5243 Project 4: ML Fairness

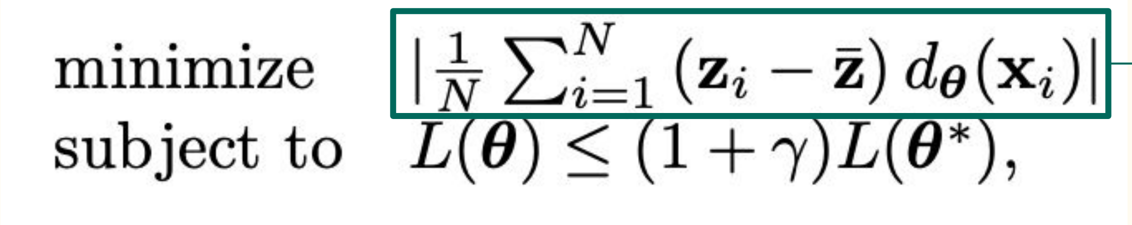
Kerry Cook;Nikhil Cherukupalli;Yuming Tu; Ziyong Zhang;Yongxin Ma

Model 1: Fairness Constraints Analysis

The goal is to build classifiers that avoid both disparate treatment and disparate impact - and accomodate the need for high performing models

- p%-rule - a way to quantify disparate impact
- Challenging to directly incorporate into a prediction task, so instead the paper introduces a new way to measure **decision boundary fairness**

Maximizing Fairness Under Accuracy Constraints:



$$\begin{array}{ll} \text{minimize} & \boxed{\left|\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \bar{\mathbf{z}}\right) d_{\boldsymbol{\theta}}(\mathbf{x}_i)\right|} \longrightarrow \\ \text{subject to} & L(\boldsymbol{\theta}) \leq (1+\gamma)L(\boldsymbol{\theta}^*), \end{array}$$

Covariance between the sensitive attribute and the distance between feature vector and decision boundary

Datasets- UCI Bank Data

Columns/variables: 17

- Binary variables: **default, housing, loan**
- Have more than 2 categories-one hot encode: **job, marital, contact, education, output**
- Numeric variables: **age, day, month**

Rows/observations: 45212

UCI Bank Classification

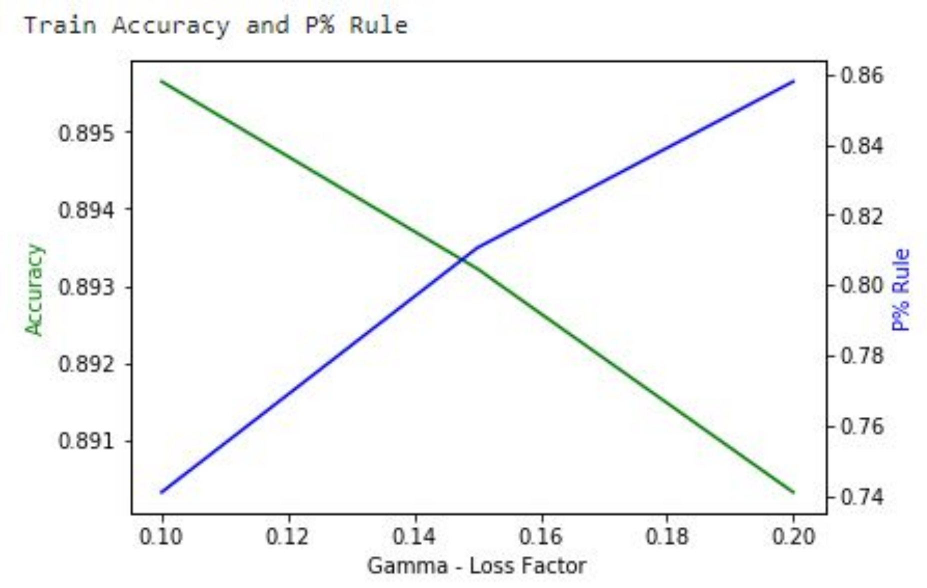
- Authors use **age** as the sensitive attribute, where ages between **25 and 60** are the **protected** group and the remaining ages are the **non-protected** group.

