

Leveraging Machine Learning to Predict E-commerce Shopping Behaviour and Enhance Recommendations



Sheikh Fuad Ahmed

20000816

Dissertation submitted in partial fulfilment of the requirements

for the degree of

Master of Science in Business Analytics

at Dublin Business School

Supervisor: Dr Vivek Kshirsagar

May 2024

DECLARATION

I Sheikh Fuad Ahmed hereby declare that the research work titled "Leveraging Machine Learning to Predict E-commerce Shopping Behaviour and Enhance Recommendations" is entirely my own effort and represents the outcome of my intellectual endeavours, except where otherwise stated and duly acknowledged through references. This research project is submitted in partial fulfilment of the requirements for the MSc in Business Analytics program at Dublin Business School.

Signed: Sheikh Fuad Ahmed

Student ID: 20000816

Date: 19th May, 2024

ACKNOWLEDGEMENT

I would like to express my deep and sincere gratitude to my dissertation supervisor Dr. Vivek Kshirsagar for his guidance, support, and invaluable feedback. It has been a privilege and a pleasure to work under his supervision.

I am also very grateful to Dublin Business School for providing me with all the necessary tools and resources to complete this project and further my studies in this course.

Thank You

ABSTRACT

This study explored how machine learning techniques could predict e-commerce shopping behaviour and enhance product recommendations. The research analysed data from user interactions, purchase histories, and sentiment analysis to develop effective models. It involved thorough data preparation, including handling missing values, processing text with TFIDF, and applying SMOTE to balance the data. Various models were tested such as Logistic Regression, AdaBoost, Random Forest, Naive Bayes, XGBoost, and Linear Support Vector Machine. The results indicated that Logistic Regression and AdaBoost were most effective for predicting shopping behaviour while Logistic Regression and Linear Support Vector Classifier excelled in sentiment analysis. These models achieved high accuracy, precision, and recall, demonstrating their practicality for real-world e-commerce applications. Implementing these models allowed e-commerce platforms to offer personalized recommendations, enhance customer satisfaction and increase sales. This study demonstrated the significant potential of machine learning to improve e-commerce strategies and enrich the overall shopping experience.

Table of Contents

1. Introduction	7
1.1 Background	7
1.2 Research Question.....	8
1.3 Research Objective.....	8
2. Literature Review	8
3. Methodology.....	13
3.1 Business Understanding	14
Objectives and Requirements	14
Data Mining Problem Definition.....	14
Preliminary Plan	15
3.2 Data Understanding.....	15
Collect Initial Data.....	15
Familiarize with Data	16
Identify Data Quality Issues	16
3.3 Data Preparation.....	17
Exploratory Data Analysis (EDA).....	18
Data Normalization.....	19
3.4 Modeling	19
E-commerce Shopping Behaviour Dataset	19
Model Selection.....	19
Model Implementation	19
Sentiment Analysis Dataset.....	20
Model Selection.....	20
Model Implementation	20
3.5 Evaluation.....	21
Performance Metrics.....	21
Model Assessment.....	22
4. Result and Discussion.....	23
4.1 Exploratory Data Analysis	23
4.2 Model Evaluation for Shopping Behaviour	26
Hyperparameter Ranges and Best Parameters.....	26
Variation in Model Performance	28
Model Deployment.....	29
4.3 Model Evaluation for Sentiment Analysis	29

Hyperparameter Ranges and Best Parameters	30
Variation in Model Performance	32
Model Deployment	33
5. Conclusion and Future Work.....	34
Conclusion.....	34
Future Work	35
Reference	36

List of Table

Table 1: Model Performance and Hyperparameters Comparison (Shopping Behaviour).....	26
Table 2: Model Performance and Hyperparameters Comparison (Sentiment Analysis).....	30

List of Figure

Figure 1: CRISP-DB Model Data Science Process Alliance, 2023.....	13
Figure 2: Distribution of Subscription Status among Customers.....	23
Figure 3: Purchase Amount Distribution by Subscription Status.....	24
Figure 4: Average Review Rating by Subscription Status.....	24
Figure 5: Purchase Frequency by Subscription Status.....	25
Figure 6: Category Preference by Subscription Status.....	25

1. Introduction

Rapid evolution of technology and the widespread adoption of internet have significantly transformed the landscape of commerce giving rise to the e-commerce industry. E-commerce defined as the buying and selling of goods and services over electronic systems through internet has become a dominant force in the global market. This transformation is not only driven by technological advancements but also by the vast amounts of data generated from online transactions and customer interactions. Leveraging this data through machine learning (ML) techniques offers significant potential for predicting shopping behaviour and enhancing product recommendations which ultimately leading to improved customer satisfaction and increased sales (Agrawal et al. 2016).

1.1 Background

E-commerce platforms generate immense amounts of data from user interactions, purchase histories, product reviews, and ratings. Analysing this data allows businesses to gain deep insights into customer preferences and behaviour patterns which can inform marketing strategies and operational decisions (Ngai et al. 2009). For instance, understanding what a customer might be interested in based on their previous interactions and reviews enables businesses to offer personalized recommendations, enhancing the shopping experience and fostering customer loyalty (Chen et al. 2012).

The COVID-19 pandemic has further accelerated the shift towards e-commerce, making the ability to predict shopping behaviour and provide personalized recommendations even more critical. With physical stores closed or operating under restrictions consumers turned to online platforms to meet their shopping needs resulting in a surge in online sales (Bhatti et al. 2020). This sudden increase in online activity has provided a wealth of data that can be used to refine ML models and improve their accuracy.

Machine learning has become an essential tool for analysing large datasets and making accurate predictions. Various ML algorithms, including logistic regression, decision trees, random forests, and support vectors, are employed to analyse consumer data and predict future behaviour (Kaur and Fadnavis 2020). These algorithms can identify patterns and trends within the data that are not immediately ostensible enabling businesses to make data-driven decisions.

In addition to analysing transactional data sentiment analysis has emerged as a valuable tool for understanding customer opinions and preferences. Sentiment analysis involves the

computational analysis of textual data, such as customer reviews and social media posts to gauge customer sentiments towards products and services (Bing Liu 2012). By integrating sentiment analysis with shopping behaviour prediction businesses can gain a more comprehensive understanding of customer needs and preferences which enabling them to offer more personalized and relevant product recommendations (Xu Goh et al. 2019).

1.2 Research Question

The primary research question addressed in this study is:

How can machine learning techniques be leveraged to predict e-commerce shopping behaviour and enhance product recommendations?

1.3 Research Objective

The main objective of this research is to develop machine learning models that can analyse e-commerce shopping behaviour and sentiment data to make accurate predictions and provide personalized recommendations. The specific objectives include:

- Analysing e-commerce shopping behaviour data to identify patterns and trends.
- Integrating sentiment analysis data to understand customer opinions and sentiments towards products.
- Providing actionable recommendations based on the analysis to enhance the customer shopping experience.

2. Literature Review

The advent of e-commerce revolutionized the way businesses operated and interacted with customers. With the ever-increasing volume of data generated by online transactions and customer interactions machine learning techniques emerged as powerful tools for predicting shopping behaviour and enhancing product recommendations (Agrawal and Schorling 2016). This literature review focuses on the application of machine learning in understanding e-commerce shopping behaviour and improving product recommendations.

E-commerce experienced substantial growth in recent years with online sales accounting for a significant portion of global retail sales (Statista 2023). The ability to analyse and leverage customer data effectively became a crucial competitive advantage for businesses operating in the e-commerce space (Grewal et al. 2017). By analysing patterns in consumer data such as browsing history, purchase history, and demographic information businesses could predict

future behaviour and tailor their marketing strategies accordingly (Zaki and Neely 2019). Several machine learning algorithms including logistic regression, decision trees, random forests, and support vector were employed for this purpose (Kaur and Fadnavis 2020).

The COVID-19 pandemic profoundly impacted the e-commerce landscape which accelerated the shift towards online shopping and forced businesses to adapt rapidly (Akhtar et al. 2020). With physical stores closed or operating under strict restrictions consumers turned to e-commerce platforms to meet their shopping needs which led to a surge in online sales (Bhatti et al. 2020). In this context the ability to accurately predict consumer shopping behaviour and provide personalized product recommendations became even more crucial (Pantano et al. 2020). Machine learning techniques played a pivotal role in helping businesses navigate the new landscape enabling them to analyse customer data and adapt their strategies accordingly (Mansurali et al. 2024). In the post-pandemic era e-commerce became an essential part of the global economy with businesses across various industries recognizing the importance of having a robust online presence (Mansurali et al. 2024). Predicting consumer shopping behaviour and providing personalized recommendations became a key differentiator for e-commerce businesses as it enhanced customer satisfaction and loyalty (Khade 2016). Machine learning techniques combined with sentiment analysis offered a powerful approach to achieving these goals enabling businesses to gain a deeper understanding of their customers and tailor their offerings accordingly (Yang et al. 2020).

Sentiment analysis emerged as a valuable tool for e-commerce businesses to understand customer opinions and preferences (Bing Liu 2012). By analysing customer reviews, social media posts, and other textual data, businesses could gain insights into product perceptions, customer satisfaction, and potential areas for improvement (Medhat et al. 2014). Integrating sentiment analysis with shopping behaviour prediction provided a more comprehensive understanding of customer needs and preferences enabling e-commerce businesses to offer personalized product recommendations.

