**A comparative study of Machine learning Approaches for Demand Forecasting in online retail**

Muhammad Ahsan

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics

Diagram

Description automatically generated with medium confidence

May 2024

Supervisor: Sam Weiss

Abstract:

The modern online retail industry sees demand forecasting as playing a crucial role in inventory management efficiency and customer satisfaction. Traditional forecasting is not effective in precisely predicting demand due to the increasing complexity and dynamic nature of online retail. Therefore, the research paper in question suggests a comparative analysis of machine-learning approaches to improving demand forecasting in online retail. This paper, therefore, aims at evaluating the performance of several distinct machine learning tools in predicting retail product demand. To reach the goal, it was essential to conduct a literature review and research different machine learning tools and the instances in which they could be used for demand forecasting. This research aims to use some of the traditional machine learning models and deep learning models to perform demand forecasting. Some of the models to be applied to transactional dataset (Online Retail II) include Auto Regressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM) and with a focus on Darts models, including NBEATSModel, BlockRNNTest, TiDEModel, and TCNModel. In addition, the research will also take into account some of the factors that might affect demand forecasting such as seasonality, trend and customer behaviour.

After conducting an extensive study and evaluation of different machine learning algorithms for demand forecasting in online retail, the results revealed some interesting insights. The ARIMA model provided unsatisfactory results because of neglection in seasonality and non-linear patterns in the data. SARIMA, with its ability to capture seasonal patterns, outperformed ARIMA in accurately predicting sales fluctuations over time. On the other hand, LSTM, being a deep learning model with memory retention capabilities, excelled in capturing complex temporal dependencies and non-linear trends within the data. However, it should be noted that SARIMA model presented a challenge in their predictions, often yielding negative values. This indicates that the forecasted demand occasionally fell below zero, which is not practically meaningful in the context of online retail. To address the limitations of traditional models, the study explored the application of the DARTS library, which encompasses advanced deep learning models designed specifically for time series forecasting. Among these, the Neural Basis Expansion for Time Series (NBeats), Block Recurrent Neural Network (BlockRNNTest), Time-series Dense Encoder (TiDE), and Temporal Convolutional Network (TCN) exhibited superior performance compared to traditional machine learning models. However, the Time-series Dense Encoder model, stood out as the most accurate and superior model for demand forecasting in online retail. The comparative study highlighted the significance of leveraging advanced machine learning models such as TiDE for accurate and reliable demand forecasting in the online retail industry. By embracing the superior predictive capabilities of the TiDE model, online retailers can make informed decisions, improve inventory planning, and ultimately drive profitability.

Introduction:

* 1. Overview:

In simple terms, this project efforts on comparing different time series models. Why? The objective is to predict the quantity needed, helping businesses improve inventory management and reduce costs with this comparative study. By figuring out which model works best, we aim to improve the accuracy of predicting how much stock is required, ultimately making inventory management more efficient and cost-effective. The main objective is to provides the context for the comparative analysis and emphasizes the need to identify the best approach for improved predictive accuracy. This comparison is essential for businesses which are looking to enhance their operations and make better decisions about stocking their products.

* 1. Background:

The accurate forecasting of demand in online retail is crucial for inventory management, resource allocation, and overall business demand forecasting plays a critical role in the supply chain management of online retail businesses. Controlling the retail market is the secret to sustainability in today's business world. By accurately predicting demand, online retailers can optimize inventory levels, minimize stockouts and overstocks, reduce costs, and enhance customer satisfaction. However, traditional forecasting methods may fall short in capturing the complexity and dynamic nature of the online retail industry. To address this challenge, machine learning techniques have gained prominence as effective tools for demand forecasting in online retail. These techniques leverage the power of advanced algorithms and large datasets to extract patterns, trends, and insights that can improve the accuracy of demand forecasts **(Ferreira, J, K., Lee, A, H, B. and Simchi‐Levi, D., 2016)**. Several studies have investigated various analytical techniques, from traditional time-series-based models to advanced neural network architectures. This academic journey involves integrating a wide range of machine learning algorithms, including, but not limited to Support Vector Machine, K-Nearest Neighbour, Gaussian Nave Bayes, regressions, Random Forest, Decision Tree Regressor, and Extreme Gradient Boosting (XGBoost) models. Notably, the Gaussian Nave Bayes algorithm emerges as a standout performer, showcasing unparalleled accuracy in the demand estimation **(Wang, J., Liu, Q, G. and Liu, L., 2019)**. This revelation has promising implications for retailers navigating the dynamic online marketplace, enabling them to anticipate and meet product demand. As the online retail space continues to redefine consumer interactions, robust demand forecasting becomes increasingly important for businesses aiming to optimize their supply chain and operational efficiency. The other study provides a detailed look at the changing landscape of time series forecasting methods, highlighting the growing use of machine learning over traditional ARIMA models. The research compares the performance of AI algorithms with ARIMA and hybrid statistical-AI models, emphasizing the superior predictive accuracy of machine learning in retail sales forecasting **(Kontopoulou, I, V. et al., 2023)**. This comparative study aims to contribute to the existing literature by examining and comparing the performance of various machine learning approaches for demand forecasting in the online retail sector. On the other hand, the work introduces a novel forecasting method that combines Long Short-Term Memory (LSTM) networks and Random Forest (RF), showcasing superior performance compared to traditional forecasting techniques **(Punia, S. et al., 2020)**. This study highlights the robustness and suitability of advanced machine learning algorithms in multi-channel retail demand forecasting, further corroborating the trend observed in contemporary research. The Decision Tree Algorithm achieved nearly 71% overall accuracy, while the Generalized Linear Model attained 64% accuracy. Notably, this study identifies the Gradient Boosted Tree as the best-fit model for sales forecasting, highlighting its superior performance compared to other models evaluated. In conclusion, the use of machine learning approaches for demand forecasting in online retail has shown significant promise and outperformed traditional methods such as ARIMA. In spite of the advancements in machine learning techniques for demand forecasting in online retail, there are still unresolved challenges. For example, there is a need to leverage real-time data and update forecasting models in a changing environment. Additionally, there is a lack of research on how to leverage demand and supply information throughout the product lifecycle, rather than relying on a single data point **(Liu, N. et al., 2013)**. So this research will more explore the potential of machine learning approaches for demand forecasting in online retail and identify the most effective models for accurate predictions on time series dataset.