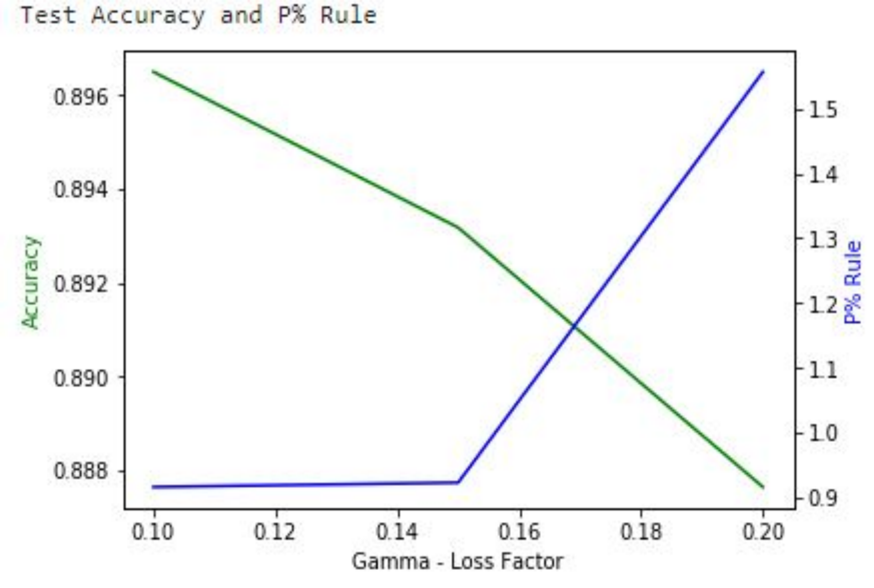
- Trained **logistic regression** classifier to **predict** whether or not a person subscribed to **term deposit** investment
    - **Age** was removed from the feature set to avoid disparate treatment

Results

Train: ~89% accuracy, p%-rule 27.3%

Test: ~90% accuracy, p%-rule 30.3%





Datasets-COMPAS: compas-scores-two-years.csv

The COMPAS dataset is used to predict the two_year_recid binary label

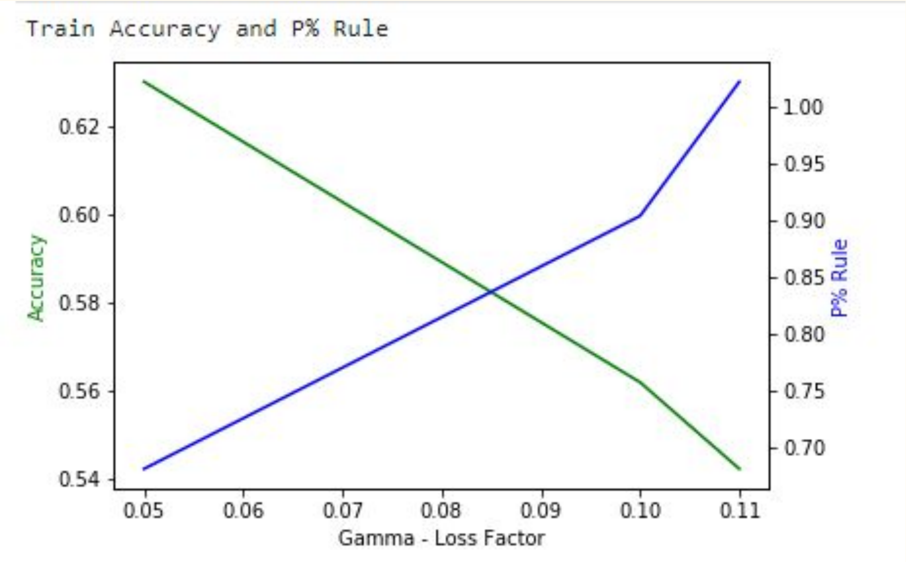
- Features: age, gender, charge degree, prior counts, and length of stay

