Classifier

```r
set.seed(42)

library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(tidymodels)
```

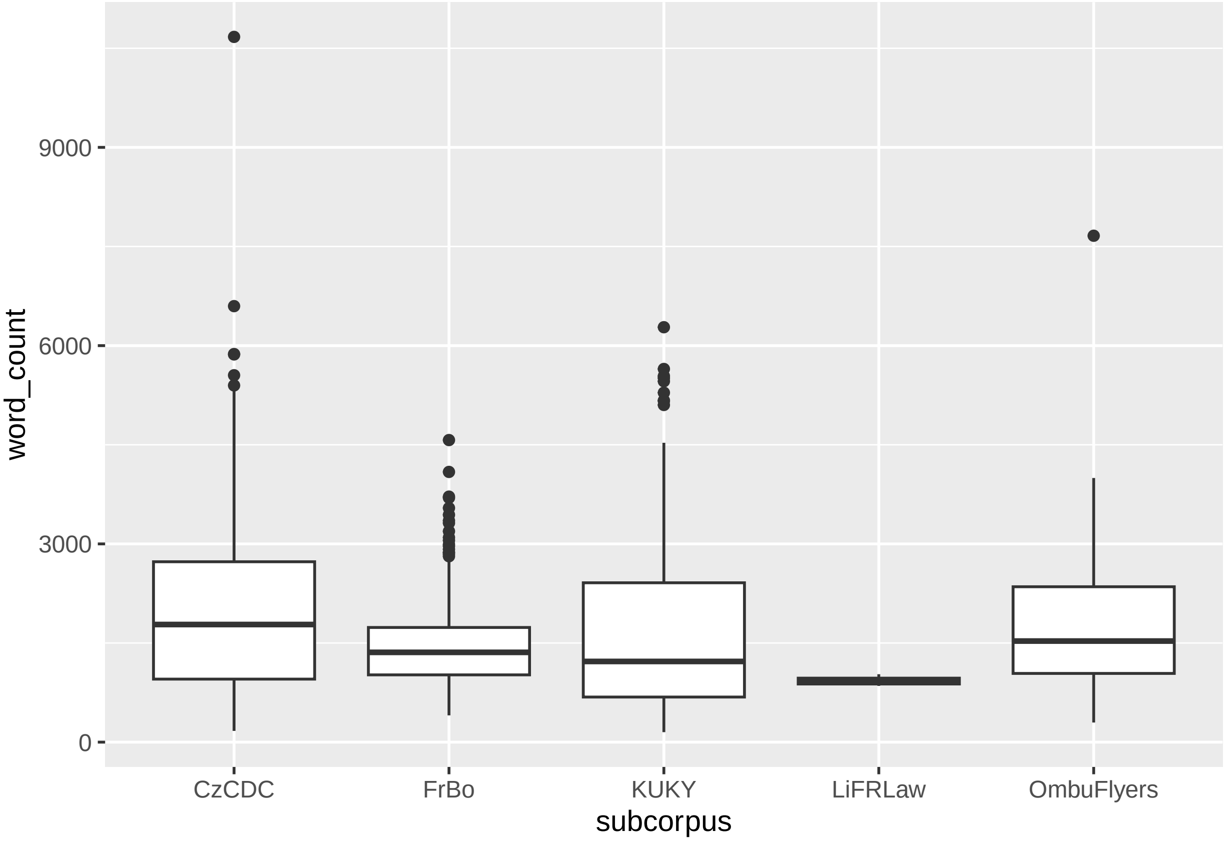
```
## -- Attaching packages -------------------------------------- tidymodels 1.2.0 --
## v broom        1.0.5     v rsample      1.2.1
## v dials        1.3.0     v tune         1.2.1
## v infer        1.0.7     v workflows    1.1.4
## v modeldata    1.4.0     v workflowsets 1.1.0
## v parsnip      1.2.1     v yardstick    1.3.2
## v recipes      1.1.0
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard()        masks purrr::discard()
## x dplyr::filter()          masks stats::filter()
## x recipes::fixed()         masks stringr::fixed()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x yardstick::precision()   masks caret::precision()
## x yardstick::recall()      masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::spec()        masks readr::spec()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()          masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

Load and tidy data

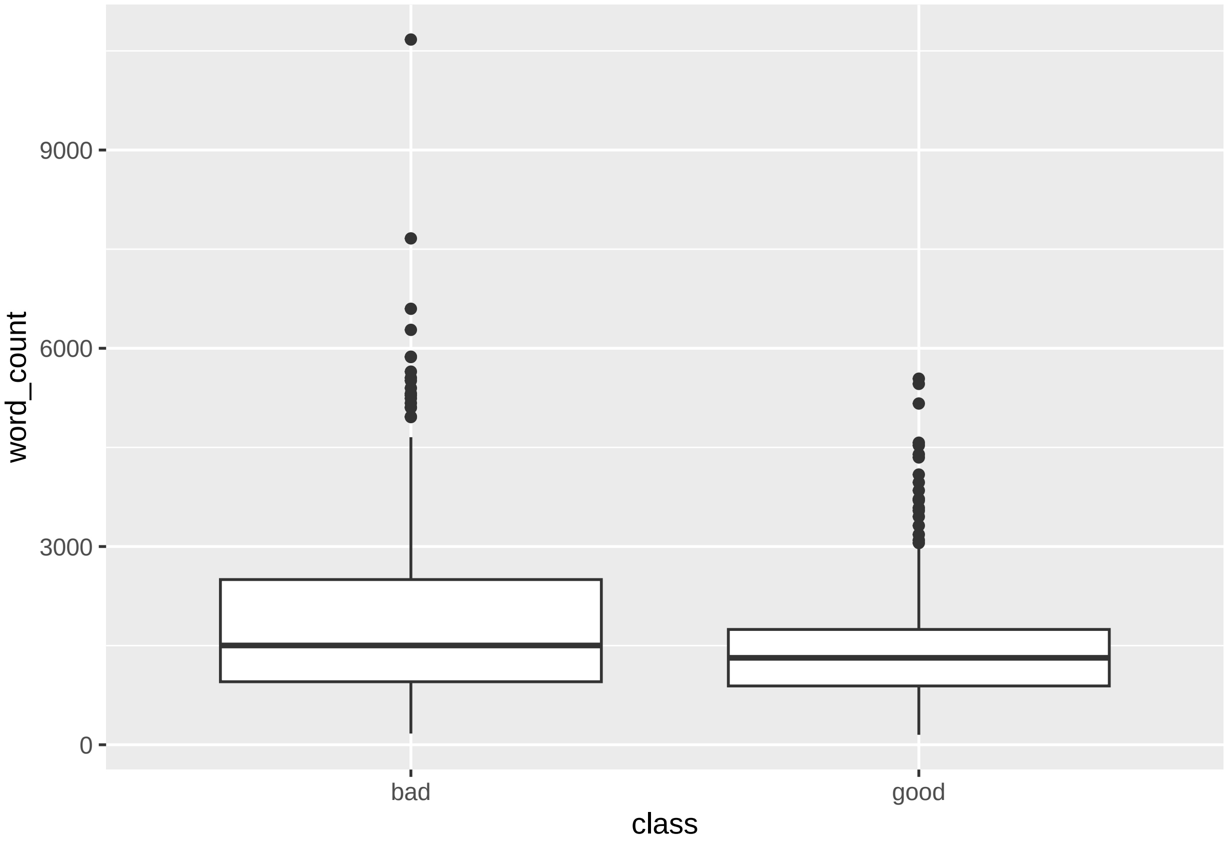
```
data <- read_csv("../measurements/measurements.csv")
```

```
## Rows: 766 Columns: 96
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (9): fpath, KUK_ID, class, FileName, FolderPath, subcorpus, DocumentTit...
## dbl (85): RuleAbstractNouns, RuleAmbiguousRegards, RuleAnaphoricReferences, ...
## lgl  (2): ClarityPursuit, SyllogismBased
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
data %>% ggplot(aes(x = subcorpus, word_count)) +
  geom_boxplot()
```



```
data %>% ggplot(aes(x = class, word_count)) +
  geom_boxplot()
```



```
data_clean <- data %>%
  select(!c(
    fpath,
    KUK_ID,
    FileName,
    FolderPath,
    # subcorpus,
    DocumentTitle,
    ClarityPursuit,
    Readability,
    SyllogismBased,
    SourceDB
  )) %>%
  # replace -1s in variation coefficients with NAs
  mutate(across(c(
    `RuleDoubleAdpos.max_allowable_distance.v`,
    `RuleTooManyNegations.max_negation_frac.v`,
    `RuleTooManyNegations.max_allowable_negations.v`,
    `RuleTooManyNominalConstructions.max_noun_frac.v`,
    `RuleTooManyNominalConstructions.max_allowable_nouns.v`,
    `RuleCaseRepetition.max_repetition_count.v`,
    `RuleCaseRepetition.max_repetition_frac.v`,
    `RulePredSubjDistance.max_distance.v`,
    `RulePredObjDistance.max_distance.v`,
    `RuleInfVerbDistance.max_distance.v`,
    `RuleMultiPartVerbs.max_distance.v`,
    `RuleLongSentences.max_length.v`,
    `RulePredAtClauseBeginning.max_order.v`,
    `mattr.v`,
    `maentropy.v`
  ), ~ na_if(.x, -1))) %>%
  # replace NAs with 0s
  replace_na(list(
    RuleGPcoordovs = 0,
    RuleGPdeverbaddr = 0,
    RuleGPpatinstr = 0,
    RuleGPdeverbsubj = 0,
    RuleGPadjective = 0,
    RuleGPpatbenperson = 0,
    RuleGPwordorder = 0,
    RuleDoubleAdpos = 0,
    RuleDoubleAdpos.max_allowable_distance = 0,
    RuleDoubleAdpos.max_allowable_distance.v = 0,
    RuleAmbiguousRegards = 0,
    RuleReflexivePassWithAnimSubj = 0,
    RuleTooManyNegations = 0,
    RuleTooManyNegations.max_negation_frac = 0,
    RuleTooManyNegations.max_negation_frac.v = 0,
    RuleTooManyNegations.max_allowable_negations = 0,
    RuleTooManyNegations.max_allowable_negations.v = 0,
    RuleTooManyNominalConstructions.max_noun_frac.v = 0,
    RuleTooManyNominalConstructions.max_allowable_nouns.v = 0,
    RuleFunctionWordRepetition = 0,
    RuleCaseRepetition.max_repetition_count.v = 0,
    RuleCaseRepetition.max_repetition_frac.v = 0,
    RuleWeakMeaningWords = 0,
    RuleAbstractNouns = 0,
    RuleRelativisticExpressions = 0,
    RuleConfirmationExpressions = 0,
    RuleRedundantExpressions = 0,
    RuleTooLongExpressions = 0,
    RuleAnaphoricReferences = 0,
    RuleLiteraryStyle = 0,
    RulePassive = 0,
    RulePredSubjDistance = 0,
    RulePredSubjDistance.max_distance = 0,
    RulePredSubjDistance.max_distance.v = 0,
    RulePredObjDistance = 0,
    RulePredObjDistance.max_distance = 0,
    RulePredObjDistance.max_distance.v = 0,
    RuleInfVerbDistance = 0,
    RuleInfVerbDistance.max_distance = 0,
    RuleInfVerbDistance.max_distance.v = 0,
    RuleMultiPartVerbs = 0,
    RuleMultiPartVerbs.max_distance = 0,
    RuleMultiPartVerbs.max_distance.v = 0,
    RuleLongSentences.max_length.v = 0,
    RulePredAtClauseBeginning.max_order.v = 0,
    RuleVerbalNouns = 0,
    RuleDoubleComparison = 0,
    RuleWrongValencyCase = 0,
    RuleWrongVerbonominalCase = 0,
    RuleIncompleteConjunction = 0
  )) %>%
  # norm data expected to correlate with text length
  mutate(across(c(
    RuleGPcoordovs,
    RuleGPdeverbaddr,
    RuleGPpatinstr,
    RuleGPdeverbsubj,
    RuleGPadjective,
    RuleGPpatbenperson,
    RuleGPwordorder,
    RuleDoubleAdpos,
    RuleAmbiguousRegards,
    RuleFunctionWordRepetition,
    RuleWeakMeaningWords,
    RuleAbstractNouns,
    RuleRelativisticExpressions,
    RuleConfirmationExpressions,
    RuleRedundantExpressions,
    RuleTooLongExpressions,
    RuleAnaphoricReferences,
    RuleLiteraryStyle,
    RulePassive,
    RuleVerbalNouns,
    RuleDoubleComparison,
    RuleWrongValencyCase,
    RuleWrongVerbonominalCase,
    RuleIncompleteConjunction,
    num_hapax,
    RuleReflexivePassWithAnimSubj,
    RuleTooManyNominalConstructions,
    RulePredSubjDistance,
    RuleMultiPartVerbs,
    RulePredAtClauseBeginning
  ), ~ .x / word_count)) %>%
  mutate(across(c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredObjDistance,
    RuleInfVerbDistance
  ), ~ .x / sent_count)) %>%
  # remove variables identified as "u counts"
  select(!c(
    RuleTooFewVerbs,
    RuleTooManyNegations,
    RuleTooManyNominalConstructions,
    RuleCaseRepetition,
    RuleLongSentences,
    RulePredAtClauseBeginning
  )) %>%
  unite("strata", c(subcorpus, class), sep = "_", remove = FALSE) %>%
  mutate(across(c(class), ~ as.factor(.x)))

# no NAs should be present now
data_clean[!complete.cases(data_clean), ]
```

```
## # A tibble: 0 x 82
## # i 82 variables: strata <chr>, class <fct>, subcorpus <chr>,
## #   RuleAbstractNouns <dbl>, RuleAmbiguousRegards <dbl>,
## #   RuleAnaphoricReferences <dbl>,
## #   RuleCaseRepetition.max_repetition_count <dbl>,
## #   RuleCaseRepetition.max_repetition_count.v <dbl>,
## #   RuleCaseRepetition.max_repetition_frac <dbl>,
## #   RuleCaseRepetition.max_repetition_frac.v <dbl>, ...
```

```
# use tidymodels::step_corr to remove high-correlating variables
```

Prepare splits and folds

```
# CHECK CONSISTENCY WITH analysis.Rmd

.split_prop <- 4 / 5 # proportion of testing data in the dataset
.no_folds <- 10 # no. of folds in v-fold cross-validation

split <- data_clean %>% initial_split(prop = .split_prop)
training_set <- training(split)
evaluation_set <- testing(split)

folds <- vfold_cv(training_set, v = .no_folds, strata = strata)

print(split)
```

```
## <Training/Testing/Total>
## <612/154/766>
```

```
print(folds)
```

```
## #  10-fold cross-validation using stratification
## # A tibble: 10 x 2
##    splits           id
##    <list>           <chr>
##  1 <split [549/63]> Fold01
##  2 <split [549/63]> Fold02
##  3 <split [549/63]> Fold03
##  4 <split [550/62]> Fold04
##  5 <split [551/61]> Fold05
##  6 <split [552/60]> Fold06
##  7 <split [552/60]> Fold07
##  8 <split [552/60]> Fold08
##  9 <split [552/60]> Fold09
## 10 <split [552/60]> Fold10
```

```r
# structure of the training set
table(training_set$subcorpus, training_set$class)
```

```
##
##              bad good
##   CzCDC      169    0
##   FrBo        57  187
##   KUKY        70   88
##   LiFRLaw      3    0
##   OmbuFlyers  38    0
```

```r
# structure of the evaluation set
table(evaluation_set$subcorpus, evaluation_set$class)
```

```
##
##              bad good
##   CzCDC       41    0
##   FrBo        22   43
##   KUKY        14   22
##   OmbuFlyers  12    0
```

Classifier helpers

Models

```r
library(vip)
```

```
##
## Attaching package: 'vip'

## The following object is masked from 'package:utils':
##
##     vi
```

```r
# decision tree libraries
library(rpart)
```

```
##
## Attaching package: 'rpart'

## The following object is masked from 'package:dials':
##
##     prune
```

```r
library(rpart.plot)
```

Null model

```r
train_null <- function(recipe, folds) {
  null_workflow <- workflow() %>% add_recipe(recipe)

  null_classification <- null_model() %>%
    set_engine("parsnip") %>%
    set_mode("classification")

  null_rs <- fit_resamples(null_workflow %>% add_model(null_classification), folds)
  cat("Null resamples:\n")
  print(null_rs)

  cat("Null metrics:\n")
  collect_metrics(null_rs) %>% print()

  return(null_rs)
}
```

Decision tree

```
train_decision_tree <- function(formula, training_set) {
  model <- rpart(formula, training_set)
  model %>% rpart.plot(type = 2, extra = 2)
  return(model)
}
```

Lasso

```
train_lasso <- function(recipe, training_set, folds) {
  lasso_tune_spec <- logistic_reg(penalty = tune(), mixture = 1) %>%
    set_mode("classification") %>%
    set_engine("glmnet")

  # cat("Lasso specification for tuning:\n")
  # print(lasso_tune_spec)

  lambda_grid <- grid_regular(penalty(), levels = 30)

  lasso_tune_wf <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(lasso_tune_spec)

  cat("Lasso tune workflow:\n")
  print(lasso_tune_wf)

  lasso_tune_rs <- tune_grid(
    lasso_tune_wf,
    folds,
    grid = lambda_grid,
    control = control_resamples(save_pred = TRUE)
  )

  # cat("Lasso tune resamples:\n")
  # print(lasso_tune_rs)

  cat("Lasso tuning metrics:\n")
  # collect_metrics(lasso_tune_rs) %>% print()
  autoplot(lasso_tune_rs) %>% print()

  lasso_tune_rs %>%
    show_best(metric = "roc_auc") %>%
    print()
  lasso_tune_rs %>%
    show_best(metric = "accuracy") %>%
    print()

  best_accuracy <- lasso_tune_rs %>%
    select_by_one_std_err(metric = "accuracy", -penalty)

  cat("Best accuracy:\n")
  print(best_accuracy)

  final_lasso <- lasso_tune_wf %>% finalize_workflow(best_accuracy)
  cat("Final workflow:\n")
  print(final_lasso)

  fitted_lasso <- fit(final_lasso, training_set)

  cat("Final coefficients:\n")
  fitted_lasso %>%
    extract_fit_parsnip() %>%
    tidy() %>%
    arrange(estimate) %>%
    print(n = 100)

  cat("Variable importance:\n")
  fitted_lasso %>%
    extract_fit_parsnip() %>%
    vi() %>%
    print(n = 100)

  return(final_lasso)
}
```

SVM

```
train_svm <- function(recipe, training_set, folds) {
  svm_spec <- svm_linear() %>%
    set_mode("classification") %>%
    set_engine("kernlab")

  svm_wf <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(svm_spec)
  cat("SVM workflow:\n")
  print(svm_wf)

  svm_rs <- fit_resamples(
    svm_wf,
    folds,
    control = control_resamples(save_pred = TRUE)
  )
  # cat("SVM resamples:\n")
  # print(svm_rs)
```

```r
  cat("SVM metrics:\n")
  collect_metrics(svm_rs) %>% print()

  svm_rs %>%
    collect_predictions() %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  print("\n")

  svm_rs %>%
    collect_predictions() %>%
    group_by(id) %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  print("\n")

  svm_rs %>%
    conf_mat_resampled(tidy = FALSE) %>%
    autoplot(type = "heatmap") %>%
    print()

  print("\n")

  final_svm <- svm_wf

  return(final_svm)
}

train_svm_rbf <- function(recipe, training_set, folds) {
  svm_spec <- svm_rbf() %>%
    set_mode("classification") %>%
    set_engine("kernlab")

  svm_wf <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(svm_spec)
  cat("SVM workflow:\n")
  print(svm_wf)

  svm_rs <- fit_resamples(
    svm_wf,
    folds,
    control = control_resamples(save_pred = TRUE)
  )
  # cat("SVM resamples:\n")
  # print(svm_rs)

  cat("SVM metrics:\n")
  collect_metrics(svm_rs) %>% print()
  svm_rs %>%
    collect_predictions() %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  print("\n")

  svm_rs %>%
    collect_predictions() %>%
    group_by(id) %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  print("\n")

  svm_rs %>%
    conf_mat_resampled(tidy = FALSE) %>%
    autoplot(type = "heatmap") %>%
    print()

  print("\n")

  final_svm <- svm_wf

  return(final_svm)
}

# not sure this works
train_svm_tune <- function(recipe, training_set, folds) {
  svm_tune_spec <- svm_linear(cost = tune()) %>%
    set_mode("classification") %>%
    set_engine("kernlab")

  cat("SVM specification for tuning:\n")
  print(svm_tune_spec)

  lambda_grid <- grid_regular(cost(), levels = 10)
  cat("SVM tuning grid:\n")
  print(lambda_grid)

  svm_tune_wf <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(svm_tune_spec)

  cat("SVM tune workflow:\n")
  print(svm_tune_wf)

  svm_tune_rs <- tune_grid(
    svm_tune_wf,
    folds,
    grid = lambda_grid,
```

```r
    control = control_resamples(save_pred = TRUE)
  )

  cat("SVM tune resamples:\n")
  print(svm_tune_rs)

  cat("SVM tuning metrics:\n")
  collect_metrics(svm_tune_rs) %>% print()
  autoplot(svm_tune_rs) %>% print()

  svm_tune_rs %>%
    show_best(metric = "roc_auc") %>%
    print()
  svm_tune_rs %>%
    show_best(metric = "accuracy") %>%
    print()

  best_accuracy <- svm_tune_rs %>%
    select_by_one_std_err(metric = "accuracy", -cost)

  cat("Best ROC AUC:\n")
  print(best_accuracy)

  final_svm <- svm_tune_wf %>% finalize_workflow(best_accuracy)

  cat("Final workflow:\n")
  print(final_svm)

  fitted_svm <- fit(final_svm, training_set)

  return(fitted_svm)
}
```

Random forest

```r
train_random_forest <- function(recipe, training_set, folds) {
  rf_spec <- rand_forest(trees = 1000) %>%
    set_mode("classification") %>%
    set_engine("ranger", importance = "impurity")

  # cat("RF specification:\n")
  # print(rf_spec)

  rf_wf <- workflow() %>%
    add_recipe(recipe) %>%
    add_model(rf_spec)

  cat("RF workflow:\n")
  print(rf_wf)

  rf_rs <- fit_resamples(
    rf_wf,
    folds,
    control = control_resamples(save_pred = TRUE)
  )
  # cat("RF resamples:\n")
  # print(rf_rs)

  cat("RF metrics:\n")
  collect_metrics(rf_rs) %>% print()


  rf_rs %>%
    collect_predictions() %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  print("\n")

  rf_rs %>%
    collect_predictions() %>%
    group_by(id) %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  print("\n")

  rf_rs %>%
    conf_mat_resampled(tidy = FALSE) %>%
    autoplot(type = "heatmap") %>%
    print()

  print("\n")

  final_rf <- rf_wf

  fitted_rf <- final_rf %>% fit(training_set)
  fitted_rf %>%
    extract_fit_parsnip() %>%
    vi() %>%
    print(n = 100)

  return(final_rf)
}
```

Recipes

```
add_corr_remove_step <- function(recipe, training_set) {
  recipe <- recipe %>% step_corr(all_numeric_predictors(), threshold = .9)

  prep <- recipe %>% prep(training = training_set)
  no <- prep %>%
    tidy() %>%
    filter(type == "corr") %>%
    pull(number)
  prep %>%
    tidy(number = no[[1]]) %>%
    print(n = 200)

  return(recipe)
}
```

All variables

```
# features excluded, because:
# - they're ucounts
# - they were selected to be excluded (unreliability or irrelevance)

formula_all <- class ~
  RuleGPcoordovs +
  RuleGPdeverbaddr +
  RuleGPpatinstr +
  RuleGPdeverbsubj +
  RuleGPadjective +
  RuleGPpatbenperson +
  RuleGPwordorder +
  RuleDoubleAdpos +
  RuleDoubleAdpos.max_allowable_distance +
  RuleDoubleAdpos.max_allowable_distance.v +
  # RuleAmbiguousRegards +
  RuleReflexivePassWithAnimSubj +
  # RuleTooFewVerbs +
  RuleTooFewVerbs.min_verb_frac +
  # RuleTooManyNegations +
  RuleTooManyNegations.max_negation_frac +
  RuleTooManyNegations.max_negation_frac.v +
  RuleTooManyNegations.max_allowable_negations +
  RuleTooManyNegations.max_allowable_negations.v +
  # RuleTooManyNominalConstructions +
  RuleTooManyNominalConstructions.max_noun_frac +
  RuleTooManyNominalConstructions.max_noun_frac.v +
  RuleTooManyNominalConstructions.max_allowable_nouns +
  RuleTooManyNominalConstructions.max_allowable_nouns +
  # RuleFunctionWordRepetition +
  # RuleCaseRepetition +
  RuleCaseRepetition.max_repetition_count +
  RuleCaseRepetition.max_repetition_count.v +
  RuleCaseRepetition.max_repetition_frac +
  RuleCaseRepetition.max_repetition_frac.v +
  RuleWeakMeaningWords +
  RuleAbstractNouns +
  RuleRelativisticExpressions +
  RuleConfirmationExpressions +
  RuleRedundantExpressions +
  RuleTooLongExpressions +
  RuleAnaphoricReferences +
  RuleLiteraryStyle +
  RulePassive +
  RulePredSubjDistance +
  RulePredSubjDistance.max_distance +
  RulePredSubjDistance.max_distance.v +
  RulePredObjDistance +
  RulePredObjDistance.max_distance +
  RulePredObjDistance.max_distance.v +
  RuleInfVerbDistance +
  RuleInfVerbDistance.max_distance +
  RuleInfVerbDistance.max_distance.v +
  RuleMultiPartVerbs +
  RuleMultiPartVerbs.max_distance +
  RuleMultiPartVerbs.max_distance.v +
  # RuleLongSentences +
  RuleLongSentences.max_length +
  RuleLongSentences.max_length.v +
  # RulePredAtClauseBeginning +
  RulePredAtClauseBeginning.max_order +
  RulePredAtClauseBeginning.max_order.v +
  RuleVerbalNouns +
  # RuleDoubleComparison +
  # RuleWrongValencyCase +
  # RuleWrongVerbonominalCase +
  # RuleIncompleteConjunction +
  sent_count +
  word_count +
  syllab_count +
  char_count +
  cli +
  ari +
  num_hapax +
  entropy +
  ttr +
  mattr +
  mattr.v +
  maentropy +
  maentropy.v +
  mamr +
  verb_dist +
  activity +
  hpoint +
  atl +
  fre +
  fkgl +
  gf +
  smog

recipe_all_base <- recipe(
  formula_all,
  data = training_set
)

# without the removal of correlating variables
```

```r
recipe_all_nocorr <- recipe_all_base %>%
  step_normalize(all_numeric_predictors())
recipe_all_nocorr
```

```
##

## -- Recipe ----------------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:    1
## predictor: 71

##

## -- Operations

## * Centering and scaling for: all_numeric_predictors()
```

```r
# with the removal of correlating variables
recipe_all <- recipe_all_nocorr %>%
  add_corr_remove_step(training_set = training_set)
```

```
## # A tibble: 10 x 2
##    terms                                    id
##    <chr>                                    <chr>
##  1 RuleCaseRepetition.max_repetition_frac.v corr_VT4kj
##  2 char_count                               corr_VT4kj
##  3 ari                                      corr_VT4kj
##  4 ttr                                      corr_VT4kj
##  5 maentropy                                corr_VT4kj
##  6 hpoint                                   corr_VT4kj
##  7 atl                                      corr_VT4kj
##  8 gf                                       corr_VT4kj
##  9 smog                                     corr_VT4kj
## 10 word_count                               corr_VT4kj
```

```r
recipe_all
```

```
##

## -- Recipe ----------------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:    1
## predictor: 71

##

## -- Operations

## * Centering and scaling for: all_numeric_predictors()

## * Correlation filter on: all_numeric_predictors()
```

No text length

```r
# features excluded, because:
# - they're ucounts
# - they were selected to be excluded (unreliability or irrelevance)

formula_notl <- class ~
  RuleGPcoordovs +
  RuleGPdeverbaddr +
  RuleGPpatinstr +
  RuleGPdeverbsubj +
  RuleGPadjective +
  RuleGPpatbenperson +
  RuleGPwordorder +
  RuleDoubleAdpos +
  RuleDoubleAdpos.max_allowable_distance +
  RuleDoubleAdpos.max_allowable_distance.v +
  # RuleAmbiguousRegards +
  RuleReflexivePassWithAnimSubj +
  # RuleTooFewVerbs +
  RuleTooFewVerbs.min_verb_frac +
  # RuleTooManyNegations +
  RuleTooManyNegations.max_negation_frac +
  RuleTooManyNegations.max_negation_frac.v +
  RuleTooManyNegations.max_allowable_negations +
  RuleTooManyNegations.max_allowable_negations.v +
  # RuleTooManyNominalConstructions +
  RuleTooManyNominalConstructions.max_noun_frac +
  RuleTooManyNominalConstructions.max_noun_frac.v +
  RuleTooManyNominalConstructions.max_allowable_nouns +
  RuleTooManyNominalConstructions.max_allowable_nouns +
  # RuleFunctionWordRepetition +
  # RuleCaseRepetition +
  RuleCaseRepetition.max_repetition_count +
  RuleCaseRepetition.max_repetition_count.v +
  RuleCaseRepetition.max_repetition_frac +
  RuleCaseRepetition.max_repetition_frac.v +
  RuleWeakMeaningWords +
  RuleAbstractNouns +
  RuleRelativisticExpressions +
  RuleConfirmationExpressions +
  RuleRedundantExpressions +
  RuleTooLongExpressions +
  RuleAnaphoricReferences +
  RuleLiteraryStyle +
  RulePassive +
  RulePredSubjDistance +
  RulePredSubjDistance.max_distance +
  RulePredSubjDistance.max_distance.v +
  RulePredObjDistance +
  RulePredObjDistance.max_distance +
  RulePredObjDistance.max_distance.v +
  RuleInfVerbDistance +
  RuleInfVerbDistance.max_distance +
  RuleInfVerbDistance.max_distance.v +
  RuleMultiPartVerbs +
  RuleMultiPartVerbs.max_distance +
  RuleMultiPartVerbs.max_distance.v +
  # RuleLongSentences +
  RuleLongSentences.max_length +
  RuleLongSentences.max_length.v +
  # RulePredAtClauseBeginning +
  RulePredAtClauseBeginning.max_order +
  RulePredAtClauseBeginning.max_order.v +
  RuleVerbalNouns +
  # RuleDoubleComparison +
  # RuleWrongValencyCase +
  # RuleWrongVerbonominalCase +
  # RuleIncompleteConjunction +
  # sent_count +
  # word_count +
  # syllab_count +
  # char_count +
  cli +
  ari +
  num_hapax +
  entropy +
  ttr +
  mattr +
  mattr.v +
  maentropy +
  maentropy.v +
  mamr +
  verb_dist +
  activity +
  hpoint +
  atl +
  fre +
  fkgl +
  gf +
  smog

recipe_notl_base <- recipe(
  formula_notl,
  data = training_set
)

# without the removal of correlating variables
recipe_notl_nocorr <- recipe_notl_base %>%
  step_normalize(all_numeric_predictors())
recipe_notl_nocorr
```

```
##
## -- Recipe ----------------------------------------------------------------------
##
## -- Inputs
## Number of variables by role

## outcome:    1
## predictor: 67

##

## -- Operations

## * Centering and scaling for: all_numeric_predictors()
```

Counts

```r
# features excluded, because:
# - they were selected to be excluded

formula_counts <- class ~
  RuleGPcoordovs +
  RuleGPdeverbaddr +
  RuleGPpatinstr +
  RuleGPdeverbsubj +
  RuleGPadjective +
  RuleGPpatbenperson +
  RuleGPwordorder +
  RuleDoubleAdpos +
  # RuleAmbiguousRegards +
  RuleReflexivePassWithAnimSubj +
  # RuleFunctionWordRepetition +
  RuleWeakMeaningWords +
  RuleAbstractNouns +
  RuleRelativisticExpressions +
  RuleConfirmationExpressions +
  RuleRedundantExpressions +
  RuleTooLongExpressions +
  RuleAnaphoricReferences +
  RuleLiteraryStyle +
  RulePassive +
  RulePredSubjDistance +
  RulePredObjDistance +
  RuleInfVerbDistance +
  RuleMultiPartVerbs +
  RuleVerbalNouns +
  # RuleDoubleComparison +
  # RuleWrongValencyCase +
  # RuleWrongVerbonominalCase +
  # RuleIncompleteConjunction +
  sent_count +
  word_count +
  syllab_count +
  char_count +
  num_hapax

recipe_counts_base <- recipe(formula_counts, data = training_set)

recipe_counts_nocorr <- recipe_counts_base %>%
  step_normalize()
recipe_counts_nocorr
```

```
##

## -- Recipe ----------------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:    1
## predictor: 28

##

## -- Operations

## * Centering and scaling for: <none>
```

```
recipe_counts <- recipe_counts_nocorr %>%
  add_corr_remove_step(training_set = training_set)
```

```
## # A tibble: 2 x 2
##   terms        id
##   <chr>        <chr>
## 1 syllab_count corr_Fw2K3
## 2 word_count   corr_Fw2K3
```

```
recipe_counts
```

```
##

## -- Recipe ----------------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:    1
## predictor: 28

##

## -- Operations

## * Centering and scaling for: <none>

## * Correlation filter on: all_numeric_predictors()
```

Indicators, averages, and coefficients

```
formula_iac <- class ~
  RuleDoubleAdpos.max_allowable_distance +
  RuleDoubleAdpos.max_allowable_distance.v +
  RuleTooFewVerbs.min_verb_frac +
  RuleTooManyNegations.max_negation_frac +
  RuleTooManyNegations.max_negation_frac.v +
  RuleTooManyNegations.max_allowable_negations +
  RuleTooManyNegations.max_allowable_negations.v +
  RuleTooManyNominalConstructions.max_noun_frac +
  RuleTooManyNominalConstructions.max_noun_frac.v +
  RuleTooManyNominalConstructions.max_allowable_nouns +
  RuleTooManyNominalConstructions.max_allowable_nouns.v +
  RuleCaseRepetition.max_repetition_count +
  RuleCaseRepetition.max_repetition_count.v +
  RuleCaseRepetition.max_repetition_frac +
  RuleCaseRepetition.max_repetition_frac.v +
  RulePredSubjDistance.max_distance +
  RulePredSubjDistance.max_distance.v +
  RulePredObjDistance.max_distance +
  RulePredObjDistance.max_distance.v +
  RuleInfVerbDistance.max_distance +
  RuleInfVerbDistance.max_distance.v +
  RuleMultiPartVerbs.max_distance +
  RuleMultiPartVerbs.max_distance.v +
  RuleLongSentences.max_length +
  RuleLongSentences.max_length.v +
  RulePredAtClauseBeginning.max_order +
  RulePredAtClauseBeginning.max_order.v +
  cli +
  ari +
  entropy +
  ttr +
  mattr +
  mattr.v +
  maentropy +
  maentropy.v +
  mamr +
  verb_dist +
  activity +
  hpoint +
  atl +
  fre +
  fkgl +
  gf +
  smog

recipe_iac_base <- recipe(formula_iac, data = training_set)

recipe_iac_nocorr <- recipe_iac_base %>%
  step_normalize()
recipe_iac_nocorr
```

```
##
## -- Recipe ---------------------------------------------------------------------
##
## -- Inputs
## Number of variables by role
## outcome:    1
## predictor: 44

##

## -- Operations

## * Centering and scaling for: <none>
```

```
recipe_iac <- recipe_iac_nocorr %>%
  add_corr_remove_step(training_set = training_set)
```

```
## # A tibble: 7 x 2
##   terms                                    id
##   <chr>                                    <chr>
## 1 RuleCaseRepetition.max_repetition_frac.v corr_fD0q0
## 2 ari                                      corr_fD0q0
## 3 maentropy                                corr_fD0q0
## 4 atl                                      corr_fD0q0
## 5 gf                                       corr_fD0q0
## 6 smog                                     corr_fD0q0
## 7 RuleLongSentences.max_length             corr_fD0q0
```

```
recipe_iac
```

```
##

## -- Recipe ----------------------------------------------------------------------

##

## -- Inputs

## Number of variables by role

## outcome:    1
## predictor: 44

##

## -- Operations

## * Centering and scaling for: <none>

## * Correlation filter on: all_numeric_predictors()
```

Evaluation

Decision tree

```
evaluate_decision_tree <- function(model, evaluation_set) {
  test_predictions <- predict(model, evaluation_set, type = "class")
  # cm <- table(evaluation_set$conti_de, test_predictions)

  cm <- confusionMatrix(
    data = test_predictions,
    reference = evaluation_set$class,
    positive = "good"
  )
  print(cm)
}
```

Tidymodels

```r
get_vi <- function(final_fit) {
  model_class <- final_fit %>%
    extract_fit_engine() %>%
    class()
  if ("glmnet" %in% model_class) {
    return(final_fit$.workflow[[1]] %>%
      extract_fit_parsnip() %>%
      vi(lambda = final_fit %>%
        extract_fit_parsnip() %>%
        tidy() %>%
        pull(penalty)))
  } else if ("ranger" %in% model_class) {
    return(
      final_fit$.workflow[[1]] %>%
        extract_fit_parsnip() %>%
        vi()
    )
  }
}

evaluate_tidymodel <- function(final_wf, split) {
  final_fitted <- last_fit(final_wf, split)

  metrics <- collect_metrics(final_fitted)
  print(metrics)

  predictions <- collect_predictions(final_fitted)
  predictions %>%
    conf_mat(truth = class, estimate = .pred_class) %>%
    autoplot(type = "heatmap") %>%
    print()
  predictions %>%
    roc_curve(truth = class, .pred_bad) %>%
    autoplot() %>%
    print()

  cat("Variable importance:\n")
  get_vi(final_fitted) %>% print(n = 100)

  return(final_fitted)
}

lasso_get_coefficients <- function(final_lasso_wf) {
  return(
    final_lasso_wf %>%
      extract_fit_parsnip() %>%
      tidy() %>%
      arrange(estimate)
  )
}

get_mismatch_details <- function(lfit, data_orig) {
  joined <- data_orig %>%
    select(KUK_ID, FileName, Readability, ClarityPursuit, subcorpus) %>%
    rowid_to_column(".row") %>%
    right_join(lfit$.predictions[[1]] %>% select(!.config), by = ".row")

  print(
    joined %>% ggplot(aes(x = .pred_good, y = class, color = subcorpus)) +
      geom_jitter(height = 0.2, width = 0)
  )

  cat("Confusion matrices by subcorpora:\n")
  joined %>%
    select(.pred_class, class, subcorpus) %>%
    table() %>%
    print()

  cat("\n")
  cat("Greatest deviations:\n")
  joined %>%
    filter(.pred_class != class) %>%
    mutate(deviation = .pred_good - 0.5) %>%
    mutate(abs_deviation = abs(deviation)) %>%
    arrange(-abs_deviation) %>%
    select(abs_deviation, .pred_class, class, subcorpus, FileName) %>%
    print(n = round(nrow(joined) / 5))
}
```

Null model

All variables

Remove correlating

```
train_null(recipe_all, folds)
```

```
## Null resamples:
## # Resampling results
## # 10-fold cross-validation using stratification
## # A tibble: 10 x 4
##    splits           id     .metrics         .notes
##    <list>           <chr>  <list>           <list>
##  1 <split [549/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [549/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [549/63]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [550/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [551/61]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [552/60]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [552/60]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [552/60]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [552/60]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [552/60]> Fold10 <tibble [3 x 4]> <tibble [0 x 3]>
## Null metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.550    10 0.0134  Preprocessor1_Model1
## 2 brier_class binary     0.248    10 0.00137 Preprocessor1_Model1
## 3 roc_auc     binary     0.5      10 0       Preprocessor1_Model1
```

```
## # Resampling results
## # 10-fold cross-validation using stratification
## # A tibble: 10 x 4
##    splits           id     .metrics         .notes
##    <list>           <chr>  <list>           <list>
##  1 <split [549/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [549/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [549/63]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [550/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [551/61]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [552/60]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [552/60]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [552/60]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [552/60]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [552/60]> Fold10 <tibble [3 x 4]> <tibble [0 x 3]>
```

Keep correlating

```
train_null(recipe_all_nocorr, folds)
```

```
## Null resamples:
## # Resampling results
## # 10-fold cross-validation using stratification
## # A tibble: 10 x 4
##    splits           id     .metrics         .notes
##    <list>           <chr>  <list>           <list>
##  1 <split [549/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [549/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [549/63]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [550/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [551/61]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [552/60]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [552/60]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [552/60]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [552/60]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [552/60]> Fold10 <tibble [3 x 4]> <tibble [0 x 3]>
## Null metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.550    10 0.0134  Preprocessor1_Model1
## 2 brier_class binary     0.248    10 0.00137 Preprocessor1_Model1
## 3 roc_auc     binary     0.5      10 0       Preprocessor1_Model1
```

```
## # Resampling results
## # 10-fold cross-validation using stratification
## # A tibble: 10 x 4
##    splits           id     .metrics         .notes
##    <list>           <chr>  <list>           <list>
##  1 <split [549/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [549/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [549/63]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [550/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [551/61]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [552/60]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [552/60]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [552/60]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [552/60]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [552/60]> Fold10 <tibble [3 x 4]> <tibble [0 x 3]>
```

Regular logistic regression

```
training_set_modif <- training_set %>%
  mutate(across(class, ~ .x == "good")) %>%
  mutate(across(RuleAbstractNouns:word_count, ~ scale(.x)))
```

All variables

