Student Dropout Prediction

Rui Chen, James Fantin, Kun Yi

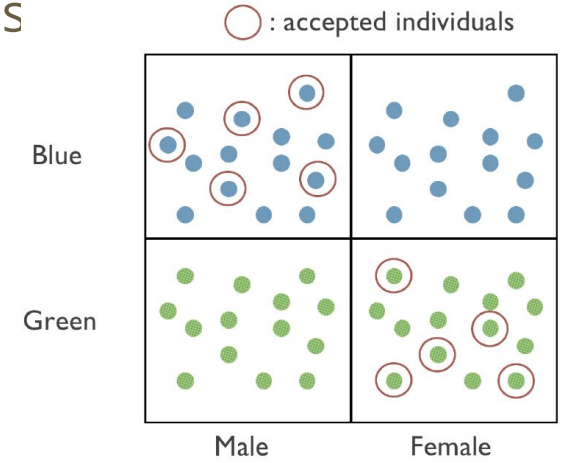
Reminder

We want to use machine learning to predict if a student will drop out or not from a University.

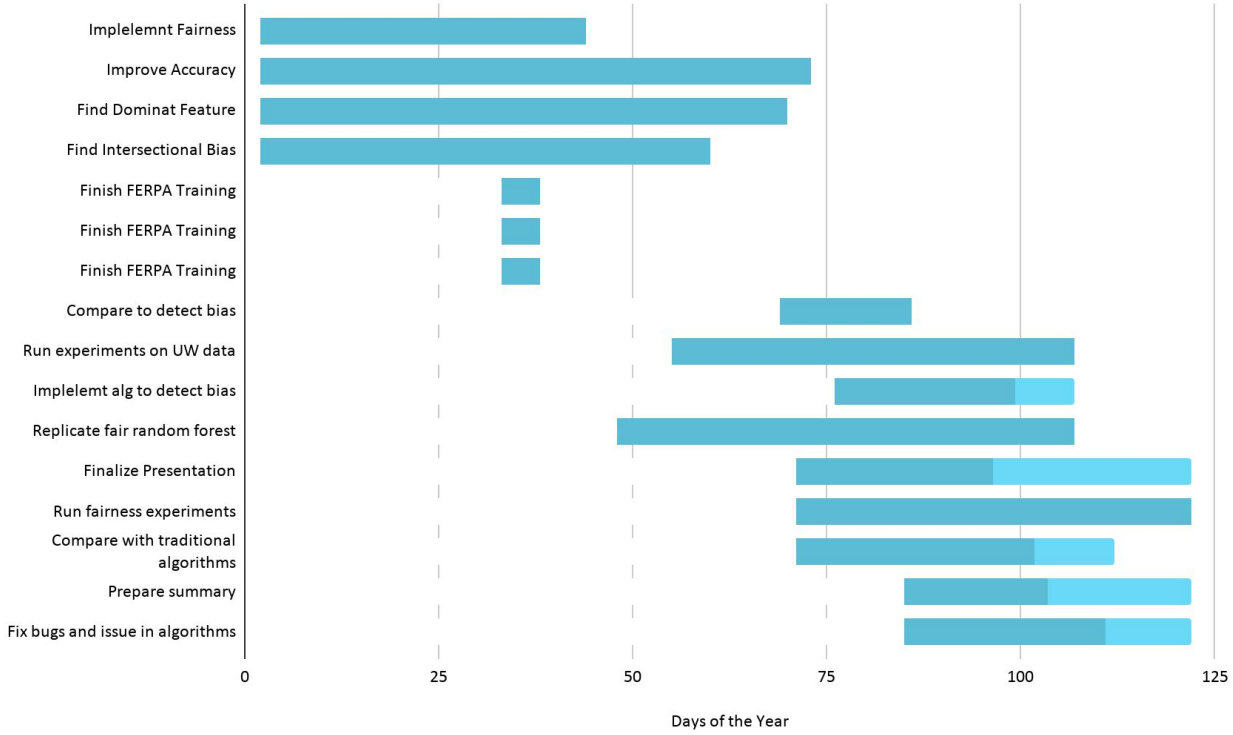
Interested in protecting fairness for demographic information about students.

Clarifications From Last Time

- What is bias?
  - An algorithm is biased if a feature such as gender affects the prediction
- What is intersectional bias?



What we have accomplished?



What we have accomplished?

