# DEMAND FORECASTING FOR E-COMMERCE PRODUCTS USING MACHINE LEARNING

MATSA VENKATA LAKSHMI KAVYA

Final Thesis Report

December 2023

**DEDICATION**

I dedicate this research project to my beloved parents, whose unwavering support and encouragement have been the driving force behind every step of my academic journey. Their boundless love and belief in my abilities have been my guiding light. I would also like to express my deepest gratitude to one of my friends, whose encouragement and support have been invaluable throughout this endeavor. This project is a reflection of the collective strength and inspiration derived from the love and encouragement of my family and friends. Thank you for being my pillars of strength.

**ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to all those who have contributed to the completion of this research. First and foremost, I extend my heartfelt thanks to my advisor Vibor Kaushik, whose guidance and expertise played a pivotal role in shaping the direction of this study. I am also grateful to the faculty members for their valuable insights and constructive feedback. Special appreciation goes to my colleagues and peers for their support and collaborative spirit. Additionally, I acknowledge the support from Liverpool John Moore’s University, which provided the resources necessary for the successful execution of this project. Lastly, my deepest thanks to my family and friends for their unwavering encouragement and understanding throughout this academic journey.

**ABSTRACT**

The surge in E-commerce activities in recent times has underscored the critical need for accurate demand forecasting to optimize inventory management and enhance operational efficiency. This research endeavors to address this imperative through the lens of machine learning, specifically employing ensemble algorithms like Linear Regression, Decision Trees, Random Forest, XGBoost, SVM, LSTM, and FNN. The purpose of the study is to advance the precision of demand forecasting for E-commerce products, acknowledging the dynamic nature of consumer behavior and market trends. The primary aim is to harness the capabilities of machine learning models to predict demand patterns, thereby facilitating proactive decision-making for inventory optimization. The study commences with a comprehensive exploration of relevant data, incorporating product attributes, historical sales data, and potential external factors influencing demand. Rigorous data cleaning and preprocessing techniques are employed to ensure the dataset's suitability for machine learning models. The subsequent phase involves a detailed exploration of relationships within the data through Exploratory Data Analysis (EDA), guiding the selection of influential features for the ensemble algorithms. Leveraging a carefully split dataset for training and testing, the ensemble models are trained to capture intricate patterns in historical data, and their efficacy is evaluated using metrics such as Mean Absolute Error (MAE) and Mean Squared Error (MSE). The results demonstrate the effectiveness of the ensemble algorithms in providing accurate demand forecasts for E-commerce products.

**TABLE OF CONTENTS**

DEDICATION 2

ACKNOWLEDGEMENTS 3

ABSTRACT 4

LIST OF FIGURES 9

LIST OF ABBREVIATIONS 10

CHAPTER 1: INTRODUCTION 11

1.1 Background of the Study 11

1.2 Problem Statement 12

1.3 Aim and Objectives 13

1.4 Research Questions 13

1.5 Scope of the Study 14

1.6 Significance of the Study 14

1.7 Structure of the Study 15

CHAPTER 2: LITERATURE REVIEW 17

2.1 Introduction 17

2.2 Traditional Demand Forecasting Methods 17

2.3 Limitations of Traditional Methods in E-commerce 18

2.4 Machine Learning in Demand Forecasting 18

2.5 Applications of Machine Learning in E-commerce 19

2.6 Challenges and Opportunities in E-commerce Demand Forecasting 20

2.7 Related Research Publications 21

CHAPTER 3: RESEARCH METHODOLOGY 25

3.1 Research Methodology 26

3.1.1 Research Design 26

3.1.2 Data Collection 26

3.1.3 Data Preprocessing 26

3.1.4 Data Exploration and Visualization 27

3.1.5 Feature Selection 27

3.1.6 Model Selection 27

3.1.7 Data Splitting 28

3.1.8 Model Training 28

3.1.9 Model Evaluation 28

3.2 Proposed Methods 29

3.2.1 Linear Regression 29

3.2.2 Decision Tree Regressor 29

3.2.3 Random Forest Regressor 30

3.2.4 XGBoost Regressor 30

3.2.5 Support Vector Machine 30

3.2.6 Long Short Term Memory 31

3.2.7 Feed Forward Neural Networks 31

3.2.8 Principal Component Analysis 32

3.3 Evaluation Metrics 32

3.3.1 Mean Squared Error 32

3.3.2 Root Mean Squared Error 32

3.3.3 Mean Absolute Error 33

3.4 Summary 33

CHAPTER 4: ANALYSIS 34

4.1 Dataset Description 34

4.2 Data Preparation 35

4.2.1 Elimination of Variables 35

4.2.2 Transformation into Numerical Variables 35

4.2.3 Rename the Columns 35

4.2.4 Creating New Features 36

4.2.5 Handling of Missing Values 36

4.2.6 Univariate Analysis 36

4.2.7 Bivariate Analysis 37

4.3 Data Visualization 37

4.3.1 Box Plot 37

4.3.2 Bar Graph 39

4.3.3 Scatter Plot 41

4.3.4 Count Plot 42

4.3.5 Heatmap 44

4.3.6 Regplot 45

4.4 Model Building 47

4.5 Interpretation of Visualizations 47

4.6 Required Resources 47

4.6.1 Hardware Resources 47

4.6.2 Software Resources 48

4.7 Results and Discussions 48

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS 49

5.1 Conclusion 49

5.2 Future Recommendations 49

REFERENCES 50

APPENDIX A: RESEARCH PROPOSAL 54

Abstract 55

List of Abbreviations 56

Table of Contents 57

1 Background 58

2 Problem Statement or Related Research or Related Work 59

3 Research Questions (If any) 62

4 Aims and Objectives 62

5 Significance of the Study 63

6 Scope of the Study 63

7 Research Methodology 64

7.1 Steps of Methodology 64

7.1.1 Data Collection 64

7.1.2 Data Preprocessing 64

7.1.3 Data Exploration and Visualization 64

7.1.4 Feature Selection 64

7.1.5 Model Selection 64

7.1.6 Data Splitting 65

7.1.7 Model Training 65

7.1.7.1 Linear Regression 65

7.1.7.2 Decision Tree Regressor 65

7.1.7.3 K-Nearest Neighbor Regressor 66

7.1.7.4 Random Forest Regressor 66

7.1.7.5 Long Short Term Memory 67

7.1.7.6 Convolutional Neural Network 67

7.1.8 Model Evaluation 67

7.1.8.1 Mean Squared Error 68

7.1.8.2 Root Mean Squared Error 68

7.1.8.3 Accuracy 68

8 Requirement Resources 68

8.1 Hardware Requirements 68

8.2 Software Requirements 68

9 Research Plan 69

REFERENCES 69

**LIST OF FIGURES**

Figure 1. Box Plot with outliers …………………………………………………………… 37

Figure 2. Box Plot without outliers ………………………………………………………....38

Figure 3. Main Category vs Selling Price …………………………………………………..39

Figure 4. Sub Category vs Selling Price ……………………………………………………40

Figure 5. Top Ten Brands in Amazon ………………………………………………………40

Figure 6. Actual Price vs Selling Price vs Actual Demand …………………………………41

Figure 7. Discount Price vs Selling Price vs Discount Demand ……………………………41

Figure 8. Sub Category vs Actual Demand …………………………………………………42

Figure 9. Main Category vs Actual Demand ………………………………………………..42

Figure 10. Main Category vs Discount Demand ……………………………………………43

Figure 11. Main Category vs Selling Demand ………………………………………………43

Figure 12. HeatMap …………………………………………………………………………44

Figure 13. Regplot for Actual and Selling Price …………………………………………….45

Figure 14. Regplot for Discount and Selling Price ………………………………………….45

Figure 15. Regplot for Ratings and Selling Price ……………………………………………46

Figure 16. Ratings vs No.of Ratings ………………………………………………………...46

Figure 17. Regplot for No. of Ratings and Selling Price ……………………………………46

Figure 18. Train Results for Ensemble Methods …………………………………………….48

Figure 19. Test Results for Ensemble Methods ……………………………………………...48

**LIST OF ABBREVATIONS**

|  |  |
| --- | --- |
| Short Forms | Description |
| ML | Machine Learning |
| ARIMA | Autoregressive Integrated Moving Average |
| LSTM | Long Short-Term Memory |
| RNN | Recurrent Neural Network |
| RMSE | Root Mean Squared Error |
| MSE | Mean Squared Error |
| MAE | Mean Absolute Error |
| CNN | Convolutional Neural Networks |
| SVM | Support Vector Machine |
| Bi LSTM | Bi-directional Long Short-Term Memory |
| Conv LSTM | Convolutional Long Short Term Memory |
| FNN | Feed Forward Neural Network |

**CHAPTER 1**

**INTRODUCTION**

Demand forecasting is a crucial aspect of ecommerce businesses, as it helps them predict the demand for their products and optimize their supply chain operations (Aamer et al., 2020). (Cai et al., 2021), The dynamic and complex business environment has posed significant challenges for business decision-making (Ren, Choi, et al., 2020). Machine learning (ML) has emerged as a powerful tool for demand forecasting, as it improves accuracy, handles complex datasets, and automates the process. ML models can be developed using various approaches such as Regression models, Gradient Boosting, FNN, (Qi et al., 2019) and LSTM. The choice of model depends on several factors such as business goals, data type, data amount and quality, and forecasting period. The implementation of ML-based demand forecasting can reduce operational costs, optimize inventory, and improve financial planning (Feizabadi, 2022). In this context, demand forecasting can help businesses predict their supply chain costs and accurately plan for seasonal fluctuations (Leung et al., 2020) (Khan et al., 2020)

* 1. **Background of the Study**

E-commerce, a rapidly growing sector, has transformed the way consumers shop, presenting both opportunities and challenges for businesses operating in this space. One of the critical challenges faced by e-commerce enterprises is the accurate prediction of product demand. The traditional methods of demand forecasting, relying on historical sales data and statistical models, often prove inadequate in capturing the complexities inherent in the dynamic e-commerce environment.

The e-commerce landscape is characterized by constantly shifting consumer preferences, fleeting trends, and a multitude of influencing factors such as promotions, discounts, and seasonality. As a result, conventional forecasting models struggle to adapt to the non-linear and unpredictable nature of consumer behavior. Inaccurate demand forecasts lead to suboptimal inventory levels, stockouts, and excess holding costs, all of which can have detrimental effects on a company's bottom line and customer satisfaction.

Recognizing the limitations of traditional forecasting approaches, there is a growing interest in exploring advanced data-driven techniques, particularly machine learning, to address the intricacies of demand forecasting in the e-commerce domain. Machine learning algorithms, with their ability to analyze vast datasets and discern patterns, offer the potential to uncover hidden insights and relationships within the data that may elude conventional methods.

Several studies have demonstrated the effectiveness of machine learning in demand forecasting across various industries, but its application within the e-commerce context is an area that requires deeper exploration. The unique challenges posed by the e-commerce sector, such as high product diversity, rapid product life cycles, and the influence of external factors like social media and online reviews, necessitate tailored approaches.

This research seeks to build upon existing knowledge by investigating the applicability and efficacy of machine learning algorithms in forecasting demand for e-commerce products. By combining historical sales data with additional relevant features, such as customer behavior and external market trends, the goal is to develop a robust forecasting framework that not only improves accuracy but also adapts dynamically to the ever-changing e-commerce landscape.

In summary, the intersection of e-commerce and machine learning presents an exciting opportunity to revolutionize demand forecasting practices. By addressing the challenges specific to the e-commerce sector, this research aims to contribute valuable insights that can empower businesses to make more informed decisions, enhance operational efficiency, and ultimately thrive in the competitive e-commerce market.

* 1. **Problem Statement**

The e-commerce industry's rapid growth has brought forth a multitude of challenges, among which accurate demand forecasting stands out as a critical factor influencing operational efficiency, customer satisfaction, and overall business success. Traditional demand forecasting methods, relying on historical data and statistical models, often fall short in capturing the dynamic and non-linear nature of consumer behavior, leading to suboptimal inventory management, increased holding costs, and potential revenue loss due to stockouts or overstock situations.

In the context of e-commerce, where product assortments are vast, consumer preferences are diverse and rapidly evolving, and external factors like promotions and seasonality play significant roles, the limitations of traditional forecasting become pronounced. The need for a forecasting approach that can adapt to these complexities and provide more accurate predictions is evident. Machine learning, with its capacity to analyze large datasets and discern complex patterns, emerges as a promising solution.

However, while there is a growing body of literature on demand forecasting using machine learning, the specific challenges and opportunities within the e-commerce sector warrant dedicated investigation. The unique characteristics of e-commerce, such as the short product life cycles, high product diversity, and the impact of online reviews and social media, require tailored machine learning models and methodologies.

Therefore, the problem at hand is twofold: first, the inadequacy of traditional forecasting methods in the e-commerce context, and second, the need for a specialized machine learning-based approach that can effectively address the challenges posed by the e-commerce environment. This research aims to bridge these gaps by developing and validating a robust machine learning framework for demand forecasting in e-commerce, with the ultimate goal of empowering businesses to make more informed decisions, optimize inventory levels, and enhance their competitiveness in the ever-evolving digital marketplace.

* 1. **Aims and Objectives**

The main aim of this research is to develop accurate and effective demand forecasting models using machine learning techniques for e-commerce products, enabling businesses to optimize inventory management and enhance operational efficiency.

The research objectives are formulated based on the aim of this study which are as follows:

* Investigate the effect of different product categories and sub-categories on pricing and discounts.
* Investigate the relationship between pricing strategies and the likelihood of a product being sold.
* Utilize Machine Learning techniques to develop a predictive model for estimating product demand and sales based on the provided attributes.
* Evaluate and compare the performance of different machine learning algorithms, such as regression models, and neural networks, in predicting demand for ecommerce products.
* Identify potential areas for future research and improvement in demand forecasting techniques for ecommerce, such as incorporating customer behavior analysis, sentiment analysis, or external data sources for more accurate predictions.
  1. **Research Questions**

Some research questions are given below-

* **RQ1:** How does the brand of a product influence its demand?
* **RQ2:** How the relationship between actual price, selling price and demand can be leveraged to make accurate demand forecasts?
* **RQ3:** How do product rating and number of ratings impact demand?
* **RQ4:** What role does the presence of a discount play in influencing demand?
* **RQ5:** What are the most suitable machine learning algorithms for accurate demand forecasting of diverse product categories within the e-commerce industry?
  1. **Scope of the Study**

The research mainly focuses on demand forecasting for e-commerce products. In scope of this research, we need to check the availability of data, the time, and resources available for this study. Identifying and creating relevant features from the collected data to enhance the predictive power of the models. Implementing and training the ensemble techniques using the pre-processed data to forecast the demand for e-commerce products. Evaluating and comparing the performance of different ensemble methods. Defining appropriate evaluation metrics to measure the accuracy and robustness of the ensemble techniques in demand forecasting. This could include metrics like Mean Squared Error, Root Mean Squared Error.

