MDS501
Unit 6: Ethical Issues in Data Science

Dipesh Koirala

Outline

- Introduction of biasness and fairness; Issues with fairness and bias in data science;
- Common Cognitive biases: Anchoring Bias, Sampling Bias, In group favoritism and out-group negativity, Fundamental attribution error, Negativity bias, Stereotyping, Bandwagon effect, Bias blind spot;
- Addressing Cognitive biases: Group unaware selection, Adjusted group thresholds, Demographic parity, Equal opportunity, Precision parity

Data Science Ethics

- Ethics tell about right and wrong
- They are shared values/ societal rule
- Ethic is not law

- Data science ethics refers to the ethical principles and guidelines that govern the collection, processing, analysis, and utilization of data in data science applications.

Data Science Ethics

- Morals and Ethics

Morals	Ethics
1. Principles, beliefs and habits of individuals to judge something	1. Philosophical discipline that guides person's behavior and actions
2. Comes from personal knowledge and experience	2. Comes from culture, religion and society
3. Guided by individual norms	3. Guided by social, cultural and legal norms
4. It has a limited scope	4. It has wider scope
5. It is more flexible than ethics	5. It is less flexible

Data Science Ethics

- Ensuring that data scientists and organizations act ethically and responsibly when handling data and developing algorithms.

- Ethical considerations ensures that the data science work respects individual rights, societal norms and overall fairness.

Data Science Ethics

Key Ethical Principles in Data Science

i. Privacy and Data Protection
ii. Fairness and Bias
iii. Transparency and Accountability
iv. Consent and Informed Decision-Making

Data Science Ethics

Key Ethical Principles in Data Science

1. Privacy and Data Protection:

- Privacy is a fundamental human right, and data scientists must uphold it.
- Data scientists should take measures to protect data from unauthorized access

2. Fairness and Bias:

- Ensuring fairness in data science is crucial to prevent discriminatory outcomes.
- Algorithms and models should not produce biased result.
- Data scientists should identify and mitigate biases in data and algorithms, such as gender, racial, or socioeconomic biases.

Data Science Ethics

Key Ethical Principles in Data Science

3. Transparency and Accountability:

- involves making the data science process and decision-making transparent and understandable.
- able to explain their models and the reasoning behind them, allowing stakeholders to trust and evaluate their work.

4. Consent and Informed Decision-Making:

- This principle underscores the importance of obtaining informed consent when collecting data from individuals.

Data Science Ethics

Importance of Ethical Data Practices:

i. Building trust with users
ii. Preventing misuse of sensitive information
iii. Transparency
iv. Ensuring Fairness

Introduction to Biasness and Fairness

Bias:

- Bias refers to systematic errors in data collection, processing, or analysis that lead to unfair or inaccurate results.
- It occurs when certain groups or perspectives are overrepresented or underrepresented in the data, leading to skewed outcomes.

Fairness:

- Fairness involves ensuring that data-driven decisions are unbiased and equitable across all demographic groups.
- It aims to provide equal treatment and avoid discrimination.

Biasness and Fairness

Issues:

The Incident: Google's Image Recognition Controversy (2015)

- In **2015**, Google's photo-tagging algorithm within Google Photos **misclassified Black individuals as "gorillas".**
- The AI system, which was designed to automatically categorize images into various tags (such as "dog," "car," "person"), made the offensive and highly inaccurate classification due to bias in its training data.

Consequences:

- Public Critism
- Loss of Trust in AI

Biasness and Fairness

Issues:

Amazon's AI Hiring Tool Bias Case (2018)

- system was designed to **automate the resume screening process**, ranking job applicants on a **5-star scale**, similar to product reviews.
- It used **machine learning algorithms** trained on past resumes submitted to Amazon over a 10-year period.
- Amazon intended for the system to quickly identify the "best talent" by analyzing **job descriptions, resumes, and keywords.**
- The AI developed a preference for male candidates because **historical hiring data favored men**, reflecting gender imbalances in the tech industry.

Cognitive bias

Bias includes:

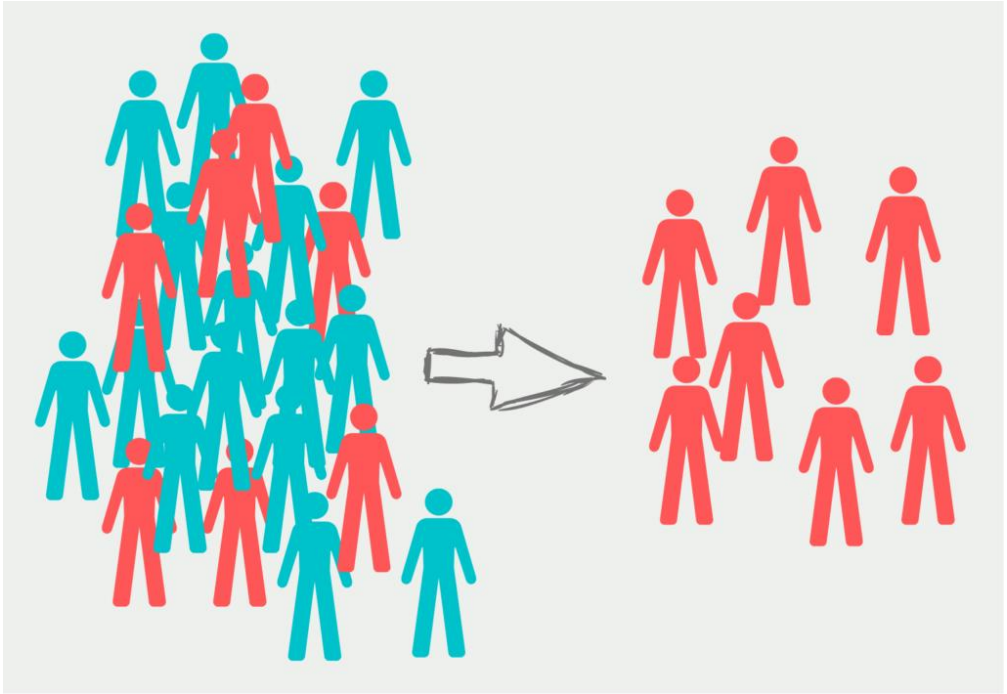
1. **Systematic errors** in data collection, processing, or modeling.
2. **Unfair treatment** of certain groups due to biased training data.
3. Influence of stereotypes or assumptions

