Vivekanand Education Society's

Institute of Technology

Since 1962

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Online Shoppers Purchasing Intention

Submitted in partial fulfilment of the requirements

of the

T.E Project in

Machine Learning

by

Raj Padvekar (39)

Kaustubh Pukale (45)

Pranav Raghuvanshi (46)

under the guidance of

Mrs. Bincy Ivin

Department of Artificial Intelligence and Data Science

Vivekanand Education Society's Institute of Technology

2024-2025

VIVEKANAND EDUCATION SOCIETY
INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Department of Artificial Intelligence and Data Science

CERTIFICATE

This is to certify that **Mr Raj Padvekar, Mr Kaustubh Pukale, Mr Pranav Raghuvanshi** of **T.E D11AD Div B** of Artificial Intelligence and Data Science studying under the University of Mumbai have satisfactorily presented the Project entitled **Online Shoppers Purchasing Intention** as a part of the T.E Mini Project for Semester-VI of Machine Learning Lab under the guidance of **Mrs Bincy Ivin** in the year 2024-2025.

Date: 07/04/2025

(Name and sign)
Head of Department

(Name and sign)
Supervisor/Guide

VESIT V.E.S. Since 1962

VIVEKANAND
EDUCATION SOCIETY
INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Department of Artificial Intelligence and Data Science

DECLARATION

We, ***Raj Padvekar, Kaustubh Pukale and Pranav Raghuvanshi*** from ***D11AD-B***, declare that this project represents our ideas in our own words without plagiarism and wherever others' ideas or words have been included, we have adequately cited and referenced the original sources.

We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our project work.

We declare that we have maintained a minimum 75% attendance, as per the University of Mumbai norms.

Yours Faithfully

1. Raj Padvekar

2. Kaustubh Pukale

3. Pranav Raghuvanshi

(Name & Signature of Students with Date)

VESIT

V.E.S.

Since 1962

VIVEKANAND EDUCATION SOCIETY

INSTITUTE OF TECHNOLOGY
(AUTONOMOUS)

(Affiliated to University of Mumbai, Approved by AICTE & Recognized by Govt. of Maharashtra)

Acknowledgement

I would like to express my sincere gratitude to all the individuals who guided and supported me throughout the development of this project on **Online Shoppers Purchasing Intention**.

My teacher, **Mrs Bincy Ivin**, for providing me with invaluable insights, technical guidance, and encouragement at every stage..

To my parents, whose constant motivation, patience, and belief in me have been a source of strength throughout this journey.

I am grateful to my friends and peers who contributed ideas and feedback that helped improve the quality and scope of this project.

Thank you everyone for playing a part in shaping this project and enhancing my learning experience in the field of data science and machine learning.

Table of Contents

Abstract

In today’s fast-growing world of e-commerce, understanding how users behave online and being able to predict whether they'll make a purchase has become more important than ever. This project explores the use of machine learning to predict the purchasing intention of online shoppers, helping businesses make smarter, data-driven decisions.

The study uses the **Online Shoppers Purchasing Intention Dataset** from the **UCI Machine Learning Repository**, which includes detailed information about user sessions—such as the number of pages visited, product views, bounce and exit rates, and even technical aspects like browser and operating system. The main goal is to determine whether a user ended up making a purchase during their session.

To build the prediction model, a full machine learning pipeline was followed, including data cleaning, exploratory analysis, feature engineering, and training various classification algorithms. Care was taken to address class imbalance and select the most meaningful features to improve accuracy.

After evaluating different models, the **Extreme Gradient Boosting Classifier** delivered the best performance. To make the model easy to use, a web-based interface was built using **Streamlit**, where users can input session details and instantly get predictions on purchase intent.

This project not only shows how machine learning can be applied to real-world problems in e-commerce but also sheds light on patterns in user behavior that could help businesses improve their services and marketing strategies.

1. Introduction

1.1 Overview of the Project

In today's digital world, online shopping has become the norm, and businesses are always searching for smarter ways to understand their customers and improve sales. This project focuses on using machine learning to predict whether a user visiting an e-commerce site is likely to make a purchase. By analyzing user behavior during their browsing session, we can help businesses personalize experiences, target the right customers, and ultimately boost their conversion rates.

We utilized the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository to conduct our analysis. This dataset captures user behavior data from over 12,000 online shopping sessions, making it highly valuable for understanding what factors influence purchasing decisions.

The dataset provides comprehensive details such as the number of pages visited during a session and the amount of time spent on various types of pages, including product and informational pages. It also includes important engagement metrics like bounce rate and exit rate, which help indicate the quality and outcome of each visit.

In addition, the dataset records whether the session occurred on a special date such as a holiday or promotional event, and includes data on how the user arrived at the site—whether directly, through a referral, or via search engines. It also distinguishes between new and returning visitors, offering insights into customer loyalty and behavior patterns.

Furthermore, technical attributes such as the browser used, operating system, and the region from which the user accessed the site are included. These features provide additional context that can be used to tailor the shopping experience or identify trends among different user segments.

The project followed a complete machine learning workflow to ensure accurate and user-friendly results.

We started with data preprocessing, where we cleaned the dataset, handled missing values, and converted categorical variables into numerical formats suitable for model training.

During the exploratory data analysis (EDA) phase, we explored the data to identify trends and understand which features had the greatest impact on purchasing decisions.

In feature engineering, we enhanced the dataset by transforming existing features and creating new ones to improve the model's predictive performance.

For model training and evaluation, we tested several machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and Extreme Gradient Boosting. Among these, Extreme Gradient Boosting provided the highest accuracy and reliability.

To make the model accessible, we built an interactive web app using Streamlit, where users can enter session data and get real-time predictions on whether a purchase is likely. Additionally, we integrated a chatbot into the app to assist users by answering queries related to the model and its predictions, enhancing the overall user experience.

This project not only highlights the practical use of machine learning in e-commerce but also provides a hands-on tool that could help businesses make informed decisions based on user behavior. It's a great example of how data can be turned into actionable insights in a real-world setting.

1.2 Scope of the Project

This project focuses on understanding user behavior by analyzing how visitors interact with an e-commerce website. Key behavioral indicators such as the number of pages visited, time spent on different page types, bounce and exit rates were explored to identify patterns that contribute to successful purchases. These behavioral signals serve as the foundation for predicting user intent.

The project utilized the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository, which contains real-world browsing data from over 12,000 sessions. This dataset enabled meaningful analysis by providing insights into actual user interactions, lending authenticity and practical relevance to the findings.

A significant part of the work involved building and comparing multiple machine learning models to predict purchasing intent. Techniques such as thoughtful feature engineering, preprocessing, and model evaluation were employed to boost predictive performance. Models like Logistic Regression, Random Forest, and XGBoost were tested, with a focus on improving reliability and accuracy.

To enhance usability, a Streamlit web application was developed, allowing users to input session details and receive immediate predictions about purchase likelihood. This makes the tool accessible and practical, even for non-technical users, thereby bridging the gap between data science and business application.

The project also highlights its business value by offering actionable insights for online retailers. It can help businesses better understand customer behavior, personalize user experiences, and increase conversion rates, thereby contributing to improved online sales strategies.

