Model Selection - readmission_1year

Eduardo Yuki Yada

Global parameters

```
k <- params$k # Number of folds for cross validation
grid_size <- params$grid_size # Number of parameter combination to tune on each model
repeats <- params$repeats
RUN_ALL_MODELS <- params$RUN_ALL_MODELS
Hmisc::list.tree(params)
```

```
##  params = list 5 (968 bytes)
## .  outcome_column = character 1= readmission_1year
## .  k = double 1= 10
## .  grid_size = double 1= 20
## .  repeats = double 1= 2
## .  RUN_ALL_MODELS = logical 1= TRUE
```

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")
predict <- stats::predict
```

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

```
cat_features_list = read_yaml(sprintf(
  "./auxiliar/significant_columns/categorical_%s.yaml",
  outcome_column
))

num_features_list = read_yaml(sprintf(
  "./auxiliar/significant_columns/numerical_%s.yaml",
  outcome_column
))

features_list = c(cat_features_list, num_features_list)
```

Minutes to run: 0

```
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))

features = base::intersect(eligible_features, features_list)

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

```
## 01. sex
## 02. age
## 03. race
## 04. education_level
## 05. patient_state
## 06. underlying_heart_disease
## 07. heart_disease
## 08. nyha_basal
## 09. prior_mi
## 10. heart_failure
## 11. af
## 12. cardiac_arrest
## 13. transplant
## 14. valvopathy
## 15. endocardites
## 16. diabetes
## 17. renal_failure
## 18. hemodialysis
## 19. copd
## 20. comorbidities_count
## 21. procedure_type_1
## 22. reop_type_1
## 23. procedure_type_new
## 24. cied_final_1
## 25. cied_final_group_1
## 26. admission_pre_t0_count
## 27. admission_pre_t0_180d
## 28. year_adm_t0
## 29. icu_t0
## 30. dialysis_t0
## 31. admission_t0_emergency
## 32. aco
## 33. antiarritmico
## 34. betabloqueador
## 35. ieca_bra
## 36. dva
## 37. digoxina
## 38. estatina
## 39. diuretico
## 40. vasodilatador
## 41. insuf_cardiaca
## 42. espironolactona
## 43. bloq_calcio
## 44. antiplaquetario_ev
## 45. insulina
## 46. anticonvulsivante
## 47. psicofarmacos
## 48. antifungico
## 49. antiviral
## 50. antiretroviral
## 51. classe_meds_qtde
## 52. meds_cardiovasc_qtde
## 53. meds_antimicrobianos
## 54. ventilacao_mecanica
## 55. cec
## 56. transplante_cardiaco
## 57. cir_toracica
## 58. outros_proced_cirurgicos
## 59. icp
## 60. intervencao_cv
## 61. angioplastia
## 62. cateterismo
## 63. eletrofisiologia
## 64. cateter_venoso_central
## 65. proced_invasivos_qtde
## 66. cve_desf
## 67. transfusao
## 68. interconsulta
## 69. equipe_multiprof
## 70. holter
## 71. teste_esforco
## 72. espiro_ergoespiro
## 73. tilt_teste
## 74. metodos_graficos_qtde
## 75. laboratorio
## 76. cultura
## 77. analises_clinicas_qtde
## 78. citologia
## 79. biopsia
## 80. histopatologia_qtde
## 81. angio_rm
## 82. angio_tc
## 83. aortografia
## 84. arteriografia
## 85. cintilografia
## 86. ecocardiograma
## 87. endoscopia
## 88. flebografia
## 89. pet_ct
## 90. ultrassom
## 91. tomografia
## 92. ressonancia
## 93. exames_imagem_qtde
## 94. dieta_parenteral
## 95. bic
## 96. mpp
## 97. 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))
```

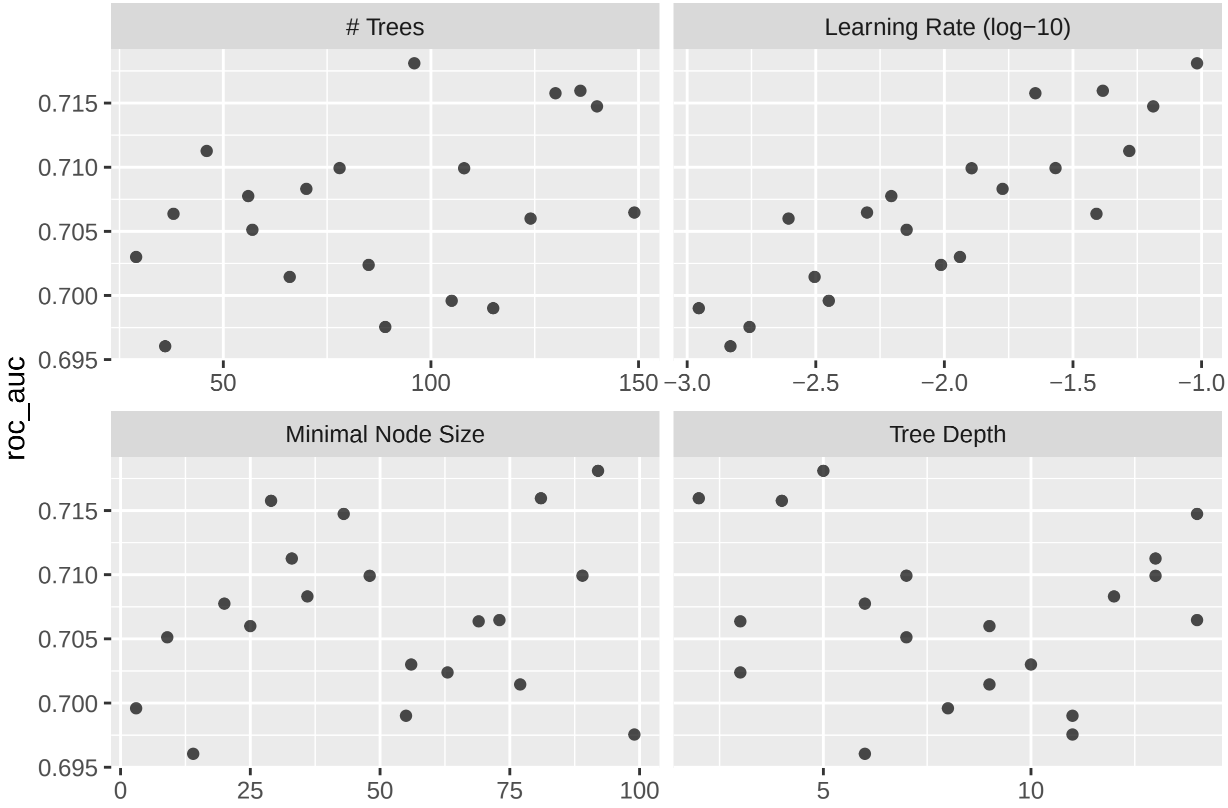
Minutes to run: 0.001

Boosted Tree (XGBoost)

