Student Dropout Prediction

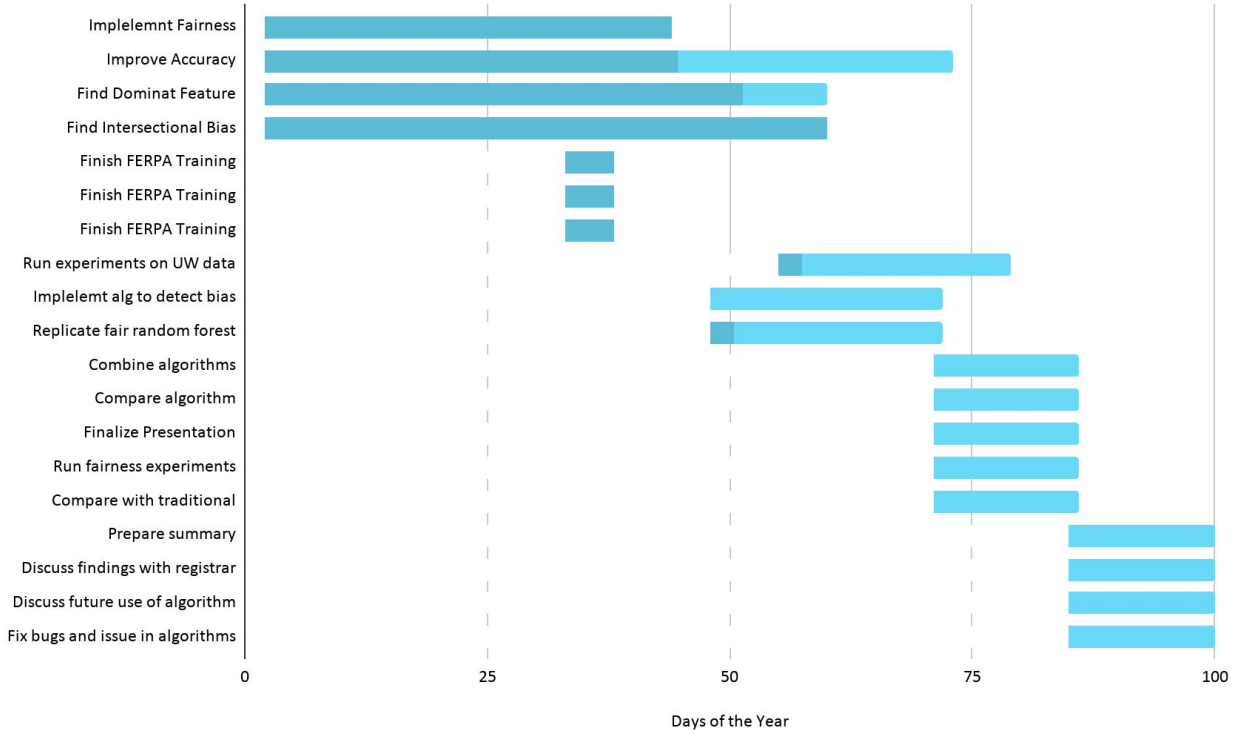
Rui Chen, James Fantin, Kun Yi

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.

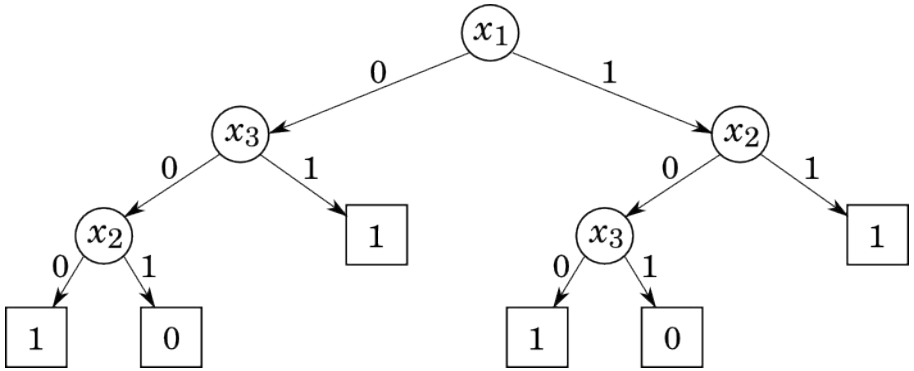
What we have accomplished?



