Bootstrap analysis: Determine Top Performers - Task1

Sage CNB Team

2025-11-11

Introduction & Goal

This notebook outlines the methodology for declaring top-performing methods in the SEA-AD DREAM Challenge, specifically by assessing for "tied" performance between submissions.

Task 1 Evaluation Overview

The primary metric used to evaluate submissions for Task 1 is the Quadratic Weighted Kappa (ADNC_QWK), computed on the predicted Alzheimer's Disease Neuropathology Consensus (ADNC) scores.

Methodology

To determine if methods are substantially different in performance, we employ a bootstrapping approach combined with a Bayes Factor (BF) calculation. This statistical framework allows us to assess the statistical equivalence of models across many resampled datasets.

1. **Bootstrapping:** We repeatedly sample with replacement (e.g. 1000-10000 times) the submitted predictions and the gold standard values. This generates a distribution of performance scores for each participant under various data scenarios.
2. **Bayes Factor (BF):** We calculate the BF for each method relative to the top-performing reference method. The BF quantifies the evidence of one model being better than another.

Determining the Top-Performer(s)

A smaller BF indicates more similar performance. We use a BF cut-off of 3 to define a tie. Any method with a BF ≤ 3 (relative to the best method) is considered not substantially different and is therefore declared a top-performer alongside the reference method.

Setup and Data Loading

Packages

```
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(knitr))
suppressPackageStartupMessages(library(synapser))

# Login to Synapse.
syn$login(silent=TRUE)
```

Helper Functions

In addition to importing the scoring functions provided by the challenge organizers, we also define two functions for data processing and analysis:

1. `get_name()`: this function resolves a Synapse `submitterid` (which can be a userID or teamID) into a human-readable username or team name. This will later help with plotting.
2. `computeBayesFactor()`: this function is an updated implementation of the BF calculation, superseding the outdated version in the `challengescoring` package. It calculates the BF for all submissions relative to a specified reference prediction index (`refPredIndex`).

```
reticulate::source_python("../evaluation/dream_evaluation.py")

get_name <- function(id) {
  name <- tryCatch({
    syn$getUserProfile(id)$userName
  }, error = function(err) {
    syn$getTeam(id)$name
  })
  name
}

computeBayesFactor <- function(bootstrapMetricMatrix,
                               refPredIndex,
                               invertBayes) {
  M <- as.data.frame(bootstrapMetricMatrix - bootstrapMetricMatrix[,refPredIndex])
  K <- apply(M ,2, function(x) {
    k <- sum(x >= 0)/sum(x < 0)

    # Logic handles whether reference column is the best set of predictions.
    if(sum(x >= 0) > sum(x < 0)){
      return(k)
    }else{
      return(1/k)
    }
  })
  K[refPredIndex] <- 0
  if(invertBayes == T){K <- 1/K}
  return(K)
}
```

Data Acquisition

The Final Round prediction files and the corresponding groundtruth file are retrieved from Synapse, where they are currently stored.

Table 1: Task 1 Final Round Leaderboard (as seen on Synapse)

id	submitterid	ADNC_QWK
9761204	gisl7	1.0000000
9761304	CMC-TJU	0.9786600
9761281	BioICAR	0.9321410
9761133	Metformin-121	0.7550492
9761109	raesalves	0.6236629

id	submitterid	ADNC_QWK
9761137	TeamGeckoAD	0.4976991
9760854	nkck	0.4951076
9761319	ADMIL	0.4418484
9761235	christina.hshi	0.3883122
9761311	Mount_Sinai_Team	0.2647668
9761249	danielr	0.2039113
9761232	TREATS	0.0754129

Bootstrapping Procedure

First, we combine all individual submission predictions on ADNC and the groundtruth values into a single dataframe. This single source will help create a more efficient resampling process for the bootstrapping analysis.