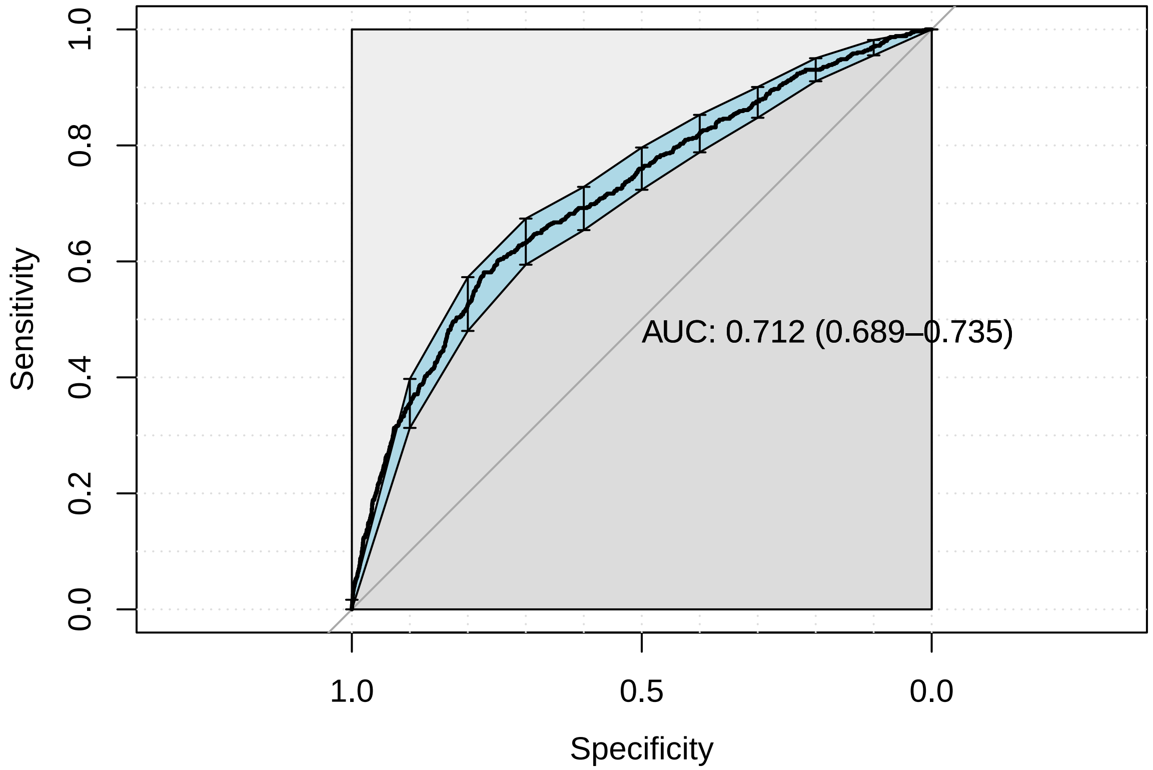
```
xgboost_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%
  step_dummy(all_nominal_predictors())

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

xgboost_grid <- grid_latin_hypercube(
  trees(range = c(25L, 150L)),
  min_n(range = c(2L, 100L)),
  tree_depth(range = c(2L, 15L)),
  learn_rate(range = c(-3, -1), trans = log10_trans()),
  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 10
##   trees min_n tree_depth learn_rate .metric .estimator  mean     n std_err .config
##   <int> <int>      <int>      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1    96    92          5     0.0960 roc_auc binary     0.718    10 0.00757 Preprocessor1_Model19
## 2   136    81          2     0.0413 roc_auc binary     0.716    10 0.00727 Preprocessor1_Model17
## 3   130    29          4     0.0226 roc_auc binary     0.716    10 0.00681 Preprocessor1_Model06
## 4   140    43         14     0.0649 roc_auc binary     0.715    10 0.00767 Preprocessor1_Model09
## 5    46    33         13     0.0524 roc_auc binary     0.711    10 0.00731 Preprocessor1_Model07
```

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

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

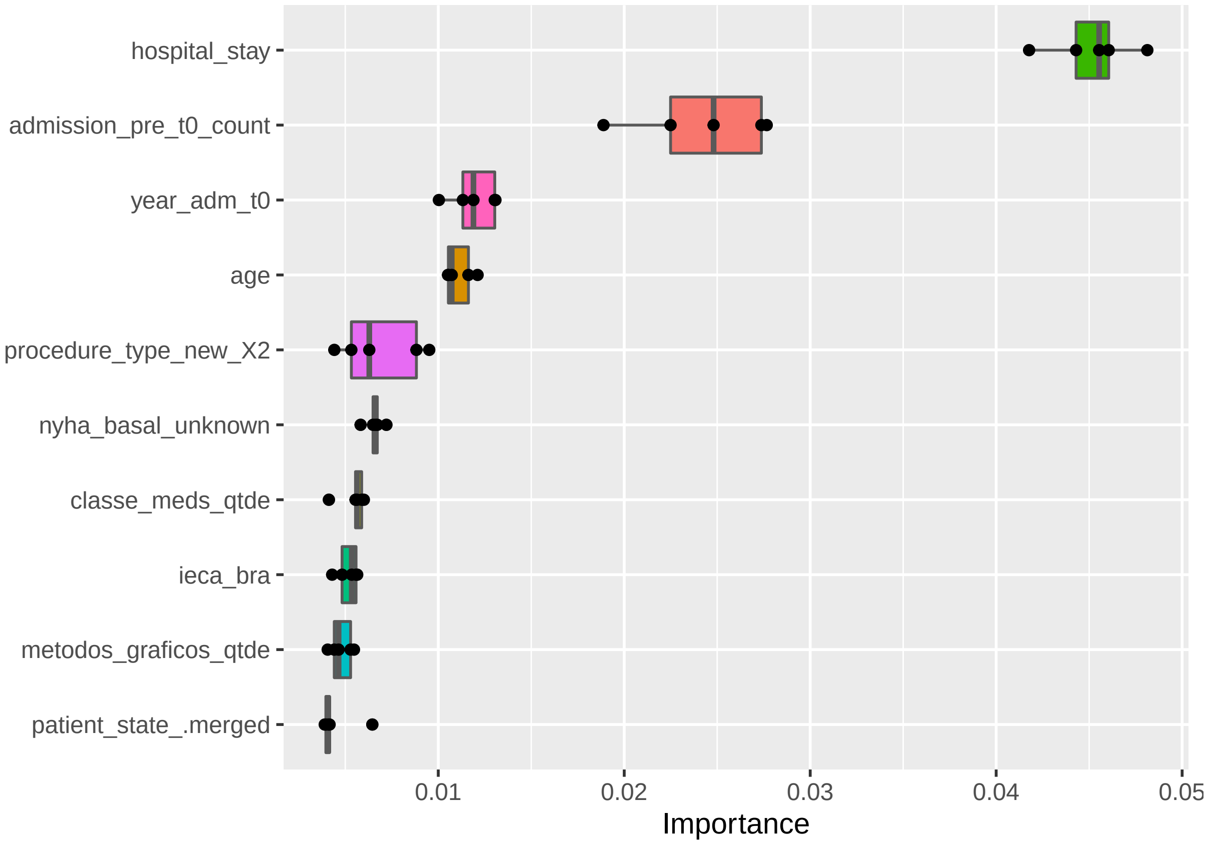


```
final_xgboost_workflow <-
  xgboost_workflow %>%
  finalize_workflow(best_xgboost)

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: 0.14"
## Confusion Matrix and Statistics
##
##     reference
## data    0    1
##    0 3188  253
##    1  939  351
##
##                Accuracy : 0.748
##                  95% CI : (0.7354, 0.7604)
##     No Information Rate : 0.8723
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2382
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7725
##             Specificity : 0.5811
##          Pos Pred Value : 0.9265
##          Neg Pred Value : 0.2721
##              Prevalence : 0.8723
##          Detection Rate : 0.6739
##    Detection Prevalence : 0.7273
##       Balanced Accuracy : 0.6768
##
##        '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(-.metric, -.estimator, -.config, -mean, -n, -std_err) %>%
  as.list
```