However, the current project scope is limited to the structured data available in the dataset. It does not account for complex emotional or social influences such as user mood, peer effects, or dynamic factors like ongoing promotions. Future enhancements could involve integrating real-time data streams, incorporating personalized recommendation systems, or applying advanced segmentation techniques to provide a more comprehensive view of customer intent and behavior.

2. Objective and Problem Statement

2.1 Clear Definition of this Problem Statement

As online shopping continues to grow, one of the biggest challenges for e-commerce businesses is figuring out which visitors are likely to make a purchase and which ones are just browsing. While traditional analytics tools offer some basic insights, they often miss the deeper behavioral patterns that signal buying intent.

This project aims to solve that problem by building a machine learning model that can predict whether a user will complete a purchase based on their browsing behavior. By looking at things like how many pages they visit, how much time they spend on product pages, how quickly they leave, where they came from, and what device they're using, the model can make an informed prediction.

With this approach, online retailers can better understand their customers, personalize the shopping experience, and make smarter business decisions that lead to higher conversion rates.

2.2 Objective of the Project

The main goal of this project is to create a system that can predict whether someone visiting an online shopping website will end up making a purchase. With online shopping becoming more popular every day, it's important for businesses to understand what makes a visitor actually buy something.

To do this, the project uses a real dataset from the UCI Machine Learning Repository that contains information about how users behave during their browsing sessions. This includes details like how many pages they visit, how much time they spend on the site, where they came from, and what device they are using.

The project involves cleaning and analyzing this data, trying out different machine learning models, and choosing the one that works best for making accurate predictions. Finally, the model is turned into a simple web app using

Streamlit, so that anyone can use it to check the chances of a user making a purchase based on their browsing behavior.

This helps businesses better understand their customers and improve how they target and serve them.

3. Dataset Description & Collection

The dataset used in this project is the **Online Shoppers Purchasing Intention Dataset**, publicly available on the UCI Machine Learning Repository. It was collected by gathering real-world browsing behavior data from an e-commerce website during a specific period in 2010.

3.1 Purpose of the Dataset

The dataset was designed to model and predict whether a visitor will make a purchase during a browsing session, based on various behavioral, transactional, and technical parameters.

3.2 Data Collection

The data was collected over a 1-year period in 2010 from a genuine e-commerce platform.Sessions were recorded anonymously while preserving important behavioral and interactional information.Data was filtered to remove bot activities, incomplete sessions, and outliers.

3.3 Dataset Statistics

The dataset used in this project consists of a total of 12,330 instances (rows), each representing a unique user session on an e-commerce website. It includes 18 features that describe various aspects of user behavior and session characteristics, excluding the target variable. The target variable, named Revenue, is a Boolean value indicating whether a purchase was made during the session—True if the user completed a purchase, and False otherwise.

3.4 Feature Description

The dataset includes 10 numerical, 8 categorical, and 1 boolean (target) feature. A brief overview of some key features is as follows:

Feature	Type	Description
Administrative	Numeric	Number of administrative pages visited
Administrative_Duration	Numeric	Time spent on administrative pages
Informational	Numeric	Number of informational pages visited
Informational_Duration	Numeric	Time spent on informational pages
ProductRelated	Numeric	Number of product-related pages visited
ProductRelated_Duration	Numeric	Time spent on product-related pages
BounceRates	Numeric	Percentage of visitors who left after the first page
ExitRates	Numeric	Percentage of exits from that page
PageValues	Numeric	Economic value of the page
SpecialDay	Numeric	Closeness of visit to a special day (e.g. Valentine’s Day)
Month	Categorical	Month of the session (e.g., Feb, Mar, etc.)
OperatingSystems	Categorical	OS used during the session
Browser	Categorical	Browser used
Region	Categorical	Geographic region of the visitor
TrafficType	Categorical	Type of traffic (e.g., referral, direct, etc.)

VisitorType	Categorical	Type of visitor (Returning, New, Other)
Weekend	Boolean	TRUE if the session was on a weekend
Revenue	Boolean	Target label: TRUE if purchase occurred

Table 3.1 Feature Description Table

3.5 Class Distribution

The dataset is highly imbalanced, with around 84.5% of user sessions not resulting in a purchase and only 15.5% ending with a successful transaction (Revenue = True). This imbalance posed challenges for model performance, making it necessary to apply techniques like SMOTE or class weighting to ensure balanced learning.

Before model training, we performed essential steps including data cleaning, label encoding for categorical features, normalization of continuous variables, and feature selection and engineering to enhance model effectiveness.

This processed dataset was then used as the core input for training and evaluating multiple machine learning models, including Logistic Regression, Random Forest, and ensemble techniques, to predict purchasing intentions accurately.

4. Literature Survey

4.1 Literature/Techniques studied

In recent years, modeling online customer purchase intention has gained significant attention due to the exponential growth of e-commerce. Various studies have proposed different techniques to enhance model accuracy and manage data complexities associated with user behavior. One common approach is data transformation using Z-score normalization, which standardizes features to have a mean of 0 and a standard deviation of 1, making it effective in mitigating the influence of outliers. Furthermore, cyclic transformation using sine and cosine functions is applied to features such as day of the week and month to preserve their cyclical properties and improve model interpretability [1].

To handle class imbalance, which is prevalent in many e-commerce datasets, the Synthetic Minority Over-sampling Technique (SMOTE) is widely adopted. SMOTE generates synthetic samples of the minority class to create a balanced dataset, thereby helping classifiers to better identify purchasing intentions [2][3]. Additionally, outlier detection plays a crucial role in preprocessing, and techniques such as the Interquartile Range (IQR) method are used to identify anomalies in the dataset by analyzing the distribution between the first and third quartiles [4].

A variety of machine learning models have been implemented to predict online shopping behavior. Random Forest, an ensemble learning method based on decision trees, is favored for its robustness and high accuracy. Gradient Boosting and its optimized version XGBoost are also commonly employed due to their ability to learn from residual errors iteratively and deliver superior predictive performance. Logistic Regression is frequently utilized for binary classification tasks due to its simplicity and interpretability, while AdaBoost focuses on improving model performance by giving more weight to previously misclassified instances [2][3]. These classifiers, when combined with advanced feature engineering techniques, contribute significantly to the prediction of online purchase behavior.

Further research has highlighted the importance of contextual and behavioral features, especially in post-pandemic scenarios. Studies have shown that user satisfaction and trust significantly influence purchase intention, particularly in developing countries where digital adaptation surged during the COVID-19 pandemic [7]. Moreover, the integration of user behavior modeling with temporal and categorical data has proven effective in improving real-time prediction systems [6].

Overall, the adoption of ensemble learning, effective data transformation techniques, and balanced dataset preparation has shown to be crucial for enhancing the predictive accuracy of models that estimate online shoppers' purchasing intentions.

4.2 Papers/Findings

Wang, Runan discusses the influence of consumer behavior on online shopping desire, examining the cultural, social, personal, and psychological factors that play a significant role in shaping a consumer's inclination to make purchases online. The paper further delves into the impact of consumer acceptance of the online shopping medium itself on their ultimate purchase intention. Additionally, it introduces and elaborates on common prediction methods utilized in the field, such as collaborative filtering and hybrid recommendation systems, highlighting their importance in understanding and forecasting online shopper behavior [1].