* 1. Research Objectives:

This research aims to hypothesis a strong and versatile framework for machine learning approaches modified to the complex domain of online retail forecasting. This research provides context for the comparative analysis and emphasizes the need to identify the best approach for improved predictive accuracy.

1. The first objective is to perform Exploratory Data Analysis on real-world online retail dataset (Online Retail II) to understand the characteristics of the data and identify significant predictor variables.
2. The second objective is to compare the performance of different machine learning techniques, such as ARIMA, SARIMA, LSTM, and DARTS forecasting methods, in terms of accuracy and robustness, taking into consideration the specific challenges and characteristics of online retail demand forecasting. And identify the most effective machine learning models for accurate demand predictions in online retail.
3. To assess the generalization ability, runtime, time, and convenience of the different machine learning models to determine their practical applicability and scalability in real-world online retail environments.
4. Critically analyses the limitations encountered in the study and provides recommendations for future development. Identify areas for improvement, propose potential extensions, and offer guidance for further exploration in online retail forecasting.

In struggling toward these objectives, this research attempt to provide a comprehensive guide for retail industry, offering practical insights into the utilization of machine learning approaches for demand forecasting. Through a thorough comparison of various forecasting techniques, the study aims to outline the strengths and weaknesses of each method, enabling decision-makers to make well-informed choices tailored to their specific business needs. Additionally, the research seeks to excite further progress in the field by analytical areas where current knowledge is lacking and suggesting paths for future research, particularly in the exploration of advanced deep learning models and their application to complex time-series forecasting scenarios. Ultimately, the goal of this study is to contribute to the ongoing improvement of demand forecasting practices in online retail, facilitating more efficient operations, improved customer satisfaction, and informed decision-making in the ever-changing world of e-commerce.

* 1. Research Questions:

This study addresses the challenge of identifying the most suitable machine learning model for predicting demand in online retail. Various aspects such as accuracy, reliability, efficiency, and simplicity of implementation will be compared. The study hypothesizes that by comparing multiple machine learning paradigms - including auto-regressive, traditional, and deep learning models - the best approach for accurately forecasting time series data related to online retail trends can be found. The objective is to provide businesses with a clear understanding of the most effective machine learning model to improve forecasting precision in online retail (**Ensafi, Y. et al., 2022**). This research seeks to provide insights that empower decision-makers to make informed choices and optimize their forecasting strategies within the evolving online retail landscape. By examining the performance of different machine learning algorithms and comparing them to traditional statistical methods in demand forecasting for online retail, this research aims to provide guidance on which approaches are most effective in accurately predicting demand in the online retail industry. Some research questions that could be explored in this study include:

1. Which machine learning model exhibits superior precision in predicting demand within the realm of online retail, and how does its effectiveness compare against alternative models in the domain? (**Ferreira, J, K., Lee, A, H, B. and Simchi‐Levi, D., 2016**)﻿
2. What algorithms best suit online retail forecasting or time series forecasting problems?
3. How effectively do Darts models, including NBEATSModel, BlockRNNTest, TiDEModel, and TCNModel, address dynamic challenges such as seasonality, sudden fluctuations in demand, and shifts in market trends within the online retail sector?
4. What are the unique strengths and limitations of individual Darts models concerning interpretability, scalability, and outlier resilience, and how do these aspects impact their real-world usability within the online retail landscape?
   1. Outline of the thesis:

Literature Review:

Several studies have been presented on demand forecasting in the online retail sector. Principally, the studies are conducted to explore different machine learning algorithms like time-series-based models, regression-based models, and neural network models. Most of the machine learning models like Linear regression and decision trees get low accuracy as you will read in this literature review. The main goal of this research is to get high accuracy for demand forecasting by using different machine learning algorithms and recognize the best-performing models for demand forecasting. The choice of the algorithm depends on the characteristics of the online retail business because each ML algorithm offers unique advantages and limitations.