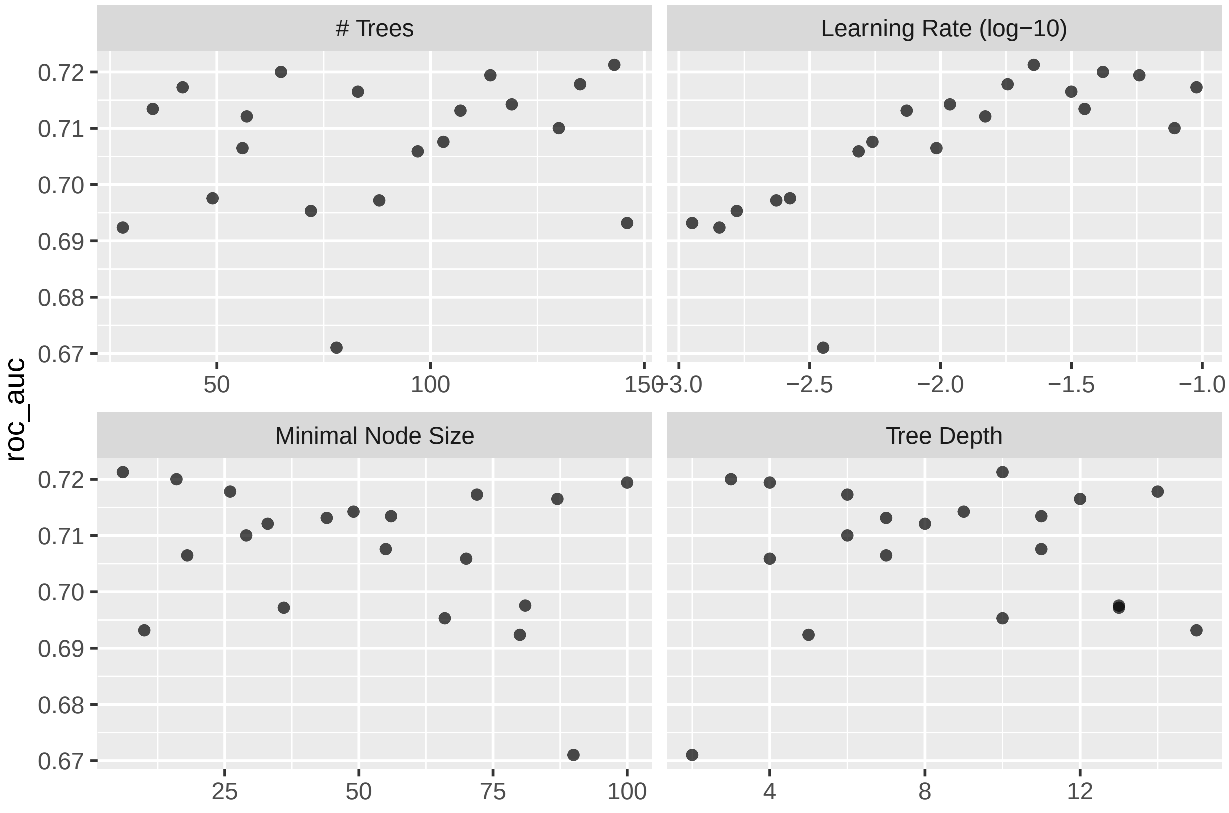
Minutes to run: 5.877

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(
  trees = tune(),
  min_n = tune(),
  tree_depth = tune(),
  learn_rate = tune(),
  sample_size = 1
) %>%
  set_engine("lightgbm",
             nthread = 8) %>%
  set_mode("classification")

lightgbm_grid <- grid_latin_hypercube(
  trees(range = c(25L, 150L)),
  min_n(range = c(2L, 100L)),
  tree_depth(range = c(2L, 15L)),
  learn_rate(range = c(-3, -1), trans = log10_trans()),
  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")
```

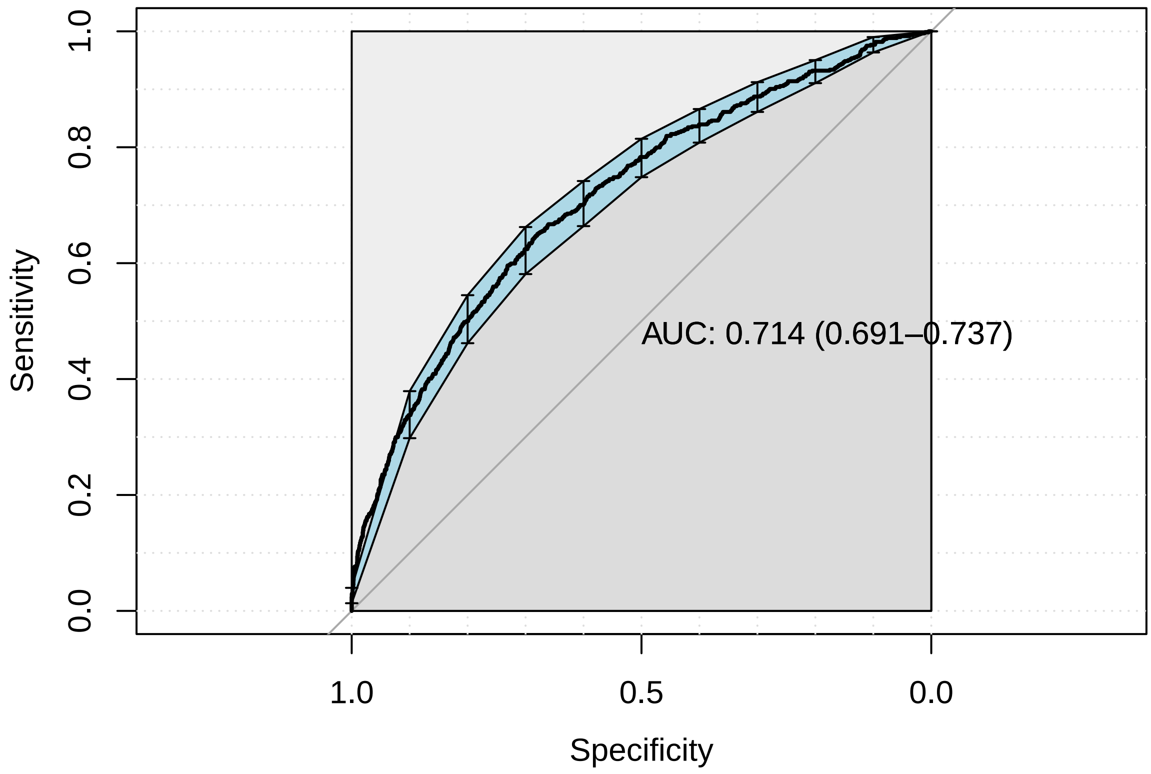
```
## # A tibble: 5 x 10
##   trees min_n tree_depth learn_rate .metric .estimator  mean     n std_err .config
##   <int> <int>      <int>      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1   143     6         10     0.0227 roc_auc binary     0.721    10 0.00741 Preprocessor1_Model01
## 2    65    16          3     0.0417 roc_auc binary     0.720    10 0.00666 Preprocessor1_Model03
## 3   114   100          4     0.0575 roc_auc binary     0.719    10 0.00703 Preprocessor1_Model20
## 4   135    26         14     0.0180 roc_auc binary     0.718    10 0.00751 Preprocessor1_Model05
## 5    42    72          6     0.0951 roc_auc binary     0.717    10 0.00744 Preprocessor1_Model15
```

```
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: 0.11"
## Confusion Matrix and Statistics
##
##     reference
## data    0    1
##    0 2804  211
##    1 1323  393
##
##                Accuracy : 0.6758
##                  95% CI : (0.6622, 0.6891)
##     No Information Rate : 0.8723
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1848
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6794
##             Specificity : 0.6507
##          Pos Pred Value : 0.9300
##          Neg Pred Value : 0.2290
##              Prevalence : 0.8723
##          Detection Rate : 0.5927
##    Detection Prevalence : 0.6373
##       Balanced Accuracy : 0.6650
##
```

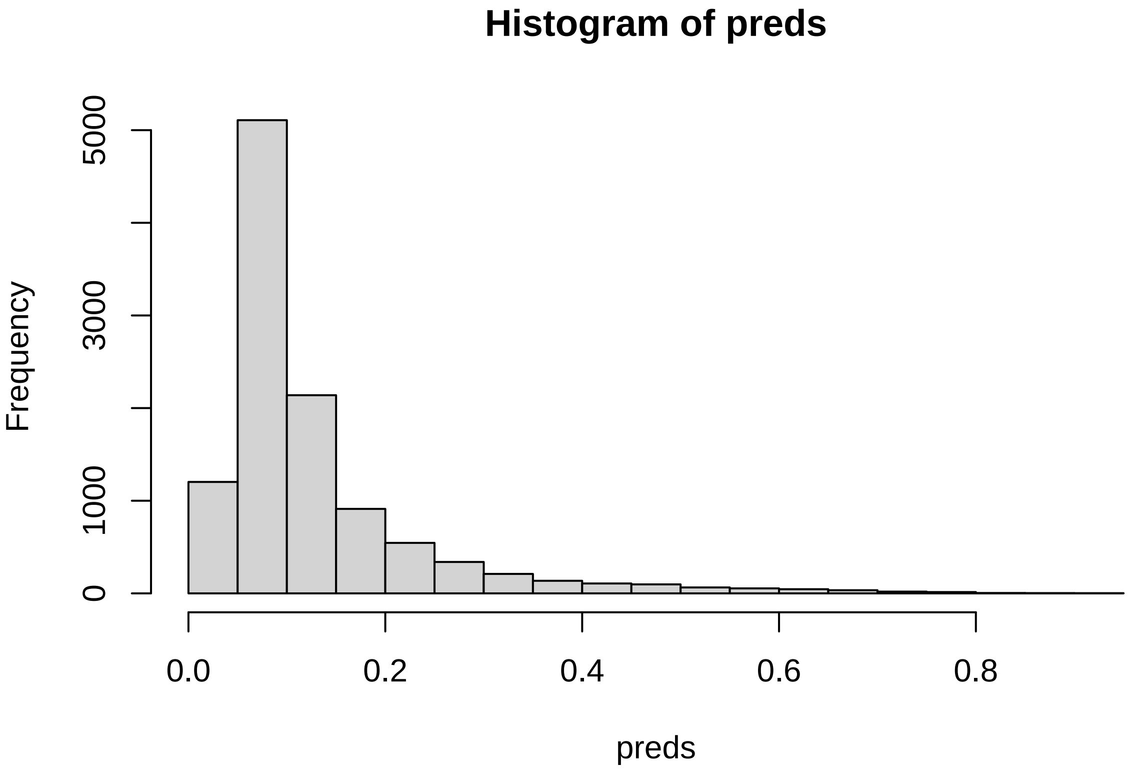
```
##        'Positive' Class : 0
##
```