A. Karakaya, İ. Karakaya, and T. Temizceri focused their research on developing a model for predicting online shoppers' purchasing intention by employing the techniques of ensemble learning. Their study revealed that addressing the common issue of imbalanced datasets through the application of the Synthetic Minority Over-sampling Technique (SMOTE) led to a notable improvement in the performance of individual classification models. Notably, the Bagging ensemble method, when utilizing PART as its base classifier, achieved a high accuracy of 92.62% in predicting whether an online shopper would make a purchase, underscoring the effectiveness of combining multiple learning algorithms for this task [2].

Satu, M.S., and Islam, S.F. investigated the process of modeling online customer purchase intention behavior by applying a variety of feature engineering and classification techniques. Their findings indicated that the Random Forest classifier, when applied to a dataset that had undergone Z-Score transformation and in conjunction with the Gain Ratio Attribute Evaluation feature selection method, yielded the most accurate predictions, reaching 92.39%. Furthermore, the Random Forest model also demonstrated the highest Area Under the Receiver Operating Characteristic curve (AUROC) of 0.974 when applied to a Square Root transformed dataset using the same feature selection method. Based on these results, the research concluded that Random Forest is a robust and highly effective machine learning classifier for accurately predicting the likelihood of online customer purchase [3].

Frazier, Andrew & Maloku, Fatbardha & Li, Xinzi & Chen, Yichun & Jung, Yeji & Zohuri, Bahman conducted a comprehensive data analysis of online shopper's purchasing intention, leveraging machine learning for prediction analytics. Their research highlighted the efficacy of an ensemble model that strategically combined the strengths of both Random Forests and XGBoost algorithms, achieving an impressive overall accuracy of 89.69% in predicting purchase intention. The exploratory data analysis conducted as part of the study shed light on significant correlations between a user's website behavior and their likelihood of making a purchase. Specifically, they found that lower exit rates and bounce rates were strong indicators of a higher propensity to purchase, and the 'page values' feature exhibited the most significant positive correlation with revenue generation [4].

Baati, K., and Mohsil, M. directed their research towards the real-time prediction of online shoppers' purchasing intention through the application of the Random Forest algorithm. Their study effectively demonstrated the advantages of addressing the inherent issue of data imbalance, as the Random Forest classifier achieved the highest accuracy of 86.78% after the implementation of the SMOTE technique to balance the training data. Moreover, the Random Forest model also exhibited the highest sensitivity (0.62) and F1 Score (0.60), further validating its suitability and effectiveness for the timely and accurate prediction of online shoppers' purchasing intention [5].

G. Sang and S. Wu explored the complex task of predicting online shoppers' purchasing intention by comparing the performance of several different classification algorithms. Their findings clearly indicated that the Multilayer Perceptron (MLP) classifier significantly outperformed both Random Forest and Support Vector Machines (SVM) in terms of predictive accuracy and F1 Score. The study also emphasized the value of incorporating detailed clickstream data along with features derived from session information, which led to a further improvement in the accuracy of predicting purchasing intention. Additionally, their research revealed that an LSTM-based recurrent neural network was particularly effective in predicting the probability of a user abandoning a website session without making a purchase [6].

García-Salirrosas, E.E.; Acevedo-Duque, Á.; Marin Chaves, V.; Mejía Henao, P.A.; and Olaya Molano, J.C. investigated the crucial factors influencing purchase intention and satisfaction among online shop users in developing countries during the unprecedented circumstances of the COVID-19 pandemic. Their research uncovered that trust and satisfaction play a direct and positive role in shaping both the perceived value of online shopping and the subsequent online purchase intention within these developing economies. Furthermore, the study established that perceived value also directly and positively impacts the online purchase intention of consumers engaging with small businesses [7].

Sakar, C.O., Polat, S.O., Katircioglu, M., et al. conducted a study focused on the real-time prediction of online shoppers' purchasing intention, employing both Multilayer Perceptron and LSTM recurrent neural networks. Consistent with the findings reported by Sang and Wu [6], this research also demonstrated that the Multilayer Perceptron (MLP) achieved higher accuracy and F1 Score in predicting purchase intention when compared to Random Forest and SVM. Moreover, their work corroborated the effectiveness of LSTM-based recurrent neural networks in accurately predicting website abandonment based on a thorough analysis of clickstream data [8].

Mootha, S., Sridhar, S., and M.S.K. Devi proposed an advanced approach that utilizes a stacking ensemble of Multi-Layer Perceptrons to predict online shoppers' purchasing intention. Their innovative model, which strategically combines the predictive power of multiple MLP classifiers, achieved a high level of accuracy, reaching 94%, in predicting whether online shoppers would make a purchase based on the details of their browsing sessions. The results of their study demonstrated that this sophisticated stacking ensemble outperformed other individual classification algorithms and existing systems in terms of its predictive capabilities [9].

Mokryn, O., Bogina, V., and Kuflik, T. explored the challenging problem of inferring the current purchase intent of website visitors who remain anonymous. Their study suggests the potential of leveraging product popularity trends and the temporal information associated with a website visit as key indicators for inferring a user's likelihood to make a purchase. The model developed in their research successfully identifies online signals that can be effectively used for the prediction of purchase intent, even in the absence of prior interaction history or identifiable user information [10].

5. Proposed Solution

This section outlines the methodology adopted for building a predictive model to determine whether an online shopper is likely to complete a purchase. The solution comprises three main components: Exploratory Data Analysis (EDA), model selection and training, and the design of key features and functionalities integrated into the system.

5.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand the underlying patterns, distribution, and relationships within the dataset. This step was crucial for formulating hypotheses, selecting relevant features, and identifying potential data quality issues.

The dataset used in this project was sourced from the UCI Machine Learning Repository and comprises a total of 12,330 user sessions. It includes 18 input features and one target variable, providing a comprehensive view of user interactions on an e-commerce platform. These features are a mix of numerical and categorical data, capturing various dimensions of user behavior and technical session metadata.

Key behavioral features include Administrative, Informational, and ProductRelated, which represent the number of pages visited in each category. Corresponding duration-based metrics like Administrative_Duration, Informational_Duration, and ProductRelated_Duration reflect the time spent on those specific page types.

Session quality is further quantified using features like BounceRates, ExitRates, and PageValues, which help in evaluating the level of user engagement. The SpecialDay feature indicates how close a session was to a special occasion or holiday, potentially influencing purchasing behavior.

Technical and temporal attributes such as Month, OperatingSystems, Browser, Region, and TrafficType provide context about when and how the session occurred. Meanwhile, VisitorType and Weekend serve as categorical indicators of user intent and behavior.

The target variable, Revenue, is a Boolean value indicating whether the session resulted in a purchase. This combination of behavioral, technical, and temporal data allows for rich analysis and effective prediction of purchasing intentions.

5.1.1 Key observations from EDA include:

The target variable (Revenue) in the dataset is highly imbalanced, with only a small proportion of user sessions resulting in a purchase. This posed a challenge during modeling, requiring specialized techniques to ensure accurate predictions.

Users who spent more time on ProductRelated pages and had higher values in the PageValues feature were significantly more likely to make a purchase. This suggests a strong correlation between detailed product exploration and buying intent. Additionally, returning visitors showed a higher tendency to convert compared to new visitors, indicating that familiarity with the website plays a role in purchase behavior.

