Machine bias in Mortgage

Jackson Hu, Maria He, Jiayu Wang, Yijia Yang, Yuning Zhao

Problems with AI Bias

ML models learn to use data to:

- Make decisions
- Perpetuate the original biases in the data
- Fairness-accuracy trade-off

Problems with AI Bias

Facebook's ad-serving algorithm discriminates by gender and race

- Amazon abandoned a project to build an AI recruitment tool, which engineers found was discriminating against female candidates.

Themis-ML

Aims to reduce bias but retain accuracy. It modifies the machine learning pipeline at various stages for discrimination detection and fairness in predictions.

Table 1: A Simple Classification Pipeline

API Interface	Function	Examples
Transformer	*Preprocess* raw data for model training.	mean-unit variance scaling, min-max scaling
Estimator	*Train* models to perform a classification task.	logistic regression, random forest
Scorer	*Evaluate* performance of different models.	accuracy, f1-score, area under the curve
Predictor	*Predict* outcomes for new data.	single-classifier prediction, ensemble prediction

Methods used to reduce machine bias

1. Remove Protected Attribute (RPA)
2. Relabel Target Variable (RTV)
3. Additive Counterfactually Fair Model (ACF)
4. Reject-option Classification (ROC)

Background on the dataset

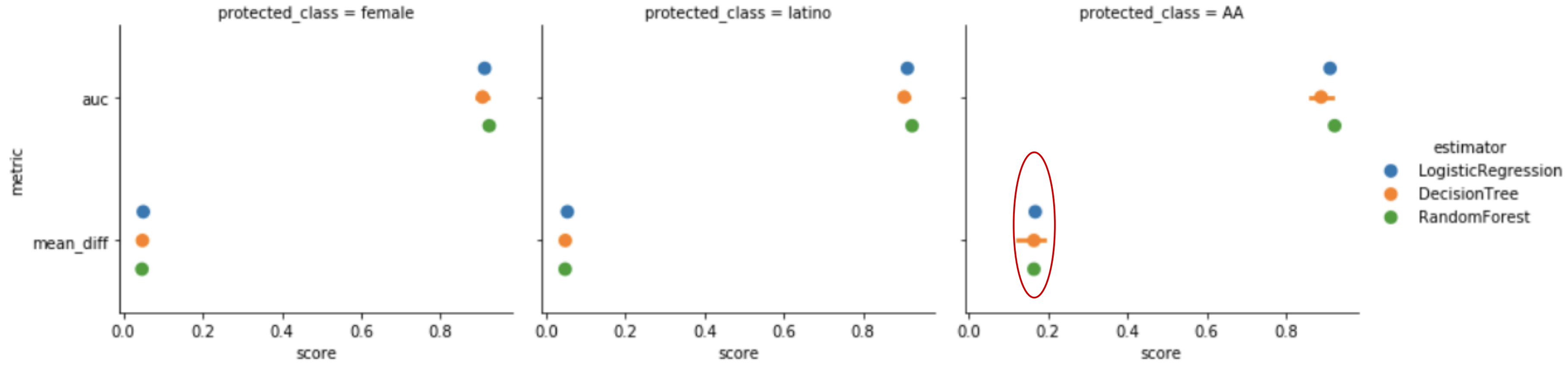
A research-ready data set of U.S. home mortgage loan applications, based on data from the federally mandated Home Mortgage Disclosure Act. In 2014, there were about 11.7 million loan records. These records include applications for home purchase, for home improvement, and for refinancing.

We explore this US national mortgage dataset and build predictions for loan approvals or denials.

Data Cleaning

- Subset data with random sampling
- Drop repeated columns
- Get dummies of categorical columns
- Come up with variables that tend to create bias:
  - Gender
  - Ethnicity
  - Race