What we have accomplished?

James:

- Developed a fair random forest algorithm based on randomly generated decision trees

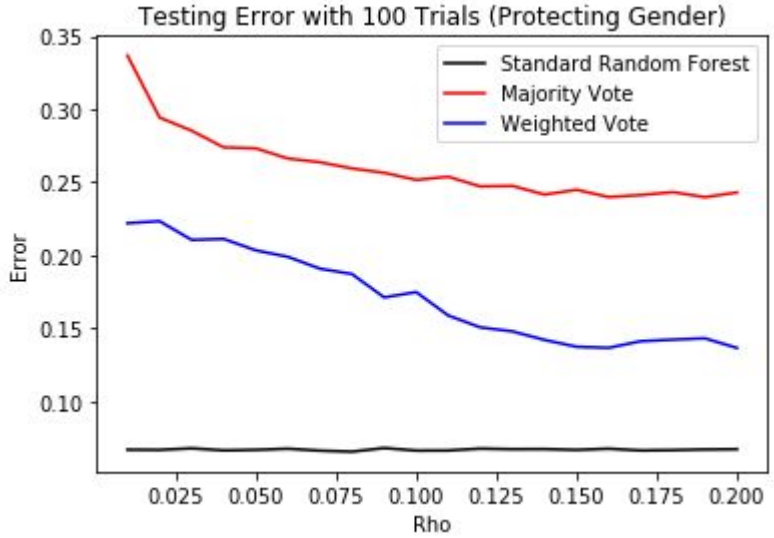


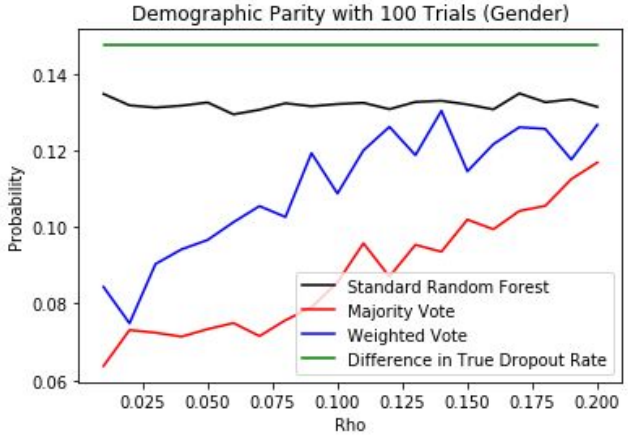
What we have accomplished?

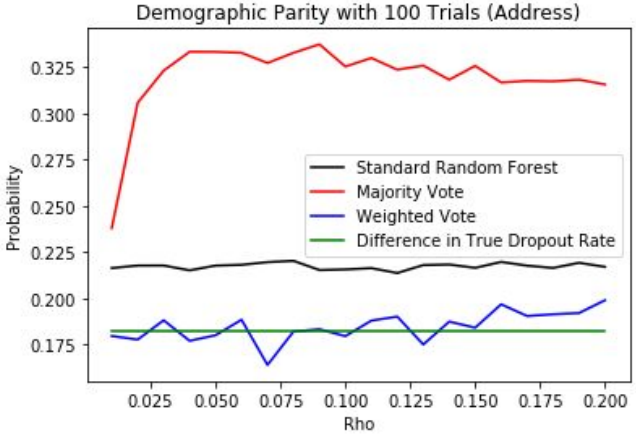
James:

- Algorithm:
  - Build many completely randomly generated trees (without demographic data)
  - Pick n trees that have discrimination < p
  - Apply ridge regression to learn weights for each tree