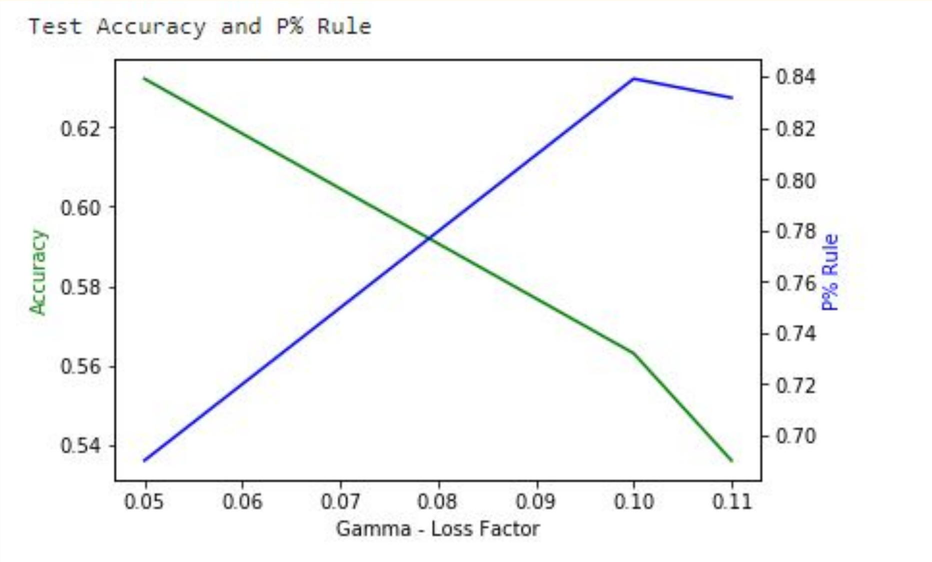
- Race: protected attribute - removed from feature set
  - Reduced to Caucasian and African-American

COMPAS Dataset Results

Train: 66% accuracy, p%-rule 59.7%

Test: 66% accuracy, p%-rule 65.7%





Model 2: Fairness Features Selection

Unlike the previous model, this method focuses on the impact of features on discriminatory predicitons, and does not focus on a specific classifier design

- Alleviate disparate impact by **identifying features** that can serve as a proxy for an individual's sensitive attribute (race/gender etc) and prevent them from influencing the decision outcome
- **Problem:** May contain information necessary to make accurate predictions

Goal: Select features that optimally satisfy **accuracy and fairness requirements**

Shapley Coefficient

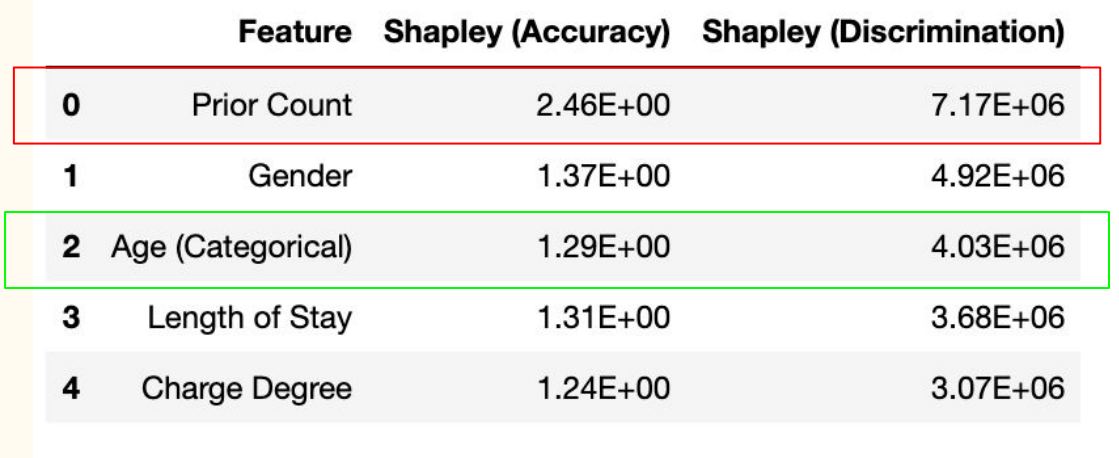
Goal: Quantify impact on both accuracy and discrimination of excluding a feature from the feature space by proposing information-theoretic accuracy and discrimination measures

Strategy:

1. Find all subsets of feature space excluding feature i
2. Calculate marginal information contribution of including i for each subset
3. Calculate weighted average over all subsets
4. Repeat for each feature

Feature Selection:

In this part, we implement the routines described in the paper. On the training data, we calculate Shapley coefficients for each of our features capturing effects on both accuracy and discrimation on our protected group (i.e. race).