Exploratory Data Analysis (EDA) a crucial step in the data mining process particularly when working with complex datasets such as those in e-commerce. It involved understanding the characteristics, patterns, and relationships within the data (Fernandes et al. 2020). EDA techniques, such as data visualization, statistical summaries, and correlation analysis, uncovered valuable insights and informed feature selection and data pre-processing strategies. One primary objective of EDA was to identify and handle missing or inconsistent data which

could significantly impact the accuracy of machine learning models (Vijay Kotu 2019). Techniques such as data visualization, statistical summaries, and outlier detection were employed to identify and address data quality issues. Additionally, EDA aided in understanding the distribution of variables, identifying potential correlations, and detecting underlying patterns or trends (Fernandes et al. 2020). Feature engineering was another important aspect of EDA particularly in the e-commerce domain (Vijay Kotu 2019). Techniques such as one-hot encoding and TFIDF were commonly used to transform raw data into a format suitable for machine learning algorithms (Bengio et al. 2013). EDA aided in identifying relevant features and determining the appropriate feature engineering techniques to improve model performance (Christof Ebert 2016).

Addressing class imbalance was a common issue in e-commerce datasets where certain customer segments or behaviour patterns might be over- or underrepresented (Chawla et al. 2002). Techniques like SMOTE (Synthetic Minority Over-sampling Technique) could be employed to balance the dataset, ensuring that machine learning models were not biased towards the majority class (Vijay Kotu 2019). Effective data preparation was a critical step in leveraging machine learning techniques for predicting e-commerce shopping behaviour and enhancing product recommendations. It involved handling missing data, encoding categorical variables, performing EDA and transforming the data into a suitable format for machine learning algorithms (Vijay Kotu 2019). Commonly used techniques for missing data imputation included mean/median substitution, regression-based imputation, and multiple imputation. Data cleaning might involve removing duplicates, handling outliers, and addressing inconsistencies in the data (Tang 2014).

The selection of appropriate machine learning models and the application of techniques such as cross-validation and hyperparameter tuning were crucial in developing robust predictive models for e-commerce shopping behaviour and sentiment analysis (Kuhn and Johnson 2013a). Logistic Regression (LR) was widely used for binary classification problems, known for its interpretability and ability to handle linearly separable data (David W. Hosmer 2013). AdaBoost combined multiple weak classifiers to create a strong classifier, often leading to improved performance (Freund and Schapire 1997). Random Forest Classifier (RFC) constructed multiple decision trees and combined their predictions, offering robustness to overfitting and the ability to handle non-linear relationships (Breiman 2001). XGBoost (XGB) a powerful gradient boosting algorithm, gained popularity for its efficiency and performance

particularly in structured data settings (Chen and Guestrin 2016). Linear Support Vector Machine (Linear SVM) was popular for binary classification tasks as it found the optimal hyperplane that maximized the margin between classes which provide good generalization performance. However, it might struggle with non-linear or high-dimensional data necessitating the use of kernel functions or feature engineering techniques (Cortes et al. 1995).

Sentiment analysis tasks involving text data, additional algorithms were commonly employed. Multinomial Naive Bayes (MultinomialNB) was a variant of the Naive Bayes algorithm that assumed independence between features and was well-suited for text classification problems (Mccallum and Nigam 1998). Linear Support Vector Classifier (Linear SVC) effectively handled high-dimensional sparse data such as text vectorized using techniques like TF-IDF (Joachims 1998). XGBoost (XGB) had also been successfully applied to sentiment analysis tasks leveraging its ability to handle both numerical and categorical features (Chen and Guestrin 2016).

To ensure the robustness and generalization performance of these models, techniques such as cross-validation and hyperparameter tuning were essential. Cross-validation was a resampling technique that partitioned the data into multiple subsets, allowing for model evaluation on different data splits and reducing the risk of overfitting (Arlot and Celisse 2010). The widely used k-fold cross-validation approach was commonly employed, with 5-fold or 10-fold being popular choices (Kuhn and Johnson 2013b). Hyperparameter tuning was the process of finding the optimal combination of model parameters that maximize performance on a given task (Bergstra et al. 2012). Grid Search was a widely used technique for hyperparameter tuning where a predefined grid of parameter values was exhaustively searched to find the best combination (Hsu et al. 2003). By implementing these machine learning models and employing techniques like cross-validation and hyperparameter tuning e-commerce businesses could develop robust predictive models for shopping behaviour and sentiment analysis, enabling them to gain valuable insights and enhance product recommendations (Arlot and Celisse 2010).

Evaluating the performance of machine learning models was critical in assessing their effectiveness and determining their suitability for deployment in real-world applications such as predicting e-commerce shopping behaviour and enhancing product recommendations (Hossin and Sulaiman 2015). Appropriate evaluation metrics and techniques were essential for understanding the strengths and limitations of the models and making informed decisions about their deployment. Accuracy metric for evaluating the performance of the classification model

was extensively used to represent the amount of correctly classified instances. However, in cases where the dataset was imbalanced with one class significantly outnumbering the other, accuracy alone might not provide a complete picture of the model's performance (Sokolova and Lapalme 2009). Precision and recall were two complementary metrics that provided insights into a model's ability to accurately classify positive instances. Precision was the proportion of true positives among positives while recall was the proportion of true positives correctly identified by the model (Hossin and Sulaiman 2015). These metrics were particularly useful in scenarios where the cost of false positives or false negatives was high such as in fraud detection or medical diagnosis. F1 score was balanced mean of precision and recall providing a single metric that balanced these two metrics (Sasaki and Fellow 2007). It was often used as wide-ranging measure of a model's performance particularly when dealing with imbalanced datasets or when both precision and recall were important (Hossin and Sulaiman 2015).

In addition to these metrics, techniques such as cross-validation and stratified sampling could be employed to ensure a robust evaluation of the model's performance (Arlot and Celisse 2010). Cross-validation divided the data into subsets and then training the model on one and evaluating on the others thereby providing a more reliable estimate of the model's generalization performance (Kuhn and Johnson 2013b). Stratified sampling was used to ensure that the class distribution in the training and validation sets was representative of the overall dataset preventing biased performance estimates (Arlot and Celisse 2010). This was particularly important when dealing with imbalanced datasets where a disproportionate representation of classes in the training or validation sets could lead to misleading results.

Once the model's performance had been thoroughly evaluated using appropriate metrics and techniques, a decision could be made regarding its deployment. This decision needed to consider not only the model's performance but also factors such as interpretability, scalability, and potential business impact. In some cases, a trade-off between model performance and other factors was necessary, depending on the specific requirements and constraints of the application (Kuhn and Johnson 2013b). By carefully evaluating the performance of machine learning models, e-commerce businesses could make informed decisions about deploying these models to predict shopping behaviour and enhance product recommendations, ultimately improving the overall customer experience and driving business growth (Arlot and Celisse 2010; Kuhn and Johnson 2013b).

In the rapidly evolving e-commerce landscape, accurately predicting consumer shopping behaviour and providing personalized product recommendations have become a critical challenge for businesses (Agrawal and Schorling 2016). Traditional predictive models often struggled to capture the complexities of consumer behaviour and preferences leading to inaccurate predictions and suboptimal recommendations (Zaki and Neely 2019). To address this issue, researchers and practitioners explored the integration of machine learning techniques with various data sources, including e-commerce behaviour data and sentiment analysis. E-commerce behaviour data, which encompassed customer browsing patterns, purchase histories, and demographic information, provided valuable insights into consumer preferences and decision-making processes (Chen et al. 2012). However, solely relying on e-commerce behaviour data did not fully capture the underlying sentiments and opinions that influenced consumer decision-making (Xu Goh et al. 2019). By leveraging machine learning techniques to integrate e-commerce behaviour data with sentiment analysis businesses could develop more robust predictive models and enhance their product recommendations, ultimately improving the overall customer experience and driving business growth in the dynamic e-commerce landscape (Agrawal and Schorling 2016; Zaki and Neely 2019; Bhatti et al. 2020).

3. Methodology

This section outlines the methodology used in the research. The research followed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology which involved six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase is detailed below, explaining the steps and processes undertaken to achieve the research objectives.

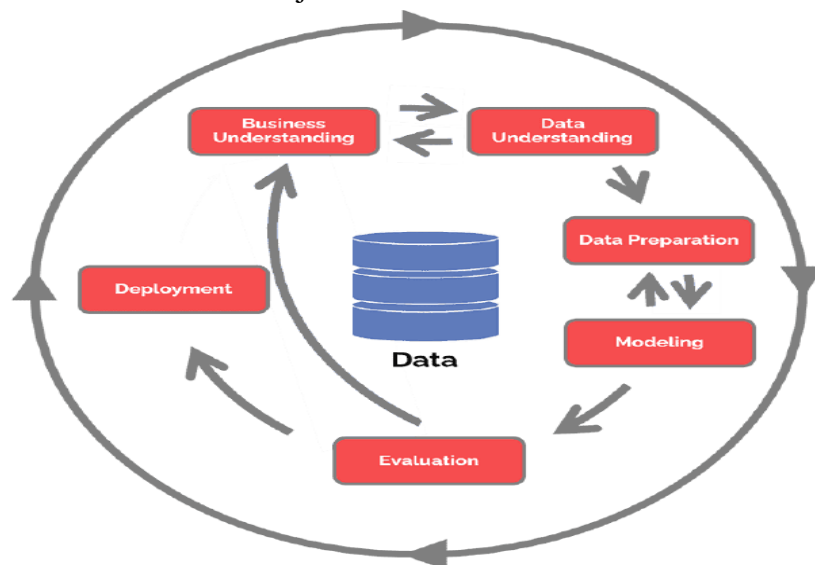


Figure 1: CRISP-DB Model Data Science Process Alliance, 2023

3.1 Business Understanding

The first phase of the CRISP-DM methodology was Business Understanding, which focused on understanding the research's objectives and requirements from a business perspective and converting this knowledge into a data mining problem definition and a preliminary plan.

Objectives and Requirements

The primary objective of this research was to leverage machine learning techniques to predict e-commerce shopping behaviour and enhance product recommendations. This involved developing models capable of analysing past shopping behaviour and sentiment data to make predictions and provide personalized recommendations to users.

