

Model Selection - readmission_60d

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

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

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

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

```
## 32. diuretico
## 33. vasodilatador
## 34. insuf_cardiaca
## 35. espironolactona
## 36. bloq_calcio
## 37. antiplaquetario_ev
## 38. insulina
## 39. anticonvulsivante
## 40. psicofarmacos
## 41. antifungico
## 42. antiviral
## 43. classe_meds_qtde
## 44. meds_cardiovasc_qtde
## 45. meds_antimicrobianos
## 46. ventilacao_mecanica
## 47. cec
## 48. transplante_cardiaco
## 49. cir_toracica
## 50. outros_proced_cirurgicos
## 51. icp
## 52. angioplastia
## 53. cateterismo
## 54. eletrofisiologia
## 55. cateter_venoso_central
## 56. proced_invasivos_qtde
## 57. cve_desf
## 58. transfusao
## 59. interconsulta
## 60. equipe_multiprof
## 61. holter
## 62. teste_esforco
## 63. espiro_ergoespiro
## 64. tilt_teste
## 65. metodos_graficos_qtde
## 66. laboratorio
## 67. cultura
## 68. analises_clinicas_qtde
## 69. citologia
## 70. biopsia
## 71. histopatologia_qtde
## 72. angio_rm
## 73. angio_tc
## 74. arteriografia
## 75. cintilografia
## 76. ecocardiograma
## 77. endoscopia
## 78. pet_ct
## 79. ultrassom
## 80. tomografia
## 81. ressonancia
## 82. exames_imagem_qtde
## 83. bic
## 84. 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")

```

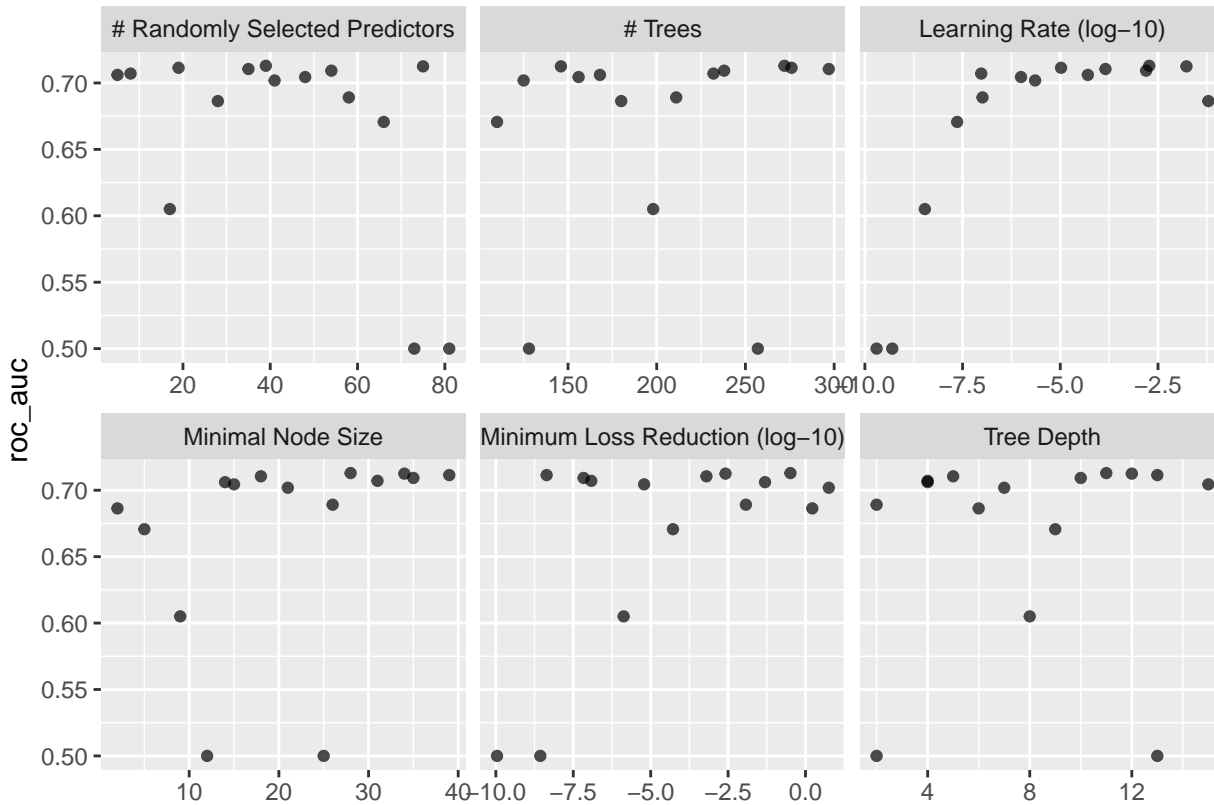
A tibble: 5 x 12

##	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	39	272	28	11	0.00189	0.326	roc_auc	binary	0.713	5	0.0183	Preproc~
##	2	75	146	34	12	0.0169	0.00262	roc_auc	binary	0.712	5	0.0183	Preproc~
##	3	19	276	39	13	0.0000105	0.00000000437	roc_auc	binary	0.711	5	0.0183	Preproc~
##	4	35	297	18	5	0.000143	0.000633	roc_auc	binary	0.710	5	0.0177	Preproc~
##	5	54	238	35	10	0.00157	0.0000000683	roc_auc	binary	0.709	5	0.0195	Preproc~

