xD

xD.gov

United States Census Bureau

Explainable Artificial Intelligence for Bias Identification and Mitigation in Demographic Models

09.18.23

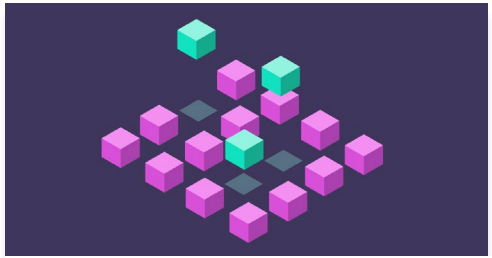
Atul Rawal, Ph.D.
Sandy L. Dietrich, Ph.D.
James McCoy

Outline

- xD Overview
- XAI Overview
- Need for XAI
- Demographics Overview & Bias in Demographics
- XAI for Demographic Research
- Use Cases
  - Example # 1 – Binary Classification
  - Example # 2 – Multiclass Classification
- Summary

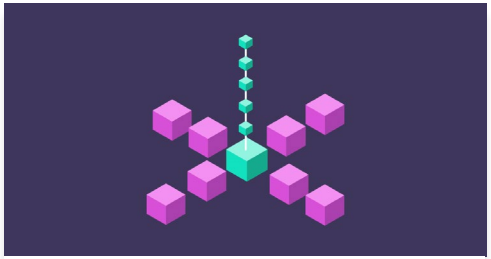
xD Overview

xD is an emerging technologies group that's advancing the delivery of data-driven services through new and transformative technologies.



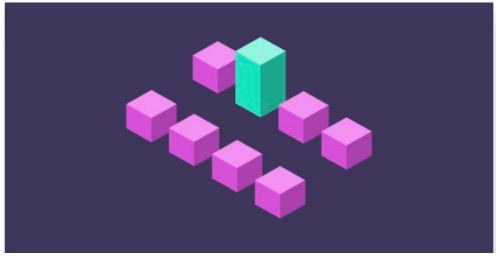
Responsible AI (RAI)

- AI/ML for labeling & classification of geographic data saving ~800,000 hours of manual labeling
- XAI & Causal learning for bias identification in geographic data
- Model Card Generator & AI Register
- Bias Toolkit



Privacy-Enhancing Technologies (PETs)

- UN Pilot for Secure Multi Party Computation
- Remote execution and Federated Learning
- Inter-Agency Multi Party Computation