	Feature	Shapley (Accuracy)	Shapley (Discrimination)
0	Prior Count	2.46E+00	7.17E+06
1	Gender	1.37E+00	4.92E+06
2	Age (Categorical)	1.29E+00	4.03E+06
3	Length of Stay	1.31E+00	3.68E+06
4	Charge Degree	1.24E+00	3.07E+06

We see that 'Prior Count' has the sharpest affect on both discrimination and accuracy, so eliminating it can prove problematic for a classifier.

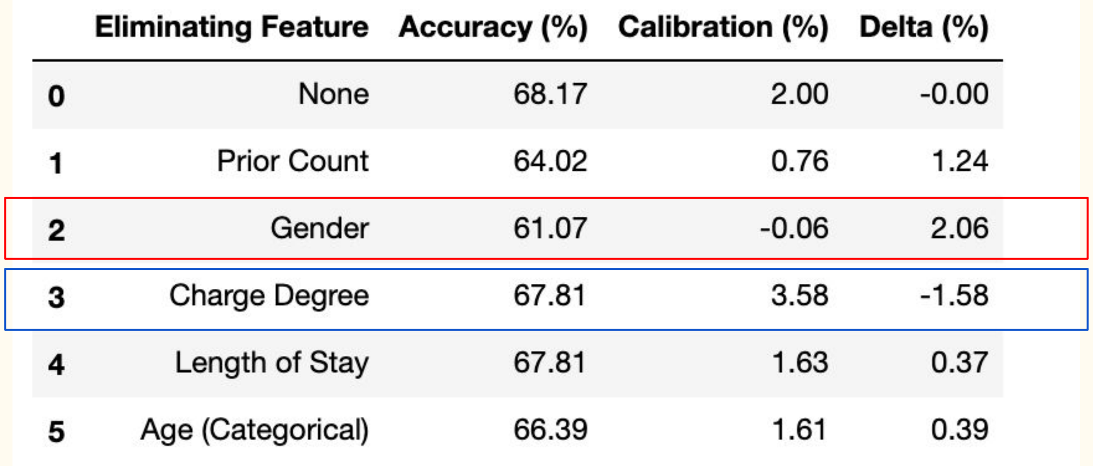
However, a feature such as 'Age (Categorical)' is relatively discriminatory but eliminating it would not seriously reduce accuracy from our results.

Classification:

- Trained an SVM model that predicts whether a person will or will not recidivate given the aforementioned features
- Calculate accuracy as well as a **calibration** - the difference between accuracy amongst the sensitive attribute groups (race)
- Trained submodels which eliminate each feature iteratively and calculate both metrics

Result:

When evaluating the results, we must compare against the baseline calibration score of 2%



	Eliminating Feature	Accuracy (%)	Calibration (%)	Delta (%)
0	None	68.17	2.00	-0.00
1	Prior Count	64.02	0.76	1.24
2	Gender	61.07	-0.06	2.06
3	Charge Degree	67.81	3.58	-1.58
4	Length of Stay	67.81	1.63	0.37
5	Age (Categorical)	66.39	1.61	0.39

We showed that eliminating 'Prior Count' should result in the strongest drop in discrimation but the results suggest that actually eliminating 'Gender' yields the greatest benefit to discrimation

Our FFS process showed that dropping 'Charge Degree' is unnecessary and the results prove that removing it from the training set results in a significantly more discriminatory classifier

Appendix

The mathematical expressions are:

$$
\begin{aligned}
&I(T;R_1,R_2) = UI(T;R_1\backslash R_2) + UI(T;R_2\backslash R_1) + SI(T;R_1,R_2) + CI(T;R_1,R_2),\\
&I(T;R_i) = UI(T;R_i\backslash R_j) + SI(T;R_1,R_2),\ i \neq j,\ i,j \in \{1,2\}.
\end{aligned}
$$

