**Can deep learning be used to build a recommendation framework for Amazon shoppers?**

Overview. Retailers are increasingly using data to improve their marketing strategies. In recent years, the industry has expanded upon traditional machine learning practices, using artificial intelligence (AI)-based techniques to better the computational accuracy and speed of models. We seek to understand data in a way not yet previously measured – generating predictions of business engagement through an automated optimization process.

Deep learning operates as the foundation of technologies like image and speech recognition. Defined by advanced mathematical properties – including some that are hidden – humans collect, analyze, and interpret vast amounts of information in a short time frame. Yet, how deep learning models perform is often more about the data fed in to “teach” the model rather than the algorithm itself. There may be few problems when data are clean and concise, but more issues arise when data are messy, incomplete, or limited. In this project, we aim to answer the question: *can a deep learning algorithm improve predictions of customer and product alignment on Amazon, given the volatile nature of social data?*

As we gather info from online shoppers, errors are likely to ensue, for example, from human input or site bugs. In such cases, deep learning algorithms will fail to detect said things and in turn, potentially falsify a user or produce inaccurate predictions of product preference. Trusting customers to provide valid credentials can have a resounding effect on the model’s performance. If Amazon recommends unrelated content, the online customer experience will be negatively impacted.

Generally, if we fail to diagnose the potential shortcomings of a deep learning model, including those related to the social nature of their data or their social impacts in application, we’re putting ourselves and those involved at risk and potentially wasting valuable resources.

Data and Methodology. Our goal is to cluster Amazon shoppers based on spending habits by applying user text (natural language processing), product images (computer vision), and web activity as training data. We will use the K-means clustering algorithm and multinomial linear regression to determine variable importance within each group.

The data collection process needs to be controlled in that sampling bias is avoided. A deep learning model, as alluded to earlier, is sensitive to whatever inputs it is given. As such, having accurate, representative samples of the marketplace should help to limit partiality. The Amazon Cloud has a plethora of filters to address this matter. We will narrow our focus by looking at individual subsets of user data (with unique customer IDs), while also considering consensus ratings and reviews for a particular item. Covering a wide spectrum of Amazon buyers allows for greater model flexibility.

Social Challenges and Impacts. The analysis above raises various social issues and challenges.

*Challenge #1: Generation and Collection*

First, customer confidentiality is vital to data generation and collection. Having large volumes of sensitive user data constitutes a need for privacy. Public fears along with commercial bylaws will immediately limit the scope of our investigation. The extent to which we use the Amazon Cloud likely comes into question, as well. If we’re to directly mitigate worries about data security, implementing privacy-protecting techniques such as encrypted algorithms, data masking, and deidentification is paramount. These will conceal our raw data and prevent hackers from obtaining anything of substance.

*Challenge #2: Cleaning and Processing*

We must ensure the quality of our training data meets the standard for deep learning. With so many data points being taken into consideration, it’s easy to overlook the intricacies of certain features. Blunders such as incorrect product labeling, omitted price values, or user mistypes can easily throw off the model and lead to poor results. These reflect normal human error. The cost and time of collecting and labeling such data can be prohibitive, so it's imperative we review these mistakes before constructing the final algorithm. To combat this issue in a dataset of such magnitude, we can implement techniques of data imputation (replace missing values with column averages, mode of a categorical variable, etc.) or remove the blank entries entirely. Again, the social nature of our data presents a unique circumstance when it comes to cleaning and processing. Both the preference (user) data and pricing/product (Amazon) data are subject to human decision-making that requires careful consideration for how it should be cleaned or re-coded.

*Challenge #3: Interpretation and Application*

Deep learning models tend to be a black box. This means that we can’t always derive our outputs or meaning in a straightforward manner. The accuracy of our model may be relatively high from a predictive standpoint, but ambiguous as to which inputs or variables matter more/less. This is due to drastic information loss after iterations of the deep learning algorithm. Corporations like Amazon that operate in a regulated retail industry might have trouble explaining their results to an auditor or, similarly, to other periphery organizations that might rely upon Amazon learning models (e.g., advertisers, marketing firms, user groups, business researchers).

Conclusion. The utilization of deep learning to predict consumer/item trends on Amazon has the potential to improve the online shopping experience. Conversely, these algorithms are sensitive to input data and can produce erroneous results if the information is mishandled. There are social challenges and impacts to consider, such as customer confidentiality and the chance for model bias. It’s critical to address these matters first to effectively use deep learning in this context.