```
lightgbm_parameters <- lightgbm_tune %>%
  show_best("roc_auc", n = 1) %>%
  select(-.metric, -.estimator, -.config, -mean, -n, -std_err) %>%
  as.list

Hmisc::list.tree(lightgbm_parameters)
```

```
##  lightgbm_parameters = list 4 (736 bytes)
## .  trees = integer 1= 143
## .  min_n = integer 1= 6
## .  tree_depth = integer 1= 10
## .  learn_rate = double 1= 0.022707
```

```
con <- file(sprintf('./auxiliar/model_selection/hyperparameters/%s.yaml', outcome_column), "w")
write_yaml(lightgbm_parameters, con)
close(con)
```

Minutes to run: 4.067

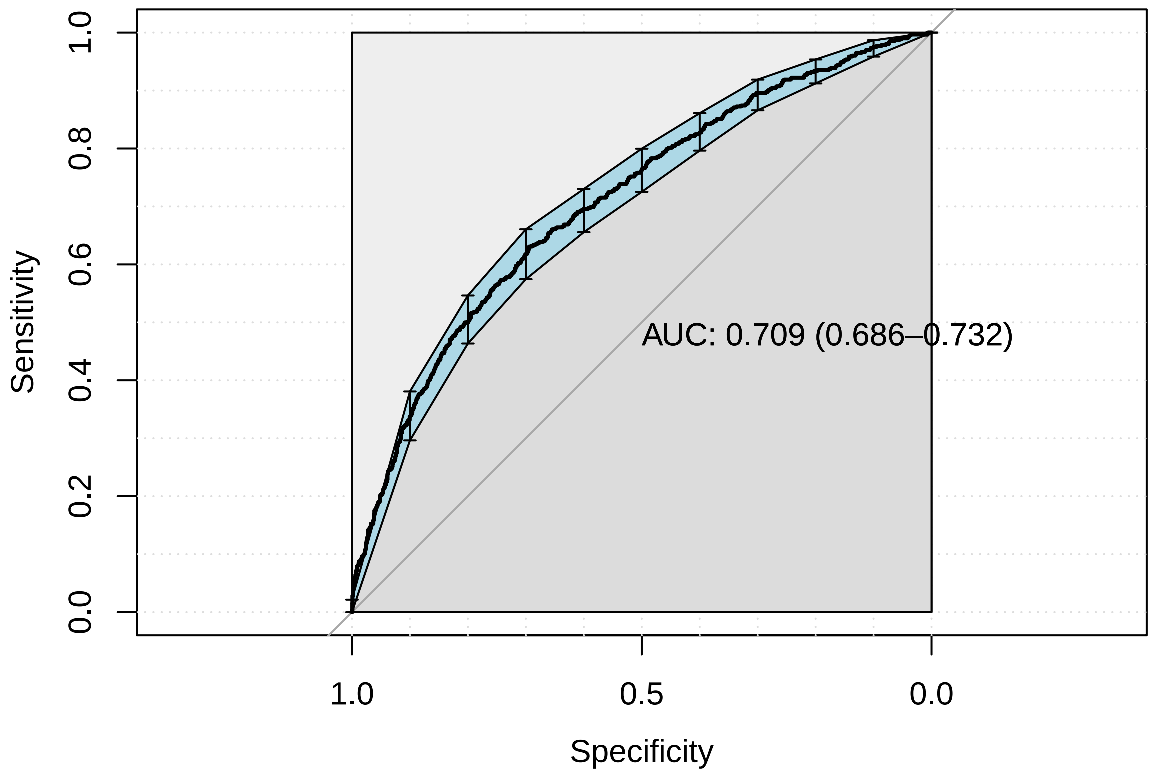


Minutes to run: 0.017

GLM

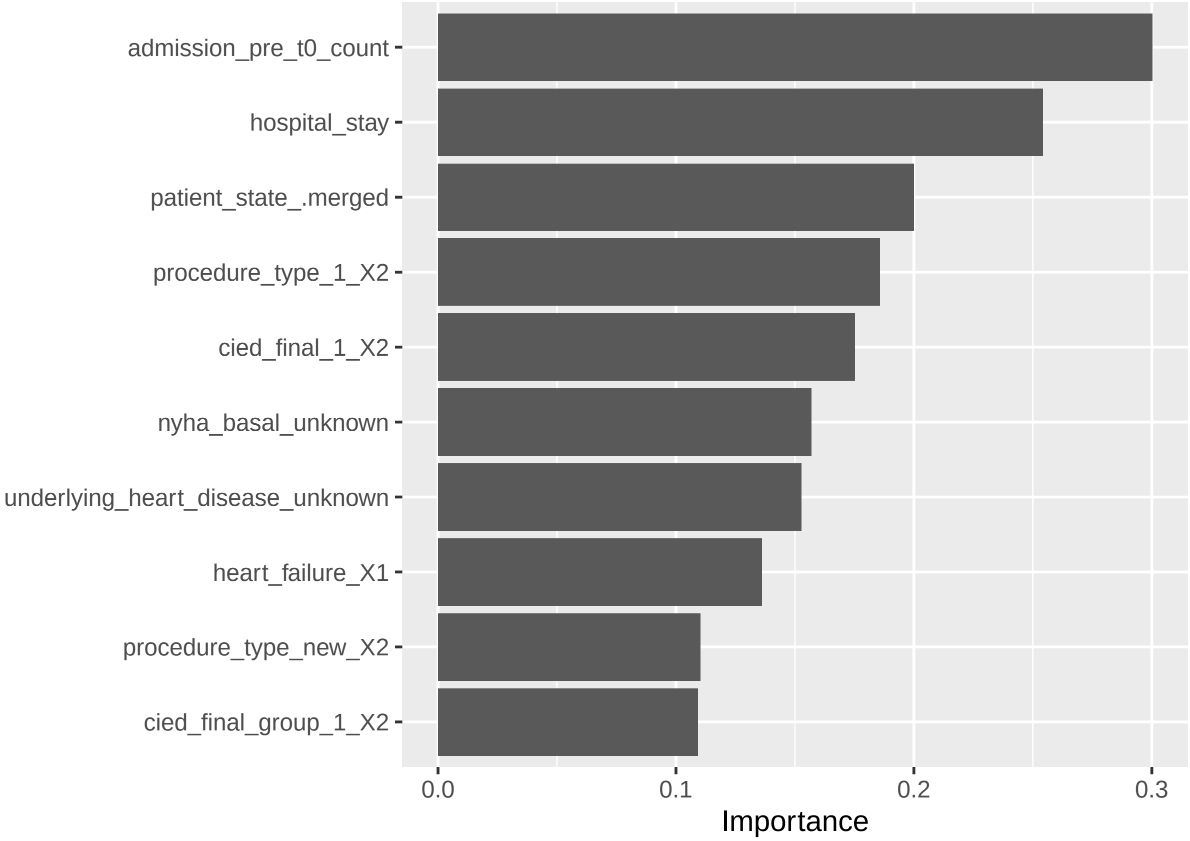
```
glmnet_recipe <-
  recipe(formula = sprintf("%s ~ .", outcome_column) %>% as.formula, data = df_train) %>%
  step_novel(all_nominal_predictors()) %>%
  step_unknown(all_nominal_predictors()) %>%
  step_other(all_nominal_predictors(), threshold = 0.05, other = ".merged") %>%
  step_dummy(all_nominal_predictors()) %>%
  step_zv(all_predictors()) %>%
  step_normalize(all_numeric_predictors())
```

