|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Leeds University  **Business School** | | A close-up of a sign  Description automatically generated | | | | | | | | |
| **Dissertation/Project Coversheet** | | | | | | | | | | |
|  | | | | | | | | | | |
| Student ID Number: | **2** | | **0** | **1** | **7** | **7** | **1** | **0** | **3** | **5** |
| Student Name | **Shraddha Dattatray Kumbhar** | | | | | | | | | |
| Module Code: | **LUBS5579M** | | | | | | | | | |
| Programme of Study: | **Business Analytics and Decision Sciences** | | | | | | | | | |
| Supervisor: | **Dr. Yuwang Chen** | | | | | | | | | |
| Title: | **Enhancing Data-Driven Decision-Making in E-Commerce Using Machine Learning Techniques.** | | | | | | | | | |
| Declared Word Count: | **9740** | | | | | | | | | |
| Please Note:  Your declared word count must be accurate, and should not mislead. Making a fraudulent statement concerning the work submitted for assessment could be considered academic malpractice and investigated as such.  If the amount of work submitted is higher than that specified by the word limit or that declared on your word count, this may be reflected in the mark awarded and noted through individual feedback given to you.  It is not acceptable to present matters of substance, which should be included in the main body of the text, in the appendices (“appendix abuse”).  It is not acceptable to attempt to hide words in graphs and diagrams; only text which is strictly necessary should be included in graphs and diagrams. | | | | | | | | | | |
| By submitting an assignment you confirm you have read and understood the University of Leeds [**Declaration of Academic Integrity**](http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf) ( <http://www.leeds.ac.uk/secretariat/documents/academic_integrity.pdf>). | | | | | | | | | | |

**Table of Contents**

[Acknowledgement 4](#_Toc176177967)

[Abstract: 6](#_Toc176177968)

[1 Introduction 7](#_Toc176177969)

[1.1Research Background 7](#_Toc176177970)

[1.2Research Objectives 9](#_Toc176177971)

[1.3Research Structure 9](#_Toc176177972)

[2 Literature Review 9](#_Toc176177973)

[2.1 Evolution of E-Commerce 9](#_Toc176177974)

[2.2 Optimizing E-commerce with Machine Learning 10](#_Toc176177975)

[2.3 ML Techniques in E-Commerce Decision-Making 12](#_Toc176177976)

[2.4 Challenges in Implementation of Forecasting and Sentiment Analysis 13](#_Toc176177977)

[2.5 Related Work in Forecasting and Sentiment Analysis 14](#_Toc176177978)

[2.5.1 Forecasting: 14](#_Toc176177979)

[2.5.2 Sentiment Analysis: 15](#_Toc176177980)

[2.6 Research Gaps 16](#_Toc176177981)

[2.6.1 Sales Forecasting 16](#_Toc176177982)

[2.6.2 Sentiment Analysis for Consumer Reviews 17](#_Toc176177983)

[3 Methodology 17](#_Toc176177984)

[3.1 Research Design 17](#_Toc176177985)

[3.2 Data Collection: 19](#_Toc176177986)

[3.2.1 Data Cleaning and Preprocessing: 19](#_Toc176177987)

[3.2.2 Machine Learning Methods: 19](#_Toc176177988)

[3.3.3 Model Evaluation: 20](#_Toc176177989)

[3.3.4 Integration and Analysis: 21](#_Toc176177990)

[3.3.5 Decision Making: 21](#_Toc176177991)

[4 Data Analysis and Results 22](#_Toc176177992)

[4.1 Forecasting Analysis: 22](#_Toc176177993)

[4.1.1 Dataset 22](#_Toc176177994)

[4.1.2 Data Analysis and Results 22](#_Toc176177995)

[4.1.3 Sales Data Visualization 24](#_Toc176177996)

[4.1.4 Sales Forecasting 28](#_Toc176177997)

[4.2 Customer Feedback Analysis 35](#_Toc176177998)

[4.2.1 Dataset 35](#_Toc176177999)

[4.2.2 Data Preprocessing: 36](#_Toc176178000)

[4.2.3 Data Cleaning: 36](#_Toc176178001)

[4.2.4 Data Visualization 42](#_Toc176178002)

[4.2.6 Statistical Tests: 46](#_Toc176178003)

[4.2.7 Model Evaluation: 47](#_Toc176178004)

[5 Discussion and Implications 52](#_Toc176178005)

[5.1 Forecasting Analysis: 52](#_Toc176178006)

[5.2 Sentiment Analysis: 53](#_Toc176178007)

[6 Conclusion 54](#_Toc176178008)

[6.1 Summary and Contribution: 54](#_Toc176178009)

[6.2 Limitations: 55](#_Toc176178010)

[6.3 Future Scope: 56](#_Toc176178011)

[7 References: 57](#_Toc176178012)

[8 Appendix: 62](#_Toc176178013)

[8.1 Literature Review Paper’s Summary Table 62](#_Toc176178014)

[8.2 Recent Approaches Paper’s Table 63](#_Toc176178015)

[8.3 Regional Forecast 64](#_Toc176178016)

[8.4 Data Exploration for Sentiment Analysis: 68](#_Toc176178017)

[8.4 Code Snippets for Forecasting Analysis 71](#_Toc176178018)

[8.5 Code Snippets for Sentiment Analysis 75](#_Toc176178019)

[9 Ethical Form 78](#_Toc176178020)

**List of Figures:**

Fig 1- Business Applications of ML in E-commerce

Fig 2- Research Framework

Fig 3- Data Visualization

Fig 4 - Daily Sales

Fig 5 – Decomposition

Fig 6 – Seasonality

Fig 7 - Monthly Sales

Fig 8- Category Sales

Fig 9- Country Sales

Fig 10- Sales Funnel

Fig 11- Correlation Heatmap

Fig 12- SARIMA Monthly Sales

Fig 13 - Books Sales

Fig 14 - Electronics Sales

Fig 15- Fashion Sales

Fig 16- Toys Sales

Fig 17- Australia Forecast

Fig 18- Canada Forecast

Fig 19- LSTM Sales Forecast

Fig 20- Source

Fig 21- Outliers detection

Fig 22- Product rating

Fig 23- Missing count

Fig 24- Missing values

Fig 25- Missing Value Count

Fig 26- Ratings Distribution

Fig 27- Discounted Prices

Fig 28- Ratings Plot

Fig 29- Top 5 Products categories

Fig 30- Bottom 5 Products categories

Fig 31- Price Comparison

Fig 32- Correlation Heatmap

Fig 33- Word Cloud

Fig 34- Statistical Tests Comparison

Fig 35- p-value Comparison

Fig 36- Review Content

Fig 37- Sentiment Count

Fig 38- Positive Word Cloud

Fig 39- Neutral Word Cloud

Fig 40- Negative Word Cloud

Fig 41- Rating and Sentiment Scores

**List of Tables:**

Table 1- Numerical Summary

Table 2 – Categorical Summary

Table 3- SARIMA Metrices

Table 4- LSTM Metrices

Table 5- Numerical Table

Table 6- Categorical Table

Table 7 - Statistics with Outliers

Table 8- Statistics without Outliers

Table 9- Statistical Tests

Table 10- Sentiment Analysis Metrices

**Abbreviations Used in the Research:**

Artificial Intelligence - AI

Artificial Neural Network - ANN

Auto-Regressive Integrated Moving Average - ARIMA

Convolution Neural Network - CNN

Data Driven Decision Making - DDDM

Decision Tree - DT

Deep Neural Network - DNN

Gradient Boosting - GB

K-Nearest Neighbour - KNN

Long Short-Term Memory - LSTM

Machine Learning - ML

Mean Absolute Error - MAE

Naïve Bayes - NB

Natural Language Processing – NLP

Neural Network - NN

Random Forest - RF

Recurrent Neural Network - RNN

Seasonal AutoRegressive Integrated Moving Average – SARIMA

Sentiment Analysis - SA

Support Vector Machine - SVM

# Acknowledgement

First and foremost, my thanks go to Dr. Yuwang Chen for his tireless efforts, guidance, and support. His encouragement has made me realise the successful completion of my study. His eminent commentaries and dedication have been helpful to retain my focus and drive, particularly through the most demanding stages of my work.

Next to my respectable supervisor is a great faculty member at LUBS. Their extensive knowledge, devotion to teaching, and further encouragement support have created a robust base for my research efforts. The vibrant intellectual and friendly ambience at LUBS has played a very vital role in the honing of my analytical and research capabilities. I am indeed very grateful to all those teachers who very generously imparted their knowledge, challenged me to think critically, and created a strong belief in academic excellence.

My friends were always there for me, giving me company and encouragement. A lot that the experience of being a postgraduate student has been so invaluable is the many discussions and brainstorming sessions combined with the insurmountable sense of success and challenge sharing experiences. Their faith in me and constant support were also important in negotiating the pitfalls of the research work.

Special thanks go out to all my family members. Your unwavering support, belief in me, and selfless love have been the basis of reasons for which I could achieve the believed-to-be unattainable significant heights. In times of doubt, when their belief in me renewed my resolve, they made this research journey possible and greatly fulfilling.

Finally, it is a collective work that was enriched with the mentorship of my advisor, the companionship of my peers, and the support of my family. I am so much thankful for having people involved who made such a steep and invaluable contribution and support to my research.

# Abstract:

The importance of DDDM in e-commerce has increased greatly with the advent of sophisticated ML techniques. However, a gap in research exists between developed and effective techniques for forecasting sales through these methods and for performing customer sentiment analysis to improve business strategies. This research attempts to close the gap with two main objectives: improving e-commerce decision processes through ML and evaluating the effectiveness of ML techniques in sales forecasting and sentiment analysis.

The research uses SARIMA and LSTM models to forecast sales and VADER for sentiment analysis. The SARIMA captures seasonal trends and is adapted to the historical sales data, which can project future trends across various categories and regions. An LSTM model was also used to evaluate its ability to predict sales in a volatile e-commerce environment by capturing complex temporal dependencies. The VADER analysis was also done on the customer reviews to categorize their sentiment as positive, neutral, or negative.

Key findings from analyzing the sales dataset indicate that SARIMA performs well in predicting trends in sales, including that the electronics category is expected to experience a 41% sales drop by 2025, hence very key in providing actionable insights into inventory and marketing strategies. The LSTM model was much less accurate, although it showed trends in sales data over time. The SA result was overwhelmingly positive for customer reviews.

The findings of this research will be useful since advanced ML-based models significantly improve e-commerce decision-making. The models will have high forecasting precision because of the valuable insights generated from consumer data, which will allow businesses to remain competitive in a constantly developing market.

**Keywords**: data-driven decision-making, machine learning, SARIMA, LSTM, sentiment analysis, e-commerce, sales forecasting, and customer satisfaction.

Enhancing Data-Driven Decision-Making in E-Commerce Using Machine Learning Techniques.

# 1 Introduction

## Research Background

Since it was invented, e-commerce has undergone a lot of changes, where successive reinventions came with constant improvements in technology and changes in consumer behaviour. One of the characteristics of the early types of e-commerce websites is basic transactional sites having minimal features that would enable one to buy some goods and services. But all this changed dramatically with the advent of sophisticated algorithms and big data analytics, which are inextricably linked to e-commerce today (Huang and Rust, 2018). Accurately generating data is important in letting e-commerce companies analyze and predict customer behaviour for more precise and insightful predictive analytics. In e-commerce, ML techniques have changed DDDM by pulling off notable gains in customer satisfaction, revenue optimization, and operational efficiency. ML algorithms can analyze large-scale datasets to uncover patterns and insights that support informed business decisions. These qualities are then applied in e-commerce sites to provide personalized recommendations, optimize pricing strategies, and improve inventory control.

The last 20 years have seen exponential growth changing e-commerce and the way goods and services are consumed. Due to e-commerce platforms, businesses may offer their goods to customers wherever they are in the world (Ghaffari et al., 2024). This increase can be attributed to several factors, including evolving consumer preferences, and growing internet usage (Muhammad Tariq, 2018). Making decisions based on data insights is crucial in keeping a competitive edge in a field that includes intense competition and volatile market conditions (Chen et al., 2012).

"Data-driven decision-making" is one of the strategic ways to enhance consumer experiences, bring efficiency to operations, and shape business strategies through data analytics. It moved from a highly intuition-dependent process to a data analysis-based process (Zineb et al., 2021). Businesses can improve their understanding of demand forecasting, marketing personalization, and customer behaviour by implementing this method (Brynjolfsson and McAfee, 2017). The data is gathered from a variety of sources, such as interactions on social media, browsing histories, product reviews, and consumer transactions (Huang and Rust, 2018). Probably the biggest challenge e-commerce poses is the management of huge quantities of data it generates daily. Plug-in of ML approaches in data-driven decisions magnifies organizational operations and assures improved customer satisfaction.

The use of ML has revolutionized e-commerce and significantly improved several operational aspects. ML includes a broad range of methods and algorithms (Zhang et al., 2023). ML-enabled e-commerce platforms have demonstrated gains in decision-making processes, productivity, and personalized user experiences. By evaluating vast amounts of data, identifying trends, and forecasting outcomes to assist firms in making better decisions. Digital platforms provide a dynamic pool of anonymous and non-anonymized users who have voluntarily joined the e-commerce platform (Micol Policarpo et al., 2021).

The advancement of e-commerce has been greatly supported by ML, which makes it possible to make better data-driven decisions. By incorporating ML algorithms into e-commerce platforms, features like dynamic pricing, fraud detection, inventory management, and personalized consumer experiences have all been enhanced. E-commerce companies can boost sales, enhance customer satisfaction, boost operational efficiency, and obtain a competitive advantage in the online market by using these strategies (Zhang et al., 2023). For instance, ML divides the customer base according to their behaviour and preferences, which produces personalized product recommendations that increase revenue and satisfy customers (Ugwu, 2024).

