was used. To assess the accuracy of the models mean_squared_error was used to find the RMSE while r2_score quantified the degree of 'how well' the model fitted the data. Further, warnings was imported to control unnecessary warning messages during execution in order to achieve a clear terminal output. Altogether, these libraries comprised an appropriate environment for performing end-to-end analysis of the data, from data preparation to the model evaluation phase, thus providing an opportunity for effective and accurate completion of the study.

3.2 Dataset Description

The dataset used for these analysis consists of sales and transfer data from numerous products from various suppliers through the years 2017, 2018, and 2019 and 2020. These are distributed across a total of 307,645 rows and 9 columns; year, month of sales, name of the supplier, item code, the description of the item, the type of the item, retail sale, retail transfer, and warehousing sale. The YEAR and MONTH tell us when the record was made; the rest of the columns are related to the product: the code, description, and type, and quantities of retail sale, transfer to retail, and to the warehouse. The ITEM TYPE column consists of options like 'WINE' 'BEER 'LIQUOR', etc; these options concern the type of product. Quantity sold refers to the number of units that was sold in the retail sales or those offered at the warehouse sales. A few values are blank in the SUPPLIER and ITEM TYPE scores while the include positive and negative values for SALES. The objective here is to scrub, modify, and transform this dataset into a form that is fit for analysis, trend determination, and modelling.

3.3 Data Cleaning

Data cleaning is very important step in order to remove or eliminate any unwanted and or un-serviceable data in the data set. From the data set, the first process of cleaning-training was conducted by inspecting for null or missing values. There were 167 missing values in the SUPPLIER column and 1 missing value in the ITEM TYPE column and 3 in the RETAIL SALES column. In this respect, for rows where SUPPLIER values were absent or blank, data imputation based on domain knowledge was made where possible; otherwise, such rows of little relevance in the analysis were totally dropped. In other cases, analogously with other missing values within the columns, blank entries were replaced either with default values, or with those employed by data imputation techniques. To eliminate duplicate data, the records were copied into a temporary table and the type 'drop_duplicates hereby the first of the duplicated records, were applied'. Negative values identified in the RETAIL SALES and WAREHOUSE SALES columns were explored to find out their root cause and the outliers in such entries were marked for review. After cleaning each post and removing any duplicates or missing values from the cleaned data, data is ready for further analysis.

3.4 Data Preprocessing and Preparation

Data preprocessing and preparation are primarily focused on the modification of the data obtained in its raw data. This step mainly covers missing values, encoding categorical variables and normalizing numerical values. First, the missing values were dealt with depending on the approach that was followed in the cleaning section. Due to the string nature of the SUPPLIER column, it was necessary to handle the missing values in this column correctly with similar supplier values or by eliminating rows with significant missing data. Since the ITEM TYPE is categorical, it was then transformed into label encoded or one hot encoded based on the need of the model. For other clearer numeric columns like RETAIL SALES, RETAIL TRANSFERS, and WAREHOUSE SALES, outliers were reviewed and when necessary, these values were either capped or their scales were made robust. Categorical variables were retained in their original format for analysis and YEAR, MONTH for time series and trends analysis. The data was therefore divided into training and testing sets with the objective of preparing the data to find sales patterns, patterns of a specific supplier, and inventory flow patterns.

3.5 Feature Engineering

Feature engineering is the conversion of new features or altering existing ones with a view to making the model better. Alternatively, given the data nature where sales and transfers are highlighted newly, seasons, moving averages and year on year sales were derived. For instance, new variable added to enable the analysis of fluctuation in sales over time was the month-to-month percentage change in sales (RETAIL SALES). Further, interactive features between ITEM TYPE and SUPPLIER were introduced to address the specific trends by supplier for each type of item. Other categorical features included the date, where one record is from a peak sales period (e.g., December compared to other months in the year) and other is from a slow period. These features are added with the hope of giving context to the analysis and enhance the model’s capability to identify trends, variations and patterns of the sales data.

3.6 Exploratory Data Analysis (EDA)

The figure 1 is histogram which illustrates the Distribution of Retail Sales. On the x axis is the scale of Retail Sales and on the y axis is the scale depicting the Frequency or how often each of the Retail Sales value occurred. We can see that the majority of values representing Retail Sales are distributed within the range of 1000 & 1250, with the most density at approximately 1100. It is also seen from the above graph that there are a few cases to specific points, for example, 0 Retail Sales or outliers. In summary, it is possible to consider this histogram as the graphic illustration of the extent and frequency of the Retail Sales, as presented by the AAE, so that the option for better understanding of the dispersion can be provided.

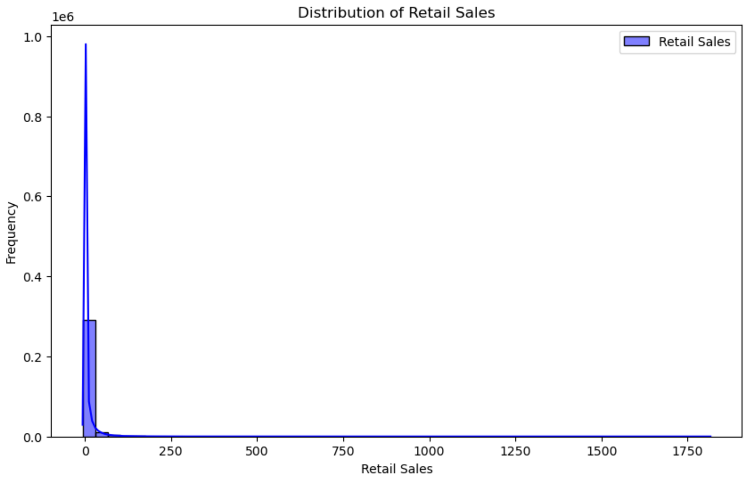


Figure 1: Distribution of Retail Sales

The histogram in figure 2 illustrates the Distribution of Warehouse Sales. The horizontal axis contains the Warehouse Sales values and the vertical axis portray the Frequency or the number of times a particular value of Warehouse Sales occurred. Based on the above graph it can be seen that most of Warehouse Sales values are likely to occur at 0 to 5000 with the most likely figure being 4500. The graph also shows that there are large fluctuations or some values are much higher or lower than the others as indicated by the vertical area of the graph with dense marks at around negative or – 5000 Warehouse Sales. In a general sense, this histogram is useful for presenting the sales data and gives the analyst an idea of the distribution of Warehouse Sales – where they are most and least common.