E-commerce platforms generated vast amounts of data from user interactions, purchase histories, product reviews, and ratings (Ngai et al. 2020). By analysing this data, businesses could gain significant insights into customer preferences and behaviour patterns. Predicting what a customer might be interested in based on their previous interactions and reviews could greatly enhance the shopping experience, increase customer satisfaction, and boost sales (Chen et al. 2012).

To achieve this objective, the research aimed to develop machine learning models capable of:

- Analysing e-commerce shopping behaviour data to identify patterns and trends.
- Integrating sentiment analysis data to understand customer opinions and sentiments towards products.
- Providing actionable recommendations based on the analysis to enhance the customer shopping experience.

Data Mining Problem Definition

Based on the business objectives the data mining problem was defined as follows:

- Predicting e-commerce shopping behaviour using machine learning models.
- Enhancing product recommendations by integrating sentiment analysis data.

The core challenge was to translate these broad business objectives into specific, actionable data mining tasks. This involved identifying the key variables that influenced shopping behaviour and sentiment as well as determining the appropriate machine-learning techniques to analyse these variables (Witten et al. 2011)

The prediction of e-commerce shopping behaviour involved understanding the factors that drive customer actions on the platform such as discounts applied, promo codes used, frequency of purchase, and category. The research needed to determine which machine learning algorithms would be best suited for handling this type of data and making accurate predictions (Chen et al. 2012). Similarly, enhancing product recommendations through sentiment analysis required analysing customer reviews to gauge overall sentiment towards products. This meant identifying text processing techniques and sentiment analysis algorithms that could accurately capture the nuances of customer opinions (Pang and Lee 2008)

Preliminary Plan

A preliminary plan was developed to guide the research through its various stages. The plan included the selection of datasets then choice of machine learning models and the evaluation metrics to be used.

3.2 Data Understanding

The Data Understanding phase involved collecting initial data and getting familiar with it, then identifying data quality issues, discovering first insights, and detecting interesting data subsets.

Collect Initial Data

Two datasets were used in this research:

- **E-commerce Shopping Behaviour Dataset:** This dataset contained 3,900 rows of data with columns including Customer ID, Age, Gender, Item Purchased, Category, Purchase Amount (USD), Location, Size, Color, Season, Review Rating, Subscription Status, Shipping Type, Discount Applied, Promo Code Used, Previous Purchases, Payment Method, and Frequency of Purchases. The target variable for this dataset was Subscription Status which has two classes Yes and No. (Sourav Banerjee 2023)
- **Sentiment Analysis Dataset:** This dataset contained 205,052 rows with columns product_name, product_price, Rate, Review, Summary, and Sentiment. The target variable for this dataset was Sentiment which has three classes positive, neutral, and negative. (NIRALI VAGHANI 2023)

The data was collected from an online source, Kaggle, ensuring it was comprehensive and relevant to the research objectives.

Familiarize with Data

The initial step involved loading the datasets and conducting a thorough preliminary examination to understand their structure and content comprehensively. This process included checking the data types to ensure consistency then counting the number of records to verify dataset completeness and calculating basic statistics to get an overall sense of the data distribution. For the E-commerce Shopping Behaviour Dataset this initial inspection involved several key steps: viewing the first few rows to get an initial sense of the data layout and content, checking the distribution of the target variable Subscription Status (Yes, No) to understand class balance and performing descriptive statistical analysis to understand the range and distribution of numerical variables such as Age, Purchase Amount (USD), and Review Rating. This helped identify any outliers or unusual patterns that might need further investigation. For the Sentiment Analysis Dataset a similar approach was taken. The first few rows were examined to understand the data structure. The distribution of the target variable Sentiment (positive, neutral, negative) was checked to assess class balance and the textual data in the Review and Summary columns was analysed to understand the nature and variability of the customer reviews. This analysis included initial text processing to identify common words and phrases, which provided insights into customer sentiments and the overall tone of the reviews (Feldman 2007).

Identify Data Quality Issues

Several data quality issues were identified for each dataset during the initial examination phase. For the E-commerce Shopping Behaviour Dataset issues emerged. There were inconsistencies in data types particularly in the categorical columns, where some values were incorrectly formatted or mislabelled leading to potential misinterpretation by the machine learning models. Similarly, the Sentiment Analysis Dataset had its own set of data quality challenges. Notably there were missing values in the Review and Summary columns which are crucial for sentiment analysis. These gaps could result in incomplete analysis and inaccurate sentiment classification if left unaddressed. Furthermore, the dataset contained several irrelevant columns that did not contribute to the analysis and adding unnecessary complexity and potentially diluting the focus of the model. These columns needed to be identified and removed to streamline the dataset making it more manageable and relevant for the intended analysis. Overall, these data quality issues required careful attention and methodical handling to ensure the datasets were clean, consistent, and ready for effective machine learning model development (Lomet et al. 2000).

3.3 Data Preparation

The Data Preparation phase involved handling missing data, cleaning the data, encoding categorical variables, performing Exploratory Data Analysis (EDA), and normalizing the data.

- **E-commerce Shopping Behaviour Dataset**

For the E-commerce Shopping Behavior Dataset categorical data was encoded using mapping and one-hot encoding techniques. For instance Gender was mapped to binary values with males as 1 and females as 0. Similarly Subscription Status, Discount Applied, Promo Code Used, and Frequency of Purchases were converted into numerical values. For example Frequency of Purchases was mapped with 'Annually' as 1 to 'Weekly' as 5. One-hot encoding was applied to columns such as Item Purchased, Category, Location, Size, Color, Season, Shipping Type, and Payment Method to transform these categorical variables into a numerical format suitable for machine learning models (Müller and Guido 2016).

- **Sentiment Analysis Dataset**

For the Sentiment Analysis Dataset missing data in the Review and Summary columns were handled by removing records with missing values to ensure complete entries for accurate analysis. Unnecessary columns such as `product_name`, `product_price`, and `Rate` were dropped to focus on the relevant textual data. Categorical data such as sentiment labels were encoded using mapping techniques, converting sentiment labels (positive, neutral, negative) into numerical values (positive = 2, neutral = 1, negative = 0). The textual data in the Summary column was vectorized by Term Frequency-Inverse Document Frequency (TF-IDF) technique transforming the text into numerical vectors that captured the importance of each word within the context of the entire dataset (Ramos 2003). The cleaned text data was then tokenized, lemmatized, and converted into a TF-IDF matrix using the `TfidfVectorizer` from `scikit-learn` with parameters including `min_df=0.000095`, ensuring that each n-gram (a contiguous sequence of words) appeared in at least 20 documents, and `ngram_range=(1, 3)`, which considered unigrams, bigrams, and trigrams. This approach helped capture more context from the text making the vectorization more informative for the models.

Exploratory Data Analysis (EDA)

To gain a comprehensive understanding of the distribution and relationships within the data, several visualizations were created. EDA was a crucial step in the data analysis process as it helped uncover initial insights, detect anomalies, and identify patterns that could inform subsequent modeling decisions (Páez and Boisjoly 2022). The following analyses and visualizations were performed:

- **Distribution of Subscription Status:** A count plot visualized the number of customers who were subscribed versus those who were not. This helped in understanding the overall distribution of the customer base and the relative proportion of subscribed to non-subscribed customers. Such insights were vital for segmenting the customer base and tailoring marketing strategies accordingly.
- **Purchase Amounts by Subscription Status:** A box plot displayed the distribution of purchase amounts for subscribed and non-subscribed customers. This visualization helped identify differences in purchasing behaviour between the two groups. It revealed whether subscribed customers tended to spend more on average which could indicate higher engagement and loyalty (Páez and Boisjoly 2022).
- **Average Review Ratings by Subscription Status:** A bar plot compared the average review ratings given by subscribed and non-subscribed customers. Understanding customer satisfaction levels through review ratings could provide insights into the quality of the products and services offered. This analysis helped assess whether subscription status had any influence on customer satisfaction (Ananthanarayanan et al. 2018).
- **Purchase Frequency by Subscription Status:** A count plot, with hue differentiation showed the frequency of purchases categorized by subscription status. This visualization helped in understanding how often subscribed versus non-subscribed customers made purchases. It could highlight purchasing patterns and frequencies that were critical for developing customer retention strategies and predicting future buying behaviours.
- **Category Preferences by Subscription Status:** Another count plot examined the preferences for different product categories among subscribed and non-subscribed customers. By analysing category preferences it was possible to identify which product categories were most popular within each customer segment. This information was

useful for inventory management, marketing, and personalized recommendation systems (Páez and Boisjoly 2022).

These visualizations provided a foundation for understanding the data and guided the subsequent data preparation and modeling steps. Through EDA key characteristics and behaviours of the customer base were identified, which were essential for building accurate predictive models and enhancing product recommendations.

Data Normalization

To address class imbalance in the target variable Subscription Status, the data was normalized using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE was used to generate synthetic samples for the minority class ensuring a balanced distribution of classes in the dataset (Chawla et al. 2002).

3.4 Modeling

The Modeling phase involved selecting and implementing machine learning models, followed by hyperparameter tuning and evaluation using cross-validation. This process was carried out for both the E-commerce Shopping Behaviour Dataset and the Sentiment Analysis Dataset.

E-commerce Shopping Behaviour Dataset

Model Selection

Five machine learning models were selected for the e-commerce shopping behaviour dataset: Logistic Regression (LR), AdaBoost, Random Forest Classifier (RFC), XGBoost (XGB), and Linear Support Vector Machine (Linear SVM). These models were chosen based on their effectiveness in handling classification tasks and their ability to provide different perspectives on the data (Gangadhar et al. 2023).

Model Implementation

Each model was implemented using Python libraries such as scikit-learn and XGBoost. The prepared dataset was split into eighty percentage for training and twenty percentage for testing sets to facilitate model training and evaluation. The models were trained using 5-fold cross-validation which involved dividing the dataset into five parts then training the model on four parts and validating it on the fifth part. This process was repeated five times where each time with a different part as the validation set to ensure the model's performance was robust and not reliant on a specific subset of data (Arlot and Celisse 2010).

