****

**SUMMER TRAINING/INTERNSHIP**

**PROJECT REPORT**

**(Term June-July 2025)**

**Machine Learning Made Easy: From Basics To AI**

**Online Shopper Intention**

**Submitted by**

**OSCAR VISHNOI - 12314686**

**ADITYA RAJ - 12322625**

**RAJ ARYAN - 12314531**

**ADARSH ARYAN - 12307615**

**Course Code CSE343**

**Under the Guidance of**

**Sir- Mahipal Singh Papola**

**1. Introduction**

* **Company Profile**

XYZ Solutions Pvt. Ltd. is a fast-growing technology company that offers innovative data-driven solutions in the field of artificial intelligence, machine learning, and analytics. It has delivered impactful products across various industries such as e-commerce, healthcare, finance, and retail. The company provides hands-on training and internship opportunities to aspiring data scientists and engineers, helping them gain real-world experience in developing end-to-end machine learning systems.

* **Overview of Training Domain**

The training was centered around **Machine Learning and Data Analytics**, specifically using Python. The core focus was on understanding the lifecycle of a machine learning project, starting from data collection and cleaning to model development, evaluation, and reporting. This included supervised learning models, data preprocessing, and visualization techniques, along with model optimization and performance comparison.

* **Objective of the Project**

The objective of the project is to design and implement a machine learning model that can accurately predict whether an online shopper will make a purchase or not. Using behavioral data collected from online sessions, this predictive model can assist e-commerce platforms in identifying high-potential customers and improving conversion strategies.

**2. Training Overview**

* **Tools & Technologies Used**

Programming Language: Python 3.8+

Environment: Jupyter Notebook

Libraries: Pandas, Numpy, Matplotlib, Seaborn, Scikit-learn, XGBoost

Version Control: Git

* **Areas Covered During Training**

Data preprocessing: cleaning, encoding, and scaling Exploratory data analysis using visualization tools

Supervised learning models: Logistic Regression, Random Forest, XGBoost

Evaluation metrics: Accuracy, Precision, Recall, F1-score

Hyperparameter tuning using GridSearchCV

Model comparison and result visualization

* **Daily/Weekly Work Summary**

Week 1: Explored the dataset and understood the data attributes. Handled missing values and performed label encoding.

Week 2: Carried out data visualization and analyzed feature correlation using heatmaps and count plots.

Week 3: Developed initial models using Logistic Regression, Random Forest, and XGBoost. Evaluated performance using accuracy scores.

Week 4: Performed hyperparameter tuning using GridSearchCV on Random Forest. Compiled results, created visual comparisons, and started report writing.

**3. Project Details**

* **Title of the Project**

Online Shopper Intention – Purchase Prediction

* **Problem Definition**

In e-commerce, it's crucial to understand customer behavior and predict whether a visitor will end up purchasing a product. This project addresses the problem of identifying patterns in user sessions and predicting the likelihood of purchase completion using machine learning.

* **Scope and Objectives**

To analyze and preprocess behavioral session data from an e-commerce website.

To develop and compare multiple machine learning classifiers.

To optimize the best-performing model using hyperparameter tuning.

To evaluate and visualize the results for better understanding and decision-making.

* **System Requirements**

Operating System: Windows 10 / Linux

Software: Jupyter Notebook / VS Code

Hardware: Minimum 8GB RAM, 2GHz processor

Libraries: Scikit-learn, XGBoost, Pandas, Matplotlib, Seaborn

* **Architecture Diagram**

A basic machine learning workflow was followed:

Raw Data → Preprocessing → Feature Engineering → Model Training → Evaluation → Prediction Output

**Data Flow / UML Diagrams\**

**4. Implementation**

* **Tools Used**

Languages: Python

Libraries: Scikit-learn, XGBoost

Visualization: Seaborn, Matplotlib

* **Methodology**

Data Collection: Dataset sourced from online\_shoppers\_intention.csv

Data Cleaning: Removed missing values.