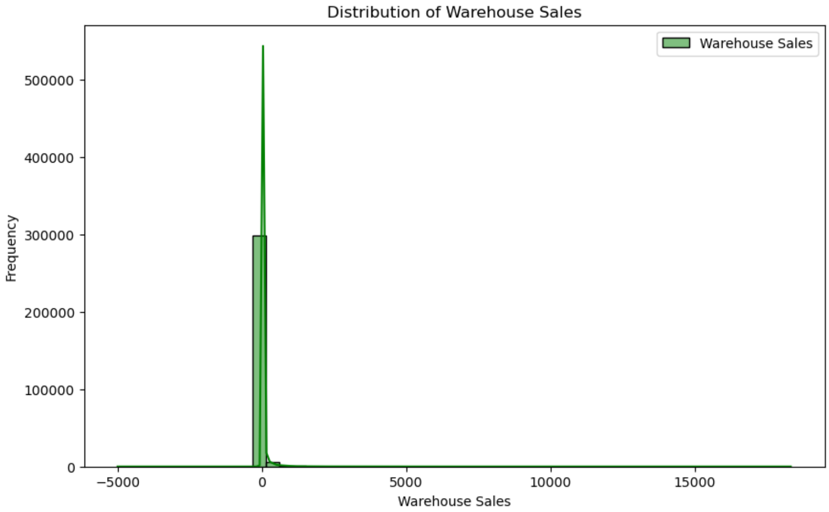


Figure 2: Distribution of Warehouse Sales

This figure 3 is presenting a bar graph in which multimedia and text annotated are given using Item types distribution. The items types are written in horizontal axis like WINE, LIQUOR, BEER, KEGS, NON-ALCOHOL, STR_SUPPLIES, REF and DUNNAGE. The vertical scale is labels as count or frequency of each item type. It also breaks the data into relative distributions of the various types of items as evident from the displayed graph above. The top bar represents WINE to depict that is belongs to the highest count or frequency among the displayed item types. The rest of the item types are shown in a descending order with LIQUOR increasing to BEER and then to KEGS. Non-ALCOHOL, STR_SUPPLIES, REF and DUNNAGE are the other categories of the item types and they exhibit lower counts than the larger categories. This bar graph as a type of visualization helps to compare and select the top item types present in the set as well as allowing for an analysis of the general distribution and composition of the item types.

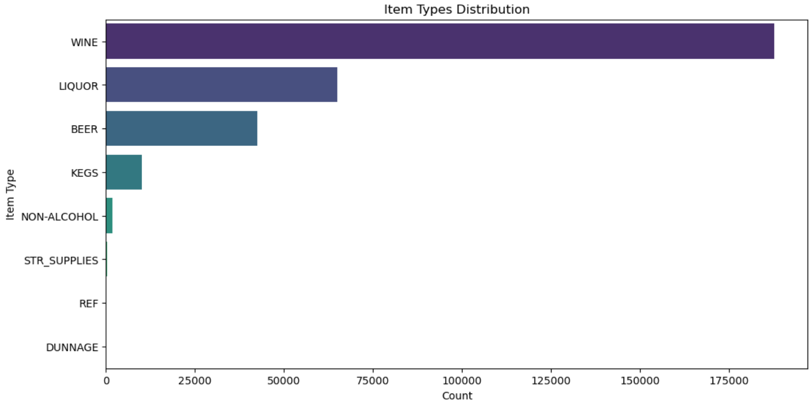


Figure 3: Item Types Distribution

Figure 4 is the bar chart of Zero Sales Ratio for Each Item Type for Retail sales. On the horizontal axis, all the item types have been shown and on the vertical axis the Zero Sales Ratio has been depicted representing the occasions when a specific type of item failed to make the sales. The above graph shows that the highest value of Zero Sales Ratio is of DUNNAGE, meaning that majority of items under DUNNAGE had no sales at all. The next highest Zero Sales Ratio is that of KEGS; followed by LIQUOR then NON-ALCOHOL. Among the item types, BEER, REF, STR_SUPPLIES, and WINE which have comparatively lower Zero Sales Ratios it can be implied that the rates of zero sales in these item types were comparatively lower. .

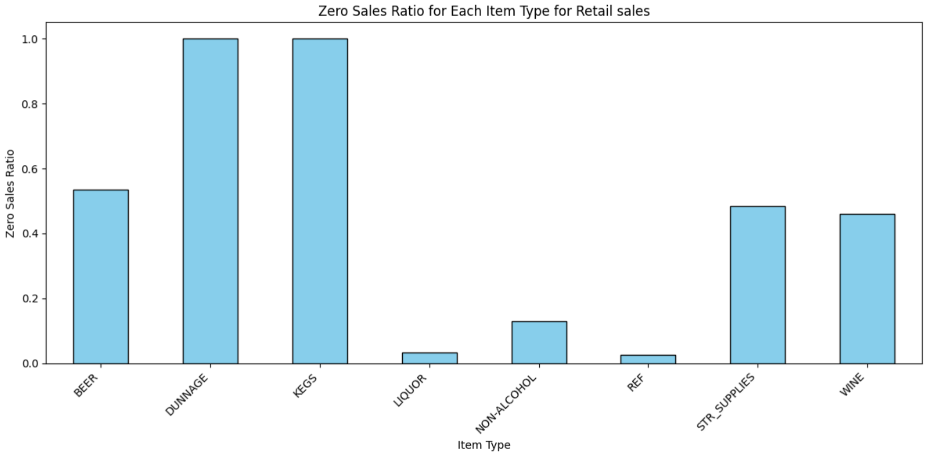


Figure 4: Zero Sales Ratio for Each Item Type for Retail sales

This Figure 5 is a bar graph that show the Zero Sales Ratio of Each Item Type for warehouse sales. Along the horizontal axis a set of different item types is provided while the vertical axis displays the Zero Sales Ratio, which illustrative the proportion of zero sales for corresponding item type. From this graph, it will be observed that the item type with the highest Zero Sales Ratio is DUNNAGE, meaning that a great number of items in DUNNAGE did not make any sale. The next highest Zero Sales Ratio is that of KEGS brand which is lower than the national average followed by LIQUOR and NON-ALCOHOL brands respectively. The fact that the Zero Sales Ratios of some item types are lower is evident, and from the results, the BEER, REF, STR_SUPPLIES, and WINE item types stand out. This visualization also facilitates the direct comparison of the Zero Sales Ratios of the various item types to better understand the sales characteristics and evaluate the performance of each group of items.

