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()
## * Use tidymodels_prefer() to resolve common conflicts.
```

Load and tidy data

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

```
## Rows: 754 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()
```
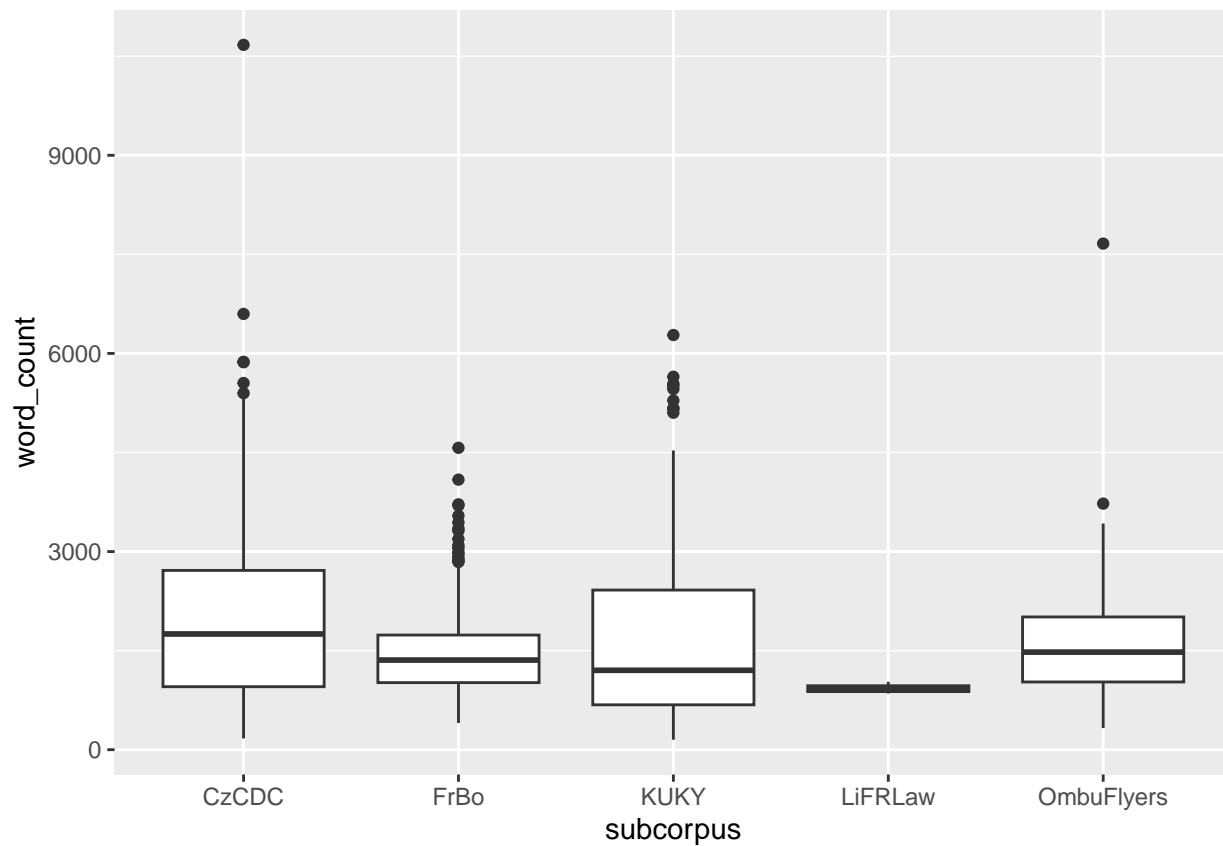


```
data %>% ggplot(aes(x = class, word_count)) +
  geom_boxplot()
```
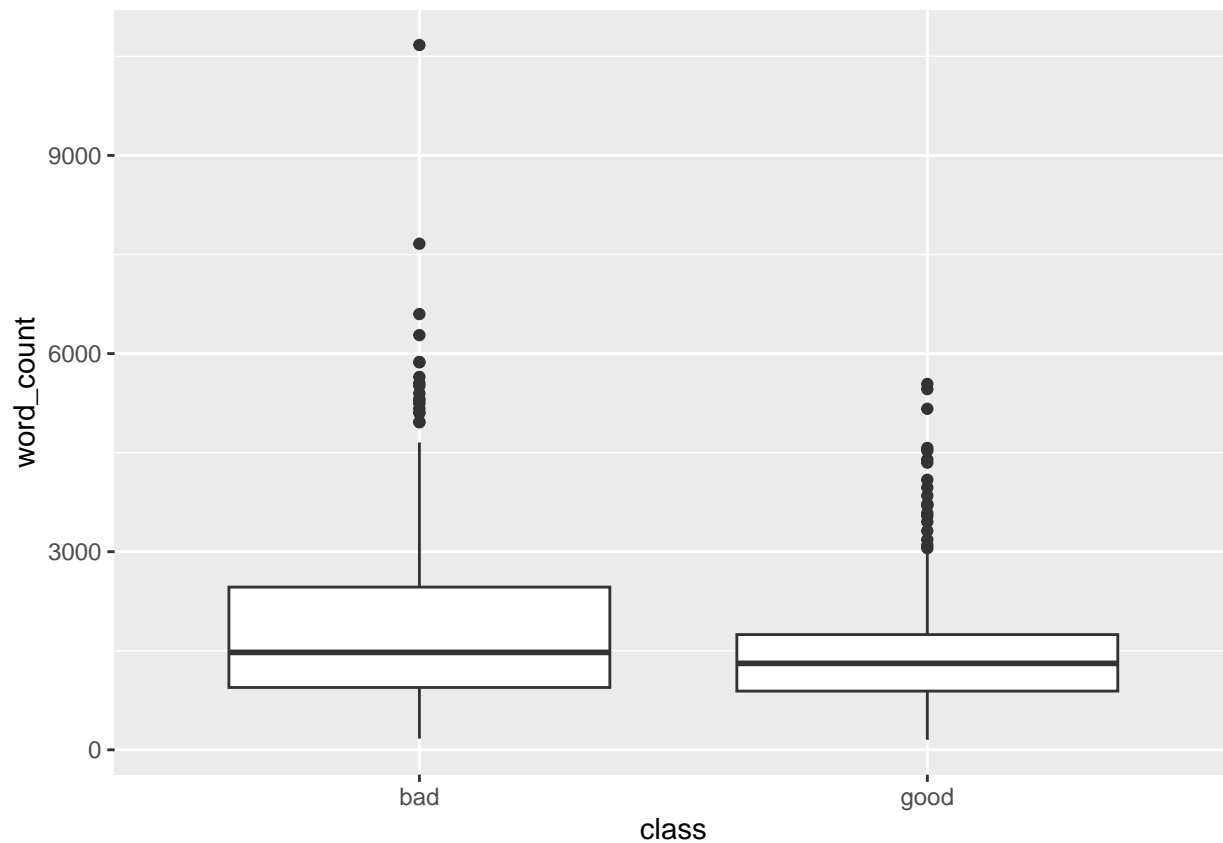
```r
data_clean <- data %>%
  select(!c(
    fpath,
    # KUK_ID,
    # FileName,
    FolderPath,
    # subcorpus,
    DocumentTitle,
    ClarityPursuit,
    Readability,
    SyllogismBased,
    SourceDB
  )) %>%
  # replace -1s in variation coefficients with NAs
  mutate(across(c(
    `RuleDoubleAdpos.max_allowable_distance.v`,
    `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 84
## # i 84 variables: KUK_ID <chr>, strata <chr>, class <fct>, FileName <chr>,
## #   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>
## <603/151/754>
```

```
print(folds)
```

```
## #  10-fold cross-validation using stratification
## # A tibble: 10 x 2
##     splits          id
##     <list>          <chr>
##   1 <split [540/63]> Fold01
##   2 <split [540/63]> Fold02
##   3 <split [541/62]> Fold03
##   4 <split [541/62]> Fold04
##   5 <split [543/60]> Fold05
##   6 <split [544/59]> Fold06
##   7 <split [544/59]> Fold07
##   8 <split [544/59]> Fold08
##   9 <split [545/58]> Fold09
## 10 <split [545/58]> Fold10
```

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

```
## 
##              bad good
##   CzCDC      175    0
##   FrBo        62  178
##   KUKY        68   90
##   LiFRLaw      2    0
##   OmbuFlyers  28    0
```

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

```
## 
##              bad good
##   CzCDC       39    0
##   FrBo        16   51
##   KUKY        14   20
##   LiFRLaw      1    0
##   OmbuFlyers  10    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")
```

```r
  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

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

Lasso

```r
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 %>%
```

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

  best_roc_auc <- lasso_tune_rs %>%
    select_by_one_std_err(metric = "roc_auc", -penalty)

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

  final_lasso <- lasso_tune_wf %>% finalize_workflow(best_roc_auc)
  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(lambda = best_roc_auc %>% pull(penalty)) %>%
    print(n = 100)

  return(final_lasso)
}
```

SVM

```r
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")
```

```r
  # 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)
}

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)
```

```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)
}

# 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,
    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

```
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)
```

```r
  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

```r
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
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()
```

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

```
## # A tibble: 11 x 2
##    terms                              id
##    <chr>                              <chr>
##  1 RuleCaseRepetition.max_repetition_frac.v corr_jzELQ
##  2 char_count                         corr_jzELQ
##  3 ari                                corr_jzELQ
##  4 ttr                                corr_jzELQ
##  5 maentropy                          corr_jzELQ
##  6 hpoint                             corr_jzELQ
##  7 atl                                corr_jzELQ
##  8 gf                                 corr_jzELQ
##  9 smog                               corr_jzELQ
## 10 word_count                         corr_jzELQ
## 11 RuleLongSentences.max_length       corr_jzELQ
```

```
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

```
# 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 +
```

```r
  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

```
# 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
```

```r
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: 24
##
## -- Operations
## * Centering and scaling for: <none>
```

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

```
## # A tibble: 0 x 2
## # i 2 variables: terms <dbl>, id <chr>
```

```r
recipe_counts
```

```
##
## -- Recipe ---------------------------------------------------------------
##
## -- Inputs
## Number of variables by role
## outcome:    1
## predictor: 24
##
## -- Operations
## * Centering and scaling for: <none>
## * Correlation filter on: all_numeric_predictors()
```

Indicators, averages, and coefficients

```r
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>
```

```r
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_tc2c2
## 2 ari                                  corr_tc2c2
## 3 maentropy                            corr_tc2c2
## 4 atl                                  corr_tc2c2
## 5 gf                                   corr_tc2c2
## 6 smog                                 corr_tc2c2
## 7 RuleLongSentences.max_length         corr_tc2c2
```

```r
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

```r
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) {
```

```r
  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")

  deviations <- joined %>%
    filter(.pred_class != class) %>%
    mutate(deviation = .pred_good - 0.5) %>%
    mutate(abs_deviation = abs(deviation)) %>%
    arrange(-abs_deviation)

  cat("Greatest deviations:\n")
  deviations %>%
    select(abs_deviation, .pred_class, class, subcorpus, FileName) %>%
    print(n = round(nrow(joined) / 5))

  cat("Highest-deviating documents names:\n")
  highest_deviation_names <- deviations %>%
    filter(abs_deviation >= 0.25) %>%
    arrange(-abs_deviation) %>%
    pull(FileName)

  print(highest_deviation_names)

  return(list(deviations = deviations, highest_deviations = highest_deviation_names))
}

plot_outlier <- function(doc_name, final_fit, dataset) {
  important_variables <- get_vi(final_fit) %>%
    head(n = 6) %>%
    filter(Importance > 0) %>%
    pull(Variable)

  dmut <- dataset %>%
    select(KUK_ID, FileName, class, all_of(important_variables)) %>%
    mutate(across(all_of(important_variables), ~ scale(.x))) %>%
    pivot_longer(
      all_of(important_variables),
      names_to = "feature", values_to = "value"
    ) %>%
```

```
    mutate(across(value, ~ .x[, 1]))

  dmut %>%
    ggplot(aes(x = class, y = value)) +
    facet_wrap(~feature) +
    geom_boxplot() +
    geom_point(
      data = dmut %>% filter(FileName == doc_name), color = "red", size = 5
    ) +
    labs(y = "measurements (scaled)")
}
```

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 [540/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [540/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [541/62]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [541/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [543/60]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [544/59]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [544/59]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [544/59]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [545/58]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [545/58]> 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     0.556    10 0.00414  Preprocessor1_Model1
## 2 brier_class binary     0.247    10 0.000453 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 [540/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [540/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [541/62]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [541/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [543/60]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
```

```
##  6 <split [544/59]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [544/59]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [544/59]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [545/58]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [545/58]> 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 [540/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [540/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [541/62]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [541/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [543/60]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [544/59]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [544/59]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [544/59]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [545/58]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [545/58]> 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     0.556    10 0.00414  Preprocessor1_Model1
## 2 brier_class binary     0.247    10 0.000453 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 [540/63]> Fold01 <tibble [3 x 4]> <tibble [0 x 3]>
##  2 <split [540/63]> Fold02 <tibble [3 x 4]> <tibble [0 x 3]>
##  3 <split [541/62]> Fold03 <tibble [3 x 4]> <tibble [0 x 3]>
##  4 <split [541/62]> Fold04 <tibble [3 x 4]> <tibble [0 x 3]>
##  5 <split [543/60]> Fold05 <tibble [3 x 4]> <tibble [0 x 3]>
##  6 <split [544/59]> Fold06 <tibble [3 x 4]> <tibble [0 x 3]>
##  7 <split [544/59]> Fold07 <tibble [3 x 4]> <tibble [0 x 3]>
##  8 <split [544/59]> Fold08 <tibble [3 x 4]> <tibble [0 x 3]>
##  9 <split [545/58]> Fold09 <tibble [3 x 4]> <tibble [0 x 3]>
## 10 <split [545/58]> 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)                                            -7.450e-01  1.810e-01
## RuleGPcoordovs                                         -2.400e-01  1.251e-01
## RuleGPdeverbaddr                                       -2.226e-01  1.292e-01
## RuleGPpatinstr                                         -2.733e-01  1.389e-01
## RuleGPdeverbsubj                                       -2.319e-01  1.162e-01
## RuleGPadjective                                         3.380e-01  1.503e-01
## RuleGPpatbenperson                                     -1.578e-01  1.345e-01
## RuleGPwordorder                                        -4.532e-02  1.462e-01
## RuleDoubleAdpos                                        -5.831e-02  1.664e-01
## RuleDoubleAdpos.max_allowable_distance                 -9.469e-02  2.614e-01
## RuleDoubleAdpos.max_allowable_distance.v                3.276e-01  2.324e-01
## RuleReflexivePassWithAnimSubj                           3.199e-02  1.368e-01
## RuleTooFewVerbs.min_verb_frac                          -1.380e+00  5.299e-01
## RuleTooManyNegations.max_negation_frac                  7.524e-03  1.974e-01
## RuleTooManyNegations.max_negation_frac.v                3.376e-02  1.578e-01
## RuleTooManyNegations.max_allowable_negations            3.160e-01  2.769e-01
## RuleTooManyNegations.max_allowable_negations.v         -1.996e-01  2.431e-01
## RuleTooManyNominalConstructions.max_noun_frac          -2.825e-01  2.312e-01
## RuleTooManyNominalConstructions.max_noun_frac.v         2.332e-01  1.565e-01
## RuleTooManyNominalConstructions.max_allowable_nouns     2.821e-01  4.935e-01
## RuleCaseRepetition.max_repetition_count                -1.497e-01  3.702e-01
## RuleCaseRepetition.max_repetition_count.v              -2.547e-01  1.920e-01
## RuleCaseRepetition.max_repetition_frac                 -2.120e-01  9.337e-01
## RuleCaseRepetition.max_repetition_frac.v                1.961e-01  9.273e-01
## RuleWeakMeaningWords                                    1.815e-02  1.388e-01
## RuleAbstractNouns                                       8.582e-02  1.390e-01
## RuleRelativisticExpressions                            -4.170e-01  1.769e-01
## RuleConfirmationExpressions                             5.597e-01  1.709e-01
## RuleRedundantExpressions                               -5.724e-02  1.752e-01
## RuleTooLongExpressions                                  1.196e-01  1.418e-01
## RuleAnaphoricReferences                                 6.321e-01  1.473e-01
## RuleLiteraryStyle                                      -2.263e-01  1.645e-01
## RulePassive                                            -6.438e-01  2.137e-01
## RulePredSubjDistance                                    3.943e-01  2.226e-01
```

