Myntra E-Commerce Dataset

Nishchay Nilabh

April 2024

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created for the purpose of collecting information about fashion items available on the Myntra platform. The specific task in mind was to gather data on item names, prices, ratings, and brand information. The goal was likely to analyze trends, pricing strategies, and customer preferences in the fashion market segment. This dataset fills the gap of easily accessible and structured data for fashion analysis on Myntra.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

I compiled this dataset independently by extracting data from Myntra through web scraping.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The dataset did not require any funding.

Any other comments?

By analyzing this dataset over time, user can gain valuable insights into market dynamics. Firstly, tracking product sales trends helps identify popular items and understand consumer preferences. Additionally, monitoring brand performance through user reviews allows businesses to measure customer satisfaction and brand loyalty. This information helps businesses to make better decisions, such as adjusting marketing strategies, optimizing product offerings, or even partnering with well-received brands.

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The instances in this dataset represent fashion items available on the Myntra platform. Each instance corresponds to a single item and includes information such as item name, brand, price, rating, and a search input indicating the category or type of item. Therefore, the dataset primarily consists of items/products.

How many instances are there in total (of each type, if appropriate)?

Each search query shows the top 250 results, and a total of 60 different search queries are present. Since all 60 categories don't have 250 products, there are a total of 13175 instances.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances scraped from Myntra's website. The larger set would be the entire range of items available on Myntra. It contains all the primary categories within the clothing domain, making it a near-accurate representation of the platform's offerings. Expanding the dataset to include additional categories would likely result in subdivisions of existing ones.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance in the dataset consists of the following data:

- **Name**: Name of the item.
- **Brand**: Brand of the item.
- **Price**: Price of the item.
- **Rating**: Rating of the item.
- **Search Input**: Search query or category used to scrape the data.

Is there a label or target associated with each instance? If so, please provide a description.

– –

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Some products did not have any rating, so the rating value was left as ”NA” for them.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

– –

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There is no immediate benchmark associated with the dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

- There may be errors in the product names due to misspellings.
- Ratings may have errors due to incorrect input or manipulation.
- Prices may vary due to discounts and promotions.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is self-contained and does not rely on external resources. All the data, including product names, ratings, prices, search queries, and brands, are gathered through web scraping from Myntra. There are no guarantees required for external resources' existence or consistency over time, and there are no restrictions, licenses, or fees associated with any external resources since none are utilized in this dataset.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No, the dataset does not contain any data that might be considered confidential. It consists of publicly available information related to products on Myntra, such as product names, ratings, prices, search queries, and brands.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No, the dataset does not contain offensive, insulting, threatening, or anxiety-inducing data. It primarily consists of product information such as names, ratings, prices, search queries, and brands, which are generally neutral in nature.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, it does not directly contain personal information about individuals.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

– –

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

– –

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

– –

Any other comments?

– –

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data associated with each instance was acquired through web scraping from the Myntra website. This involved extracting information such as product names, ratings, prices, search queries, and brands directly from the website's pages. Since the data was scraped from the website itself, it can be considered directly observable.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected using a software program specifically designed for web scraping. This program utilized web scraping libraries, such as Beautiful Soup and Selenium, to extract information from the HTML structure of Myntra's website. The program was configured to navigate through the website, locate relevant elements (such as product names, ratings, prices, etc.), and extract the desired data.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The sampling strategy used was deterministic. I selected 60 categories covering various types of clothes and then extracted the 250 most relevant products from each category.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The data collection process involved web scraping from Myntra. Since web scraping is typically automated, there weren't individuals directly involved in collecting the data. Therefore, no compensation was required.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The data was gathered over the span of a week, mainly due to daily scraping limits imposed by the Myntra website.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No, there were no ethical review processes conducted for this dataset.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No, the dataset does not directly relate to people. It contains information about products from Myntra, an online shopping platform.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

– –

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

– –

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

– –

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

– –

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

– –

Any other comments?

– –

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Yes, basic preprocessing was performed:

- **Stop-word removal**:
  Removed stop-words using nltk library from the product name column.
- **Remove special characters, convert to lowercase**:
  Converted names to lowercase and removed special characters using regex.

- **Missing value replacement**:
  Replaced NaN values in ratings column with 0.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
Yes, the "raw" data is also present in the dataset GitHub link.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
Yes, all the libraries needed for preprocessing are present in requirements.txt file.

Any other comments?
– –

Uses

Has the dataset been used for any tasks already? If so, please provide a description.
– –

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
– –

What (other) tasks could the dataset be used for?
The dataset can be used for the following tasks

- Brand image analysis
- Price optimization strategies
- Understanding user behavior and preferences

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
The dataset exhibits bias mirroring real-world occurrences, with a majority of 4-5 star reviews and certain brands having higher product count compared to others.

Are there tasks for which the dataset should not be used? If so, please provide a description.
Yes, the tasks are as follows

- **Generalizing beyond Myntra**: The dataset is specific to Myntra, so it may not be suitable for generalizations about other e-commerce platforms.
- **Making causal inferences**: The dataset may not provide enough information to establish causal relationships between variables, as it lacks contextual and external factors.
- **Assessing product quality**: Ratings alone may not sufficiently capture product quality, as they are subjective and may be influenced by various factors beyond product attributes.

And, just 13k sample points are not enough to make any robust conclusion. Maybe, with a lot more data points, we can see a stronger/deeper trend between brands,prices,etc.

Any other comments?
– –

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset is public and any third party is free to use it.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
It has been uploaded on the GitHub link (It does not have a DOI).

When will the dataset be distributed?
It has already been uploaded.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
– –

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
– –

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
– –

Any other comments?
– –

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset will not be maintained, however the code, to generate more data seamlessly, has been given given in the GitHub repository.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Email address : n.nishchay@iitg.ac.in

Is there an erratum? If so, please provide a link or other access point.

– –

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

– –

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

– –

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

– –

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Yes, users are free to generate more data. There will not be any verification/validation for that data.

Any other comments?

– –

Code Link

GitHub Repository: `https://github.com/Rockhopper130/mmdp_dataset`