* 1. A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company:

In this research paper by (Mitra et al., 2022), the authors conducted a comparative study on demand forecasting models for the retail sector, recognizing the pivotal role of sales forecasting and demand planning in optimizing supply chain performance. The study focused on several machine learning models, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Gradient Boosting, Adaptive Boosting (AdaBoost), and Artificial Neural Network (ANN). Notably, the hybrid model combining Random Forest, XGBoost, and Linear Regression (LR) emerged as the top performer regarding accuracy. The findings underscore the potential of machine learning approaches to enhance sales forecasting precision in retail. The proposed hybrid model outperforms other models and carries significant implications for optimizing capacity, labour, inventory, and overall supply chain efficiency, offering valuable insights for practitioners in the field.

* 1. Time-series forecasting of seasonal items sales using machine learning:

This research by (Ensafi et al., 2022) looked into predicting furniture sales using different models. They compared methods like SARIMA, Triple Exponential Smoothing, Prophet, LSTM, and CNN. The Stacked LSTM method was the most accurate, with Prophet and CNN performing well. The study emphasizes the importance of sales forecasting in the furniture industry. The literature review explores various approaches for accuracy, considering related papers in sales forecasting. To measure precision, they used methods like RMSE and MAPE. Overall, the research sheds light on effective ways to predict furniture sales, focusing on finding the most accurate forecasting model.

* 1. Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach:

This research by (Panarese et al., 2022) introduces a machine learning-based platform designed for enhancing sales forecasting in the retail sector. The study employs machine learning techniques to optimize sales forecasting by recognizing the retail industry's integration of IT innovation for improved customer experiences. The methods applied involve decision trees (DT) and boosting techniques, focusing on the XGBoost regression model. The results demonstrate the development of a prototype platform tailored for managing trading companies, showcasing the XGBoost algorithm as the top performer in sales forecasting. Notably, the XGBoost regression model contributes to a remarkable 15-20% improvement in forecasting accuracy, highlighting its efficacy in advancing the capabilities of sales prediction platforms within the retail sector. This research presents a promising step toward leveraging machine learning for more accurate and efficient sales forecasting, with the potential to significantly impact the operations of trading companies in the retail landscape.

* 1. Intelligent Sales Prediction Using Machine Learning Techniques:

This paper by Cheriyan et al. (2018) researches the realm of intelligent sales prediction, leveraging machine learning techniques and data mining to elevate the accuracy and efficiency of forecasting sales trends through a meticulous analysis of various data mining techniques, including the Generalized Linear Model, Decision Tree, and Gradient Boost Algorithm. The model performance results show that the Gradient Boost Algorithm achieves the highest overall accuracy of 98% in sales forecasting, followed by the Decision Tree Algorithm with approximately 71% accuracy and the Generalized Linear Model with 64% accuracy. The study reveals that the Gradient Boost Algorithm emerges as the top performer with an impressive 98% overall accuracy in sales forecasting. This research envisions future improvements in sales prediction models, aiming to explore additional machine learning algorithms for enhanced system intelligence without manual intervention. The significance of these findings lies in their contribution to the field of sales forecasting, showcasing the effectiveness of machine learning algorithms and data mining techniques in fortifying the accuracy and reliability of predictions.

* 1. Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail:

This paper by (Punia et al) introduces a novel forecasting method that combines long short-term memory (LSTM) networks and random forest (RF) for demand forecasting in multi-channel retail. The proposed method exhibits a unique ability to model complex relationships of both temporal and regression types, providing a competitive edge in accuracy compared to other forecasting methods. Evaluation of a real-world multivariate dataset from a multi-channel retailer, benchmarked against various techniques, including neural networks, multiple regression, ARIMAX, LSTM networks, and RF, reveals the new method's statistical superiority in bias, accuracy, and variance. The study not only ranks explanatory variables by their relative importance but also highlights the forecasting method's robustness for extended horizons across online and offline channels in multi-channel retail. This research contributes significantly to Operational Research Management Science (ORMS) and Predictive Analytics, opening avenues for future exploration into advanced neural networks, comparative analyses with other deep learning models, and testing the proposed method on data from non-traditional fulfilment channels.