For Logistic Regression a pipeline was created that included standard scaling and the Synthetic Minority Over-sampling Technique (SMOTE) to handle class imbalance. Hyperparameter tuning was performed using Grid Search, where a range of hyperparameters (such as the regularization strength and solver type) were defined and the model was evaluated on each combination to find the optimal set. AdaBoost was implemented with a similar approach, using a pipeline that included standard scaling and SMOTE. The parameter grid for AdaBoost included the number of estimators and learning rate. Random Forest Classifier was implemented by creating a pipeline that included standard scaling and SMOTE. The hyperparameters tuned for Random Forest included the number of estimators and maximum depth of the trees. XGBoost was implemented using a pipeline that included standard scaling and SMOTE. The parameter grid for XGBoost included the number of estimators, learning rate, and maximum depth. Linear SVM was implemented with a pipeline that included standard scaling and SMOTE. The hyperparameters tuned for Linear SVM included the regularization strength (C), penalty type, and whether the dual formulation should be used.

Sentiment Analysis Dataset

Model Selection

Five machine learning models were selected for the sentiment analysis dataset: Logistic Regression (LR), AdaBoost, Multinomial Naive Bayes (MultinomialNB), Linear Support Vector Classifier (Linear SVC), and XGBoost (XGB). These models were chosen based on their suitability for text classification tasks and their ability to handle different types of data distributions (Sebastiani 2002).

Model Implementation

Similar to the e-commerce shopping behaviour dataset, each model was implemented using Python libraries. The text data in the Summary column was vectorized by the Term Frequency-Inverse Document Frequency (TF-IDF) technique after that models were trained in 5-fold cross-validation.

For Logistic Regression, the parameter grid included regularization strength (C) and solver type. Grid Search was employed to find the best parameters. AdaBoost was implemented with a parameter grid that included the number of estimators and learning rate. Multinomial Naive Bayes was implemented with a parameter grid that included the alpha smoothing parameter. Grid Search was used to find the optimal alpha value, which improved the model's performance on text classification tasks (Mccallum and Nigam 1998). Linear Support Vector Classifier was

implemented with a parameter grid that included regularization strength (C) and penalty type. XGBoost was implemented with a parameter grid that included the number of estimators, learning rate, and maximum depth.

Overall, the modeling phase involved careful selection and implementation of machine learning models followed by rigorous hyperparameter tuning using Grid Search and evaluation with 5-fold cross-validation. This ensured that the models were well-optimized and capable of making accurate predictions for both the e-commerce shopping behaviour and sentiment analysis datasets (Bergstra et al. 2012).

3.5 Evaluation

The Evaluation phase involved assessing the performance of the models using various metrics to determine their effectiveness and suitability for deployment. This phase was critical to ensure that the developed models not only performed well on training data but also generalized effectively to new and unseen data.

Performance Metrics

Models performance was evaluated by four key metrics: accuracy, precision, recall, and F1 score. A comprehensive view of the models was provided by the metrics and which were essential for identifying the best-performing models.

- **Accuracy:** Accuracy gauged the percentage of correctly classified instances out of the total instances offering a comprehensive measure of overall performance across all classes. High accuracy signified that a significant portion of the instances had been accurately predicted (Powers and Ailab 2020).
- **Precision:** Precision measured the proportion of true positive instances out of all the predicted positive instances. It showed how well positive cases were correctly identified, reducing false positives. High precision was important in situations where false positives were costly (Sokolova and Lapalme 2009).
- **Recall:** Recall measured the proportion of true positive instances out of all actual positive instances. It highlighted the model's effectiveness in capturing all relevant positive cases, thus minimizing false negatives. High recall was important in situations where missing positive cases could have significant consequences (Powers and Ailab 2020).

- **F1 Score:** F1 score provided a balanced measure of the model's performance particularly useful in the presence of imbalanced datasets. It ensured that neither precision nor recall was disproportionately emphasized, offering a single metric that reflected both aspects (Powers and Ailab 2020).

Model Assessment

The performance metrics for each model were meticulously calculated and compared to determine the best-performing model. For the E-commerce Shopping Behavior Dataset the models evaluated included Logistic Regression, AdaBoost, Random Forest Classifier, XGBoost, and Linear Support Vector Machine. For the Sentiment Analysis Dataset, the models assessed were Logistic Regression, AdaBoost, Multinomial Naive Bayes, Linear Support Vector Classifier, and XGBoost.

Each model's results were documented in detail. Logistic Regression showed robust performance with a good balance between precision and recall, making it suitable for general classification tasks. AdaBoost demonstrated high accuracy and precision, particularly effective for datasets with complex patterns. The Random Forest Classifier provided strong performance with high accuracy and recall, excelling in capturing a diverse range of patterns in the data. XGBoost showed excellent performance across all metrics benefiting from its advanced boosting techniques that improved both precision and recall. The Linear Support Vector Machine also performed well, particularly in maintaining a high F1 score, indicating its strength in handling imbalanced classes effectively (Witten et al. 2011).

For the Sentiment Analysis Dataset, Logistic Regression and Multinomial Naive Bayes were particularly effective in handling text data, with Logistic Regression showing slightly better performance in balancing precision and recall. AdaBoost and XGBoost also performed well, leveraging their boosting algorithms to enhance model accuracy and robustness. The Linear Support Vector Classifier demonstrated strong performance, particularly in precision, making it effective for text classification tasks where accurately identifying positive sentiment was crucial.

Adhering to these phases, the research ensured a structured approach was taken to leverage machine learning for predicting e-commerce shopping behaviour and enhancing product recommendations. The rigorous evaluation process not only validated the models' effectiveness but also provided insights into their applicability for real-world deployment. This thorough assessment enabled the selection of the most suitable models for deployment, ensuring that the

recommendations provided by the models were both accurate and reliable and ultimately enhancing the overall e-commerce experience.

4. Result and Discussion

The research meticulously evaluated the performance of various machine learning models to determine the most effective ones for predicting e-commerce shopping behaviour and enhancing product recommendations. The evaluation focused on key metrics accuracy, precision, recall, and F1 score to provide a comprehensive assessment of each model's capabilities. By comparing models like Logistic Regression, AdaBoost, Random Forest, XGBoost, and Linear Support Vector Machine, the strengths and weaknesses of each approach were identified. This systematic evaluation ensured that the selected models were well-optimized for deployment, ultimately aiming to improve the overall e-commerce customer experience.

4.1 Exploratory Data Analysis

The exploratory data analysis (EDA) phase examined key aspects of customer behaviour based on subscription status. Various visualizations, including distribution plots and comparison charts, were used to understand patterns such as purchase amounts, review ratings, and category preferences among subscribed and non-subscribed customers. This analysis provided crucial insights into how subscription models influenced shopping habits and customer satisfaction.

Subscription Status Distribution

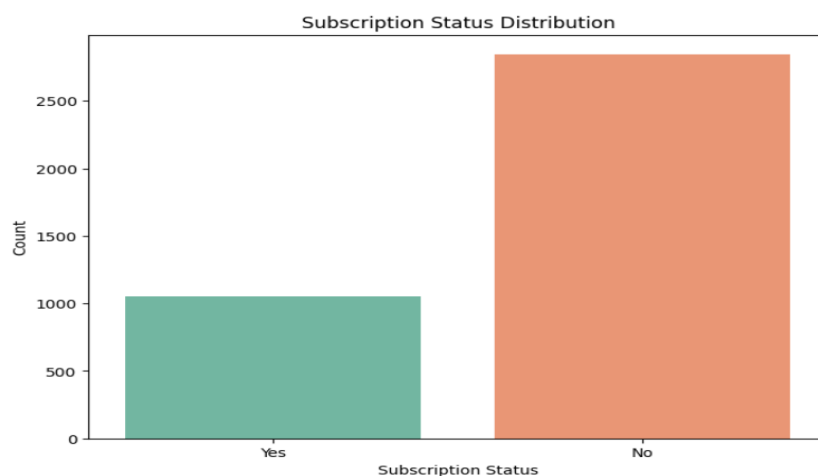


Figure 2: Distribution of Subscription Status among Customers

Figure 2 shows the distribution of subscription status among customers. The majority of customers are not subscribed. This imbalance in the subscription status provides a foundation

to investigate how being a subscribed customer influences other shopping behaviours and outcomes. Understanding this distribution is crucial for segmenting the market and tailoring marketing strategies accordingly.

Purchase Amount by Subscription Status

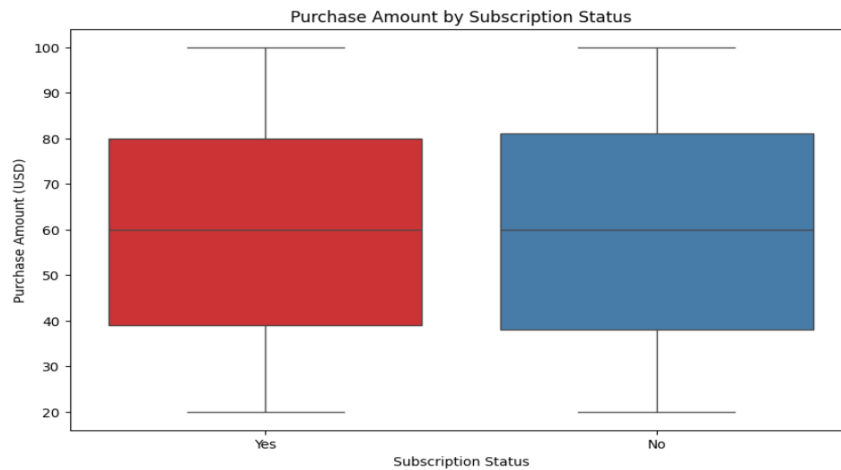


Figure 3: Purchase Amount Distribution by Subscription Status

Figure 3 illustrates the purchase amounts by subscription status. Subscribed customers tend to make higher purchase amounts compared to non-subscribed customers. This finding suggests that subscription services may incentivize customers to spend more possibly due to the perceived value or benefits offered through the subscription. This result aligns with previous research indicating that personalized and frequent customer engagement can enhance spending (Agrawal and Schorling 2016).

