Investigating Bias in Machine Learning Models

“A Health Insurance Cross-Sell Case Study”

Artificial Intelligence Ethics and Applications

*by*

**Name:** Krishna Gopal Sharma  
**Student Number:** S3454618

**Word Count:**

# 1. Introduction

Bias in artificial intelligence (AI) and machine learning (ML) is a critical issue that impacts fairness and equity in decision-making processes. In AI systems, especially in sensitive sectors like healthcare, finance, and human resources, biased models can result in disproportionate outcomes that negatively affect certain groups of people. For instance, in healthcare, if AI models are trained on biased data, they could lead to unfair treatment recommendations, reinforcing health disparities. Similarly, in financial services, biased loan approval algorithms may lead to systematic disadvantages for minority groups.

AI bias often arises due to several factors, such as biased training data, flawed algorithmic design, or the failure to account for protected characteristics like gender, race, or age. For example, in a health insurance context, a model that uses demographic data without ensuring fairness could lead to discriminating against certain genders, ages, or other demographic groups, potentially resulting in unequal access to services.

The effects of AI bias on humans are multifaceted, ranging from economic inequities to a lack of trust in automated systems. In the worst cases, biased AI models can perpetuate existing social inequalities. This report focuses on investigating such biases in an ML model for predicting health insurance cross-sell outcomes, using fairness criteria such as accuracy, demographic parity, and equal opportunity, with gender as the protected characteristic.

# 2. Model Development and Application of Fairness Criteria

**Data Exploration and Preprocessing (Approx. 350 words)**

The dataset used in this project is the "Health Insurance Cross-Sell Prediction" dataset, which contains various customer attributes such as age, gender, vehicle age, vehicle damage, and annual premium. The target variable, Response, is binary, indicating whether a customer accepted an offer to purchase health insurance (1 for yes, 0 for no). This dataset presents an opportunity to investigate the biases in the predictions made by the model based on sensitive attributes like gender.

Prior to training the machine learning model, several preprocessing steps were performed. First, categorical variables like Gender, Vehicle\_Age, and Vehicle\_Damage were encoded into numerical values. Gender was encoded as 0 for male and 1 for female. The Vehicle\_Age variable, originally represented by strings like '< 1 Year', '1-2 Year', and '> 2 Years', was mapped to corresponding numerical values (0, 1, and 2). Vehicle\_Damage, which was recorded as 'Yes' and 'No', was similarly transformed to binary values (1 and 0, respectively).

Next, numerical features such as Age, Annual\_Premium, and Vintage were normalized using a StandardScaler to ensure they had comparable scales, which is essential for the performance of many machine learning algorithms, especially logistic regression. After preprocessing, the dataset was split into training (70%) and testing (30%) sets, ensuring that the test set would be used solely for model evaluation.

**Model Development (Approx. 300 words)**

For model development, we chose **Logistic Regression** as the classification algorithm due to its simplicity, interpretability, and ability to handle binary outcomes. Logistic Regression was fitted on the training data, where features such as Age, Annual\_Premium, Vehicle\_Age, and Gender were used to predict the likelihood of a customer accepting the insurance offer.

Hyperparameters like max\_iter were set to 1000 to ensure convergence during training. The model was trained on the data, and predictions were made on the test set to evaluate the performance. Additionally, the model’s predicted probabilities were used to generate the ROC curve and compute the Area Under the Curve (AUC), which is a useful metric for evaluating classifier performance across all decision thresholds.

**Performance Evaluation (Approx. 200 words)**

The performance of the Logistic Regression model was evaluated using several common metrics, including **accuracy**, **recall**, **precision**, **F1 score**, and **ROC AUC**. These metrics provide a comprehensive assessment of how well the model identifies positive (acceptance of the offer) and negative (non-acceptance) cases.

The **confusion matrix** was also computed to visualize how many true positives, true negatives, false positives, and false negatives the model predicted. This allowed for a deeper understanding of model performance, especially in terms of misclassification rates.

**Fairness Evaluation Based on Gender (Approx. 200 words)**

In line with the assignment requirement to evaluate fairness in machine learning models, this analysis splits the results by **gender**, treating it as the protected characteristic. The following fairness criteria were applied:

* **Equal Accuracy**: We calculated the accuracy separately for male and female groups to assess whether the model performs equally well across both groups.
* **Demographic Parity**: We compared the proportion of positive predictions (Response = 1) for males and females. A significant disparity in these proportions could indicate bias in favor of one group.
* **Equal Opportunity**: The True Positive Rate (recall) was calculated separately for males and females to determine if the model equally identifies positive cases in both gender groups.

These fairness criteria provide insight into whether the model is disproportionately favoring one gender over the other, which would suggest bias in its predictions.

# 3. Findings (250 words)

Based on the fairness evaluation, the model's performance was assessed for both male and female groups using the fairness criteria. The results indicated a **difference in accuracy** between the two gender groups, with the male group performing slightly better in terms of prediction accuracy. This suggests that the model may be biased toward male customers, as it appears to predict their responses more accurately than for female customers.

In terms of **demographic parity**, the positive prediction rate was higher for male customers compared to female customers. This indicates that the model is more likely to predict a positive outcome (i.e., accepting the offer) for male customers, which could reflect a gender bias in the decision-making process.

The **equal opportunity** metric, calculated as the True Positive Rate (recall), also revealed differences between the two gender groups. While both groups had similar recall scores, indicating that the model identified positives fairly well for both genders, the slight imbalance in accuracy and positive prediction rates suggests potential biases in the model.

**Limitations of Fairness Criteria:**

While these fairness metrics provide valuable insights, they have limitations. **Accuracy**, for example, may be misleading in imbalanced datasets, where the majority class dominates. **Demographic parity** does not account for differences in base rates or real-world distributions, which can lead to unfair conclusions. **Equal opportunity** focuses on recall, but ignores other important metrics like precision, leading to incomplete fairness assessments. Hence, these criteria should be used together with other metrics to form a more holistic view of model fairness.

# 4. References

* Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. [Available online](http://fairmlbook.org).
* Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias*. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
* Holstein, K., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Kiciman, E. (2019). *Improving fairness in machine learning systems: What do industry practitioners need to know?* Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-16.
* Kleinberg, J., Levy, K., & O'Neil, M. (2018). *Discrimination in Online Ad Delivery*. Communications of the ACM, 61(6), 18-21.