Table 2: Preview of Model Predictions and Corresponding Groundtruth Values

donor_id	gisl7	CMC-TJU	BioICAR	Metformin-121	raesalves	TeamGeckoAD	nkck	ADMIL	christina.hshi	Mount_Sinai_Team	danielr	TREATS
D672	High	High	High	High	Intermediate	Low	High	Low	High	Low	High	High	High
D713	High	High	High	High	High	Intermediate	High	Low	Intermediate	Low	Low	High	High
D632	High	High	High	High	High	Intermediate	Not AD	High	High	Intermediate	High	High	High
D823	Low	Low	Low	Low	Low	Intermediate	Low	Not AD	Intermediate	High	Intermediate	High	Low
D810	Intermediate	Intermediate	Intermediate	Intermediate	Intermediate	Low	Intermediate	Intermediate	Intermediate	Low	High	High	Intermediate
D880	High	High	High	High	Intermediate	Intermediate	High	Low	Intermediate	High	High	High	High

Next, we perform a pre-bootstrapping verification to ensure the bootstrapping logic correctly reproduces the published Final Round scores *prior* to any resampling. This serves as a critical check on the bootstrapping and scoring functions.

```
bs.check <- sapply(names(pred_filenames), function(team) {
  apply(matrix(1:nrow(truth), nrow(truth), 1), 2, function(ind) {
    cohen_kappa_score(
      submissions$truth[ind],
      submissions[[team]][ind],
      weights="quadratic"
    )
  })
})
```

Table 3: Comparison of "Bootstrapping" Scores (Prior to Resampling) and Original Scores

submitterid	bf_calculated_score	original_score	scores_match
gisl7	1.0000000	1.0000000	TRUE
CMC-TJU	0.9786600	0.9786600	TRUE
BioICAR	0.9321410	0.9321410	TRUE
Metformin-121	0.7550492	0.7550492	TRUE

submitterid	bf_calculated_score	original_score	scores_match
raesalves	0.6236629	0.6236629	TRUE
TeamGeckoAD	0.4976991	0.4976991	TRUE
nkck	0.4951076	0.4951076	TRUE
ADMIL	0.4418484	0.4418484	TRUE
christina.hshi	0.3883122	0.3883122	TRUE
Mount_Sinai_Team	0.2647668	0.2647668	TRUE
danielr	0.2039113	0.2039113	TRUE
TREATS	0.0754129	0.0754129	TRUE

The core bootstrapping process is then executed:

- Resample the rows (data points) of the combined dataframe 10,000 times ($N = 10,000$) with replacement
- For each of the N bootstrap samples, we re-score every submission using the primary challenge metric (Quadratic Weighted Kappa, `QWK`) on ADNC

This will produce a matrix of 10,000 bootstrapped scores per submission.

```
# Set seed for reproducible results (since we're doing a random sample)
set.seed(202511)

# Run bootstrapping.
N <- 10000
bs_indices <- matrix(1:nrow(truth), nrow(truth), N) %>%
  apply(2, sample, replace = TRUE)

bs <- sapply(names(pred_filenames), function(team) {
  apply(bs_indices, 2, function(ind) {
    cohen_kappa_score(
      submissions$truth[ind],
      submissions[[team]][ind],
      weights="quadratic"
    )
  })
})
```

Bayes Factor Calculation

The resulting $10,000 \times$ number of submissions matrix of bootstrapped scores is used to calculate the BF for each submission.

The top-performing model (highest `ADNC_QWK` score) is automatically set as the reference prediction ($BF = 0$) against which all other methods are compared.

As established earlier, a BF ≤ 3 indicates that a method's performance is not substantially different from the top performer.

submission	bayes
gisl7	0.000000
CMC-TJU	7.025682
BioICAR	22.419204
Metformin-121	2499.000000
raesalves	Inf

