Adjusted SCM base regression training

Brian K. Masinde

```
# Clean workspace
rm(list = ls())

# Load libraries
library(rpart)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(mlflow)
library(reticulate)
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
```

```
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
library(purrr)
```

```
##
## Attaching package: 'purrr'

## The following object is masked from 'package:data.table':
##
##     transpose

## The following object is masked from 'package:caret':
##
##     lift
```

```
library(themis)
```

```
## Loading required package: recipes

##
## Attaching package: 'recipes'

## The following object is masked from 'package:stats':
##
##     step
```

```
library(doMC)
```

```
## Loading required package: foreach

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when

## Loading required package: iterators

## Loading required package: parallel
```

```
library(here)
```

```
## here() starts at /Users/masinde/Projects/causal_fairness_Ph_IbF
```

Inputs

```r
base_train <- read.csv(here("data", "base_train.csv"))
base_validation <- read.csv(here("data", "base_validation.csv"))
```

```r
# Combining train and validation datasets to one
# Because we are going to use CV to train the models later
# naming it df_base_train2 to remain consistent with df naming
df_base_train2  <- rbind(base_train, base_validation)

cat("number of rows in combined train data:", nrow(df_base_train2), sep = " ")
```

```
## number of rows in combined train data: 7184
```

Import trained model for wind

```r
base_wind_model <- readRDS(here("adjusted SCM/new base models",
                                "dec_base_wind_model_tuned.rds"))

base_rain_model <- readRDS(here("adjusted SCM/new base models",
                                "dec_base_rain_model_tuned.rds"))
```

Interaction terms (moderators)

```r
# Define wind and rain interaction variables
wind_fractions <- c("blue_ss_frac", "yellow_ss_frac", "orange_ss_frac", "red_ss_frac")
rain_fractions <- c("blue_ls_frac", "yellow_ls_frac", "orange_ls_frac", "red_ls_frac")

# Predict using model "base_wind_model"
# To get variable: wind_max_pred

df_base_train2[["wind_max_pred"]] <- predict(base_wind_model, newdata = df_base_train2)

df_base_train2[["rain_total_pred"]] <- predict(base_rain_model, newdata = df_base_train2)

# Compute wind interaction terms dynamically

for (col in wind_fractions) {
  print(col)
  new_col_name <- paste0("wind_", col)
  df_base_train2 [[new_col_name]] <- df_base_train2 [[col]] * df_base_train2 [["wind_max_pred"]]
}
```

```
## [1] "blue_ss_frac"
## [1] "yellow_ss_frac"
## [1] "orange_ss_frac"
## [1] "red_ss_frac"
```

```r
# Multiply rain fractions by rain_total
for (col in rain_fractions) {
  new_col_name <- paste0("rain_", col)
  df_base_train2 [[new_col_name]] <- df_base_train2 [[col]] * df_base_train2 [["rain_total_pred"]]
}
```