```
glm(
  formula_all,
  data = training_set_modif,
  family = binomial(link = "logit")
) %>% summary()
```

```
##
## Call:
## glm(formula = formula_all, family = binomial(link = "logit"),
##     data = training_set_modif)
##
## Coefficients: (1 not defined because of singularities)
##                                                        Estimate Std. Error
## (Intercept)                                          -5.815e-01  1.671e-01
## RuleGPcoordovs                                       -5.074e-02  1.260e-01
## RuleGPdeverbaddr                                     -2.489e-01  1.320e-01
## RuleGPpatinstr                                       -1.270e-01  1.316e-01
## RuleGPdeverbsubj                                     -1.933e-01  1.148e-01
## RuleGPadjective                                       3.952e-01  2.386e-01
## RuleGPpatbenperson                                   -1.703e-01  1.295e-01
## RuleGPwordorder                                      -1.446e-01  1.550e-01
## RuleDoubleAdpos                                       6.323e-02  1.617e-01
## RuleDoubleAdpos.max_allowable_distance               -2.776e-02  2.707e-01
## RuleDoubleAdpos.max_allowable_distance.v              1.041e-01  2.222e-01
## RuleReflexivePassWithAnimSubj                        -8.326e-02  1.423e-01
## RuleTooFewVerbs.min_verb_frac                        -1.797e+00  5.367e-01
## RuleTooManyNegations.max_negation_frac                1.358e-01  2.071e-01
## RuleTooManyNegations.max_negation_frac.v             -4.608e-02  1.559e-01
## RuleTooManyNegations.max_allowable_negations          2.424e-01  2.638e-01
## RuleTooManyNegations.max_allowable_negations.v       -1.448e-01  2.330e-01
## RuleTooManyNominalConstructions.max_noun_frac        -3.317e-01  2.176e-01
## RuleTooManyNominalConstructions.max_noun_frac.v       7.527e-02  1.634e-01
## RuleTooManyNominalConstructions.max_allowable_nouns  3.154e-01  5.022e-01
## RuleCaseRepetition.max_repetition_count              -2.595e-01  3.832e-01
## RuleCaseRepetition.max_repetition_count.v           -2.389e-01  1.916e-01
## RuleCaseRepetition.max_repetition_frac               8.332e-01  1.099e+00
## RuleCaseRepetition.max_repetition_frac.v             1.219e+00  1.079e+00
## RuleWeakMeaningWords                                -1.196e-01  1.351e-01
## RuleAbstractNouns                                    1.056e-01  1.366e-01
## RuleRelativisticExpressions                         -2.598e-01  1.369e-01
## RuleConfirmationExpressions                          1.833e-01  1.570e-01
## RuleRedundantExpressions                            -1.947e-01  1.623e-01
## RuleTooLongExpressions                               2.882e-01  1.552e-01
## RuleAnaphoricReferences                              5.204e-01  1.548e-01
## RuleLiteraryStyle                                   -4.104e-01  1.616e-01
## RulePassive                                         -4.972e-01  2.051e-01
## RulePredSubjDistance                                 4.758e-01  2.172e-01
## RulePredSubjDistance.max_distance                   -5.392e-01  2.923e-01
## RulePredSubjDistance.max_distance.v                 -6.081e-02  2.127e-01
## RulePredObjDistance                                  2.251e-04  2.551e-01
## RulePredObjDistance.max_distance                    -3.251e-01  2.803e-01
## RulePredObjDistance.max_distance.v                   3.876e-02  1.916e-01
## RuleInfVerbDistance                                  1.657e-01  2.624e-01
## RuleInfVerbDistance.max_distance                     3.270e-01  1.385e-01
## RuleInfVerbDistance.max_distance.v                  -2.439e-01  1.855e-01
## RuleMultiPartVerbs                                   5.539e-01  2.528e-01
## RuleMultiPartVerbs.max_distance                      8.468e-02  2.252e-01
## RuleMultiPartVerbs.max_distance.v                    1.599e-01  2.190e-01
## RuleLongSentences.max_length                         3.448e+00  9.828e-01
## RuleLongSentences.max_length.v                       8.485e-01  2.205e-01
## RulePredAtClauseBeginning.max_order                 -2.599e-01  3.283e-01
## RulePredAtClauseBeginning.max_order.v                2.779e-02  2.618e-01
## RuleVerbalNouns                                     -6.928e-02  1.587e-01
## sent_count                                           1.298e+00  7.708e-01
## word_count                                          -5.628e+00  3.832e+00
## syllab_count                                        -1.337e+01  6.339e+00
## char_count                                           1.854e+01  8.225e+00
## cli                                                 -8.734e-01  2.335e+00
## ari                                                 -5.628e+00  1.956e+00
## num_hapax                                            5.712e-01  9.716e-01
## entropy                                             -6.519e-01  3.855e-01
## ttr                                                 -1.092e+00  1.293e+00
## mattr                                               -1.207e+00  1.121e+00
## mattr.v                                             -4.288e-01  4.514e-01
## maentropy                                            9.184e-01  1.166e+00
## maentropy.v                                          9.324e-01  6.971e-01
## mamr                                                -1.154e-01  2.997e-01
## verb_dist                                            3.170e-01  3.314e-01
## activity                                             1.668e+00  5.612e-01
## hpoint                                              -1.182e+00  8.745e-01
## atl                                                  8.325e-01  2.690e+00
## fre                                                 -2.980e+00  1.045e+00
## fkgl                                                        NA         NA
## gf                                                  -2.400e+00  2.475e+00
## smog                                                 1.635e+00  2.006e+00
##                                                     z value Pr(>|z|)
## (Intercept)                                           -3.479 0.000503 ***
## RuleGPcoordovs                                        -0.403 0.687185
## RuleGPdeverbaddr                                     -1.885 0.059432 .
## RuleGPpatinstr                                       -0.965 0.334677
## RuleGPdeverbsubj                                     -1.683 0.092298 .
## RuleGPadjective                                       1.656 0.097703 .
## RuleGPpatbenperson                                   -1.315 0.188646
## RuleGPwordorder                                      -0.933 0.350771
## RuleDoubleAdpos                                       0.391 0.695761
## RuleDoubleAdpos.max_allowable_distance               -0.103 0.918321
## RuleDoubleAdpos.max_allowable_distance.v              0.469 0.639328
## RuleReflexivePassWithAnimSubj                        -0.585 0.558582
## RuleTooFewVerbs.min_verb_frac                        -3.348 0.000814 ***
## RuleTooManyNegations.max_negation_frac                0.656 0.512087
## RuleTooManyNegations.max_negation_frac.v             -0.296 0.767594
## RuleTooManyNegations.max_allowable_negations          0.919 0.358160
## RuleTooManyNegations.max_allowable_negations.v       -0.621 0.534471
## RuleTooManyNominalConstructions.max_noun_frac        -1.525 0.127325
## RuleTooManyNominalConstructions.max_noun_frac.v       0.461 0.644988
## RuleTooManyNominalConstructions.max_allowable_nouns   0.628 0.530051
## RuleCaseRepetition.max_repetition_count              -0.677 0.498276
## RuleCaseRepetition.max_repetition_count.v            -1.247 0.212388
## RuleCaseRepetition.max_repetition_frac                0.758 0.448318
## RuleCaseRepetition.max_repetition_frac.v              1.129 0.258693
## RuleWeakMeaningWords                                 -0.885 0.376126
## RuleAbstractNouns                                     0.773 0.439470
## RuleRelativisticExpressions                          -1.898 0.057734 .
## RuleConfirmationExpressions                           1.167 0.243117
## RuleRedundantExpressions                             -1.199 0.230455
## RuleTooLongExpressions                                1.857 0.063326 .
## RuleAnaphoricReferences                               3.362 0.000775 ***
## RuleLiteraryStyle                                    -2.540 0.011083 *
## RulePassive                                          -2.424 0.015345 *
## RulePredSubjDistance                                  2.191 0.028487 *
## RulePredSubjDistance.max_distance                    -1.845 0.065042 .
## RulePredSubjDistance.max_distance.v                  -0.286 0.774961
## RulePredObjDistance                                   0.001 0.999296
## RulePredObjDistance.max_distance                     -1.160 0.246052
## RulePredObjDistance.max_distance.v                    0.202 0.839646
## RuleInfVerbDistance                                   0.631 0.527832
## RuleInfVerbDistance.max_distance                      2.361 0.018208 *
## RuleInfVerbDistance.max_distance.v                   -1.315 0.188458
## RuleMultiPartVerbs                                    2.191 0.028448 *
## RuleMultiPartVerbs.max_distance                       0.376 0.706919
## RuleMultiPartVerbs.max_distance.v                     0.730 0.465362
## RuleLongSentences.max_length                          3.508 0.000451 ***
## RuleLongSentences.max_length.v                        3.848 0.000119 ***
## RulePredAtClauseBeginning.max_order                  -0.792 0.428556
## RulePredAtClauseBeginning.max_order.v                 0.106 0.915457
## RuleVerbalNouns                                      -0.437 0.662408
## sent_count                                            1.684 0.092098 .
## word_count                                           -1.469 0.141952
## syllab_count                                         -2.110 0.034877 *
## char_count                                            2.255 0.024155 *
## cli                                                  -0.374 0.708383
## ari                                                  -2.877 0.004012 **
## num_hapax                                             0.588 0.556610
## entropy                                              -1.691 0.090784 .
## ttr                                                  -0.845 0.398068
## mattr                                                -1.077 0.281681
## mattr.v                                              -0.950 0.342143
## maentropy                                             0.788 0.430877
## maentropy.v                                           1.338 0.181024
## mamr                                                 -0.385 0.700324
## verb_dist                                             0.957 0.338746
## activity                                              2.972 0.002957 **
## hpoint                                               -1.351 0.176635
## atl                                                   0.309 0.756963
## fre                                                  -2.853 0.004337 **
## fkgl                                                     NA       NA
## gf                                                   -0.970 0.332153
## smog                                                  0.815 0.415107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 842.12  on 611  degrees of freedom
## Residual deviance: 424.47  on 541  degrees of freedom
## AIC: 566.47
##
## Number of Fisher Scoring iterations: 6
```

Indicators, averages, and coefficients

```
glm(
  formula_iac,
  data = training_set_modif,
  family = binomial(link = "logit")
) %>% summary()
```

```
##
## Call:
## glm(formula = formula_iac, family = binomial(link = "logit"),
##     data = training_set_modif)
##
## Coefficients: (1 not defined because of singularities)
##                                                          Estimate Std. Error
## (Intercept)                                             -0.452532   0.134377
## RuleDoubleAdpos.max_allowable_distance                   0.153689   0.192495
## RuleDoubleAdpos.max_allowable_distance.v                -0.114459   0.167523
## RuleTooFewVerbs.min_verb_frac                           -1.539441   0.426885
## RuleTooManyNegations.max_negation_frac                   0.040402   0.178987
## RuleTooManyNegations.max_negation_frac.v                 0.063467   0.130559
## RuleTooManyNegations.max_allowable_negations             0.096269   0.236561
## RuleTooManyNegations.max_allowable_negations.v          -0.198630   0.201009
## RuleTooManyNominalConstructions.max_noun_frac           -0.351172   0.178675
## RuleTooManyNominalConstructions.max_noun_frac.v          0.139525   0.137715
## RuleTooManyNominalConstructions.max_allowable_nouns      0.219309   0.413569
## RuleTooManyNominalConstructions.max_allowable_nouns.v -0.218766   0.189946
## RuleCaseRepetition.max_repetition_count                0.053659   0.302008
## RuleCaseRepetition.max_repetition_count.v             -0.325508   0.169448
## RuleCaseRepetition.max_repetition_frac                 0.458775   0.922474
## RuleCaseRepetition.max_repetition_frac.v               0.718221   0.906236
## RulePredSubjDistance.max_distance                     -0.562731   0.275941
## RulePredSubjDistance.max_distance.v                    0.037959   0.179267
## RulePredObjDistance.max_distance                      -0.259888   0.245379
## RulePredObjDistance.max_distance.v                     0.005293   0.164510
## RuleInfVerbDistance.max_distance                       0.214965   0.118217
## RuleInfVerbDistance.max_distance.v                    -0.374875   0.150446
## RuleMultiPartVerbs.max_distance                        0.151781   0.208376
## RuleMultiPartVerbs.max_distance.v                      0.173853   0.185069
## RuleLongSentences.max_length                           3.111818   0.890676
## RuleLongSentences.max_length.v                         0.624271   0.181781
## RulePredAtClauseBeginning.max_order                   -0.101123   0.359959
## RulePredAtClauseBeginning.max_order.v                 -0.125394   0.217829
## cli                                                   -0.797606   1.761512
## ari                                                   -4.234860   1.336233
## entropy                                               -0.167785   0.307403
## ttr                                                   -0.393476   0.326889
## mattr                                                 -0.891455   0.870774
## mattr.v                                               -0.575654   0.399181
## maentropy                                              0.599774   0.885082
## maentropy.v                                            1.133037   0.631452
## mamr                                                   0.029908   0.228002
## verb_dist                                              0.439288   0.270594
## activity                                               1.977103   0.398249
## hpoint                                                -0.404004   0.359116
## atl                                                    1.612271   1.915494
## fre                                                   -2.095035   0.545251
## fkgl                                                         NA         NA
## gf                                                    -1.876752   2.118482
## smog                                                   0.646687   1.695271
##                                                       z value Pr(>|z|)    
## (Intercept)                                            -3.368 0.000758 ***
## RuleDoubleAdpos.max_allowable_distance                  0.798 0.424634    
## RuleDoubleAdpos.max_allowable_distance.v               -0.683 0.494453    
## RuleTooFewVerbs.min_verb_frac                          -3.606 0.000311 ***
## RuleTooManyNegations.max_negation_frac                  0.226 0.821417    
## RuleTooManyNegations.max_negation_frac.v                0.486 0.626883    
## RuleTooManyNegations.max_allowable_negations            0.407 0.684044    
## RuleTooManyNegations.max_allowable_negations.v         -0.988 0.323073    
## RuleTooManyNominalConstructions.max_noun_frac          -1.965 0.049365 *  
## RuleTooManyNominalConstructions.max_noun_frac.v         1.013 0.310992    
## RuleTooManyNominalConstructions.max_allowable_nouns     0.530 0.595914    
## RuleTooManyNominalConstructions.max_allowable_nouns.v  -1.152 0.249433    
## RuleCaseRepetition.max_repetition_count                 0.178 0.858980    
## RuleCaseRepetition.max_repetition_count.v              -1.921 0.054733 .  
## RuleCaseRepetition.max_repetition_frac                  0.497 0.618955    
## RuleCaseRepetition.max_repetition_frac.v                0.793 0.428050    
## RulePredSubjDistance.max_distance                      -2.039 0.041418 *  
## RulePredSubjDistance.max_distance.v                     0.212 0.832306    
## RulePredObjDistance.max_distance                       -1.059 0.289542    
## RulePredObjDistance.max_distance.v                      0.032 0.974333    
## RuleInfVerbDistance.max_distance                        1.818 0.069003 .  
## RuleInfVerbDistance.max_distance.v                     -2.492 0.012711 *  
## RuleMultiPartVerbs.max_distance                         0.728 0.466368    
## RuleMultiPartVerbs.max_distance.v                       0.939 0.347526    
## RuleLongSentences.max_length                            3.494 0.000476 ***
## RuleLongSentences.max_length.v                          3.434 0.000594 ***
## RulePredAtClauseBeginning.max_order                    -0.281 0.778766    
## RulePredAtClauseBeginning.max_order.v                  -0.576 0.564849    
## cli                                                    -0.453 0.650695    
## ari                                                    -3.169 0.001528 ** 
## entropy                                                -0.546 0.585193    
## ttr                                                    -1.204 0.228706    
## mattr                                                  -1.024 0.305953    
## mattr.v                                                -1.442 0.149278    
## maentropy                                               0.678 0.497995    
## maentropy.v                                             1.794 0.072759 .  
## mamr                                                    0.131 0.895637    
## verb_dist                                               1.623 0.104500    
## activity                                                4.964 6.89e-07 ***
## hpoint                                                 -1.125 0.260590    
## atl                                                     0.842 0.399956    
## fre                                                    -3.842 0.000122 ***
## fkgl                                                       NA       NA    
## gf                                                     -0.886 0.375674    
## smog                                                    0.381 0.702858    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 842.12  on 611  degrees of freedom
## Residual deviance: 502.46  on 568  degrees of freedom
## AIC: 590.46
## 
## Number of Fisher Scoring iterations: 6
```

Counts

```
glm(
  formula_counts,
  data = training_set_modif,
  family = binomial(link = "logit")
) %>% summary()
```

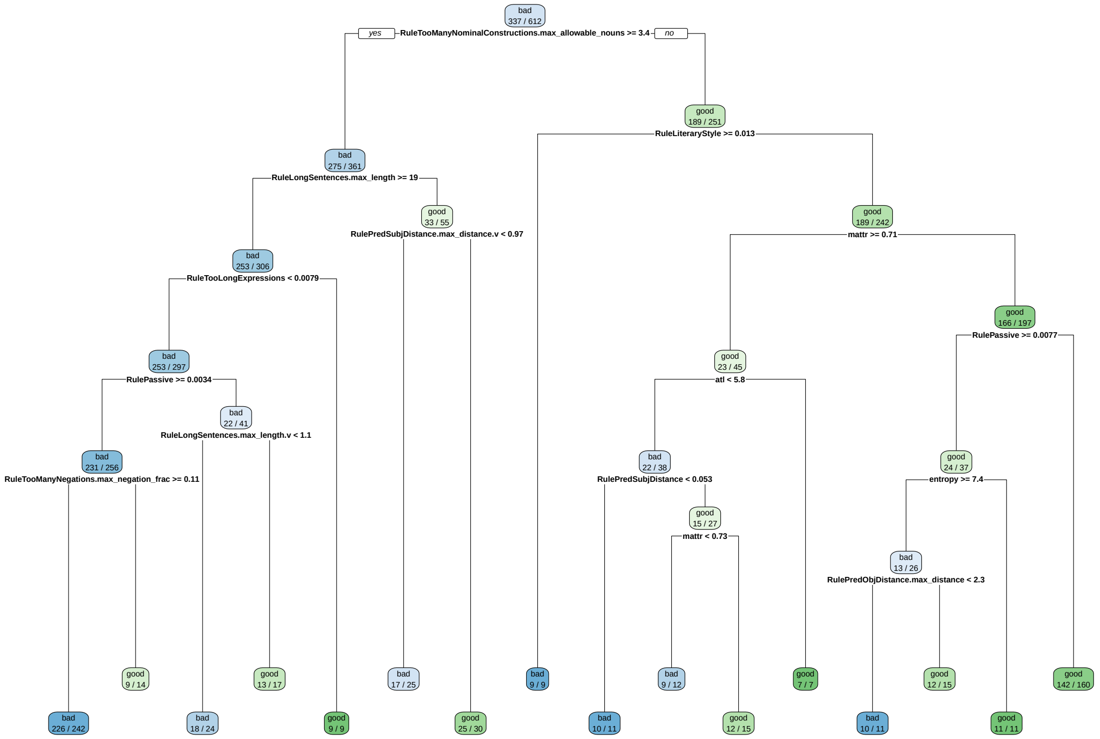
```
## 
## Call:
## glm(formula = formula_counts, family = binomial(link = "logit"), 
##     data = training_set_modif)
## 
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                    -0.48980    0.12417  -3.945 7.99e-05 ***
## RuleGPcoordovs                 -0.02693    0.10339  -0.260 0.794499    
## RuleGPdeverbaddr               -0.24009    0.11055  -2.172 0.029870 *  
## RuleGPpatinstr                -0.04447    0.09841  -0.452 0.651321
## RuleGPdeverbsubj              -0.19249    0.12937  -1.488 0.136774
## RuleGPadjective                0.21364    0.17015   1.256 0.209258
## RuleGPpatbenperson            -0.07276    0.09844  -0.739 0.459841
## RuleGPwordorder               -0.19871    0.11969  -1.660 0.096863 .
## RuleDoubleAdpos               -0.12260    0.11105  -1.104 0.269616
## RuleReflexivePassWithAnimSubj  0.02322    0.10779   0.215 0.829408
## RuleWeakMeaningWords          -0.06538    0.10696  -0.611 0.541037
## RuleAbstractNouns             -0.01576    0.11206  -0.141 0.888158
## RuleRelativisticExpressions   -0.22035    0.12580  -1.752 0.079842 .
## RuleConfirmationExpressions    0.14181    0.12686   1.118 0.263644
## RuleRedundantExpressions      -0.22443    0.14833  -1.513 0.130264
## RuleTooLongExpressions         0.36750    0.11623   3.162 0.001568 **
## RuleAnaphoricReferences        0.33398    0.11934   2.799 0.005134 **
## RuleLiteraryStyle             -0.48480    0.12558  -3.861 0.000113 ***
## RulePassive                   -0.56990    0.14435  -3.948 7.88e-05 ***
## RulePredSubjDistance           0.19828    0.13807   1.436 0.150991
## RulePredObjDistance            0.20756    0.14615   1.420 0.155553
## RuleInfVerbDistance            0.05512    0.14772   0.373 0.709032
## RuleMultiPartVerbs             0.37500    0.15199   2.467 0.013616 *
## RuleVerbalNouns                0.13503    0.12274   1.100 0.271277
## sent_count                     1.70513    0.44154   3.862 0.000113 ***
## word_count                    -3.65338    1.82665  -2.000 0.045496 *
## syllab_count                   0.27311    3.29165   0.083 0.933876
## char_count                     1.03853    3.85562   0.269 0.787656
## num_hapax                     -0.19284    0.16742  -1.152 0.249389
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 842.12  on 611  degrees of freedom
## Residual deviance: 529.92  on 583  degrees of freedom
## AIC: 587.92
##
## Number of Fisher Scoring iterations: 6
```

Decision tree

```
library(rpart) # decision trees for classification and regression
library(rpart.plot) # visualization of decision trees created with rpart
```

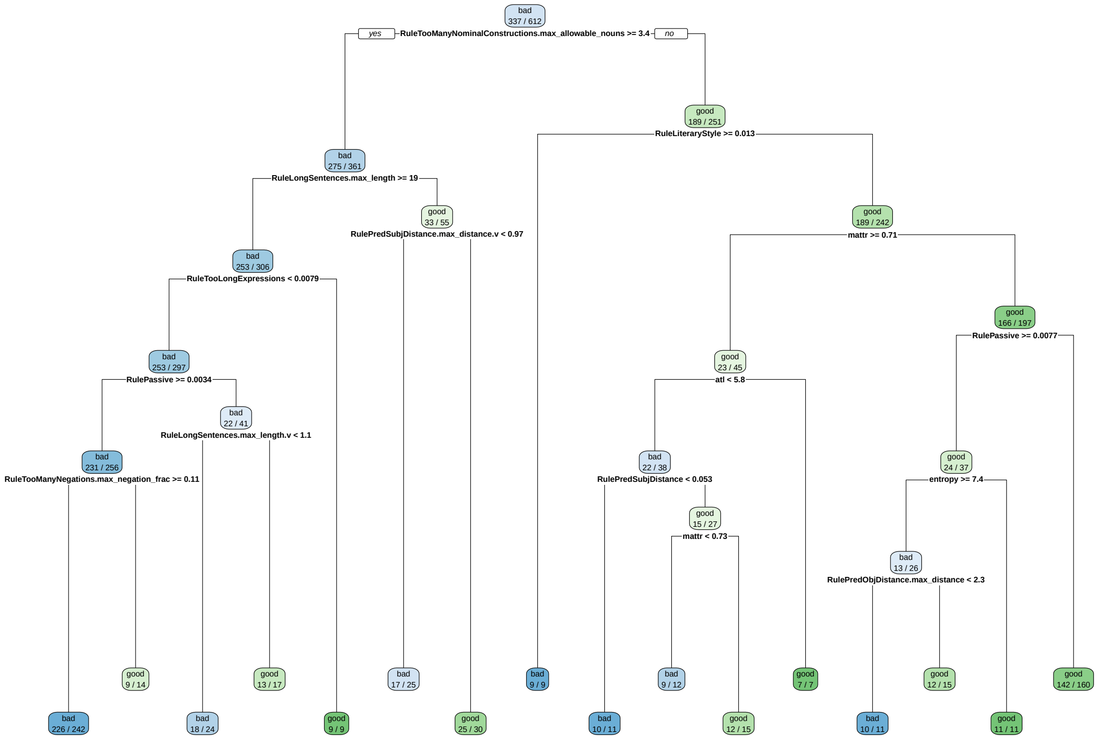
All variables

```
model_dt_all <- train_decision_tree(formula_all, training_set)
```



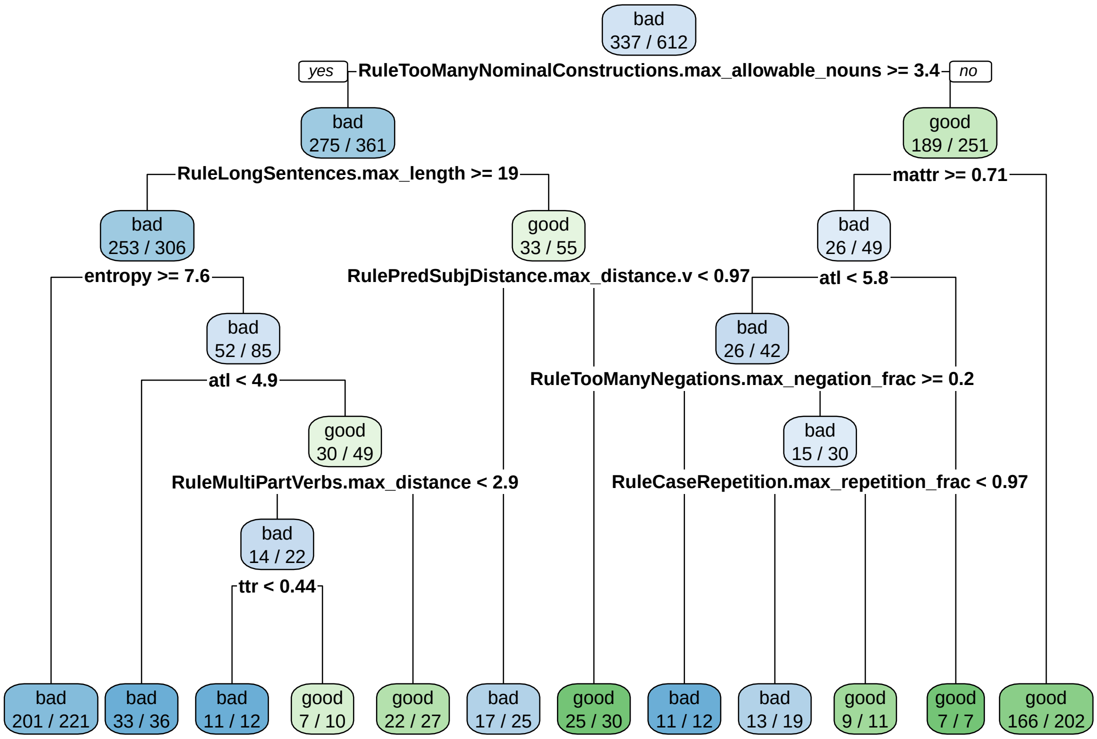
No TL

```r
model_dt_notl <- train_decision_tree(formula_notl, training_set)
```



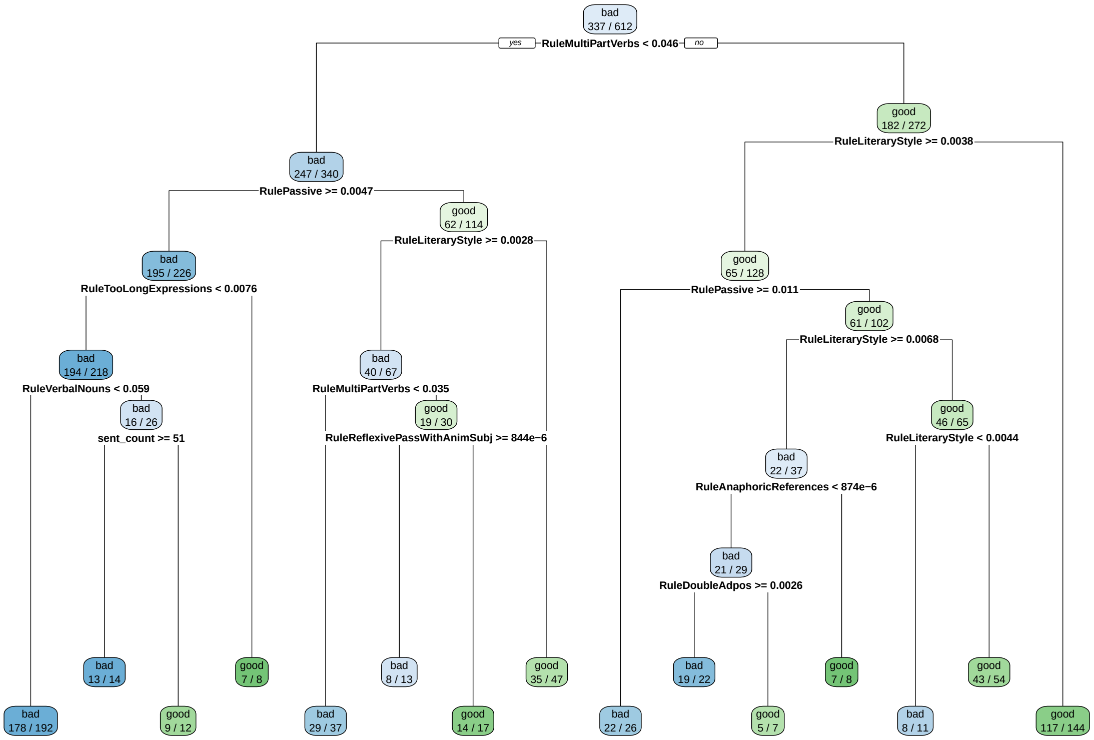
IAC

```
model_dt_iac <- train_decision_tree(formula_iac, training_set)
```



Counts

```
model_dt_counts <- train_decision_tree(formula_counts, training_set)
```



Lasso

All variables

Remove correlating

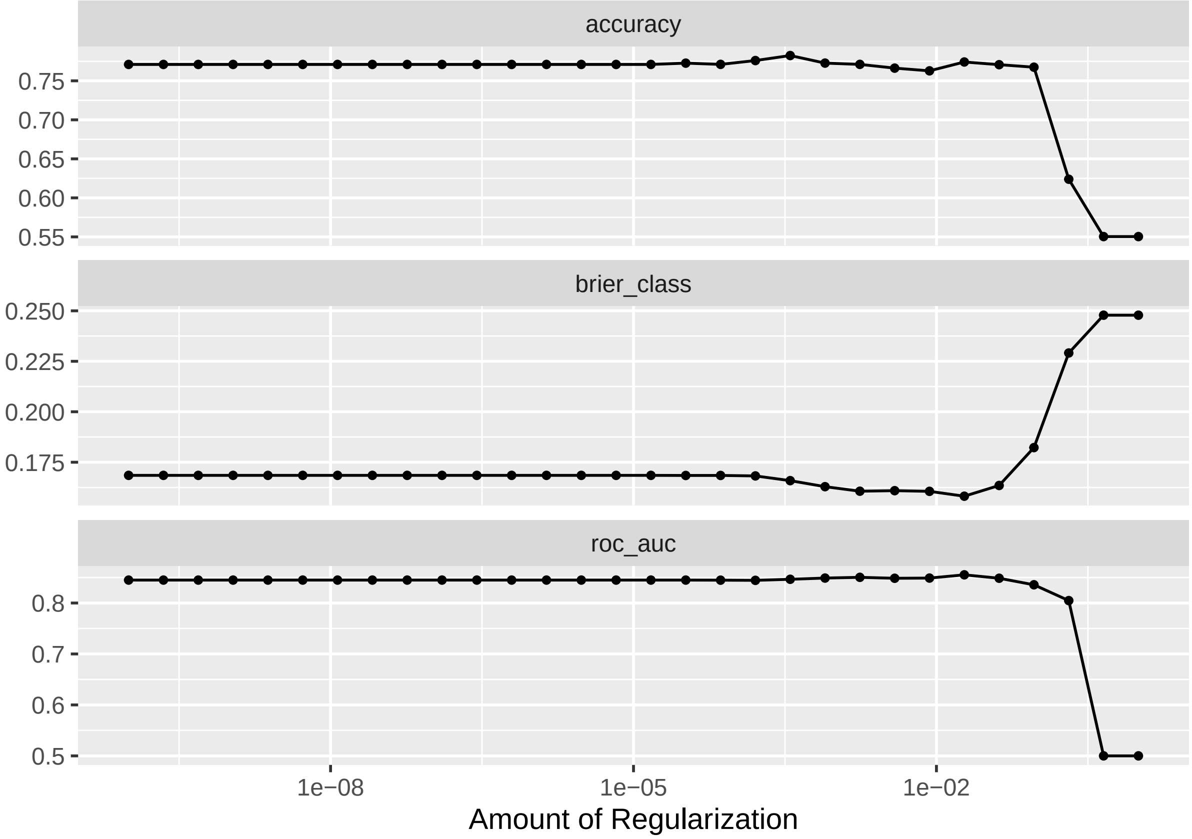
```
# train_lasso(recipe_all, training_set, folds)
```

Keep correlating

```
model_lasso_all <- train_lasso(recipe_all_nocorr, training_set, folds)
```

```
## Lasso tune workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = tune()
##   mixture = 1
##
```

```
## Computational engine: glmnet
##
## Lasso tuning metrics:
```

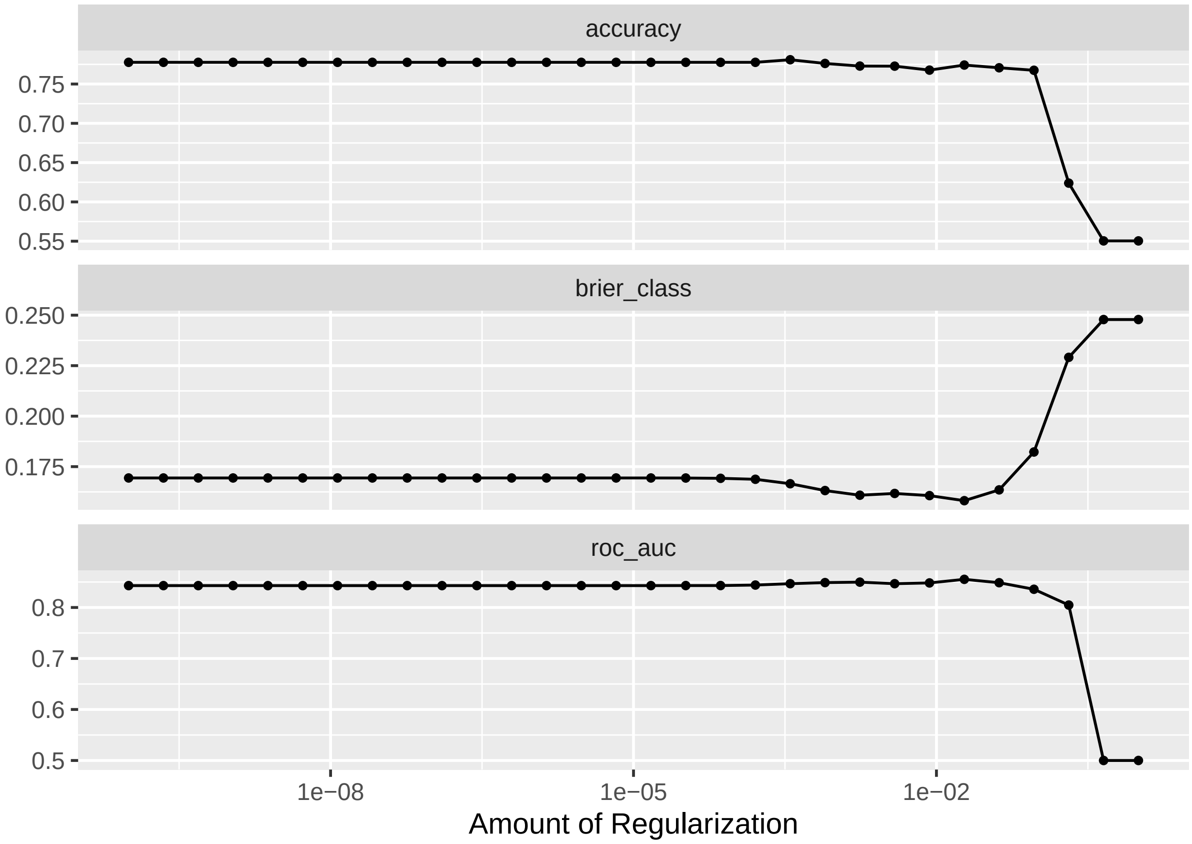