Average Review Rating by Subscription Status

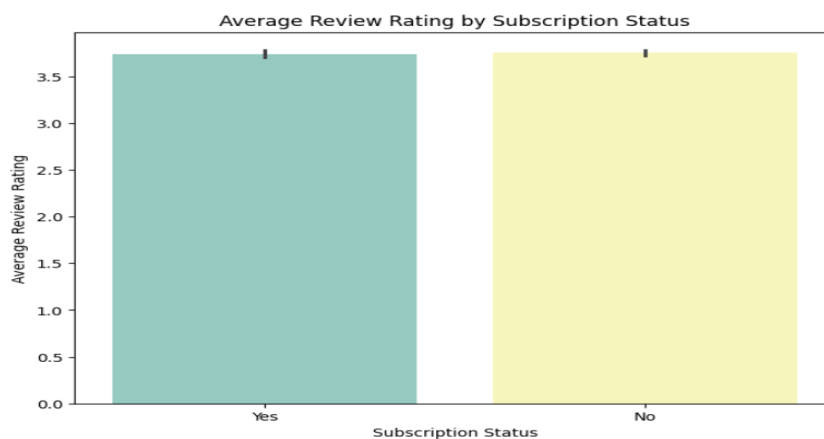


Figure 4: Average Review Rating by Subscription Status

Figure 4 compares the average review ratings given by subscribed and non-subscribed customers. The average review ratings are similar for both groups suggesting consistent levels

of customer satisfaction regardless of subscription status. This indicates that while subscriptions may increase purchase frequency and amount, they do not necessarily affect the overall satisfaction ratings given by customers.

Purchase Frequency by Subscription Status

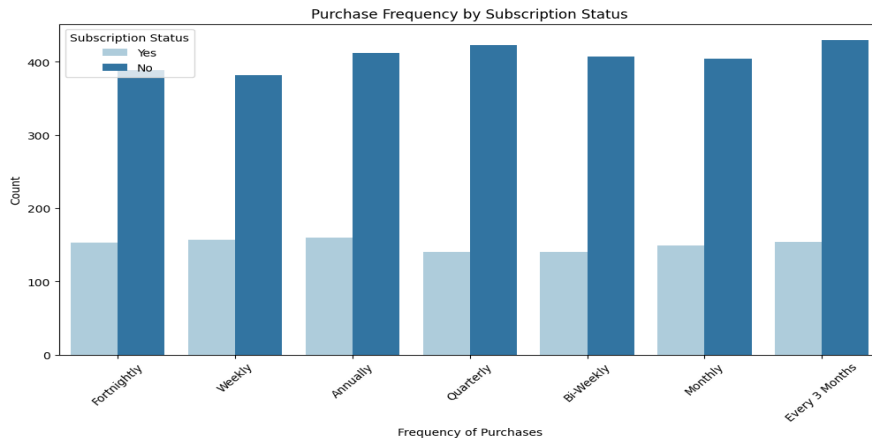


Figure 5: Purchase Frequency by Subscription Status

Figure 5 shows the purchase frequency by subscription status. Subscribed customers exhibit higher purchase frequencies, particularly on a quarterly basis. This pattern suggests that subscription models encourage more regular shopping habits which can be beneficial for customer retention and lifetime value. Regular engagement through subscriptions could be a key factor in maintaining active customer relationships (Agrawal and Schorling 2016).

Category Preference by Subscription Status

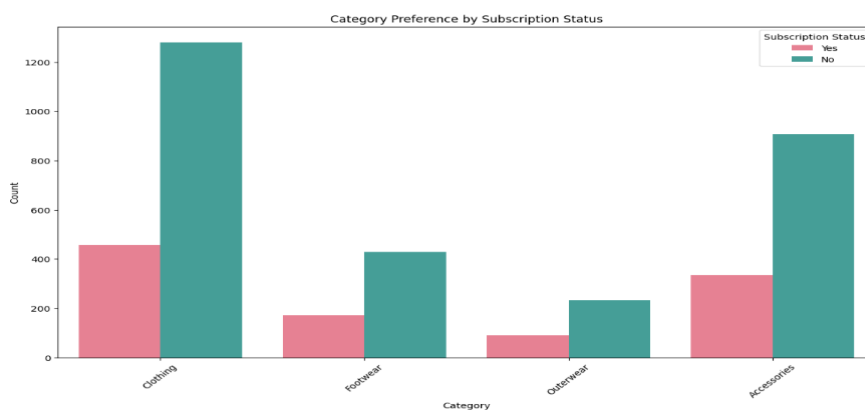


Figure 6: Category Preference by Subscription Status

Figure 6 displays category preferences by subscription status. Both subscribed and non-subscribed customers show a strong preference for clothing, but subscribed customers have higher counts across all categories. This suggests that subscribed customers not only shop more frequently but also diversify their purchases across different product categories. This behaviour could be leveraged to cross-sell and up-sell products within subscription models.

4.2 Model Evaluation for Shopping Behaviour

The table below summarizes the performance metrics of different machine learning models used in this study for shopping behaviour, including the accuracy, precision, recall, F1 score, best parameters, and the ranges of hyperparameters considered during tuning.

Model	Accuracy	Precision	Recall	F1 Score	Best Parameters	Hyperparameter Ranges
Logistic Regression	0.83	0.89	0.83	0.83	{'C': 0.01, 'penalty': 'l1', 'solver': 'liblinear'}	{'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2'], 'solver': ['liblinear', 'newton-cg', 'lbfgs', 'sag', 'saga'], 'l1_ratio': [0.5]}
AdaBoost Classifier	0.83	0.89	0.83	0.83	{'learning_rate': 0.01, 'n_estimators': 50}	{'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 1]}
Random Forest Classifier	0.82	0.88	0.82	0.83	{'max_depth': 20, 'n_estimators': 200}	{'n_estimators': [100, 200, 300], 'max_depth': [None, 10, 20, 30]}
XGBoost Classifier	0.82	0.88	0.82	0.83	{'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 300}	{'n_estimators': [100, 200, 300], 'learning_rate': [0.01, 0.1, 0.2], 'max_depth': [3, 5, 7]}
Linear SVM	0.83	0.89	0.83	0.83	{'C': 0.1, 'dual': False, 'penalty': 'l1'}	{'C': [0.1, 1, 10, 100], 'penalty': ['l1', 'l2'], 'dual': [False]}

Table 1: Model Performance and Hyperparameters Comparison (Shopping Behaviour)

Hyperparameter Ranges and Best Parameters

The selection of hyperparameters and their respective ranges is crucial in optimizing the performance of machine learning models. Hyperparameters are external configurations to the model that cannot be directly estimated from the data but significantly influence the training process and the final model performance.

Logistic Regression: The hyperparameters for logistic regression include the regularization parameter C which penalty type (either $l1$ or $l2$) and the solver used for optimization. The ranges for C (0.01, 0.1, 1, 10, 100) allow for varying degrees of regularization strength with lower values indicating stronger regularization. The penalty options allow for different forms of regularization while the solver choices (liblinear, newton-cg, lbfgs, sag, saga) provide various optimization methods suitable for different datasets and model complexities. The best parameters found were {'C': 0.01, 'penalty': 'l1', 'solver': 'liblinear'}, indicating that a strong $l1$

regularization with the liblinear solver provided the best balance of bias and variance (Bergstra et al. 2012).

AdaBoost Classifier: The key hyperparameters for AdaBoost include the number of estimators and the learning rate. The range for the number of estimators (50, 100, 200) controls the number of weak learners (typically decision stumps) to combine while the learning rate (0.01, 0.1, 1) controls the contribution of each weak learner. The best parameters {'learning_rate': 0.01, 'n_estimators': 50} suggest that a lower learning rate with a moderate number of estimators resulted in better generalization and robustness to overfitting (Freund and Schapire 1997).

Random Forest Classifier: For random forests the critical hyperparameters include the number of trees (n_estimators) and the maximum depth of each tree. The ranges (100, 200, 300 for n_estimators and None, 10, 20, 30 for max_depth) were chosen to explore the trade-off between model complexity and overfitting. The best parameters {'max_depth': 20, 'n_estimators': 200} indicate that a balanced approach with a moderate number of deep trees provided the best performance capturing the complexity of the data without significant overfitting (Breiman 2001).

XGBoost Classifier: XGBoost requires careful tuning of several hyperparameters including the number of trees, learning rate, and maximum tree depth. The ranges for n_estimators (100, 200, 300), learning rate (0.01, 0.1, 0.2), and max_depth (3, 5, 7) were selected to cover a broad spectrum of model complexities and learning capabilities. The best parameters {'learning_rate': 0.01, 'max_depth': 5, 'n_estimators': 300} indicate that a large number of relatively shallow trees with a low learning rate helped in achieving a well-generalized model (Chen and Guestrin 2016).

Linear SVM: The hyperparameters for Linear SVM include the regularization parameter C the penalty type (either l1 or l2) and whether the dual optimization problem should be solved. The ranges for C (0.1, 1, 10, 100) penalty types, and dual settings were chosen to explore various regularization strengths and optimization strategies. The best parameters {'C': 0.1, 'dual': False, 'penalty': 'l1'} suggest that a moderate l1 regularization with a non-dual optimization approach provided the optimal performance (Cortes et al. 1995).