James:

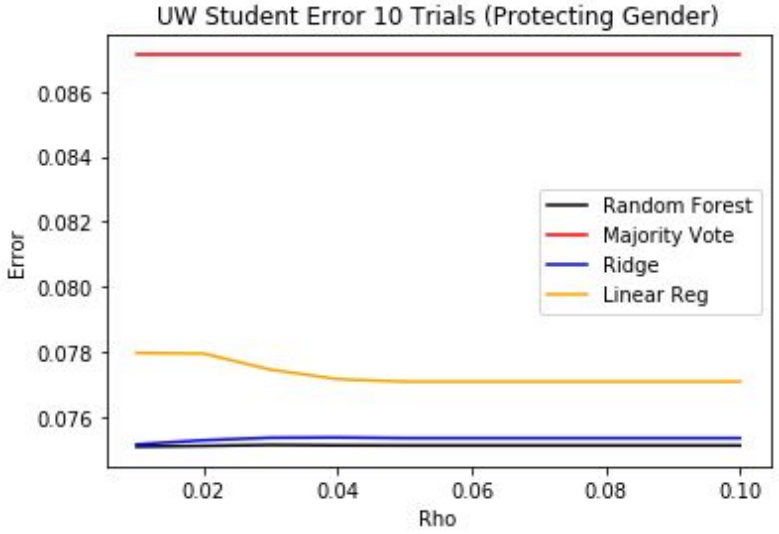
- Finished building competing fair forest algorithms
- Finished our distributed fair random forest algorithm
- Compared our algorithm with the existing fair forest algorithms

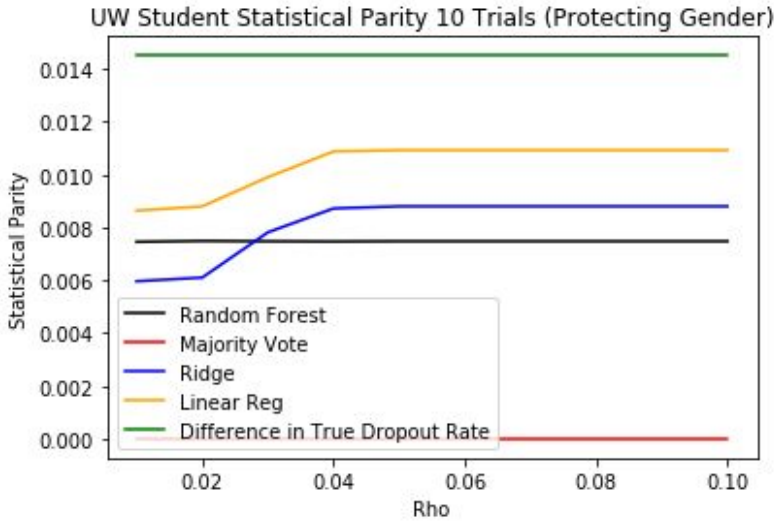
What we have accomplished?

Our Distributed Fair Random Forest Algorithm

- Assume a third party hold private demographic data
- A data center holds remaining data and builds a model
  - Build completely random decision trees
  - Third party determines which ones are fair
  - The fair trees are weighted based on accuracy to make decisions

Current Results





Current Results On UW Data

DFRF is our Distributed Fair Random Forest Model

FRF, FDT, FAHT are competing fair algorithms

RF is a standard random forest

Model	Error Rate	Statistical Parity
DFRF	0.075172969	**0.005964156**
FRF	**0.074395516**	0.046874894
FDT	0.119231385	0.075656425
FAHT	0.084257619	0.026245917
RF	0.075104586	0.007457310

What we will accomplish

James

- Finish optimizing a few algorithms for our data
- Work on final presentation and setting up demo

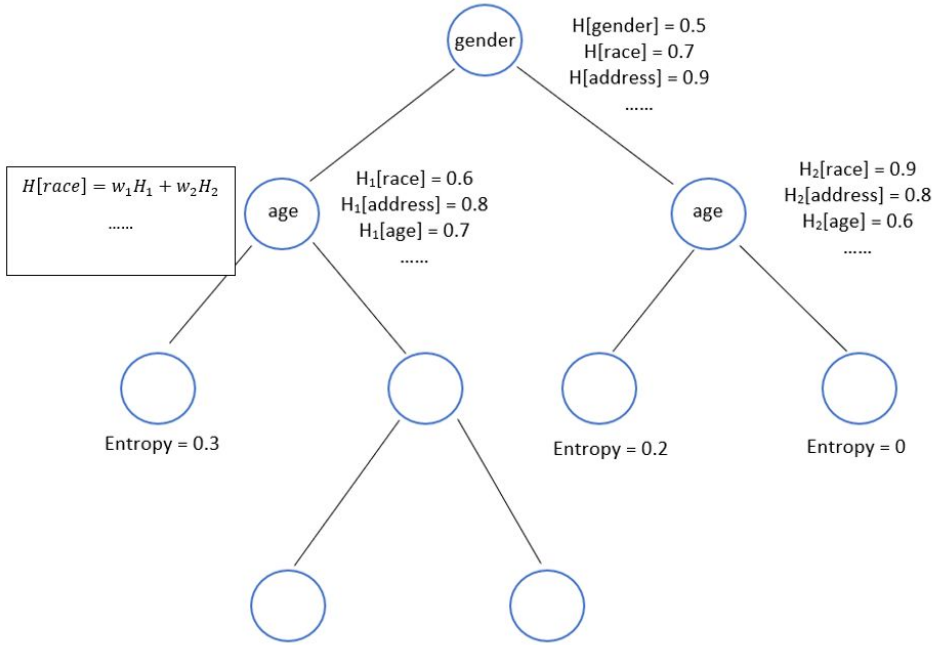
What we have accomplished?