The out of scope for this research would be conducting a survey of e-commerce businesses to understand their demand forecasting needs. We will not use the traditional time series methods like ARIMA and Exponential Smoothing. While model performance is being evaluated, in-depth hyperparameter tuning for each model won’t be the focus of the study. Basic hyperparameter tuning will be used for a comparative analysis.

* 1. **Significance of the Study**

This research might aid in the importance of forecasting and prediction methods, and it could also have an impact on efforts to maximise the demand for e-commerce products based on different factors.

The findings of this research might lead to enhanced demand prediction accuracy, customer behaviour insights, optimized pricing and discounting strategies, supply chain and inventory management.

* By understanding the demand for different products, businesses can optimize their inventory levels and reduce costs.
* By forecasting demand, businesses can set process that are more likely to be profitable.
* By understanding the factors that influence demand, businesses can target their marketing campaigns more effectively.

The research has the potential to revolutionize demand forecasting in e-commerce by harnessing machine learning techniques on a dataset rich with attributes that capture brand dynamics, pricing information and many more. This can lead to informed decisions, improved customer satisfaction and greater operational efficiency within the e-commerce landscape.

* 1. **Structure of the Study**

The research study follows a structured framework to comprehensively address key components of the research. In the introduction, the background of the e-commerce industry is presented, highlighting the challenges in demand forecasting. The problem statement elucidates the limitations of traditional methods, paving the way for the research questions and objectives. The literature review explores existing demand forecasting methodologies in e-commerce and delves into the applications of machine learning, identifying gaps in current research. The theoretical framework establishes the foundational principles of demand forecasting and machine learning, laying the groundwork for the development of a tailored forecasting framework. The methodology section outlines the data collection process, model selection, and experimental design. Data analysis and results delve into the implementation of machine learning models, performance evaluation, and comparative analyses. The discussion section interprets results, discusses implications for e-commerce, addresses challenges, and suggests avenues for future research. The conclusion summarizes findings, highlights contributions, and provides practical recommendations. The study aims to contribute valuable insights to academia and offer actionable strategies for enhancing demand forecasting in the dynamic e-commerce landscape.

In the Introduction, the contextual background of the e-commerce industry is established, emphasizing the dynamic nature of consumer behavior and the consequential challenges in accurately forecasting product demand. The problem statement underscores the limitations of traditional forecasting methods within this complex environment, setting the stage for the study’s objectives. Research questions are formulated to guide the investigation, and the significance of the research is highlighted to underscore its potential impact on the operational efficiency of e-commerce businesses.

The literature review critically examines existing demand forecasting techniques in e-commerce, shedding light on both traditional and emerging methodologies. A comprehensive analysis of machine learning applications in demand forecasting is undertaken, identifying gaps in the current body of knowledge. This section serves as the theoretical foundation for the study, elucidating the principles of demand forecasting and the application of machine learning models within the specific context of the e-commerce sector.

The methodology section outlines the research approach, detailing the data collection process, machine learning model selection, and the rationale behind these choices. It also provides insights into feature engineering and the integration of external factors influencing demand. This section is crucial in ensuring the replicability and rigor of the research. The subsequent data analysis and results section delves into the practical implementation of machine learning models, evaluating their performance and conducting a comparative analysis with traditional forecasting methods. The interpretation of results provides valuable insights into the effectiveness of the machine learning framework and its implications for e-commerce businesses.

Moving to the discussion section, the findings are contextualized within the existing literature, drawing comparisons with previous studies and offering insights into the practical implications for businesses. The challenges and limitations encountered during the study are candidly addressed, providing a holistic view of the research’s scope and constraints. Opportunities for future research are explored, guiding scholars and practitioners toward potential areas of advancement in the field.

The conclusion encapsulates the study’s key findings, emphasizing its contributions to both academic knowledge and practical applications. The research’s significance in improving demand forecasting in the e-commerce industry is re-emphasized, and practical recommendations are offered for businesses seeking to implement machine learning-based forecasting models. This structured framework aims to provide a cohesive narrative, ensuring a comprehensive understanding of the research’s methodology, findings, and broader implications.

**CHAPTER 2**

**LITERATURE REVIEW**

The literature review serves as a comprehensive survey of existing scholarly works relevant to the research topic. It offers a critical analysis of previous studies, identifying gaps, trends, and debates in the field. This section not only establishes the context for the current research but also provides a foundation for theoretical and conceptual frameworks. By synthesizing and evaluating existing knowledge, the literature review aids in shaping the research questions and methodology, ultimately contributing to the advancement of the academic discourse within the chosen subject area.

**2.1 Introduction**

In the realm of e-commerce, demand forecasting is a pivotal aspect that influences inventory management and operational efficiency. Traditional forecasting methods, reliant on historical data and statistical models, face challenges in the dynamic e-commerce environment characterized by rapidly changing consumer preferences and diverse product offerings. These limitations have spurred interest in leveraging machine learning for demand forecasting. Studies highlight the potential of machine learning algorithms, such as regression models and neural networks, to improve accuracy by capturing intricate patterns and adapting to the dynamic nature of the e-commerce sector. However, the unique challenges of e-commerce, including short product life cycles and high product diversity, necessitate a tailored approach. This study contributes to this evolving field by developing and validating a machine learning-based framework specifically designed to address the complexities of demand forecasting in the context of e-commerce products.

**2.2 Traditional Demand Forecasting Methods**

Traditional demand forecasting methods have long been employed as foundational tools in various industries, including e-commerce. These methods typically rely on historical sales data and statistical models to project future demand patterns. Common approaches encompass time series analysis, moving averages, and exponential smoothing. While these methods provide a baseline for forecasting, they often encounter challenges in capturing the complexity of the e-commerce environment. The limitations arise from difficulties in adapting to rapidly changing consumer behaviors, handling diverse product assortments, and incorporating the influence of external factors. As the e-commerce sector continues to evolve, there is a growing recognition of the need for more dynamic and adaptable forecasting approaches, prompting a shift towards the exploration of advanced techniques, notably machine learning algorithms.

**2.3 Limitations of Traditional Methods in E-commerce**

Traditional demand forecasting methods face notable limitations when applied to the dynamic landscape of e-commerce. These methods, rooted in historical data and statistical models, struggle to adapt swiftly to the ever-changing consumer preferences and diverse product assortments inherent to e-commerce platforms. The short product life cycles, rapid introduction of new products, and the influence of external variables, such as promotions and seasonality, pose challenges for traditional methods. The conventional approaches often fall short in capturing the intricate patterns and nuances of demand fluctuations, leading to suboptimal forecasting accuracy. These limitations underscore the need for more flexible and sophisticated techniques, prompting an exploration of advanced methodologies, particularly machine learning, to enhance the efficacy of demand forecasting in the e-commerce sector.

**2.4 Machine Learning in Demand Forecasting**

Machine learning (ML) has emerged as a transformative approach to demand forecasting, offering unparalleled adaptability and accuracy in the e-commerce domain. Unlike traditional methods, ML leverages algorithms that can dynamically analyze vast datasets, discern complex patterns, and adapt to changing market conditions.

Regression models, such as linear regression and logistic regression, have been utilized to model the relationship between demand and various factors, such as product attributes, pricing, and promotions. These models have shown promise in capturing nonlinear relationships and improving forecasting accuracy.

Decision trees and random forests have also been employed in demand forecasting for ecommerce. These algorithms are effective in handling categorical variables and interactions among features, providing insights into the important factors influencing demand. Support vector machines (SVMs) have been used to handle both linear and nonlinear relationships in demand data and have demonstrated good performance in forecasting demand for ecommerce products.

Neural networks, including feedforward neural networks and recurrent neural networks (RNNs), have gained significant attention in demand forecasting for ecommerce. These algorithms can capture complex patterns and dependencies in the data, making them well-suited for modeling the dynamics of demand. RNNs, in particular, have shown promise in capturing temporal dependencies and seasonality in ecommerce demand data.

**2.5 Applications of Machine Learning in E-commerce**

Machine learning (ML) has found diverse applications in revolutionizing various facets of the e-commerce landscape, contributing significantly to operational efficiency and customer experience. In the realm of demand forecasting, ML algorithms play a pivotal role in enhancing accuracy and adaptability. Beyond demand forecasting, the applications of ML in e-commerce encompass:

1. **Personalized Recommendations:** ML algorithms analyze user behavior, purchase history, and preferences to deliver personalized product recommendations, contributing to increased customer engagement and satisfaction.
2. **Fraud Detection and Security:** ML models detect anomalous patterns and behaviors to enhance fraud detection and strengthen the security of online transactions, safeguarding both consumers and e-commerce platforms.
3. **Dynamic Pricing Strategies:** ML enables e-commerce businesses to implement dynamic pricing strategies by analyzing market conditions, competitor pricing, and customer behavior, optimizing pricing for competitiveness and profitability.
4. **Supply Chain Optimization:** ML assists in optimizing supply chain processes by predicting demand patterns, improving inventory management, and streamlining logistics for more efficient and cost-effective operations.
5. **Chatbots and Customer Service:** Natural Language Processing (NLP) powered by ML enables the development of intelligent chatbots, enhancing customer service by providing instant responses, addressing queries, and facilitating smoother interactions.
6. **Image and Voice Recognition:** ML applications include image and voice recognition technologies, allowing for visual and voice search functionalities, simplifying the user experience and product discovery.
7. **Customer Segmentation and Targeting:** ML analyzes customer data to identify distinct segments, enabling targeted marketing campaigns and personalized communication strategies tailored to specific customer groups.
8. **Predictive Analytics for Marketing:** ML models predict consumer behavior, allowing e-commerce platforms to optimize marketing campaigns, allocate resources effectively, and improve the overall return on investment in marketing efforts.
9. **Enhanced Search Capabilities:** ML-driven search algorithms improve the accuracy of product searches by understanding user intent, leading to more relevant search results and facilitating quicker and more satisfying user experiences.
10. **Inventory Management and Forecasting:** Beyond demand forecasting, ML aids in inventory management by predicting stockouts, optimizing reorder points, and minimizing excess inventory, contributing to cost reduction and improved resource allocation.

**2.6 Challenges and Opportunities in E-commerce Demand Forecasting**

Recognizing and addressing these challenges while capitalizing on the opportunities presented by machine learning in demand forecasting is crucial for advancing the effectiveness of e-commerce operations in an increasingly dynamic market.

1. **Dynamic Consumer Behavior:** E-commerce faces the challenge of rapidly evolving consumer preferences and behavior, making it difficult for traditional forecasting methods to adapt swiftly to changing trends.
2. **Short Product Life Cycles:** The e-commerce landscape often involves products with short life cycles. Forecasting accurately for products with limited market presence poses challenges for traditional models designed for stable, long-term trends.
3. **High Product Diversity:** E-commerce platforms offer a diverse range of products, each with its own demand patterns. Traditional methods struggle to handle the intricacies of forecasting demand for a vast and varied product assortment.
4. **Influence of External Factors:** Factors such as promotions, discounts, and seasonality significantly impact e-commerce demand. Incorporating these external variables into forecasting models requires sophisticated techniques not often present in traditional approaches.
5. **Data Quality and Integration:** E-commerce relies on data from various sources, and ensuring the quality and seamless integration of these diverse datasets for accurate forecasting poses a substantial challenge.

Below are few of the Opportunities in E-Commerce Demand Forecasting:

1. **Machine Learning Advancements:** The advancement of machine learning provides an opportunity to overcome traditional forecasting limitations. ML algorithms excel in adapting to dynamic environments, capturing complex patterns, and accommodating diverse product portfolios.
2. **Big Data Analytics:** The wealth of data generated by e-commerce platforms presents an opportunity for leveraging big data analytics. Analyzing vast datasets enables a more comprehensive understanding of consumer behavior and demand patterns.
3. **Real-time Data Processing:** Machine learning facilitates real-time data processing, allowing for instantaneous adjustments to demand forecasts based on the latest market trends and consumer behaviors.
4. **Personalization and Customer Segmentation:** ML algorithms enable personalized forecasting models tailored to individual customer segments. This personalization enhances the accuracy of predictions, especially in markets with distinct and varied consumer preferences.
5. **Integration of External Factors:** ML models offer the capability to seamlessly integrate external factors, such as promotions and seasonality, into forecasting algorithms, providing a more holistic and accurate prediction of demand fluctuations.
6. **Continuous Learning and Adaptability:** ML models are designed for continuous learning and adaptability, making them well-suited for environments where consumer behaviors and market dynamics evolve rapidly.
7. **Improved Inventory Management:** ML-driven demand forecasting contributes to more effective inventory management by preventing stockouts, minimizing excess inventory, and optimizing reorder points, ultimately reducing holding costs.
8. **Enhanced Decision-Making:** Accurate demand forecasting through machine learning empowers e-commerce businesses with more informed decision-making capabilities, enabling strategic planning, resource allocation, and marketing strategies based on reliable predictions.

**2.7 Related Research Publications**

(Vavliakis et al., 2021), In this research, they propose a sales forecasting system for e-commerce platforms using a hybrid model that combines ARIMA, LSTM, and XGBoost algorithms.They concluded that the proposed hybrid model outperformed the other models in terms of RMSE, MSE and MAE. The gaps observed in the study include the need for further testing of the proposed algorithm in real-world scenarios and the improvement of the simulation framework with additional features such as out-of-stock periods, web traffic sources, and customer profiles. The authors also mentioned the plan to compare their system with more sales forecasting models and datasets in future work.

(Hsieh, 2019), The purpose of the research was to explore different forecasting methods for predicting e-commerce sales during a price war. The study used data from a medical equipment seller on Amazon and analyzed sales volume data, number of views, and competitors' sales data. Three forecasting methods were employed: exponential smoothing, linear trend, and seasonal variation. The study concluded that the exponential smoothing method was the most appropriate for responding to a price war due to its ability to focus on recent observations while considering past characteristics. The study also identified the presence of seasonality in the data. However, the research observed a limitation in accessing social media interaction data, which could have improved the accuracy of the sales forecast. Additionally, the study suggested that incorporating public opinion analysis and other factors like inventory cost and delivery time could further enhance the forecasting model.

(Ji et al., 2019), The purpose of the research in the paper is to propose a three-stage forecasting model called the C-A-XGBoost model. The researchers aim to improve the accuracy of sales forecasting by combining clustering techniques with the XGBoost algorithm and it outperformed other models in terms of forecasting accuracy. The gaps observed in the paper include the lack of detailed explanation of the basic theory behind the algorithm and the limited exploration of parameter optimization. The researchers focus more on the procedures of the algorithm rather than the underlying theory.

(Pan & Zhou, 2020), The purpose of the study is to investigate the application of CNN in data mining and sales forecasting for e-commerce. The study aims to automatically extract effective features from structured time series data using CNN and to predict the sales volume of goods based on this data and it concluded that the improvement in accuracy of sales forecasting, and the potential for predicting the sales volume of commodities without requiring manual intervention. The study does not explicitly mention any limitations or gaps in the research. However, potential areas for improvement or further research could include the scalability and generalizability of the proposed method across different e-commerce platforms.