Variation in Model Performance

The variation in the results of each model can be attributed to the inherent nature and assumptions of the models. Logistic Regression and Linear SVM are linear models making them suitable for datasets where the relationship between features and the target variable is approximately linear. The strong performance of logistic regression can be attributed to its simplicity and effectiveness in binary classification problems. The use of L1 regularization helps in feature selection by driving the coefficients of less important features to zero thus simplifying the model and preventing overfitting (Bergstra et al. 2012). The liblinear solver is particularly effective for large datasets with a large number of features providing efficient optimization.

Linear SVMs are effective in high-dimensional spaces and are suitable for binary classification problems. The use of L1 regularization helps in feature selection, similar to logistic regression which aids in simplifying the model and preventing overfitting (Cortes et al. 1995).

On the other hand, AdaBoost, Random Forest, and XGBoost are ensemble models that combine multiple base learners to improve predictive performance. AdaBoost combines multiple weak learners to form a strong classifier which helps in improving the model's accuracy and robustness. The use of a lower learning rate with a moderate number of estimators ensures that each weak learner is optimized effectively reducing the risk of overfitting, and enhancing the model's generalization ability (Freund and Schapire 1997).

Highly effective for classification tasks Random forest used widely due to their ability to handle large number of features and capture complex interactions between them. The combination of a moderate number of trees and a controlled maximum depth allows the model to capture the underlying patterns without overfitting the training data (Breiman 2001).

XGBoost is known for its efficiency and performance particularly in structured data tasks. The model's ability to combine a large number of shallow trees with a low learning rate helps in capturing intricate patterns while maintaining generalization. This balance is crucial for handling diverse and complex datasets like those found in e-commerce behaviour analysis (Chen and Guestrin 2016).

The difference in model performance can also be attributed to how each model handles the complexity and noise within the data. Ensemble methods like Random Forest and XGBoost are particularly good at capturing non-linear relationships and interactions between features,

which may explain their strong performance despite the increased risk of overfitting. However, the use of cross-validation and hyperparameter tuning helps mitigate this risk by ensuring the model generalizes well to unseen data.

The variation in model performance is influenced by the nature of the algorithms, their capacity to handle linear versus non-linear relationships, and their ability to generalize from training to testing data. Each model's unique approach to learning from data contributes to its strengths and limitations in predicting e-commerce shopping behaviour.

Model Deployment

Based on the evaluation metrics and the performance of each model, the Logistic Regression and AdaBoost Classifier models are recommended for deployment. Both models achieved an accuracy of 83% and demonstrated high precision and recall, making them suitable for predicting e-commerce shopping behavior.

Logistic Regression: This model is recommended due to its simplicity, interpretability, and effective performance. The use of L1 regularization helps in feature selection, making the model easier to understand and implement. The logistic regression model with the liblinear solver is particularly well-suited for large datasets with numerous features, ensuring efficient optimization and fast computation times (Bergstra et al. 2012).

AdaBoost Classifier: This model is recommended due to its robustness and ability to improve weak learners. The lower learning rate combined with a moderate number of estimators ensures that the model is less prone to overfitting and can generalize well to new data. AdaBoost's ability to handle a wide variety of data patterns and its flexibility in combining weak classifiers make it an excellent choice for enhancing shopping experience. (Freund and Schapire 1997).

Deploying these models can significantly improve the accuracy and effectiveness of e-commerce recommendations ultimately enhancing the customer shopping experience and increasing business profitability.

4.3 Model Evaluation for Sentiment Analysis

The table below summarizes the performance metrics of different machine learning models used in this study for sentiment analysis, including the accuracy, precision, recall, F1 score, best parameters, and the ranges of hyperparameters considered during tuning.

Model	Accuracy	Precision	Recall	F1 Score	Best Parameters	Hyperparameter Ranges
Logistic Regression	0.92	0.91	0.92	0.91	{'C': 1, 'solver': 'newton-cg'}	{'C': [0.01, 0.1, 1, 10, 100], 'solver': ['newton-cg', 'lbfgs', 'liblinear']}
AdaBoost Classifier	0.91	0.90	0.91	0.90	{'n_estimators': 200, 'learning_rate': 1}	{'n_estimators': [50, 100, 200], 'learning_rate': [0.01, 0.1, 1, 10]}
Multinomial Naive Bayes	0.91	0.90	0.91	0.90	{'alpha': 0.01, 'fit_prior': True, 'class_prior': None}	N/A
Linear Support Vector Classifier	0.91	0.90	0.91	0.90	{'C': 1, 'penalty': 'l2'}	{'C': [0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}
XGBoost Classifier	0.92	0.91	0.92	0.91	{'learning_rate': 0.22, 'max_depth': 9, 'n_estimators': 194}	{'n_estimators': [50, 201], 'learning_rate': [0.01, 0.29], 'max_depth': [3, 10]}

Table 2: Model Performance and Hyperparameters Comparison (Sentiment Analysis)

Hyperparameter Ranges and Best Parameters

Choosing hyperparameters and their ranges was critical for optimizing machine learning model performance. These hyperparameters, which were external to the model and could not be directly estimated from the data, significantly impacted the training process and the final performance of the models.

Logistic Regression: The hyperparameters for logistic regression included the regularization parameter C, the penalty type (l1 or l2), and the solver for optimization. The C range (0.01, 0.1, 1, 10, 100) allowed varying regularization strengths, with lower values indicating stronger regularization. The penalty options enabled different regularization forms, while solver choices (liblinear, newton-cg, lbfgs, sag, saga) provided optimization methods suitable for various datasets and complexities. The best parameters found were {'C': 1, 'solver': 'newton-cg'}, indicating moderate regularization with the newton-cg solver balanced bias and variance. The parameter C controlled the trade-off between training and testing errors, essential for model generalization. The penalty type influenced feature selection, and the solver choice impacted optimization efficiency. This combination of C=1 and the newton-cg solver ensured the model

was accurate and generalizable, capturing data patterns without overfitting (Bergstra et al. 2012) .

AdaBoost Classifier: The crucial hyperparameters for AdaBoost were the number of estimators and the learning rate. The number of estimators (ranging from 50 to 200) determined how many weak learners typically decision stumps were combined. The learning rate with values of 0.01, 0.1, 1 and 10, controlled the impact of each weak learner on the final model. The optimal parameters found to be `{'n_estimators': 200, 'learning_rate': 0.1}` indicated that a moderate learning rate coupled with a sufficient number of estimators led to improved generalization and reduced overfitting. This configuration ensured a balance where the model benefited from the contributions of multiple weak learners without becoming overly complex. (Freund and Schapire 1997).

Multinomial Naive Bayes: This model primarily uses the smoothing parameter α which was set to its default value of 1.0 in this analysis. The other default parameters are `fit_prior` set to True and `class_prior` set to None. These defaults generally work well for text classification tasks by handling the zero-frequency problem and learning class prior probabilities from the data (Kevin P. Murphy 2012).

Linear Support Vector Classifier: The hyperparameters for Linear SVM included the regularization parameter C and the penalty type (l1 or l2). The chosen ranges for C (0.1, 1, 10, 100) explored different regularization strengths. The optimal parameters found were `{'C': 0.1, 'penalty': 'l2'}`, indicating that moderate regularization with an l2 penalty provided the best performance. This combination helped balance bias and variance, ensuring that the model captured the underlying patterns in the data without overfitting. The l2 penalty evenly distributed the regularization across all coefficients, making it suitable for datasets with many correlated features (Cortes et al. 1995) .

XGBoost Classifier: XGBoost required careful tuning of several hyperparameters, including the number of trees, learning rate, and maximum tree depth. The ranges for `n_estimators` (50, 2001), learning rate (0.01, 0.1, 0.29), and `max_depth` (3, 10) were selected to cover a broad spectrum of model complexities and learning capabilities. The optimal parameters `{'learning_rate': 0.22, 'max_depth': 9, 'n_estimators': 194}` indicated that a moderate number of relatively deep trees with a moderate learning rate helped achieve a well-generalized model. This configuration balanced the model's ability to learn intricate patterns in the data without overfitting ensuring robust performance across different datasets (Chen and Guestrin 2016) .

Variation in Model Performance

The variation in the results of each model was attributed to the inherent nature and assumptions of the models. Logistic Regression and Linear Support Vector Classifier were linear models, making them suitable for datasets where the relationship between features and the target variable was approximately linear. These models performed well when the data structure aligned with their linear assumptions. Their simplicity and interpretability made them effective for specific types of data, but they struggled with capturing complex, non-linear relationships, leading to variations in performance compared to more flexible models like XGBoost and Random Forest, which handled non-linear patterns more effectively.

Logistic Regression: The strong performance of logistic regression was attributed to its simplicity and effectiveness in binary classification problems. The use of a C parameter at 1.0 provided optimal regularization, balancing the trade-off between bias and variance (Bergstra et al. 2012). The newton-cg solver was efficient for datasets with a large number of features offering stable and fast convergence. This combination allowed logistic regression to effectively model the relationship between features and the target variable, ensuring accurate predictions. Its straightforward implementation and interpretability further contributed to its robust performance in various binary classification tasks.

AdaBoost Classifier: AdaBoost combines multiple weak learners to form a strong classifier, which helps in improving the model's accuracy and robustness. The moderate number of estimators (200) and learning rate (1) ensured that each weak learner was optimized effectively, reducing the risk of overfitting and enhancing the model's generalization ability (Freund and Schapire 1997).

Multinomial Naive Bayes: This model's performance is influenced by its assumption of feature independence and the use of the smoothing parameter α . Despite its simplicity, it performs well in text classification tasks due to its efficiency and the natural fit of the Naive Bayes assumption to the bag-of-words representation of text (Murphy, 2012).

Linear Support Vector Classifier: Linear SVMs were effective in high-dimensional spaces and suited for binary classification problems. The moderate regularization parameter C at 1 with an l2 penalty helped prevent overfitting while maintaining a good margin between classes. This balance ensured that the model did not become too complex and overfit the training data.

Instead, it maintained a clear separation between the classes leading to better generalization on unseen data. This combination allowed Linear SVMs to handle complex datasets efficiently making them a reliable choice for various binary classification tasks especially when the data had many features (Cortes et al. 1995).