The unique information of R1 with respect to T, denoted by UI(T;R1 \ R2), represents the information content related to T that is only available in R1. The shared information of R1 and R2, denoted by SI(T;R1,R2), represents the information content related to T that both R1 and R2 possess. Finally, the synergistic information of R1 and R2, denoted by CI(T;R1,R2), represents the information content that can be obtained only if both R1 and R2 are available

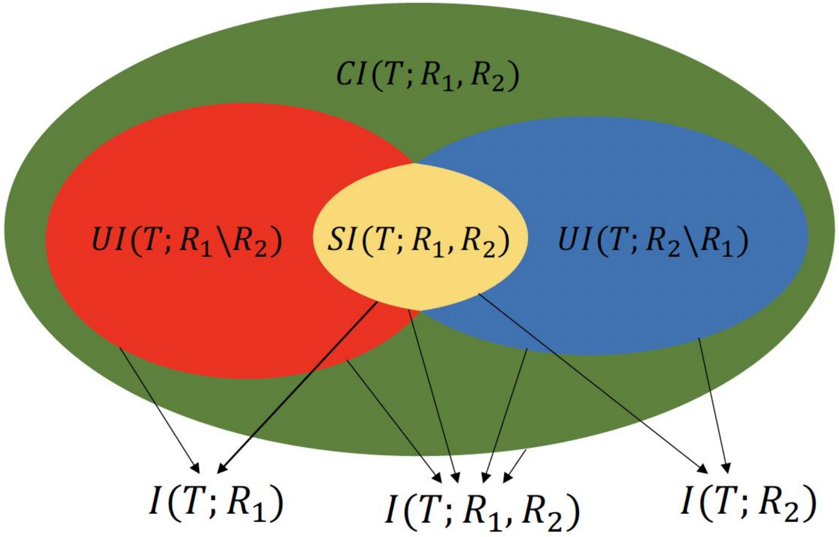
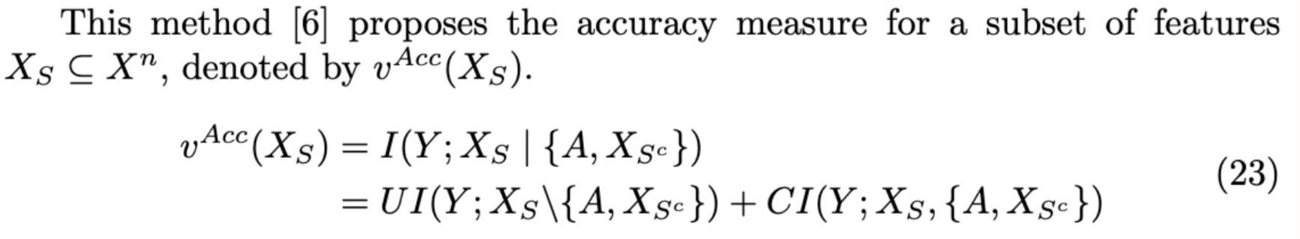


Figure 1: Decomposition of Information.

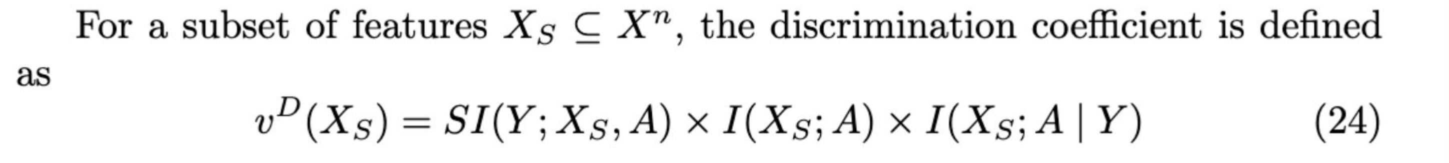
Quantifying accuracy effect



This method [6] proposes the accuracy measure for a subset of features $X_S \subseteq X^n$, denoted by $v^{Acc}(X_S)$.

$$\begin{aligned} v^{Acc}(X_S) &= I(Y; X_S \mid \{A, X_{S^c}\}) \\ &= UI(Y; X_S \backslash \{A, X_{S^c}\}) + CI(Y; X_S, \{A, X_{S^c}\}) \end{aligned} \tag{23}$$

