Bias and Fairness in Healthcare AI

(Identifying and mitigating inherent AI biases to ensure equitable outcomes)

Presenter

Ankit Virmani
(Senior Cloud Data Architect- ML and Data platforms)

Introductions and Agenda

Ankit Virmani

Senior Cloud Data Architect: Data and ML Platforms

Experience: Big data and ML Operations across healthcare, finance and consumer industries. Worked at Google, Amazon and Deloitte.

Passions: Data engineering, horror movies and dogs!

Agenda

- ❑ Introduction to Bias in the overall MLOps lifecycle
- ❑ Deep Dive into bias examples and mitigations (both technical and functional)
- ❑ Examples of identifying bias
- ❑ Functional Framework/Putting structure to the thought
- ❑ Recap and Questions

Introduction to Bias in the overall MLOps lifecycle

Some Terms used throughout

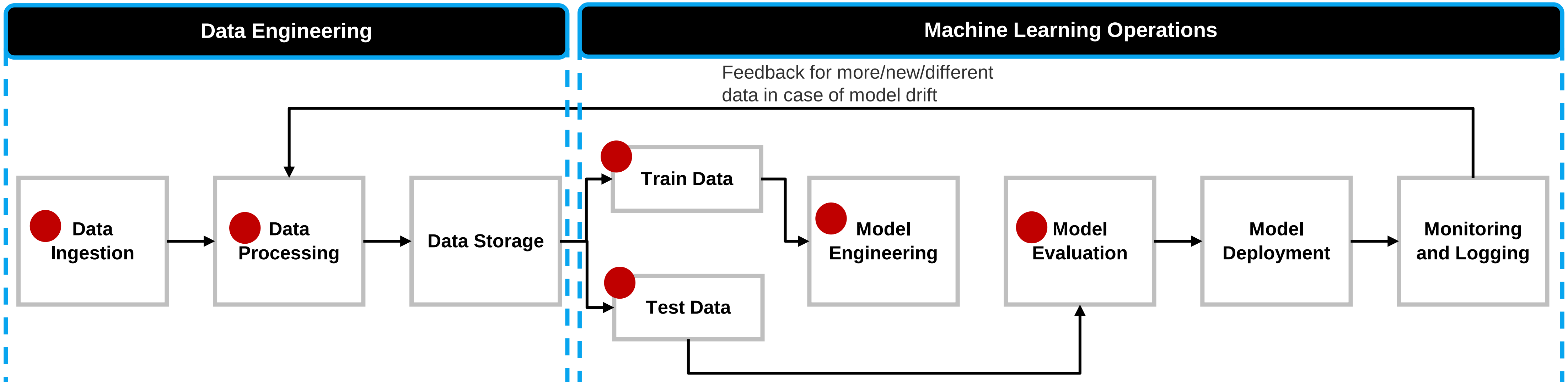
Term	Description
MLOps	End to end lifecycle of a Machine Learning project- all the way from data engineering to model development, model deployment, model usage and model production
Data Engineering	Practice that involves designing, building, and maintaining the systems and infrastructure required for collecting, storing, processing, and analyzing data.
Bias	Inclination or prejudice (intentional or unintentional) for or against certain data in the MLOps cycle
Feature Engineering	Part of machine learning cycle where the features for a given machine learning model are identified based on business requirements and data analysis. This could also include weighting the features to identify the ones that have the most causation between features and labels in supervised learning
Explainable AI	The methodology in which the model predictions our output can be explained by the model inputs
Synthetic Data	The simulated data that's artificially manufactured rather than generated by real-world events. It's created algorithmically and is used as a stand-in for test data sets of production or operational data, to validate mathematical models and to train machine learning (ML) models

Different origins/stages of bias and potential mitigation approaches

Bias manifests itself throughout the MLOps cycle as seen in the slide- all the way from data sourcing to model creation. It is not always easy to identify that bias has occurred unless there are significant business/technical impacts of the model running in production, but some mitigation strategies can potentially help reduce the impact

Stage	Root Cause/Origin of Bias	Metrics to calculate Bias!
Data Ingestion **Data Processing** **Data Split (test/train)**	• Inherent bias in the source data • Misalignment between features and the business context • Gathering data from population belonging to a certain group • Missing/unexpected data/feature values • Non-random split of data between test and train	1 Difference in Population Size 2 Difference in Positive Proportions in True Labels
Feature Engineering **Model Training** **Model Validation** **Model Evaluation**	• Features influencing the learning of ML models, producing imbalanced/non-ideal models • Incorrect Imputation techniques • Encoding categorical variables • Biased annotation for supervised models • Sampling Bias, algorithmic selection, feedback loops • Evaluating model performance on not-previously established evaluation criteria	3 Accuracy Difference. 4 Difference in Positive Proportions in Predicted Labels. 5 Recall Difference 6 Specificity Difference