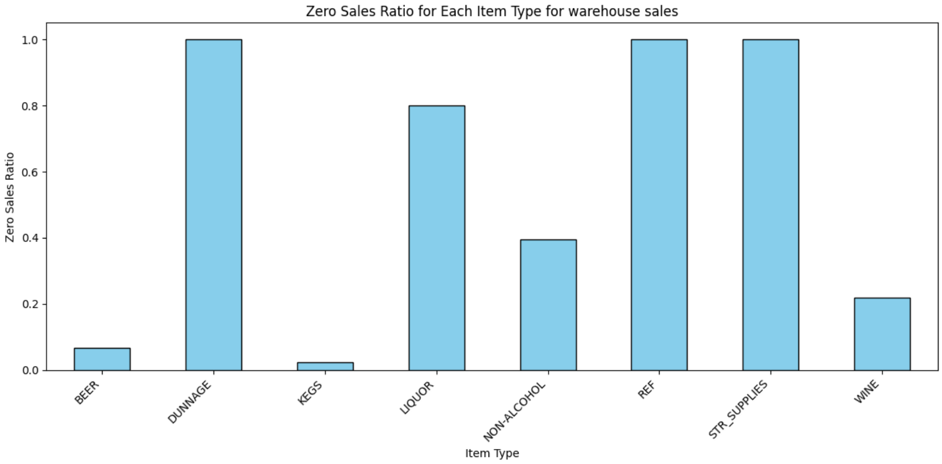


Figure 5: Zero Sales Ratio for Each Item Type for warehouse sales

The Figure 6 is a line graph showing the Monthly Retail Sales Trend in a scale of 12 months. On the horizontal direction, we have the month while on the vertical direction we have the Total Retail Sales. The line graph presents variation in retail sales throughout the year. This is reflected in sales figures as the firm begins the year with relatively small sales which gradually ramps up to mid-year before rising steeply up to the 7th month. After the peak, the graph features a gradual falling off the sales levels up to the later months of the year. That is because the nature of the line and shifts in sales volume from one month to another will determine the significant business periods of high and low sales and this information is relevant when making decisions on business strategies, inventory updates and marketing.

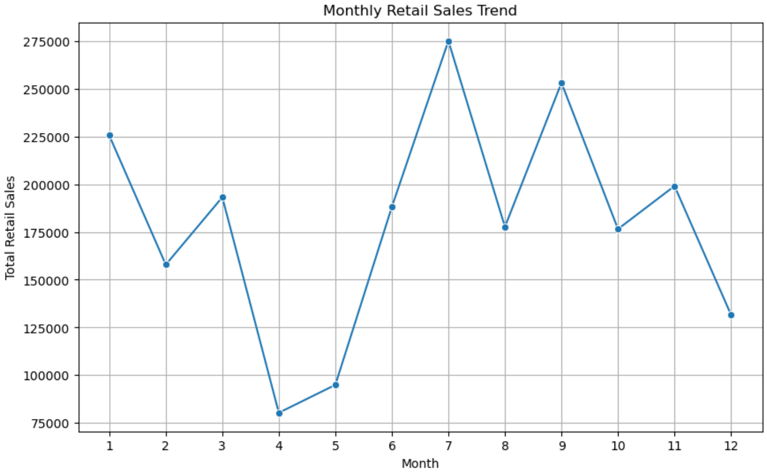


Figure 6: Monthly Retail Sales Trend

The Figure 7 presented is a bar chart on the Seasonal Retail Sales Trend. On the x-axis, the values are presented as different seasons: Winter, Spring, Summer, and Autumn On the y-axis there are Total Retail Sales. The graph shows the obvious fluctuations in the retail sales according to the season. Sales made during winter period is the highest and the least is made during the summer season. Annual retail sales by season and month The difference between winter and spring retail sales is quite large when compared to the other seasons With winter having much higher sales than the Spring. The following bar graph works well in depicting the sales trend in the retail business throughout the different seasons which can be useful to companies in order to come up with their sales and supply projections, and hiring in their companies. Variation in sales indicates that the retailers require anticipation of such disparity and devise strategies on how to maximize on the periods of high demand while at the same time minimizing on the low demands period.

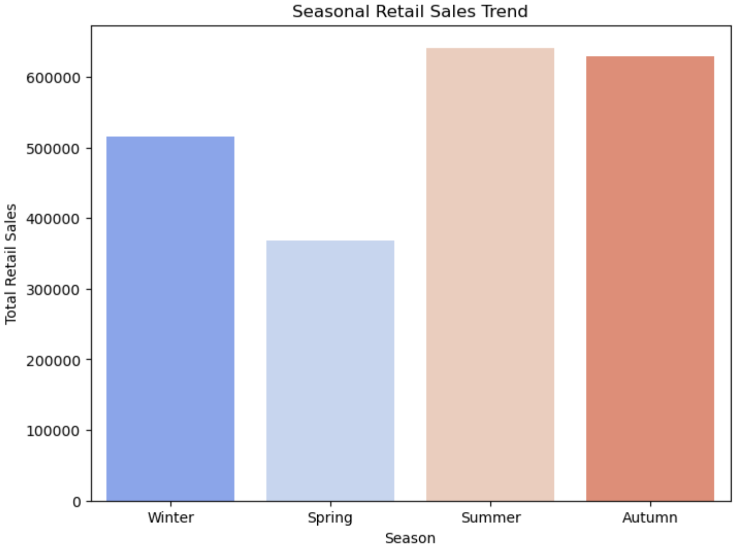


Figure 7: Seasonal Retail Sales Trend

The Figure 8 given is a histogram that describes the profit distribution of the organization. The horizontal axis embodies profitability values obtained from subtracting Warehouse Sales from Retail Sales. As presented on the chart, the vertical axis represents the count or frequency of each of the profitability values. The data plotted on the graph above is positively skewed where the higher frequency of profitability values is oriented towards the negative side of the scale between -15 000 and -5000. At ~$0 the histogram exhibits a spike; there are numerous items where the sales price in retail stores is nearly equal to warehouse costs and these could be break even or low margin SKUs. Extending towards the right, the graph also reveals that a few of these enterprises possess slightly higher profitability values, extending forward to around 5,000. These are the better or higher margin products in the product offer or portfolio.