```
glmnet_spec <-
  logistic_reg(penalty = 0) %>%
  set_mode("classification") %>%
  set_engine("glmnet")

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

glm_fit <- glmnet_workflow %>%
  fit(df_train)

glmnet_auc <- validation(glm_fit, df_test)
```

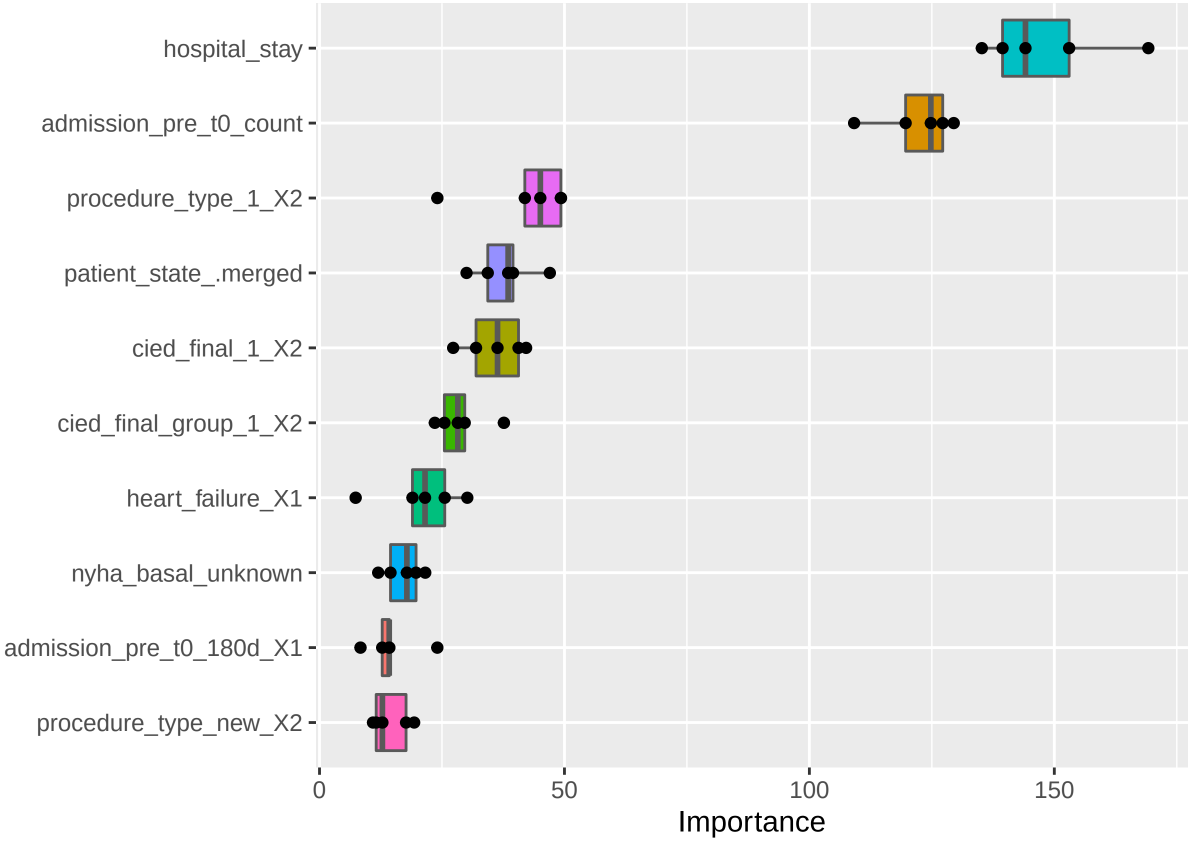


```
## [1] "Optimal Threshold: 0.12"
## Confusion Matrix and Statistics
##
##     reference
## data    0    1
##    0 2866  223
##    1 1261  381
##
##                Accuracy : 0.6863
##                  95% CI : (0.6729, 0.6995)
##     No Information Rate : 0.8723
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1876
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6945
##             Specificity : 0.6308
##          Pos Pred Value : 0.9278
##          Neg Pred Value : 0.2320
##              Prevalence : 0.8723
##          Detection Rate : 0.6058
##    Detection Prevalence : 0.6529
##       Balanced Accuracy : 0.6626
##
##        'Positive' Class : 0
##
```

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

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



```
extract_vip(glm_fit, pred_wrapper = pfun_glmnet,
            reference_class = "1", method = 'permute')
```



Minutes to run: 2.09

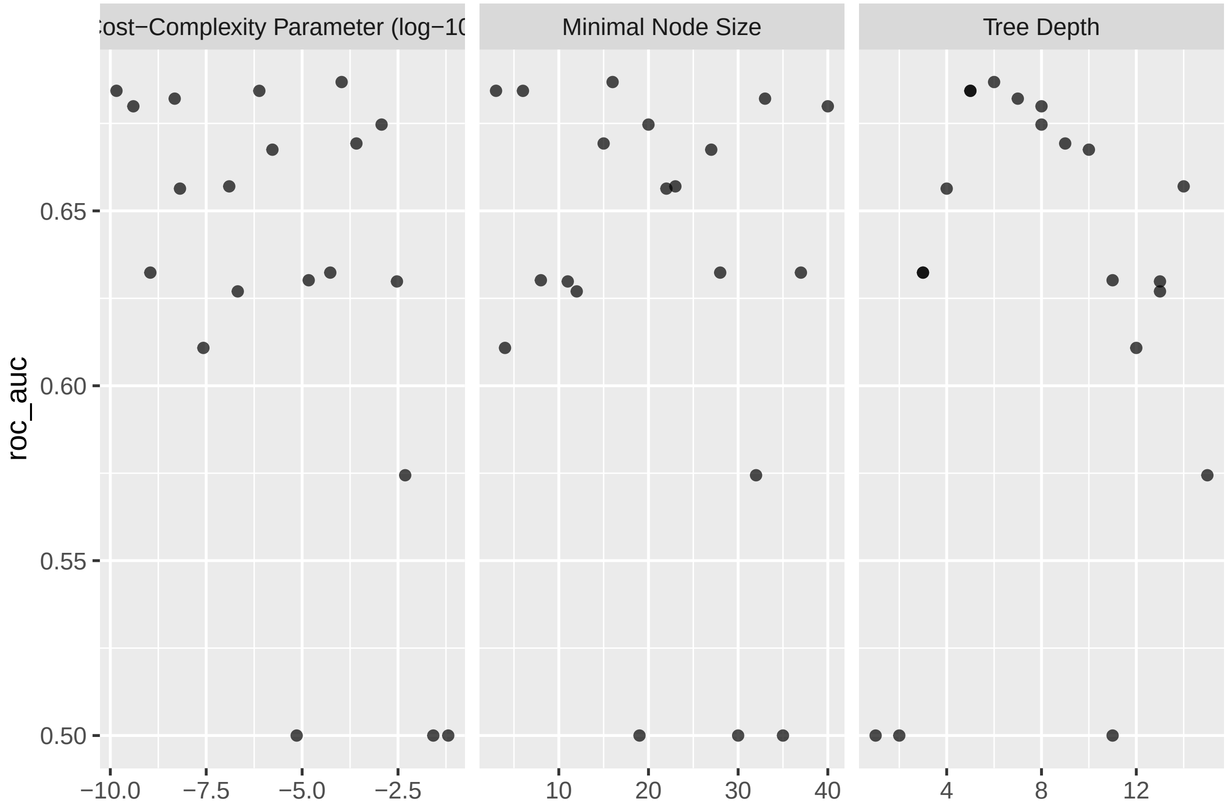
Decision Tree

