Extract summary statistics for leaderboard and final round

G Bhatti

9/19/2024

# Calculate the total number of valid submissions during the leaderboard phase

df <- read\_csv(here("data/submissions/Job-393694313420778661233189284.csv"))

## Rows: 183 Columns: 35  
## ── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (16): name, etag, projectId, status, entityid, dockerrepositoryname, doc...  
## dbl (17): ROW\_ID, ROW\_VERSION, id, createdOn, createdBy, modifiedOn, evaluat...  
## lgl (2): submitteralias, orgSagebionetworksSynapseWorkflowOrchestratorSubmi...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Process submission data and test data  
df\_select\_submissions <- df |>   
 filter(submission\_status == "SCORED") |>   
 filter(!createdBy %in% c("1420476","3379638","3453428","3503156", "3504542","3505047","3505276"))   
  
df\_select\_submissions |>   
 group\_by(evaluationid) |>   
 summarise(n\_submitters=n\_distinct(submitterid),  
 n\_submissions=n())

## # A tibble: 2 × 3  
## evaluationid n\_submitters n\_submissions  
## <dbl> <int> <int>  
## 1 9615595 7 37  
## 2 9615596 5 12

## Get the final rankings for the leaderboard phase

ft1 <- df\_select\_submissions |>   
 group\_by(submitterid) |>   
 mutate(submission\_datetime = as.POSIXlt(createdOn/1000, origin = "1970-01-01")) |>   
 slice\_min(rmse,n=1,with\_ties = FALSE) |>   
 ungroup() |>   
 mutate(Leaderboard\_Rank = rank(rmse)) |>   
 select(Rank = Leaderboard\_Rank,   
 Participant = Submitter,   
 `RMSE (wks)` = rmse,  
 `MAE (wks)`=mae,  
 Correlation=cor,  
 `Sub-challenge`=evaluationid) |>   
 mutate(Participant=str\_replace(Participant,"\\(","\\(@"),  
 `Sub-challenge`=if\_else(`Sub-challenge`=="9615595","SC1","SC2")) |>   
 arrange(Rank) |>   
 qflextable() |>   
 align(align = "center", part = "all") |>   
 colformat\_double(j = c("RMSE (wks)","MAE (wks)","Correlation"),digits = 2) |>   
 add\_header\_row(values = c("Ranking during the leaderboard round"), colwidths = c(6))   
  
  
  
save\_as\_docx(ft1, path = here("results/tablex.docx"))

## How did the Robust Placental clock fare in the leaderboard round?

## Get the final rankings for the final round

all\_predictions<- read\_csv(here("data/submissions/all\_predictions.csv"))

## Rows: 384 Columns: 13  
## ── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): ID  
## dbl (12): ga\_3320980\_9615595, ga\_3320980\_9615596, ga\_3449866\_9615595, ga\_344...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

ano<- read\_csv(here("data/prb/anoall.csv") )

## Rows: 384 Columns: 22  
## ── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (7): Selection, LNDC, abruptio, Group, Sample, Sex, Group2  
## dbl (14): Main\_Index, Methylation, Del\_GA\_Calc, Age, Parity, HGTinch, Pre\_Pr...  
## lgl (1): MPFD  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

ano<- ano |> select(Sample,Del\_GA\_Calc) |>   
 left\_join(all\_predictions,by=c("Sample"="ID"))  
  
calculate\_test\_metrics<- function(x){  
 y<-ano$Del\_GA\_Calc  
 test\_cor<- cor(x,y)  
 test\_rmse<-rmse(x,y)  
 test\_mae<-mae(x,y)  
 return(c(test\_cor=test\_cor,test\_rmse=test\_rmse,  
 test\_mae=test\_mae))  
}  
ga\_cols<- ano |> select(starts\_with("ga\_"))  
  
test\_metrics<- ga\_cols |>   
 map\_dfr(~calculate\_test\_metrics(.),.id="column\_name") |>   
 separate(col=column\_name,into = c("GA","submitterid","evaluationid"),sep = "\_") |>   
 mutate(submission\_stamp = paste(submitterid, evaluationid, sep = "\_")) |>   
 select(-GA,-submitterid,-evaluationid)  
   
df\_test <- read\_csv(here("data/submissions/Test\_data\_evaluation.csv") )

## Rows: 12 Columns: 7  
## ── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): dockerrepositoryname  
## dbl (5): submitterid, createdBy, createdOn, evaluationid, Test\_rmse  
## dttm (1): submission\_datetime  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

ft2 <- df\_test |>   
 mutate(submission\_stamp = paste(submitterid, evaluationid, sep = "\_")) |>  
 left\_join(test\_metrics,by=c("submission\_stamp")) |>   
 group\_by(submitterid) |>   
 slice\_min(Test\_rmse,n=1,with\_ties = FALSE) |>   
 ungroup() |>   
 mutate(Leaderboard\_Rank = rank(test\_rmse)) |>   
 left\_join(df\_select\_submissions |>   
 select(submitterid,Submitter) |>   
 filter(!duplicated(submitterid)),  
 by="submitterid") |>   
 select(Rank = Leaderboard\_Rank,   
 Participant = Submitter,   
 `RMSE (wks)` = test\_rmse,  
 `MAE (wks)`=test\_mae,  
 Correlation=test\_cor,  
 `Sub-challenge`=evaluationid) |>   
 mutate(Participant=str\_replace(Participant,"\\(","\\(@"),  
 `Sub-challenge`=if\_else(`Sub-challenge`=="9615595","SC1","SC2")) |>   
 arrange(Rank) |>   
 qflextable() |>   
 align(align = "center", part = "all") |>   
 colformat\_double(j = c("RMSE (wks)","MAE (wks)","Correlation"),digits = 2) |>   
 add\_header\_row(values = c("Ranking after the final evaluation"), colwidths = c(6))   
   
save\_as\_docx(ft2, path = here("results/tabley.docx"))