ML improves inventory management by assisting companies in maintaining ideal stock levels, improving the precision of demand forecasts, and decreasing the possibility of stockouts (Anon, 2024). The ethical and successful application of ML in e-commerce depends on these problems being resolved (Rahul, 2024).

## Research Objectives

This research aims to enhance DDDM in the e-commerce sector through ML. By making use of two different datasets, it attempts to achieve the following critical research objectives:

* Evaluating the applications of ML that helps e-commerce companies make better decisions.
* Models like SARIMA and LSTM are used for sales forecasting and for evaluating sales values from past sales data and seasonal trends.
* Exploring how NLP analysis of customer reviews can improve marketing strategies and lead to effective product recommendations

## Research Structure

The abstract summarizes the research objectives, methodologies used, and results. The introduction stages the framework by stating the problem statement with the objective of the research. The literature review shows the existing literature gaps. Methodology chapters formulate the research design, techniques for data collection, and ML used in the case of SARIMA, LSTM, and NLP. This will be followed by the result of Sales Forecasting and Sentiment Analysis. The discussion will include the implications of the findings. The conclusion shall state the summary of the key contributions, future scope, and limitations.

# 2 Literature Review

## 2.1 Evolution of E-Commerce

The term "e-commerce" refers to both the online buying and selling of goods and services as well as the use of computers to increase a business's overall efficiency (Oudan, 2011). IBM claims that e-commerce benefits businesses in ways other than increased operational effectiveness. It allows you to establish previously unfeasible virtual relationships with staff members, vendors, suppliers, and customers. It also supports your company's resource efficiency, product sales, and first-rate customer service. They were mainly aimed at being adept at technology customers and businesses experimenting with online sales (Laudon and Traver, 2021). Businesses like Amazon and eBay were innovators in this area, setting standards for online shopping and creating the framework for future growth. Mobile commerce surged with the proliferation of smartphones and seamless mobile payment solutions letting consumers shop anywhere at their convenience.

The introduction of SSL (Secure Sockets Layer) encryption technology improved the protection of confidential data and made online shopping more accessible and secure for a wider range of users. Today, e-commerce has become part of global economics since it can adjust to new consumer behaviour and technologies.

## 2.2 Optimizing E-commerce with Machine Learning

ML, a branch of AI, is the process that enables algorithms to learn from data and form predictions without having to be specifically coded for every task. In the e-commerce sector, ML plays a critical role in automating and optimization of various business operations. This technology makes it possible to evaluate large datasets, allowing for the identification of trends and the guidance of decisions that increase profitability, customer satisfaction, and productivity. ML models can produce insights that guide strategic choices and operational enhancements by evaluating the data. These models handle and analyze data using regression analysis, classification, clustering, NLP, and other methods.

Algorithms that analyze customer data, such as prior purchases and browsing behaviours, provide personalized product recommendations to enhance the shopping experience and increase revenue (Ricci et al., 2011). To maximize revenue, dynamic pricing models adjust prices in real-time in response to market and demand (Kleinberg et al., 2018, Chen et al., 2012). By forecasting future demand using previous data and market trends, ML also improves inventory management by lowering overstocking and stockouts (Kumar and Reinartz, 2018, (Taylor and Letham, 2017).

AI-powered chatbots and virtual assistants enhance customer service by responding to queries and elevating complicated problems, while sentiment analysis of client feedback enables companies to proactively address problems (Huang and Rust, 2018, (Liu, 2022). Additionally, anomaly detection and clustering techniques used in ML can identify suspicious activities in real-time, which helps in fraud detection (Ngai et al., 2011). Businesses in the e-commerce sector may improve DDDM using ML, which will raise profitability, customer satisfaction, and efficiency.

Customer Segmentation Product Analysis Sales Forecasting

A black and white logo

Description automatically generated  A person holding a magnifying glass

Description automatically generated

Rating Quality **A light bulb and a shopping cart

Description automatically generated A black and white symbol

Description automatically generated** Price Watch

**ML in E-commerce**

 Smart Suggest

Price Forecasting A hand holding a coin

Description automatically generated 

Intelligent Pricing

**Fig 1-Business Applications of ML in E-commerce**

## 2.3 ML Techniques in E-Commerce Decision-Making

Interpretation of the models is a critical factor to take into consideration when using ML models for e-commerce decision-making. Recent research has shown that logistic regression is still a common methodology in e-commerce due to its ease of use and simplicity. Because each variable in logistic regression is categorical, it is particularly useful for forecasting customer attrition. Businesses can use logistic regression to identify the critical elements influencing client retention and implement focused measures to enhance it because of its ease of interpretation (De and Prabu, 2022).

DTs have become prominent because of their appropriate performance measures and easy interpretation in fraud detection and client segmentation. While DT frequently performs better than other statistical techniques, their effectiveness can be greatly increased when combined with ensemble techniques like bagging (Breiman, 1996). Because it can handle big datasets, RF is an ensemble strategy that performs better than many other models, yielding better performance measures like Area Under the Curve (AUC) and Receiver Operating Characteristics (ROC) values (Baghla and Gupta, 2022). These methods, which predict future patterns based on past sales data, are highly beneficial for initiatives like inventory control and sales forecasting.

In e-commerce, SVM is commonly utilized for classification purposes such as fraud detection and consumer classification. Kernel techniques enhance SVM sometimes outperforming ANNs depending on the dataset and data transformations used (Cortes and Vapnik, 1995). Because SVMs are effective at spotting complex patterns in transaction data that indicate fraudulent conduct, they provide great fraud detection capabilities. CNNs are used in e-commerce to research customer sentiment.

Applications of predictive analytics that make use of methods like Holt-Winters exponential smoothing and ARIMA include inventory management and sales forecasting. ARIMA models are a valuable tool for helping businesses forecast future sales based on past patterns when applied to time-series data. The Holt-Winters method extends exponential smoothing to include seasonality in the data and generate more accurate forecasts (Taylor and Letham, 2017). These methods are essential for maintaining optimal inventory levels and planning advertising strategies.

SA is another crucial application of ML in e-commerce, utilizing NLP techniques to analyze consumer reviews and feedback. This helps businesses understand customer satisfaction and improve their products and services accordingly. By analyzing the tone of customer feedback, businesses can enhance their client retention tactics and proactively address problems (Pang and Lee, 2008).

Two boosting algorithms that have shown great potential in increasing model accuracy by combining weak classifiers into strong classifiers are AdaBoost and XGBoost. These methods can increase performance, but they are complex and require extensive preparation to handle data correctly (Chen and Guestrin, 2016). These algorithms perform particularly well in scenarios requiring a high degree of precision, including client segmentation and sales forecasting. Naïve Bayes is used in sentiment analysis despite its shortcomings since it performs effectively with big datasets.

## 2.4 Challenges in Implementation of Forecasting and Sentiment Analysis

Although forecasting and SA provide numerous opportunities for enhancing decision-making in e-commerce, several challenges result from these techniques, which must be addressed. For instance, conventional techniques like ARIMA and Holt-Winters have very weak nonlinear relationships and seasonal patterns, often requiring a good deal of historical data to achieve satisfactory predictions (George and Jenkins, 2015). Although ML techniques, such as RF and GBT, handle nonlinearities and interactions that improve the accuracy of forecasts, they require significant computational resources and deep expertise in hyperparameter tuning. While the algorithm analysis for Walmart and Amazon sales forecasting is not as challenging as choosing and fine-tuning the performance of the best models, studies like (Elias and Singh, 2019), opine. Certain methods that follow lexicon-based approaches or are based on ML, like NB and SVM, may allow the context to be misunderstood at times due to nuances of a language, thus leading to misclassification in sentiment. Also, the hybrid technique, which combines lexical and ML techniques, is quite efficient but very hard to implement and often requires a lot of preprocessing. This process gets further complicated because of the variation in languages and the requirement for specific techniques of NLP. According to (Iswanto and Poerwoto, 2018), Indonesian text documents require specific techniques of NLP. There are substantial benefits from these advanced techniques available, but their implementation shall require a careful thought process over the challenges to bring an enhancement in e-commerce operations.

## 2.5 Related Work in Forecasting and Sentiment Analysis

### 2.5.1 Forecasting:

In the context of e-commerce, forecasting is the process of predicting future sales, demand, and market trends using historical data and statistical techniques. Forecasting in e-commerce is crucial, and the outcomes have a big influence on how decisions are made. Furthermore, businesses can improve their supply chain management system and people management by employing sales forecasts for the e-commerce platform to gain a better understanding of their financial status (Singh et al., 2020). Time-series forecasting has long employed conventional techniques like ARIMA and Holt-Winters exponential smoothing. ARIMA models can be used to project future sales based on historical patterns since they are effective at capturing the linear correlations present in time-series data (George and Jenkins, 2015). For data with strong seasonal trends, the Holt-Winters technique which includes factors for level, trend, and seasonality is very useful (Holt, 2004). In recent years, machine learning techniques have been added to forecasting models to increase their accuracy and durability.

Using methods such as RF and GBTs to handle non-linear correlations and interactions between variables has significantly improved forecast accuracy (Chen and Guestrin, 2016). Based on the findings of (Elias and Singh, 2019), have conducted a study on Walmart sales forecasting using ML algorithms. The success of this study was largely dependent on the application of numerous categorization algorithms to the sales data. Using the MAE assessment R^2 score, the researcher compared three different algorithms. This study aims to maximize MAE and R^2 score by adjusting each model's hyperparameters to assess the algorithm's correctness. In this study, the methods used were GB, RF and Extremely Randomized Tree. According to the results, the RF approach is the most effective. It has the lowest MAE assessment score (1979.4) and a high R^2 (0.94) score, which indicates its great accuracy when compared to the other algorithms (Singh et al., 2020).

(YU and LE, 2016) conducted series of studies on the Amazon sales forecast using statistical techniques. This study focused on forecasting sales in the future using the available data. Another study discusses the application of ML algorithms to explain sales forecasts. This study aims to evaluate the top model currently on the market and covers the major ML models that are frequently used in sales prediction. This paper identified the problem of how to mix intelligence-driven and data-driven models to select the optimal model based on business knowledge. This study employed the DT, NN, NB, RF, and SVM as its methodology. The neural network is at 70%, the decision tree is at 76%, the random forest is at 85%, the NB is at 83%, and the SVM is at the lowest 59%. Consequently, with its high accuracy model of 85%, the RF approach is the optimal one to select (Bohanec et al., 2017). Furthermore, NN’s particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have found broad use due to their ability to decide significant time-related trends in sales data (Hochreiter and Schmidhuber, 1997). Because these models understand long-term trends and dependencies, they are highly effective at long-term forecasting. Hybrid models that combine statistical and ML techniques are also gaining popularity.

### 2.5.2 Sentiment Analysis:

SA has become a vital tool in the e-commerce industry for understanding customer opinions and feedback (Ahmad Azrir et al., 2024). Text data derived from customer reviews is frequently processed and analyzed using NLP techniques. Text data is preprocessed using fundamental NLP techniques like tokenization, stemming, and lemmatization to prepare it for additional analysis (Bird et al., 2009). Researchers have suggested three primary approaches to address the SA problem the hybrid method, the ML method, and the lexicon-based approach. SentiWordNet and WordNetAffect are two vocabulary-based approaches that use a vocabulary of phrases to do SA. On the other hand, the corpus-based approach applies a variety of algorithms based on k-NN, Conditional Random Fields (CRF), and Hidden Markov Models (HMM) to the statistical analysis of a set of documents (Taneja et al., 2024).

The ML approach, also known as the supervised learning approach, is based on traditional ML and deep learning [DL] models. In this method, an ML or DL algorithm is fed a labelled dataset to train a SA model. There is widespread recognition of this sentiment classification technique. In SA, the NB classifier and SVM, two of the most popular classical ML techniques are used. DNN, CNN, RNN are the DL models that have been used for SA the most lately (Taneja et al., 2024). This method's first step analyzes the text using a lexicon-based technique, and the output is fed into an ML algorithm in the second phase as input training data. The effective dictionary is expanded using the results of the first step (Taneja et al., 2024).

User reviews are categorized by review mining as good, neutral, or bad. The results of the review prospecting step are considered during the review summarizing process, which automatically creates a summary for related review (Huang et al., 2019). ML models can be used to automatically classify a review's sentiment based on the words and phrases it contains (Ahmad Azrir et al., 2024). Before being fed into an SA model, text data is frequently cleaned and prepared using preprocessing techniques like stopword removal and stemming. The authors (Iswanto and Poerwoto, 2018) looked at how stemming and stopword removal affected SA accuracy in Indonesian text documents. The computerized SA of Twitter documents for the Indonesian language obtained up to 85.50% recall and accuracy. The study found that these techniques did not affect the accuracy of SA. After collecting Twitter data using Tweepy, the research's authors (Tyagi and Tripathi, 2019) used N-gram modelling and a KNN technique to categorize attitudes as positive, negative, or neutral. According to (Sairamvinay Vijayaraghavan and Basu, 2020) findings, DL algorithms are preferable to other ML algorithms when it comes to predicting the sentiment of reviews.