```
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()
```

```
## # A tibble: 40 x 9
##    cost_complexity tree_depth min_n .metric  .estimator  mean     n  std_err .config
##              <dbl>      <int> <int> <chr>    <chr>      <dbl> <int>    <dbl> <chr>
##  1        2.67e- 8         12     4 accuracy binary     0.848    10 0.00210  Preprocessor1_Model01
##  2        2.67e- 8         12     4 roc_auc  binary     0.611    10 0.00631  Preprocessor1_Model01
##  3        1.07e- 4          6    16 accuracy binary     0.869    10 0.00108  Preprocessor1_Model02
##  4        1.07e- 4          6    16 roc_auc  binary     0.687    10 0.00542  Preprocessor1_Model02
##  5        4.77e- 9          7    33 accuracy binary     0.867    10 0.00102  Preprocessor1_Model03
##  6        4.77e- 9          7    33 roc_auc  binary     0.682    10 0.00619  Preprocessor1_Model03
##  7        1.45e-10          5     3 accuracy binary     0.871    10 0.000925 Preprocessor1_Model04
##  8        1.45e-10          5     3 roc_auc  binary     0.684    10 0.00512  Preprocessor1_Model04
##  9        7.66e- 7          5     6 accuracy binary     0.871    10 0.000925 Preprocessor1_Model05
## 10        7.66e- 7          5     6 roc_auc  binary     0.684    10 0.00513  Preprocessor1_Model05
## # ... with 30 more rows
## # i Use `print(n = ...)` to see more rows
```

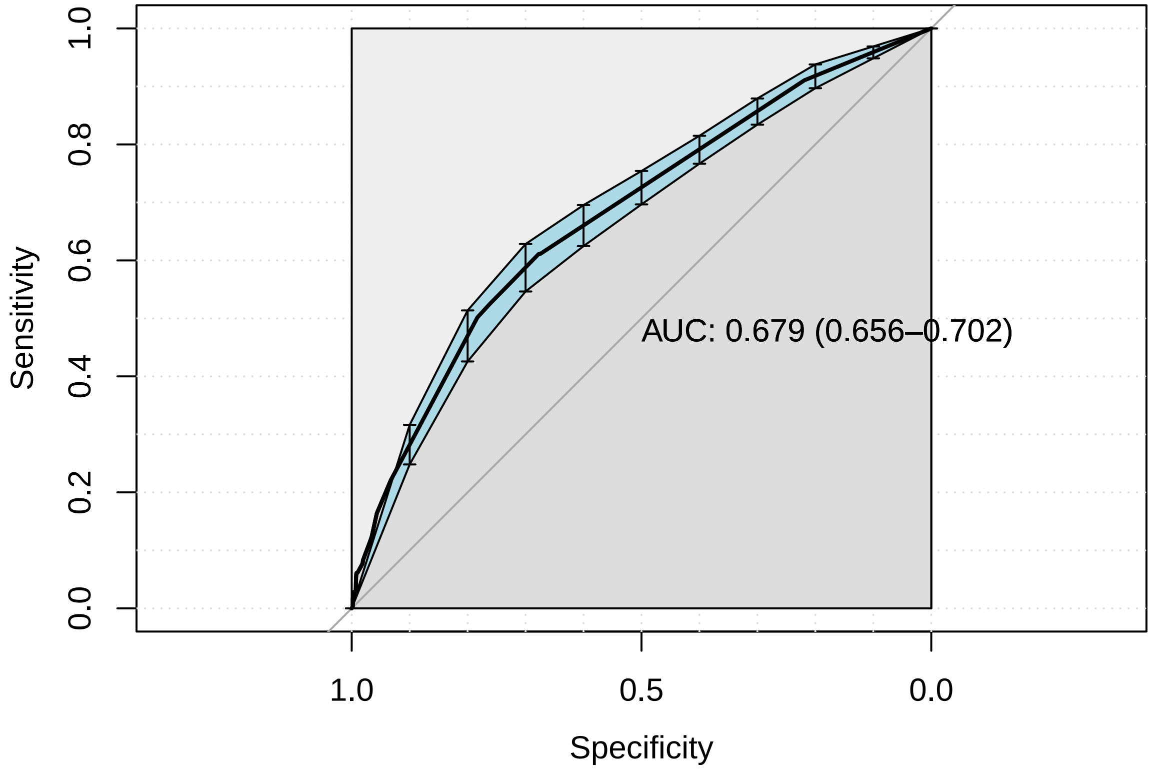
```
autoplot(tree_tune, metric = "roc_auc")
```



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

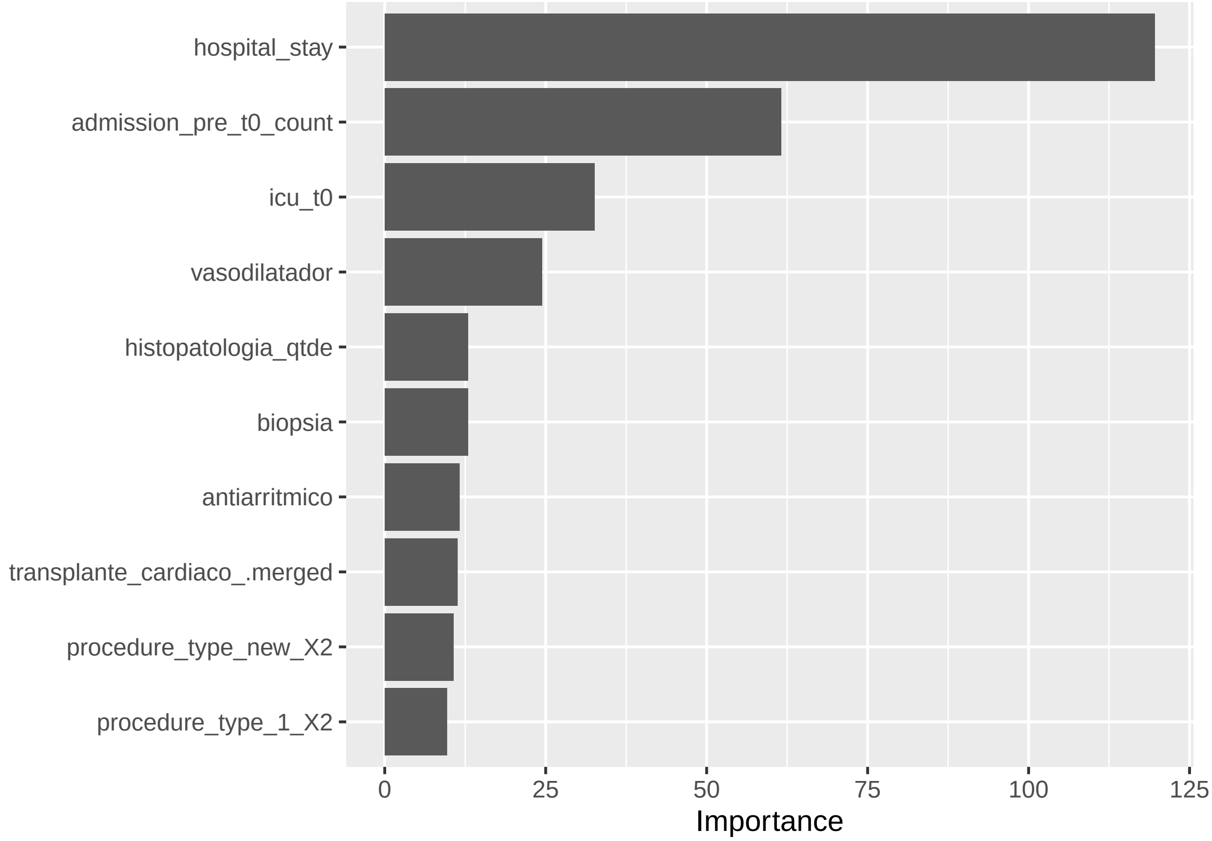
```
## # A tibble: 5 x 9
##   cost_complexity tree_depth min_n .metric .estimator  mean     n std_err .config
##             <dbl>      <int> <int> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1        1.07e- 4          6    16 roc_auc binary     0.687    10 0.00542 Preprocessor1_Model02
## 2        1.45e-10          5     3 roc_auc binary     0.684    10 0.00512 Preprocessor1_Model04
## 3        7.66e- 7          5     6 roc_auc binary     0.684    10 0.00513 Preprocessor1_Model05
## 4        4.77e- 9          7    33 roc_auc binary     0.682    10 0.00619 Preprocessor1_Model03
## 5        3.98e-10          8    40 roc_auc binary     0.680    10 0.00410 Preprocessor1_Model20
```

```r
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)
```



```
## [1] "Optimal Threshold: 0.11"
## Confusion Matrix and Statistics
##
##     reference
## data    0    1
##    0 2796  235
##    1 1331  369
##
##                Accuracy : 0.669
##                  95% CI : (0.6554, 0.6824)
##     No Information Rate : 0.8723
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1625
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6775
##              Specificity : 0.6109
##           Pos Pred Value : 0.9225
##           Neg Pred Value : 0.2171
##               Prevalence : 0.8723
##           Detection Rate : 0.5910
##     Detection Prevalence : 0.6407
##        Balanced Accuracy : 0.6442
##
##         'Positive' Class : 0
##
```

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