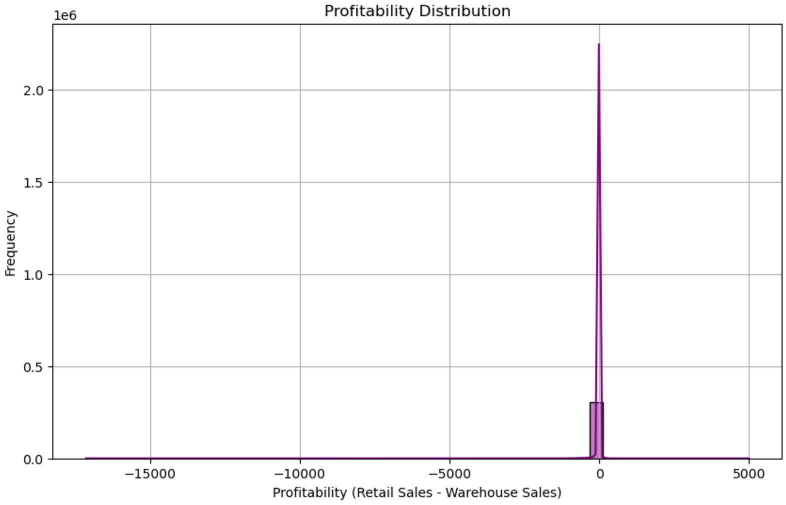


Figure 8: Profitability Distribution

The Figure 9 given above is the heatmap that shows the Retail Sales based month and type of item. Where, the horizontal axis defines the months and the vertical axis recounts the various categories of items. The data for retail sales values for each item type and specific months is also incorporated in the heatmap, with cells colored according to the scale on the right side of the figure. It entails a heatmap that gives detailed information of the performance of retail sales by various products for the year. The progression in different tones of the same hue from dark to light is used to depict differences in value wherein dark represents large values, while light shade represents the small values as detected from the sales data. The information displayed in this heatmap makes it easy to notice and solve the sales issues. For instance, it is very clear that WINE, LIQUOR, and BEER have had the highest retail sales and there exists seasonal differences among the months it operates. It also shows that the DUNNAGE and KEGS item types have had either low or zero sales weekly throughout the entire year.

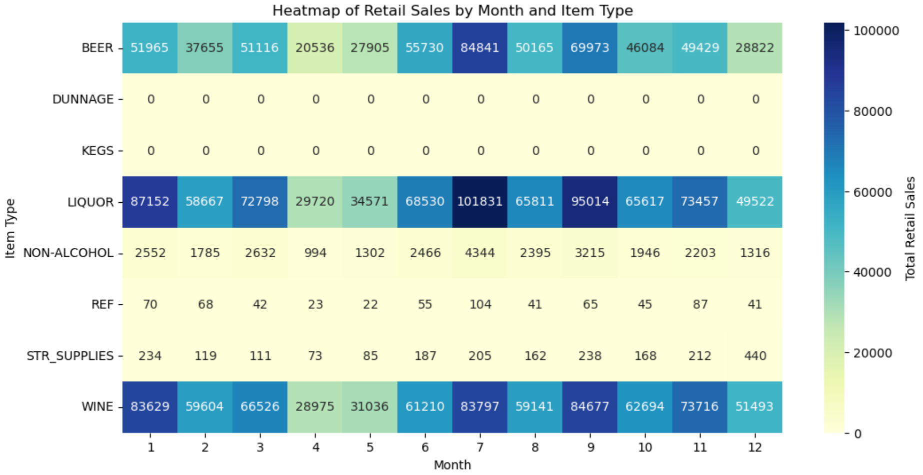


Figure 9: Heatmap of Retail Sales by Month and Item Type

Figure 10 shows a scatter plot focusing on the Retail Sales and the Profitability by Item Type. The horizontal axis would be the Retail Sales, the vertical axis would be the Profitability which can be obtained from the formula Retail Sales- Warehouse Sales. The scatter plot shows graphically a set of data points in which each point represents an item type. As seen from the figure above, the different item types are represented by different colored shapes making comparison easier. Hence, from the presented graph different patterns and correlations regarding the relations between retail sells and profitability are observed. It proves that the higher retail sales are, the higher is the probability to have higher profitability, the item types WINE, BEER and LIQUOR are located in the highest right corner proving high both sales and profitability. On the other hand, the DUNNAGE and REF this item types have such low retail selling and profitability levels.

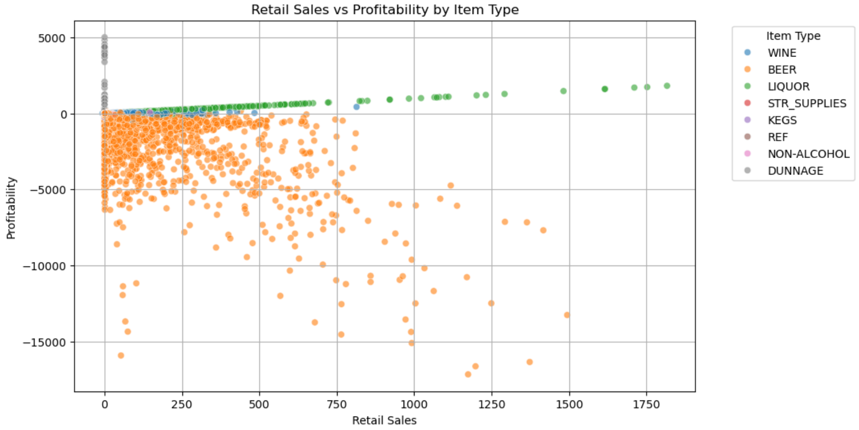


Figure 10: Retail Sales vs Profitability by Item Type

4 Design Specification

The figure 11 below presents a suggested work-flow chart of a typical predictive analytics project in the FMCG (Fast-Moving Consumer Goods) business with key areas of improvement on the sales, customer behavior and the supply chain. The process starts with a "Dataset" and goes through several steps: This includes “Data Cleaning,” “Data Preprocessing,” “Feature Engineering,” last but not the least “Data Splitting.” These initial stages entails the process of data pre-processing or transforming and shaping it in a way which will be suitable of modelling. The next step is called “Model evaluation” where the performance of different machine learning models is compared, the models included are Linear Regression, Random Forest, Decision Tree, SVR, KNN. These models are then trained in the "Model Training" phase, which aims to create three domains: “A Prediction of sales,” “Analyzing customer behaviour,” and “Supply Chain optimization”. The “Prediction of sales” domain is used after training of models has been done and the trained models are used to predict future sales so that the FMCG company can make better decisions. Domain “Analyzing customer behaviour” is aimed at studying customers’ behaviour patterns that can be useful for future marketing disappearances and products development. Last but not least, the “Optimization of Supply Chain” domain utilizes the findings of the other domains in order to enhance supply chain responsiveness and make products available to customers at particular points of time. The following is a comprehensive diagram which gives an over view of structured work flow on how to use predictive analytics in fmcg industry. The presented framework focuses on data cleaning, feature creation, model building, and domain-specific investigation, which should offer an end-to-end solution for improving the sales of a given product, the behavior of customers combined with supply chain management, thereby increasing the efficiency and efficacy of an FMCG organization.

