Supervised ML (Assignment4)

Assignment Project

Problem Statement

An e-commerce company, **ShopSmart**, wants to better understand the behaviour of visitors browsing its online store. Every day, thousands of users navigate through product pages, spend varying amounts of time on the website, and interact differently with content. Currently, the company is unable to accurately predict which visitors are likely to complete a purchase, resulting in inefficient marketing strategies and lost revenue opportunities.

To address this challenge, the company has collected data from **12,330 individual user sessions** over a period of one year. Each session represents a unique visitor to ensure unbiased insights, avoiding the influence of specific users, promotional campaigns, seasonal trends, or special events. The dataset includes a mix of **numerical and categorical features** that capture detailed information about users' browsing behaviour and interactions with the website.

You are hired as a **Machine Learning Engineer** to build a predictive model that can determine whether a visitor is likely to make a purchase based on their session behaviour. Your task involves performing **exploratory data analysis (EDA)**, applying appropriate **feature preprocessing and transformations**, and developing a **Decision Tree based classification model**. Since the dataset is **imbalanced**, model performance must be evaluated using the **F1 score**, with a benchmark value of **0.55** used to assess the effectiveness of the solution.

Note - Make use of pruning to improve the performance of Decision Tree Classifier.

Dataset Description

Feature Name	Description
Administrative	Number of pages visited related to account management (login, profile, etc.)
Administrative_Duration	Total time (in seconds) spent on administrative pages
Informational	Number of informational pages visited (FAQ, policies, etc.)
Informational_Duration	Total time (in seconds) spent on informational pages
ProductRelated	Number of product-related pages visited
ProductRelated_Duration	Total time (in seconds) spent on product-related pages
BounceRates	Percentage of visitors who leave the site after viewing one page
ExitRates	Percentage of exits from a page
PageValues	Average value of pages visited before completing a transaction
SpecialDay	Closeness of the visit date to a special day (e.g., Valentine's Day, Mother's Day)
Month	Month of the visit (Feb, Mar, May, etc.)
OperatingSystems	Operating system used by the visitor (encoded)
Browser	Browser used by the visitor (encoded)
Region	Geographic region of the visitor (encoded)
TrafficType	Source of traffic (e.g., direct, referral, ads)
VisitorType	Type of visitor (`Returning_Visitor`, `New_Visitor`, `Other`)
Weekend	Whether the visit occurred on a weekend
Revenue	**Target variable** - whether the visitor made a purchase