```
# extract_vip(final_tree_fit, pred_wrapper = predict,
#             reference_class = "1", use_matrix = FALSE,
#             method = 'permute')
```

Minutes to run: 5.454

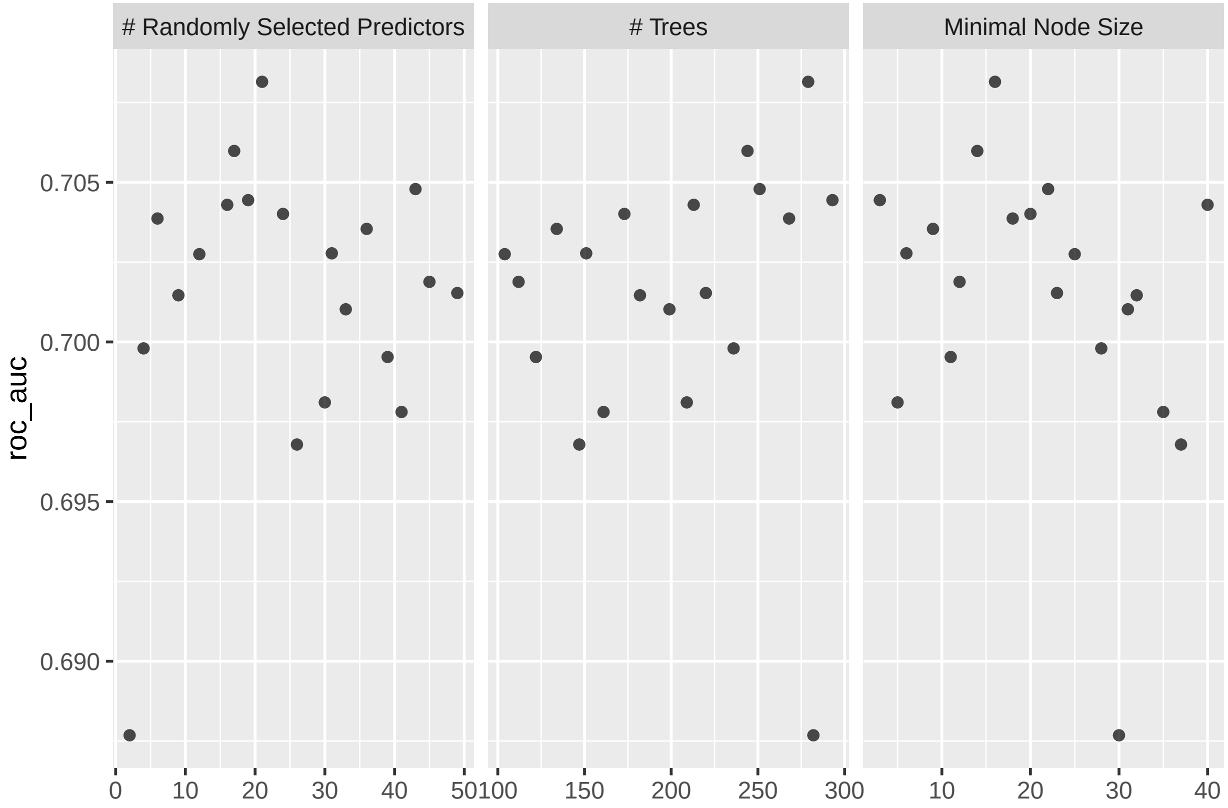
Random Forest

```
rf_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())

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, 50L)),
                                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()
```

```
## # A tibble: 40 x 9
##     mtry trees min_n .metric  .estimator  mean     n  std_err .config
##    <int> <int> <int> <chr>    <chr>      <dbl> <int>    <dbl> <chr>
##  1     9   182    32 accuracy binary     0.874    10 0.000945 Preprocessor1_Model01
##  2     9   182    32 roc_auc  binary     0.701    10 0.00890  Preprocessor1_Model01
##  3    45   112    12 accuracy binary     0.875    10 0.00133  Preprocessor1_Model02
##  4    45   112    12 roc_auc  binary     0.702    10 0.00928  Preprocessor1_Model02
##  5    12   104    25 accuracy binary     0.874    10 0.000853 Preprocessor1_Model03
##  6    12   104    25 roc_auc  binary     0.703    10 0.00862  Preprocessor1_Model03
##  7    26   147    37 accuracy binary     0.875    10 0.00142  Preprocessor1_Model04
##  8    26   147    37 roc_auc  binary     0.697    10 0.00852  Preprocessor1_Model04
##  9    43   251    22 accuracy binary     0.875    10 0.00132  Preprocessor1_Model05
## 10    43   251    22 roc_auc  binary     0.705    10 0.00842  Preprocessor1_Model05
## # ... with 30 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```

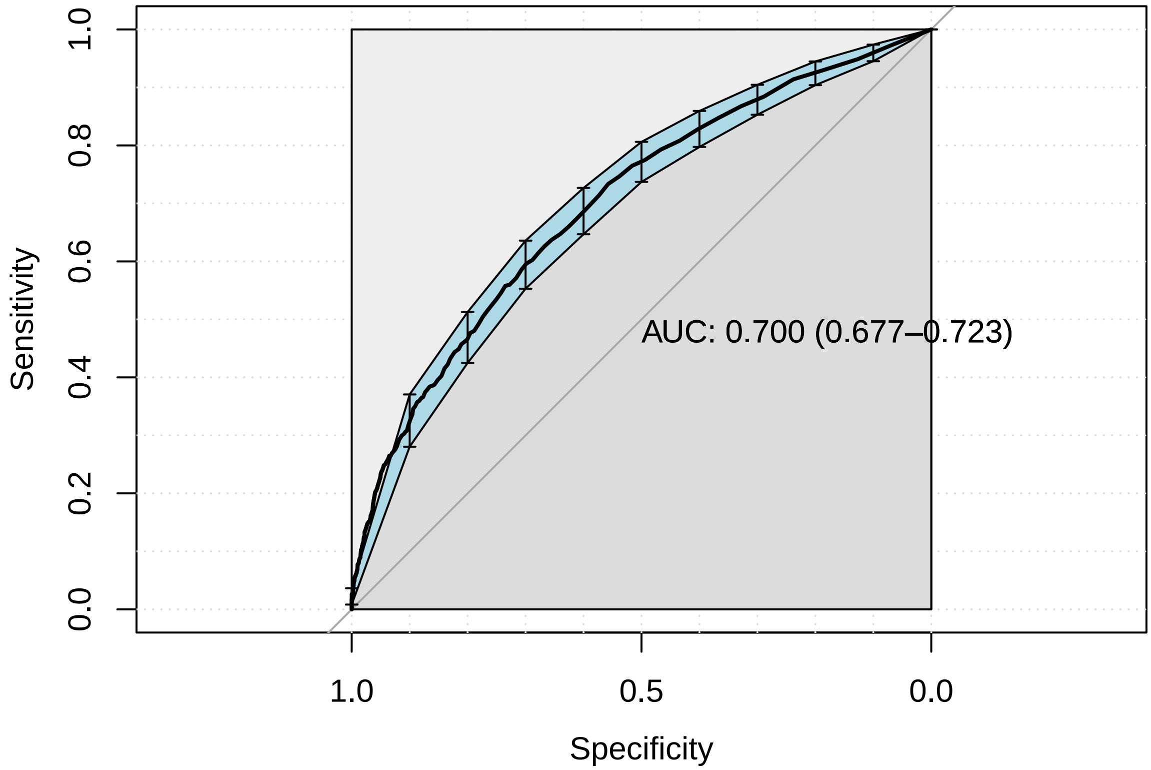
```
autoplot(rf_tune, metric = "roc_auc")
```



