

Model Selection - death_2year

Eduardo Yuki Yada

Global parameters

```
k = 5 # Number of folds for cross validation
grid_size = 15 # Number of parameter combination to tune on each model
```

Minutes to run: 0

Imports

```
library(tidyverse)
library(yaml)
library(tidymodels)
library(usemodels)
library(vip)
library(bonsai)
library(lightgbm)
library(caret)
library(pROC)

source("aux_functions.R")
```

Minutes to run: 0

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 <- mutate(df, across(where(is.character), as.factor))
```

Minutes to run: 0.006

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

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

Minutes to run: 0

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$categories_columns, columns_list$numerical_columns)) %>%
  setdiff(c(exception_columns, correlated_columns))

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

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

```
## 01. sex
## 02. age
## 03. education_level
## 04. underlying_heart_disease
## 05. heart_disease
## 06. nyha_basal
## 07. hypertension
## 08. prior_mi
## 09. heart_failure
## 10. af
## 11. cardiac_arrest
## 12. valvopathy
## 13. diabetes
## 14. renal_failure
## 15. hemodialysis
## 16. stroke
## 17. copd
## 18. comorbidities_count
## 19. procedure_type_1
## 20. reop_type_1
## 21. procedure_type_new
## 22. cied_final_1
## 23. cied_final_group_1
## 24. admission_pre_t0_count
## 25. admission_pre_t0_180d
## 26. year_adm_t0
## 27. icu_t0
## 28. dialysis_t0
## 29. admission_t0_emergency
## 30. aco
## 31. antiarritmico
```

```

## 32. ieca_bra
## 33. dva
## 34. digoxina
## 35. estatina
## 36. diuretico
## 37. vasodilatador
## 38. insuf_cardiaca
## 39. espironolactona
## 40. antiplaquetario_ev
## 41. insulina
## 42. psicofarmacos
## 43. antifungico
## 44. antiviral
## 45. classe_meds_qtde
## 46. meds_cardiovasc_qtde
## 47. meds_antimicrobianos
## 48. vni
## 49. ventilacao_mecanica
## 50. transplante_cardiaco
## 51. outros_proced_cirurgicos
## 52. icp
## 53. angioplastia
## 54. cateterismo
## 55. cateter_venoso_central
## 56. proced_invasivos_qtde
## 57. transfusao
## 58. interconsulta
## 59. equipe_multiprof
## 60. holter
## 61. teste_esforco
## 62. tilt_teste
## 63. metodos_graficos_qtde
## 64. laboratorio
## 65. cultura
## 66. analises_clinicas_qtde
## 67. citologia
## 68. histopatologia_qtde
## 69. angio_tc
## 70. cintilografia
## 71. ecocardiograma
## 72. endoscopia
## 73. flebografia
## 74. pet_ct
## 75. ultrassom
## 76. tomografia
## 77. ressonancia
## 78. exames_imagem_qtde
## 79. bic
## 80. hospital_stay

```

Minutes to run: 0

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) %>% dplyr::select(all_of(c(features, outcome_column)))
df_test <- testing(df_split) %>% dplyr::select(all_of(c(features, outcome_column)))

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

Minutes to run: 0.001

Boosted Tree (XGBoost)

```
xgboost_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  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())

xgboost_spec <- boost_tree(
  mtry = tune(),
  trees = tune(),
  min_n = tune(),
  tree_depth = tune(),
  learn_rate = tune(),
  loss_reduction = tune()
) %>%
  set_engine("xgboost",
             nthread = 8) %>%
  set_mode("classification")

xgboost_grid <- grid_latin_hypercube(
  finalize(mtry(), df_train),
  trees(range = c(100L, 300L)),
  min_n(),
  tree_depth(),
  learn_rate(),
  loss_reduction(),
  size = grid_size
)

xgboost_workflow <-
  workflow() %>%
  add_recipe(xgboost_recipe) %>%
  add_model(xgboost_spec)

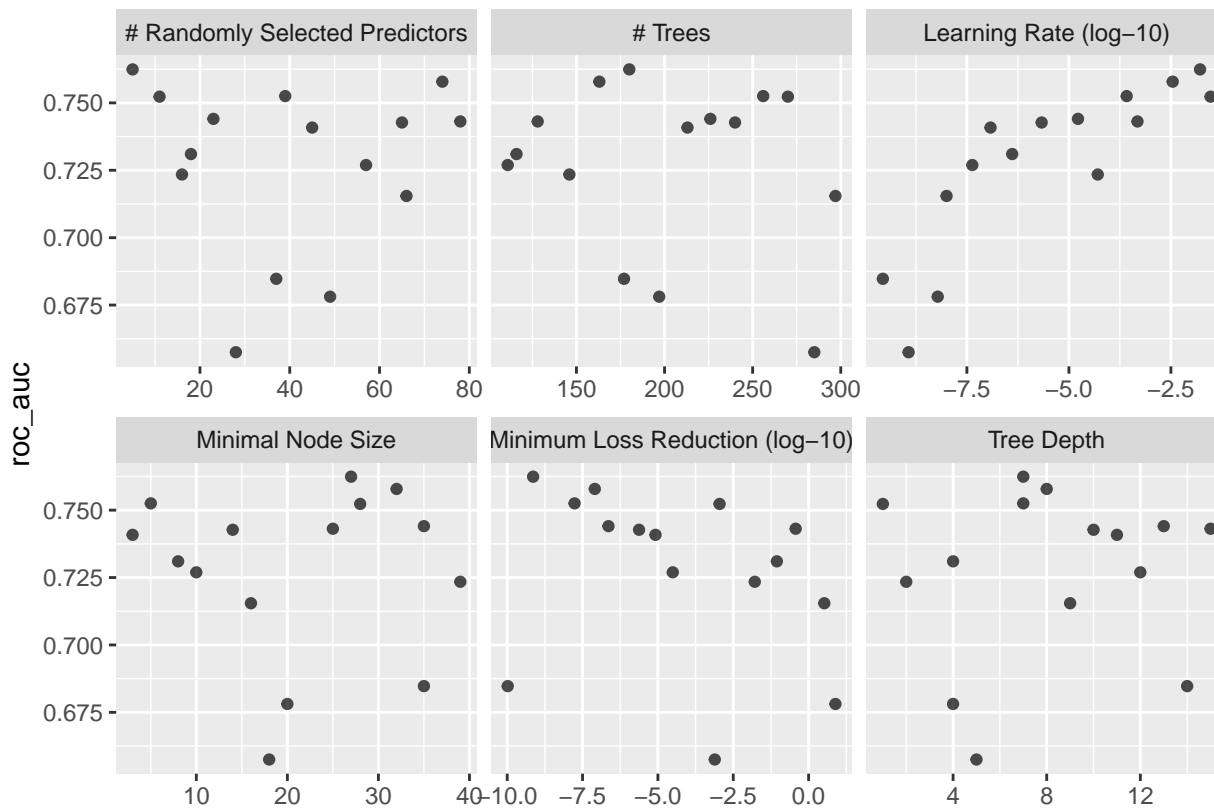
xgboost_tune <-
  xgboost_workflow %>%
  tune_grid(resamples = df_folds,
            grid = xgboost_grid)

xgboost_tune %>%
  show_best("roc_auc")
```

```
## # A tibble: 5 x 12
##   mtry trees min_n tree_depth learn_rate loss_reduction .metric .estimator mean n std_err .config
##   <int> <int> <int>    <int>    <dbl>    <dbl> <chr>    <chr>    <dbl> <int>  <dbl> <chr>
## 1     5   180    27        7  0.0164    7.24e-10 roc_auc binary    0.762    5 0.00758 Preproc~
## 2    74   163    32        8  0.00349    8.06e- 8 roc_auc binary    0.758    5 0.00369 Preproc~
## 3    39   256     5        7  0.000260    1.72e- 8 roc_auc binary    0.753    5 0.00506 Preproc~
## 4    11   270    28        1  0.0296    1.11e- 3 roc_auc binary    0.752    5 0.00714 Preproc~
## 5    23   226    35       13  0.0000167    2.29e- 7 roc_auc binary    0.744    5 0.00413 Preproc~
```