```
## # A tibble: 5 x 7
##    penalty .metric .estimator  mean     n std_err .config
##      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.0189   roc_auc binary     0.855    10  0.0192 Preprocessor1_Model25
## 2 0.00174  roc_auc binary     0.850    10  0.0180 Preprocessor1_Model22
## 3 0.000788 roc_auc binary     0.849    10  0.0170 Preprocessor1_Model21
## 4 0.00853  roc_auc binary     0.849    10  0.0201 Preprocessor1_Model24
## 5 0.0418   roc_auc binary     0.849    10  0.0162 Preprocessor1_Model26
## # A tibble: 5 x 7
##     penalty .metric  .estimator  mean     n std_err .config
##       <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.000356  accuracy binary     0.782    10  0.0163 Preprocessor1_Model20
## 2 0.000161  accuracy binary     0.776    10  0.0163 Preprocessor1_Model19
## 3 0.0189    accuracy binary     0.774    10  0.0172 Preprocessor1_Model25
## 4 0.000788  accuracy binary     0.773    10  0.0160 Preprocessor1_Model21
## 5 0.0000329 accuracy binary     0.773    10  0.0175 Preprocessor1_Model17
## Best accuracy:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0924 Preprocessor1_Model27
## Final workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.0923670857187388
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 72 x 3
##    term                                                 estimate penalty
##    <chr>                                                   <dbl>   <dbl>
##  1 (Intercept)                                         -0.230     0.0924
##  2 smog                                                -0.191     0.0924
##  3 RuleLiteraryStyle                                   -0.168     0.0924
##  4 gf                                                  -0.0184    0.0924
##  5 entropy                                             -0.0165    0.0924
##  6 maentropy                                           -0.00435   0.0924
##  7 ari                                                 -0.000272  0.0924
##  8 RuleGPcoordovs                                       0         0.0924
##  9 RuleGPdeverbaddr                                     0         0.0924
## 10 RuleGPpatinstr                                       0         0.0924
## 11 RuleGPdeverbsubj                                     0         0.0924
## 12 RuleGPadjective                                      0         0.0924
## 13 RuleGPpatbenperson                                   0         0.0924
## 14 RuleGPwordorder                                      0         0.0924
## 15 RuleDoubleAdpos                                      0         0.0924
## 16 RuleDoubleAdpos.max_allowable_distance               0         0.0924
## 17 RuleDoubleAdpos.max_allowable_distance.v             0         0.0924
## 18 RuleReflexivePassWithAnimSubj                        0         0.0924
## 19 RuleTooFewVerbs.min_verb_frac                        0         0.0924
## 20 RuleTooManyNegations.max_negation_frac               0         0.0924
## 21 RuleTooManyNegations.max_negation_frac.v             0         0.0924
## 22 RuleTooManyNegations.max_allowable_negations         0         0.0924
## 23 RuleTooManyNegations.max_allowable_negations.v       0         0.0924
## 24 RuleTooManyNominalConstructions.max_noun_frac        0         0.0924
## 25 RuleTooManyNominalConstructions.max_noun_frac.v      0         0.0924
## 26 RuleTooManyNominalConstructions.max_allowable_nouns  0         0.0924
## 27 RuleCaseRepetition.max_repetition_count              0         0.0924
## 28 RuleCaseRepetition.max_repetition_count.v            0         0.0924
## 29 RuleCaseRepetition.max_repetition_frac               0         0.0924
## 30 RuleCaseRepetition.max_repetition_frac.v             0         0.0924
## 31 RuleWeakMeaningWords                                 0         0.0924
## 32 RuleAbstractNouns                                    0         0.0924
## 33 RuleRelativisticExpressions                          0         0.0924
## 34 RuleConfirmationExpressions                          0         0.0924
## 35 RuleRedundantExpressions                             0         0.0924
## 36 RuleTooLongExpressions                               0         0.0924
## 37 RuleAnaphoricReferences                              0         0.0924
## 38 RulePassive                                          0         0.0924
## 39 RulePredSubjDistance                                 0         0.0924
## 40 RulePredSubjDistance.max_distance                    0         0.0924
## 41 RulePredSubjDistance.max_distance.v                  0         0.0924
## 42 RulePredObjDistance                                  0         0.0924
## 43 RulePredObjDistance.max_distance                     0         0.0924
## 44 RulePredObjDistance.max_distance.v                   0         0.0924
## 45 RuleInfVerbDistance                                  0         0.0924
## 46 RuleInfVerbDistance.max_distance                     0         0.0924
## 47 RuleInfVerbDistance.max_distance.v                   0         0.0924
## 48 RuleMultiPartVerbs                                   0         0.0924
## 49 RuleMultiPartVerbs.max_distance                      0         0.0924
## 50 RuleMultiPartVerbs.max_distance.v                    0         0.0924
## 51 RuleLongSentences.max_length                         0         0.0924
## 52 RuleLongSentences.max_length.v                       0         0.0924
## 53 RulePredAtClauseBeginning.max_order                  0         0.0924
## 54 RulePredAtClauseBeginning.max_order.v                0         0.0924
## 55 RuleVerbalNouns                                      0         0.0924
## 56 sent_count                                           0         0.0924
## 57 word_count                                           0         0.0924
## 58 syllab_count                                         0         0.0924
## 59 char_count                                           0         0.0924
## 60 cli                                                  0         0.0924
## 61 num_hapax                                            0         0.0924
## 62 ttr                                                  0         0.0924
## 63 mattr                                                0         0.0924
## 64 mattr.v                                              0         0.0924
## 65 maentropy.v                                          0         0.0924
## 66 verb_dist                                            0         0.0924
## 67 hpoint                                               0         0.0924
## 68 fre                                                  0         0.0924
## 69 fkgl                                                 0         0.0924
## 70 mamr                                                 0.0576    0.0924
## 71 atl                                                  0.100     0.0924
## 72 activity                                             0.408     0.0924
## Variable importance:
## # A tibble: 71 x 3
##    Variable                                            Importance Sign
##    <chr>                                                    <dbl> <chr>
##  1 char_count                                            13.8     POS
##  2 syllab_count                                           9.84    NEG
##  3 ari                                                    5.09    NEG
##  4 word_count                                             4.36    NEG
##  5 RuleLongSentences.max_length                           3.32    POS
##  6 fre                                                    2.55    NEG
##  7 gf                                                     2.25    NEG
##  8 RuleTooFewVerbs.min_verb_frac                          1.74    NEG
##  9 activity                                               1.64    POS
## 10 smog                                                   1.53    POS
## 11 sent_count                                             1.21    POS
## 12 RuleCaseRepetition.max_repetition_frac.v               1.20    POS
## 13 mattr                                                  1.19    NEG
## 14 hpoint                                                 1.19    NEG
## 15 ttr                                                    1.06    NEG
## 16 atl                                                    1.02    POS
## 17 maentropy.v                                            0.900   POS
## 18 maentropy                                              0.892   POS
## 19 RuleLongSentences.max_length.v                         0.830   POS
## 20 RuleCaseRepetition.max_repetition_frac                 0.821   POS
## 21 cli                                                    0.791   NEG
## 22 entropy                                                0.598   NEG
## 23 num_hapax                                              0.547   POS
## 24 RuleMultiPartVerbs                                     0.534   POS
## 25 RulePredSubjDistance.max_distance                      0.519   NEG
## 26 RuleAnaphoricReferences                                0.516   POS
## 27 RulePassive                                            0.492   NEG
## 28 RulePredSubjDistance                                   0.466   POS
## 29 RuleLiteraryStyle                                      0.410   NEG
## 30 mattr.v                                                0.405   NEG
## 31 RuleGPadjective                                        0.392   POS
## 32 verb_dist                                              0.327   POS
## 33 RuleInfVerbDistance.max_distance                       0.322   POS
## 34 RulePredObjDistance.max_distance                       0.320   NEG
## 35 RuleTooManyNominalConstructions.max_noun_frac          0.319   NEG
## 36 RuleTooManyNominalConstructions.max_allowable_nouns    0.291   POS
## 37 RuleTooLongExpressions                                 0.290   POS
## 38 RuleRelativisticExpressions                            0.257   NEG
## 39 RulePredAtClauseBeginning.max_order                    0.255   NEG
## 40 RuleCaseRepetition.max_repetition_count                0.249   NEG
## 41 RuleGPdeverbaddr                                       0.246   NEG
## 42 RuleInfVerbDistance.max_distance.v                     0.243   NEG
## 43 RuleCaseRepetition.max_repetition_count.v              0.236   NEG
## 44 RuleTooManyNegations.max_allowable_negations           0.230   POS
## 45 RuleRedundantExpressions                               0.195   NEG
## 46 RuleGPdeverbsubj                                       0.189   NEG
## 47 RuleConfirmationExpressions                            0.186   POS
## 48 RuleInfVerbDistance                                    0.166   POS
## 49 RuleGPpatbenperson                                     0.162   NEG
## 50 RuleMultiPartVerbs.max_distance.v                      0.157   POS
## 51 RuleGPwordorder                                        0.142   NEG
## 52 RuleTooManyNegations.max_negation_frac                 0.134   POS
## 53 RuleTooManyNegations.max_allowable_negations.v         0.133   NEG
## 54 RuleGPpatinstr                                         0.125   NEG
## 55 RuleWeakMeaningWords                                   0.118   NEG
## 56 RuleAbstractNouns                                      0.103   POS
## 57 mamr                                                   0.102   NEG
## 58 RuleDoubleAdpos.max_allowable_distance.v               0.0976  POS
## 59 RuleMultiPartVerbs.max_distance                        0.0891  POS
## 60 RuleReflexivePassWithAnimSubj                          0.0819  NEG
## 61 RuleTooManyNominalConstructions.max_noun_frac.v        0.0799  POS
## 62 RulePredSubjDistance.max_distance.v                    0.0700  NEG
## 63 RuleDoubleAdpos                                        0.0563  POS
## 64 RuleVerbalNouns                                        0.0556  NEG
## 65 RuleTooManyNegations.max_negation_frac.v               0.0552  NEG
## 66 RuleGPcoordovs                                         0.0487  NEG
## 67 RuleDoubleAdpos.max_allowable_distance                 0.0357  NEG
## 68 RulePredAtClauseBeginning.max_order.v                  0.0334  POS
## 69 RulePredObjDistance.max_distance.v                     0.0322  POS
## 70 RulePredObjDistance                                    0.00271 POS
## 71 fkgl                                                   0       NEG
```

No TL

```
model_lasso_notl <- train_lasso(recipe_notl_nocorr, training_set, folds)
```

```
## Lasso tune workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = tune()
##   mixture = 1
##
## Computational engine: glmnet
##
## Lasso tuning metrics:
```



```
## # A tibble: 5 x 7
##    penalty .metric .estimator  mean     n std_err .config
##      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.0189   roc_auc binary     0.855    10  0.0192 Preprocessor1_Model25
## 2 0.00174  roc_auc binary     0.850    10  0.0178 Preprocessor1_Model22
## 3 0.000788 roc_auc binary     0.849    10  0.0165 Preprocessor1_Model21
## 4 0.0418   roc_auc binary     0.849    10  0.0162 Preprocessor1_Model26
## 5 0.00853  roc_auc binary     0.848    10  0.0200 Preprocessor1_Model24
## # A tibble: 5 x 7
##    penalty .metric  .estimator  mean     n std_err .config
##      <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 3.56e- 4 accuracy binary     0.781    10  0.0146 Preprocessor1_Model20
## 2 1   e-10 accuracy binary     0.778    10  0.0160 Preprocessor1_Model01
## 3 2.21e-10 accuracy binary     0.778    10  0.0160 Preprocessor1_Model02
## 4 4.89e-10 accuracy binary     0.778    10  0.0160 Preprocessor1_Model03
## 5 1.08e- 9 accuracy binary     0.778    10  0.0160 Preprocessor1_Model04
## Best accuracy:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0924 Preprocessor1_Model27
## Final workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.0923670857187388
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 68 x 3
##    term                                                 estimate penalty
##    <chr>                                                   <dbl>   <dbl>
##  1 (Intercept)                                         -0.230     0.0924
##  2 smog                                                -0.191     0.0924
##  3 RuleLiteraryStyle                                   -0.168     0.0924
##  4 gf                                                  -0.0184    0.0924
##  5 entropy                                             -0.0165    0.0924
##  6 maentropy                                           -0.00435   0.0924
##  7 ari                                                 -0.000272  0.0924
##  8 RuleGPcoordovs                                       0         0.0924
##  9 RuleGPdeverbaddr                                     0         0.0924
## 10 RuleGPpatinstr                                       0         0.0924
## 11 RuleGPdeverbsubj                                     0         0.0924
## 12 RuleGPadjective                                      0         0.0924
## 13 RuleGPpatbenperson                                   0         0.0924
## 14 RuleGPwordorder                                      0         0.0924
## 15 RuleDoubleAdpos                                      0         0.0924
## 16 RuleDoubleAdpos.max_allowable_distance               0         0.0924
## 17 RuleDoubleAdpos.max_allowable_distance.v             0         0.0924
## 18 RuleReflexivePassWithAnimSubj                        0         0.0924
## 19 RuleTooFewVerbs.min_verb_frac                        0         0.0924
## 20 RuleTooManyNegations.max_negation_frac               0         0.0924
## 21 RuleTooManyNegations.max_negation_frac.v             0         0.0924
## 22 RuleTooManyNegations.max_allowable_negations         0         0.0924
## 23 RuleTooManyNegations.max_allowable_negations.v       0         0.0924
## 24 RuleTooManyNominalConstructions.max_noun_frac        0         0.0924
## 25 RuleTooManyNominalConstructions.max_noun_frac.v      0         0.0924
## 26 RuleTooManyNominalConstructions.max_allowable_nouns  0         0.0924
## 27 RuleCaseRepetition.max_repetition_count              0         0.0924
## 28 RuleCaseRepetition.max_repetition_count.v            0         0.0924
## 29 RuleCaseRepetition.max_repetition_frac               0         0.0924
## 30 RuleCaseRepetition.max_repetition_frac.v             0         0.0924
## 31 RuleWeakMeaningWords                                 0         0.0924
## 32 RuleAbstractNouns                                    0         0.0924
## 33 RuleRelativisticExpressions                          0         0.0924
## 34 RuleConfirmationExpressions                          0         0.0924
## 35 RuleRedundantExpressions                             0         0.0924
## 36 RuleTooLongExpressions                               0         0.0924
## 37 RuleAnaphoricReferences                              0         0.0924
## 38 RulePassive                                          0         0.0924
## 39 RulePredSubjDistance                                 0         0.0924
## 40 RulePredSubjDistance.max_distance                    0         0.0924
## 41 RulePredSubjDistance.max_distance.v                  0         0.0924
## 42 RulePredObjDistance                                  0         0.0924
## 43 RulePredObjDistance.max_distance                     0         0.0924
## 44 RulePredObjDistance.max_distance.v                   0         0.0924
## 45 RuleInfVerbDistance                                  0         0.0924
## 46 RuleInfVerbDistance.max_distance                     0         0.0924
## 47 RuleInfVerbDistance.max_distance.v                   0         0.0924
## 48 RuleMultiPartVerbs                                   0         0.0924
## 49 RuleMultiPartVerbs.max_distance                      0         0.0924
## 50 RuleMultiPartVerbs.max_distance.v                    0         0.0924
## 51 RuleLongSentences.max_length                         0         0.0924
## 52 RuleLongSentences.max_length.v                       0         0.0924
## 53 RulePredAtClauseBeginning.max_order                  0         0.0924
## 54 RulePredAtClauseBeginning.max_order.v                0         0.0924
## 55 RuleVerbalNouns                                      0         0.0924
## 56 cli                                                  0         0.0924
## 57 num_hapax                                            0         0.0924
## 58 ttr                                                  0         0.0924
## 59 mattr                                                0         0.0924
## 60 mattr.v                                              0         0.0924
## 61 maentropy.v                                          0         0.0924
## 62 verb_dist                                            0         0.0924
## 63 hpoint                                               0         0.0924
## 64 fre                                                  0         0.0924
## 65 fkgl                                                 0         0.0924
## 66 mamr                                                 0.0576    0.0924
## 67 atl                                                  0.100     0.0924
## 68 activity                                             0.408     0.0924
## Variable importance:
## # A tibble: 67 x 3
##    Variable                                            Importance Sign
##    <chr>                                                    <dbl> <chr>
##  1 ari                                                    4.40    NEG
##  2 RuleLongSentences.max_length                           3.15    POS
##  3 atl                                                    2.41    POS
##  4 gf                                                     2.12    NEG
##  5 fre                                                    1.75    NEG
##  6 RuleTooFewVerbs.min_verb_frac                          1.68    NEG
##  7 activity                                               1.64    POS
##  8 smog                                                   1.45    POS
##  9 cli                                                    1.30    NEG
## 10 RuleCaseRepetition.max_repetition_frac.v               1.30    POS
## 11 RuleCaseRepetition.max_repetition_frac                 0.932   POS
## 12 mattr                                                  0.888   NEG
## 13 RuleLongSentences.max_length.v                         0.827   POS
## 14 ttr                                                    0.733   NEG
## 15 maentropy.v                                            0.672   POS
## 16 RulePassive                                            0.522   NEG
## 17 RuleAnaphoricReferences                                0.506   POS
## 18 RuleMultiPartVerbs                                     0.493   POS
## 19 RulePredSubjDistance                                   0.482   POS
## 20 maentropy                                              0.471   POS
## 21 RuleGPadjective                                        0.464   POS
## 22 RulePredSubjDistance.max_distance                      0.438   NEG
## 23 RuleTooManyNominalConstructions.max_allowable_nouns    0.421   POS
## 24 RuleLiteraryStyle                                      0.417   NEG
## 25 hpoint                                                 0.398   NEG
## 26 verb_dist                                              0.385   POS
## 27 num_hapax                                              0.370   POS
## 28 RulePredObjDistance.max_distance                       0.362   NEG
## 29 RuleTooManyNominalConstructions.max_noun_frac          0.343   NEG
## 30 RuleInfVerbDistance.max_distance.v                     0.337   NEG
## 31 RuleInfVerbDistance.max_distance                       0.327   POS
## 32 RuleTooLongExpressions                                 0.294   POS
## 33 mattr.v                                                0.271   NEG
## 34 RuleCaseRepetition.max_repetition_count                0.258   NEG
## 35 RuleGPdeverbaddr                                       0.256   NEG
## 36 RuleRelativisticExpressions                            0.246   NEG
## 37 RuleCaseRepetition.max_repetition_count.v              0.238   NEG
## 38 RulePredAtClauseBeginning.max_order                    0.228   NEG
## 39 entropy                                                0.222   NEG
## 40 RuleGPdeverbsubj                                       0.205   NEG
## 41 RuleConfirmationExpressions                            0.203   POS
## 42 RuleRedundantExpressions                               0.191   NEG
## 43 RuleTooManyNegations.max_allowable_negations           0.181   POS
## 44 RuleGPpatbenperson                                     0.158   NEG
## 45 RuleTooManyNegations.max_negation_frac                 0.155   POS
## 46 RuleWeakMeaningWords                                   0.148   NEG
## 47 RuleInfVerbDistance                                    0.137   POS
## 48 RuleGPwordorder                                        0.130   NEG
## 49 RuleMultiPartVerbs.max_distance.v                      0.129   POS
## 50 RuleTooManyNegations.max_allowable_negations.v         0.125   NEG
## 51 RuleTooManyNegations.max_negation_frac.v               0.110   NEG
## 52 RuleAbstractNouns                                      0.104   POS
## 53 RuleTooManyNominalConstructions.max_noun_frac.v        0.103   POS
## 54 RuleGPpatinstr                                         0.0969  NEG
## 55 RulePredSubjDistance.max_distance.v                    0.0941  NEG
## 56 RuleMultiPartVerbs.max_distance                        0.0679  POS
## 57 RuleReflexivePassWithAnimSubj                          0.0665  NEG
## 58 mamr                                                   0.0618  NEG
## 59 RulePredObjDistance.max_distance.v                     0.0426  POS
## 60 RuleGPcoordovs                                         0.0318  NEG
## 61 RulePredAtClauseBeginning.max_order.v                  0.0304  POS
## 62 RuleDoubleAdpos                                        0.0295  POS
## 63 RuleDoubleAdpos.max_allowable_distance.v               0.0277  POS
## 64 RuleDoubleAdpos.max_allowable_distance                 0.0272  NEG
## 65 RulePredObjDistance                                    0.0228  POS
## 66 RuleVerbalNouns                                        0.00959 NEG
## 67 fkgl                                                   0       NEG
```

Indicators, averages, and coefficients

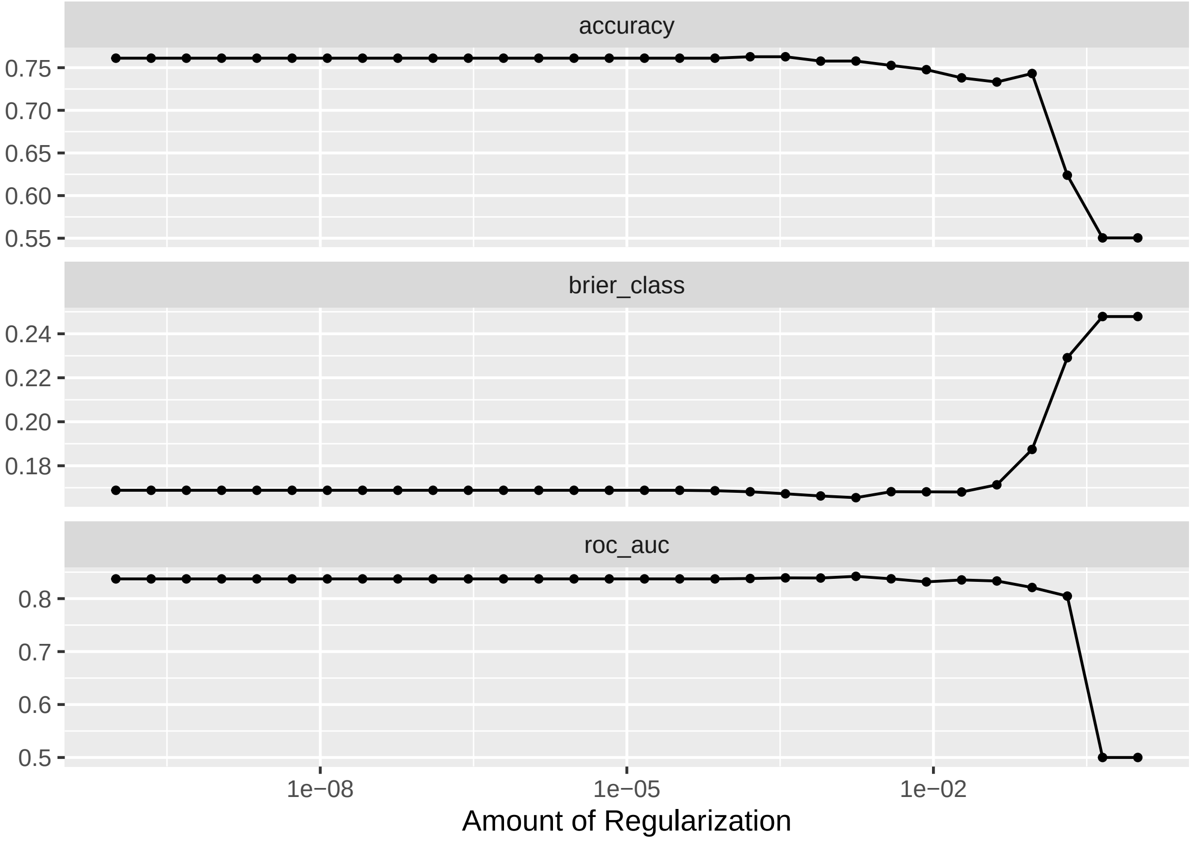
Remove correlating

```
# train_lasso(recipe_iac, training_set, folds)
```

Keep correlating

```
model_lasso_iac <- train_lasso(recipe_iac_nocorr, training_set, folds)
```

```
## Lasso tune workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = tune()
##   mixture = 1
##
## Computational engine: glmnet
##
## Lasso tuning metrics:
```



```
## # A tibble: 5 x 7
##    penalty .metric .estimator  mean     n std_err .config
##      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00174  roc_auc binary     0.842    10  0.0168 Preprocessor1_Model22
## 2 0.000356 roc_auc binary     0.839    10  0.0160 Preprocessor1_Model20
## 3 0.000788 roc_auc binary     0.839    10  0.0164 Preprocessor1_Model21
## 4 0.000161 roc_auc binary     0.838    10  0.0156 Preprocessor1_Model19
## 5 0.00386  roc_auc binary     0.837    10  0.0179 Preprocessor1_Model23
## # A tibble: 5 x 7
##    penalty .metric  .estimator  mean     n std_err .config
##      <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 1.61e- 4 accuracy binary     0.763    10  0.0132 Preprocessor1_Model19
## 2 3.56e- 4 accuracy binary     0.763    10  0.0138 Preprocessor1_Model20
## 3 1   e-10 accuracy binary     0.761    10  0.0137 Preprocessor1_Model01
## 4 2.21e-10 accuracy binary     0.761    10  0.0137 Preprocessor1_Model02
## 5 4.89e-10 accuracy binary     0.761    10  0.0137 Preprocessor1_Model03
## Best accuracy:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1 0.00386 Preprocessor1_Model23
## Final workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model ------------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.00385662042116347
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 45 x 3
##    term                                                   estimate penalty
##    <chr>                                                     <dbl>   <dbl>
##  1 RuleTooFewVerbs.min_verb_frac                         -16.1     0.00386
##  2 RuleCaseRepetition.max_repetition_frac                -14.2     0.00386
##  3 RuleTooManyNominalConstructions.max_noun_frac          -6.66    0.00386
##  4 mattr                                                  -6.42    0.00386
##  5 RuleCaseRepetition.max_repetition_count.v              -1.90    0.00386
##  6 ttr                                                    -1.09    0.00386
##  7 RuleTooManyNominalConstructions.max_allowable_nouns.v  -0.991   0.00386
##  8 RuleTooManyNegations.max_allowable_negations.v         -0.867   0.00386
##  9 RuleInfVerbDistance.max_distance.v                     -0.778   0.00386
## 10 entropy                                                -0.576   0.00386
## 11 ari                                                    -0.167   0.00386
## 12 gf                                                     -0.140   0.00386
## 13 RuleDoubleAdpos.max_allowable_distance.v               -0.138   0.00386
## 14 RulePredSubjDistance.max_distance.v                    -0.0890  0.00386
## 15 fre                                                    -0.0449  0.00386
## 16 smog                                                   -0.0307  0.00386
## 17 RulePredSubjDistance.max_distance                      -0.0230  0.00386
## 18 RulePredObjDistance.max_distance                       -0.0213  0.00386
## 19 hpoint                                                 -0.00122 0.00386
## 20 RuleTooManyNegations.max_negation_frac.v                0       0.00386
## 21 RuleTooManyNegations.max_allowable_negations            0       0.00386
## 22 RuleCaseRepetition.max_repetition_count                 0       0.00386
## 23 RulePredObjDistance.max_distance.v                      0       0.00386
## 24 RuleMultiPartVerbs.max_distance                         0       0.00386
## 25 RulePredAtClauseBeginning.max_order.v                   0       0.00386
## 26 cli                                                     0       0.00386
## 27 mattr.v                                                 0       0.00386
## 28 maentropy                                               0       0.00386
## 29 mamr                                                    0       0.00386
## 30 fkgl                                                    0       0.00386
## 31 RuleDoubleAdpos.max_allowable_distance                  0.00441 0.00386
## 32 RulePredAtClauseBeginning.max_order                     0.00681 0.00386
## 33 verb_dist                                               0.0325  0.00386
## 34 RuleTooManyNominalConstructions.max_allowable_nouns     0.0332  0.00386
## 35 RuleLongSentences.max_length                            0.0354  0.00386
## 36 RuleInfVerbDistance.max_distance                        0.100   0.00386
## 37 RuleMultiPartVerbs.max_distance.v                       0.155   0.00386
## 38 RuleTooManyNegations.max_negation_frac                  0.479   0.00386
## 39 RuleLongSentences.max_length.v                          1.10    0.00386
## 40 atl                                                     1.90    0.00386
## 41 RuleTooManyNominalConstructions.max_noun_frac.v         2.11    0.00386
## 42 RuleCaseRepetition.max_repetition_frac.v                4.98    0.00386
## 43 maentropy.v                                             9.14    0.00386
## 44 activity                                               11.4     0.00386
## 45 (Intercept)                                            18.4     0.00386
## Variable importance:
## # A tibble: 44 x 3
##    Variable                                              Importance Sign
##    <chr>                                                      <dbl> <chr>
##  1 RuleCaseRepetition.max_repetition_frac.v                49.6     POS
##  2 maentropy.v                                             46.4     POS
##  3 RuleTooFewVerbs.min_verb_frac                           39.6     NEG
##  4 RuleCaseRepetition.max_repetition_frac                  33.7     POS
##  5 mattr                                                   19.5     NEG
##  6 mattr.v                                                 17.2     NEG
##  7 activity                                                16.6     POS
##  8 RuleTooManyNominalConstructions.max_noun_frac           13.8     NEG
##  9 ttr                                                      4.91    NEG
## 10 RuleTooManyNominalConstructions.max_noun_frac.v          3.68    POS
## 11 maentropy                                                3.60    POS
## 12 RuleCaseRepetition.max_repetition_count.v                2.97    NEG
## 13 atl                                                      2.13    POS
## 14 RuleLongSentences.max_length.v                           1.90    POS
## 15 RuleTooManyNominalConstructions.max_allowable_nouns.v    1.33    NEG
## 16 RuleTooManyNegations.max_allowable_negations.v           1.19    NEG
## 17 mamr                                                     1.05    POS
## 18 RuleInfVerbDistance.max_distance.v                       0.923   NEG
## 19 RuleTooManyNegations.max_negation_frac                   0.851   POS
## 20 ari                                                      0.816   NEG
## 21 entropy                                                  0.382   NEG
## 22 RuleTooManyNegations.max_allowable_negations             0.377   POS
## 23 RuleMultiPartVerbs.max_distance.v                        0.351   POS
## 24 RuleDoubleAdpos.max_allowable_distance.v                 0.291   NEG
## 25 gf                                                       0.285   NEG
## 26 RuleTooManyNegations.max_negation_frac.v                 0.276   POS
## 27 RulePredAtClauseBeginning.max_order.v                    0.233   NEG
## 28 RuleLongSentences.max_length                             0.223   POS
## 29 RuleTooManyNominalConstructions.max_allowable_nouns      0.203   POS
## 30 RuleCaseRepetition.max_repetition_count                  0.198   POS
## 31 fre                                                      0.173   NEG
## 32 RulePredSubjDistance.max_distance                        0.127   NEG
## 33 RuleInfVerbDistance.max_distance                         0.106   POS
## 34 hpoint                                                   0.0650  NEG
## 35 RulePredObjDistance.max_distance                         0.0644  NEG
## 36 verb_dist                                                0.0525  POS
## 37 RulePredSubjDistance.max_distance.v                      0.0480  POS
## 38 cli                                                      0.0475  NEG
## 39 RulePredAtClauseBeginning.max_order                      0.0357  NEG
## 40 RuleDoubleAdpos.max_allowable_distance                   0.0303  POS
## 41 RuleMultiPartVerbs.max_distance                          0.0229  POS
## 42 RulePredObjDistance.max_distance.v                       0.00554 POS
## 43 fkgl                                                        0       NEG
## 44 smog                                                        0       NEG
```

Counts

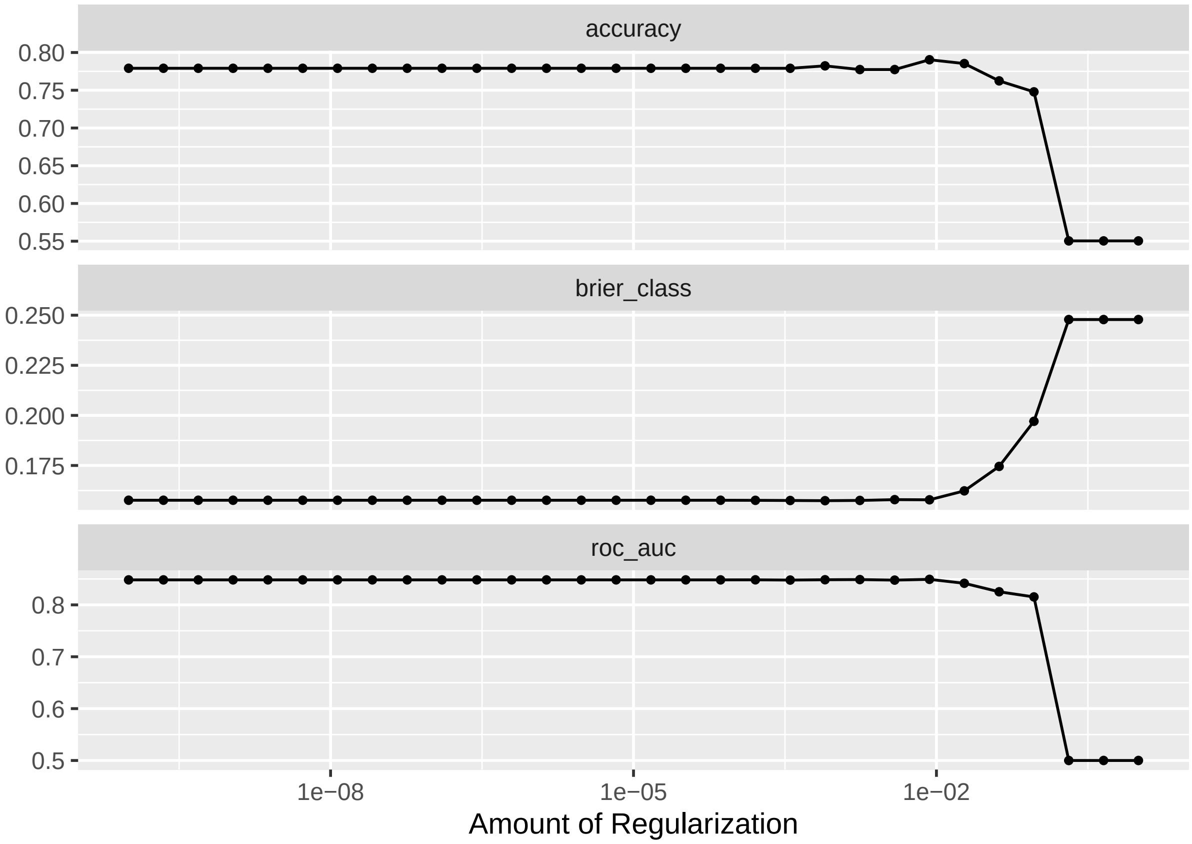
Remove correlating

```
# train_lasso(recipe_counts, training_set, folds)
```

Keep correlating

```
model_lasso_counts <- train_lasso(recipe_counts_nocorr, training_set, folds)
```

```
## Lasso tune workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = tune()
##   mixture = 1
##
## Computational engine: glmnet
##
## Lasso tuning metrics:
```



```
## # A tibble: 5 x 7
##        penalty .metric .estimator  mean     n std_err .config
##          <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00853      roc_auc binary     0.849    10  0.0192 Preprocessor1_Model24
## 2 0.00174      roc_auc binary     0.849    10  0.0188 Preprocessor1_Model22
## 3 0.000788     roc_auc binary     0.848    10  0.0190 Preprocessor1_Model21
## 4 0.000161     roc_auc binary     0.848    10  0.0188 Preprocessor1_Model19
## 5 0.0000000001 roc_auc binary     0.848    10  0.0186 Preprocessor1_Model01
## # A tibble: 5 x 7
##    penalty .metric  .estimator  mean     n std_err .config
##      <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 8.53e- 3 accuracy binary     0.790    10  0.0171 Preprocessor1_Model24
## 2 1.89e- 2 accuracy binary     0.785    10  0.0205 Preprocessor1_Model25
## 3 7.88e- 4 accuracy binary     0.782    10  0.0179 Preprocessor1_Model21
## 4 1   e-10 accuracy binary     0.779    10  0.0172 Preprocessor1_Model01
## 5 2.21e-10 accuracy binary     0.779    10  0.0172 Preprocessor1_Model02
## Best accuracy:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0189 Preprocessor1_Model25
## Final workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.018873918221351
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 29 x 3
##    term                          estimate penalty
##    <chr>                            <dbl>   <dbl>
##  1 RuleRedundantExpressions      -616.        0.0189
##  2 RuleRelativisticExpressions   -332.        0.0189
##  3 RuleGPdeverbsubj              -149.        0.0189
##  4 RuleLiteraryStyle             -123.        0.0189
##  5 RulePassive                   -119.        0.0189
##  6 RuleGPdeverbaddr               -92.8       0.0189
##  7 (Intercept)                     -1.69      0.0189
##  8 word_count                      -0.000438  0.0189
##  9 RuleGPcoordovs                   0         0.0189
## 10 RuleGPpatinstr                   0         0.0189
## 11 RuleGPpatbenperson               0         0.0189
## 12 RuleGPwordorder                  0         0.0189
## 13 RuleDoubleAdpos                  0         0.0189
## 14 RuleReflexivePassWithAnimSubj    0         0.0189
## 15 RuleWeakMeaningWords             0         0.0189
## 16 RuleAbstractNouns                0         0.0189
## 17 RuleConfirmationExpressions      0         0.0189
## 18 RulePredObjDistance              0         0.0189
## 19 syllab_count                     0         0.0189
## 20 char_count                       0         0.0189
## 21 num_hapax                        0         0.0189
## 22 sent_count                       0.00502   0.0189
## 23 RuleInfVerbDistance              0.912     0.0189
## 24 RuleVerbalNouns                  5.83      0.0189
## 25 RulePredSubjDistance            18.2       0.0189
## 26 RuleMultiPartVerbs              34.1       0.0189
## 27 RuleTooLongExpressions          60.5       0.0189
## 28 RuleGPadjective                113.        0.0189
## 29 RuleAnaphoricReferences        157.        0.0189
## Variable importance:
## # A tibble: 28 x 3
##    Variable                      Importance Sign
##    <chr>                              <dbl> <chr>
##  1 RuleRedundantExpressions      2170.         NEG
##  2 RuleRelativisticExpressions    563.         NEG
##  3 RuleGPdeverbaddr               487.         NEG
##  4 RuleConfirmationExpressions    410.         POS
##   5 RuleAnaphoricReferences        349.        POS
##   6 RuleGPdeverbsubj               336.        NEG
##   7 RuleGPadjective                311.        POS
##   8 RuleGPpatbenperson             170.        NEG
##   9 RuleTooLongExpressions         161.        POS
##  10 RuleGPwordorder                157.        NEG
##  11 RulePassive                    124.        NEG
##  12 RuleLiteraryStyle              121.        NEG
##  13 RuleGPcoordovs                  87.9       NEG
##  14 RuleGPpatinstr                  48.6       NEG
##  15 RuleDoubleAdpos                 35.3       NEG
##  16 RuleMultiPartVerbs              27.0       POS
##  17 RuleWeakMeaningWords            26.5       NEG
##  18 RuleReflexivePassWithAnimSubj   18.6       POS
##  19 RulePredSubjDistance            16.0       POS
##  20 RuleVerbalNouns                  7.89      POS
##  21 RuleAbstractNouns                3.67      NEG
##  22 num_hapax                        2.87      NEG
##  23 RulePredObjDistance              0.866     POS
##  24 RuleInfVerbDistance              0.412     POS
##  25 sent_count                       0.0306    POS
##  26 word_count                       0.00242   NEG
##  27 syllab_count                     0.000220  POS
##  28 char_count                       0.00000347 POS
```

SVM

All variables

Remove correlating

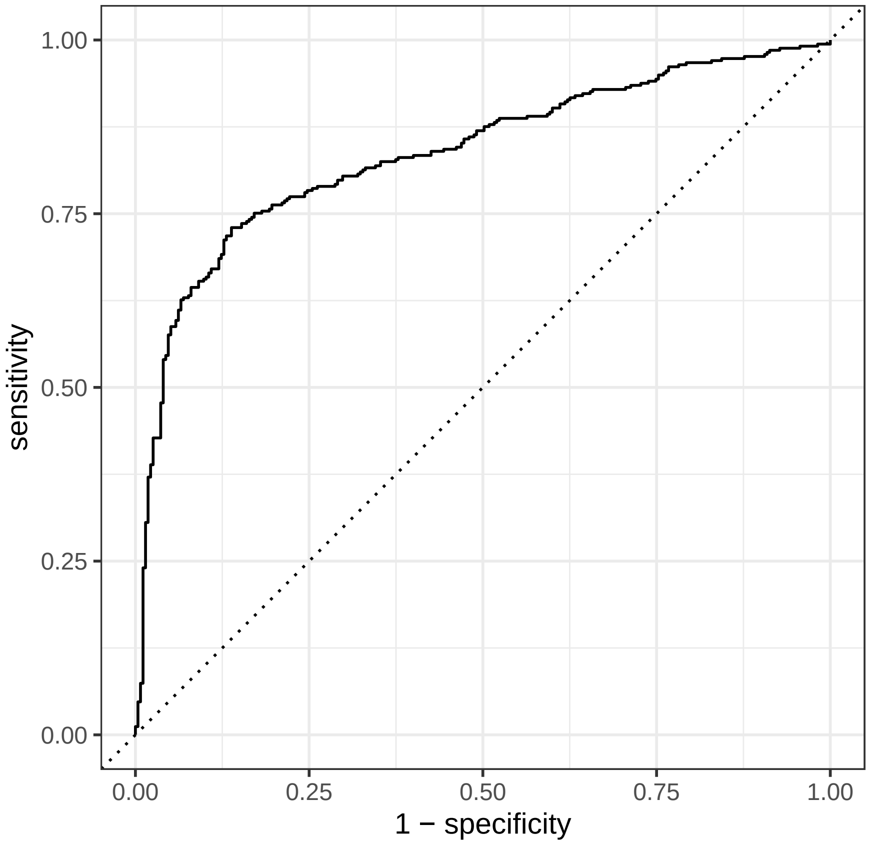
```
# train_svm(recipe_all, training_set, folds)
```

Keep correlating

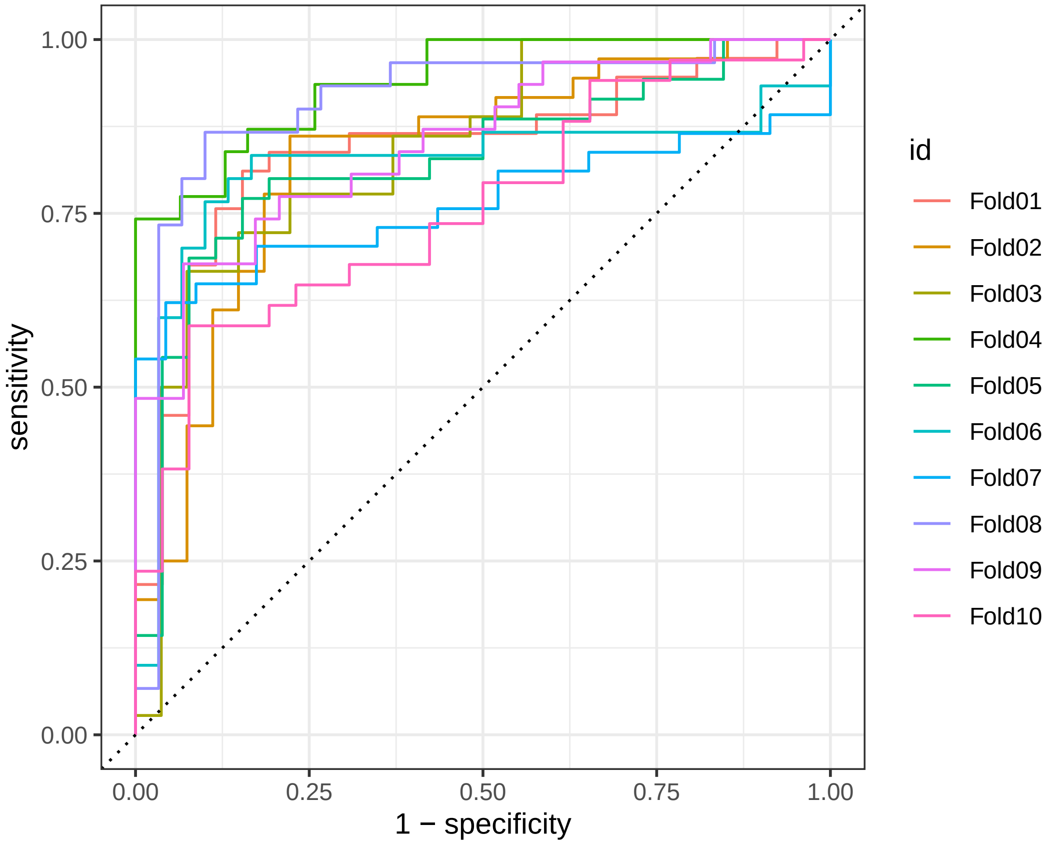
```
model_svm_all <- train_svm(recipe_all_nocorr, training_set, folds)
```

