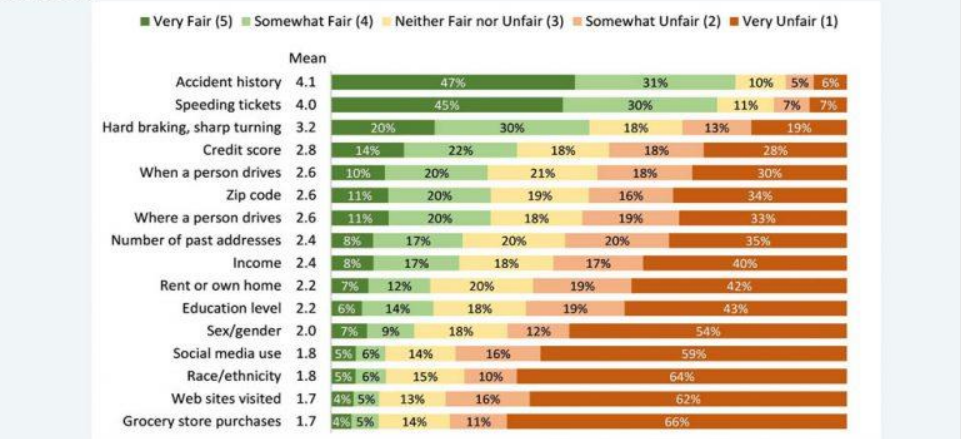
Scenario for Roundtable Discussion: Bias, Fairness, and the Modelling Lifecycle

Background: This scenario is a blend of two scenarios that we have used for internal educational events. In these events, participants would review the scenario and discuss answers to questions about identifying potential sources of bias, and developing a plan to test for whether social bias has impacted the model. The story itself is fictional, and does not describe a predictive model that is in use at our company. However, the story is inspired by situations we have encountered across various projects; it is intended to stimulate discussion on how social bias might impact a predictive model, and how to mitigate this bias.

Introduction

Past traffic violations are a commonly-used rating factor in Auto Insurance. Indeed, it is intuitive to expect that a driver who has committed more traffic violations will be more likely to cause an accident in the future. The *Actuarial Review* article "Sense & Sensitivity: Should fairness be a reason to eliminate predictive insurance rating factors?" included a poll asking the public about which factors they considered to be fair in decisions around auto insurance. "Speeding tickets" was considered the second-most fair factor in the poll, following "Accident history."

Figure 1. How Americans rate the fairness of companies using various types of data in car insurance decisions.



In the CAS Research Paper "Understanding Potential Influences of Racial Bias on P&C Insurance: Four Rating Factors Explored," the authors identified Motor Vehicle Records (MVR) as one potential source of racial bias. They cited several studies that demonstrated that Black or other minority drivers are subjected to traffic stops at a disproportionately higher rate than White drivers. The authors concluded that there was not strong enough evidence to suggest that this would result in a racial bias in the MVR, but given the bias in some components of the MVR they recommended further study to assess its impact in insurance rating.

In this discussion, let's think about how we could study this further in the context of a predictive modelling project.

General Questions (20 minutes)

As you read through this scenario, look out for examples of decisions that require human judgment. For each example, consider:

- Whose judgment is involved in making the decision?
- How might this decision impact the result, from a model performance and/or bias perspective?
- Is it possible to monitor the impact of this decision, and if so, how?

There are also some more specific questions embedded in the description of the project.

The Modelling Project

An insurer's predictive modelling team has been approached by the Claims Special Investigation Unit (SIU) team to build a predictive model to identify potentially fraudulent Personal Auto claims. Claims identified by the model will be referred to a triage team, who will decide whether the claim warrants further investigation to determine whether or not there is sufficient evidence of fraud. The model will be used in Ontario, which has a Human Rights Code that prohibits use of various protected grounds (including race) in models that have a customer-facing impact; this prohibition extends to data that may act as a proxy for protected grounds, such as census data about race. The Pricing team has also expressed an interest in understanding the high-level insights coming out of the model; they intend to use these insights to inform business adjustments to rating differentials that are associated with higher levels of fraud.

The SIU team has provided data on its triage team's past decisions around whether to accept or reject claims that have been referred to it for fraud investigation. Past referrals have been produced by a combination of an older predictive model, business rules, and manual referrals from claims adjusters. The modelling team uses these data as a response variable to train a logistic GLM. The SIU team is interested in using a model for which a high percentage of referred claims will be accepted by the triage team; i.e., their preferred model performance metric for this use case is precision.

The first version of the model includes the predictor "number of major convictions," which is derived from information about the insured's traffic violations: the Pricing team has defined groupings of non-criminal traffic convictions into "minor," "major," and "serious."

Given that "number of major convictions" was predictive in the model, the SIU team has requested that the modelling team also test the "number of minor convictions" variable in the model to see if it improves predictive power. The modelling team adds this to the list of predictors, and finds that it is statistically significant. However, the significance of "number of major convictions" drops; the modelling team investigates and concludes that this is due to moderate correlation between the variables. They decide to remove "number of major convictions" from the second version of the model.