```
best_xgboost <- xgboost_tune %>%
  select_best("roc_auc")

autoplot(xgboost_tune, metric = "roc_auc")
```

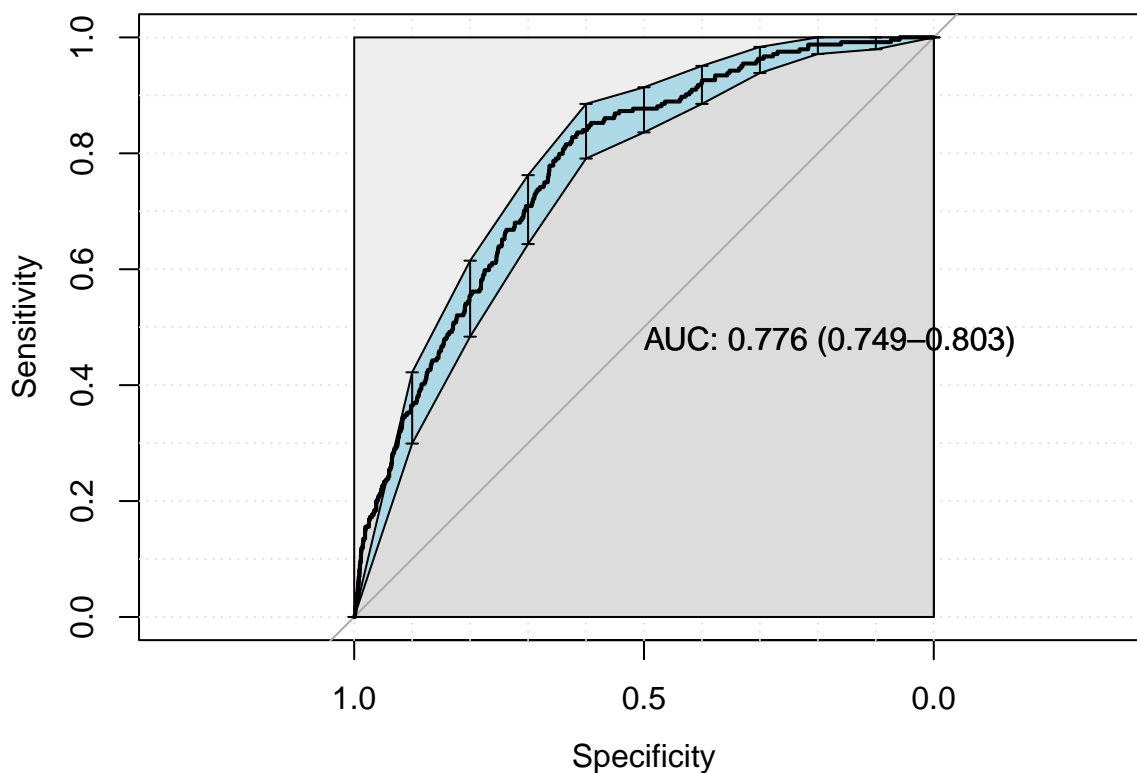


```
final_xgboost_workflow <-
  xgboost_workflow %>%
  finalize_workflow(best_xgboost)

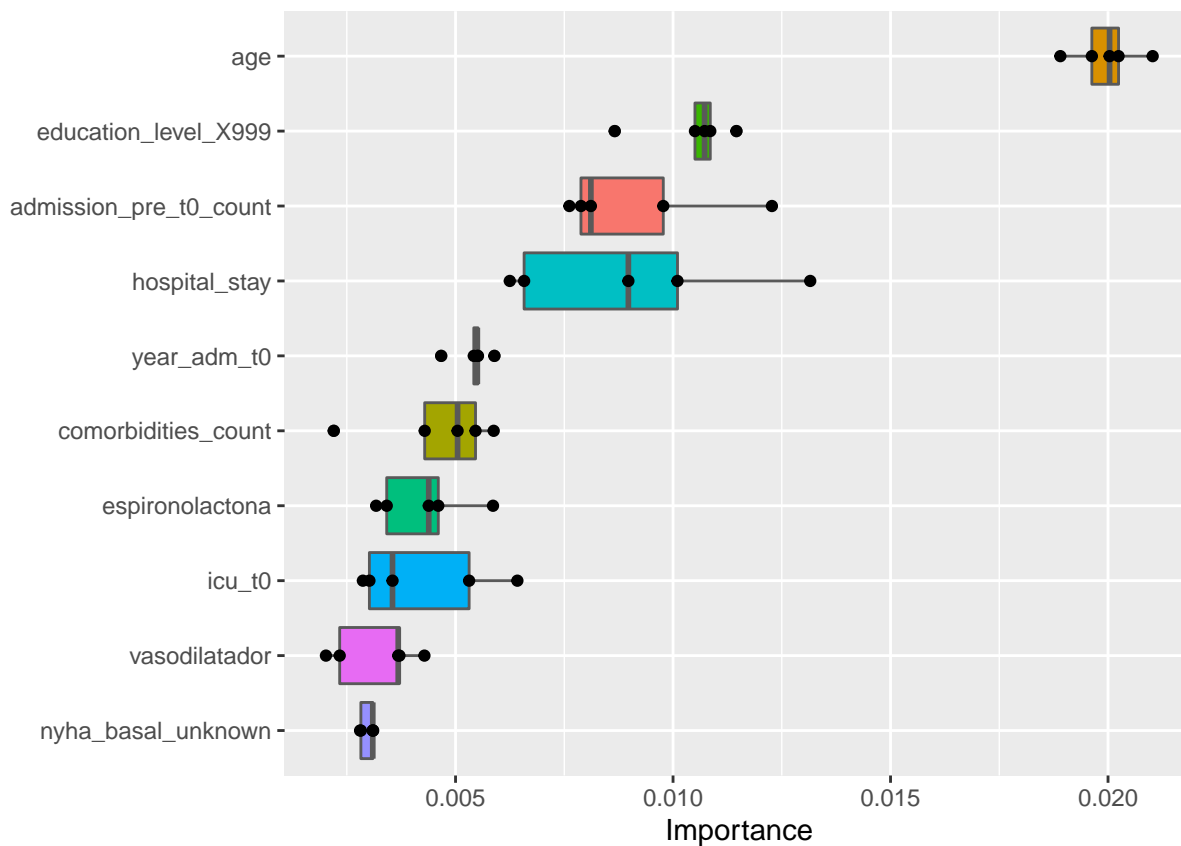
last_xgboost_fit <-
  final_xgboost_workflow %>%
  last_fit(df_split)

final_xgboost_fit <- extract_workflow(last_xgboost_fit)

xgboost_auc <- validation(final_xgboost_fit, df_test)
```



```
## [1] "Optimal Threshold: 0.06"
## Confusion Matrix and Statistics
##
##      reference
## data    0    1
##    0 2799   42
##    1 1687  202
##
##              Accuracy : 0.6345
##              95% CI   : (0.6206, 0.6482)
##    No Information Rate : 0.9484
##    P-Value [Acc > NIR] : 1
##
##              Kappa   : 0.1079
##
## Mcnemar's Test P-Value : <2e-16
##
##              Sensitivity : 0.6239
##              Specificity : 0.8279
##              Pos Pred Value : 0.9852
##              Neg Pred Value : 0.1069
##              Prevalence : 0.9484
##              Detection Rate : 0.5918
##              Detection Prevalence : 0.6006
##              Balanced Accuracy : 0.7259
##
##              'Positive' Class : 0
##
extract_vip(final_xgboost_fit, pred_wrapper = predict,
            reference_class = "0")
```



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

saveRDS(
  xgboost_parameters,
  file = sprintf(
    "./auxiliar/model_selection/hyperparameters/xgboost_%s.rds",
    outcome_column
  )
)
```

Minutes to run: 2.209

Boosted Tree (LightGBM)

```
lightgbm_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  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_spec <- boost_tree(
  mtry = tune(),
  trees = tune(),
  min_n = tune(),
  tree_depth = tune(),
  learn_rate = tune(),
  loss_reduction = tune(),
  sample_size = 1
) %>%
```

```

set_engine("lightgbm",
           nthread = 8) %>%
set_mode("classification")

lightgbm_grid <- grid_latin_hypercube(
  finalize(mtry(), df_train),
  trees(range = c(100L, 300L)),
  min_n(),
  tree_depth(),
  learn_rate(),
  loss_reduction(),
  size = grid_size
)

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

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

lightgbm_tune %>%
  show_best("roc_auc")

```

```

## # A tibble: 5 x 12
##   mtry trees min_n tree_depth  learn_rate loss_reduction .metric .estima~1 mean    n std_err .config
##   <int> <int> <int>    <int>      <dbl>      <dbl> <chr>    <chr>    <dbl> <int>    <dbl> <chr>
## 1    48   298   38         7 0.00000347      2.63e-9 roc_auc binary    0.772     5 0.00551 Prepro~
## 2    63   257   22        13 0.000281       1.38e-4 roc_auc binary    0.769     5 0.00432 Prepro~
## 3    10   147   33         5 0.0106         5.06e-6 roc_auc binary    0.769     5 0.00867 Prepro~
## 4    24   160    7        12 0.0729         1.12e+1 roc_auc binary    0.769     5 0.00664 Prepro~
## 5    68   279   30        10 0.0000000117    5.75e-2 roc_auc binary    0.768     5 0.00441 Prepro~
## # ... with abbreviated variable name 1: .estimator

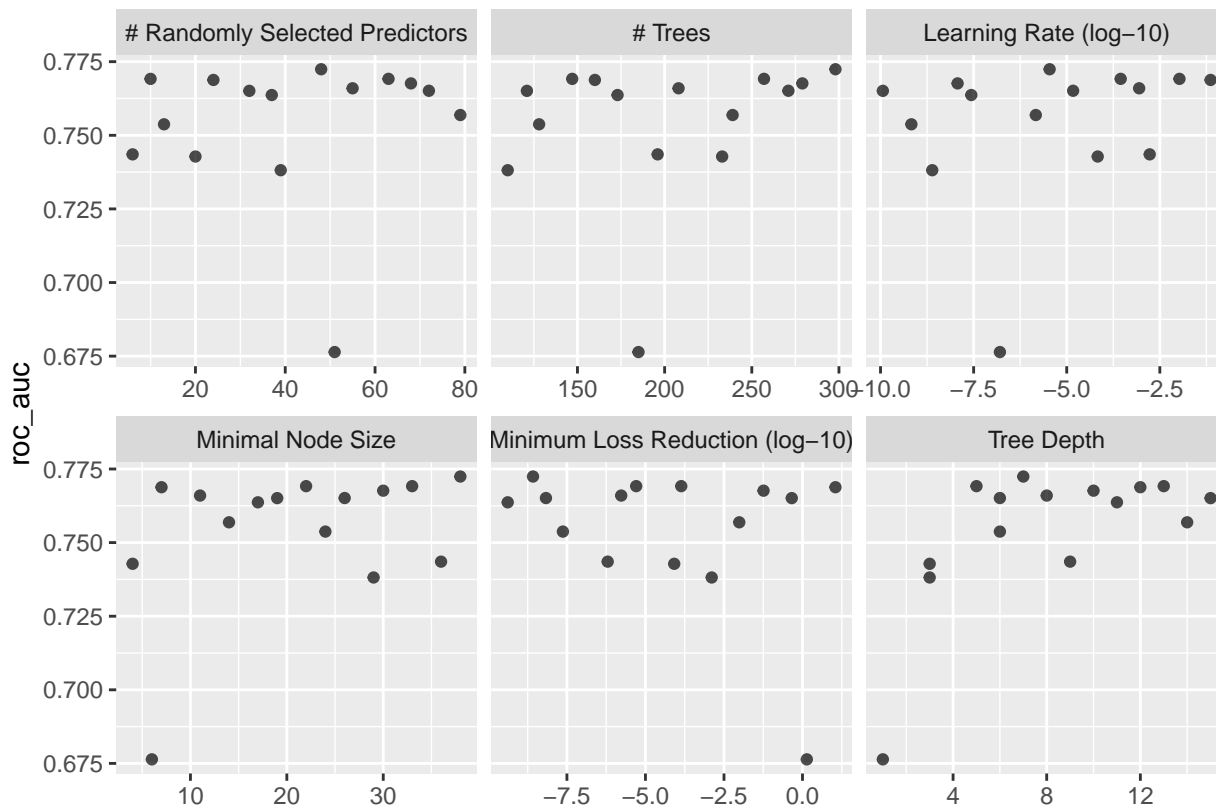
```

```

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

autoplot(lightgbm_tune, metric = "roc_auc")

