Project 4: Algorithm Implementation and Evaluation

Group 10
Fairness-Aware Classifier with Prejudice Remover Regularizer
Information Theoretic Measures for Fairness-Aware Feature Selection

Data Preprocessing

- Data Split : Train : Validation : Test = 5 : 1 : 1
- Target Variable : **two_year_recid**
- Non-sensitive Features : **age, c_charge_degree, score_text, priors_count, is_violent_recid, length_of_stay**
- Sensitive Feature : **race (African-American vs. Caucasian)**

Fairness-Aware Classifier with Prejudice Remover Regularizer

- This algorithm proposes measures to quantity the indirect prejudice and a regularization approach that is applicable to any prediction algorithm with probabilistic discriminative models.
- The indirect prejudice is defined as the statistical dependence between the sensitive feature and the target variable.

- (indirect) Prejudice Index:

$$\mathrm{PI} = \sum_{(y,s)\in\mathcal{D}} \hat{\Pr}[y,s] \ln \frac{\hat{\Pr}[y,s]}{\hat{\Pr}[y]\hat{\Pr}[s]}.$$

Baseline Model

- Logistic Regression

$$\mathcal{M}[y|\mathbf{x}, s; \boldsymbol{\Theta}] = y\sigma(\mathbf{x}^\top \mathbf{w}_s) + (1-y)(1-\sigma(\mathbf{x}^\top \mathbf{w}_s))$$

- Results:

	Validation	Test
Average Accuracy	0.700	0.688
Calibration Score	0.040	0.055

Logistic Regression with Prejudice Remover Regularizer

- Objective Function to Minimize

$$-\mathcal{L}(\mathcal{D};\boldsymbol{\Theta}) + \eta \mathrm{R}(\mathcal{D},\boldsymbol{\Theta}) + \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_2^2$$

Two regularizers: Enforce fair classification

Avoid overfitting

- Prejudice Remover Regularizer

$$\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \sum_{y\in\{0,1\}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}] \ln \frac{\hat{\Pr}[y|s_i]}{\hat{\Pr}[y]}$$

- Estimated Probability of y Given s

$$\hat{\Pr}[y|s] \approx \frac{\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D} \text{ s.t. } s_i=s} \mathcal{M}[y|\mathbf{x}_i,s;\boldsymbol{\Theta}]}{|\{(\mathbf{x}_i,s_i)\in\mathcal{D} \text{ s.t. } s_i=s\}|}$$

- Estimated Probability of y

$$\hat{\Pr}[y] \approx \frac{\sum_{(\mathbf{x}_i,s_i)\in\mathcal{D}} \mathcal{M}[y|\mathbf{x}_i,s_i;\boldsymbol{\Theta}]}{|\mathcal{D}|}$$

Model Evaluation

Accuracy	eta = 0	eta = 1	eta = 2	eta = 3	eta = 4	eta = 5	eta = 10	eta = 15	eta = 20	eta = 25
Train	0.7041	0.7066	0.706	0.7048	0.705	0.7052	0.7058	0.7067	0.7042	0.7058
Validation	0.6963	0.6883	0.6873	0.6856	0.6811	0.6844	0.6816	0.6816	0.6828	0.6828
Test	0.689	0.6798	0.6827	0.6757	0.6775	0.6763	0.678	0.6751	0.6694	0.6751

Calibration	eta = 0	eta = 1	eta = 2	eta = 3	eta = 4	eta = 5	eta = 10	eta = 15	eta = 20	eta = 25
Train	0.0187	0.0181	0.0168	0.0101	0.0105	0.011	0.008	0.0126	0.0176	0.0135
Validation	0.0467	0.0355	0.024	0.0207	0.0162	0.023	0.022	0.022	0.0196	0.0196
Test	0.0525	0.0154	0.0166	0.0027	0.0062	0.0085	0.0019	0.0031	0.0193	0.0031

Information Theoretic Measures for Fairness-aware Feature Selection

- *Information theory is the scientific study of quantification of information in a message/ event*
- *How to select features in a fair way that minimizes discrimination while maintaining accuracy*

Algorithm

Calculate the Shapley Values i.e., the marginal accuracy and discrimination for each feature using combined accuracy and discrimination

Definition 5. Let $\mathcal{P}$ denote the power set. Given a *characteristic function* $v(\cdot) : \mathcal{P}([n]) \to \mathbb{R}$, the Shapley value function $\phi_{(\cdot)} : [n] \to \mathbb{R}$ is defined as:

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

Given the characteristic functions $v^{Acc}(\cdot)$ and $v^{D}(\cdot)$, the corresponding Shapley value functions are denoted by $\phi^{Acc}_{(\cdot)}$ and $\phi^{D}_{(\cdot)}$. We refer to these as *marginal accuracy coefficient* and *marginal discrimination coefficient*, respectively.