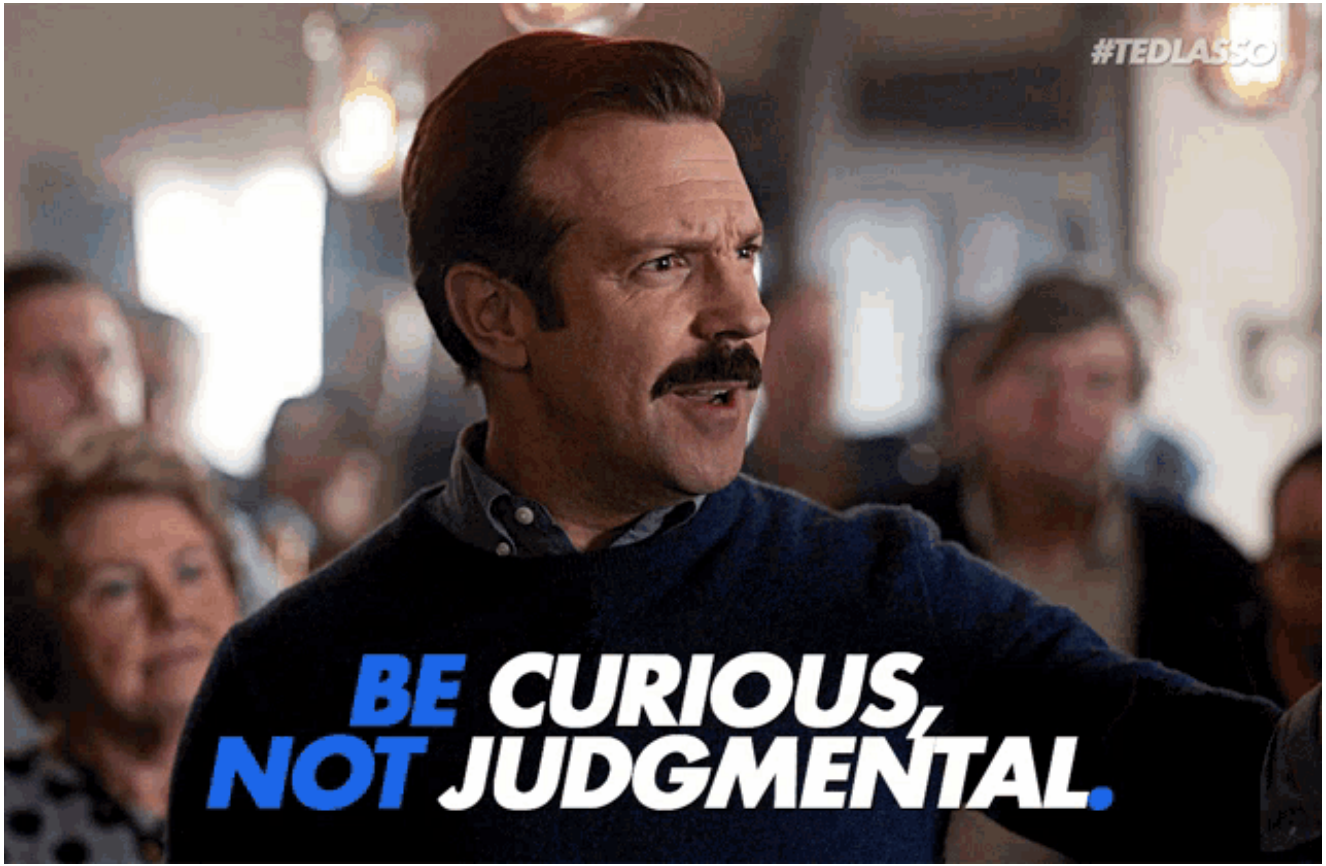


Incubation & Transformation

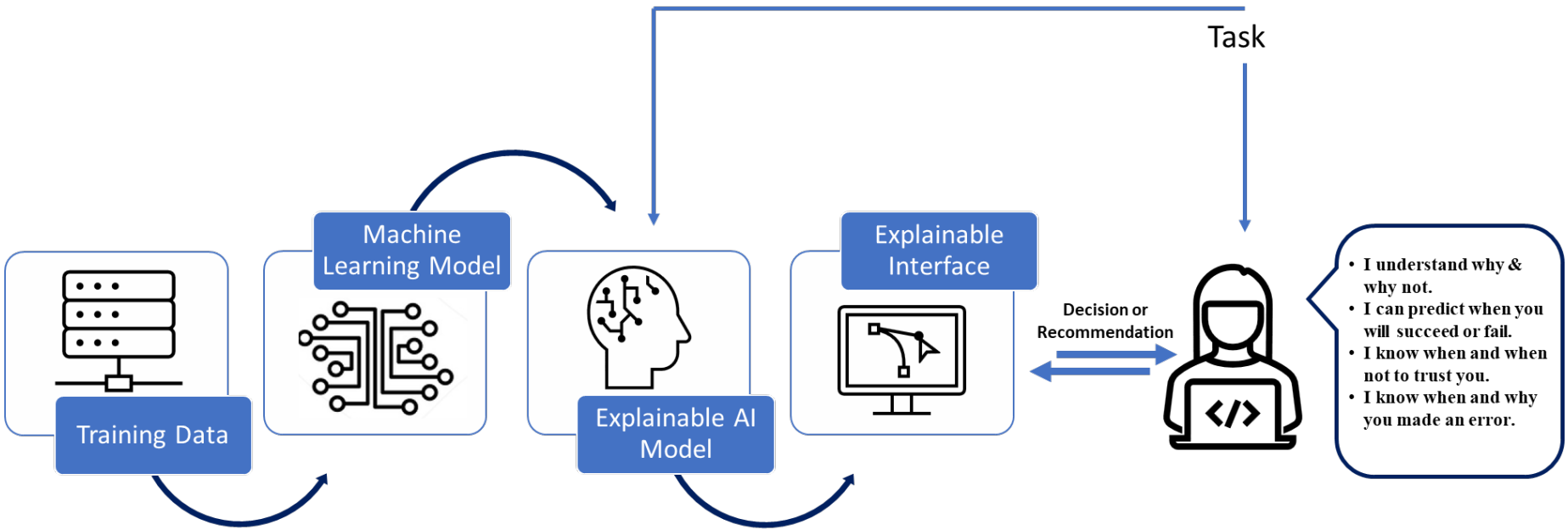
- Developer Experience at Census
- Bias in Infrastructure
- DAO for Equitable Government participation
- Privacy Preserving Record Linkage for Health