```

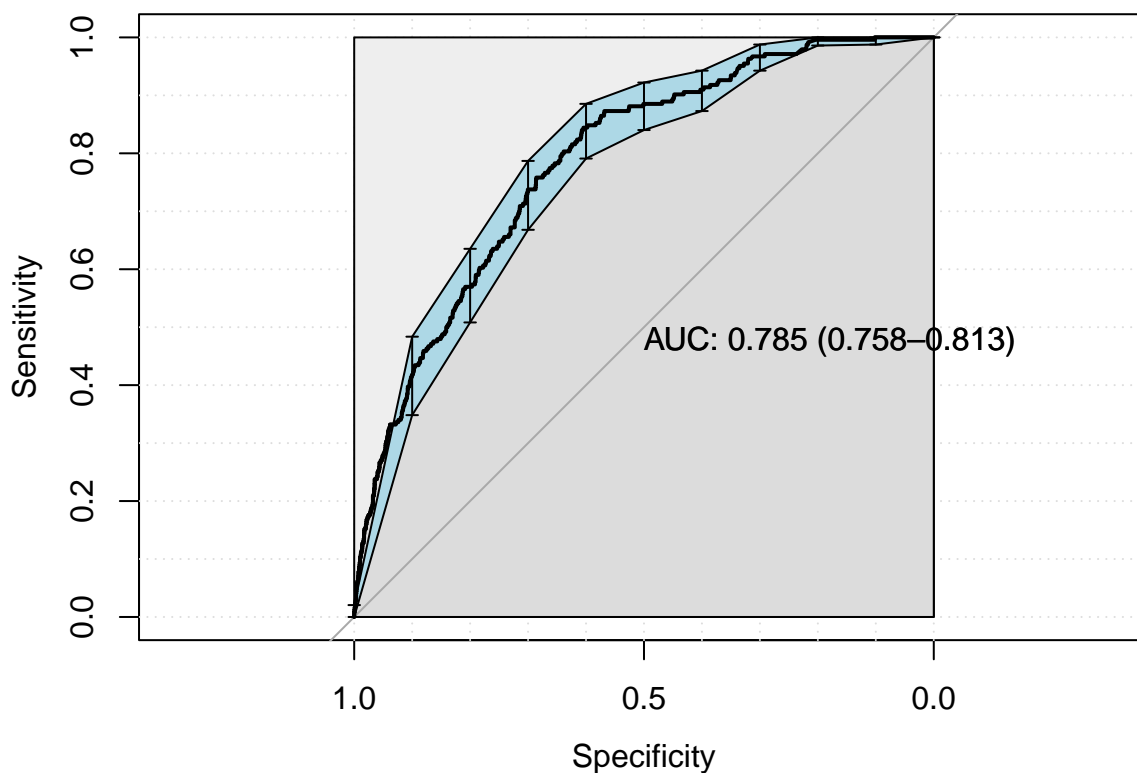



```
final_lightgbm_workflow <-
  lightgbm_workflow %>%
  finalize_workflow(best_lightgbm)

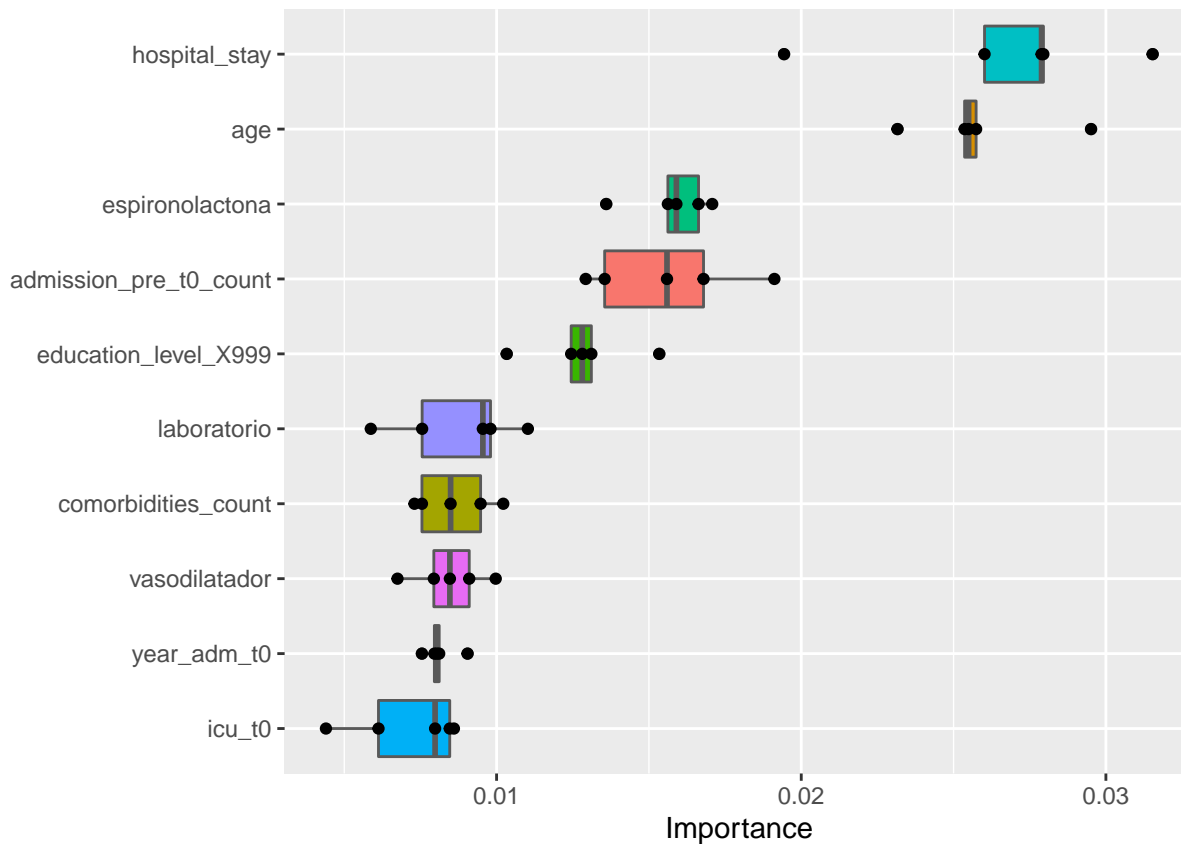
last_lightgbm_fit <-
  final_lightgbm_workflow %>%
  last_fit(df_split)

final_lightgbm_fit <- extract_workflow(last_lightgbm_fit)

lightgbm_auc <- validation(final_lightgbm_fit, df_test)
```



```
## [1] "Optimal Threshold: 0.05"
## Confusion Matrix and Statistics
##
##      reference
## data    0    1
## 0  2712   38
## 1  1774  206
##
##              Accuracy : 0.6169
##              95% CI   : (0.6029, 0.6308)
##      No Information Rate : 0.9484
##      P-Value [Acc > NIR] : 1
##
##              Kappa : 0.1028
##
##  McNemar's Test P-Value : <2e-16
##
##              Sensitivity : 0.6045
##              Specificity : 0.8443
##      Pos Pred Value : 0.9862
##      Neg Pred Value : 0.1040
##      Prevalence : 0.9484
##      Detection Rate : 0.5734
##      Detection Prevalence : 0.5814
##      Balanced Accuracy : 0.7244
##
##      'Positive' Class : 0
##
pfun_lightgbm <- function(object, newdata) predict(object, data = newdata)
extract_vip(final_lightgbm_fit, pred_wrapper = pfun_lightgbm,
            reference_class = "1")
```