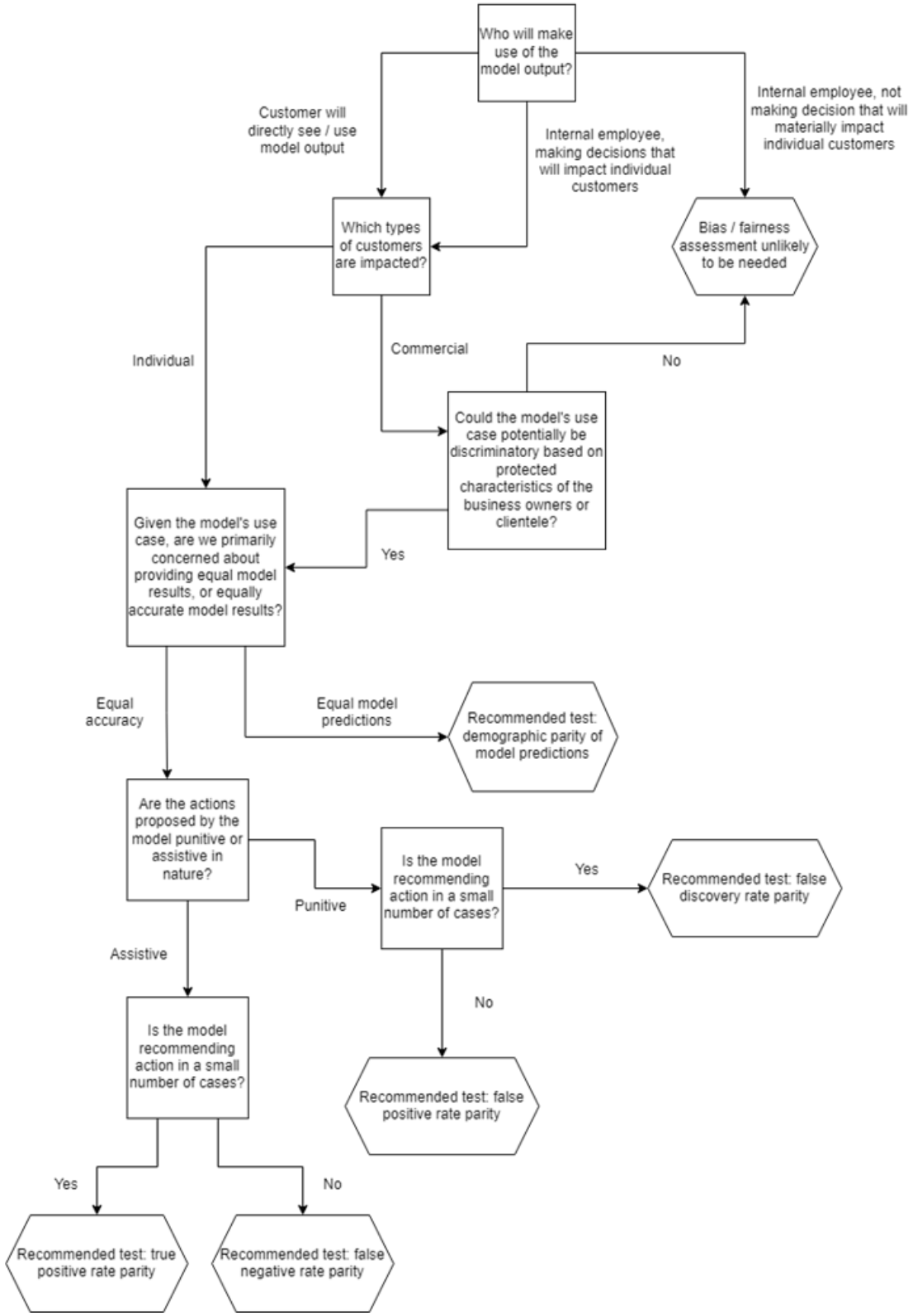
The modelling team now decides to test both models to see if they exhibit evidence of racial bias. They do not have access to data on the race of individual customers, but they do have access to census data on the racial breakdown within granular geographic units, so they will test the model to see if it shows a racial bias at the geographic level. (The census data is not currently used as a predictor in the model.)

In reporting back to the SIU, the modelling team summarizes their findings as: "The model has demonstrated that individuals with minor convictions are more likely to commit fraud." This high-level insight is also conveyed to the Pricing team. The SIU decides to use the model with the "minor convictions" predictor as its method of referring claims to the triage team. The model is deployed, and the modelling team sets up a monthly process for monitoring the precision of the model.

Discussion point (10 minutes): using the flowchart in Appendix A, what metric(s) would you select to test these models for bias? Do you have an a priori hypothesis about which of the two models might show a higher bias using this metric?

Discussion point (10 minutes): suppose that the test does show evidence of a material bias, based on the metric you selected above. Using the flowchart in Appendix B, what method would you test to see if it can mitigate the bias?