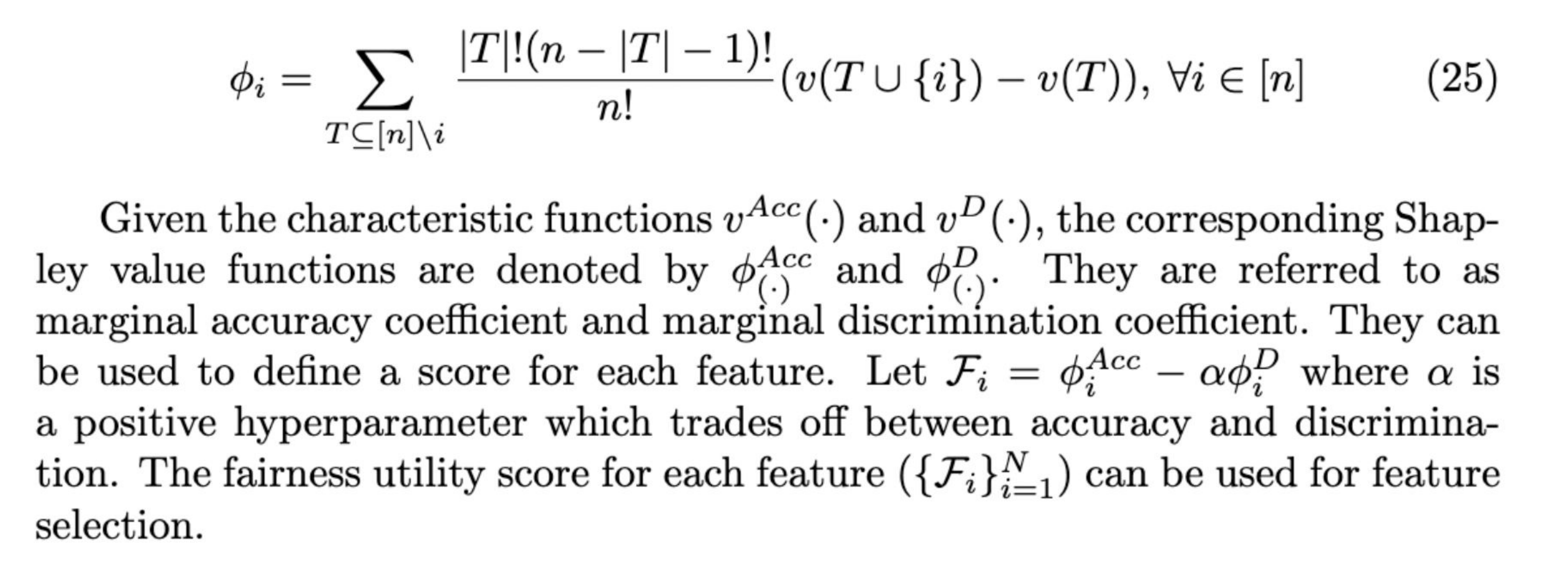
Quantifying discriminatory effect



For a subset of features $X_S \subseteq X^n$, the discrimination coefficient is defined as

$$v^D(X_S) = SI(Y; X_S, A) \times I(X_S; A) \times I(X_S; A \mid Y) \tag{24}$$

Fairness utility score calculation



$$\phi_i = \sum_{T \subseteq [n] \setminus i} \frac{|T|!(n - |T| - 1)!}{n!} (v(T \cup \{i\}) - v(T)), \ \forall i \in [n] \tag{25}$$

Given the characteristic functions $v^{Acc}(\cdot)$ and $v^{D}(\cdot)$, the corresponding Shapley value functions are denoted by $\phi^{Acc}_{(\cdot)}$ and $\phi^{D}_{(\cdot)}$. They are referred to as marginal accuracy coefficient and marginal discrimination coefficient. They can be used to define a score for each feature. Let $\mathcal{F}_i = \phi_i^{Acc} - \alpha \phi_i^{D}$ where α is a positive hyperparameter which trades off between accuracy and discrimination. The fairness utility score for each feature ($\{\mathcal{F}_i\}_{i=1}^{N}$) can be used for feature selection.