* 1. A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks:

This paper by (Kontopoulou et al., 2023) navigates the evolving landscape of time series forecasting, contrasting traditional ARIMA models with contemporary artificial intelligence (AI) approaches. Acknowledging ARIMA models' historical prominence for their mathematical simplicity, the study emphasizes a recent shift towards AI techniques, revealing that, in general, AI algorithms outperform ARIMA models. Significant exceptions exist, prompting the exploration of hybrid statistical-AI models that combine the strengths of both approaches, consistently showcasing superior performance. In stock index forecasting, LSTM and XGBoost algorithms prove more effective than ARIMA, except for datasets with significantly lower values. The findings underscore the potential superiority of AI algorithms in time series forecasting, emphasizing the need for further evaluation and analysis. The paper suggests future research directions, focusing on understanding cases where ARIMA models excel and optimizing hybrid statistical-AI models for enhanced forecasting performance.

* 1. Conclusion:

In synthesis, the exploration of various methodologies for demand forecasting in the online retail domain underscores the crucial role of machine learning (ML) and artificial intelligence (AI) techniques in optimizing predictive accuracy. The literature review clarifies a important trend towards leveraging ML algorithms such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks for enhanced sales prediction. These studies collectively advocate for the effectiveness of hybrid models, combining diverse ML approaches to harness the strengths of individual algorithms (**Tsai, C. et al., 2013**).

Firstly, while existing studies have showcased the efficiency of hybrid ML models, there remains a need to investigate deeper into the specific characteristics and challenges of online retail datasets (**Adulyasak, Y. et al., 2023**). This research aims to bridge this gap by performing thorough Exploratory Data Analysis (EDA) on the dataset (Online Retail II), thereby explaining the underlying patterns and identifying significant predictor variables. Secondly, while previous research has compared various ML techniques for demand forecasting, including ARIMA, SARIMA, LSTM, Random Forest and XGBoost methods, there is a lack of agreement on the most effective models for online retail demand prediction (**Resul, T. and Sule, O., 2020**). This study seeks to fill this gap by conducting a comprehensive comparison of these techniques with DARTS library model such as NBeats, RNN, TiDE, and TCN, considering their accuracy and robustness in addressing the unique challenges of online retail forecasting. Furthermore, while some studies have evaluated the performance of ML models, there is limited understanding into their generalization ability, runtime, and practical applicability in real-world online retail environments. This research endeavours to assess these aspects to control the scalability and convenience of different ML models for practical implementation (**Ensafi, Y. et al., 2022**). Lastly, while previous research has underwritten valuable findings, there remains a need for critical analysis of limitations encountered and recommendations for future development. This study aims to fulfil this objective by classifying areas for development, proposing potential extensions, and offering guidance for further exploration in online retail forecasting.

In essence, by addressing these gaps and objectives, this research paper aims to provide a comprehensive understanding of machine learning approaches for demand forecasting in online retail, offering practical insights for enhancing predictive accuracy and optimizing supply chain performance in the digital commerce landscape.

Research Methodology:

Over the years, various data mining and machine learning techniques have been proposed by analysts and researchers to tackle prediction challenges that have significant impacts on our daily routines. Among these challenges, forecasting product demand emerges as a critical task for organizations and businesses, enabling them to make informed decisions crucial for their expansion and development. This study focuses on employing the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology as a guiding framework for data mining endeavours, with the aim of generating robust forecasts in alignment with business objectives. The CRISP-DM methodology offers a structured approach that streamlines the research process, facilitating more substantial and expedited outcomes. Comprising six distinct stages, this methodology provides a systematic roadmap for the preparation and execution of the study, thereby enhancing its clarity and coherence.

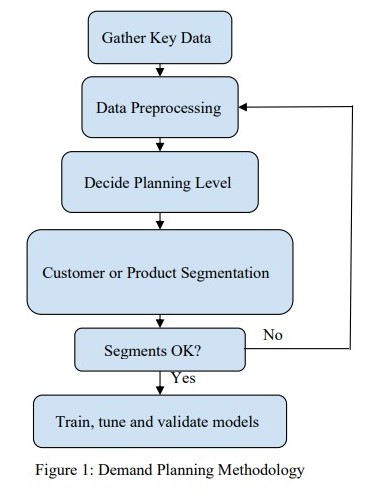


Figure 1: Demand Planning Methodology

* 1. CRISP-DM Methodology:

Data mining represents an innovative approach that necessitates a diverse range of skills and expertise. Presently, there is no universally embraced framework for conducting data mining projects, highlighting the substantial influence of individuals or teams spearheading such initiatives. The necessity for a standardized methodology in data mining is apparent, one capable of guiding the translation of business or organizational challenges into actionable data mining tasks, suggesting appropriate data transformations and techniques, and providing a systematic means to evaluate feasibility and document acquired knowledge.

