RDS (DS-UA 202) Spring 2022: Homework 1 Solutions

Problem 1 (20 points): Fairness from the point of view of different stakeholders

(a) (5 points) For each criterion A-E below, explain in 1-2 sentences which stakeholders would benefit from a model that optimizes that metric, why. If you believe that a criterion is not reasonable in this case, state so.

A. Accuracy

Accuracy benefits multiple stakeholders, including the software vendor, the decision maker, and the general public. In a sense, high accuracy, like calibration, is a necessary condition that a risk instrument should meet to be considered useful.

B. Positive predictive value

Positive predictive value (PPV), defined as TP / (TP + FP), is higher when the true positives constitute a high proportion of the positive class. PPV benefits society, in the sense that we don't spend resources on incarcerating individuals who do not go on to reoffend (FP). It also clearly benefits the prisoners who would not go on to reoffend themselves.

C. False positive rate

False positive rate (FPR), defined as FP / N, is lower when fewer of the defendants who would not go on to reoffend are classified as high risk. Minimizing the FPR benefits low-risk defendants. However, a decision maker may be willing to incur a higher FPR to help lower the false negative rate (FNR).

D. False negative rate

False negative rate (FNR), defined as FN / P, is lower when fewer of the defendants who would go on to reoffend are classified as low risk. A low FNR benefits society by ensuring that individuals who are likely to recidivate do not have a chance to do so. It also benefits a decision maker whose reputation will suffer if an individual whom they release goes on to commit a crime. Finally, it benefits the software vendor that caters to this type of a decision maker.

E. Statistical parity (demographic parity among the individuals receiving any prediction)

Statistical parity (demographic parity among the individuals receiving any prediction), does not make sense in this application. Note that we can ``optimize'' statistical parity at many levels of accuracy; statistical parity alone is not unambiguously better for any group. (A system that predicts high risk for every single individual has perfect statistical

parity.) Furthermore, we are not interested in equalizing outcomes between populations per se. Rather, fairness here amounts to balancing the kinds of error that different demographic groups incur - a goal that, as we know from the literature, cannot be met directly and so requires trade-offs.

(b) (6 points) Consider a hypothetical scenario in which *TechCorp*, a large technology company, is hiring for data scientist roles. Alex, a recruiter at *TechCorp*, uses a resume screening tool called *Prophecy* to help identify promising candidates. *Prophecy* takes applicant resumes as input and returns them in ranked (sorted) order, with the more promising applicants (according to the tool) appearing closer to the top of the ranked list. Alex takes the output of the *Prophecy* tool under advisement when deciding whom to invite for a job interview.

For each type of bias:

- Give an example of how this type of bias may arise in the scenario described above;
- Name a stakeholder group that may be harmed by this type of bias; and
- Propose an intervention that may help mitigate this type of bias.

A. Pre-existing bias

- **Example:** If very few female candidates were hired in the past, then historical data on which *Prophecy's* predictive analytic is trained won't have much information about how female employees perform. The model would then prioritize male candidates over female candidates - male candidates would appear higher up in the ranked list that is shown to Alex. We may observe similar behavior for other groups that are historically underrepresented in the workforce: non-White individuals, individuals with disabilities, or individuals who belong to multiple underrersented or historically disadvantaged groups (e.g., Black women) will be less likely to appear at the top of the ranked list.
- **Stakeholders:** Members of historically underrepresented groups.
- **Intervention:** *Prophecy* and *TechCorp* should work together to ensure that additional data is included into the training set that shows on-the-job performance of members of historically underrepresented groups. If such data is not available, consider using an alternative selection procedure. For example, select candidates who meet the explicit requirements for the job, and randomly select a subset of these to interview.

B. Technical

- **Example**: The technical system may be interpolating values of attributes for which values are missing under systematically incorrect assumptions. For example, if age is missing more frequently for older job applicants than for younger job applicants, yet the system uses mean imputation to fill in missing age values, then age will be skewed to younger values when interpolated. (This is a situation where age is not missing at random, i.e., whether it's missing depends on its actual value.) If age is an important feature in *Prophecy's* model, and if higher age values are better (they result in ranking the candidate higher in the list), then this will systematically downgrade individuals who did not specify their age.
- **Stakeholder:** Individuals who did not specify age values.
- **Intervention**: *Prophecy* should implement a more sophisticated data interpolation method that does not make the assumption that data is missing at random.

- **Example**: Another example of technical bias that is specific to rankings is that only the few highest-ranked results will in fact be considered by the recruiter. If the top positions are systematically occupied by candidates who are very similar in terms of their demographics that the company currently employs (due to pre-existing bias), then these are going to be the candidates that the recruiter selects.
- **Stakeholder:** Any job applicant may be harmed by this, particularly applicants who come from demographic groups that are underrepresented in the workforce.
- **Intervention:** Recruiters should be instructed to look beyond the top-10, and consider candidates who are ranked lower in the list yet meet the qualifications for the job. *Prophecy* should randomize candidates at the top of the list.

C. Emergent

- **Example**: Recruiters have been using *Prophecy's* recommendations for a while. Because Prophecy consistently places the same "type" of a candidate at the top of the ranked list, recruiters have come to believe that these are, in fact, the preferable candidates. Such candidates are then hired, and become part of the training data for future iterations of *Prophecy*'s model. We see a feedback loop.
- **Stakeholder:** Any job applicant may be harmed by this, particularly applicants who come from demographic groups that are underrepresented in the workforce.
- **Intervention:** Recruiters should be instructed to take *Prophecy's* recommendations with a grain of salt: look beyond the top-10, and consider candidates who are ranked lower in the list yet meet the qualifications for the job. *Prophecy* should randomize candidates at the top of the list.

(c) (9 points) Consider a hypothetical scenario in which an admissions officer at *Best University* is evaluating applicants based on 3 features: SAT score, high school GPA, and family income bracket (low, medium, high). We discussed several equality of opportunity (EO) doctrines in class and in the "Fairness and Friends" comic: formal, substantive / luck egalitarian, and substantive / Rawlsian.

A. In a selection procedure that is fair according to formal EO, which of these features would the admissions officer use? Briefly justify your answer.

The admissions officer would use applicants' SAT score and high school GPA. Under formal EO, the admissions officer should use only qualifications that are relevant for success at *Best University*. SAT score and high school GPA are plausibly relevant for the probability of success at *Best University*, but family income is not relevant.

B. Suppose that income-based differences are observed in applicants' SAT scores: the median SAT score is lower for applicants from low-income families, as compared to those from medium- and high-income families. Which EO doctrine(s) is/are consistent with the goal of correcting such differences in the applicant pool? Briefly justify your answer.

Substantive/Rawlsian EO and substantive/luck-egalitarian EO are consistent with the goal of correcting these differences. Rawlsian EO would argue that the advantage of being born into a high-income family has snowballed into an advantage in SAT performance. Rawlsian EO would seek to ensure that higher SAT scores due to these advantages do not translate into a better chance of admissions at *Best University*. Luck-egalitarian EO would argue that the difference in SAT scores between income brackets is a matter of "brute luck." Under luck-egalitarian EO, matters of brute luck should not affect the outcomes of a competition.

C. Describe an applicant selection procedure that is fair according to luck-egalitarian EO.

Applicants could be evaluated against other applicants from the same family income bracket. For example, *Best University* could rank applicants by SAT score and high school GPA within each income bracket (low, medium, high). The top *n* applicants from each bracket could be offered admission to *Best University*.