When AI Goes Wrong

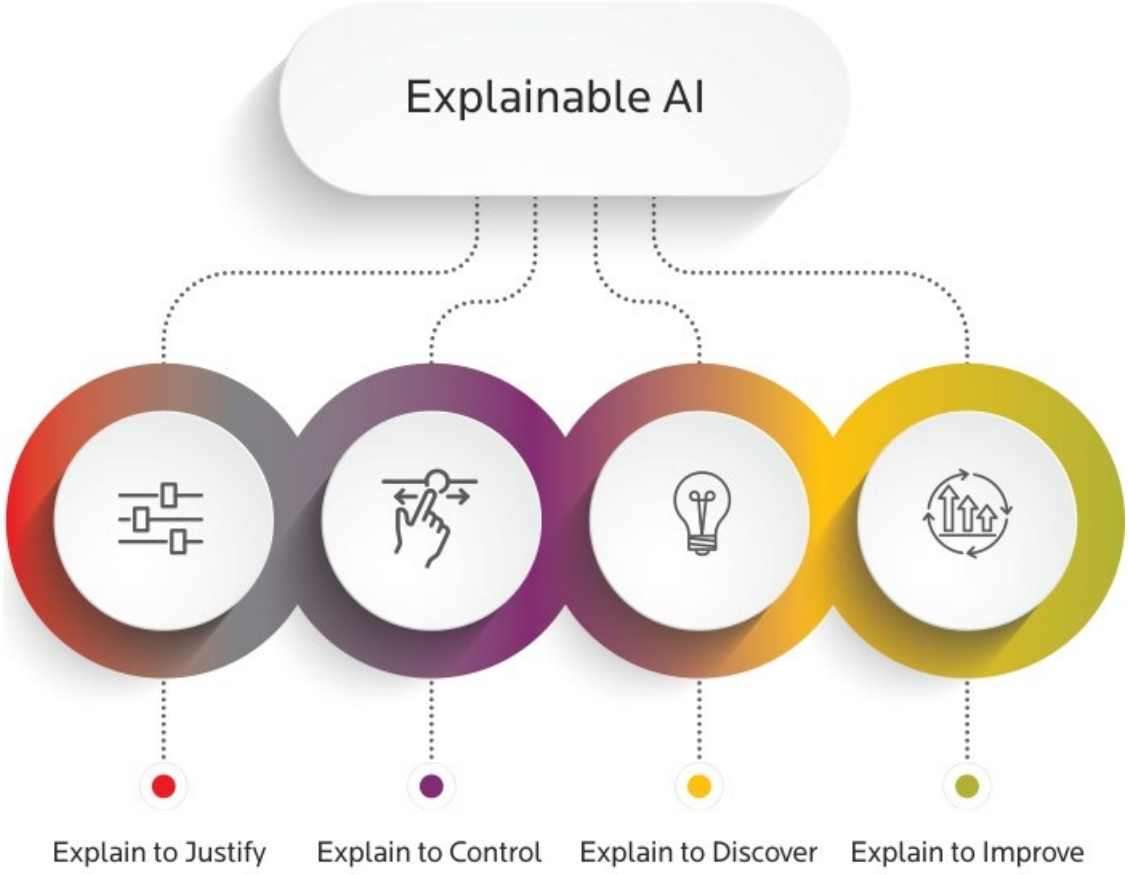
Motivation



XAI Overview

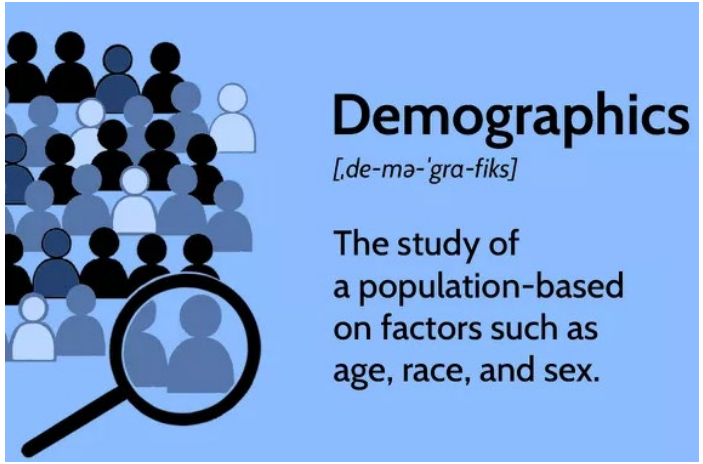


Why is XAI Needed?



Demographics

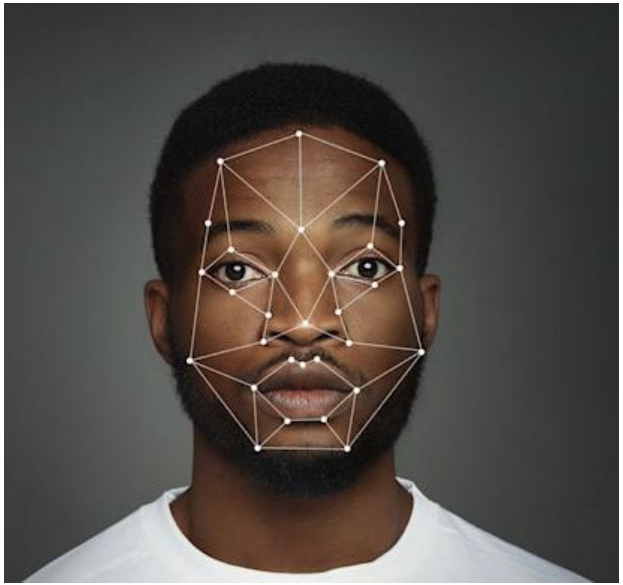
- Demographic Analysis is the collection and analysis of broad characteristics about groups of people & population.
- Statistics that describe populations and their characteristic.
  - Sensitive Attributes
  - Potential for Bias
- The combination of the internet, big data, and artificial intelligence is greatly amplifying the usefulness and application of demographics as a tool for marketing and business strategy.



