Project 4 Group 7

Machine Learning Fairness

Kechen Lu
Fei Li
Siyu Li

Model & Algorithms

Baseline Models: Logistic Regression (Without Constraints)

Algorithm 1: Information Theoretic Measures for Fairness-aware Feature selection

Algorithm 2: Learning Fair Representations

Baseline Model: Logistic Regression (Without Constraints)

- shows the model's accuracy and calibration for each of the data sets (training, validation, and test).
- used as a reference point for model performance without any fairness constraints applied.

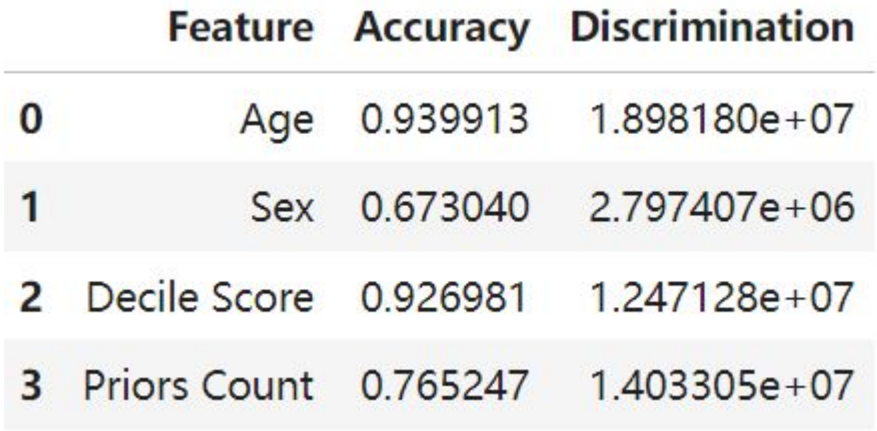
Algorithm 1: Information Theoretic Measures for Fairness-aware Feature Selection(FFS)

- identify and select features that contribute to the accuracy of predictions while minimizing discriminatory impact, for example, race.
- employs Shapley value analysis from cooperative game theory to quantify the marginal impact of each feature, assessing both accuracy and discrimination contributions.
- **Shapley value functions** to calculate marginal accuracy and marginal discrimination

Evaluation

- Quantify the accuracy and discrimination impact of subsets of features based on information-theoretic measures.
- Calculate a fairness-utility score for each feature based on its contribution to both accurate predictions and nondiscriminatory outcomes.
- Using a hyperparameter, alpha, trades off between accuracy and discrimination
- Utilizes Shapley analysis to determine the marginal impact of each feature.

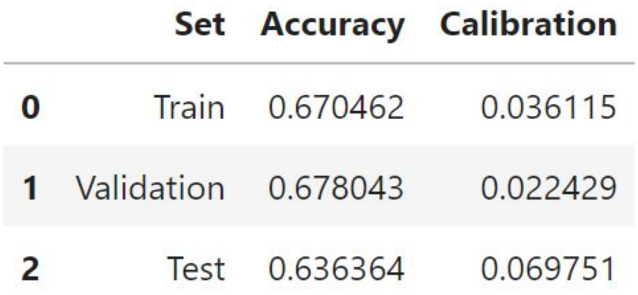
Result:



	Feature	Accuracy	Discrimination
0	Age	0.939913	1.898180e+07
1	Sex	0.673040	2.797407e+06
2	Decile Score	0.926981	1.247128e+07
3	Priors Count	0.765247	1.403305e+07

- 'Age' and 'Decile Score' have the greatest impact on model accuracy,
- 'Age' has the strongest impact on discrimination.
- Dropping 'Age' seems to be a good choice.

Result:



	Set	Accuracy	Calibration
0	Train	0.670462	0.036115
1	Validation	0.678043	0.022429
2	Test	0.636364	0.069751

- **Feature Impact**: Features such as 'Age' were identified to have a significant impact on both accuracy and discrimination. The Shapley values indicated 'Age' contributed greatly to unfair bias.
- **Model Adjustments**: Based on the fairness-utility scores, 'Age' was removed from the features set to reduce bias.
- **Model Performance After Adjustment:** the logistic regression model showed a marginal decline in accuracy, the calibration scores improved, reflecting a reduction in bias and an increase in fairness across racial groups.