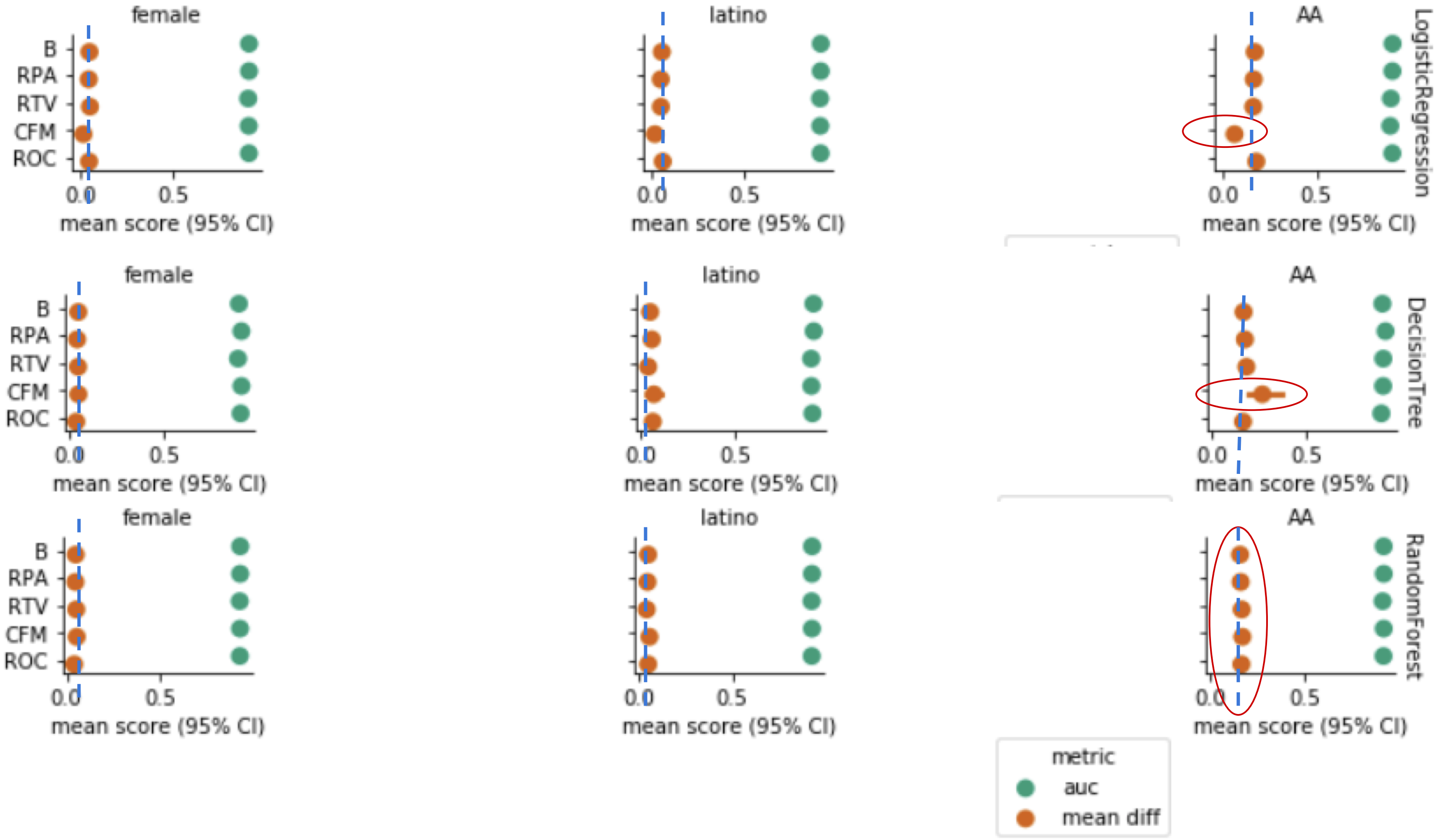
Establish Baseline Metrics



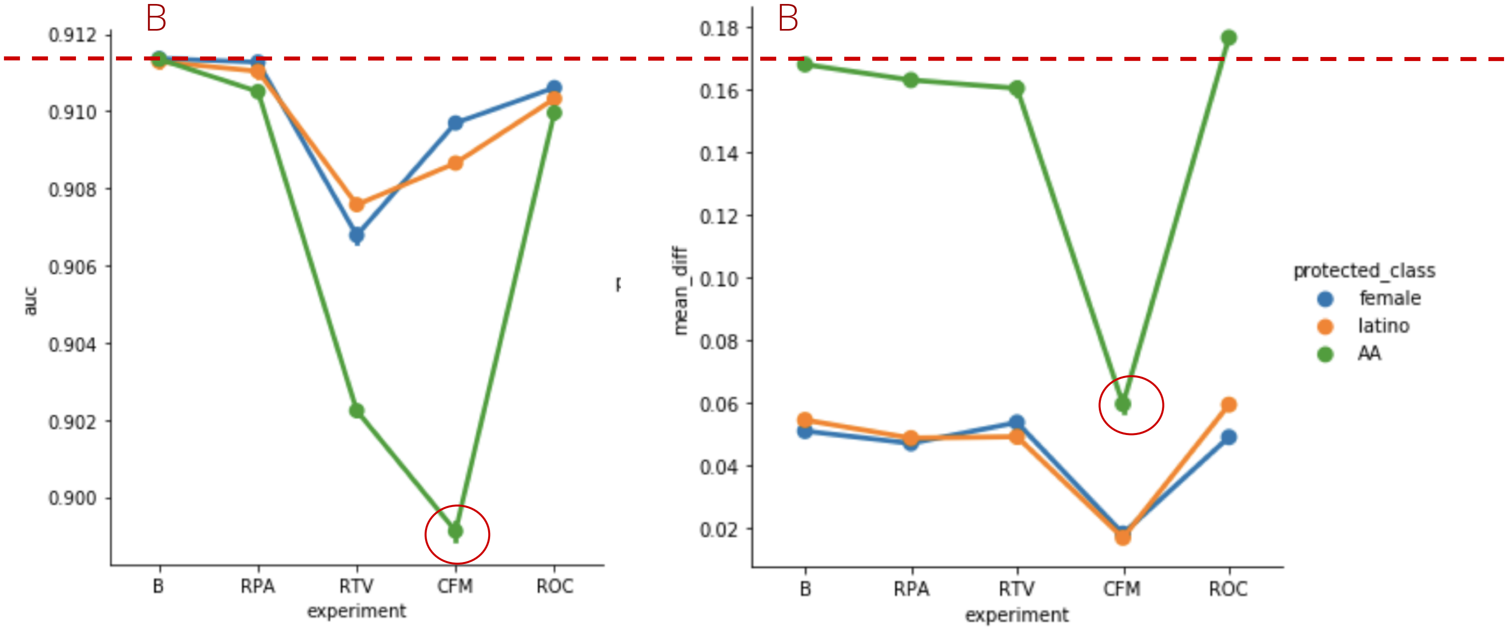
Female
```
mean difference: 3.84 - 95% CI [3.74, 3.94]
normalized mean difference: 5.33 - 95% CI [5.23, 5.43]
```
Latino
```
mean difference: 7.08 - 95% CI [6.93, 7.23]
normalized mean difference: 8.11 - 95% CI [7.96, 8.26]
```
AA
```
mean difference: 18.19 - 95% CI [18.02, 18.37]
normalized mean difference: 21.51 - 95% CI [21.34, 21.68]
```

Groups of Interest - Protected Class

- **protected class = female:**
    - best utility measured by auc (higher is better) = 0.923: RandomForest
    - best fairness measured by mean_diff (lower is better) = 0.047: RandomForest
- **protected class = latino:**
    - best utility measured by auc (higher is better) = 0.923: RandomForest
    - best fairness measured by mean_diff (lower is better) = 0.048: RandomForest
- **protected class = African American:**
    - best utility measured by auc (higher is better) = 0.923: RandomForest
    - best fairness measured by mean_diff (lower is better) = 0.156: RandomForest


female
latino
AA
LogisticRegression
DecisionTree
RandomForest
B
RPA
RTV
CFM
ROC
0.0
0.5
mean score (95% CI)
metric
auc
mean diff

Findings


B
B
0.912
0.910
0.908
0.906
0.904
0.902
0.900
auc
B
RPA
RTV
CFM
ROC
experiment
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
mean_diff
B
RPA
RTV
CFM
ROC
experiment
protected_class
female
latino
AA

Learnings & Recommendations

- Learnings:
  - Method result varies
  - Trade-off
- Recommendations:
  - Collect more features
  - Run models on larger dataset

Limitation

- “Fairness”
  - Relabelling
- Limitations within bias adjusting methods
  - ACF
  - ROC

Q&A

Thank You!

LogisticRegression DecisionTree RandomForest