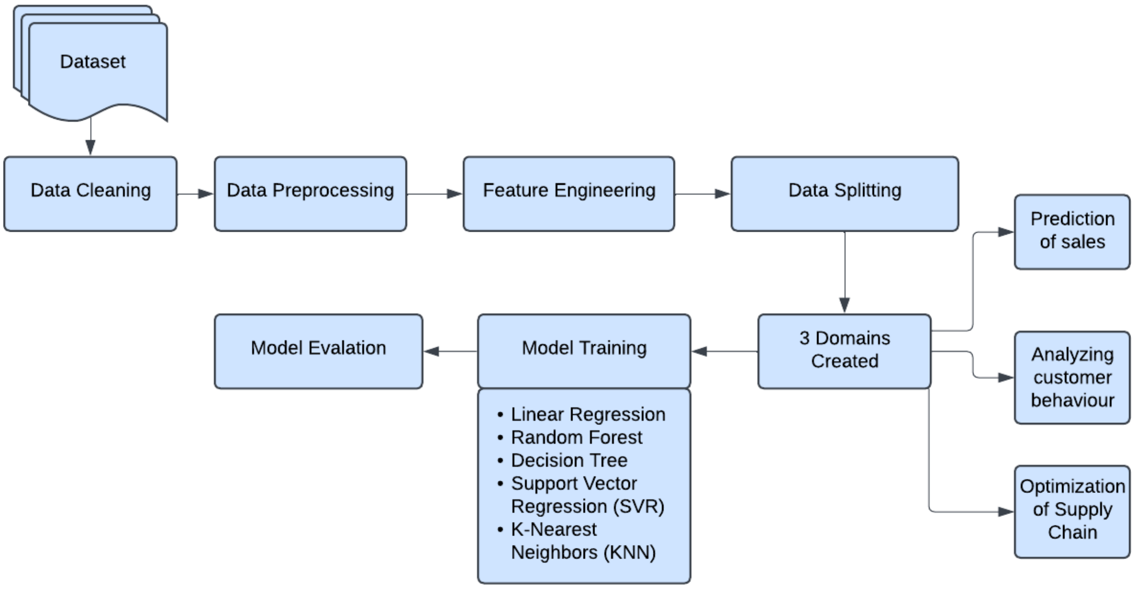


Figure 11: Proposed Workflow Diagram

5 Implementation

5.1 Implementation of Sales Prediction Models

5.1.1 Implementation for Linear Regression

Linear Regression was used as the baseline model because of its simplicity and easily understandable parameters. It operates with the identification of the line that in some way reduced the sum of squares of the prediction errors or residuals. With the training data the model was estimated rigs to the training set in order to discover the characteristics of relation between the independent variables and the target variable which is the retail sales. The model fit on the test set was then assessed with RMSE and R^2 to set a basis for comparing with other complex models.

5.1.2 Implementation for Random Forest

Random forest was used to enhance the prediction accuracy because it is an ensemble learning method. It develops several decision trees during training and integrates decision of all the developed trees during prediction, in order to avoid overemphasized training effect. The model was trained using RandomForestRegressor in scikit-learn environment with default parameters with random_state set given to ensure reproducibility. Consequently, the prediction was made on test set, where performance was evaluated by RMSE and R^2 Coefficients, which makes the model brings valuable insights for non-lineal data.

5.1.3 Implementation for Decision Tree

In order to determine whether the model is able to split data depending on feature thresholds, Decision Tree Regressor, a non-parametric supervised learning approach, was applied. The model used here develops a tree structure in which each node represents a decision rule for splitting the data. The trained model was used in predicting retail sales using the given dataset and then tested using RMSE and R^2. It was however tested for vulnerability to overfitting against other ensemble methods but is simple and easy to interpret.

5.1.4 Implementation for Support Vector Regression (SVR)

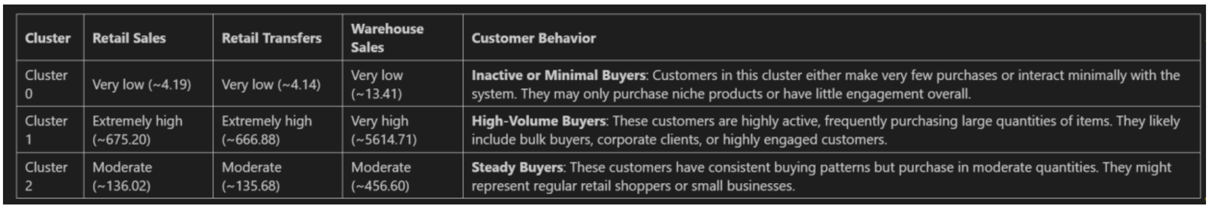
SVR was used to demonstrate its potential to map data into higher dimensions by kernel functions and to find a hyperplane whence minimum prediction risk could be obtained. A scikit-learn SVR with a radial basis function kernel was used to fit the data. Additionally, for SVR, neither the kernel nor the regularization parameter were implemented from hyperparameters; they were capitalized on their default settings. The model performance was evaluated on the test set using RMSE and R² and high performance on small and non-linear data was demonstrated.

5.1.5 Implementation for K-Nearest Neighbors (KNN)

In this study, cross validation and KNN was used in order to assess the closeness of the data points and predict by averaging the k-closest training points. This model based on KNeighborsRegressor and was trained on the current dataset; accuracy was further calculated by RMSE and R². The nonlinearity of the data was evaluated on KNN's, but its vulnerability to k-choice and the computational complexity in large data sets was highlighted.

