Student Name: Ishita Patel

Student ID: 501326439

Date of Submission: September 23, 2024

CIND820: Big Data Analytics Project

The boom in online shopping has been a permanent foundation of modern commerce thanks to the ease and convenience it extends to consumers worldwide. The four key drivers of online shopping behavior over in-store retail are digital technology, e-commerce platforms, growth on the Internet, and assortment expansion. A growing number of customers are making purchases on internet sites, so companies must learn what causes a customer to buy or walk away. By understanding these behaviors, companies can fine-tune their websites and sharpen their marketing tactics to deliver better overall customer experiences and drive conversions. Online buyer behavior is often studied to identify patterns that help businesses shape their website design, product offerings, or marketing strategies.

A combination of classification and regression themes and click stream analysis would be used as it allows for the prediction of outcomes like purchase decisions and continuous values, such as session duration. This method offers an understanding of user behavior and buying habits, aiding in refining marketing tactics and enhancing user experience by anticipating shopper intentions from their clickstream activities.

The main focus of this project revolves around the following research queries; "What types of shopping habits indicate a likelihood of intending to make a purchase? Additionally, how can we create a model that predicts whether a customer will finalize a purchase?" By addressing these questions companies can confidently decide how to enhance their websites and implement tailored marketing tactics for boosting conversion rates.

This research project utilizes the Online Shoppers Purchasing Intention Dataset from the UC Irvine Machine Learning Repository, which comprises 12,330 user session entries on a shopping platform. The dataset consists of 10 features relating to aspects like duration spent on pages and

bounce rates, as well as 8 categorical features such as customers' browser type and geographic location of the customer.

To solve the problem of classifying purchasing intentions, the class imbalance will be addressed by oversampling the minority class of sessions that resulted in purchases, improving model performance. A correlation matrix and a Random Forest Classifier will be used for feature selection, identifying key factors contributing to successful purchases. The data will be split into a training set (70%) and a test set (30%) for model evaluation. Prediction models, including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), and Multilayer Perceptron (MLP), will be created to anticipate buying decisions. Each model's performance will be assessed using accuracy scores and confusion matrix outcomes, providing insights into their effectiveness in forecasting purchasing patterns.

For this project, Python and its related libraries, like pandas and NumPy, are utilized for data analysis and model creation purposes. The preprocessing of data will involve managing missing values, scaling features, and encoding variables. These research findings can provide insights for companies aiming to enhance their online shopping platforms and boost sales through an understanding and prediction of customer buying patterns.

This project aims to offer advice that can be used to make business choices that will improve customer satisfaction and boost sales, on online retail sites.

References

- Sakar, C. O., & Kastro, Y. (n.d.). *Online Shoppers Purchasing Intention Dataset Data Set*. UCI Machine Learning Repository: Online Shoppers Purchasing Intention Dataset Data set. https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

- Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing & Applications*, *31*(10), 6893–6908. https://doi.org/10.1007/s00521-018-3523-0