Bias Analysis & Mitigation Techniques

FAIRNESS - STRUCTURED DATA-

Sample Bias-

Definition: Sampling bias occurs when some members of a population are systematically more likely to be selected in a sample than others. Sampling bias is a threat to external validity – it limits the generalizability of your findings to a broader group of people.

Input: Tabular data should be used for the sample bias test which contains continuous and categorical features. The desired features should be selected to perform this test.

Output: Sample bias can be reduced or eliminated by examining the domain of each feature and making sure we have balanced evenly distributed data covering all of it. It shows the confidence in favor of sample bias for selected columns.

Example: For example, A group of people wherein gender data distribution is not even. The data has a 70% of male and 30% of female distribution which led to biased in the sample.

Pre-Prediction Label Bias-

Definition: The pre-processing label bias is to reduce bias by manipulating the training data prior to training the algorithm.

Input: Categorical and continuous features should be selected along with target feature to perform the hypothesis test to check the p-value.

Output: It shows the p-value for the test which reflects the degree of data compatibility with the null hypothesis.

Example: A group of people having a class income level > 50K wherein out of the total, 80% of Male and 20% of females have a class level more than > 50K despite the fact that they have an even distribution of 50%-50% in the dataset. This led to label bias in the data.

Exclusion Bias- Exclusion bias results from the exclusion of particular groups from the sample. We delete some feature(s) thinking that they are irrelevant to our labels/outputs based on pre-existing beliefs.

During data pre-processing, features that are considered irrelevant end up being removed. This can consist of removing null values, outliers, or other extraneous data points. The removal process may lead to exclusion bias and the removed features may end up being underrepresented when the data is applied to a real-world problem and resulting in the loss of the true accuracy of the data collected.

Input: Tabular data should have the combination of categorical and numerical features.

Output: perform the correlation test basis on numerical and categorical features such as Kendall, Eta, Pearson, etc. to see how features are correlated with the target variable.

Example: For example, imagine you have a dataset of customer sales in America and Canada. 98% of the customers are from America, so you choose to delete the location data thinking it is irrelevant. However, this means the model will not pick up on the fact that Canadian customers spend two times more.

Post-Predictions Label bias-

Definition- Post prediction bias analysis can help identify biases that might have emanated from biases in the data, or from biases introduced by the classification and prediction algorithms. These analyses take into consideration the data, including the labels, and the predictions of a model. It is defined as the rate of how many unseen points a model labeled accurately over the total number of observations.

Input: The protected and sensitive feature to be selected along with the target feature. The accuracy metric needs to be selected to generate the report accordingly.

Output: This test is performed to check the accuracy metrics such as Precision and Recall, TPR, and FPR. This assesses performance by analyzing predicted labels or by comparing the predictions with the observed target values in the data with respect to groups with different attributes.

Example: A group of people wherein out of the total, 20% of female income class predicted correctly whereas 30% of male income class predicted incorrectly. This led to post prediction label bias in the sample.

Definition of RAI Toolkit Capabilities like Fairness, Transparency and Robustness and Other Common terminology

Fairness:

Fairness is an area of machine learning that studies how to ensure that biases in the data and model inaccuracies do not lead to models that treat individuals unfavorably on the basis of characteristics such as e.g., race, gender, disabilities, and sexual or political orientation. A fairness metric checks whether a classifier produces the same result for one individual as it does for another individual who is identical to the first, except with respect to one or more sensitive attributes. Fairness in machine learning is:

- Transparent about the potential bias inherent in data
- For all people regardless of demographics
- Always avoids engaging in unfair competition

Transparency:

Transparency is about how much it is possible to understand a system's inner workings “in theory”. It can also mean a way of providing explanations of algorithmic models and decisions that are comprehensible to the user. Transparency in machine learning system ensures that:

- AI is understandable/explainable by humans
- Decisions are accurate, reliable, reproducible
- Outcomes are interpretable by common sense
- System is auditable and actions are traceable

Increasing transparency can benefit stakeholders in different ways:

- End users: can better understand why certain results are being generated.
- Developers: can more easily debug, tune, and optimize machine learning models.
- Project managers: can better comprehend how an otherwise technical project works

Robustness:

The robustness is the property that characterizes how effective your algorithm is while being tested on the new independent (but similar) dataset. In the other words, the robust algorithm is the one, the testing error which is close to the training error. In other words, robustness is the idea that a model's prediction is stable to small variations in the input, hopefully, because its prediction is based on reliable abstractions of the real task that mirror how a human would perform the task. That is, small invisible noise should not flip the prediction. Robustness ensures that system should be

- Resilient to attack or abusive/malicious use
- Secure by design
- Human privacy is protected
- Integrity ensured, controlled access

Glossary (Common Terminology used in Toolkit):

- **Protected Feature:** Personal characteristic that is protected by law, like age, sex, gender identity, race, or disability. These are known as 'protected characteristics' or 'protected attributes' or 'protected features'. Protected attributes are often presented as categorical features that need to be encoded before feeding them into a machine learning algorithm.
- **Sensitive Feature:** Sensitive features or important features are those which can largely impact the outcome of the machine learning model/system.
- **Continuous Feature:** A continuous feature can take all possible values in range like height, weight & time, etc.
- **Categorical Feature:** A categorical feature can only take a specific value amongst the set of all possible values like gender, age group & education level, etc.
- **Correlation:** Correlation explains how one or more variables are related to each other. These variables can be input data features that have been used to forecast our target variable. Correlation is a statistical technique that determines how one variable moves/changes in relation to the other variable. It has a value ranging from -1 to +1.

Reference links:

https://towardsdatascience.com/5-types-of-bias-how-to-eliminate-them-in-your-machine-learning-project-75959af9d3a0

https://developers.google.com/machine-learning/glossary#selection_bias

https://towardsdatascience.com/identifying-and-correcting-label-bias-in-machine-learning-ed177d30349e

https://blog.taus.net/9-types-of-data-bias-in-machine-learning