Testing Classification XGBOOST (unadjusted SCM)

Brian K. Masinde

```
# Environment: Cleaning environment
rm(list = ls())

# Libraries: Load
library(rpart)
library(caret)
library(pROC) # For AUC calculation
library(dplyr)
library(data.table)
library(mlflow)
library(purrr)
library(here)
```

Inputs

Importing test data and models

```
df_base_test <- read.csv(here("data", "base_test.csv"))

nrow(df_base_test)
```

```
## [1] 1797
```

```
## Import model: wind_max ~ track_min_dist + island_groups
#   Decision tree model
base_wind_model <- readRDS(here("unadjusted SCM/new base models",
                                "dec_base_wind_model_tuned.rds"))

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

# Import model:
#  damage(categorical = 1 => 10, else 0) ~ track_min_dist + wind_max_pred...
#  XGBoost model
base_class_full_model <- readRDS(here("unadjusted SCM/new base models",
                                "damage_fit_class_full.rds"))
```

```
# 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")
```

```r
# Predict to get wind_max_pred
#   model to use: base_wind_model

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

# ALWAYS MAKE SURE YOU'RE PREDICTING ON THE RIGHT MODEL
df_base_test[["rain_total_pred"]] <- predict(base_rain_model, newdata = df_base_test)

# Compute wind interaction terms dynamically
for (col in wind_fractions) {
  print(col)
  new_col_name <- paste0("wind_", col)
  df_base_test[[new_col_name]] <- df_base_test[[col]] * df_base_test[["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_pred
for (col in rain_fractions) {
  new_col_name <- paste0("rain_", col)
  df_base_test[[new_col_name]] <- df_base_test[[col]] * df_base_test[["rain_total_pred"]]
}
```

```r
df_base_test$damage_binary_2 <- factor(df_base_test$damage_binary,
                                       levels = c("0", "1"),  # Your current levels
                                       labels = c("Damage_below_10", "Damage_above_10"))  # New valid l
```

```r
# predict for damage_binary
# Make probability predictions for classification
y_preds_probs <- predict(base_class_full_model, newdata = df_base_test, type = "prob")[,2]  # Probabili
#y_preds_probs
```

```r
# AUC
# Compute AUC (better for classification)
auc_value <- auc(roc(df_base_test$damage_binary_2, y_preds_probs))
```

```
## Setting levels: control = Damage_below_10, case = Damage_above_10
```

```
## Setting direction: controls < cases
```

```r
auc_value
```

```
## Area under the curve: 0.9064
```

```r
# extracting probability that y_pred == 1
#y_preds_prob_1 <- y_preds_prob[ ,2]
```

```r
## assigning final class based on threshold
threshold = 0.3
y_pred <- ifelse(y_preds_probs > threshold, 1, 0)

y_pred  <- factor(y_pred, levels = c("0", "1"),  # Your current levels
                                    labels = c("Damage_below_10", "Damage_above_10"))  # New valid l
# using table function
conf_matrix <- confusionMatrix(as.factor(y_pred),
                    df_base_test$damage_binary_2,
                    positive = "Damage_above_10"
                    )
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Damage_below_10 Damage_above_10
##   Damage_below_10            1610              46
##   Damage_above_10              86              55
##
##                Accuracy : 0.9265
##                  95% CI : (0.9135, 0.9382)
##     No Information Rate : 0.9438
##     P-Value [Acc > NIR] : 0.9990351
##
##                   Kappa : 0.4163
##
##  Mcnemar's Test P-Value : 0.0006875
##
##             Sensitivity : 0.54455
##             Specificity : 0.94929
##          Pos Pred Value : 0.39007
##          Neg Pred Value : 0.97222
##              Prevalence : 0.05620
##          Detection Rate : 0.03061
##    Detection Prevalence : 0.07846
##       Balanced Accuracy : 0.74692
##
##        'Positive' Class : Damage_above_10
##
```

```r
# RESULTS FOR TABLE #3
# confusion matrix by regions
# Make sure the grouping variable is a factor
# Make sure island_groups is a factor
df_base_test$island_groups <- as.factor(df_base_test$island_groups)

# Loop through each group and generate a confusion matrix
for (grp in levels(df_base_test$island_groups)) {

  # Subset data for the current group
  group_indices <- df_base_test$island_groups == grp
  y_true_group <- df_base_test$damage_binary_2[group_indices]
```

```
  y_pred_group <- y_pred[group_indices]

  # Generate and print confusion matrix
  cat("Confusion Matrix for Island Group:", grp, "\n")
  print(confusionMatrix(y_pred_group, y_true_group, positive = "Damage_above_10"))
  cat("\n")
}
```

```
## Confusion Matrix for Island Group: Luzon
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Damage_below_10 Damage_above_10
##   Damage_below_10            1162              33
##   Damage_above_10              48              25
##
##                Accuracy : 0.9361
##                  95% CI : (0.9212, 0.9489)
##     No Information Rate : 0.9543
##     P-Value [Acc > NIR] : 0.9987
##
##                   Kappa : 0.3485
##
##  Mcnemar's Test P-Value : 0.1198
##
##             Sensitivity : 0.43103
##             Specificity : 0.96033
##          Pos Pred Value : 0.34247
##          Neg Pred Value : 0.97238
##              Prevalence : 0.04574
##          Detection Rate : 0.01972
##    Detection Prevalence : 0.05757
##       Balanced Accuracy : 0.69568
##
##        'Positive' Class : Damage_above_10
##
##
## Confusion Matrix for Island Group: Mindanao
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Damage_below_10 Damage_above_10
##   Damage_below_10             102               3
##   Damage_above_10               2               3
##
##                Accuracy : 0.9545
##                  95% CI : (0.8971, 0.9851)
##     No Information Rate : 0.9455
##     P-Value [Acc > NIR] : 0.4412
##
##                   Kappa : 0.5217
##
##  Mcnemar's Test P-Value : 1.0000
```

```
##
##             Sensitivity : 0.50000
##             Specificity : 0.98077
##          Pos Pred Value : 0.60000
##          Neg Pred Value : 0.97143
##              Prevalence : 0.05455
##          Detection Rate : 0.02727
##    Detection Prevalence : 0.04545
##       Balanced Accuracy : 0.74038
##
##        'Positive' Class : Damage_above_10
##
##
## Confusion Matrix for Island Group: Visayas
## Confusion Matrix and Statistics
##
##                  Reference
## Prediction        Damage_below_10 Damage_above_10
##   Damage_below_10             346              10
##   Damage_above_10              36              27
##
##                Accuracy : 0.8902
##                  95% CI : (0.8563, 0.9185)
##     No Information Rate : 0.9117
##     P-Value [Acc > NIR] : 0.9453285
##
##                   Kappa : 0.4824
##
##  Mcnemar's Test P-Value : 0.0002278
##
##             Sensitivity : 0.72973
##             Specificity : 0.90576
##          Pos Pred Value : 0.42857
##          Neg Pred Value : 0.97191
##              Prevalence : 0.08831
##          Detection Rate : 0.06444
##    Detection Prevalence : 0.15036
##       Balanced Accuracy : 0.81774
##
##        'Positive' Class : Damage_above_10
##
```

```r
# logging in mflow:
# Logging the model and parameter using MLflow

# 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))

# set experiment
# Logging metrics for model training and the parameters used
mlflow_set_experiment(experiment_name = "Attempt 2: R - U-SCM - XGBOOST classification -CV (Test metircs
```

```
## [1] "378093463115413703"
```

```
# 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 5ff528b99913417ea43~ 378093463115~ crawlin~ masinde RUNNI~ 2025-07-24 15:10:38
## # 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)

# Extract the best parameters (remove AUC column)
#best_params_model <- best_params %>% # Remove AUC column if present
#    select(-AUC)

parameters_used  <- base_class_full_model$bestTune

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

# Log the model type as a parameter
mlflow_log_param("model_type", "undj-scm-xgboost-classification")

# predicting
threshold = 0.3
y_preds_probs <- predict(base_class_full_model, newdata = df_base_test, type = "prob")[,2]  # Probabili
y_pred <- ifelse(y_preds_probs > threshold, 1, 0)

y_pred  <- factor(y_pred, levels = c("0", "1"),  # Your current levels
                                      labels = c("Damage_below_10", "Damage_above_10"))  # New valid l

# summarize results
conf_matrix <- confusionMatrix(as.factor(y_pred),
                    df_base_test$damage_binary_2,
                    positive = "Damage_above_10"
                    )

# accuracy
accuracy  <- conf_matrix$overall['Accuracy']
```

```
# Positive class = 1, precision, recall, and F1
# Extract precision, recall, and F1 score
precision <- conf_matrix$byClass['Precision']
recall <- conf_matrix$byClass['Recall']
f1_score <- conf_matrix$byClass['F1']
auc_value <- auc(roc(df_base_test$damage_binary_2, y_preds_probs))
```

```
## Setting levels: control = Damage_below_10, case = Damage_above_10
```

```
## Setting direction: controls < cases
```

```
# Log parameters and metrics
# mlflow_log_param("model_type", "scm-xgboost-classification")
mlflow_log_metric("accuracy", accuracy)
```

```
## 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.
```

```
mlflow_log_metric("F1", f1_score)
mlflow_log_metric("Precision", precision)
mlflow_log_metric("Recall", recall)
#mlflow_log_metric("AUC", auc_value)


# Save model
#saveRDS(model, file = file.path(path_2_folder, "spam_clas_model.rds"))

# 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 5ff528b99913417ea43~ 378093463115~ crawlin~ masinde FINIS~ 2025-07-24 15:10:38
## # i 7 more variables: end_time <dttm>, artifact_uri <chr>,
## #   lifecycle_stage <chr>, run_id <chr>, metrics <list>, params <list>,
## #   tags <list>
```

Table 2 Results: Recall, Precision and F1 results

```
cat("Recall of the unadjusted causal model:", recall, sep = " ", "\n")
```

```
## Recall of the unadjusted causal model: 0.5445545
```

```
cat("Precision of the unadjusted causal model:", precision, sep = " ", "\n")
```

```
## Precision of the unadjusted causal model: 0.3900709
```

```r
cat("F1_score of the unadjusted causal model:", f1_score, sep = " ", "\n")
```

```
## F1_score of the unadjusted causal model: 0.4545455
```

OLD CODE