An interesting trend was observed in the month of November, which saw increased purchasing activity. This aligns with typical e-commerce patterns, such as seasonal sales and promotions during holidays. Visual tools like histograms, correlation heatmaps, and boxplots were used to explore these patterns, guiding both feature engineering and model selection.

Many features—such as Administrative_Duration, Informational_Duration, ProductRelated_Duration, and PageValues—exhibited right-skewness, meaning that while most users spent minimal time on these pages, a smaller group spent significantly more time. Furthermore, features like Informational_Duration, SpecialDay, and PageValues were zero-inflated, indicating that many users had no interaction with these aspects during their sessions.

Notably, users who eventually made a purchase tended to engage slightly more with administrative pages, such as account or checkout sections. However, product-related activity was the most decisive behavioral signal—buyers viewed more product pages and spent longer on them, making it a strong indicator of purchase intent.

The PageValues feature emerged as the most powerful predictor, showing a clear distinction between buyers and non-buyers. Higher values were closely associated with purchasing behavior. Additionally, buyers interacted slightly more with informational content, suggesting that being well-informed may lead to higher conversion rates.

Lastly, purchases were marginally more common around special days, such as holidays or promotional events. While the impact was subtle, it points to the role of timing and seasonal trends in influencing customer decisions.

5.2 Model Selection and Training

Based on the EDA findings, the data was preprocessed and used to train several machine learning models. The goal was to evaluate and compare different algorithms to identify the most accurate and reliable model for predicting purchase intention.

5.2.1 Data Preprocessing

The preprocessing phase began with handling the VisitorType column, which contained categories such as "New_Visitor", "Returning_Visitor", and "Other". One-hot encoding was applied to convert these categories into separate binary columns, making them compatible with machine learning algorithms. The Weekend column, originally a Boolean feature, was mapped to numerical values—False to 0 and True to 1—to facilitate better model interpretation.

For the Month column, each month name was mapped to its respective numerical value (e.g., January to 1, February to 2, etc.). To account for the cyclical nature of months, sine and cosine transformations were applied. This ensured that months like December and January, which are temporally close, were treated as such by the model.

Several features related to user behavior and session engagement—including Administrative,Administrative_Duration, Informational, Informational_Duration, ProductRelated,ProductRelated_Duration,BounceRates,ExitRates,and PageValues—were found to be heavily skewed. To mitigate the influence of outliers and extreme values, a log1p transformation was applied to these features. However, Informational and Informational_Duration were dropped later due to redundancy and limited contribution to the predictive power of the model.

Outliers in key continuous features like ProductRelated_Duration, BounceRates, and ExitRates were managed using the Interquartile Range (IQR) method. Features such as PageValues and TrafficType were excluded from outlier treatment due to their statistical stability and importance in prediction.

Categorical features like SpecialDay, OperatingSystems, and Browser were grouped into broader categories to reduce sparsity. These, along with the Region column, were then one-hot encoded to convert them into numerical format. Post-encoding, all resulting Boolean features (e.g., SpecialDay_1, OperatingSystems_2) were converted to integers to maintain consistency in data types across the dataset.

The final preprocessed dataset consisted of a comprehensive set of engineered features, including numerical variables, one-hot encoded categorical columns, cyclic month indicators, and technical session identifiers. The target variable remained Revenue, indicating whether a purchase occurred.

For model training, a 70-30 train-test split was used to ensure an unbiased evaluation. ColumnTransformer was employed to apply StandardScaler exclusively to numerical features, ensuring that all inputs were standardized to a common scale. To maintain transformation consistency in deployment, the preprocessor was saved using joblib.

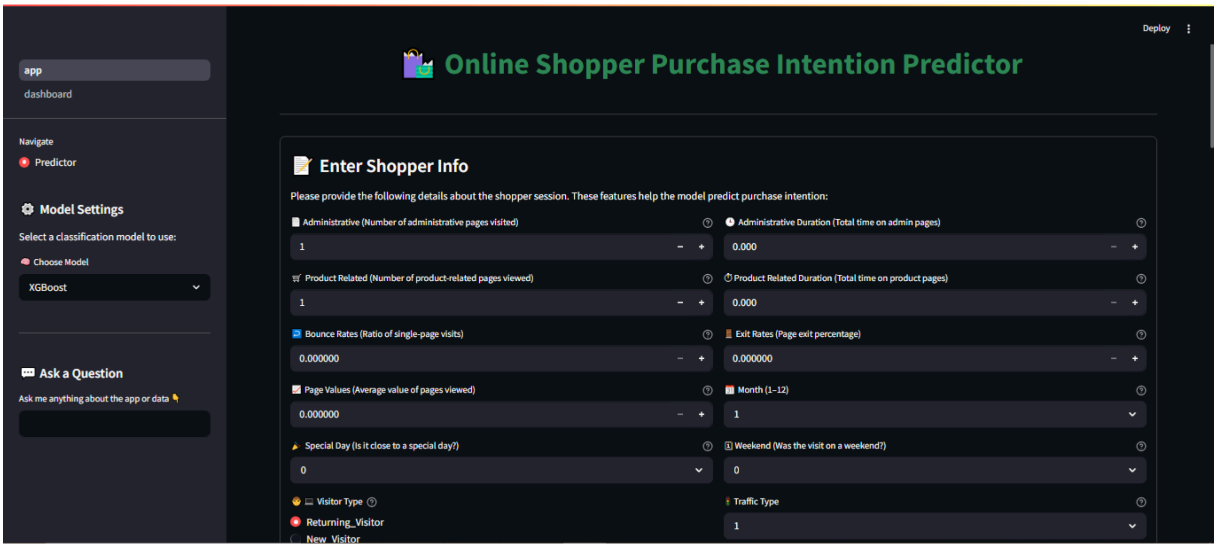
To address the issue of class imbalance, the SMOTE (Synthetic Minority Oversampling Technique) was applied to the training set. This technique generated synthetic examples of the minority class, helping the model learn from a balanced distribution and preventing bias toward the majority class.

5.3 Key Features and Functionalities`

This project brings together two powerful tools into one easy-to-use web application. First, it features a prediction system that can tell whether a visitor is likely to make a purchase based on how they browse the site. By simply entering details like time spent on different pages, bounce rates, and session behavior, the app uses a trained machine learning model to quickly predict the chances of a purchase.

But it doesn't stop there. The app also comes with an interactive analytical dashboard that helps make sense of all that user data. It includes clear and engaging visuals that show how users interact with the website. You can explore things like which pages people spend the most time on, when they're more likely to buy, and how different behaviors influence the outcome.

To further enhance user interaction and support, the application also features an AI-powered chatbot that responds to queries related to the model and its usage. Integrated with the Gemini API, the chatbot provides intelligent, context-aware answers, making the platform more accessible for both technical and non-technical users. Whether you're aiming to improve campaign effectiveness, streamline the user journey, or understand customer behavior in depth, this application offers the tools and insights needed to make informed, impactful decisions.



