Predicting Customer Purchase Intentions in E-commerce

Student Name: Leticia Andrade Vieira

Student Number: 2025304

Date: 29/10/2025

GitHub Link: https://github.com/leticiavieira03/-CA1-Strategic-Thinking-student2025304

Table of content:

Introduction:

In recent years, the rapid growth of e-commerce has transformed how people buy products and services. This shift has created a new type of consumer: more demanding, connected, and seeking personalized experiences. With millions of users visiting online stores every day, companies are increasingly focused on understanding customer behaviour to improve marketing strategies and boost sales performance.

Each online interaction — from pages visited to time spent browsing — generates valuable data that can reveal important behavioural patterns. By analysing this data, businesses can predict the likelihood of a customer completing a purchase and design more effective marketing campaigns.

This project, "Predicting Customer Purchase Intentions in E-commerce Websites," aims to apply data analytics and machine learning techniques to predict the probability of purchase based on user behaviour. The dataset used is from the UCI Machine Learning Repository, 'containing information about user sessions such as page views, bounce rates, and visitor types. The first phase of this Capstone Project focuses on data understanding and exploratory analysis, while the second phase will involve predictive model development and evaluation.

Business understanding

This project aims to find out whether it is possible to predict if a visitor will make a purchase based on their browsing activity.

This topic is important because businesses spend a lot of time and money attracting visitors, but only a small percentage of them end up buying something. Being able to identify which visitors are most likely to purchase can help companies create more effective marketing strategies, improve customer experience, and increase sales.

Objectives:

The main objective of the first phase of this Capstone Project is to explore and understand customer behavior data from e-commerce websites to identify the key factors that influence purchase intentions. This stage focuses on **data exploration and preparation**, forming the foundation for the predictive modelling.

this assignment aims to explore the following objectives:

1- Review relevant literature on customer purchase behavior and predictive analytics in e-commerce

2- Collect, clean, and understand the dataset obtained from the UCI Machine Learning Repository.

3- Conduct exploratory data analysis (EDA) to identify behavioral patterns and relationships between variables that may influence purchase decisions.

4- Summarize initial findings and highlight the main factors affecting customer purchase intentions.

Problem Definition:

This report aims to identify and analyze the key challenges related to understanding customer purchase behavior in e-commerce.
The growing reliance on online shopping has generated vast amounts of user interaction data; however, many organizations fail to utilize this information effectively to understand customer intentions and improve marketing performance. As a result, businesses often struggle to predict which visitors are most likely to complete a purchase, leading to missed opportunities and inefficient use of marketing resources.
Understanding this issue is crucial, as the ability to anticipate customer needs and behaviors provides a significant competitive advantage in the digital marketplace. This project will explore how data-driven analysis can help identify behavioral patterns that influence purchasing decisions.
The challenges of this project include:

- **Data quality and availability:** ensuring the dataset contains sufficient, reliable, and representative information for meaningful analysis.
- **Avoiding bias:** maintaining objectivity in interpreting results and focusing strictly on factual data patterns.
- **Scalability and applicability:** ensuring that the insights obtained can be adapted to different e-commerce contexts.

The context of the problem and the importance of this to be addressed are:

The rapid growth of e-commerce has transformed how consumers interact with online stores. Companies now have access to large volumes of data generated by user interactions, such as page views, session duration, and visitor types. However, despite the availability of this data, many organizations struggle to interpret it effectively and understand customer purchasing behavior.
Accurately identifying which users are likely to make a purchase is critical for improving marketing strategies, enhancing user experience, and increasing conversion rates. By analyzing behavioral data, companies can make **data-driven decisions**, optimize resource allocation, and gain a **competitive advantage** in the digital marketplace.
Addressing this problem is essential, as it allows businesses to **anticipate customer needs**, design targeted marketing campaigns and maximize the efficiency of online sales strategies.

Scope:

Over the first semester, the scope of the project is to explore the following topics and try to answer the following questions:

i. **Customer Behavior Analysis:** Which factors influence whether a visitor completes a purchase? How do variables such as page views, session duration, and visitor type affect purchase intentions?

ii. **Patterns in Online Shopping Data:** What trends or patterns can be identified from the dataset? Are there specific behaviors that indicate a higher likelihood of purchase?

iii. **Data Quality and Reliability:** Are there missing, inconsistent, or biased data points

that need to be addressed before further analysis?

At the end of this first-semester phase, the project aims to provide a **comprehensive understanding of the dataset**, uncover preliminary insights about customer behavior, and establish a solid foundation for further predictive modeling in subsequent stages.