```
## RulePredSubjDistance.max_distance                     -1.012e+00  2.931e-01
## RulePredSubjDistance.max_distance.v                   -1.976e-01  2.184e-01
## RulePredObjDistance                                   -8.086e-02  2.625e-01
## RulePredObjDistance.max_distance                       7.289e-05  2.677e-01
## RulePredObjDistance.max_distance.v                     1.401e-01  2.041e-01
## RuleInfVerbDistance                                    3.368e-02  2.753e-01
## RuleInfVerbDistance.max_distance                       2.202e-01  1.496e-01
## RuleInfVerbDistance.max_distance.v                    -2.026e-01  1.937e-01
## RuleMultiPartVerbs                                     1.494e-02  2.541e-01
## RuleMultiPartVerbs.max_distance                       -1.061e-01  2.964e-01
## RuleMultiPartVerbs.max_distance.v                      1.130e-01  2.000e-01
## RuleLongSentences.max_length                           3.248e+00  1.031e+00
## RuleLongSentences.max_length.v                         6.332e-01  2.078e-01
## RulePredAtClauseBeginning.max_order                   -3.824e-03  2.924e-01
## RulePredAtClauseBeginning.max_order.v                  5.747e-02  2.805e-01
## RuleVerbalNouns                                        8.998e-02  1.607e-01
## sent_count                                             2.153e+00  7.629e-01
## word_count                                            -6.383e+00  4.245e+00
## syllab_count                                          -1.769e+01  7.498e+00
## char_count                                             2.415e+01  9.812e+00
## cli                                                    1.176e+00  2.277e+00
## ari                                                   -5.458e+00  2.195e+00
## num_hapax                                              2.081e-01  1.032e+00
## entropy                                               -6.733e-01  3.723e-01
## ttr                                                   -7.841e-01  1.398e+00
## mattr                                                  1.997e-01  1.095e+00
## mattr.v                                               -6.532e-01  4.290e-01
## maentropy                                             -5.555e-01  1.127e+00
## maentropy.v                                            1.223e+00  6.357e-01
## mamr                                                  -1.388e-02  3.085e-01
## verb_dist                                              5.231e-01  3.120e-01
## activity                                               2.222e+00  5.840e-01
## hpoint                                                -2.681e+00  9.692e-01
## atl                                                   -1.442e+00  2.690e+00
## fre                                                   -2.928e+00  1.125e+00
## fkgl                                                          NA         NA
## gf                                                    -1.604e+00  2.569e+00
## smog                                                   7.215e-01  2.070e+00
##                                                       z value Pr(>|z|)
## (Intercept)                                            -4.115 3.88e-05 ***
## RuleGPcoordovs                                         -1.919 0.055016 .
## RuleGPdeverbaddr                                       -1.724 0.084788 .
## RuleGPpatinstr                                         -1.967 0.049147 *
## RuleGPdeverbsubj                                       -1.996 0.045932 *
## RuleGPadjective                                         2.249 0.024515 *
## RuleGPpatbenperson                                     -1.173 0.240835
## RuleGPwordorder                                        -0.310 0.756534
## RuleDoubleAdpos                                        -0.350 0.725974
## RuleDoubleAdpos.max_allowable_distance                 -0.362 0.717230
## RuleDoubleAdpos.max_allowable_distance.v               1.410 0.158655
## RuleReflexivePassWithAnimSubj                          0.234 0.815085
## RuleTooFewVerbs.min_verb_frac                          -2.604 0.009227 **
## RuleTooManyNegations.max_negation_frac                 0.038 0.969602
## RuleTooManyNegations.max_negation_frac.v               0.214 0.830607
```

```
## RuleTooManyNegations.max_allowable_negations              1.141 0.253791
## RuleTooManyNegations.max_allowable_negations.v           -0.821 0.411585
## RuleTooManyNominalConstructions.max_noun_frac            -1.222 0.221761
## RuleTooManyNominalConstructions.max_noun_frac.v           1.490 0.136113
## RuleTooManyNominalConstructions.max_allowable_nouns       0.572 0.567497
## RuleCaseRepetition.max_repetition_count                  -0.404 0.685877
## RuleCaseRepetition.max_repetition_count.v                -1.326 0.184741
## RuleCaseRepetition.max_repetition_frac                   -0.227 0.820359
## RuleCaseRepetition.max_repetition_frac.v                  0.211 0.832535
## RuleWeakMeaningWords                                      0.131 0.895931
## RuleAbstractNouns                                         0.617 0.536968
## RuleRelativisticExpressions                              -2.357 0.018414 *
## RuleConfirmationExpressions                               3.274 0.001059 **
## RuleRedundantExpressions                                 -0.327 0.743903
## RuleTooLongExpressions                                    0.844 0.398902
## RuleAnaphoricReferences                                   4.291 1.78e-05 ***
## RuleLiteraryStyle                                        -1.376 0.168900
## RulePassive                                              -3.013 0.002589 **
## RulePredSubjDistance                                      1.772 0.076462 .
## RulePredSubjDistance.max_distance                        -3.454 0.000553 ***
## RulePredSubjDistance.max_distance.v                      -0.905 0.365651
## RulePredObjDistance                                      -0.308 0.758036
## RulePredObjDistance.max_distance                          0.000 0.999783
## RulePredObjDistance.max_distance.v                        0.686 0.492450
## RuleInfVerbDistance                                       0.122 0.902644
## RuleInfVerbDistance.max_distance                          1.472 0.140955
## RuleInfVerbDistance.max_distance.v                       -1.046 0.295643
## RuleMultiPartVerbs                                        0.059 0.953099
## RuleMultiPartVerbs.max_distance                          -0.358 0.720408
## RuleMultiPartVerbs.max_distance.v                         0.565 0.572133
## RuleLongSentences.max_length                              3.149 0.001637 **
## RuleLongSentences.max_length.v                            3.047 0.002315 **
## RulePredAtClauseBeginning.max_order                      -0.013 0.989567
## RulePredAtClauseBeginning.max_order.v                     0.205 0.837633
## RuleVerbalNouns                                           0.560 0.575633
## sent_count                                                2.822 0.004775 **
## word_count                                               -1.504 0.132636
## syllab_count                                             -2.359 0.018325 *
## char_count                                                2.461 0.013854 *
## cli                                                       0.517 0.605356
## ari                                                      -2.487 0.012896 *
## num_hapax                                                 0.202 0.840113
## entropy                                                  -1.808 0.070534 .
## ttr                                                      -0.561 0.574936
## mattr                                                     0.182 0.855343
## mattr.v                                                  -1.523 0.127801
## maentropy                                                -0.493 0.622089
## maentropy.v                                               1.924 0.054382 .
## mamr                                                     -0.045 0.964128
## verb_dist                                                 1.677 0.093636 .
## activity                                                  3.804 0.000142 ***
## hpoint                                                   -2.766 0.005680 **
## atl                                                      -0.536 0.591953
## fre                                                      -2.603 0.009236 **
```

```
## fkgl                                                              NA       NA
## gf                                                            -0.624 0.532353
## smog                                                           0.349 0.727388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 828.48  on 602  degrees of freedom
## Residual deviance: 409.00  on 532  degrees of freedom
## AIC: 551
##
## Number of Fisher Scoring iterations: 7
```

Indicators, averages, and coefficients

```r
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)                                           -0.485315   0.135304
## RuleDoubleAdpos.max_allowable_distance                 0.124592   0.197691
## RuleDoubleAdpos.max_allowable_distance.v              -0.124484   0.172429
## RuleTooFewVerbs.min_verb_frac                         -1.074415   0.427968
## RuleTooManyNegations.max_negation_frac                 0.011705   0.171404
## RuleTooManyNegations.max_negation_frac.v               0.117015   0.129299
## RuleTooManyNegations.max_allowable_negations           0.230022   0.240110
## RuleTooManyNegations.max_allowable_negations.v        -0.328178   0.207044
## RuleTooManyNominalConstructions.max_noun_frac         -0.385240   0.187521
## RuleTooManyNominalConstructions.max_noun_frac.v        0.151785   0.131303
## RuleTooManyNominalConstructions.max_allowable_nouns    0.474443   0.404167
## RuleTooManyNominalConstructions.max_allowable_nouns.v -0.211813   0.186629
## RuleCaseRepetition.max_repetition_count                0.046290   0.290951
## RuleCaseRepetition.max_repetition_count.v             -0.376412   0.163926
## RuleCaseRepetition.max_repetition_frac                 0.068135   0.793295
## RuleCaseRepetition.max_repetition_frac.v               0.385322   0.781831
## RulePredSubjDistance.max_distance                     -0.597528   0.275426
## RulePredSubjDistance.max_distance.v                   -0.089574   0.180387
## RulePredObjDistance.max_distance                      -0.171092   0.275808
## RulePredObjDistance.max_distance.v                     0.074113   0.170041
## RuleInfVerbDistance.max_distance                       0.142428   0.116819
## RuleInfVerbDistance.max_distance.v                    -0.350950   0.150909
## RuleMultiPartVerbs.max_distance                       -0.002272   0.240486
## RuleMultiPartVerbs.max_distance.v                      0.144803   0.177500
## RuleLongSentences.max_length                           2.612078   0.881488
```

```
## RuleLongSentences.max_length.v                                0.518320   0.174887
## RulePredAtClauseBeginning.max_order                           0.014461   0.309973
## RulePredAtClauseBeginning.max_order.v                        -0.083262   0.226972
## cli                                                          -0.153504   1.729663
## ari                                                          -3.641986   1.421287
## entropy                                                      -0.133483   0.299296
## ttr                                                          -0.305837   0.341453
## mattr                                                         0.101935   0.840677
## mattr.v                                                      -0.503228   0.373581
## maentropy                                                    -0.450348   0.846024
## maentropy.v                                                   0.857332   0.574551
## mamr                                                         -0.134925   0.238009
## verb_dist                                                     0.408672   0.270109
## activity                                                      1.996512   0.391700
## hpoint                                                       -0.249828   0.366353
## atl                                                           1.175201   1.893936
## fre                                                          -1.506646   0.574457
## fkgl                                                                NA         NA
## gf                                                           -1.050525   2.139429
## smog                                                          0.146754   1.710213
##                                                              z value Pr(>|z|)
## (Intercept)                                                   -3.587 0.000335 ***
## RuleDoubleAdpos.max_allowable_distance                         0.630 0.528541
## RuleDoubleAdpos.max_allowable_distance.v                      -0.722 0.470330
## RuleTooFewVerbs.min_verb_frac                                 -2.511 0.012056 *
## RuleTooManyNegations.max_negation_frac                         0.068 0.945557
## RuleTooManyNegations.max_negation_frac.v                       0.905 0.365468
## RuleTooManyNegations.max_allowable_negations                   0.958 0.338070
## RuleTooManyNegations.max_allowable_negations.v                -1.585 0.112952
## RuleTooManyNominalConstructions.max_noun_frac                 -2.054 0.039939 *
## RuleTooManyNominalConstructions.max_noun_frac.v                1.156 0.247688
## RuleTooManyNominalConstructions.max_allowable_nouns            1.174 0.240444
## RuleTooManyNominalConstructions.max_allowable_nouns.v         -1.135 0.256402
## RuleCaseRepetition.max_repetition_count                        0.159 0.873591
## RuleCaseRepetition.max_repetition_count.v                     -2.296 0.021663 *
## RuleCaseRepetition.max_repetition_frac                         0.086 0.931555
## RuleCaseRepetition.max_repetition_frac.v                       0.493 0.622122
## RulePredSubjDistance.max_distance                             -2.169 0.030047 *
## RulePredSubjDistance.max_distance.v                           -0.497 0.619496
## RulePredObjDistance.max_distance                              -0.620 0.535040
## RulePredObjDistance.max_distance.v                             0.436 0.662941
## RuleInfVerbDistance.max_distance                              1.219 0.222762
## RuleInfVerbDistance.max_distance.v                            -2.326 0.020041 *
## RuleMultiPartVerbs.max_distance                              -0.009 0.992462
## RuleMultiPartVerbs.max_distance.v                             0.816 0.414621
## RuleLongSentences.max_length                                  2.963 0.003044 **
## RuleLongSentences.max_length.v                                2.964 0.003039 **
## RulePredAtClauseBeginning.max_order                           0.047 0.962789
## RulePredAtClauseBeginning.max_order.v                        -0.367 0.713739
## cli                                                          -0.089 0.929282
## ari                                                          -2.562 0.010393 *
## entropy                                                      -0.446 0.655605
## ttr                                                          -0.896 0.370416
## mattr                                                         0.121 0.903491
```

```
## mattr.v                                      -1.347 0.177967
## maentropy                                    -0.532 0.594511
## maentropy.v                                   1.492 0.135652
## mamr                                         -0.567 0.570788
## verb_dist                                     1.513 0.130283
## activity                                       5.097 3.45e-07 ***
## hpoint                                        -0.682 0.495281
## atl                                            0.621 0.534924
## fre                                           -2.623 0.008723 **
## fkgl                                              NA       NA
## gf                                            -0.491 0.623405
## smog                                           0.086 0.931617
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 828.48  on 602  degrees of freedom
## Residual deviance: 502.98  on 559  degrees of freedom
## AIC: 590.98
##
## Number of Fisher Scoring iterations: 6
```

Counts

```glm
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)                   -0.409019   0.110283  -3.709 0.000208 ***
## RuleGPcoordovs                -0.185174   0.104529  -1.772 0.076477 .
## RuleGPdeverbaddr              -0.152961   0.112171  -1.364 0.172682
## RuleGPpatinstr                -0.003148   0.102088  -0.031 0.975399
## RuleGPdeverbsubj              -0.299533   0.132567  -2.259 0.023854 *
## RuleGPadjective                0.235592   0.129553   1.819 0.068987 .
## RuleGPpatbenperson            -0.076535   0.100870  -0.759 0.448002
## RuleGPwordorder               -0.073571   0.118355  -0.622 0.534194
## RuleDoubleAdpos               -0.113453   0.108640  -1.044 0.296345
## RuleReflexivePassWithAnimSubj  0.107873   0.108506   0.994 0.320141
## RuleWeakMeaningWords           0.029967   0.107677   0.278 0.780781
## RuleAbstractNouns              0.083220   0.109713   0.759 0.448140
## RuleRelativisticExpressions   -0.464180   0.156740  -2.961 0.003062 **
## RuleConfirmationExpressions    0.191687   0.117427   1.632 0.102597
## RuleRedundantExpressions      -0.232668   0.166078  -1.401 0.161228
## RuleTooLongExpressions         0.123506   0.105835   1.167 0.243223
```