● Predictor

⚙ Model Settings

Select a classification model to use:

Choose Model

XGBoost

Ask a Question

Ask me anything about the app or data 👇

Online Shopper Purchase Intention Predictor

Enter Shopper Info

Please provide the following details about the shopper session. These features help the model predict purchase intention:

Administrative (Number of administrative pages visited): 1

Administrative Duration (Total time on admin pages): 0.000

Product Related (Number of product-related pages viewed): 1

Product Related Duration (Total time on product pages): 0.000

Bounce Rates (Ratio of single-page visits): 0.000000

Exit Rates (Page exit percentage): 0.000000

Page Values (Average value of pages viewed): 0.000000

Month (1–12): 1

Special Day (Is it close to a special day?): 0

Weekend (Was the visit on a weekend?): 0

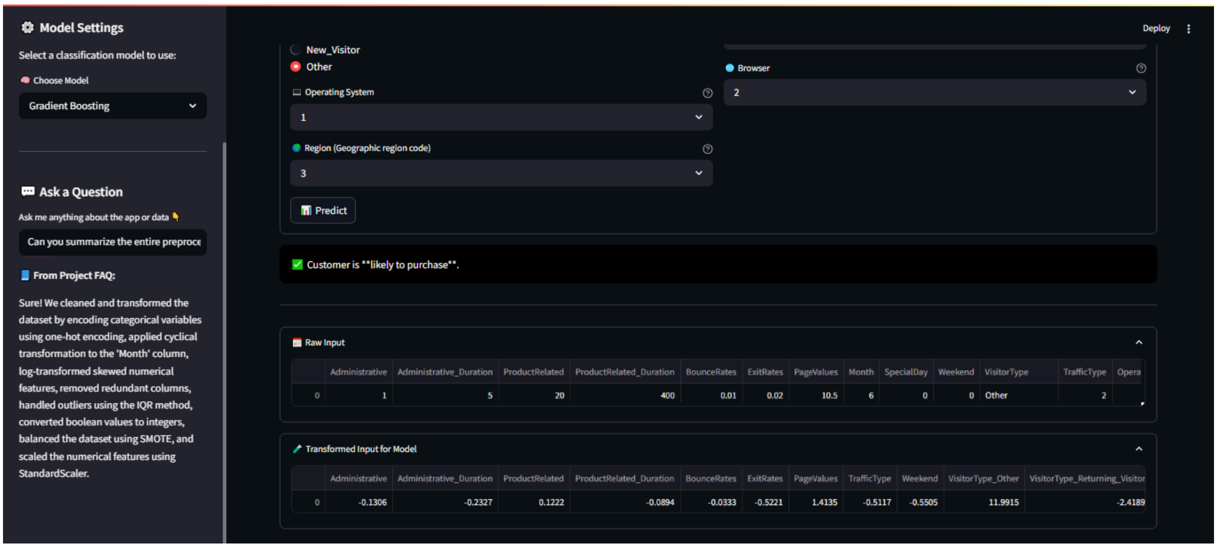
Visitor Type

● Returning_Visitor

○ New_Visitor

Traffic Type: 1

Fig 5.1 App Homepage



⚙ Model Settings

Select a classification model to use:

Choose Model

Gradient Boosting

Ask a Question

Ask me anything about the app or data 👇

Can you summarize the entire preproce

From Project FAQ:

Sure! We cleaned and transformed the dataset by encoding categorical variables using one-hot encoding, applied cyclical transformation to the 'Month' column, log-transformed skewed numerical features, removed redundant columns, handled outliers using the IQR method, converted boolean values to integers, balanced the dataset using SMOTE, and scaled the numerical features using StandardScaler.

○ New_Visitor

● Other

Browser: 2

Operating System: 1

Region (Geographic region code): 3

✅ Customer is **likely to purchase**.

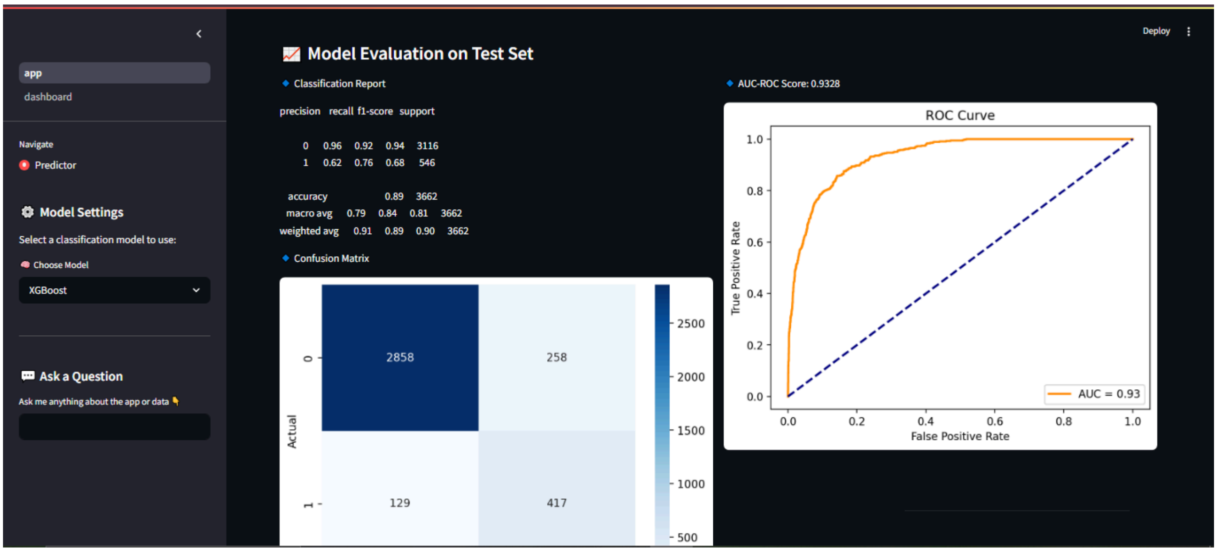
Raw Input

	Administrative	Administrative_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	Month	SpecialDay	Weekend	VisitorType	TrafficType	Opera
0	1	5	20	400	0.01	0.02	10.5	6	0	0	Other	2	

Transformed Input for Model

	Administrative	Administrative_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	TrafficType	Weekend	VisitorType_Other	VisitorType_Returning_Visitor
0	-0.1306	-0.2327	0.1222	-0.0894	-0.0333	-0.5221	1.4135	-0.5117	-0.5505	11.9915	-2.4189

Fig 5.2 Model’s Prediction



```
app
dashboard

Navigate
Predictor

Model Settings
Select a classification model to use:
Choose Model
XGBoost

Ask a Question
Ask me anything about the app or data
```