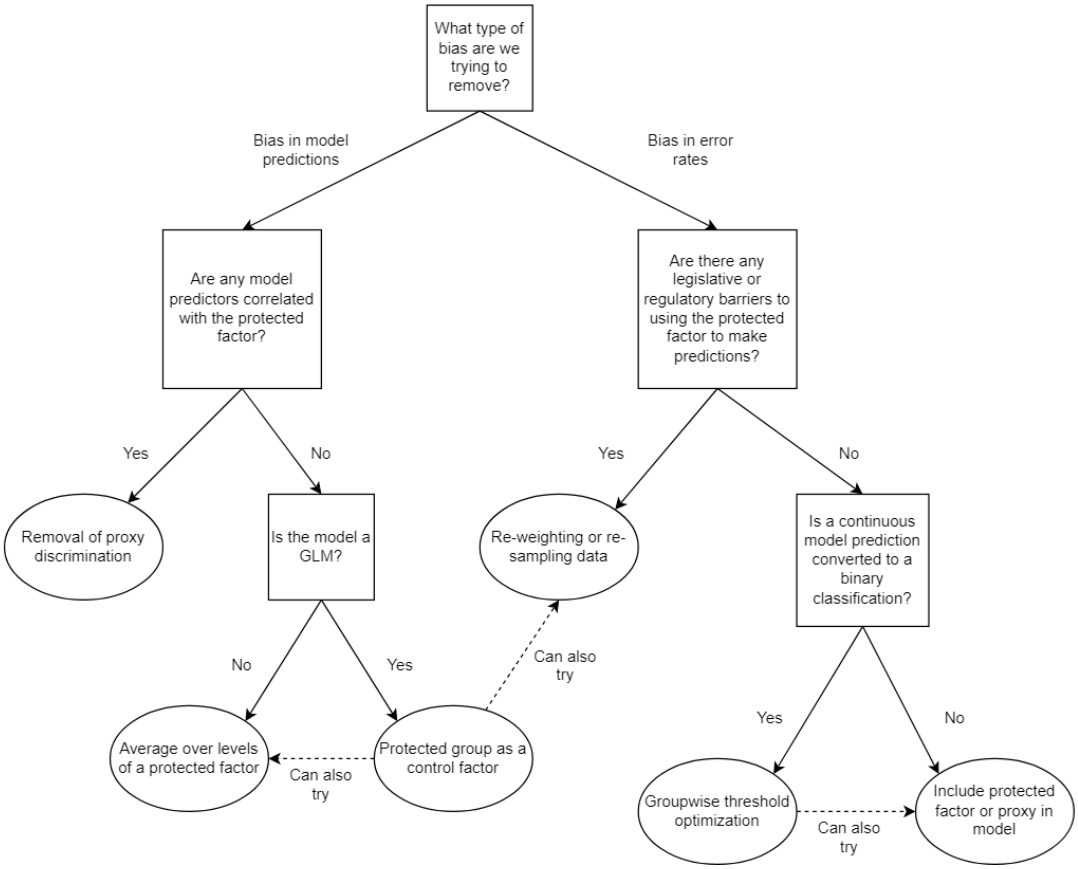
Appendix A: Detection Metric Flowchart



Metrics described in the diagram on the preceding page include:

- **Demographic parity** means that the model makes the same predictions for both classes. (See page 13 of the CAS research paper *Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance*.)
- **True positive rate parity** means that the true positive rates are the same for both classes. It is also known as **recall parity** or **equal opportunity**. (See page 14 of *Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance*)
- The metrics **false negative rate parity**, **false positive rate parity**, and **false discovery rate parity** are defined similarly: the corresponding error rate is the same for both classes.
- A model that exhibits both true positive rate parity and false positive rate parity satisfies the equalized odds criterion. (See pages 14-15 of *Methods for Quantifying Discriminatory Effects on Protected Classes in Insurance*)

Appendix B: Simple Bias Mitigation Methods



This is a list of some methods that we recommend our team try if they detect an undesired bias in the model, including guidance on situations in which each technique is appropriate.

- **Removal of proxy discrimination** means refitting the model after removing a suspected proxy variable from the list of predictors.
- **Average over levels of a protected factor** means that the protected factor is used as a predictor when fitting the model, but predictions are made without using the true value of the protected factor. The final output is produced by making predictions for each level of the protected factor, and taking the average.
- **Including protected group as a control factor** means that the protected factor is used as a predictor when fitting the model, but the protected group variable is not used when making model predictions (e.g. by manually deleting the coefficient from the model).
- **Re-weighting or re-sampling data** means defining weights (or sampling data) in a way that ensures that the weighted average response for each group is equal.
- **Groupwise threshold optimization** means that the classification threshold is set differently for each group such that parity is achieved for the relevant bias detection metric.
- **Including protected factor or proxy in model** means that the factor is included as a predictor when fitting the model, and is used when making predictions as well.