Demographic Bias

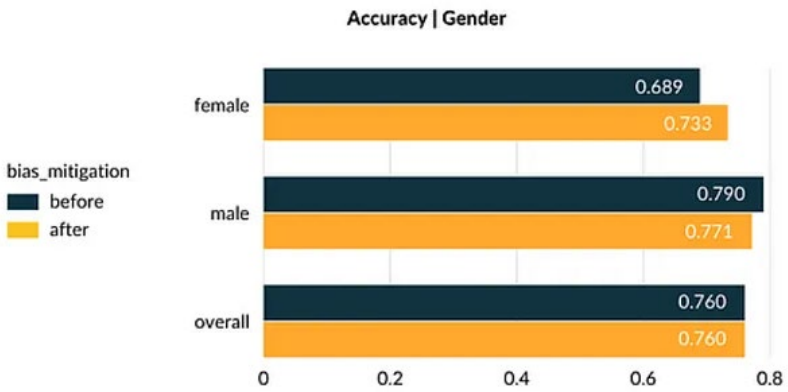
- Unequal representation or treatment based on demographic factors.
- Can be detrimental and misleading in crucial domains, like medical diagnostics and treatment plans.
- Example :

  Machine learning model suggests varied treatments for two patients solely due to their demographic details, even when they have the same medical condition.



XAI for Demographic Models

- Explaining/interpreting predictions, and recommended actions to stakeholders.
- Aims to create more understandable, interpretable, and reliable models, by improving the quality of predictions.
- Bias identification & mitigation.



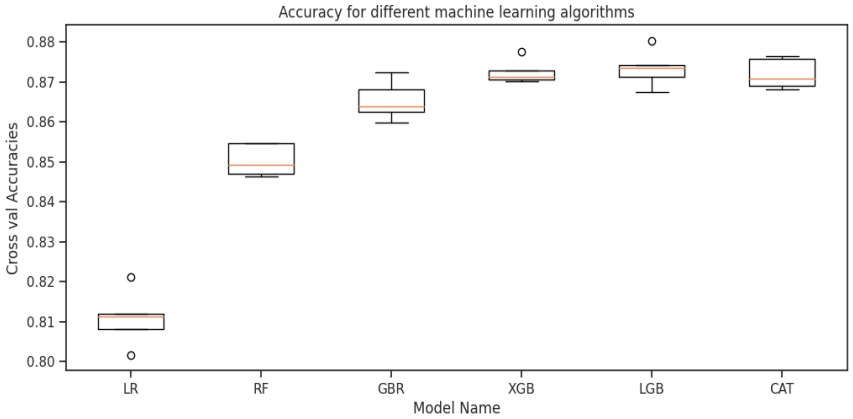
Use Case Examples

- Example 1 – 36,000 Samples
  - Binary Classification.
  - 1994 Census database for prediction of individual income of </> 50K.
  - Features: Age, Sex, Race, Education, Occupation.
- Example 2 – 500,000 Samples
  - Multiclass Classification.
  - 2015 – 2019 ACS data for prediction of individual income categories
    - <$25 K, $25K-$50K, , $50K-$100K, $100K-$150K, and >$150K
  - Features: Age, Sex, Race, Ethnicity, Martial Status, Education, Occupation, Citizenship, Employment Status, Language, Years in the US, English.

Example # 1

- Six ML models for income classification.
  - Classes: Above or below $50K

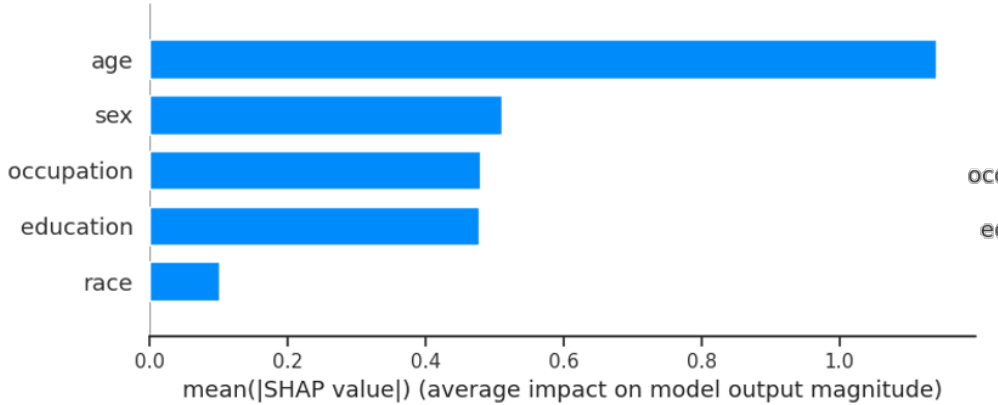
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.8108	0.80	0.80	0.80
Gradient Boosting	0.8653	0.86	0.86	0.86
Light Gradient Boosting	0.8734	0.87	0.87	0.87
CAT Forest	0.8716	0.89	0.89	0.89
Extreme Gradient Boosting	0.8752	0.89	0.89	0.89
Random Forest	0.8504	0.98	0.98	0.98