(Giri et al., 2019), The purpose of the study was to predict the quantity of new product using historical sales data of female apparel clothing and their corresponding images. The study aimed to achieve this by using a deep neural network V3 inception model for extracting the feature vectors of the images and then applying nonlinear Neural Network (MLP-Multilayer Perceptron) regression. The study found that the model performed quite well despite the small size of the dataset. However, the study also highlighted the limitation of using a smaller dataset and suggested that the model's performance could be further improved by increasing the number of trained images.

(Rai et al., 2019), The purpose of the study is to investigate demand prediction for e-commerce advertisements, specifically in the Consumer-to-Consumer (C2C) e-commerce market.The study also aims to compare the performance of state-of-the-art linear models, decision tree ensembles, and deep learning methods such as LSTM, Multi-Layer perceptron (MLP), Gated Recurrent Units (GRU) for demand prediction and concluded that deep learning models outperformed and one of the limitations in this study is they did not consider factors such as competing products, lost sales due to stockouts, customer feedback, and UI/UX features of websites, which could potentially affect users' perception about the advertisements and lead to better results.

(Ren, Chan, et al., 2020), The study aims to address the challenges faced by traditional forecasting methods and inventory planning in the fashion retail industry. Along with the traditional methods, they have also used artificial neural networks. Gaps observed in the paper include the need for further research on reducing noise and extracting useful data from big data, capturing customer needs to improve demand forecasting performance, and effectively utilizing user-generated information for demand forecasting in the fashion industry.

(Z. Li & Zhang, 2022), The study aimed to compare the performance of the proposed model with other models such as LGBM and LSTM using real sales data and they concluded that ConvLSTM model outperforme the LGBM and LSTM models.he study did not address potential challenges or obstacles in implementing the proposed ConvLSTM model in real-world e-commerce platforms.

(Swami et al., 2020), The purpose of the study is to predict the total sales for every product and store in the next month given the past data. The methods used in the study include eXtreme Gradient Boosting (XGBoost) and Long Short Term Memory (LSTM) based network architecture for learning tasks. The conclusions drawn from the study are that XGBoost fared better than LSTM over the dataset, attributed to its relatively higher sparsity.The gaps observed in the study include the need for a more rigorous approach for hyperparameter selections to tune the customized LSTM-based architecture.

(Zhu, 2021), The study aims to address the gap in existing research, which often overlooks the impact of consumer sentiment on sales prediction. They have used BiLSTM and CNN and the study cknowledges the need for improvement in the normalization methods used in feature engineering to address potential unbalanced weights of different attributes.

(Zohdi et al., 2022a), The purpose of the study was to investigate demand forecasting in supply chain management using machine learning algorithms based on customer information. The researchers aimed to compare the performance of different machine learning algorithms, including K-nearest neighbors, decision tree, gradient boosting, multi-layer perceptron, and extreme learning machine, in forecasting demand based on customer data. Extreme learning machine (ELM) and multi-layer perceptron (MLP) algorithms outperformed the other algorithms in terms of forecasting accuracy. The study did not explicitly mention any specific gaps. However, potential areas for further research or improvement could include the exploration of hybrid models combining different algorithms.

, The purpose of the study is to develop a machine learning and data mining-based method for predicting sales on online platforms. The study aims to construct a sales prediction model suitable for online products and evaluate the adaptability of the model in different types of online products. The conclusions drawn from the study are that the deep learning model, particularly the convolutional neural network (CNN) model, showed good prediction accuracy and generalization ability.The conclusions drawn from the study are that the deep learning model, particularly the convolutional neural network (CNN) model, showed good prediction accuracy and generalization ability.

(Kharfan et al., 2021), The purpose of the study was to develop a methodology for forecasting the demand of newly designed and launched seasonal products in the fashion retail industry. The study aimed to address the challenge of limited sales history for these new products by identifying significant features and building a forecasting model. The accuracy and bias of the proposed methodology for the validation and test sets. The results of the techniques used in the forecasting model were compared to assess their effectiveness. These gaps could include the need for more extensive validation of the proposed methodology, the exploration of alternative forecasting techniques, or the consideration of additional factors that may impact demand forecasting in the fashion retail industry.

**CHAPTER 3**

**RESEARCH METHODOLOGY**

The demand forecasting landscape within the e-commerce sector is characterized by dynamic consumer behaviors, diverse product assortments, and the need for real-time adaptability. As traditional forecasting methods encounter limitations in capturing the intricacies of this environment, the utilization of machine learning (ML) emerges as a promising solution (Kulshrestha & Saini, 2020). Before delving into the details of our research methodology, it is essential to establish the context and rationale guiding our approach.

1. Contextualizing Demand Forecasting in E-commerce: The e-commerce industry operates within a highly competitive and rapidly evolving landscape. Accurate demand forecasting is paramount for optimizing inventory management, ensuring customer satisfaction, and maximizing operational efficiency. Traditional methods, rooted in historical data analysis, face challenges in responding to the dynamic nature of consumer preferences and the vast array of products offered by e-commerce platforms.
2. Significance of Machine Learning in Demand Forecasting: Acknowledging the limitations of traditional approaches, our research adopts a machine learning-centric methodology. Machine learning models, equipped with the ability to analyze large datasets, discern intricate patterns, and adapt to changing market conditions, hold the potential to revolutionize demand forecasting in e-commerce. This shift aims to enhance the precision, adaptability, and real-time responsiveness required in the contemporary e-commerce landscape
3. Research Objectives: The primary objectives of this study are twofold: first, to assess the effectiveness of machine learning models in demand forecasting for e-commerce products, and second, to provide practical insights and guidelines for the integration of machine learning-based forecasting into e-commerce operational workflows. By addressing these objectives, we aspire to contribute valuable knowledge to the intersection of machine learning and e-commerce demand forecasting.

As we embark on the exploration of our research methodology, the subsequent section will detail the systematic approach undertaken to achieve these objectives. From data collection and model selection to performance evaluation and ethical considerations, each aspect of the methodology is designed to provide a comprehensive and actionable framework for advancing demand forecasting practices within the dynamic realm of e-commerce.

**3.1 Research Methodology**

The successful realization of accurate and adaptable demand forecasting within the e-commerce sector hinges on a meticulous and well-structured research methodology. In this section, we elucidate the systematic approach undertaken to address the intricacies of our research topic. The methodology is designed to navigate the challenges posed by dynamic consumer behaviors, diverse product assortments, and the need for real-time responsiveness inherent in the e-commerce landscape.

**3.1.1 Research Design**

Our study adopts a quantitative research design, leveraging historical data from e-commerce platforms to develop and evaluate machine learning models. This design facilitates a rigorous analysis of the relationships between various variables and enables the identification of patterns essential for accurate demand forecasting.

**3.1.2 Data Collection**

The data collection process is integral to our research methodology, focusing on gathering comprehensive information necessary for demand forecasting in the e-commerce domain using machine learning. Primary data sources include historical sales data, customer interactions, and detailed product attributes obtained from reputable e-commerce platforms. Historical sales data provides insights into past purchasing patterns, while customer interactions, encompassing browsing history and feedback, offer nuanced perspectives on consumer behavior. Detailed product attributes contribute to a granular understanding of each product. Supplementary data, such as promotional calendars and external events, is also incorporated to account for real-time influences on demand. By leveraging these diverse data sources, our aim is to create a robust dataset that captures the complexities of the e-commerce landscape, facilitating the development of accurate and adaptive machine learning models for demand forecasting.

* + 1. **Data Preprocessing**

Data preprocessing is a crucial step in our research methodology, ensuring the quality and uniformity of the dataset for effective demand forecasting in e-commerce using machine learning. The process involves rigorous cleaning to eliminate inaccuracies and outliers, imputation to handle missing values, and normalization to standardize the data. Historical sales data, customer interactions, and detailed product attributes are subjected to these preprocessing techniques to enhance the accuracy of machine learning models. By cleaning and standardizing the dataset, we aim to create a robust foundation that allows for meaningful insights into consumer behavior, product trends, and external influences, ultimately contributing to the precision of the demand forecasting models in the dynamic e-commerce environment.

* + 1. **Data Exploration and Visualization**

Data exploration and visualization form a pivotal phase in our research methodology, facilitating a comprehensive understanding of the e-commerce dataset for demand forecasting using machine learning. Leveraging statistical and graphical techniques, we delve into the intricacies of historical sales data, customer interactions, and product attributes. Descriptive statistics and visualizations, including histograms, scatter plots, and time series analyses, are employed to unravel patterns, trends, and potential outliers within the dataset. Exploratory techniques aid in identifying relationships between variables and discerning temporal dynamics, guiding subsequent model development. This visual exploration not only provides valuable insights into consumer behavior and product trends but also serves as a foundation for the selection and engineering of features critical to the accuracy of machine learning models in predicting demand within the dynamic e-commerce landscape.

* + 1. **Feature Selection**

Feature selection is a pivotal aspect of our research methodology, aimed at enhancing the precision of demand forecasting in e-commerce through machine learning. The process involves identifying and prioritizing relevant features from the dataset, including historical sales data, customer interactions, and product attributes. Techniques such as statistical tests, correlation analyses, and machine learning algorithms are employed to discern the most influential variables. By selecting key features that significantly impact demand patterns, we aim to streamline model complexity, mitigate overfitting, and improve the interpretability of the machine learning models. This strategic feature selection process ensures that the subsequent models focus on the most salient aspects of the data, optimizing their efficacy in capturing and predicting demand fluctuations within the dynamic e-commerce environment.

* + 1. **Model Selection**

Model selection is a critical component of our research methodology, seeking to optimize the accuracy and adaptability of demand forecasting models in the e-commerce domain using machine learning. Various machine learning models are considered, ranging from regression models to decision trees and neural networks, each chosen based on their ability to capture the complex patterns inherent in e-commerce demand data. The selection process involves evaluating model performance on historical sales data, customer interactions, and product attributes. By comparing and contrasting the strengths and weaknesses of different models, we aim to identify the most suitable algorithm for accurately predicting demand fluctuations in response to dynamic consumer behaviors and external influences within the e-commerce landscape. The chosen model is intended to offer a robust foundation for forecasting, providing insights that align with the unique challenges posed by the diverse and rapidly changing nature of e-commerce products.

* + 1. **Data Splitting**

Data splitting is a crucial step in our research methodology, employed to ensure robust model evaluation and generalizability in the context of demand forecasting for e-commerce products using machine learning. The dataset, comprising historical sales data, customer interactions, and product attributes, is divided into training and testing sets. The training set is used to train the machine learning models, allowing them to learn patterns and relationships within the data. The testing set, which the models have not seen during training, serves as an independent dataset to assess the models' performance and their ability to generalize to new, unseen data. This data splitting strategy is vital for avoiding overfitting, providing a realistic evaluation of model effectiveness, and ensuring the applicability of our machine learning approach to real-world scenarios within the dynamic e-commerce landscape.

* + 1. **Model Training**

Model training is a pivotal phase in our research methodology, essential for developing accurate demand forecasting models in the e-commerce domain using machine learning. Leveraging the carefully curated training dataset, which includes historical sales data, customer interactions, and detailed product attributes, the selected machine learning algorithms are systematically trained. During this process, the models learn to recognize patterns, relationships, and intricacies within the data. Iterative adjustments are made to model parameters to optimize performance, ensuring adaptability to the dynamic nature of the e-commerce landscape. By exposing the models to diverse scenarios and historical patterns, our aim is to enhance their predictive capabilities, enabling accurate and responsive forecasting of demand fluctuations in the rapidly evolving and complex e-commerce environment.

* + 1. **Model Evaluation**

Model evaluation is a critical component of our research methodology, serving as the yardstick for assessing the performance and reliability of machine learning models in the context of demand forecasting for e-commerce products. Using metrics such as Mean Asolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), the trained models are rigorously evaluated on a separate testing dataset, ensuring an unbiased assessment of their predictive accuracy. This evaluation phase gauges the models' ability to generalize to new, unseen data and provides insights into their robustness and adaptability. By comparing the performance of machine learning models against established benchmarks and, where applicable, traditional forecasting methods, we aim to validate the efficacy of our approach and identify the most effective model for accurate and dynamic demand forecasting within the e-commerce landscape.

* 1. **Proposed Methods**

The proposed methods in our research methodology encompass a comprehensive approach to demand forecasting for e-commerce products using machine learning. I employ a range of machine learning algorithms, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoosting, Support Vector Machine, Long Short Term Memory, and Feedforward Neural Networks.

* + 1. **Linear Regression**

Linear regression is a fundamental component of our research methodology for demand forecasting in e-commerce using machine learning. This regression model is employed to establish relationships between dependent variables, such as product demand, and independent variables, including historical sales data, customer interactions, and various product attributes. The model aims to quantify and understand the linear dependencies within the dataset, providing a baseline for predicting future demand based on observed patterns. By optimizing the regression parameters during the model training phase, we seek to develop a robust linear regression model that accurately captures the underlying trends in e-commerce demand, contributing to the overall accuracy and effectiveness of our machine learning-based forecasting approach.

* + 1. **Decision Tree Regressor**

The Decision Tree Regressor plays a pivotal role in our research methodology for demand forecasting in the e-commerce sector using machine learning. This algorithm is employed to model the complex relationships and decision-making processes inherent in the demand patterns of e-commerce products. Decision trees recursively split the dataset based on the most significant features, creating a tree-like structure that allows for the prediction of numerical outcomes, such as demand levels. By training the Decision Tree Regressor on historical sales data, customer interactions, and detailed product attributes, we aim to capture nonlinear dependencies and intricate patterns within the data. The flexibility of decision trees in handling diverse features makes them valuable for understanding the multifaceted nature of e-commerce demand, contributing to the accuracy and adaptability of our forecasting models.

* + 1. **Random Forest Regressor**

Random Forest Regressor serves as a powerful and versatile algorithm. Comprising an ensemble of decision trees, this model is adept at capturing complex relationships within the dataset, which includes historical sales data, customer interactions, and detailed product attributes. The Random Forest Regressor mitigates overfitting by aggregating predictions from multiple decision trees, resulting in a robust and accurate forecasting tool. Through the ensemble learning approach, it leverages the collective knowledge of individual trees to enhance predictive performance and adaptability to the dynamic nature of e-commerce demand. The model is carefully trained, tuned, and evaluated to ensure its efficacy in providing precise and responsive predictions for demand fluctuations in the diverse and rapidly changing landscape of e-commerce products.

* + 1. **XGBoost Regressor**

XGBoost Regression emerges as a key algorithm. Employed on a dataset comprising historical sales data, customer interactions, and detailed product attributes, XGBoost, an efficient gradient boosting framework, excels in capturing intricate patterns and dependencies within the data. By iteratively boosting weak learners, XGBoost enhances model performance and adaptability, making it well-suited for the dynamic nature of e-commerce demand. This methodology involves careful parameter tuning and model evaluation to harness the predictive power of XGBoost, aiming to provide accurate and responsive forecasts that contribute to a nuanced understanding of demand fluctuations in the complex landscape of e-commerce products.