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

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

Minutes to run: 2.925

GLM

```
glmnet_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  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_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())

glmnet_spec <-
  logistic_reg(penalty = tune(), mixture = tune()) %>%
  set_mode("classification") %>%
  set_engine("glmnet")

glmnet_grid <- grid_latin_hypercube(penalty(),
  mixture(),
```

```
size = grid_size)
```

```
glmnet_workflow <-  
  workflow() %>%  
  add_recipe(glmnet_recipe) %>%  
  add_model(glmnet_spec)
```

```
glmnet_tune <-  
  glmnet_workflow %>%  
  tune_grid(resamples = df_folds,  
            grid = glmnet_grid)
```

```
glmnet_tune %>%  
  collect_metrics()
```

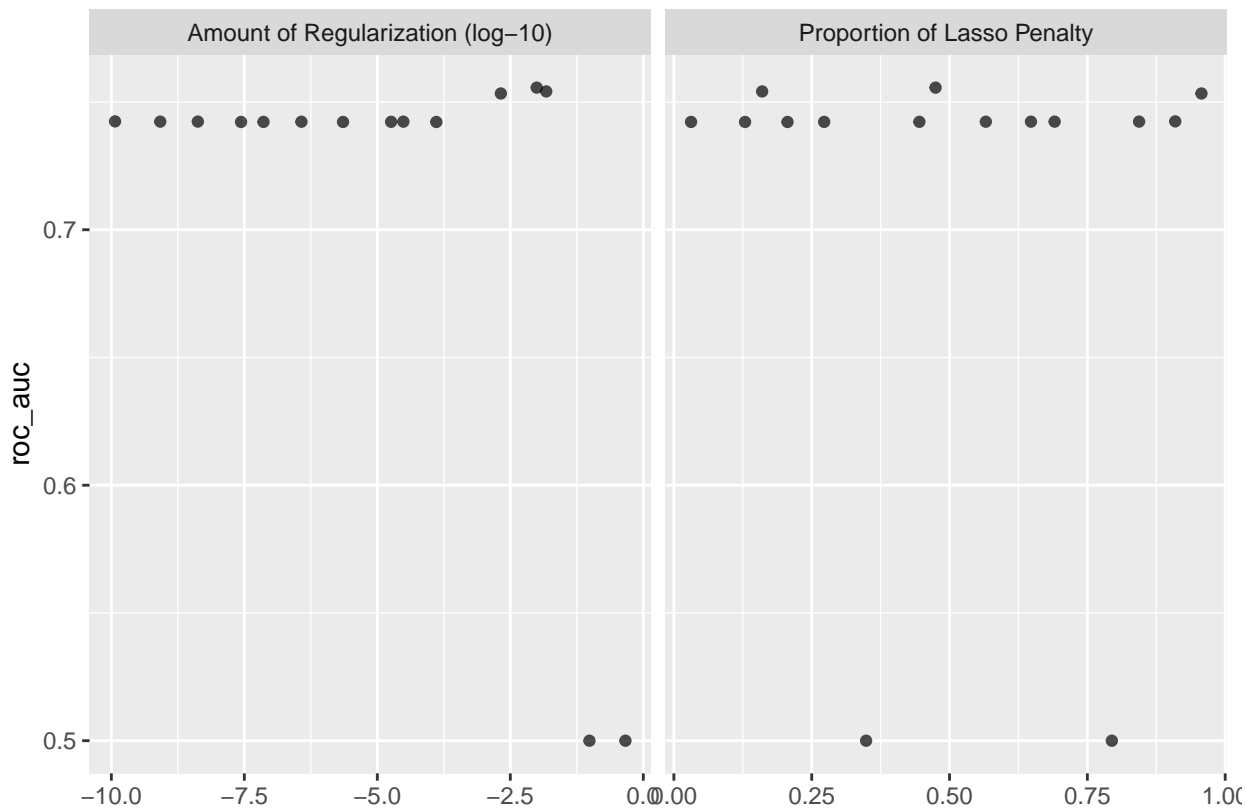
```
## # A tibble: 30 x 8
```

##		penalty	mixture	.metric	.estimator	mean	n	std_err	.config
##		<dbl>	<dbl>	<chr>	<chr>	<dbl>	<int>	<dbl>	<chr>
##	1	0.00000227	0.0312	accuracy	binary	0.953	5	0.00108	Preprocessor1_Model01
##	2	0.00000227	0.0312	roc_auc	binary	0.742	5	0.0175	Preprocessor1_Model01
##	3	0.000129	0.129	accuracy	binary	0.953	5	0.00108	Preprocessor1_Model02
##	4	0.000129	0.129	roc_auc	binary	0.742	5	0.0175	Preprocessor1_Model02
##	5	0.0150	0.160	accuracy	binary	0.954	5	0.00111	Preprocessor1_Model03
##	6	0.0150	0.160	roc_auc	binary	0.754	5	0.00996	Preprocessor1_Model03
##	7	0.0000000274	0.206	accuracy	binary	0.953	5	0.00108	Preprocessor1_Model04
##	8	0.0000000274	0.206	roc_auc	binary	0.742	5	0.0174	Preprocessor1_Model04
##	9	0.0000000725	0.273	accuracy	binary	0.953	5	0.00108	Preprocessor1_Model05
##	10	0.0000000725	0.273	roc_auc	binary	0.742	5	0.0174	Preprocessor1_Model05

```
## # ... with 20 more rows
```

```
## # i Use 'print(n = ...)' to see more rows
```

```
autoplot(glmnet_tune, metric = "roc_auc")
```



```

glmnet_tune %>%
  show_best("roc_auc")

## # A tibble: 5 x 8
##   penalty mixture .metric .estimator mean      n std_err .config
##   <dbl>   <dbl> <chr>   <chr>   <dbl> <int>   <dbl> <chr>
## 1 9.92e- 3  0.475 roc_auc binary   0.756     5 0.00669 Preprocessor1_Model08
## 2 1.50e- 2  0.160 roc_auc binary   0.754     5 0.00996 Preprocessor1_Model03
## 3 2.10e- 3  0.957 roc_auc binary   0.753     5 0.0112  Preprocessor1_Model15
## 4 1.17e-10  0.909 roc_auc binary   0.742     5 0.0174  Preprocessor1_Model14
## 5 4.23e- 9  0.844 roc_auc binary   0.742     5 0.0174  Preprocessor1_Model13

best_glmnet <- glmnet_tune %>%
  select_best("roc_auc")

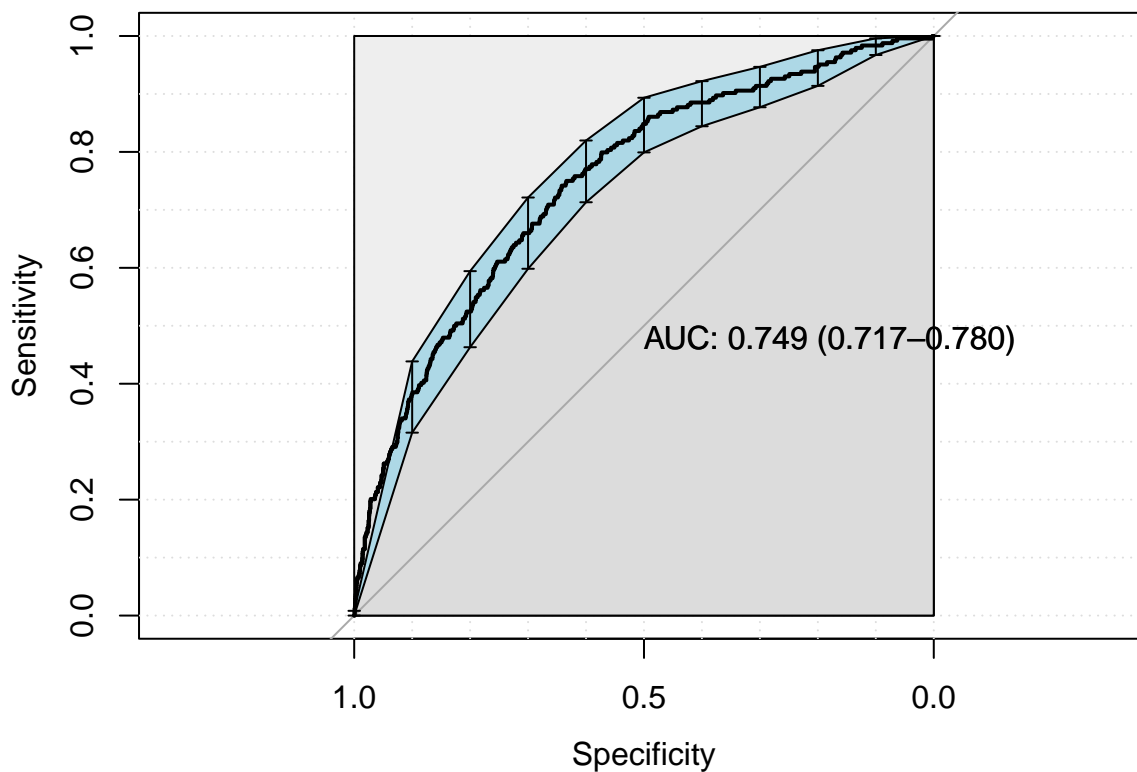
final_glmnet_workflow <-
  glmnet_workflow %>%
  finalize_workflow(best_glmnet)

last_glmnet_fit <-
  final_glmnet_workflow %>%
  last_fit(df_split)

final_glmnet_fit <- extract_workflow(last_glmnet_fit)

glmnet_auc <- validation(final_glmnet_fit, df_test)

