On the Long-term Impact of Algorithmic Decision Policies

Appendix

1 The Effort Function

We define $\epsilon_{s,k}(x_k, x'_k,)$—depending on the type of feature k—as follows:

- **Non-monotone numerical feature:** Suppose feature k is numerical, but it is not clear which direction of change should increase the probability of the instance $\mathbf{x}$ being labeled as positive. An example of this type of feature in the education context is *extracurricular activities*—depending on other factors this may be increase or decrease one's performance in school. For this type of feature, we assume change in either direction requires effort, and define $\epsilon_{s,k}(x_k, x'_k,)$ as follows:

$$\epsilon_{s,k}(x_k, x'_k) = |Q_{s,k}(x'_k) - Q_{s,k}(x_k)|.$$

- **Ordinal feature:** We define $\epsilon_{s,k}(x_k, x'_k,)$ similar to numerical features—depending on whether we consider the attribute monotone or not.

- **Categorical feature:** Suppose feature k is categorical and can take on n_k different values $\{v_1, \cdots, v_{n_k}\}$ (example: marital status). We define $\epsilon_{s,k}$ via n_K^2 constants, $c_{i,j}$ for $1 \leq i, j \leq n_k$, with $c_{i,j}$ specifying the effort required to change the value of feature k from v_i to v_j. Throughout our simulations and for simplicity, we assume there exists a constant c such that $c_{i,j} \equiv c$ for all $1 \leq i, j \leq n_k$.

- **(Conditionally) immutable feature:** We call feature k *(conditionally) immutable* if there exist two values $x_k \neq x'_k$, where the change from x_k to x'_k is considered impossible. For example, race is an immutable feature (one cannot be expected to change their race). Age is conditionally immutable (one cannot be expected to become younger). In this case we define our effort function as follows: $\epsilon_{s,k}(x_k, x'_k) = \infty$.

2 The Student Performance Data Set

The student performance data set (Cortez & Silva, 2008) contains information about student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features. The data set consists of 649 instances/student, with each instance consisting of 32 features. The task is to predict the student's final grade (value from 0 to 20) in Portuguese. Out of the 32 features, we choose only features that are considered mutable in at least one direction, that is, the student can exert effort and change the feature value. We dropped all immutable features—except gender—to be able to find a social model for every student. (Since the data set is very small, this would not have been possible had we kept the immutable features). This results in a total of 23 features out of which 10 are binary and the rest are numerical. We then perform a 70:30 train-test split, with the train set consisting of 454 instances and the test set consisting of 195 instances.

	school	address	traveltime	studytime	schoolsup	famsup	paid	activities	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	Intercept
Ridge Regression, Regularization Const. = 0.1	1.4407	0.3136	-0.2481	1.6016	-1.4664	0.0881	-0.7904	0.1965	2.5093	0.5606	-0.4103	0.5203	-0.9247	-0.0061	-1.0606	-0.5881	-0.8108	-1.5757	8.9708
Ridge Regression, Regularization Const. = 200	0.6065	0.2795	-0.0926	0.4125	-0.2505	0.0571	-0.1035	0.1211	0.6293	0.2433	-0.2105	0.1164	-0.2048	-0.1341	-0.2757	-0.3072	-0.2371	-0.1182	10.5303

(a) Mutable feature weights (and intercept) common to both models.

	Fedu	Fjob (at_home)	Fjob (health)	Fjob (other)	Fjob (services)	Fjob (teacher)	Medu	Mjob (at_home)	Mjob (health)	Mjob (other)	Mjob (services)	Mjob (teacher)	Pstatus	age	failures	famsize	guardian (father)	guardian (mother)	guardian (other)	nursery	reason (course)	reason (home)	reason (other)	reason (reputation)	sex
Ridge Regression, Regularization Const. = 200	0.2781	-0.0262	0.0007	0.0048	-0.1533	0.174	0.2966	-0.2095	0.1352	-0.1355	0.0145	0.1953	0.0362	-0.0753	-0.5692	-0.1352	0.13	-0.0402	-0.0899	0.0053	-0.1258	0.0601	-0.217	0.2827	0.3614

(b) Immutable feature weights; only applicable to Ridge Regression with regularization const = 200.

Figure 1: Weights assigned by 2 models, one trained with only mutable features (top row in Figure 1a) and the other trained with both mutable and immutable features (bottom row in Figure 1a and the model in Figure 1b).