```
## RuleAnaphoricReferences          0.464418   0.118392    3.923 8.76e-05 ***
## RuleLiteraryStyle               -0.481297   0.128396   -3.749 0.000178 ***
## RulePassive                     -0.994309   0.145213   -6.847 7.53e-12 ***
## RulePredSubjDistance             0.436763   0.136887    3.191 0.001419 **
## RulePredObjDistance             -0.168758   0.139622   -1.209 0.226787
## RuleInfVerbDistance              0.358416   0.143624    2.496 0.012577 *
## RuleMultiPartVerbs               0.411360   0.145989    2.818 0.004836 **
## RuleVerbalNouns                  0.312812   0.115151    2.717 0.006597 **
## num_hapax                        0.127466   0.113074    1.127 0.259625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 828.48  on 602  degrees of freedom
## Residual deviance: 555.75  on 578  degrees of freedom
## AIC: 605.75
##
## Number of Fisher Scoring iterations: 5
```

Decision tree

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

All variables

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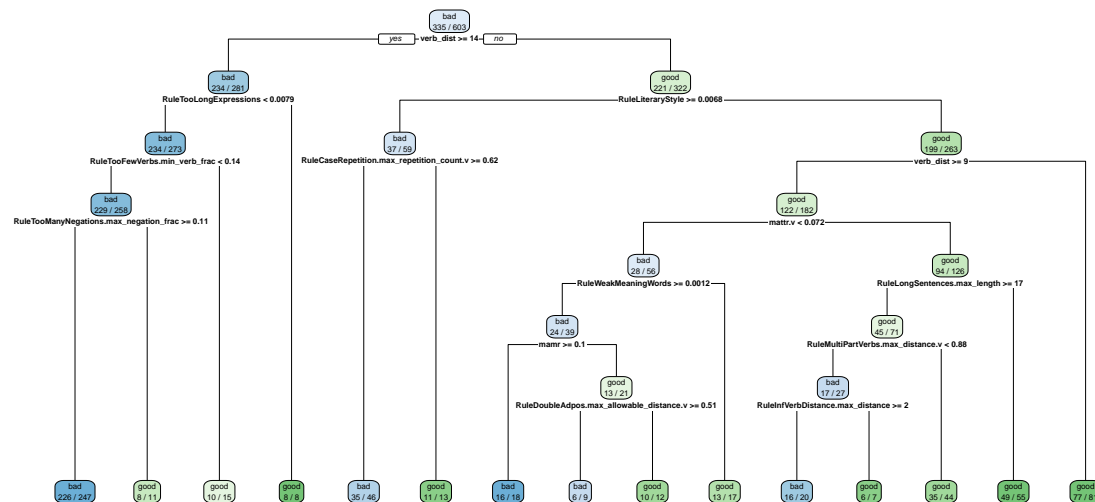

No TL

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

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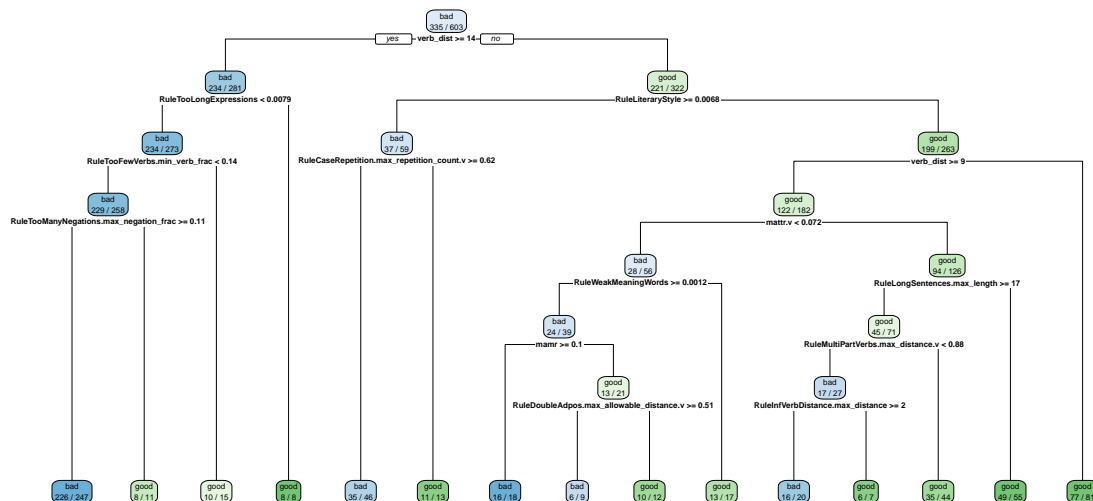
```r
model_dt_iac <- train_decision_tree(formula_iac, training_set)
```
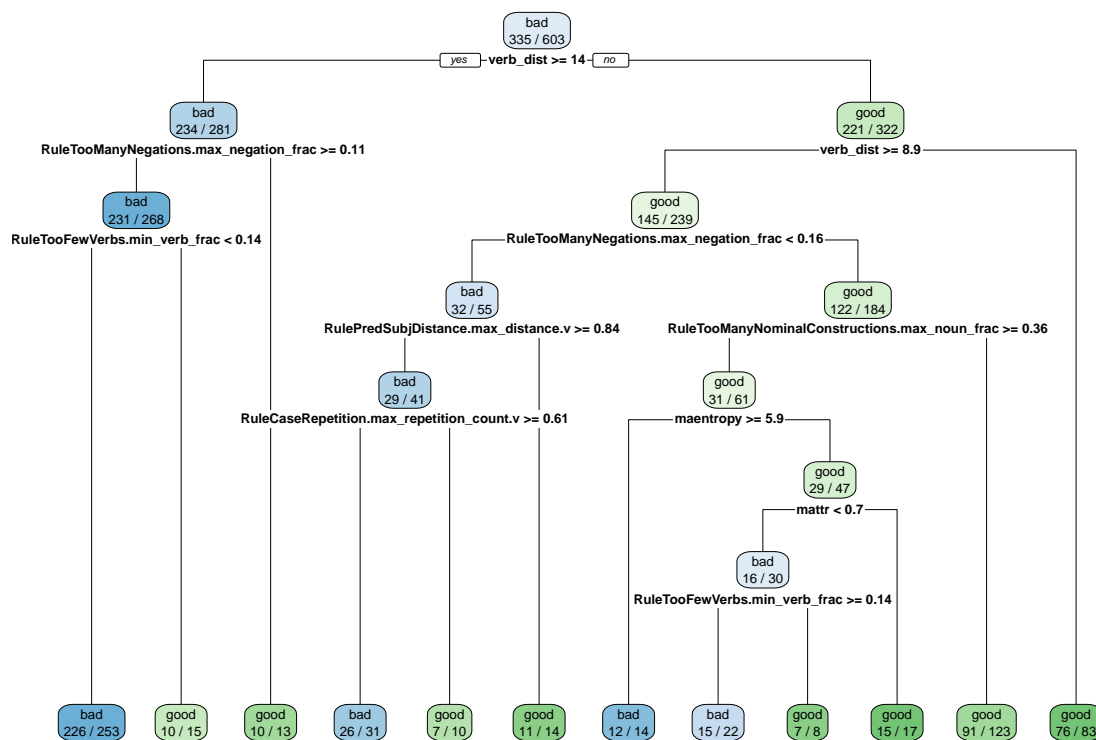

Counts

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

bad
335 / 603
yes — RulePassive >= 0.0076 — *no*

bad
199 / 252

good
215 / 351

RuleVerbalNouns < 0.059

RulePredSubjDistance < 0.055

good
31 / 53

bad
88 / 155

good
148 / 196

RuleMultiPartVerbs < 0.035

RulePassive >= 0.002

RuleLiteraryStyle >= 0.0068

good
30 / 46

bad
81 / 121

bad
22 / 40

RuleAnaphoricReferences < 0.0016

RuleAnaphoricReferences < 0.0012

RuleDoubleAdpos >= 0.0059

good
18 / 33

bad
76 / 104

good
16 / 29

RuleGPwordorder < 0.0013

RuleLiteraryStyle >= 0.0022

RuleTooLongExpressions >= 0.0011

bad
12 / 21

RuleLiteraryStyle >= 0.0079

bad
177 / 199

bad
6 / 7

bad
12 / 18

good
12 / 15

good
12 / 13

bad
69 / 86

good
11 / 18

good
12 / 17

good
27 / 34

bad
9 / 11

bad
10 / 13

good
6 / 8

good
7 / 8

good
130 / 156

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:
```
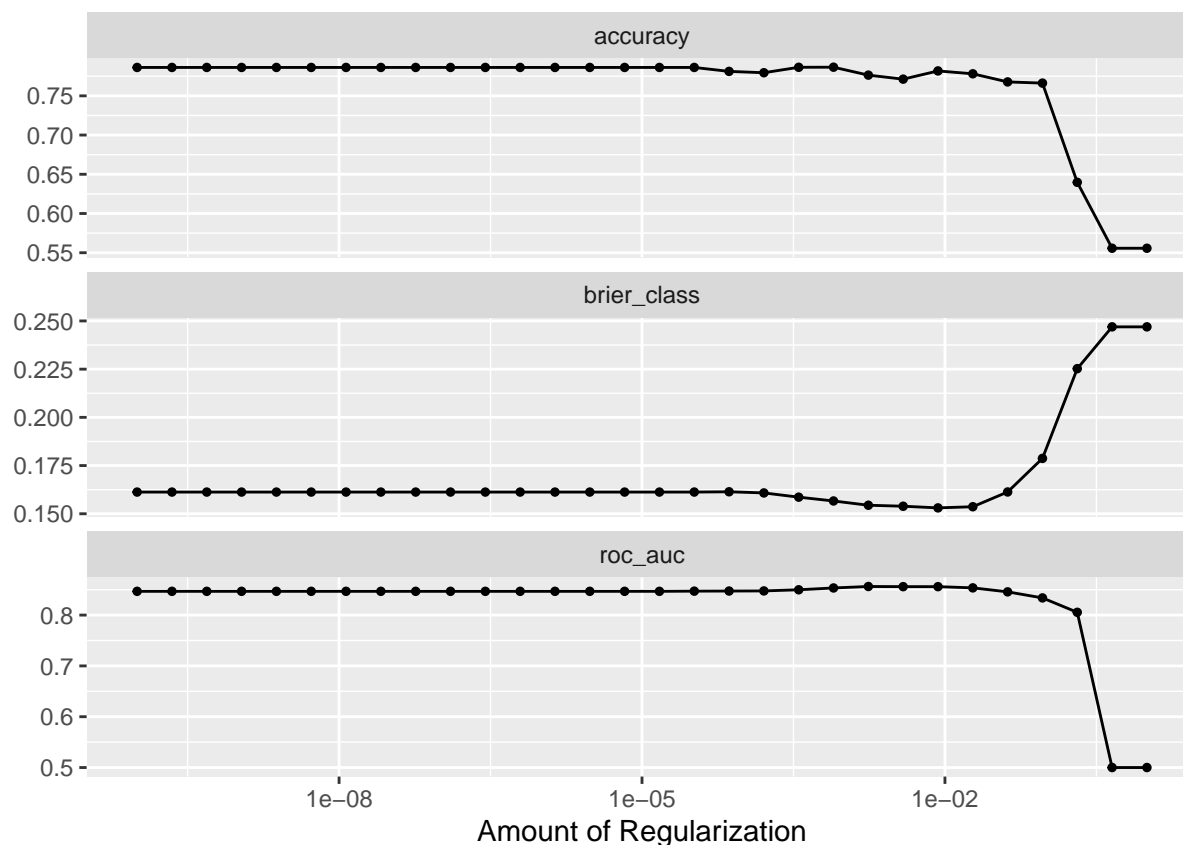
```
## # A tibble: 5 x 7
##    penalty .metric .estimator  mean     n std_err .config
##      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00174  roc_auc binary     0.856    10 0.0120  Preprocessor1_Model22
## 2 0.00386  roc_auc binary     0.856    10 0.0111  Preprocessor1_Model23
## 3 0.00853  roc_auc binary     0.856    10 0.00828 Preprocessor1_Model24
## 4 0.0189   roc_auc binary     0.854    10 0.00798 Preprocessor1_Model25
## 5 0.000788 roc_auc binary     0.853    10 0.0129  Preprocessor1_Model21
## # A tibble: 5 x 7
##    penalty .metric  .estimator  mean     n std_err .config
##      <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 7.88e- 4 accuracy binary     0.786    10  0.0168 Preprocessor1_Model21
## 2 3.56e- 4 accuracy binary     0.786    10  0.0183 Preprocessor1_Model20
## 3 1    e-10 accuracy binary    0.786    10  0.0162 Preprocessor1_Model01
## 4 2.21e-10 accuracy binary     0.786    10  0.0162 Preprocessor1_Model02
## 5 4.89e-10 accuracy binary     0.786    10  0.0162 Preprocessor1_Model03
## Best ROC AUC:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0418 Preprocessor1_Model26
## Final workflow:
## == Workflow ======================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor --------------------------------------------------
```

```
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model -----------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.0417531893656041
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 72 x 3
##    term                                                 estimate penalty
##    <chr>                                                   <dbl>   <dbl>
##  1 (Intercept)                                            -0.294  0.0418
##  2 RuleLiteraryStyle                                      -0.267  0.0418
##  3 smog                                                   -0.182  0.0418
##  4 RulePassive                                            -0.173  0.0418
##  5 maentropy                                              -0.162  0.0418
##  6 entropy                                               -0.0937  0.0418
##  7 RuleGPcoordovs                                               0  0.0418
##  8 RuleGPdeverbaddr                                             0  0.0418
##  9 RuleGPpatinstr                                               0  0.0418
## 10 RuleGPdeverbsubj                                             0  0.0418
## 11 RuleGPadjective                                              0  0.0418
## 12 RuleGPpatbenperson                                           0  0.0418
## 13 RuleGPwordorder                                              0  0.0418
## 14 RuleDoubleAdpos                                              0  0.0418
## 15 RuleDoubleAdpos.max_allowable_distance                       0  0.0418
## 16 RuleDoubleAdpos.max_allowable_distance.v                     0  0.0418
## 17 RuleReflexivePassWithAnimSubj                                0  0.0418
## 18 RuleTooFewVerbs.min_verb_frac                                0  0.0418
## 19 RuleTooManyNegations.max_negation_frac                       0  0.0418
## 20 RuleTooManyNegations.max_negation_frac.v                     0  0.0418
## 21 RuleTooManyNegations.max_allowable_negations                 0  0.0418
## 22 RuleTooManyNegations.max_allowable_negations.v               0  0.0418
## 23 RuleTooManyNominalConstructions.max_noun_frac                0  0.0418
## 24 RuleTooManyNominalConstructions.max_noun_frac.v              0  0.0418
## 25 RuleTooManyNominalConstructions.max_allowable_nouns          0  0.0418
## 26 RuleCaseRepetition.max_repetition_count                      0  0.0418
## 27 RuleCaseRepetition.max_repetition_count.v                    0  0.0418
## 28 RuleCaseRepetition.max_repetition_frac                       0  0.0418
## 29 RuleCaseRepetition.max_repetition_frac.v                     0  0.0418
## 30 RuleWeakMeaningWords                                         0  0.0418
## 31 RuleAbstractNouns                                            0  0.0418
## 32 RuleRelativisticExpressions                                 0  0.0418
## 33 RuleConfirmationExpressions                                 0  0.0418
## 34 RuleRedundantExpressions                                    0  0.0418
## 35 RuleTooLongExpressions                                      0  0.0418
## 36 RulePredSubjDistance                                        0  0.0418
## 37 RulePredSubjDistance.max_distance                           0  0.0418
```

```
## 38 RulePredSubjDistance.max_distance.v                0      0.0418
## 39 RulePredObjDistance                               0      0.0418
## 40 RulePredObjDistance.max_distance                  0      0.0418
## 41 RulePredObjDistance.max_distance.v                0      0.0418
## 42 RuleInfVerbDistance                               0      0.0418
## 43 RuleInfVerbDistance.max_distance                  0      0.0418
## 44 RuleInfVerbDistance.max_distance.v                0      0.0418
## 45 RuleMultiPartVerbs                                0      0.0418
## 46 RuleMultiPartVerbs.max_distance                   0      0.0418
## 47 RuleMultiPartVerbs.max_distance.v                 0      0.0418
## 48 RuleLongSentences.max_length                      0      0.0418
## 49 RuleLongSentences.max_length.v                    0      0.0418
## 50 RulePredAtClauseBeginning.max_order               0      0.0418
## 51 RulePredAtClauseBeginning.max_order.v             0      0.0418
## 52 RuleVerbalNouns                                   0      0.0418
## 53 sent_count                                        0      0.0418
## 54 word_count                                        0      0.0418
## 55 syllab_count                                      0      0.0418
## 56 char_count                                        0      0.0418
## 57 cli                                               0      0.0418
## 58 ari                                               0      0.0418
## 59 num_hapax                                         0      0.0418
## 60 ttr                                               0      0.0418
## 61 mattr                                             0      0.0418
## 62 mattr.v                                           0      0.0418
## 63 maentropy.v                                       0      0.0418
## 64 mamr                                              0      0.0418
## 65 verb_dist                                         0      0.0418
## 66 hpoint                                            0      0.0418
## 67 fre                                               0      0.0418
## 68 fkgl                                              0      0.0418
## 69 gf                                                0      0.0418
## 70 RuleAnaphoricReferences                      0.0539      0.0418
## 71 atl                                           0.381      0.0418
## 72 activity                                      0.541      0.0418
## Variable importance:
## # A tibble: 71 x 3
##    Variable                                Importance Sign
##    <chr>                                        <dbl> <chr>
##  1 activity                                     0.541 POS
##  2 atl                                          0.381 POS
##  3 RuleLiteraryStyle                            0.267 NEG
##  4 smog                                         0.182 NEG
##  5 RulePassive                                  0.173 NEG
##  6 maentropy                                    0.162 NEG
##  7 entropy                                     0.0937 NEG
##  8 RuleAnaphoricReferences                     0.0539 POS
##  9 RuleGPcoordovs                                   0 NEG
## 10 RuleGPdeverbaddr                                 0 NEG
## 11 RuleGPpatinstr                                   0 NEG
## 12 RuleGPdeverbsubj                                 0 NEG
## 13 RuleGPadjective                                  0 NEG
## 14 RuleGPpatbenperson                               0 NEG
## 15 RuleGPwordorder                                  0 NEG
```

```
## 16 RuleDoubleAdpos                                          0       NEG
## 17 RuleDoubleAdpos.max_allowable_distance                   0       NEG
## 18 RuleDoubleAdpos.max_allowable_distance.v                 0       NEG
## 19 RuleReflexivePassWithAnimSubj                            0       NEG
## 20 RuleTooFewVerbs.min_verb_frac                            0       NEG
## 21 RuleTooManyNegations.max_negation_frac                   0       NEG
## 22 RuleTooManyNegations.max_negation_frac.v                 0       NEG
## 23 RuleTooManyNegations.max_allowable_negations             0       NEG
## 24 RuleTooManyNegations.max_allowable_negations.v           0       NEG
## 25 RuleTooManyNominalConstructions.max_noun_frac            0       NEG
## 26 RuleTooManyNominalConstructions.max_noun_frac.v          0       NEG
## 27 RuleTooManyNominalConstructions.max_allowable_nouns      0       NEG
## 28 RuleCaseRepetition.max_repetition_count                  0       NEG
## 29 RuleCaseRepetition.max_repetition_count.v                0       NEG
## 30 RuleCaseRepetition.max_repetition_frac                   0       NEG
## 31 RuleCaseRepetition.max_repetition_frac.v                 0       NEG
## 32 RuleWeakMeaningWords                                     0       NEG
## 33 RuleAbstractNouns                                        0       NEG
## 34 RuleRelativisticExpressions                              0       NEG
## 35 RuleConfirmationExpressions                              0       NEG
## 36 RuleRedundantExpressions                                 0       NEG
## 37 RuleTooLongExpressions                                   0       NEG
## 38 RulePredSubjDistance                                     0       NEG
## 39 RulePredSubjDistance.max_distance                        0       NEG
## 40 RulePredSubjDistance.max_distance.v                      0       NEG
## 41 RulePredObjDistance                                      0       NEG
## 42 RulePredObjDistance.max_distance                         0       NEG
## 43 RulePredObjDistance.max_distance.v                       0       NEG
## 44 RuleInfVerbDistance                                      0       NEG
## 45 RuleInfVerbDistance.max_distance                         0       NEG
## 46 RuleInfVerbDistance.max_distance.v                       0       NEG
## 47 RuleMultiPartVerbs                                       0       NEG
## 48 RuleMultiPartVerbs.max_distance                          0       NEG
## 49 RuleMultiPartVerbs.max_distance.v                        0       NEG
## 50 RuleLongSentences.max_length                             0       NEG
## 51 RuleLongSentences.max_length.v                           0       NEG
## 52 RulePredAtClauseBeginning.max_order                      0       NEG
## 53 RulePredAtClauseBeginning.max_order.v                    0       NEG
## 54 RuleVerbalNouns                                          0       NEG
## 55 sent_count                                               0       NEG
## 56 word_count                                               0       NEG
## 57 syllab_count                                             0       NEG
## 58 char_count                                               0       NEG
## 59 cli                                                      0       NEG
## 60 ari                                                      0       NEG
## 61 num_hapax                                                0       NEG
## 62 ttr                                                      0       NEG
## 63 mattr                                                    0       NEG
## 64 mattr.v                                                  0       NEG
## 65 maentropy.v                                              0       NEG
## 66 mamr                                                     0       NEG
## 67 verb_dist                                                0       NEG
## 68 hpoint                                                   0       NEG
## 69 fre                                                      0       NEG
```