```
## SVM workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: svm_linear()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Linear Support Vector Machine Model Specification (classification)
##
## Computational engine: kernlab
##
## SVM metrics:
```

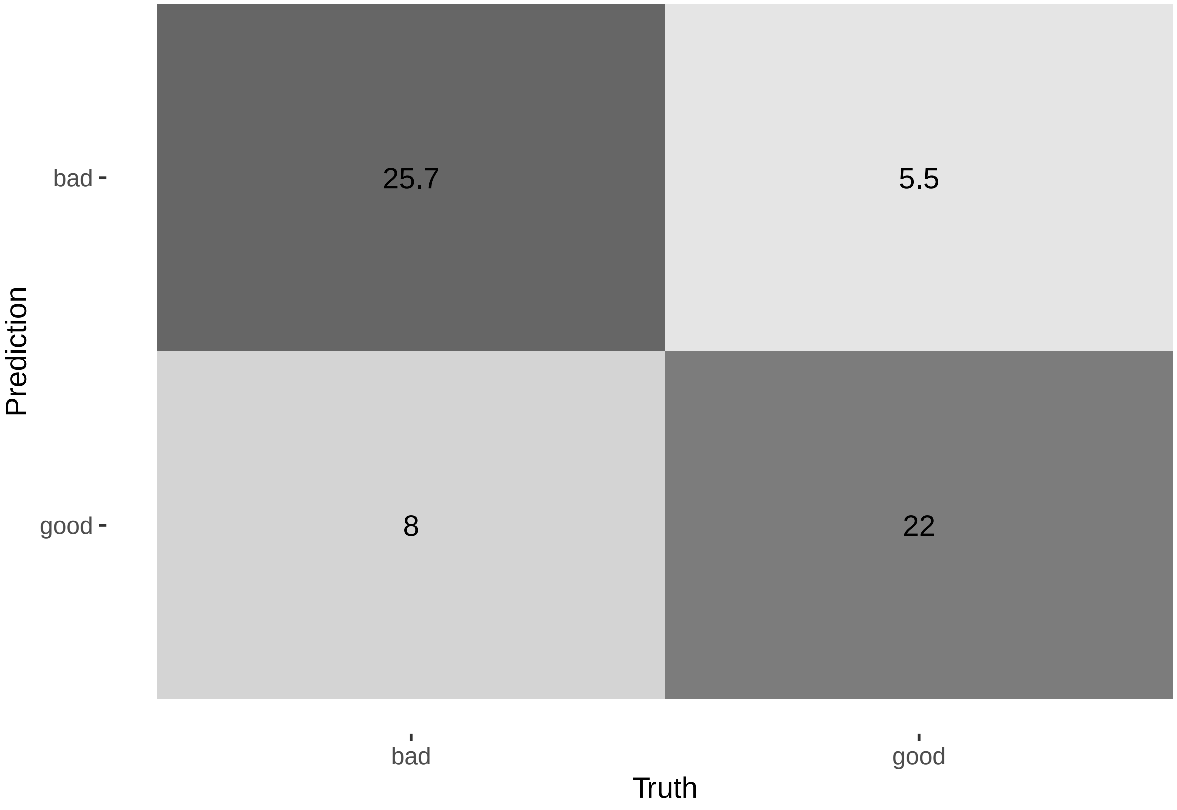
```
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.779    10 0.0174  Preprocessor1_Model1
## 2 brier_class binary     0.167    10 0.00766 Preprocessor1_Model1
## 3 roc_auc     binary     0.839    10 0.0177  Preprocessor1_Model1
```



```
## [1] "\n"
```


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

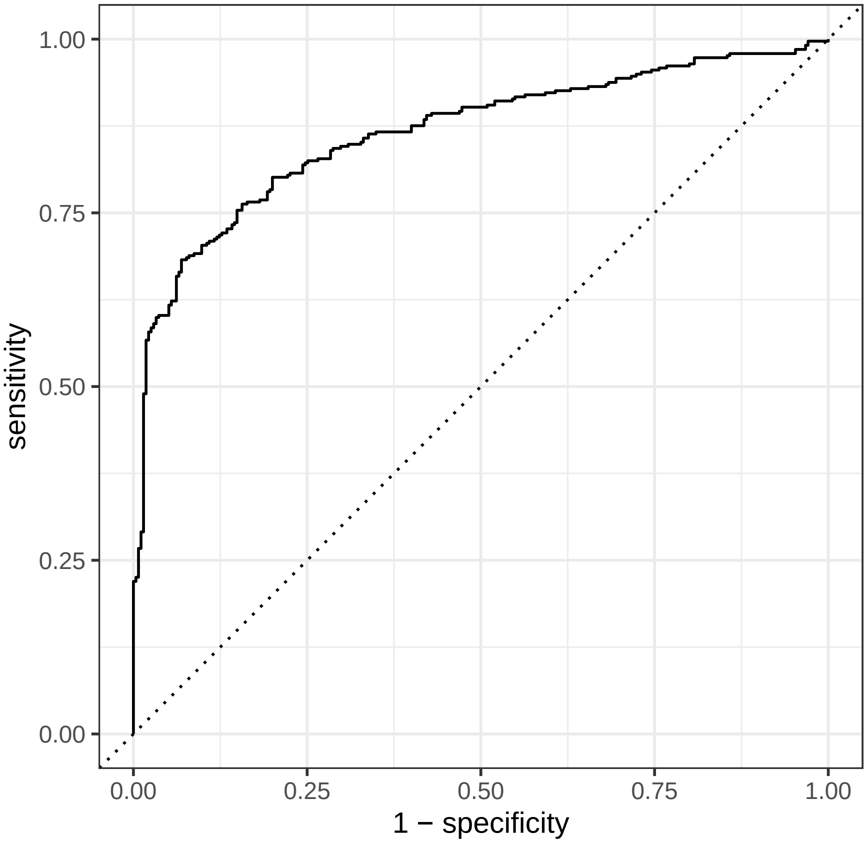
```
## [1] "\n"
```


Prediction
bad
good
25.7
5.5
8
22
bad
good
Truth

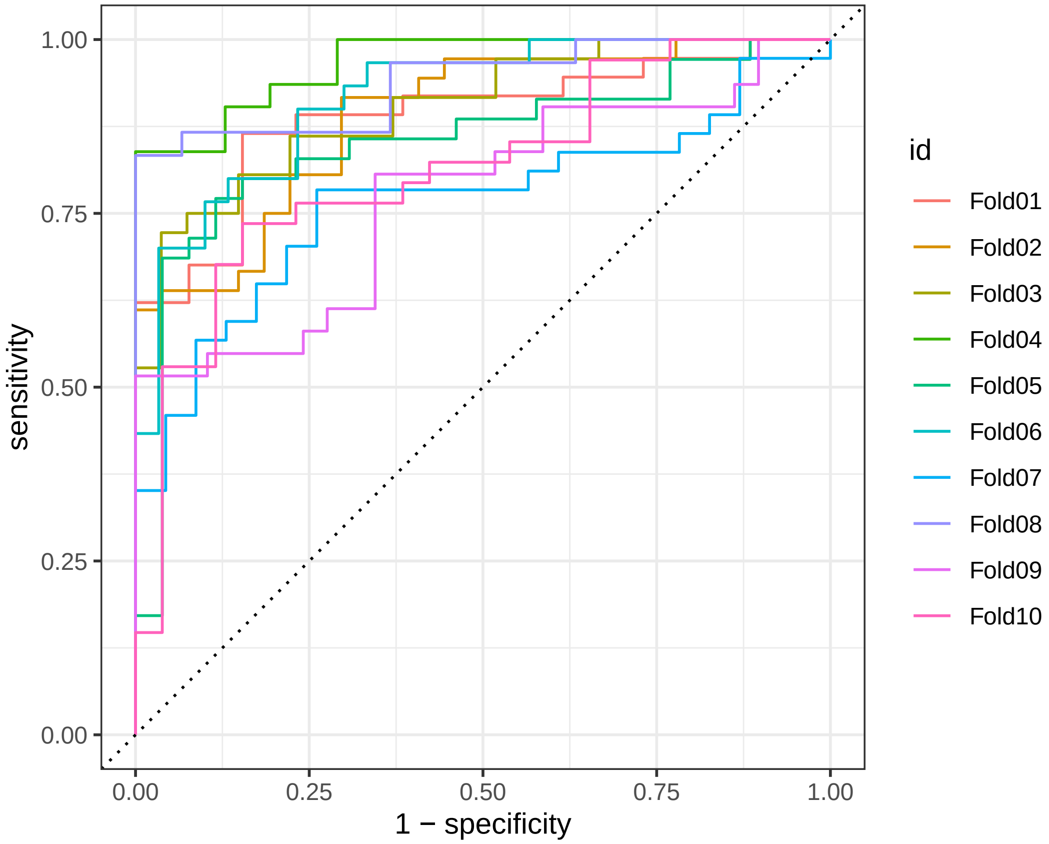
```
## [1] "\n"
```

```
model_svm_rbf_all <- train_svm_rbf(recipe_all_nocorr, training_set, folds)
```

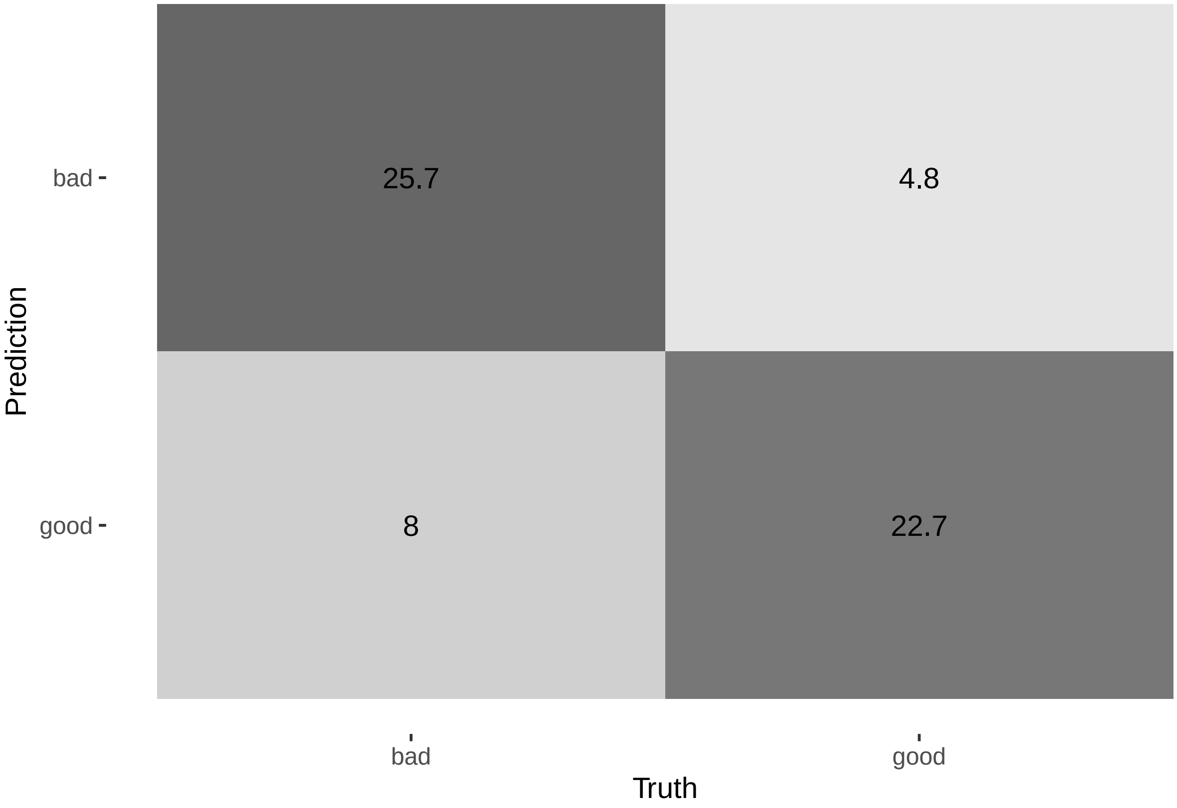
```
## SVM workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: svm_rbf()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Radial Basis Function Support Vector Machine Model Specification (classification)
##
## Computational engine: kernlab
##
## SVM metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.791    10  0.0204 Preprocessor1_Model1
## 2 brier_class binary     0.146    10  0.0123 Preprocessor1_Model1
## 3 roc_auc     binary     0.871    10  0.0215 Preprocessor1_Model1
```



```
## [1] "\n"
```


sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

```
## [1] "\n"
```


Prediction
bad
good
25.7
4.8
8
22.7
bad
good
Truth

```
## [1] "\n"
```

Random forest

All variables

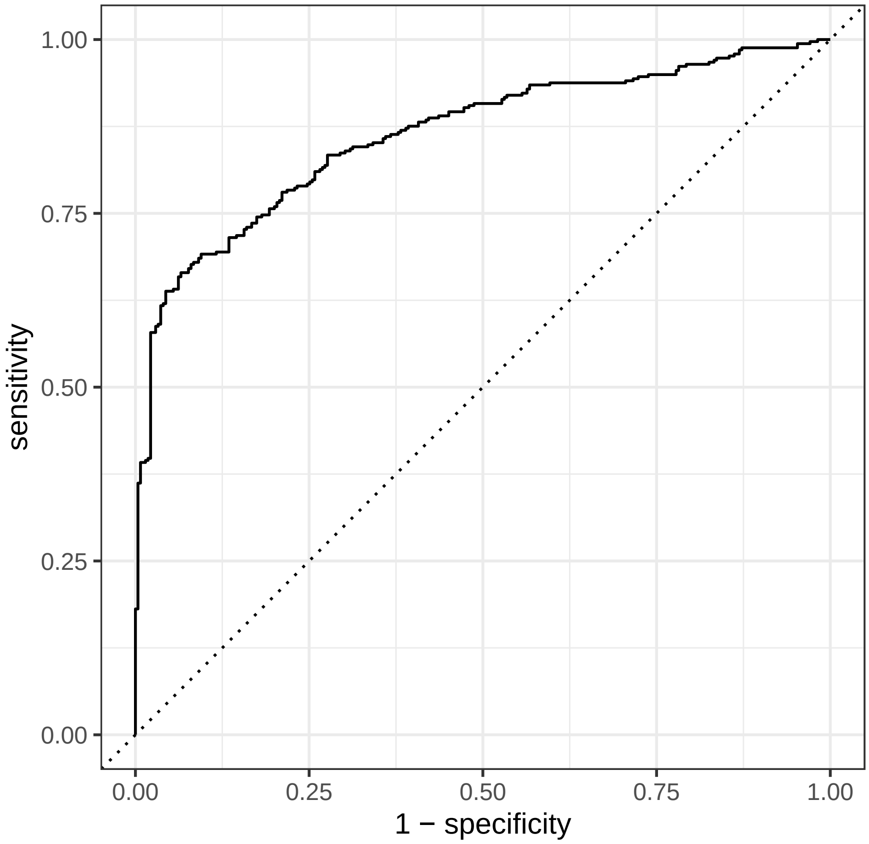
Remove correlating

```
# train_random_forest(recipe_all, training_set, folds)
```

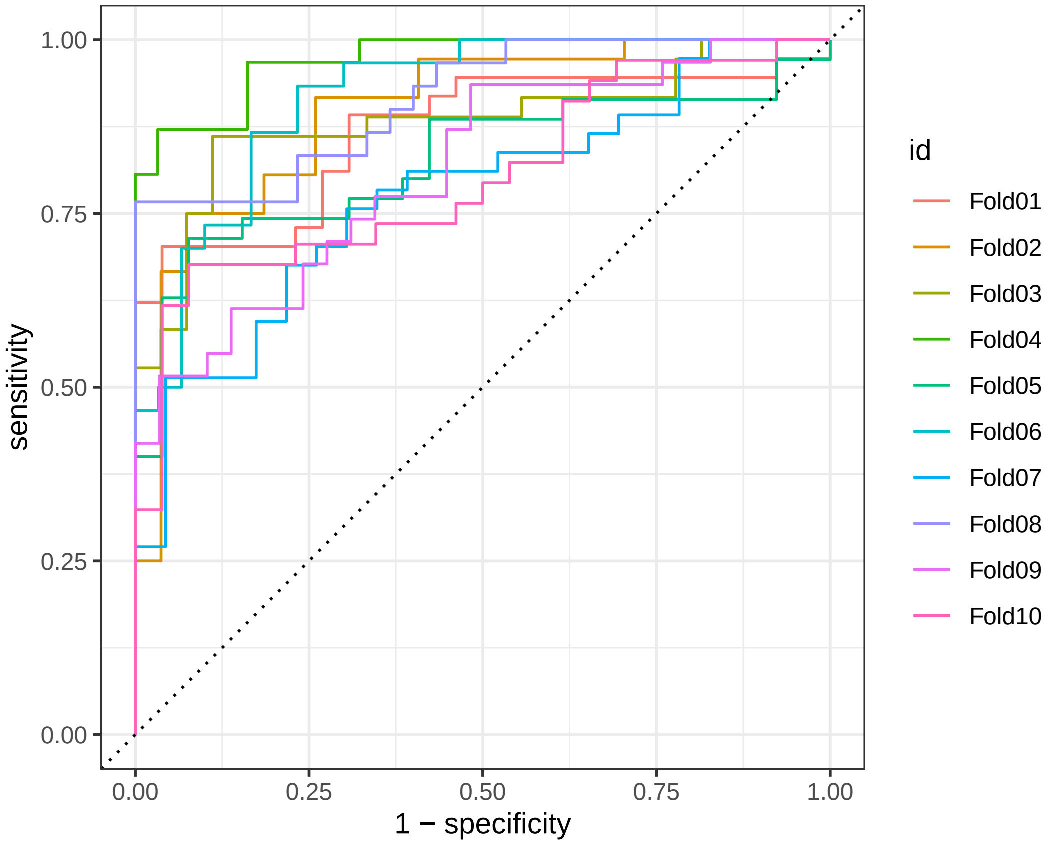
Keep correlating

```
model_rf_all <- train_random_forest(recipe_all_nocorr, training_set, folds)
```

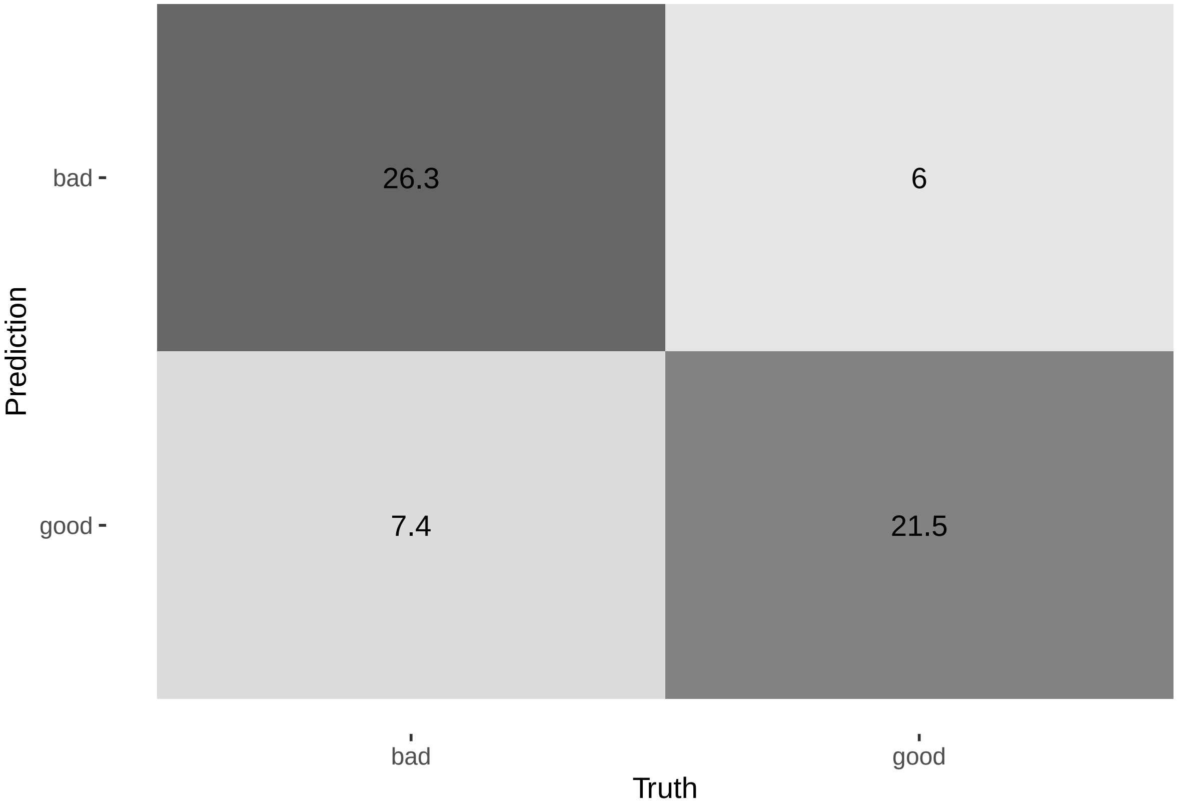
```
## RF workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   trees = 1000
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
##
## RF metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.781    10 0.0180  Preprocessor1_Model1
## 2 brier_class binary     0.149    10 0.00944 Preprocessor1_Model1
## 3 roc_auc     binary     0.867    10 0.0194  Preprocessor1_Model1
```


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity

```
## [1] "\n"
```


sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

```
## [1] "\n"
```

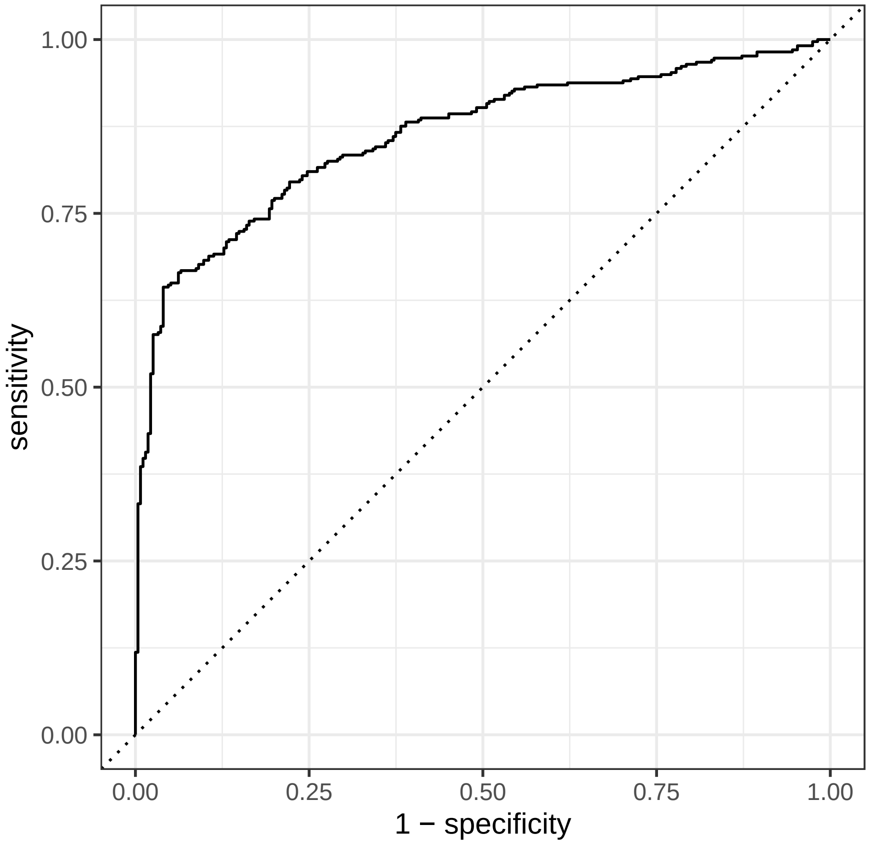

Prediction
bad
good
26.3
6
7.4
21.5
bad
good
Truth

```
## [1] "\n"
## # A tibble: 71 x 2
##    Variable                                            Importance
##    <chr>                                                    <dbl>
##  1 activity                                                12.9
##  2 verb_dist                                               12.2
##  3 RuleTooManyNominalConstructions.max_allowable_nouns     11.8
##  4 RuleLongSentences.max_length                            11.1
##  5 ari                                                     10.3
##  6 RuleTooFewVerbs.min_verb_frac                           10.1
##  7 smog                                                     9.21
##  8 RuleLiteraryStyle                                        8.96
##  9 RulePredAtClauseBeginning.max_order                      8.78
## 10 gf                                                       8.46
## 11 RulePassive                                              6.82
## 12 fkgl                                                     5.75
## 13 mamr                                                     5.49
## 14 RuleMultiPartVerbs                                       5.28
## 15 atl                                                      5.02
## 16 RulePredAtClauseBeginning.max_order.v                    4.77
## 17 maentropy                                                4.36
## 18 mattr                                                    4.09
## 19 RuleTooManyNegations.max_negation_frac                   4.06
## 20 RuleTooManyNominalConstructions.max_noun_frac            3.86
## 21 RuleVerbalNouns                                          3.79
## 22 entropy                                                  3.73
## 23 RuleTooLongExpressions                                   3.69
## 24 RulePredSubjDistance                                     3.53
## 25 RuleAnaphoricReferences                                  3.49
## 26 cli                                                      3.33
## 27 maentropy.v                                              3.27
## 28 RuleCaseRepetition.max_repetition_count.v                3.25
## 29 RuleLongSentences.max_length.v                           3.21
## 30 RulePredSubjDistance.max_distance                        3.17
## 31 mattr.v                                                  3.07
## 32 RuleDoubleAdpos.max_allowable_distance.v                 2.92
## 33 RulePredObjDistance                                      2.77
## 34 RuleTooManyNegations.max_negation_frac.v                 2.76
## 35 word_count                                               2.76
## 36 RuleInfVerbDistance.max_distance                         2.73
## 37 RuleCaseRepetition.max_repetition_frac                   2.71
## 38 RulePredSubjDistance.max_distance.v                      2.69
## 39 RuleMultiPartVerbs.max_distance                          2.57
## 40 RuleCaseRepetition.max_repetition_frac.v                 2.56
## 41 RuleInfVerbDistance.max_distance.v                       2.54
## 42 RuleTooManyNegations.max_allowable_negations.v           2.48
## 43 RuleCaseRepetition.max_repetition_count                  2.40
## 44 RulePredObjDistance.max_distance                         2.37
## 45 RulePredObjDistance.max_distance.v                       2.37
## 46 char_count                                               2.35
## 47 num_hapax                                                2.33
## 48 fre                                                      2.32
## 49 ttr                                                      2.31
## 50 RuleTooManyNegations.max_allowable_negations             2.31
## 51 syllab_count                                              2.24
## 52 RuleInfVerbDistance                                       2.22
## 53 sent_count                                                2.21
## 54 RuleDoubleAdpos                                           2.18
## 55 RuleMultiPartVerbs.max_distance.v                         2.15
## 56 RuleTooManyNominalConstructions.max_noun_frac.v           2.06
## 57 RuleAbstractNouns                                         1.98
## 58 RuleDoubleAdpos.max_allowable_distance                    1.95
## 59 RuleWeakMeaningWords                                      1.77
## 60 RuleReflexivePassWithAnimSubj                             1.58
## 61 hpoint                                                    1.52
## 62 RuleGPwordorder                                           1.48
## 63 RuleGPpatinstr                                            1.24
## 64 RuleGPdeverbaddr                                          1.17
## 65 RuleRelativisticExpressions                               1.03
## 66 RuleGPdeverbsubj                                          0.933
## 67 RuleGPpatbenperson                                        0.843
## 68 RuleGPcoordovs                                            0.830
## 69 RuleConfirmationExpressions                               0.268
## 70 RuleRedundantExpressions                                  0.249
## 71 RuleGPadjective                                           0.216
```

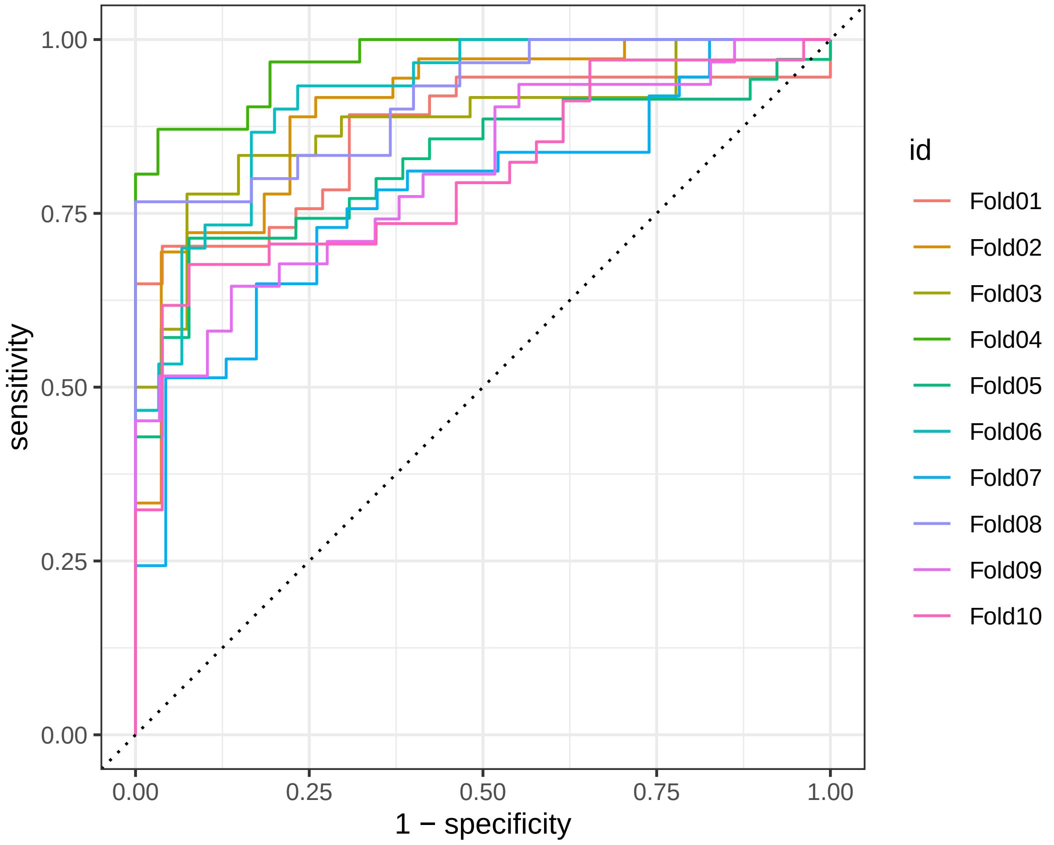
No TL

```
model_rf_notl <- train_random_forest(recipe_notl_nocorr, training_set, folds)
```

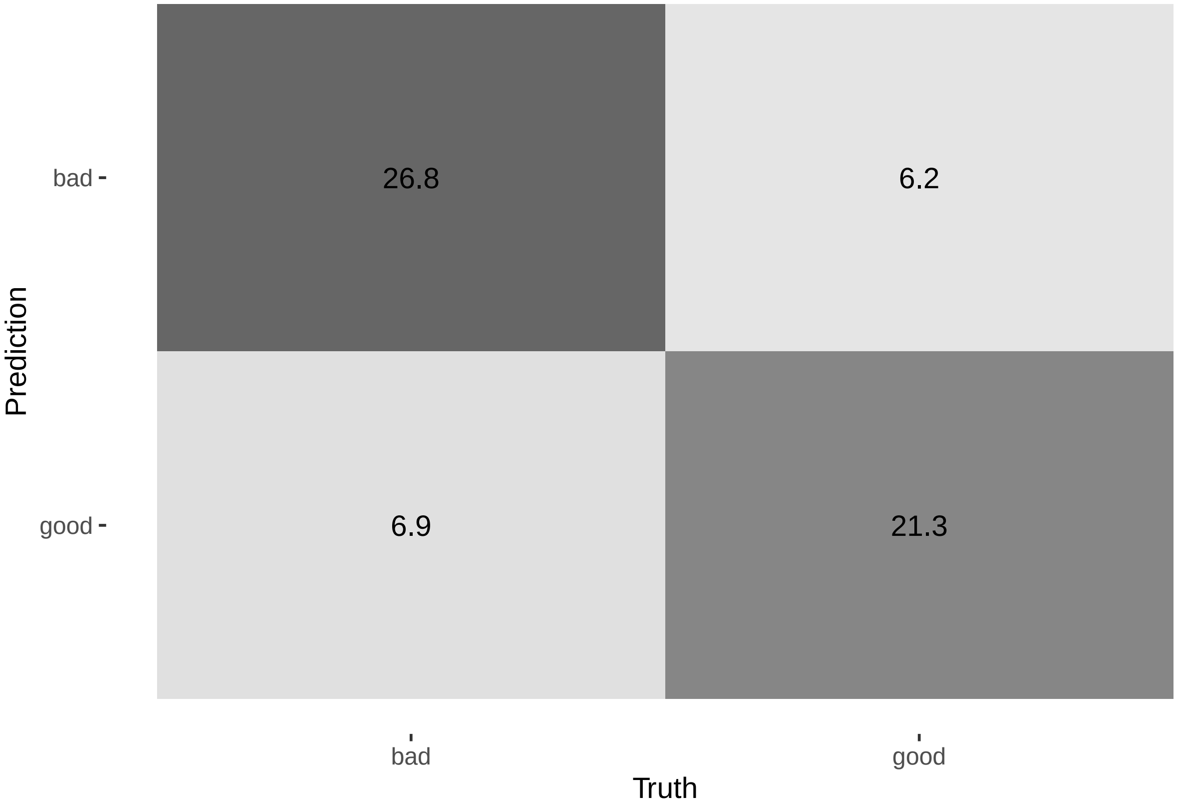
```
## RF workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   trees = 1000
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
##
## RF metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.785    10 0.0202  Preprocessor1_Model1
## 2 brier_class binary     0.150    10 0.00938 Preprocessor1_Model1
## 3 roc_auc     binary     0.866    10 0.0195  Preprocessor1_Model1
```


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity

```
## [1] "\n"
```


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

```
## [1] "\n"
```

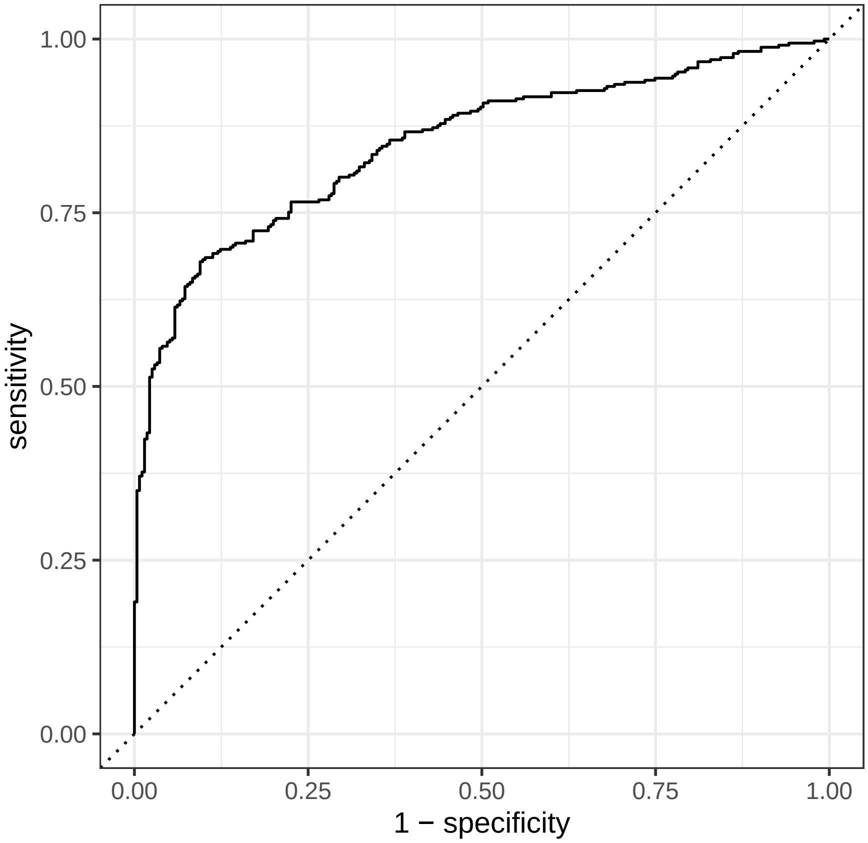

bad
good
Prediction
26.8
6.2
6.9
21.3
bad
good
Truth

```
## [1] "\n"
## # A tibble: 67 x 2
##    Variable                                            Importance
##    <chr>                                                    <dbl>
##  1 activity                                                14.6
##  2 RuleTooManyNominalConstructions.max_allowable_nouns     13.0
##  3 verb_dist                                               11.9
##  4 RuleTooFewVerbs.min_verb_frac                           11.5
##  5 gf                                                      10.1
##  6 ari                                                     10.0
##  7 RuleLongSentences.max_length                             9.48
##  8 smog                                                     9.37
##  9 RuleLiteraryStyle                                        8.57
## 10 RulePredAtClauseBeginning.max_order                      7.91
## 11 RulePassive                                              7.63
## 12 fkgl                                                     5.49
## 13 atl                                                      5.42
## 14 mamr                                                     5.38
## 15 RuleMultiPartVerbs                                       4.62
## 16 maentropy                                                4.55
## 17 RuleTooManyNegations.max_negation_frac                   4.52
## 18 RulePredAtClauseBeginning.max_order.v                    4.44
## 19 RuleVerbalNouns                                          4.34
## 20 mattr                                                    4.28
## 21 entropy                                                  4.12
## 22 RuleAnaphoricReferences                                  4.05
## 23 RuleTooLongExpressions                                   4.03
## 24 RuleTooManyNominalConstructions.max_noun_frac            3.80
## 25 RulePredSubjDistance                                     3.73
## 26 maentropy.v                                              3.40
## 27 cli                                                      3.38
## 28 RuleLongSentences.max_length.v                           3.35
## 29 RuleDoubleAdpos.max_allowable_distance.v                 3.30
## 30 RulePredSubjDistance.max_distance                        3.27
## 31 mattr.v                                                  2.98
## 32 RuleTooManyNegations.max_negation_frac.v                 2.96
## 33 num_hapax                                                2.95
## 34 RuleCaseRepetition.max_repetition_frac.v                 2.89
## 35 RulePredSubjDistance.max_distance.v                      2.87
## 36 RuleCaseRepetition.max_repetition_count.v                2.86
## 37 RuleInfVerbDistance.max_distance                         2.79
## 38 RulePredObjDistance                                      2.76
## 39 RuleCaseRepetition.max_repetition_frac                   2.75
## 40 RuleInfVerbDistance.max_distance.v                       2.73
## 41 RuleCaseRepetition.max_repetition_count                  2.72
## 42 RuleTooManyNegations.max_allowable_negations             2.71
## 43 ttr                                                      2.67
## 44 RuleMultiPartVerbs.max_distance                          2.61
## 45 RulePredObjDistance.max_distance                         2.53
## 46 RuleTooManyNegations.max_allowable_negations.v           2.50
## 47 RuleMultiPartVerbs.max_distance.v                        2.49
## 48 RulePredObjDistance.max_distance.v                       2.33
## 49 RuleInfVerbDistance                                      2.23
## 50 RuleDoubleAdpos                                          2.21
## 51 RuleDoubleAdpos.max_allowable_distance                   2.16
## 52 hpoint                                                   2.09
## 53 fre                                                      2.09
## 54 RuleAbstractNouns                                        2.05
## 55 RuleTooManyNominalConstructions.max_noun_frac.v          2.00
## 56 RuleWeakMeaningWords                                     1.88
## 57 RuleReflexivePassWithAnimSubj                            1.60
## 58 RuleGPwordorder                                          1.58
## 59 RuleGPdeverbaddr                                         1.32
## 60 RuleGPpatinstr                                           1.28
## 61 RuleRelativisticExpressions                              0.966
## 62 RuleGPdeverbsubj                                         0.905
## 63 RuleGPpatbenperson                                       0.820
## 64 RuleGPcoordovs                                           0.811
## 65 RuleRedundantExpressions                                 0.339
## 66 RuleGPadjective                                          0.305
## 67 RuleConfirmationExpressions                              0.305
```

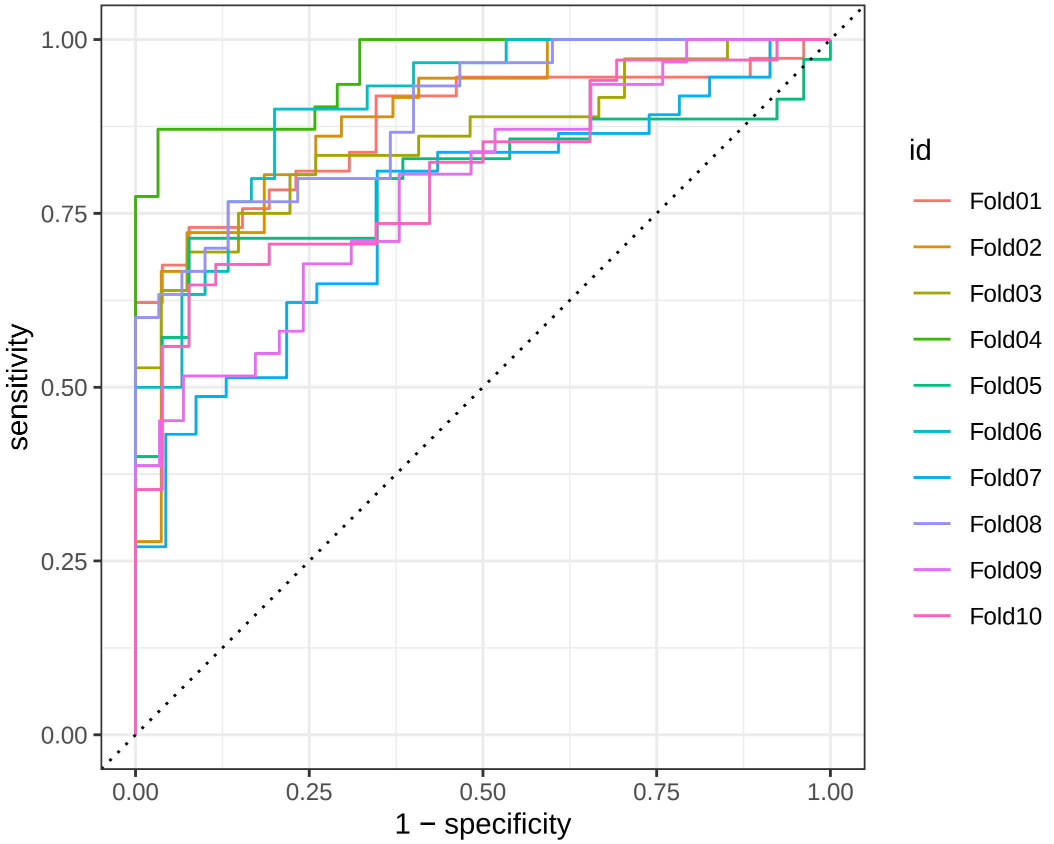
IAC