3 Trained Models

We trained the following models on the student performance data set:

- **Neural network:** A shallow neural network with one hidden layer (ReLu activation) containing 100 nodes. Loss function with L2 regularization with regularization strength = 10. Regularization strength and number of nodes in the hidden layer were found using grid search by doing a 3 fold cross validation and taking the parameters that resulted in the maximum average test accuracy.
- **Linear regressor:** Least squares solver. Finds parameters B such that L2 norm of $|Bx - Y|$ is minimized.
- **Decision Tree:** Decision Tree Regressor with maximum depth of 5 to avoid overfitting. Max depth parameter was chosen using grid search by doing a 3-fold cross validation and choosing the parameter that maximised the average test set accuracy. Criterion for splitting was minimization of MSE.

4 Fairness Notions for Regression

- **Positive residual difference** (Calders et al., 2013) is computed by taking the absolute difference of mean positive residuals across groups:

$$\left| \frac{1}{|G_1^+|} \sum_{i \in G_1} \max\{0, (\hat{y}_i - y_i)\} - \frac{1}{|G_2^+|} \sum_{i \in G_2} \max\{0, (\hat{y}_i - y_i)\} \right|.$$

- **Negative residual difference** (Calders et al., 2013) is computed by taking the absolute difference of mean negative residuals across groups:

$$\left| \frac{1}{|G_1^-|} \sum_{i \in G_1} \max\{0, (y_i - \hat{y}_i)\} - \frac{1}{|G_2^-|} \sum_{i \in G_2} \max\{0, (y_i - \hat{y}_i)\} \right|.$$

5 Why Existing Notions of Fairness Fail to Capture Effort-Reward Disparity

Figure 1 shows an example of 2 ridge regressions, both trained on the student performance dataset (described in section 2), but one has access to only mutable features and the other has access to

both mutable and (conditionally) immutable features. For simplicity, let's call them "mutable model" and "combined model" respectively. Both the "mutable model" and "combined model" have similar error distributions on the dataset with Mean Averaged Errors (MAE) of 2.028 and 2.046 on the entire population. They also have similar errors across sub-groups defined based on the value of sensitive feature s (for the student dataset, s corresponds to gender); with MAEs for the sub-group with $s = 1$ (females) being 1.999 and 2.067 and MAEs for sub-group with $s = 0$ (males) being 2.068 and 2.016 for "mutable model" and "combined model" respectively. Lastly, both the models also have comparable measures of existing fairness notions defined in section 4 with positive residuals of 0.296 and 0.228 and negative residuls of 0.237 and 0.249 respectively.

However, when evaluated for the effort-reward unfairness, "mutable model" and "combined model" perform differently with measures of 0.043 and 0.532 respectively. One of the reasons for such contrasting values is the different weights each model assigns to the mutable features (shown in Figure 1a). For example, consider a student at a benefit level of $b_{initial}$ (assuming benefit function = predicted value by the model) subject to predictions by the "mutable model" (top row in Figure 1a), were to imitate a role model having value of the continuous feature, "studytime", greater by 1 unit. Assume, for simplicity, that all other feature values of the role model and the student are same. Say the effort exerted to make this change is e which brings the student to a benefit level of b (=new predicted value by the model), thus making utility, $u_{mutable} = b - b_{initial} - e$. Now say the same student were subject to predictions by the "combined model" (bottom row in Figure 1a) and were to immitate the same role model (having "studytime" greater by 1 unit and having all other features same as the student) as in the previous case. Since both the models have similar prediction errors, we can assume that the student has a similar prediction value as in the previous case (thus being at the same benefit level of $b_{initial}$). The effort is independent of the model, so effort in this case remains e. However, since the weight assigned by "combined model" to "studytime" is 0.25x the weight assigned by "mutable model" (see Figure 1a), increasing "studytime" by 1 unit will result in a new benefit level of $b^{'} (< b)$. Thus utility in this case, $u_{combined} = b^{'} - b_{initial} - e$. Since $b_{initial}$, b, $b^{'}$ and e are all positive values, $u_{mutable} > u_{combined}$. Thus, the values of utility can differ considerably for 2 models even though their error distributions across the population may be very similar. Our notion of effor-reward unfairness captures this disparity while existing notions of fairness might not.

References

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. Controlling attribute effect in linear regression. In *Proceedings of the International Conference on Data Mining*, pp. 71–80. IEEE, 2013.

Cortez, P. and Silva, A. M. G. Using data mining to predict secondary school student performance. 2008.