* + 1. **Support Vector Machine**

Support Vector Machine (SVM) Regression plays a crucial role. Employed on a dataset that encompasses historical sales data, customer interactions, and detailed product attributes, SVM Regression seeks to establish a hyperplane that best captures the relationships within the data. By optimizing kernel functions and tuning parameters, this methodology aims to leverage the strengths of SVM in capturing complex patterns and non-linear dependencies. (Han et al., 2021)SVM can be summarized as to solve a quadratic programming problem, which tends to cause curse of dimensionality with the increase of training dataset size.The goal is to harness the predictive power of SVM Regression to enhance forecasting precision and offer valuable insights into the nuances of demand patterns in the e-commerce sector.

* + 1. **Long Short Term Memory**

Long Short-Term Memory (LSTM) networks are integral components. Employed on a dataset comprising historical sales data, customer interactions, and detailed product attributes, LSTM, a specialized type of recurrent neural network (RNN), excels in capturing sequential dependencies and long-term patterns within the time-series data inherent to e-commerce demand (J. Li et al., 2021). By leveraging memory cells and gates, LSTM can effectively learn and retain information over extended periods, making it well-suited for dynamic and evolving demand scenarios. (Bandara et al., 2019) LSTM exploits the non-linear demand relationships available in an E-commerce product assortment hierarchy. This approach aims to provide accurate and adaptive forecasting capabilities, contributing to a nuanced understanding of temporal dynamics and enhancing the precision of demand predictions within the intricate landscape of e-commerce products.

* + 1. **Feed Forward Neural Networks**

Feedforward Neural Networks (FNN), also known as multilayer perceptrons, are an integral part of our research methodology for demand forecasting in the e-commerce domain using machine learning. Comprising an input layer, one or more hidden layers, and an output layer, FNNs process information in a unidirectional manner, making them effective for capturing complex relationships within the dataset, including historical sales data, customer interactions, and detailed product attributes. The nodes (neurons) in each layer are connected through weighted edges, and activation functions introduce non-linearity to the model, enabling it to learn intricate patterns. Through careful design, including selecting the number of layers and nodes, choosing activation functions, and optimizing weights, FNNs aim to model the intricate demand patterns in the e-commerce landscape. This methodology involves training and evaluating the FNN to provide accurate and responsive demand forecasts, contributing to a nuanced understanding of consumer behaviors and product trends within the dynamic e-commerce environment.

* + 1. **Principal Component Analysis**

Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in data analysis and machine learning. Its primary objective is to transform a high-dimensional dataset into a lower-dimensional space while retaining the essential variance within the data. PCA achieves this by identifying the principal components, which are orthogonal vectors that capture the maximum variance in the original dataset. These components are ordered by the amount of variance they explain, allowing for the prioritization of information retention. By projecting data points onto these principal components, PCA facilitates the visualization of patterns and relationships within the data, aiding in simplifying complex datasets and mitigating the curse of dimensionality. This technique is particularly valuable for feature extraction, noise reduction, and understanding the underlying structure of data, making it a fundamental tool in exploratory data analysis and preprocessing for various analytical and machine learning tasks.

**3.3 Evaluation Metrics**

In this research, I have used Mean Squared Error, Root Mean Squared Error, and Accuracy to evaluate the performance of ensembled Machine learning model and neural networks.

**3.3.1 Mean Squared Error**

The Mean Squared Error (MSE) is a pivotal evaluation metric and it quantifies the average squared differences between predicted and actual demand values, providing a measure of the overall accuracy and precision of the forecasting models. By squaring the errors, MSE penalizes larger discrepancies more heavily, offering insights into the model's performance in capturing the magnitudes of deviations. As part of our methodology, minimizing MSE is a key objective during model training, ensuring that the developed machine learning models effectively reduce the squared errors and, consequently, enhance their ability to provide accurate and reliable predictions of demand fluctuations in the complex and dynamic landscape of e-commerce products.

* + 1. **Root Mean Squared Error**

Root Mean Squared Error (RMSE) serves as a crucial evaluation metric. RMSE is a variation of the Mean Squared Error (MSE) and represents the square root of the average squared differences between predicted and actual demand values. This metric provides a more interpretable scale by taking the square root, aligning with the original unit of the target variable. RMSE is particularly valuable in our methodology as it not only measures the overall accuracy of the forecasting models but also offers insights into the magnitude of errors, emphasizing the impact of larger discrepancies. Minimizing RMSE is a key goal during model training, ensuring that the developed machine learning models provide accurate and reliable predictions of demand fluctuations in the intricate and dynamic landscape of e-commerce products.

* + 1. **Mean Absolute Error**

The Mean Absolute Error (MAE) calculates the average absolute differences between predicted and actual demand values, providing a straightforward measure of the accuracy and precision of the forecasting models. As part of our methodology, minimizing MAE is a central objective during model training, ensuring that the developed machine learning models effectively reduce the magnitude of errors. This metric is particularly valuable for its ease of interpretation, representing the average absolute deviation between predicted and actual values. By emphasizing the absolute values of errors, MAE offers insights into the overall accuracy of the models in capturing demand fluctuations within the complex and dynamic landscape of e-commerce products.

* 1. **Summary**

In our research methodology for demand forecasting in the e-commerce domain using machine learning, a comprehensive and iterative approach is employed. The methodology encompasses the utilization of diverse algorithms, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, Support Vector Machine (SVM) Regression, Feedforward Neural Networks, and Long Short-Term Memory (LSTM) networks. The dataset, comprising historical sales data, customer interactions, and detailed product attributes, undergoes rigorous preprocessing, feature selection, and data splitting. Model training involves carefully chosen algorithms, and evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are employed for quantitative assessment. The proposed methods aim to capture complex demand patterns, with LSTM networks emphasizing sequential dependencies, and feedforward neural networks and traditional regression models focusing on capturing intricate relationships. The research methodology prioritizes model accuracy and adaptability to the dynamic nature of e-commerce demand, offering a nuanced understanding of consumer behaviors and product trends within this complex landscape.

**CHAPTER 4**

**ANALYSIS**

Analysis marks a pivotal juncture where the culmination of our work is rigorously examined and interpreted. Having traversed the intricacies of data collection, preprocessing, and the implementation of various machine learning algorithms, this phase encapsulates a comprehensive evaluation of the forecasting models. The extensive array of algorithms employed, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, Support Vector Machine (SVM) Regression, Feedforward Neural Networks, and Long Short-Term Memory (LSTM) networks, has been meticulously selected to capture the diverse and dynamic nature of demand patterns in the e-commerce landscape. The evaluation is grounded in a suite of performance metrics, prominently featuring Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE), which collectively provide a quantitative assessment of the accuracy and precision of our models. This analysis not only dissects the individual performance of each algorithm but also undertakes a comparative study to discern the most effective models for predicting demand across a spectrum of e-commerce products. As we delve into this phase, we aim not only to unravel the predictive capabilities of our models but also to extract actionable insights that can empower stakeholders in the e-commerce domain to make informed decisions based on reliable demand forecasts.

**4.1 Dataset Description**

The dataset utilized in this research project is sourced from Kaggle and is titled "Amazon Products Sales Dataset 2023". Comprising a comprehensive collection of product information from Amazon, this dataset is a rich resource for investigating and forecasting demand patterns in the e-commerce domain. The dataset includes a diverse array of product attributes, such as product titles, descriptions, ratings, and pricing information. It is worth noting that all data in this dataset is anonymized and does not contain any personally identifiable information (PII). This dataset consists of 551585 rows with 9 columns. All the columns in dataset are of object datatype. Numerical column is of object datatype due to the INR symbol.

Dataset Link: <https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset>

Please find the below detailed dataset description used in this project –

1. name – The Name of the Product
2. main\_category – The Main Category of the Product belongs to
3. sub\_category – The Sub Category of the Product belongs to
4. image – The image of the product looks like
5. link – Product link from the amazon website
6. ratings – The ratings given by amazon customers to a particualar product
7. no\_of\_ratings – The number of ratings given to a particular product in amazon shopping
8. actual\_price – The actual price of the product
9. discount\_price – The discount price of the product

**4.2 Data Preparation**

Data Preparation is a crucial step in the research, which includes data exploration, data cleaning, feature engineering, data transformation, and data splitting.

**4.2.1 Elimination of Variables**

Once the dataset is collected, exploratory data analysis has been done on Amazon Products Sales Dataset 2023 to understand the structure and characterisitics of the dataset. As per the research, I have dropped few features where it has unique values and it’s not required for the prediction. Please find the below features –

1. Image
2. Link

**4.2.2 Transformation in to Numerical Variables**

In the dataset, every feature is of object type. So, I have converted few columns such as ratings, no\_of\_ratings, actual\_price, and discount price to numerical columns.

1. Ratings and no\_of\_ratings columns contains in-appropriate values. Hence, I have replaced them with zeros as each and every product may or may not contain rating. (e.g. If product launched was new and in such cases, we cannot expect ratings) and converted ratings column to float and no\_of\_ratings column to int.
2. Actual price and discount\_price columns contains values in the format (₹20,000). Hence, I have removed those and converted it to int.

**4.2.3 Rename the Columns**

I have renamed few features to be specific as I have removed the INR symbol from the actual price and discount price columns. Hence, I renamed them to actual\_price\_in\_INR and discount\_price\_in\_INR respectively.

**4.2.4 Creating New Features**

From the dataset, we have a feature called name in which it contains the title of the product. I have created a new feature called “ brand” out of name column by fetching the first word of the entire string.

I have created new feature called "selling\_price\_in\_INR” by taking the difference of actual\_price\_in\_INR and discount\_price\_in\_INR columns.

Also created another new features called “actual\_demand”, “discount\_demand”, “selling\_demand”. Actual demand was calculated by taking the min value, 25% value, 75% value of actual\_price\_in\_INR by grouping it with sub\_category and brand. So, that we will be able to know for a particular sub category, this brand contains the pricing range. If the actual price is greater than or equal to min and less than or equal to 25% value then I considered that as low demand. If the actual price is greater than 25% value and less than 75% value then I considered that as moderate demand and if the actual price is greater than 75% values then its high demand based on the sub category and brand. This calculation is just to get clarity on what pricing basis, products is being sold. Similarly, I calculated it for discount\_demand and selling\_demand based on the discount\_price\_in\_INR and selling\_price\_in\_INR respectively.

**4.2.5 Handling of Missing Values**

Missing Values has been identified in the dataset for ratings, no\_of\_ratings, actual\_price\_in\_INR, discount\_price\_in\_INR columns.

Ratings column has been imputed with the median value based on sub\_category and brand. Rows with null values in actual\_price\_in\_INR column has been removed as the product can’t be sold if the price was not provided. Discount\_price\_in\_INR column has been imputed with zeros as the discount price was not compulsory for each and every product.

**4.2.6 Univariate Analysis**

Univariate analysis is a statistical method that involves the analysis of a single variable (or univariate data) at a time. It is the simplest form of statistical analysis and provides insights into the distribution, central tendency, and variability of a single variable. Univariate analysis is often the first step in the exploratory data analysis process and helps to understand the characteristics of individual variables.

For all the numerical variables in the dataset, box plot has been visually represented the distribution of data and highlighted outliers, median, and quartiles.

For all the categorical variables in the dataset, count plots and bar plots have been visualized.

**4.2.7 Bivariate Analysis**

Bivariate analysis involves the simultaneous analysis of two variables to explore the relationships between them. It is a step beyond univariate analysis, allowing researchers to understand how two variables vary in relation to each other. Bivariate analysis is crucial for uncovering patterns, associations, and dependencies between pairs of variables.

Scatter plots, Heatmap, and box plots have visualized for the numerical and categorical variables.

**4.3 Data Visualization**

Data visualization is the graphical representation of data to help uncover patterns, trends, and insights that may not be immediately apparent in raw data. It plays a crucial role in exploratory data analysis, communication of findings, and decision-making.

**4.3.1 Box Plot**

A box plot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It provides a visual summary of key statistics, including the median, quartiles, and potential outliers.

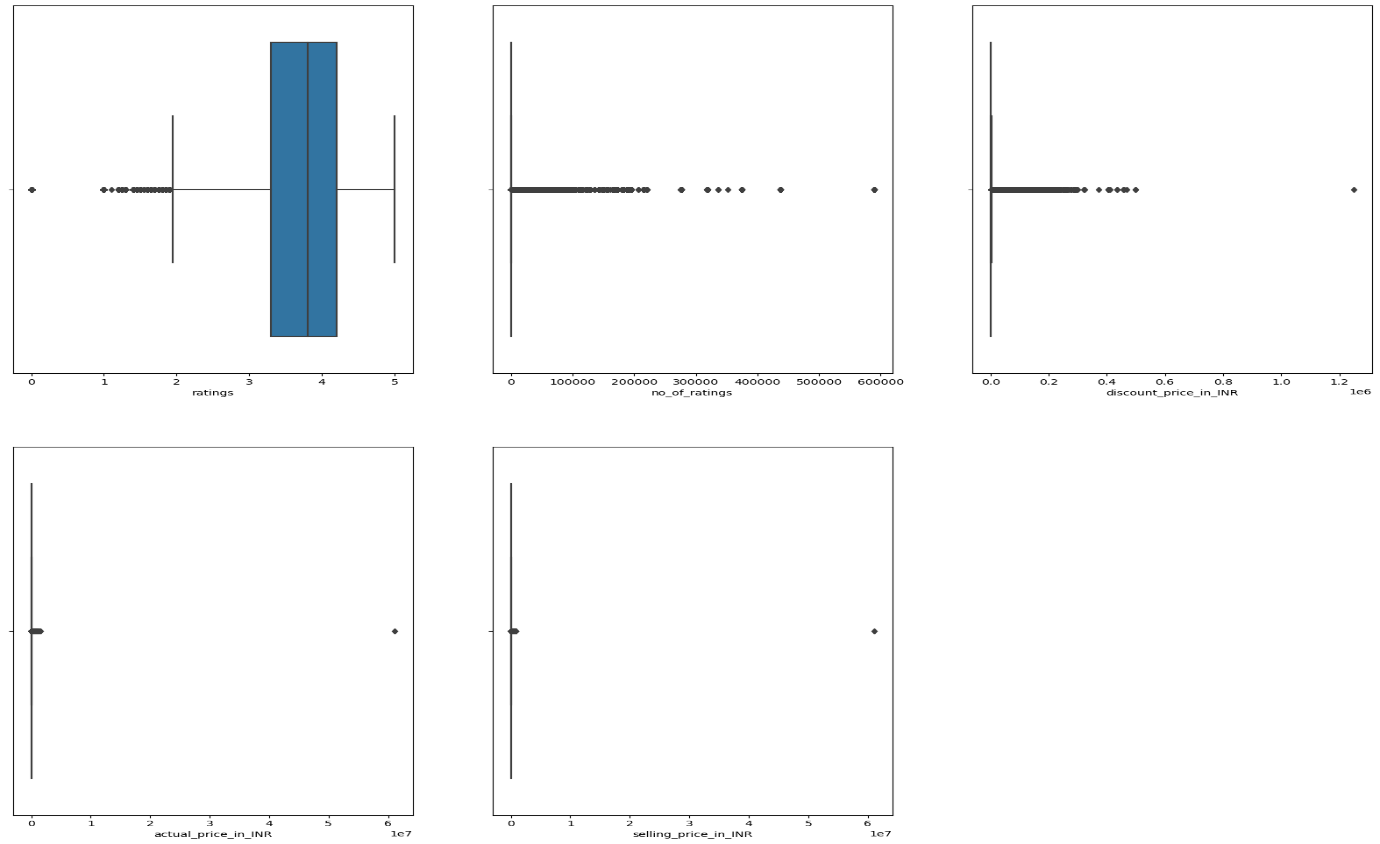


Figure 1. Box Plot with outliers

The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The length of the box indicates the spread of the middle 50% of the data. The line inside the box represents the median, which is the middle value of the dataset when it is sorted in ascending order. The whiskers extend from the box to the minimum and maximum values within a certain range. The range is typically calculated as 1.5 times the IQR. Individual data points beyond the whiskers are considered outliers and are plotted individually. Box plots are useful for identifying the central tendency and spread of a dataset, as well as detecting potential outliers.

