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

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

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

Which products are similar, and can we generate a suggested list of items for users based on this data?

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 is 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 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 is 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 deal with missing values. In performing some initial exploratory data analysis, we determined that dropping missing data as opposed to imputing them while retaining most of the dataset variability was feasible.

We will also create a filtered dataset that contains only purchased entries. Our initial exploratory data analysis also uncovered that less than 2% of all entries were purchases. Since many of our questions pertain specifically to purchases, this smaller dataset would make working with the dataset easier and filtering/cleaning the data would only need to be done once. Calculating the purchases per views for items will still require us to step through the entire dataset in chunks, however this can be done on the initial dataset offline.

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 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.

In answering which products are similar to others, we will make a clustering algorithm using K-Means to

We will also create a clustering model using the K-Means algorithm 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 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.

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.

1 Milestones Completed

The data cleaning ended up being much more difficult than anticipated due to the size of the dataset. Downloading the dataset took a considerable amount of time, and finding a way to crawl through the dataset was a hurdle as well. We created a sample spanning the entire time period for the dataset to gain some initial insight. Doing this, we were able to determine that less than 2% of the entries were purchases. Since most of our findings had to do with purchases specifically, it was beneficial to eliminate all other view and add/remove from cart entries. Doing so still allowed us to view almost 7 million entries.

The next step was to deal with missing values. We were able to determine that the percentage of missing values was negligible enough to drop rather than impute them and still maintain the variability and integrity of the dataset.

After these steps were taken, we had a compressed dataset of around 200Mb with 5.7 million entries to work with.

Because of these factors, the dataset was not fully cleaned until 4/19 with our initial goal of completing it by 3/31, however the valuable insights that were made by cleaning the dataset have been helpful in refocusing our efforts to answer the questions at hand.

Headway was also made on the K-Means algorithm to cluster products based on their brand and category code. Initially, we were going to create a predictive model that could look at user behavior and determine if the user would purchase a given product. Given that around 94% of all events are views, we have determined that this type of predictive model would not be very insightful (since guessing 'view' every time could result in an accuracy over 90%).

The K-Means clustering algorithm, however, will be useful in creating clusters of comparable products based on the category_code and brand of a given product.

To create this algorithm, we first needed to get a list of all unique category_code/brand pairs using the drop_duplicates method in Pandas dataframe object. In doing so, we were able to isolate only 148,602 unique item pairs. These item pairs are made up of only 139 unique category_code values and 4,081 brand values. We will be able to use these unique values to create a two-feature K-Means algorithm and create some nice visualizations using the category_code and brand values as x and y coordinates on a coordinate plane.

2 Milestones To-do

We still have much to accomplish for our project. Now that the data is clean, we can focus on analyzing the data to answer the questions we had laid out.

First, we need to finish grouping the different attributes, specifically by brands and item categories. This grouping will allow us to answer which brands, or item categories exhibit the highest volume of sales during the specified period, and which features of brands and item categories demonstrate the highest purchases per view ratio.

We also have started creating a K-means clustering algorithm. We will continue working to complete this algorithm to generate a suggested list of items for users.

We will finish by creating visualizations that will help with understanding our findings.

Results So Far

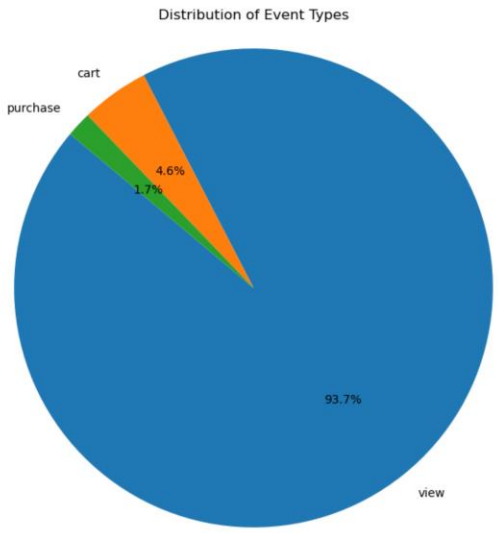
Some initial changes in the direction of our product occurred when we were able to look at the data more closely. Initially, we believed we would be able to determine insight on specific products, however since these products are represented by a product ID number that is specific to an unknown vendor, we determined that this feature would not prove to be as insightful.