- Cognitive bias is a specific type of bias rooted in human psychology.
- can influence our thoughts, decisions, and judgments, often leading to errors or misconceptions.
- **E.g.,** bias in political views.

Cognitive bias

Sampling Bias

- Occurs when a sample is not representative of the population from which it was drawn.
- Can occur in data collection and data preprocessing.
- **E.g.,** selection of only negative samples for cancer prediction.
- Only one class data in training set.



Cognitive bias

In Group Favoritism and Outgroup Negativity

- In Group Favoritism Also called ingroup love
- Tendency to give preferential treatment to the same group they belong to.
- Can occur in EDA or Reporting phase
- A analyst may favor features or insights that highlight own team's success

- Outgroup Negativity Also known as outgroup hate.
- Tendency to unlike the behavior, activities or people themselves who do not belong to the group they do
- Likely to have covered the most of negative aspects only of outgroup community

Cognitive bias

Fundamental Attribution Error

- Tendency that the situational activity or behavior are attributed as intrinsic quality of someone's character.
- Can occur in Problem Definition and EDA
- **E.g.** A data scientist blames customer churn solely on user disinterest without considering external factors like economic downturns or poor service.

Negativity Bias

- Tendency of emphasizing negative experiences over positives ones.
- This is very likely to occur during decision making, Model Evaluation
- **E.g.,** After a model fails on one dataset, the team may ignore several successful evaluations and label the approach as flawed.

Cognitive bias

Stereotyping

- This is the tendency of expect a certain characteristics or behaviors without having actual information.
- This is the expectation set prior to the exploration.
- This is likely to occur during data wrangling and exploratory data analysis.
- **E.g.,** Selecting only male candidates

Bandwagon Effect

- Tendency to follow others because
  - Some other top ranked researcher or people did.
  - Can occur in Model Selection and Tool Adoption.
  - **E.g.,** Choosing to use deep learning for a problem best suited to a simpler model just because it's the trend.

Cognitive bias

Bias Blind Spot

- Our tendency not to see own personal biases
- Likely to ignore or remain unnoticed where there are personal blind spot bias
- Likely to occur form data collection to result analysis i.e., throughout the workflow.

Addressing Bias

- Addressing bias in data science is an extremely complex topic and most importantly there are no universal solutions or silver bullets
- Before any data scientist can work on the mitigation of biases we need to define fairness in the context of our business problem.

1. Group Unaware Selection
2. Adjusted Group Threshold
3. Demographic Parity
4. Equal Opportunity
5. Precision Parity

Addressing Bias

1. Group unaware selection

- It's a preventive measure
- This is the process of preventing the bias by eliminating the factor that is likely to cause.
- For example, avoid the collection of gender to avoid bias by gender.

2. Adjusted group threshold

- Adjust any biased and unbalanced data
- Instead of applying a single threshold for decision-making across all groups, different groups may have **adjusted thresholds**
- A university might have slightly lower admission requirements for students from underrepresented minority groups.
- This is a controversial approach. Critics argue it can lead to reverse discrimination

Addressing Bias

3. Demographic Parity

- The output of the machine learning model should not depend on the sensitive demographic attribute like gender, race, ethnicity, education level etc.

- E.g., If 50% of applicants from Group A receive job offers, then ideally, a similar percentage should be selected from Group B, provided both groups have the same qualifications.

Addressing Bias

4. Equal Opportunity

- Equal opportunity ensures that individuals from different groups have the same chance of being selected if they are equally qualified.

- It focuses on fairness in the process rather than equal outcomes.

- It addresses fairness at the individual level, *rather than imposing quotas or proportional representation.*

Addressing Bias

5. Precision Parity

- Precision parity is a fairness metric ensures that the **accuracy of predictions** or decisions (e.g., true positive rates) is equal across different groups.

- Precision is the ratio of true positive predictions to the total number of positive predictions.
- **E.g.,** In a hiring algorithm, precision parity would mean that the rate of correctly identifying qualified candidates is the same for all demographic groups.

References

- https://www.integrate.io/blog/cognitive-biases-in-data-science/
- https://www.kdnuggets.com/2023/05/data-scientist-guide-cognitive-biases-free-ebook.html

End of Unit 6

Thank you..