Comparing the Effectiveness of User and Item-based Recommender Systems for Medium-sized E-Commerce Businesses

Cinny Lin (ycl461), Yingrong Mao (ym1780), Zhihan Yang (zy1229)

Introduction

As technology advances and logistics improve, more people are able to buy as well as sell products using online shopping sites. However, the dominance of gigantic e-commerce platforms such as Amazon and Taobao makes it especially difficult for their smaller counterparts to survive. We identified that a good recommender system is essential for any platform to become popular today, as seen in Spotify for music and Netflix for film streaming service, or Tiktok and Twitter for video and text-based social media [Polatidis and Georgiadis]. Therefore, we believe that building a recommender system for smaller e-commerce platforms will be able to help them increase revenue and become more competitive in regional areas.

Although Amazon never revealed its statistics, McKinsey estimated that 35% of what consumers purchase on the e-commerce giant came from product recommendations. This approximation again suggests that a well-designed recommendation algorithm can substantially drive product sales for medium ecommerce businesses. We thought that we could use geographic data and behavioral patterns to infer consumers' demographics, and perform customer segmentation based on these given and deducted data to give product recommendations. Interestingly, we found that Amazon has suggested the opposite, explicitly stating that "the better way was to base product recommendations not on similarities between customers but on correlations between products".

Therefore, our project would focus on comparing the recommendation effectiveness based on two different approaches -- consumer-based and item-based. Consumer-based approaches could be customer segmented clusters and individual-level personalization ("Recommended for you", "Browsing history", "Your recently viewed items and featured recommendations", "Recommended for you based on a previous purchase"). On the other hand, item-based recommendations include trend analysis ("Best selling in this category"), time prediction ("Here is a newer version of this item), or item clusters ("Frequently bought together", "Related to items you've viewed", "Customers who bought this item also bought") **[1]**. Our assumption is that consumer-based recommendation systems may work just as well for smaller e-commerce platforms than the item-based approach, but with less computation power and cost.

[1]: quoted sentences in parenthesis are based on what is displayed on Amazon

Dataset and Features

The data we will use in this project to understand medium-sized e-commerce business is generously provided by the Brazilian e-commerce platform, [Olist](), obtained from [Kaggle](). The data is divided into multiple datasets for easier organization. Please refer to the appendix for the visual representation of the relationship between different datasets.

The datasets include 3,095 vendors selling 32,951 products ranging from 72 categories to 96,096 unique consumers from 4,119 cities in Brazil.

In terms of the consumers, we have features including their geographic location *(state, city, zip code from "customer")*, purchase history *(ordered item, price, shipping limit date from "items")* and purchase behavior *(payment method, number of installments from "order_payment" and review score from "reviews")* **[2]**.

In terms of the products, we have features including the product itself *(price, size, description length of the product, from "products")* and consumption of the product *(purchase time, review rating, location of consumer and seller, from "orders", "order_reviews", "customers", "sellers")*.

[2]: words in parenthesis refer to the column names, and words quoted refer to dataset name simplified, for example, "customer" refers to "olist_order_customer_dataset"

Explanation of Method Used

In terms of consumer-based recommendation, we plan to focus on customer segmentation. This approach is an unsupervised classification problem, which means that we could use k-means clustering and logistic regression. As detailed above, the data features we can use to fit our model include customers' geographic location, purchase history and purchase behavior.

When it comes to item-based recommendation, we can do trend analysis and sales prediction. Implementation of these may involve identifying surge in consumption of a particular product category and seeing overall product sales pattern, which is in essence a regression problem. This means that we could employ linear regression, decision tree-based boosting, Ridge and Lasso regression, neural network, and ensembling. Data features we could use to fit these models include data about the product itself and data regarding the product's consumption, as detailed in the previous section. K-nearest neighbors and SVM are also models we could use for both recommendation systems since they are both classification and regression models.

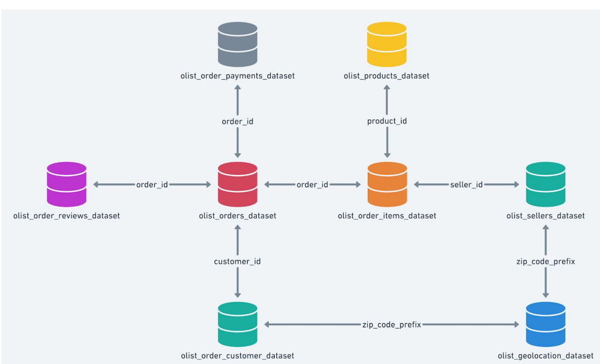
Eventually, we would like to evaluate the prediction results between user-to-user and item-to-item collaborative filtering. We would not predict which exact product, but instead would be predicting the product category and price range consumers would purchase. For evaluation, we want to see the percentage in which the user-to-user recommender system predicted similar products as item-to-item ones.

Limitations and Further Work

Fraud detection is also a serious issue in online shopping, and smaller e-commerce platforms are especially vulnerable to profit loss from fraudulent transactions. For implementation, the data patterns we could look at are larger than average orders of the same product, multiple shipping addresses, high payment sequences from different credit cards etc. These data features can be inferred from orders and order_payments datasets. Due to limitations on our dataset not being labelled, we may encounter difficulty in performing fraud detection, we however recognize that this problem is worth-considering in future implementations.

An additional set of data also provided by Olist and available on Kaggle is the [marketing funnel dataset]. Connecting this with the original set of data we were working on, we have a more comprehensive range of features in terms of sellers, which include their geographic location *(state, city, zip code from "sellers")*, the way they learned about Olist *(origin in "marketing_qualified_leads")*, the type of products they sell, their business size, and assessment of their behavior *(respectively business_segment, lead_type, lead_behaviour_profile in "closed_deal")*. Analyzing that dataset could be potentially useful in understanding how medium-sized e-commerce platforms can market themselves more efficiently. It is important to note that sellers are just as important as consumers to Olist, as both parties are customers of Olist.

Appendix



Works Cited Page

1. Hardesty, Larry. "The History of Amazon's Recommendation Algorithm." *Amazon Science*, Amazon Science, 20 Aug. 2020, www.amazon.science/the-history-of-amazons-recommendation-algorithm.
2. MacKenzie, Ian, et al. "How Retailers Can Keep up with Consumers." *McKinsey & Company*, McKinsey & Company, 14 Feb. 2018, www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers.
3. Olist. "Brazilian E-Commerce Public Dataset by Olist." *Kaggle*, 29 Nov. 2018, www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_sellers_dataset.csv.
4. Polatidis, Nikolaos and Georgiadis, Christos. "Recommender System: The Importance of Personalization in E-Business Environments".ResearchGate, 4(4):32-36, Oct. 2013, https://www.researchgate.net/publication/275998003_Recommender_Systems_The_Importance_of_Personalization_in_E-Business_Environments