5.2 Analysing Customer Behaviour by K means Clustering

In this section, the customer behaviour was offered with the help of K-Means clustering technique for the analysis of the data, which is one of the most actively used unsupervised machine learning techniques that can be applied for the segmentation of data based on similarities. The analysis was conducted using three key features: Retail sales, retail transfers and warehouse sales, which relate to customer purchasing and the flow of transactions in the retailers' stream. First, for data preparation, missing value cases were managed by replacing the empty set by zeros with the intention of maintaining the data quality. In order to convert all the features on to a single scale so that no feature dominates over the other one in clustering, the dataset was scaled using StandardScaler of sklearn Preprocessing library. And then, through the elbow method, the number of clusters influencing a algorithm was also identified. This method computes the inertia (within-cluster sum of squares) based on the number of clusters which is in turn selected when the rate of decrease of the inertia is smaller than before, forming an elbow point. According to this method, bigger number of clusters equal to 3 were revealed to be the most suitable for this dataset. Next, K-Mean clustering was done with the number of clusters are 3 and the clusters were inserted in another form known as Customer_Cluster in the data set as shown below in Figure 12. Purchases customers were split into three groups which created clusters for segmentation. This segmentation gave some customer insight and separated customers into those buying a lot (cluster 1), those buying frequently (cluster 2), or those in a category of low to no buying (cluster 0). Thus, based on these clusters, specific measures for marketing, inventory and sales promotion could be tailored. For instance, the early or large consumers or customers may be offered volume discounts or bonuses and other loyal consumers accordingly. Such insights make it easier for businesses to adjust their strategies, helping customers be happier, and increasing efficiency of the supply chain.



Cluster	Retail Sales	Retail Transfers	Warehouse Sales	Customer Behavior
Cluster 0	Very low (~4.19)	Very low (~4.14)	Very low (~13.41)	**Inactive or Minimal Buyers**: Customers in this cluster either make very few purchases or interact minimally with the system. They may only purchase niche products or have little engagement overall.
Cluster 1	Extremely high (~675.20)	Extremely high (~666.88)	Very high (~5614.71)	**High-Volume Buyers**: These customers are highly active, frequently purchasing large quantities of items. They likely include bulk buyers, corporate clients, or highly engaged customers.
Cluster 2	Moderate (~136.02)	Moderate (~135.68)	Moderate (~456.60)	**Steady Buyers**: These customers have consistent buying patterns but purchase in moderate quantities. They might represent regular retail shoppers or small businesses.

Figure 12: Cluster Interpretation for Customers

5.3 Optimization of Supply Chain

For optimization of supply chain, following strategies were incorporated and 3 strategies of reinforcement learning combined, which included predicted sales, customer buying patterns. or the proportion of one type or another of product), the first strategic concept was directed to the enablement of optimal distribution depending on the clients. Customers can be divided into specific groups based on the predicted sales and clustering results of the customer behavior analysis; thus, the supply chain suits different types of customers. For example, major customers (Cluster 1) were characterized as the segments that usually make many purchases. In order to meet their requirements, the supply chain can be so designed that there is sufficient stock, lower prices which are in exchange with larger quantities and consumer incentives that will increase sales. Instead, steady buyers (Cluster 2) should be provided with consistent stock, however not flooding the market. There were various forms that could be used to enhance sales with this segment including cross selling or upselling. For Inactive customers or Cluster 0, additional interactions are necessary through campaign means by offering to sell low-frequency and high-frequency products together.

In addition, the demand fluctuations varying according to seasons were also modeled in the optimization strategy. Seasonal forecasts and their influence on various clusters could be managed to distribute stock proactively between warehouses and stores, so that the requirements for particular goods would be met without the danger of stock shortening or, vice versa, stock overwhelming. Also, the environmental concern was added through the consideration of the recycles material like Pepsi glass bottles which affect the organizations' revenue from customer to the recycling firms. Another supply chain solution that was undertaken was the incorporation of seasonality into the forecast systems. By making sales data cumulative by season and by item, the 75 percentile of sales allowed for rational evaluation of when to buy more or less of particular items. Finally, techniques named reinforcement learning (RL) were used for information and inventory management. The RL agent plays the role of an Order Manager who operates on the stochastic environment of the inventory system determining appropriate moments and quantities for restocking so as to maximize the profits on stocks while keeping in check with the holding costs and penalties for stockout situations. This not only manages stocks well, but it also leads to improvement in the overall functionality of supply flows.

6 Evaluation

6.1 Case Study 1: Sales Prediction Models

In this case, the different sorts of predictive models were compared to assess their effectiveness in engaging in sales forecasts, by focusing on the accuracy and effectiveness of each approach. The goal was to understand which of these models could be used for the forecast of next sales periods efficiently using the history. Five different models were tested: Linear Regression and Random Forest, Decision Tree, SVR, and KNN. The evaluation was carried out using two key performance metrics: RMSE and R2.

The results suggest that the Random Forest model achieved the highest level of accuracy in general with least RMSE of 5.74 and highest R^2 of 0.96 as well. The Linear Regression model was also significantly effective with the RMSE of 5.76 and the R^2 of 0.96 which also make it reasonable to consider it for making sales predictions in place of Random Forest. Next is the Decision Tree model, which also gives fairly good results compared to actual values, RMSE of 7.67 and R² of 0.93. On the other hand, using the Support Vector Regression (SVR) model yielded a relatively high RMSE of 18.53, and a low model reliability of 0.62. The K-Nearest Neighbors (KNN) model seemed to perform very closely to Decision Tree with a slightly higher RMSE of 7.83 and R² of 0.93.

Table 1: Performance Evaluation of Sales Prediction Models (RMSE and R²)

Model	RMSE	R²
Linear Regression	5.7593	0.9632
Random Forest	5.7441	0.9634
Decision Tree	7.6727	0.9347
Support Vector Regression	18.5278	0.6195
K-Nearest Neighbors	7.8328	0.9320

Figure 12 presents a comparison of different machine learning models based on two performance metrics: In this case we will be using RMSE and R2 Score. The first graph at the left presents the RMSE's for all the models with the models that have values closer to the lower axis on the right presenting the better performances. The right-hand graph represents R2 scores, which are the statistics between 0 and 1 where the closer to 1 means the improved goodness of fit of the model predictions to the actual values. The models being compared cover: the basic K-Nearest Neighbors (KNN) model, the Support Vector Regression model, the Decision Trees model, the Random Forest model and the Linear Regression model. The Random Forest model has the highest R2 score of out all the modelled here.