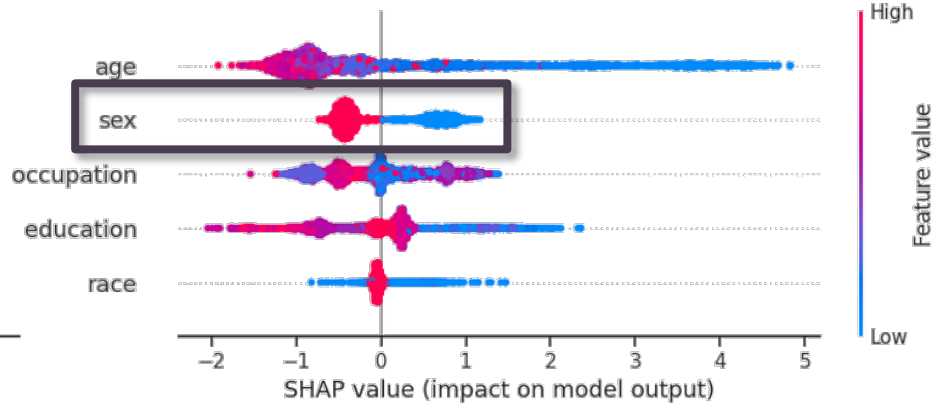


Example # 1

Explainability via feature relevance & beeswarm plot.

- Feature relevance highlights the bias in age, sex.
- Beeswarm highlights the higher values (females) in red are less likely to earn over 50K in 1964 than males.

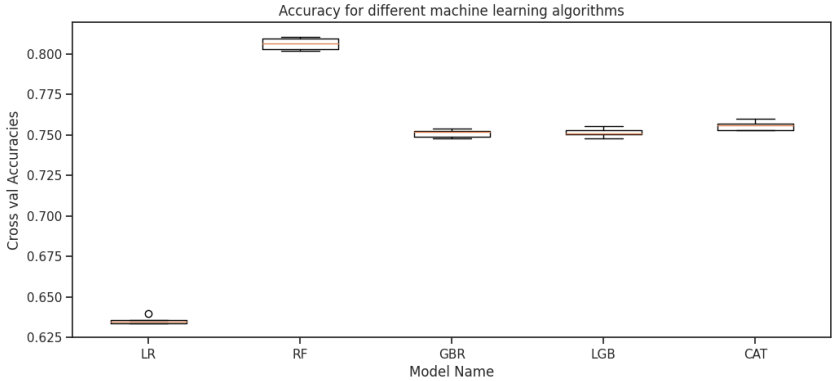




Example # 2

- Five ML models for income classification.
  - Classes: <$25 K, $25K-$50K, , $50K-$100K, $100K-$150K, and >$150K