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

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

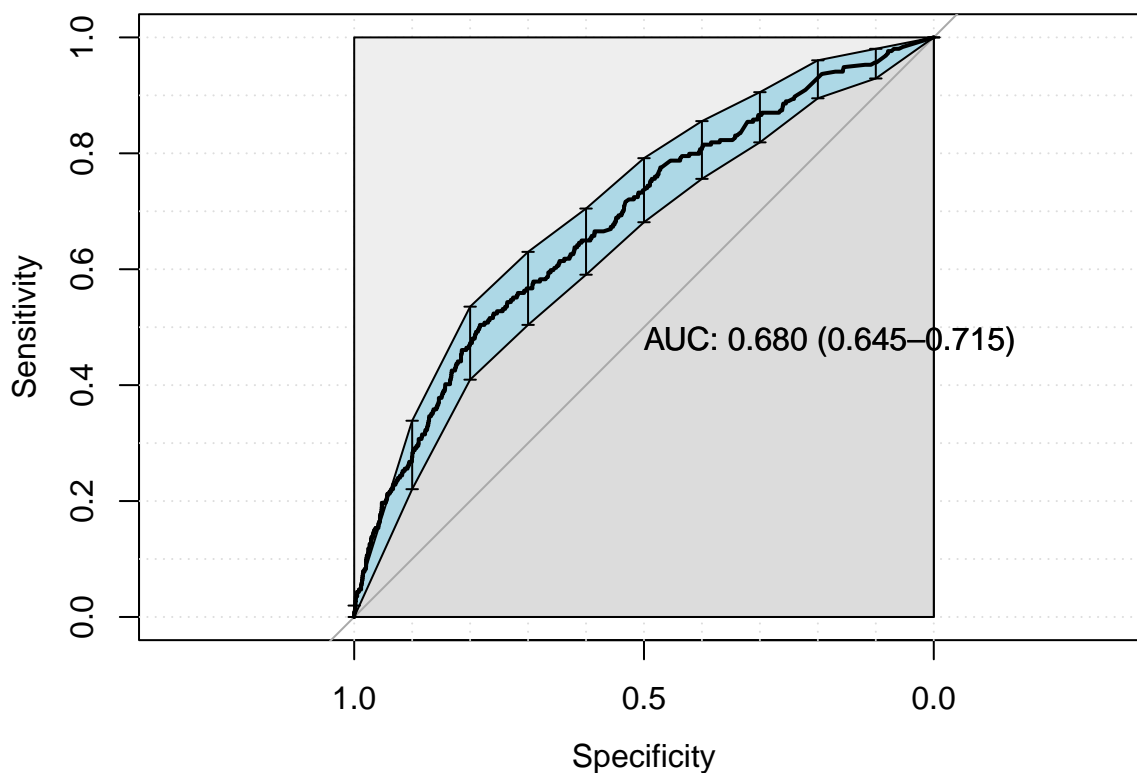


```
final_xgboost_workflow <-
  xgboost_workflow %>%
  finalize_workflow(best_xgboost)

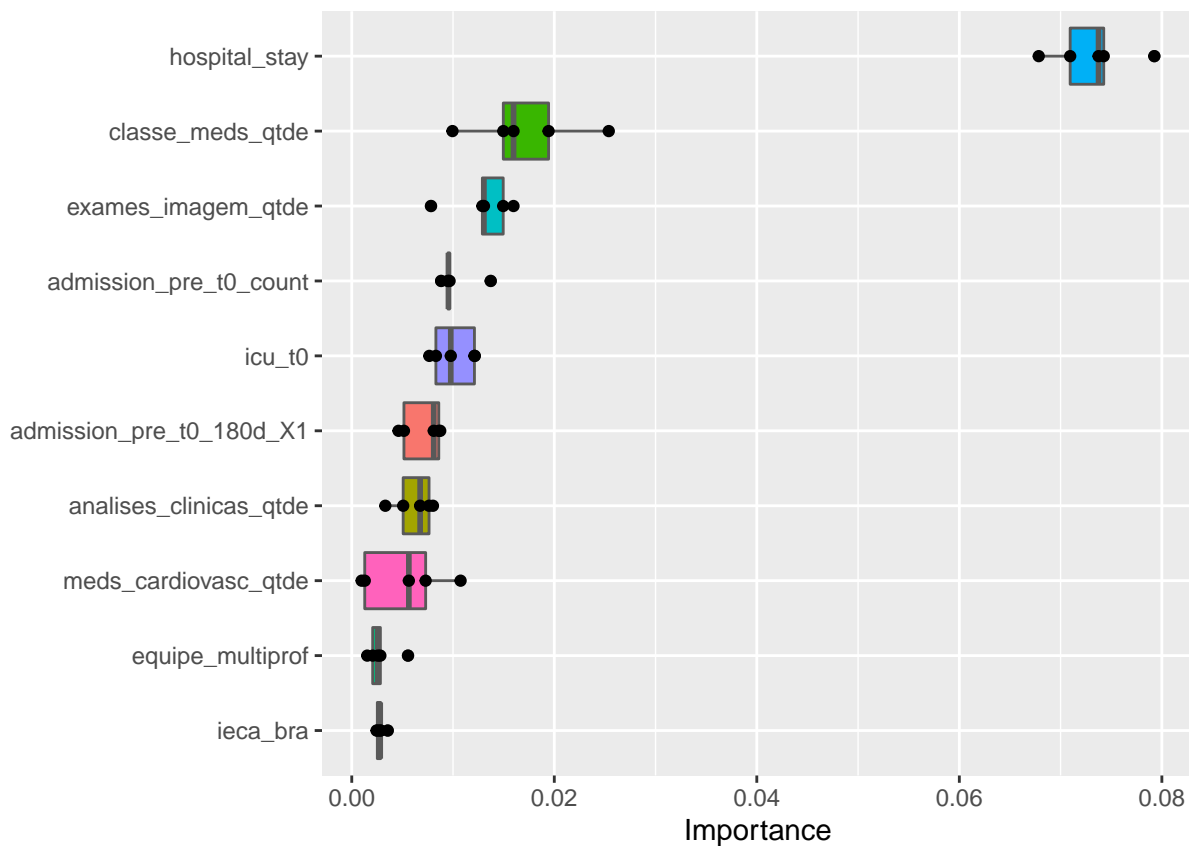
last_xgboost_fit <-
  final_xgboost_workflow %>%
  last_fit(df_split)

final_xgboost_fit <- extract_workflow(last_xgboost_fit)

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



```
## [1] "Optimal Threshold: 0.33"
## Confusion Matrix and Statistics
##
##      reference
## data    0    1
##    0 3505 126
##    1  971 128
##
##              Accuracy : 0.7681
##              95% CI   : (0.7558, 0.78)
##    No Information Rate : 0.9463
##    P-Value [Acc > NIR] : 1
##
##              Kappa   : 0.1117
##
## Mcnemar's Test P-Value : <2e-16
##
##              Sensitivity : 0.7831
##              Specificity : 0.5039
##              Pos Pred Value : 0.9653
##              Neg Pred Value : 0.1165
##              Prevalence : 0.9463
##              Detection Rate : 0.7410
##              Detection Prevalence : 0.7677
##              Balanced Accuracy : 0.6435
##
##              '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: 2.068

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)

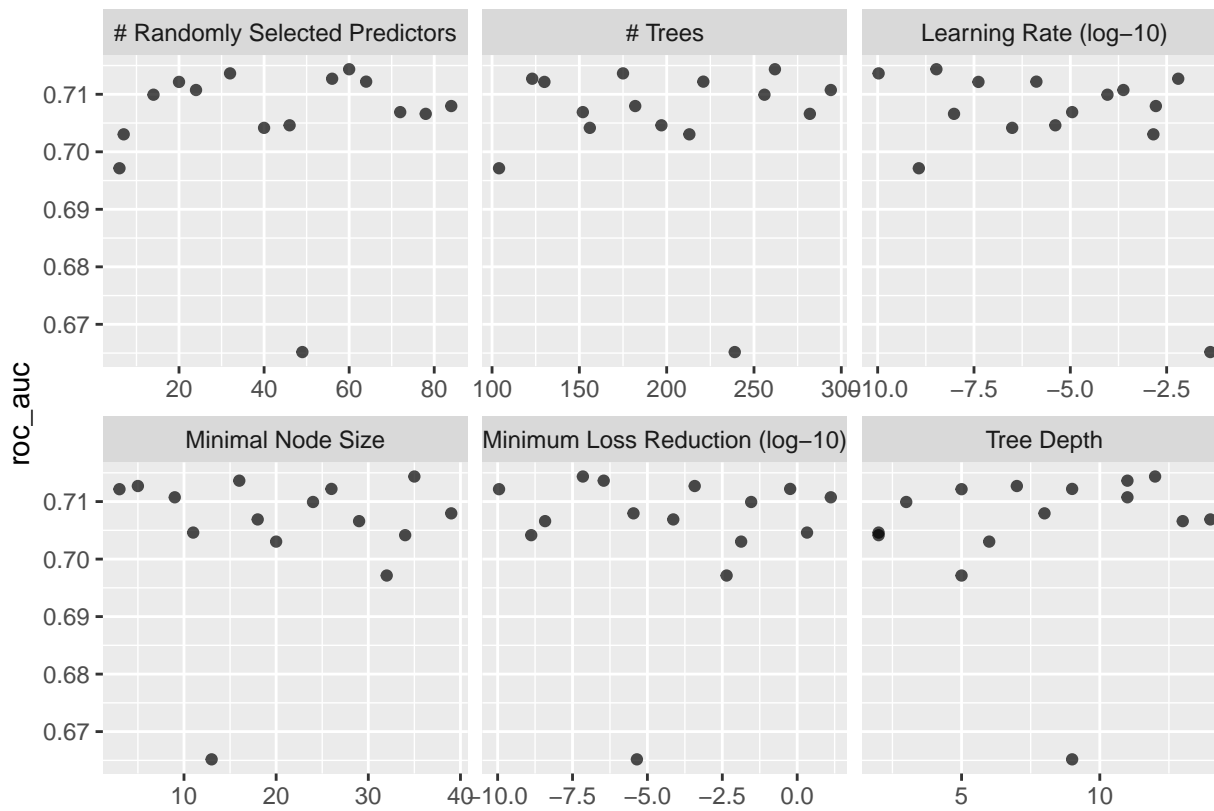
lightgbm_tune %>%
  show_best("roc_auc")