Here is an image of the first five entries of our cleaned dataset for reference:

	event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
0	2019-10-01 00:02:14+00:00	purchase	1004856	2053013555631882655	electronics.smartphone	samsung	130.76	543272936	8187d148-3c41-46d4-b0c0-9c08cd9dc564
1	2019-10-01 00:04:37+00:00	purchase	1002532	2053013555631882655	electronics.smartphone	apple	642.69	551377651	3c80f0d6-e9ec-4181-8c5c-837a30be2d68
2	2019-10-01 00:07:07+00:00	purchase	13800054	2053013557418656265	furniture.bathroom.toilet	santeri	54.42	555332717	1dea3ee2-2ded-42e8-8e7a-4e2ad6ae942f
3	2019-10-01 00:09:26+00:00	purchase	4804055	2053013554658804075	electronics.audio.headphone	apple	189.91	524601178	2af9b570-0942-4dcd-8f25-4d84fba82553
4	2019-10-01 00:09:54+00:00	purchase	4804056	2053013554658804075	electronics.audio.headphone	apple	161.98	551377651	3c80f0d6-e9ec-4181-8c5c-837a30be2d68

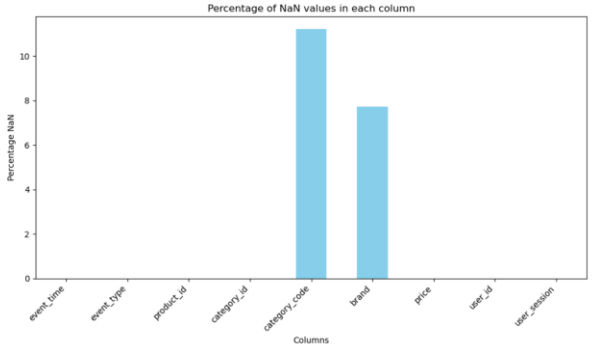
Initial exploratory data analysis was performed on a sample of 50,000 entries taken randomly from each month's data. This sample was convenient as we could gain insight without having to traverse the massive dataset of 67 million entries with a size of 20Gb compressed using gzip. Using this dataset will also come in handy for the next steps for our project.

This sample dataset revealed that a vast majority of entries are views with less than 2% of all entries being purchases. The following is a visual representation for the sample dataset.



This was fortunate, as most of our questions pertained to purchases and the views and cart actions were not as useful. Filtering out only these entries still generated a dataset with almost 7 million entries and a compressed size of only 256Mb.

To clean the data, we also looked at the proportion of missing values over all values for each column. We were able to see that "category_code" and "brand" were the two columns with the most missing values, with "user_session" only having 4.



Knowing this, we focused on what we should do about these values, whether to impute them or drop them. To better determine how many values were missing, we generated a table showing missing and valid values for these two categories:

	Valid category_code	NaN category_code	Total
Valid brand	5,707,926	611,751	6,319,677
NaN brand	372,838	156,309	529,147
Total	6,080,764	768,060	6,848,824

We determined that dropping all rows with missing values would give us a total of 5,707,926 entries which would account for around 83% of the existing data. Given these figures, we determined most of the remaining data would still retain most of the variability of the purchased data and decided to drop rather than impute rows with missing values.

Additionally, we were able to gain some initial insights into the brand and category_code features. These features will help create a K-Means clustering algorithm intended for suggesting items for users.

We discovered that there were only 148,602 unique category/brand pairs:

[81]:

	category_code	brand
0	electronics.smartphone	samsung
1	electronics.smartphone	apple
2	furniture.bathroom.toilet	santeri
3	electronics.audio.headphone	apple
6	appliances.environment.air_heater	oasis
...	...	...
5707822	apparel.shoes	panasonic
5707829	construction.tools.welding	rock
5707830	electronics.camera.photo	xiaomi
5707875	apparel.shoes.keds	peg-perego
5707909	sport.bicycle	remax

148602 rows × 2 columns

Additionally, we were able to see that only 139 categories were present in the purchased data with 4,081 different brands.

	category_code
0	accessories.bag
1	accessories.umbrella
2	accessories.wallet
3	apparel.belt
4	apparel.costume
...	...
134	sport.snowboard
135	sport.tennis
136	sport.trainer
137	stationery.cartrige
138	stationery.paper

139 rows × 1 columns

	brand
0	a-case
1	a-derma
2	a-elita
3	a-mega
4	a-toys
...	...
4076	zte
4077	zubr
4078	zuru
4079	zvezda
4080	zyxel

4081 rows × 1 columns

This information is not only useful for the clustering algorithm but will also be extremely useful in answering many of the questions we are setting out to answer.

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].