A group of blue rectangular objects

Description automatically generated

Figure 2. Box Plot without outliers

**4.3.2 Bar Graph**

A bar graph (or bar chart) is a visual representation of data where individual bars or columns represent different categories or groups. The length of each bar is proportional to the value it represents. Bar graphs are effective for comparing the magnitudes of different categories or showing how a single category changes over time.

The primary elements of a bar graph are the bars or columns, each representing a category or group. The x-axis typically represents the categories or groups, while the y-axis represents the values or frequencies.The length or height of each bar corresponds to the value it represents. Longer bars indicate larger values.Labels along the x-axis identify the categories or groups being compared. These labels are often textual and provide context to the data.On the y-axis, numerical values may be labeled to indicate the scale or magnitude of the data.

In the below figure, which main category has the highest selling price has been visualized through bar plot.

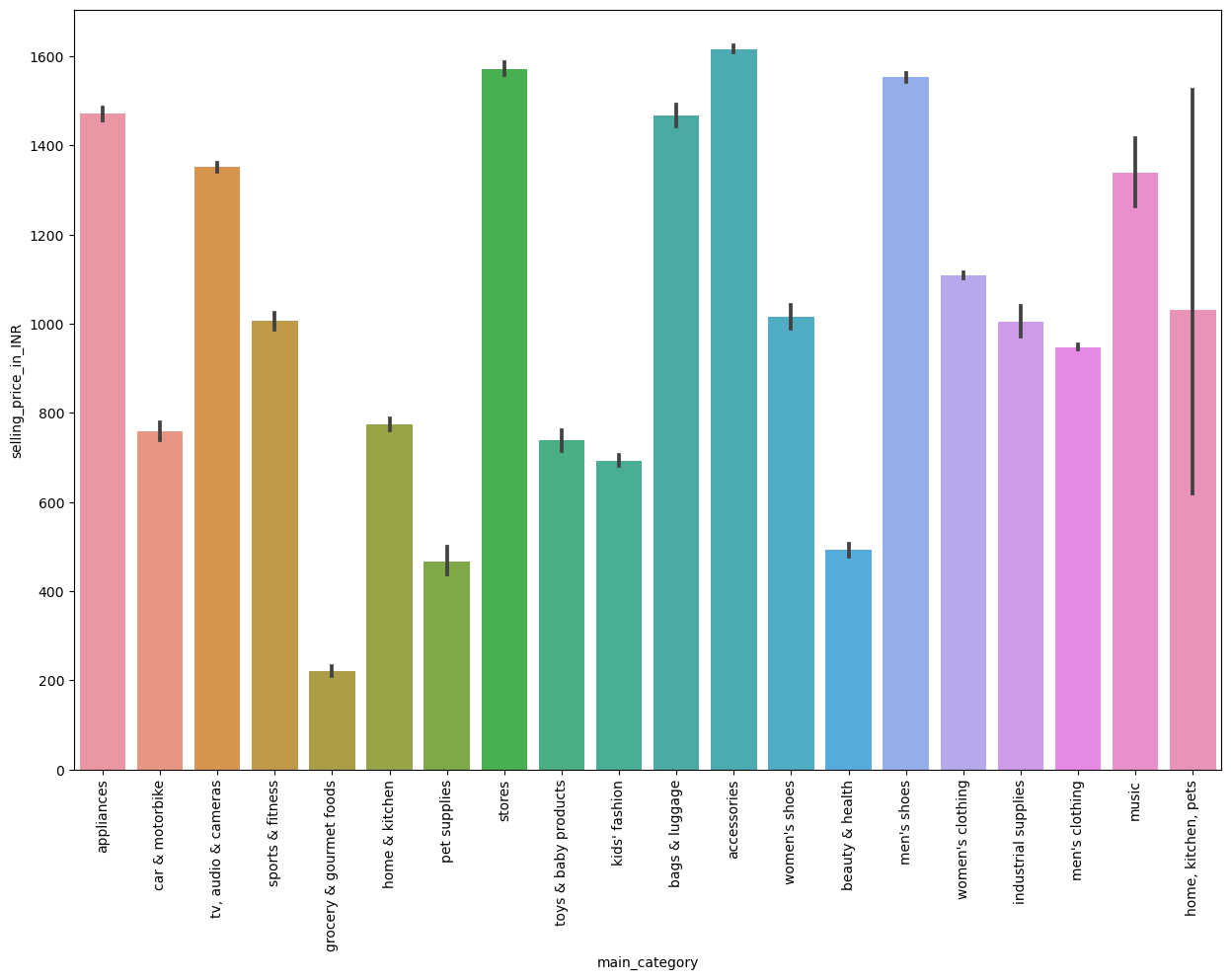


Figure 3. Main Category vs Selling Price

In the below figure, which sub category has the highest selling price has been visualized through bar plot.

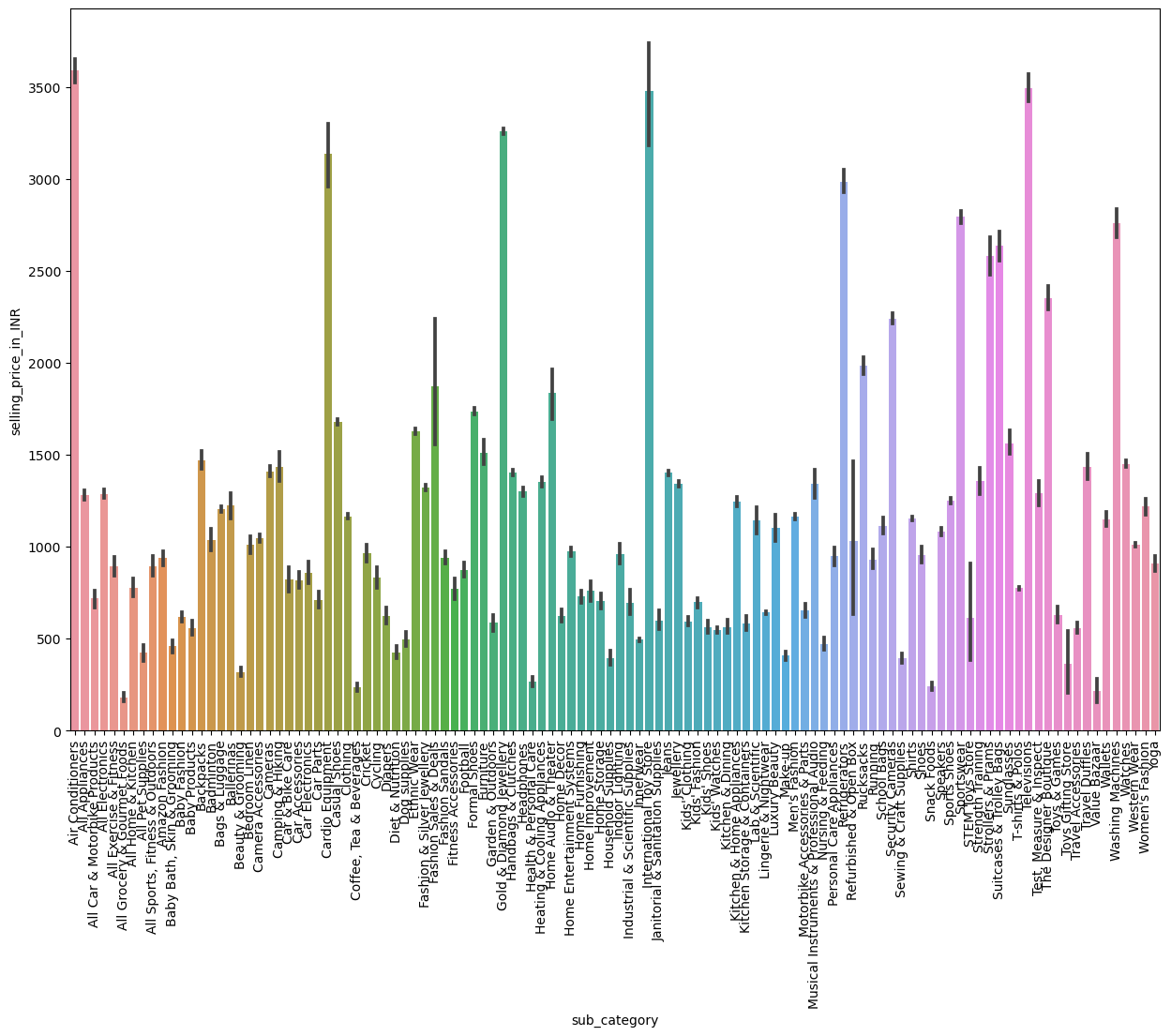


Figure 4. Sub Category vs Selling Price

In the below figure, it shows the top ten brands that have been visualized using bar graph.

A graph of different colored bars

Description automatically generated

Figure 5. Top Ten Brands in Amazon

**4.3.3 Scatter Plot**

A scatter plot is a type of data visualization that displays individual data points on a two-dimensional graph. Each point represents the values of two variables, and the position of the point is determined by the values of those variables. Scatter plots are useful for visually identifying relationships or patterns between two continuous variables.

Below figures are based on the calculated demand for each sub category and brand in terms of actual price and discount price respectively vs selling price.



Figure 6. Actual Price vs Selling Price vs Actual Demand



Figure 7. Discount Price vs Selling Price vs Discount Demand

**4.3.4 Count Plot**

A count plot is a type of bar plot that displays the counts of unique values or categories in a dataset. It is particularly useful for visualizing the distribution of categorical variables and identifying the relative frequencies of different categories in the dataset.

A line of colored lines

Description automatically generated with medium confidence

Figure 8. Sub Category vs Actual Demand

A graph of different colored bars

Description automatically generated

Figure 9. Main Category vs Actual Demand

For the figures shown for count plot, as I have already calculated the demand based on actual price, discount price, and selling price. Now I would like to visualize it through count plot in terms of main category vs all demands.

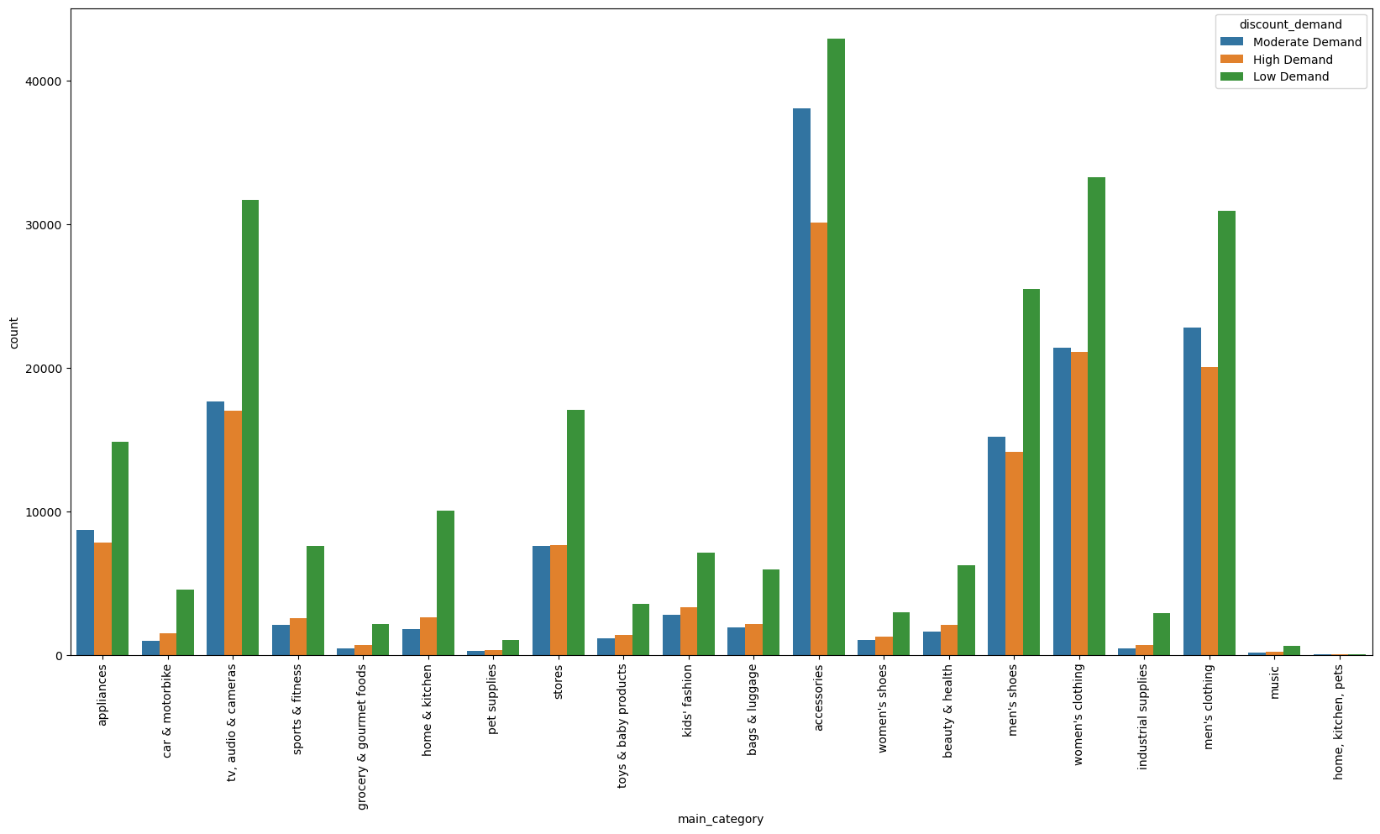


Figure 10. Main Category vs Discount Demand

A graph of different colored bars

Description automatically generated

Figure 11. Main Category vs Selling Demand

**4.3.5 Heatmap**

A heatmap is a graphical representation of data where values in a matrix are represented as colors. They are commonly used to visualize the correlation between variables, the intensity of a phenomenon across two dimensions. Heatmaps are versatile and can be used for various purposes beyond correlation matrices. They are especially effective for visualizing patterns and structures in large datasets.