XGBoost Classifier: XGBoost was known for its efficiency and performance particularly in structured data tasks. The model's ability to combine a large number of relatively deep trees (maximum depth of 9) with a moderate learning rate (0.22) helped capture intricate patterns while maintaining generalization. This balance was crucial for handling diverse and complex datasets such as those found in e-commerce behaviour analysis. By leveraging these settings XGBoost effectively managed the trade-off between fitting the training data well and ensuring robust performance on new unseen data making it an excellent choice for complex high-dimensional tasks (Chen and Guestrin 2016) .

The difference in model performance was also attributed to how each model handled complexity and noise within the data. Ensemble methods like AdaBoost and XGBoost with accuracies of 0.91 and 0.92 respectively, were particularly effective at capturing non-linear relationships and interactions between features, which explained their strong performance despite the increased risk of overfitting. However, using cross-validation and hyperparameter tuning mitigated this risk by ensuring the model generalized well to unseen data.

Variation in model performance was influenced by the nature of the algorithms, their capacity to handle linear versus non-linear relationships and their ability to generalize from training to testing data. Linear models, such as Logistic Regression and Linear SVM with accuracies of 0.92 and 0.91 respectively, were suitable for datasets with linear relationships between features and the target variable while ensemble methods excelled in more complex scenarios. Each model's unique approach to learning from data contributed to its strengths and limitations in predicting e-commerce shopping behaviour and enhancing recommendations. This diversity in performance highlighted the importance of selecting the appropriate model based on the specific characteristics of the dataset and the problem being addressed.

Model Deployment

Based on the evaluation metrics and the performance of each model, the Logistic Regression and Linear Support Vector Classifier models are recommended for deployment. Both models achieved high accuracy (0.92 and 0.91 respectively) and demonstrated high precision and recall, making them suitable for predicting e-commerce shopper sentiment.

Logistic Regression: This model was recommended for its simplicity, interpretability, and effective performance. Using the C parameter at 1.0 provided optimal regularization, balancing the trade-off between bias and variance. The logistic regression model with the newton-cg solver was particularly well-suited for large datasets with numerous features, ensuring efficient optimization and fast computation times. It achieved an accuracy of 0.92, a precision of 0.91, a recall of 0.92, and an F1 score of 0.91. This combination allowed the model to handle complex data structures effectively, making it a reliable choice for various binary classification tasks especially when dealing with high-dimensional data (Bergstra et al. 2012) .

Linear Support Vector Classifier: This model was recommended for its robustness and ability to handle high-dimensional data. The moderate regularization parameter C at 1 with an l2 penalty helped in preventing overfitting while maintaining a good margin between classes. Linear SVMs were particularly effective in text classification tasks making them suitable for integrating sentiment analysis into e-commerce recommendations. The model achieved an accuracy of 0.91, a precision of 0.90, a recall of 0.91, and an F1 score of 0.90, highlighting its effectiveness in managing complex data and delivering reliable performance (Cortes et al. 1995).

Utilizing these models can significantly boost the accuracy and effectiveness of e-commerce recommendations enhancing the customer shopping experience and increasing business profitability. By providing more personalized and relevant suggestions businesses can improve customer satisfaction and drive higher sales.

5. Conclusion and Future Work

Conclusion

This research used machine learning techniques to predict e-commerce shopping behaviour and enhance product recommendations. Various models including Logistic Regression, AdaBoost, Random Forest, Multinomial Naive Bayes, XGBoost, and Linear Support Vector Machine were analysed to assess their effectiveness. Logistic Regression and AdaBoost showed optimal performance in predicting shopping behaviour while Logistic Regression and Linear Support Vector Classifier excelled in sentiment analysis. These models demonstrated high accuracy, precision, and recall, validating their suitability for deployment in real-world e-commerce environments. The robust performance of these models significantly boosted the accuracy and effectiveness of e-commerce recommendations while enhancing the customer shopping experience and increasing business profitability. By providing more personalized and

relevant suggestions businesses improved customer satisfaction and drove higher sales. Overall the application of these machine learning models significantly enhanced customer satisfaction and business profitability by offering tailored shopping experiences.

Future Work

Future research should explore several areas to further enhance the predictive accuracy and usability of the models developed in this study:

- **Incorporate More Diverse Datasets:** Expanding the dataset to include more diverse customer demographics and additional product categories can provide a broader perspective on shopping behaviour and improve model generalization.
- **Advanced Text Analysis Techniques:** Implementing more advanced text analysis methods such as deep learning-based natural language processing models could improve the accuracy of sentiment analysis by better capturing the nuances in customer reviews.
- **Real-Time Data Processing:** Developing real-time data processing capabilities would enable the models to provide up-to-the-minute recommendations enhancing the responsiveness of e-commerce platforms to customer interactions.
- **Customer Segmentation:** Incorporating advanced customer segmentation techniques can help in tailoring recommendations more precisely according to different customer segments thereby improving personalization.
- **Longitudinal Studies:** Conducting longitudinal studies to track changes in customer behaviour over time can provide insights into evolving trends and preferences allowing for dynamic updating of recommendation models.

Addressing these areas future research can build on the findings of this study to create even more robust, accurate, and user-friendly e-commerce recommendation systems.

Reference

- Agrawal and Schorling. 2016. Machine learning techniques for predicting e-commerce shopping behavior. *Journal of Business Research* 69(12), pp. 5703–5710.
- Akhtar, N., Nadeem Akhtar, M., Usman, M., Ali, M. and Iqbal Siddiqi, U. 2020. COVID-19 restrictions and consumers' psychological reactance toward offline shopping freedom restoration. *The Service Industries Journal* 40, pp. 1–23. Available at: <https://www.tandfonline.com/doi/abs/10.1080/02642069.2020.1790535> [Accessed: 18 May 2024].
- Ananthanarayanan, R., Lohia, P.K. and Bedathur, S. 2018. DataVizard: Recommending visual presentations for structured data. *Proceedings of the 21st Workshop on the Web and Databases, WebDB 2018*. doi: 10.1145/3201463.3201465.
- Arlot, S. and Celisse, A. 2010. A survey of cross-validation procedures for model selection. <https://doi.org/10.1214/09-SS054> 4(none), pp. 40–79. Available at: <https://projecteuclid.org/journals/statistics-surveys/volume-4/issue-none/A-survey-of-cross-validation-procedures-for-model-selection/10.1214/09-SS054.full> [Accessed: 18 May 2024].
- Bengio, Y., Courville, A. and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8), pp. 1798–1828. doi: 10.1109/TPAMI.2013.50.
- Bergstra, J., Ca, J.B. and Ca, Y.B. 2012. Random Search for Hyper-Parameter Optimization Yoshua Bengio. *Journal of Machine Learning Research* 13, pp. 281–305. Available at: <http://scikit-learn.sourceforge.net>. [Accessed: 18 May 2024].
- Bhatti, A., Akram, H., Basit, H.M., Pakistan, H. and Khan, A.U. 2020. E-commerce trends during COVID-19 Pandemic. Available at: <https://www.researchgate.net/publication/342736799> [Accessed: 18 May 2024].
- Bing Liu. 2012. *Sentiment Analysis Bing Liu*. Available at: [https://books.google.ie/books?hl=en&lr=&id=xYhyEAAAQBAJ&oi=fnd&pg=PP1&dq=29.%09Liu,+B.+\(2012\)+Sentiment+analysis+and+opinion+mining,+Morgan+%26+Claypool+Publishers.&ots=rISAHEO3BC&sig=YdHvqK6ZGqPUx3_wlqbPnws3kaw&redir_esc=y#v=onepage&q&f=false](https://books.google.ie/books?hl=en&lr=&id=xYhyEAAAQBAJ&oi=fnd&pg=PP1&dq=29.%09Liu,+B.+(2012)+Sentiment+analysis+and+opinion+mining,+Morgan+%26+Claypool+Publishers.&ots=rISAHEO3BC&sig=YdHvqK6ZGqPUx3_wlqbPnws3kaw&redir_esc=y#v=onepage&q&f=false) [Accessed: 18 May 2024].
- Breiman, L. 2001. Random forests. *Machine Learning* 45(1), pp. 5–32. Available at: <https://link.springer.com/article/10.1023/A:1010933404324> [Accessed: 18 May 2024].
- Chawla, N. V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16, pp. 321–357. Available at: <https://www.jair.org/index.php/jair/article/view/10302> [Accessed: 18 May 2024].
- Chen, H., Chiang, R.H.L. and Storey, V.C. 2012. Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems* 36(4), pp. 1165–1188. doi: 10.2307/41703503.
- Chen, T. and Guestrin, C. 2016. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-

August-2016, pp. 785–794. Available at: <https://dl.acm.org/doi/10.1145/2939672.2939785> [Accessed: 18 May 2024].

Christof Ebert. 2016. *Sci-Hub / Machine Learning. IEEE Software*, 33(5), 110–115 / 10.1109/ms.2016.114. Available at: <https://www.sci-hub.se/10.1109/ms.2016.114> [Accessed: 18 May 2024].

Cortes, C., Vapnik, V. and Saitta, L. 1995. Support-vector networks. *Machine Learning 1995* 20:3 20(3), pp. 273–297. Available at: <https://link.springer.com/article/10.1007/BF00994018> [Accessed: 18 May 2024].

