Online Shopper Purchase Prediction Report

By: Zul Fauzi Abas

# Introduction

In the evolving landscape of e-commerce, understanding and predicting online shopper behavior presents a significant opportunity for businesses. The goal of this project is to develop a predictive model capable of determining whether a website visitor will make a purchase based on their browsing behavior. Achieving a model accuracy that surpasses the baseline by at least 10% is our primary objective. This report outlines the approach, data analysis, model development, and findings of our project.

# Data Overview

The dataset for this project was sourced from Kaggle and focuses specifically on features related to users' browsing behavior. It is assumed that this dataset accurately reflects the broader population of online shoppers, providing a reliable foundation for predicting purchase intentions.

# Methodology

Our approach to developing the predictive model comprised several key stages:

Exploratory Data Analysis (EDA): Initial data exploration revealed patterns, outliers, and skewness across various features, alongside a noticeable class imbalance.

Data Processing: We addressed these issues by removing outliers using the Interquartile Range (IQR) method, correcting skewness through log transformation, and correcting class imbalance with the Synthetic Minority Over-sampling Technique (SMOTE).

Model Development: Several machine learning models were tested, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Trees, and Random Forests. The choice of models was informed by their suitability for classification tasks.

Model Tuning: The best-performing model was then fine-tuned to optimize its predictive accuracy.

# Findings

## Exploratory Data Analysis

The EDA phase highlighted the presence of outliers and right-skewed distributions across multiple features, potentially leading to model overfitting and inaccurate predictions. Additionally, the observed class imbalance posed a risk of bias towards the majority class.

## Data Preparation

The data preparation steps effectively mitigated these issues, ensuring the dataset was well-suited for model training. Outliers were removed, skewness was normalized, and class imbalance was rectified, laying the groundwork for accurate and reliable model predictions.

## Model Evaluation

Among the tested models, Logistic Regression demonstrated the optimal balance between training and testing performance, indicating a strong generalization capability without overfitting. After fine-tuning, the model achieved an accuracy of 84.3%, significantly exceeding our initial goal and demonstrating a substantial improvement over the baseline prediction accuracy.

# Conclusion and Future Work

The project successfully developed a predictive model that significantly improves the accuracy of online purchase predictions, exceeding the predefined objective. This model has the potential to enhance e-commerce strategies by enabling more personalized and efficient customer interactions.

Moving forward, the focus will be on further refining the model, exploring additional features, and investigating new applications. Ensuring ethical data use and adapting to emerging e-commerce trends will remain paramount as we continue to leverage predictive analytics to improve the online shopping experience.