Base regression model training

```r
# BASE REGRESSION MODEL TRAINING AND TUNING USING CV
# CV folds and models
n_folds <- 7
n_models <- 25  # adjust depending on search space size, affects seeds length


# Reproducibility: Defining seeds (a little bit complicated because of parallel processing)

#  Generate a reproducible list of seeds
set.seed(1234)

seeds_list <- vector(mode = "list", length = n_folds + 1)
for (i in 1:n_folds) {
  seeds_list[[i]] <- sample.int(1000000, n_models)  # one seed per model per fold
}
seeds_list[[n_folds + 1]] <- sample.int(1000000, 1)  # for final model


# Set up train control with 10-fold cross-validation
train_control <- trainControl(
  method = "cv",
  number = n_folds,
  summaryFunction = defaultSummary,
  search = "random", # random selection of the expanded grid
  seeds = seeds_list
)

# Detect and register the number of available cores (use all but one)
num_cores <- parallel::detectCores() - 2
registerDoMC(cores = num_cores)  # Enable parallel processing

# Measure the time for a code block to run
system.time({
# Train the model using grid search with 7-fold CV
base_xgb_reg_model <- train(
  damage_perc ~  track_min_dist +
    wind_max_pred + # This was missing in the Dataiku workflow
    rain_total_pred +
    roof_strong_wall_strong +
    roof_strong_wall_light +
    roof_strong_wall_salv +
    roof_light_wall_strong +
    roof_light_wall_light +
```

```
    roof_light_wall_salv +
    roof_salv_wall_strong +
    roof_salv_wall_light +
    roof_salv_wall_salv +
    wind_blue_ss_frac +
    wind_yellow_ss_frac +
    wind_orange_ss_frac +
    wind_red_ss_frac +
    rain_blue_ls_frac +
    rain_yellow_ls_frac +
    rain_orange_ls_frac +
    rain_red_ls_frac +
    island_groups,  # Confounder adjustment
  data = df_base_train2,
  method = "xgbTree",
  trControl = train_control,
  tuneLength = n_models,  #  this replaces tuneGrid
  metric = "RMSE"  # Optimize based on RMSE
)
Sys.sleep(2)  # This is just an example to simulate a delay
})
```

```
##    user  system elapsed
## 334.143   2.707  44.033
```

```
# Print best parameters
print(base_xgb_reg_model$bestTune)
```

```
##   nrounds max_depth        eta  gamma colsample_bytree min_child_weight
## 1     100         6 0.02113197 3.9941        0.5383644                14
##   subsample
## 1 0.7606618
```

```
# best parameters
best_params <- base_xgb_reg_model$bestTune
damage_fit_reg_min <- train(damage_perc ~ track_min_dist +
                              wind_max_pred +
                              rain_total_pred +
                              roof_strong_wall_strong +
                              roof_strong_wall_light +
                              roof_strong_wall_salv +
                              roof_light_wall_strong +
                              roof_light_wall_light +
                              roof_light_wall_salv +
                              roof_salv_wall_strong +
                              roof_salv_wall_light +
                              roof_salv_wall_salv +
                              wind_blue_ss_frac +
                              wind_yellow_ss_frac +
                              wind_orange_ss_frac +
                              wind_red_ss_frac +
                              rain_blue_ls_frac +
```

```
                              rain_yellow_ls_frac +
                              rain_orange_ls_frac +
                              rain_red_ls_frac +
                              island_groups    # Confounder adjustment
                              , # Confounder adjustment
                              method = "xgbTree",
                              trControl = trainControl(method = "none"),
                              tuneGrid = best_params, # Use the best parameters here
                              metric = "RMSE",
                              data = df_base_train2
                          )
```

```
# Sanity Check
# RMSE on the trainset (training + validation)
# Compute RMSE

damage_pred  <- predict(damage_fit_reg_min, newdata = df_base_train2)
rmse_value <- rmse(df_base_train2$damage_perc, damage_pred)
rmse_value
```

```
## [1] 6.252095
```

```
# Define bin edges
# Define bin edges
bins <- c(0.00009, 1, 10, 50, 100)

# Assign data to bins
bin_labels <- cut(df_base_train2$damage_perc, breaks = bins, include.lowest = TRUE, right = TRUE)

# Create a data frame with actual, predicted, and bin labels
data <- data.frame(
  actual = df_base_train2$damage_perc,
  predicted = damage_pred,
  bin = bin_labels
)

# Calculate RMSE per bin
unique_bins <- levels(data$bin) # Get unique bin labels
rmse_by_bin <- data.frame(bin = unique_bins, rmse = NA, count = NA) # Initialize results data frame

for (i in seq_along(unique_bins)) {
  bin_data <- data[data$bin == unique_bins[i], ] # Filter data for the current bin
  rmse_by_bin$rmse[i] <- sqrt(mean((bin_data$actual - bin_data$predicted)^2, na.rm = TRUE)) # Calculate
  rmse_by_bin$count[i] <- nrow(bin_data) # Count observations in the bin
}

# Display RMSE by bin
print(rmse_by_bin)
```

```
##         bin      rmse count
## 1 [9e-05,1]  2.991399  5960
## 2    (1,10]  5.048177  4813
```

```
## 3   (10,50] 14.917181  4297
## 4  (50,100] 45.722353  4063
```

```
# set tracking URI
mlflow_set_tracking_uri("http://127.0.0.1:5000")

# Ensure any active run is ended
suppressWarnings(try(mlflow_end_run(), silent = TRUE))

# Logging metrics for model training and the parameters used
mlflow_set_experiment(experiment_name = "Attempt2: SCM - XGBOOST base regression - CV (Training metircs)
```

```
## [1] "140261814914201194"
```

```
# Ensure that MLflow has only one run. Start MLflow run once.
run_name <- paste("XGBoost Run", Sys.time())  # Unique name using current time


# Start MLflow run
mlflow_start_run(nested = FALSE)
```

```
## Warning: `as_integer()` is deprecated as of rlang 0.4.0
## Please use `vctrs::vec_cast()` instead.
## This warning is displayed once every 8 hours.
```

```
## # A tibble: 1 x 13
##   run_uuid             experiment_id run_name user_id status start_time
##   <chr>                <chr>         <chr>    <chr>   <chr>  <dttm>
## 1 c66dc37257074582803~ 140261814914~ mysteri~ masinde RUNNI~ 2025-07-24 17:44:07
## # i 7 more variables: artifact_uri <chr>, lifecycle_stage <chr>, run_id <chr>,
## #   end_time <lgl>, metrics <lgl>, params <lgl>, tags <list>
```

```
# Ensure the run ends even if an error occurs
#on.exit(mlflow_end_run(), add = TRUE)


# -------- best parameters ---------------
best_params <- base_xgb_reg_model$bestTune

# Log each of the best parameters in MLflow
for (param in names(best_params)) {
  mlflow_log_param(param, best_params[[param]])
}

# obtain predicted values
train_predictions <- predict(damage_fit_reg_min, newdata = df_base_train2)


# Define bin edges
# Define bin edges
bins <- c(0.00009, 1, 10, 50, 100)
```

```r
# Assign data to bins
bin_labels <- cut(df_base_train2$damage_perc, breaks = bins, include.lowest = TRUE, right = TRUE)

# Create a data frame with actual, predicted, and bin labels
data <- data.frame(
  actual = df_base_train2$damage_perc,
  predicted = train_predictions,
  bin = bin_labels
)

# Calculate RMSE per bin
unique_bins <- levels(data$bin) # Get unique bin labels
rmse_by_bin <- data.frame(bin = unique_bins, rmse = NA, count = NA) # Initialize results data frame

for (i in seq_along(unique_bins)) {
  bin_data <- data[data$bin == unique_bins[i], ] # Filter data for the current bin
  rmse_by_bin$rmse[i] <- sqrt(mean((bin_data$actual - bin_data$predicted)^2, na.rm = TRUE)) # Calculate
  rmse_by_bin$count[i] <- nrow(bin_data) # Count observations in the bin
}

# Display RMSE by bin
print(rmse_by_bin)
```

```
##         bin      rmse count
## 1 [9e-05,1]  2.991399  5960
## 2    (1,10]  5.048177  4813
## 3   (10,50] 14.917181  4297
## 4  (50,100] 45.722353  4063
```

```r
as.data.frame(rmse_by_bin)
```

```
##         bin      rmse count
## 1 [9e-05,1]  2.991399  5960
## 2    (1,10]  5.048177  4813
## 3   (10,50] 14.917181  4297
## 4  (50,100] 45.722353  4063
```

```r
RMSE_1 <- rmse_by_bin[1, "rmse"]
RMSE_10 <-  rmse_by_bin[2, "rmse"]
RMSE_50 <- rmse_by_bin[3, "rmse"]
RMSE_100 <- rmse_by_bin[4, "rmse"]

# Log binned RMSE metrics
mlflow_log_metric("RMSE_1", RMSE_1)
```

```
## Warning: `as_double()` is deprecated as of rlang 0.4.0
## Please use `vctrs::vec_cast()` instead.
## This warning is displayed once every 8 hours.
```

```r
mlflow_log_metric("RMSE_10", RMSE_10)
mlflow_log_metric("RMSE_50", RMSE_50)
```

```r
mlflow_log_metric("RMSE_100", RMSE_100)

# End MLflow run
mlflow_end_run()
```

```
## # A tibble: 1 x 13
##   run_uuid             experiment_id run_name user_id status start_time
##   <chr>                <chr>         <chr>    <chr>   <chr>  <dttm>
## 1 c66dc37257074582803~ 140261814914~ mysteri~ masinde FINIS~ 2025-07-24 17:44:07
## # i 7 more variables: end_time <dttm>, artifact_uri <chr>,
## #   lifecycle_stage <chr>, run_id <chr>, metrics <list>, params <list>,
## #   tags <list>
```

```r
# save the trained rds file
path <- here("adjusted SCM/new base models")
saveRDS(damage_fit_reg_min, file = file.path(path, paste0("base_reg_model", ".rds")))
```

OLD CODE