Machine Learning and Artificial Intelligence: Ethics & Fairness

Richard Wanjohi, Ph.D

Data Scientist

H&R Block

Jan, 2023

My Descriptive Stats

- Data Scientist: H& R Block
  - Previously: ECCO Select/USDA, Commerce Bank, Westar Energy, Walmart, Inc
- Adjunct Professor
  - Data Science & Analytics: University of Wisconsin, SDSU, UNLV et. al (2021-)
  - Big Data Analytics: Rockhurst University (2018-2021)
- Ph.D & MS in Statistics, University of Arkansas, Fayetteville
- Interest: Deep Learning, LLM

Introduction

- We are living in data-driven world. Data, data every where!

Introduction

- We are living in data-driven world. Data, data every where!
- Data is collected via:
    - Experiments or trials, Observations, Polls, Surveys, Studies etc
- Other sources of data:
    - Social Media
    - Transactional data
    - Service data: Weather channels, Stock Exchange etc
    - Machine data: from industrial equipment, sensors, and web logs

Introduction

- We are living in data-driven world. Data, data every where!
- Data is collected via:
  - Experiments or trials, Observations, Polls, Surveys, Studies etc
- Other sources of data:
  - Social Media
  - Transactional data
  - Service data: Weather channels, Stock Exchange etc
  - Machine data: from industrial equipment, sensors, and web logs
- Use data to make informed decisions:

  Data $\Rightarrow$ Information $\Rightarrow$ Knowledge $\Rightarrow$ Insight $\Rightarrow$ Wisdom

Introduction

- Machine learning (ML), a subfield of Artificial Intelligence (AI), is a field of computer science concerned with programs that learn
- Learn:

Introduction

- Machine learning (ML), a subfield of Artificial Intelligence (AI), is a field of computer science concerned with programs that learn
- Learn:
  - Use of historical **Data** (with inputs and outputs) & Mathematical functions (Algorithms)
    - **Data** + Algorithms $\Longrightarrow$ **Model**

Introduction

- Machine learning (ML), a subfield of Artificial Intelligence (AI), is a field of computer science concerned with programs that learn
- Learn:
  - Use of historical **Data** (with inputs and outputs) & Mathematical functions (Algorithms)
    - **Data** + Algorithms $\Longrightarrow$ **Model**
  - Predict outputs given new and unseen inputs, in future.
    - **Model** + new inputs $\Longrightarrow$ Predictions (new outputs)

Introduction

- Machine learning (ML), a subfield of Artificial Intelligence (AI), is a field of computer science concerned with programs that learn
- Learn:
  - Use of historical **Data** (with inputs and outputs) & Mathematical functions (Algorithms)
    - **Data** + Algorithms $\Longrightarrow$ **Model**
  - Predict outputs given new and unseen inputs, in future.
    - **Model** + new inputs $\Longrightarrow$ Predictions (new outputs)
- the reliability of the information obtained and models built largely depends on **quality** and **quantity** of data.

Introduction

- ML/AI impacts **everything**: from Social Media to Agriculture to Healthcare to Education to Retail to Insurance to Banking & Finance

Introduction

- ML/AI impacts **everything**: from Social Media to Agriculture to Healthcare to Education to Retail to Insurance to Banking & Finance
- While ML/AI system has the potential to improve lives, it can also be a source of harm.

Introduction

- ML/AI impacts **everything**: from Social Media to Agriculture to Healthcare to Education to Retail to Insurance to Banking & Finance
- While ML/AI system has the potential to improve lives, it can also be a source of harm.
- ML applications have discriminated against individuals on the basis of race, sex, religion, socioeconomic status, and other categories.

General ML procedure

- Formulate problem statement. Clear objective/goal. Translating a real-life problem into a machine learning problem.
- Obtain and prep credible data (quality and quantity)
- Determine what methodology or approach to use, why, when and how (skillset).
    - Build, validate and test your model(s)
    - Communicate results
    - Deploy models to production
- Toolset to use. From ingestion to **data prep** to **EDA** to **inference** to **modeling** to **deployment**.
  How do the tools integrate with existing systems?

Where do problems arise?

Some Notable Use-Cases

AI/ML is used, among many$^{\infty}$ others, to:

- Determine who gets loan or not
- Identify who is a risk borrower, default on a loan, insurance risk,... etc?
- determine which candidate get accepted to a university or gets a job

Some Notable Use-Cases

AI/ML is used, among many$^{\infty}$ others, to:

- Determine who gets loan or not
- Identify who is a risk borrower, default on a loan, insurance risk,... etc?
- determine which candidate get accepted to a university or gets a job
- Identify email is spam or not
- In cancer diagnostics - MRI scans

Some Notable Use-Cases

AI/ML is used, among many$^{\infty}$ others, to:

- Determine who gets loan or not
- Identify who is a risk borrower, default on a loan, insurance risk,... etc?
- determine which candidate get accepted to a university or gets a job
- Identify email is spam or not
- In cancer diagnostics - MRI scans
- Identifying hot spots for crimes (Arrest data vs Crime data??)
- the likelihood that a convicted criminal will relapse into criminal behavior (Parole decision)
- Facial recognition for possible criminals etc

Source of harm: Data

ML models only “see” the world through the data used for training

- Data bias is complex. It is a type of error in which certain elements of a dataset are more heavily weighted and/or represented than others.
- The data reflect the biases of the systems and people who generated it.