```
model_rf_iac <- train_random_forest(recipe_iac_nocorr, training_set, folds)
```

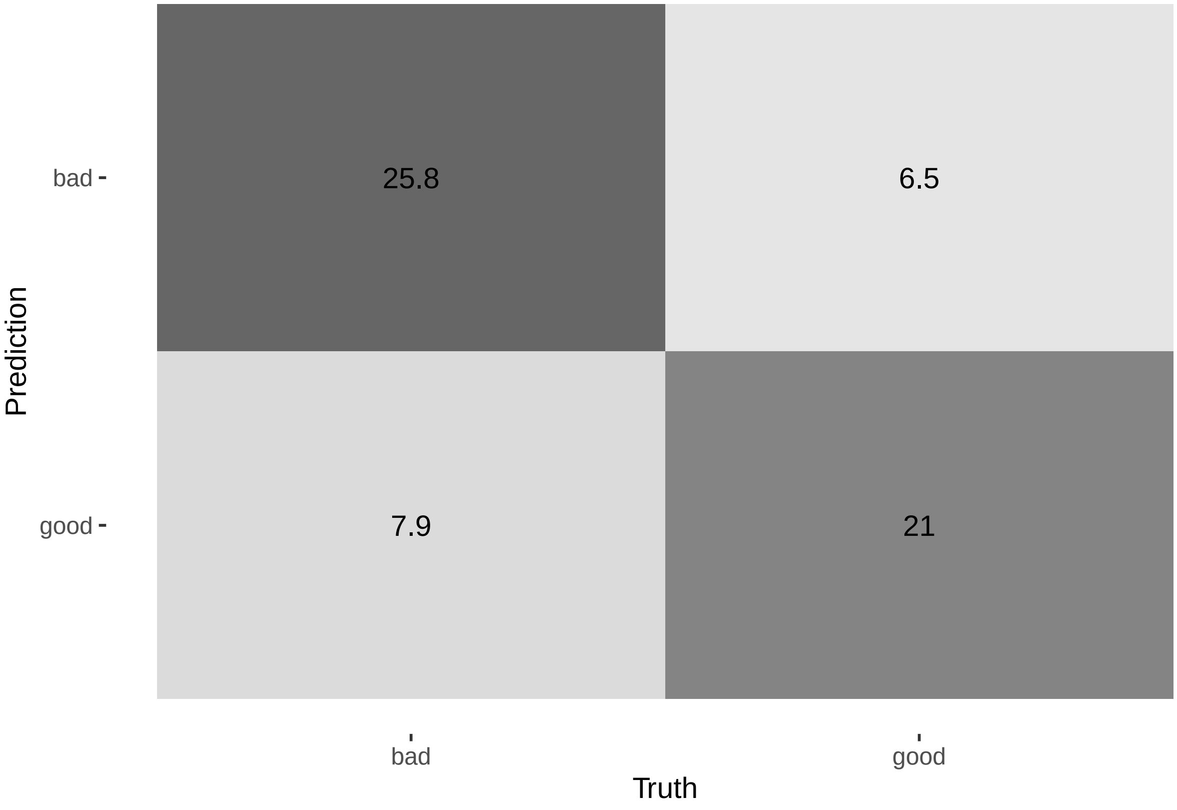
```
## RF workflow:
## == Workflow ====================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ----------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   trees = 1000
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
##
## RF metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.764    10 0.0159  Preprocessor1_Model1
## 2 brier_class binary     0.156    10 0.00897 Preprocessor1_Model1
## 3 roc_auc     binary     0.853    10 0.0200  Preprocessor1_Model1
```


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 − specificity

```
## [1] "\n"
```


sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

```
## [1] "\n"
```


Prediction
bad
good
25.8
6.5
7.9
21
bad
good
Truth

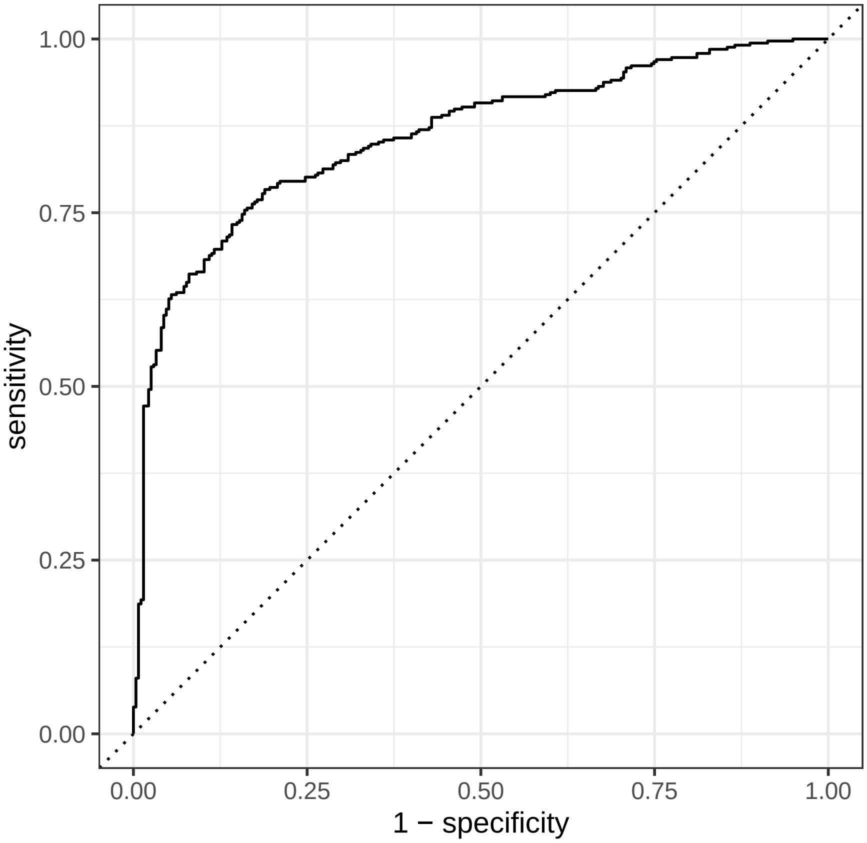
```
## [1] "\n"
## # A tibble: 44 x 2
##    Variable                                              Importance
##    <chr>                                                      <dbl>
##  1 RuleTooManyNominalConstructions.max_allowable_nouns        15.5
##  2 activity                                                   15.5
##  3 verb_dist                                                  15.1
##  4 RuleTooFewVerbs.min_verb_frac                              13.2
##  5 RuleLongSentences.max_length                               12.1
##  6 smog                                                       11.3
##  7 gf                                                         11.0
##  8 ari                                                        10.4
##  9 RulePredAtClauseBeginning.max_order                         9.69
## 10 mamr                                                        6.56
## 11 atl                                                         6.47
## 12 fkgl                                                        6.17
## 13 RuleTooManyNegations.max_negation_frac                      6.02
## 14 entropy                                                     5.96
## 15 RuleTooManyNominalConstructions.max_noun_frac               5.76
## 16 maentropy                                                   5.58
## 17 mattr                                                       5.47
## 18 RulePredAtClauseBeginning.max_order.v                       5.26
## 19 cli                                                         5.06
## 20 RuleTooManyNominalConstructions.max_allowable_nouns.v       4.69
## 21 maentropy.v                                                 4.68
## 22 RuleLongSentences.max_length.v                              4.63
## 23 RuleDoubleAdpos.max_allowable_distance.v                    4.53
## 24 mattr.v                                                     4.37
## 25 RulePredSubjDistance.max_distance                           4.07
## 26 RuleTooManyNegations.max_negation_frac.v                    4.07
## 27 RuleInfVerbDistance.max_distance.v                          4.03
## 28 RuleInfVerbDistance.max_distance                            4.01
## 29 ttr                                                         4.00
## 30 RuleCaseRepetition.max_repetition_count.v                   3.96
## 31 RulePredSubjDistance.max_distance.v                         3.67
## 32 RuleMultiPartVerbs.max_distance                             3.66
## 33 RuleTooManyNegations.max_allowable_negations               3.65
## 34 RuleCaseRepetition.max_repetition_frac                      3.62
## 35 RulePredObjDistance.max_distance                            3.57
## 36 RuleCaseRepetition.max_repetition_frac.v                    3.56
## 37 RuleCaseRepetition.max_repetition_count                     3.46
## 38 RuleMultiPartVerbs.max_distance.v                           3.46
## 39 RuleTooManyNegations.max_allowable_negations.v              3.46
## 40 fre                                                         3.42
## 41 RulePredObjDistance.max_distance.v                          3.32
## 42 hpoint                                                      3.09
## 43 RuleTooManyNominalConstructions.max_noun_frac.v             2.85
## 44 RuleDoubleAdpos.max_allowable_distance                      2.73
```

Counts

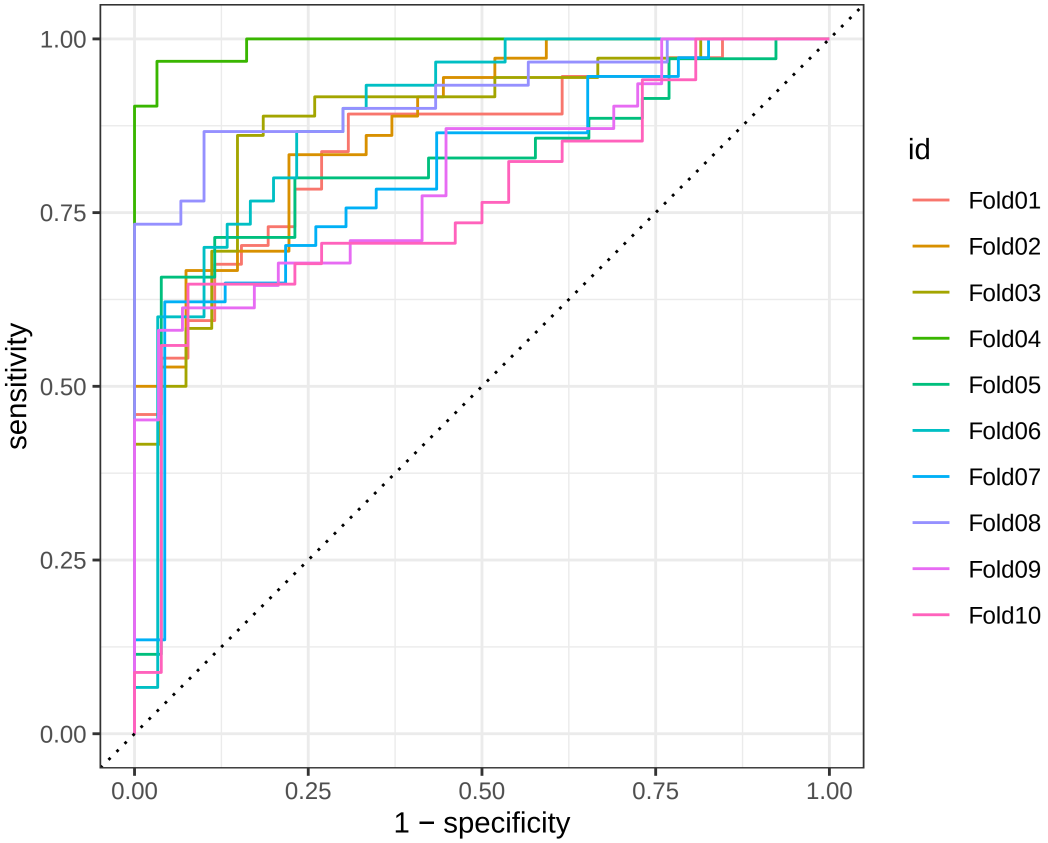
```
model_rf_counts <- train_random_forest(recipe_counts_nocorr, training_set, folds)
```

```
## RF workflow:
```

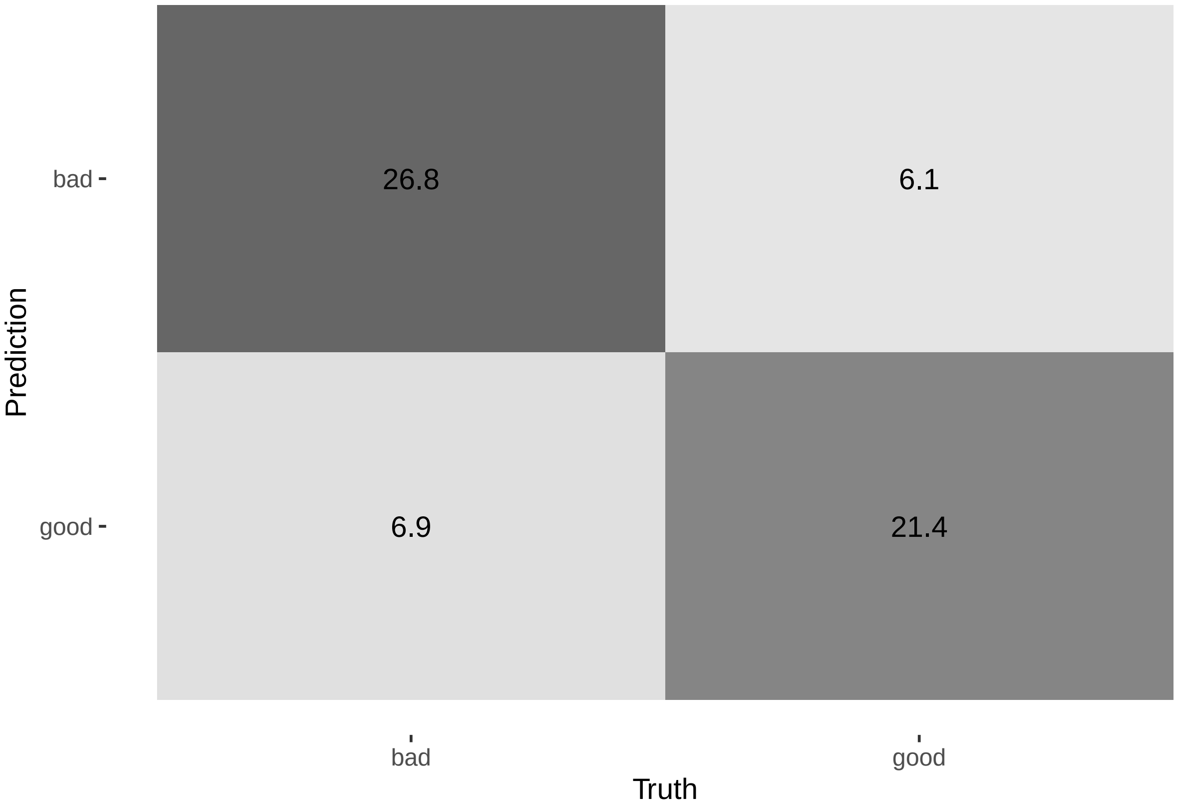
```
## == Workflow ===================================================================
## Preprocessor: Recipe
## Model: rand_forest()
##
## -- Preprocessor ---------------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model ----------------------------------------------------------------------
## Random Forest Model Specification (classification)
##
## Main Arguments:
##   trees = 1000
##
## Engine-Specific Arguments:
##   importance = impurity
##
## Computational engine: ranger
##
## RF metrics:
## # A tibble: 3 x 6
##   .metric     .estimator  mean     n std_err .config
##   <chr>       <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy    binary     0.787    10 0.0199  Preprocessor1_Model1
## 2 brier_class binary     0.155    10 0.00814 Preprocessor1_Model1
## 3 roc_auc     binary     0.862    10 0.0207  Preprocessor1_Model1
```



```
## [1] "\n"
```


sensitivity
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1 − specificity
id
Fold01
Fold02
Fold03
Fold04
Fold05
Fold06
Fold07
Fold08
Fold09
Fold10

```
## [1] "\n"
```



```
## [1] "\n"
## # A tibble: 28 x 2
##    Variable                         Importance
##    <chr>                                 <dbl>
##  1 RuleMultiPartVerbs                    30.7
##  2 RuleLiteraryStyle                     28.3
##  3 RulePassive                           28.0
##  4 RulePredSubjDistance                  20.0
##  5 RuleInfVerbDistance                   15.2
##  6 sent_count                            12.7
##  7 RuleVerbalNouns                       11.6
##  8 word_count                            10.6
##  9 num_hapax                              8.93
## 10 char_count                             8.75
## 11 RuleTooLongExpressions                 8.48
## 12 RulePredObjDistance                    8.26
## 13 syllab_count                           8.26
## 14 RuleDoubleAdpos                        7.74
## 15 RuleAbstractNouns                      6.96
## 16 RuleAnaphoricReferences                6.64
## 17 RuleGPwordorder                        6.49
## 18 RuleWeakMeaningWords                   5.91
## 19 RuleReflexivePassWithAnimSubj          5.76
## 20 RuleGPdeverbsubj                       3.72
## 21 RuleGPpatinstr                         3.42
## 22 RuleGPdeverbaddr                       2.99
## 23 RuleGPpatbenperson                     2.16
## 24 RuleGPcoordovs                         1.86
## 25 RuleRelativisticExpressions        1.84
## 26 RuleConfirmationExpressions        1.36
## 27 RuleRedundantExpressions           0.550
## 28 RuleGPadjective                    0.550
```

Evaluations

Decision tree

All variables

```
evaluate_decision_tree(model_dt_all, evaluation_set)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   68   21
##       good  21   44
##
##                Accuracy : 0.7273
##                  95% CI : (0.6497, 0.7958)
##     No Information Rate : 0.5779
##     P-Value [Acc > NIR] : 8.678e-05
##
##                   Kappa : 0.441
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.6769
##             Specificity : 0.7640
##          Pos Pred Value : 0.6769
##          Neg Pred Value : 0.7640
##              Prevalence : 0.4221
##          Detection Rate : 0.2857
##    Detection Prevalence : 0.4221
##       Balanced Accuracy : 0.7205
##
##        'Positive' Class : good
##
```

No TL

```
evaluate_decision_tree(model_dt_notl, evaluation_set)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   68   21
##       good  21   44
##
##                Accuracy : 0.7273
##                  95% CI : (0.6497, 0.7958)
##     No Information Rate : 0.5779
##     P-Value [Acc > NIR] : 8.678e-05
##
##                   Kappa : 0.441
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.6769
##             Specificity : 0.7640
##          Pos Pred Value : 0.6769
##          Neg Pred Value : 0.7640
##              Prevalence : 0.4221
##          Detection Rate : 0.2857
##    Detection Prevalence : 0.4221
##       Balanced Accuracy : 0.7205
##
##        'Positive' Class : good
##
```

IAC

```
evaluate_decision_tree(model_dt_iac, evaluation_set)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   62   21
##       good  27   44
##
##                Accuracy : 0.6883
##                  95% CI : (0.6088, 0.7604)
##     No Information Rate : 0.5779
##     P-Value [Acc > NIR] : 0.003172
##
##                   Kappa : 0.369
##
##  Mcnemar's Test P-Value : 0.470486
##
##             Sensitivity : 0.6769
##             Specificity : 0.6966
##          Pos Pred Value : 0.6197
##          Neg Pred Value : 0.7470
##              Prevalence : 0.4221
##          Detection Rate : 0.2857
##    Detection Prevalence : 0.4610
##       Balanced Accuracy : 0.6868
##
##        'Positive' Class : good
##
```

Counts

```r
evaluate_decision_tree(model_dt_counts, evaluation_set)
```

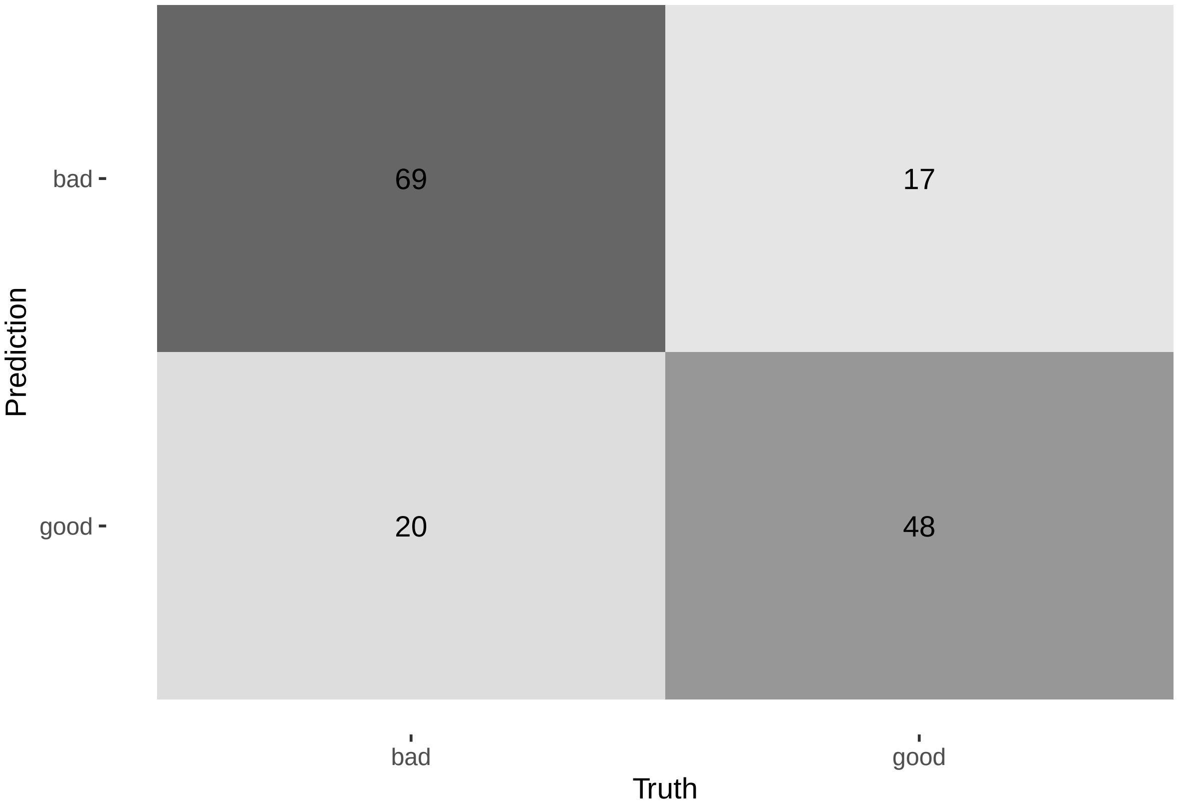
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   65   16
##       good  24   49
##
##                Accuracy : 0.7403
##                  95% CI : (0.6635, 0.8075)
##     No Information Rate : 0.5779
##     P-Value [Acc > NIR] : 2.051e-05
##
##                   Kappa : 0.4763
##
##  Mcnemar's Test P-Value : 0.2684
##
##             Sensitivity : 0.7538
##             Specificity : 0.7303
##          Pos Pred Value : 0.6712
##          Neg Pred Value : 0.8025
##              Prevalence : 0.4221
##          Detection Rate : 0.3182
##    Detection Prevalence : 0.4740
##       Balanced Accuracy : 0.7421
##
##        'Positive' Class : good
##
```

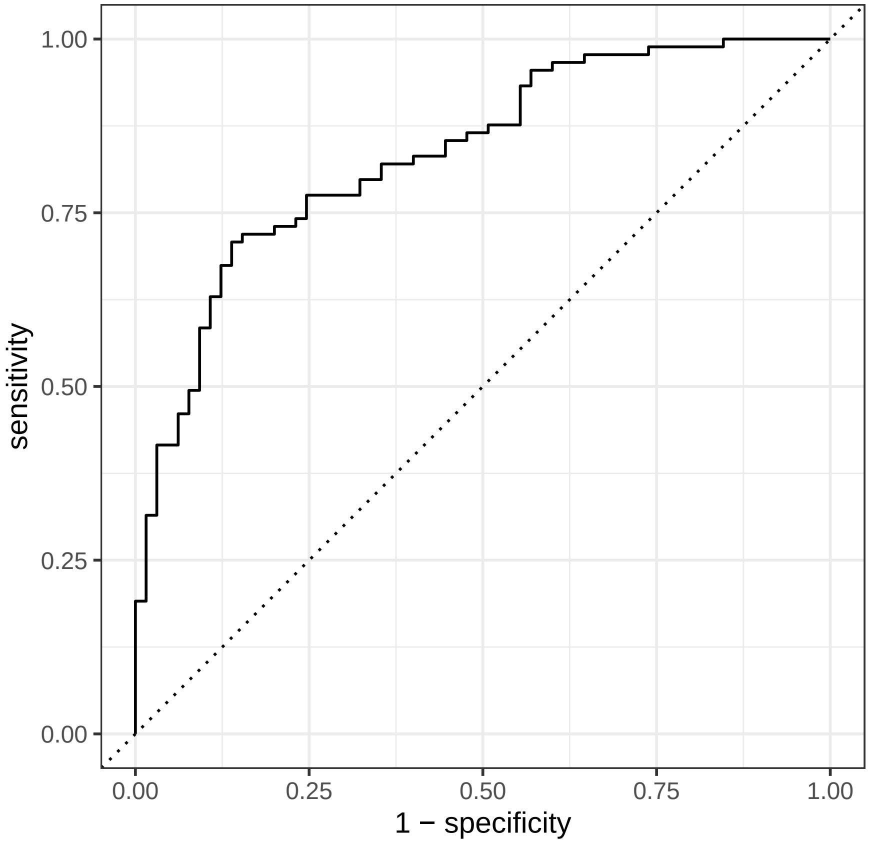
Lasso

All

```r
lfit_lasso_all <- model_lasso_all %>% evaluate_tidymodel(split)
```

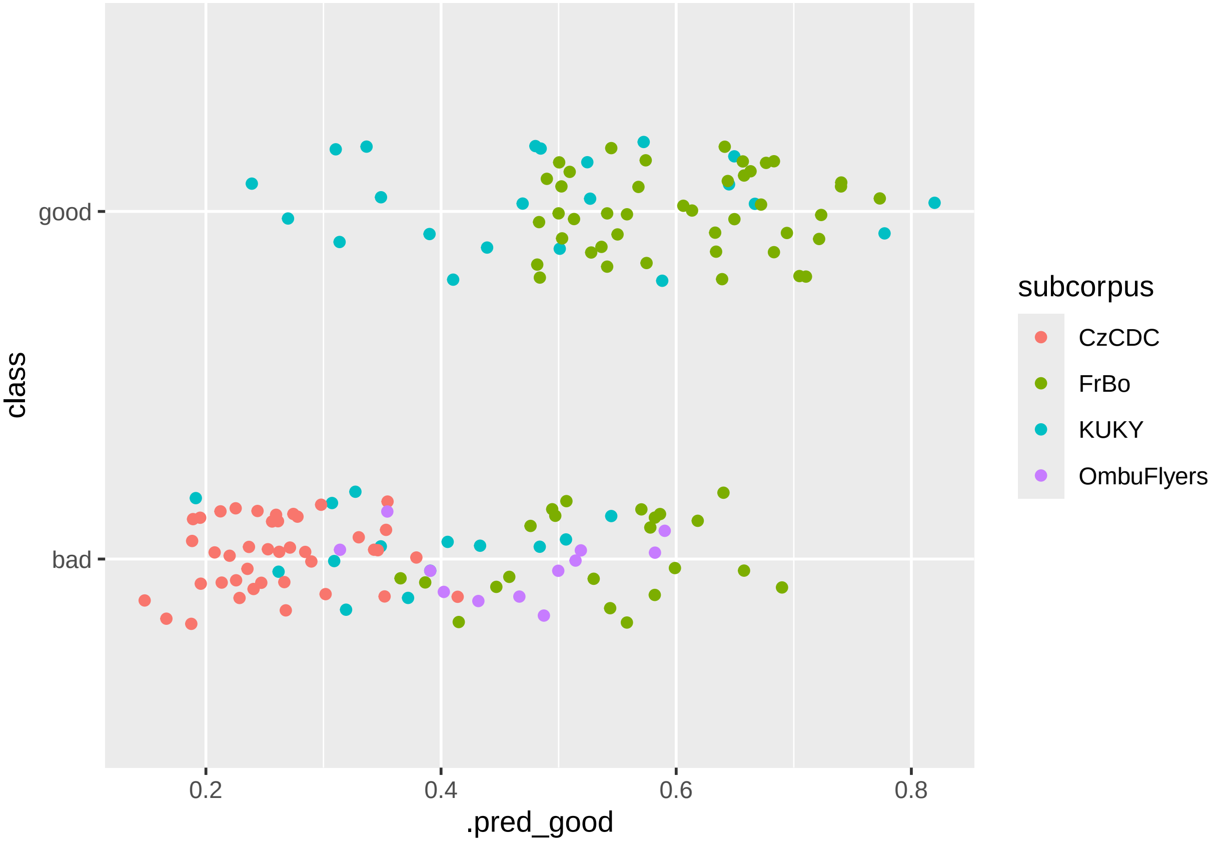
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.760 Preprocessor1_Model1
## 2 roc_auc     binary         0.835 Preprocessor1_Model1
## 3 brier_class binary         0.178 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 71 x 3
##    Variable                                            Importance Sign
##    <chr>                                                    <dbl> <chr>
##  1 activity                                              0.408    POS
##  2 smog                                                  0.191    NEG
##  3 RuleLiteraryStyle                                     0.168    NEG
##  4 atl                                                   0.100    POS
##  5 mamr                                                  0.0576   POS
##  6 gf                                                    0.0184   NEG
##  7 entropy                                               0.0165   NEG
##  8 maentropy                                             0.00435  NEG
##  9 ari                                                   0.000272 NEG
## 10 RuleGPcoordovs                                        0        NEG
## 11 RuleGPdeverbaddr                                      0        NEG
## 12 RuleGPpatinstr                                        0        NEG
## 13 RuleGPdeverbsubj                                      0        NEG
## 14 RuleGPadjective                                       0        NEG
## 15 RuleGPpatbenperson                                    0        NEG
## 16 RuleGPwordorder                                       0        NEG
## 17 RuleDoubleAdpos                                       0        NEG
## 18 RuleDoubleAdpos.max_allowable_distance                0        NEG
## 19 RuleDoubleAdpos.max_allowable_distance.v              0        NEG
## 20 RuleReflexivePassWithAnimSubj                         0        NEG
## 21 RuleTooFewVerbs.min_verb_frac                         0        NEG
## 22 RuleTooManyNegations.max_negation_frac                0        NEG
## 23 RuleTooManyNegations.max_negation_frac.v              0        NEG
## 24 RuleTooManyNegations.max_allowable_negations          0        NEG
## 25 RuleTooManyNegations.max_allowable_negations.v        0        NEG
## 26 RuleTooManyNominalConstructions.max_noun_frac         0        NEG
## 27 RuleTooManyNominalConstructions.max_noun_frac.v       0        NEG
## 28 RuleTooManyNominalConstructions.max_allowable_nouns   0        NEG
## 29 RuleCaseRepetition.max_repetition_count               0        NEG
## 30 RuleCaseRepetition.max_repetition_count.v             0        NEG
## 31 RuleCaseRepetition.max_repetition_frac                0        NEG
## 32 RuleCaseRepetition.max_repetition_frac.v              0        NEG
## 33 RuleWeakMeaningWords                                  0        NEG
## 34 RuleAbstractNouns                                     0        NEG
## 35 RuleRelativisticExpressions                           0        NEG
## 36 RuleConfirmationExpressions                           0        NEG
## 37 RuleRedundantExpressions                              0        NEG
## 38 RuleTooLongExpressions                                0        NEG
## 39 RuleAnaphoricReferences                               0        NEG
## 40 RulePassive                                           0        NEG
## 41 RulePredSubjDistance                                  0        NEG
## 42 RulePredSubjDistance.max_distance                     0        NEG
## 43 RulePredSubjDistance.max_distance.v                   0        NEG
## 44 RulePredObjDistance                                   0        NEG
## 45 RulePredObjDistance.max_distance                      0        NEG
## 46 RulePredObjDistance.max_distance.v                    0        NEG
## 47 RuleInfVerbDistance                                   0        NEG
## 48 RuleInfVerbDistance.max_distance                      0        NEG
## 49 RuleInfVerbDistance.max_distance.v                    0        NEG
## 50 RuleMultiPartVerbs                                    0        NEG
## 51 RuleMultiPartVerbs.max_distance                       0        NEG
## 52 RuleMultiPartVerbs.max_distance.v                     0        NEG
## 53 RuleLongSentences.max_length                          0        NEG
## 54 RuleLongSentences.max_length.v                        0        NEG
## 55 RulePredAtClauseBeginning.max_order                   0        NEG
## 56 RulePredAtClauseBeginning.max_order.v                 0        NEG
## 57 RuleVerbalNouns                                       0        NEG
## 58 sent_count                                            0        NEG
## 59 word_count                                            0        NEG
## 60 syllab_count                                          0        NEG
## 61 char_count                                            0        NEG
## 62 cli                                                   0        NEG
## 63 num_hapax                                             0        NEG
## 64 ttr                                                   0        NEG
## 65 mattr                                                 0        NEG
## 66 mattr.v                                               0        NEG
## 67 maentropy.v                                           0        NEG
## 68 verb_dist                                             0        NEG
## 69 hpoint                                                0        NEG
## 70 fre                                                   0        NEG
## 71 fkgl                                                  0        NEG
```

```
lfit_lasso_all %>% get_mismatch_details(data)
```



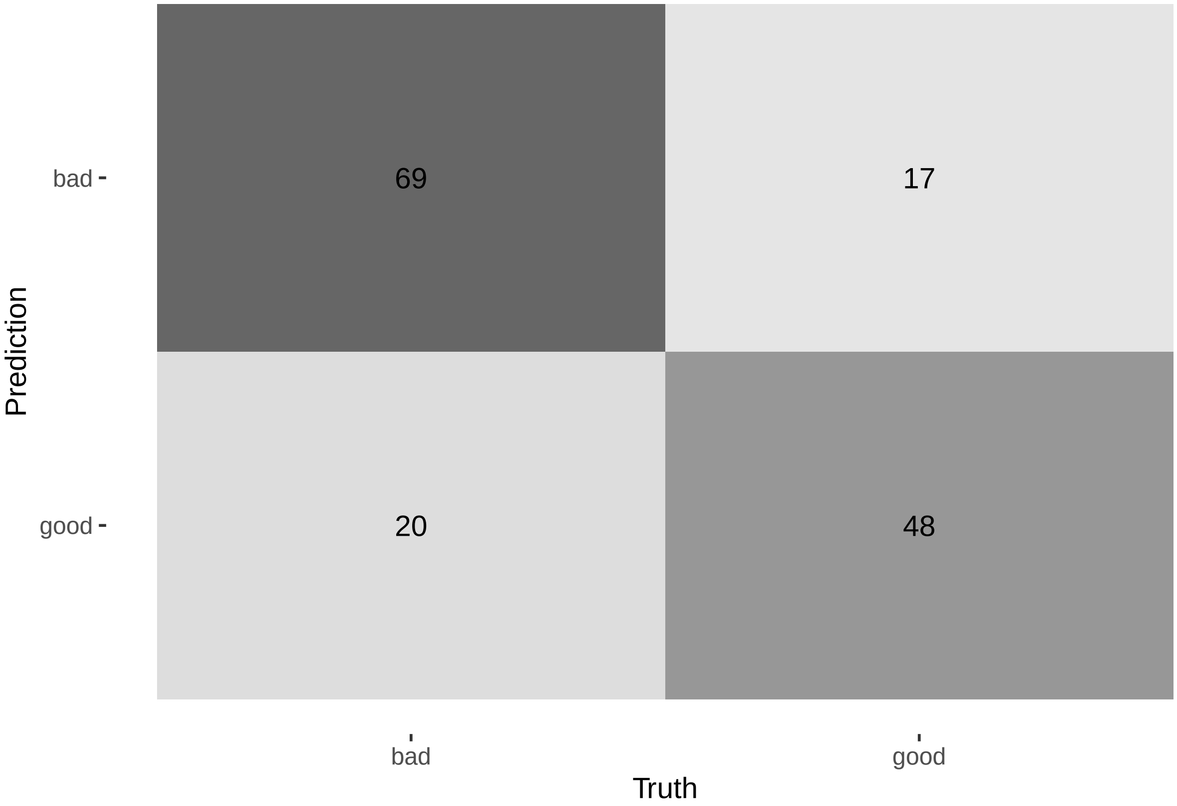
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   41    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##           class
## .pred_class bad good
##        bad    8    5
##        good  14   38
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##        bad   12   12
##        good   2   10
##
## , , subcorpus = OmbuFlyers
##
##           class
## .pred_class bad good
##        bad    8    0
##        good   4    0
##
##
## Greatest deviations:
## # A tibble: 37 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.261  bad         good  KUKY       Odvolani_proti_rozhodnuti_o_nepov~
##  2        0.230  bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  3        0.190  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  4        0.190  bad         good  KUKY       MV_Odneti_trvaleho_pobytu_Kru_po
##  5        0.186  bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
##  6        0.163  bad         good  KUKY       Odvolani
##  7        0.158  good        bad   FrBo       orig_Co je to EIA_final
##  8        0.151  bad         good  KUKY       AK_JH_Podani_US_podpis
##  9        0.140  good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
## 10        0.118  good        bad   FrBo       orig_znalci, znalecké posudky
## 11        0.110  bad         good  KUKY       invalidní důchod_1399-23_původní
## 12        0.0989 good        bad   FrBo       64
## 13        0.0902 good        bad   OmbuFlyers Soudni-poplatky
## 14        0.0897 bad         good  KUKY       Ockovani_JSm
## 15        0.0862 good        bad   FrBo       orig_Sousedské vztahy
## 16        0.0819 good        bad   OmbuFlyers Detsky-domov
## 17        0.0819 good        bad   FrBo       orig_Jak probíhá správní řízení
## 18        0.0818 good        bad   FrBo       orig_Jak zajistit, aby skládka do~
## 19        0.0780 good        bad   FrBo       orig_územní řízení
## 20        0.0704 good        bad   FrBo       orig_Co je to a jak probíhá integ~
## 21        0.0608 bad         good  KUKY       důchod-dorovnávací přídavek_1298-~
## 22        0.0581 good        bad   FrBo       orig_Jak využít svého práva být i~
## 23        0.0447 good        bad   KUKY       Pravni rada_uver SVJ
## 24        0.0438 good        bad   FrBo       149
## 25        0.0306 bad         good  KUKY       4842_2023_VOP
## 26        0.0298 good        bad   FrBo       142
## 27        0.0197 bad         good  KUKY       6525_2022_VOP
## 28        0.0189 good        bad   OmbuFlyers Studny
## 29        0.0182 bad         good  FrBo       red_Pozemkové úpravy_final
## 30        0.0166 bad         good  FrBo       156
## 31        0.0160 bad         good  FrBo       red_Jaké jsou povinnosti veřejnýc~
## # i 6 more rows
```

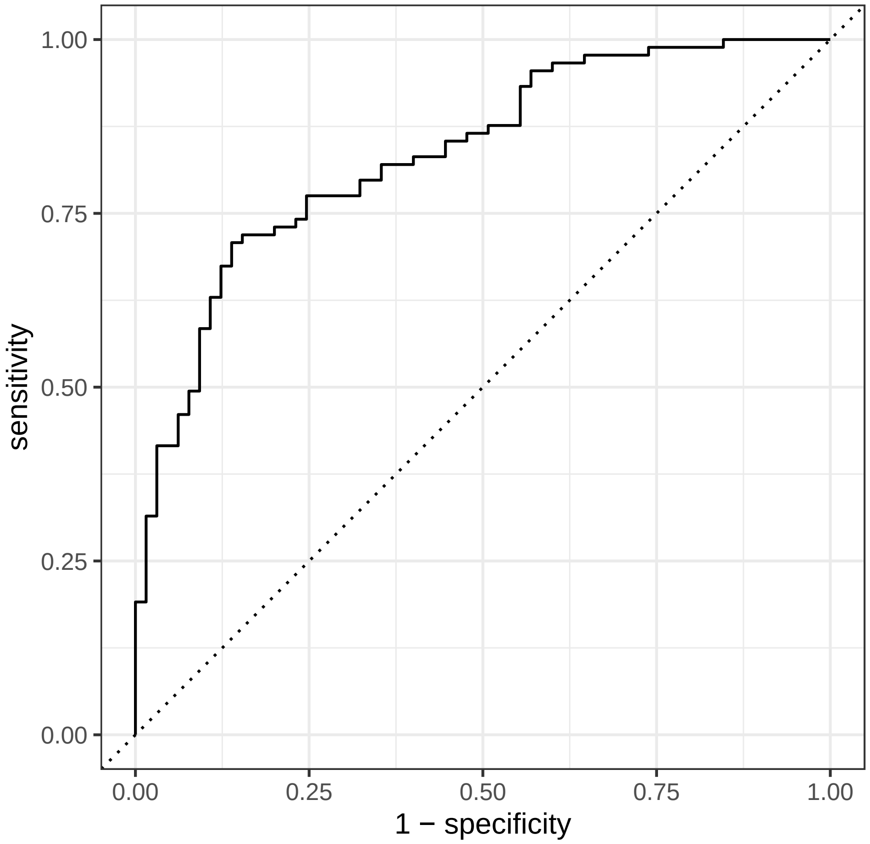
```
# lfit_lasso_all %>%
#   lasso_get_coefficients() %>%
#   print(n = 100)
```