Rui

- Found a new criteria to decide feature in each level: Weighted addition
  - Nodes may have different sizes
- Verified the hypothesis of result of decision tree detector

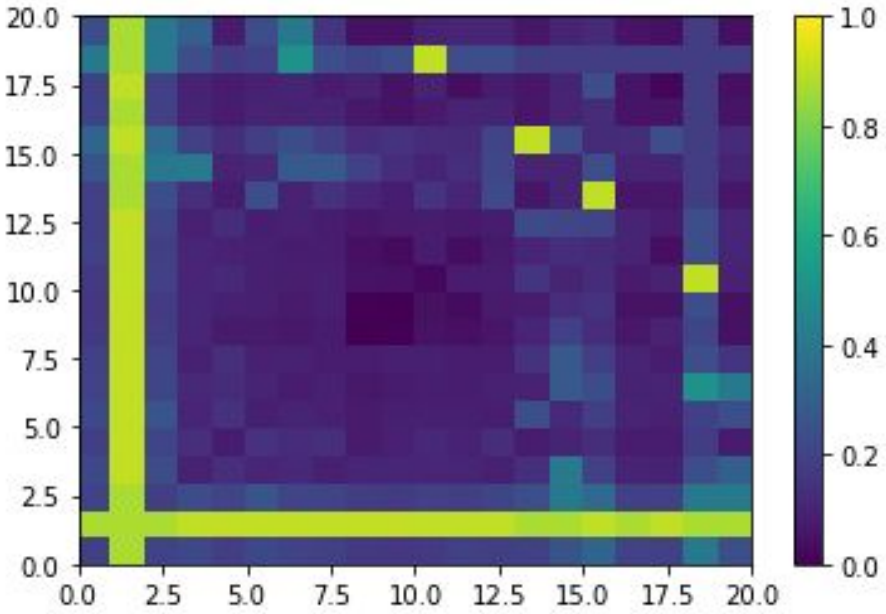
$$D(f_i, f_j) \geq D(f_a, f_b)$$
$$if$$
$$i, j \leq a, b$$

where *D* is disparity between two features.



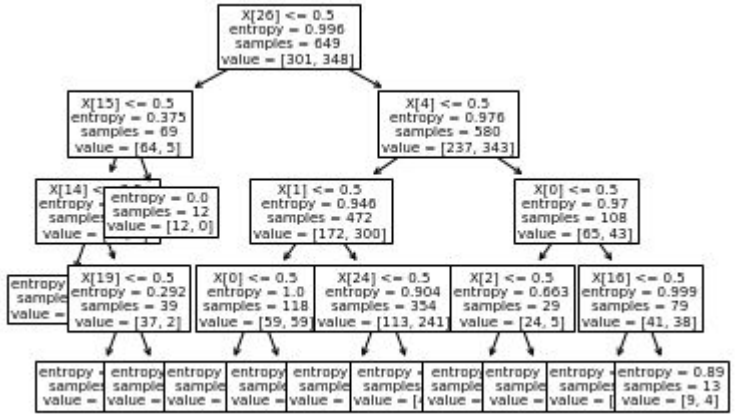
Current Result

Colormap with first 20 features



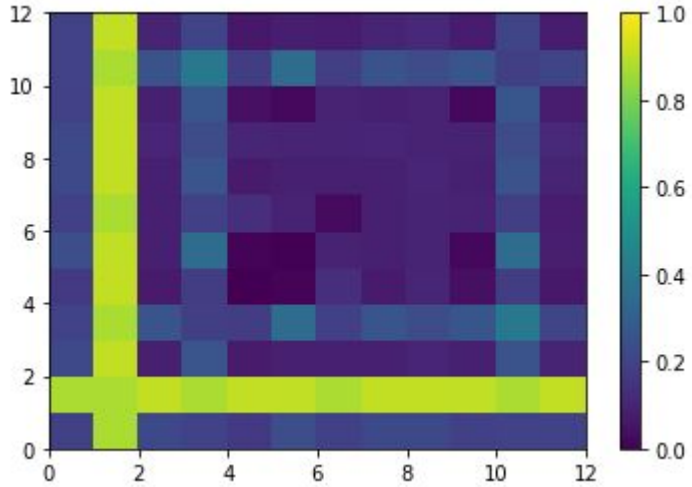
Standard decision tree VS. decision tree detector

