Trends and Analysis of eCommerce Data

Francesca Barrios
CSPB 4502
University of Colorado Boulder
frba7936@colorado.edu

Laura Brown
CSPB 4502
University of Colorado Boulder
labr1398@colorado.edu

Seth Ely
CSPB 4502
University of Colorado Boulder
seel6470@colorado.edu

Darby Hansen
CSPB 4502
University of Colorado Boulder
daha2377@colorado.edu

Problem Statement/Motivation

In the dynamic landscape of online eCommerce retail, understanding consumer behavior and market trends is pivotal for success. Our group's objective is to leverage insights obtained from a dataset sourced from a multi-category online store, covering the period from October 2019 to April 2020. We aim to uncover valuable trends and patterns across various brands and categories, reflecting real-world marketing scenarios.

Here are some key questions our group seeks to answer with this analysis:

Which brands, items, or item categories exhibit the highest volume of sales during the specified period?

Are there discernible trends in sales performance across different brands and categories?

Among the features of brands, items, or item categories, which ones demonstrate the highest purchases per view ratio?

What factors contribute to the rate of purchases per view of these brands, items, or item categories?

Can brand type lead to customer retention (customer's purchase frequency after initial purchase)?

Can seasonal patterns be identified within the 7-month window of the dataset?

How do seasonal variations impact the sales and purchasing behaviors of consumers across individual brands and categories?

Can we find any trending patterns within our item sets outside of seasonal trend behavior – which can be used to predict popularity purchase windows?

Answering these questions can help us to understand consumer needs and behavior. For example, uncovering which brands or items consistently lead in sales can shed light on consumer preferences and brand loyalty within the online retail space.

Identifying brands or items with high purchases per view ratios can reveal effective marketing strategies or product attributes that drive purchases, providing actionable insights for optimizing sales and marketing.

Discovering seasonal trends can unveil fluctuations in consumer demand influenced by factors like holidays, weather changes, or cultural events, enabling retailers to plan strategies and offerings accordingly.

By looking into these factors, we anticipate discovering compelling insights that can inform marketing strategies, inventory management decisions, or overall business operations in the online eCommerce domain.

Literature Survey

The world of eCommerce is a dynamic landscape, constantly evolving and presenting new opportunities and challenges. It is also a foundational wealth of data which can present information to anticipate customer needs and drive profits and sales.

Most major eCommerce corporations are very protective of their data, considering it proprietary. The analysis of this data is what helps them to create recommendation algorithms, identify seasonal trends and sales patterns, or forecast consumer behavior with precision.

By harnessing the power of big data and advanced analytics, businesses can tailor their strategies to meet the evolving needs and preferences of their customers. Whether it's predicting the next big trend, optimizing pricing strategies, or personalizing the shopping experience, data-driven insights are the cornerstone of success in the digital marketplace.

One such comprehensive analysis is presented in "Global eCommerce Market Analysis & Trends" by Grand View Research. This study found that the global e-commerce market was worth around USD 9.09 trillion in 2019 and was expected to grow significantly, reaching USD 27.15 trillion by 2027, driven by factors like increasing small and medium enterprises and smartphone usage. The analysis indicated that Asia Pacific dominates this market, with key players including Alibaba, Amazon, and Walmart shaping its trajectory. (Grand View Research, 2020).

Since most companies keep their data private, the dataset we found does not include the name or organizational information where the data originates. However, the public availability of this dataset through the OpenCDP Project posted on Kaggle has allowed widespread analysis and insight collection by many individuals seeking to gain similar insights as our group.

One such project, by user Tshephisho Sefara, analyzed this data to create a predictive model using XGBoost to predict whether or not a user would add an item to their cart given the date/time as well as other features present in the data. (Sefara, 2020)

With his detailed analysis, it's no wonder his project has the most upvotes on the hosted site, Kaggle, and while we seek to find similar trends and insights as this user did, we believe we can find more compelling trends and prediction points than the 68% accuracy his final model was able to produce. We are interested in answering questions similar to this study, and while many other users have gathered similar insights, we do not seek to replicate their work but gather our own information and create more compelling insights with greater correlation.

Proposed Work

We will first need to clean our data through several means. The dataset includes a feature for the date and time of the event. We will need to parse the data into two features, one for the date and one for the time. This will enable us to answer our third question more easily by making the date more readily accessible and interpretable for the tools we will be using.