```



```

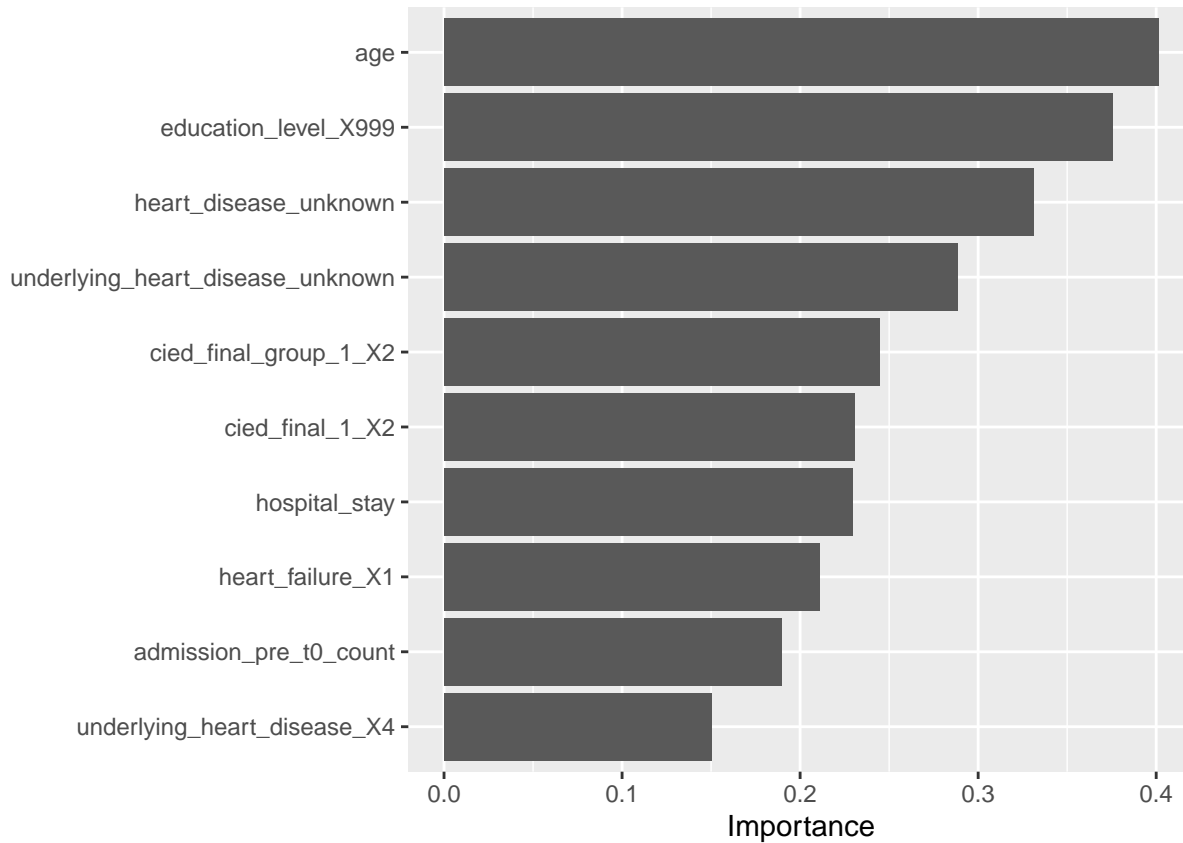
## [1] "Optimal Threshold: 0.04"
## Confusion Matrix and Statistics
##
##      reference
## data    0    1
## 0 2844   61
## 1 1642  183

```

```
##
##           Accuracy : 0.64
##           95% CI : (0.6261, 0.6537)
##       No Information Rate : 0.9484
##       P-Value [Acc > NIR] : 1
##
##           Kappa : 0.0945
##
##  McNemar's Test P-Value : <2e-16
##
##       Sensitivity : 0.6340
##       Specificity : 0.7500
##       Pos Pred Value : 0.9790
##       Neg Pred Value : 0.1003
##       Prevalence : 0.9484
##       Detection Rate : 0.6013
##       Detection Prevalence : 0.6142
##       Balanced Accuracy : 0.6920
##
##       'Positive' Class : 0
##
```

```
pfun_glmnet <- function(object, newdata) predict(object, newx = newdata)

extract_vip(final_glmnet_fit, pred_wrapper = pfun_glmnet,
             reference_class = "1", method = 'model')
```



1.266

Minutes to run:

Decision Tree

```
tree_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  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_zv(all_predictors())

tree_spec <-
  decision_tree(cost_complexity = tune(),
                tree_depth = tune(),
                min_n = tune()) %>%
  set_mode("classification") %>%
  set_engine("rpart")

tree_grid <- grid_latin_hypercube(cost_complexity(),
                                  tree_depth(),
                                  min_n(),
                                  size = grid_size)

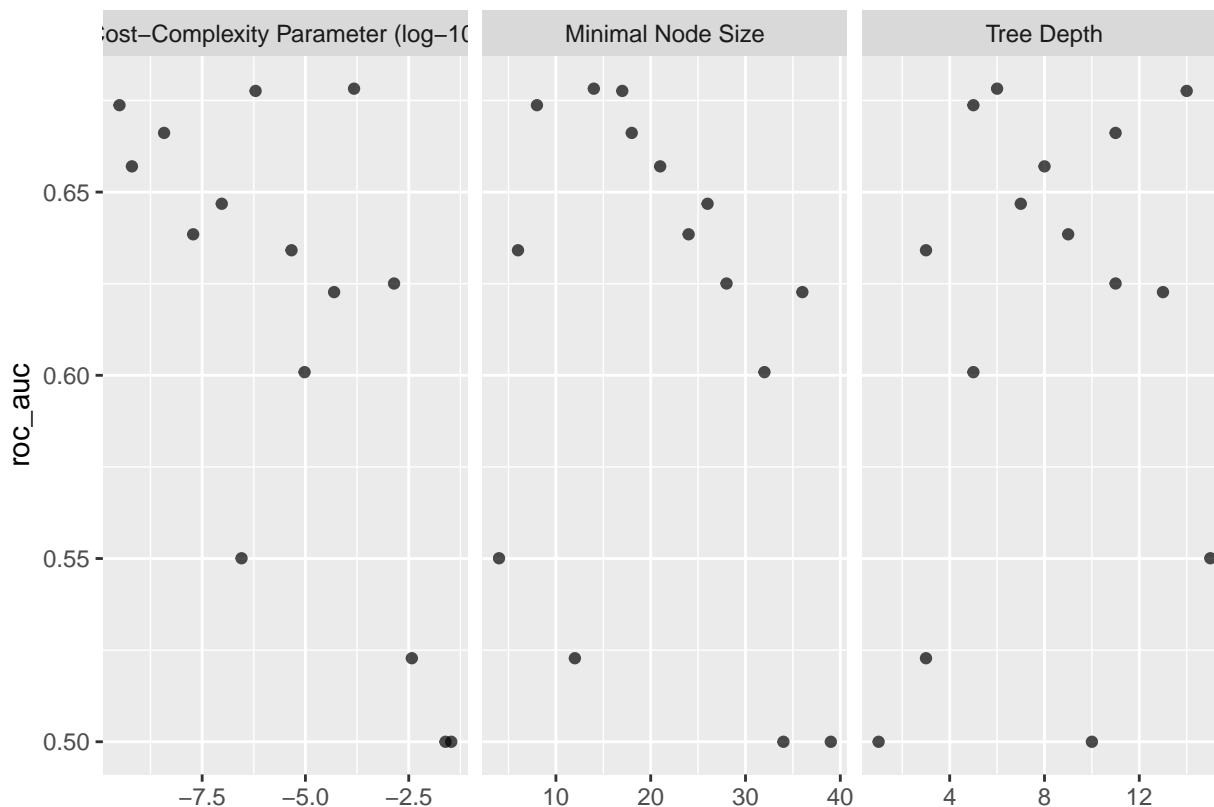
tree_workflow <-
  workflow() %>%
  add_recipe(tree_recipe) %>%
  add_model(tree_spec)

tree_tune <-
  tree_workflow %>%
  tune_grid(resamples = df_folds,
            grid = tree_grid)

tree_tune %>%
  collect_metrics()

autoplot(tree_tune, metric = "roc_auc")

