


Project 4: Handling Conditional Discrimination and Information Theoretic Measures for Fairness-Aware Feature Selection



Jackson Zhao, Danielle Solomon, Tianyi Xia, Peng
Jiang, Nicolette Auld-Griffith



Project Summary

Goal: To run two machine learning algorithms on a dataset and maximize predictive accuracy

Algorithm 1: Handling Conditional Discrimination

Algorithm 2: Information Theoretic Measures for Fairness-Aware Feature Selection

Dataset

COMPAS dataset (Correctional Offender Management Profiling for Alternative Sanctions)

<https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis>

- A database containing the criminal history, jail and prison time, demographics, and COMPAS risk scores for defendants from Broward County from 2013 and 2014.
- The ground truth on whether or not these individuals actually recidivated within two years after the screening is also being collected.
- Recidivism is defined as a new arrest within two years.
- ProPublica's analysis shows that the COMPAS risk scores are discriminatory against race and gender.

Handling Conditional Discrimination: Objective

Overall Goal: Instead of relying on the algorithm to deal with the fairness issue, we aim to modify the data before training the model (i.e. pre-processing). The goal is to balance the dataset so that the inherent/unexplainable discrimination is avoided. To achieve this, we used Local Massaging and Preferential sampling.

Local Massaging: Objective is to adjust labels within each age category to mitigate bias and balance the probability of recidivism between males and females.

Preferential Sampling: Objective is to modify the dataset composition by deleting and duplicating instances to create a more balanced dataset that does not reinforce existing biases.

Handling Conditional Discrimination: Data Preprocessing

Input X: X_columns = ['race', 'juv_fel_count', 'decile_score', 'juv_misd_count', 'juv_other_count', 'priors_count', 'days_b_screening_arrest', 'c_days_from_compas', 'c_charge_degree', 'is_recid', 'r_days_from_arrest', 'is_violent_recid', 'score_text', 'v_decile_score', 'v_score_text', 'start', 'end', 'event']

Output Y: ['two_year_recid']

Sensitive Feature: Gender (Female, Male)

Handling Conditional Discrimination: Local Massaging Methodology

- 1) **Calculate Delta Values:** Determine the required number of label adjustments for males and females in each age category to address biases.
- 2) **Train Model:** Use XGBoost to predict the probability of recidivism based on available features, ensuring categorical variables are handled properly.
 - a) Output: Probability of recidivism (``prob_recid``), which is crucial for identifying which instances are close to the decision boundary.
- 3) **Identify Instances Near Decision Boundary:** Use predicted probabilities to find individuals who are closest to the decision boundary.
 - a) Deleted 1/2 of the records for males and females based on their proximity to the decision boundary and duplicating the opposite (correctly labeled ones).
- 4) **Adjust Labels:** Modify labels of selected instances based on delta values to achieve a more balanced representation of outcomes.
 - a) Female Adjustment: Increase the probability of recidivism for females with the highest probability of recidivism who are currently labeled as non-recidivists by changing labels from 0 to 1 for those closest to the decision boundary (``nlargest`` based on ``prob_recid``).
 - b) Male Adjustment: Decrease the probability of recidivism for males with the lowest probability of recidivism who are currently labeled as recidivists by changing labels from 1 to 0 for those closest to the decision boundary (``nsmallest`` based on ``prob_recid``).
- 5) **Evaluate Impact:** Assess changes using both performance metrics (accuracy, F1-score, etc.) and fairness metrics (demographic parity, equality of opportunity) to ensure the adjustments improve fairness without unduly sacrificing accuracy.

Handling Conditional Discrimination: Preferential Sampling Methodology

- 1) **Calculate Delta Values:** Similar to massaging, delta values guide how many instances need to be deleted or duplicated.
 - a) Delete instances for each gender that reinforce biases (`nsmallest` for males and `nlargest` for females, based on `prob_recid`).
 - b) Duplicate instances that correct biases (`nlargest` for males and `nsmallest` for females, based on `prob_recid`).
- 2) **Train Model:** Use XGBoost to predict the probability of recidivism based on available features, ensuring categorical variables are handled properly.
 - a) Output: Probability of recidivism (`prob_recid`), which is crucial for identifying which instances are close to the decision boundary.
- 3) **Identify Instances Near Decision Boundary:** Use predicted probabilities to find individuals who are closest to the decision boundary.
 - a) Deleted 1/2 of the records for males and females based on their proximity to the decision boundary and duplicating the opposite (correctly labeled ones).
- 4) ***Modify Dataset Composition:** Delete and duplicate instances to structurally adjust the dataset, aiming for a composition that reflects reduced bias.
 - a) Delete instances that are likely reinforcing bias (e.g., males predicted not to recidivate with high confidence but actually do, and vice versa for females).
 - b) Duplicate instances that help counteract bias (e.g., males predicted not to recidivate with low confidence but actually do not, and vice versa for females).
- 5) **Evaluate Impact:** Assess changes using both performance metrics (accuracy, F1-score, etc.) and fairness metrics (demographic parity, equality of opportunity) to ensure the adjustments improve fairness without unduly sacrificing accuracy.