```
## 70 fkgl                                                    0        NEG
## 71 gf                                                      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:
```
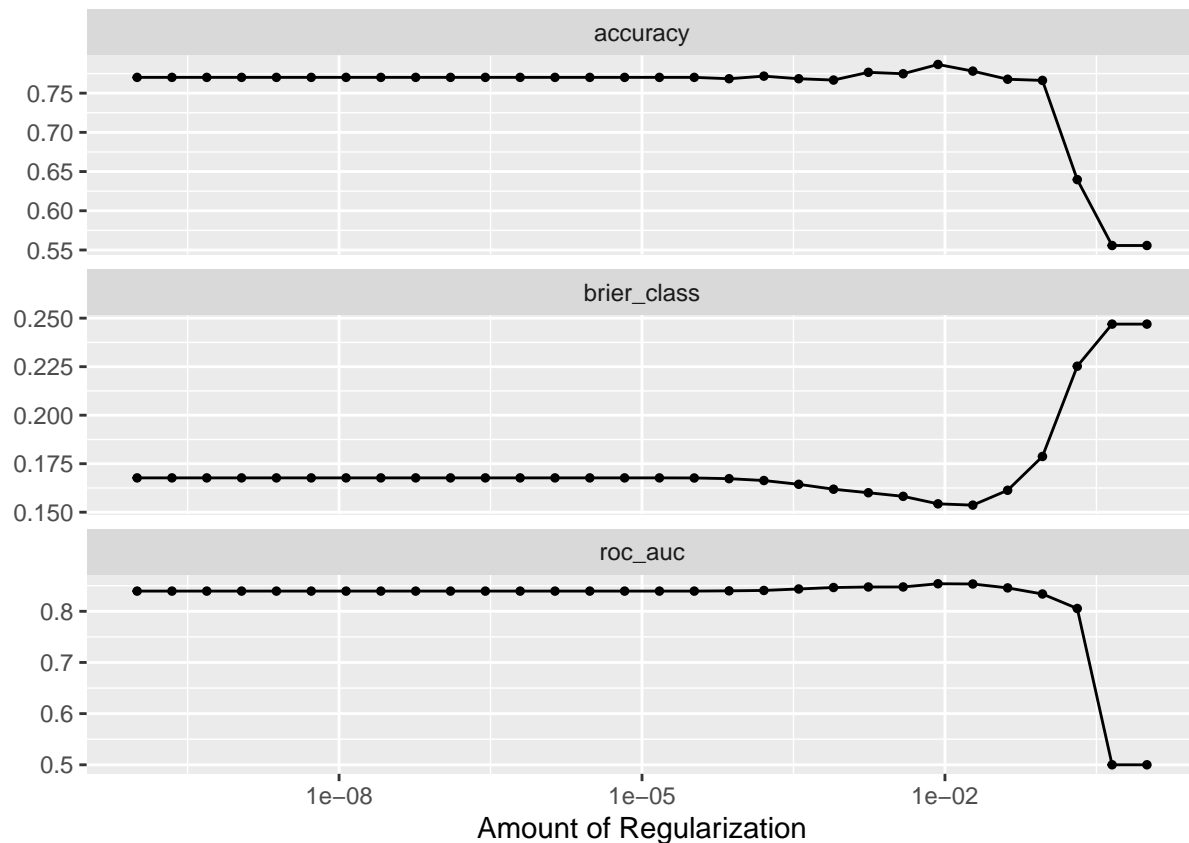
```
## # A tibble: 5 x 7
##    penalty .metric .estimator  mean     n std_err .config
##      <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00853  roc_auc binary     0.854    10 0.00856 Preprocessor1_Model24
## 2 0.0189   roc_auc binary     0.853    10 0.00804 Preprocessor1_Model25
## 3 0.00386  roc_auc binary     0.848    10 0.0110  Preprocessor1_Model23
## 4 0.00174  roc_auc binary     0.848    10 0.0125  Preprocessor1_Model22
## 5 0.000788 roc_auc binary     0.846    10 0.0139  Preprocessor1_Model21
## # A tibble: 5 x 7
##    penalty .metric  .estimator  mean     n std_err .config
##      <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00853  accuracy binary     0.787    10  0.0163 Preprocessor1_Model24
## 2 0.0189   accuracy binary     0.778    10  0.0145 Preprocessor1_Model25
## 3 0.00174  accuracy binary     0.777    10  0.0153 Preprocessor1_Model22
## 4 0.00386  accuracy binary     0.775    10  0.0171 Preprocessor1_Model23
## 5 0.000161 accuracy binary     0.772    10  0.0182 Preprocessor1_Model19
## Best ROC AUC:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0418 Preprocessor1_Model26
## Final workflow:
## == Workflow ============================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor --------------------------------------------------------
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model ---------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.0417531893656041
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 68 x 3
##    term                                        estimate penalty
##    <chr>                                          <dbl>   <dbl>
##  1 (Intercept)                                  -0.294   0.0418
##  2 RuleLiteraryStyle                            -0.267   0.0418
##  3 smog                                         -0.182   0.0418
##  4 RulePassive                                  -0.173   0.0418
##  5 maentropy                                    -0.162   0.0418
##  6 entropy                                      -0.0937  0.0418
##  7 RuleGPcoordovs                                0       0.0418
##  8 RuleGPdeverbaddr                              0       0.0418
##  9 RuleGPpatinstr                                0       0.0418
## 10 RuleGPdeverbsubj                              0       0.0418
```

```
## 11 RuleGPadjective                                      0          0.0418
## 12 RuleGPpatbenperson                                   0          0.0418
## 13 RuleGPwordorder                                      0          0.0418
## 14 RuleDoubleAdpos                                      0          0.0418
## 15 RuleDoubleAdpos.max_allowable_distance               0          0.0418
## 16 RuleDoubleAdpos.max_allowable_distance.v             0          0.0418
## 17 RuleReflexivePassWithAnimSubj                        0          0.0418
## 18 RuleTooFewVerbs.min_verb_frac                        0          0.0418
## 19 RuleTooManyNegations.max_negation_frac               0          0.0418
## 20 RuleTooManyNegations.max_negation_frac.v             0          0.0418
## 21 RuleTooManyNegations.max_allowable_negations         0          0.0418
## 22 RuleTooManyNegations.max_allowable_negations.v       0          0.0418
## 23 RuleTooManyNominalConstructions.max_noun_frac        0          0.0418
## 24 RuleTooManyNominalConstructions.max_noun_frac.v      0          0.0418
## 25 RuleTooManyNominalConstructions.max_allowable_nouns  0          0.0418
## 26 RuleCaseRepetition.max_repetition_count              0          0.0418
## 27 RuleCaseRepetition.max_repetition_count.v            0          0.0418
## 28 RuleCaseRepetition.max_repetition_frac               0          0.0418
## 29 RuleCaseRepetition.max_repetition_frac.v             0          0.0418
## 30 RuleWeakMeaningWords                                 0          0.0418
## 31 RuleAbstractNouns                                    0          0.0418
## 32 RuleRelativisticExpressions                          0          0.0418
## 33 RuleConfirmationExpressions                          0          0.0418
## 34 RuleRedundantExpressions                             0          0.0418
## 35 RuleTooLongExpressions                               0          0.0418
## 36 RulePredSubjDistance                                 0          0.0418
## 37 RulePredSubjDistance.max_distance                    0          0.0418
## 38 RulePredSubjDistance.max_distance.v                  0          0.0418
## 39 RulePredObjDistance                                  0          0.0418
## 40 RulePredObjDistance.max_distance                     0          0.0418
## 41 RulePredObjDistance.max_distance.v                   0          0.0418
## 42 RuleInfVerbDistance                                  0          0.0418
## 43 RuleInfVerbDistance.max_distance                     0          0.0418
## 44 RuleInfVerbDistance.max_distance.v                   0          0.0418
## 45 RuleMultiPartVerbs                                   0          0.0418
## 46 RuleMultiPartVerbs.max_distance                      0          0.0418
## 47 RuleMultiPartVerbs.max_distance.v                    0          0.0418
## 48 RuleLongSentences.max_length                         0          0.0418
## 49 RuleLongSentences.max_length.v                       0          0.0418
## 50 RulePredAtClauseBeginning.max_order                  0          0.0418
## 51 RulePredAtClauseBeginning.max_order.v                0          0.0418
## 52 RuleVerbalNouns                                      0          0.0418
## 53 cli                                                  0          0.0418
## 54 ari                                                  0          0.0418
## 55 num_hapax                                            0          0.0418
## 56 ttr                                                  0          0.0418
## 57 mattr                                                0          0.0418
## 58 mattr.v                                              0          0.0418
## 59 maentropy.v                                          0          0.0418
## 60 mamr                                                 0          0.0418
## 61 verb_dist                                            0          0.0418
## 62 hpoint                                               0          0.0418
## 63 fre                                                  0          0.0418
## 64 fkgl                                                 0          0.0418
```

```
## 65 gf                                                       0        0.0418
## 66 RuleAnaphoricReferences                                  0.0539   0.0418
## 67 atl                                                      0.381    0.0418
## 68 activity                                                 0.541    0.0418
## Variable importance:
## # A tibble: 67 x 3
##    Variable                                          Importance Sign
##    <chr>                                                  <dbl> <chr>
##  1 activity                                               0.541 POS
##  2 atl                                                    0.381 POS
##  3 RuleLiteraryStyle                                      0.267 NEG
##  4 smog                                                   0.182 NEG
##  5 RulePassive                                            0.173 NEG
##  6 maentropy                                              0.162 NEG
##  7 entropy                                                0.0937 NEG
##  8 RuleAnaphoricReferences                                0.0539 POS
##  9 RuleGPcoordovs                                          0     NEG
## 10 RuleGPdeverbaddr                                        0     NEG
## 11 RuleGPpatinstr                                         0     NEG
## 12 RuleGPdeverbsubj                                       0     NEG
## 13 RuleGPadjective                                        0     NEG
## 14 RuleGPpatbenperson                                     0     NEG
## 15 RuleGPwordorder                                        0     NEG
## 16 RuleDoubleAdpos                                        0     NEG
## 17 RuleDoubleAdpos.max_allowable_distance                0     NEG
## 18 RuleDoubleAdpos.max_allowable_distance.v              0     NEG
## 19 RuleReflexivePassWithAnimSubj                         0     NEG
## 20 RuleTooFewVerbs.min_verb_frac                         0     NEG
## 21 RuleTooManyNegations.max_negation_frac                0     NEG
## 22 RuleTooManyNegations.max_negation_frac.v              0     NEG
## 23 RuleTooManyNegations.max_allowable_negations          0     NEG
## 24 RuleTooManyNegations.max_allowable_negations.v        0     NEG
## 25 RuleTooManyNominalConstructions.max_noun_frac         0     NEG
## 26 RuleTooManyNominalConstructions.max_noun_frac.v       0     NEG
## 27 RuleTooManyNominalConstructions.max_allowable_nouns   0     NEG
## 28 RuleCaseRepetition.max_repetition_count               0     NEG
## 29 RuleCaseRepetition.max_repetition_count.v             0     NEG
## 30 RuleCaseRepetition.max_repetition_frac                0     NEG
## 31 RuleCaseRepetition.max_repetition_frac.v              0     NEG
## 32 RuleWeakMeaningWords                                  0     NEG
## 33 RuleAbstractNouns                                     0     NEG
## 34 RuleRelativisticExpressions                           0     NEG
## 35 RuleConfirmationExpressions                           0     NEG
## 36 RuleRedundantExpressions                              0     NEG
## 37 RuleTooLongExpressions                                0     NEG
## 38 RulePredSubjDistance                                  0     NEG
## 39 RulePredSubjDistance.max_distance                     0     NEG
## 40 RulePredSubjDistance.max_distance.v                   0     NEG
## 41 RulePredObjDistance                                   0     NEG
## 42 RulePredObjDistance.max_distance                      0     NEG
## 43 RulePredObjDistance.max_distance.v                    0     NEG
## 44 RuleInfVerbDistance                                   0     NEG
## 45 RuleInfVerbDistance.max_distance                      0     NEG
## 46 RuleInfVerbDistance.max_distance.v                    0     NEG
```

```
## 47 RuleMultiPartVerbs                                         0       NEG
## 48 RuleMultiPartVerbs.max_distance                            0       NEG
## 49 RuleMultiPartVerbs.max_distance.v                          0       NEG
## 50 RuleLongSentences.max_length                               0       NEG
## 51 RuleLongSentences.max_length.v                             0       NEG
## 52 RulePredAtClauseBeginning.max_order                        0       NEG
## 53 RulePredAtClauseBeginning.max_order.v                      0       NEG
## 54 RuleVerbalNouns                                            0       NEG
## 55 cli                                                        0       NEG
## 56 ari                                                        0       NEG
## 57 num_hapax                                                  0       NEG
## 58 ttr                                                        0       NEG
## 59 mattr                                                      0       NEG
## 60 mattr.v                                                    0       NEG
## 61 maentropy.v                                                0       NEG
## 62 mamr                                                       0       NEG
## 63 verb_dist                                                  0       NEG
## 64 hpoint                                                     0       NEG
## 65 fre                                                        0       NEG
## 66 fkgl                                                       0       NEG
## 67 gf                                                         0       NEG
```