## 2.6 Research Gaps

### 2.6.1 Sales Forecasting

While many large e-commerce sites forecast sales accurately because of advanced prediction models, there is a noticeable lack of study in the literature about the application of these algorithms to broader scenarios. More precisely, there hasn't been much research done on using and adjusting predictive analytics across different datasets. Previous research has primarily focused on large-scale data, often ignoring smaller datasets or datasets from diverse sources. Furthermore, little research has been done on the potential combinations of various ML algorithms and traditional statistical methods to increase the accuracy and durability of sales forecasting models. It signifies that more in-depth analysis is necessary to evaluate the efficiency of hybrid models under various data conditions.

#### 2.6.1.1Possibility of Research:

This research focuses on applying the SARIMA and LSTM models to obtain sales predictions over the next few years. In SARIMA, one can extract seasonal components with the analysis of seasonal patterns and historic sales data, thus providing a statistical framework for prediction. Meanwhile, the LSTM can make use of its architecture to support learning processes with long-term data dependencies, which makes this model very competent for time series forecasting. These models are applied to this dataset to obtain robust, reliable sales predictions in this research. The SARIMA and LSTM models enable an e-commerce business to react dynamically regarding market demand.

### 2.6.2 Sentiment Analysis for Consumer Reviews

Research on the application of ML to SA in customer reviews about products is scant, particularly in the setting of combining such analysis into larger frameworks for enhancing DDDM in e-commerce using ML techniques. Most of the existing studies have been based on traditional methods for SA, rather than investigating the potential added value coming from more advanced ML algorithms. Hence, the gap underscores the need for advanced ML models to be integrated into SA to improve insight coverage and accuracy toward more informed, strategic, and e-commerce-based decision-making processes.

#### 2.6.2.2 Possibility of Research:

To evaluate the sentiment in customer reviews and comments, this study will employ NLP models. To extract deeper insights from text data, the study will make use of ML and other advanced SA approaches. Although there is a lot of literature on SA techniques, little is known about how these techniques may be applied to actual business operations. Even though analytics is the focus of a lot of studies, there aren't many useful techniques for applying SA. For marketing campaigns to be successful, it is important to understand consumer mindset.

# 3 Methodology

## 3.1 Research Design

In terms of the literature research, a framework is designed for the research regarding the enchainment of DDDM in e-commerce with the aid of ML techniques, which can be outlined as follows: data preprocessing, exploratory data analysis, text preprocessing, implementation of models of ML, evaluation of the models, and integration of insights towards decision-making.

**1**

**Sales Dataset Data Cleaning and EDA**

**Preprocessing**

**2 Descriptive Statistics**

**Cleaning Data Visualisation   
 Transformation**

**Text Preprocessing Removing white spaces**

**3**

**Sales Forecasting SARIMA, LSTM**

**Machine Learning   
 Methods**

**4 Sentiment Analysis NLP**

**Model Evaluation Metrics MSE,MAE, Accuracy, Precision, F1 Score, AUC,ROC**

**Validation Train- test split**

**5**

**Integration and Sales Insights  
 Analysis**

**6 Sentiment Insights**

**Decision Making**

**Fig 2- Research Framework**

## 3.2 Data Collection:

The datasets in the sales category may be gathered using transactional data, which captures information regarding the sale, the product sold, and customers' history of transactions.

### 3.2.1 Data Cleaning and Preprocessing:

* *Cleaning and Transformation:* To handle missing values, eliminate duplicates, and correct inconsistencies in data types to get a consistent and error-free dataset.
* *Removing White Spaces:* This will be done to remove all white spaces within the text for consistency of data.

A diagram and pie chart

Description automatically generated

Fig 3 – Data Visualization (Saleem et al., 2021)

* *Descriptive Statistics and Data Visualization:*

It is the process of exploratory analysis for comprehending data distribution, anomaly detection, and uncovering underlying patterns. Descriptive statistics and different types of visualization, like scatter plot, box plots etc. are the techniques included.

### 3.2.2 Machine Learning Methods:

1. *Sales Forecasting:*

* Models used: the SARIMA and the LSTM for sales forecasting.
* SARIMA: This will be applied to handle seasonality in sales data and help capture seasonal patterns with trend components.
* LSTM—This type of Recurrent Neural Network (RNN) effectively captures long-term dependencies in time series data, is suitable for RNN is proficient at capturing long-term dependencies existing in time series data and hence ideal for making forecasts based on historical sales data.

1. *Sentiment Analysis:*

* NLP involves the treatment of customer reviews and feedback with NLP techniques to derive sentiment insights.

### 3.3.3 Model Evaluation:

* These metrics, which may be used to evaluate the performance of the trained model, include Accuracy, Precision, MSE, MAE, Recall, and F1 Score, ROC AUC.
* *Accuracy-* An ML metric that measures the proportion of correct predictions in total predictions (Blog, 2023).

* *Precision-* Precision is the fraction of the number of correctly predicted positive classes to the total number of classes that were predicted as positive (Blog, 2023).
* *Recall-* Recall defines the ratio of the number of true positive predictions to all actual positive samples in the dataset (Blog, 2023).
* *F1 Score-* The F1 score is a measure of how accurate the model is in classifying instances as either positive or negative for the precision and recall values. It ranges from 0 to 1, where 1 is best. (Blog, 2023).
* *MAE:* The mean absolute error is just the average of the absolute differences between the actual and predicted values in the dataset (Chugh, 2020)

Where,

y^​i​ = predicted value of y

yi​ = mean of y

* *ROC AUC-* The two most common metrics in machine learning, which are used in binary classification models, are AUC and ROC curves. The ROC-AUC can be seen as an integrated measure of performance across all possible thresholds of classification (Blog, 2023).
* *MSE-* The mean squared error is just an average of the squares of the differences between the values at hand and the predicted ones in each dataset (Chugh, 2020).

### 3.3.4 Integration and Analysis:

1. *Sales insights* from the sales forecasting models drive inventory management, marketing strategies, and financial planning.
2. *Sentiment Insight:* This could be obtained from SA to help improve customer experience, handle complaints, and enhance products.

### 3.3.5 Decision Making:

It helps derive insights from sales forecasting and sentiment analysis into driving data-driven decision-making processes.

# 4 Data Analysis and Results

## 4.1 Forecasting Analysis:

### 4.1.1 Dataset

For the sales forecasting the Customer360Insights.csv dataset was chosen because it contains the e-commerce transactions of customers with 2000 rows and 23 columns and has data ranging from 2019 to 2023. It contains both categorical and numerical data, giving information on customer engagement and behaviour.

* ***Customer Demographics:*** This includes FullName, Gender, Age, CreditScore, and MonthlyIncome. These variables capture demographic information about the customer base.
* ***Geographical Data:*** It contains country, state, and city that will help measure the regional sales performance of an organization.
* ***Product Information:*** Category, Product, Cost, and Price enable trend analysis in products, profitability assessment, and inventory optimization.
* ***Transactional Data:*** This contains SessionStart, CartAdditionTime, OrderConfirmation, OrderConfirmationTime, PaymentMethod, and SessionEnd.
* ***Details After Purchase:*** It contains OrderReturn and ReturnReason columns which can be used for return rate calculations.

### 4.1.2 Data Analysis and Results

#### 4.1.2.1 Data Exploration:

The dataset is relevant to customer interaction and sales transactions. From Table 1 it can be infer that the main columns were 'CustomerID', 'Age', and 'Price,' along with several timestamps. The columns with null values were 'OrderConfirmationTime,' 'PaymentMethod,' 'OrderReturn,' and 'ReturnReason.' No missing values were observed, and there were no duplicate rows in the dataset.

It ended up with 1700 entries after the date columns had been converted into datetime format and the handling of missing values in the dataset was done by dropping the rows having N.A. values. Numerical columns were inputted using medians, while categorical columns used modes to fill. Some key statistics include the mean age as 44.6, the average price being 207, the most common category 'toys,' and the leading campaign for that was 'Instagram-ads.'.

**Table 1**- **Numerical Summary**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Statistics | CustomerID | Age | CreditScore | MonthlyIncome | Cost | Price | Quantity |
| count | 1700 | 1700 | 1700 | 1700 | 1700 | 1700 | 1700 |
| mean | 1614.432 | 44.605 | 690.251 | 5606.441 | 133.16 | 206.985 | 3.549 |
| std | 351.946 | 16.394 | 51.068 | 1448.135731 | 257.84 | 388.323 | 1.69 |
| min | 1001 | 18 | 600 | 3001 | 5 | 8. | 1 |
| 25% | 1313.75 | 31 | 645 | 4428.25 | 15 | 25 | 2 |
| 50% | 1619.5 | 44 | 692 | 5689 | 30 | 50 | 4 |
| 75% | 1922.25 | 59 | 733 | 6871 | 60 | 100 | 5 |
| max | 2200 | 72 | 780 | 7999 | 1000 | 1500 | 6 |

**Table 2 – Categorical Summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Count | unique | Top Value | Frequency |
| Full Name | 1700 | 1103 | Crystal Mitchell | 4 |
| Gender | 1700 | 2 | Female | 879 |
| Country | 1700 | 9 | Canada | 222 |
| State | 1700 | 18 | Maharashtra | 123 |
| City | 1700 | 52 | Yokohama | 57 |
| Category | 1700 | 5 | Toys | 360 |
| Product | 1700 | 20 | Magazine | 105 |
| CampaignSchema | 1700 | 6 | Instagram-ads | 300 |
| PaymentMethod | 1700 | 4 | Credit Card | 583 |
| ReturnReason | 1700 | 5 | Poor Value | 1519 |

### 

### 4.1.3 Sales Data Visualization

A graph of blue lines with red numbers

Description automatically generated

Fig 4 - Daily Sales

The graph provides the daily sales trend from 2019 to 2023, reflecting the fluctuations and seasonality that are present. Sales peaks are identified when the nature of crucial values seen, such as 409, 9, 72, and 171, are variations that connote serious sales events.

#### Decomposition

A graph of time and time

Description automatically generated with medium confidence A graph of a number of blue lines

Description automatically generated with medium confidence

Fig 5 - Decomposition Fig 6 - Seasonality

For the decomposition of time series of sales, an additive model was preferred since the data does not vary much over time. In this trend for sales, it crests about 800 units, the trend is flat and peaks at about 600 units, and some minor seasonal oscillations and residuals go up and down between -200 to 200 units.

A graph with numbers and lines

Description automatically generated

Fig 7 - Monthly Sales

This plot shows a trend in the monthly sales. There is an observed sales trend from December 2018, with peaks covering some months like December 2021 with 13,749 units and then dropping to 9968 in March 2022.

A graph with blue rectangles

Description automatically generated

Fig 8- Category Sales

The chart represents the sales data by category. It can be viewed that the largest category is electronics, with 214,390 items sold, and the lowest is toys, with 8,625. This chart details how much revenue is divided among categories.

A graph of sales by country

Description automatically generated

Fig 9- Country Sales

In the sales by country representation, Canada leads with the best sales at 45,173, while the UK has the lowest at 27,439. There is quite a huge difference between the two, therefore spotlighting Canada's strong market performance as opposed to the UK's.

A graph with blue bars

Description automatically generated with medium confidence

Fig 10- Sales Funnel

This Sales Funnel chart indicates a constant volume of events flowing through 1,700 visits, 1,700 adds to the cart, and 1,700 purchases.

A screenshot of a graph

Description automatically generated

Fig 11- Correlation Heatmap

The correlation heatmap expresses how some of the key variables in a dataset relate to one another. More specifically, 'Cost' and 'Price' are perfectly positively correlated at 1, indicating they move together either up or down. All other correlations should be kept very low for example, 'MonthlyIncome' and 'Quantity' correlate only 0.037.

### 4.1.4 Sales Forecasting

#### 4.1.4.1 SARIMA MODEL:

One of the creations for handling seasonal data patterns is the Seasonal Auto-Regressive Integrated Moving Average (SARIMA). It includes seasonal lags and differencing to deal with periodic fluctuations. For this reason, considering the possibility of modelling seasonal spikes and trends, SARIMA would be very useful in sales forecasting within e-commerce (Kusawa, 2023).

A number with a white background

Description automatically generated with medium confidence

For the SARIMA model, statsmodels implementation of SARIMAX was used. This model is fitted to the historic sales data by a grid search for the optimal parameters of the model. Cross-validation using TimeSeriesSplit has also been done to check the stability of the model across different splits of data. For the sake of robustness MAE, RMSE, and MAPE were taken into consideration while evaluating this model. Resampling the dataset at a monthly frequency and summing up the sales data to prepare it so that it can be effective for modelling was done. An order of (1,1,1) and seasonal\_order of (1,1,1,12) for the SARIMA model to capture seasonal effects and trends over time were implemented. This is relevant when modelling seasonal changes that are usually inherent within e-commerce sales. The fitted model was used to generate out-of-sample forecasts until the end of 2024. This consisted of predicted sales values, along with confidence intervals that showed ranges of possible sales outcomes.

A graph with numbers and lines

Description automatically generated

Fig 12- SARIMA Monthly Sales

|  |  |
| --- | --- |
| Metric | Value |
| MAE | 2407.915 |
| RMSE | 3094.472 |
| MAPE | 52.163 |
| Cross-Validation MAE | 3392.431 |
| Cross-Validation RMSE | 3962.842 |
| Cross-Validation MAPE | 73.038 |

Table 3- SARIMA Metrices

The plot discloses the forecast, which includes observed and forecasted sales from January 2019 to December 2024 and some 2025 values. The blue line shows the observed sales and the orange line represents the forecasted sales, which is surrounded by a shaded area, forming a confidence interval. Such visualization helps uncover future trends in sales, very vital in strategic planning and decision-making on inventory and marketing.