Handling Conditional Discrimination: Results

Local Massaging

Accuracy: 98.47%

F1 score: 98.29%

Positive prediction (female): 47.50%

Positive prediction (male): 35.92%

Equal opportunity (female): 1.0

Equal opportunity (male): 1.0

Preferential Sampling

Accuracy: 99.70%

F1 score: 99.60%

Positive prediction (female): 46.24%

Positive prediction (male): 37.06%

Equal opportunity (female): 1.0

Equal opportunity (male): 1.0

Information Theoretic Measures for Fairness-Aware Feature Selection: Objective

Overall Goal: Select features in a fair way that minimizes discrimination while maintaining accuracy. We achieved this by calculating the Shapley Values i.e., the marginal accuracy for each feature to deduce the marginal impact. The Shapley value ensures each feature gains as much or more as they would have from acting independently.

Information Theoretic Measures for Fairness-Aware Feature Selection: Data Preprocessing

Features: sex (Male = 1, Female = 0), age (Age < 25, 25 < Age < 45, or Age > 45), charge_degree (Misdemeanor = 1, Felony = 0), priors_count (0, 1–3, or > 3), length_of_stay (\leq 1 week, \leq 3 months, or > 3 months)

Output Y: ['two_year_recid']

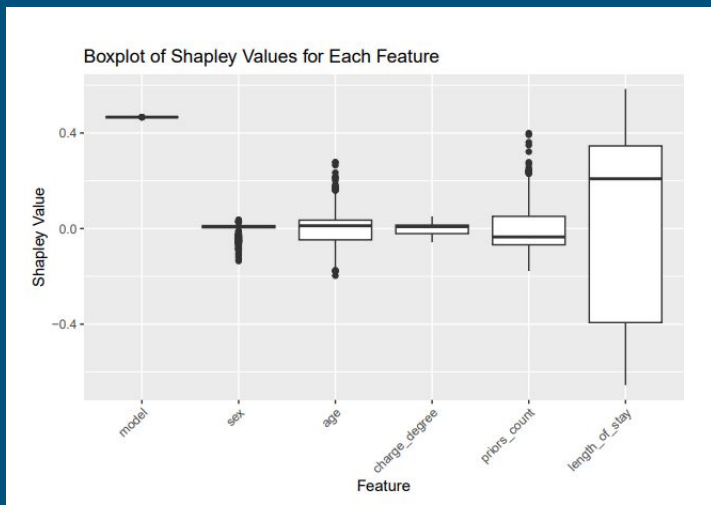
Sensitive Feature/Protected Attribute: Race (African American = 0, Caucasian A = 1)

Information Theoretic Measures for Fairness-Aware Feature Selection: Methodology

- 1) **Test/Training Data:** We randomly split the whole dataset into training and test subsets. We also organized our data into the protected group and not protected group.
 - a) Protected group: sensitive variable = 1 → Caucasian
 - b) Not protected group: sensitive variable = 0 → African American
- 2) **Train Model:** We then trained a classifier with the all features as the input, and prediction of Y as the output
 - a) Used random.forest and XGBoost
- 3) **Accuracy:** Applied our measures of accuracy to this dataset to determine which features exhibit the strongest proxies for discrimination
- 4) **Shapley Calculation:** Calculated Shapley values to determine the impact of each feature

Information Theoretic Measures for Fairness-Aware Feature Selection: Results

- **Features with least impact:** Charge degree and gender
 - least informative features for the prediction task
- **Strongest proxies for discrimination:** Age, priors count and length of stay
- Highest level of accuracy for the protected group



##	Random.Forest	XGBoost
## Model Accuracy	0.926016260	0.88455285
## Protected	0.929521277	0.90026596
## Not Protected	0.920502092	0.85983264
## Calibration	0.009019185	0.04043332