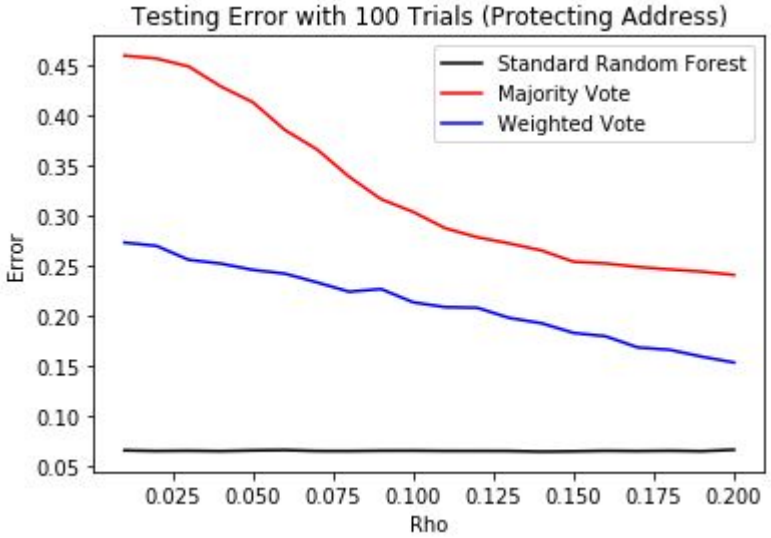
Current Results

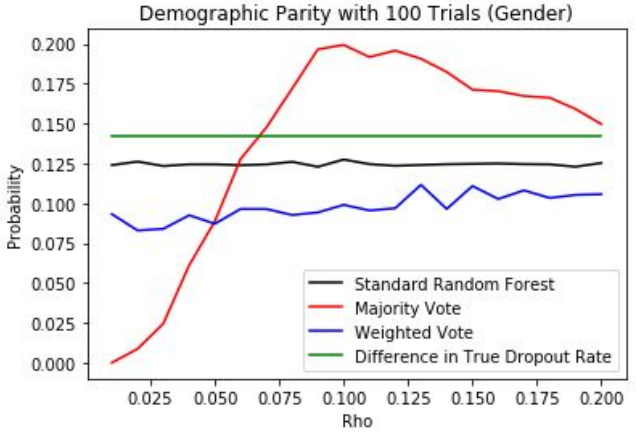


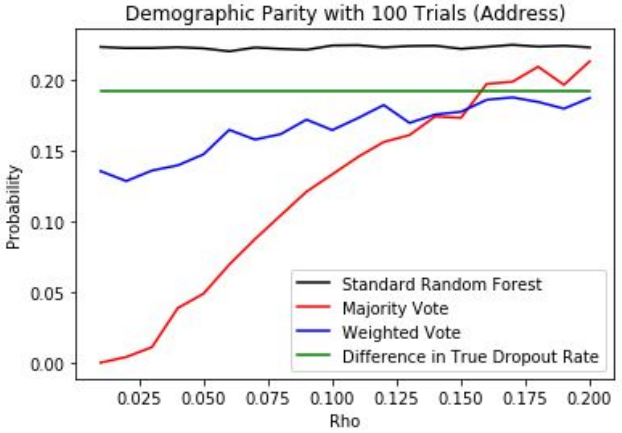




Current Results







What we will accomplish

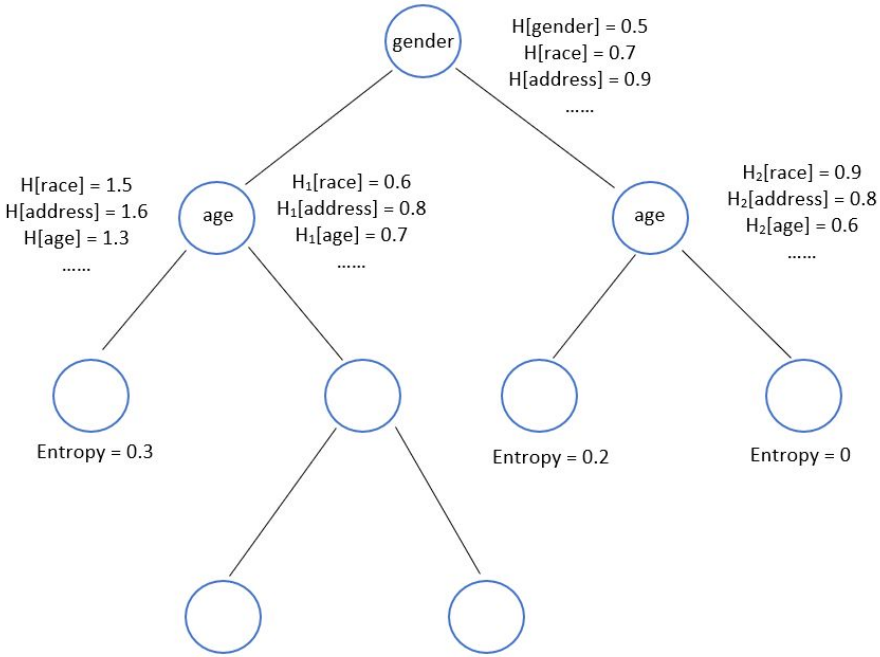
James

- Implement existing fair algorithms to compare against
- Investigate intersectional bias in algorithm, look at steps to reduce intersectional bias
- Combine with Rui to have an algorithm that can detect and remove bias

What we have accomplished?

Rui

- Build intersectional bias detector based on standard decision tree
- Split each parent node by the feature with best purity (lowest entropy or impurity)
- Stop split until max depth or ϱ% pure node
- Only tested with binary features
- Randomly process 70% instances in 20 trial



Current Result

From Portugal Dataset,

- Dominant feature is

  (Want_to_take_higher_education & Mother_work_at_home & Gender),

  might exist intersectional bias

10	list	6	[43, 21, 17, 37, 41, 31]
11	list	6	[43, 44, 17, 41, 29, 18]
12	list	6	[43, 17, 21, 33, 41, 35]
13	list	6	[43, 44, 21, 26, 18, 19]
14	list	6	[43, 21, 17, 26, 19, 39]
15	list	6	[43, 21, 17, 37, 32, 23]
16	list	6	[43, 17, 21, 33, 40, 38]
17	list	6	[43, 21, 17, 26, 18, 39]
18	list	6	[43, 21, 17, 38, 41, 19]
19	list	6	[43, 21, 26, 44, 39, 34]

What we will accomplish

Rui

- Test detector with both binary and continuous features
- Each leaf as a cluster, compare two similar clusters to define the dominant feature that cause different dropout rate between the two clusters
- Design an algorithm that will automatically detect intersectional bias

What we have accomplished?

Kun

