E-commerce Purchase Intent Prediction Report

1. Introduction

This project aims to predict whether a visitor to an e-commerce website will make a purchase. Using the Online Shoppers Purchasing Intention dataset, we analyse session-based behavioural data to build machine learning models capable of classifying purchase intent.

2. Dataset Overview

The dataset contains 12,330 rows and 18 features, including page durations, bounce rates, traffic sources, month, visitor type, and weekend flag. The target variable is 'Revenue', indicating whether a purchase was made (True) or not (False).

3. Data Exploration

We examined the dataset for null values, data types, and class distribution. The target variable 'Revenue' is imbalanced, with significantly fewer positive (purchase) cases. Count plots and
descriptive statistics were used to understand the feature distribution

4. Data Preprocessing

Data cleaning and transformation steps included:
- No missing values were found, so no imputation was necessary.
- Converted categorical variables such as 'Month', 'Visitor Type', and 'Weekend' using one-hot encoding.
- Applied feature scaling using Standard Scaler to standardize numerical attributes.
- Addressed class imbalance using SMOTE, which synthetically increased the number of positive class samples

5. Model Building

Three machine learning models were built and trained on the processed data:

- **Logistic Regression:**
  A simple linear model that achieved an accuracy score of 0.84.

- **Support Vector Machine (SVM):** A powerful classifier for high-dimensional data, achieving an accuracy score of 0.88.

- **Random Forest**:
  An ensemble model combining multiple decision trees, which achieved the best performance with an accuracy of 0.93.

6. Model Evaluation

- **Confusion Matrix:**
  Confusion matrices were generated for each model to evaluate prediction accuracy in terms of true positives, true negatives, false positives, and false negatives.

- **ROC Curve and AUC:**
  ROC curves were plotted for each model, and the Area Under the Curve (AUC) was used as a key metric:
  - Logistic Regression: AUC = 0.92
  - SVM: AUC = 0.94
  - Random Forest: AUC = 0.99

The Random Forest model performed the best across all evaluation metrics.

7.Conclusion

The Random Forest model outperformed others in terms of accuracy and ROC-AUC. With proper preprocessing and class balancing, we achieved high-performance classification on user session data. This project demonstrates the effectiveness of ML models in predicting user behaviour in e-commerce settings.