## # A tibble: 5 x 12
##   mtry trees min_n tree_depth learn_rate loss_reduction .metric .estimator mean n std_err .config
##   <int> <int> <int>    <int>    <dbl>         <dbl> <chr>    <chr>    <dbl> <int>  <dbl> <chr>
## 1     60   262    35      12  3.35e- 9    7.01e- 8 roc_auc binary  0.714     5  0.0177 Preproc~
## 2     32   175    16      11  1.06e-10    3.48e- 7 roc_auc binary  0.714     5  0.0180 Preproc~
## 3     56   123     5       7  6.32e- 3    3.80e- 4 roc_auc binary  0.713     5  0.0191 Preproc~
## 4     64   221    26       9  1.32e- 6    5.84e- 1 roc_auc binary  0.712     5  0.0180 Preproc~
## 5     20   130     3       5  4.12e- 8    1.11e-10 roc_auc binary  0.712     5  0.0170 Preproc~

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

autoplot(lightgbm_tune, metric = "roc_auc")

```

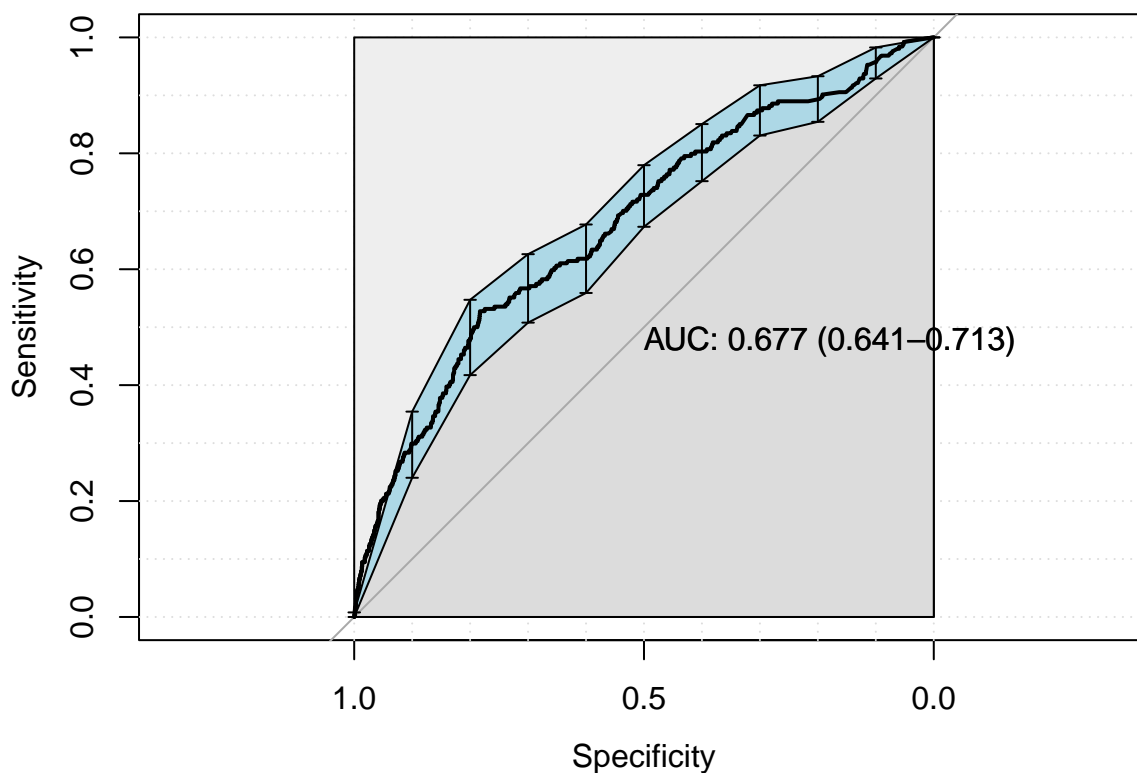



```
final_lightgbm_workflow <-
  lightgbm_workflow %>%
  finalize_workflow(best_lightgbm)

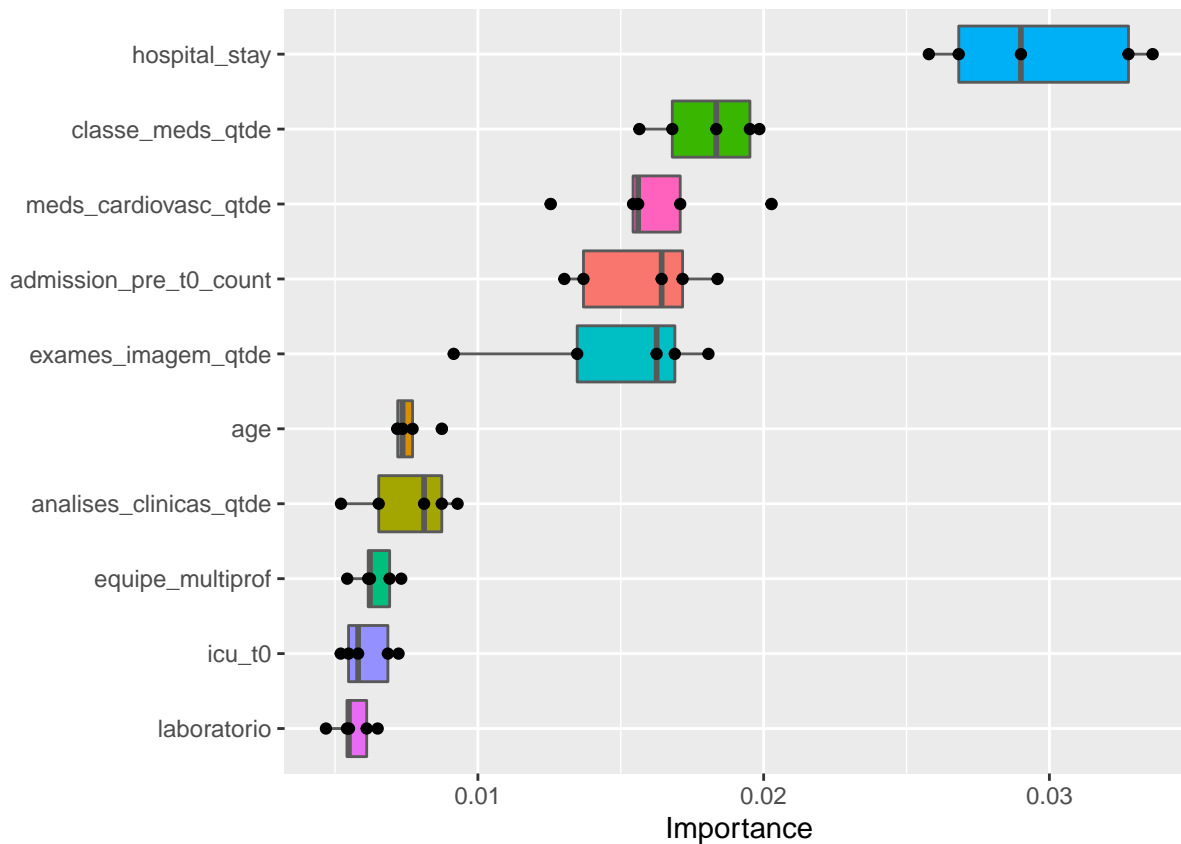
last_lightgbm_fit <-
  final_lightgbm_workflow %>%
  last_fit(df_split)

final_lightgbm_fit <- extract_workflow(last_lightgbm_fit)