- Finished transfer testing with traditional algorithms on two different dataset found last semester.
- Finished single bias detection on transfer testing.
- Finished intersectional bias detection on transfer testing.
- Contact school for data, and start cleaning data with group members.

What we will accomplish

Kun

- Contact school for more information about dataset.
- Run experiments on University dataset to discover bias
- Run fairness experiments using the completely fair random forest algorithm. Compare with traditional algorithms.
- Summarize data results in well formatted tables.

Confidence in Success

We are currently a bit behind our planned schedule but are confident in our success.

- Had hoped to have better accuracy with our initial fair random forest algorithm
- Taking longer to detect fairness in trees
- Has taken a long time to work with the University to get data

Reflection

Successes:

- Able to detect intersectional bias in our datasets
- Have an initial fair random forest algorithm

Roadblocks:

- University has taken a long time to give us data
- Researching new algorithms and methods is difficult
- Struggling to balance fairness and accuracy

Reflection

Changes to plan:

- No longer pursuing transfer learning due to lack of research and time constraints
- Assume that we have access to private student demographic information

Lessons Learned

- It is difficult to implement algorithms from research papers (sometimes vague)
- It is only worth our time to read research articles from top-tier conferences
- Working with Universities can take an extremely long time
- Look at all resources available before you implement an algorithm (it may already exist)

Questions