```



```

tree_tune %>%
  show_best("roc_auc")

best_tree <- tree_tune %>%
  select_best("roc_auc")

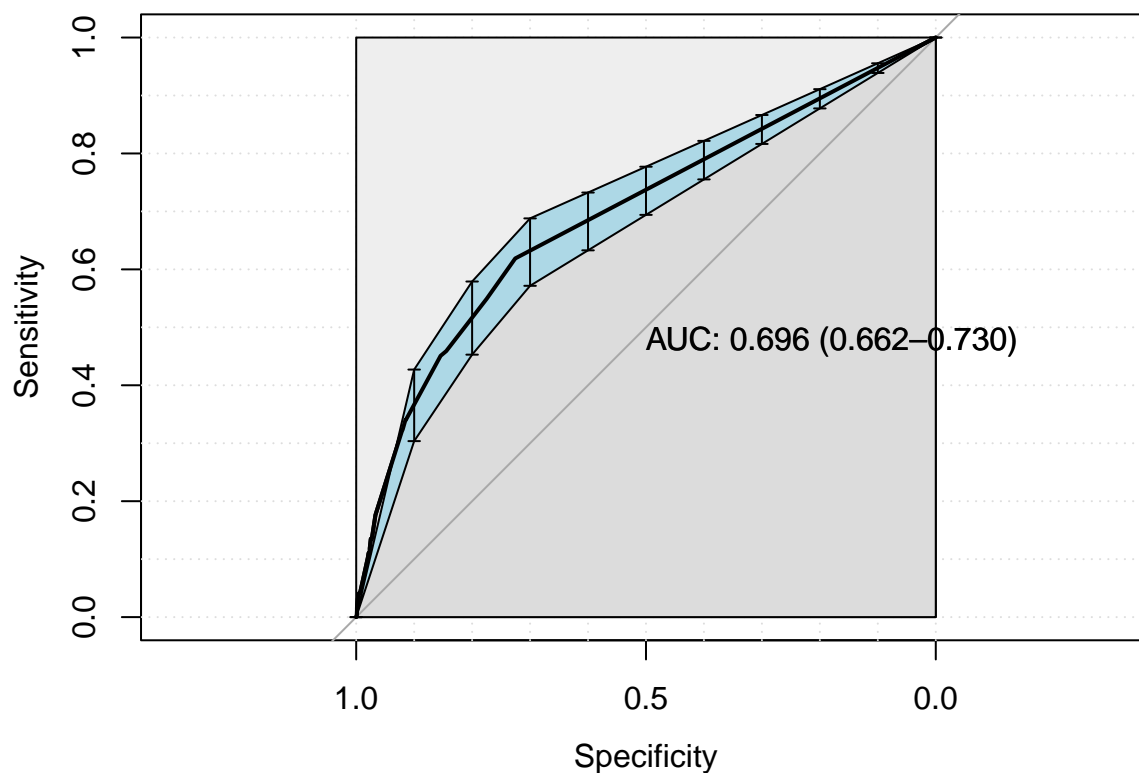
final_tree_workflow <-
  tree_workflow %>%
  finalize_workflow(best_tree)

last_tree_fit <-
  final_tree_workflow %>%
  last_fit(df_split)

final_tree_fit <- extract_workflow(last_tree_fit)

tree_auc <- validation(final_tree_fit, df_test)

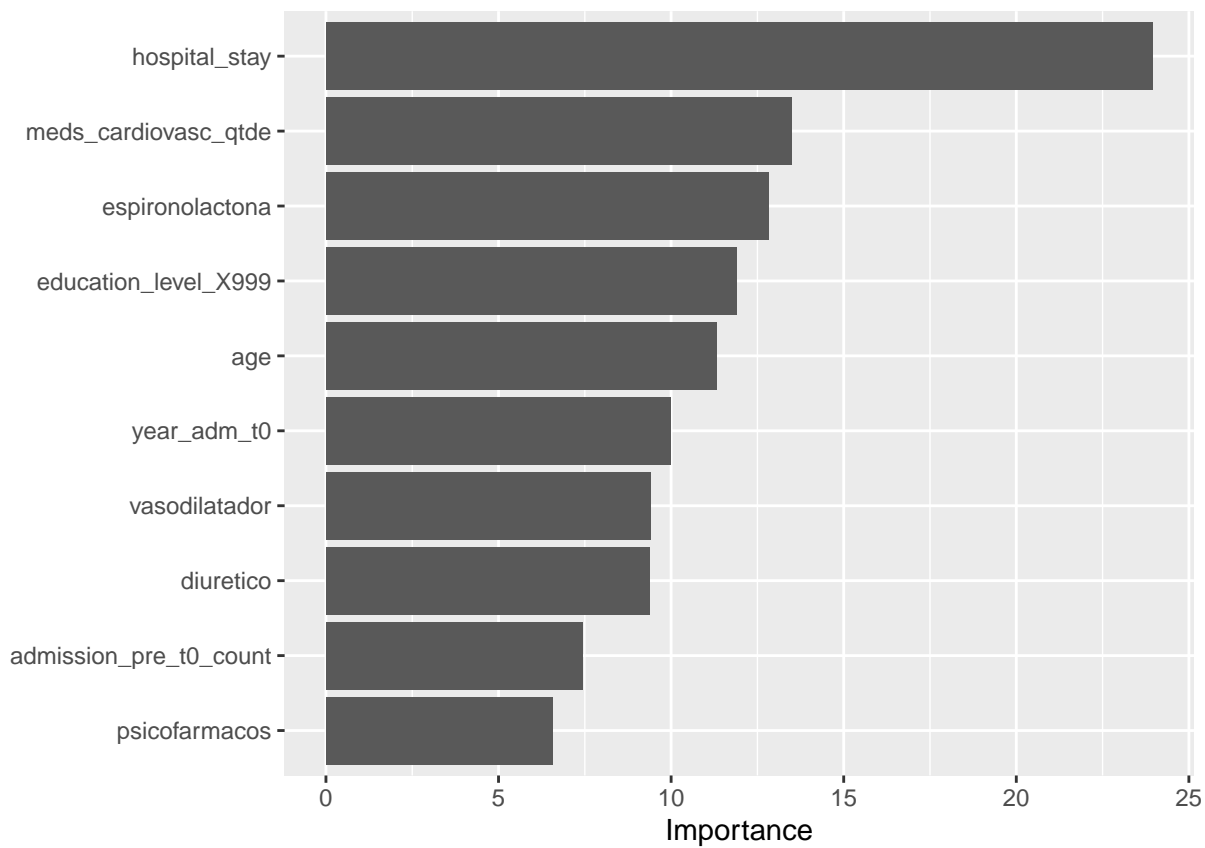
```



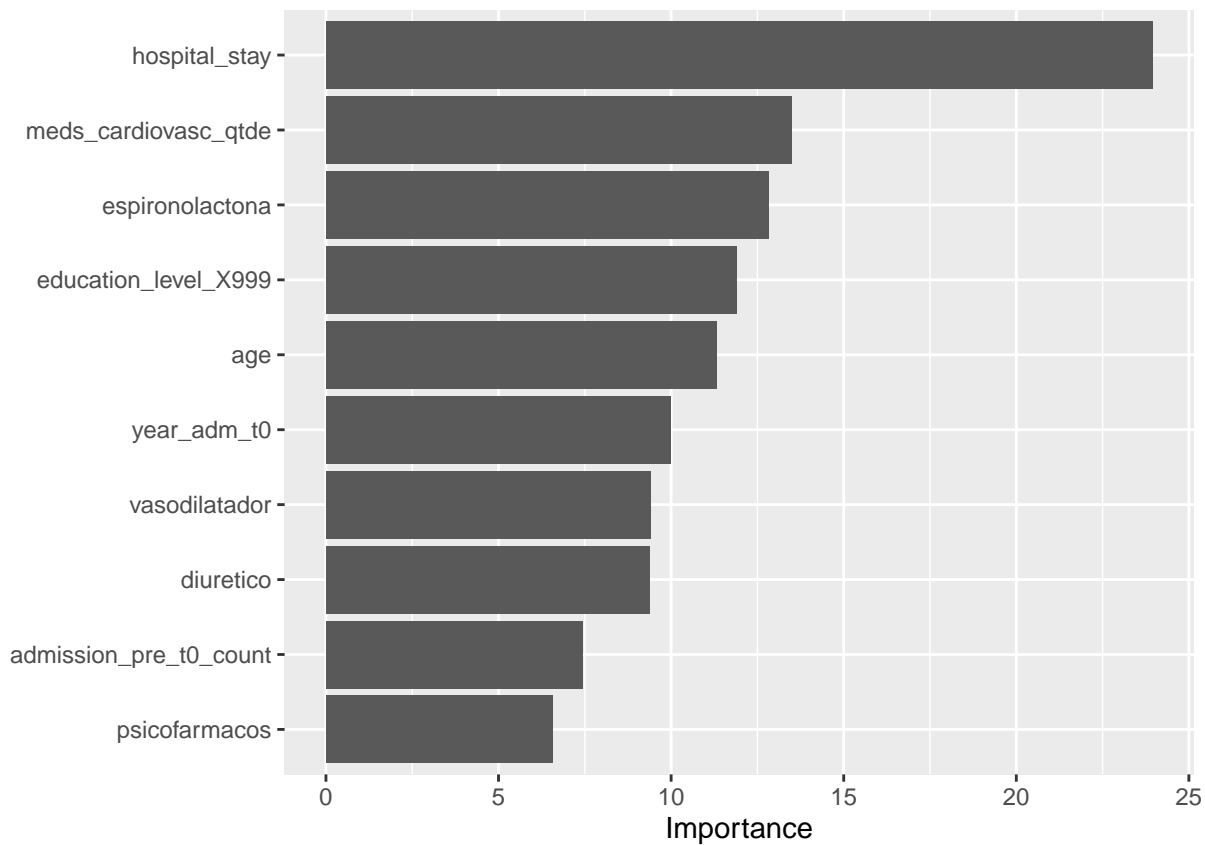
```

if (tree_auc$auc > 0.55) {
  final_tree_fit %>%
    extract_fit_parsnip() %>%
    vip()
}