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



```
## [1] "Optimal Threshold: 0.06"
## Confusion Matrix and Statistics
##
##      reference
## data    0    1
##    0 3505  120
##    1  971  134
##
##              Accuracy : 0.7693
##              95% CI   : (0.7571, 0.7813)
##    No Information Rate : 0.9463
##    P-Value [Acc > NIR] : 1
##
##              Kappa   : 0.1204
##
## Mcnemar's Test P-Value : <2e-16
##
##              Sensitivity : 0.7831
##              Specificity : 0.5276
##              Pos Pred Value : 0.9669
##              Neg Pred Value : 0.1213
##              Prevalence : 0.9463
##              Detection Rate : 0.7410
##              Detection Prevalence : 0.7664
##              Balanced Accuracy : 0.6553
##
##              '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: 2.718

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)
```

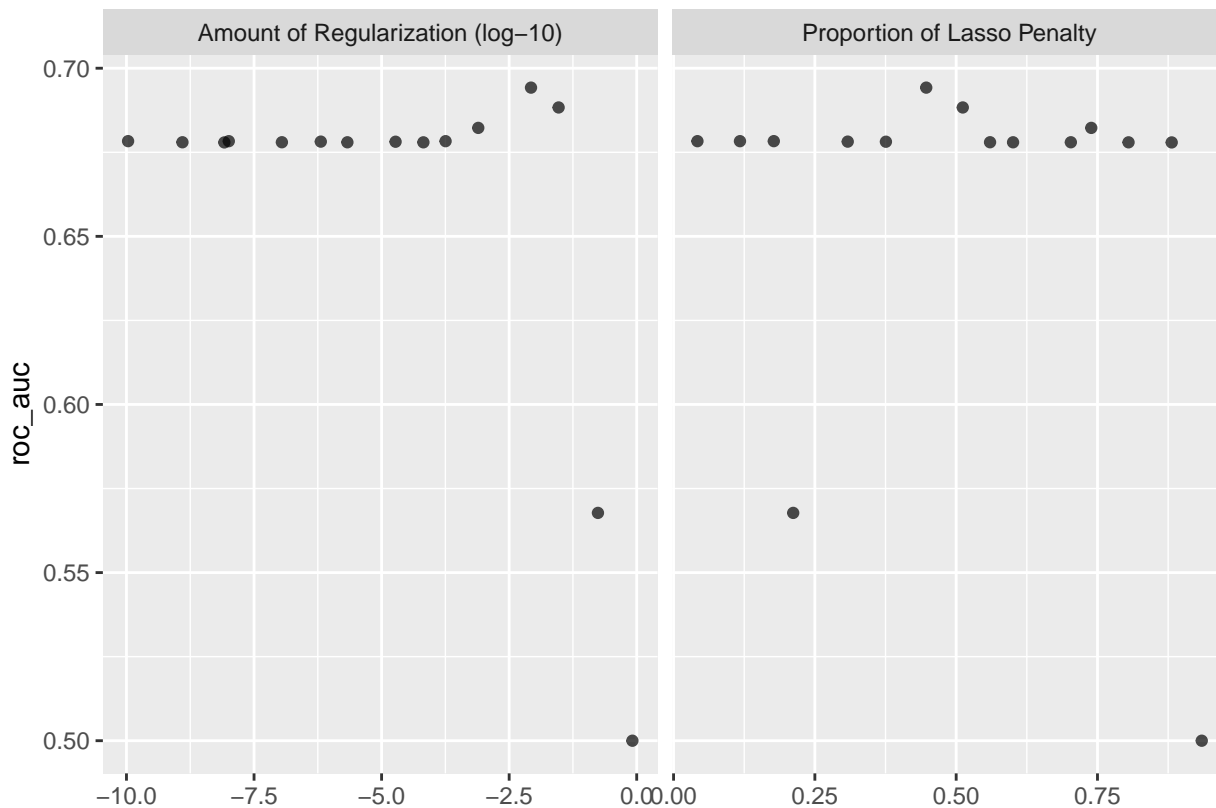
```
glmnet_workflow <-  
  workflow() %>%  
  add_recipe(glmnet_recipe) %>%  
  add_model(glmnet_spec)
```

```
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 1.79e- 4 0.0423 accuracy binary    0.941     5 0.00240 Preprocessor1_Model01  
## 2 1.79e- 4 0.0423 roc_auc  binary    0.678     5 0.0110  Preprocessor1_Model01  
## 3 1.01e- 8 0.117  accuracy binary    0.941     5 0.00240 Preprocessor1_Model02  
## 4 1.01e- 8 0.117  roc_auc  binary    0.678     5 0.0109  Preprocessor1_Model02  
## 5 1.08e-10 0.177  accuracy binary    0.941     5 0.00240 Preprocessor1_Model03  
## 6 1.08e-10 0.177  roc_auc  binary    0.678     5 0.0109  Preprocessor1_Model03  
## 7 1.74e- 1 0.212  accuracy binary    0.941     5 0.00221 Preprocessor1_Model04  
## 8 1.74e- 1 0.212  roc_auc  binary    0.568     5 0.0425  Preprocessor1_Model04  
## 9 6.44e- 7 0.308  accuracy binary    0.941     5 0.00240 Preprocessor1_Model05  
## 10 6.44e- 7 0.308  roc_auc  binary    0.678     5 0.0109  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 8.48e- 3  0.447 roc_auc binary   0.694     5 0.0148 Preprocessor1_Model07
## 2 2.95e- 2  0.512 roc_auc binary   0.688     5 0.0124 Preprocessor1_Model08
## 3 7.85e- 4  0.739 roc_auc binary   0.682     5 0.0116 Preprocessor1_Model12
## 4 1.08e-10  0.177 roc_auc binary   0.678     5 0.0109 Preprocessor1_Model03
## 5 1.01e- 8  0.117 roc_auc binary   0.678     5 0.0109 Preprocessor1_Model02

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

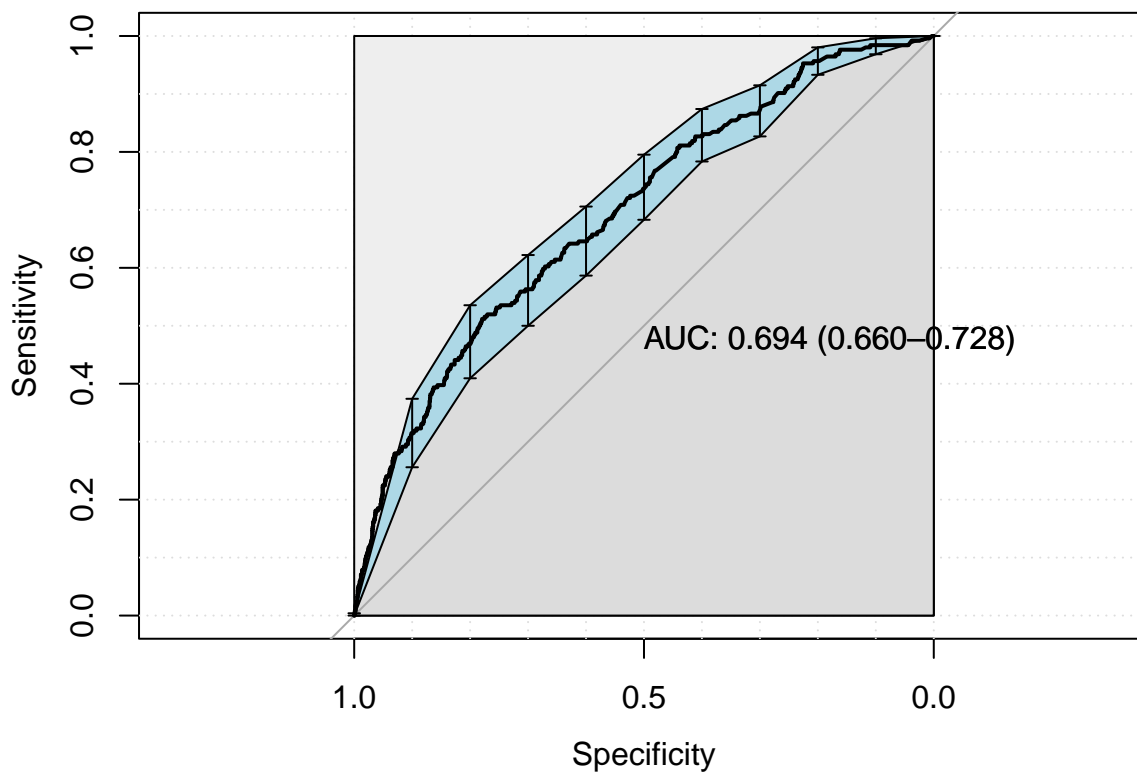
final_glmnet_workflow <-
  glmnet_workflow %>%
  finalize_workflow(best_glmnet)

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)