No TL

```
lfit_lasso_notl <- model_lasso_notl %>% evaluate_tidymodel(split)
```

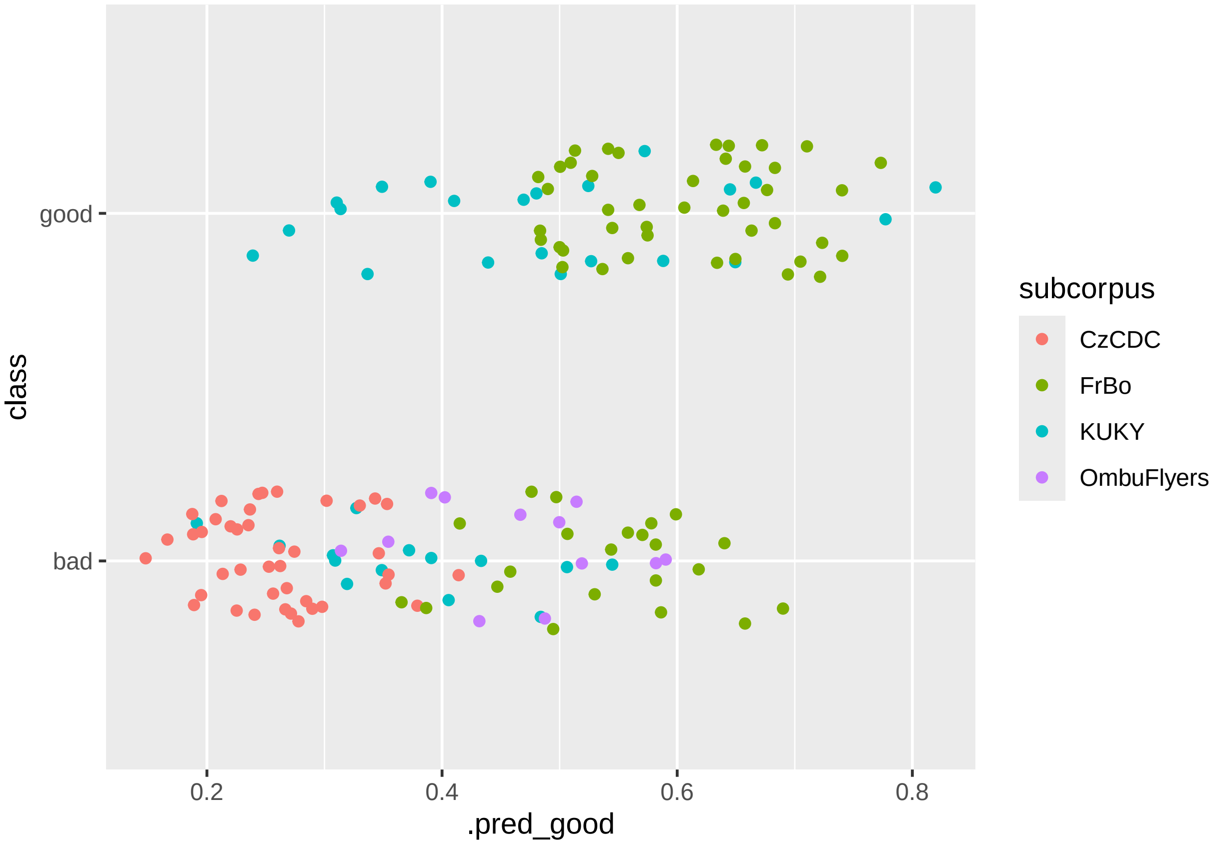
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.760 Preprocessor1_Model1
## 2 roc_auc     binary         0.835 Preprocessor1_Model1
## 3 brier_class binary         0.178 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 67 x 3
##    Variable                                            Importance Sign
##    <chr>                                                    <dbl> <chr>
##  1 activity                                              0.408    POS
##  2 smog                                                  0.191    NEG
##  3 RuleLiteraryStyle                                     0.168    NEG
##  4 atl                                                   0.100    POS
##  5 mamr                                                  0.0576   POS
##  6 gf                                                    0.0184   NEG
##  7 entropy                                               0.0165   NEG
##  8 maentropy                                             0.00435  NEG
##  9 ari                                                   0.000272 NEG
## 10 RuleGPcoordovs                                        0        NEG
## 11 RuleGPdeverbaddr                                      0        NEG
## 12 RuleGPpatinstr                                        0        NEG
## 13 RuleGPdeverbsubj                                      0        NEG
## 14 RuleGPadjective                                       0        NEG
## 15 RuleGPpatbenperson                                    0        NEG
## 16 RuleGPwordorder                                       0        NEG
## 17 RuleDoubleAdpos                                       0        NEG
## 18 RuleDoubleAdpos.max_allowable_distance                0        NEG
## 19 RuleDoubleAdpos.max_allowable_distance.v              0        NEG
## 20 RuleReflexivePassWithAnimSubj                         0        NEG
## 21 RuleTooFewVerbs.min_verb_frac                         0        NEG
## 22 RuleTooManyNegations.max_negation_frac                0        NEG
## 23 RuleTooManyNegations.max_negation_frac.v              0        NEG
## 24 RuleTooManyNegations.max_allowable_negations          0        NEG
## 25 RuleTooManyNegations.max_allowable_negations.v        0        NEG
## 26 RuleTooManyNominalConstructions.max_noun_frac         0        NEG
## 27 RuleTooManyNominalConstructions.max_noun_frac.v       0        NEG
## 28 RuleTooManyNominalConstructions.max_allowable_nouns   0        NEG
## 29 RuleCaseRepetition.max_repetition_count               0        NEG
## 30 RuleCaseRepetition.max_repetition_count.v             0        NEG
## 31 RuleCaseRepetition.max_repetition_frac                0        NEG
## 32 RuleCaseRepetition.max_repetition_frac.v              0        NEG
## 33 RuleWeakMeaningWords                                  0        NEG
## 34 RuleAbstractNouns                                     0        NEG
## 35 RuleRelativisticExpressions                           0        NEG
## 36 RuleConfirmationExpressions                           0        NEG
## 37 RuleRedundantExpressions                              0        NEG
## 38 RuleTooLongExpressions                                0        NEG
## 39 RuleAnaphoricReferences                               0        NEG
## 40 RulePassive                                           0        NEG
## 41 RulePredSubjDistance                                  0        NEG
## 42 RulePredSubjDistance.max_distance                     0        NEG
## 43 RulePredSubjDistance.max_distance.v                   0        NEG
## 44 RulePredObjDistance                                   0        NEG
## 45 RulePredObjDistance.max_distance                      0        NEG
## 46 RulePredObjDistance.max_distance.v                    0        NEG
## 47 RuleInfVerbDistance                                   0        NEG
## 48 RuleInfVerbDistance.max_distance                      0        NEG
## 49 RuleInfVerbDistance.max_distance.v                    0        NEG
## 50 RuleMultiPartVerbs                                    0        NEG
## 51 RuleMultiPartVerbs.max_distance                       0        NEG
## 52 RuleMultiPartVerbs.max_distance.v                     0        NEG
## 53 RuleLongSentences.max_length                          0        NEG
## 54 RuleLongSentences.max_length.v                        0        NEG
## 55 RulePredAtClauseBeginning.max_order                   0        NEG
## 56 RulePredAtClauseBeginning.max_order.v                 0        NEG
## 57 RuleVerbalNouns                                       0        NEG
## 58 cli                                                   0        NEG
## 59 num_hapax                                             0        NEG
## 60 ttr                                                   0        NEG
## 61 mattr                                                 0        NEG
## 62 mattr.v                                               0        NEG
## 63 maentropy.v                                           0        NEG
## 64 verb_dist                                             0        NEG
## 65 hpoint                                                0        NEG
## 66 fre                                                   0        NEG
## 67 fkgl                                                  0        NEG
```

```
lfit_lasso_notl %>% get_mismatch_details(data)
```



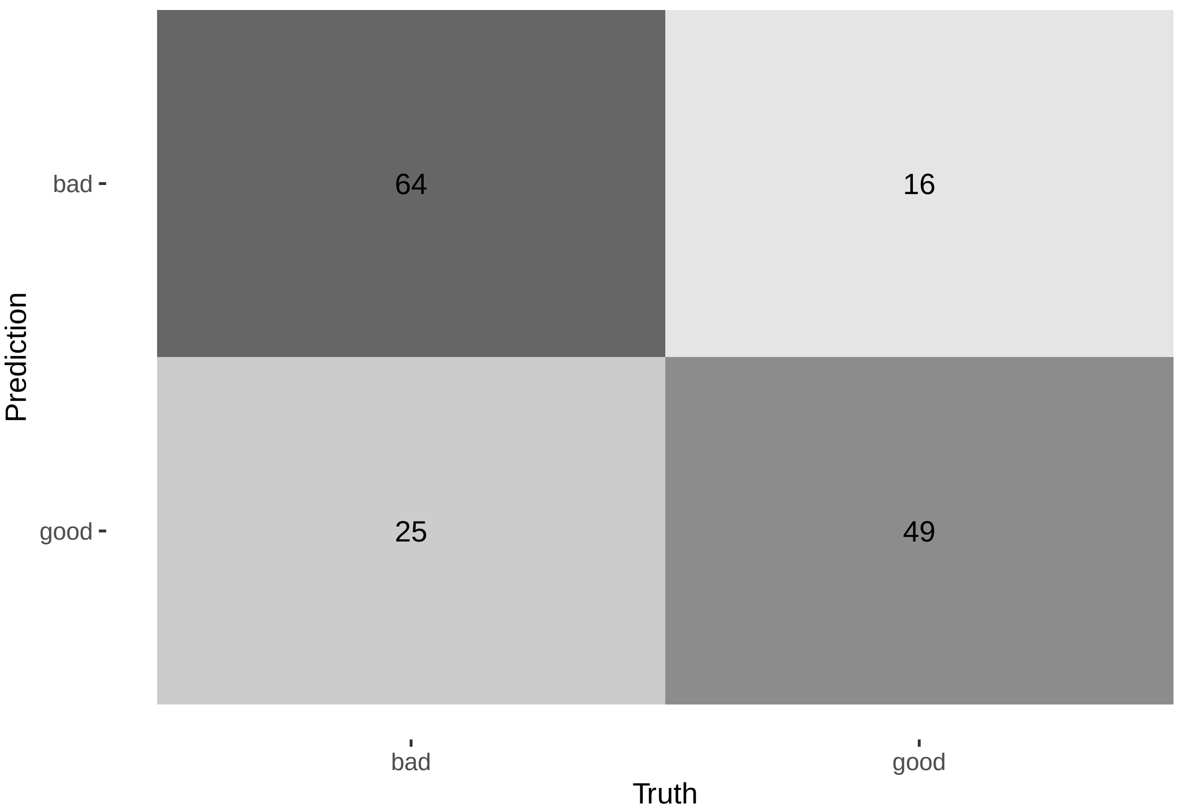
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   41    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    8    5
##        good  14   38
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   12   12
##        good   2   10
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    8    0
##        good   4    0
##
##
## Greatest deviations:
## # A tibble: 37 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.261  bad         good  KUKY       Odvolani_proti_rozhodnuti_o_nepov~
##  2        0.230  bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  3        0.190  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  4        0.190  bad         good  KUKY       MV_Odneti_trvaleho_pobytu_Kru_po
##  5        0.186  bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
##  6        0.163  bad         good  KUKY       Odvolani
##  7        0.158  good        bad   FrBo       orig_Co je to EIA_final
##  8        0.151  bad         good  KUKY       AK_JH_Podani_US_podpis
##  9        0.140  good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
## 10        0.118  good        bad   FrBo       orig_znalci, znalecké posudky
## 11        0.110  bad         good  KUKY       invalidní důchod_1399-23_původní
## 12        0.0989 good        bad   FrBo       64
## 13        0.0902 good        bad   OmbuFlyers Soudni-poplatky
## 14        0.0897 bad         good  KUKY       Ockovani_JSm
## 15        0.0862 good        bad   FrBo       orig_Sousedské vztahy
## 16        0.0819 good        bad   OmbuFlyers Detsky-domov
## 17        0.0819 good        bad   FrBo       orig_Jak probíhá správní řízení
## 18        0.0818 good        bad   FrBo       orig_Jak zajistit, aby skládka do~
## 19        0.0780 good        bad   FrBo       orig_územní řízení
## 20        0.0704 good        bad   FrBo       orig_Co je to a jak probíhá integ~
## 21        0.0608 bad         good  KUKY       důchod-dorovnávací přídavek_1298-~
## 22        0.0581 good        bad   FrBo       orig_Jak využít svého práva být i~
## 23        0.0447 good        bad   KUKY       Pravni rada_uver SVJ
## 24        0.0438 good        bad   FrBo       149
## 25        0.0306 bad         good  KUKY       4842_2023_VOP
## 26        0.0298 good        bad   FrBo       142
## 27        0.0197 bad         good  KUKY       6525_2022_VOP
## 28        0.0189 good        bad   OmbuFlyers Studny
## 29        0.0182 bad         good  FrBo       red_Pozemkové úpravy_final
## 30        0.0166 bad         good  FrBo       156
## 31        0.0160 bad         good  FrBo       red_Jaké jsou povinnosti veřejnýc~
## # i 6 more rows
```

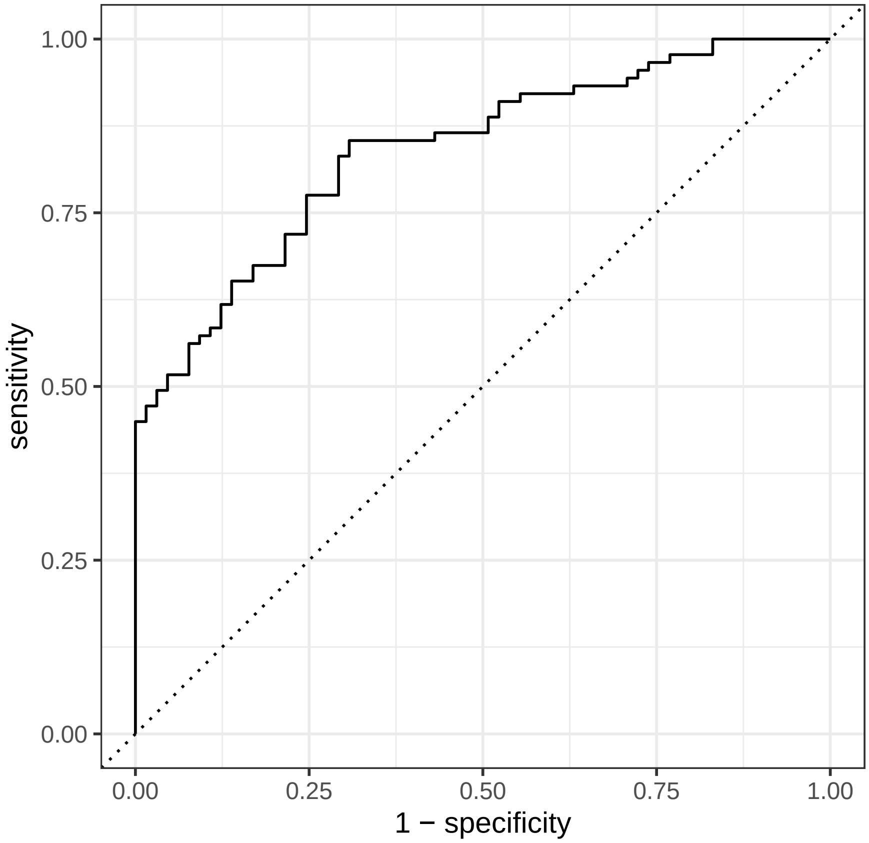
```
# lfit_lasso_notl %>%
#   lasso_get_coefficients() %>%
#   print(n = 100)
```

IAC

```
lfit_lasso_iac <- model_lasso_iac %>% evaluate_tidymodel(split)
```

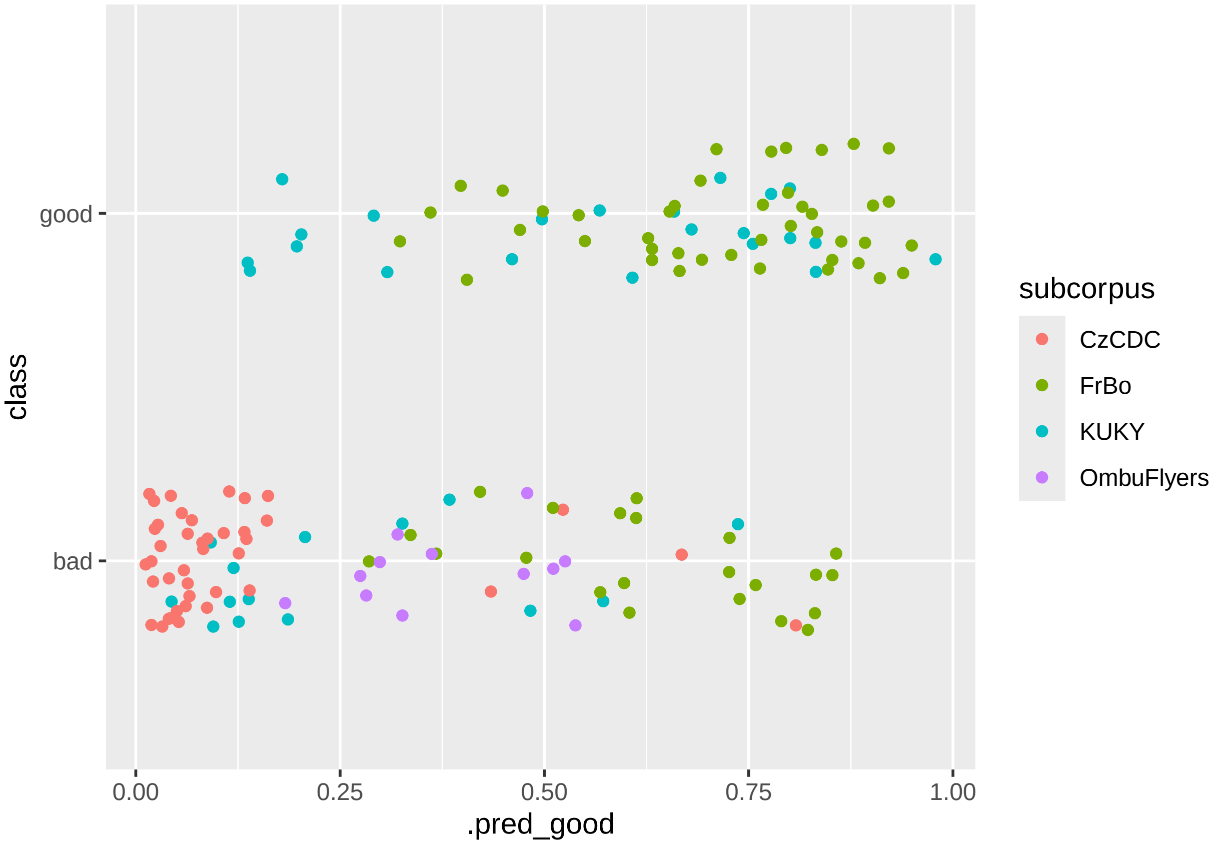
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.734 Preprocessor1_Model1
## 2 roc_auc     binary         0.840 Preprocessor1_Model1
## 3 brier_class binary         0.164 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 44 x 3
##    Variable                                              Importance Sign
##    <chr>                                                      <dbl> <chr>
##  1 RuleTooFewVerbs.min_verb_frac                           16.1     NEG
##  2 RuleCaseRepetition.max_repetition_frac                  14.2     NEG
##  3 activity                                                11.4     POS
##  4 maentropy.v                                              9.14    POS
##  5 RuleTooManyNominalConstructions.max_noun_frac            6.66    NEG
##  6 mattr                                                    6.42    NEG
##  7 RuleCaseRepetition.max_repetition_frac.v                 4.98    POS
##  8 RuleTooManyNominalConstructions.max_noun_frac.v          2.11    POS
##  9 atl                                                      1.90    POS
## 10 RuleCaseRepetition.max_repetition_count.v                1.90    NEG
## 11 RuleLongSentences.max_length.v                           1.10    POS
## 12 ttr                                                      1.09    NEG
## 13 RuleTooManyNominalConstructions.max_allowable_nouns.v    0.991   NEG
## 14 RuleTooManyNegations.max_allowable_negations.v           0.867   NEG
## 15 RuleInfVerbDistance.max_distance.v                       0.778   NEG
## 16 entropy                                                  0.576   NEG
## 17 RuleTooManyNegations.max_negation_frac                   0.479   POS
## 18 ari                                                      0.167   NEG
## 19 RuleMultiPartVerbs.max_distance.v                        0.155   POS
## 20 gf                                                       0.140   NEG
## 21 RuleDoubleAdpos.max_allowable_distance.v                 0.138   NEG
## 22 RuleInfVerbDistance.max_distance                         0.100   POS
## 23 RulePredSubjDistance.max_distance.v                      0.0890  NEG
## 24 fre                                                      0.0449  NEG
## 25 RuleLongSentences.max_length                             0.0354  POS
## 26 RuleTooManyNominalConstructions.max_allowable_nouns      0.0332  POS
## 27 verb_dist                                                0.0325  POS
## 28 smog                                                     0.0307  NEG
## 29 RulePredSubjDistance.max_distance                        0.0230  NEG
## 30 RulePredObjDistance.max_distance                         0.0213  NEG
## 31 RulePredAtClauseBeginning.max_order                      0.00681 POS
## 32 RuleDoubleAdpos.max_allowable_distance                   0.00441 POS
## 33 hpoint                                                   0.00122 NEG
## 34 RuleTooManyNegations.max_negation_frac.v                 0       NEG
## 35 RuleTooManyNegations.max_allowable_negations             0       NEG
## 36 RuleCaseRepetition.max_repetition_count                  0       NEG
## 37 RulePredObjDistance.max_distance.v                       0       NEG
## 38 RuleMultiPartVerbs.max_distance                          0       NEG
## 39 RulePredAtClauseBeginning.max_order.v                    0       NEG
## 40 cli                                                      0       NEG
## 41 mattr.v                                                  0       NEG
## 42 maentropy                                                0       NEG
## 43 mamr                                                     0       NEG
## 44 fkgl                                                     0       NEG
```

```
lfit_lasso_iac %>% get_mismatch_details(data)
```



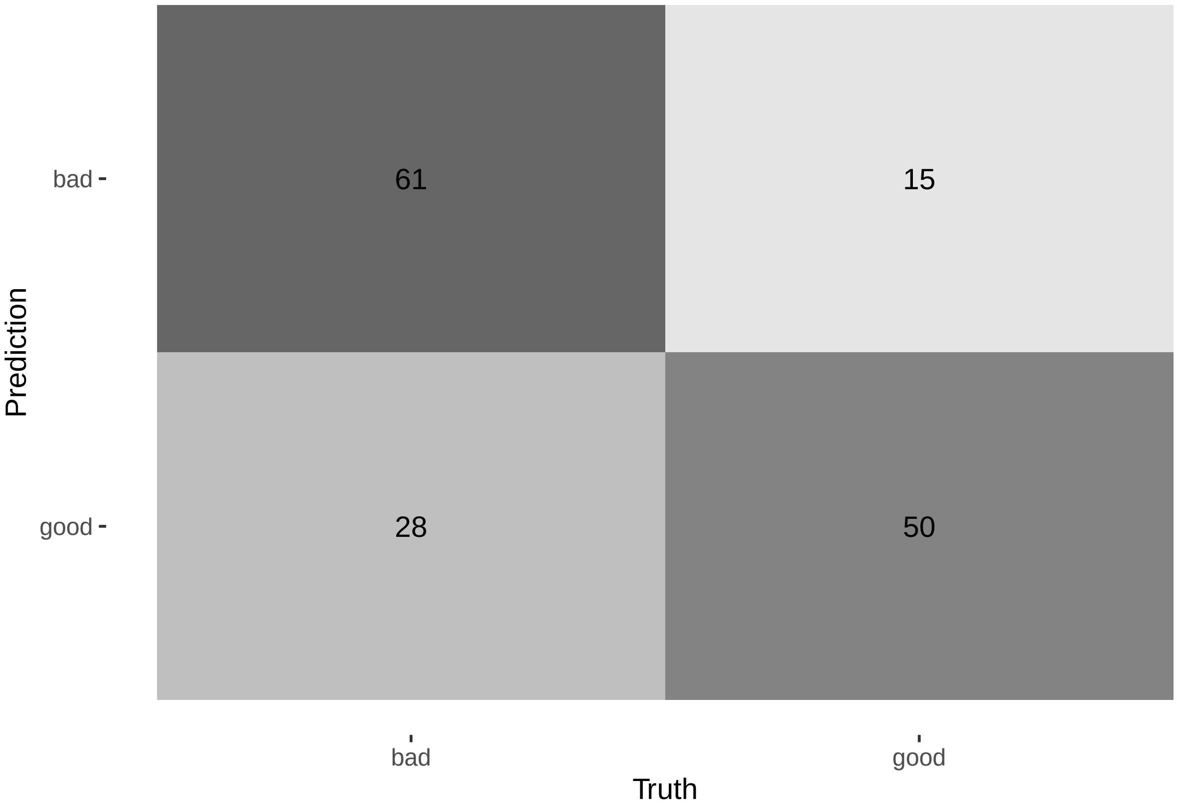
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   38    0
##        good   3    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    5    7
##        good  17   36
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   12    9
##        good   2   13
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    9    0
##         good   3    0
##
##
## Greatest deviations:
## # A tibble: 41 x 5
##    abs_deviation .pred_class class subcorpus FileName
##            <dbl> <fct>       <fct> <chr>     <chr>
##  1        0.363  bad         good  KUKY      0217_6Afs_2000035_20210219141328__~
##  2        0.360  bad         good  KUKY      Mestsky_urad_Vyzva_k_zaplaceni_nak~
##  3        0.357  good        bad   FrBo      orig_Jaké otázky (ne)můžete položi~
##  4        0.352  good        bad   FrBo      orig_Co je to EIA_final
##  5        0.332  good        bad   FrBo      orig_Zastupitelstvo_o čem a jak ro~
##  6        0.331  good        bad   FrBo      orig_Jak probíhá správní řízení
##  7        0.322  good        bad   FrBo      64
##  8        0.321  bad         good  KUKY      Odvolani_proti_rozhodnuti_o_nepovo~
##  9        0.308  good        bad   CzCDC     2-2825-08_1
## 10        0.303  bad         good  KUKY      Odvolani
## 11        0.297  bad         good  KUKY      MV_Odneti_trvaleho_pobytu_Kru_po
## 12        0.290  good        bad   FrBo      142
## 13        0.259  good        bad   FrBo      149
## 14        0.239  good        bad   FrBo      orig_územní řízení
## 15        0.237  good        bad   KUKY      Dopis_studentské brigády
## 16        0.227  good        bad   FrBo      orig_znalci, znalecké posudky
## 17        0.226  good        bad   FrBo      orig_Jak zajistit, aby skládka dod~
## 18        0.209  bad         good  KUKY      29 A 80-2021_20231122101241
## 19        0.192  bad         good  KUKY      AK_JH_Podani_US_podpis
## 20        0.177  bad         good  FrBo      14
## 21        0.168  good        bad   CzCDC     3-376-98
## 22        0.139  bad         good  FrBo      red_pravni_nastroje_ochrany_ovzdusi
## 23        0.113  good        bad   FrBo      orig_Certifikáty autorizovaných in~
## 24        0.112  good        bad   FrBo      orig_Správní exekuce
## 25        0.104  good        bad   FrBo      orig_Kdy a jak požadovat náhradu š~
## 26        0.102  bad         good  FrBo      red_Jaké právní nástroje můžete vy~
## 27        0.0976 good        bad   FrBo      orig_Jak využít svého práva být in~
## 28        0.0948 bad         good  FrBo      red_Les - co smíme a co je zakázáno
## 29        0.0928 good        bad   FrBo      orig_Co je to a jak probíhá integr~
## 30        0.0720 good        bad   KUKY      Pravni rada_uver SVJ
## 31        0.0684 good        bad   FrBo      68
## # i 10 more rows
```

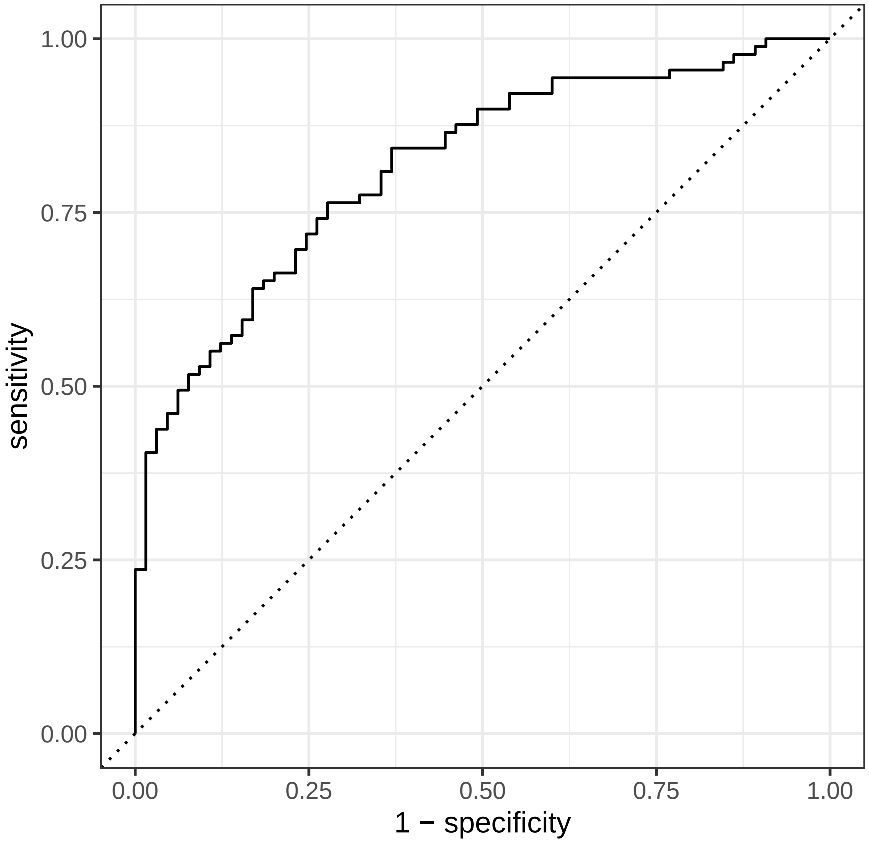
```
# lfit_lasso_iac %>%
#   lasso_get_coefficients() %>%
#   print(n = 100)
```

Counts

```
lfit_lasso_counts <- model_lasso_counts %>% evaluate_tidymodel(split)
```

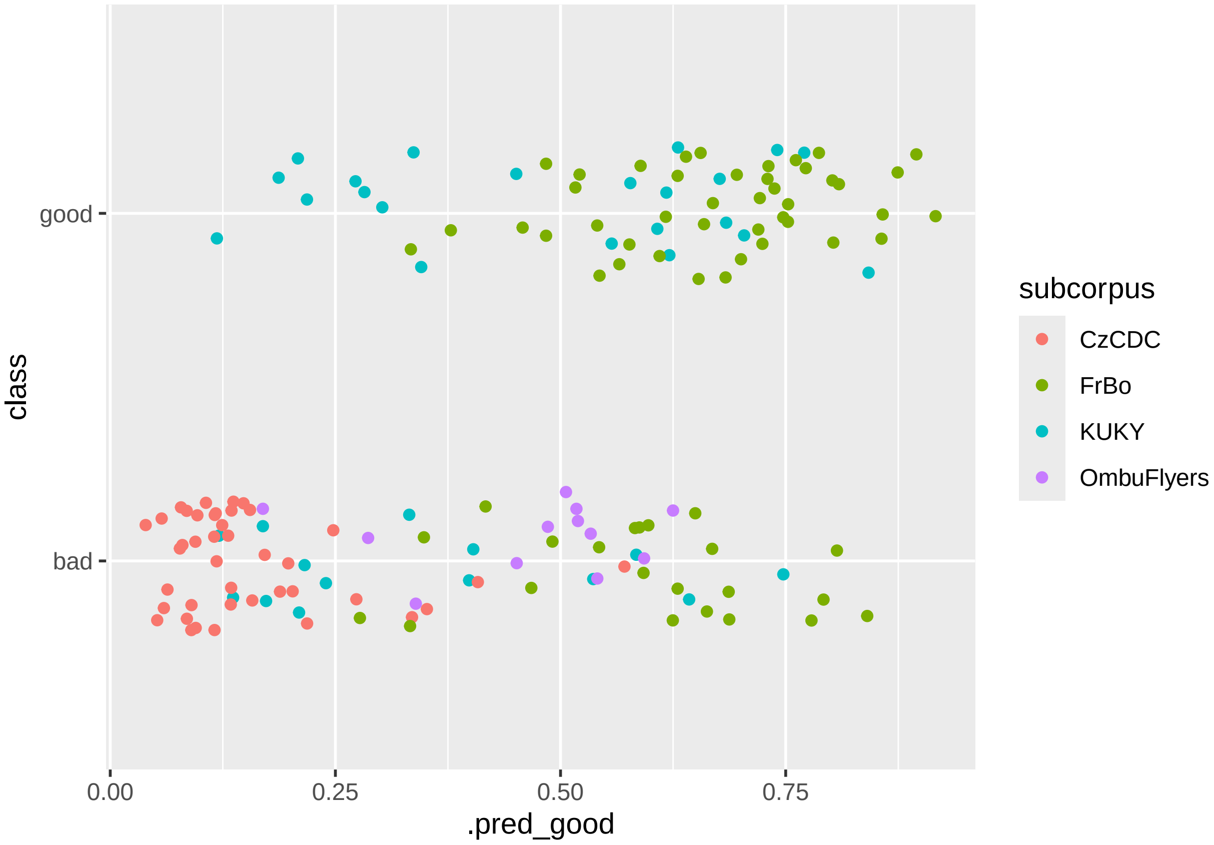
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.721 Preprocessor1_Model1
## 2 roc_auc     binary         0.816 Preprocessor1_Model1
## 3 brier_class binary         0.174 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 28 x 3
##    Variable                      Importance Sign
##    <chr>                              <dbl> <chr>
##  1 RuleRedundantExpressions      616.       NEG
##  2 RuleRelativisticExpressions   332.       NEG
##  3 RuleAnaphoricReferences       157.       POS
##  4 RuleGPdeverbsubj              149.       NEG
##  5 RuleLiteraryStyle             123.       NEG
##  6 RulePassive                   119.       NEG
##  7 RuleGPadjective               113.       POS
##  8 RuleGPdeverbaddr               92.8      NEG
##  9 RuleTooLongExpressions         60.5      POS
## 10 RuleMultiPartVerbs             34.1      POS
## 11 RulePredSubjDistance           18.2      POS
## 12 RuleVerbalNouns                 5.83     POS
## 13 RuleInfVerbDistance             0.912    POS
## 14 sent_count                      0.00502  POS
## 15 word_count                      0.000438 NEG
## 16 RuleGPcoordovs                  0        NEG
## 17 RuleGPpatinstr                  0        NEG
## 18 RuleGPpatbenperson              0        NEG
## 19 RuleGPwordorder                 0        NEG
## 20 RuleDoubleAdpos                 0        NEG
## 21 RuleReflexivePassWithAnimSubj   0        NEG
## 22 RuleWeakMeaningWords            0        NEG
## 23 RuleAbstractNouns               0        NEG
## 24 RuleConfirmationExpressions     0        NEG
## 25 RulePredObjDistance             0        NEG
## 26 syllab_count                    0        NEG
## 27 char_count                      0        NEG
## 28 num_hapax                       0        NEG
```

```
lfit_lasso_counts %>% get_mismatch_details(data)
```



```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   40    0
##        good   1    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    6    5
##        good  16   38
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   10   10
##        good   4   12
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    5    0
##        good   7    0
##
##
## Greatest deviations:
## # A tibble: 43 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1         0.382 bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  2         0.341 good        bad   FrBo       orig_Co je to EIA_final
##  3         0.313 bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
##  4         0.307 good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  5         0.292 good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
##  6         0.292 bad         good  KUKY       AK_JH_Podani_US_podpis
##  7         0.282 bad         good  KUKY       invalidní důchod_1399-23_původní
##  8         0.279 good        bad   FrBo       orig_Co je to a jak probíhá integ~
##  9         0.247 good        bad   KUKY       Dopis vysvětlující dopis klientovi
## 10         0.228 bad         good  KUKY       Odvolani
## 11         0.218 bad         good  KUKY       1732_2023_VOP
## 12         0.198 bad         good  KUKY       Odvolani_proti_rozhodnuti_o_nepov~
## 13         0.187 good        bad   FrBo       orig_znalci, znalecké posudky
## 14         0.187 good        bad   FrBo       orig_Sousedské vztahy
## 15         0.168 good        bad   FrBo       orig_Jak probíhá správní řízení
## 16         0.166 bad         good  FrBo       190
## 17         0.163 bad         good  KUKY       29 A 80-2021_20231122101241
## 18         0.163 good        bad   FrBo       149
## 19         0.155 bad         good  KUKY       důchod-dorovnávací přídavek_1298-~
## 20         0.150 good        bad   FrBo       orig_územní řízení
## 21         0.143 good        bad   KUKY       Pravni rada_uver SVJ
## 22         0.130 good        bad   FrBo       orig_Jak zajistit, aby skládka do~
## 23         0.125 good        bad   OmbuFlyers Ochrana-osob-omezenych-na-svobode
## 24         0.125 good        bad   FrBo       64
## 25         0.122 bad         good  FrBo       red_Co je to úřední deska a jak j~
## 26         0.0975 good       bad   FrBo       orig_pravni_nastroje_ochrany_ovzd~
## 27         0.0928 good       bad   OmbuFlyers Studny
## 28         0.0922 good       bad   FrBo       orig_Jaké právní nástroje můžete ~
## 29         0.0876 good       bad   FrBo       142
## 30         0.0841 good       bad   KUKY       UOOUOsobniUdajePuvodne
## 31         0.0826 good       bad   FrBo       orig_Vyvlastnění podle zákona o u~
## # i 12 more rows
```

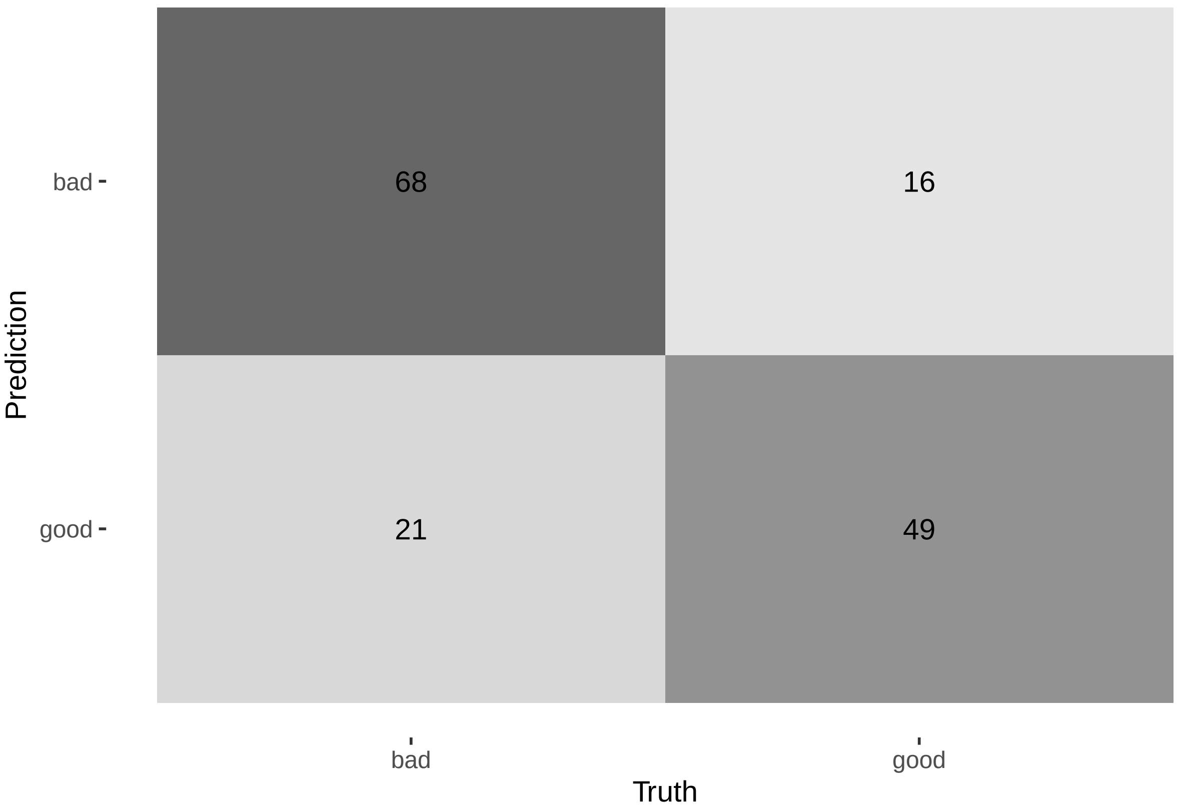
```
# lfit_lasso_counts %>%
#   lasso_get_coefficients() %>%
#   print(n = 100)
```

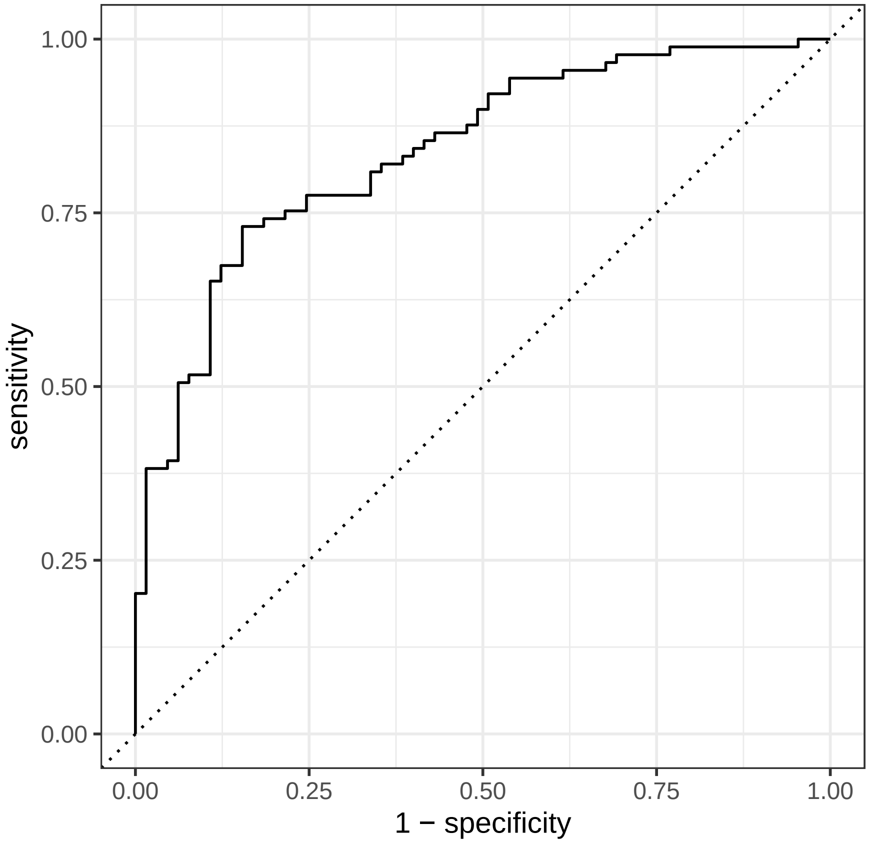
Random forest

All