```

```
extract_vip(final_tree_fit, pred_wrapper = predict,
            reference_class = "0", use_matrix = FALSE,
            method = 'model')
```

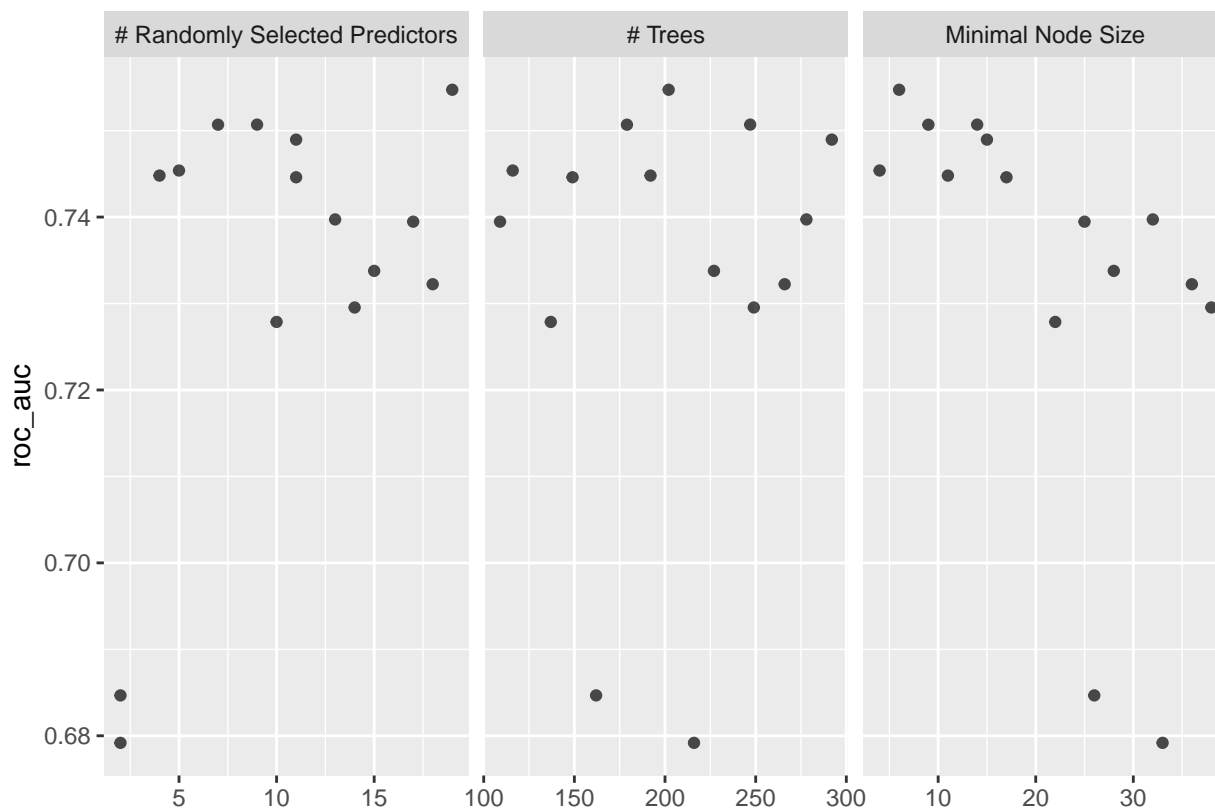


1.791

Minutes to run:

Random Forest

```
rf_recipe <-  
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula,  
    data = df_train) %>%  
  step_nominal(all_nominal_predictors()) %>%  
  step_unknown(all_nominal_predictors()) %>%  
  step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%  
  step_dummy(all_nominal_predictors()) %>%  
  step_zv(all_predictors()) %>%  
  step_impute_mean(all_numeric_predictors())  
  
rf_spec <-  
  rand_forest(mtry = tune(),  
    trees = tune(),  
    min_n = tune()) %>%  
  set_mode("classification") %>%  
  set_engine("randomForest",  
    probability = TRUE,  
    nthread = 8)  
  
rf_grid <- grid_latin_hypercube(mtry(range = c(1L, 20L)),  
  trees(range = c(100L, 300L)),  
  min_n(),  
  size = grid_size)  
  
rf_workflow <-  
  workflow() %>%  
  add_recipe(rf_recipe) %>%  
  add_model(rf_spec)  
  
rf_tune <-  
  rf_workflow %>%  
  tune_grid(resamples = df_folds,  
    grid = rf_grid)  
  
rf_tune %>%  
  collect_metrics()  
  
autoplot(rf_tune, metric = "roc_auc")
```



```
rf_tune %>%
  show_best("roc_auc")

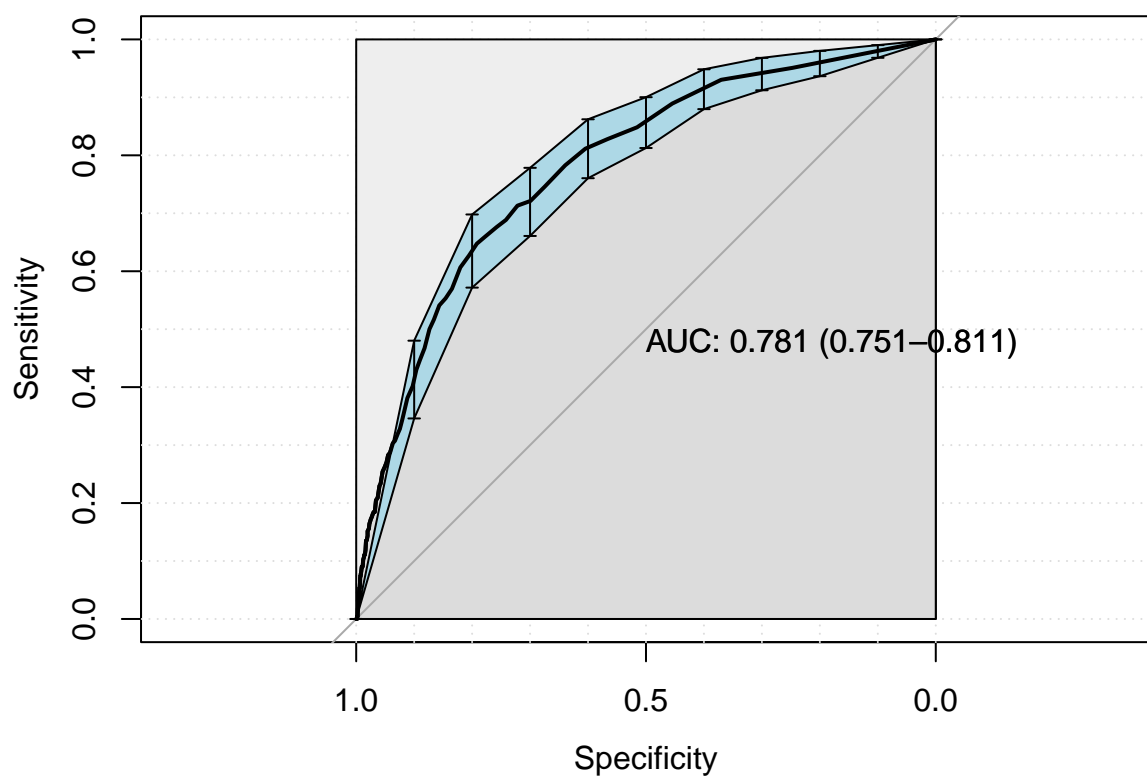
best_rf <- rf_tune %>%
  select_best("roc_auc")

final_rf_workflow <-
  rf_workflow %>%
  finalize_workflow(best_rf)

last_rf_fit <-
  final_rf_workflow %>%
  last_fit(df_split)

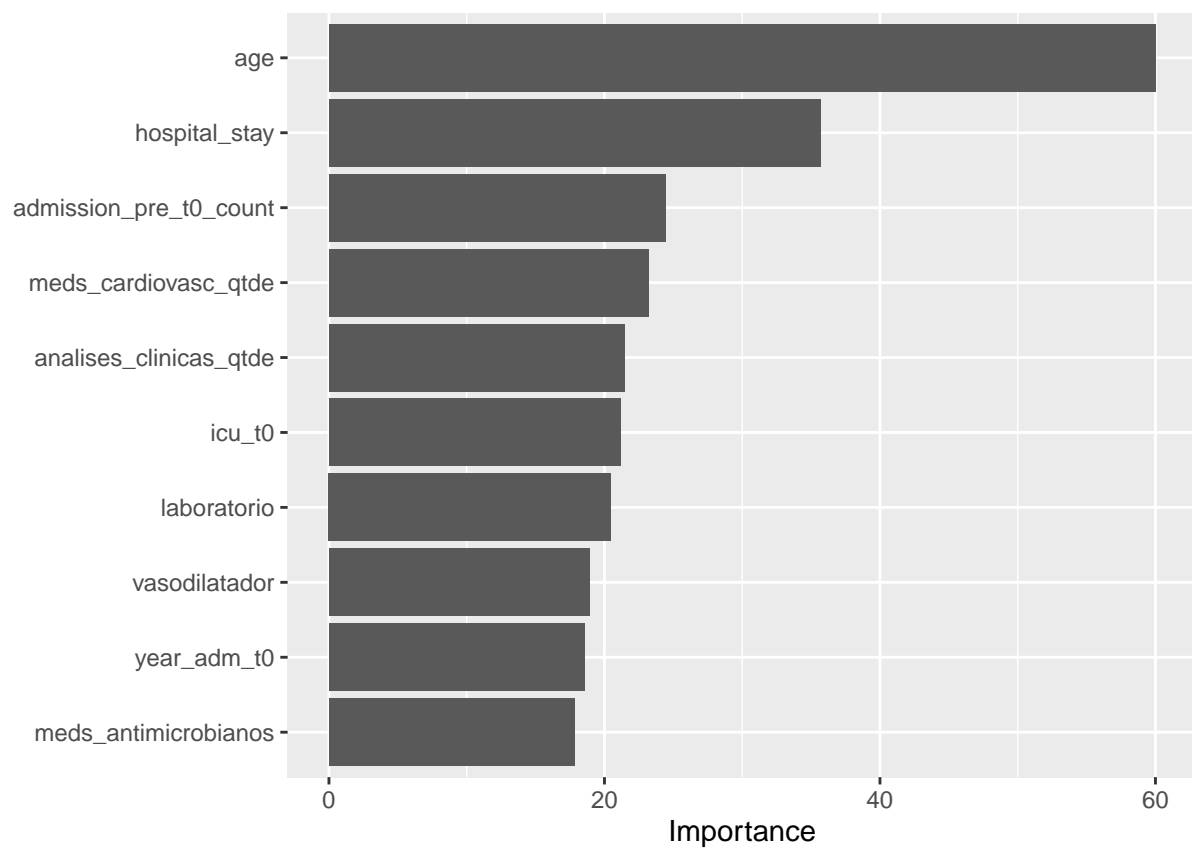
final_rf_fit <- extract_workflow(last_rf_fit)

rf_auc <- validation(final_rf_fit, df_test)
```



```
pfun_rf <- function(object, newdata) predict(object, data = newdata)

extract_vip(final_rf_fit, pred_wrapper = predict,
            reference_class = "1", use_matrix = FALSE,
            method = 'model')
```