Indicators, averages, and coefficients

Remove correlating

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

Keep correlating

```r
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:
```
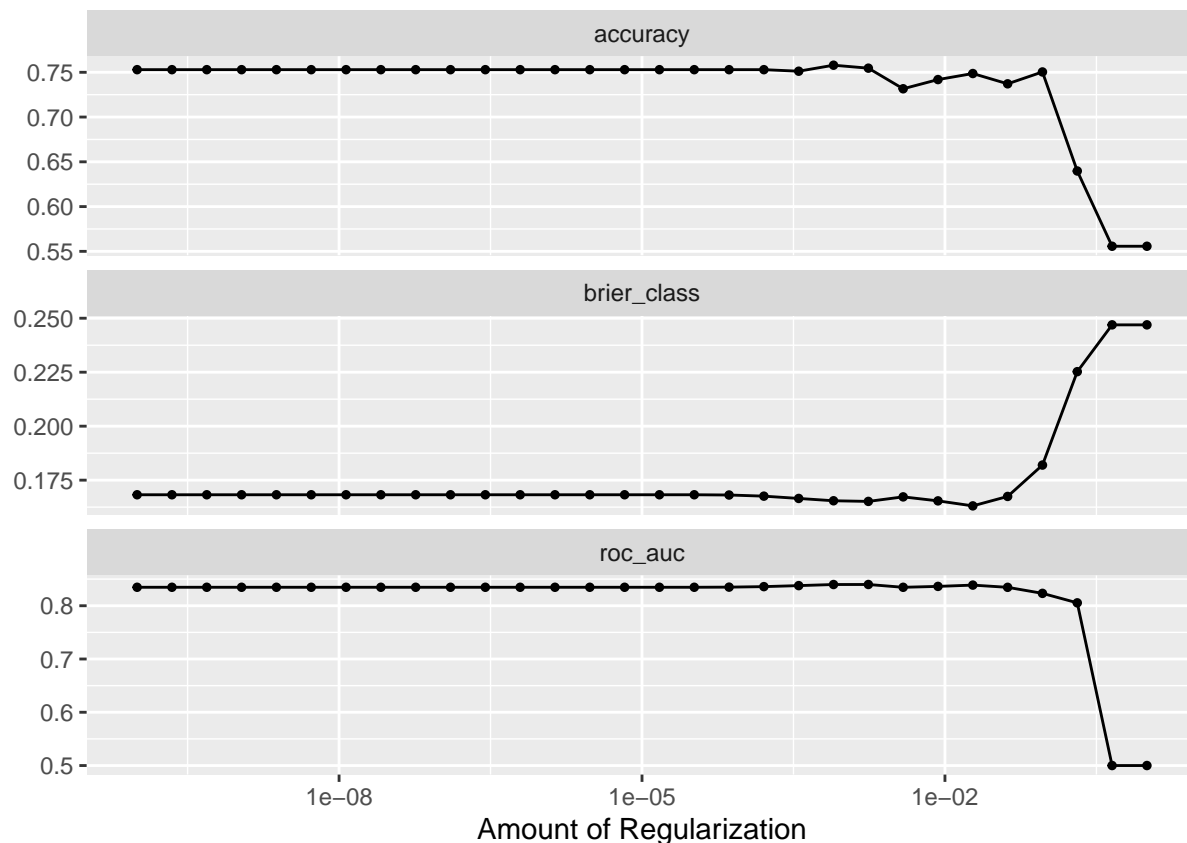
```
## # A tibble: 5 x 7
##   penalty .metric .estimator  mean     n std_err .config
##     <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00174  roc_auc binary     0.840    10 0.0121  Preprocessor1_Model22
## 2 0.000788 roc_auc binary     0.840    10 0.0141  Preprocessor1_Model21
## 3 0.0189   roc_auc binary     0.839    10 0.00890 Preprocessor1_Model25
## 4 0.000356 roc_auc binary     0.838    10 0.0153  Preprocessor1_Model20
## 5 0.00853  roc_auc binary     0.836    10 0.00864 Preprocessor1_Model24
## # A tibble: 5 x 7
##   penalty .metric .estimator  mean     n std_err .config
##     <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 7.88e- 4 accuracy binary    0.758    10  0.0125 Preprocessor1_Model21
## 2 1.74e- 3 accuracy binary    0.755    10  0.0138 Preprocessor1_Model22
## 3 1   e-10 accuracy binary    0.753    10  0.0117 Preprocessor1_Model01
## 4 2.21e-10 accuracy binary    0.753    10  0.0117 Preprocessor1_Model02
## 5 4.89e-10 accuracy binary    0.753    10  0.0117 Preprocessor1_Model03
## Best ROC AUC:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0418 Preprocessor1_Model26
## Final workflow:
## == Workflow ========================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor ----------------------------------------------------
```

```
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model --------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.0417531893656041
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 45 x 3
##    term                                                estimate penalty
##    <chr>                                                  <dbl>   <dbl>
##  1 maentropy                                              -1.41   0.0418
##  2 RuleTooManyNegations.max_allowable_negations.v        -0.345   0.0418
##  3 entropy                                               -0.258   0.0418
##  4 smog                                                  -0.111   0.0418
##  5 RuleTooManyNominalConstructions.max_allowable_nouns.v -0.0104  0.0418
##  6 gf                                                 -0.000136   0.0418
##  7 RuleDoubleAdpos.max_allowable_distance                     0   0.0418
##  8 RuleDoubleAdpos.max_allowable_distance.v                   0   0.0418
##  9 RuleTooFewVerbs.min_verb_frac                              0   0.0418
## 10 RuleTooManyNegations.max_negation_frac                     0   0.0418
## 11 RuleTooManyNegations.max_negation_frac.v                   0   0.0418
## 12 RuleTooManyNegations.max_allowable_negations               0   0.0418
## 13 RuleTooManyNominalConstructions.max_noun_frac              0   0.0418
## 14 RuleTooManyNominalConstructions.max_noun_frac.v            0   0.0418
## 15 RuleTooManyNominalConstructions.max_allowable_nouns        0   0.0418
## 16 RuleCaseRepetition.max_repetition_count                    0   0.0418
## 17 RuleCaseRepetition.max_repetition_count.v                  0   0.0418
## 18 RuleCaseRepetition.max_repetition_frac                     0   0.0418
## 19 RuleCaseRepetition.max_repetition_frac.v                   0   0.0418
## 20 RulePredSubjDistance.max_distance                          0   0.0418
## 21 RulePredSubjDistance.max_distance.v                        0   0.0418
## 22 RulePredObjDistance.max_distance                           0   0.0418
## 23 RulePredObjDistance.max_distance.v                         0   0.0418
## 24 RuleInfVerbDistance.max_distance                           0   0.0418
## 25 RuleInfVerbDistance.max_distance.v                         0   0.0418
## 26 RuleMultiPartVerbs.max_distance                            0   0.0418
## 27 RuleMultiPartVerbs.max_distance.v                          0   0.0418
## 28 RuleLongSentences.max_length                               0   0.0418
## 29 RuleLongSentences.max_length.v                             0   0.0418
## 30 RulePredAtClauseBeginning.max_order                        0   0.0418
## 31 RulePredAtClauseBeginning.max_order.v                      0   0.0418
## 32 cli                                                        0   0.0418
## 33 ari                                                        0   0.0418
## 34 ttr                                                        0   0.0418
## 35 mattr                                                      0   0.0418
## 36 mattr.v                                                    0   0.0418
## 37 maentropy.v                                                0   0.0418
```

```
## 38 mamr                                                        0         0.0418
## 39 verb_dist                                                   0         0.0418
## 40 hpoint                                                      0         0.0418
## 41 fre                                                         0         0.0418
## 42 fkgl                                                        0         0.0418
## 43 atl                                                         1.03      0.0418
## 44 (Intercept)                                                 4.33      0.0418
## 45 activity                                                    5.25      0.0418
## Variable importance:
## # A tibble: 44 x 3
##    Variable                                          Importance Sign
##    <chr>                                                  <dbl> <chr>
##  1 activity                                            5.25     POS
##  2 maentropy                                           1.41     NEG
##  3 atl                                                 1.03     POS
##  4 RuleTooManyNegations.max_allowable_negations.v      0.345    NEG
##  5 entropy                                             0.258    NEG
##  6 smog                                                0.111    NEG
##  7 RuleTooManyNominalConstructions.max_allowable_nouns.v 0.0104 NEG
##  8 gf                                                  0.000136 NEG
##  9 RuleDoubleAdpos.max_allowable_distance              0        NEG
## 10 RuleDoubleAdpos.max_allowable_distance.v            0        NEG
## 11 RuleTooFewVerbs.min_verb_frac                       0        NEG
## 12 RuleTooManyNegations.max_negation_frac              0        NEG
## 13 RuleTooManyNegations.max_negation_frac.v            0        NEG
## 14 RuleTooManyNegations.max_allowable_negations        0        NEG
## 15 RuleTooManyNominalConstructions.max_noun_frac       0        NEG
## 16 RuleTooManyNominalConstructions.max_noun_frac.v     0        NEG
## 17 RuleTooManyNominalConstructions.max_allowable_nouns 0        NEG
## 18 RuleCaseRepetition.max_repetition_count             0        NEG
## 19 RuleCaseRepetition.max_repetition_count.v           0        NEG
## 20 RuleCaseRepetition.max_repetition_frac              0        NEG
## 21 RuleCaseRepetition.max_repetition_frac.v            0        NEG
## 22 RulePredSubjDistance.max_distance                   0        NEG
## 23 RulePredSubjDistance.max_distance.v                 0        NEG
## 24 RulePredObjDistance.max_distance                    0        NEG
## 25 RulePredObjDistance.max_distance.v                  0        NEG
## 26 RuleInfVerbDistance.max_distance                    0        NEG
## 27 RuleInfVerbDistance.max_distance.v                  0        NEG
## 28 RuleMultiPartVerbs.max_distance                     0        NEG
## 29 RuleMultiPartVerbs.max_distance.v                   0        NEG
## 30 RuleLongSentences.max_length                        0        NEG
## 31 RuleLongSentences.max_length.v                      0        NEG
## 32 RulePredAtClauseBeginning.max_order                 0        NEG
## 33 RulePredAtClauseBeginning.max_order.v               0        NEG
## 34 cli                                                 0        NEG
## 35 ari                                                 0        NEG
## 36 ttr                                                 0        NEG
## 37 mattr                                               0        NEG
## 38 mattr.v                                             0        NEG
## 39 maentropy.v                                         0        NEG
## 40 mamr                                                0        NEG
## 41 verb_dist                                           0        NEG
## 42 hpoint                                              0        NEG
```

```
## 43 fre                                                    0        NEG
## 44 fkgl                                                   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:
```
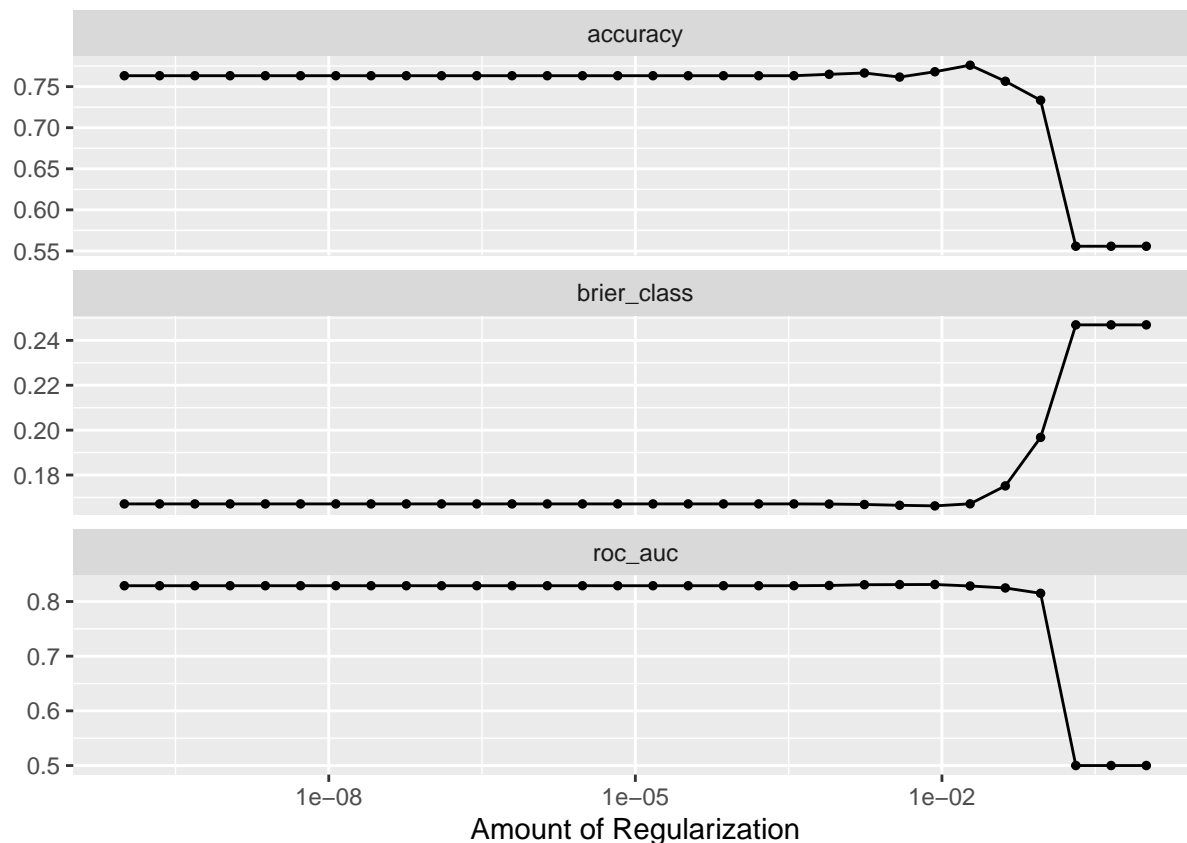
```
## # A tibble: 5 x 7
##        penalty .metric .estimator  mean     n std_err .config
##          <dbl> <chr>   <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.00853       roc_auc binary    0.831    10  0.0137 Preprocessor1_Model24
## 2 0.00386       roc_auc binary    0.831    10  0.0145 Preprocessor1_Model23
## 3 0.00174       roc_auc binary    0.831    10  0.0154 Preprocessor1_Model22
## 4 0.000788      roc_auc binary    0.829    10  0.0156 Preprocessor1_Model21
## 5 0.0000000001  roc_auc binary    0.829    10  0.0157 Preprocessor1_Model01
## # A tibble: 5 x 7
##        penalty .metric  .estimator  mean     n std_err .config
##          <dbl> <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 0.0189        accuracy binary    0.776    10  0.0128 Preprocessor1_Model25
## 2 0.00853       accuracy binary    0.768    10  0.0121 Preprocessor1_Model24
## 3 0.00174       accuracy binary    0.767    10  0.0125 Preprocessor1_Model22
## 4 0.000788      accuracy binary    0.765    10  0.0119 Preprocessor1_Model21
## 5 0.0000000001  accuracy binary    0.763    10  0.0121 Preprocessor1_Model01
## Best ROC AUC:
## # A tibble: 1 x 2
##   penalty .config
##     <dbl> <chr>
## 1  0.0418 Preprocessor1_Model26
## Final workflow:
## == Workflow ======================================================
## Preprocessor: Recipe
## Model: logistic_reg()
##
## -- Preprocessor --------------------------------------------------
```

```
## 1 Recipe Step
##
## * step_normalize()
##
## -- Model ------------------------------------------------------------------------
## Logistic Regression Model Specification (classification)
##
## Main Arguments:
##   penalty = 0.0417531893656041
##   mixture = 1
##
## Computational engine: glmnet
##
## Final coefficients:
## # A tibble: 25 x 3
##    term                        estimate penalty
##    <chr>                          <dbl>   <dbl>
##  1 RuleRelativisticExpressions   -140.   0.0418
##  2 RulePassive                   -122.   0.0418
##  3 RuleLiteraryStyle             -102.   0.0418
##  4 (Intercept)                   -1.30   0.0418
##  5 RuleGPcoordovs                    0   0.0418
##  6 RuleGPdeverbaddr                  0   0.0418
##  7 RuleGPpatinstr                    0   0.0418
##  8 RuleGPdeverbsubj                  0   0.0418
##  9 RuleGPadjective                   0   0.0418
## 10 RuleGPpatbenperson                0   0.0418
## 11 RuleGPwordorder                   0   0.0418
## 12 RuleDoubleAdpos                   0   0.0418
## 13 RuleReflexivePassWithAnimSubj     0   0.0418
## 14 RuleWeakMeaningWords              0   0.0418
## 15 RuleAbstractNouns                 0   0.0418
## 16 RuleConfirmationExpressions       0   0.0418
## 17 RuleRedundantExpressions          0   0.0418
## 18 RuleTooLongExpressions            0   0.0418
## 19 RulePredObjDistance               0   0.0418
## 20 num_hapax                         0   0.0418
## 21 RuleInfVerbDistance            1.02   0.0418
## 22 RuleVerbalNouns                5.24   0.0418
## 23 RulePredSubjDistance           14.6   0.0418
## 24 RuleMultiPartVerbs             24.0   0.0418
## 25 RuleAnaphoricReferences        39.8   0.0418
## Variable importance:
## # A tibble: 24 x 3
##    Variable                    Importance Sign
##    <chr>                            <dbl> <chr>
##  1 RuleRelativisticExpressions      140.  NEG
##  2 RulePassive                      122.  NEG
##  3 RuleLiteraryStyle                102.  NEG
##  4 RuleAnaphoricReferences          39.8  POS
##  5 RuleMultiPartVerbs               24.0  POS
##  6 RulePredSubjDistance             14.6  POS
##  7 RuleVerbalNouns                  5.24  POS
##  8 RuleInfVerbDistance              1.02  POS
```

```
##  9 RuleGPcoordovs                         0    NEG
## 10 RuleGPdeverbaddr                        0    NEG
## 11 RuleGPpatinstr                          0    NEG
## 12 RuleGPdeverbsubj                        0    NEG
## 13 RuleGPadjective                         0    NEG
## 14 RuleGPpatbenperson                      0    NEG
## 15 RuleGPwordorder                         0    NEG
## 16 RuleDoubleAdpos                         0    NEG
## 17 RuleReflexivePassWithAnimSubj           0    NEG
## 18 RuleWeakMeaningWords                    0    NEG
## 19 RuleAbstractNouns                       0    NEG
## 20 RuleConfirmationExpressions             0    NEG
## 21 RuleRedundantExpressions                0    NEG
## 22 RuleTooLongExpressions                  0    NEG
## 23 RulePredObjDistance                     0    NEG
## 24 num_hapax                               0    NEG
```