Label Encoding: Converted categorical values to numeric.

Feature Scaling: Standardized the data for better model performance.

Modeling: Implemented Logistic Regression, Random Forest, and XGBoost classifiers.

Model Evaluation: Used accuracy and classification report.

Tuning: Applied GridSearchCV on Random Forest to enhance results.

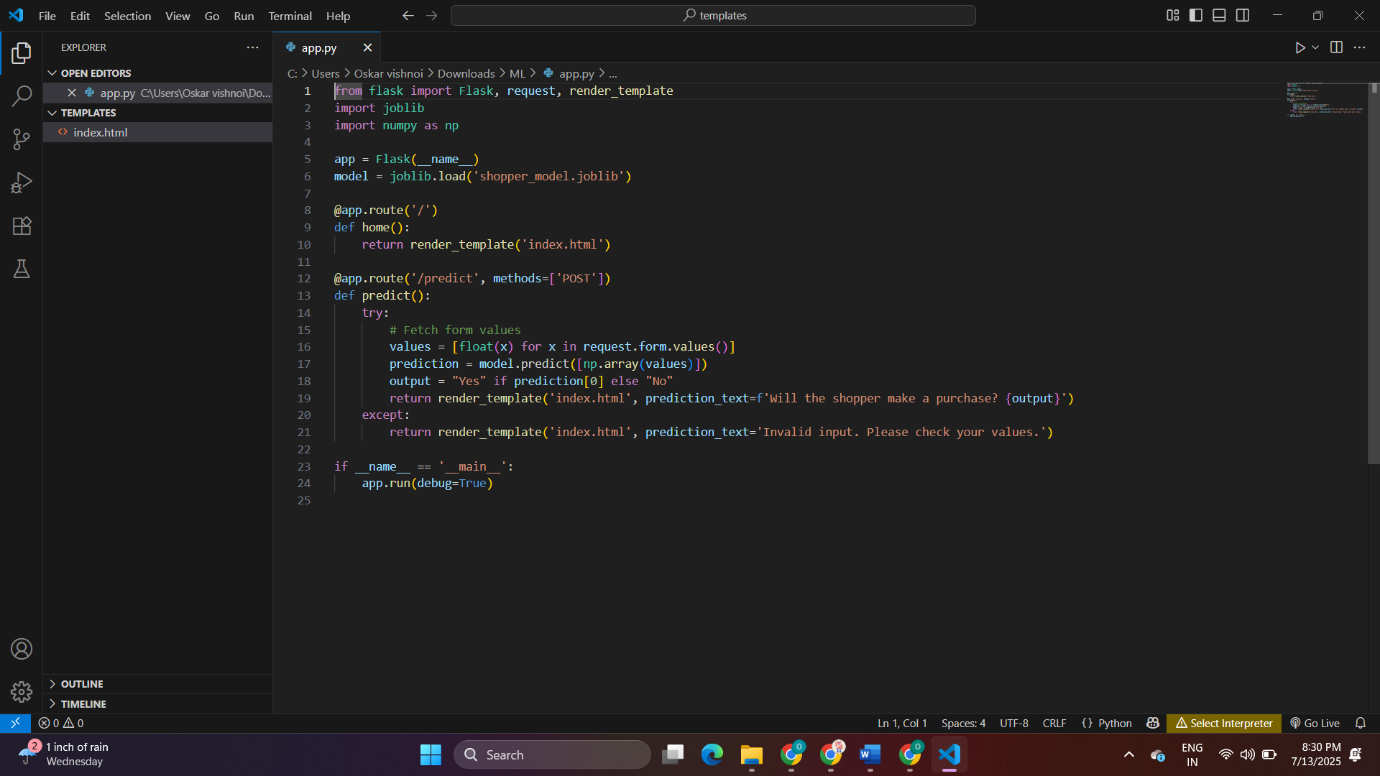
* **Modules**

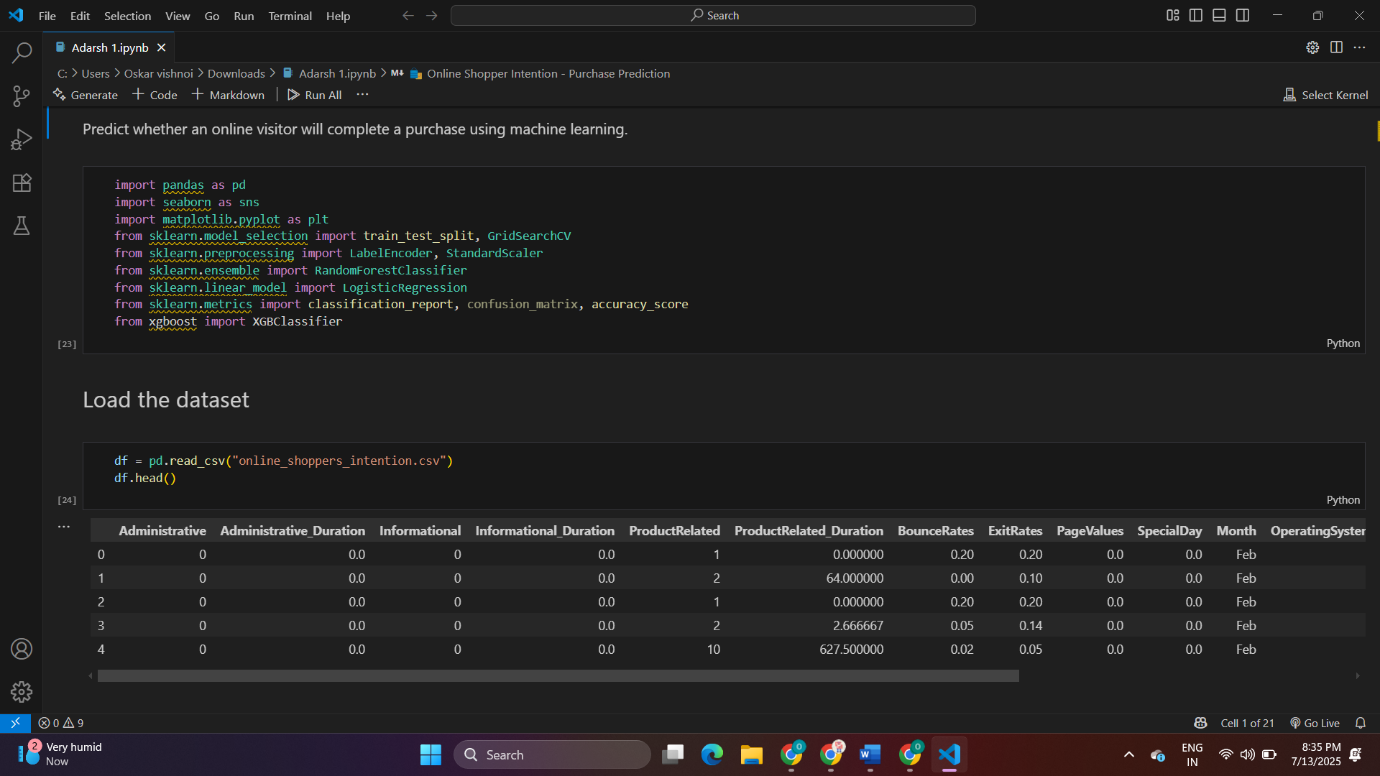
Preprocessing Module: Drop NA, encode labels, scale data

