Dafni Tziakouri 12/04/2024

Ethical Data Science: 2nd Report - Discrimination

Exercise 1: ML Loop

A 2016 investigation found a racial disparity in Amazon's same-day delivery service: in many cities, Black residents were about half as likely as White residents to live in a ZIP code where the service was offered.

Consider a hypothetical machine learning system that uses data available to Amazon to determine which ZIP codes would be profitable targets for same-day delivery service. Since the rollout happens gradually, data from already-serviced neighborhoods helps inform the decision of where to expand.

Give three distinct reasons why racial disparities might arise in the predictions of such a system. Place each of these reasons in one (or more) of the stages of the machine learning loop: real world, measurement, learning, action, and feedback.

Three distinct reasons why racial disparities might arise in the predictions of the hypothetical machine learning system, placed in different stages of the ML Loop are the following:

- **Real world stage - Historical Bias in Data collection:**

  The historical data used to train the machine learning model may reflect existing racial disparities in access to services and resources. For example, if historically Black neighborhoods have been underserved by delivery services due to various socio-economic factors, the data would inherently contain biases reflecting these disparities. In this stage, the disparities are rooted in the real-world conditions that influence the collection of data. Historical biases in data collection lead to skewed representations of different demographic groups, including racial ones.

- **Measurement Stage - Proxy* Variables as Indicators:**

  The system might use proxy variables that inadvertently correlate with race to determine profitability. For instance, if the model relies on factors such as median income or housing prices to predict demand for same-day delivery, it may indirectly perpetuate racial disparities since these factors often correlate with race due to systemic inequalities. During the measurement stage, the system relies on observable features or proxies to make predictions. If these proxies are themselves influenced by racial disparities, the predictions will reflect and potentially exacerbate existing inequalities.

  *Proxy: a variable that is not in itself directly relevant, but that serves in place of an unobservable

- **Action Stage - Reinforcement of Existing Patterns:**

  The machine learning model's decisions on where to expand the same-day delivery service may inadvertently reinforce existing racial disparities. If the model predominantly selects ZIP codes with higher concentrations of White residents due to historical biases or skewed profitability metrics, it perpetuates the disparity by allocating resources unequally. At the action stage, the model translates predictions into decisions or actions. If these actions favor certain demographic groups over others due to biases in the data or algorithms, they perpetuate and potentially exacerbate existing disparities.

In summary, racial disparities in the predictions of the hypothetical machine learning system can arise from biases in data collection, the use of proxy variables that correlate with race, and the reinforcement of existing patterns through decision-making processes.

Exercise 2: The naive statistician

Google's image generation algorithm failure which resulted in the inability to create images of Caucasians is an example of an unintended consequence caused by how the algorithm was tuned.

Write short report (max 1 page) defining the problems you identify and a possible solution.

Hints:

• Which features should we consider? Is this selection generalizable to all countries?

• Demographics can be measured at different geographical levels. Which is the relationship between demographic measures and the content of the query?

• What about asking for a historical picture?

• How do we treat fact versus fiction (user intent, expected use of the image, etc.)? If someone wants to have a Black queen in Bridgerton, why not? If that same image is going to illustrate a news article, maybe not.

In this approach there many problems that occur. Firstly, it focuses solely on racial proportions, neglecting other aspects of diversity such as ethnicity, age, gender, and physical abilities. This oversimplified representation fails to capture the richness and complexity of human diversity, potentially leading to biased or stereotypical depictions. Secondly, demographic proportions vary not only between countries but also within regions and communities. Relying on global demographic data may overlook local nuances and result in inaccurate representations. Moreover, historical demographic data may not adequately reflect current societal compositions, especially in rapidly changing regions. Thirdly, the approach assumes that adding racial descriptors based on demographic estimates aligns with user preferences. However, it overlooks the diverse intentions behind image requests. For instance, a user seeking an image for historical accuracy may have different expectations than one creating fictional content. This can lead to inappropriate or misleading image outputs. Lastly, the approach lacks ethical considerations regarding the potential reinforcement of stereotypes or discriminatory practices. It assigns racial attributes based on probabilities risks perpetuating biases and marginalizing underrepresented groups. Additionally, it raises concerns about privacy and consent when inferring demographic information without explicit user consent.

To address these problems, a comprehensive approach to image generation should be adopted. It will be helpful to expand the scope of diversity beyond race to include factors such as ethnicity, age, gender, physical abilities, and cultural attributes. Also, with incorporate a diverse range of images in the training dataset will ensure equitable representation across various dimensions of human diversity. Furthermore, a good idea is to develop region-specific demographic models that account for cultural and geographical variations in population composition. By utilize real-time data sources and community feedback can contribute to continuously update demographic estimates and ensure accurate representations. In addition, implementing natural language processing techniques to interpret user queries and discern the underlying intent and context can be useful. For instance, tailor image outputs based on contextual cues, user preferences, and intended use cases, ensuring relevance, accuracy, and alignment with ethical guidelines. Lastly, it is necessary to establish clear ethical guidelines for image generation algorithms, emphasizing fairness, inclusivity, and respect for diversity. Also, by promote transparency in algorithmic decision-making processes, it enables users to understand how demographic attributes are inferred and utilized in image generation.

In conclusion, enhancing diversity representation in image generation algorithms requires a multifaceted approach that transcends simplistic statistical solutions. By embracing diversity as a fundamental principle and integrating contextual understanding, user intent, and ethical considerations into algorithmic design, it can follow the way towards more inclusive and equitable image generation systems that celebrate the richness and complexity of human diversity.