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     0.766    10 0.0159  Preprocessor1_Model1
## 2 brier_class binary     0.167    10 0.00536 Preprocessor1_Model1
## 3 roc_auc     binary     0.837    10 0.0120  Preprocessor1_Model1
```
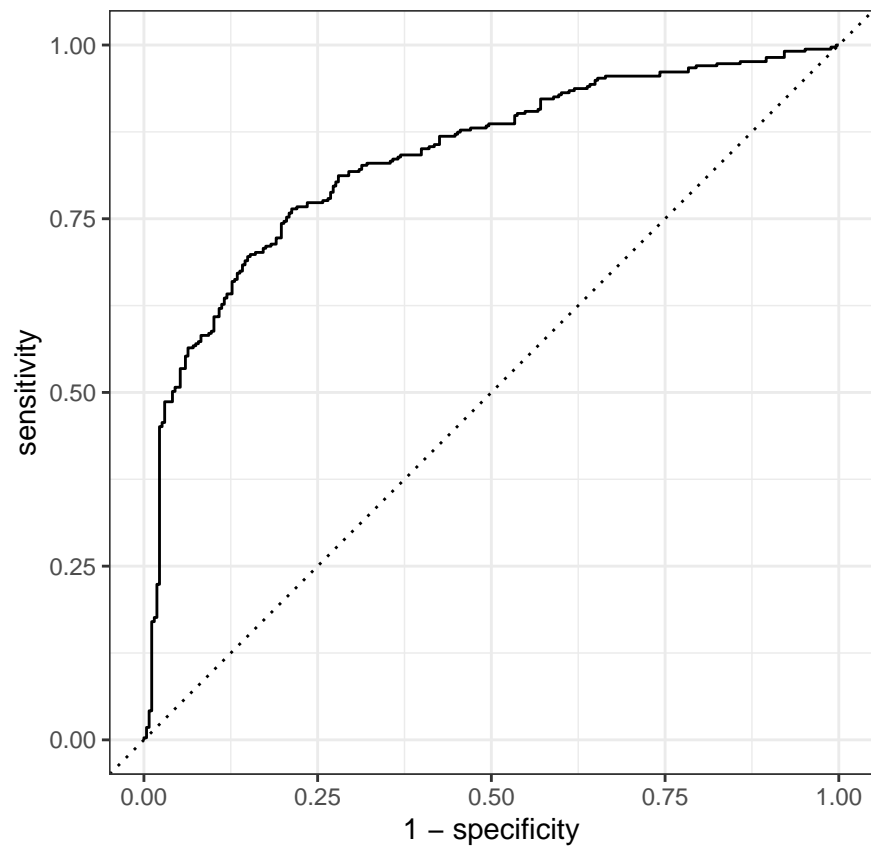
```
## [1] "\n"
```
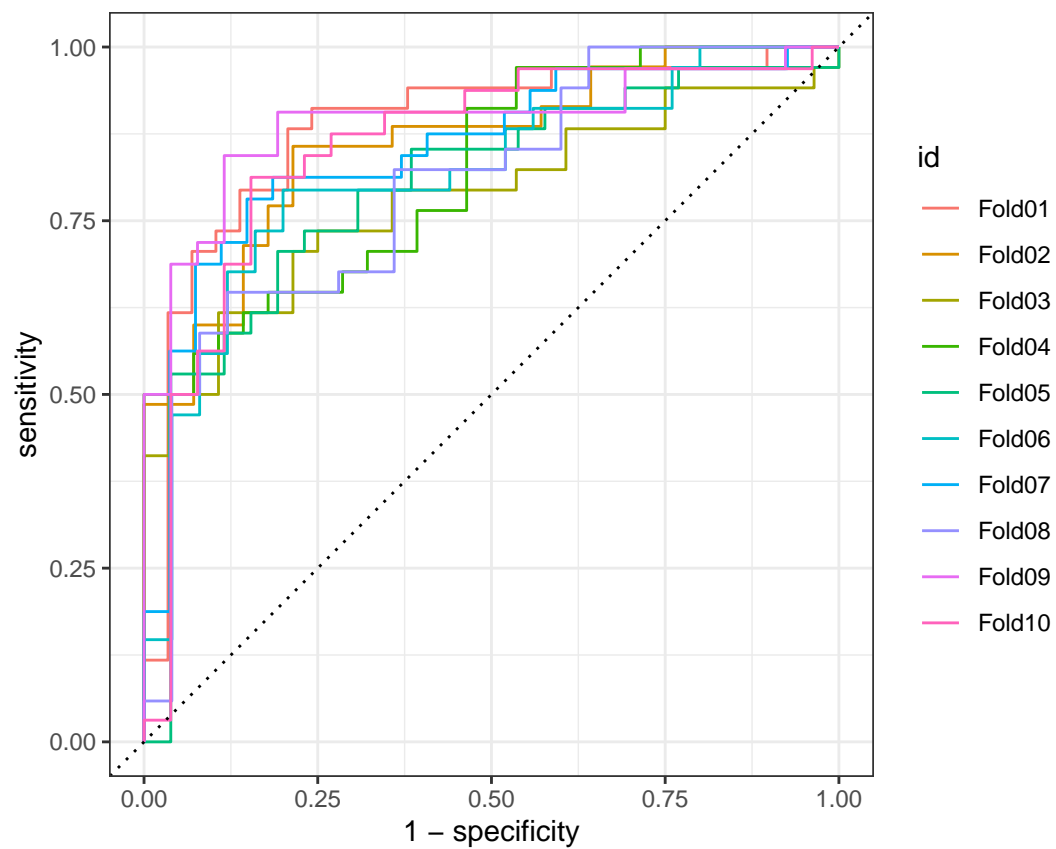
[1] "\n"
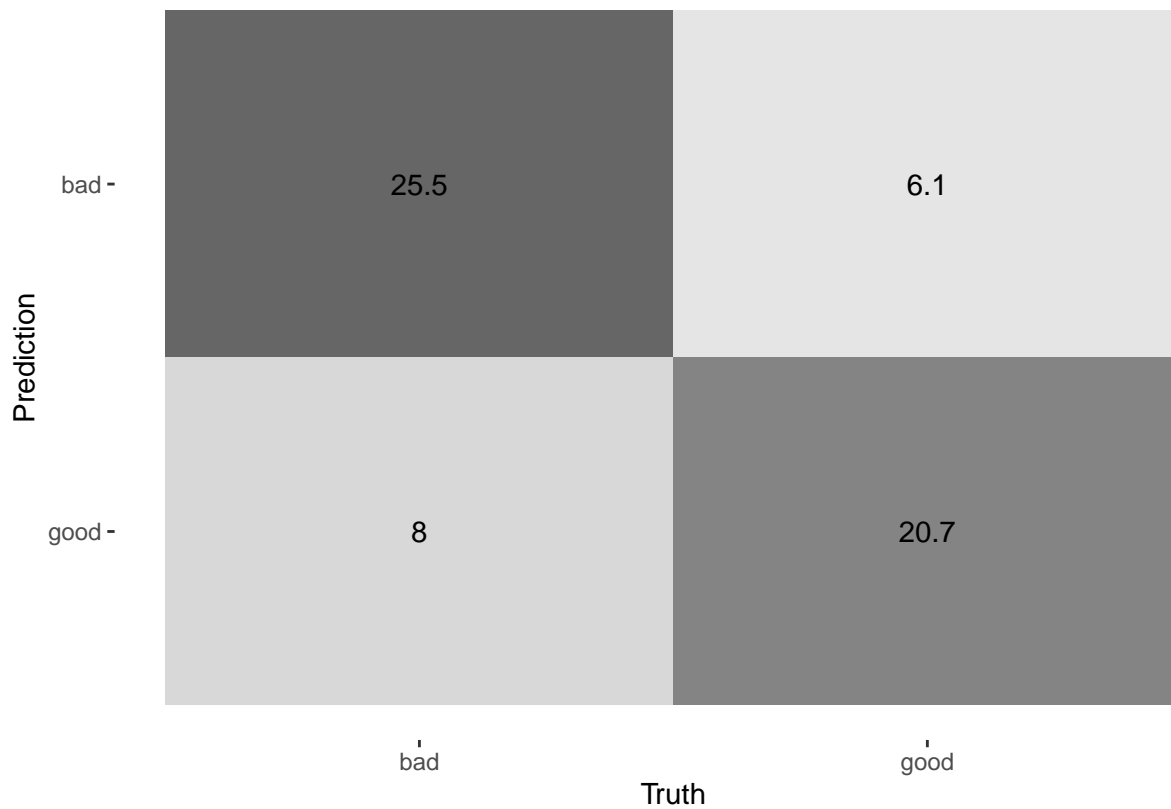
```
## [1] "\n"
```

```
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     0.781    10 0.0139  Preprocessor1_Model1
## 2 brier_class binary     0.145    10 0.00561 Preprocessor1_Model1
## 3 roc_auc     binary     0.869    10 0.00928 Preprocessor1_Model1
```
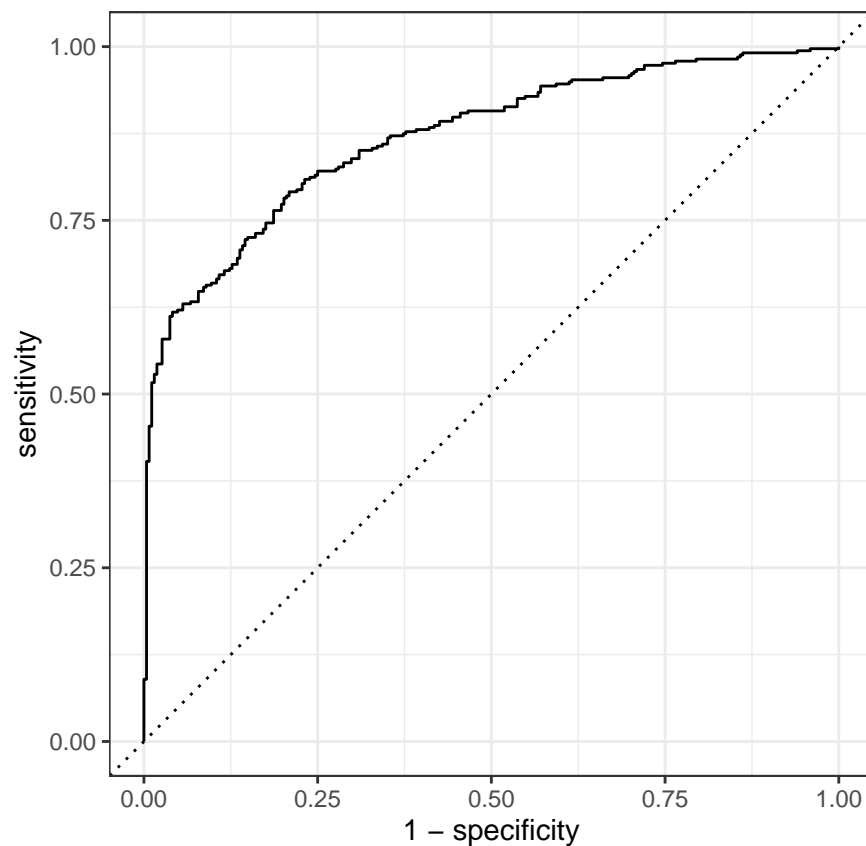


```
## [1] "\n"
```
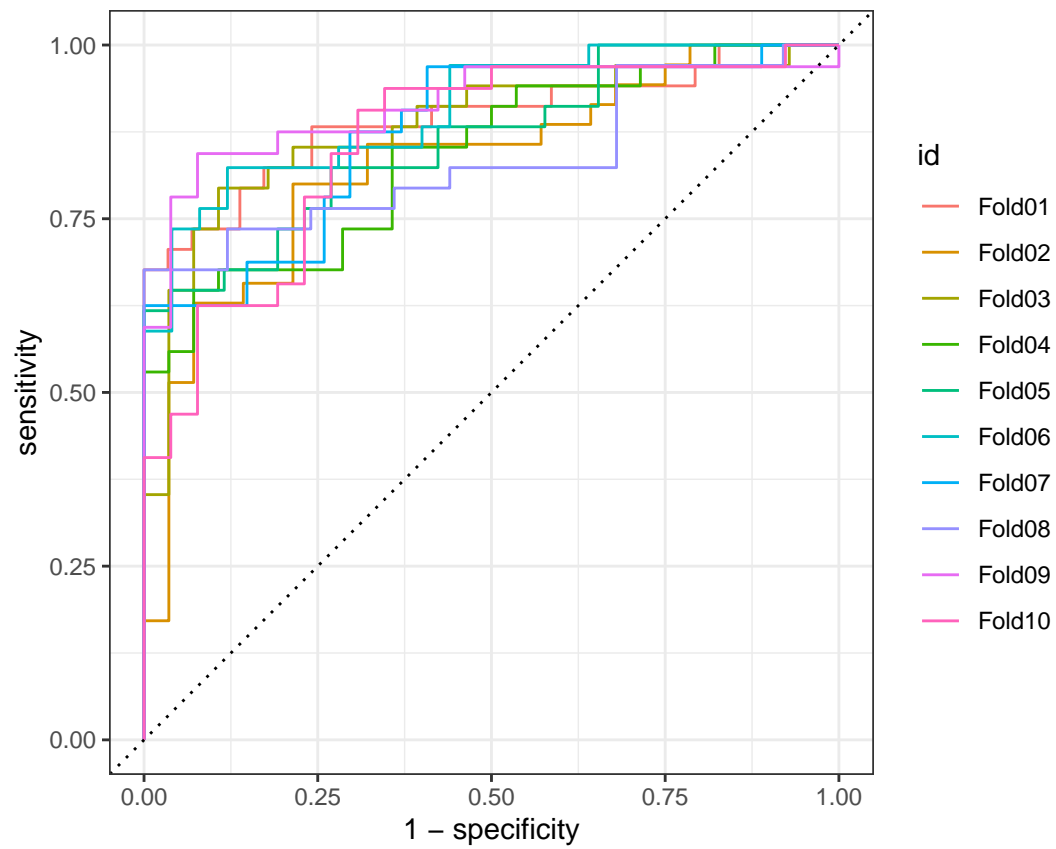
[1] "\n"
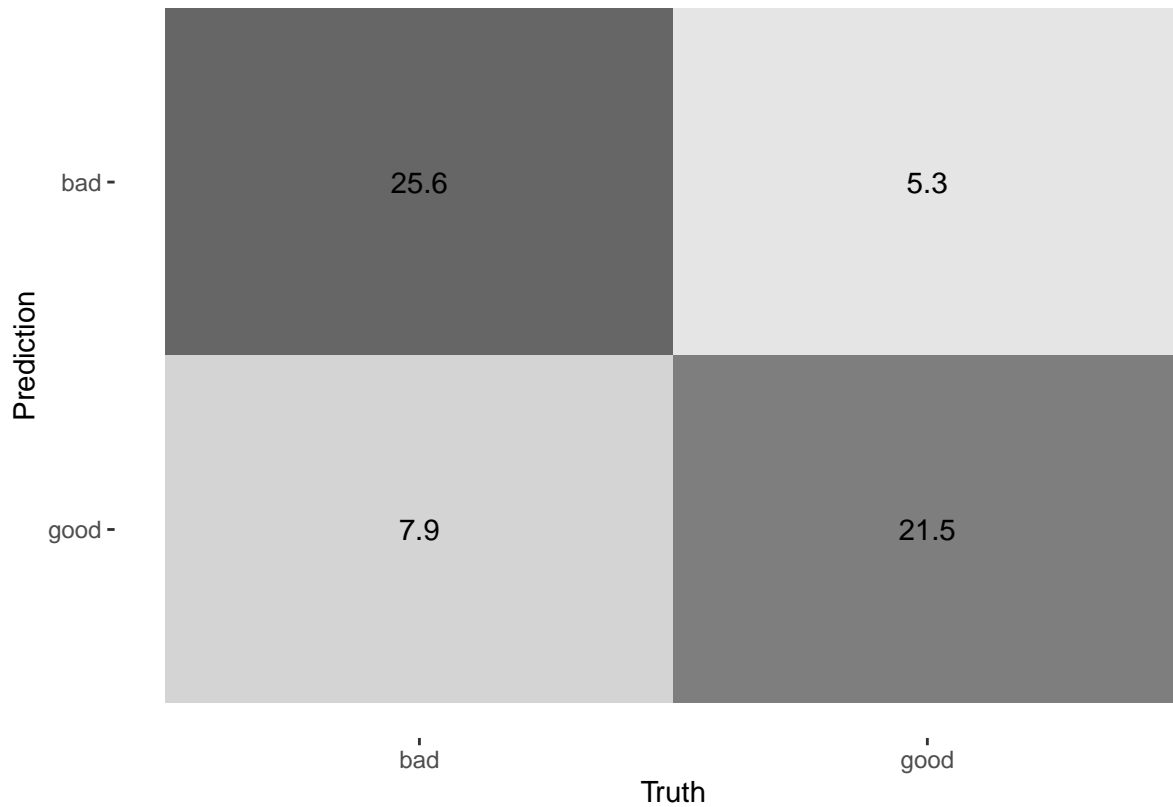
```
## [1] "\n"
```

Random forest

All variables

Remove correlating

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

Keep correlating