Potential Mitigation Strategies for Bias



Strategies

Data Ingestion
- Have a solid understanding of data
- Deep inspection, timeframe monitoring
- Data SME with ML team- vital

Data Processing
- Identify suitable imputation
- Understand the outlier groups before removing them

Train Data / Test Data
- Simple random Sampling
- Stratified random sampling
- Cross validation

Model Engineering
- Appropriate model selection followed by cross validation
- Sensitive feature removal (gender, ethnicity etc.)

Model Evaluation
- Establish performance indicators before model creation
- Cross-validation

Technologies

Data Engineering
Open Source: DataProfiler, OPenRefine, Apache Nifi, Apache Spark, Apache Beam, CDAP
Commercial: Public cloud offerings for managed Spark, Beam, other tools like Talend, Trifacta etc

Train/Test Data and Model Engineering
Open Source: IBM Watson OpenScale, Google's What-if Tool, Aequitas, Fairlearn, AI Fairness 360

Model Evaluation to Monitoring and Logging
Open Source: Tensorflow Fairness Indicators, Scikit-learn, MLFlow
Commerical: GCP Vertex AI etc

Summary of real life use cases in Healthcare AI with practical solutions

Context	Deep Dive	Final Conclusions/Solutions if achieved!
National healthcare underserved black patients over white patients An algorithm assigned black patients the same risk as white patients, even when the black patients were sicker	ML algorithm had used healthcare cost as a proxy for overall health ~ less healthcare costs for black patients == black patients are healthier-- NOT true!	The manufacturer of the algorithm fixed the ML model by removing healthcare cost as an indicator for overall health.
AI can, advertently or inadvertently detect a person's race from medical images	model was learning to do this by matching known health outcomes with racial information	AI can accurately predict self-reported race, even from corrupted, cropped, and noised medical images, often when clinical experts cannot, creates an enormous risk for all model deployments in medical imaging. Regulators and lawmakers should weigh this new research when deploying specific AI tools throughout the healthcare system—tools that could inadvertently perpetuate biases inherent in the data—by enacting requirements for explicit testing and monitoring of model development and performance on demographic subgroups
Black patients are more likely than white patients to have their pain dismissed and untreated	model was able to predict pain better than the human diagnosis and despite imaging not showing the expected level of disease severity	The study hypothesized that underserved patients with disabling pain but without severe radiographic disease could be less likely to receive surgical treatments and more likely to be offered non-specific therapies for pain. This approach could lead to overuse of pharmacological remedies\
When GPT4 was asked to diagnose sore throat, chest pain, and breathing difficulties, it ranked possible diagnoses differently depending on a potential patient's gender or race.	GPT-4, it made the correct diagnosis (mono) 100% when the patient was white, but only 86% of the time for Black men, 73% for Hispanic men, and 74% for Asian men.	GPT-4 and similar AI models will need to be improved significantly before they can be applied to patient care management. There will also likely need to be safeguards built into the technology before it's used for clinical decision making.
VBAC algorithm was designed to help healthcare providers assess the likelihood of safely giving birth through natural delivery. The algorithm considers many things, such as the woman's age, the reason for the previous C-section, and how long ago it happened	VBAC was skewed because it that predicted Black/African American and Hispanic/Latino women were less likely to have a successful vaginal birth after a C-section than non-Hispanic White women	After years of work by researchers, advocates, and clinicians, changes were made to the algorithm. The new version of the algorithm no longer considers race or ethnicity when predicting the risk of complications from VBAC. This means that doctors can make decisions based on more accurate and impartial information that works for all women, providing more equitable care regardless of race or ethnicity

Putting it all together

Know your data!	Be aware of your data- functionally. Have a data SME who understands the business context of the data. Utilize their knowledge during data engineering, feature engineering phases
Remove sensitive groups from training	Avoid sensitive groups like gender, socioeconomic position, ethnic traits, regional preferences, and so on if they interfere with the interpretation of the data and model
Train regularly	Update training data on a regular interval to ensure that the ML model can absorb and learn new data patterns. Use triggers like model drift to retrain the models with new training data
Use the right models	AI and ML teams should be aware of when to apply which ML algorithm. Select the most appropriate machine learning model for the data at hand
Conduct Bias checks before deploying in production and regularly after that	Conduct bias testing as a part of your machine learning project lifecycle to discover bias at an early stage before it causes real-world system damage