Coolblue's Choice: good recommendations or secret revenue making?

Adapted from: Gebru, Morgenstern, Vecchione, Vaughan, Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.

Hidde Devenijns

Iris Heuten

Sanne Jansen

Iris van Walraven

Course: Online Data Collection & Managament

Tilburg University

Dr. Hannes Datta

0. Introduction

In this document we examine the differences in price of the laptops recommended by Coolblue (named Coolblue's choice) versus all the laptops on the website. By scraping the data of the Coolblue website we want to find out if the laptops in the Coolblue's choice category are higher in price compared to the average prices of all the laptops they offer.

1. Motivation

1.1 *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.*

Coolblue is a rapidly growing e-commerce company, with a wide range of products, e.g.: consumer electronics, household appliances, travel items, etc. In 2021 Coolblue achieved a turnover of 2.3 billion euros (Coolblue, 2022). Coolblue places a lot of emphasis on customer satisfaction, and ultimately on customer loyalty. It offers a lot of digital touchpoints, which can be seen as important factors of the customer experience. One example is the label 'Coolblue's choice', see the picture with the label below.

The dataset was created to see whether there was a difference in price provided by the laptops recommended by Coolblue itself (Coolblue's choice) compared to all the laptops that can be found on the Coolblue website. Because Coolblue is a commercial website whose aim is to make money, we wonder if the products they recommend are higher in price, because people may be more inclined to buy recommended products and they thus achieve more revenue. To look at this, we compare the average prices of all laptops Coolblue offers in the 'laptop' category, to the average price of a laptop in the 'Coolblue's Choice' category.

There has been chosen to scrape Coolblue instead of the Coolblue API, because the API of Coolblue costs money. Scraping Coolblue, however, is free. With the Coolblue API we were also not able to get access to all the data we wanted. The publicly available data is limited by the API.

The reason why the laptop page of Coolblue was chosen to scrape is because, compared to other categories, Coolblue offers a high assortment of laptops. This means we can get a clear overview of the average prices of all laptops compared to the one's in Coolblue's choice (see 3.3).

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

This dataset is created by team 10, consisting of Hidde Devenijns, Iris Heuten, Sanne Jansen, and Iris van Walraven, for the course Online Data Collection and Management (Spring) at Tilburg University. The instructor for this course is dr. Hannes Datta and is part of the Master's program Marketing Analytics.

1.3 *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.*

There was no funding made available for the creation of this dataset.

2. Composition

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

Two categories are examined: all laptops in the category 'all laptops' and the category 'Coolblue's choice' laptops. The category 'all laptops' consists of 996 different laptops, and the category 'Coolblue's choice'

* https://arxiv.org/abs/1803.09010

consists of 43 laptops. For each laptop, the laptop price, laptop title, and laptop URL was scraped.

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

The dataset should have contained 996 laptops in total. However, when scraping, we noticed we only got a return of 440 laptops. This is due to a mistake in the Coolblue website we found: the page range in the 'all laptops' category only goes to 20. At an average of 22 laptops per page, this means only (20 x 22 =) 440 laptops were available for scraping. We tried setting the page count higher, and also changing the URL to page 21 instead of 20 to test if there were actually more pages. Changing the number showed the last page, page 20 again. This means that people who visit the Coolblue website won't see all offered laptops in the 'all laptop' category due to the limit of the pages. 43 Of these laptops are also found in the subcategory called Coolblue's choice.

2.3 *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 has two categories, of which one is a subcategory of the other. This means that the items found in the sub-category can also be found in the main category. There is no 'larger' set in the laptop category because all available laptops are taken from the two different categories (all laptops vs. Coolblue's choice). There is, however, a larger set in total: Cooblue offers more than just laptops, and while our sample focuses on the laptop category, more samples can be taken from the website to compare against each other. For example, refrigerators could be scraped and compared to the Coolblue's choice refrigerators to draw a better conclusion about the differences in price for more categories. In the end, we want to draw a conclusion about the average prices of *category x* compared *to Coolblue's choice of category x* (in which x can translate to different product categories Cooblue offers, such as refrigerators.

2.4 *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 (laptop) consists of a title, which include the name, brand, version of the laptop. Moreover it consists of a price. This is the retail price at which the laptop is being sold to consumers. Furthermore it consists of a link to the laptop page on the website of Coolblue.

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

Each laptop is labeled with a title, a price and a URL. Moreover 43 of the 440 laptops do have the Coolblue's choice label.

