CS519 Group Project - Gamma

Phase 2 Report

Indronil Bhattacharjee, Roscoe Hill, and Erica Flores

Motivation

E-commerce defines the activity of buying or selling goods through an online, digital platform which has become a critical part of the global economy. Retailers of every size, ranging from small home-based businesses to large corporate entities utilize online selling platforms to increase their reach to new customers. This is particularly important for small businesses, who can avoid the large overhead costs of a brick and mortar storefront.

According to the article "38 eCommerce Statistics of 2023" (Forbes Advisor Online):

- By 2026, the e-commerce market is expected to total over $8.1 trillion
- By 2026, 24% of retail purchases are expected to take place online
- 20.8% of retail purchases are expected to take place online in 2023

Problem Definition

E-commerce does come with the caveat of trying to sell an item based on listing information and photos to a customer who cannot physically interact with the item before deciding to purchase. It is critical for a business involved in e-commerce to consider which attributes of their products or service are most critical to translating an online shopping session into a purchase.

Machine Learning Tasks

By running publicly available E-commerce data through various machine learning models, we aim to compile a report of recommendations for that business in order to maximize sales.

The first ML (machine learning) task that we will cover is Regression. We will use regression to determine whether free shipping or offering coupons increases sales more. We can also use Regression to determine which product category is most successful and should be advertised more.

A second ML task that we can focus on is classification. Possible classification attributes could be classifying transactions as "successful" or "unsuccessful" based on whether a purchase is made. And recommending similar items to customers based on a previous purchase.

Lastly, a ML task that we deem important is clustering. We could possibly cluster customers based on demographic data to determine an ideal audience. We can also use clustering to determine which products have similar performance.

Related Works

There is no shortage of other teams in the world also trying to tackle the same problem of how to maximize machine learning with raw data to gain profit. Other related works such as the Linear Regression Project by Akarsh654 focusing on time spent on a website and length of membership correlation is a great example. There are many attributes one can tackle when it comes to E-commerce for analysis.

References

(1) Akarsh654, Machine Learning Project, 2020, GitHub Repository https://github.com/Akarsh654/Machine-Learning-Projects/blob/master/Linear%20Regression/Ecommerce/Ecommerce%20Project.ipynb

(2) Baluch, A. "38 eCommece Statistics of 2023", Forbes Advisor, 8 February 2023, https://www.forbes.com/advisor/business/ecommerce-statistics/#sources_section