```



```

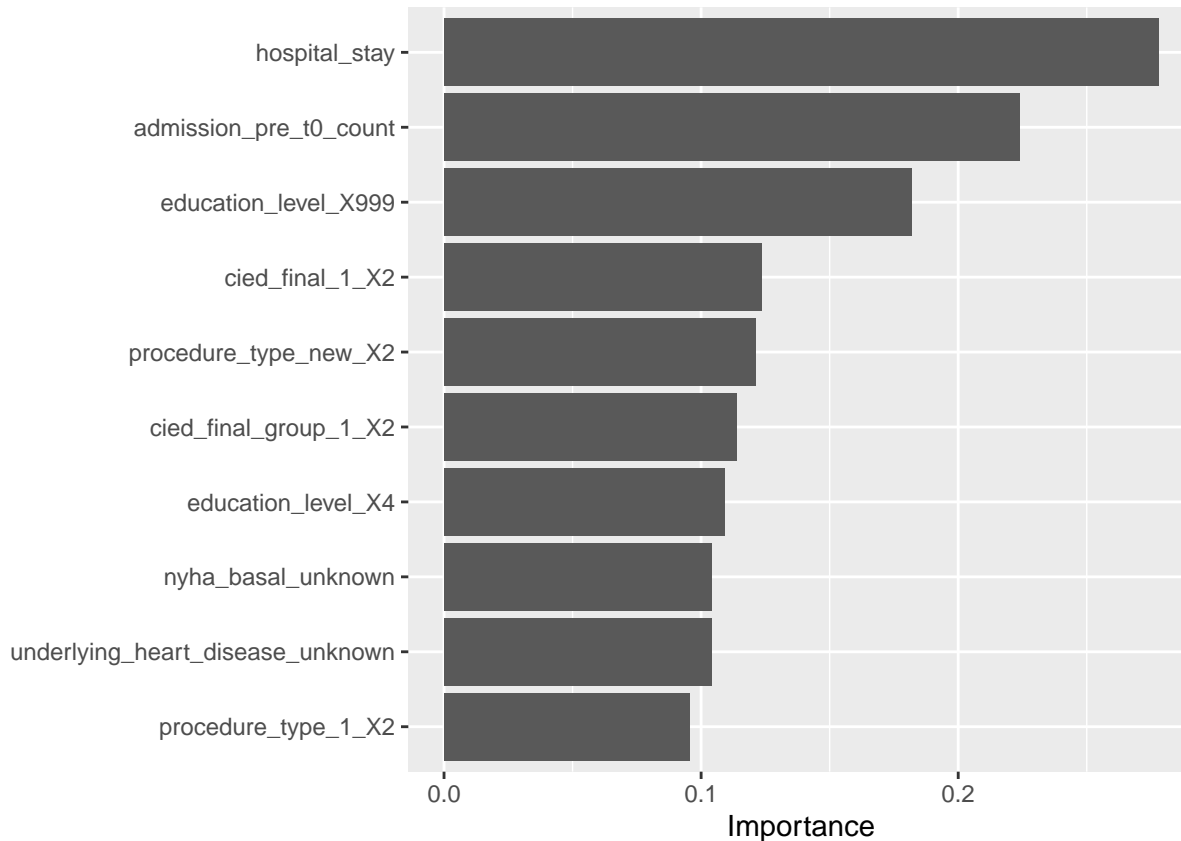
## [1] "Optimal Threshold: 0.06"
## Confusion Matrix and Statistics
##
##      reference
## data    0    1
## 0 3489 124
## 1  987 130

```

```
##
##           Accuracy : 0.7651
##           95% CI : (0.7528, 0.7771)
## No Information Rate : 0.9463
## P-Value [Acc > NIR] : 1
##
##           Kappa : 0.1119
##
## Mcnemar's Test P-Value : <2e-16
##
##           Sensitivity : 0.7795
##           Specificity : 0.5118
##           Pos Pred Value : 0.9657
##           Neg Pred Value : 0.1164
##           Prevalence : 0.9463
##           Detection Rate : 0.7376
##           Detection Prevalence : 0.7638
##           Balanced Accuracy : 0.6457
##
##           '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')
```



1.029

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)

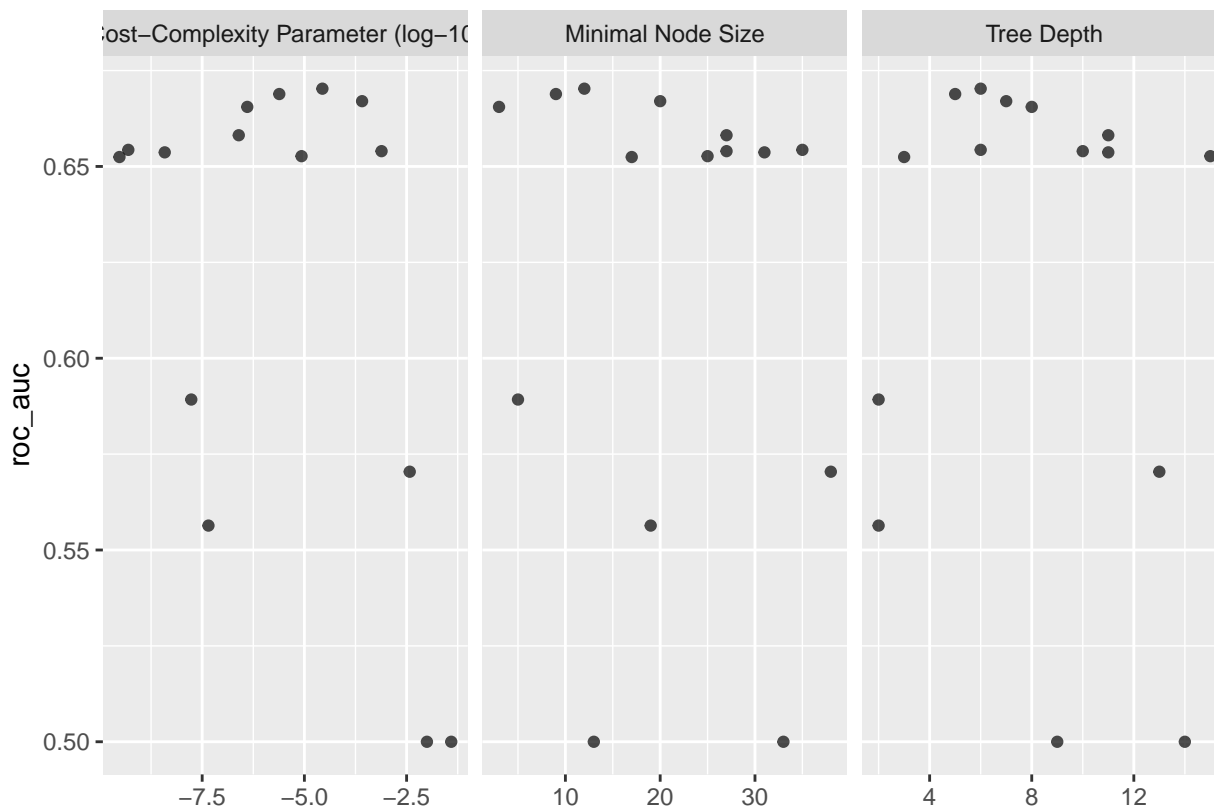
tree_workflow <-
  workflow() %>%
  add_recipe(tree_recipe) %>%
  add_model(tree_spec)