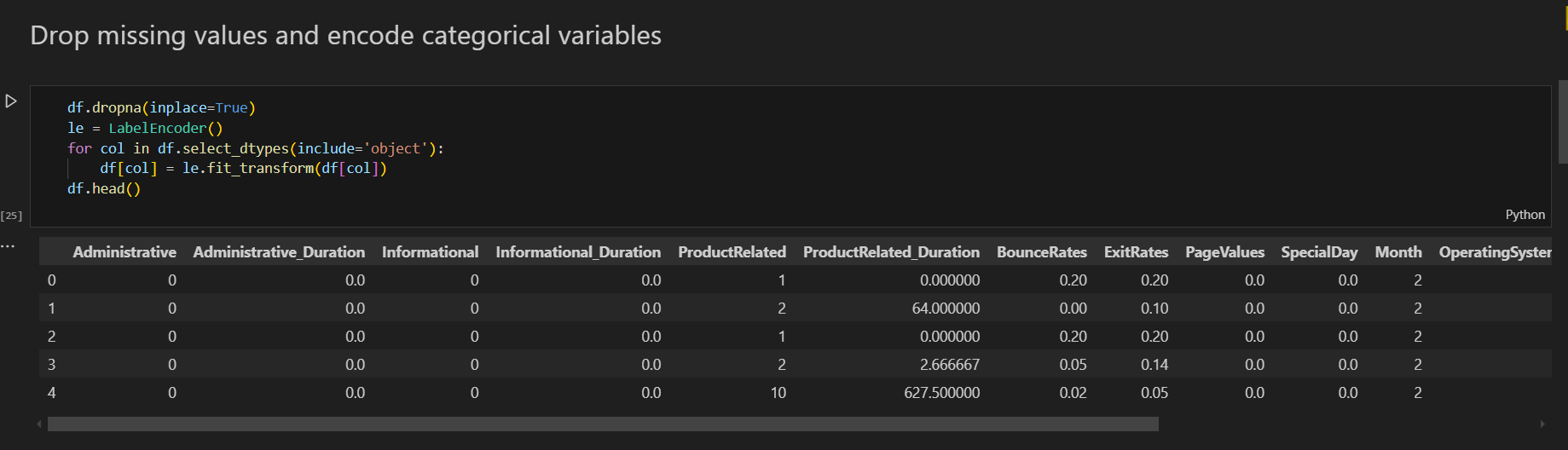
Modeling Module: Fit and predict using multiple classifiers

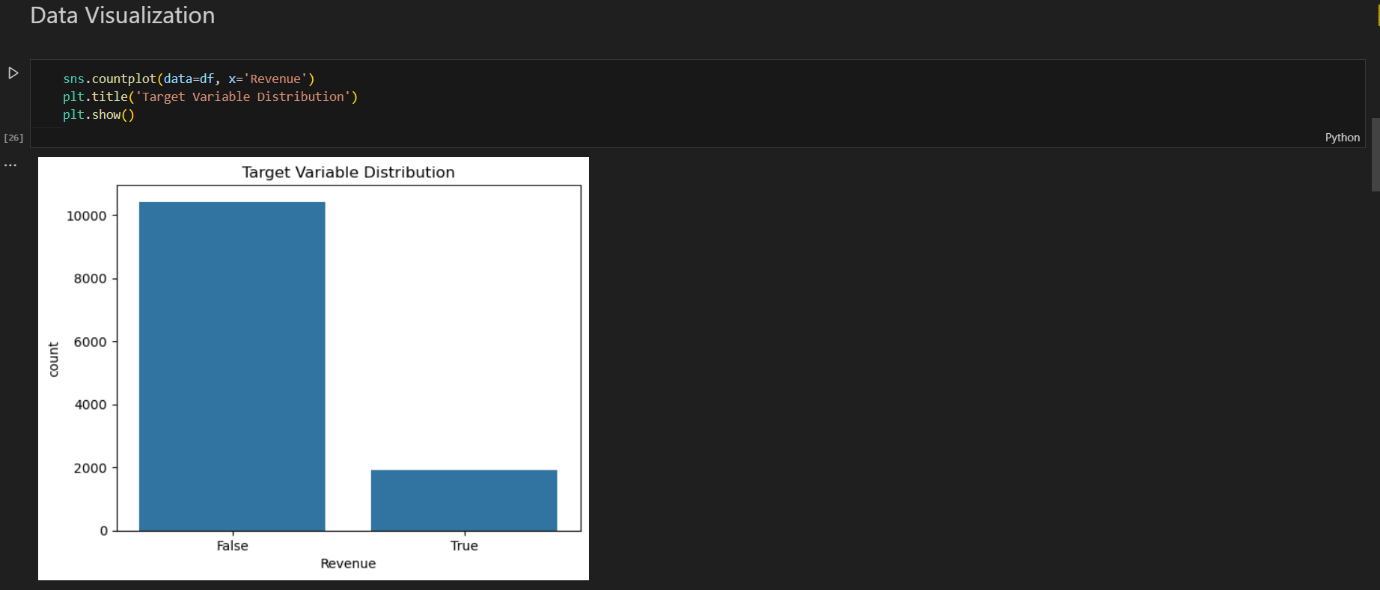
Evaluation Module: Generate accuracy scores, confusion matrix