The second step in cleaning our data is to use inference-based processes to fill in the missing values in 'category_code' based on brand names. A quick google search of the brand name will tell us what industry the products fall under. This step is similar to those in the literature, allowing previous analyses of category purchases without skewing data.

Our next step is to remove duplicate entries. This is a replication of what has been done in previous studies to not skew the data.

The final step is to transform our data. We will do this by using different groupby functions to group the brands, items, and item categories together to create a more condensed table. This is similar to what has previously been done with this data and it makes it easier to analyze a data set of this size.

With the data cleaned as outlined above, we should be able to filter and visualize the data to show relevant insights for each of the three questions for each of the three features (item, item categories, and brand).

We will also create predictive models, one using K Nearest Neighbor to create clusters useful for making "for you" suggestions, and one using a Decision Tree Classifier to predict user behavior to determine if a given set of values would result in a purchase, or otherwise.

Data Set

The project dataset is hosted on Kaggle, found here: https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store.

Title: "ECommerce behavior data from multicategory store"

This dataset contains 7 months of behavior data from a multi category online store, ranging from October 2019 to April 2020. The data was originally collected by Open CDP project, and due to the size of the total dataset, is embedded as seven datasets, one for each month.

It contains 67,501,979 total values and a storage size of 16.45 Gb. Because of the size of the dataset, it is split into seven individual datasets, one for each month which we will have to join as part of our data cleaning. All of us plan to download the entirety of the dataset and as we collaborate on the project together.

The dataset contains nine features:

event_time: Continuous data type representing the date and time of the event.

event_type: Nominal data type indicating different types of events (view, purchase, remove from cart).

product_id: Nominal data type identifying unique products.

category_id: Nominal data type denoting product category identifiers.

category_code: Nominal data type describing the category for the product.

brand: Nominal data type specifying brand name of the product.

price: Continuous data type representing product prices.

user_id: Nominal data type identifying unique users.

user_session: Nominal type data representing user sessions.

We will use these features accordingly to gather the insights needed to answer our questions.

Evaluation Methods

Our main evaluation methods will be to determine the accuracy, relevance and interpretability of the insights gained.

To assess the accuracy of our findings this, we would like to create a predictive model that can use the features given in the dataset to predict user purchase versus view or item removal. Evaluating this model can be done by interpreting the accuracy, precision, and recall of the model after hyperparameter tuning. We will also create a clustering model using K-Nearest Neighbor to gather greater insight into user behavior as well as the ability to create custom “for you” lists that would be useful in a real-world marketing application. This will be assessed using using equivalent metrics.

Furthermore, we will evaluate the relevance of our findings by comparing them with the anticipated real-world needs of contemporary eCommerce retailers, ensuring alignment with the dynamic landscape of the current eCommerce market.

Finally, we will evaluate our analysis based on how interpretable it is for the intended audience. We will improve the interpretability of our findings by using charts, graphs, and visualizations. Additionally, we will use feature selection to identify the features that present the strongest correlation to create a compelling narrative from our insights.

Tools

The project will be written in Jupyter Notebook, Python v.3 and hosted on github with consequent tools described below:

Pandas Dataframes for any kind of initial data exploration such as sifting for null values, removing nulls and duplicates, and groupby, filtering and correlation computations during analysis.

Sklearn packages for any kind of clustering, training on trends and evaluation metrics.

TabPy (Tableau) import for graph visualizations of trend analysis, as well as bar graph and histograms for comparisons

Subsequent tools may be used and will be listed in the project progress report.

Milestones

By March 31st, our team will have our dataset downloaded, cleaned and functional; which will give two weeks to have this pre-processing step finished.

Most of our team’s analysis and questions we plan to “mine” will be completed in the following three weeks, and it is by that time we will give an update on our progress in the April 22nd report.

The remaining week and a half will be spent on finishing the evaluation, finalizing intuitive visualizations, and uploading the presentation video for final submission.

REFERENCES

[1] Kechinov, Michael, 2020. "eCommerce behavior data from multi category store." Retrieved from [https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store].

[2] Grand View Research, 2020. "Global eCommerce Market Analysis & Trends." Retrieved from [https://www.grandviewresearch.com/industry-analysis/e-commerce-market].

[3] Tshephisho, Sefara, 2020. "eCommerce behaviour using XGBoost." Retrieved from [https://www.kaggle.com/code/tshephisho/ecommerce-behaviour-using-xgboost/notebook#Know-your-Customers].