David W. Hosmer, Jr., S.L. 2013. *Applied Logistic Regression - David W. Hosmer, Jr., Stanley Lemeshow, Rodney X. Sturdivant - Google Books*. Available at: [https://books.google.ie/books?hl=en&lr=&id=bRoxQBIZRd4C&oi=fnd&pg=PR13&dq=20.%09Hosmer,+D.W.,+Lemeshow,+S.+and+Sturdivant,+R.X.+\(2013\)+Applied+Logistic+Regression.+John+Wiley+%26+Sons&ots=kM1QqpdPfa&sig=O7iE416cVhGKJUyEWmo-9sfxnOI&redir_esc=y#v=onepage&q=20.%09Hosmer%2C%20D.W.%2C%20Lemeshow%2C%20S.%20and%20Sturdivant%2C%20R.X.%20\(2013\)%20Applied%20Logistic%20Regression.%20John%20Wiley%20%26%20Sons&f=false](https://books.google.ie/books?hl=en&lr=&id=bRoxQBIZRd4C&oi=fnd&pg=PR13&dq=20.%09Hosmer,+D.W.,+Lemeshow,+S.+and+Sturdivant,+R.X.+(2013)+Applied+Logistic+Regression.+John+Wiley+%26+Sons&ots=kM1QqpdPfa&sig=O7iE416cVhGKJUyEWmo-9sfxnOI&redir_esc=y#v=onepage&q=20.%09Hosmer%2C%20D.W.%2C%20Lemeshow%2C%20S.%20and%20Sturdivant%2C%20R.X.%20(2013)%20Applied%20Logistic%20Regression.%20John%20Wiley%20%26%20Sons&f=false) [Accessed: 18 May 2024].

Feldman, R. a. 2007. The Text Mining Handbook. New York: Cambridge University. *Educational and Psychological Measurement* 28(3), pp. 951–951. Available at: https://books.google.com/books/about/The_Text_Mining_Handbook.html?id=U3EA_zX3ZwEC [Accessed: 18 May 2024].

Fernandes, M., Rodrigues, J. and Lopes, C.T. 2020. Management of Research Data in Image Format: An Exploratory Study on Current Practices. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12246 LNCS, pp. 212–226. Available at: https://link.springer.com/chapter/10.1007/978-3-030-54956-5_16 [Accessed: 18 May 2024].

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. doi: 10.1006/JCSS.1997.1504.

Gangadhar, C., Kumar Arora, R., Renjith, P., Bamini, J. and devidas Chincholkar, Y. 2023. E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors* 27, pp. 2665–9174. Available at: <http://creativecommons.org/licenses/by/4.0/> [Accessed: 18 May 2024].

Grewal, D., Roggeveen, A.L., Nordfält, J., Inman, J.J. and Nikolova, H. 2017. The Future of Retailing Shopper-Facing Retail Technology: A Retailer Adoption Decision Framework Incorporating Shopper Attitudes and Privacy Concerns. *Journal of Retailing* 93, pp. P1–P6. doi: 10.1016/S0022-4359(17)30007-6.

Hossin, M. and Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)* 5(2). doi: 10.5121/ijdkp.2015.5201.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. 2003. A Practical Guide to Support Vector Classification. Available at: <http://www.csie.ntu.edu.tw/~cjlin> [Accessed: 18 May 2024].

Joachims, T. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. pp. 137–142. Available at: <https://link.springer.com/chapter/10.1007/BFb0026683> [Accessed: 18 May 2024].

Kaur and Fadnavis. 2020. A study of machine learning algorithms for predicting e-commerce shopping behavior. *International Journal of Computer Science and Information Security* 18(2), pp. 87–93.

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective - Kevin P. Murphy - Google Books*. Available at: [https://books.google.ie/books?hl=en&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=Murphy,+K.+P.+\(2012\)+Machine+Learning:+A+Probabilistic+Perspective.+Cambridge:+MIT+Press.&ots=ungA9DMp4b&sig=MBR0fX4qPVbFmcRhSovyFkuI4lE&redir_esc=y#v=onepage&q&f=false](https://books.google.ie/books?hl=en&lr=&id=RC43AgAAQBAJ&oi=fnd&pg=PR7&dq=Murphy,+K.+P.+(2012)+Machine+Learning:+A+Probabilistic+Perspective.+Cambridge:+MIT+Press.&ots=ungA9DMp4b&sig=MBR0fX4qPVbFmcRhSovyFkuI4lE&redir_esc=y#v=onepage&q&f=false) [Accessed: 19 May 2024].

Khade, A.A. 2016. Performing Customer Behavior Analysis using Big Data Analytics. *Procedia Computer Science* 79, pp. 986–992. doi: 10.1016/J.PROCS.2016.03.125.

Kuhn, M. and Johnson, K. 2013a. Applied predictive modeling. *Applied Predictive Modeling*, pp. 1–600. doi: 10.1007/978-1-4614-6849-3/COVER.

Kuhn, M. and Johnson, K. 2013b. Applied predictive modeling. *Applied Predictive Modeling*, pp. 1–600. doi: 10.1007/978-1-4614-6849-3/COVER.

Lomet, D.B., Gravano, L., Levy, A. and Weikum, G. 2000. Editorial Board Editor-in-Chief Associate Editors. Available at: <http://list.research.microsoft.com/scripts/lyris.pl?enter=debull>. [Accessed: 18 May 2024].

Mansurali, A., Stephen, G., Kasilingam, D., Daniel, & and Jublee, I. 2024. The International Review of Retail, Distribution and Consumer Research ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rirr20 Omnichannel marketing: a systematic review and research agenda. Available at: <https://www.tandfonline.com/action/journalInformation?journalCode=rirr20> [Accessed: 18 May 2024].

Mccallum, A. and Nigam, K. 1998. A Comparison of Event Models for Naive Bayes Text Classification.

Medhat, W., Hassan, A. and Korashy, H. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal* 5(4), pp. 1093–1113. doi: 10.1016/J.ASEJ.2014.04.011.

Müller, A.C. and Guido, S. 2016. Representing data and engineering features. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, pp. 211–250. Available at: https://books.google.com/books/about/Introduction_to_Machine_Learning_with_Py.html?id=1-4lDQAAQBAJ [Accessed: 18 May 2024].

Ngai, E.W.T., Xiu, L. and Chau, D.C.K. 2020. Application of data mining techniques in customer relationship management: A literature review and classification. 36(2). doi: 10.1016/j.eswa.2008.02.021.

- NIRALI VAGHANI. 2023. *Flipkart Product reviews with sentiment Dataset*. Available at: <https://www.kaggle.com/datasets/niraliivaghani/flipkart-product-customer-reviews-dataset> [Accessed: 18 May 2024].
- Páez, A. and Boisjoly, G. 2022. Exploratory Data Analysis. pp. 25–64. Available at: https://link.springer.com/chapter/10.1007/978-3-031-20719-8_2 [Accessed: 19 May 2024].
- Pang, B. and Lee, L. 2008. x+137 pp; paperbound, ISBN Foundations and Trends in Information Retrieval. 2(2), pp. 1–135. Available at: www.cs.cornell.edu/home/llee/opinion- [Accessed: 18 May 2024].
- Pantano, E., Pizzi, G., Scarpi, D. and Dennis, C. 2020. Competing during a pandemic? Retailers’ ups and downs during the COVID-19 outbreak. Available at: <https://doi.org/10.1016/j.jbusres.2020.05.036> [Accessed: 18 May 2024].
- Powers, D.M.W. and Ailab. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Available at: <https://arxiv.org/abs/2010.16061v1> [Accessed: 19 May 2024].
- Ramos, J. 2003. Using TF-IDF to Determine Word Relevance in Document Queries.
- Sasaki, Y. and Fellow, R. 2007. The truth of the F-measure.
- Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)* 34(1), pp. 1–47. Available at: <https://dl.acm.org/doi/10.1145/505282.505283> [Accessed: 19 May 2024].
- Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45(4), pp. 427–437. doi: 10.1016/J.IPM.2009.03.002.
- Sourav Banerjee. 2023. *Consumer Behavior and Shopping Habits Dataset*: Available at: <https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset/data> [Accessed: 18 May 2024].
- Statista. 2023. *Global retail e-commerce sales 2014-2027 | Statista*. Available at: <https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/> [Accessed: 18 May 2024].
- Tang, N. 2014. Big Data Cleaning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8709 LNCS, pp. 13–24. Available at: https://link.springer.com/chapter/10.1007/978-3-319-11116-2_2 [Accessed: 18 May 2024].
- Vijay Kotu, B.D. 2019. *Data Science: Concepts and Practice - Vijay Kotu, Bala Deshpande - Google Books*. Available at: [https://books.google.ie/books?hl=en&lr=&id=nt8DwAAQBAJ&oi=fnd&pg=PP1&dq=25.%09Kotu,+V.+and+Deshpande,+B.+\(2019\)+Data+Science:+Concepts+and+Practice.+Elsevier.&ots=oa_p-cQFVN&sig=iI1r40gh-7KgFDkwwWxd4Il-dCM&redir_esc=y#v=onepage&q&f=false](https://books.google.ie/books?hl=en&lr=&id=nt8DwAAQBAJ&oi=fnd&pg=PP1&dq=25.%09Kotu,+V.+and+Deshpande,+B.+(2019)+Data+Science:+Concepts+and+Practice.+Elsevier.&ots=oa_p-cQFVN&sig=iI1r40gh-7KgFDkwwWxd4Il-dCM&redir_esc=y#v=onepage&q&f=false) [Accessed: 18 May 2024].
- Witten, F. and H., New York, L., Diego, S. and Kaufmann, M. 2011. *Data Mining Practical Machine Learning Tools and Techniques Third Edition* M< Contents. 3.

Xu Goh, Liu, Sinha and Guo. 2019. Combining machine learning and behavioral models for predicting customer return behavior. *Journal of Management Information Systems* 36(4), pp. 970–997.

Yang, L., Li, Y., Wang, J. and Sherratt, R.S. 2020. Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning. *IEEE Access* 8, pp. 23522–23530. doi: 10.1109/ACCESS.2020.2969854.

Zaki, M.J. and Neely, M. 2019. Predicting consumer behavior in e-commerce: Machine learning techniques and applications. *International Journal of Electronic Commerce* 23(4), pp. 472–499.