A screenshot of a color chart

Description automatically generated

Figure 12. HeatMap

**4.3.6 Regplot**

Regplot is used to create a scatter plot with a linear regression fit. It combines a scatter plot with a regression line to help visualize the relationship between two variables and estimate a linear trend.

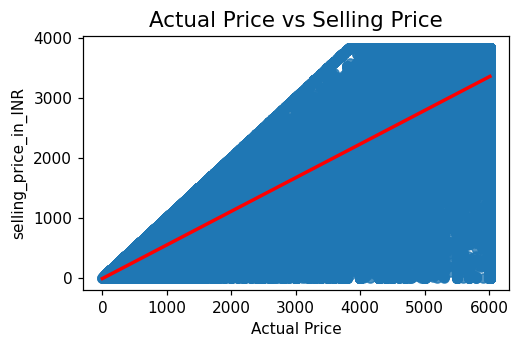


Figure 13. Regplot for Actual and Selling Price

From the below figure, it clearly says that if discount price increases then selling price definitely increases.

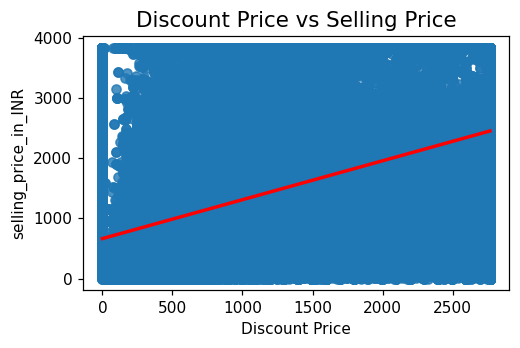


Figure 14. Regplot for Discount and Selling Price

From the below figure 15, it says that selling price does not vary based on ratings. From the below figure 16, regplot says that increase in ratings will definitely increase the number of ratings.

A blue graph with a red line

Description automatically generated A graph with a red line

Description automatically generated

Figure 15. Regplot for Ratings and Selling Price Figure 16. Ratings vs No.of Ratings

From the below figure, it says that selling price does not vary with the number of ratings.

A blue graph with a red line

Description automatically generated

Figure 17. Regplot for No. of Ratings and Selling Price

**4.4 Model Building**

Initially, the dataset was split in to train and test dataset in the ratio 70:30. As we have many columns in the dataset, I have used principal component analysis to reduce the dimensionality and it has provided us the n-components to build the model. Ensemble models such as Linear Regression, Decision Trees, Random Forest Regressor, XGBoosting, SVM, LSTM, and FNN has been used and trained the models on the train data. Hyper parameter tuning has been performed on algorithms to get the better results.

**4.5 Interpretation of Visualizations**

Let us interpret the results obtained from various visualizations. From Figure1., I observed that actual price\_in\_INR, discount\_price\_in\_INR, and selling\_price\_in\_INR has infinte values. Hence, dropped them and handled the outliers and that can be clearly seen in Figure 2.

From Figure 3., It clearly says that accessories has the highest selling price followed by stores, men’s shoes, appliances, and bags&luggage are the top five and grocery & gourmet foods has the least selling price in terms of main category.

From Figure 4., Air Conditioners has the highest selling price followed by televisions, international toy stores, gold and diamond jewellery, and cardio equipment are the top five and grocery & gourmet foods has the least selling price in terms of sub category.

From Figure 6. and Figure 7., If the actual price increases then selling price increases. But if discount price is greater then or equal to nearly 1200 then the products have high demand.

From Figure 9., Accessories has the low demand in terms of actual price if discount was not provided.

From Figure 14., If discount price increases then selling price increases. From Figure 16., If number of ratings increases then ratings also increases.

From Figure 15., There will be no change in selling price if the rating increases or decreases. Number of ratings will have no impact on selling price.

**4.6 Required Resources**

We require hardware resources and software resources.

**4.6.1 Hardware Resources**

* Processor: Intel Core i5 or higher
* RAM: 16 GigaBytes
* Operating System: Microsoft Windows 10

**4.6.2 Software Resources**

* Technology Used: Python 3.9
* Package Manager: Anaconda
* IDE: Jupyter Notebook
* Browser: Google Chrome/ Microsoft Edge
* Microsoft Office 365
* Machine Learning Library: scikit-learn

**4.7 Results and Discussions**

We have used ensembled techniques and will compare the results. Based on the below Figure 18., we can compare the train results of all the applied algorithms. Comparitively, Decision Tree has r2-score of 98% followed by Random Forest (98%), XGBoosting (94%), and Linear Regression (89%) algorithms.

If we compare the test results of all the applied algorithms, Decision Tree has r2-score of 98% followed by Random Forest (95%), Feed Forward Neural networks (94%), XGBoosting (93%), Linear Regression (89%), and LSTM (28%).

We have the minimised Mean Squared Error and Root Mean Squared Error for Feed Forward Neural Networks.

A white background with black numbers

Description automatically generated

Figure 18. Train Results for Ensemble Methods

A table with numbers and a test

Description automatically generated with medium confidence

Figure 19. Test Results for Ensemble Methods

**CHAPTER 5**

**CONCLUSIONS AND RECOMMENDATIONS**

**5.1 Conclusions**

Random Forest outperformed the Linear Regression, Decision Trees, XGBoosting, and SVM in Machine Learning Algorithms. Whereas in Neural Networks, Feed Forward Neural Networks outperformed the LSTM. For Train Dataset, Decision Trees outperformed all the other machine learning algorithms applied. Overall, if we compare the results, Random Forest outperformed all the other algorithms applied on the dataset.

**5.2 Future Recommendations**

For future advancements, integrating image recognition into the dataset holds great potential. This approach can enhance product identification and contribute to more accurate price predictions. Leveraging convolutional neural networks (CNNs) or similar techniques to extract visual features promises to capture nuanced patterns, offering a holistic perspective for improved machine learning models in E-commerce demand forecasting. Exploring the inclusion of visual data can lead to more comprehensive and sophisticated predictive analytics.

**REFERENCES**

Aamer, A., Eka Yani, L. P., & Alan Priyatna, I. M. (2020). Data Analytics in the Supply Chain Management: Review of Machine Learning Applications in Demand Forecasting. *Operations and Supply Chain Management: An International Journal*, 1–13. https://doi.org/10.31387/oscm0440281

Ali Salamai, A., Abdulrahman Ageeli, A., & M. El-kenawy, E.-S. (2022). Forecasting E-Commerce Adoption Based on Bidirectional Recurrent Neural Networks. *Computers, Materials & Continua*, *70*(3), 5091–5106. <https://doi.org/10.32604/cmc.2022.021268>

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). *Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology* (pp. 462–474). <https://doi.org/10.1007/978-3-030-36718-3_39>

Cai, W., Song, Y., & Wei, Z. (2021). Multimodal Data Guided Spatial Feature Fusion and Grouping Strategy for E-Commerce Commodity Demand Forecasting. *Mobile Information Systems*, *2021*, 1–14. <https://doi.org/10.1155/2021/5568208>

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., & Schmidt-Thieme, L. (2021). *Do We Really Need Deep Learning Models for Time Series Forecasting?*

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058. <https://doi.org/10.1016/J.JJIMEI.2022.100058>

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications*, *25*(2), 119–142. <https://doi.org/10.1080/13675567.2020.1803246>

Gasparin, A., Lukovic, S., & Alippi, C. (2022). Deep learning for time series forecasting: The electric load case. *CAAI Transactions on Intelligence Technology*, *7*(1), 1–25. <https://doi.org/10.1049/cit2.12060>

Giri, C., & Chen, Y. (2022). Deep Learning for Demand Forecasting in the Fashion and Apparel Retail Industry. *Forecasting*, *4*(2), 565–581. <https://doi.org/10.3390/forecast4020031>

Giri, C., Thomassey, S., Balkow, J., & Zeng, X. (2019). Forecasting New Apparel Sales Using Deep Learning and Nonlinear Neural Network Regression. *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, 1–6. <https://doi.org/10.1109/ICESI.2019.8863024>

Han, Z., Zhao, J., Leung, H., Ma, K. F., & Wang, W. (2021). A Review of Deep Learning Models for Time Series Prediction. *IEEE Sensors Journal*, *21*(6), 7833–7848. <https://doi.org/10.1109/JSEN.2019.2923982>

Hsieh, P.-H. (2019). *A Study of Models for Forecasting E-Commerce Sales During a Price War in the Medical Product Industry* (pp. 3–21). <https://doi.org/10.1007/978-3-030-22335-9_1>

Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., & Zhang, H. (2019). Deep Learning with Long Short-Term Memory for Time Series Prediction. *IEEE Communications Magazine*, *57*(6), 114–119. <https://doi.org/10.1109/MCOM.2019.1800155>

Jain, A., Karthikeyan, V., B, S., BR, S., K, S., & S, B. (2020). Demand Forecasting for E-Commerce Platforms. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 1–4. <https://doi.org/10.1109/INOCON50539.2020.9298395>

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise. *Mathematical Problems in Engineering*, *2019*, 1–15. <https://doi.org/10.1155/2019/8503252>

Khan, M. A., Saqib, S., Alyas, T., Ur Rehman, A., Saeed, Y., Zeb, A., Zareei, M., & Mohamed, E. M. (2020). Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. *IEEE Access*, *8*, 116013–116023. <https://doi.org/10.1109/ACCESS.2020.3003790>

Kharfan, M., Chan, V. W. K., & Firdolas Efendigil, T. (2021). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. *Annals of Operations Research*, *303*(1–2), 159–174. <https://doi.org/10.1007/s10479-020-03666-w>

Kulshrestha, S., & Saini, M. L. (2020). Study for the Prediction of E-Commerce Business Market Growth using Machine Learning Algorithm. *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 1–6. <https://doi.org/10.1109/ICRAIE51050.2020.9358275>

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2021). An Experimental Review on Deep Learning Architectures for Time Series Forecasting. *International Journal of Neural Systems*, *31*(03), 2130001. <https://doi.org/10.1142/S0129065721300011>

Leung, K. H., Mo, D. Y., Ho, G. T. S., Wu, C. H., & Huang, G. Q. (2020). Modelling near-real-time order arrival demand in e-commerce context: a machine learning predictive methodology. *Industrial Management & Data Systems*, *120*(6), 1149–1174. <https://doi.org/10.1108/IMDS-12-2019-064>

Li, J., Cui, T., Yang, K., Yuan, R., He, L., & Li, M. (2021). Demand Forecasting of E-Commerce Enterprises Based on Horizontal Federated Learning from the Perspective of Sustainable Development. *Sustainability*, *13*(23), 13050. <https://doi.org/10.3390/su132313050>

Li, Z., & Zhang, N. (2022). Short-Term Demand Forecast of E-Commerce Platform Based on ConvLSTM Network. *Computational Intelligence and Neuroscience*, *2022*, 1–10. <https://doi.org/10.1155/2022/5227829>

Moukafih, Y., Hafidi, H., & Ghogho, M. (2019). Aggressive Driving Detection Using Deep Learning-based Time Series Classification. *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–5. <https://doi.org/10.1109/INISTA.2019.8778416>

Nithin, S. S. J., Rajasekar, T., Jayanthy, S., Karthik, K., & Rithick, R. R. (2022). Retail Demand Forecasting using CNN-LSTM Model. *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 1751–1756. <https://doi.org/10.1109/ICEARS53579.2022.9752283>

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S., Reuter, U., Gama, J., & Gandomi, A. (2020). Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods. *Mathematics*, *8*(10), 1799. <https://doi.org/10.3390/math8101799>

Pan, H., & Zhou, H. (2020). Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce. *Electronic Commerce Research*, *20*(2), 297–320. <https://doi.org/10.1007/s10660-020-09409-0>

Qi, Y., Li, C., Deng, H., Cai, M., Qi, Y., & Deng, Y. (2019). A Deep Neural Framework for Sales Forecasting in E-Commerce. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 299–308. <https://doi.org/10.1145/3357384.3357883>

Rai, S., Gupta, A., Anand, A., Trivedi, A., & Bhadauria, S. (2019). Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. <https://doi.org/10.1109/ICCCNT45670.2019.8944783>

Ramadevi, B., & Bingi, K. (2022). Chaotic Time Series Forecasting Approaches Using Machine Learning Techniques: A Review. *Symmetry*, *14*(5), 955. <https://doi.org/10.3390/sym14050955>

Ren, S., Chan, H.-L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. *Annals of Operations Research*, *291*(1–2), 761–777. <https://doi.org/10.1007/s10479-019-03148-8>

Ren, S., Choi, T.-M., Lee, K.-M., & Lin, L. (2020). Intelligent service capacity allocation for cross-border-E-commerce related third-party-forwarding logistics operations: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *134*, 101834. <https://doi.org/10.1016/j.tre.2019.101834>

Singh, K., Booma, P. M., & Eaganathan, U. (2020). E-Commerce System for Sale Prediction Using Machine Learning Technique. *Journal of Physics: Conference Series*, *1712*(1), 012042. <https://doi.org/10.1088/1742-6596/1712/1/012042>

Smirnov, P. S., & Sudakov, V. A. (2021). Forecasting new product demand using machine learning. *Journal of Physics: Conference Series*, *1925*(1), 012033. <https://doi.org/10.1088/1742-6596/1925/1/012033>

Swami, D., Shah, A. D., & Ray, S. K. B. (2020). *Predicting Future Sales of Retail Products using Machine Learning*.

Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep Learning for Time Series Forecasting: A Survey. *Big Data*, *9*(1), 3–21. <https://doi.org/10.1089/big.2020.0159>

Vavliakis, K., Siailis, A., & Symeonidis, A. (2021). Optimizing Sales Forecasting in e-Commerce with ARIMA and LSTM Models. *Proceedings of the 17th International Conference on Web Information Systems and Technologies*, 299–306. <https://doi.org/10.5220/0010659500003058>

Yin, X., & Tao, X. (2021). Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. *Scientific Programming*, *2021*, 1–9. <https://doi.org/10.1155/2021/2179692>

Zhu, H. (2021). A Deep Learning Based Hybrid Model for Sales Prediction of E-commerce with Sentiment Analysis. *2021 2nd International Conference on Computing and Data Science (CDS)*, 493–497. <https://doi.org/10.1109/CDS52072.2021.00091>

Zohdi, M., Rafiee, M., Kayvanfar, V., & Salamiraad, A. (2022a). Demand forecasting based machine learning algorithms on customer information: an applied approach. *International Journal of Information Technology*, *14*(4), 1937–1947. <https://doi.org/10.1007/s41870-022-00875-3>

Zohdi, M., Rafiee, M., Kayvanfar, V., & Salamiraad, A. (2022b). Demand forecasting based machine learning algorithms on customer information: an applied approach. *International Journal of Information Technology*, *14*(4), 1937–1947. <https://doi.org/10.1007/s41870-022-00875-3>

**APPENDIX A**

**RESEARCH PROPOSAL**

# DEMAND FORECASTING FOR E-COMMERCE PRODUCTS USING MACHINE LEARNING

MATSA VENKATA LAKSHMI KAVYA

Research Proposal

AUGUST 2023

# Abstract

Effective demand forecasting plays a pivotal role in the retail industry by optimizing inventory management, reducing costs, and enhancing customer satisfaction. In this research paper, we present a comprehensive study on demand forecasting for E-Commerce products, utilizing machine learning techniques. The study begins by exploring historical sales data, product attributes, external factors and customer behavior. This study aims to develop accurate and robust predictive models. The research investigates state-of-the-art machine learning algorithms like Linear Regression, Support Vector Machine, Random Forest Algorithms.