For example, actual sales were highest in August 2021, ranging at 13,749 units, while forecasted sales were around 6,730 units in October 2024. From October 2023, 9066 to January 2024, 649 there is an increase in sales and after that again increase in 2025. The dispersion in probable sales is represented by the shading of a confidence interval, which gives a more realistic slant toward the business strategy. With SARIMA within the mechanism of forecasting, e-commerce will leverage historical data in optimizing their operations through predictive insights. It therefore places advanced forecasting at the core of efficient and effective e-commerce business operations.

#### 4.1.4.2 Forecasting Results

##### 4.1.4.2.1 Category Forecast

A graph with blue lines

Description automatically generated

Fig 13 - Books Sales

The plot represents a trend from 2020 to 2024, the sale of these books peaked in 2024 at 2390 units sold. A little going down, the forecast for 2025 gives this at 1932 units. This shows the prediction of sales of books using the SARIMA model.

A graph showing the sales of electronics

Description automatically generated with medium confidence

Fig 14 - Electronics Sales

The plot shows the future prediction of sales for the electronics category. The trend of historical sales peaks in 2022 at 54,520 units. The SARIMA model further projects the sales decrease to 32,165 units in 2025. Such analysis is not only indicative of future trends but has also located areas of improvement in the sales strategy.

A graph with a line and numbers

Description automatically generated

Fig 15- Fashion Sales

This is the fashion category sales forecast. The historical trend indicated that the sales were variable from the year 2020 to 2024, peaking in 2021 with 5,270 units. The forecast for 2025 was slightly lower, where an approximate sale of about 4,744 units.

A graph with blue lines

Description automatically generated

Fig 16- Toys Sales

The following is the visual analysis for the sales forecast for toys. The peak of the historical data reached 1915 units in 2022 and decreased to 1530 units in 2023, but the forecasted quantity will increase to 1718 units in 2025.

##### 4.1.4.2.2 Regional Forecast

A graph with a line

Description automatically generated

Fig 17- Australia Forecast

The above graph shows the sales forecast for the Australia model based on trends through 2020-2024. A historical series in blue bounces around with a peak of 13,605 in 2022. The forecasts sales in 2025 are slightly higher at 7,042.

A graph with blue lines

Description automatically generated

Fig 18- Canada Forecast

The visual of the actual sales from 2020 up to 2024, peaking in 2022 at 11,965 units, with a forecast for 2025 to be around 7,308 units for Canada is shown.

**Further Regional Forecast for other countries can be found in the appendix.**

#### 4.1.5 LSTM Model

LSTM is a sort of RNN designed to overcome the vanishing gradient problem occurring in traditional RNNs. This peculiarity makes LSTM very appropriate for tasks on sequential data, like time-series forecasting, which is very common in e-commerce sales forecasting (Wikipedia Contributors, 2018).

The LSTM model provides the capture of complex patterns in data that are very relevant for good performance in forecasting. Specifically, in time series analysis, the algorithm runs very efficiently since it can learn and hold temporal dependencies.

The resampling of the sales data is done on a monthly frequency, removing missing values to guarantee continuity is done., and MinMax scaling is applied to the data, so it's well-prepared for any type of time series analysis. Then we have the data that is in form where time series analysis can take place. Later, the data will be scaled between 0 and 1 for normalizing input.

This ensures the sliding window approach whereby only the last 12 months' sales data are part of any input sequence in the dataset. Seasonal and temporal patterns are important to be captured in the data for an accurate forecast. This LSTM model shall be created by an LSTM layer of size 50, followed by a dense output layer. The ADAM optimizer will be used together with the mean squared error as the loss function. Adam is a method for efficient stochastic optimization. It borrows from two popular optimization methods: AdaGrad and RMSProp. This makes an LSTM model perform even better. (Chang et al., 2018). During training, the weights should be adjusted step by step to obtain the least possible prediction error, with the number of epochs running at 200 and a batch size of 32.

A graph with blue lines and numbers

Description automatically generated

Fig 19- LSTM Sales Forecast

The results are for applying the LSTM model on sales forecasting against the e-commerce dataset over the past years. The observed sales data has wide variations marked by well-defined peaks and downs. In the blue line, the indication is of actual sales, which fluctuates. For example, in January 2022, it peaked at 13,749, then went down to 2,666 in mid-2022. These fluctuations show the unpredictability of sales patterns within markets. The red dotted line shows forecasted sales, which hold steadier between 6,000 and 7,000, with a peak of 6,963 in January 2024. With these forecasts, the business can fine-tune inventory and marketing plans. This will enhance operational efficiency, maximize customer satisfaction, and give the business a competitive advantage. This sets the highly fluctuating observed sales data against smoother and more stable forecasted values.

|  |  |
| --- | --- |
| Metric | Value |
| MAE | 2138.62 |
| RMSE | 2503.85 |
| MAPE | 44.23% |
| R^2 | 0.04 |

Table 4- LSTM Metrices

## 4.2 Customer Feedback Analysis

### 4.2.1 Dataset

The dataset used in this study is sourced from an extensive set of open-source sales data for e-commerce products on Amazon, available on Kaggle. This dataset contains information obtained from over 1000 product ratings and reviews from the official website of the company. It includes product information on prices, ratings, reviews, and categories. The dataset has 1465 rows and 15 columns, which can be divided into categories and numerical variables.

|  |  |  |
| --- | --- | --- |
| Variable Type | Variable Name | Description |
| Numerical Variable | Discounted Price | Final price after discounts in Indian Rupees (₹). |
|  | Actual Price | Initial cost before discounts in Indian Rupees (₹). |
|  | Rating | Average rating of the product from 1-5 and total rating count. |

Table 5- Numerical Table

|  |  |
| --- | --- |
| ****Variable**** | Description |
| Product ID | Unique product identifier |
| Product Name | Product name |
| Category | including main category and subcategories |
| Discount % | Discount percentage |
| About the Product | Product features and details |
| User ID | Unique user identifier |
| Username | Reviews users have provided |
| Review ID | Unique review identifier |
| Review Title | Review title |
| Review Content | Full review text |
| Image Link | URL of product image |
| Product Link | URL of the Amazon product page. |

Table 6- Categorical Table

### 4.2.2 Data Preprocessing:

It is a crucial phase in the data analysis process because it ensures that the dataset is consistent and ready for analysis. This specific dataset needed to be preprocessed to manage missing values, correct invalid rating entries, and fix format issues with the price column. Several steps were necessary to transform the unstructured data into a format suitable for statistical analysis and ML models.

### 4.2.3 Data Cleaning:

The dataset needed to go through several cleaning steps to ensure data integrity and consistency. Initially, the 'discounted\_price' and 'actual\_price' columns were eliminated from commas and currency symbols for simpler numerical analysis.

Before converting the 'discount\_percentage' column from a string to a float type, the percentage sign was removed. Furthermore, there was an unusual string **"|"** in the "rating" column. I checked the Amazon page to view the rating and found that B08L12N5H1's product ID has a rating of 4. For this reason, I will also give the item a 4.0 rating. It was converted to a numerical data type before this was deleted and replaced with the default rating of 4.0.

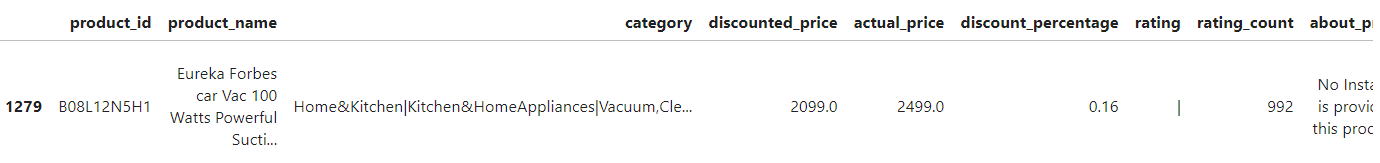


Fig 20- Source***:*** [*https://www.amazon.in/Eureka-Forbes-Vacuum-Cleaner-Washable/dp/B08L12N5H1*](https://www.amazon.in/Eureka-Forbes-Vacuum-Cleaner-Washable/dp/B08L12N5H1)

Additionally, commas were removed from the 'rating\_count' column before its conversion to a numerical format. These modifications ensured that every numerical column was in the proper format for analysis and modelling.

#### 4.2.3.1 Outliers Detection:

The results of modelling and data analysis can be significantly impacted by outliers. Consequently, it was essential to handle and identify dataset outliers. The Interquartile Range (IQR) approach was applied to search for outliers in the 'discounted\_price' and 'rating' columns.

***This is how the IQR is calculated:*** Data points that are outside of the range of Q3 + 1.5 \* IQR or below Q1 - 1.5 \* IQR are called outliers. By removing these outliers, we improved the robustness of the research and ensured that the data used for it accurately reflected the overall patterns.

A comparison of a bar graph

Description automatically generated

Fig 21- Outliers detection

**Table 7 - Statistics with Outliers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Statistics | discounted\_price | actual\_price | discount\_percentage | rating | rating\_count |
| Count | 1465 | 1465 | 1465 | 1465 | 1463 |
| Mean | 3125.31 | 5444.99 | 0.477 | 4.10 | 18295.54 |
| Std | 6944.30 | 10874.83 | 0.216 | 0.29 | 42753.86 |
| Min | 39.00 | 39.00 | 0.00 | 2.00 | 2.00 |
| 25% | 325.00 | 800.00 | 0.32 | 4.00 | 1186.00 |
| 50% | 799.00 | 1650.00 | 0.50 | 4.10 | 5179.00 |
| 75% | 1999.00 | 4295.00 | 0.63 | 4.30 | 17336.50 |
| Max | 77990.00 | 139900.00 | 0.94 | 5.00 | 426973.00 |

**Table 8- Statistics without Outliers**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Statistics | discounted\_price | actual\_price | discount\_percentage | rating | rating\_count |
| Count | 1176 | 1176 | 1176 | 1176 | 1176 |
| Mean | 950.004 | 2129.953 | 0.495 | 4.123 | 19223.82 |
| Std | 884.204 | 2414.82 | 0.218 | 0.229 | 43754.84 |
| Min | 39 | 39 | 0 | 3.6 | 4 |
| 25% | 299 | 699 | 0.37 | 4.0 | 1229 |
| 50% | 599 | 1299 | 0.53 | 4.1 | 5672.5 |
| 75% | 1399 | 2499 | 0.65 | 4.3 | 17398.75 |
| Max | 4499 | 19999 | 0.94 | 4.7 | 426973 |

#### 4.2.3.2 Data Transformation:

Further modifications were made to enable insightful analysis. Two distinct columns, 'category\_1' and 'category\_2', were generated from the 'category' column which included hierarchical categories divided by '|'. This separation made a more detailed analysis of product categories possible.

A new category column called "rating\_rank" was made using the 'rating' column. Based on rating results, products were divided into five groups: "Very Poor," "Poor," "Average," "Good," "Very Good," and "Excellent." This adjustment made it possible to understand the product rating distribution in an improved manner.

A screenshot of a computer

Description automatically generated

Fig 22- Product rating

#### 4.2.3.3 Handling Missing Values:

Maintaining the integrity of the dataset requires careful handling of missing values. It was found that there were some missing values in the 'rating\_count' column after the dataset was examined.

A screenshot of a computer program

Description automatically generated

Fig 23- Missing count

A screenshot of a computer

Description automatically generated

Fig 24- Missing values

A graph of a bar

Description automatically generated

Fig 25- Missing Value Count

These missing values were imputed using the 'rating\_count' column's median value. Since the median ensures a more balanced test and reduces the impact of outliers, it is the best choice for imputation. This method prevents missing values from distorting the analysis results and preserves the dataset's overall quality. There are no duplicates in the dataset.

#### 4.2.3.4 Data Normalization:

Standardizing numerical columns ensured that the data was all the same size, for ML approaches. A mean of zero and a standard deviation of one required scaling of the 'discounted\_price' and 'actual\_price' columns. Standardizing these columns improves performance and speeds up the convergence point of many ML algorithms.

A graph with a line going up

Description automatically generated

Fig 26- Ratings Distribution

Most of the reviews were between 4.0-4.375. None of the products had a rating less than 2.0. This shows that the distribution is left-skewed.

### 4.2.4 Data Visualization

A graph showing a distribution of discounted prices

Description automatically generated

Fig 27- Discounted Prices

The plot shows that most products have a lower price. Over 1,200 products are within the modal price range in the sample of less than ₹10,000. A clear decline can be witnessed in the number of products with price increments. For example, only a few products cost more than ₹50,000, and fewer than 100 cost more than ₹20,000. This means the online store is targeted at selling products at reasonable prices that appeal to more customers hunting for low-cost alternatives.

A bar graph with numbers and a number of bars

Description automatically generated with medium confidence

Fig 28- Ratings Plot

The rating bar plot makes it easier to view the product rating distribution in the dataset. 244 goods have this ranking; 4.1 has been the most frequently given rating for any goods. Very few ratings go below 3.5. Most products in this distribution have obtained positive ratings and hence show overall very good customer happiness.

A graph with different colored rectangles

Description automatically generated

Fig 29- Top 5 Products categories

A graph with different colored squares

Description automatically generated with medium confidence

Fig 30- Bottom 5 Products categories

From the pictures, there are products under their respective categories. The top five categories had: USB cables with 233 products, a smart watch with 76, smartphones at 68, smart televisions at 63, and in-ear headphones with 52. The bottom five categories, which include impact vacuum bags, data cards, multimedia speakers, colorful paper and battery chargers had one product each. The disparity truly shows the closeness of the platform toward inventory strategy.

A graph showing the difference between the price and rating

Description automatically generated

Fig 31- Price Comparison

From the scatter plot it is observed that most of the products priced below ₹20,000 come under the rating tags from 4.0 to 4.5 Ratings above 4.0 showing a big variation in satisfaction across price brackets.