In order to ensure a systematic and comprehensive research process, this study will adopt the Cross-Industry Standard Process for Data Mining methodology. The CRISP-DM methodology is a widely accepted framework for data mining projects, consisting of six sequential phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The use of CRISP-DM methodology ensures that all aspects of the research process, from understanding the business context to deploying the final model, are covered in a structured and organized manner (**Ofoegbu, K., (no date)**). This methodology will enable a systematic approach to analysing demand forecasting in online retail and exploring the effectiveness of different machine learning algorithms. However, the primary function of the CRISP-DM methodology is to facilitate communication and collaboration among team members with diverse expertise and experiences to ensure an integrated and productive project. It brings together diverse tools and individuals with varying expertise and experiences, to form a cohesive and efficient project team. By following the CRISP-DM methodology, this study aims to achieve a clear understanding of the business problem and goals related to demand forecasting in online retail and develop accurate prediction models using machine learning algorithms.

The CRISP-DM data mining model offers a comprehensive framework for the entire lifecycle of a data mining project, outlining its stages, tasks, and outcomes. This lifecycle is structured into six phases, as depicted in Figure 1. The order of these phases is adaptable, with arrows indicating the primary dependencies between stages, signifying the result of each step or the subsequent key activity in the phase. The outer circle in Figure 1 represents the cyclical nature of data mining. Even post-implementation of the solution, data mining continues as an ongoing process. Information gained throughout the project, in conjunction with the chosen approach, may lead to new and often more precise business concerns. Subsequent data mining endeavours can then benefit from this accumulated experience. The stages of the CRISP-DM model have been utilized throughout this project and will now be explained within the context of demand forecasting.

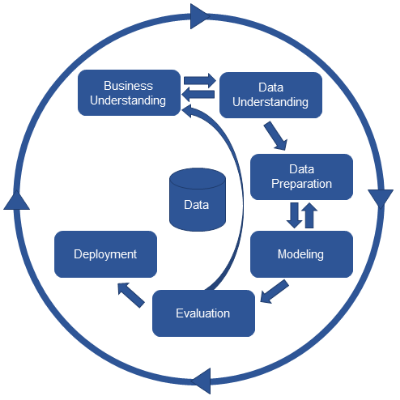


Figure 2: CRISP-DM workflow

The stages outlined in the CRISP-DM model have been systematically applied throughout the duration of this project. Hereafter, an explanation of these stages will be presented within the context of demand forecasting.

* + 1. Business Understanding:

In the preliminary phase of this research endeavor, the central objective revolves around attaining a comprehensive understanding of the business goals entwined with demand forecasting, with a specific lens focused on predicting the requisite quantity for a gift store's inventory. This endeavor is fundamentally rooted in the alignment of demand forecasting efforts with the overarching strategic objectives of the gift store. A multifaceted approach is adopted, entailing several key considerations integral to this phase. Firstly, the process begins with a meticulous examination aimed at identifying and selecting the top 10 products within the gift store's inventory landscape that hold paramount significance. This selection is not arbitrary but guided by the recognition that understanding the demand dynamics surrounding these specific products will yield actionable insights pivotal for streamlining inventory management practices and optimizing sales strategies. Subsequently, a critical aspect entails the precise delineation of the key metrics indispensable for effective demand forecasting. Of particular significance is the task of accurately predicting the quantity required for the identified top 10 products. This necessitates a granular understanding of various metrics and indicators, including the frequency of demand forecasts and discerning any discernible patterns in seasonal variations.

Moreover, there exists a pressing need to delve into the potential business impact that accurate demand forecasting for the top 10 products can wield. This entails a comprehensive exploration of how such forecasting initiatives can catalyze transformative changes across critical business dimensions. Areas of consideration encompass the optimization of inventory management protocols and the strategic deployment of promotional tactics aimed at enhancing operational efficiency and bolstering revenue streams.Integral to this phase is also the active engagement with key stakeholders vested in the gift store's operations, including store managers and inventory personnel. By soliciting insights from these stakeholders, this research ensures alignment with their unique requirements and expectations, thereby fostering a collaborative environment conducive to the attainment of shared objectives. Furthermore, a thorough examination of historical sales data pertaining to the selected top 10 products is undertaken to unravel underlying patterns, trends, and fluctuations in demand. This historical backdrop serves as a foundational cornerstone upon which robust forecasting models can be constructed, leveraging insights gleaned from past performance to inform future projections. Lastly, due consideration is accorded to external factors capable of exerting influence on demand dynamics, such as seasonal events or holidays. Incorporating these external variables into the forecasting framework is deemed essential to fortify the predictive accuracy of the models, thereby ensuring their efficacy in real-world scenarios.

* + 1. Data Understanding:

Commencing with the basics of understanding data, the first step involves collecting it. Then, we dive deeper to get to know the data better. This study looks for any issues like inconsistencies and explore the data to find initial insights or interesting groups of information. The main aim is to draw conclusions and uncover hidden knowledge by thoroughly understanding the details and patterns within the dataset.