Source of harm: Data

ML models only "see" the world through the data used for training

- Data bias is complex. It is a type of error in which certain elements of a dataset are more heavily weighted and/or represented than others.
- The data reflect the biases of the systems and people who generated it.

Some important questions (Quality of data):

(a) Where did the data come from?

(b) How and why was data collected?

(c) Is there some incentive to distort or spin results to support some self- serving position?

Human bias in Data

1. Selection (Sample/Representation) bias: Occurs when a dataset does not reflect the realities of the environment in which a model will run
   - Coverage bias: Data is not selected in a representative fashion.
   - Non-response bias : Participation gaps in the data-collection process.
   - Sampling bias: Proper randomization is not used during data collection

Human bias in Data

1. Selection (Sample/Representation) bias: Occurs when a dataset does not reflect the realities of the environment in which a model will run
    - Coverage bias: Data is not selected in a representative fashion.
    - Non-response bias : Participation gaps in the data-collection process.
    - Sampling bias: Proper randomization is not used during data collection

Example

- AI facial recognition system trained primarily on images of white men will have low accuracy levels for women and people of different ethnicities.

Human bias in Data

2. Prejudice bias: as a result of cultural influences or stereotypes. (Appearances, social class, status, gender)

   Example

   - using data about medical professionals that includes only female nurses and male doctors
   - Recruitment screening system that discriminated against women (Amazon 2015)

Human bias in Data

4. Prejudice bias: as a result of cultural influences or stereotypes. (Appearances, social class, status, gender)

   Example

   - using data about medical professionals that includes only female nurses and male doctors
   - Recruitment screening system that discriminated against women (Amazon 2015)

5. Reporting bias: When the frequency of events, properties, and/or outcomes captured in a data set does not accurately reflect their real-world frequency.

Bias in Deployment

6. Observer (confirmation/implicit) bias: the effect of seeing what you expect to see or want to see in data
   Effects: choosing source data or model results that align with currently held beliefs or hypotheses.

Bias in Deployment

8. Observer (confirmation/implicit) bias: the effect of seeing what you expect to see or want to see in data
   Effects: choosing source data or model results that align with currently held beliefs or hypotheses.

9. Deployment bias: When the problem the model is intended to solve is different from the way it is actually used.

   Example:
   AI model developed to predict cost of care, and would distinguish high and low cost patients.
   Model is used to predict healthcare needs instead of healthcare cost...
   (Result: black patient needing to be as twice as sick as white patient to benefit for same healthcare program)

Bias in Deployment

10. Exclusion bias: as a result of excluding some feature(s) from our dataset usually under the umbrella of cleaning our data

Bias in Deployment

10. Exclusion bias: as a result of excluding some feature(s) from our dataset usually under the umbrella of cleaning our data
11. Measurement bias:
    - Oversimplification of more complex construct
    - Measurement & accuracy varies across groups/locations etc

Example:
Predict the likelihood that a defendant will re-offend?
Minority communities are more highly policed & models often include proxy variables such as "arrest" to measure 'crime' or 'riskiness'

Impact in the Society

Biased AI systems can:

- Unfairly allocate opportunities, resources or information
- Infringe on civil liberties
- Pose a detriment to the safety of individuals
- Fail to provide the same quality of service to some people as others
- Negatively impact a person's wellbeing such as by being derogatory or offensive.

Mitigation

Data and data prep:

- Training data for machine learning projects **has to be representative of the real world**
- Where possible, combine inputs from multiple sources to ensure data diversity.
- Enlist the help of someone with domain expertise to review your collected and/or annotated data.

Mitigation

Data and data prep:

- Training data for machine learning projects **has to be representative of the real world**
- Where possible, combine inputs from multiple sources to ensure data diversity.
- Enlist the help of someone with domain expertise to review your collected and/or annotated data.
- Create a gold standard for your data labeling.
- Make clear guidelines for data labeling expectations so data labelers are consistent.
- Use multi-pass annotation for any project where data accuracy may be prone to bias.

Mitigation

Statistical/demographic parity:

- Covering all the cases you expect your model to be exposed to.
- Apply Equal opportunity fairness: Ensures that the proportion of correctly selected is the same across groups
- Check that the model has equal accuracy for each group.
- Apply "Fairness through unawareness" (Group unaware) : Removes all group membership information from the dataset.

Thank You

Richard.Wanjohi

@WanjohiRG

richard.wanjohi@hrblock.com

rgwanjohi@gmail.com