A diagram of a heatmap

Description automatically generated

Fig 32- Correlation Heatmap

The correlation heatmap points to a very strong positive relationship of 0.96 between the actual price and the discounted price. The ratings show a very weak positive relation of 0.12 with both the actual and discounted prices, while the rating count return is almost negligible with other variables indicating distinct relationships in the dataset.

A close up of words

Description automatically generated

Fig 33- Word Cloud

The word cloud of the content of the review shows positive adjectives that have been highly used, including "good," "product," and "easy." Commonly used words that identify rated products include words such as "phone" and "cable." Their issues are rarely identified by terms such as "issue" and "problem," so therefore, proving that there are few negative experiences.

**Data Exploration can be found in the appendix.**

### 4.2.6 Statistical Tests:

|  |  |  |
| --- | --- | --- |
| Test | Statistic | p-value |
| ANNOVA | 259.80 | 1.465814e-107 |
| Chi-Square | 4809.80 | 9.898843e-01 |
| Correlation | 0.12 | 2.992638e-06 |

Table 9- Statistical Tests

The statistical tests reveal some very important facts about the data set. According to the results of the ANOVA test, there is a significant difference in Actual Cost the p-value is less than 1.47e-107, and the F-statistic is 259.80. Chi-Square test results state that categories and ratings do not correlate significantly; the chi2 statistic is 4809.80, while the p-value is 0.99. Correlation Analysis shows a slightly positive relationship between actual cost and its rating. The correlation coefficient is 0.12, with a p-value less than 2.99e-06.

A graph showing a comparison of statistical tests

Description automatically generated

Fig 34- Statistical Tests Comparison

A graph of a comparison of values

Description automatically generated

Fig 35- p-value Comparison

### 4.2.7 Model Evaluation:

#### 4.2.7.1 Feedback Insights

In the present research, customer reviews are analyzed by using an e-commerce dataset for NLP. In the case of sentiment analysis, VADER (Valence Aware Dictionary and sentiment Reasoner) was applied to the dataset since it is an efficient handler of texts from social media and reviews of products. This was very important since VADER reliably supports the correct classification of customer reviews into positive, negative, and neutral sentiments. This can give insight into customer opinions and behaviour, hence better marketing strategies, recommendations of products, and improvements in customer satisfaction in e-commerce.

Customer reviews were classified as positive, neutral, or negative by turning them into cleaned textual data and applying sentiment scores. The output displays a list of the reviews processed, assigned sentiment, and compound score, which is a sum of sentiment intensities for each review. As an example, high-scoring reviews of 0.9574 and 0.9895, like "Seems sturdy,", "Charging is fine too No complaints…,", and "I ordered this cable to connect my phone to Android auto car…", have been put under good reviews.

A screenshot of a computer

Description automatically generated

Fig 36- Review Content

A graph with a green bar

Description automatically generated

Fig 37- Sentiment Count

The bar chart and sentiment count of customer reviews give an idea of how customers are, on average, satisfied with these products. There are 1,449 positive reviews and just 15 negative reviews for the same product, along with 1 neutral review. A huge skew toward the positive reveals, in general, that customers do tend to feel satisfied with their purchase. Those reviews are highly positive, which at the same time denotes the good quality of the products and their performance, which unequivocally reflects positively on the company. Such a positive feeling is very important for branding and could be strategically used in marketing for customer attraction and retention.

A close up of words

Description automatically generated

Fig 38- Positive Word Cloud

***Word Cloud: Positive Reviews****:*

* ***Prominent Words****:* "product," "good," "use," "cable," "one," "quality," "phone," "tv"
* ***Insights****:* Customers frequently mention the quality and usability of the products, indicating these are key aspects driving positive sentiment. Terms like "good," "great," and "best" suggest high satisfaction with product performance and value.

A close-up of words

Description automatically generated

Fig 39- Neutral Word Cloud

***Neutral Reviews Word Cloud:***

***More Prominent Words:*** “quick”, “product”, “packing”, “opened”, “just”, “good”

**Insights:** Terms used in neutral reviews are general and do not carry strong sentiment, hence they indicate neither significant satisfaction nor dissatisfaction. Words like "quick" and "packing" are some indications that the comments in the review might have been about the packaging or delivery.

A close up of words

Description automatically generated

Fig 40- Negative Word Cloud

***Word Cloud: Negative Reviews***

**Most Used Words:** "product," "bad," "service," "quality," "unit," "buy," "battery," "cable"

**Insights:** Bad reviews discuss product quality and service. Words like "bad," "poor," "problem," and "battery" identify areas for correction.

A graph with red dots and blue dots

Description automatically generated

Fig 41- Rating and Sentiment Scores

The scatter plot illustrates that the correlation between the ratings and sentiment scores was strong, with 1,449 reviews being positive and around ratings of 4.0 or higher. There were 15 negative reviews which fell around ratings under 3.5.

SA is the main facilitator of tactical decisions about research. For example, if the business knew about the change in sentiments by customers, it could alter ad campaigns and product portfolios before it is already too late to be relevant in the changing market. This would not only avoid the pitfalls of the business with such a data-driven ideology but also enhance the shopping experiences of the consumer through customization and satisfaction. These inbuilt sentiment analyses work as a correcting mechanism in case the initial strategies may not exactly meet customer expectations. While continuously analyzing the feedback, a company will be able to make necessary changes. Thus following through such an iterative process of strategies, customer-centricity would be borne out at the end eventually, customer satisfaction leads to business success.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | | Value | |
| Correlation (Rating, Compound) | | 0.197989 | |
| Accuracy | | 0.99 | |
| Precision | | 0.99 | |
| Recall | | 1.00 | |
| F1 Score | | 0.99 | |
| MSE | | 0.01 | |
| MAE | | 0.01 | |
| ROC AUC | | 1.00 | |
|  | **Predicted Negative** | | **Predicted Positive** |
| Actual Negative | 2 | | 3 |
| Actual Positive | 0 | | 288 |

Table 10- Sentiment Analysis Metrices

The analysis reveals that the model performed well, with an accuracy of 99%, a precision of 99%, a recall of 100%, and the very good quality of the test dataset is attested by the high ROC AUC of 1.00. The low MSE and MAE, both at 0.01, further confirm how close the predictions are to the real values, hence improving decision-making in e-commerce.

## 5 Discussion and Implications

### 5.1 Forecasting Analysis:

The models, like SARIMA and LSTM, are provided with the fundamental tools an e-commerce company needs to accurately predict future sales trends. For example, the SARIMA model is predicting a decline in electronics sales of 41%, from 54,520 units sold in 2022 down to 32,165 units in 2025. Meanwhile, toys are supposed to have increased by 12% from 1,530 units sold in 2023 to 1,718 units in 2025, and special promotions or inventory changes may be in order. Sales in the UK would fall from 27,439 units to 5,209 units by 2025, and hence there is more marketing effort to be made in that area.

Sales in the fashion category are forecasted to decline slightly from 5,270 units in 2021 to 4,744 units in 2025. However, sales of toys should be increased slightly to 1,718 units by 2025 from 1,530 units in 2023, which was down from 1,915 units in 2021.

The LSTM model predictions stabilize that sales will remain between the ranges of 6,000-7,000 units in the year 2024, even though the past sales had big ups and downs, peaking at 13,749 units in January 2021.

These are the means through which any e-commerce business can make sound decisions, mitigate risks, and be competitive by the projection of the trends likely to take place in the market.

### 5.2 Sentiment Analysis:

1. ***Identifying the Significantly Positive Elements:***

**Marketing Strategy:** The ads will be based on positive sentiment, comprising elements such as "product," "quality," "good," and "use," of which people are talking the most. An ad with features like this would attract more customers and at the same time create more reliability. For example, the ad for a "good" or "great" consumable product that is durable and usable will have people following the quality.

**Suggestions to the Users:** Users giving a high rating to a product can then suggest it to other users. For example, if a particular kind of cable for the phone continues getting appreciated, further, the same brand of accessories or those behaving similarly can be suggested to make the shopping experience better

1. ***Managing Negative emotions:***

**Marketing Strategy:** Accepting the challenges on issues, such as "poor quality," "bad service," and problems with "battery" or "unit," provides businesses with the opportunity to come up with solutions. For instance, if the clients are not happy with the life of the battery, marketing can reflect the improved performance of the battery in the next versions of the products or concerns and general solutions in their marketing papers.

**Enhancements to the Product:** How to pinpoint the exact issues that lead to negative reviews. Businesses can focus on problem-specific solutions to enhance customer support and service policies to reduce bad experiences if the common complaint is related to "service".

1. ***Real-time Feedback and Modification:***

**Marketing Adaptation:** Sentiment analysis is a continuous process, allowing for real-time tweaking of marketing strategies. If a newly launched product receives negative feedback, then changes can be promptly made to avoid an increase in the issue.

**Product Development:** Constant analyses help to alter product characteristics to best-fit customer needs. Through the insights given via reviews, one can predict the development of future products and to a very large extent organically align them with the wants and needs that are expressed by the customer.

# 6 Conclusion

## 6.1 Summary and Contribution:

In this research, I explored the application of advanced ML SARIMA and LSTM models could be applied and provide significant improvements to decision-making in the e-commerce sector. In this research, I was able to illustrate how these models could predict future sales trends, inform strategic business decisions, and understand customer sentiments by using a large historical dataset of e-commerce transactions.

The SARIMA model, which identified seasonal patterns, was employed to project sales across several product categories and regions. It showed significant insights regarding future trends in sales. These forecasts estimate a business's insightful data that facilitates better inventory control and resource allocation that considers market dynamics.

The model was tested against an LSTM that could capture complex temporal dependencies in sales data to find its potential in predicting future sales, even though e-commerce markets are highly volatile. In this regard, the errors for the performance metrics MAE, RMSE, and R² showed this model was not very accurate but did turn up some very fundamental insights on stability and patterns of sales over time. This shows the necessity of integrating various methods of ML to arrive at more accurate predictions in a very dynamic market environment.

Other than the forecasting, customer SA was also done using the VADER sentiment analysis tool. This was for the review of customers so they could be ranked into positive, neutral, or negative sentiments. This very strong positive sentiment underlines high degrees of customer satisfaction and gives businesses some important insight into sharpening their product offerings and marketing strategies. This means facilitating better customer engagement, stronger brand loyalty, and better customer experiences.

The contributions of this research are many. First, it underlines the critical role of integrating advanced ML models like SARIMA and LSTM into e-commerce operations to enable better sales forecasting and strategic planning. It contributes by showing that, precisely, with the application of these models, businesses can be better positioned to track market trends, optimize inventory levels, and change their marketing efforts accordingly to achieve future demand. It tries to underline the critical value of sentiment analysis in e-commerce by explaining ways in which firms can use consumer feedback to improve their products and strategies so that they realize an increase in customer satisfaction and competitiveness.

This research makes important contributions toward the use of data in driving decisions within the context of e-commerce. It gives practical insights into how techniques of ML and NLP can help achieve operational efficiency, respond to changing marketplaces much more aware, and deliver better customer experience. Entailed with forecasting analysis and sentiment analysis capabilities, this research has the makings of a solid framework for e-commerce businesses targeting growth and success in the increasingly competitive digital marketplace.

## 6.2 Limitations:

This study applied SA, LSTM, and SARIMA techniques to enhance e-commerce decision-making. However, it was limited because it was predicated on historical sales data, which may not have been able to predict sudden market changes like a downturn. For example, SARIMA predicted that by 2025, electronics sales have decreased from 54,520 units in 2022 to 32,165 units; yet this prediction may prove to be inaccurate if a sudden shift occurs in the market. Preprocessing the data can introduce biases that influence the findings, such as how outliers and missing values are handled. Even with its low MAE of 2138.62, RMSE of 2503.85, MAPE of 44.23%, and R² of 0.04, the LSTM model had difficulties making accurate predictions in a variety of data distributions. SA had drawbacks too. The class imbalances, in which good evaluations predominated over negative ones, led to an overestimation of the model's accuracy. Furthermore, the model is overfitted, measurements like ROC AUC, while ideal at 1.00, may not accurately represent performance in the actual world. Due to the class imbalance, there were many more positive evaluations than negative one of the sentiment analysis's limitations was the potential for overfitting the model's accuracy.

Moreover, even measures with an ideal value of 1.00, such as ROC AUC, may not work well in practical situations if the model is overfitted. According to the study's assumption, businesses would use the findings to inform decisions in the real world, even though organizational structure and subject-matter expertise varied. These kinds of issues draw attention to the necessity of rigorously evaluating real-world applications and continuously developing models.

## 6.3 Future Scope:

In this research, metrics like MAE, RMSE, accuracy, precision, and recall evaluate the performance of the ML model results for SARIMA, LSTM, and SA. While these metrics proved useful in gaining insights, there is scope for betterment. Model improvement in terms of accuracy could be done using cross-validation methods so that the models built are more robust and generalised. Moreover, class imbalance in the data would need to be handled in the case of SA; otherwise, the results may not be reliable. The generative AI harvesting in DDDM remains promising in the future. Generative AI generates artificial data for better model training, simulates future scenarios, and provides personalised marketing. Such integration can improve predictive accuracy for the model very significantly, resulting in informed decisions regarding inventory management, product development, and customer engagement in this dynamic sector of e-commerce.

# 7 References:

Ahmad Azrir, A.H.I., Naveen, P. and Haw, S.C. 2024. Sentiment Analysis using Machine Learning Models on Shopee Reviews. *Journal of System and Management Sciences*. **14**(2).