Model Evaluation on Test Set

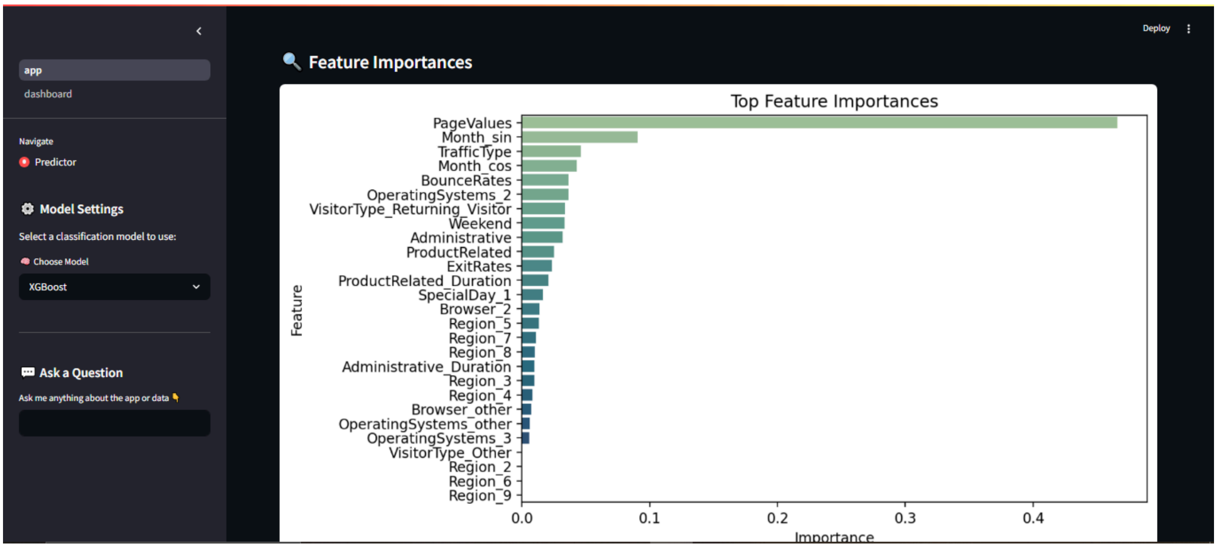
Classification Report

```
precision   recall  f1-score  support

   0    0.96    0.92    0.94    3116
   1    0.62    0.76    0.68     546

  accuracy              0.89    3662
 macro avg    0.79    0.84    0.81    3662
weighted avg    0.91    0.89    0.90    3662
```

Confusion Matrix

AUC-ROC Score: 0.9328

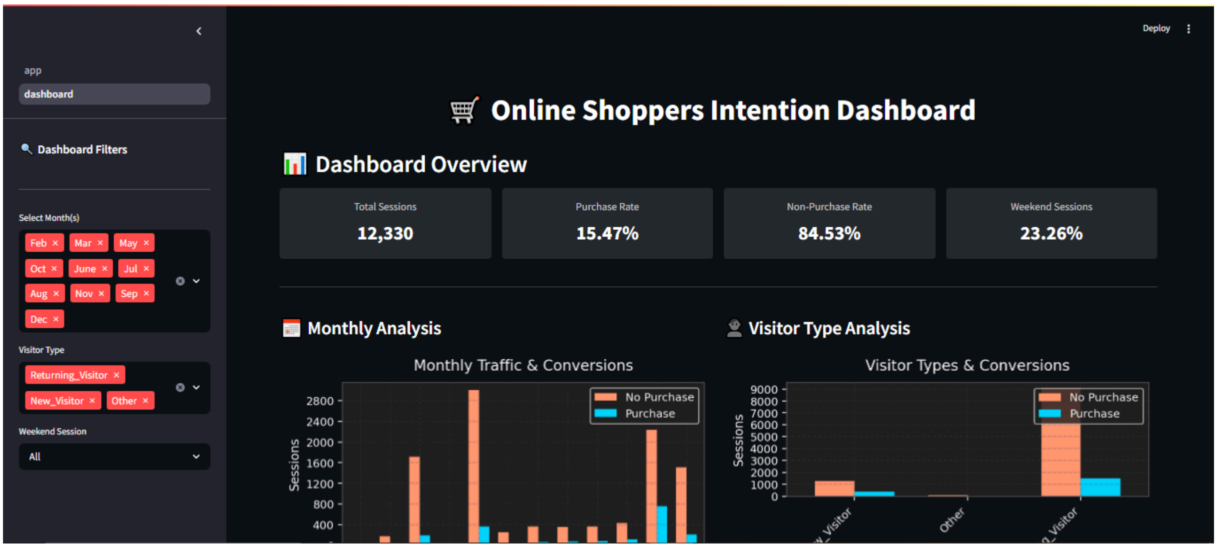
Fig 5.3 Model Evaluation



```
app
dashboard

Navigate
Predictor

Model Settings
Select a classification model to use:
Choose Model
XGBoost

Ask a Question
Ask me anything about the app or data
```

Feature Importances

Fig 5.4 Model’s Feature Importance



🔍 Dashboard Filters

Select Month(s)

Feb × Mar × May × Oct × June × Jul × Aug × Nov × Sep × Dec ×

Visitor Type

Returning_Visitor × New_Visitor × Other ×

Weekend Session

All

🛒 Online Shoppers Intention Dashboard

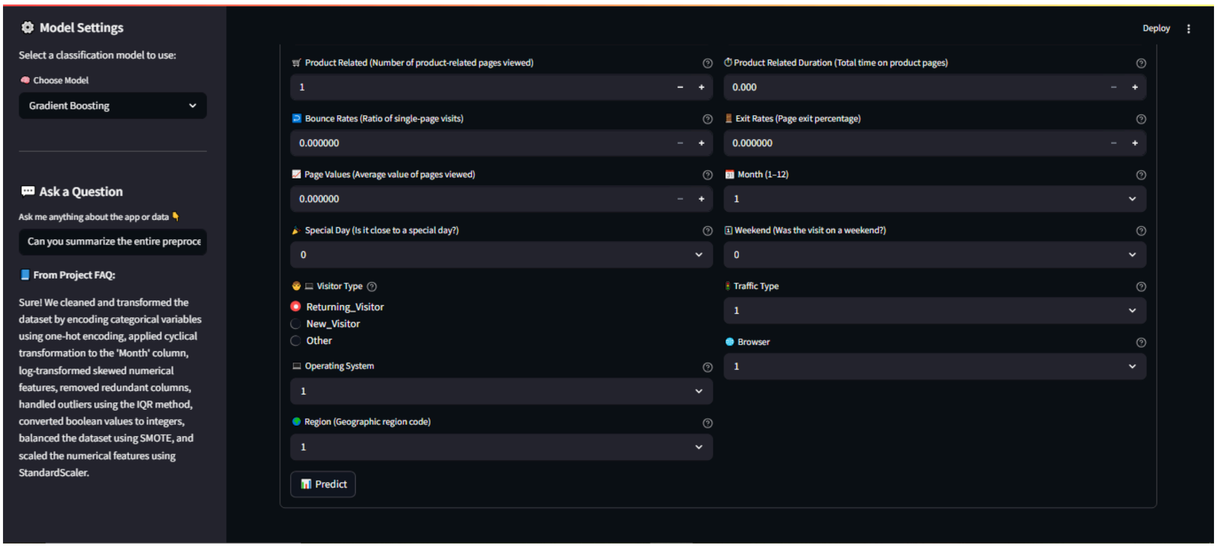
📊 Dashboard Overview

Total Sessions	Purchase Rate	Non-Purchase Rate	Weekend Sessions
12,330	15.47%	84.53%	23.26%

Monthly Analysis

Visitor Type Analysis

Fig 5.5 Analytical Dashboard



⚙ Model Settings

Select a classification model to use:

Choose Model

Gradient Boosting

💬 Ask a Question

Ask me anything about the app or data 👇

Can you summarize the entire preproce

📘 From Project FAQ:

Sure! We cleaned and transformed the dataset by encoding categorical variables using one-hot encoding, applied cyclical transformation to the 'Month' column, log-transformed skewed numerical features, removed redundant columns, handled outliers using the IQR method, converted boolean values to integers, balanced the dataset using SMOTE, and scaled the numerical features using StandardScaler.