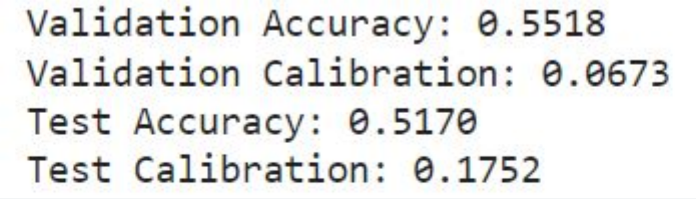
Algorithm 2: Learning Fair Representations

The LFR algorithm is designed to learn fair representations of data that are useful for making predictions while minimizing discrimination or bias related to a sensitive attribute (like race or gender). The algorithm transforms the input features into a new space that makes the sensitive attributes less predictive of the outcome, thereby aiming to ensure fairness in the predictions

Evaluation Process

- **Data Splitting**: The dataset is divided into training, validation, and test sets.
- **Model Training:** The LFR model is trained on the training data.
- **Model Tuning:** Hyperparameters are optimized to achieve the best balance between prediction accuracy and fairness.
- **Prediction and Evaluation:** The trained model is used to predict outcomes on both validation and test datasets. Metrics such as accuracy and calibration are calculated.

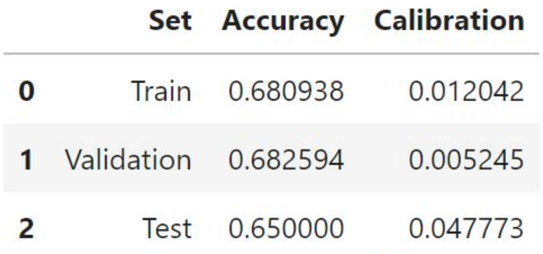
Result:



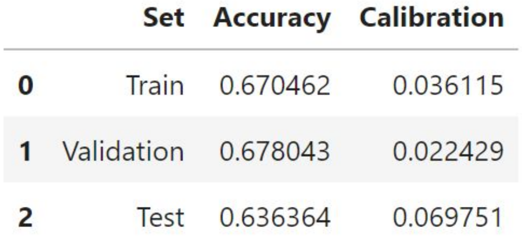
```
Validation Accuracy: 0.5518
Validation Calibration: 0.0673
Test Accuracy: 0.5170
Test Calibration: 0.1752
```

Accuracy: while the LFR model sacrifices some accuracy for fairness, it might still need tuning as the decrease in accuracy is notable.

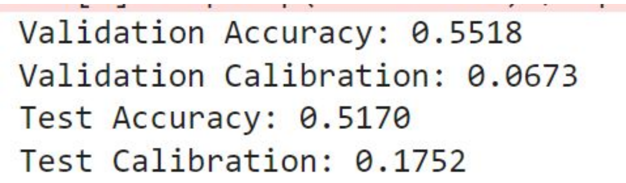
Fairness: the LFR model does not show better calibration compared to the baseline



	Set	Accuracy	Calibration
0	Train	0.680938	0.012042
1	Validation	0.682594	0.005245
2	Test	0.650000	0.047773



	Set	Accuracy	Calibration
0	Train	0.670462	0.036115
1	Validation	0.678043	0.022429
2	Test	0.636364	0.069751



```
Validation Accuracy: 0.5518
Validation Calibration: 0.0673
Test Accuracy: 0.5170
Test Calibration: 0.1752
```

Performance Comparison:

Accuracy: The baseline model is the most accurate across all three data splits (train, validation, test), with Algorithm 1 (FFS) following closely behind. Algorithm 2 (LFR) lags in accuracy on both validation and test sets.

Calibration: The calibration values show that the baseline model is less fair compared to Algorithm 1 (FFS), and Algorithm 2 (LFR) has the highest calibration value, which might suggest it's the least fair among the three according to this specific fairness metric.

Summary

- The baseline model provides the best accuracy.
- Algorithm 1 (FFS) offers a balance between maintaining high accuracy and incorporating fairness, though with a slight decrease in both compared to the baseline.
- Algorithm 2 (LFR) shows a more significant drop in accuracy and fairness based on calibration.

Thank you!