Anon 2024. Full Service eCommerce Agency. *oscprofessionals*. [Online]. [Accessed 12 July 2024]. Available from: https://www.oscprofessionals.com/blog/how-ai-transforming-ecommerce/.

Baghla, S. and Gupta, G. 2022. Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce. *Microprocessors and Microsystems*. **94**, p.104680.

Bird, S., Klein, E. and Loper, E. 2009. *Natural Language Processing with Python*. ‘O’Reilly Media, Inc.’

Blog, D.C. 2023. Understanding F1 Score, Accuracy, ROC-AUC, and PR-AUC Metrics for Models. *Deepchecks*. [Online]. Available from: https://deepchecks.com/f1-score-accuracy-roc-auc-and-pr-auc-metrics-for-models/.

Bohanec, M., Kljajić Borštnar, M. and Robnik-Šikonja, M. 2017. Explaining machine learning models in sales predictions. *Expert Systems with Applications*. **71**, pp.416–428.

Breiman, L. 1996. Bagging predictors. *Machine Learning*. **24**(2), pp.123–140.

Brynjolfsson, E. and McAfee, A. 2017. The Business of Artificial Intelligence. *Harvard Business Review*. [Online]. Available from: https://hbr.org/2017/07/the-business-of-artificial-intelligence.

Chang, Z., Zhang, Y. and Chen, W. 2018. Effective Adam-Optimized LSTM Neural Network for Electricity Price Forecasting. *IEEE Xplore*. [Online], pp.245–248. Available from: https://ieeexplore.ieee.org/abstract/document/8663710?casa\_token=zQC1Sx3bkQcAAAAA:CoZ9MrTNv166whhApf8Ykdg\_wRPqJPinJt50uC\_JpD\_j1sxLr87yFK4BCR5NIflcns1oskaFPQ.

Chen, H., Chiang, R.H.L. and Storey, V.C. 2012. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*. **36**(4), pp.1165–1188.

Chen, T. and Guestrin, C. 2016. XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16*., pp.785–794.

Chugh, A. 2020. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*. [Online]. Available from: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.

Cortes, C. and Vapnik, V. 1995. Support-vector networks. *Machine Learning*. **20**(3), pp.273–297.

Dee, C. 2023. 10 ways AI is transforming ecommerce. *Algolia*. [Online]. Available from: https://www.algolia.com/blog/ecommerce/10-ways-ai-is-transforming-ecommerce/.

Elias, N.S. and Singh, S. 2019.FORECASTING of WALMART SALES using MACHINE LEARNING ALGORITHMS.

Freund, Y. and Schapire, R.E. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. **55**(1), pp.119–139.

George and Jenkins, G.M. 2015. *Time Series Analysis*. John Wiley & Sons.

Ghaffari, S., Behnam Yousefimehr and Mehdi Ghatee 2024. Generative-AI in E-Commerce: Use-Cases and Implementations. *IEEE*. (2640-5768).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*. **63**(11), pp.139–144.

Holt, C.C. 2004. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting*. **20**(1), pp.5–10.

Huang, L., Dou, Z., Hu, Y. and Huang, R. 2019. Textual Analysis for Online Reviews: A Polymerization Topic Sentiment Model. *IEEE Access*. **7**, pp.91940–91945.

Huang, M.-H. and Rust, R.T. 2018. Artificial Intelligence in Service. *Journal of Service Research*. **21**(2), pp.155–172.

Iswanto, B.H. and Poerwoto, V. 2018. Sentiment analysis on Bahasa Indonesia tweets using Unibigram models and machine learning techniques. *IOP Conference Series: Materials Science and Engineering*. **434**, p.012255.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Sunstein, C.R. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*. **10**, pp.113–174.

Kusawa, S. 2023. Unraveling SARIMA Model: A Comprehensive Guide - Data Magic AI Blog. *https://datamagiclab.com/*. [Online]. Available from: https://datamagiclab.com/unraveling-sarima-model-a-comprehensive-guide/.

Laudon, K.C. and Traver, C.G. 2021. *E-commerce : business, technology, society*. Harlow, England ; New York: Pearson.

Liu, B. 2022. *Sentiment Analysis and Opinion Mining*. Springer Nature.

M, S. 2013. Logistic Regression in E-commerce: Applications and LimitationsJournal of Business Research, 66(9), 1320-1330.

Micol Policarpo, L., da Silveira, D.E., da Rosa Righi, R., Antunes Stoffel, R., da Costa, C.A., Victória Barbosa, J.L., Scorsatto, R. and Arcot, T. 2021. Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*. **41**, p.100414.

Muhammad Tariq, M. 2018. Information and Communication Technology (ICT) and Environmental Sustainability in Developed and Developing Countries. *ProQuest*. **12**(3), pp.758–783.

Ngai, E.W.T., Hu, Y., Wong, Y.H., Chen, Y. and Sun, X. 2011. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*. **50**(3), pp.559–569.

Oudan, R. 2011. Strategic Decision-Making In the Emerging Field Of E-commerce. *International Journal of Management & Information Systems (IJMIS)*. **14**(2).

Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*. **2**(1–2), pp.1–135.

Policarpo, L.M. 2021. *Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review* (D. E. da Silveira, R. da Rodrigo da, R. A. Stoffel, C. A. da Costa, J. L. Victória Barbosa, R. Scorsatto, & T. Arcot, eds.).

Rahul, M. 2024. Expert Analysis Reveals Untapped Potential In Elevating Customer Experience And Sales On E-Commerce Using A/B Testing. *Free Press Journal*. [Online]. [Accessed 12 July 2024]. Available from: https://www.freepressjournal.in/business/expert-analysis-reveals-untapped-potential-in-elevating-customer-experience-and-sales-on-e-commerce-using-ab-testing.

Sairamvinay Vijayaraghavan and Basu, D. 2020. Sentiment Analysis in Drug Reviews using Supervised Machine Learning Algorithms. *arXiv (Cornell University)*.

Saleem, H., Bin Muhammad, K., Hussain Nizamani, A., Saleem, S. and Butt, J. 2021. INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND TECHNOLOGY (IJARET), © IAEME Publication. *INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND TECHNOLOGY (IJARET)*. **12**(0976-6499).

Singh, K., Booma, P.M. and Eaganathan, U. 2020. E-Commerce System for Sale Prediction Using Machine Learning Technique. *Journal of Physics: Conference Series*. **1712**, p.012042.

Taneja, K., Jyoti Vashishtha and Saroj Ratnoo 2024. Transformer Based Unsupervised Learning Approach for Imbalanced Text Sentiment Analysis of E-Commerce Reviews. *Procedia computer science*. **235**, pp.2318–2331.

Taylor, S.J. and Letham, B. 2017. Forecasting at Scale. *The American Statistician*. **72**(1), pp.37–45.

Tyagi, P. and Tripathi, R.C. 2019. A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data. *SSRN Electronic Journal*.

Ugwu, G. 2024. Why Machine Learning Is Good for Retail and E-commerce. *North South Tech*. [Online]. [Accessed 12 July 2024]. Available from: https://northsouthtech.co/blog/why-machine-learning-is-good-for-retail-and-e-commerce/.

Wikipedia Contributors 2018. Long short-term memory. *Wikipedia*. [Online]. Available from: https://en.wikipedia.org/wiki/Long\_short-term\_memory.

YU, J. and LE, X. 2016. Sales Forecast for Amazon Sales Based on Different Statistics Methodologies. *DEStech Transactions on Economics and Management*. (iceme-ebm).

Zhang, X., Guo, F., Chen, T., Pan, L., Beliakov, G. and Wu, J. 2023. A Brief Survey of Machine Learning and Deep Learning Techniques for E-Commerce Research. *Journal of Theoretical and Applied Electronic Commerce Research*. **18**(4), pp.2188–2216.

Zhang, Y. and Wallace, B.C. 2015. A Sensitivity Analysis of (and Practitioners’ Guide to) Convolutional Neural Networks for Sentence Classification.

Zineb, E.F., Najat, R. and Jaafar, A. 2021. An Intelligent Approach for Data Analysis and Decision Making in Big Data: A Case Study on E-commerce Industry. *International Journal of Advanced Computer Science and Applications*. **12**(7).

# 8. Appendix:

## 8.1 Literature Review Paper’s Summary Table

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Paper/Author | Machine Learning Methods | Business Applications |
| Personalised Recommendations | Ricci et al., 2011 | Content-based and Collaborative Filtering | Personalized recommendations |
| Inventory Management | Kumar and Reinartz, 2018 | Just-In-Time (JIT), Economic Order Quantity (EOQ) | Inventory management |
|  | Baghla and Gupta, 2022 | Random Forests | Inventory control |
|  | Taylor and Letham, 2017 | Holt-Winters Exponential Smoothing, ARIMA | Inventory management |
| Customer Service | Huang and Rust, 2018 | Rule-based systems, Decision Trees | Customer service |
|  | Ngai et al., 2011 | Statistical analysis | Fraud detection |
|  | Breiman, 1996 | Decision Trees, Bagging | Fraud detection |
|  | Cortes and Vapnik, 1995 | Support Vector Machines (SVM), Kernel Techniques | Fraud detection |
|  | Duchemin and Matheus, 2021 | Support Vector Machines (SVM), Neural Networks, Bagging, Boosting | Fraud detection |
|  | De and Prabu, 2022 | Logistic Regression, Decision Trees, Random Forest | Fraud detection |
| Customer Sentiment | Krizhevsky et al., 2012 | Convolutional Neural Networks (CNNs), Ensemble approaches | Customer sentiment, Visual product searches |
|  | Kim, 2014 | Convolutional Neural Networks (CNNs) | Sentiment analysis |
|  | Poria et al., 2016 | Aspect-based Sentiment Analysis, BERT | Customer sentiment analysis |
|  | Pang and Lee, 2008 | Natural Language Processing (NLP) | Sentiment analysis |
| Sales Forecasting | Taylor and Letham, 2017 | Holt-Winters Exponential Smoothing, ARIMA | Sales forecasting |
|  | Baghla and Gupta, 2022 | Random Forests | Sales forecasting |
|  | Chen and Guestrin, 2016 | AdaBoost, XGBoost | Sales forecasting |
|  | Hinton et al., 2006 | Deep Belief Networks(DBNs) | Predictive Analysis |
| Client Segmentation | Breiman, 1996 | Decision Trees, Bagging | Client segmentation |
|  | Cortes and Vapnik, 1995 | Support Vector Machines (SVM), Kernel Techniques | Consumer classification |
|  | Chen and Guestrin, 2016 | AdaBoost, XGBoost | Client segmentation |
|  | Baghla and Gupta, 2022 | Random Forests, Gradient Boosting Trees (GBT) | Client segmentation |
| Customer Attrition | De and Prabu, 2022 | Logistic Regression | Customer attrition forecasting |
|  | Duchemin and Matheus, 2021 | Support Vector Machines (SVM), Neural Networks, Bagging, Boosting | Churn prediction |
|  | Krizhevsky et al.,  2012 | Convolutional Neural Networks (CNNs) | Sales estimation |

Table 11

## 8.2 Recent Approaches Paper’s Table

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Paper/Author | Machine Learning Method | Approach |
| Recent Approaches |  |  |  |
| Inventory Management | Baghla and Gupta, 2022 | Random Forests | Inventory control |
| Fraud Detection | Cortes and Vapnik, 1995 | Support Vector Machines (SVM), Kernel Techniques | Fraud detection |
|  | Duchemin and Matheus, 2021 | SVM, Neural Networks, Bagging, Boosting | Fraud detection |
|  | De and Prabu, 2022 | Logistic Regression, Decision Trees, Random Forest | Fraud detection |
| Customer Sentiment | Krizhevsky et al., 2012 | Convolutional Neural Networks (CNNs), Ensemble approaches | Customer sentiment, Visual product searches |
|  | Kim, 2014 | Convolutional Neural Networks (CNNs) | Sentiment analysis |
|  | Poria et al., 2016 | Aspect-based Sentiment Analysis, BERT | Customer sentiment analysis |
| Sales Forecasting | Baghla and Gupta, 2022 | Random Forests | Sales forecasting |
|  | Chen and Guestrin, 2016 | AdaBoost, XGBoost | Sales forecasting |
| Client Segmentation | Cortes and Vapnik, 1995 | Support Vector Machines (SVM), Kernel Techniques | Consumer classification |
|  | Chen and Guestrin, 2016 | AdaBoost, XGBoost | Client segmentation |
|  | Baghla and Gupta, 2022 | Random Forests, Gradient Boosting Trees (GBT) | Client segmentation |
| Customer Attrition | Duchemin and Matheus, 2021 | SVM, Neural Networks, Bagging, Boosting | Churn prediction |

Table 12

## 8.3 Regional Forecast

A graph with blue lines

Description automatically generated

Fig 42- China Forecast

The graph is the sales forecast for China next year, based on historical sales data from 2020 to 2024. The blue trend line in the graph indicates the trend of historical sales, which has huge fluctuations, peaking at 11,792 in 2022 and lows of 5,074 in 2023. The red dot forecasts 5,021 sales in 2025, which means it will continue to drop a little.

A graph with a line and numbers

Description automatically generated

Fig 43- India Forecast

This graph represents the sales forecast for India in the upcoming year 2025 based on previous data available from 2020-2024. The blue line indicates that, in 2023, the historical sales trends peaked at 10911 and drastically came down to 4101 in the year 2024. The red dotted line projects further decline in 2025 to 3080.

