ORIE 5741 Project Proposal

Team Members: Lisa Lyu (zl879), Yuanqi Sun (ys2274), Wenxuan Zhang (wz443)

Project Overview

The increase in e-commerce usage in recent years has opened up substantial potential in the market. However, despite this growth, conversion rates, which reflect the percentage of visitors who make a purchase out of the total number of visitors to the website, have not kept pace, prompting the need for seeking solutions that offer personalized promotions to online shoppers. In traditional retail settings, salespeople use their accumulated experience to provide shoppers with a variety of customized options. This experiential knowledge significantly impacts time efficiency, conversion rates, and overall sales performance. To translate this personalized service into the digital domain, the application of machine learning models to predict and replicate customer behaviors becomes paramount.

With this content, our project aims to answer the following question: can we accurately predict a visitor's purchasing intent based on their online interactions and other behavioral attributes? Addressing this question will enable us to refine the e-commerce marketing strategies, enhance website design, and elevate the overall customer journey. Ultimately, our objective is to optimize conversion rates and revenue generation in the online retail landscape through proactive and data-driven strategies.

Dataset Introduction

The dataset we will use is “Online Shoppers Purchasing Intention Dataset” from UCI machine learning repository. This dataset consists of 12,330 sessions, where each session belongs to a different customer in a 1-year period. This duration ensures diversity by mitigating biases towards specific campaigns, special occasions, user profiles, or time periods, as the data was systematically crawled from various e-commerce platforms.

The target value we want to predict is a binary value, indicating whether the visitor has been finalized with the transaction. In this dataset, 84.5% (10,422) are negative class samples that did not end with shopping, and the rest (1908) are positive samples. The dataset offers a rich array of both numerical and categorical features potentially correlated with customer intent. Key numerical features encompass the visitor's browsing behavior, including metrics such as the number of distinct page types visited during the session, the cumulative time spent on each page category, and various metrics from Google Analytics commonly employed in web activity analysis derived from URL information. Categorical features include the time and geographical information of the visitor, as well as visitor characteristics such as type and the specific browser and operating system employed during the session.

Link to the dataset:
https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset

Possible Approaches

- Exploratory Data Analysis
  When exploring our data, several visualization charts could be made. We would use seaborn packages to explore the count of revenue by different features. For example, we can explore in different months, what is the number of finalized transactions. This exploration would allow us to figure out some internal patterns such as seasonality, user preferences, and regional differences. What's more, a correlation analysis is also important. By quantifying the relationships between features and visualizing these connections through a heatmap, we can identify the most pertinent attributes for model development. This meticulous exploration not only enriches our understanding of the dataset but also enhances the robustness of our predictive models.
- Processing the Data When Splitting to Traning and Testing
  Given the dataset, we realized that it is imbalanced. There are 80% of customers hold a False result, which means these people didn't finalize the transaction. And there are only 20% of customers who finalized the transaction. Thus, approaches such as oversampling, undersampling, or class weighting might be used.
- Since this is a classification problem, there are several models we could try based on what we have learned from lectures. SVM is one of the models we can try. We will identify the most correlated features and try kernelized SVM, which could also be used after we fully explore the data pattern and distribution. Logistic regression is another choice. The output would be the probability that a customer would finalize the transaction once he/she stays on the website. The last possible model is decision trees or random forest, depending on the performance of the model. Trees can classify data more precisely, but would also lead to an overfitting problem. Thus, using a bagging method by doing the majority votes would reduce the variance.