Enhancing Fairness in AI Decision-Making Systems at BluePermanente

Alex Lavaee[1]

[1]EquiTech Corp.

1 Introduction

As a technical consulting firm specializing in the fintech and insurtech sectors, we understand the critical importance of fairness in AI decision-making systems. The case of Jerry vs. BluePermanente highlights significant concerns regarding potential biases within your risk-prediction algorithm which aims to predict a higher risk score for patients with greater health care needs. Our recommendations aim to address these issues by enhancing the fairness and transparency of your AI systems. We believe that by implementing these measures, BluePermanente can not only improve patient trust but also ensure compliance with regulatory standards and ethical principles.

2 Data Collection

The first step in creating a model that reduces bias and maximizes fairness is ensuring the dataset used for training the risk-prediction model is diverse and representative of the entire population. Include underrepresented groups in the training data to improve the model's accuracy and fairness for these groups. Address any existing gaps in the dataset related to socio-economic factors that might correlate with race, affecting health care needs and costs.

It is also important to understand the practical implications or reality of the situation where many groups do not have satisfactory representation in off the shelf datasets due to historical racism. However, it is important to acknowledge and contribute to the betterment of existing datasets by enlisting more inclusive data collection strategies for the future.

3 Bias Evaluation and Detection Measures

Before creating a risk-prediction algorithm we must introduce methods to evaluate the performance of statistical models. We suggest employing standard statistical measures of fairness, or group fairness criteria.

3.1 Group Fairness

Since we are trying to (indirectly) predict a risk score by predicting an insurance cost for an insurance provider, we can represent the true insurance cost we have in our dataset with the

continuous variable Y. We can then represent known characteristics as X. A summary of X is shown in Table 1. A subset of X includes certain characteristics that should be considered while evaluating the fairness of machine learning models. Examples of such characteristics include but are not limited to: gender, race, and age. We will refer to these characteristics as "sensitive characteristics," and denote one of them by A. Finally, we denote the predictions of our model by R.

Category	**Description**
Comorbidities	Indicators for specific illnesses at time $T-1$
Demographics	Basic demographic information
Costs	Costs claimed from the patients' insurance payer over time $T-1$
Biomarkers/Medication	Results from various tests and biomarkers at time $T-1$

Table 1: Categorization of Features Available for Risk-Prediction Model

3.1.1 Independence

For a model to satisfy independence the prediction rate of our model should be equal for people belonging to different subgroups in A or in other words, (R, A) should satisfy statistical independence.

$$P(R = r \mid A = a) = P(R = r \mid A = b) \quad \forall r \in R \quad \forall a, b \in A$$

Typically we should set a reasonable threshold for this metric using a predefined slack ϵ, where $\epsilon > 0$. Intuitively, this tells us that sensitive characteristics should not be good predictors of Y:

$$P(R = r \mid A = a) \geq P(R = r \mid A = b) - \epsilon \quad \forall r \in R \quad \forall a, b \in A$$

However, we argue that the independence metric does NOT apply to this scenario since we want our model to be sensitive to changes in race, gender, and age since they play an important role in representing the medical conditions of individuals.

3.1.2 Separation

For a model to satisfy separation we require the models' predictions R to be statistically independent of A given Y:

$$P(R = r \mid Y = q, A = a) = P(R = r \mid Y = q, A = b) \quad \forall r \in R \quad q \in Y \quad \forall a, b \in A$$

Intuitively, this means that the evaluation performance of the model is not affected for different target values Y (in our case cost) across the different subgroups for a sensitive characteristic A.

We believe the separation metric adequately captures the desire to have a well-performant model on different sensitive characteristics because it ensures the performance of the model is consistent across subgroups of a sensitive characteristic.

3.1.3 Sufficiency

For a model to satisfy sufficiency we require the true values Y to be statistically independent of A given the models predictions R:

$$P(Y = q \mid R = r, A = a) = P(Y = q \mid R = r, A = b) \quad \forall q \in Y \quad r \in R \quad \forall a, b \in A$$

Intuitively this is the opposite of separation. If the model predicts a value r given a characteristic a, it should be as confident as when the model predicts the same value r given a characteristic b.

We believe the sufficiency metric adequately captures the desire to have a model where the confidence does not change when it predicts a value across subgroups of a sensitive characteristic.

4 Model Development

4.1 Fairness-aware Model Training

Implement fairness-aware machine learning techniques during the model training phase. Techniques such as re-weighting training instances, optimizing for a fairness metric directly, or using adversarial learning can help reduce bias. One approach is to modify the loss function to penalize predictions that result in unfair outcomes.

5 Model Deployment

5.1 Regular Monitoring and Updating

Implement a continuous monitoring system to regularly assess the fairness and performance of the algorithm. Update the model as necessary to adapt to changes in population demographics and health care practices.

6 Conclusion

The fairness of AI decision-making systems, especially in critical sectors like healthcare, is paramount. By taking proactive steps to ensure fairness, transparency, and accountability, BluePermanente can lead by example in the responsible use of AI. Implementing the recommended measures will not only address the concerns raised in the case of Jerry vs. BluePermanente but also enhance the overall equity and effectiveness of your healthcare services.

7 Sources

- Fairness in Machine Learning: A Survey
- Fairness in Machine Learning
- Hidden Technical Debts for Fair Machine Learning in Financial Services

- Fairness and Machine Learning