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")

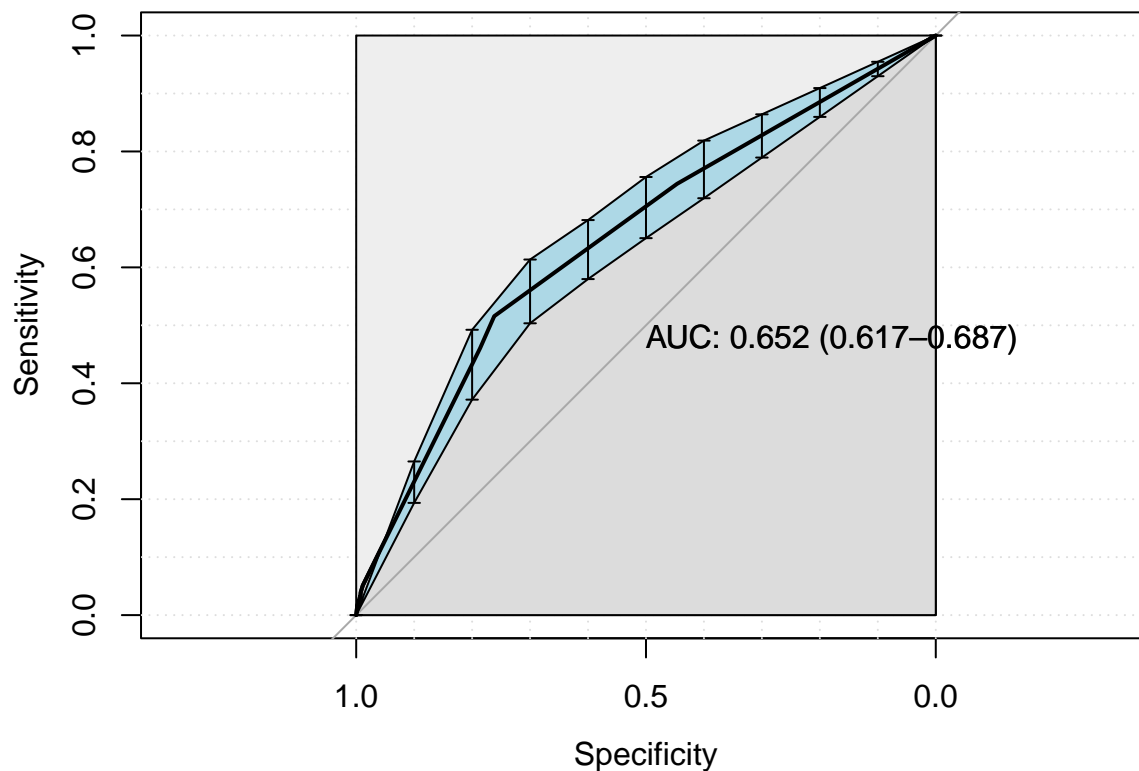
final_tree_workflow <-
  tree_workflow %>%
  finalize_workflow(best_tree)

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)

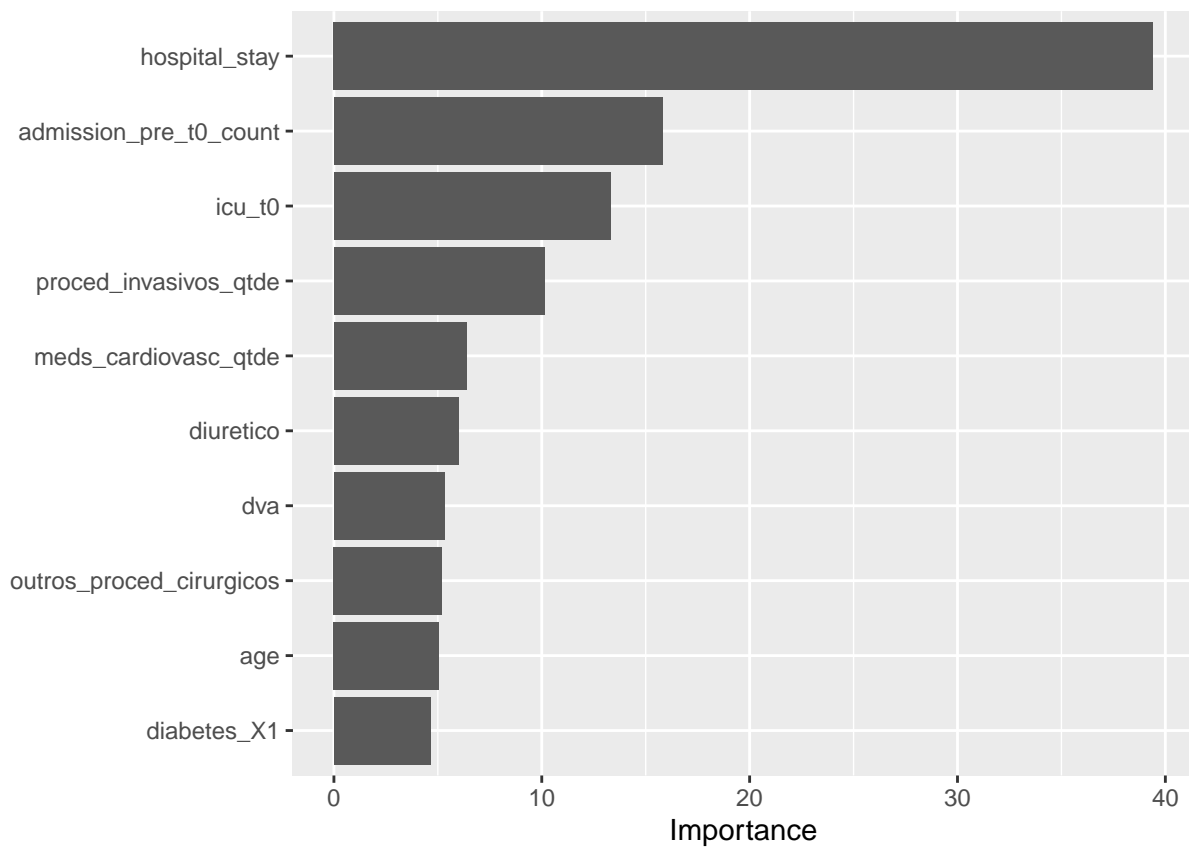
```



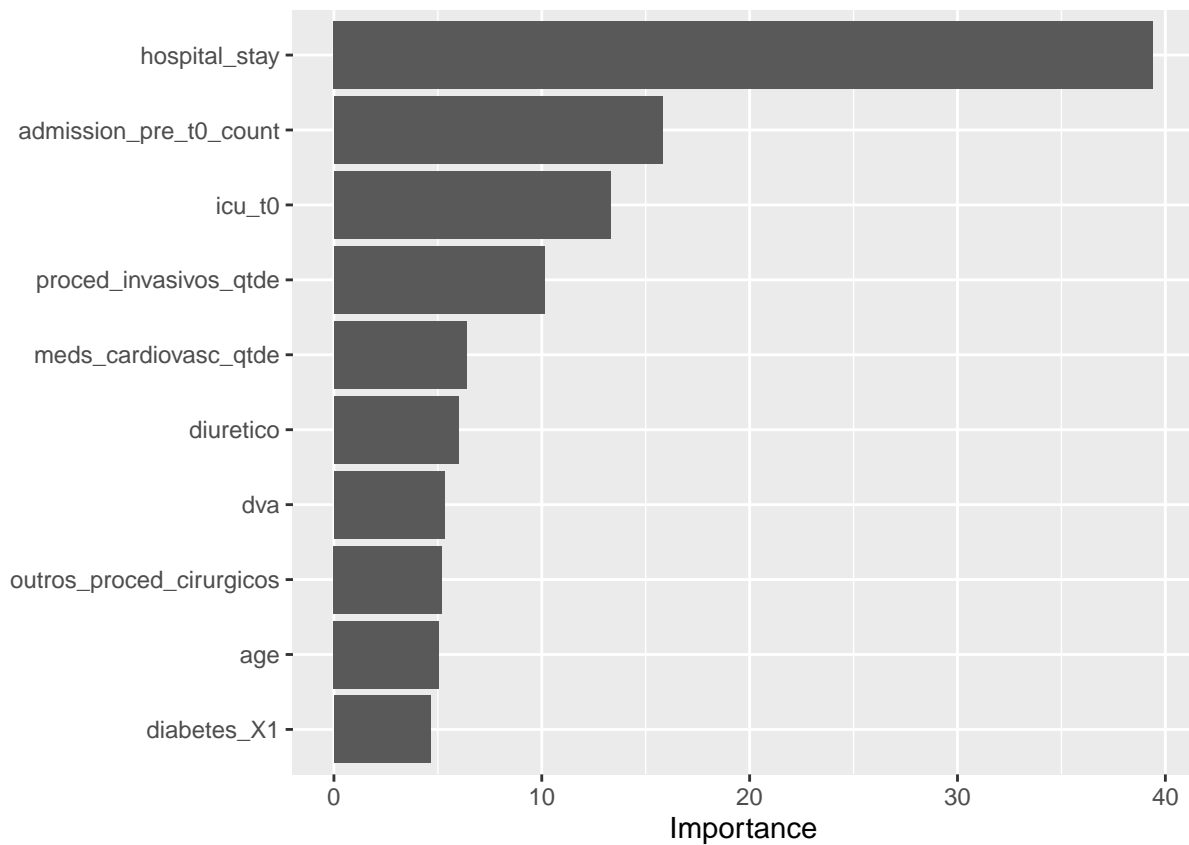
```