```
lfit_rf_all <- model_rf_all %>% evaluate_tidymodel(split)
```

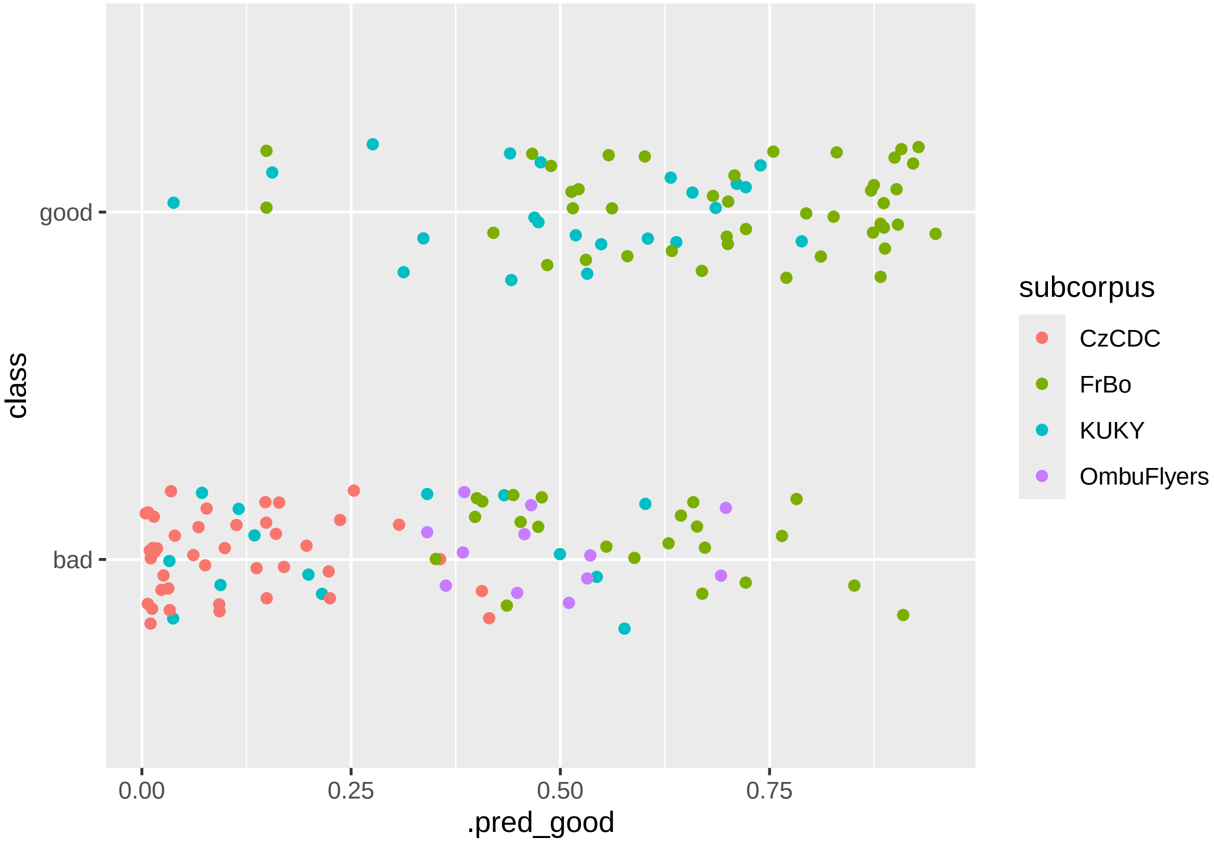
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.760 Preprocessor1_Model1
## 2 roc_auc     binary         0.838 Preprocessor1_Model1
## 3 brier_class binary         0.165 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 71 x 2
##    Variable                                            Importance
##    <chr>                                                    <dbl>
##  1 verb_dist                                               13.1
##  2 RuleLongSentences.max_length                            12.6
##  3 RuleTooManyNominalConstructions.max_allowable_nouns     12.4
##  4 activity                                                12.1
##  5 RuleTooFewVerbs.min_verb_frac                           10.9
##  6 ari                                                     10.6
##  7 gf                                                       9.07
##  8 RuleLiteraryStyle                                        8.58
##  9 smog                                                     8.00
## 10 RulePredAtClauseBeginning.max_order                      7.89
## 11 RulePassive                                              7.32
## 12 mamr                                                     5.61
## 13 atl                                                      5.32
## 14 fkgl                                                     5.24
## 15 RuleMultiPartVerbs                                       4.49
## 16 RulePredAtClauseBeginning.max_order.v                    4.30
## 17 mattr                                                    4.08
## 18 RuleTooManyNegations.max_negation_frac                   4.04
## 19 maentropy                                                3.92
## 20 RuleVerbalNouns                                          3.92
## 21 RuleTooLongExpressions                                   3.79
## 22 RuleTooManyNominalConstructions.max_noun_frac            3.75
## 23 entropy                                                  3.72
## 24 maentropy.v                                              3.59
## 25 RuleAnaphoricReferences                                  3.45
## 26 RulePredSubjDistance                                     3.43
## 27 cli                                                      3.29
## 28 RuleLongSentences.max_length.v                           3.18
## 29 RuleDoubleAdpos.max_allowable_distance.v                 3.17
## 30 mattr.v                                                  3.02
## 31 RulePredSubjDistance.max_distance                        2.97
## 32 RuleCaseRepetition.max_repetition_count.v                2.93
## 33 word_count                                               2.83
## 34 RuleCaseRepetition.max_repetition_frac.v                 2.80
## 35 RulePredObjDistance                                      2.74
## 36 RuleInfVerbDistance.max_distance                         2.74
## 37 RuleCaseRepetition.max_repetition_frac                   2.72
## 38 RuleCaseRepetition.max_repetition_count                  2.69
## 39 RuleTooManyNegations.max_negation_frac.v                 2.66
## 40 num_hapax                                                2.58
## 41 RulePredSubjDistance.max_distance.v                      2.58
## 42 RuleTooManyNegations.max_allowable_negations             2.49
## 43 RuleInfVerbDistance.max_distance.v                       2.48
## 44 ttr                                                      2.45
## 45 RuleMultiPartVerbs.max_distance.v                        2.40
## 46 RulePredObjDistance.max_distance                         2.38
## 47 RulePredObjDistance.max_distance.v                       2.38
## 48 RuleMultiPartVerbs.max_distance                          2.35
## 49 char_count                                               2.30
## 50 syllab_count                                             2.29
## 51 RuleDoubleAdpos                                             2.21
## 52 RuleInfVerbDistance                                         2.14
## 53 fre                                                         2.13
## 54 RuleTooManyNegations.max_allowable_negations.v              2.10
## 55 RuleAbstractNouns                                           2.10
## 56 RuleTooManyNominalConstructions.max_noun_frac.v             1.98
## 57 sent_count                                                  1.94
## 58 RuleDoubleAdpos.max_allowable_distance                      1.91
## 59 hpoint                                                      1.78
## 60 RuleWeakMeaningWords                                        1.72
## 61 RuleReflexivePassWithAnimSubj                               1.57
## 62 RuleGPwordorder                                             1.47
## 63 RuleGPpatinstr                                              1.17
## 64 RuleGPdeverbaddr                                            1.16
## 65 RuleRelativisticExpressions                                 1.04
## 66 RuleGPdeverbsubj                                            0.920
## 67 RuleGPpatbenperson                                          0.877
## 68 RuleGPcoordovs                                              0.790
## 69 RuleRedundantExpressions                                    0.269
## 70 RuleGPadjective                                             0.246
## 71 RuleConfirmationExpressions                                 0.229
```

```
lfit_rf_all %>% get_mismatch_details(data)
```

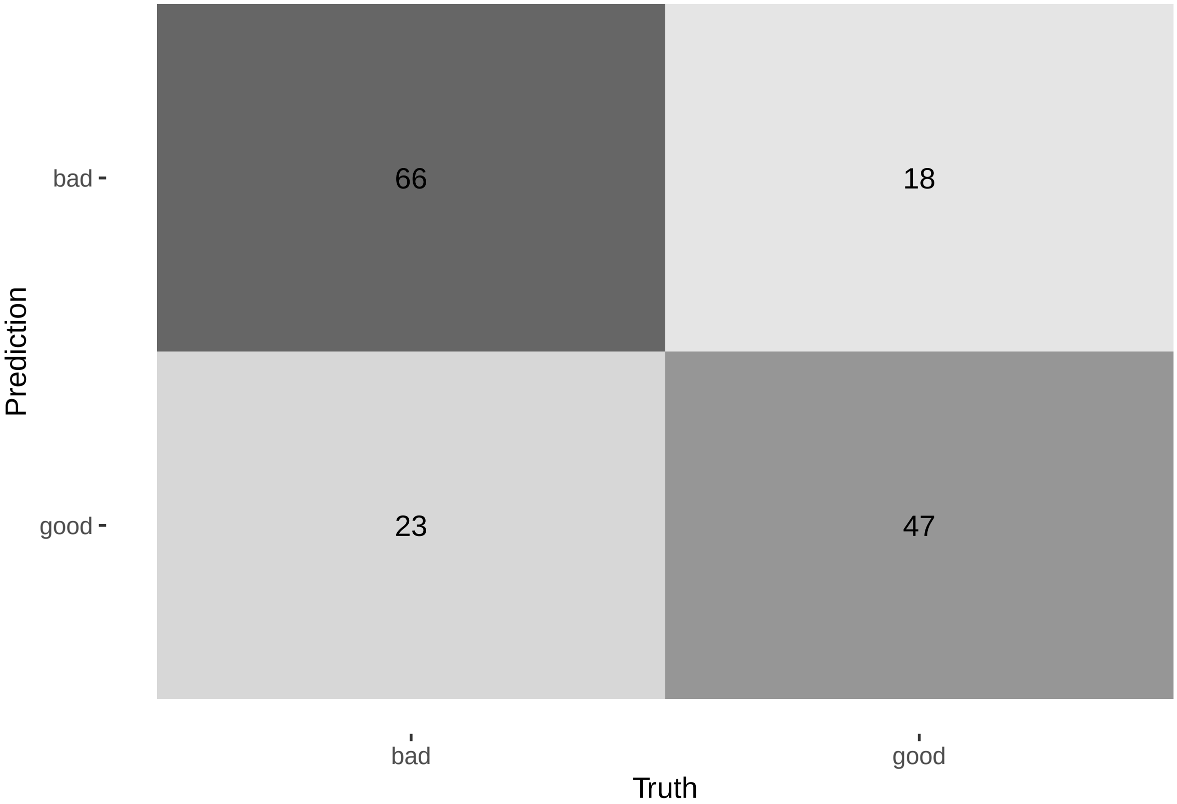


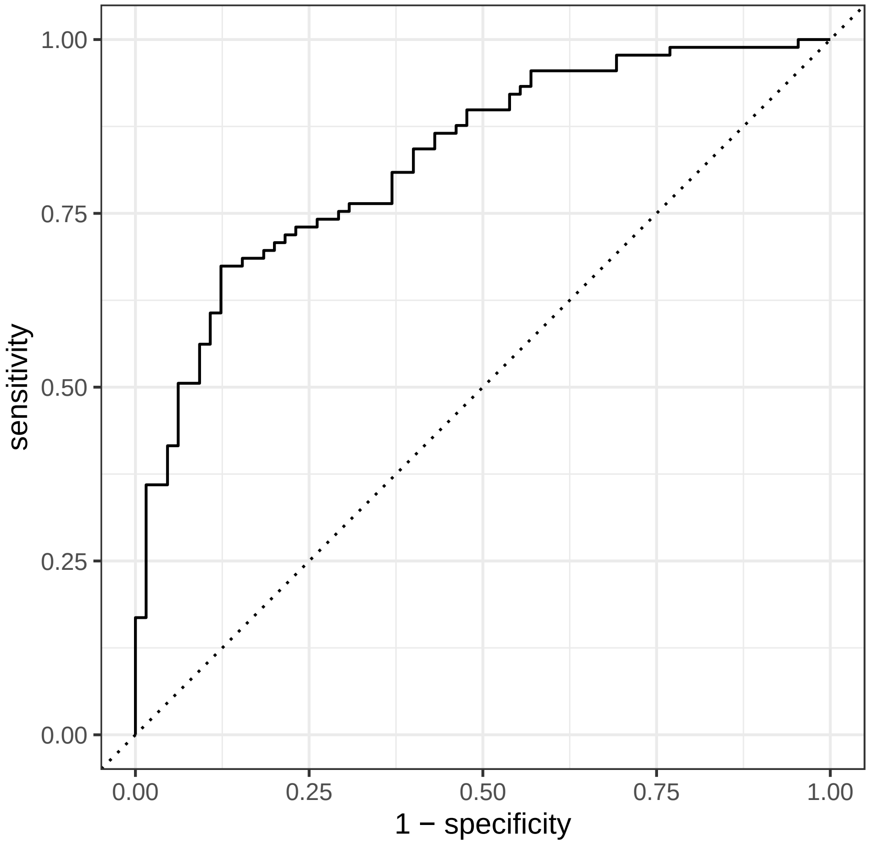
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   41    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    9    6
##        good  13   37
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   11   10
##        good   3   12
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    7    0
##        good   5    0
##
##
## Greatest deviations:
## # A tibble: 37 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.462  bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  2        0.410  good        bad   FrBo       orig_Jak zajistit, aby skládka do~
##  3        0.351  good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
##  4        0.351  bad         good  FrBo       red_Mohou spolky ve správních žal~
##  5        0.351  bad         good  FrBo       red_Mohou spolky ve správních žal~
##  6        0.344  bad         good  KUKY       Odvolani
##  7        0.282  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  8        0.265  good        bad   FrBo       orig_Jak probíhá správní řízení
##  9        0.224  bad         good  KUKY       invalidní důchod_1399-23_původní
## 10        0.222  good        bad   FrBo       142
## 11        0.198  good        bad   OmbuFlyers Soudni-poplatky
## 12        0.192  good        bad   OmbuFlyers Studny
## 13        0.187  bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
## 14        0.173  good        bad   FrBo       orig_územní řízení
## 15        0.170  good        bad   FrBo       orig_Jak využít svého práva být i~
## 16        0.164  bad         good  KUKY       AK_JH_Podani_US_podpis
## 17        0.163  good        bad   FrBo       64
## 18        0.159  good        bad   FrBo       orig_Kdy a jak požadovat náhradu ~
## 19        0.144  good        bad   FrBo       orig_Co je to a jak probíhá integ~
## 20        0.129  good        bad   FrBo       orig_znalci, znalecké posudky
## 21        0.102  good        bad   KUKY       Duchody
## 22        0.0885 good        bad   FrBo       orig_Sousedské vztahy
## 23        0.0800 bad         good  FrBo       red_pravni_nastroje_ochrany_ovzdu~
## 24        0.0767 good        bad   KUKY       Dopis vysvětlující dopis klientovi
## 25        0.0601 bad         good  KUKY       29 A 80-2021_20231122101241
## 26        0.0585 bad         good  KUKY       4842_2023_VOP
## 27        0.0550 good        bad   FrBo       orig_Certifikáty autorizovaných i~
## 28        0.0436 good        bad   KUKY       Pravni rada_uver SVJ
## 29        0.0358 good        bad   OmbuFlyers Detsky-domov
## 30        0.0336 bad         good  FrBo       red_Pozemkové úpravy_final
## 31        0.0322 good        bad   OmbuFlyers Katastr-nemovitosti
## # i 6 more rows
```

No TL

```
lfit_rf_notl <- model_rf_notl %>% evaluate_tidymodel(split)
```

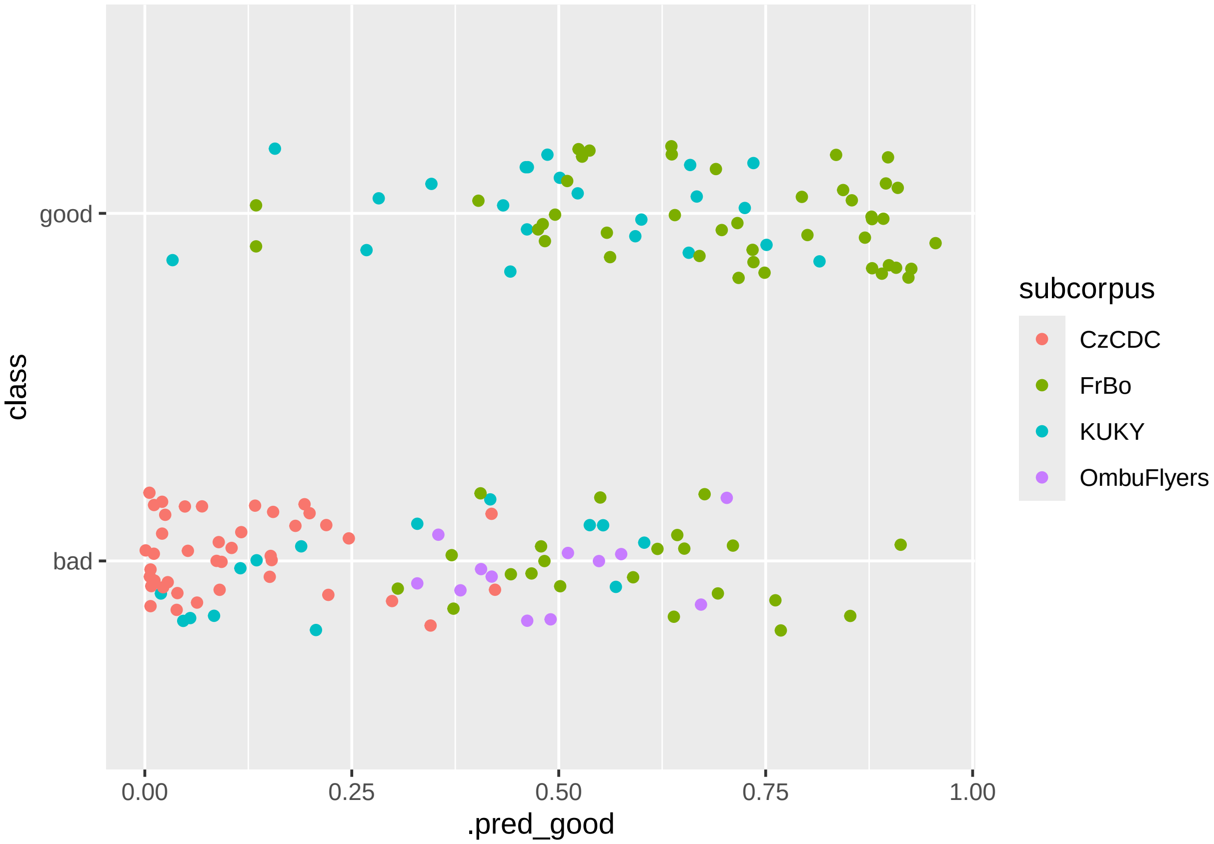
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.734 Preprocessor1_Model1
## 2 roc_auc     binary         0.828 Preprocessor1_Model1
## 3 brier_class binary         0.167 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 67 x 2
##    Variable                                            Importance
##    <chr>                                                    <dbl>
##  1 verb_dist                                                13.5
##  2 activity                                                 13.5
##  3 RuleTooManyNominalConstructions.max_allowable_nouns      13.0
##  4 RuleTooFewVerbs.min_verb_frac                            11.1
##  5 RuleLongSentences.max_length                             11.1
##  6 gf                                                       10.4
##  7 smog                                                     10.4
##  8 ari                                                       8.97
##  9 RuleLiteraryStyle                                         8.86
## 10 RulePredAtClauseBeginning.max_order                       8.25
## 11 RulePassive                                               6.67
## 12 fkgl                                                      5.51
## 13 mamr                                                      5.36
## 14 RulePredAtClauseBeginning.max_order.v                     5.27
## 15 atl                                                       5.06
## 16 maentropy                                                 4.48
## 17 entropy                                                   4.28
## 18 RuleTooManyNegations.max_negation_frac                    4.26
## 19 RuleMultiPartVerbs                                        4.26
## 20 RuleTooManyNominalConstructions.max_noun_frac             4.12
## 21 mattr                                                     4.11
## 22 RuleTooLongExpressions                                    4.07
## 23 RuleAnaphoricReferences                                   3.95
## 24 RuleVerbalNouns                                          3.76
## 25 RulePredSubjDistance                                     3.70
## 26 RulePredSubjDistance.max_distance                        3.32
## 27 maentropy.v                                              3.22
## 28 mattr.v                                                  3.18
## 29 cli                                                      3.13
## 30 RuleLongSentences.max_length.v                           3.12
## 31 ttr                                                      2.99
## 32 RuleCaseRepetition.max_repetition_count.v                2.98
## 33 RuleDoubleAdpos.max_allowable_distance.v                 2.94
## 34 RuleCaseRepetition.max_repetition_frac.v                 2.89
## 35 RulePredObjDistance                                      2.83
## 36 RuleCaseRepetition.max_repetition_frac                   2.82
## 37 RulePredSubjDistance.max_distance.v                      2.80
## 38 RuleTooManyNegations.max_allowable_negations             2.76
## 39 RuleInfVerbDistance.max_distance.v                       2.74
## 40 RuleInfVerbDistance.max_distance                         2.73
## 41 RuleTooManyNegations.max_negation_frac.v                 2.71
## 42 num_hapax                                                2.57
## 43 RuleMultiPartVerbs.max_distance                          2.56
## 44 RuleTooManyNegations.max_allowable_negations.v           2.55
## 45 RuleCaseRepetition.max_repetition_count                  2.54
## 46 fre                                                      2.50
## 47 RulePredObjDistance.max_distance.v                       2.48
## 48 RuleMultiPartVerbs.max_distance.v                        2.46
## 49 RulePredObjDistance.max_distance                         2.36
## 50 RuleDoubleAdpos                                          2.25
## 51 RuleInfVerbDistance                                      2.15
## 52 RuleDoubleAdpos.max_allowable_distance                   2.07
## 53 RuleTooManyNominalConstructions.max_noun_frac.v          2.06
## 54 RuleWeakMeaningWords                                     2.05
## 55 hpoint                                                   1.95
## 56 RuleAbstractNouns                                        1.93
## 57 RuleReflexivePassWithAnimSubj                            1.60
## 58 RuleGPwordorder                                          1.59
## 59 RuleGPpatinstr                                           1.40
## 60 RuleGPdeverbaddr                                         1.17
## 61 RuleRelativisticExpressions                              0.943
## 62 RuleGPdeverbsubj                                         0.862
## 63 RuleGPpatbenperson                                       0.841
## 64 RuleGPcoordovs                                           0.836
## 65 RuleGPadjective                                          0.346
## 66 RuleRedundantExpressions                                 0.318
## 67 RuleConfirmationExpressions                              0.277
```

```
lfit_rf_notl %>% get_mismatch_details(data)
```

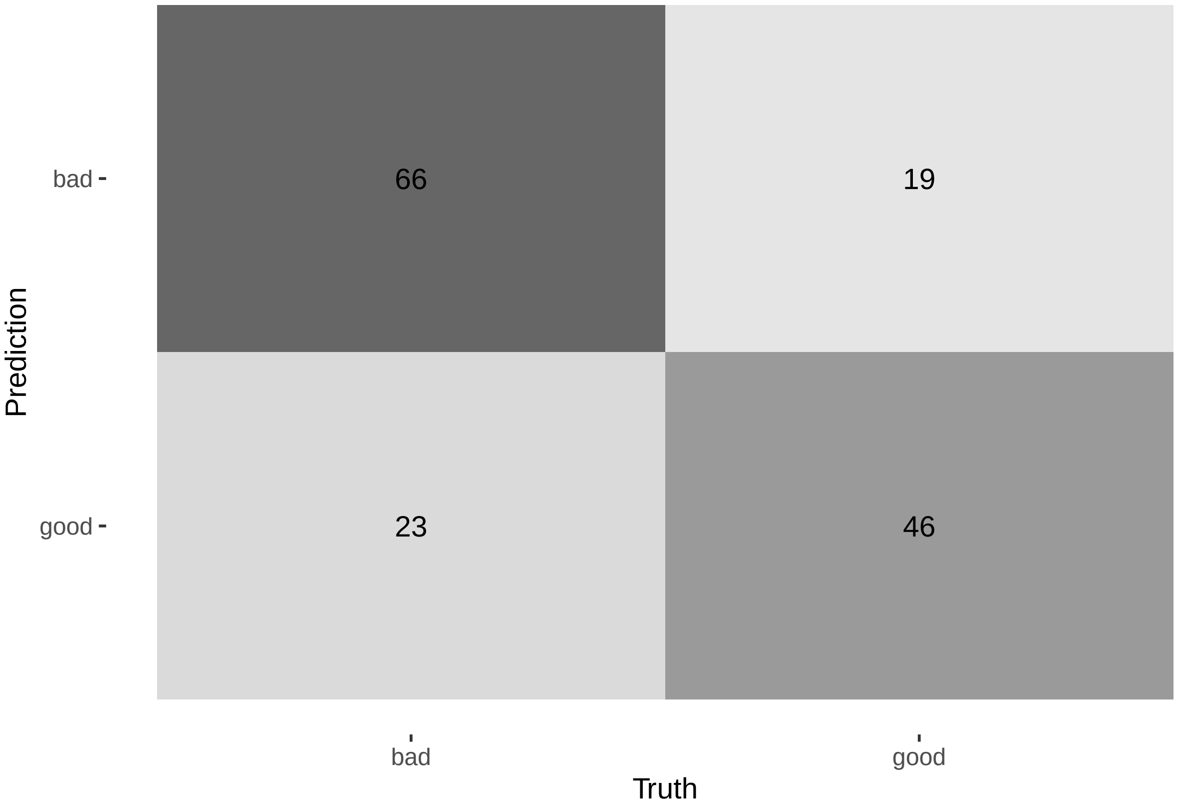


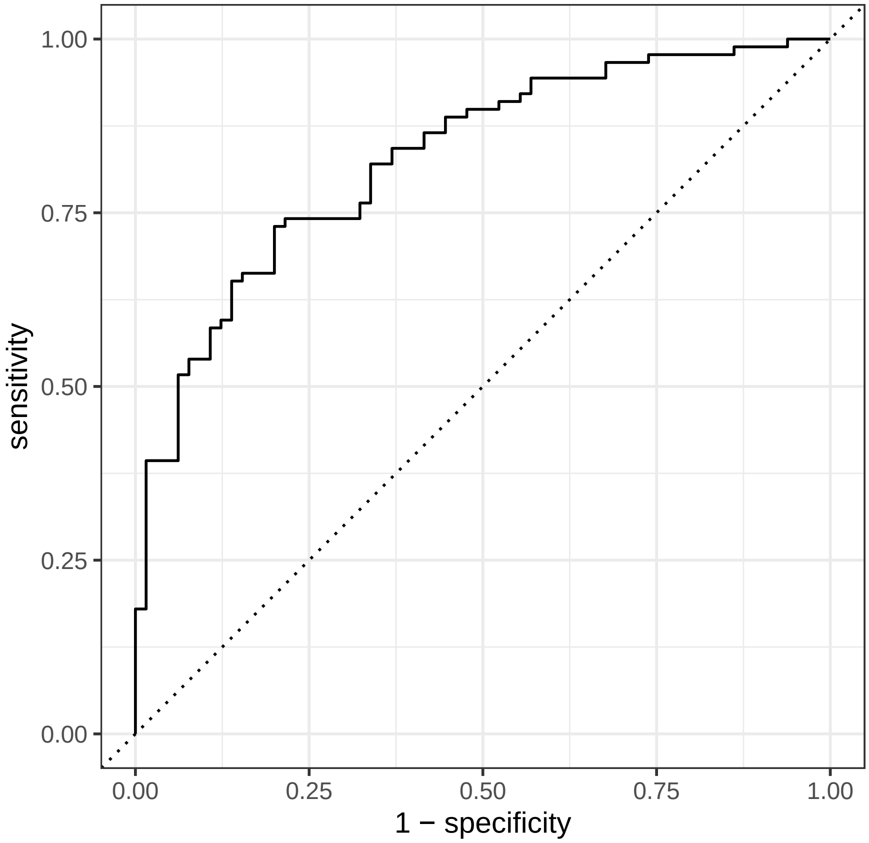
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   41    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    8    7
##        good  14   36
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   10   11
##        good   4   11
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    7    0
##        good   5    0
##
##
## Greatest deviations:
## # A tibble: 41 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1         0.466 bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  2         0.413 good        bad   FrBo       orig_Jak zajistit, aby skládka do~
##  3         0.366 bad         good  FrBo       red_Mohou spolky ve správních žal~
##  4         0.366 bad         good  FrBo       red_Mohou spolky ve správních žal~
##  5         0.352 good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
##  6         0.343 bad         good  KUKY       Odvolani
##  7         0.268 good        bad   FrBo       orig_Jak probíhá správní řízení
##  8         0.262 good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  9         0.232 bad         good  KUKY       invalidní důchod_1399-23_původní
## 10         0.217 bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
## 11         0.210 good        bad   FrBo       orig_územní řízení
## 12         0.203 good        bad   OmbuFlyers Studny
## 13         0.192 good        bad   FrBo       142
## 14         0.176 good        bad   FrBo       orig_Jak využít svého práva být i~
## 15         0.172 good        bad   OmbuFlyers Soudni-poplatky
## 16         0.154 bad         good  KUKY       AK_JH_Podani_US_podpis
## 17         0.152 good        bad   FrBo       64
## 18         0.143 good        bad   FrBo       orig_znalci, znalecké posudky
## 19         0.139 good        bad   FrBo       orig_Kdy a jak požadovat náhradu ~
## 20         0.119 good        bad   FrBo       orig_Co je to a jak probíhá integ~
## 21         0.103 good        bad   KUKY       Duchody
## 22         0.0970 bad        good  FrBo       red_pravni_nastroje_ochrany_ovzdu~
## 23         0.0899 good       bad   FrBo       orig_Sousedské vztahy
## 24         0.0755 good       bad   OmbuFlyers Detsky-domov
## 25         0.0690 good       bad   KUKY       Dopis vysvětlující dopis klientovi
## 26         0.0671 bad        good  KUKY       29 A 80-2021_20231122101241
## 27         0.0584 bad        good  KUKY       4842_2023_VOP
## 28         0.0536 good       bad   KUKY       Pravni rada_uver SVJ
## 29         0.0502 good       bad   FrBo       orig_Certifikáty autorizovaných i~
## 30         0.0486 good       bad   OmbuFlyers Katastr-nemovitosti
## 31         0.0398 bad        good  KUKY       Odvolani_proti_rozhodnuti_o_nepov~
## # i 10 more rows
```

IAC

```
lfit_rf_iac <- model_rf_iac %>% evaluate_tidymodel(split)
```

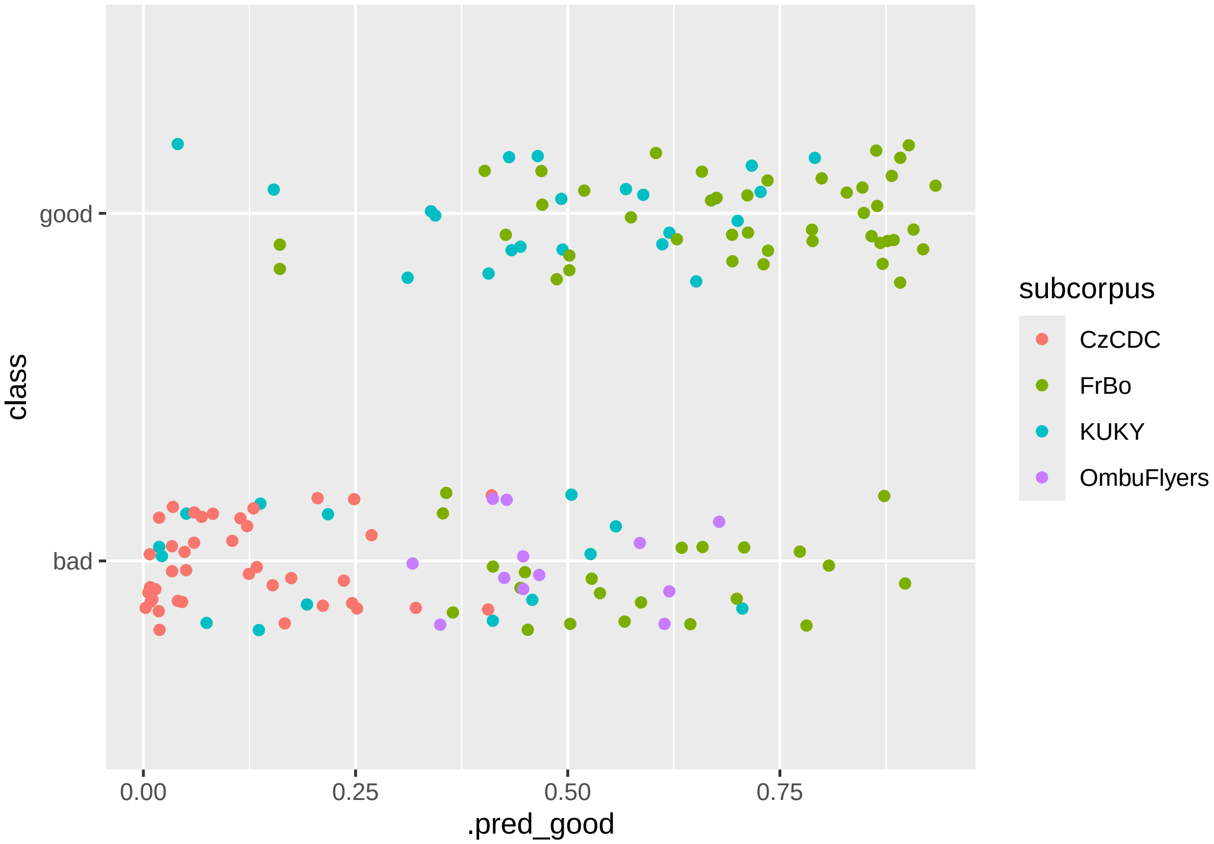
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.727 Preprocessor1_Model1
## 2 roc_auc     binary         0.828 Preprocessor1_Model1
## 3 brier_class binary         0.168 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 44 x 2
##    Variable                                              Importance
##    <chr>                                                      <dbl>
##  1 RuleTooManyNominalConstructions.max_allowable_nouns        15.8
##  2 activity                                                   15.3
##  3 verb_dist                                                  13.7
##  4 RuleTooFewVerbs.min_verb_frac                              13.5
##  5 RuleLongSentences.max_length                               12.4
##  6 ari                                                        11.4
##  7 gf                                                         11.4
##  8 smog                                                       10.7
##  9 RulePredAtClauseBeginning.max_order                         9.53
## 10 mamr                                                        6.63
## 11 fkgl                                                        6.48
## 12 atl                                                         6.39
## 13 RuleTooManyNegations.max_negation_frac                      6.02
## 14 maentropy                                                   5.98
## 15 RuleTooManyNominalConstructions.max_noun_frac               5.69
## 16 entropy                                                     5.62
## 17 mattr                                                       5.42
## 18 RulePredAtClauseBeginning.max_order.v                       5.05
## 19 maentropy.v                                                 4.95
## 20 cli                                                         4.70
## 21 RuleTooManyNominalConstructions.max_allowable_nouns.v       4.67
## 22 RuleLongSentences.max_length.v                              4.56
## 23 RuleInfVerbDistance.max_distance.v                          4.23
## 24 RulePredSubjDistance.max_distance                           4.22
## 25 mattr.v                                                     4.21
## 26 RuleDoubleAdpos.max_allowable_distance.v                    4.20
## 27 ttr                                                         4.04
## 28 RuleInfVerbDistance.max_distance                            3.94
## 29 RuleTooManyNegations.max_negation_frac.v                    3.93
## 30 RuleCaseRepetition.max_repetition_count.v                   3.90
## 31 RuleCaseRepetition.max_repetition_frac                      3.90
## 32 RulePredSubjDistance.max_distance.v                         3.82
## 33 RuleTooManyNegations.max_allowable_negations                3.70
## 34 RuleCaseRepetition.max_repetition_frac.v                    3.66
## 35 RulePredObjDistance.max_distance.v                          3.63
## 36 RulePredObjDistance.max_distance                            3.46
## 37 RuleTooManyNegations.max_allowable_negations.v              3.44
## 38 RuleMultiPartVerbs.max_distance                             3.41
## 39 RuleCaseRepetition.max_repetition_count                     3.31
## 40 hpoint                                                      3.22
## 41 RuleMultiPartVerbs.max_distance.v                           3.17
## 42 fre                                                         3.11
## 43 RuleTooManyNominalConstructions.max_noun_frac.v             3.08
## 44 RuleDoubleAdpos.max_allowable_distance                      3.08
```

```
lfit_rf_iac %>% get_mismatch_details(data)
```

