**Structural Bias in Machine Learning Model**

A bias can be encountered in any Machine Learning model, this basically may influence the outcome generated by the machine learning model [1]. Biases in models must be removed since they provide us with an impartial outcome. A bias-free machine learning model cannot exist as it requires a certain amount of bias to model the data and to analyze predictions. However, the aim is to reduce these biases occurring in our model. In the case of training models for Criminal Recidivism, there can be various bias which may be encountered. To explicate this, some of the biases may be against certain races, where people of a particular race may be impartially evaluated for gauging the recidivism score. Another such bias may occur in the gender of any offender, where a person of a particular gender may be more biased/likely to be categorized as a recidivist.

There can likely be the following –

* Sample Bias – This is an inevitable type of bias arising due to the randomness and irregularities in the data samples. This occurs during the training phase of the model. This can be an intuitive way of the model to pick up the more frequently occurring values of a particular attribute. [2]

For instance, in the dataset used to train our model, the number of cases, where the tuple has value “Single” in the “Marital Status” column is much higher than other values. This inevitably creates an occurrence of sample bias, where the model would be inclined to categorize most of the “single” values to higher recidivism.

* Algorithmic Bias – This is the type of bias which is introduced by the algorithmic phase of the machine learning model and is not present due to anomalies in data samples. Data Scientists strive to attain a perfect balance between high variance and high bias.

Here, in our model, the Random Forest Classifier introduces bias, when training and testing the given dataset. This is an inherent property of the algorithm.

* Measurement Bias – It occurs when we select the features we wish to incorporate in the model. It may be the way these features/attributes are used in the machine learning phase. [2]

A striking example of this, is the use of this for criminal recidivism, where any priorly committed crimes or crimes committed by relatives/friends may also taint the outcome of the model.

Thus, attributes like “Agency Type”, “Custody Status”, “Legal Status”, etc. may create a measurement bias for criminals in evaluating their score text.

* Prejudice Bias – This type of bias is mainly due to the influence of social stereotypes and orthodox opinions. It mainly occurs on training data, where prejudice against a particular culture, gender, ethnicity or any such factor may make the model biased while generating the output. If the algorithm is exposed to a more even handed data distribution, then the statistical relationship between such potentially prejudiced attributes can be avoided.

In the case of Criminal Recidivism, the attributes like “Ethnic Code”, “Sex Code”, “Marital Status”, etc. can contribute towards prejudice bias of the model.

References:

[1] Gordon, Diana & desJardins, Marie. (1995). Evaluation and Selection of Biases in Machine Learning. Machine Learning. 20. 5-22. 10.1007/BF00993472.

[2] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. “A Survey on Bias and Fairness in Machine Learning.” ArXiv abs/1908.09635 (2019): n. pag.