A graph with a line going up

Description automatically generated

Fig 44- Italy Forecast

It displays the sales forecast for Italy, using a sales history from 2020 to 2024. The trend peaked at 12,167 in 2022. The sales for the year 2025 are then projected with a red dot to be far higher than that, at 10,219 units.

A graph with a line

Description automatically generated

Fig 45- Japan Forecast

The visual for sales forecasting in the case of Japan, including historical data from 2020-2024. In blue, one finds the historical sales data peaking at 9,539 units sold in the year 2021. The red dot forecasts a slight decline to 4,516 units sold in 2025.

A graph with blue lines and numbers

Description automatically generated

Fig 46- Spain Forecast

The graph representing the sales forecast for Spain next year shows key trends in historic and forecasted sales. In the period spanning from 2020 to 2024, sales were variable, reaching a peak of 9,432 units in 2021 before hitting a decline to 5,720 units in 2024. The forecast for 2025 is an even deeper decline, with 5,209 units.

A graph with numbers and lines

Description automatically generated

Fig 47- UK Forecast

The graph indicates that in historical sales, the blue trend peaked in 2020 at 7540 units before dropping to 4052 units in 2023. It's projected to reach 4749 in 2025

A graph with a line

Description automatically generated

Fig 48- USA Forecast

Above is the graphical representation of sales forecasting for the USA. It shows major ups and downs, peaking in 2023 at 16,921, then falling back in 2024 to 4,126. The projected sales for 2025, increase to 12,785.

## 8.4 Data Exploration for Sentiment Analysis:

A graph with blue and white stripes

Description automatically generated

Fig 53

From the above visual, it can be observed that Office Products from the main category is highly rated with 4.31 and the Car & Motorbikes with the lowest rating of 3.80.

A bar graph with different colored lines

Description automatically generated

Fig 54

In the rating distribution graph by product sub-category, the product with the highest number of reviews is **power accessories**. The car Accessories has the most lowly rated product.

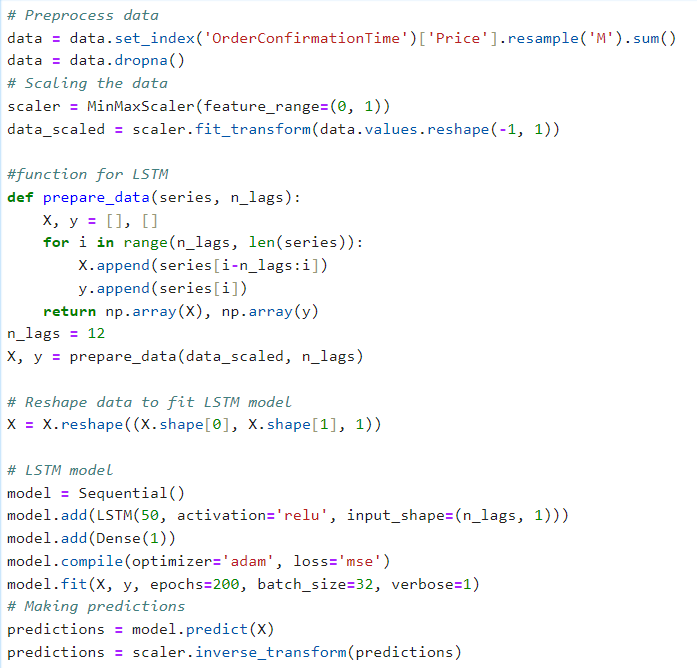
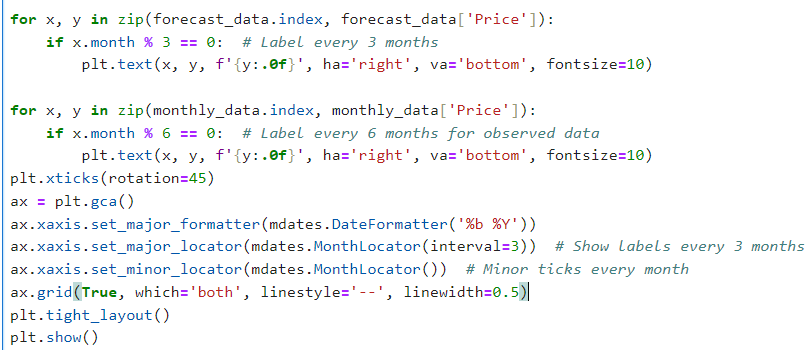
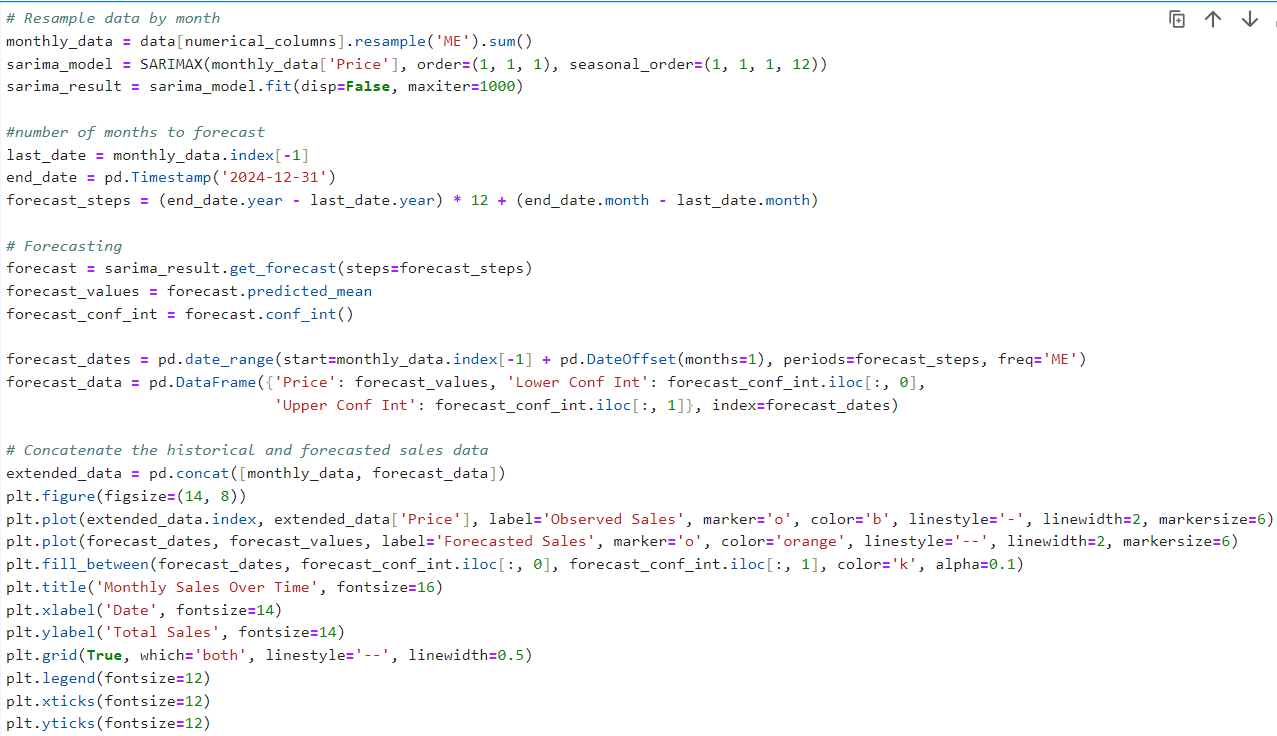
A graph with numbers and text

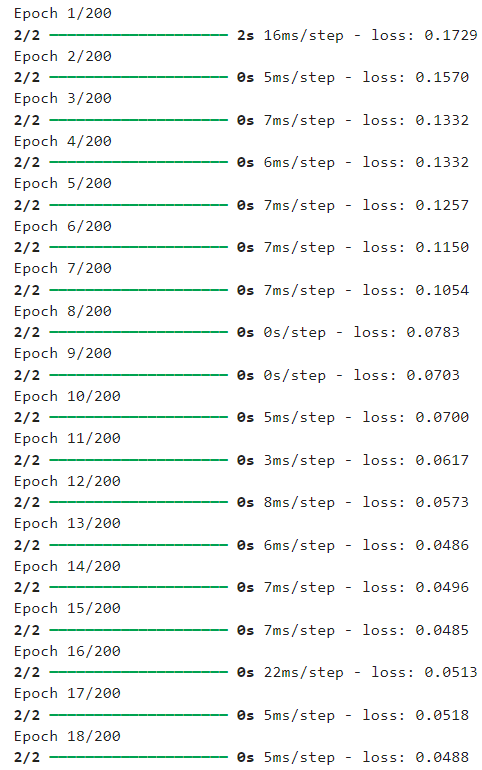
Description automatically generated

Fig 55

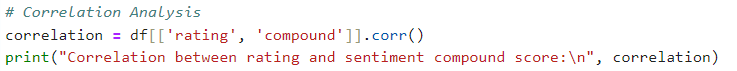
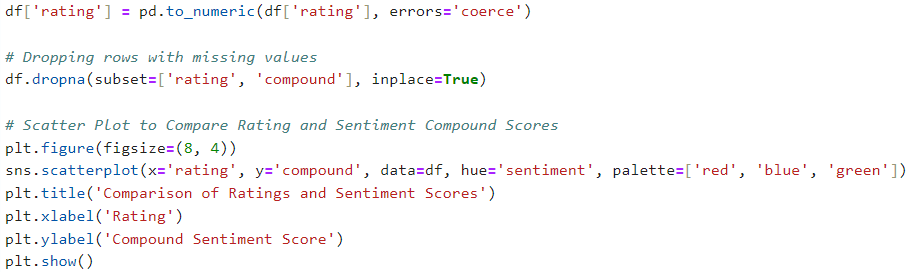
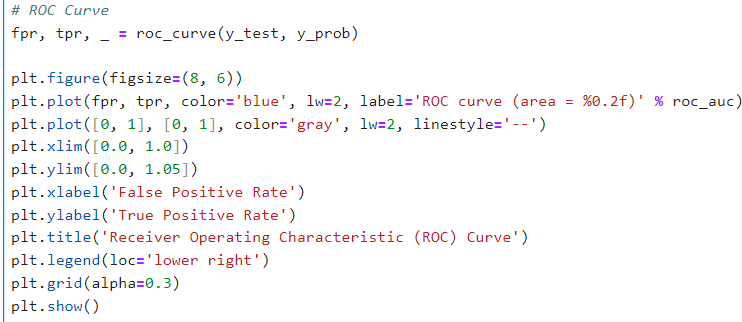
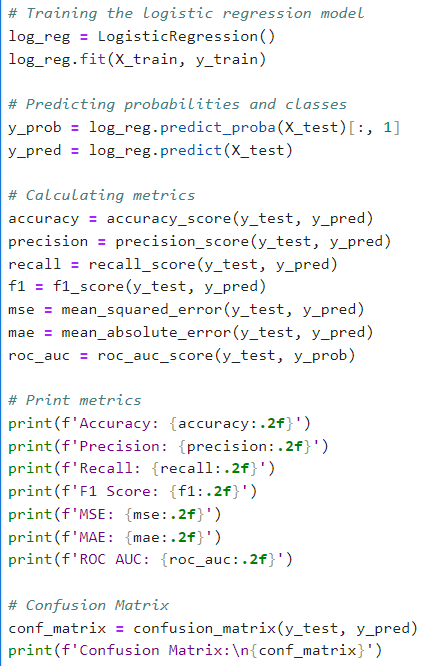
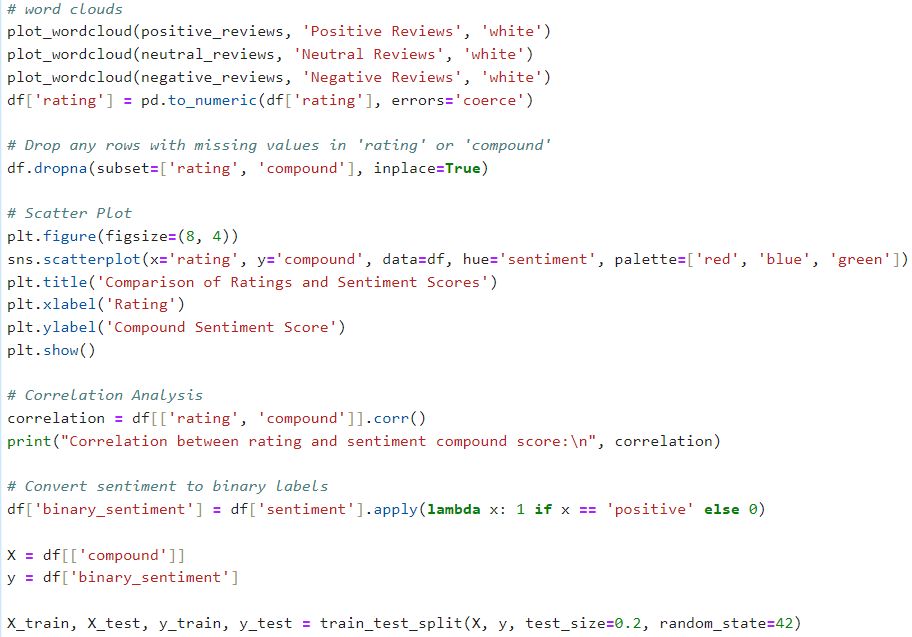
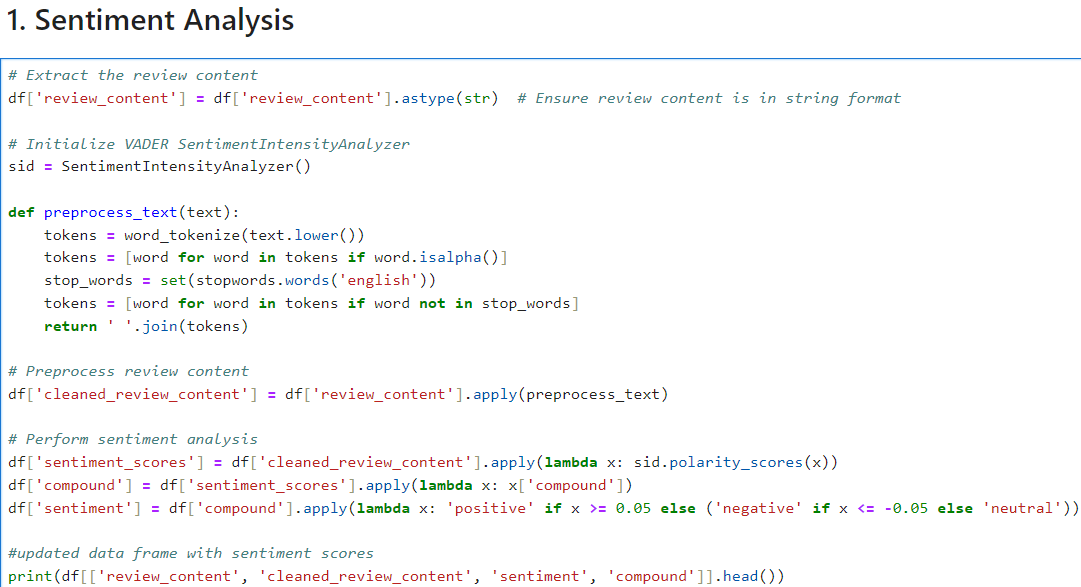
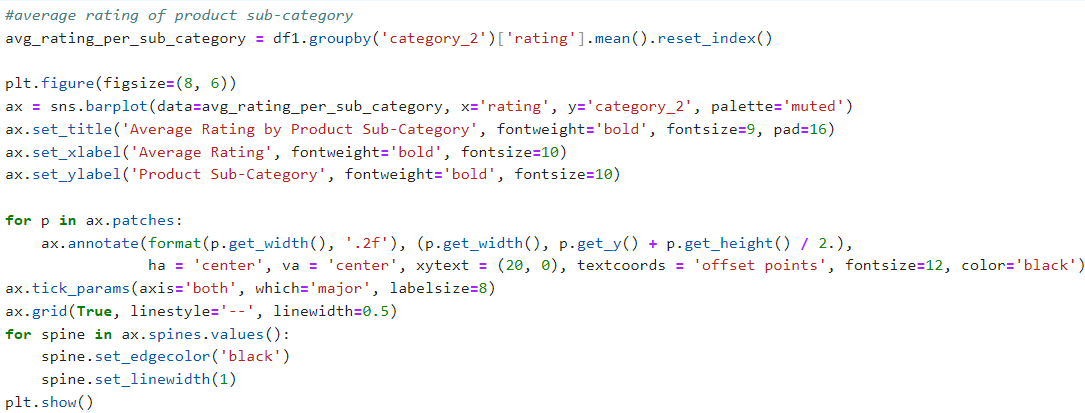
The above bar plot presents the top 10 most active reviewers. "Amazon Customer" has posted the most reviews at 630.