- Standard decision tree
  - Use weighted entropy for splitting criteria
  - Different feature in one level
  - Features can be reused
  - Result:
- Decision tree detector
  - Weighted entropy for splitting criteria
  - Same feature in one level
  - Features cannot be reused
  - Result top 10 features: [26, 4, 0, 24, 1, 9, 14, 22, 2, 19]



Current Result

Colormap result from Standard Decision Tree:



What we will accomplish

Rui

- Keep researching other splitting criteria to get better result satisfy our hypothesis
- Work on final presentation and setting up demo

What we have accomplished?

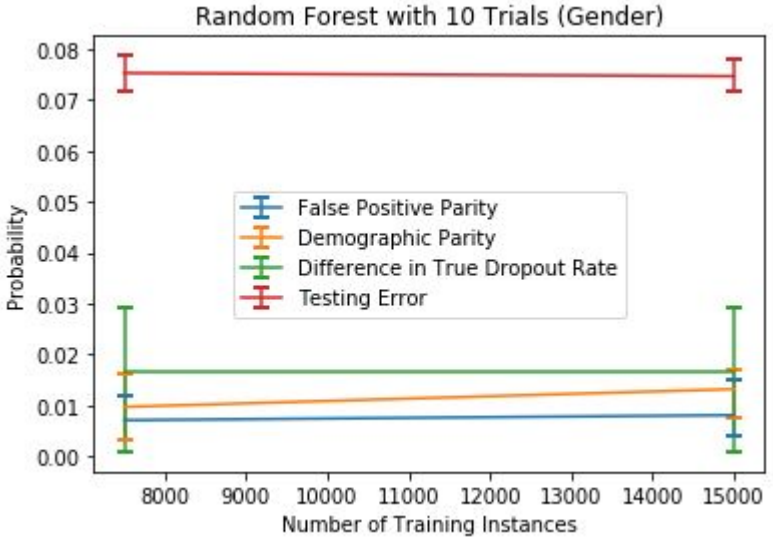
Kun

- Contacted school for more information about dataset.
- Contacted registrar office for the remote access on the dataset
- Cleaned the dataset for testing the algorithm
- Runned the test on the dataset using the random forest algorithm

Current Result

Kun

- Total dropout rate after cleaned data set is around 8.8%.
- Female dropout rate is slightly lower than male dropout rate.



What we will accomplish

Kun

- Prepare data results in well formatted tables
- Discuss findings on University dataset with registrar
- Work on final presentation and setting up demo

COVID-19 Changes

- Moved to using Zoom for all meetings
- Delayed time accessing University dataset
  - Needed to setup virtual access from our computers
  - Took us 3 weeks to figure set up with the registrar’s office

Confidence in Success

- We are currently a bit behind our planned schedule but are confident in our success.
- 100% confident on developing a distributed fair random forest algorithm
- 90% confident on developing a detecting intersectional bias algorithm
- 100% that our algorithm has good prediction results compared to the traditional algorithm

Reflection

Successes:

- Compared our algorithm on the University data set with other fair algorithms and show ours is more fair
- Good result of prediction rate than traditional algorithm
- Good initial results suggesting the first several features in a tree are more likely to show intersectional bias

Reflection

Roadblocks:

- Took a long time to get access to the University data
- Researching new algorithms and methods is difficult
- Proving a new method works is challenging

Reflection

Changes to plan:

- No longer looking to remove intersectional bias
  - Only detecting intersectional bias
  - Changing our detection scheme to only look at the first set of features in a decision tree
- No longer trying to get more data from University

Lessons Learned

- Working with Universities can take an extremely long time
- It is worth it to email authors of papers you read
  - Received help on an algorithm I was implementing from their paper
- Running experiments can take an extremely long time on large data sets

Questions?

Please feel free to email us if you have any questions about our algorithm or result

Email: jfantin1@uwyo.edu, rchen2@uwyo.edu, kyi1@uwyo.edu