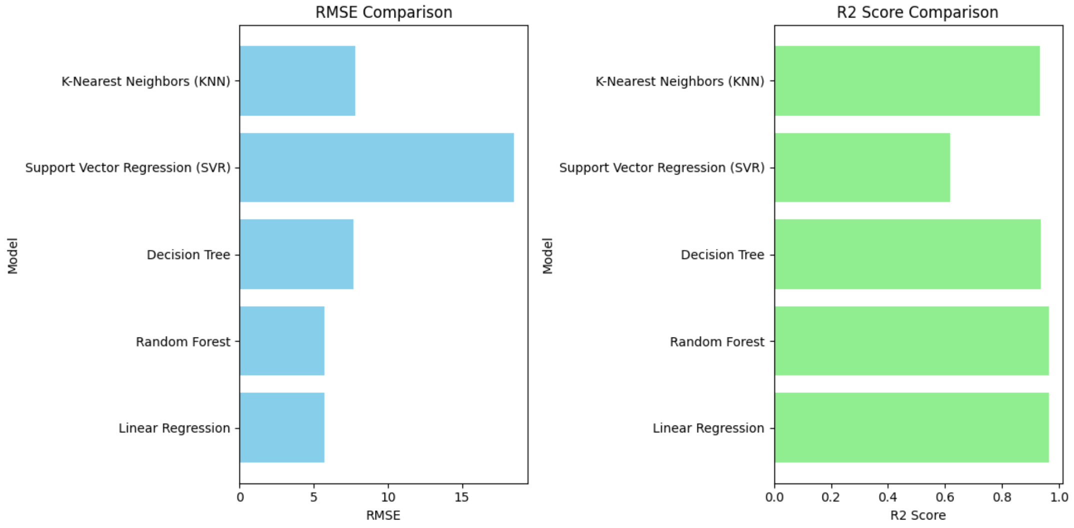


Figure 12: RMSE and R2 Comparison

6.2 Case Study 2: Customer Behaviour by K means Clustering

The plot shown in this figure is used to determine the best number of clusters using the Elbow technique that measures inertia, the amount of variation within clusters. The horizontal axis is the number of clusters, the vertical axis is the inertia of the corresponding number of clusters. In the graph, the inertia value rises when the number of clusters is still 2, meaning much variation occurs within the cluster. With increase in the number of clusters, there is steep decrease in inertia which shows that there is decrease in within cluster variance. However, at a certain point, the rate of decrease in inertia is decreasing and, therefore, its form takes an 'elbow' form. This elbow point usually indicates the number of clusters as any incremental number of clusters past this point does not enhance the clustering quality. The elbow is estimated to be at about 4 or 5 clusters and therefore it could be perhaps optimal for this particular dataset. This information can be useful to data analyst and researchers who may be in a position to decide the number of clusters to use in your analysis or in your algorithms.

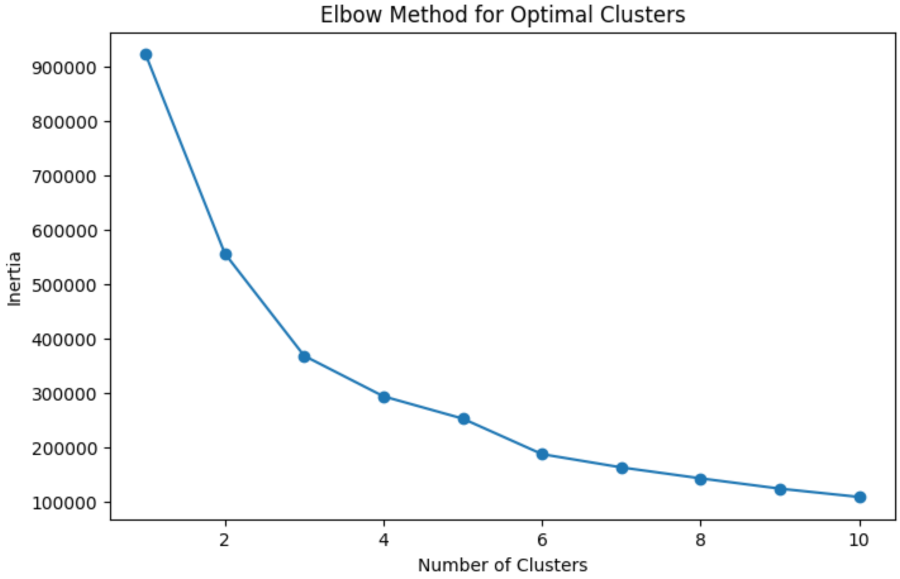


Figure 13: Elbow Method for Optimal Clusters

Table 2: Centroids of Customer Clusters Based on Retail Sales, Retail Transfers, and Warehouse Sales

RETAIL SALES	RETAIL TRANSFERS	WAREHOUSE SALES	Cluster
4.192908	4.138506	13.412819	Cluster 0
675.197359	666.876883	5614.711169	Cluster 1
136.020922	135.689039	456.603903	Cluster 2

6.3 Case Study 3: Optimization of Supply Chain

The following table shows seasonal sales for different ranges of products with their respective stock choices. The sales figures point how awareness of seasonality for inventory planning efficiency is of paramount cruciality. Specific examples of brands and en_GB Article Title: "TITO'S HANDMADE VODKA" and "CORONA EXTRA LOOSE NR" have increased overall retail sales in the autumn period and therefore, the message "Stock Up" was provided. This is so because when elaborating on the sales patterns of both varieties across the calendar year, it is clear that, for instance, the 'ZUKUA TERRE SICILIANE MOSCATO' and 'MARCHESI DI BAROLO BAROLO,' sell very little, or even negatively in the winter season; it is recommended to 'Stock Down.' The seasonal sales data shows the importance of effective inventory management in line with sales data to avoid situations where certain products run out of stock or when certain products take too long to move off the shelves. To make these decisions even more effective, using the 75th percentile which helps isolate those products with the highest sale in a particular season from the remaining items. From understanding those products that are frequently in high demand, firms are in a position to stock products that these are available in large quantities within the periods of high demand while at the same time avoiding instances whereby they order for products that may take so long to sell in the market. Making accurate provision and proper scheduling of resources in each fiscal year can be facilitated from this process, relative supply chain management can be enhanced, and thereby overall profitability can be increased.