🛒 Product Related (Number of product-related pages viewed): 1

⏱ Product Related Duration (Total time on product pages): 0.000

Bounce Rates (Ratio of single-page visits): 0.000000

Exit Rates (Page exit percentage): 0.000000

Page Values (Average value of pages viewed): 0.000000

Month (1–12): 1

Special Day (Is it close to a special day?): 0

Weekend (Was the visit on a weekend?): 0

Visitor Type

● Returning_Visitor
○ New_Visitor
○ Other

Traffic Type: 1

Browser: 1

Operating System: 1

Region (Geographic region code): 1

Fig 5.6 Chat Bot Support

6. Design & Development Approach

6.1 System architecture

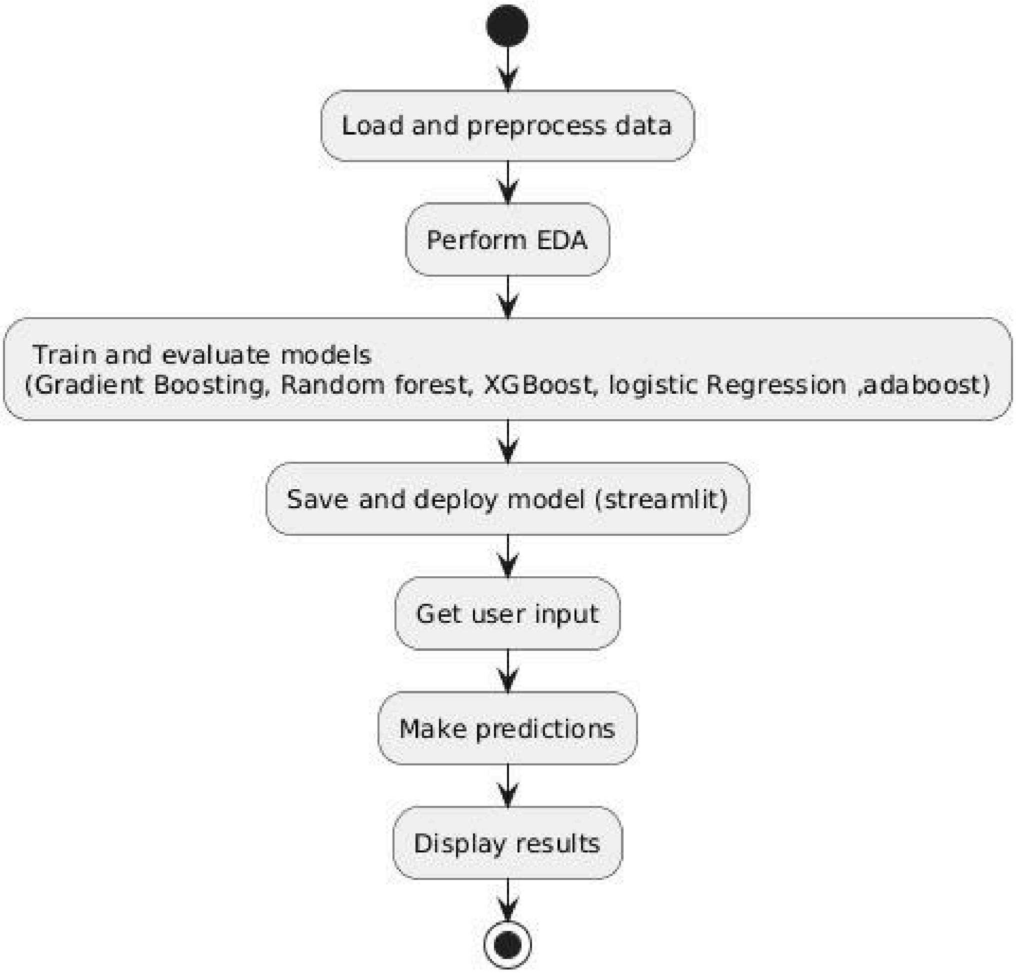

Load and preprocess data
Perform EDA
Train and evaluate models
(Gradient Boosting, Random forest, XGBoost, logistic Regression ,adaboost)
Save and deploy model (streamlit)
Get user input
Make predictions
Display results

Fig 6.1: System Architecture

6.2 Technologies and tools used

The Online Shoppers Purchasing Intention system leverages a range of modern technologies and tools to ensure efficient data processing, model training, and user interaction. Python serves as the core programming language, supporting various stages of the machine learning pipeline. For data preprocessing and analysis, libraries such as **Pandas** and **NumPy** are employed, while **Matplotlib** and **Seaborn** are used for Exploratory Data Analysis (EDA) and visualization. To build and evaluate predictive models, machine learning frameworks like **Scikit-learn**, **XGBoost**, and **Random Forest** classifiers are utilized. The trained models are then deployed using the **streamlit** web framework, enabling real-time interaction through a web-based interface. Data from the user is collected via the interface, and the backend logic makes predictions using the deployed models. Additionally, tools such as **Jupyter Notebook** aid in iterative development and testing, and **Git** is used for version control throughout the development process.

7 Results and Discussion:

Train Set:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.8820	0.8866	0.8761	0.8813
Adaboost	0.8892	0.8998	0.8758	0.8876
Gradient Boosting	0.9260	0.9249	0.9273	0.9261
XGBoost	0.9218	0.9226	0.9209	0.9217
Logistic Regression	0.8580	0.8907	0.8162	0.8518

Table 7.1 Train Set Classification Report

Test Set

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.8695	0.5420	0.8040	0.6475
Adaboost	0.8831	0.5815	0.7711	0.6630
Gradient Boosting	0.8930	0.6149	0.7546	0.6776
XGBoost	0.8943	0.6178	0.7637	0.6830
Logistic Regression	0.8779	0.5656	0.7821	0.6564

Table 7.2 Test Set Classification Report

Stratified K Fold Scores

Model	Fold Accuracies	Mean Accuracy	Std. Dev
Random Forest	[0.8799, 0.8869, 0.8816, 0.8767, 0.8751]	0.8800	0.0041
Adaboost	[0.8845, 0.8918, 0.8951, 0.8861, 0.8841]	0.8883	0.0044
Gradient Boosting	[0.9037, 0.9045, 0.9025, 0.8984, 0.8992]	0.9017	0.0024
XGBoost	[0.9041, 0.9045, 0.9070, 0.8988, 0.9050]	0.9039	0.0027
Logistic Regression	[0.8976, 0.8869, 0.8914, 0.8902, 0.8837]	0.8900	0.0047