1. **Dataset Name**: Designated as the "Online Retail II Transaction Dataset."
2. **Source:** Originating from a UK-based and registered online retail entity, encompassing transactions spanning from 01/12/2009 to 09/12/2011.
3. **Description:** This dataset summarizes real-time online retail transactions spanning a two-year duration, representative of a company specializing in the sale of unique all-occasion giftware. Particularly, a significant portion of its clientele comprises wholesalers.
4. **Dataset Characteristics**:

* Multivariate: Signifying the presence of multiple variables or features associated with each transaction.
* Sequential: Ordered chronologically, reflecting a time series of transactions.
* Time-Series: Captures transactional activities over a defined period, thereby classified as a time-series dataset.
* Subject Area**:** Business domain.

1. **Feature Type:**

* Integer: Encompassing integer values denoting quantities, customer IDs, and similar attributes.
* Real: Comprising real numbers indicative of prices, monetary values, and related metrics.

1. **Number of Instances:** A substantial dataset comprising 1,067,371 transactions.
2. **Number of Features:** While an exact count of features is unspecified, the dataset harbours multiple dimensions of information.
3. **Dataset Information:** Representing transactions conducted by customers, the dataset constitutes sequential and time-series data, relating interactions with the online retail establishment over a span of two years. Each transaction is anticipated to encompass of details encompassing product particulars, purchase quantities, pricing details, customer identifiers, and timestamps.
4. **Additional Information:**

* Specific to a UK-based online retail entity.
* Transactional period spans from 01/12/2009 to 09/12/2011.
* Specialization in unique all-occasion giftware.
* Majority of wholesale customers.

1. **Missing Values:** Affirmative, indicating the presence of missing data points within the dataset.
   * 1. Data Preparation:

The data preparation phase is crucial for refining collected data to ensure accurate demand forecasting for the gift store. It involves tasks aimed at improving dataset quality and relevance. This includes selecting suitable data types, identifying outliers, addressing missing values, ensuring non-negative quantity, integrating a holidays library, creating a new dataset, and refining attributes. The initial step in data preparation involves meticulous selection of appropriate data types for each variable within the dataset. This critical phase ensures that the data is structured in a manner that aligns seamlessly with the requirements of the forecasting model, thereby fostering streamlined and effective analytical procedures. This study used a dataset name as Online Retail ll which is shown in Figure 3:



Figure 3: Dataset before preprocessing

Beginning with the selection of appropriate data types for each variable, this step ensures that the dataset is structured in a manner helpful to efficient analysis and unified integration with forecasting models. Subsequently, outlier detection mechanisms are employed to identify and address inconsistent data points that could potentially skew model predictions. By implementing robust outlier handling techniques, the dataset's integrity is bolstered, thereby enhancing the reliability of forecasting outcomes.

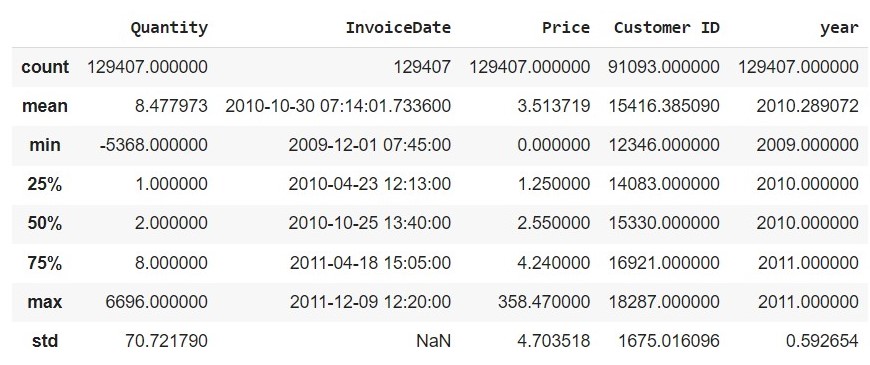


Figure 4 : Description of the dataset

* + 1. Data Cleaning and Preprocessing:

The code provided in the data preparation phase clarifies the practical implementation of these preparatory tasks. Through a series of data cleaning and preprocessing steps, including outlier removal, handling of missing values, and attribute creation, the dataset is meticulously curated to align with the objectives of demand forecasting (**Kumar, A., Shankar, R. and Aljohani, R, N., 2020**). Each code snippet is accompanied by explanatory comments, elucidating the rationale behind the respective data manipulation techniques employed. This transparent approach not only facilitates reproducibility but also underscores the rigor and thoroughness inherent in the data preparation process. The dataset description provides a comprehensive overview of the variables included in the dataset, offering insights into their respective distributions and characteristics. Examining the "Quantity" variable reveals a wide range of values, from negative to positive integers, indicating both purchases and potentially returns. The presence of negative quantities suggests a need for further investigation into the data collection process to understand the context of such occurrences. Upon detection of such values, a notification is printed, indicating the need for further analysis. Then, corrective measures are implemented to address inconsistencies in the dataset. Rows with positive quantities, indicative of purchases, are analysed to identify entries where the corresponding invoice contains the 'C' identifier, denoting cancellations or returns. These entries are deemed inconsistent and subsequently removed to maintain data coherence. Furthermore, attention is directed towards rows featuring negative quantities, potentially signifying returns, yet lacking the 'C' identifier in the corresponding invoice. Such inconsistencies are rectified by eliminating these entries, ensuring that negative quantities are appropriately associated with cancellation transactions. This particular approach to data validation and cleaning safeguards against inaccuracies and inconsistencies, thereby bolstering the dataset's suitability for subsequent analyses, including demand forecasting within online retail (**Gupta, P. et al., 2021**). Moving on to the "InvoiceDate" variable, the provided mean date and time indicate the central tendency of transaction timestamps, albeit with an unspecified standard deviation, which warrants caution regarding potential inconsistencies or missing values in this temporal data.

Regarding "Price," the dataset showcases a diverse range of item prices, as reflected in the minimum, maximum, and mean values. The presence of zero-priced items raises questions about their nature, potentially indicating promotional offers or data entry errors that require scrutiny during the data preparation phase. "Customer ID" serves as a unique identifier for customers, with the count of unique IDs suggesting the breadth of the customer base. The variability in customer IDs may also indicate varying purchasing patterns or levels of engagement across different customer segments. Then the preprocessing steps applied to the "Description" column, the aim is to enhance the clarity and consistency of product names for improved understanding and analysis. The string manipulation functions are utilized to standardize the format of product names. Specifically, the “replace()” method is employed to remove any periods ('.') present in the product names, ensuring uniformity. Additionally, the names are converted to uppercase using the “upper()” function to standardize the casing. The strip() method is then applied to remove any leading or trailing white spaces from the names. Following this standardization process, rows containing product names with a question mark ('?') are identified and subsequently removed from the dataset. This step aims to eliminate entries with ambiguous or potentially erroneous product names, thereby enhancing the overall quality of the dataset (**Verstraete, G., Aghezzaf, E. and Desmet, B., 2019**). Missing product names (identified by null values in the "Description" column) are addressed through a data imputation process based on the associated "StockCode." For each missing product name, the most frequent product name corresponding to the same "StockCode" is determined using the mode() function. If a suitable replacement is found, it is assigned to the missing entry. However, if no matching product name is available for a given "StockCode," the corresponding row is removed from the dataset. To ensure consistency, the "Description" column is explicitly converted to a string data type, facilitating uniform handling and compatibility with subsequent analyses. Overall, these preprocessing steps contribute to refining the "Description" column, thereby enhancing the interpretability and reliability of the dataset for further analysis, such as demand forecasting in the context of online retail.

Furthermore, particular attention is directed towards handling missing and incorrect values within the dataset. Techniques such as data imputation and removal of incomplete data points are applied to moderate the adverse effects of missing or erroneous data on model performance. Moreover, a rigorous rule is enforced to ensure that quantity values remain non-negative throughout the dataset (**Kontopoulou, I, V. et al., 2023**). This normalization step aligns with the practicality of product quantities and contributes to the coherence of the dataset. An integral aspect of data preparation involves incorporating contextual information, such as the impact of holidays on sales, into the dataset. Leveraging a holidays library, this study improves the dataset with pertinent temporal context, thereby empowering the forecasting model with insights into seasonal variations in demand. Moreover, holiday data is incorporated into the dataset to account for their influence on sales and demand patterns in online retail. Initially, holiday information is retrieved for selected countries between 2009 and 2011, with dates and holiday names appended to a consolidated list. This data is then organized into a DataFrame and filtered based on the specified date range, ensuring relevance to the dataset. Duplicate entries are removed to maintain data integrity. Alongside, date-related features, including month, day of the year, and weekday, are extracted from transaction dates to provide temporal context. Then, the holiday data is merged with the main dataset, enabling the identification of transaction dates coinciding with holidays. New columns representing each unique holiday are created, with values indicating holiday observance. Finally, redundant columns are eliminated, resulting in a refined dataset enriched with holiday information, poised for subsequent analysis, particularly in demand forecasting within online retail. Lastly, the "Year" variable highlights the temporal distribution of transactions, with the mean year indicating the central tendency of transaction years within the dataset (**Cheriyan, S. et al., 2018**).