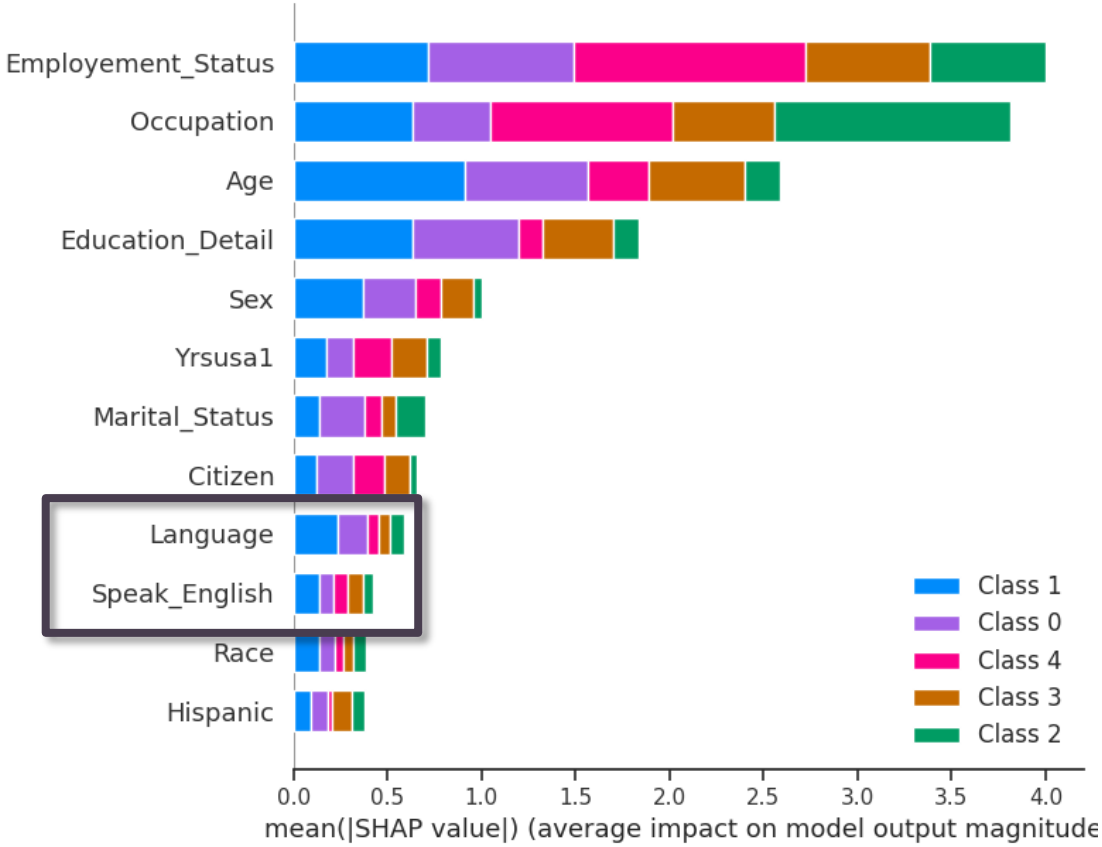
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.6355	0.46	0.64	0.52
Gradient Boosting	0.7510	0.78	0.78	0.78
Light Gradient Boosting	0.7514	0.78	0.78	0.78
CAT Boosting	0.7556	0.79	0.80	0.79
Random Forest	0.8062	0.99	0.99	0.99



Example # 2

Ranked list of features based on impact on the income classification.

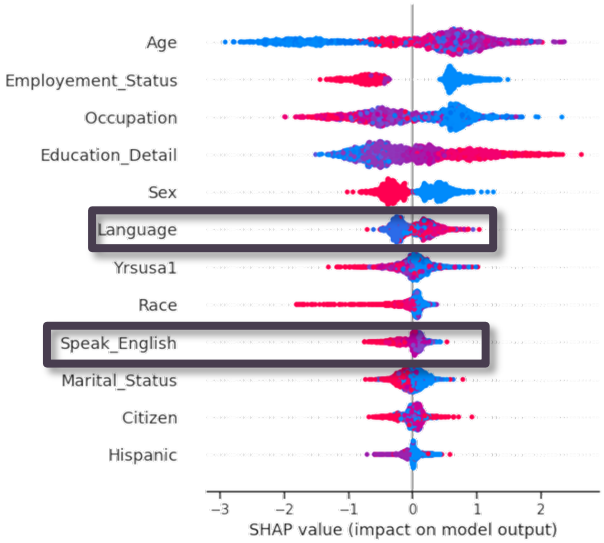
- Features can have varying impact on the different income classes.