The weights $\frac{|T|!(n-|T|-1)!}{n!}$ in the definition of the Shapley value function are chosen so that the following lemma holds:

Lemma 3. [41] Shapley value is the unique aggregation function satisfying the following properties:

- Symmetry: If $v(T \cup \{i\}) = v(T \cup \{j\})$, for all $T \subseteq [n] \setminus \{i, j\}$, $\implies \phi_i = \phi_j$.
- Efficiency: $\sum_{i \in [n]} \phi_i = v([n])$.
- Monotonicity: Given two characteristic functions $v^{(1)}(\cdot)$, $v^{(2)}(\cdot)$, and the corresponding Shapley value functions $\phi^{(1)}_{(\cdot)}$, $\phi^{(2)}_{(\cdot)}$, if $v^{(1)}(T \cup \{i\}) - v^{(1)}(T) \geq v^{(2)}(T \cup \{i\}) - v^{(2)}(T), \forall T \subseteq [n] \implies \phi^{(1)}_i \geq \phi^{(2)}_i$.

Algorithm

	Feature	Accuracy	Discrimination
0	c_charge_degree	0.981916	84457.538802
1	age_cat	1.195308	107576.605927
2	sex	0.903141	78210.136834
3	is_violent_recid	0.744201	72271.230359
4	priors_count	1.185834	108126.985005
5	length_of_stay	1.038081	101778.528671

The obtained outcome demonstrates that Age and Priors Counts exhibit the most significant influence on accuracy while also serving as strong indicators for discrimination.

Calculating F-score..

Upon discovering this, we then calculate the fairness-utility scores of each feature to make a more informed decision.

@ alpha = .000001 , is_violent_recid should be removed

	Feature	Accuracy	Discrimination	F_score
0	c_charge_degree	0.981916	84457.538802	0.897458
1	age_cat	1.195308	107576.605927	1.087731
2	sex	0.903141	78210.136834	0.824930
3	is_violent_recid	0.744201	72271.230359	0.671930
4	priors_count	1.185834	108126.985005	1.077707
5	length_of_stay	1.038081	101778.528671	0.936302

Accuracy after Logistic Regression = 58.55%

@ alpha = .00001 & .0001 , length_of_stay should be removed

	Feature	Accuracy	Discrimination	F_score
0	c_charge_degree	0.981916	84457.538802	0.137340
1	age_cat	1.195308	107576.605927	0.119542
2	sex	0.903141	78210.136834	0.121039
3	is_violent_recid	0.744201	72271.230359	0.021489
4	priors_count	1.185834	108126.985005	0.104564
5	length_of_stay	1.038081	101778.528671	0.020295

	Feature	Accuracy	Discrimination	F_score
0	c_charge_degree	0.981916	84457.538802	-7.463838
1	age_cat	1.195308	107576.605927	-9.562353
2	sex	0.903141	78210.136834	-6.917873
3	is_violent_recid	0.744201	72271.230359	-6.482922
4	priors_count	1.185834	108126.985005	-9.626865
5	length_of_stay	1.038081	101778.528671	-9.139772

Accuracy after Logistic Regression = 63.84%

Using calibration..

	Eliminating Feature	Accuracy	Calibration
0	base	0.665152	0.015793
1	c_charge_degree	0.662121	0.016577
2	age_cat	0.660606	0.026210
3	sex	0.654545	0.009297
4	is_violent_recid	0.596970	0.018033
5	priors_count	0.624242	0.010977
6	length_of_stay	0.657576	0.014673

Therefore, we remove both length_of_stay & sex.

Accuracy after Logistic Regression = 64.29%

Conclusion

- By using the logistic regression with prejudice remover regularizer, we are able to penalize the model for replying to heavily on the sensitive feature and thus to decrease the influence of indirect prejudice by adjusting the regularization coefficient, while maintaining the predictive accuracy.
- By using the Information Theoretic Measures for Fairness-aware Feature Selection on the baseline model, we are able to improve accuracy by ~2% when accounting for the tradeoff between accuracy and fairness