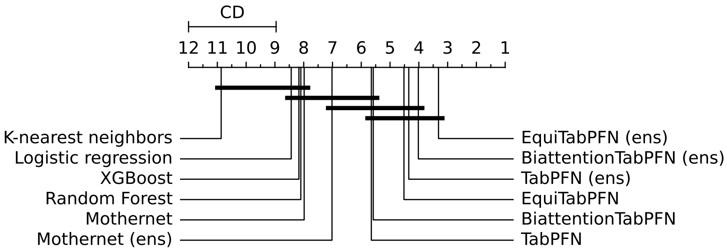

Figure 1: Critical diagram on the 30 real-world datasets from OpenML-CC18 benchmark.

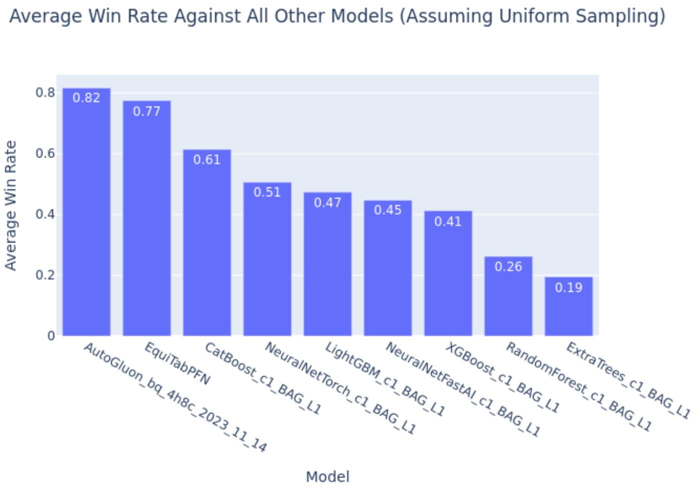

Figure 2: Average winrate on TabRepo Benchmark.

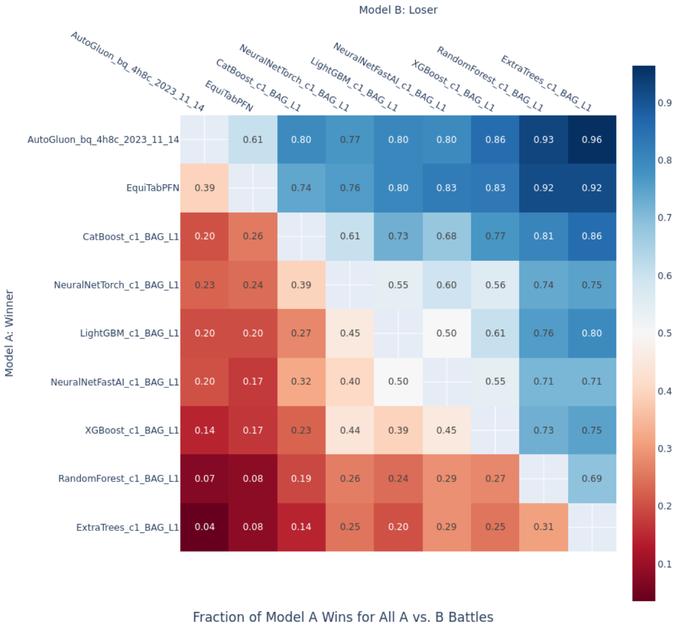

Figure 3: Pairwise winrate on TabRepo Benchmark.

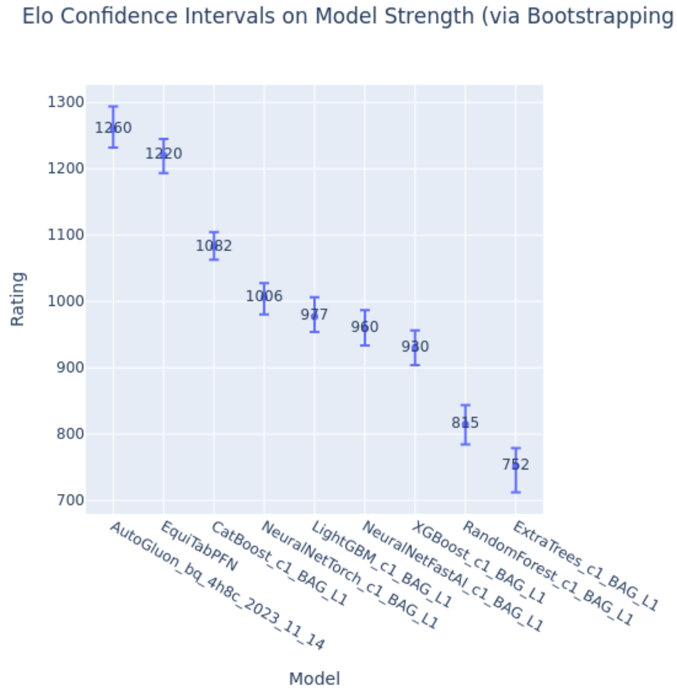

Figure 4: Elo ratings on TabRepo Benchmark: higher is better.

| Method | Mean acc. (↑) | Mean rank (↓) |
|---|---:|---:|
| EquiTabPFN | **0.81** | **3.94** |
| TabPFN | **0.81** | 4.39 |
| CatBoost | 0.80 | 5.29 |
| XGBoost | 0.80 | 6.72 |
| SAINT | 0.78 | 7.47 |
| NODE | 0.78 | 7.75 |
| rtdl$_R esNet$ | 0.78 | 8.06 |
| RandomForest | 0.77 | 8.06 |
| LinearModel | 0.74 | 9.34 |
| rtdl-FTTransformer | 0.75 | 10.10 |
| DANet | 0.75 | 10.66 |
| DecisionTree | 0.74 | 11.74 |
| rtdl-MLP | 0.70 | 12.16 |
| LightGBM | 0.74 | 12.72 |
| MLP | 0.69 | 12.99 |
| KNN | 0.70 | 13.25 |
| SVM | 0.70 | 13.52 |
| STG | 0.62 | 16.80 |
| VIME | 0.57 | 17.41 |
| TabNet | 0.62 | 17.66 |

Table 1: Mean accuracy and rank on TabZilla on 61 classifications datasets (including binary and multiclassification).

| Method | Mean acc. (↑) | Mean rank (↓) |
|---|---:|---:|
| EquiTabPFN | **0.77** | **3.06** |
| XGBoost | 0.76 | 4.69 |
| TabPFNModel | 0.75 | 5.27 |
| CatBoost | 0.75 | 5.40 |
| SAINT | 0.72 | 7.13 |
| rtdl$_{R}esNet$ | 0.72 | 7.19 |
| NODE | 0.71 | 7.46 |
| RandomForest | 0.71 | 8.77 |
| LinearModel | 0.66 | 9.83 |
| DANet | 0.68 | 10.27 |
| LightGBM | 0.68 | 11.04 |
| DecisionTree | 0.66 | 11.65 |
| rtdl-FTTransformer | 0.64 | 12.23 |
| KNN | 0.62 | 12.35 |
| SVM | 0.60 | 13.88 |
| rtdl-MLP | 0.56 | 14.08 |
| MLP | 0.56 | 14.42 |
| VIME | 0.50 | 16.46 |
| TabNet | 0.51 | 17.38 |
| STG | 0.50 | 17.42 |

Table 2: Mean rank against methods in TabZilla on 26 multi-classification datasets (e.g. datasets with at least 3 classes).