ॐ : OM BHAGAWATHEY VASUDEVAYA NAMAH: ॐ

Fairness

Machine learning Model evaluation is not only just calculating loss metrics → Before it goes to production, it is critical to audit training data & evaluate predictions for bias.

Prediction Bias

→ This value indicating how far apart the average of predictions is average from the labels in the dataset

ML-Bias

→ stereotyping, prejudice or favoritism towards some things, people or groups over others. These bias can affect collection & interpretation of data

1. automation bias
2. confirmation bias.
3. Experimenters bias
4. group attribution bias
5. Implicit bias
6. In-group bias

⑦ Out-group homogeneity bias.

② - Systematic error introduced by a sampling or reporting procedure.

① Coverage bias, ② Non-response bias, ③ participation bias
④ reporting bias, ⑤ Sampling bias, ⑥ Selection bias.

Machine learning models are not inherently objective. ML practitioners train models by feeding them a dataset of training examples, and human involvement in the formation and curation of this data can make a model's predictions to bias.

① Reporting bias →

occurs when the frequency of events, properties or outcomes captured in a dataset does not accurately reflect the real world frequency. This bias can arise because people tend to focus on documenting that are unusual or especially memorable than ordinary are not recorded.

Eg:

Sentiment Analysis model is trained to predict whether book reviews are positive or negative based on corpus

⇒ Majority of reviews in training dataset reflect extreme opinions (either they love or hated)

② Historical Bias

Eg: A city housing dataset from the 1960's contains home-price data that reflects discriminatory lending practices in effect during that decade.

③ Automation bias

is tendency to favour results generated by automated systems over those generated by non-automated systems. irrespective of error rates of each.

Eg: Spreadsheet manipulation

④ Selection Bias

If dataset doesn't reflect real-world distribution like coverage bias, non-response bias, sampling bias.

→ Coverage bias

If data is not selected in a representative fashion.

Eg: model trained to predict future sales of a new product based on phone surveys conducted with a sample of customers who bought the product and left the people/customers who opt to buy competing product were not surveyed.

→ Non response bias

also known participation bias occurs if data ends by being unrepresentative due to participation gaps in data-collection process.

→ Sampling bias

⇒ occurs if proper randomization is not used during data collection.

* Group attribution bias

is a tendency to generalize what is true of individuals to the entire group to which they belong.

⇒ In group bias

is a preference for members of your own group you also belong or for characteristics that you also share.

⇒ Outgroup homogeneity bias is a tendency to stereotype individual members of a group to which you do not belong to their characteristics as more uniform.

* Implicit bias

Occurs when assumptions are made on one's own model of thing and personnal experience that don't necessarily apply more generally.

Eg: head shake gesture "yes" and some place is "no"

* Confirmation bias.

Individuals consciously process data in ways that affirm preexisting beliefs & hypothesis

Eg: Arguments of drugs.

* Experimenter's bias

Builder keep training a model until it produces a result that aligns with their original hypothesis

Identifying Bias

Keep issues of fairness in mind & audit for potential sources of bias.

Missing feature values

If dataset is missing one or more feature values for large number of examples.

Indicator that certain key characteristics of your dataset are under-represented.

⇒ Unexpected feature bias

"It introduces bias when in dataset you see unprecedented feature values that stand out as especially uncharacteristic or unusual.

⇒ Data Skew

Remove drug - adoptability, it is not sufficient to look overall accuracy.

Mitigating Bias

Once the source of bias is identified in training data.

There are two strategies

(1) Augmenting the training data.

(2) Augmenting the models loss function

(1) Augmenting the training data

The most straight forward way to address the problem is often the collect additional data.

(2) Augmenting the Models optimization function

"log loss" to penalize incorrect model predictions.

Log loss doesn't take subgroup membership in consideration

Instead we use optimization functions designed to penalize error in fairness-aware fashion.

Tensorflow Model Remediation library provides utilities for applying two different bias-mitigation techniques during Model training

=> MinDiff → it aims to balance the error for two different slices of data. By adding a penalty for differences in the prediction distributions for the two groups

=> Counterfactual Logit Pairing aims to ensure that changing a sensitive attribute of a given example doesn't alter the models prediction for that example.

Evaluating for Fairness

Aggregate performance Metrics

① ~~accuracy~~ precision
② recall
③ accuracy.

} precision will not not capture issues

Demographic parity

Majority Group & Minority group

The best way to check the demographic parity fairness or any benefit

of demographic parity both majority and minority are selected in same proportion

Equality of opportunity

Counterfactual Fairness

When we dont have datas for demographic parity or Equality of opportunity group. Here there would be candiates with marked as eligible or not eligible those candiates are picked by model then we can say model is Counterfactual Fairness

Benefits
→ It is used to evaluate the predictions of Model

Demographic parity & Equality of opportunity area the groups in aggregate

Majority → Demographic like place.

Minority → Demographic like place where they are less considered.