Data & Code Overview

In the second update of our model, we have carried forward the insurance data, which includes key demographic, vehicle, and claim history features. In this step, we will understand and articulate the model's decision logic and the possible counterfactual outcomes for individuals. We will define the metrics that we will be excluding from the model to avoid a possible biased/discriminant outcome.

Overview of Schools of Ethics in Discrimination

The ethical frameworks governing discrimination in AI shape how we approach fairness in our insurance model. The Negative Rights/Libertarian approach emphasizes freedom from discriminatory treatment, focusing primarily on removing protected attributes to prevent direct harm. This view aligns with what philosophers like Nozick would consider the "minimal state" of AI ethics — ensuring systems don't actively discriminate[1]. In contrast, the Equal Opportunities/Positive Rights perspective recognizes that simply removing protected features is insufficient when historical inequalities remain embedded in seemingly neutral data. This framework calls for proactive measures to ensure all individuals have equitable outcomes, regardless of background.

Utilitarianism offers a different lens, suggesting we balance overall model performance against potential harms to specific groups. This approach might accept some disparate impact if the aggregate benefit is substantial, though this raises questions about whose utility is prioritized. Finally, the Dehumanization/Dignity perspective (rooted in Kantian ethics) focuses on respecting human autonomy in AI systems through transparency and explainability, rejecting the treatment of people as mere data points[2].

These ethical frameworks inform our standards for feature selection — from simple blindness (removing protected attributes) to more sophisticated approaches like blindness with proxies (removing correlated features), relevance assessment (both statistical and conceptual), and conscientiousness (ongoing monitoring for fairness).

Our Standard

In light of the understanding of the schools of ethics in discrimination, we have decided to exclude a few features from our analysis and model to ensure that there is no bias in the outcomes of the models. We have outlined our rationale for selection of these features based on these principles.

For our model, we initially shortlisted the top 10 most important features of which we decided to drop the features of home value and income. We decided to exclude home value on the grounds of blindness with proxies as the value of one's property can be an indicator of their socio-economic status, leading to potential unfair bias. On the same grounds, we will be excluding years on the job (YOJ) from our model for future working, as it can be a proxy to reflect stability. Income has been removed to uphold equal opportunity, ensuring financial status does not influence predictions in any way.

On the legal rules around protected features[3], we decided to not consider gender for our model, to ensure no discrimination against a certain gender. On the grounds of blindness, we have also excluded the following features from the model: number of children, whether the driver is a single parent, type of car, education and occupation. All these features may reflect the familial/socio-economic status of the applicant. Car type also becomes conceptually irrelevant since we have the value of the car included in the model (bluebook). We have also eliminated marital status to ensure equal opportunity.

We have, however, also decided to retain the features of age and whether the driver's license has been revoked. Although age is a protected feature, in our case it can be justified as a business necessity. It is a legally relevant factor for risk assessment, as drivers under the age of 25 are statistically more likely to get into accidents[4], which is why companies like Hertz charge them higher insurance[5]. As for number of licenses revoked, although it is a small dataset for whom the value is positive, it is highly relevant to risk assessment, ensuring the model reflects real-life driving concerns.

Potential Objections & Future Solutions

Our approach to preventing discrimination through feature selection raises several important concerns. Removing high-importance features like income and car type may significantly reduce model accuracy. As Barocas and Selbst note, "removing information may sometimes harm the very groups we seek to protect"[6]. To address this, we propose a balanced approach: using ethically neutral features for baseline predictions while incorporating sensitive features only with strict fairness constraints to maintain prediction quality without introducing bias.

The transparency provided by XAI methods, while beneficial, risks oversimplifying complex decisions and potentially revealing sensitive information. We recommend implementing tiered explanations tailored to different stakeholders, with all explanations undergoing fairness audits to prevent revealing protected characteristics.

Removing multiple features (home value, sex, income, car type, etc.) may create an underspecified model that fails to capture legitimate risk factors. This could lead to underfitting and unfair treatment of low-risk individuals within high-risk groups. We propose developing fairness-aware composite features that capture relevant risk information without encoding protected characteristics[7].

Traditional insurance risk assessment relies heavily on socioeconomic factors, creating tension between ethical ideals and business requirements. We recommend gradual implementation with parallel testing to quantify the "cost of fairness" against benefits in customer satisfaction and reduced legal risk.

Finally, counterfactual explanations might suggest impossible changes to customers, creating frustration rather than empowerment. We will implement feasibility constraints in our DiCE implementation, ensuring explanations focus on actionable changes within an individual's control[8].

By addressing these objections with thoughtful solutions, we balance non-discrimination imperatives with practical business considerations, recognizing that our ethical choices involve necessary trade-offs in the pursuit of fair algorithmic decision-making.

References:

[1] Nozick, R. (1974). Anarchy, State, and Utopia. Basic Books.

[2] Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 1(1).

[3] DeLucia, S.-A. (2023, May 19). *Evaluating model fairness*. Arize AI. https://arize.com/blog/evaluating-model-fairness/#:~:text=Let's%20dissect%20each.-,Sensitive%20Group%20Bias,address%20these%20concerns%20as%20well.

[4] Tefft, B. C. (2018, June 14). *Rates of motor vehicle crashes, injuries and deaths in relation to driver age, United States, 2014-2015 - AAA Foundation for Traffic Safety*. AAA Foundation for Traffic Safety -. https://aaafoundation.org/rates-motor-vehicle-crashes-injuries-deaths-relation-driver-age-united-states-2014-2015/

[5] *Hertz*. Car Rental: Save More on Rental Cars, Vans & Trucks. (n.d.). https://www.hertz.com/rentacar/reservation/

[6] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104, 671-732.

[7] Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? CHI Conference on Human Factors in Computing Systems Proceedings.

[8] Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2).