* + 1. Modelling:

During the modelling stage, a particular selection and implementation of various modelling techniques are undertaken, aiming to optimize their performance through careful parameter tuning. This phase commonly involves encountering multiple techniques suitable for addressing the same data mining problem, each necessitating specific data formats for effective application (**Gupta, D, S., Fuehrer, F. and Jeyachandra, C, B., 2014**). The processes of data planning and modelling are closely intertwined, as challenges related to data quality or the need for additional insights frequently emerge during the modelling process (**Cappiello, C., Samá, W. and Vitali, M., 2018).** In the context of this research, seven distinct types of modelling techniques are considered for prediction purposes such as ARIMA, SARIMA, LSTM and four models of DARTS library. These techniques encompass a diverse range of methodologies, each offering unique strengths and capabilities for addressing the complexities inherent in the dataset. Through a systematic exploration of these techniques, the aim is to identify the most suitable approach that yields robust and accurate predictions in the domain of online retail demand forecasting (**Wang, J., Liu, Q, G. and Liu, L., 2019**). The ARIMA model provided unsatisfactory results because of neglection in seasonality and non-linear patterns in the data. SARIMA, with its ability to capture seasonal patterns, outperformed ARIMA in accurately predicting sales fluctuations over time. On the other hand, LSTM, being a deep learning model with memory retention capabilities, excelled in capturing complex temporal dependencies and non-linear trends within the data. However, it should be noted that SARIMA model presented a challenge in their predictions, often yielding negative values. This indicates that the forecasted demand occasionally fell below zero, which is not practically meaningful in the context of online retail. To address the limitations of traditional models, the study explored the application of the DARTS library, which encompasses advanced deep learning models designed specifically for time series forecasting. Among these NBeats, BlockRNNTest, TiDE, and TCN exhibited superior performance compared to traditional machine learning models. However, the Time-series Dense Encoder model (TiDE), stood out as the most accurate and superior model for demand forecasting in online retail. The comparative study highlighted the significance of leveraging advanced machine learning models such as TiDE for accurate and reliable demand forecasting in the online retail industry. By embracing the superior predictive capabilities of the TiDE model, online retailers can make informed decisions, improve inventory planning, and ultimately drive profitability. In this project, seven distinct types of techniques are considered for predictions.

References:

1. Adulyasak, Y. et al. (2023) "Retail Analytics in the New Normal: The Influence of Artificial Intelligence and the Covid-19 Pandemic," arXiv (Cornell University). Available at: https://doi.org/10.48550/arxiv.2312.00046.
2. Ensafi, Y. et al. (2022) Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. Available at: https://doi.org/10.1016/j.jjimei.2022.100058.
3. Ferreira, J, K., Lee, A, H, B. and Simchi‐Levi, D. (2016) Analytics for an Online Retailer: Demand Forecasting and Price Optimization. Available at: https://doi.org/10.1287/msom.2015.0561.
4. Kontopoulou, I, V. et al. (2023) A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. Available at: https://doi.org/10.3390/fi15080255.
5. Liu, N. et al. (2013) Sales Forecasting for Fashion Retailing Service Industry: A Review. Available at: https://doi.org/10.1155/2013/738675.
6. Punia, S. et al. (2020) Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. Available at: https://doi.org/10.1080/00207543.2020.1735666.
7. Wang, J., Liu, Q, G. and Liu, L. (2019) A Selection of Advanced Technologies for Demand Forecasting in the Retail Industry. Available at: <https://doi.org/10.1109/icbda.2019.8713196>.
8. Resul, T. and Sule, O. (2020) Demand Prediction Using Machine Learning Methods and Stacked Generalization. Available at: <https://doi.org/10.48550/arXiv.2009.09756>.
9. Tsai, C. et al. (2013) "A comparative study of hybrid machine learning techniques for customer lifetime value prediction," Kybernetes, 42(3),p. 357-370. Available at: <https://doi.org/10.1108/03684921311323626>.
10. Kumar, A., Shankar, R. and Aljohani, R, N. (2020) A big data driven framework for demand-driven forecasting with effects of marketing-mix variables. Available at: <https://www.sciencedirect.com/science/article/pii/S0019850118306606>.
11. Gupta, P. et al. (2021) Implementation of Demand Forecasting – A Comparative Approach. Available at: <https://doi.org/10.1088/1742-6596/1714/1/012003>.
12. Verstraete, G., Aghezzaf, E. and Desmet, B. (2019) A data-driven framework for predicting weather impact on high-volume low-margin retail products. Available at: <https://doi.org/10.1016/j.jretconser.2019.02.019>.
13. Cappiello, C., Samá, W. and Vitali, M. (2018) Quality awareness for a Successful Big Data Exploitation. Available at: <https://doi.org/10.1145/3216122.3216124>.
14. Gupta, D, S., Fuehrer, F. and Jeyachandra, C, B. (2014) Production Forecasting in Unconventional Resources using Data Mining and Time Series Analysis. Available at: <https://doi.org/10.2118/171588-ms>.