Furthermore, the research introduces the integration of advanced machine learning algorithms such as LSTM (Long Short-Term Memory) and XGBoost, to harness the predictive capabilities of deep learning and ensemble methods. A comparative analysis is conducted to evaluate the performance of these models in accurately forecasting product demand across different categories. The research paper also addresses challenges related to data preprocessing, feature engineering, and model tuning, providing insights into best practices for achieving optimal forecasting results. The goal is to provide e-commerce businesses with actionable insights into selecting and deploying suitable forecasting models that can adapt to dynamic market conditions and evolving consumer preferences.

The proposed approach showcases promising results, paving the way for improved inventory management strategies, streamlined supply chain operations, and heightened customer satisfaction. As e-commerce continues to reshape the retail landscape, the insights derived from this research hold significant implications for industry practitioners and researchers seeking to optimize demand forecasting methodologies.

# List of Abbreviations

CNN ………………………. Convolutional Neural Networks

KNN ………………………. K-Nearest Neighbor

MSE ………………………. Mean Squared Error

RMSE …………………….. Root Mean Squared Error

**Table of Contents**

[Abstract 55](#_Toc75518344)

[1. Background](#_Toc75518345) 58

[2. Problem Statement OR Related Research OR Related Work 5](#_Toc75518346)9

[3. Research Questions (If any) 6](#_Toc75518347)2

[4. Aim and Objectives 6](#_Toc75518348)2

[5. Significance of the Study 6](#_Toc75518349)3

[6. Scope of the Study](#_Toc75518350) 63

[7. Research Methodology](#_Toc75518351) 64

[8. Requirements Resources](#_Toc75518352) 68

[9. Research Plan](#_Toc75518353) 69

[References](#_Toc75518354) 69

# 1. Background

E-Commerce is the buying and selling of goods and services over the internet (Ali Salamai et al., 2022). It is a rapidly growing industry, with global sales reaching over $4.9 trillion in 2021. E-Commerce has several advantages over traditional brick-and-mortar businesses, including accessibility, cost savings and convenience.

E-Commerce is one of the main 10 businesses which have an extremely huge measure of information stockpiling. With the admittance to how much information they will want to establish a game changing climate for the Web based business industry. This industry has been burning through a huge number of dollars in publicizing, virtual entertainment, arranging tied down information and substantially more to create more deals however they didn't understand that with machine learning they will be able to move forward against their rivals. Machine Learning is an enormous tree limb which have numerous specializations, for example, information mining, man-made consciousness, expanded reality and expectation. For this exploration, will be just zeroing in on the expectation utilizing machine learning.(Singh et al., 2020).

Demand Forecasting is the process of predicting future demand for products or services. It is an important tool for e-commerce businesses, as it can help them to optimize inventory levels, set prices strategically, and plan marketing campaigns. In this research, we will be using machine learning algorithms to forecast demand for e-commerce products and can be used to identify fraudulent transactions. This can help e-commerce businesses protect themselves from financial losses. We additionally might want to scrutinize the capacity of neural networks models for undertaking of sales determining which is broadly experienced in various applications(Ensafi et al., 2022).

# 2. Problem Statement or Related Research or Related Work

In the rapidly evolving landscape of e-commerce, the accurate prediction of product demand is paramount for businesses to optimize inventory management, enhance customer satisfaction, and improve operational efficiency. Traditional methods often fall short in capturing the intricate dynamics of online consumer behaviour, which include factors such as brand, category, pricing, discounts, and ratings. To address this challenge, this research project aims to develop and evaluate machine learning models for demand forecasting in the e-commerce domain.

The goal is to design and implement machine learning models that accurately forecast the demand for e-commerce products over a specified period. This involves exploring a range of machine learning algorithms, including Linear regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Random Forest, along with deep learning models such as Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). The project seeks to identify the most effective models for demand forecasting, provide actionable insights for e-commerce businesses to make informed decisions about inventory management, pricing strategies and resource allocation.

By addressing this problem, the research aims to bridge the gap between data-driven insights and real-world e-commerce operations, contributing to the advancement of demand forecasting practices in the digital retail sector.

**2.1 Related Research**

In this research, they have used variety of datasets for their experiments, including the MNIST dataset, the CIFAR-10 dataset, and the ImageNet dataset. They have applied various deep learning models like CNN, RNN, LSTM Networks. The authors found that deep learning models can achieve state-of-the-art performance on time series prediction tasks. However, they also noted that the performance of the different models can vary depending on the characteristics of data.

(Nithin et al., 2022) In this research, they have used the daily sales data of 100 retail products. They have applied CNN to extract features from the historical sales data, and a LSTM network to capture the temporal dependencies in the data. The CNN-LSTM model trained on a training dataset and achieved high accuracy in retail demand forecasting.

(Smirnov & Sudakov, 2021) In this research, they have used sales data for 100 products. The products are from different categories such as electronics, clothing, and food. They proposed a data augmentation technique to improve the performance of the random forest regressor. This algorithm achieved high accuracy in forecasting the demand for new products.

(Giri & Chen, 2022) In this research, they have used historical sales data of 100 fashion and apparel products. A deep learning model is trained on preprocessed data. They have used LSTM model. They achieved a high accuracy in demand forecasting for fashion and apparel products by using LSTM model.

(Zohdi et al., 2022b) In this research, they have used historical demand data for 100 products from an online retailer. They have used KNN, Decision Tree, Gradient Boosting algorithms and those algorithms are evaluated using MSE, MAE, Coefficient of determination (R^2). All these algorithms achieved the best performance in terms of MSE, MAE, R^2.

(Nosratabadi et al., 2020) In this research, they have conducted a systematic literature review of 188 papers published in top economics journals from 2010 to 2019. They are classified as time series analysis, econometrics, financial risk management and other applications. They have used SVM, Random forests and neural networks in econometrics. Finally, they concluded that machine learning and deep learning methods have the potential to revolutionize economics research.

(Moukafih et al., 2019) In this research, they have used UAH-DriveSet, which is a public dataset that provides a large amount of naturalistic driving data obtained from smartphones via a driving monitoring application. They have used LSTM Fully convolutional network to classify driving session as aggressive or non-aggressive. They can learn long-term dependencies in time series data, and it is used to extract spatial features from the time series data. It has achieved an accuracy of 95.88% in detecting aggressive driving behavior.

(Torres et al., 2021) In this research, they begin by defining time series forecasting and deep learning algorithms. It also discusses the advantages of using deep learning for time series forecasting such as its ability to learn complex temporal patterns and its ability to handle non-linear data. They have used feedforward neural networks, RNN, and CNN. Finally, they concluded that deep learning models have been shown to outperform other methods in many cases.

(Ensafi et al., 2022) In this research, they have used sales history of a furniture store for a period of 10 years. They have used Seasonal Autoregressive Integrated Moving Average (SARIMA), Triple Exponential Smoothing (TES), Prophet and LSTM. The results of the study show that the LSTM model outperforms in terms of accuracy.

(Singh et al., 2020) In this research, they have used historical sales data from an e-commerce platform. They used Random Forest algorithm to predict the sales of products in the e-commerce platform. They also proposed a systematic pre-processing framework to overcome challenges in e-commerce. They concluded that the proposed model is a promising method for predicting sales in an e-commerce platform.

(Lara-Benítez et al., 2021) In this research, they have conducted an extensive experimental study on 12 different forecasting problems using 50000 time series. They evaluated deep learning algorithms such as LSTM, Gated recurrent unit (GRU), CNN, Deep autoregressive model (DAR), DeepAR, Prophet, and transformer. They concluded that LSTM and CNN are the best performing architectures.

(Yin & Tao, 2021) In this research, they have used sales data from an e-commerce platform. It includes information such as product category, price, brand and many more. They have used CNN and RNN algorithms and concluded that they achieve a high accuracy in predicting merchandise sales.

(Hua et al., 2019) In this research, they have used traffic data collected from a telecommunication network. They have used LSTM model; it was trained to minimize the mean squared error between the predicted traffic data and the actual traffic data. It has achieved a high accuracy in time series prediction.

(Ramadevi & Bingi, 2022) In this research, they have used CNN, Wavelet neural networks (WNNs), Fuzzy neural networks (FNNs), LSTM, Extreme learning machines (ELMs) and SVM algorithms. They have evaluated these algorithms using root mean squared error and the mean absolute error. They concluded that machine learning algorithms can be effective for chaotic time series forecasting.

(Elsayed et al., 2021) In this research, they have used multiple datasets like air quality, electricity demand, exchange rate, financial time series, gas demand, hydrology, machine health, solar irradiance, traffic flow, and wind speed. They have used simple gradient boosting regression tree and deep learning models like LSTM, prophet, ConvLSTM, DeepAR, Exponential smoothing, wave net and whisker but finally, they concluded that gradient descent model able to achieve better performance than deep learning models.

(Gasparin et al., 2022) In this research, they have used hourly electric load data from the Swiss power grid. They have used deep learning models like CNN, RNN, LSTM and the results show that they have achieved a high accuracy in electric load forecasting.

# 3. Research Questions (If any)

Some research questions are given below-

* **RQ1:** How does the brand of a product influence its demand?
* **RQ2:** How the relationship between actual price, selling price and demand can be leveraged to make accurate demand forecasts?
* **RQ3:** How do product rating and number of ratings impact demand?
* **RQ4:** What role does the presence of a discount play in influencing demand?
* **RQ5:** What are the most suitable machine learning algorithms for accurate demand forecasting of diverse product categories within the e-commerce industry?

# 4. Aim and Objectives

The main aim of this research is to develop accurate and effective demand forecasting models using machine learning techniques for e-commerce products, enabling businesses to optimize inventory management and enhance operational efficiency.

The research objectives are formulated based on the aim of this study which are as follows:

* Investigate the effect of different product categories and sub-categories on pricing and discounts.
* Investigate the relationship between pricing strategies and the likelihood of a product being sold.
* Utilize Machine Learning techniques to develop a predictive model for estimating product demand and sales based on the provided attributes.
* To evaluate and compare the performance of different forecasting models using appropriate metrics (e.g., Mean Squared Error, Root Mean Squared Error).

# 5. Significance of the Study

This research might aid in the importance of forecasting and prediction methods, and it could also have an impact on efforts to maximise the demand for e-commerce products based on different factors.

The findings of this research might lead to enhanced demand prediction accuracy, customer behaviour insights, optimized pricing and discounting strategies, supply chain and inventory management.

* By understanding the demand for different products, businesses can optimize their inventory levels and reduce costs.
* By forecasting demand, businesses can set process that are more likely to be profitable.
* By understanding the factors that influence demand, businesses can target their marketing campaigns more effectively.

The research has the potential to revolutionize demand forecasting in e-commerce by harnessing machine learning techniques on a dataset rich with attributes that capture brand dynamics, pricing information and many more. This can lead to informed decisions, improved customer satisfaction and greater operational efficiency within the e-commerce landscape.

# 6. Scope of the Study

The research mainly focuses on demand forecasting for e-commerce products. In scope of this research, we need to check the availability of data, the time, and resources available for this study. Identifying and creating relevant features from the collected data to enhance the predictive power of the models. Implementing and training the ensemble techniques using the pre-processed data to forecast the demand for e-commerce products. Evaluating and comparing the performance of different ensemble methods. Defining appropriate evaluation metrics to measure the accuracy and robustness of the ensemble techniques in demand forecasting. This could include metrics like Mean Squared Error, Root Mean Squared Error.

The out of scope for this research would be conducting a survey of e-commerce businesses to understand their demand forecasting needs. We will not use the traditional time series methods like ARIMA and Exponential Smoothing. While model performance is being evaluated, in-depth hyperparameter tuning for each model won’t be the focus of the study. Basic hyperparameter tuning will be used for a comparative analysis.

# 7. Research Methodology

As a proof of notion, this project aims to show the potential of applying machine learning approaches to maximize the demand for e-commerce products. In this research, we will use Amazon Products Sales Dataset from the Kaggle. The collected dataset needs to be pre-processed, and then we visualize the dataset with the help of Exploratory Data Analysis. Next, we will split the dataset into training and testing with the ration of 70:30. Now, we will apply machine learning algorithms such as Linear Regression, K-Nearest Neighbors (KNN) Regressor, Random Forest Regressor, Decision Tree Regressor, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). Each model’s performance can be calculated in terms of Accuracy, Mean Squared Error (MSE), Root Mean Squared Error (RMSE). Once the performance is evaluated, we will compare the proposed models and choose the best fit model for the chosen data. and convenience.

**7.1 Steps of Methodology**

Steps of Methodology has been explained below -.

* + 1. **Data Collection**

The first step is to collect a dataset for the products that we want to forecast demand for. In this research, we will use the “Amazon Products Sales Dataset 2023” from Kaggle in which consists of category, sub-category, actual price, discount, images, URL, title of the product, ratings, number of ratings.

* + 1. **Data Preprocessing**

Once we collect the dataset, we need to clean and prepare the dataset. Handle missing values, outliers, and any inconsistencies.

* + 1. **Data Exploration and Visualization**

After data preprocessing, visualize the distribution of variables, exploring trends among the features.

* + 1. **Feature Selection**

Extract relevant features from the dataset that might impact demand forecasting. Features should be measurable.

* + 1. **Model Selection**

Choose a set of machine learning models such as Linear regressor, K-Nearest Neighbors Regressor, Decision Tree Regressor, Random Forest Regressor, and Long Short-Term Memory.

* + 1. **Data Splitting**

In this research, we will divide the dataset into training and validation sets. 70% of the data will be training set and 30% of the data will be the validation set. Predictive models were built using the training set and their efficacy was assessed using the validation set.

* + 1. **Model Training**

Train each machine learning on the training data and fine-tune hyperparameters using cross-validation. In this research we are using Linear Regression, K-Nearest Neighbors (KNN) Regressor, Random Forest Regressor, Decision Tree Regressor, Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN). We will explain the models in detail.

* + - 1. **Linear Regression**

Linear Regression assumes that dependent and independent variables have a linear connection. The goal of regression analysis is to determine an optimal line of calculation that characterizes the connection between independent and dependent variables. It predicts relationships between continuous variables using a statistical technique. It can only support linear solutions. The X-axis in linear regression represents the independent variable, while the Y-axis represents the dependent one. One input is required for simple linear regression. Linear regression with many predictors is known as multiple linear regression. When there are large number of features with less datasets, linear regressions may outperform Decision trees/Random forests.