```r
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     0.780    10 0.00932 Preprocessor1_Model1
## 2 brier_class binary     0.148    10 0.00422 Preprocessor1_Model1
## 3 roc_auc     binary     0.865    10 0.00889 Preprocessor1_Model1
```
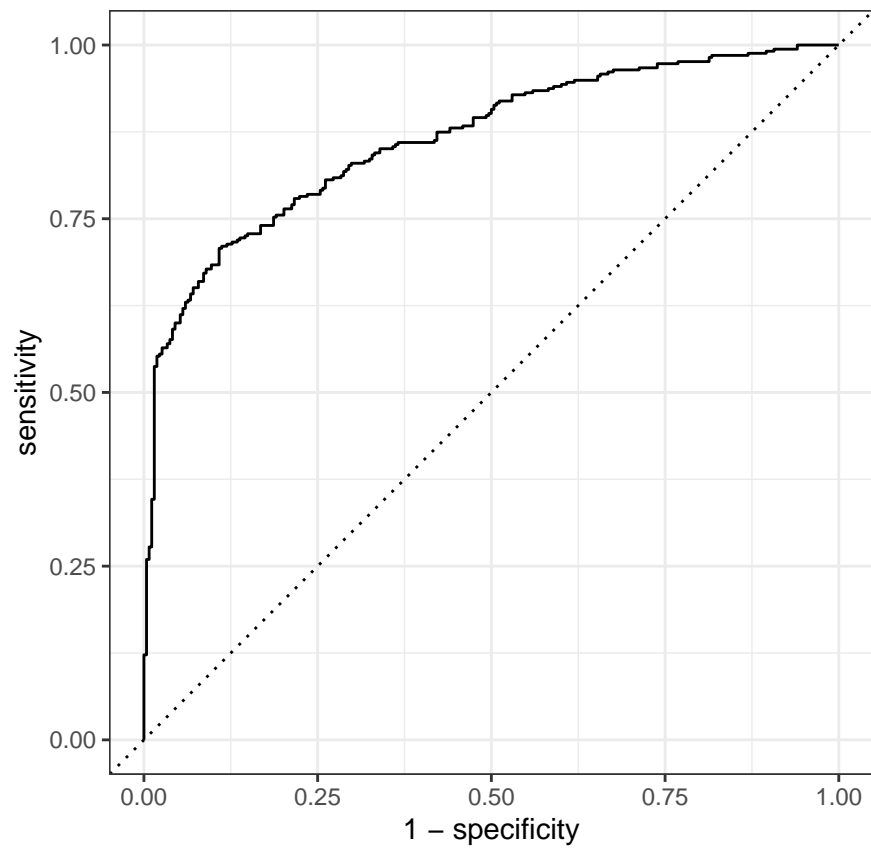
```
## [1] "\n"
```
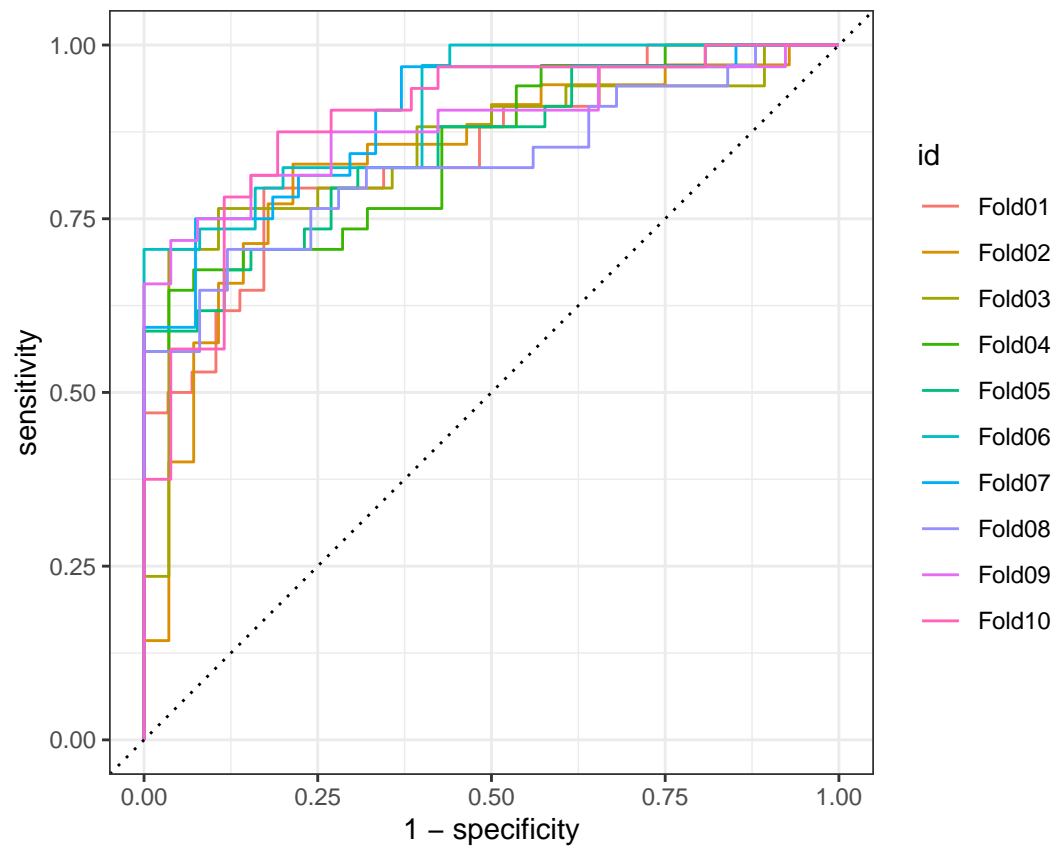
[1] "\n"
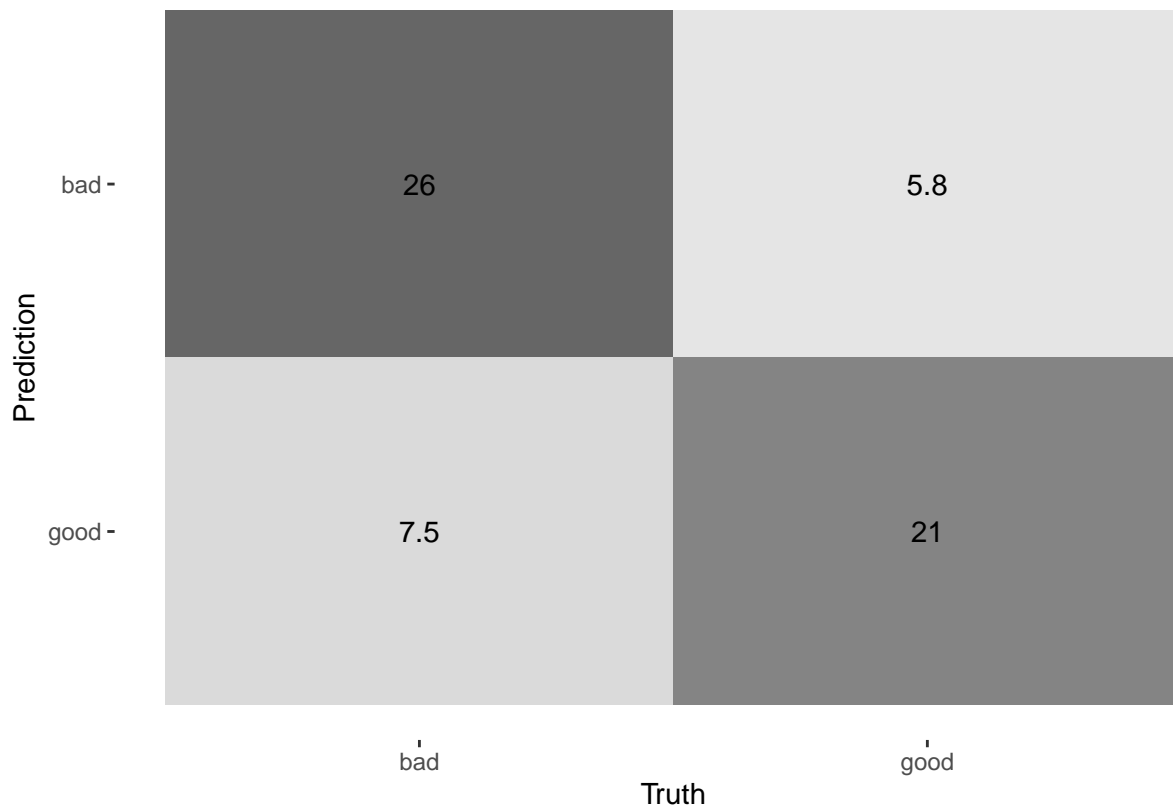
```
## [1] "\n"
## # A tibble: 71 x 2
##    Variable                                        Importance
##    <chr>                                                <dbl>
##  1 activity                                             15.6
##  2 RuleTooFewVerbs.min_verb_frac                        13.9
##  3 verb_dist                                            13.1
##  4 RuleLongSentences.max_length                         10.5
##  5 RuleTooManyNominalConstructions.max_allowable_nouns  10.3
##  6 ari                                                   9.55
##  7 RulePassive                                           9.29
##  8 RulePredAtClauseBeginning.max_order                   8.92
##  9 gf                                                    7.24
## 10 smog                                                  7.01
## 11 atl                                                   6.85
## 12 RuleLiteraryStyle                                     6.42
## 13 fkgl                                                  5.59
## 14 RuleTooManyNegations.max_negation_frac                4.76
## 15 maentropy                                             4.58
## 16 RuleTooLongExpressions                                4.10
## 17 mamr                                                  4.00
## 18 RulePredAtClauseBeginning.max_order.v                 3.90
## 19 RuleTooManyNominalConstructions.max_noun_frac         3.75
## 20 mattr                                                 3.71
## 21 RuleVerbalNouns                                       3.67
## 22 RulePredSubjDistance                                  3.59
## 23 RuleCaseRepetition.max_repetition_count.v             3.58
## 24 RuleMultiPartVerbs                                    3.41
## 25 maentropy.v                                           3.39
## 26 RuleCaseRepetition.max_repetition_frac.v              3.38
## 27 cli                                                   3.30
## 28 RuleLongSentences.max_length.v                        3.27
## 29 RuleCaseRepetition.max_repetition_frac                3.04
## 30 RulePredSubjDistance.max_distance                     2.98
## 31 RuleAnaphoricReferences                               2.96
## 32 mattr.v                                               2.92
## 33 RulePredSubjDistance.max_distance.v                   2.86
## 34 entropy                                               2.70
## 35 RulePredObjDistance.max_distance                      2.64
## 36 RuleDoubleAdpos                                       2.61
## 37 RuleInfVerbDistance                                   2.56
## 38 RuleTooManyNegations.max_allowable_negations.v        2.49
## 39 RuleMultiPartVerbs.max_distance                       2.47
## 40 RuleCaseRepetition.max_repetition_count               2.46
## 41 RulePredObjDistance                                   2.43
## 42 RuleTooManyNominalConstructions.max_noun_frac.v       2.39
## 43 RuleTooManyNegations.max_allowable_negations          2.37
## 44 fre                                                   2.36
## 45 RuleInfVerbDistance.max_distance                      2.35
## 46 RuleTooManyNegations.max_negation_frac.v              2.28
## 47 RuleMultiPartVerbs.max_distance.v                     2.23
## 48 RuleInfVerbDistance.max_distance.v                    2.18
## 49 ttr                                                   2.11
## 50 RuleAbstractNouns                                     2.05
```