KNN

```
# knn_recipe <-  
#   recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%  
#   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_zv(all_predictors()) %>%  
#   step_impute_mean(all_numeric_predictors())  
#  
# knn_spec <-  
#   nearest_neighbor(neighbors = tune(),  
#                     weight_func = tune(),  
#                     dist_power = tune()) %>%  
#   set_mode("classification") %>%  
#   set_engine("kknn")  
#  
# knn_grid <- grid_latin_hypercube(neighbors(),  
#                                   weight_func(),  
#                                   dist_power(),  
#                                   size = grid_size)  
#  
# knn_workflow <-  
#   workflow() %>%  
#   add_recipe(knn_recipe) %>%  
#   add_model(knn_spec)  
#  
# knn_tune <-  
#   knn_workflow %>%  
#   tune_grid(resamples = df_folds,  
#             grid = knn_grid)  
#  
# knn_tune %>%  
#   collect_metrics()  
#  
# autoplot(knn_tune, metric = "roc_auc")  
#  
# knn_tune %>%  
#   show_best("roc_auc")  
#  
# best_knn <- knn_tune %>%  
#   select_best("roc_auc")  
#  
# final_knn_workflow <-  
#   knn_workflow %>%  
#   finalize_workflow(best_knn)  
#  
# last_knn_fit <-  
#   final_knn_workflow %>%  
#   last_fit(df_split)  
#  
# final_knn_fit <- extract_workflow(last_knn_fit)  
#  
# knn_auc = validation(final_knn_fit, df_test)
```

Minutes to run: 0

SVM

```
# sum_recipe <-
#   recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
#   step_nominal(all_nominal_predictors()) %>%
#   step_unknown(all_nominal_predictors()) %>%
#   step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%
#   step_dummy(all_nominal_predictors()) %>%
#   step_zv(all_predictors()) %>%
#   step_impute_mean(all_numeric_predictors())
#
# sum_spec <-
#   sum_rbf(cost = tune(), rbf_sigma = tune()) %>%
#   set_mode("classification") %>%
#   set_engine("kernlab")
#
# sum_grid <- grid_latin_hypercube(cost(),
#                                   rbf_sigma(),
#                                   size = grid_size)
#
# sum_workflow <-
#   workflow() %>%
#   add_recipe(sum_recipe) %>%
#   add_model(sum_spec)
#
# sum_tune <-
#   sum_workflow %>%
#   tune_grid(resamples = df_folds,
#             grid = grid_size)
#
# sum_tune %>%
#   collect_metrics()
#
# autoplot(sum_tune, metric = "roc_auc")
#
# sum_tune %>%
#   show_best("roc_auc")
#
# best_sum <- sum_tune %>%
#   select_best("roc_auc")
#
# final_sum_workflow <-
#   sum_workflow %>%
#   finalize_workflow(best_sum)
#
# last_sum_fit <-
#   final_sum_workflow %>%
#   last_fit(df_split)
#
# final_sum_fit <- extract_workflow(last_sum_fit)
#
# sum_auc = validation(final_sum_fit, df_test)
```

Minutes to run: 0

Models Comparison

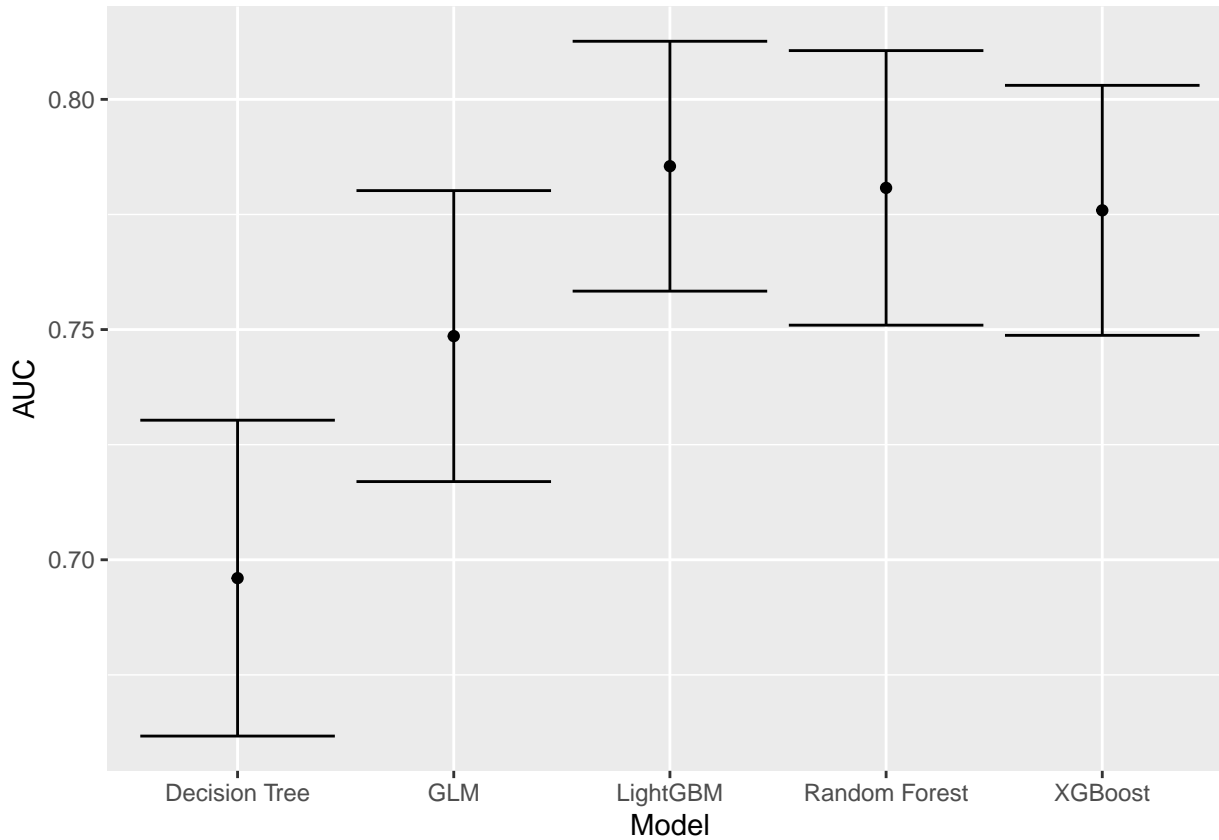
```
df_auc <- tibble::tribble(
  ~Model, ~`AUC`, ~`Lower Limit`, ~`Upper Limit`,
  'XGBoost', as.numeric(xgboost_auc$auc), xgboost_auc$ci[1], xgboost_auc$ci[3],
```

```

'LightGBM', as.numeric(lightgbm_auc$auc), lightgbm_auc$ci[1], lightgbm_auc$ci[3],
'GLM', as.numeric(glmnet_auc$auc), glmnet_auc$ci[1], glmnet_auc$ci[3],
'Decision Tree', as.numeric(tree_auc$auc), tree_auc$ci[1], tree_auc$ci[3],
'Random Forest', as.numeric(rf_auc$auc), rf_auc$ci[1], rf_auc$ci[3]
) %>%
  mutate(Target = outcome_column)

df_auc %>%
  ggplot(aes(x = Model, y = AUC, ymin = `Lower Limit`, ymax = `Upper Limit`)) +
    geom_point() +
    geom_errorbar()

```



```

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

```

Minutes to run: 0.002