2020-2021 SIOP Machine Learning Competition Awards

Nick Koenig & Isaac Thompson

April 2021

Overview

1. Overview and Winners Announced
2. Winning Teams Present (4th, 3rd, 2nd, & 1st)
3. Panel Discussion, Q&A, and Closing Remarks

Why a Competition

Disrupt or be disrupted

The AI revolution is here. If we are not keeping up we are falling behind.

Meritocracy

Let the best ideas win.

Open Sourcing Everything

Open data, open registration, automated ranking, & open winning code.

Future Proof

The ‘best’ method may arrive 2023. The data and scores provide an open ‘benchmark’ to test these new methods.

Open Registration: 2020-2021 Stats

Participants
- 200+ teams (600 people)

Number of submissions
- 1500 +

First ever ML comp?
- 60%

University vs industry
- 53% vs 47%

Education
- 60% Ph.D., 40% MS & BS

Programming language
- 60% used R
- 66% used Python
- 27% used both
- 0% SPSS, Matlab, SAS, etc.

SIOP 2021
A VIRTUAL EVENT
The Machine Learning Competition

Open Competition

- Hosted competition on EvalAI, open source AI challenge platform.

- Teams created names, and submitted to get automated feedback

- A live public leaderboard showed progress

- 57 unique teams interacted with portal.

Rishabh:
Rishabh is a visiting research scholar at Georgia Tech and one of the lead developers of EvalAI Project.

Deshraj:
Co-founder of caliper.ai, MS in CS from Georgia Tech.

Open Solutions

- All winning solutions written in open source and fully reproducible code
- All made completely available on Github

Open Data

Data

- ~ 50k employee selection records

Objective

- To maximize both validity and fairness

Goal of the Competition

- The goal of the competition was to build an algorithm that effectively predicted performance and turnover while balancing adverse impact
- The weights were obtained by running a simulation of 4 million potential combinations and examining the distribution among them to settle on weights that led to scores that could realistically fall between 1-100

$$FinalScore = OverallAccuracy - Unfairness$$

$$OverallAccuracy = (PercentageTrueTopPerformers * 25) + (PercentageTrueRetained * 25) \\ + (PercentageTrueTopPerformersandRetained * 50)$$

$$Unfairness = ABS(1 - AIRatio) * 100$$

Modern Hire

About the Data

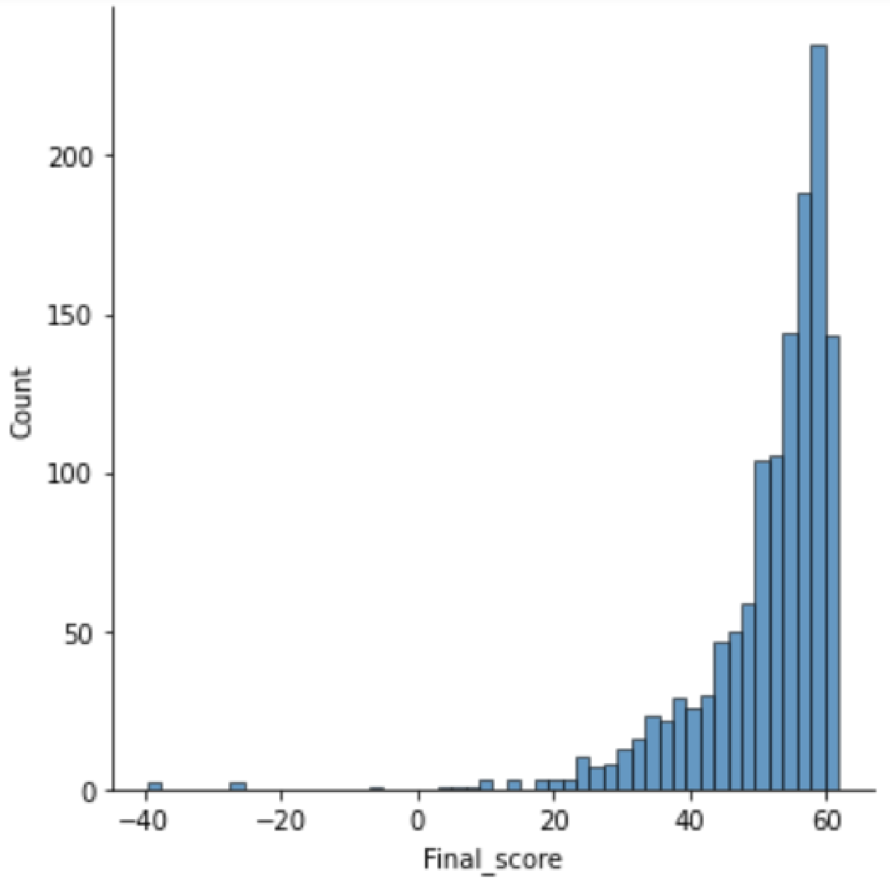
- This marvelous dataset was provided to us by Walmart
- These data were collected as part of a validation study of a pre-employment test for entry-level positions in retail stores.
- There were some small modifications made for the purposes of the ML Competition
  - All PII was removed
  - Protected_Group was a contrived variable to be used as a surrogate for ethnicity/gender
  - Actual item content was removed
- In addition to the assessment data the following were included:
  - Supervisor ratings of job performance (n = 12,390)
  - Turnover (n = 48,602)
- Split
  - Training Set (n = 44,102)
  - Public Leaderboard Holdout (n = 2,250)
  - Private Leaderboard Holdout (n = 2,250)

What does all this mean?