Methodology

The project will follow a structured data analysis process divided into two main phases.

1. Data Collection and Preparation (Semester 1)

The dataset used for this project is the *Online Shoppers Purchasing Intention Dataset* from the UCI Machine Learning Repository. This dataset contains information about user sessions on an e-commerce website, including metrics such as page views, bounce rates, exit rates, time spent on site, and visitor type.

During this phase, the data will be cleaned and pre-processed to remove missing values, inconsistencies, and irrelevant variables. Exploratory Data Analysis (EDA) will be conducted to understand the relationships between features and the target variable ("Revenue"), which indicates whether a purchase occurred.

Visualization techniques such as histograms, correlation heatmaps, and box plots will be used to identify trends and patterns.

Data Sources

The dataset used in this project is the **Online Shoppers Purchasing Intention Dataset**, obtained from the **UCI Machine Learning Repository**. It contains 12,330 records of user sessions collected from a real online retail store over a 12-month period.

The dataset includes features such as the number of pages visited, duration of each session, bounce and exit rates, visitor type (new or returning), operating system, browser, and month of visit. The target variable, *Revenue*, indicates whether a purchase was made during the session (1 for purchase, 0 for no purchase).

The data is publicly available and anonymized, meaning no personally identifiable information (PII) is included. Permission for academic use is granted under the dataset's open license on the UCI repository.

<u>Inclusions of the project:</u>

-
- Definition of the problem and project objectives.
- Collection and understanding of the **Online Shoppers Purchasing Intention Dataset**.
- Cleaning and preparation of the dataset for analysis.
- Exploratory Data Analysis (EDA) to identify patterns and trends in customer behavior.
- Analysis of key variables such as **page views, session duration, bounce rates, and visitor type**.
- Identification of factors that may influence the likelihood of purchase.
- Summarization of initial findings and insights from the dataset.
- Documentation of the methodology and results for presentation and reporting.

In-depth analysis:

	First semester	Second semester
Focus	extensive literature review and data set collection	Analysis of
Aims	Prepare the material to move to the second-semester .	provide ample time for additional review if required, which will ensure the report's quality.

Final consideration on the project scope:

By the end of semester two, aim to deliver a comprehensive academic report that includes the following:

1. **Comprehensive Research:** By the end of the two-semester project, the goal is to deliver a well-researched academic report that integrates **literature review, primary and secondary research, data analysis, and predictive modeling**.
2. **First Semester Deliverables:**
3. Thorough **exploratory data analysis (EDA)** of the UCI Online Shoppers Purchasing Intention Dataset.
4. Data cleaning, preparation, and descriptive statistics.
5. Identification of key patterns and trends in customer behavior, such as page views, session duration, bounce rates, and visitor type.
6. Documentation of methodology and initial findings.
7. **Second Semester Deliverables:**
8. Development and training of **predictive machine learning models** to estimate the likelihood of a visitor completing a purchase.
9. Evaluation of model performance using appropriate metrics to ensure reliability and accuracy.
10. **Actionable, data-driven recommendations** for e-commerce marketing strategies, based on model insights.
11. Comprehensive discussion of findings, limitations, and potential applications in real-world e-commerce contexts.
12. **Overall Goal:** The project aims to provide a **full-cycle, evidence-based analysis**, combining data exploration, predictive modeling, and strategic recommendations to improve understanding of customer purchase behavior and enhance marketing decision-making.

Potential data for the project:

Data Source	Data amount	Permission
www.	Full	open resources allowed by their terms and conditions
Dat	Full	open resources allowed by their terms and conditions

	Full	open resources allowed by their terms and conditions

Table 3: Data sources

The data that has been found as potential for the project are from open resources, so they are allowed to be used according to their terms and conditions.

Ethical considerations:

While this project on predicting customer purchase intentions in e-commerce does not involve the use of sensitive personal data, ethical considerations remain a fundamental part of the research process.
The dataset used — the **Online Shoppers Purchasing Intention Dataset** from the UCI Machine Learning Repository — contains only anonymized and aggregated information, ensuring that no individual user can be identified.
The project will prioritize **transparency, accuracy, and integrity** in the use and interpretation of data. All analyses will be conducted objectively, avoiding manipulation or misrepresentation of results.
All third-party materials, including datasets and academic references, will be appropriately cited following **Harvard Referencing guidelines**, ensuring proper acknowledgment of the contributions of others and the avoidance of plagiarism.
Furthermore, the report will emphasize the importance of **responsible data handling and ethical use of information** for academic and educational purposes, promoting respect for data privacy and responsible analytics practices in the field of data science.

References: