

Final Model - death_30days

Eduardo Yuki Yada

Global parameters

```
k <- 5 # Number of folds for cross validation
grid_size <- 30 # Number of parameter combination to tune on each model
max_auc_loss <- 0.01 # Max accepted loss of AUC for reducing num of features
```

Imports

```
library(tidyverse)
library(yaml)
library(tidymodels)
library(usemodels)
library(vip)
library(kableExtra)
library(SHAPforxgboost)
library(xgboost)
library(Matrix)
library(mltools)
library(bonsai)
library(lightgbm)
library(pROC)
library(caret)
library(themis)

source("aux_functions.R")

select <- dplyr::select
```

Loading data

```
load('dataset/processed_data.RData')
load('dataset/processed_dictionary.RData')

columns_list <- yaml.load_file("./auxiliar/columns_list.yaml")

outcome_column <- params$outcome_column
features_list <- params$features_list

df[columns_list$outcome_columns] <- lapply(df[columns_list$outcome_columns], factor)
df <- mutate(df, across(where(is.character), as.factor))

dir.create(file.path("./auxiliar/final_model/hyperparameters/"),
          showWarnings = FALSE,
          recursive = TRUE)

dir.create(file.path("./auxiliar/final_model/performance/"),
          showWarnings = FALSE,
          recursive = TRUE)
```

```
dir.create(file.path("./auxiliar/final_model/selected_features/"),
          showWarnings = FALSE,
          recursive = TRUE)
```

Eligible features

```
eligible_columns <- df_names %>%
  filter(momento.aquisicao == 'Admissão t0') %>%
  .$variable.name

exception_columns <- c('death_intraop', 'death_intraop_1', 'disch_outcomes_t0')

correlated_columns = c('year_procedure_1', # com year_adm_t0
                      'age_surgery_1', # com age
                      'admission_t0', # com admission_pre_t0_count
                      'atb', # com meds_antimicrobianos
                      'classe_meds_cardio_qtde', # com classe_meds_qtde
                      'suporte_hemod', # com proced_invasivos_qtde,
                      'radiografia', # com exames_imagem_qtde
                      'ecg' # com metodos_graficos_qtde
                     )

eligible_features <- eligible_columns %>%
  base::intersect(c(columns_list$categorical_columns, columns_list$numerical_columns)) %>%
  setdiff(c(exception_columns, correlated_columns))

if (is.null(features_list)) {
  features = eligible_features
} else {
  features = base::intersect(eligible_features, features_list)
}
```

Starting features:

```
gluedown::md_order(features, seq = TRUE, pad = TRUE)
```

1. sex
2. age
3. education_level
4. underlying_heart_disease
5. heart_disease
6. nyha_basal
7. hypertension
8. prior_mi
9. heart_failure
10. af
11. valvopathy
12. diabetes
13. renal_failure
14. hemodialysis
15. cancer
16. comorbidities_count
17. procedure_type_1
18. reop_type_1
19. procedure_type_new
20. cied_final_1
21. cied_final_group_1
22. admission_pre_t0_count
23. admission_pre_t0_180d
24. year_adm_t0

25. icu_t0
26. antiaritmico
27. antihipertensivo
28. betabloqueador
29. dva
30. diuretico
31. vasodilatador
32. espironolactona
33. antiplaquetario_ev
34. insulina
35. psicofarmacos
36. antifungico
37. classe_meds_qtde
38. meds_cardiovasc_qtde
39. meds_antimicrobianos
40. vni
41. ventilacao_mecanica
42. intervencao_cv
43. cateter_venoso_central
44. proced_invasivos_qtde
45. transfusao
46. interconsulta
47. equipe_multiprof
48. holter
49. metodos_graficos_qtde
50. laboratorio
51. cultura
52. analises_clinicas_qtde
53. citologia
54. histopatologia_qtde
55. angio_tc
56. angiografia
57. cintilografia
58. ecocardiograma
59. flebografia
60. ultrassom
61. tomografia
62. ressonancia
63. exames_imagem_qtde
64. bic
65. hospital_stay

Train test split (70%/30%)

```
set.seed(42)

if (outcome_column == 'readmission_30d') {
  df_split <- readRDS("dataset/split_object.rds")
} else {
  df_split <- initial_split(df, prop = .7, strata = all_of(outcome_column))
}

df_train <- training(df_split) %>% select(all_of(c(features, outcome_column)))
df_test <- testing(df_split) %>% select(all_of(c(features, outcome_column)))

df_folds <- vfold_cv(df_train, v = k,
                      strata = all_of(outcome_column))
```

Feature Selection

```
model_fit_wf <- function(df_train, features, outcome_column, hyperparameters){  
  model_recipe <-  
    recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula,  
           data = df_train %>% select(all_of(c(features, outcome_column)))) %>%  
    step_novel(all_nominal_predictors()) %>%  
    step_unknown(all_nominal_predictors()) %>%  
    step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged")  
  
  model_spec <-  
    do.call(boost_tree, hyperparameters) %>%  
    set_engine("lightgbm") %>%  
    set_mode("classification")  
  
  model_workflow <-  
    workflow() %>%  
    add_recipe(model_recipe) %>%  
    add_model(model_spec)  
  
  model_fit_rs <- model_workflow %>%  
    fit_resamples(df_folds)  
  
  model_fit <- model_workflow %>%  
    fit(df_train)  
  
  model_auc <- validation(model_fit, df_test, plot = F)  
  
  raw_model <- parsnip::extract_fit_engine(model_fit)  
  
  feature_importance <- lgb.importance(raw_model, percentage = TRUE)  
  
  cv_results <- collect_metrics(model_fit_rs) %>% filter(.metric == 'roc_auc')  
  
  return(  
    list(  
      cv_auc = cv_results$mean,  
      cv_auc_std_err = cv_results$std_err,  
      importance = feature_importance,  
      auc = as.numeric(model_auc$auc),  
      auc_lower = model_auc$ci[1],  
      auc_upper = model_auc$ci[3]  
    )  
  )  
}  
  
hyperparameters <- readRDS(  
  sprintf(  
    "./auxiliar/model_selection/hyperparameters/lightgbm_%s.rds",  
    outcome_column  
  )  
)  
  
hyperparameters$sample_size <- 1  
  
full_model <- model_fit_wf(df_train, features, outcome_column, hyperparameters)  
  
sprintf('Full Model CV Train AUC: %.3f' ,full_model$cv_auc)  
  
## [1] "Full Model CV Train AUC: 0.791"
```

```

sprintf('Full Model Test AUC: %.3f' ,full_model$auc)

## [1] "Full Model Test AUC: 0.783"

Features with zero importance on the initial model:
unimportant_features <- setdiff(features, full_model$importance$Feature)

unimportant_features %>%
  gluedown::md_order()

1. underlying_heart_disease
2. heart_disease
3. nyha_basal
4. hemodialysis
5. intervencao_cv

trimmed_features <- full_model$importance$Feature
hyperparameters$mtry <- min(hyperparameters$mtry, length(trimmed_features))
trimmed_model <- model_fit_wf(df_train, trimmed_features,
                                outcome_column, hyperparameters)

sprintf('Trimmed Model CV Train AUC: %.3f' ,trimmed_model$cv_auc)

## [1] "Trimmed Model CV Train AUC: 0.752"

sprintf('Trimmed Model Test AUC: %.3f' ,trimmed_model$auc)

## [1] "Trimmed Model Test AUC: 0.782"

selection_results <- tibble::tribble(
  ~`Number of Features`, ~`CV AUC`, ~`CV AUC Std Error`, ~`AUC Loss`, ~`Least Important Feature`,
  length(features), full_model$cv_auc, full_model$cv_auc_std_err, 0, tail(full_model$importance$Feature, 1)
)

if (full_model$cv_auc - trimmed_model$cv_auc < max_auc_loss) {
  current_features <- trimmed_features
  current_model <- trimmed_model
  current_least_important <- tail(current_model$importance$Feature, 1)
  current_auc_loss <- full_model$cv_auc - current_model$cv_auc

  selection_results <- selection_results %>%
    add_row(`Number of Features` = length(trimmed_features),
            `CV AUC` = current_model$cv_auc,
            `CV AUC Std Error` = current_model$cv_auc_std_err,
            `AUC Loss` = current_auc_loss,
            `Least Important Feature` = current_least_important)
} else {
  current_features <- features
  current_model <- full_model
  current_least_important <- tail(current_model$importance$Feature, 1)
  current_auc_loss <- 0
}

while (current_auc_loss < max_auc_loss) {
  last_feature_dropped <- current_least_important

  current_features <- setdiff(current_features, current_least_important)
  hyperparameters$mtry <- min(hyperparameters$mtry, length(current_features))
  current_model <- model_fit_wf(df_train, current_features, outcome_column, hyperparameters)
  current_least_important <- tail(current_model$importance$Feature, 1)

  current_auc_loss <- full_model$cv_auc - current_model$cv_auc
}

```

```

selection_results <- selection_results %>%
  add_row(`Number of Features` = length(current_features),
         `CV AUC` = current_model$cv_auc,
         `CV AUC Std Error` = current_model$cv_auc_std_err,
         `AUC Loss` = current_auc_loss,
         `Least Important Feature` = current_least_important)

# print(c(length(current_features), current_auc_loss))
}

selection_results %>% niceFormatting(digits = 4, label = 1)

```

Table 1:

Number of Features	CV AUC	CV AUC Std Error	AUC Loss	Least Important Feature
65	0.7908	0.0332	0.0000	vni
64	0.7719	0.0421	0.0189	underlying_heart_disease

```

if (exists('last_feature_dropped')) {
  selected_features <- c(current_features, last_feature_dropped)
} else {
  selected_features <- current_features
}

con <- file(sprintf('./auxiliar/final_model/selected_features/%s.yaml', outcome_column), "w")
write_yaml(selected_features, con)
close(con)

feature_selected_model <- model_fit_wf(df_train, selected_features,
                                         outcome_column, hyperparameters)

sprintf('Selected Model CV Train AUC: %.3f', feature_selected_model$cv_auc)

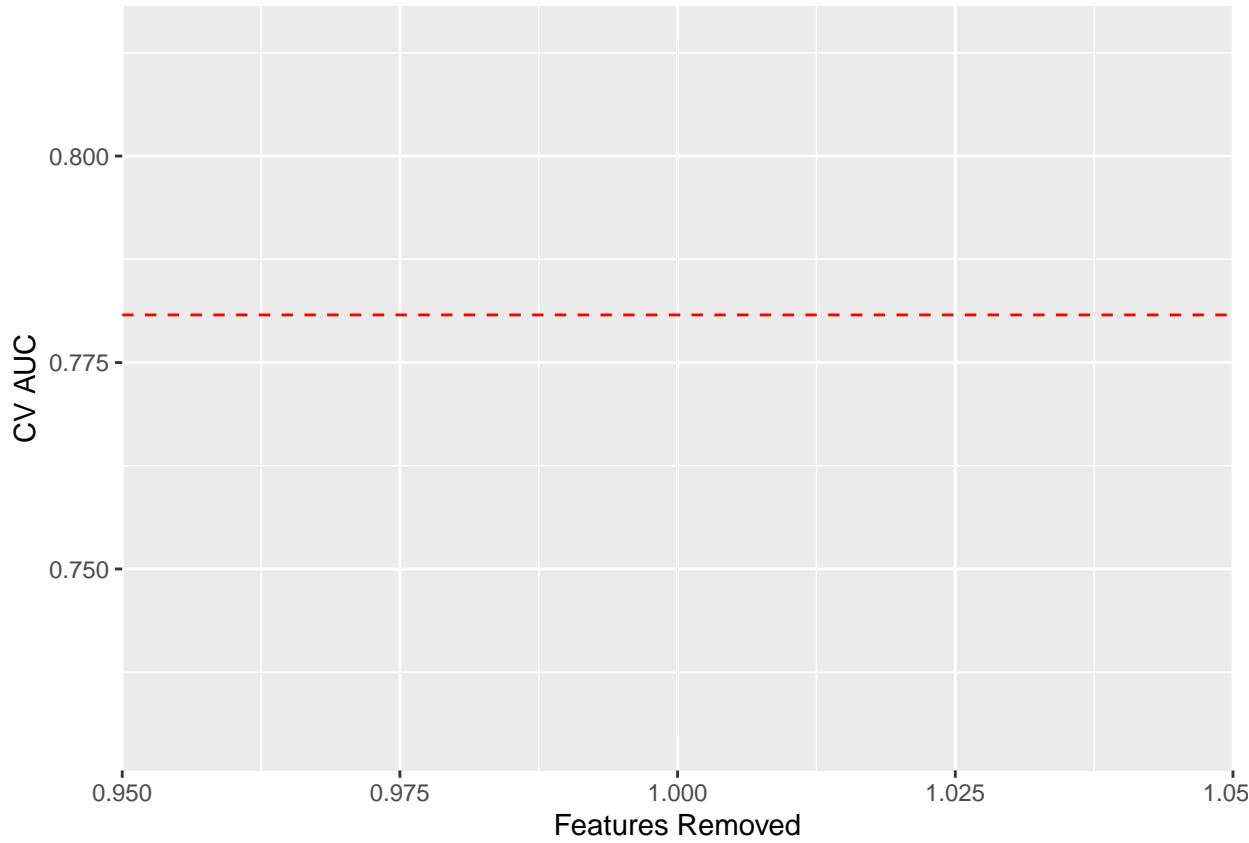
## [1] "Selected Model CV Train AUC: 0.763"
sprintf('Selected Model Test AUC: %.3f', feature_selected_model$auc)

## [1] "Selected Model Test AUC: 0.790"

selection_results <- selection_results %>%
  filter(`Number of Features` < length(features)) %>%
  mutate(`Features Removed` = length(features) - `Number of Features`,
        `CV AUC Low` = `CV AUC` - `CV AUC Std Error`,
        `CV AUC High` = `CV AUC` + `CV AUC Std Error`)

selection_results %>%
  ggplot(aes(x = `Features Removed`, y = `CV AUC`,
             ymin = `CV AUC Low`, ymax = `CV AUC High`)) +
  geom_line() +
  geom_ribbon(alpha = .3) +
  geom_hline(yintercept = full_model$cv_auc - max_auc_loss,
             linetype = "dashed", color = "red")

```



```
# selection_results %>%
#   filter(`Number of Features` < length(features)) %>%
#   mutate(`Features Removed` = length(features) - `Number of Features`) %>%
#   ggplot(aes(x = `Features Removed`, y = `AUC Loss`)) +
#   geom_line()
```

Hyperparameter tuning

Selected features:

```
gluedown::md_order(selected_features, seq = TRUE, pad = TRUE)
```

1. sex
2. age
3. education_level
4. underlying_heart_disease
5. heart_disease
6. nyha_basal
7. hypertension
8. prior_mi
9. heart_failure
10. af
11. valvopathy
12. diabetes
13. renal_failure
14. hemodialysis
15. cancer
16. comorbidities_count
17. procedure_type_1
18. reop_type_1
19. procedure_type_new
20. cied_final_1
21. cied_final_group_1

```

22. admission_pre_t0_count
23. admission_pre_t0_180d
24. year_adm_t0
25. icu_t0
26. antiarritmico
27. antihipertensivo
28. betabloqueador
29. dva
30. diuretico
31. vasodilatador
32. espironolactona
33. antiplaquetario_ev
34. insulina
35. psicofarmacos
36. antifungico
37. classe_meds_qtd
38. meds_cardiovasc_qtd
39. meds_antimicrobianos
40. ventilacao_mecanica
41. intervencao_cv
42. cateter_venoso_central
43. proced_invasivos_qtd
44. transfusao
45. interconsulta
46. equipe_multiprof
47. holter
48. metodos_graficos_qtd
49. laboratorio
50. cultura
51. analises_clinicas_qtd
52. citologia
53. histopatologia_qtd
54. angio_tc
55. angiografia
56. cintilografia
57. ecocardiograma
58. flebografia
59. ultrassom
60. tomografia
61. ressonancia
62. exames_imagem_qtd
63. bic
64. hospital_stay

```

Standard

```

lightgbm_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula,
         data = df_train %>% select(all_of(c(selected_features, outcome_column)))) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%
  step_dummy(all_nominal_predictors())

```

```

lightgbm_smote_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula,
         data = df_train %>% select(all_of(c(selected_features, outcome_column)))) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%
  step_dummy(all_nominal_predictors())

```

```

step_impute_mean(all_numeric_predictors()) %>%
step_smote(!!sym(outcome_column))

lightgbm_upsample_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula,
         data = df_train %>% select(all_of(c(selected_features, outcome_column)))) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%
  step_dummy(all_nominal_predictors()) %>%
  step_upsample(!!sym(outcome_column))

lightgbm_tuning <- function(recipe) {

  lightgbm_spec <- boost_tree(
    mtry = tune(),
    trees = tune(),
    min_n = tune(),
    tree_depth = tune(),
    learn_rate = tune(),
    loss_reduction = tune(),
    sample_size = 1.0
  ) %>%
    set_engine("lightgbm") %>%
    set_mode("classification")

  lightgbm_grid <- grid_latin_hypercube(
    mtry(range = c(1L, length(selected_features))),
    trees(range = c(100L, 300L)),
    min_n(),
    tree_depth(),
    learn_rate(),
    loss_reduction(),
    size = grid_size
  )
}

lightgbm_workflow <-
  workflow() %>%
  add_recipe(recipe) %>%
  add_model(lightgbm_spec)

lightgbm_tune <-
  lightgbm_workflow %>%
  tune_grid(resamples = df_folds,
            grid = lightgbm_grid)

lightgbm_tune %>%
  show_best("roc_auc") %>%
  niceFormatting(digits = 5, label = 4)

best_lightgbm <- lightgbm_tune %>%
  select_best("roc_auc")

lightgbm_tune %>%
  collect_metrics() %>%
  filter(.metric == "roc_auc") %>%
  select(mean, mtry:tree_depth) %>%
  pivot_longer(mtry:tree_depth,
              values_to = "value",
              names_to = "parameter"
  ) %>%

```

```

ggplot(aes(value, mean, color = parameter)) +
  geom_point(alpha = 0.8, show.legend = FALSE) +
  facet_wrap(~parameter, scales = "free_x") +
  labs(x = NULL, y = "AUC")

final_lightgbm_workflow <-
  lightgbm_workflow %>%
  finalize_workflow(best_lightgbm)

last_lightgbm_fit <-
  final_lightgbm_workflow %>%
  last_fit(df_split)

final_lightgbm_fit <- extract_workflow(last_lightgbm_fit)

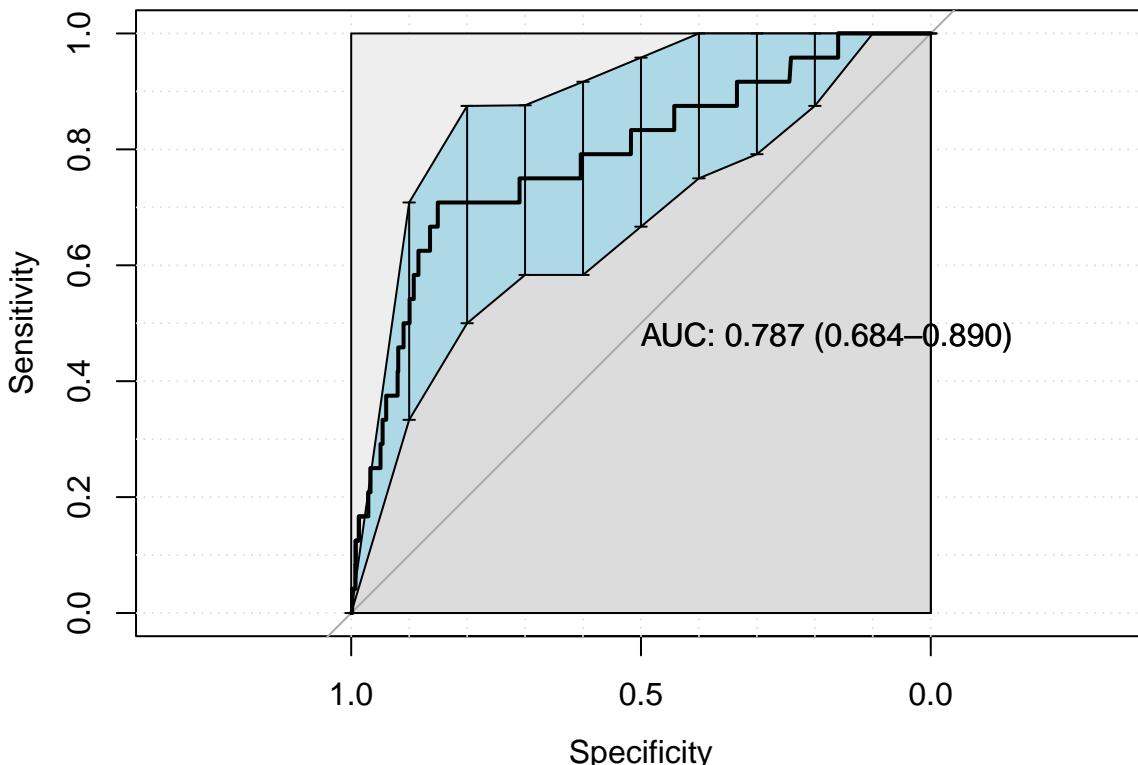
lightgbm_auc <- validation(final_lightgbm_fit, df_test)

lightgbm_parameters <- lightgbm_tune %>%
  show_best("roc_auc", n = 1) %>%
  select(trees, mtry, min_n, tree_depth, learn_rate, loss_reduction) %>%
  as.list

return(list(auc = as.numeric(lightgbm_auc$auc),
            auc_lower = lightgbm_auc$ci[1],
            auc_upper = lightgbm_auc$ci[3],
            parameters = lightgbm_parameters,
            fit = final_lightgbm_fit))
}

standard_results <- lightgbm_tuning(lightgbm_recipe)

```



```

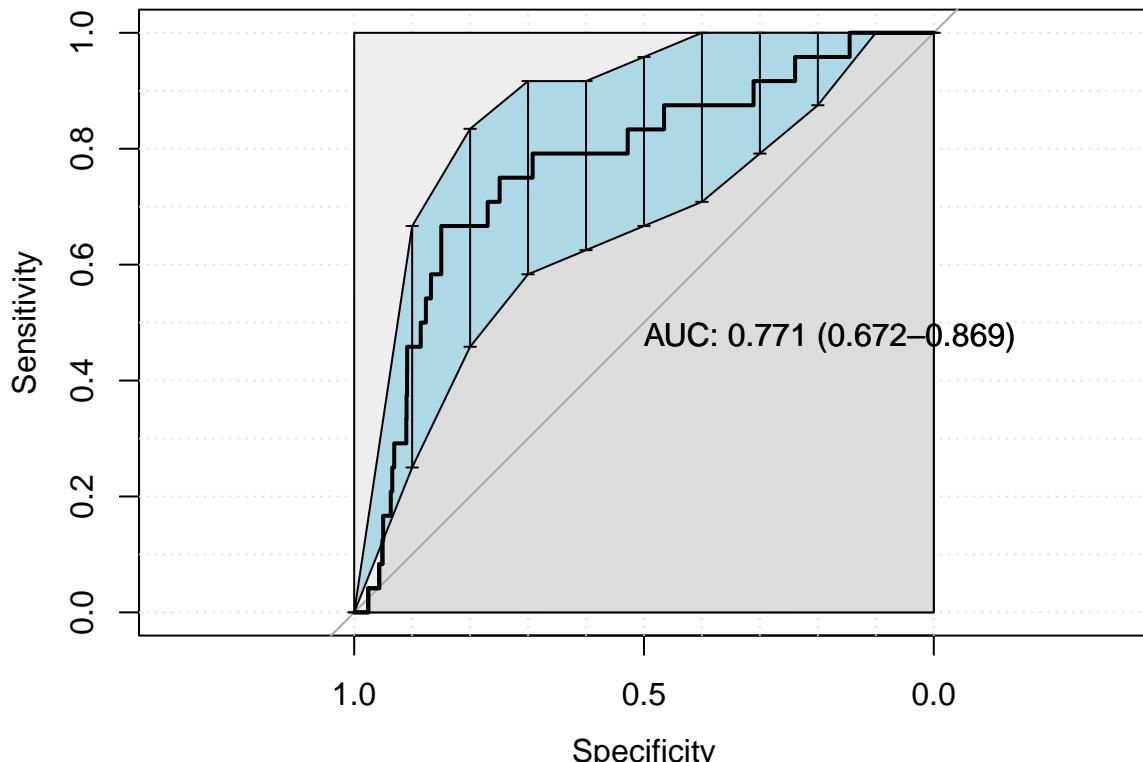
## [1] "Optimal Threshold: 0.00"
## Confusion Matrix and Statistics

```

```

## reference
## data    0    1
##    0 4003    7
##    1  703   17
##
##           Accuracy : 0.8499
##           95% CI : (0.8394, 0.86)
## No Information Rate : 0.9949
## P-Value [Acc > NIR] : 1
##
##           Kappa : 0.0362
##
## McNemar's Test P-Value : <2e-16
##
##           Sensitivity : 0.85062
##           Specificity : 0.70833
## Pos Pred Value : 0.99825
## Neg Pred Value : 0.02361
## Prevalence : 0.99493
## Detection Rate : 0.84630
## Detection Prevalence : 0.84778
## Balanced Accuracy : 0.77947
##
## 'Positive' Class : 0
##
smote_results <- lightgbm_tuning(lightgbm_smote_recipe)

```



```

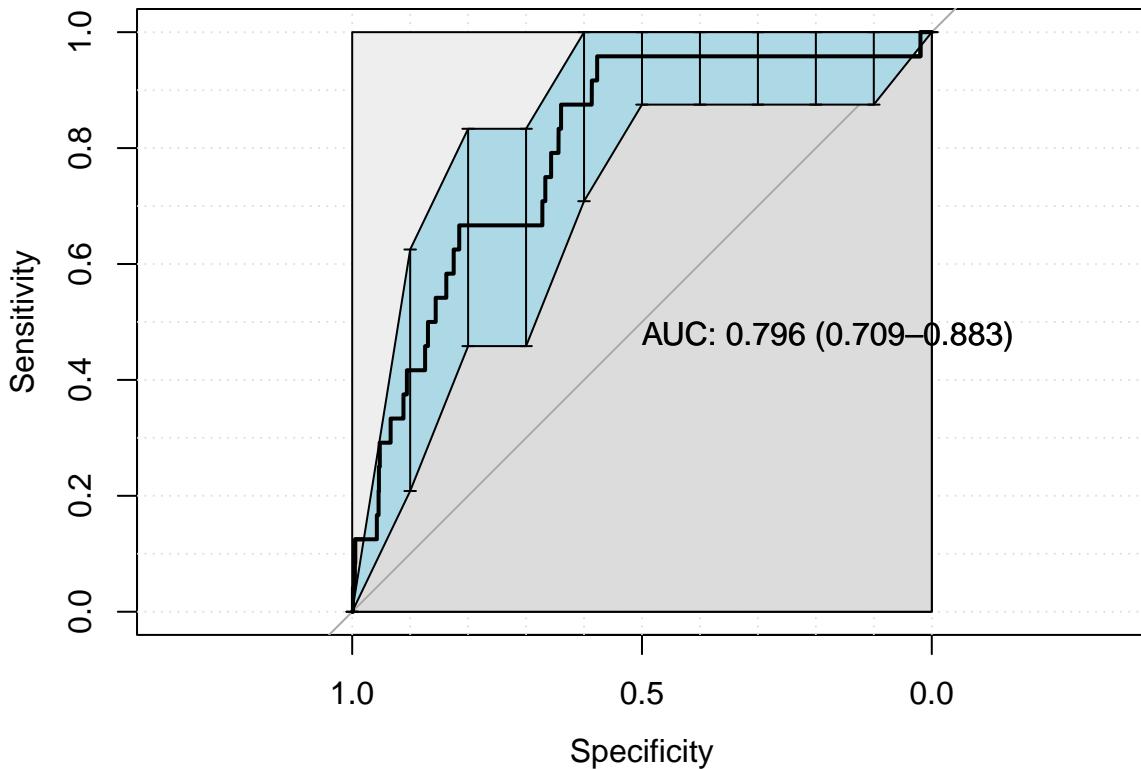
## [1] "Optimal Threshold: 0.50"
## Confusion Matrix and Statistics
##
## reference
## data    0    1

```

```

##      0 3999     8
##      1 707    16
##
##          Accuracy : 0.8488
##  95% CI : (0.8383, 0.8589)
##  No Information Rate : 0.9949
##  P-Value [Acc > NIR] : 1
##
##          Kappa : 0.0333
##
##  Mcnemar's Test P-Value : <2e-16
##
##          Sensitivity : 0.84977
##          Specificity : 0.66667
##  Pos Pred Value : 0.99800
##  Neg Pred Value : 0.02213
##          Prevalence : 0.99493
##          Detection Rate : 0.84545
##  Detection Prevalence : 0.84715
##          Balanced Accuracy : 0.75822
##
##          'Positive' Class : 0
##
upsample_results <- lightgbm_tuning(lightgbm_upsample_recipe)

```



```

## [1] "Optimal Threshold: 0.39"
## Confusion Matrix and Statistics
##
##      reference
## data      0      1
##      0 2717      1
##      1 1989     23
##
```

```

##          Accuracy : 0.5793
## 95% CI : (0.5651, 0.5934)
## No Information Rate : 0.9949
## P-Value [Acc > NIR] : 1
##
##          Kappa : 0.0127
##
## McNemar's Test P-Value : <2e-16
##
##          Sensitivity : 0.57735
##          Specificity : 0.95833
## Pos Pred Value : 0.99963
## Neg Pred Value : 0.01143
##          Prevalence : 0.99493
## Detection Rate : 0.57442
## Detection Prevalence : 0.57463
## Balanced Accuracy : 0.76784
##
## 'Positive' Class : 0
##
final_lightgbm_fit <- standard_results$fit
lightgbm_parameters <- standard_results$parameters

# saveRDS(
#   lightgbm_parameters,
#   file = sprintf(
#     "./auxiliar/final_model/hyperparameters/lightgbm_%s.rds",
#     outcome_column
#   )
# )

```

SHAP values

```

lightgbm_model <- parsnip::extract_fit_engine(final_lightgbm_fit)

trained_rec <- prep(lightgbm_recipe, training = df_train)

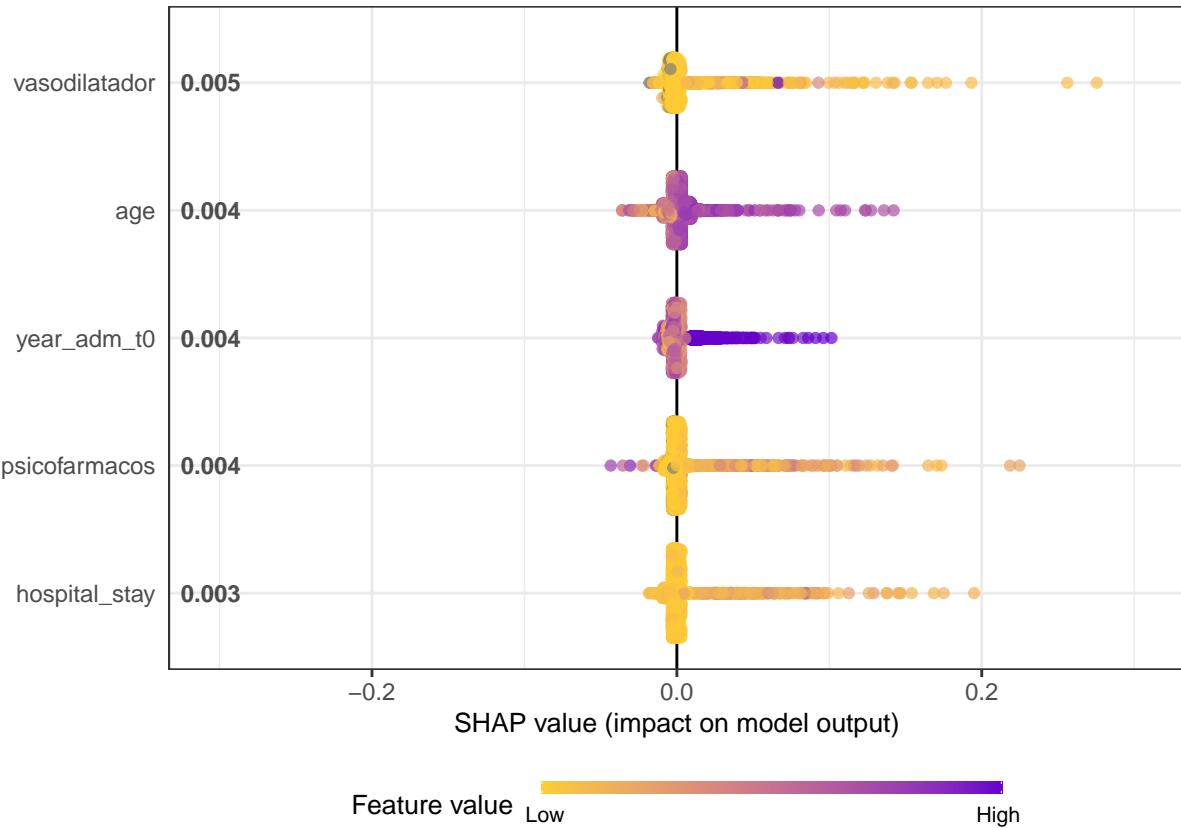
df_train_baked <- bake(trained_rec, new_data = df_train)
df_test_baked <- bake(trained_rec, new_data = df_test)

matrix_train <- as.matrix(df_train_baked %>% select(-all_of(outcome_column)))
matrix_test <- as.matrix(df_test_baked %>% select(-all_of(outcome_column)))

n_plots <- min(5, length(selected_features))

shap.plot.summary.wrap1(model = lightgbm_model, X = matrix_train,
                       top_n = n_plots, dilute = F)

```



```

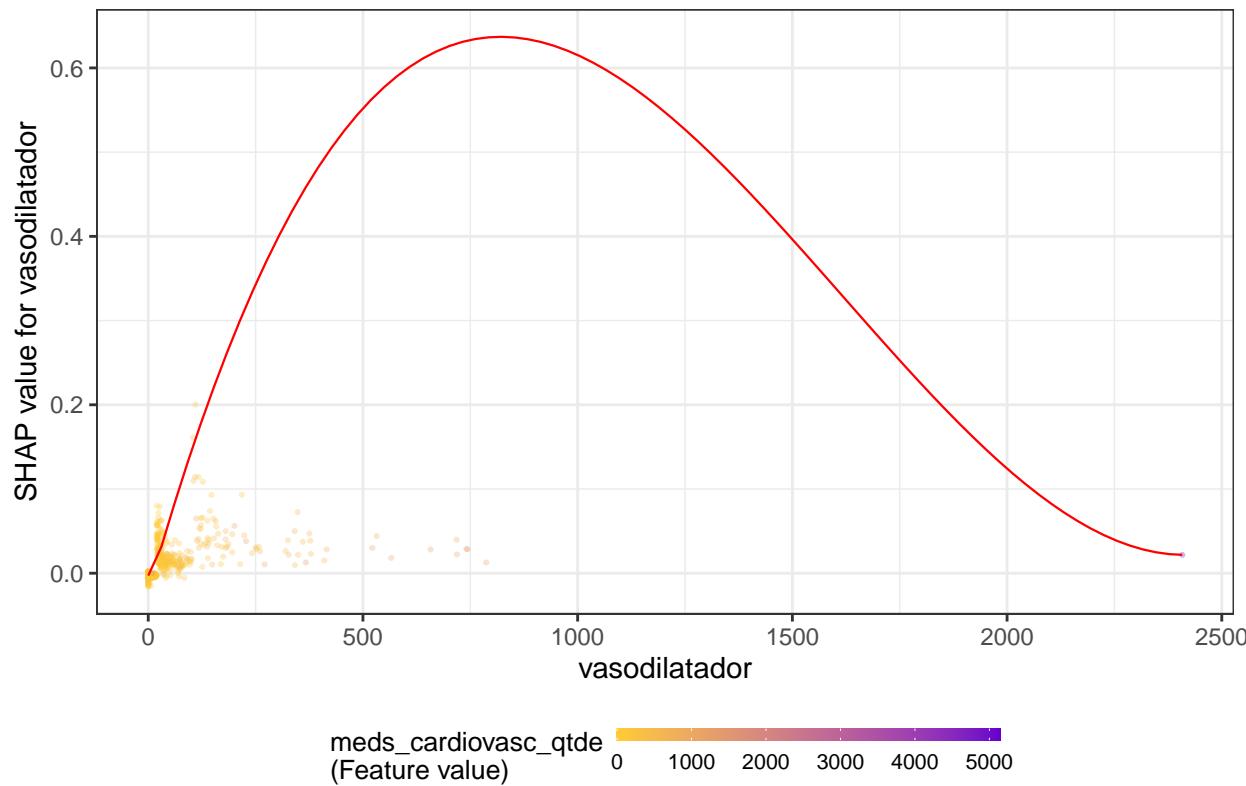
shap <- shap.prep(lightgbm_model, X_train = matrix_test)

for (x in shap.importance(shap, names_only = TRUE)[1:n_plots]) {
  p <- shap.plot.dependence(
    shap,
    x = x,
    color_feature = "auto",
    smooth = TRUE,
    jitter_width = 0.01,
    alpha = 0.3
  ) +
    labs(title = x)
  print(p)
}

## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 1055 rows containing non-finite values (stat_smooth).
## Warning: Removed 1055 rows containing missing values (geom_point).

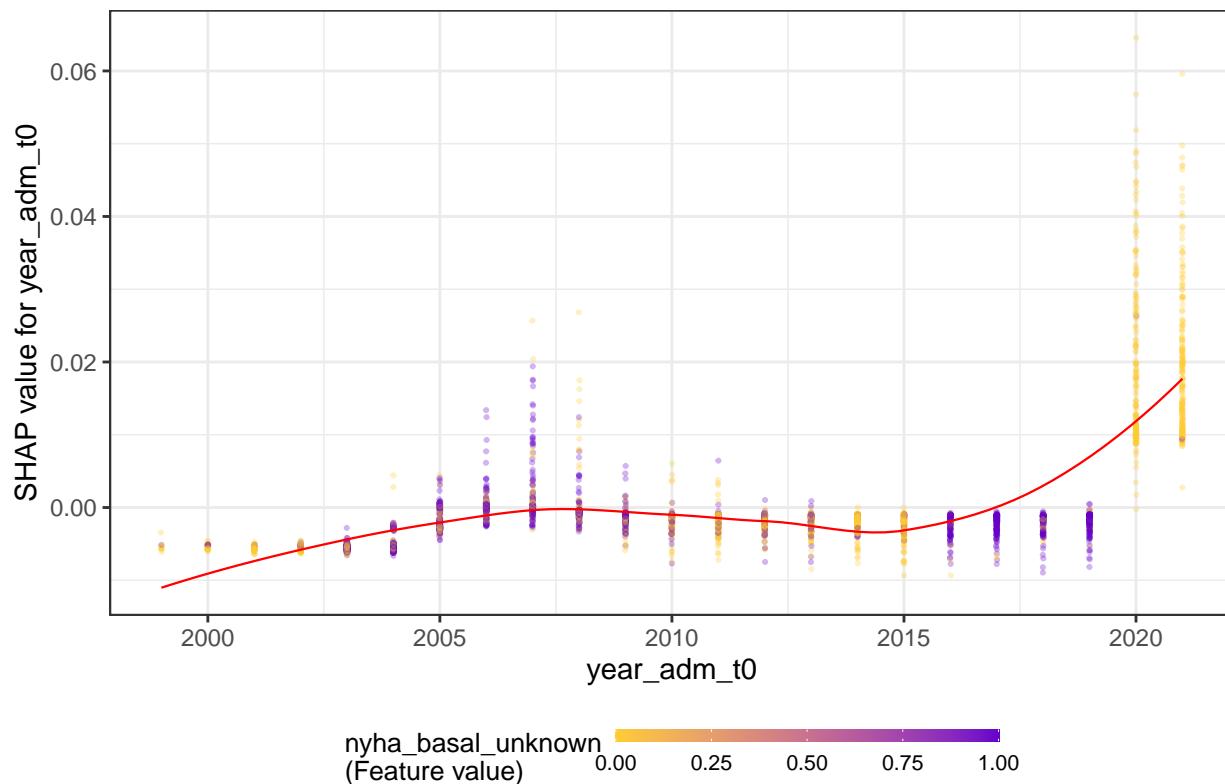
```

vasodilatador

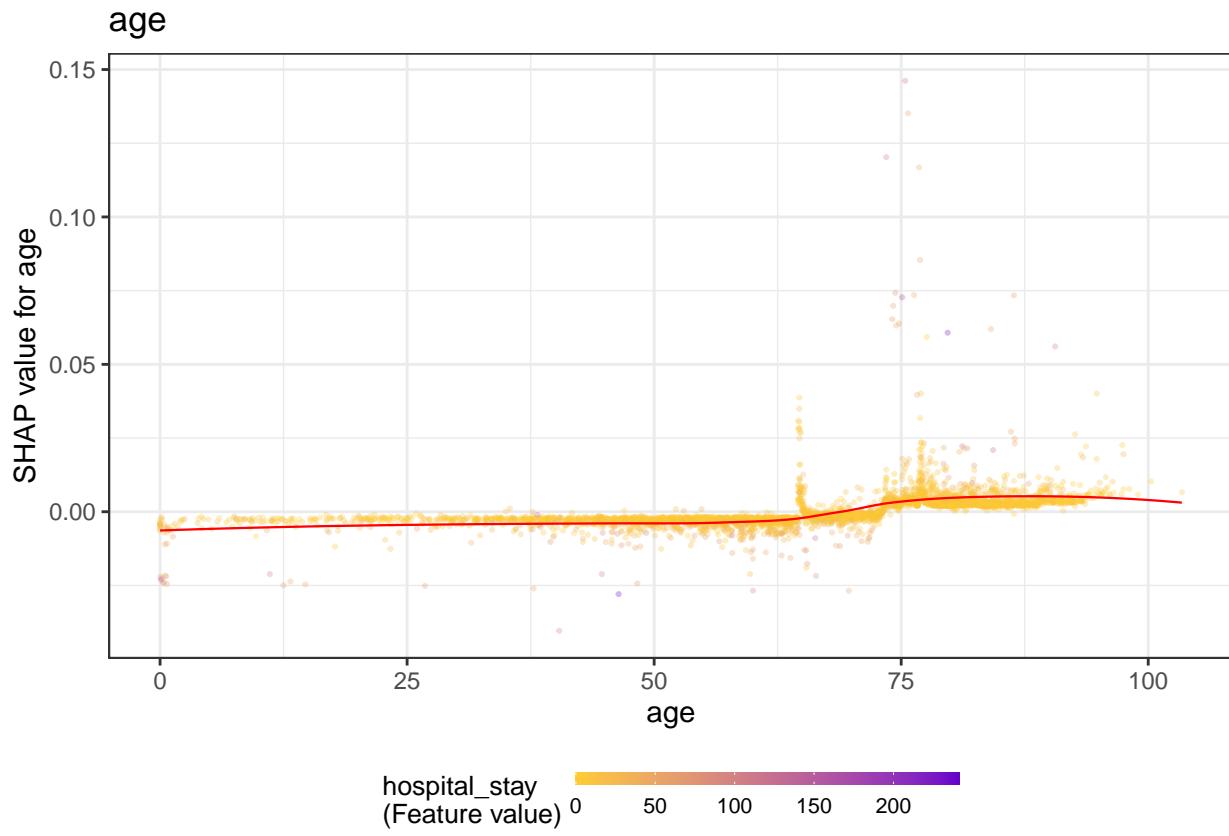


```
## `geom_smooth()` using formula 'y ~ x'  
## Warning: Removed 7 rows containing non-finite values (stat_smooth).  
## Warning: Removed 7 rows containing missing values (geom_point).
```

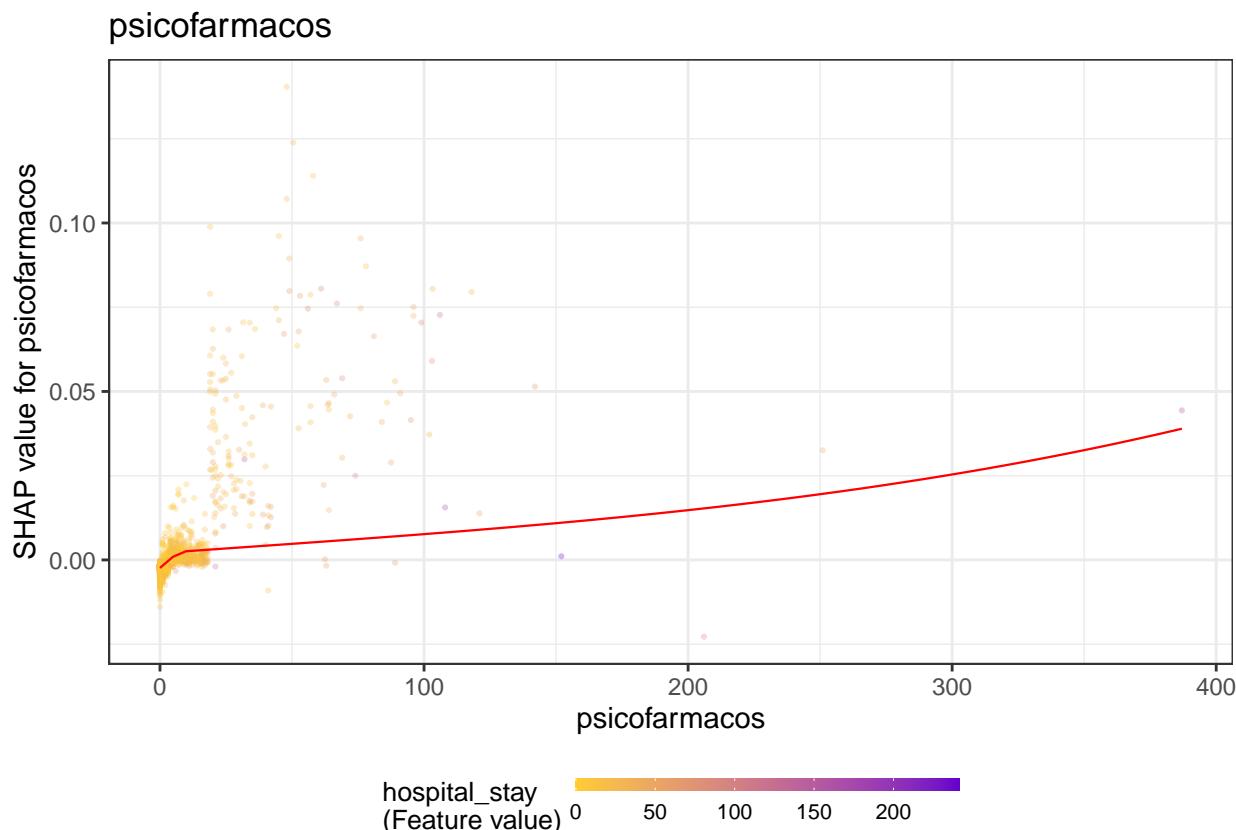
year_adm_t0



```
## `geom_smooth()` using formula 'y ~ x'
```

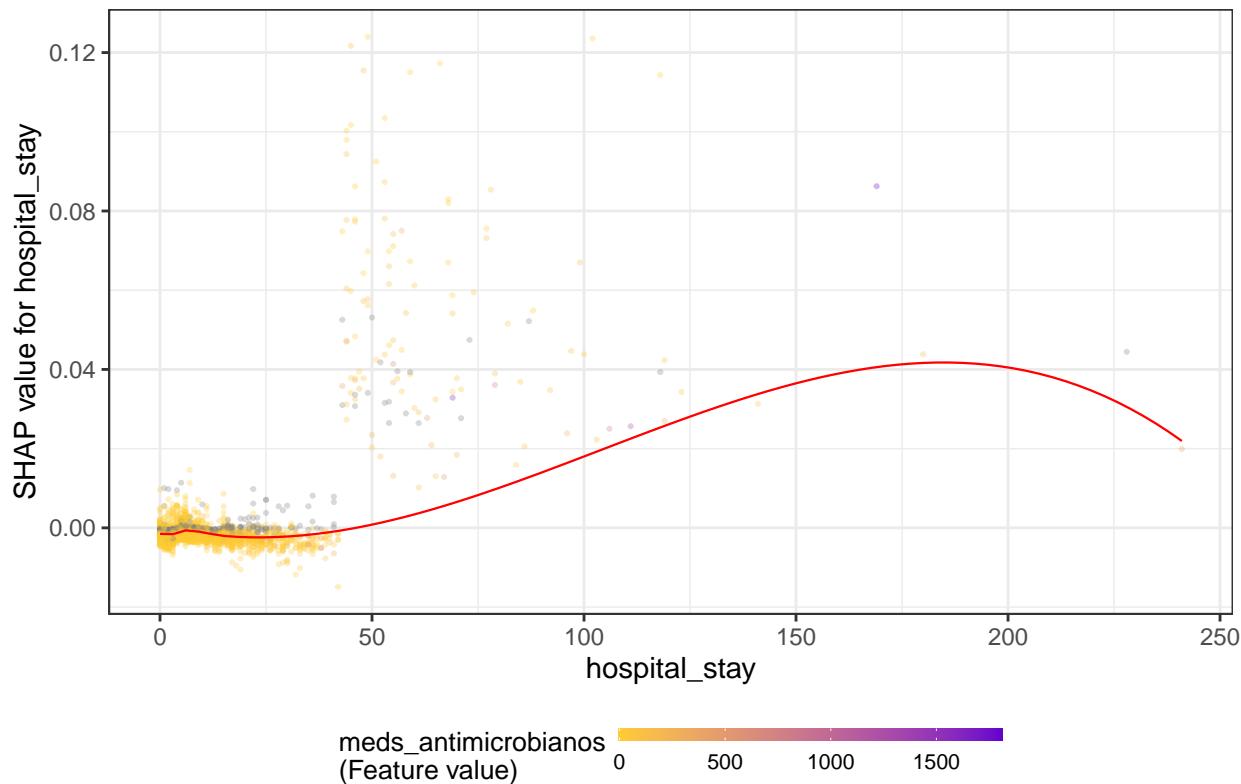


```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 1055 rows containing non-finite values (stat_smooth).
## Warning: Removed 1055 rows containing missing values (geom_point).
```



```
## `geom_smooth()` using formula 'y ~ x'
```

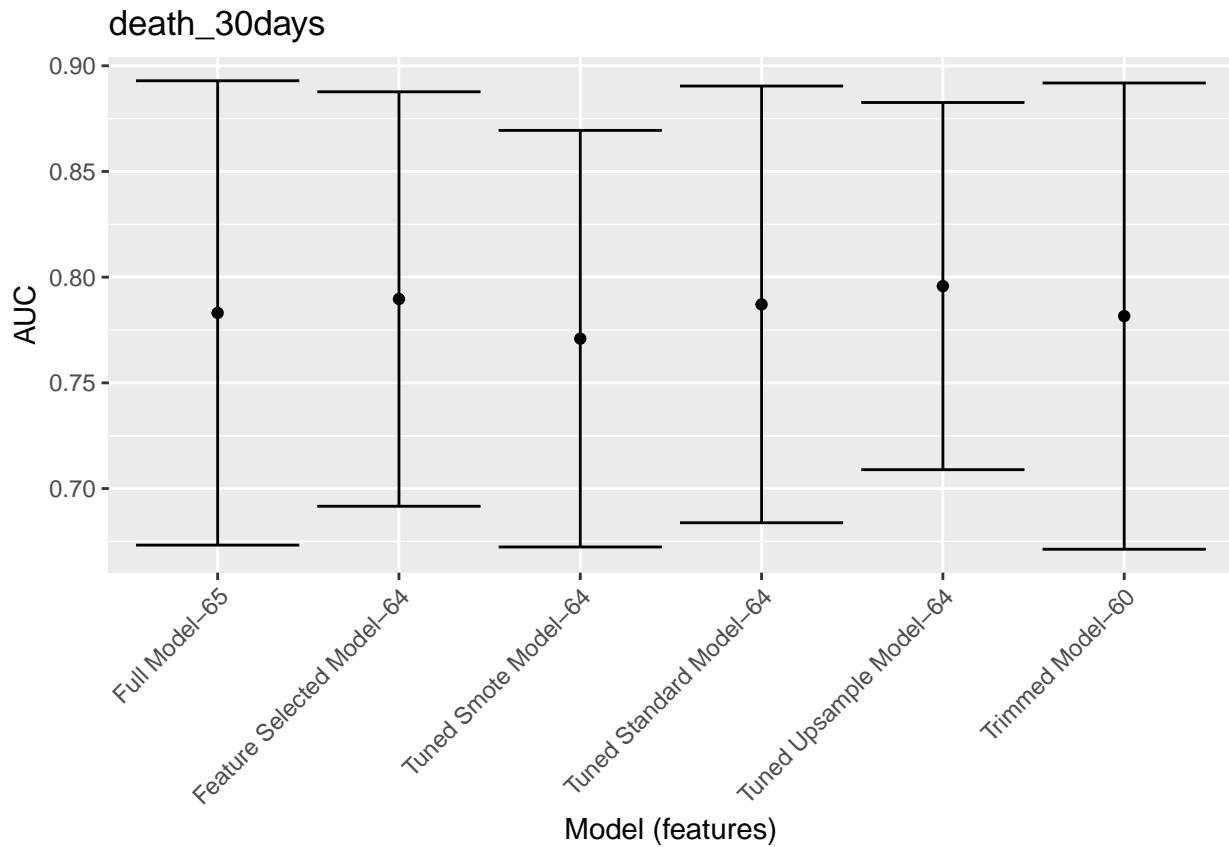
hospital_stay



Models Comparison

```
df_auc <- tibble::tribble(
  ~Model, ~`AUC`, ~`Lower Limit`, ~`Upper Limit`, ~`Features`,
  'Full Model', full_model$auc, full_model$auc_lower, full_model$auc_upper, length(features),
  'Trimmed Model', trimmed_model$auc, trimmed_model$auc_lower, trimmed_model$auc_upper, length(trimmed_features),
  'Feature Selected Model', feature_selected_model$auc, feature_selected_model$auc_lower, feature_selected_model$auc_upper, length(selected_features),
  'Tuned Standard Model', standard_results$auc, standard_results$auc_lower, standard_results$auc_upper, length(standard_features),
  'Tuned Smote Model', smote_results$auc, smote_results$auc_lower, smote_results$auc_upper, length(selected_features),
  'Tuned Upsample Model', upsample_results$auc, upsample_results$auc_lower, upsample_results$auc_upper, length(upsample_features)
) %>%
  mutate(Target = outcome_column,
    `Model (features)` = fct_reorder(paste0(Model, " - ", Features), -Features))

df_auc %>%
  ggplot(aes(
    x = `Model (features)` ,
    y = AUC,
    ymin = `Lower Limit` ,
    ymax = `Upper Limit` )
  ) +
  geom_point() +
  geom_errorbar() +
  labs(title = outcome_column) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
saveRDS(df_auc, sprintf("./auxiliar/final_model/performance/%s.RData", outcome_column))
```