Ethical Data Science Analysis of Amazon E-Commerce Data

Data collected from https://www.kaggle.com/datasets/grikomsn/amazon-cell-phones-reviews

Data Cleaning & Preprocessing

Data cleaning and preparation are inevitable activities before the meaningful processing of a particular dataset. This step guarantees the truthful and sound nature of a dataset. It allows the researcher to cogitate on the ethical aspects of data processing, including the privacy and social justice-related impacts of data collection and formatting. In the case of a product dataset available on the product, the cleaning process would involve looking at missing values, any inconsistencies that would be present, and a decision to ensure that cleaning the product attributes makes it analytically valid and ethically defensible.

On technical grounds, various issues related to the dataset are typical of real-world data science. There are incomplete prices, a mismatch in the naming of brands (e.g., Motorola and motorola), or incomplete products. For example, some of the rows would contain an original price and nothing in the discounted price column, indicating at least partial failure of the scraping, or anomalies in how data is presented to the user by Amazon. Brand names in particular may need to be normalized, since inconsistencies may unnaturally make it appear that there is more diversity among products than there is. Cleaning is also necessary to eliminate distortion of brand-level performance analyses due to inconsistencies.

An additional preprocessing step is the processing of missing values or anomalous values. For example, an item priced at “0 Rs” might be a scrape or a sale item. One has to decide whether or not to leave such records out, estimate them or leave them as they are. Every decision has analytic implications as well as moral implications. Keep any data that may influence the result by hiding information, which will bias the results. On the other hand, by misrepresenting

the truth of the data, one risks using unwarranted values by imputation. Best practices encourage clear documentation of all the preprocessing steps so that the analysis can be reproducible and transparent- two principles that are highly rated in ethical data science.

On an ethical front, preprocessing also demands asking whether the dataset was built in a way that will not compromise the integrity of the consumer-generated information. No personal data, ratings and review count reflect the crowd's voice. Improper preprocessing, like randomly discarding low-rated products to make the analyses look better, might de facto disenfranchise some consumer experiences. Ethically, data scientists should make choices regarding cleaning that do not elide or skew marginalized voices in the data. For example, niche brands with underrepresented reviews can be shown as an outlier, again giving large companies an unfair advantage. Therefore, the ethical preprocessing has to have a tradeoff between statistical rigidity and fairness of representation.

Privacy issues also enter the preprocessing stage, especially involving potentially linkable data. The product IDs (asin) and links to product pages could be used to wheel on to live Amazon listings, although by the time get there there may be another listing or user with the same name or identifier. By being cross-referenced against other scraped data sources that contain customer reviews, there is a chance that it would be able to re-identify or trace individual patterns of reviewer activity. The right preprocessing here may be masking or restricting the exposure of such identifiers during the end-analysis, which may limit re-identification risk. Although maintaining product-level traceability can be useful to researchers, a trade-off on privacy should be recognized and considered carefully.

In a social justice sense, preprocessing decisions can guide the possibility of analysis reinforcing or disrupting extant injustices in online marketplaces. As an illustration, products

with few reviews can be disregarded as track records that are not statistically correct. However, such products can be supplied by smaller, minority-owned enterprises that do not have the advertising capabilities to create mass appeal. Not including them in preprocessing is a cause of digital marginalization. The alternative explanation of the socially just approach can be using weightings or stratified analysis so that underrepresented brands can stay visible in the data. This is evidence that the choices made in preprocessing technical databases directly determine whether or not marginalized players are represented in the developing insight.

In addition, the preprocessing emphasizes the asymmetry of power between vocabulary collectors, platforms, and consumers. As a powerful platform, Amazon in its product listing format emphasizes some types of data in favor of others, ratings and reviews, consumer feedback (not detailed feedback) and ethical sourcing information. By cleansing and preprocessing the dataset, researchers adopt those structural biases. Responsible data science must be aware of these in-built inequalities instead of assuming that the data set is neutral data representing reality.

Data cleaning/preprocessing can be beyond a technical issue and be related to ethical, privacy and social justice aspects. Each choice is critical to the data's narrative, including decisions on addressing missing prices, standardizing brand names, and which approach should be taken with low-rated products. The idea of realizing responsible preprocessing is to provide primary importance to transparency, secure privacy, and fair reflection of all the involved stakeholders, especially smaller brands and vulnerable consumers. Being the cornerstone of the project, this step will provide not only technically valid analyses but also refer to the ethical values of responsible data science.