Table 7.3 Stratified K Fold Score

AUC ROC Curve

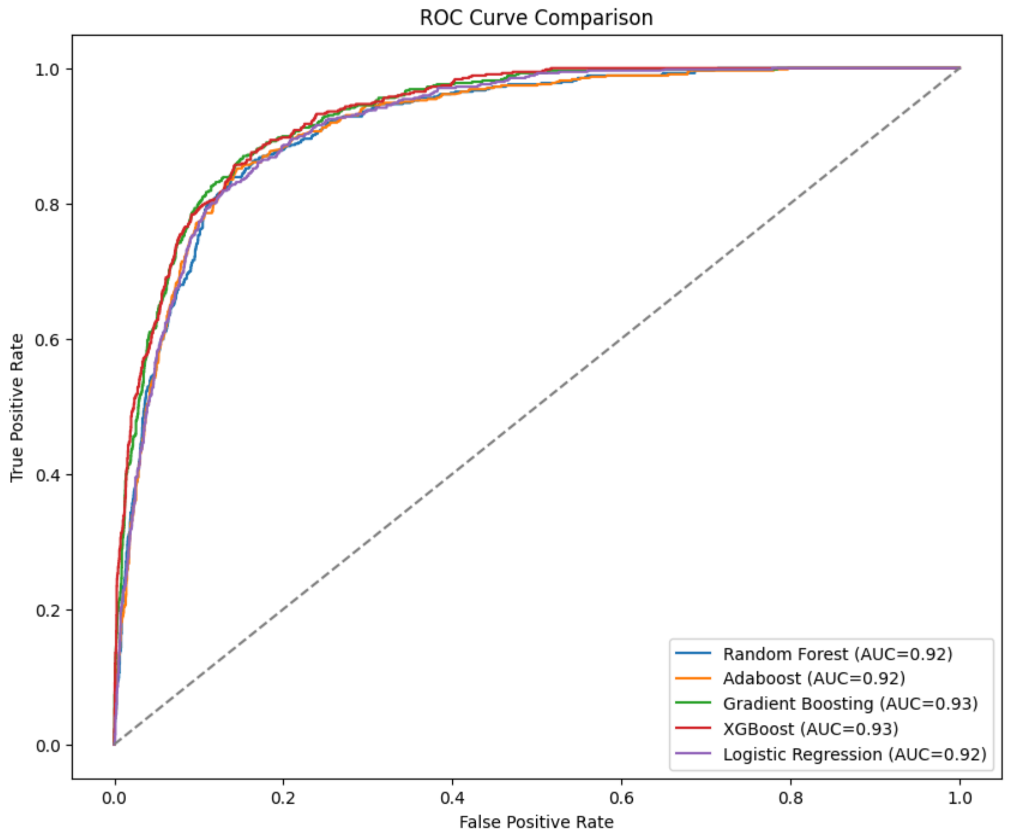


Fig 7.1: Auc Roc Curve

8.Conclusion and Future Work

Conclusion

The development of the Online Shoppers Purchasing Intention system highlights the effectiveness of machine learning in analyzing user behavior and predicting purchasing decisions in e-commerce environments. Through comprehensive data preprocessing, exploratory data analysis (EDA), and the application of machine learning models such as, Random Forest, and XGBoost, the project was able to build a robust and accurate prediction model. Among the models evaluated, ensemble-based approaches demonstrated superior performance in terms of accuracy and reliability.

The model was successfully integrated into a user-friendly web application using the streamlit framework, enabling real-time interaction and prediction. This deployment provides an accessible and practical interface for both academic exploration and potential business use.

Future Work

While the current system successfully meets its intended objective, there are several opportunities for enhancement in future iterations. One of the key improvements could be the integration with live e-commerce platforms, allowing real-time session data to be collected and processed for on-the-fly predictions. This would significantly enhance the system’s responsiveness and utility in practical applications.

Further improvements in feature engineering could involve incorporating domain-specific variables, identifying temporal behavior patterns, or including user demographic information. These additions could boost the predictive accuracy and provide more personalized insights into customer behavior.

In terms of model optimization, the adoption of hyperparameter tuning methods such as Grid Search, Random Search, or more advanced AutoML frameworks can automate and refine the model selection process, potentially leading to better-performing models with minimal manual intervention.

To support wider adoption and reliability, the system can be made scalable and production-ready by leveraging technologies such as Docker and Kubernetes, enabling efficient containerization and orchestration. Furthermore, deploying the system on cloud platforms like AWS, Azure, or Google Cloud would allow for broader accessibility and easier management of computational resources.

The user experience can also be enhanced through the development of a more interactive and visually rich frontend using modern frameworks such as React or Dash. This would make the application more engaging and accessible to a wider audience, including non-technical users.

Lastly, as the system processes user-related data, it is essential to emphasize data privacy, anonymization, and compliance with data protection regulations like GDPR. Ensuring robust data governance practices will not only build user trust but also safeguard against legal risks.

Overall, this project lays a strong foundation for building intelligent decision-making tools in e-commerce, with the potential to evolve into sophisticated recommendation systems and user behavior analytics platforms in the future.

REFERENCES:

[1] Wang, Runan. (2023). Research And Analysis of Online Shopper Intention. Journal of Education,
Humanities and Social Sciences. 16. 46-52. 10.54097/ehss.v16i.9496.

[2] A. Karakaya, İ. Karakaya and T. Temizceri, "An Online Shoppers Purchasing Intention Model Based on Ensemble Learning," 2023 4th International Informatics and Software Engineering Conference (IISEC), Ankara, Turkiye, 2023, pp. 1-4, doi: 10.1109/IISEC59749.2023.10391024

[3] Satu, M.S., Islam, S.F. Modeling online customer purchase intention behavior applying different feature engineering and classification techniques. Discov Artif Intell 3, 36 (2023). https://doi.org/10.1007/s44163-023-00086-0

[4] Frazier, Andrew & Maloku, Fatbardha & Li, Xinzi & Chen, Yichun & Jung, Yeji & Zohuri, Bahman. (2022). Data Analysis of Online Shopper's Purchasing Intention Machine Learning for Prediction Analytics. Journal of Economics & Management Research. 3(3): 1-8. 1-8. 10.47363/JESMR/2022(3)162.

[5] Baati K, Mohsil M. Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. Artificial Intelligence Applications and Innovations. 2020 May 6;583:43–51. doi: 10.1007/978-3-030-49161-1_4. PMCID: PMC7256375.

[6] G. Sang and S. Wu, "Predicting the Intention of Online Shoppers' Purchasing," 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Wuhan, China, 2022, pp. 333-337, doi: 10.1109/AEMCSE55572.2022.00074.

[7] García-Salirrosas, E.E.; Acevedo-Duque, Á.; Marin Chaves, V.; Mejía Henao, P.A.; Olaya Molano, J.C. Purchase Intention and Satisfaction of Online Shop Users in Developing Countries during the COVID-19 Pandemic. Sustainability 2022, 14, 6302. https://doi.org/ 10.3390/su14106302 DOI: https://doi.org/10.3390/su14106302

[8] Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. Neural Comput & Applic 31, 6893–6908 (2019).https://doi.org/10.1007/s00521-018-3523-0DOI:https://doi.org/10.1007/s00521-018-3523-0

[9] S. Mootha, S. Sridhar and M. S. K. Devi, "A Stacking Ensemble of Multi Layer Perceptrons to Predict Online Shoppers' Purchasing Intention," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 721-726, doi: 10.1109/ISRITI51436.2020.9315447. DOI: https://doi.org/10.1109/ISRITI51436.2020.9315447

[10] Mokryn, O., Bogina, V., Kuflik, T. (2019). Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. Electronic Commerce Research and Applications, 34, 100 836.