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')
```

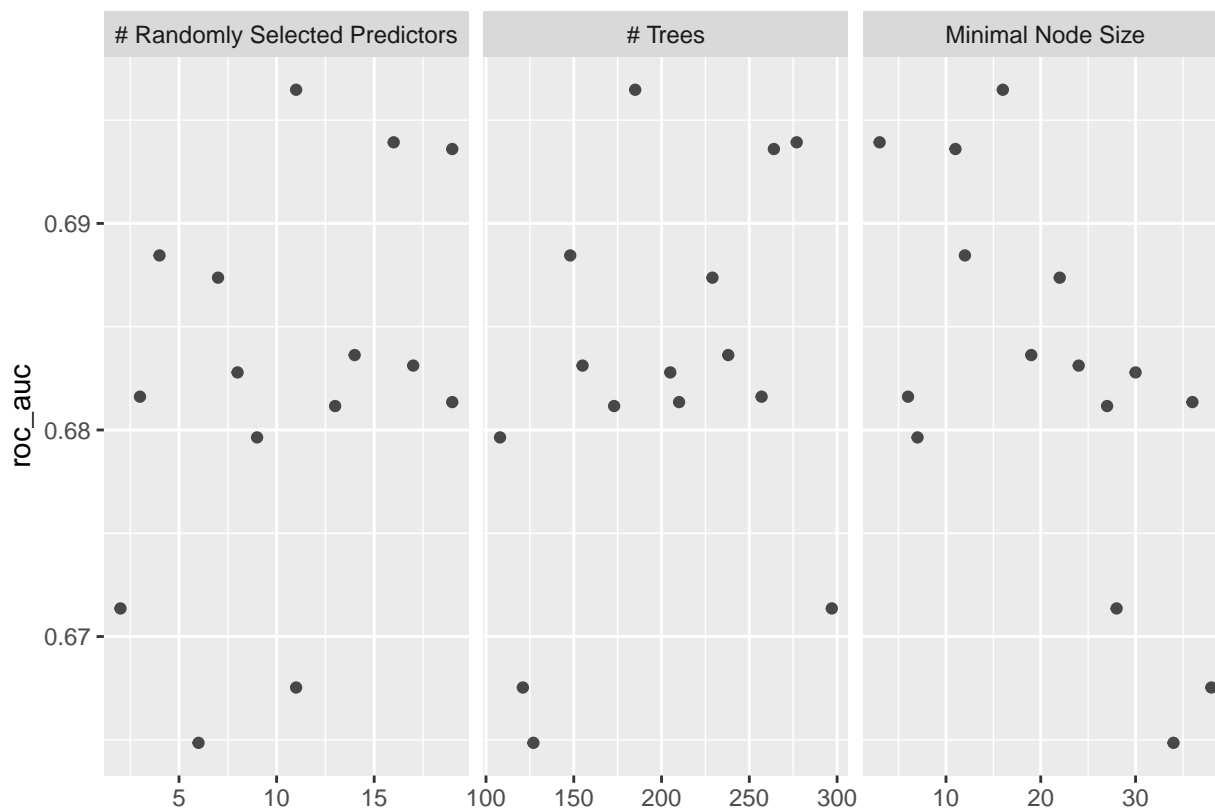


1.663

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")

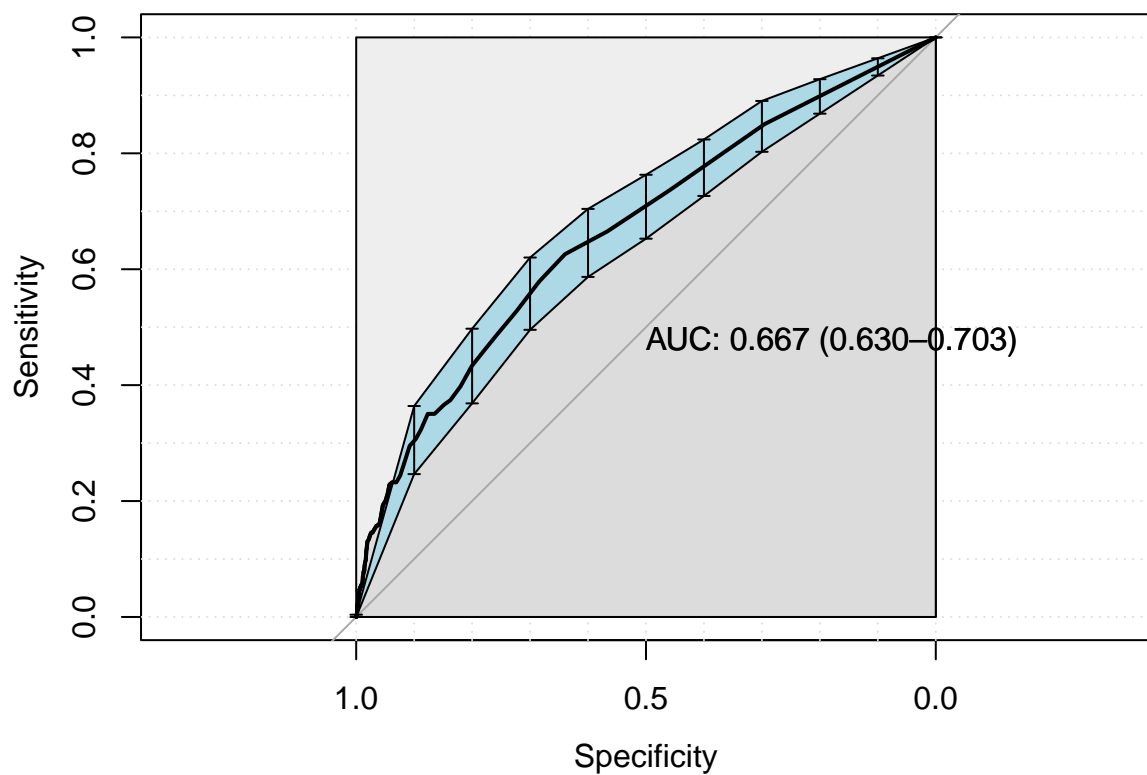
best_rf <- rf_tune %>%
  select_best("roc_auc")

final_rf_workflow <-
  rf_workflow %>%
  finalize_workflow(best_rf)

last_rf_fit <-
  final_rf_workflow %>%
  last_fit(df_split)