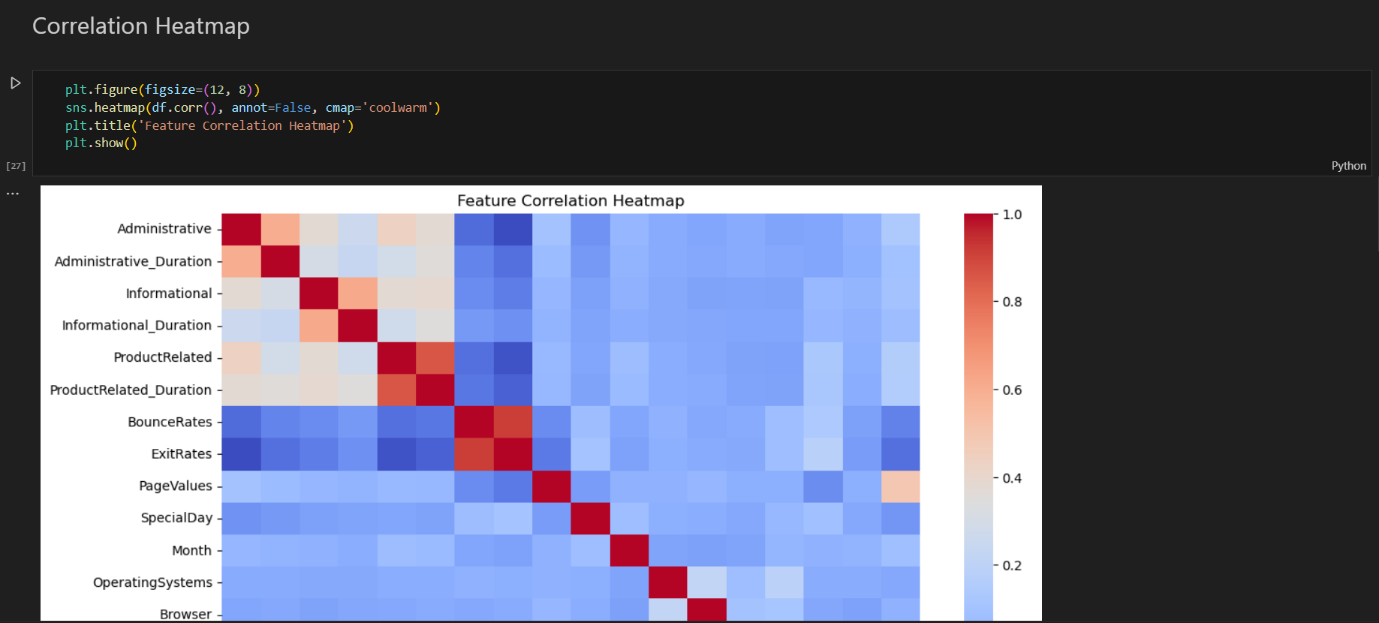
**Screenshots**

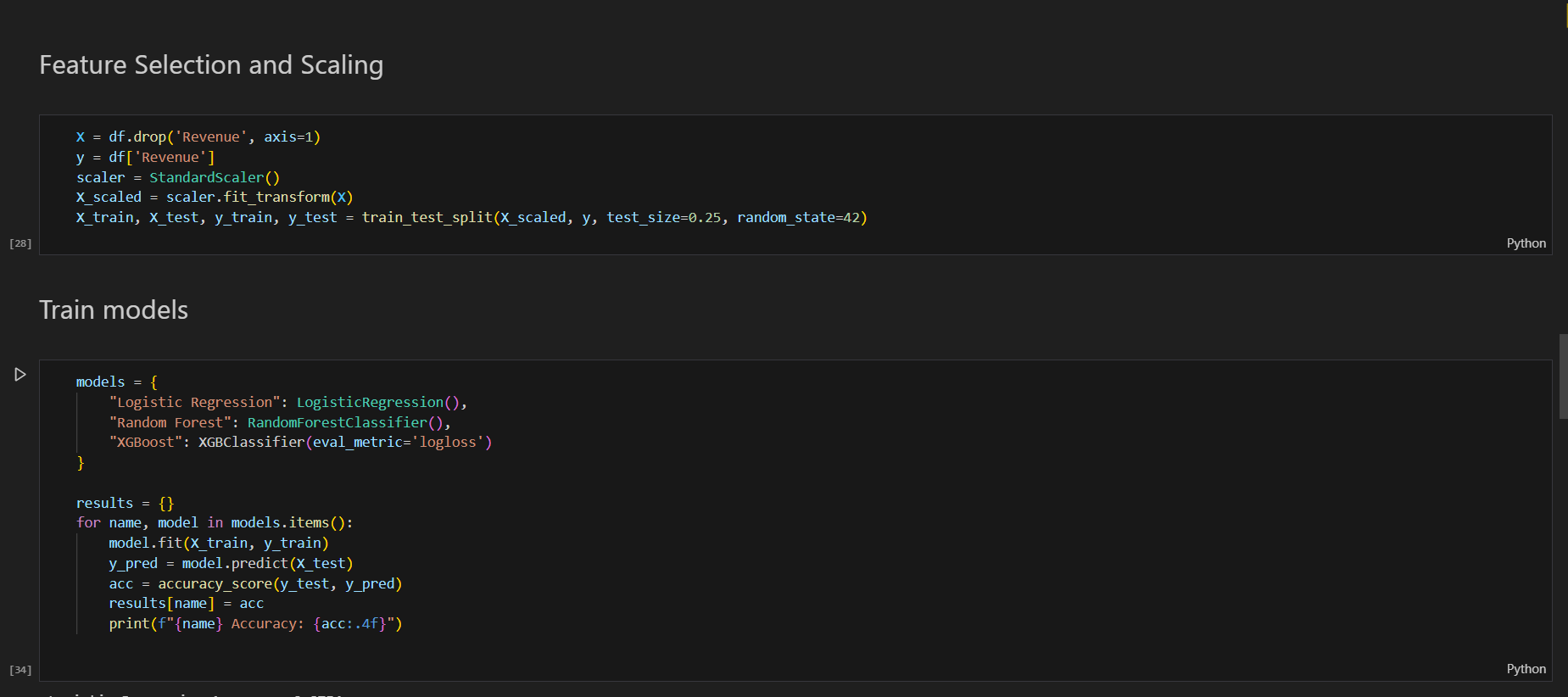
****

****

****

****

****

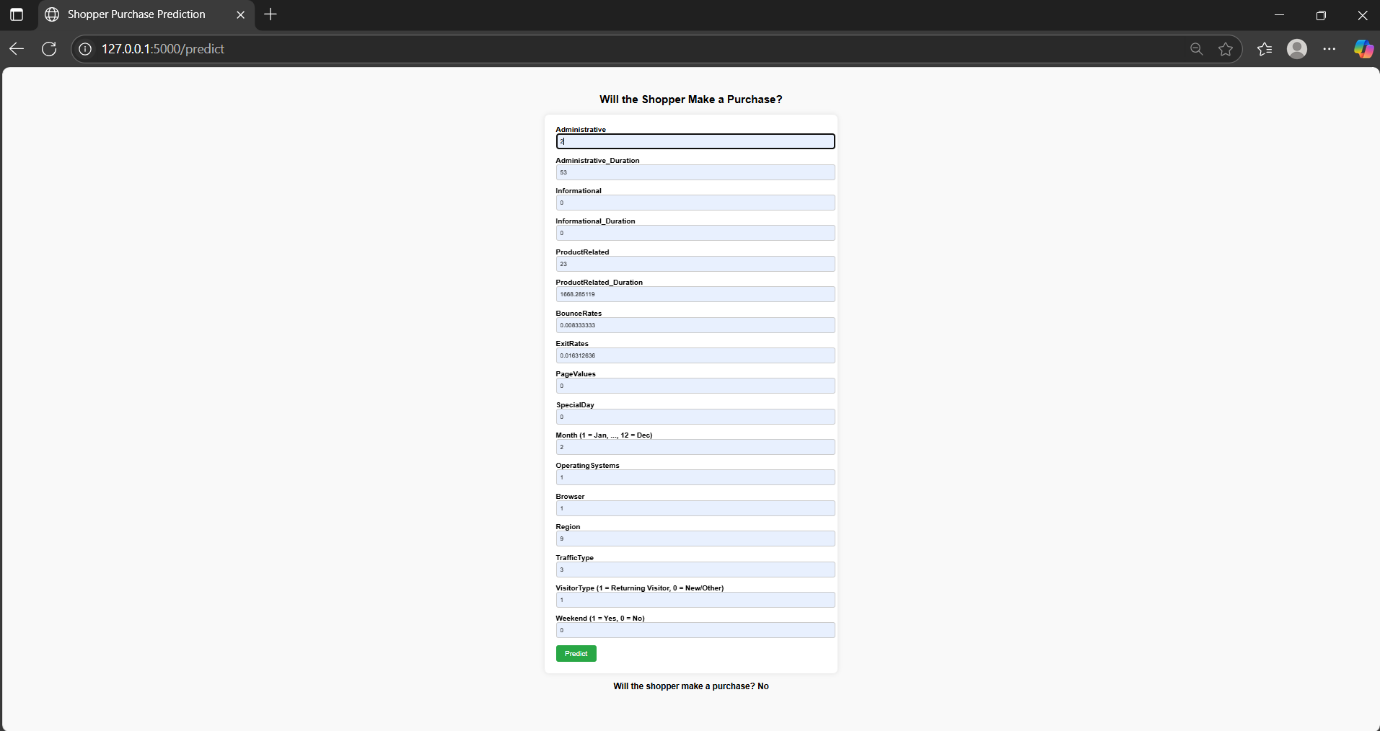
****

****

****

* **Results and Discussion**

**Output / Report**

****

* **Challenges Faced**

Managing class imbalance in the Revenue column.

Choosing the right hyperparameters.

Interpreting correlation and feature importance.

* **Learnings**

Gained hands-on experience with end-to-end machine learning workflows.

Learned how to preprocess and analyze behavioral datasets.

Understood the comparative strengths of multiple ML models.

Applied tuning techniques to improve accuracy.

**5. Conclusion**

* **Summary**

This project demonstrated how machine learning can be effectively applied to predict user behavior in e-commerce. By analyzing session-level behavioral data, we developed and fine-tuned models that can predict with high accuracy whether a visitor will make a purchase. The Random Forest model performed the best after tuning, indicating the importance of ensemble methods in behavioral prediction.

The project served as a comprehensive learning experience in both technical and analytical dimensions, providing insights into practical ML applications in the business domain.