Synopsis

Project Title: Modeling Online Purchase Intent Through Behavioural Clustering and Classification

Authors: Ane Novrup Larsen, Nathasja Skov Fink Nielsen

Problem Statement

E-commerce platforms attract thousands of daily visitors, yet only 4% complete a purchase[1]. Traditional machine learning models that predict purchasing intention typically rely on raw session metrics such as time spent on pages, bounce rates, and product views.[2] However, online shoppers exhibit highly diverse behaviors. Research by Huseynov and Yıldırım (2016)[3] categorises these visitors into five distinct segments: 'opportunist customers,' 'transient customers,' 'need-based shoppers,' 'skeptical newcomers,' and 'repetitive purchasers.' Consequently, many visitors simply browse without immediate intent to buy.

Research Question

Does segmenting online visitors based on their navigational engagement (Unsupervised Learning) improve the accuracy of predicting purchase intention (Supervised Learning) compared to using raw traffic metrics alone?

Proposed Approach

This project will use the *Online Shoppers Purchasing Intention Dataset* from the UCI Machine Learning Repository[4], which contains session-level behavioural data such as page visit durations, bounce and exit rates, page values, special-day proximity, and visitor types.

1. Data Preparation: The dataset will be preprocessed by One-Hot encoding categorical features, and standardising numeric variables. Given the rarity of purchase events, measures will be taken to prevent the model from biasing towards the majority class. Exploratory Data Analysis (EDA) will be conducted to understand behavioural patterns and class distribution between purchasing and non-purchasing sessions.

2. Unsupervised Learning (Segmentation): We will primarily apply a K-Means clustering approach to identify distinct groups of visitors with similar browsing patterns. However, acknowledging that shoppers might form irregular (concave) groups, we may potentially evaluate DBSCAN as an alternative. To prevent data leakage, the clustering model will be fitted only

[1] McDowell, W.C., Wilson, R.C., & Kile, C. (2016). *An Examination of Retail Website Design and Conversion Rate.* Journal of Business Research, 69(11), 4837–4842.

[2] Allenbrand, C. (2023). *Clicking through the Clickstream.* Intelligent Information Management, 15(3), 180–215.

[3] Huseynov, F., & Yıldırım, S. (2016) *Behavioural segmentation analysis of online consumer audience in Turkey by using real e-commerce transaction data.* International Journal of Economics and Business Research, 14(1), 12–28.

[4] UCI Machine Learning Repository – Online Shoppers Dataset

on the training set, and the resulting centroids (in the case of K-Means) will be used to assign labels to the test set.

3. Supervised Learning (Prediction): Two predictive models—one baseline (using raw behavioural metrics) and one enhanced (including cluster labels as an additional feature)—will be trained using Logistic Regression, Decision Trees, and Random Forest classifiers. Model performance will be assessed using standard evaluation metrics such as log-loss, Youden's Index, precision, recall, F1-score and ROC-AUC.

4. Evaluation The comparison between models with and without behavioural segmentation will determine whether unsupervised segmentation contributes to improved predictive performance.