## 8.4 Code Snippets for Forecasting Analysis





## 8.5 Code Snippets for Sentiment Analysis



## 9 Ethical Form

**Ethical Approval Form**

Internal research ethics application form for taught student modules (where University ethical approval is in place for the module)

For module LUBS5579 covered by University of Leeds ethical approval reference AREA 17-055.

|  |  |
| --- | --- |
| Student ID | 201771035 |
| Your name | SHRADDHA DATTATRAY KUMBHAR |
| Degree Programme | MSc Business Analytics and Decision Sciences |
| Provisional title/ topic area | **Enhancing Data-Driven Decision-Making in E-Commerce Using Machine Learning Techniques.** |
| Name of dissertation supervisor | **Dr. Yu-wang Chen** |

|  |  |
| --- | --- |
| Are you planning to conduct fieldwork with (data on) human participants for your dissertation? | Please tick the relevant box |
| **Yes** (This includes online research methods and secondary data analysis). |  |
| **No**, I am conducting library based research or content/ media analysis only. | ü |

If you ticked ‘no’ you do not need to take further action in respect of ethical approval. Please proceed to the declarations on page 8 and 9.

If you ticked ‘yes’ you need to complete the rest of this form.

You MUST submit discuss your research design and the ethical issues it raises with your dissertation supervisor and receive their signed approval **before you approach any participants or collect any data**.

You MUST attach a copy of your research proposal to this form.

You MUST include a copy of your ethics form (signed by your supervisor), together with your research proposal, as an appendix to your final dissertation submission.

**INTERNAL RESEARCH ETHICS APPLICATION**

**Part A: Compliance with the module’s block ethical approval**

Ethical review is required for all research involving human participants, including research undertaken by students within a taught student module. Further details of the University of Leeds ethical review requirements are provided in the *Research Ethics Policy* available at:

<http://ris.leeds.ac.uk/ResearchEthicsPolicies> and at [www.leeds.ac.uk/ethics](http://www.leeds.ac.uk/ethics).

|  |  |  |
| --- | --- | --- |
| 1. **Will your dissertation involve any of the following?** | Yes | No |
| New data collected by administering questionnaires/interviews for quantitative analysis |  | ü |
| New data collected by qualitative methods |  | ü |
| New data collected from observing individuals or populations |  | ü |
| Working with aggregated or population data |  | ü |
| Using already published data or data in the public domain |  | ü |
| Any other research methodology, please specify: |  | ü |

|  |  |  |
| --- | --- | --- |
| 1. **Will any of the participants be from any of the following groups?** (Tick as appropriate) | Yes | No |
| Children under 16 |  | ü |
| Adults with learning disabilities |  | ü |
| Adults with other forms of mental incapacity or mental illness |  | ü |
| Adults in emergency situations |  | ü |
| Prisoners or young offenders |  | ü |
| Prisoners or young offenders |  | ü |
| Those who could be considered to have a particularly dependent relationship with the investigator, e.g. members of staff, students |  | ü |
| Other vulnerable groups, please specify: |  | ü |

|  |  |  |
| --- | --- | --- |
| 1. **Will the project/dissertation/fieldwork involve any of the following:**   (You may select more than one) | Yes | No |
| Patients and users of the NHS (including NHS patients treated under contracts with private sector) |  | ü |
| Individuals identified as potential participants because of their status as relatives or carers of patients and users of the NHS |  | ü |
| The use of, or potential access to, NHS premises or facilities |  | ü |
| NHS staff - recruited as potential research participants by virtue of their professional role |  | ü |
| A prison or a young offender institution in England and Wales (and is health related) |  | ü |

If you have answered ‘yes’ to ANY of the above questions in 2 or 3 then you will need to apply for full ethical review, a faculty committee level process. This can take up to 6-8 weeks, so it is important that you consult further with your supervisor for guidance with this application as soon as possible. Please now complete and sign the final page of this document. The application form for full ethical review and further information about the process are available at <http://ris.leeds.ac.uk/uolethicsapplication>.

If you answered ‘no’ to ALL of the questions in sections 2 and 3 please continue to part B.

**INTERNAL RESEARCH ETHICS APPLICATION**

**Part B: Ethical considerations within block ethical approval**

|  |  |  |
| --- | --- | --- |
| 1. **Will the research touch on sensitive topics or raise other challenges?** | Yes | No |
| Will the study require the cooperation of a gatekeeper for initial access to groups or individuals who are taking part in the study (eg students at school, members of self-help groups, residents of a nursing home)? |  | ü |
| Will participants be taking part in the research without their knowledge and consent (eg covert observation of people in non-public places)? |  | ü |
| Will the study involve discussion of sensitive topics (eg sexual activity, drug use)? |  | ü |
| Could the study induce psychological stress or anxiety or cause harm or have negative consequences beyond the risks encountered in normal life? |  | ü |
| Are there any potential conflicts of interest? |  | ü |
| Does any relationship exist between the researcher(s) and the participant(s), other than that required by the activities associated with the project (e.g., fellow students, staff, etc)? |  | ü |
| Does the research involve any risks to the researchers themselves, or individuals not directly involved in the research? |  | ü |

If you have answered ‘yes’ to any of the questions in (5), please describe the ethical issues raised and your plans to resolve them on a separate page. Agree this with your supervisor and submit it with this form. Again, you MAY be referred for light touch or full ethical review.

|  |  |  |
| --- | --- | --- |
| 1. **International Research**   Does your research involve participants outside of the UK? | Yes | No |
| Are any of your research participants located outside of the UK, e.g., will you be gathering data through Skype interviews with participants located overseas? |  | ü |
| Will any of the fieldwork or research require you to travel outside of the UK to collect data? |  | ü |

If you have answered ‘yes’ to either part of question (5), please describe the ethical issues raised with: gaining consent and gathering data from participants located overseas, securely storing and transferring data from the field back to the UK, any cultural issues that may be relevant. Please outline your plans to resolve this on a separate page and ensure that you have completed a risk assessment form. Agree this with your supervisor and submit it with this form.

You MAY be referred for light touch or full ethical review if you are unable to demonstrate that you have resolved the ethical issues relating to international research.

|  |  |  |
| --- | --- | --- |
| 1. **Personal** **safety**   Where will any fieldwork/ interviews/ focus groups take place? | Yes | No |
| At the university or other public place (please specify below). |  | ü |
| At my home address |  | ü |
| At the research subject's home address |  | ü |
| Some other location (please specify below). |  | ü |

If you conduct fieldwork anywhere except at the university or other public place you need to review security issues with your supervisor and have them confirmed by the Module Leader who may refer you for light touch or full ethical review. Write a brief statement indicating any security/personal safety issues arising for you and/or for your participants, explaining how these will be managed. Agree this with your supervisor and submit it with this form.

Please note that conducting fieldwork at the research subject's home address will require strong justification and is generally not encouraged.

A risk assessment is required before any data is gathered for any dissertation project, please view the Health and Safety advice on the module’s Minerva pages.

|  |  |  |
| --- | --- | --- |
| 1. **Anonymity** | Yes | No |
| Is there any potential for data to be traced back to individuals or organisations, for instance because it has been unanonymised or anonymised in such a way that there remains risk (eg highlighting people’s positions within an organisation, which may reveal them). |  | ü |

If you have answered ‘yes’ to question 7, please discuss this further with your supervisor. You need to provide a strong justification for this decision on a separate sheet. **This application will need to be reviewed by the dissertation Module Leader and may require a full ethical review.**

1. **Data management issues**

|  |  |  |  |
| --- | --- | --- | --- |
| Will the research involve any of the following activities at any stage (including identification of potential research participants)? | | Yes | No |
| Examination of personal records by those who would not normally have access | |  | ü |
| Sharing data with other organisations | |  | ü |
| Use of personal addresses, postcodes, faxes, e-mails or telephone numbers | |  | ü |
| Publication of direct quotations from respondents | |  | ü |
| Publication of data that might allow identification of individuals to be identified | |  | ü |
| Use of audio/visual recording devices | |  | ü |
| Storage of personal data on any of the following: | |  | ü |
|  | FLASH memory or other portable storage devices |  | ü |
| Home or other personal computers |  | ü |
| Private company computers |  | ü |
| Laptop computers |  | ü |

If you have answered ‘yes’ to any of the questions under 8, you must ensure that you follow the University of Leeds Information Protection Policy: <http://www.leeds.ac.uk/informationsecurity> and the Research Data Management Policy: <http://library.leeds.ac.uk/research-data-policies#activate-tab1_university_research_data_policy>.

You are obliged to provide a copy of your anonymised data to your supervisor for their records and to destroy other copies of your data when your degree has been confirmed.

**Dissertation Research Ethical Approval: Declaration**

|  |  |
| --- | --- |
| For students | Please tick as appropriate |
| **Option 1:** I **will NOT** conduct fieldwork with (data on) human participants for my dissertation. | ü |
| **Option 2:** I **will** conduct fieldwork with (data on) human participants for my dissertation. |  |

For **options 1 and 2** - I confirm that:

The research ethics form is accurate to the best of my knowledge.

I have consulted the University of Leeds Research Ethics Policy available at <http://ris.leeds.ac.uk/ResearchEthicsPolicies>.

I understand that ethical approval will only apply to the project I have outlined in this application and that I will need to re-apply, should my plans change substantially.

For **option 2** only:

I am aware of the University of Leeds protocols for ethical research, in particular in respect to protocols on **informed consent**, **verbal consent**, **reimbursement for participants** **and low risk observation**. If any are applicable to me, signing this form confirms that I will carry out my work in accordance with them. http://ris.leeds.ac.uk/PlanningResearch

Student’s signature: **Shraddha Dattatray Kumbhar**

Date: **27th August 2024**

|  |  |  |
| --- | --- | --- |
| For supervisors | Yes | No |
| No further action required | | |
| I confirm that the dissertation is in line with the module’s block ethical approval (Part A & question 8). | ü |  |
| I have discussed the ethical issues arising from the research with the student and agree that these have been accurately and fully addressed. | ü |  |
| I have reviewed the student’s research proposal. | ü |  |
| I have reviewed the student’s Risk Assessment Form. | | |
| Further actions required | | |
| Refer to dissertation Module Leader for further review / discussion. |  |  |
| The dissertation falls outside the module’s block ethical approval and the student was advised to apply for full ethical review. |  |  |

Supervisor’s signature: …… ............................................………….………..

Date: ……28/08/2024………………….……................................................………………..