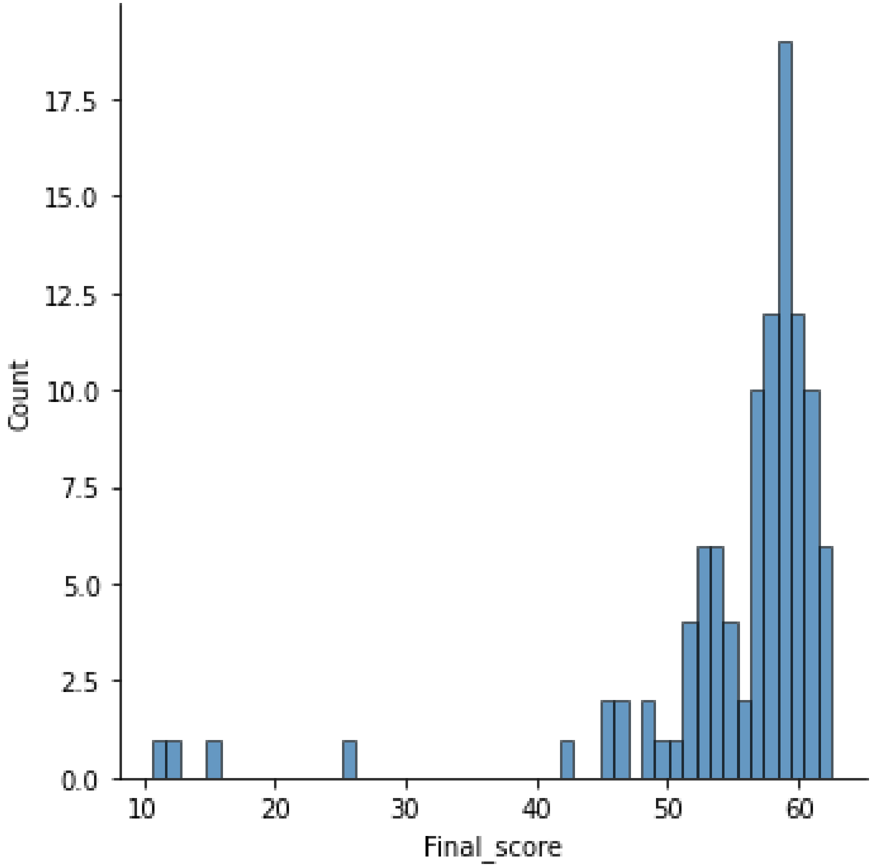
- Community
  - Over 200 teams worked ‘together’ one of the largest employer's data sets, to solve one of the toughest and most applicable problems in the field: **bringing fairness and validity to selection algorithms.**
- Less talk more action
  - Accelerate the conversation around fairness and AI, yet not focused on talk, **“code wins arguments”**. What are we missing as a field to solve the fairness issue?
- No excuses
  - Code base that could work in solving fairness for many assessment batteries, **available to all.**
- Living project
  - Data/code/results public, collaborators will **continue to innovate**, comparing new methods on the same data and competition benchmarks.

Final Score Distribution

Public Leaderboard



Private Leaderboard



Best Predictions on the Public Leaderboard

- **True Top Performers Hired % Correct:**
  - 64% - *Team Procrustination*
- **True Retained Hired % Correct:**
  - 72.6% - *Sam Cannon's Student Loan Debt*
- **True Top Performers and Retained Hired % Correct:**
  - 63.33% - *go ahead, make my data & raaka*
- **Best Selection Ratio (AI):**
  - 100% - 19 different teams

Top Scores

Public Leaderboard

Place	Team Name	Final Score
1	go ahead, make my data	61.72
2	#GreedyCow	61.63
3	Don't Ask Us y	61.40
4	Data_and_Confused	60.96
5	Waffles	60.91
6	raaka	60.91
7	Team Procrustination	60.64
8	Axiom Consulting Partners	60.63
9	gooners/Lets_Be_Fair	60.23
10	Ensemble Methods	60.19

Private Leaderboard

Place	Team Name	Final Score
1	?????	62.53
2	?????	62.50
3	?????	61.09
4	?????	60.72
5	gooners/Lets_Be_Fair	60.51
6	Chopin with Determination	60.47
7	sal	60.38
8	Waffles	60.27
9	Models For All	59.77
10	Who R we?	59.65

4th Place Team

60.72

COLORADO STATE UNIVERSITY

3rd Place Team

61.09

2nd Place Team

62.50

1st Place Team

62.53

Winning Team Presentations

Discussion Questions

- **Messy/extra data**
  - Like most real-world situations, there was incomplete data, in this case far more turnover data than performance data. Did any of you utilize the extra turnover for your models?
- **Fairness**
  - Competition on fairness... what part of your pipeline addressed that problem... did you all use MOO, other libraries for fairness, adversarial debiasing (why or why not) and what worked?
- **Race on the prediction side**
  - Winning solutions used race predictions. This was used to make outcomes more fair. Should the field use this type of data? Pros/cons/ ethical considerations, legality
- **Assessment Battery**
  - Did anyone treat the scales as scales or were the features that tended to work the best individual items?
- **Takeaway**
  - What is your top main takeaway/learning from this process?

Wrap Up

- Cross-validation!!
- Pros of a multidisciplinary approach.
- Machine learning competition basis for continued innovation.
  - New methods apples to apples
  - 200 teams!
- Fairness (AIR =1.0) is possible but what do we need as a field to realize it?
  - The most fair results outperformed the least fair results on all criterion
  - Ethical/legal, algorithmic, the way we collect this data?
- Non-linear complex models outperformed basic models.
- Tools that are built to focus on fairness did not win the competition.
  - Univariate models solve the problems

Thank You