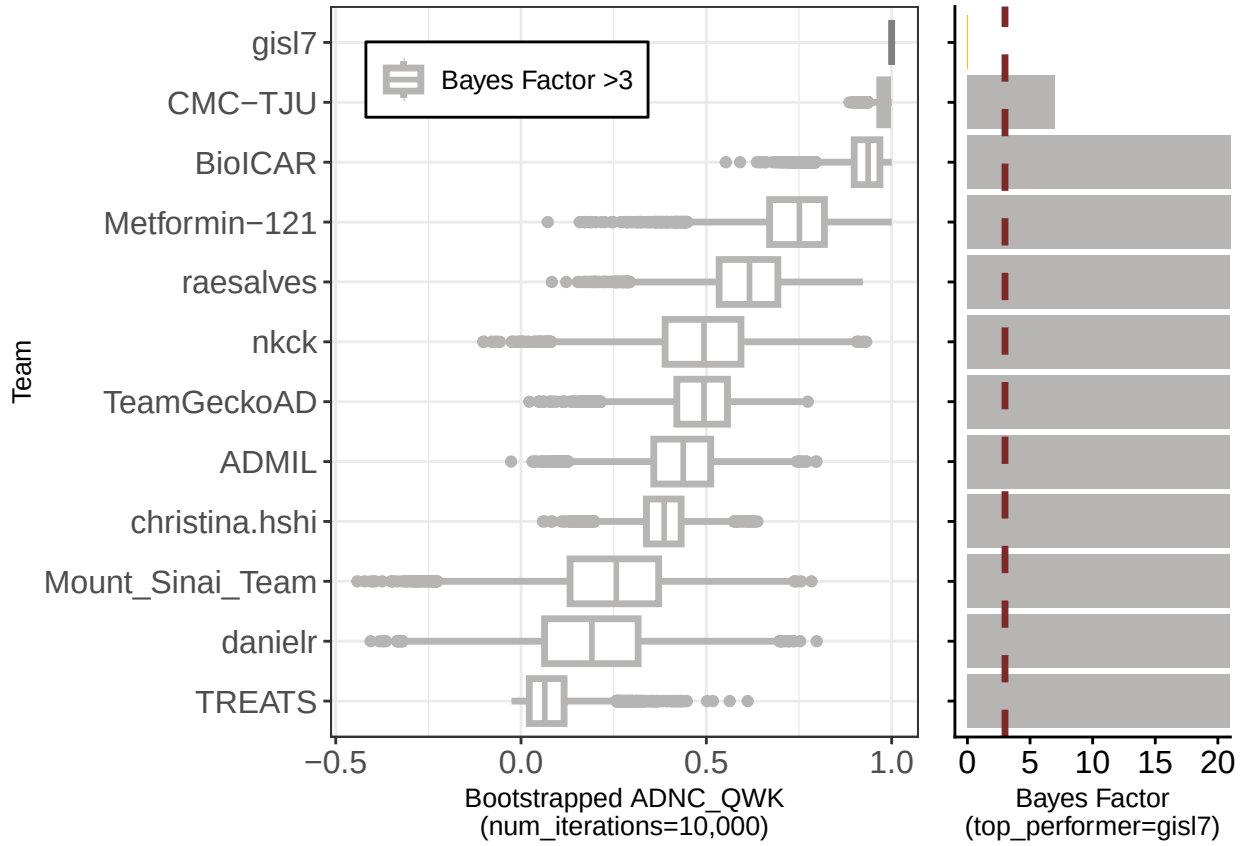
submission	bayes
TeamGeckoAD	Inf
nkck	Inf
ADMIL	Inf
christina.hshi	Inf
Mount_Sinai_Team	Inf
danielr	Inf
TREATS	Inf

Results & Top Performer Determination

Finally, the results are visualized using a combined plot showing two views:

1. **Bootstrapped Score Distribution (left):** boxplot showing the distribution of the 10,000 bootstrapped `ADNC_QWK` scores for each team. The color indicates the BF category.
2. **Bayes Factor Visualization (right):** bar plot showing the calculated BF relative to the top-performer. The horizontal dashed line indicates the BF ≤ 3 cut-off used to determine a tie.

Methods falling into the "BF ≤ 3" category, including the top performer ($BF = 0$), will be declared "Top Performers" for this challenge task.



Conclusion

The objective of this analysis was to use the Bayes factor to identify submissions that are statistically indistinguishable from the overall top performer, using a threshold of BF ≤ 3 to define a tie.

The analysis confirms that the reference submission, Team gisl7 (BF = 0), is the only team whose performance is not substantially different from itself. Since no other submissions yielded a Bayes factor ≤ 3, **team gisl7 is declared the sole "Top Performer" for Task 1 of the SEA-AD DREAM Challenge**.