2.6 *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.*

All of the data of all the laptops is available on the Coolblue website. Prices, title, and URLs are available for every product.

2.7 *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.*

Not applicable

2.8 *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 are no recommended data splits

2.9 *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. There are also no official archival versions of the complete dataset, and no restrictions associated.

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

The dataset does not contain data that might be considered confidential. All data that is found on the website is open. We also looked at the terms and conditions page on the website, but nothing was mentioned about scraping the website or being prohibited to do so.

2.11 *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 any offensive, insulting, threatening or anxiety-causing data.

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

The dataset does not relate to people.

2.13 *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.*

The dataset does not identify any subpopulations of people. It *does* identify a subpopulation of laptops: laptops that are also found in a different category recommended by the website itself, called 'Coolblue's choice'.

2.14 *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.*

It is not possible to identify individuals from the dataset, because the dataset is based on laptops and it's prices, titles and URLs.

2.15 *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.*

The dataset does not contain any data that might be considered sensitive.

3. Collection Process

3.1 *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by sub- jects (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.*

All the data that was scraped was directly observable on the Coolblue website. The Coolblue's choice label is, according to their website, "*The best product for you according to our customers and specialists. This product has the highest customer satisfaction because it is almost never returned by our customers. This way you know for sure you make a choice that makes you happy.*"

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

The data was collected by the webscraping program of Jupyter Notebook, which is written in the Python programming language. Within Jupyter Notebook, Selenium is used to scroll through the pages. By comparing the scraped data to the data that can be visibly seen on the Coolblue website, we made sure that the scraper contained all the right information.

3.3 *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 was to choose a product category that offered a lot of options, as well as different models and price categories. Due to this, we found laptops a good choice. Cooblue offers a wide range of laptops, regarding price and brand. Because the category contains a big number of laptops, the subcategory also contains a nice number of options (440 vs. 43).

If, for example, we had chosen the category of *Airfryers*, only 75 items were available for the sample compared to 13 in the *Coolblue's choice for Airfryers* category.

3.4 *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)?*

Only the students of our team, who are students of the Online Data Collection and Management course at Tilburg University are involved in the data collection. None of us were compensated for our work in terms of financial compensation. We were compensated by gaining new, in-depth knowledge of web scraping.

3.5 *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 time- frame in which the data associated with the instances was created.*

The data was collected at once. Since there was no need to analyze user pattern or behaviour, we were able to collect the data in one instance. Because we wanted to compare two categories with each other, there was no need to scrape the categories several times at different times because the Coolblue's choice does not differ quickly over time.

3.6 *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.*

There were no ethical review processes conducted

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

The dataset does not relate to people.

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

We obtained the dataset directly via the Coolblue website.

3.9 *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 other- wise reproduce, the exact language of the notification itself.*

No individuals were notified about the data collection.

3.10 *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.*

No individuals were asked for consent to scrape their data. Because only product information is scraped the

only individuals that could have been notified were the owners of the company website.

3.11 *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).*

No consent was obtained from the individuals.

3.12 *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.*

No analysis on the potential impact of the dataset and its use on data subjects were conducted.

4. Preprocessing, cleaning, labeling

4.1 *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 remain- der of the questions in this section.*

No preprocessing/cleaning/labeling of the data was done, because all the data was already labeled when it was scraped.

4.2 *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.*

No data besides the csv file of the scraped data was saved.

4.3 *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

The code used to preprocess the data is included in the 'Coolblue.py' file that is provided together with this documentation (

5. Uses

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

Apart from team 10 for the course Online Data Collection and Management (Spring) at Tilburg University, the dataset has not yet been used by other people or researchers.

5.2 *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.*

No, there is no repository that links to systems or papers that use the dataset.

5.3 *What (other) tasks could the dataset be used for?*

As already described in the motivation, the dataset could be used to find out whether the average price is different for laptops with the 'Coolblue's choice' label compared to all laptops. With this information researchers can determine what kind of products they should choose if they would also like to use such labels on their website.

5.4 *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?*

No, there is nothing about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses.

5.5 *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset should not be used to copy Coolblue's marketing strategy of having a Coolblue's Choice label. Moreover, the dataset should not be used by different companies to implement price changes, because they want to set a better price than Coolblue.

Sources:

Coolblue (2022). Coolblue achieves record turnover of 2.3 billion. Accessed at April 2022, from https://aboutcoolblue.com/news/coolblue-achieves-record-turnover-of-23-billion/