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

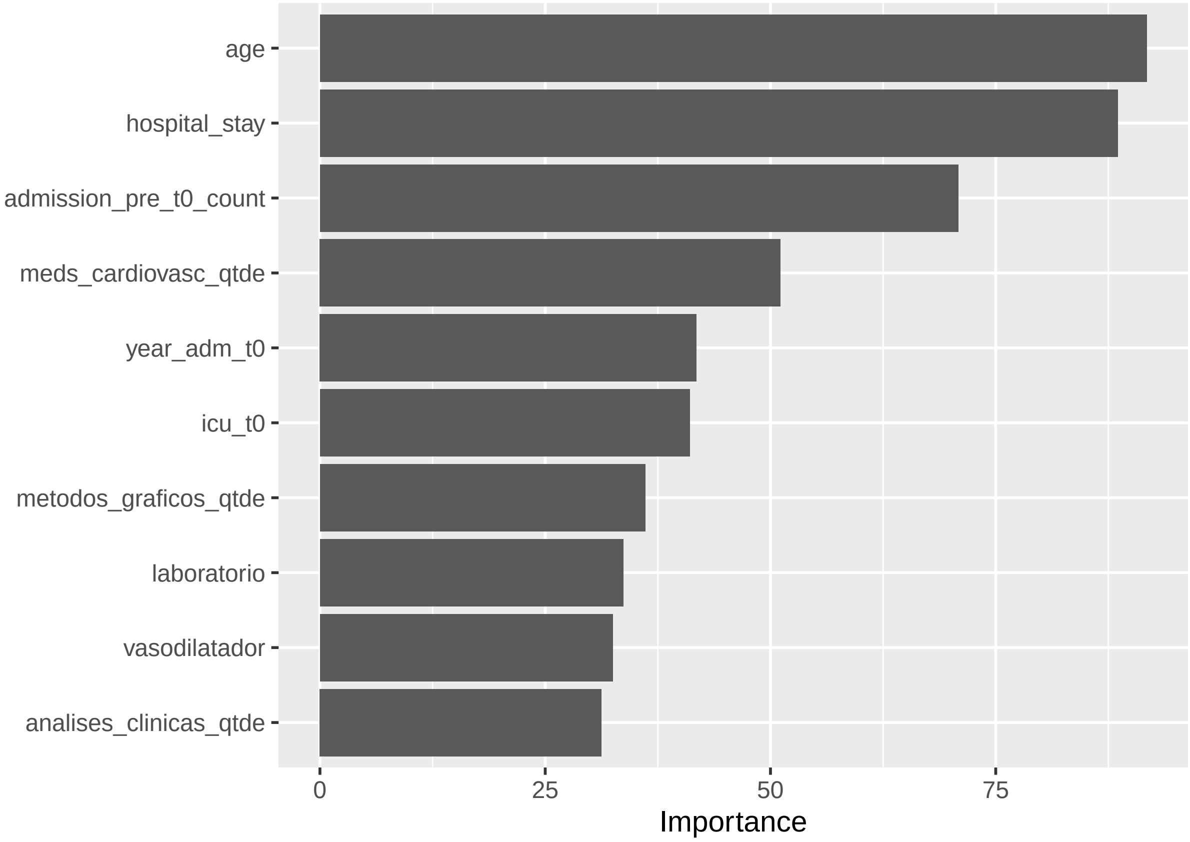
```
## # A tibble: 5 x 9
##    mtry trees min_n .metric .estimator  mean     n std_err .config
##   <int> <int> <int> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1    21   279    16 roc_auc binary     0.708    10 0.00807 Preprocessor1_Model12
## 2    17   244    14 roc_auc binary     0.706    10 0.00886 Preprocessor1_Model17
## 3    43   251    22 roc_auc binary     0.705    10 0.00842 Preprocessor1_Model05
## 4    19   293     3 roc_auc binary     0.704    10 0.00868 Preprocessor1_Model19
## 5    16   213    40 roc_auc binary     0.704    10 0.00851 Preprocessor1_Model18
```

```
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)
```



```
## [1] "Optimal Threshold: 0.08"
## Confusion Matrix and Statistics
##
##     reference
## data    0    1
##    0 2884  244
##    1 1243  360
##
##                Accuracy : 0.6857
##                  95% CI : (0.6722, 0.6989)
##     No Information Rate : 0.8723
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1728
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6988
##             Specificity : 0.5960
##          Pos Pred Value : 0.9220
##          Neg Pred Value : 0.2246
##              Prevalence : 0.8723
##          Detection Rate : 0.6096
##    Detection Prevalence : 0.6612
##       Balanced Accuracy : 0.6474
##
##        'Positive' Class : 0
##
```

```
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')
```

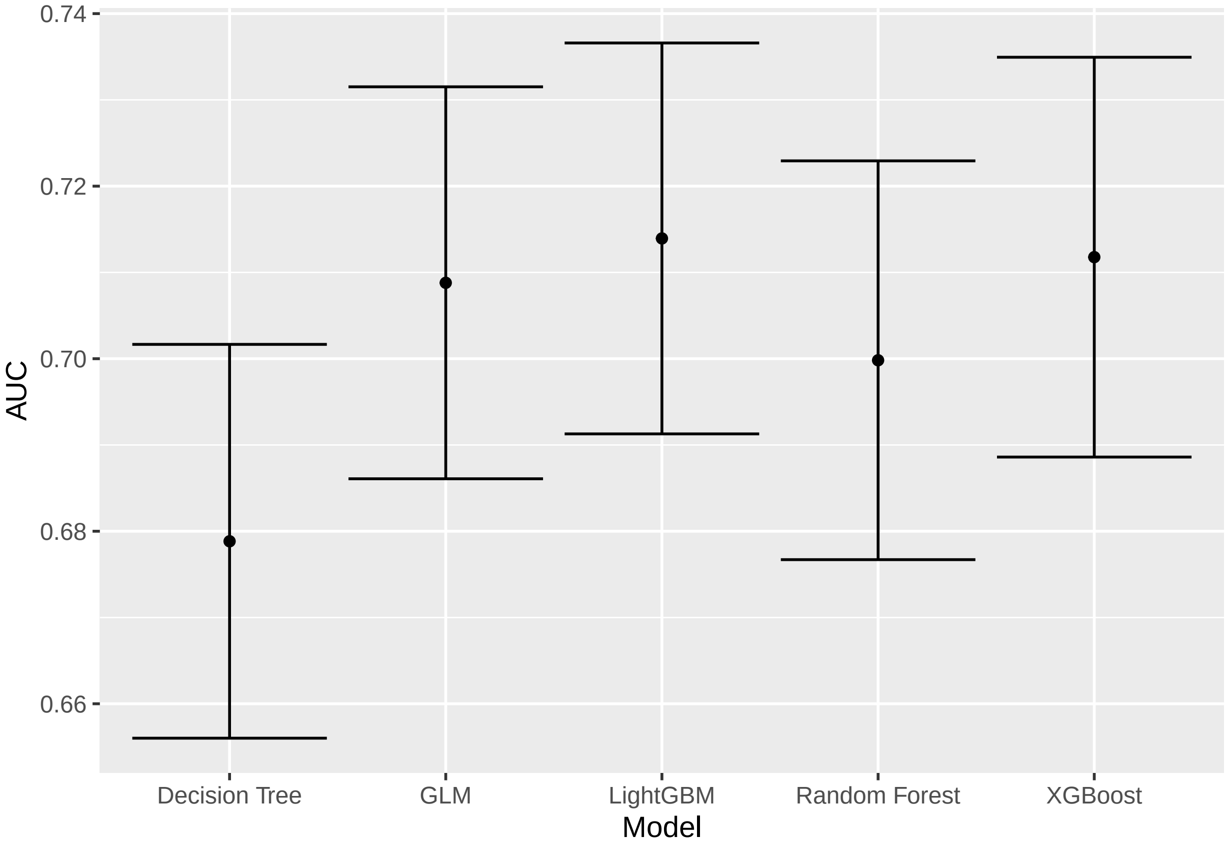


```
# extract_vip(final_rf_fit, pred_wrapper = predict,
#             reference_class = "1", use_matrix = FALSE,
#             method = 'permute')
```

Minutes to run: 67.883

Models Comparison

```
if (RUN_ALL_MODELS) {
  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$ci[3]
  ) %>%
    mutate(Target = outcome_column)
} else {
  df_auc <- tibble::tribble(
    ~Model, ~`AUC`, ~`Lower Limit`, ~`Upper Limit`,
    'LightGBM', as.numeric(lightgbm_auc$auc), lightgbm_auc$ci[1], lightgbm_auc$ci[3]
  ) %>%
    mutate(Target = outcome_column)
}

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



```
write_csv(df_auc, sprintf("./auxiliar/model_selection/performance/%s.csv", outcome_column))
```

Minutes to run: 0.002