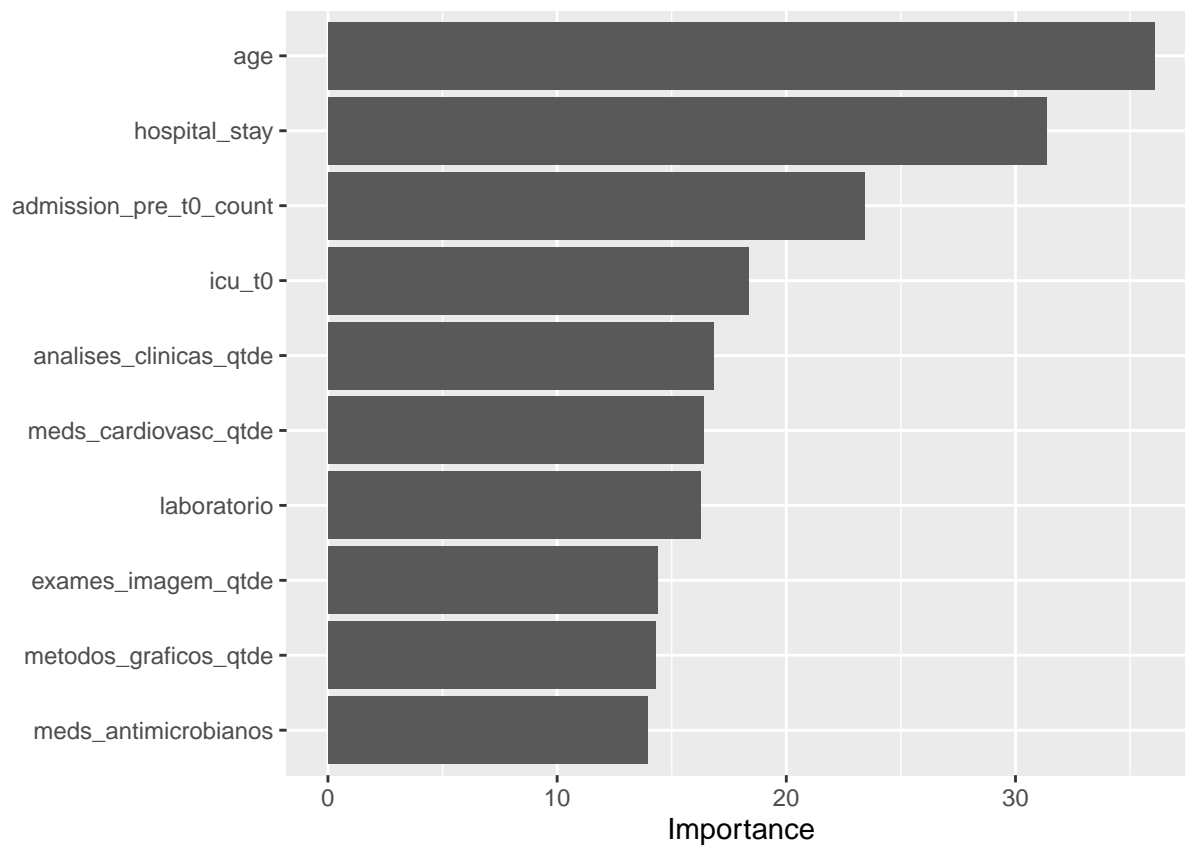
final_rf_fit <- extract_workflow(last_rf_fit)

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

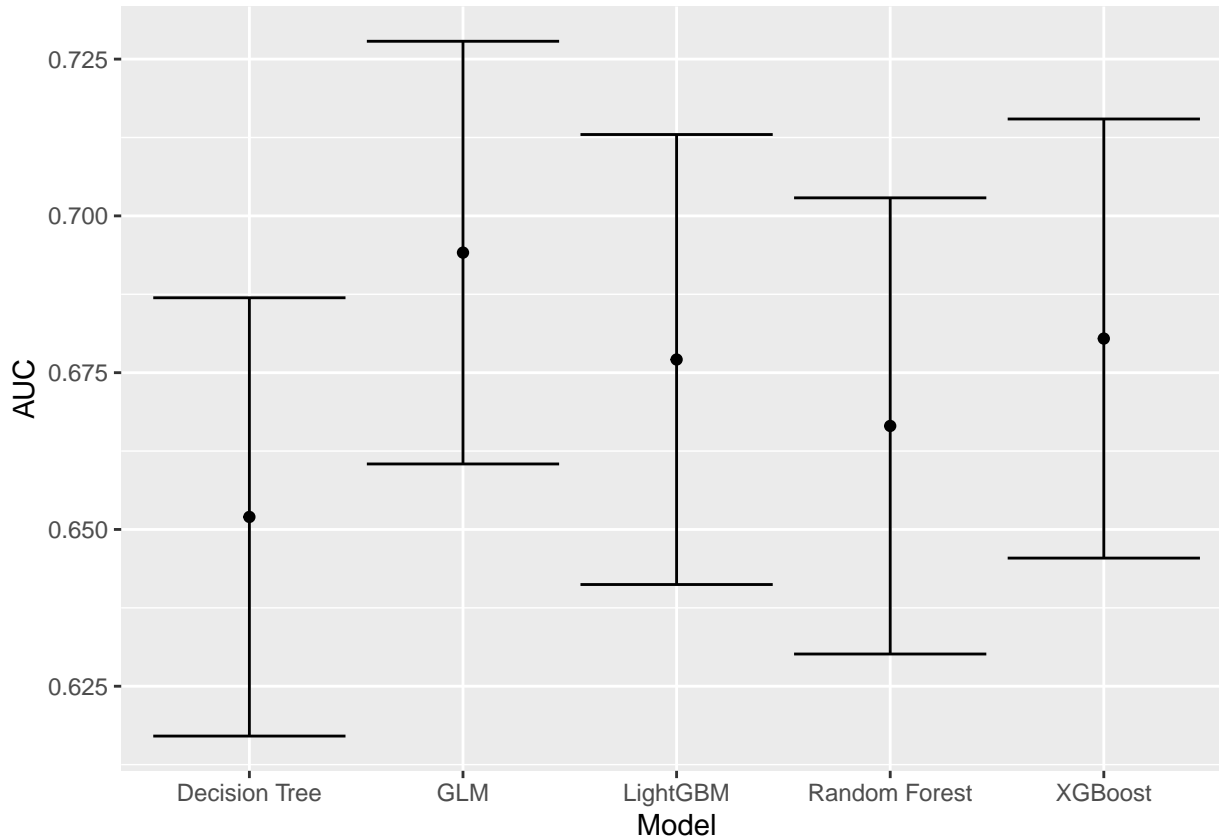
```
df_auc <- tibble::tribble(
  ~Model, ~`AUC`, ~`Lower Limit`, ~`Upper Limit`,
  'XGBoost', as.numeric(xgboost_auc$auc), xgboost_auc$ci[1], xgboost_auc$ci[3],
```

```

'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