

Average Win Rate Against All Other Models (Assuming Uniform Sampling)

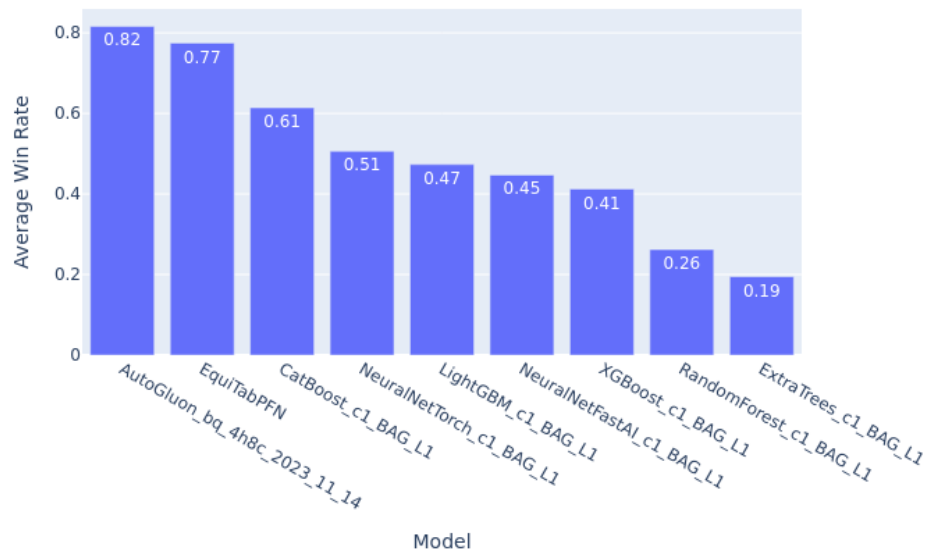


Figure 7. Average winrate on TabRepo Benchmark.

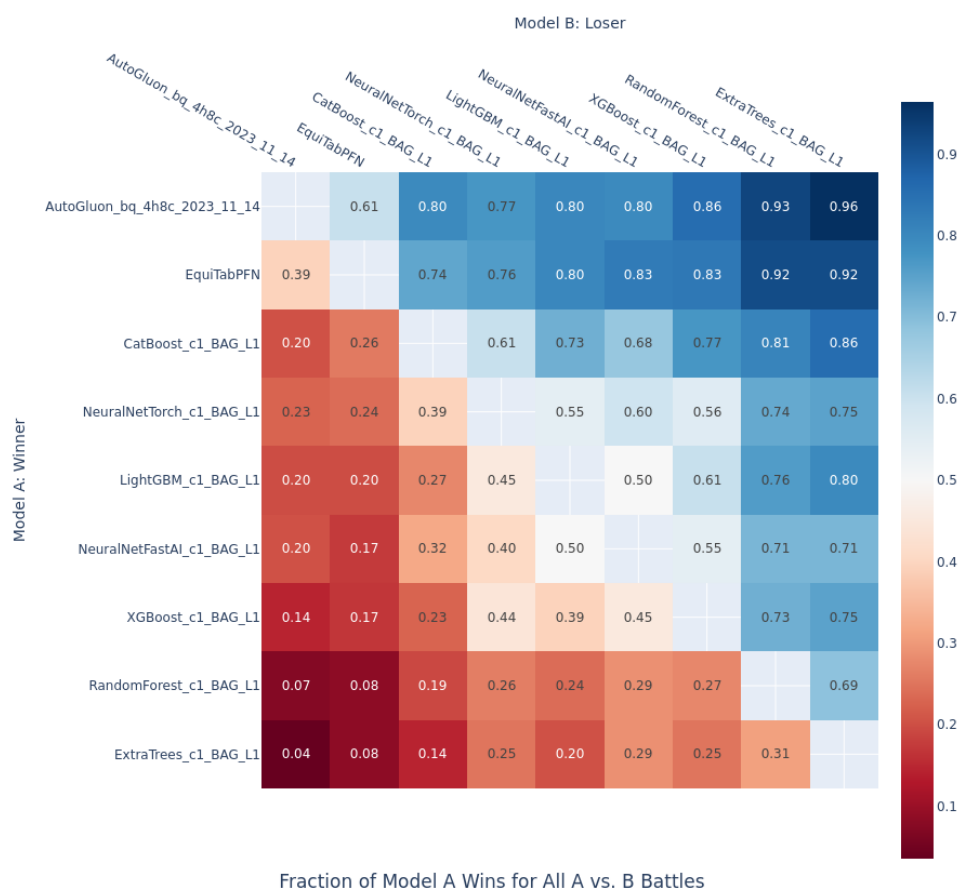


Figure 8. Pairwise winrate on TabRepo Benchmark.

Elo Confidence Intervals on Model Strength (via Bootstrapping)

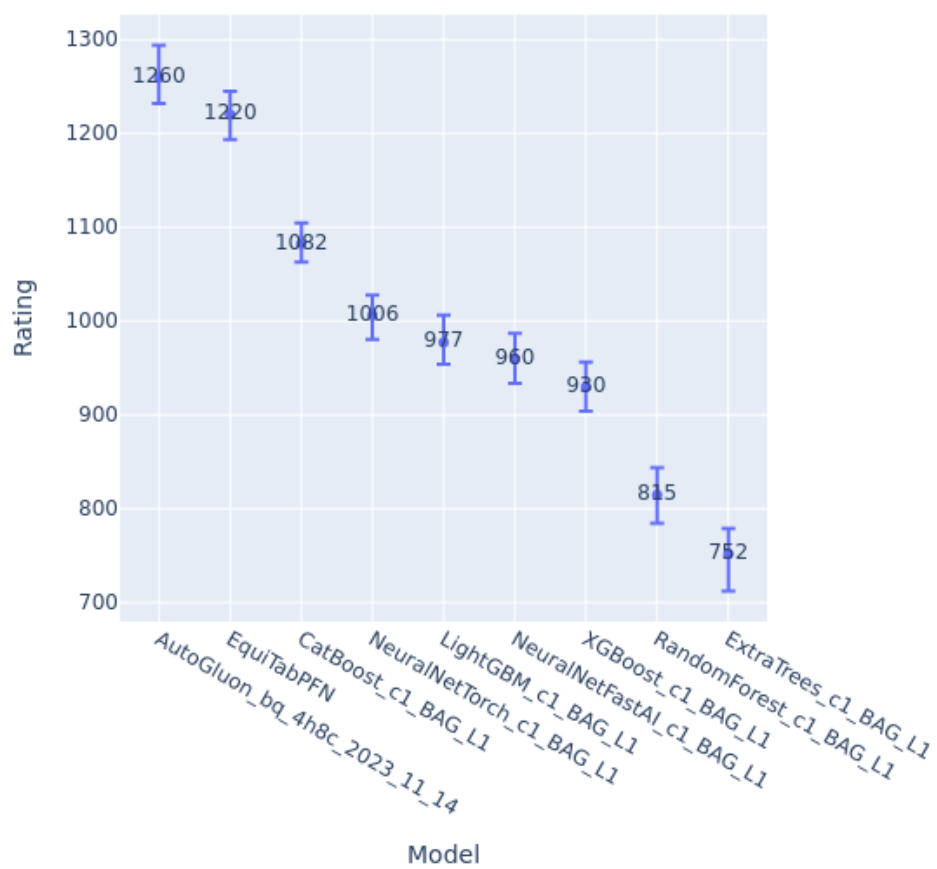


Figure 9. Elo ratings on TabRepo Benchmark: higher is better.