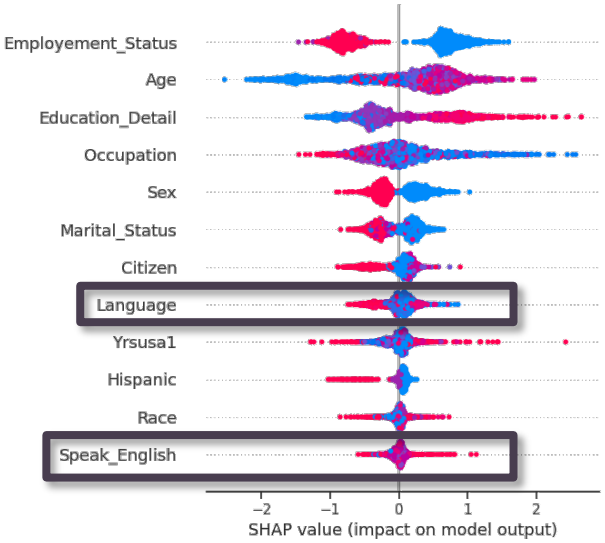
Example # 2

Beeswarm plots highlight potential bias in the spoken language as English speakers are:

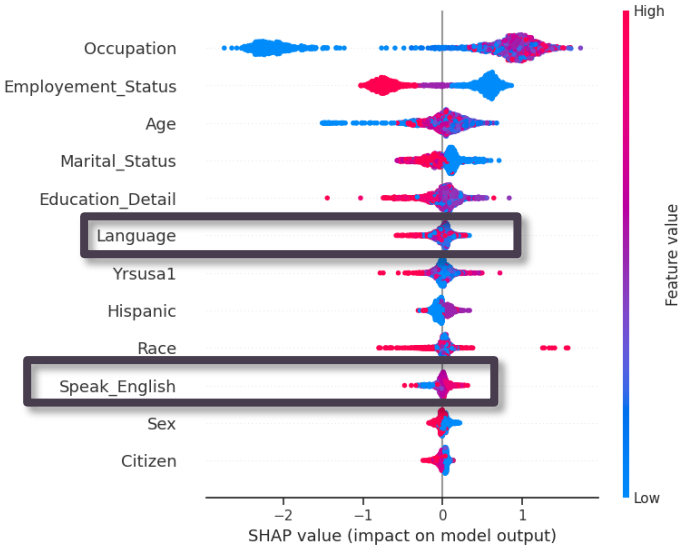
- Less likely to earn less than $25K.
- More likely to earn $25K - $50K than other individuals who speak other languages.
- More likely to earn $50K - $100K than other individuals who speak other languages.



< $25K



$25K - $50K



$50K - $100K

Summary

- Utilization of AI in demographic applications is increasingly vulnerable to bias scrutiny.
- The incorporation of XAI as a must-have feature in demographic use of AI will help alleviate bias.

Questions/Discussion

Get in touch
atul.rawal@census.gov
xD.gov

Atul Rawal, Ph.D.

Education

Albright College, B.Sc. in Theoretical Physics
Joint School of Nanoscience & Nanoengineering, Ph.D. in Nanoengineering
Howard University, Ph.D. in Electrical Engineering (On Leave)
Towson University, Ph.D. in Computer Science (Fall 2024)

Expertise

Artificial Intelligence/Machine Learning, Explainable AI (XAI), Causal learning, Computational Biology, Molecular Dynamics (MD), Quantum Mechanics (QM), Protein Engineering/Dynamics.

Hobbies

Soccer, Lacrosse, Motorcycles, Hiking/Chasing waterfalls with Red. Started a journey to visit all 63 National Parks by the time I turn 40.