David Wang

501208639

Tamer Abdou, Ph.D.

May 15, 2023

Online commerce is an industry that has greatly benefited from the ever increasing development of today’s technologically connected world. Since the COVID-19 pandemic, businesses across Canada have reported trends of shoppers turning to online purchases over retail stores across different industries.

As such, developing an understanding of the patterns and mindsets of online shoppers would allow businesses to further increase sales. Being able to identify what separates a customer who purchases an item from those who leave the website without a purchase would allow businesses to develop their marketing and websites in a way to maximize purchases made on a website. The goal is then to predict which patterns of online shopping behavior lead to successful purchases. Such findings could help drive marketing decisions, website design and/or product design with the goal of increasing sales in businesses.

The data was sourced from the *Online Shoppers Purchasing Intention Dataset* from the UCI Machine Learning Repository. It contains 12,330 records comprising 10,422 (84.5%) sessions that did not lead to a purchase and 1,905 (15.5%) sessions that led to a customer purchase. There are 10 numerical attributes and 8 categorical attributes that describe the browsing behavior of each potential customer, such as the types of pages they would visit and the duration of time spent on specific types of pages. (can expand) External factors, such as the operating systems, browsers, and regions of the customer are also included. Additionally, time-based factors such as proximity to a holiday were recorded to determine if it had an effect on the shopping behavior of customers.

The goal is to build a classification model that will successfully be able to classify which customer behaviors led to a purchase and which did not. I intend to first balance the dataset by oversampling for successful purchases, as the disparity between records of outcomes will affect prediction performance. Afterwards, I will select for features using a correlation matrix and a Random Forest Classifier in order to improve classification results. Afterwards, I will split the data into training (70%) and test (30%) datasets. I then plan to use multiple algorithms (logistic regression, Naive Bayes, Support Vector Machines (SVM), Multilayer Perceptron (MLP)) to classify whether specific customer behaviors led to purchases. Finally, I plan on scoring the performance of the models using accuracy scores generated from confusion matrices in order to compare between performances of feature selection methods and classification algorithms. The technologies I intend to use will be Python Jupyter notebooks and the pandas library for general data analysis. The data will be imported from the UCI Machine Learning Repository in a .csv format.

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. (2018, May 9). *Real-time prediction of online shoppers’ purchasing intention using Multilayer Perceptron and LSTM recurrent neural networks - neural computing and applications*. SpringerLink. https://link.springer.com/article/10.1007/s00521-018-3523-0
* Zanzana, S., & Martin, J. (2023, February 21). Retail e-commerce and COVID-19: How online sales evolved as in-person shopping resumed. https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2023002-eng.htm