ONLINE SHOPPING PURCHASE INTENTION

GROUP NUMBER 2

- Sophia Nelson - 6428678
- Sarah Daniel - 6909062
- Reethender Vedira - 6910677

UNIVERSITY OF SURREY

PROJECT AIM AND OBJECTIVES

Problem Statement
This project aims to identify the key behavioral and technical factors that influence whether a customer completes an online purchase, to improve the conversion rate for an e-commerce platform.

Hypotheses
- Higher user engagement (e.g., more time on product pages, higher page value scores) increases the likelihood of making a purchase.
- Lower bounce and exit rates are associated with a higher probability of conversion.
- The month of the visit and the type of content viewed (product vs. informational pages) significantly influence purchase behavior.

Aim
To apply machine learning techniques to predict online purchase behavior, identify key drivers of conversion, and provide actionable insights to optimise e-commerce strategies.

Objectives
- **Data Understanding and Preparation**: Explore the dataset, clean, and transform it for machine learning use.
- **Exploratory Data Analysis**: Analyse patterns and visualize relationships between features and purchase intent.
- **Modeling**: Train and evaluate multiple machine learning models (Logistic Regression, Random Forest, XGBoost) to predict purchase intent.
- **Evaluation**: Assess model performance using metrics such as Precision-Recall AUC, Recall, ROC AUC and F1 score among others.
- **Insights and Recommendations**: Provide actionable recommendations based on model findings to improve conversion rates.
- **Reporting**: Present findings and recommendations clearly and concisely to ecommerce/business managers, marketers, data analysts.

CRISP-DM FRAMEWORK APPLICATION

Core Business Objectives:
Identify the key behavioural and technical factors that most influence whether a customer completes a purchase online.
Use insights from shopper behaviour to recommend data-driven strategies that can improve the sales conversion rate.

Success Criteria:
Insights from the analysis enable targeted marketing strategies that seek to increase the online conversion rate for the long term.

Source of data:
The dataset *'Online Shoppers Purchasing Intention'* was collected from the external source University of California, Irvine (UCI) repository.
The dataset captures anonymised session-level behavioral data of visitors to an e-commerce website.

Initial observations:
The dataset contains 12,330 instances and 17 features. The data contains a mix of quantitative and categorical/symbolic variables.

Key variables:
The target variable is 'Revenue' and the values in the column are stored as a Boolean (True/False).

Some of the important predictor variables include page value score, exit rate, bounce rate and visit month.

Modelling approach:
The project focuses on a binary classification problem to predict the likelihood that a user session will result in a conversion.

Firstly, a baseline model for Logistic regression will be used, followed by a parsimonious model for feature reduction and interpretability.

To complement this, two tree-based models (Random Forest and XGBoost) will be used to maximise prediction accuracy and to verify whether those same predictors are still the most influential and to identify potential non-linear relationships that the logistic regression model may have missed.

Tools & Techniques:
R, PowerBI, tidyverse, caret, preprocessing (scaling, encoding), Variable Inflation Factor, ROSE for balancing, K-fold (n=10).

Evaluation:
confusion matrix, Roc Curve (overall ability separate classes), PR Curve (better for class imbalance), AUC, precision, recall, F1 score, Accuracy), feature importance, SHAP values.

Reporting:
Provide actionable insights for business decision-makers to optimise marketing strategies that result in enhanced online conversion.

BENEFITS OF CRISP - DM

- **Structured Approach**
  - CRISP-DM provided a clear, systematic process, helping us stay focused on the business objectives and ensuring a thorough analysis. It guided us through each stage of the project in a logical and organized manner.
- **Improved Data Quality**
  - By following the Data Understanding and Data Preparation steps rigorously, we ensured that the data was properly cleaned, balanced, and ready for modeling. This led to better model performance and more reliable insights.
- **Effective Model Selection and Evaluation**
  - The iterative nature of CRISP-DM allowed us to experiment with different machine learning models, tuning them based on performance metrics. This led to the identification of the most accurate model for predicting purchase intent.
- **Actionable Business Insights**
  - By following CRISP-DM's Evaluation and Deployment steps, we could link model findings directly to business strategies, such as improving website engagement and reducing exit rates, ensuring the insights were practical and actionable.
- **Flexibility and Iteration**
  - The iterative nature of CRISP-DM allowed us to revisit and refine earlier steps as needed, improving the overall quality of the analysis and ensuring that the final model was the best fit for the problem.