12 March, 2024

UC San Diego

CSE-291J Assignment – 2

Agenda

1. Evaluation of the fairness and bias involved in the ACSPublicCoverage dataset
2. Analysed the dataset
3. Defined the different fairness metrics
4. Performed different intervention strategies

Dataset Description

- Dataset: U.S. Census microdata
- Where does it come from/How was it collected?
    - Dataset is derived from the U.S. Census Bureau's American Community Survey (ACS).
- What is it for?
    - To help ML research build models related to demographics, employment, income, education, and other societal metrics.
    - Has been processed into ML-ready formats
- How to access?
    - Folktables Library

Exploratory Data Analysis

Code-Explanation

Dataset Details

1 year Data of 2018 census

States Considered: 5 ("CA","MA","TX", "NY", "GA")

Number of features: 19

Number of rows: 364726

Feature Analysis

Sensitive Attributes:

1. Race
2. Marital Status
3. Disability
4. Sex

Non-Sensitive Attributes:

1. Schooling status
2. Military status
3. Principal income
4. State
5. Citizenship status
6. Nativity

and more

Sensitive Feature Analysis

1. Sex: ‘Male’, ‘Female’
2. Race: 'White alone', 'Black or African American alone', 'American Indian alone', 'Alaska Native alone', 'American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races', 'Asian alone', 'Native Hawaiian and Other Pacific Islander alone', 'Some Other Race alone','Two or More Races'
3. Marital Status: 'Married', 'Widowed', 'Divorced', 'Separated', 'Never married or under 15 years old'
4. Disability: 'With a disability', 'Without a disability'

Fairness Metrics & Relevance

1. TPR Difference
2. Demographic Parity
3. Equalised Odds
4. PPV
5. Demographic Mistreatment*

*Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment, Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi

Fairness Metrics & Relevance Contd..

Metric	Relevance to the dataset
TPR Difference	Checks if all groups have equal access to public coverage classification
Demographic Parity Difference	Ensures all groups have equal likelihood of being classified as covered
Equalized Odds	Ensures model does not favour any group unfairly
PPV	Ensures that positive predictions are equally reliable across groups
Disparate Mistreatment	Prevents systematic misclassification of vulnerable groups

Correlation between Sensitive attributes

How much can “unfairness” in your predictions be explained by dataset characteristics? Can you fix them with dataset-based interventions?

Correlation between Sensitive attributes and target variables

Baseline Models

Code- Explanation

Dataset Intervention

Dataset Intervention : Massaging

Modifying a dataset or model to improve its fairness

Steps Followed:

1. Measure metrics across sensitive groups
2. Rank instances based on the bias
3. Modify labels to reduce disparities
4. Retrain classifier on the massaged data
5. Re-evaluate Fairness

Dataset Intervention : Oversampling

Steps followed

1. Analysed the counts of data pertaining to each category
2. Then sampled in the way that each category has equal counts
3. Picked the Train and Test split from the resampled data
4. Then used this new resampled data with each classifier model
5. Re-evaluated all the fairness metrics

Model Intervention

Post processing Intervention

Post Processing Intervention

Models considered: Logistic Regression, Decision Tree, Random Forest, XGBoost

Steps Followed:

1. Train different classification models
2. Adjusts decision thresholds for different demographic groups
3. Compare performance across different thresholds

Experiments performed: 0.5, 0.3, 0.7, 60th Percentile, 90th Percentile

Comparison between Interventions

Comparison Analysis : Logistic Regression

Metrics	Baseline	Dataset Intervention	Model Intervention	Post Processing Intervention
Accuracy	74%	72%	71%	63%
Demographic Parity Difference	78.16%	70.50%	17.8%	0.006%
TPR Difference	82.59%	62.70%	17.28%	6%
Disparate Mistreatment Difference	5.08%	9.40%	21.32%	2.73%
Equalised Odds Difference	17.89%	1.88%	10.9%	6.88%

Comparison Analysis : Decision Tree

Metrics	Baseline	Dataset Intervention	Model Intervention	Post Processing Intervention
Accuracy	72%	99%	77%	52%
Demographic Parity Difference	39.42%	37.83%	54.80%	18.38%
TPR Difference	35.70%	2.53%	55.35%	18.19%
Disparate Mistreatment Difference	0.99%	1.49%	0.05%	19.36%
Equalised Odds Difference	16.59%	3.42%	25.41%	12.28%

Comparison Analysis : Random Forest

Metrics	Baseline	Dataset Intervention	Model Intervention	Post Processing Intervention
Accuracy	77%	99%	-	72%
Demographic Parity Difference	49.17%	37.96%	-	0.07%
TPR Difference	47.38%	2.56%	-	10.6%
Disparate Mistreatment Difference	2.93%	1.38%	-	1.35%
Equalised Odds Difference	25.68%	3.26%	-	12.53%

Comparison Analysis : XGBoost

Metrics	Baseline	Dataset Intervention	Model Intervention	Post Processing Intervention
Accuracy	79%	90%	78%	74%
Demographic Parity Difference	53.55%	43.88%	55.32%	0.006%
TPR Difference	52.13%	25.56%	55.9%	15.19%
Disparate Mistreatment Difference	1.47%	10.05%	1.57%	1.60%
Equalised Odds Difference	26.07%	21.51%	28.44%	5.43%

Observations

- What types of interventions are most appropriate for your task (e.g. legal, practical to deploy, etc.)?
  - Massaging is not practical for a large dataset
  - Oversampling increases the final dataset size
  - Post processing intervention has legal problems
- What are the tradeoffs between them (e.g. how are other metrics negatively impacted by a particular intervention, etc.)
  - Overuse of dataset interventions might introduce biases themselves

Research Paper

Research Paper

Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi

Research Paper

Contributions of the paper for fairness evaluation:

- Propose **disparate mistreatment** to measure fairness based on different misclassification rates across groups
- Misclassification rate is defined as:

$$P(\hat{y} \neq y | z = 0) \neq P(\hat{y} \neq y | z = 1)$$

- Incorporate this into model training

Research Paper

Contributions of the paper for fairness Improvement:

- Propose a formal framework to incorporate fairness constraints into decision boundary-based classifiers
- They propose a convex optimization approach to minimize differences in false positive and false negative rates between protected and non-protected groups.
- Allows for fairness without explicitly using sensitive attributes in decision-making, addressing legal and ethical concerns
- Integrated with Classifiers like: Logistic Regression, XGBoost using COMPAS dataset
- The approach improves fairness, it introduces minor accuracy trade-offs

Observations

Is it more effective than other intervention strategies you tried? Why or why not?

- Effective and appropriate
- Minimal compromise on accuracy
- Fairness improvement over basic intervention techniques

Conclusion

- We considered baseline models that were highly biased
- Dataset intervention was best for Accuracy and Some Fairness
- Post-Processing Intervention was the Most Effective but might not be always ethical
- XGBoost performed the best amongst all the models considered
- The research paper implemented has been the best performance we achieved


Thank you!