* + - 1. **Decision Tree Regressor**

Decision trees are non-parametric supervised machine learning methods used for classification and regression. It is a structure like a flowchart in which decisions and decision-making processes are visually and explicitly represented. They are effective in capturing non-linear relationships which can be difficult to achieve with other algorithms like Support Vector Machine and Linear Regression. The outputs are easy to read without requiring statistical knowledge or complex concepts. Some people believe decision trees more closely mirror human decision-making than others like regression and classification approaches. Trees can be displayed graphically and can be easily interpreted by non-experts. Decision trees can easily handle qualitative (categorical) features without the need to create dummy variables. In general cases, Decision trees will have better average accuracy. Decision trees handle collinearity better than Linear Regressor. It supports automatic feature interaction. Decision Trees don’t have the same level of predictive accuracy as some of other regression and classification approaches. They can be non-robust. For example, a small change in the data can cause a large change in the final estimated tree. As the tree grows, it becomes prone to overfitting and requires pruning.

* + - 1. **K-Nearest Neighbor Regressor**

Like decision trees it is simple and easy to explain to laypeople. It is non-parametric. Therefore, it doesn’t have any assumptions on the data distribution. KNN is more of an exception to the general machine learning workflow. It doesn’t have a training/validation/test set. The model created with KNN is available in a labeled dataset, placed in metric space. If you want to classify any object, the model must read through all the data and compare the distances of the closest objects. It is easy to implement for multi-class problems. Just supply the ‘k’ a value that is equivalent to the number of classes and you are good to go. When working with KNN, you just need to provide two parameters, k (the numbers of neighbors to consider) and the choice of Distance Function (e.g., Euclidean, Manhattan distance). It can be used for classification and regression. We don’t need to fit a model in advance, just provide the data point and it will give you the prediction. It is slow with a larger dataset. If it is going to classify a new sample, it will have to read the whole dataset, hence, it becomes very slow as the dataset increases. KNN is more appropriate to use when you have a small number of inputs. If the number of variables grows, the KNN algorithm will have a hard time predicting the output of a new data point. Feature inputs need to be scaled. KNN is very sensitive to outliers. Since it is an instance-based algorithm based on the distance criteria, if we have some outliers in the data, it is liable to create a biased outcome. It is not capable of treating or dealing with missing values. If we have an imbalanced class data, the algorithm might wrongly pick the majority class.

* + - 1. **Random Forest Regressor**

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. A classification algorithm consisting of many decision trees combined to get a more accurate result as compared to a single tree. This avoids and prevents overfitting by using multiple trees. This gives accurate and precise results. It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing. It consumes more computation. The process of generation and analyzing is time-consuming. This has complex visualization as it determines the pattern behind the data.

* + - 1. **Long Short-Term Memory**

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is specifically designed to handle sequential data. (Jain et al., 2020) LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well suited for tasks such as language translation, speech recognition, and time series forecasting. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. LSTM can by default retain the information for a long period of time. Using LSTM, forecasting models can predict future values based on previous, sequential data. This provides greater accuracy for demand forecasters which results in better decision making for the business. LSTM could take inputs with different lengths. This feature is especially useful when LSTM is used to build general forecasting models for specific customers or industries. The different gates inside LSTM boost its capability for capturing non-linear relationships for forecasting. Causal factors generally have non-linear impact on demand. When these factors are used as part of the input variable, the LSTM could learn the nonlinear relationship for forecasting. The LSTM requires more computation than other recurrent neural networks. The main reason is that it has more parameters, which are used for demand forecasts.

* + - 1. **Convolutional Neural Network**

The CNN model will utilize two layers. The first is a 1-D convolutional layer and the second is a maximum pooling layer. In the wake of going through these layers, the result is then smoothed to be deciphered by a secret thick layer and afterward estimating the interest. The 1-D CNN layer recognizes the examples between the timesteps. The info is in three-dimensional shape comprising of tests, timesteps and highlights. The information is preprocessed to reshape and resample from 2-D information as [samples, timesteps] into three-dimensional information as [samples, timesteps, features]. Similar three-dimensional information will be utilized for the LSTM model.(Nithin et al., 2022)

* + 1. **Model Evaluation**

In Machine Learning, it is crucial to be able to assess the efficiency of models. There are several metrics by which one might assess a machine learning model’s efficacy. In this analysis, we compare the K-Nearest Neighbor, Decision Tree, Random Forest, Long Short-Term Memory and Convolutional Neural Networks and their performance can be evaluated using below techniques.

* + - 1. **Mean Squared Error**

It is computed by averaging squares of variance between actual and predicted values of the target variable, which may be used to assess the accuracy of a forecasting technique.

* + - 1. **Root Mean Squared Error**

An effective statistic for assessing a model’s predictive power is RMSE (Root Mean Squared Error). It affects how much inaccuracy is added to these predictions by the model. RMSE was selected as it has the advantage being measured on the same scale as the output being forecasted.

* + - 1. **Accuracy**

The accuracy of a machine learning model indicates how often it correctly predicted outcomes. The accuracy of a model is measured by how many classes it properly predicts relative to how many predictions it makes in total. There are two other ways to evaluate a model’s efficacy, but this is one of them. This efficiency analysis will compare the relative merits of several methods.

# 8. Requirements Resources

We have two types of required resources. They are Hardware Requirements and Software Requirements.

* 1. **Hardware Requirements**
* Processor: Intel Core i5 or higher
* RAM: 8 Gigabytes or higher
* GB Storage space or higher
  1. **Software Requirements**
* Language: Python
* Package Manager: Anaconda
* IDE: Jupyter
* Browser: Google Chrome/ Microsoft Edge
* Microsoft Office 365

# 9. Research Plan

A screenshot of a graph

Description automatically generated

# References

Aamer, A., Eka Yani, L. P., & Alan Priyatna, I. M. (2020). Data Analytics in the Supply Chain Management: Review of Machine Learning Applications in Demand Forecasting. *Operations and Supply Chain Management: An International Journal*, 1–13. https://doi.org/10.31387/oscm0440281

Ali Salamai, A., Abdulrahman Ageeli, A., & M. El-kenawy, E.-S. (2022). Forecasting E-Commerce Adoption Based on Bidirectional Recurrent Neural Networks. *Computers, Materials & Continua*, *70*(3), 5091–5106. https://doi.org/10.32604/cmc.2022.021268

Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). *Sales Demand Forecast in E-commerce Using a Long Short-Term Memory Neural Network Methodology* (pp. 462–474). https://doi.org/10.1007/978-3-030-36718-3\_39

Cai, W., Song, Y., & Wei, Z. (2021). Multimodal Data Guided Spatial Feature Fusion and Grouping Strategy for E-Commerce Commodity Demand Forecasting. *Mobile Information Systems*, *2021*, 1–14. https://doi.org/10.1155/2021/5568208

Elsayed, S., Thyssens, D., Rashed, A., Jomaa, H. S., & Schmidt-Thieme, L. (2021). *Do We Really Need Deep Learning Models for Time Series Forecasting?*

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058. https://doi.org/10.1016/J.JJIMEI.2022.100058

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications*, *25*(2), 119–142. https://doi.org/10.1080/13675567.2020.1803246

Gasparin, A., Lukovic, S., & Alippi, C. (2022). Deep learning for time series forecasting: The electric load case. *CAAI Transactions on Intelligence Technology*, *7*(1), 1–25. https://doi.org/10.1049/cit2.12060

Giri, C., & Chen, Y. (2022). Deep Learning for Demand Forecasting in the Fashion and Apparel Retail Industry. *Forecasting*, *4*(2), 565–581. https://doi.org/10.3390/forecast4020031

Giri, C., Thomassey, S., Balkow, J., & Zeng, X. (2019). Forecasting New Apparel Sales Using Deep Learning and Nonlinear Neural Network Regression. *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, 1–6. https://doi.org/10.1109/ICESI.2019.8863024

Han, Z., Zhao, J., Leung, H., Ma, K. F., & Wang, W. (2021). A Review of Deep Learning Models for Time Series Prediction. *IEEE Sensors Journal*, *21*(6), 7833–7848. https://doi.org/10.1109/JSEN.2019.2923982

Hsieh, P.-H. (2019). *A Study of Models for Forecasting E-Commerce Sales During a Price War in the Medical Product Industry* (pp. 3–21). https://doi.org/10.1007/978-3-030-22335-9\_1

Hua, Y., Zhao, Z., Li, R., Chen, X., Liu, Z., & Zhang, H. (2019). Deep Learning with Long Short-Term Memory for Time Series Prediction. *IEEE Communications Magazine*, *57*(6), 114–119. https://doi.org/10.1109/MCOM.2019.1800155

Jain, A., Karthikeyan, V., B, S., BR, S., K, S., & S, B. (2020). Demand Forecasting for E-Commerce Platforms. *2020 IEEE International Conference for Innovation in Technology (INOCON)*, 1–4. https://doi.org/10.1109/INOCON50539.2020.9298395

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise. *Mathematical Problems in Engineering*, *2019*, 1–15. https://doi.org/10.1155/2019/8503252

Khan, M. A., Saqib, S., Alyas, T., Ur Rehman, A., Saeed, Y., Zeb, A., Zareei, M., & Mohamed, E. M. (2020). Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. *IEEE Access*, *8*, 116013–116023. https://doi.org/10.1109/ACCESS.2020.3003790

Kharfan, M., Chan, V. W. K., & Firdolas Efendigil, T. (2021). A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches. *Annals of Operations Research*, *303*(1–2), 159–174. https://doi.org/10.1007/s10479-020-03666-w

Kulshrestha, S., & Saini, M. L. (2020). Study for the Prediction of E-Commerce Business Market Growth using Machine Learning Algorithm. *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 1–6. https://doi.org/10.1109/ICRAIE51050.2020.9358275

Lara-Benítez, P., Carranza-García, M., & Riquelme, J. C. (2021). An Experimental Review on Deep Learning Architectures for Time Series Forecasting. *International Journal of Neural Systems*, *31*(03), 2130001. https://doi.org/10.1142/S0129065721300011

Leung, K. H., Mo, D. Y., Ho, G. T. S., Wu, C. H., & Huang, G. Q. (2020). Modelling near-real-time order arrival demand in e-commerce context: a machine learning predictive methodology. *Industrial Management & Data Systems*, *120*(6), 1149–1174. https://doi.org/10.1108/IMDS-12-2019-0646

Li, J., Cui, T., Yang, K., Yuan, R., He, L., & Li, M. (2021). Demand Forecasting of E-Commerce Enterprises Based on Horizontal Federated Learning from the Perspective of Sustainable Development. *Sustainability*, *13*(23), 13050. https://doi.org/10.3390/su132313050

Li, Z., & Zhang, N. (2022). Short-Term Demand Forecast of E-Commerce Platform Based on ConvLSTM Network. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/5227829

Moukafih, Y., Hafidi, H., & Ghogho, M. (2019). Aggressive Driving Detection Using Deep Learning-based Time Series Classification. *2019 IEEE International Symposium on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–5. https://doi.org/10.1109/INISTA.2019.8778416

Nithin, S. S. J., Rajasekar, T., Jayanthy, S., Karthik, K., & Rithick, R. R. (2022). Retail Demand Forecasting using CNN-LSTM Model. *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 1751–1756. https://doi.org/10.1109/ICEARS53579.2022.9752283

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S., Reuter, U., Gama, J., & Gandomi, A. (2020). Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods. *Mathematics*, *8*(10), 1799. https://doi.org/10.3390/math8101799

Pan, H., & Zhou, H. (2020). Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce. *Electronic Commerce Research*, *20*(2), 297–320. https://doi.org/10.1007/s10660-020-09409-0

Qi, Y., Li, C., Deng, H., Cai, M., Qi, Y., & Deng, Y. (2019). A Deep Neural Framework for Sales Forecasting in E-Commerce. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 299–308. https://doi.org/10.1145/3357384.3357883

Rai, S., Gupta, A., Anand, A., Trivedi, A., & Bhadauria, S. (2019). Demand prediction for e-commerce advertisements: A comparative study using state-of-the-art machine learning methods. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. https://doi.org/10.1109/ICCCNT45670.2019.8944783

Ramadevi, B., & Bingi, K. (2022). Chaotic Time Series Forecasting Approaches Using Machine Learning Techniques: A Review. *Symmetry*, *14*(5), 955. https://doi.org/10.3390/sym14050955

Ren, S., Chan, H.-L., & Siqin, T. (2020). Demand forecasting in retail operations for fashionable products: methods, practices, and real case study. *Annals of Operations Research*, *291*(1–2), 761–777. https://doi.org/10.1007/s10479-019-03148-8

Ren, S., Choi, T.-M., Lee, K.-M., & Lin, L. (2020). Intelligent service capacity allocation for cross-border-E-commerce related third-party-forwarding logistics operations: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *134*, 101834. https://doi.org/10.1016/j.tre.2019.101834

Singh, K., Booma, P. M., & Eaganathan, U. (2020). E-Commerce System for Sale Prediction Using Machine Learning Technique. *Journal of Physics: Conference Series*, *1712*(1), 012042. https://doi.org/10.1088/1742-6596/1712/1/012042

Smirnov, P. S., & Sudakov, V. A. (2021). Forecasting new product demand using machine learning. *Journal of Physics: Conference Series*, *1925*(1), 012033. https://doi.org/10.1088/1742-6596/1925/1/012033

Swami, D., Shah, A. D., & Ray, S. K. B. (2020). *Predicting Future Sales of Retail Products using Machine Learning*.

Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep Learning for Time Series Forecasting: A Survey. *Big Data*, *9*(1), 3–21. https://doi.org/10.1089/big.2020.0159

Vavliakis, K., Siailis, A., & Symeonidis, A. (2021). Optimizing Sales Forecasting in e-Commerce with ARIMA and LSTM Models. *Proceedings of the 17th International Conference on Web Information Systems and Technologies*, 299–306. https://doi.org/10.5220/0010659500003058

Yin, X., & Tao, X. (2021). Prediction of Merchandise Sales on E-Commerce Platforms Based on Data Mining and Deep Learning. *Scientific Programming*, *2021*, 1–9. https://doi.org/10.1155/2021/2179692

Zhu, H. (2021). A Deep Learning Based Hybrid Model for Sales Prediction of E-commerce with Sentiment Analysis. *2021 2nd International Conference on Computing and Data Science (CDS)*, 493–497. https://doi.org/10.1109/CDS52072.2021.00091

Zohdi, M., Rafiee, M., Kayvanfar, V., & Salamiraad, A. (2022a). Demand forecasting based machine learning algorithms on customer information: an applied approach. *International Journal of Information Technology*, *14*(4), 1937–1947. https://doi.org/10.1007/s41870-022-00875-3

Zohdi, M., Rafiee, M., Kayvanfar, V., & Salamiraad, A. (2022b). Demand forecasting based machine learning algorithms on customer information: an applied approach. *International Journal of Information Technology*, *14*(4), 1937–1947. https://doi.org/10.1007/s41870-022-00875-3

**Refer: Harvard Referencing Guide**