Ethical Data Science Analysis of Amazon E-Commerce Data

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

Introduction & Dataset Overview

The popularity of data science has helped researchers, businesses, and decision-makers apply data-derived information in their decision-making processes across sectors. The data employed in this research contains data on the products existing on Amazon as of December 2019. In particular, it has 720 rows and 10 columns, including such attributes as the ID of the product (asin), brand, title, price, original price, rating, number of reviews, and carefully selected links. Although this dataset can perhaps be seen off-the-bat as a simple set of e-commerce product data, the potential use cases of this dataset go much further than just market analytics. A more philosophical approach to the data at hand can help us understand the morality of data collection, privacy of the consumers, consumer rights and social justice concerns regarding access to the digital market and fairness.

As a descriptive tool, the dataset is beneficial in providing information related to the behavior of consumers as well as their performance in terms of brands. The companies could utilize it to realize the trends in the market, reveal which brands are winning the most reviews and ratings, and adjust their prices so that they do not become unprofitable. Similarly, customers could learn how higher prices relate to better product ratings or how some brands perform better. But beyond the technical applications, there is a follow-up on what working with such data implies.

The ethical issues are front and center even in data gathering. As this data is scraped off of Amazon, whether it was gathered consentually by Amazon must be asked or whether it violates Amazon ToS. Web scraping tends to exist in a gray ethical space- it is not necessarily

illegal but can violate the understanding between users and websites. As ethical data scientists, it is also necessary to consider whether data collection procedures do not violate the rights of the platform and consumers who do not directly participate but share their reviews and ratings and are involved in the dataset. Although there are no personal identifiers, consumers have a specific digital footprint- e.g. the number of reviews they leave behind or the cumulative remarks they leave behind. The ethical use of this data would be to list the source of the information and ensure transparency regarding the uses required to perform an analysis.

Second, privacy consideration is a factor in the comprehension of the dataset. Although there are no names, addresses or direct identifiers of consumers, aggregate consumer behavior is recorded. This questions the so-called risk of re-identification; even data supposedly deanonymized at source may indirectly contain information about individual people when combined with other information. For example, suppose a reviewer routinely comments or leaves ratings about a niche brand or product. In that case, it is possible to use such information to cross-reference with other data and potentially identify that person. In this way, the dataset serves as a reminder that even in aggregated or anonymized data, there is a risk to privacy, and careful treatment of consumer information must also relate to information that consumers have created.

Third, it has important social justice repercussions in the data set. Reviews and ratings have an influential impact on Amazon in that, they determine how products should be ranked, recommended and eventually bought. Established brands with a reputation and a consumer base will automatically do more reviews, enhancing their market control. In the meantime, lesser-established or newer brands can have difficulty creating a profile, despite making quality products. This creates an inherent bias in terms of replicating inequalities in online retailing.

Regarding social justice, this points to how smaller sellers (many of whom might be minority-owned or in areas of the developing world without the resources to engage in mass-market campaigns) are unwittingly disenfranchised by functioning as a digital monopoly.

The ethical and social justice issues are also associated with pricing practices listed in the dataset. Numerous products are marked with a current price and the original price indicating that there are offers to be promoted. Nevertheless, studies based on online marketplaces have reported that the sellers sometimes exaggerate the original price so that the discount may seem higher than it actually is. Such a practice has the potential to misinform consumers, particularly those of lower income who are more susceptible to such marketing strategies that focus on the issue of affordability. By examining trends of pricing and discounting in the data set, researchers will be able to provide illumination of this fairness and transparency in prices or abuse of the consumer confidence model.

Lastly, this data set will allow us to consider how ethical data science can balance business value and ethical responsibility. Commercial analysis may be purely commercial, emphasizing only which products can bring the biggest revenue or which brands hold the market. Nevertheless, such morally-informed critique should also challenge the implications of such findings on consumer protection, brand diversity, and equitable accessibility to technology. A case in point is whether consumers get good quality products at the right prices, and whether the marketing strategies distort the perception that prices are a good value. Do big-end brands push out smaller brands, restraining consumer choice? These questions can place the project in the context of ethics, privacy, and social justice more broadly in data science.

Lastly, this data set is not just a set of product information. It is a conglomeration of consumer behaviour, businesses, and their moral obligation in the digital economy. As we move

forward in the other segments of this project, we will not solely perform statistical analysis on the data, but also question critically the ethical aspects, privacy and social justice implications, brought about by the data and the inferences drawn about it. Such a dual style will make the analysis technically viable and aligned with responsible and socially conscious data science concepts.