Table 3: Seasonal Sales and Stock Decisions

SEASON	ITEM DESCRIPTION	RETAIL SALES	Stock Decision
Autumn	TITO'S HANDMADE VODKA - 1.75L	7875.56	Stock Up
Autumn	CORONA EXTRA LOOSE NR - 12OZ	6717.00	Stock Up
Autumn	HEINEKEN LOOSE NR - 12OZ	5283.00	Stock Up
Autumn	MILLER LITE 30PK CAN - 12OZ	4278.00	Stock Up
Autumn	BUD LIGHT 30PK CAN	3843.00	Stock Up
...	...	...	...
Winter	ZUKUA TERRE SICILIANE MOSCATO 750	0.00	Stock Down
Winter	MARCHESI DI BAROLO BAROLO 31343 - 750ML	-0.33	Stock Down
Winter	LA TERRE CHARD - 750ML	-0.43	Stock Down
Winter	BON APPETIT MUSCAT - 750ML	-0.58	Stock Down
Winter	SANCHO BARON - 750ML	-0.67	Stock Down

7 Conclusion and Future Works

7.1 Conclusion

This research work provides a holistic view on how to achieve the best supply chain through supply chain big data analytics mainly with the help of data mining algorithms. Using sales prediction models, clustering of customers, and seasonal sales forecasting, the study offers prescriptive suggestion to improve stock control and decision-making processes. Customer behaviour has been clustered using KMeans and different buyer segments have been identified to improve business practices, for instance, extending volume discounts for high volume buyers or reactivating customers who have not purchased from the business as often. In addition, an aspect of the seasonal sales analysis was stock control decisions related to consumer demand patterns; where the control for the 75 th percentile was underscored as a means of guiding businesses to maximize their procurements of high demand items during the various seasons. RL integration with the system is taken up to another level to learn ways of managing inventory to reduce holding costs and to avoid shortages. Altogether, the results suggest that the sophisticated analysis that allows for prediction of supply chain key parameters carries a great potential of bringing up to a new level of efficiency and profitability within the supply chain.

7.2 Limitations

However, there are limitations to this study, which are present as follows. First, the analysis largely depends on past sales data, which may not be adjusted for important changes in the market and specific disturbances, including economic crises or global COVID-like interruptions of supplies. Further, the specificity of the study is based mainly on certain aspects of customer behavior attributes and sales data, and thus lacks the complete picture of the consumer choice structure. As much as using KMeans clustering is useful it does so under the premise that the customer segments are well defined and consistent which can rarely be true in constantly evolving markets. Last but not the least, the performance of the RL model critically depends on the realism of the simulation environment which might not capture all the stochasticity in inventory management.

7.3 Implications

The results of this research have significant implications to any organisation desiring to improve its supply chain. Through the use of machine learning solutions such as sales prediction, customer segmentation and reinforcement learning, companies can act more efficiently based on the insights derived from the result of the model. Examining the importance of matching inventory density with customers' usage and purchase cycles in striking an optimal balance between over stocking and stock out. Moreover, it will also help businesses to offer better customer engagement strategies because of behavioral differences and therefore increase customer loyalty and sales.

7.4 Future Work

More research could be done along several lines in future work in this area. Thus, one possible further development is the addition of more data sources to the existing model, including up-to- date information on market fluctuations, weather or social media activity, to improve the sales and inventory forecasting. Further, considering the other types of clustering like the hierarchical clustering or using deep learning to develop better clustering algorithms may useful for considering more rich information of customer behavior. They also highlighted some directions for future work including using the reinforcement learning in multi-agent environment to represent different departments or stakeholders. Finally, extending the study to other industries and adding it with other related supply chain processes such as supply ordering and supply delivery would provide a more comprehensive solution to the overall supply chain management.

References

1. Feliu Juan, M., 2024. *Enhancing demand forecasting: an analysis of factors impacting sales and implementation of improved methodologies for accurate prediction* (Bachelor's thesis, Universitat Politècnica de Catalunya).

2. Kukard, W.A., 2023. Transforming an Industry: Leveraging Emerging Technologies in the Smart Home Environment to Create Value for the FMCG Industry.
3. Sazu, M.H. and Jahan, S.A., 2022. The impact of big data analytics on supply chain management practices in fast moving consumer goods industry: evidence from developing countries. *International Journal of Business Reflections*, *3*(1).
4. Wikensten, A., 2022. Managing Product Allocation in a Scarce and Uncertain Supply Chain: Activities in a FMCG Company.
5. Punia, S. and Shankar, S., 2022. Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowledge-Based Systems*, *258*, p.109956.
6. Bartwal, D., Sindhwani, R. and Vaidya, O.S., 2024. Improving forecast accuracy for seasonal products in FMCG industry: integration of SARIMA and regression model. *International Journal of Industrial and Systems Engineering*, *46*(2), pp.259-279.
7. Regina, T. and Jodiawan, P., 2021. Proposed Improvement of Forecasting Using Time Series Forecasting of Fast Moving Consumer Goods. *JIEMS (Journal of Industrial Engineering and Management Systems)*, *14*(1).
8. Godoy, J.N.V., Arias, R. and Franco, H., 2022, November. Demand Forecasting of Fast-Moving Consumer Goods by Deep Learning-Based Time Series Analysis. In *International Conference on Smart Technologies, Systems and Applications* (pp. 3-17). Cham: Springer Nature Switzerland.
9. Chaowai, K. and Chutima, P., 2024. Demand Forecasting and Ordering Policy of Fast-Moving Consumer Goods with Promotional Sales in a Small Trading Firm. *Engineering Journal*, *28*(4), pp.21-40.
10. Gijare, J.S. and Kumar, A., MARKET BASKET ANALYSIS FOR A LIFESTYLE STORE, FOCUSING ON THE PURCHASING PATTERNS OF CUSTOMERS.
11. Günesen, S.N., Şen, N., Yıldırım, N. and Kaya, T., 2021, January. Customer churn prediction in FMCG sector using machine learning applications. In *IFIP International Workshop on Artificial Intelligence for Knowledge Management* (pp. 82-103). Cham: Springer International Publishing.
12. Mahdi, M. and Jabbari, M., 2024. Predicting customer churn in the fast-Moving consumer goods segment of the retail industry using deep learning. *Mathematics and Computational Sciences*, *5*(3), pp.58-79.
13. Rithesh, B., Use of Predictive Analysis, Forecasting the Trends in the FMCG.