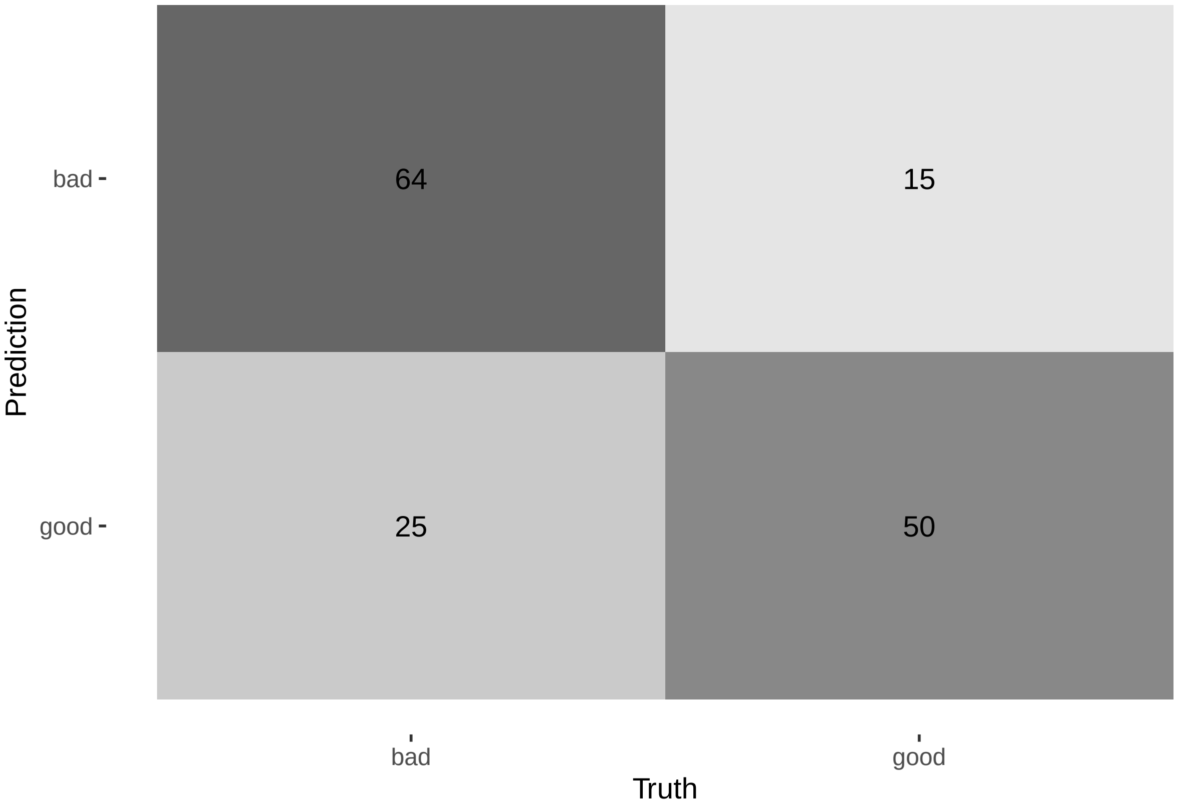


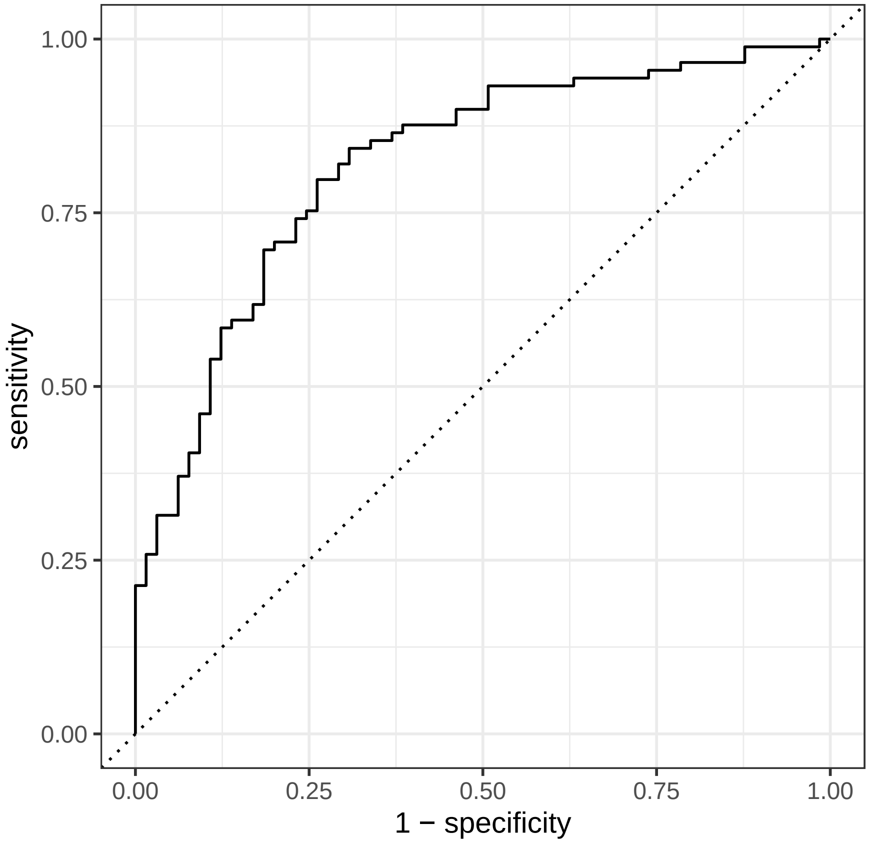
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   41    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    7    7
##        good  15   36
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   10   12
##        good   4   10
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    8    0
##        good   4    0
##
##
## Greatest deviations:
## # A tibble: 42 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.460  bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  2        0.398  good        bad   FrBo       orig_Jak zajistit, aby skládka do~
##  3        0.373  good        bad   FrBo       orig_Jak probíhá správní řízení
##  4        0.346  bad         good  KUKY       Odvolani
##  5        0.339  bad         good  FrBo       red_Mohou spolky ve správních žal~
##  6        0.339  bad         good  FrBo       red_Mohou spolky ve správních žal~
##  7        0.308  good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
##  8        0.281  good        bad   FrBo       orig_územní řízení
##  9        0.274  good        bad   FrBo       orig_Kdy a jak požadovat náhradu ~
## 10        0.208  good        bad   FrBo       142
## 11        0.206  good        bad   KUKY       Duchody
## 12        0.199  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
## 13        0.189  bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
## 14        0.178  good        bad   OmbuFlyers Studny
## 15        0.161  bad         good  KUKY       invalidní důchod_1399-23_původní
## 16        0.159  good        bad   FrBo       orig_znalci, znalecké posudky
## 17        0.156  bad         good  KUKY       AK_JH_Podani_US_podpis
## 18        0.145  good        bad   FrBo       orig_Jak využít svého práva být i~
## 19        0.134  good        bad   FrBo       64
## 20        0.120  good        bad   OmbuFlyers Soudni-poplatky
## 21        0.114  good        bad   OmbuFlyers Detsky-domov
## 22        0.0978 bad         good  FrBo       red_pravni_nastroje_ochrany_ovzdu~
## 23        0.0933 bad         good  KUKY       Odvolani_proti_rozhodnuti_o_nepov~
## 24        0.0864 good        bad   FrBo       orig_Certifikáty autorizovaných i~
## 25        0.0850 good        bad   OmbuFlyers Katastr-nemovitosti
## 26        0.0730 bad         good  FrBo       red_Les - co smíme a co je zakázá~
## 27        0.0691 bad         good  KUKY       Mestsky_urad_Vyzva_k_zaplaceni_na~
## 28        0.0670 good        bad   FrBo       68
## 29        0.0661 bad         good  KUKY       4842_2023_VOP
## 30        0.0567 good        bad   KUKY       Pravni rada_uver SVJ
## 31        0.0557 bad         good  KUKY       29 A 80-2021_20231122101241
## # i 11 more rows
```

Counts

```
lfit_rf_counts <- model_rf_counts %>% evaluate_tidymodel(split)
```

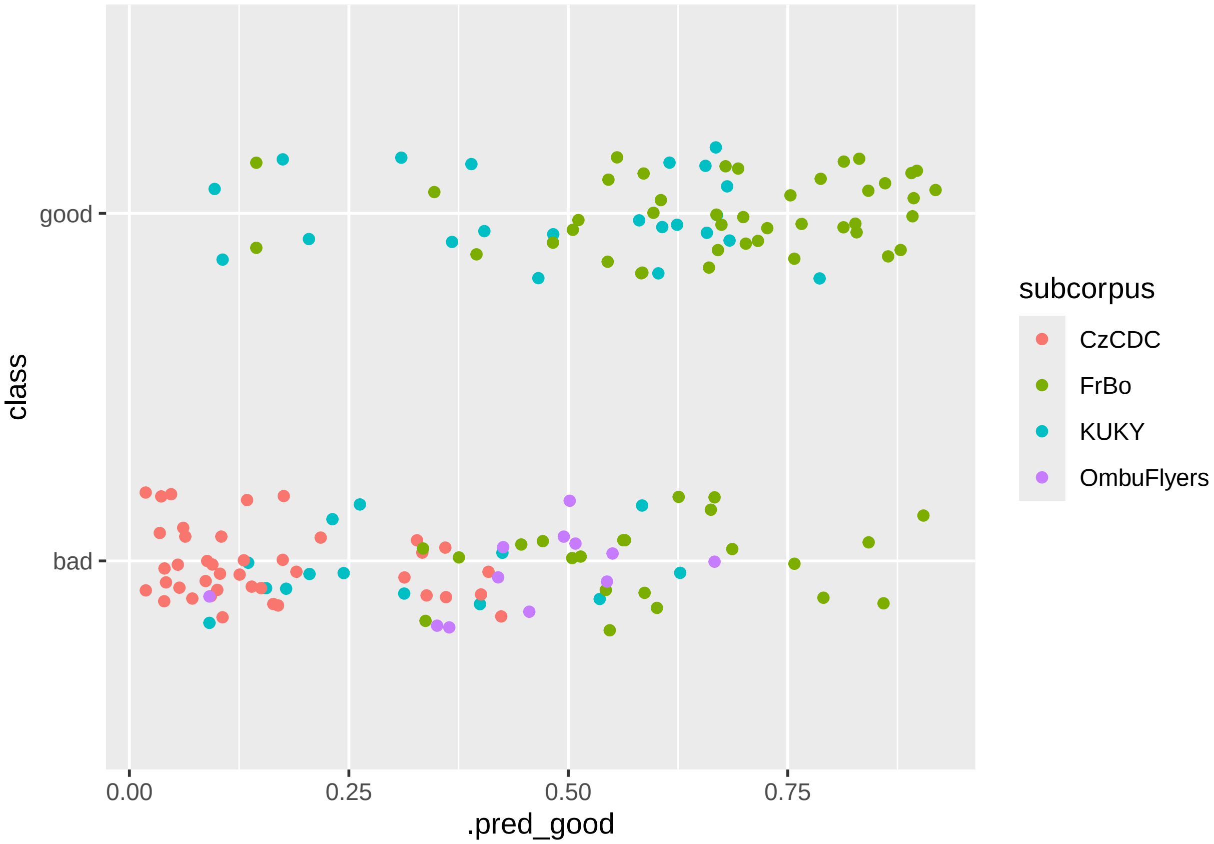
```
## # A tibble: 3 x 4
##   .metric     .estimator .estimate .config
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.740 Preprocessor1_Model1
## 2 roc_auc     binary         0.820 Preprocessor1_Model1
## 3 brier_class binary         0.175 Preprocessor1_Model1
```





```
## Variable importance:
## # A tibble: 28 x 2
##    Variable                      Importance
##    <chr>                              <dbl>
##  1 RuleMultiPartVerbs                29.9
##  2 RulePassive                       28.8
##  3 RuleLiteraryStyle                 27.8
##  4 RulePredSubjDistance              19.9
##  5 RuleInfVerbDistance               15.0
##  6 sent_count                        13.0
##  7 RuleVerbalNouns                   11.6
##  8 word_count                        10.2
##  9 num_hapax                          9.20
## 10 RulePredObjDistance                8.86
## 11 char_count                         8.81
## 12 RuleTooLongExpressions             8.61
## 13 syllab_count                       8.32
## 14 RuleDoubleAdpos                    7.65
## 15 RuleAbstractNouns                  7.29
## 16 RuleGPwordorder                    7.13
## 17 RuleAnaphoricReferences            6.61
## 18 RuleWeakMeaningWords               5.55
## 19 RuleReflexivePassWithAnimSubj      5.33
## 20 RuleGPdeverbsubj                   3.51
## 21 RuleGPpatinstr                     3.38
## 22 RuleGPdeverbaddr                   3.02
## 23 RuleGPpatbenperson                 2.19
## 24 RuleGPcoordovs                     1.89
## 25 RuleRelativisticExpressions        1.89
## 26 RuleConfirmationExpressions        1.27
## 27 RuleGPadjective                    0.597
## 28 RuleRedundantExpressions           0.570
```

```
lfit_rf_counts %>% get_mismatch_details(data)
```



```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   41    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad    5    5
##        good  17   38
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   11   10
##        good   3   12
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    7    0
##        good   5    0
##
##
## Greatest deviations:
## # A tibble: 40 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.405  good        bad   FrBo       orig_Co je to a jak probíhá integ~
##  2        0.403  bad         good  KUKY       Mestsky_urad_PRIKAZ_REV2
##  3        0.394  bad         good  KUKY       0217_6Afs_2000035_20210219141328_~
##  4        0.359  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  5        0.355  bad         good  FrBo       red_Mohou spolky ve správních žal~
##  6        0.355  bad         good  FrBo       red_Mohou spolky ve správních žal~
##  7        0.342  good        bad   FrBo       orig_Jaké otázky (ne)můžete polož~
##  8        0.325  bad         good  KUKY       invalidní důchod_1399-23_původní
##  9        0.296  bad         good  KUKY       AK_JH_Podani_US_podpis
## 10        0.291  good        bad   FrBo       64
## 11        0.258  good        bad   FrBo       orig_Jak zajistit, aby skládka do~
## 12        0.190  bad         good  KUKY       Odvolani
## 13        0.187  good        bad   FrBo       orig_Jak probíhá správní řízení
## 14        0.167  good        bad   FrBo       orig_Sousedské vztahy
## 15        0.167  good        bad   OmbuFlyers Socialni-sluzby
## 16        0.162  good        bad   FrBo       orig_Jaké právní nástroje můžete ~
## 17        0.153  bad         good  FrBo       red_Co je to úřední deska a jak j~
## 18        0.132  bad         good  KUKY       1732_2023_VOP
## 19        0.127  good        bad   KUKY       Dopis vysvětlující dopis klientovi
## 20        0.126  good        bad   FrBo       149
## 21        0.110  bad         good  KUKY       29 A 80-2021_20231122101241
## 22        0.105  bad         good  FrBo       orig_Nástroje občana při kontrole~
## 23        0.101  good        bad   FrBo       orig_Co je to EIA_final
## 24        0.0956 bad         good  KUKY       4842_2023_VOP
## 25        0.0870 good        bad   FrBo       142
## 26        0.0841 good        bad   KUKY       UOOUOsobniUdajePuvodne
## 27        0.0646 good        bad   FrBo       orig_Změny v zákoně o EIA
## 28        0.0626 good        bad   FrBo       orig_znalci, znalecké posudky
## 29        0.0504 good        bad   OmbuFlyers Zvlastni-opravneni
## 30        0.0472 good        bad   FrBo       orig_Certifikáty autorizovaných i~
## 31        0.0441 good        bad   OmbuFlyers Studny
## # i 9 more rows
```

Variable importances

```
prepare_vi_for_comparison <- function(final_fit) {
  model_vi <- get_vi(final_fit) %>%
    arrange(-Importance) %>%
    rowid_to_column("rank") %>%
    mutate(across(rank, ~ if_else(Importance == 0, NA, .x))) %>%
    select(rank, Variable, Importance)
}

importances <- full_join(
  prepare_vi_for_comparison(lfit_lasso_all),
  prepare_vi_for_comparison(lfit_lasso_notl),
```

```r
  by = "Variable",
  suffix = c(
    ".lasso.all",
    ".lasso.notl"
  )
) %>%
  full_join(
    prepare_vi_for_comparison(lfit_lasso_iac),
    by = "Variable",
  ) %>%
  full_join(
    prepare_vi_for_comparison(lfit_lasso_counts),
    by = "Variable",
    suffix = c(
      ".lasso.iac",
      ".lasso.counts"
    )
  ) %>%
  full_join(
    prepare_vi_for_comparison(lfit_rf_all),
    by = "Variable"
  ) %>%
  full_join(
    prepare_vi_for_comparison(lfit_rf_notl),
    by = "Variable",
    suffix = c(
      ".rf.all",
      ".rf.notl"
    )
  ) %>%
  full_join(
    prepare_vi_for_comparison(lfit_rf_iac),
    by = "Variable"
  ) %>%
  full_join(
    prepare_vi_for_comparison(lfit_rf_counts),
    by = "Variable",
    suffix = c(
      ".rf.iac",
      ".rf.counts"
    )
  ) %>%
  select(Variable, everything())
importances_df <- importances %>%
  select(-Variable) %>%
  select(starts_with("rank")) %>%
  as.data.frame()
rownames(importances_df) <- importances %>% pull(Variable)
print(importances_df)
```

```
##                                                        rank.lasso.all
## activity                                                            1
## smog                                                                2
## RuleLiteraryStyle                                                   3
## atl                                                                4
## mamr                                                               5
## gf                                                                 6
## entropy                                                            7
## maentropy                                                          8
## ari                                                                9
## RuleGPcoordovs                                                    NA
## RuleGPdeverbaddr                                                  NA
## RuleGPpatinstr                                                    NA
## RuleGPdeverbsubj                                                  NA
## RuleGPadjective                                                   NA
## RuleGPpatbenperson                                                NA
## RuleGPwordorder                                                   NA
## RuleDoubleAdpos                                                   NA
## RuleDoubleAdpos.max_allowable_distance                            NA
## RuleDoubleAdpos.max_allowable_distance.v                          NA
## RuleReflexivePassWithAnimSubj                                     NA
## RuleTooFewVerbs.min_verb_frac                                     NA
## RuleTooManyNegations.max_negation_frac                            NA
## RuleTooManyNegations.max_negation_frac.v                          NA
## RuleTooManyNegations.max_allowable_negations                      NA
## RuleTooManyNegations.max_allowable_negations.v                    NA
## RuleTooManyNominalConstructions.max_noun_frac                     NA
## RuleTooManyNominalConstructions.max_noun_frac.v                   NA
## RuleTooManyNominalConstructions.max_allowable_nouns               NA
## RuleCaseRepetition.max_repetition_count                           NA
## RuleCaseRepetition.max_repetition_count.v                         NA
## RuleCaseRepetition.max_repetition_frac                            NA
## RuleCaseRepetition.max_repetition_frac.v                          NA
## RuleWeakMeaningWords                                              NA
## RuleAbstractNouns                                                 NA
## RuleRelativisticExpressions                                       NA
## RuleConfirmationExpressions                                       NA
## RuleRedundantExpressions                                          NA
## RuleTooLongExpressions                                            NA
## RuleAnaphoricReferences                                           NA
## RulePassive                                                       NA
## RulePredSubjDistance                                              NA
## RulePredSubjDistance.max_distance                                 NA
## RulePredSubjDistance.max_distance.v                               NA
## RulePredObjDistance                                               NA
## RulePredObjDistance.max_distance                                  NA
## RulePredObjDistance.max_distance.v                                NA
## RuleInfVerbDistance                                               NA
## RuleInfVerbDistance.max_distance                                  NA
## RuleInfVerbDistance.max_distance.v                                NA
## RuleMultiPartVerbs                                                NA
## RuleMultiPartVerbs.max_distance                                   NA
## RuleMultiPartVerbs.max_distance.v                                 NA
## RuleLongSentences.max_length                                      NA
## RuleLongSentences.max_length.v                                    NA
## RulePredAtClauseBeginning.max_order                               NA
## RulePredAtClauseBeginning.max_order.v                             NA
## RuleVerbalNouns                                                   NA
## sent_count                                                        NA
## word_count                                                        NA
## syllab_count                                                      NA
## char_count                                                        NA
## cli                                                               NA
## num_hapax                                                         NA
## ttr                                                               NA
## mattr                                                             NA
## mattr.v                                                           NA
## maentropy.v                                                       NA
## verb_dist                                                         NA
## hpoint                                                            NA
## fre                                                               NA
## fkgl                                                              NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v             NA
##                                                        rank.lasso.notl
## activity                                                             1
## smog                                                                 2
## RuleLiteraryStyle                                                    3
## atl                                                                  4
## mamr                                                                 5
## gf                                                                   6
## entropy                                                              7
## maentropy                                                            8
## ari                                                                  9
## RuleGPcoordovs                                                      NA
## RuleGPdeverbaddr                                                    NA
## RuleGPpatinstr                                                      NA
## RuleGPdeverbsubj                                                    NA
## RuleGPadjective                                                     NA
## RuleGPpatbenperson                                                  NA
## RuleGPwordorder                                                     NA
## RuleDoubleAdpos                                                     NA
## RuleDoubleAdpos.max_allowable_distance                              NA
## RuleDoubleAdpos.max_allowable_distance.v                            NA
## RuleReflexivePassWithAnimSubj                                       NA
## RuleTooFewVerbs.min_verb_frac                                       NA
## RuleTooManyNegations.max_negation_frac                              NA
## RuleTooManyNegations.max_negation_frac.v                            NA
## RuleTooManyNegations.max_allowable_negations                        NA
## RuleTooManyNegations.max_allowable_negations.v                      NA
## RuleTooManyNominalConstructions.max_noun_frac                       NA
## RuleTooManyNominalConstructions.max_noun_frac.v                     NA
## RuleTooManyNominalConstructions.max_allowable_nouns                 NA
## RuleCaseRepetition.max_repetition_count                             NA
## RuleCaseRepetition.max_repetition_count.v                           NA
## RuleCaseRepetition.max_repetition_frac                              NA
## RuleCaseRepetition.max_repetition_frac.v                            NA
## RuleWeakMeaningWords                                                NA
## RuleAbstractNouns                                                   NA
## RuleRelativisticExpressions                                         NA
## RuleConfirmationExpressions                                         NA
## RuleRedundantExpressions                                            NA
## RuleTooLongExpressions                                              NA
## RuleAnaphoricReferences                                            NA
## RulePassive                                                        NA
## RulePredSubjDistance                                               NA
## RulePredSubjDistance.max_distance                                  NA
## RulePredSubjDistance.max_distance.v                                NA
## RulePredObjDistance                                                NA
## RulePredObjDistance.max_distance                                   NA
## RulePredObjDistance.max_distance.v                                 NA
## RuleInfVerbDistance                                                NA
## RuleInfVerbDistance.max_distance                                   NA
## RuleInfVerbDistance.max_distance.v                                 NA
## RuleMultiPartVerbs                                                 NA
## RuleMultiPartVerbs.max_distance                                    NA
## RuleMultiPartVerbs.max_distance.v                                  NA
## RuleLongSentences.max_length                                       NA
## RuleLongSentences.max_length.v                                     NA
## RulePredAtClauseBeginning.max_order                                NA
## RulePredAtClauseBeginning.max_order.v                              NA
## RuleVerbalNouns                                                    NA
## sent_count                                                         NA
## word_count                                                         NA
## syllab_count                                                       NA
## char_count                                                         NA
## cli                                                                NA
## num_hapax                                                          NA
## ttr                                                                NA
## mattr                                                              NA
## mattr.v                                                            NA
## maentropy.v                                                        NA
## verb_dist                                                          NA
## hpoint                                                             NA
## fre                                                                NA
## fkgl                                                               NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v              NA
##                                                       rank.lasso.iac
## activity                                                           3
## smog                                                              28
## RuleLiteraryStyle                                                 NA
## atl                                                                9
## mamr                                                              NA
## gf                                                                20
## entropy                                                           16
## maentropy                                                         NA
## ari                                                               18
## RuleGPcoordovs                                                    NA
## RuleGPdeverbaddr                                                  NA
## RuleGPpatinstr                                                    NA
## RuleGPdeverbsubj                                                  NA
## RuleGPadjective                                                   NA
## RuleGPpatbenperson                                                NA
## RuleGPwordorder                                                   NA
## RuleDoubleAdpos                                                   NA
## RuleDoubleAdpos.max_allowable_distance                            32
## RuleDoubleAdpos.max_allowable_distance.v                          21
## RuleReflexivePassWithAnimSubj                                     NA
## RuleTooFewVerbs.min_verb_frac                                      1
## RuleTooManyNegations.max_negation_frac                            17
## RuleTooManyNegations.max_negation_frac.v                          NA
## RuleTooManyNegations.max_allowable_negations                      NA
## RuleTooManyNegations.max_allowable_negations.v                    14
## RuleTooManyNominalConstructions.max_noun_frac                      5
## RuleTooManyNominalConstructions.max_noun_frac.v                    8
## RuleTooManyNominalConstructions.max_allowable_nouns               26
## RuleCaseRepetition.max_repetition_count                           NA
## RuleCaseRepetition.max_repetition_count.v                         10
## RuleCaseRepetition.max_repetition_frac                             2
## RuleCaseRepetition.max_repetition_frac.v                           7
## RuleWeakMeaningWords                                              NA
## RuleAbstractNouns                                                 NA
## RuleRelativisticExpressions                                       NA
## RuleConfirmationExpressions                                       NA
## RuleRedundantExpressions                                          NA
## RuleTooLongExpressions                                            NA
## RuleAnaphoricReferences                                           NA
## RulePassive                                                       NA
## RulePredSubjDistance                                              NA
## RulePredSubjDistance.max_distance                                 29
## RulePredSubjDistance.max_distance.v                               23
## RulePredObjDistance                                               NA
## RulePredObjDistance.max_distance                                  30
## RulePredObjDistance.max_distance.v                                NA
## RuleInfVerbDistance                                               NA
## RuleInfVerbDistance.max_distance                                  22
## RuleInfVerbDistance.max_distance.v                                15
## RuleMultiPartVerbs                                                NA
## RuleMultiPartVerbs.max_distance                                   NA
## RuleMultiPartVerbs.max_distance.v                                 19
## RuleLongSentences.max_length                                      25
## RuleLongSentences.max_length.v                                    11
## RulePredAtClauseBeginning.max_order                               31
## RulePredAtClauseBeginning.max_order.v                             NA
## RuleVerbalNouns                                                   NA
## sent_count                                                        NA
## word_count                                                        NA
## syllab_count                                                      NA
## char_count                                                        NA
## cli                                                               NA
## num_hapax                                                         NA
## ttr                                                               12
## mattr                                                              6
## mattr.v                                                           NA
## maentropy.v                                                        4
## verb_dist                                                         27
## hpoint                                                            33
## fre                                                               24
## fkgl                                                              NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v             13
##                                                       rank.lasso.counts
## activity                                                             NA
## smog                                                                 NA
## RuleLiteraryStyle                                                     5
## atl                                                                  NA
## mamr                                                                 NA
## gf                                                                   NA
## entropy                                                              NA
## maentropy                                                            NA
## ari                                                                  NA
## RuleGPcoordovs                                                       NA
## RuleGPdeverbaddr                                                      8
## RuleGPpatinstr                                                       NA
## RuleGPdeverbsubj                                                      4
## RuleGPadjective                                                       7
## RuleGPpatbenperson                                                   NA
## RuleGPwordorder                                                      NA
## RuleDoubleAdpos                                                      NA
## RuleDoubleAdpos.max_allowable_distance                               NA
## RuleDoubleAdpos.max_allowable_distance.v                             NA
## RuleReflexivePassWithAnimSubj                                        NA
## RuleTooFewVerbs.min_verb_frac                                        NA
## RuleTooManyNegations.max_negation_frac                               NA
## RuleTooManyNegations.max_negation_frac.v                             NA
## RuleTooManyNegations.max_allowable_negations                         NA
## RuleTooManyNegations.max_allowable_negations.v                       NA
## RuleTooManyNominalConstructions.max_noun_frac                        NA
## RuleTooManyNominalConstructions.max_noun_frac.v                      NA
## RuleTooManyNominalConstructions.max_allowable_nouns                  NA
## RuleCaseRepetition.max_repetition_count                              NA
## RuleCaseRepetition.max_repetition_count.v                            NA
## RuleCaseRepetition.max_repetition_frac                               NA
## RuleCaseRepetition.max_repetition_frac.v                             NA
## RuleWeakMeaningWords                                                 NA
## RuleAbstractNouns                                                    NA
## RuleRelativisticExpressions                                           2
## RuleConfirmationExpressions                                          NA
## RuleRedundantExpressions                                              1
## RuleTooLongExpressions                                                9
## RuleAnaphoricReferences                                               3
## RulePassive                                                           6
## RulePredSubjDistance                                                 11
## RulePredSubjDistance.max_distance                                    NA
## RulePredSubjDistance.max_distance.v                                  NA
## RulePredObjDistance                                                  NA
## RulePredObjDistance.max_distance                                     NA
## RulePredObjDistance.max_distance.v                                   NA
## RuleInfVerbDistance                                                  13
## RuleInfVerbDistance.max_distance                                     NA
## RuleInfVerbDistance.max_distance.v                                   NA
## RuleMultiPartVerbs                                                   10
## RuleMultiPartVerbs.max_distance                                      NA
## RuleMultiPartVerbs.max_distance.v                                    NA
## RuleLongSentences.max_length                                         NA
## RuleLongSentences.max_length.v                                       NA
## RulePredAtClauseBeginning.max_order                                  NA
## RulePredAtClauseBeginning.max_order.v                                NA
## RuleVerbalNouns                                                      12
## sent_count                                                           14
## word_count                                                           15
## syllab_count                                                         NA
## char_count                                                           NA
## cli                                                                  NA
## num_hapax                                                            NA
## ttr                                                                  NA
## mattr                                                                NA
## mattr.v                                                              NA
## maentropy.v                                                          NA
## verb_dist                                                            NA
## hpoint                                                               NA
## fre                                                                  NA
## fkgl                                                                 NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v                NA
##                                                       rank.rf.all rank.rf.notl
## activity                                                        4            2
## smog                                                            9            7
## RuleLiteraryStyle                                               8            9
## atl                                                            13           15
## mamr                                                           12           13
## gf                                                              7            6
## entropy                                                        23           17
## maentropy                                                      19           16
## ari                                                             6            8
## RuleGPcoordovs                                                 68           64
## RuleGPdeverbaddr                                               64           60
## RuleGPpatinstr                                                 63           59
## RuleGPdeverbsubj                                               66           62
## RuleGPadjective                                                70           65
## RuleGPpatbenperson                                             67           63
## RuleGPwordorder                                                62           58
## RuleDoubleAdpos                                                51           50
## RuleDoubleAdpos.max_allowable_distance                         58           52
## RuleDoubleAdpos.max_allowable_distance.v                       29           33
## RuleReflexivePassWithAnimSubj                                  61           57
## RuleTooFewVerbs.min_verb_frac                                   5            4
## RuleTooManyNegations.max_negation_frac                         18           18
## RuleTooManyNegations.max_negation_frac.v                       39           41
## RuleTooManyNegations.max_allowable_negations                   42           38
## RuleTooManyNegations.max_allowable_negations.v                 54           44
## RuleTooManyNominalConstructions.max_noun_frac                  22           20
## RuleTooManyNominalConstructions.max_noun_frac.v                56           53
## RuleTooManyNominalConstructions.max_allowable_nouns             3            3
## RuleCaseRepetition.max_repetition_count                        38           45
## RuleCaseRepetition.max_repetition_count.v                      32           32
## RuleCaseRepetition.max_repetition_frac                         37           36
## RuleCaseRepetition.max_repetition_frac.v                       34           34
## RuleWeakMeaningWords                                           60           54
## RuleAbstractNouns                                              55           56
## RuleRelativisticExpressions                                    65           61
## RuleConfirmationExpressions                                    71           67
## RuleRedundantExpressions                                       69           66
## RuleTooLongExpressions                                         21           22
## RuleAnaphoricReferences                                        25           23
## RulePassive                                                    11           11
## RulePredSubjDistance                                           26           25
## RulePredSubjDistance.max_distance                              31           26
## RulePredSubjDistance.max_distance.v                            41           37
## RulePredObjDistance                                            35           35
## RulePredObjDistance.max_distance                               46           49
## RulePredObjDistance.max_distance.v                             47           47
## RuleInfVerbDistance                                            52           51
## RuleInfVerbDistance.max_distance                               36           40
## RuleInfVerbDistance.max_distance.v                             43           39
## RuleMultiPartVerbs                                             15           19
## RuleMultiPartVerbs.max_distance                                48           43
## RuleMultiPartVerbs.max_distance.v                              45           48
## RuleLongSentences.max_length                                    2            5
## RuleLongSentences.max_length.v                                 28           30
## RulePredAtClauseBeginning.max_order                            10           10
## RulePredAtClauseBeginning.max_order.v                          16           14
## RuleVerbalNouns                                                20           24
## sent_count                                                     57           NA
## word_count                                                     33           NA
## syllab_count                                                   50           NA
## char_count                                                     49           NA
## cli                                                            27           29
## num_hapax                                                      40           42
## ttr                                                            44           31
## mattr                                                          17           21
## mattr.v                                                        30           28
## maentropy.v                                                    24           27
## verb_dist                                                       1            1
## hpoint                                                         59           55
## fre                                                            53           46
## fkgl                                                           14           12
## RuleTooManyNominalConstructions.max_allowable_nouns.v          NA           NA
##                                                        rank.rf.iac
## activity                                                        2
## smog                                                            8
## RuleLiteraryStyle                                              NA
## atl                                                            12
## mamr                                                           10
## gf                                                              7
## entropy                                                        16
## maentropy                                                      14
## ari                                                             6
## RuleGPcoordovs                                                 NA
## RuleGPdeverbaddr                                               NA
## RuleGPpatinstr                                                 NA
## RuleGPdeverbsubj                                               NA
## RuleGPadjective                                                NA
## RuleGPpatbenperson                                             NA
## RuleGPwordorder                                                NA
## RuleDoubleAdpos                                                NA
## RuleDoubleAdpos.max_allowable_distance                         44
## RuleDoubleAdpos.max_allowable_distance.v                       26
## RuleReflexivePassWithAnimSubj                                  NA
## RuleTooFewVerbs.min_verb_frac                                   4
## RuleTooManyNegations.max_negation_frac                         13
## RuleTooManyNegations.max_negation_frac.v                       29
## RuleTooManyNegations.max_allowable_negations                   33
## RuleTooManyNegations.max_allowable_negations.v                 37
## RuleTooManyNominalConstructions.max_noun_frac                  15
## RuleTooManyNominalConstructions.max_noun_frac.v                43
## RuleTooManyNominalConstructions.max_allowable_nouns             1
## RuleCaseRepetition.max_repetition_count                        39
## RuleCaseRepetition.max_repetition_count.v                      30
## RuleCaseRepetition.max_repetition_frac                         31
## RuleCaseRepetition.max_repetition_frac.v                       34
## RuleWeakMeaningWords                                           NA
## RuleAbstractNouns                                              NA
## RuleRelativisticExpressions                                    NA
## RuleConfirmationExpressions                                    NA
## RuleRedundantExpressions                                       NA
## RuleTooLongExpressions                                         NA
## RuleAnaphoricReferences                                        NA
## RulePassive                                                    NA
## RulePredSubjDistance                                           NA
## RulePredSubjDistance.max_distance                              24
## RulePredSubjDistance.max_distance.v                            32
## RulePredObjDistance                                            NA
## RulePredObjDistance.max_distance                               36
## RulePredObjDistance.max_distance.v                             35
## RuleInfVerbDistance                                            NA
## RuleInfVerbDistance.max_distance                               28
## RuleInfVerbDistance.max_distance.v                             23
## RuleMultiPartVerbs                                             NA
## RuleMultiPartVerbs.max_distance                                38
## RuleMultiPartVerbs.max_distance.v                              41
## RuleLongSentences.max_length                                    5
## RuleLongSentences.max_length.v                                 22
## RulePredAtClauseBeginning.max_order                             9
## RulePredAtClauseBeginning.max_order.v                          18
## RuleVerbalNouns                                                NA
## sent_count                                                     NA
## word_count                                                     NA
## syllab_count                                                   NA
## char_count                                                     NA
## cli                                                            20
## num_hapax                                                      NA
## ttr                                                            27
## mattr                                                          17
## mattr.v                                                        25
## maentropy.v                                                    19
## verb_dist                                                       3
## hpoint                                                         40
## fre                                                            42
## fkgl                                                           11
## RuleTooManyNominalConstructions.max_allowable_nouns.v          21
##                                                       rank.rf.counts
## activity                                                          NA
## smog                                                              NA
## RuleLiteraryStyle                                                  3
## atl                                                               NA
## mamr                                                              NA
## gf                                                                NA
## entropy                                                           NA
## maentropy                                                         NA
## ari                                                               NA
## RuleGPcoordovs                                                    24
## RuleGPdeverbaddr                                                  22
## RuleGPpatinstr                                                    21
## RuleGPdeverbsubj                                                  20
## RuleGPadjective                                                   27
## RuleGPpatbenperson                                                23
## RuleGPwordorder                                                   16
## RuleDoubleAdpos                                                   14
## RuleDoubleAdpos.max_allowable_distance                            NA
## RuleDoubleAdpos.max_allowable_distance.v                          NA
## RuleReflexivePassWithAnimSubj                                     19
## RuleTooFewVerbs.min_verb_frac                                     NA
## RuleTooManyNegations.max_negation_frac                            NA
## RuleTooManyNegations.max_negation_frac.v                          NA
## RuleTooManyNegations.max_allowable_negations                      NA
## RuleTooManyNegations.max_allowable_negations.v                    NA
## RuleTooManyNominalConstructions.max_noun_frac                     NA
## RuleTooManyNominalConstructions.max_noun_frac.v                   NA
## RuleTooManyNominalConstructions.max_allowable_nouns               NA
## RuleCaseRepetition.max_repetition_count                           NA
## RuleCaseRepetition.max_repetition_count.v                         NA
## RuleCaseRepetition.max_repetition_frac                            NA
## RuleCaseRepetition.max_repetition_frac.v                          NA
## RuleWeakMeaningWords                                              18
## RuleAbstractNouns                                                 15
## RuleRelativisticExpressions                                       25
## RuleConfirmationExpressions                                       26
## RuleRedundantExpressions                                          28
## RuleTooLongExpressions                                            12
## RuleAnaphoricReferences                                           17
## RulePassive                                                        2
## RulePredSubjDistance                                               4
## RulePredSubjDistance.max_distance                                 NA
## RulePredSubjDistance.max_distance.v                               NA
## RulePredObjDistance                                               10
## RulePredObjDistance.max_distance                                  NA
## RulePredObjDistance.max_distance.v                                NA
## RuleInfVerbDistance                                                5
## RuleInfVerbDistance.max_distance                                  NA
## RuleInfVerbDistance.max_distance.v                                NA
## RuleMultiPartVerbs                                                 1
## RuleMultiPartVerbs.max_distance                                   NA
## RuleMultiPartVerbs.max_distance.v                                 NA
## RuleLongSentences.max_length                                      NA
## RuleLongSentences.max_length.v                                    NA
## RulePredAtClauseBeginning.max_order                               NA
## RulePredAtClauseBeginning.max_order.v                             NA
## RuleVerbalNouns                                                    7
## sent_count                                                         6
## word_count                                                         8
## syllab_count                                                      13
## char_count                                                        11
## cli                                                               NA
## num_hapax                                                          9
## ttr                                                               NA
## mattr                                                             NA
## mattr.v                                                           NA
## maentropy.v                                                       NA
## verb_dist                                                         NA
## hpoint                                                            NA
## fre                                                               NA
## fkgl                                                              NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v             NA
```

```
importances %>%
  mutate(
    mean_rank = rowMeans(
      select(importances, starts_with("rank")),
      na.rm = TRUE
    ),
    no_of_irrelevance = rowSums(
      select(importances, starts_with("rank")) %>% is.na()
    )
  ) %>%
  select(Variable, mean_rank, no_of_irrelevance) %>%
  arrange(mean_rank) %>%
  print(n = 100)
```

```
## # A tibble: 72 x 3
##    Variable                                          mean_rank no_of_irrelevance
##    <chr>                                                 <dbl>             <dbl>
##  1 activity                                               2.17                 2
##  2 RuleTooFewVerbs.min_verb_frac                          3.5                  4
##  3 RuleLiteraryStyle                                      5.17                 2
##  4 RulePassive                                            7.5                  4
##  5 verb_dist                                              8                    4
##  6 RuleTooManyNominalConstructions.max_allowable_no~      8.25                 4
##  7 gf                                                     8.67                 2
##  8 mamr                                                   9                    3
##  9 RuleLongSentences.max_length                           9.25                 4
## 10 smog                                                   9.33                 2
## 11 ari                                                    9.33                 2
## 12 atl                                                    9.5                  2
## 13 RuleMultiPartVerbs                                    11.2                  4
## 14 fkgl                                                  12.3                  5
## 15 maentropy                                             13                    3
## 16 entropy                                               14.3                  2
## 17 RulePredAtClauseBeginning.max_order                   15                    4
## 18 mattr                                                 15.2                  4
## 19 RuleTooManyNominalConstructions.max_noun_frac         15.5                  4
## 20 RuleVerbalNouns                                       15.8                  4
## 21 RuleTooLongExpressions                                16                    4
## 22 RulePredAtClauseBeginning.max_order.v                 16                    5
## 23 RuleTooManyNegations.max_negation_frac                16.5                  4
## 24 RulePredSubjDistance                                  16.5                  4
## 25 RuleAnaphoricReferences                               17                    4
## 26 RuleTooManyNominalConstructions.max_allowable_no~     17                    6
## 27 maentropy.v                                           18.5                  4
## 28 word_count                                            18.7                  5
## 29 RuleLongSentences.max_length.v                        22.8                  4
## 30 cli                                                   25.3                  5
## 31 sent_count                                            25.7                  5
## 32 RuleCaseRepetition.max_repetition_count.v             26                    4
## 33 RuleCaseRepetition.max_repetition_frac                26.5                  4
## 34 RulePredObjDistance                                   26.7                  5
## 35 RuleDoubleAdpos.max_allowable_distance.v              27.2                  4
## 36 RuleCaseRepetition.max_repetition_frac.v              27.2                  4
## 37 RulePredSubjDistance.max_distance                     27.5                  4
## 38 mattr.v                                               27.7                  5
## 39 ttr                                                   28.5                  4
## 40 RuleInfVerbDistance.max_distance.v                    30                    4
## 41 char_count                                            30                    6
## 42 RuleInfVerbDistance                                   30.2                  4
## 43 num_hapax                                             30.3                  5
## 44 RuleInfVerbDistance.max_distance                      31.5                  4
## 45 syllab_count                                          31.5                  6
## 46 RulePredSubjDistance.max_distance.v                   33.2                  4
## 47 RuleTooManyNegations.max_negation_frac.v              36.3                  5
## 48 RuleTooManyNegations.max_allowable_negations.v        37.2                  4
## 49 RuleTooManyNegations.max_allowable_negations          37.7                  5
## 50 RuleGPdeverbsubj                                      38                    4
## 51 RuleRelativisticExpressions                           38.2                  4
## 52 RuleMultiPartVerbs.max_distance.v                     38.2                  4
## 53 RuleDoubleAdpos                                       38.3                  5
## 54 RuleGPdeverbaddr                                      38.5                  4
## 55 RuleTooManyNominalConstructions.max_noun_frac.v       40                    4
## 56 RulePredObjDistance.max_distance                      40.2                  4
## 57 RuleCaseRepetition.max_repetition_count               40.7                  5
## 58 RuleRedundantExpressions                              41                    4
## 59 fre                                                   41.2                  4
## 60 RuleAbstractNouns                                     42                    5
## 61 RuleGPadjective                                       42.2                  4
## 62 RulePredObjDistance.max_distance.v                    43                    5
## 63 RuleMultiPartVerbs.max_distance                       43                    5
## 64 RuleWeakMeaningWords                                  44                    5
## 65 RuleGPwordorder                                       45.3                  5
## 66 RuleReflexivePassWithAnimSubj                         45.7                  5
## 67 RuleDoubleAdpos.max_allowable_distance                46.5                  4
## 68 hpoint                                                46.8                  4
## 69 RuleGPpatinstr                                        47.7                  5
## 70 RuleGPpatbenperson                                    51                    5
## 71 RuleGPcoordovs                                      52                  5
## 72 RuleConfirmationExpressions                          54.7                5
```