```
## 51 word_count                                    2.00
## 52 RuleWeakMeaningWords                           2.00
## 53 RuleDoubleAdpos.max_allowable_distance.v       1.98
## 54 num_hapax                                      1.94
## 55 syllab_count                                   1.92
## 56 sent_count                                     1.91
## 57 char_count                                     1.89
## 58 RuleDoubleAdpos.max_allowable_distance         1.86
## 59 RulePredObjDistance.max_distance.v             1.82
## 60 RuleGPwordorder                                1.58
## 61 hpoint                                         1.45
## 62 RuleGPcoordovs                                 1.38
## 63 RuleReflexivePassWithAnimSubj                  1.36
## 64 RuleGPpatinstr                                 1.09
## 65 RuleGPdeverbaddr                               1.01
## 66 RuleRelativisticExpressions                    0.943
## 67 RuleGPdeverbsubj                               0.806
## 68 RuleGPpatbenperson                             0.724
## 69 RuleConfirmationExpressions                    0.526
## 70 RuleGPadjective                                0.252
## 71 RuleRedundantExpressions                       0.0978
```

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     0.780    10 0.0108  Preprocessor1_Model1
## 2 brier_class binary     0.147    10 0.00427 Preprocessor1_Model1
## 3 roc_auc     binary     0.865    10 0.00841 Preprocessor1_Model1
```

```
## [1] "\n"
```
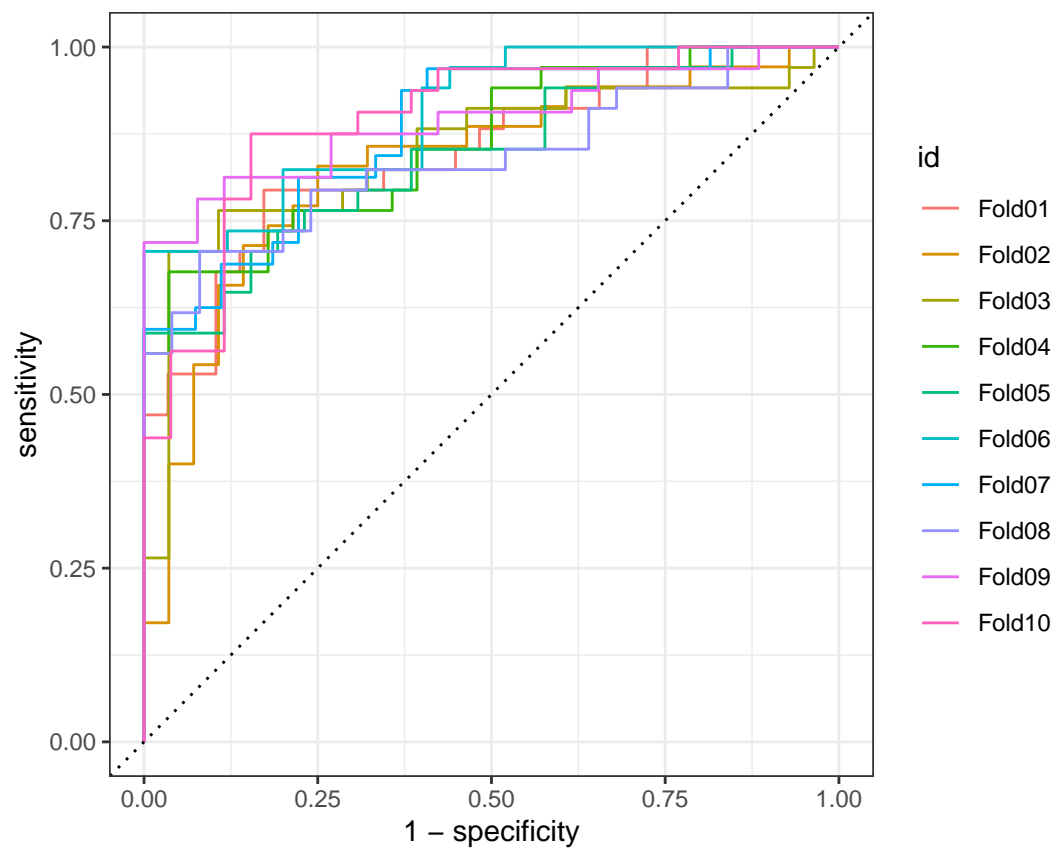
[1] "\n"
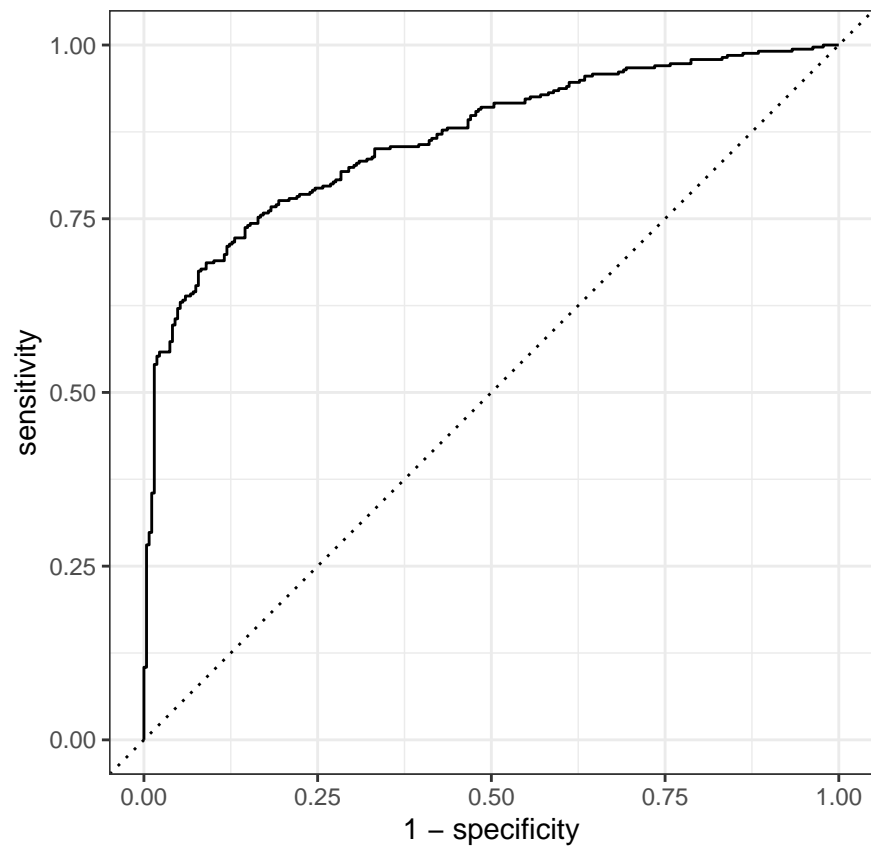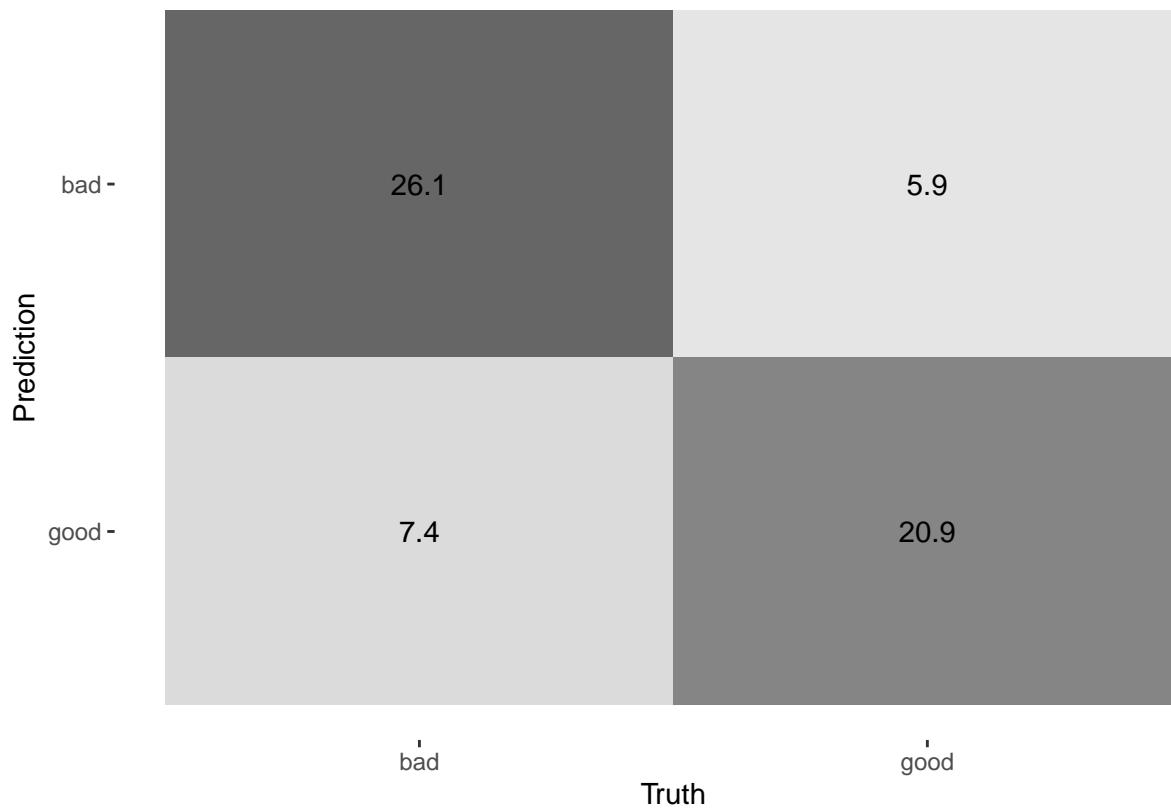
```
## [1] "\n"
## # A tibble: 67 x 2
##    Variable                                            Importance
##    <chr>                                                    <dbl>
##  1 activity                                                  16.8
##  2 verb_dist                                                 15.5
##  3 RuleTooFewVerbs.min_verb_frac                             13.2
##  4 RuleLongSentences.max_length                              11.1
##  5 RuleTooManyNominalConstructions.max_allowable_nouns       11.1
##  6 RulePredAtClauseBeginning.max_order                        9.69
##  7 ari                                                        9.28
##  8 RulePassive                                                9.03
##  9 RuleLiteraryStyle                                          7.09
## 10 gf                                                         7.04
## 11 smog                                                       6.35
## 12 atl                                                        5.72
## 13 fkgl                                                       5.03
## 14 RuleTooManyNegations.max_negation_frac                     4.96
## 15 mamr                                                       4.30
## 16 mattr                                                      4.26
## 17 RuleTooLongExpressions                                     4.25
## 18 RulePredAtClauseBeginning.max_order.v                      4.23
## 19 RuleMultiPartVerbs                                         4.15
## 20 RuleVerbalNouns                                            4.11
## 21 maentropy                                                  3.95
## 22 RuleTooManyNominalConstructions.max_noun_frac              3.86
## 23 RuleCaseRepetition.max_repetition_count.v                  3.57
## 24 maentropy.v                                                3.53
## 25 entropy                                                    3.49
## 26 RulePredSubjDistance                                       3.42
## 27 RulePredSubjDistance.max_distance                          3.35
## 28 RuleLongSentences.max_length.v                             3.27
## 29 RuleCaseRepetition.max_repetition_frac.v                   3.26
## 30 cli                                                        3.26
## 31 RuleAnaphoricReferences                                    3.11
## 32 mattr.v                                                    3.04
## 33 RulePredObjDistance.max_distance                           2.83
## 34 RuleCaseRepetition.max_repetition_frac                     2.77
## 35 RuleTooManyNegations.max_allowable_negations.v             2.64
## 36 RuleCaseRepetition.max_repetition_count                    2.60
## 37 RuleInfVerbDistance                                        2.56
## 38 RuleInfVerbDistance.max_distance                           2.54
## 39 RuleMultiPartVerbs.max_distance                            2.48
## 40 RulePredObjDistance                                        2.47
## 41 ttr                                                        2.45
## 42 RuleTooManyNegations.max_allowable_negations               2.42
## 43 num_hapax                                                  2.42
## 44 RuleTooManyNegations.max_negation_frac.v                   2.41
## 45 RuleTooManyNominalConstructions.max_noun_frac.v            2.37
## 46 RuleInfVerbDistance.max_distance.v                         2.37
## 47 RuleDoubleAdpos                                            2.33
## 48 fre                                                        2.28
## 49 RuleDoubleAdpos.max_allowable_distance.v                   2.22
## 50 RulePredSubjDistance.max_distance.v                        2.22
```

```
## 51 RulePredObjDistance.max_distance.v                    2.16
## 52 RuleMultiPartVerbs.max_distance.v                     2.12
## 53 RuleWeakMeaningWords                                  2.10
## 54 RuleAbstractNouns                                     2.05
## 55 RuleDoubleAdpos.max_allowable_distance                1.84
## 56 hpoint                                                1.73
## 57 RuleGPcoordovs                                        1.61
## 58 RuleGPwordorder                                       1.44
## 59 RuleReflexivePassWithAnimSubj                         1.41
## 60 RuleGPpatinstr                                        1.09
## 61 RuleGPdeverbaddr                                      1.04
## 62 RuleRelativisticExpressions                           0.996
## 63 RuleGPpatbenperson                                    0.845
## 64 RuleGPdeverbsubj                                      0.738
## 65 RuleConfirmationExpressions                           0.612
## 66 RuleGPadjective                                       0.365
## 67 RuleRedundantExpressions                              0.0837
```

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     0.767    10 0.0120  Preprocessor1_Model1
## 2 brier_class binary     0.153    10 0.00403 Preprocessor1_Model1
## 3 roc_auc     binary     0.856    10 0.00838 Preprocessor1_Model1
```
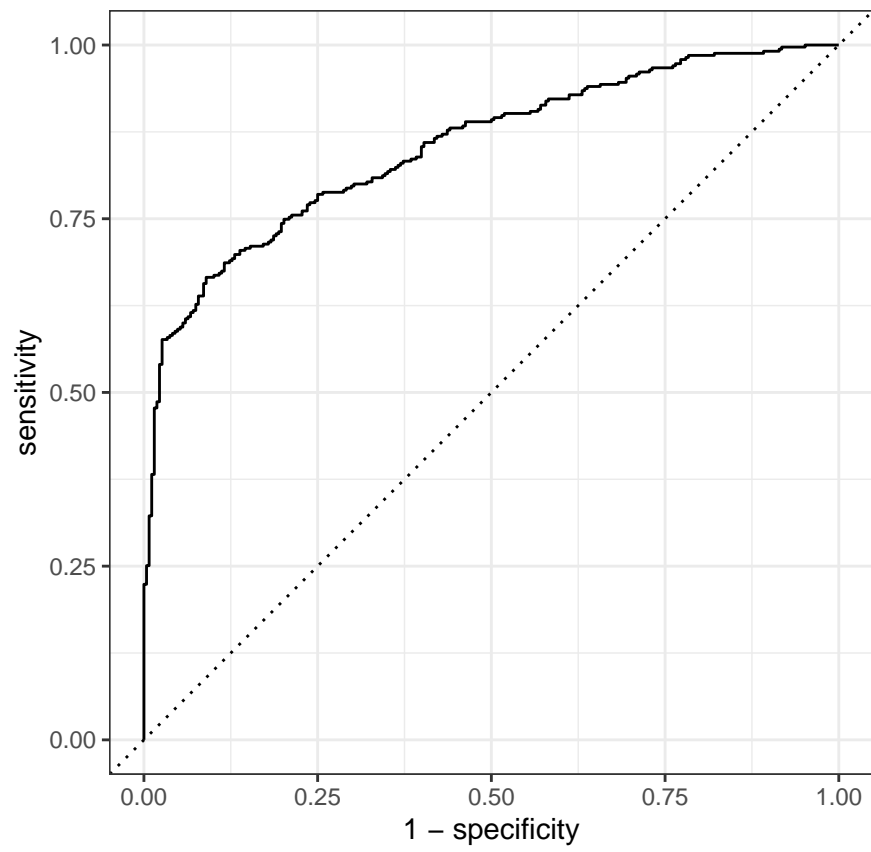
```
## [1] "\n"
```
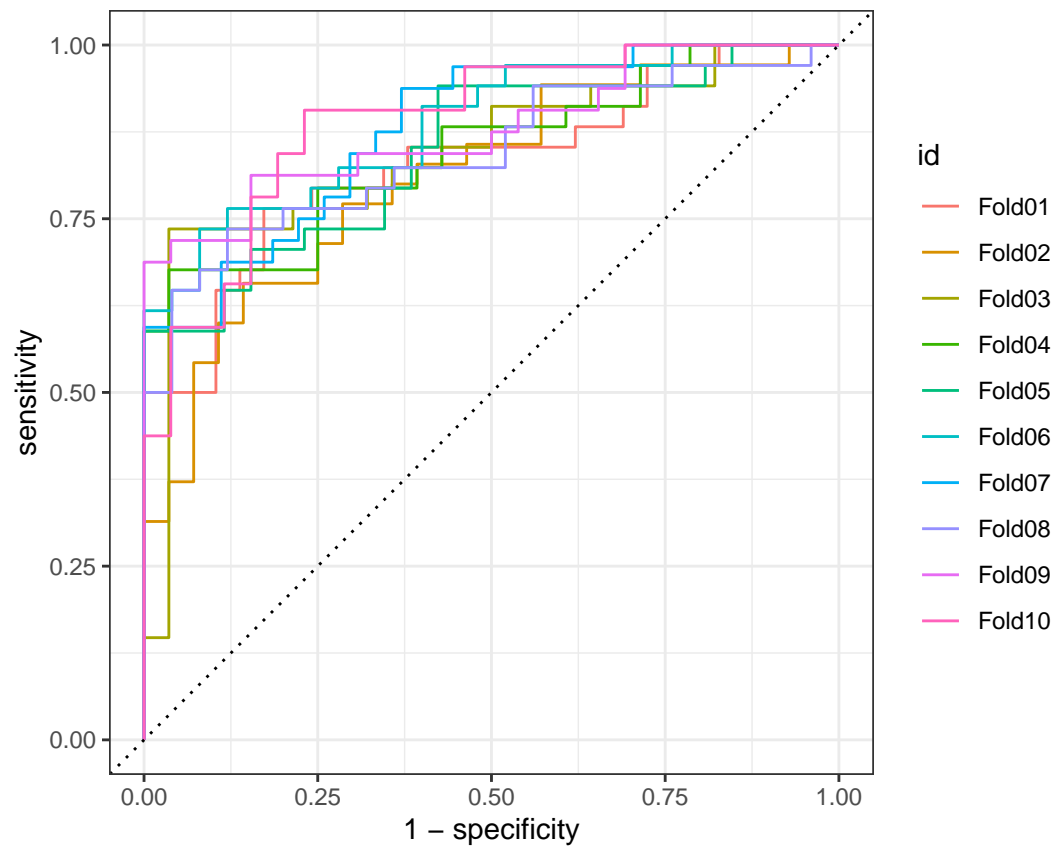
[1] "\n"
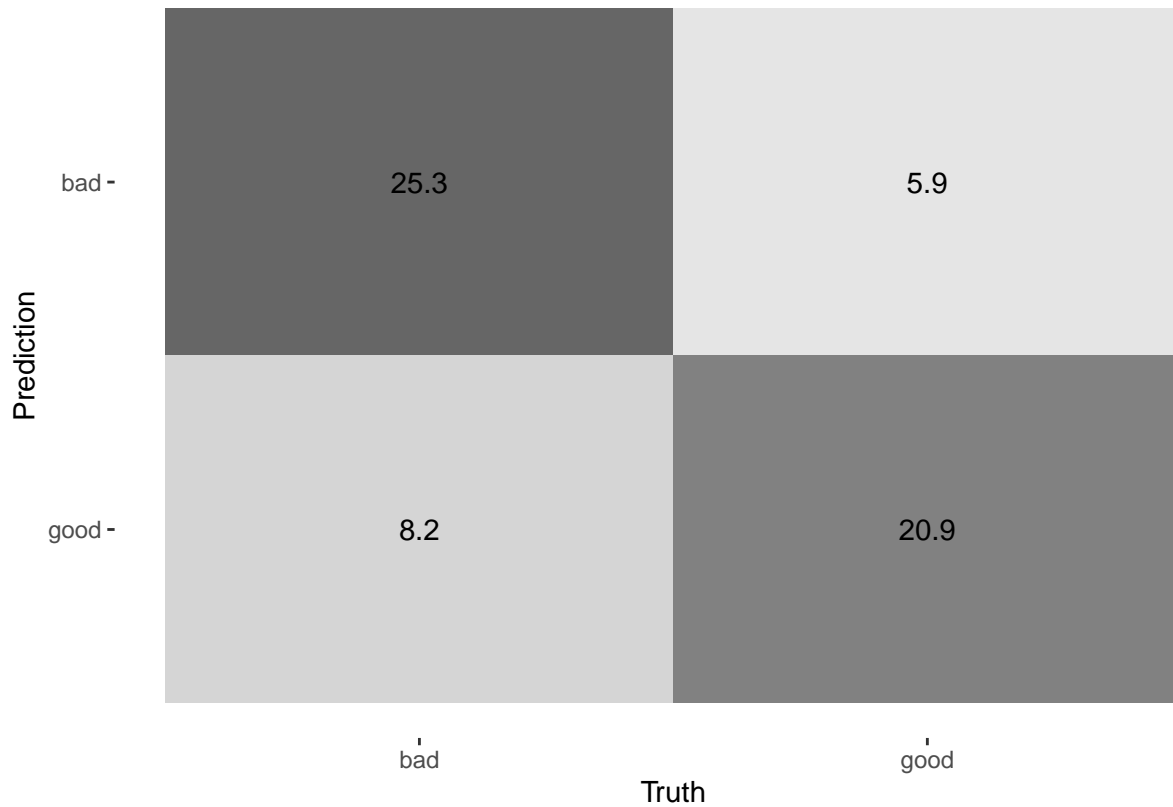
```
## [1] "\n"
## # A tibble: 44 x 2
##    Variable                                              Importance
##    <chr>                                                      <dbl>
##  1 activity                                                    19.0
##  2 verb_dist                                                   16.8
##  3 RuleTooFewVerbs.min_verb_frac                               16.8
##  4 RuleTooManyNominalConstructions.max_allowable_nouns         12.6
##  5 RuleLongSentences.max_length                                11.7
##  6 ari                                                         10.9
##  7 RulePredAtClauseBeginning.max_order                         10.0
##  8 gf                                                          9.03
##  9 atl                                                         7.72
## 10 smog                                                        7.72
## 11 RuleTooManyNegations.max_negation_frac                      6.27
## 12 RuleTooManyNominalConstructions.max_noun_frac               5.77
## 13 maentropy                                                   5.70
## 14 fkgl                                                        5.64
## 15 mattr                                                       5.35
## 16 RuleTooManyNominalConstructions.max_allowable_nouns.v       5.30
## 17 mamr                                                        5.20
## 18 RuleLongSentences.max_length.v                              4.82
## 19 RulePredAtClauseBeginning.max_order.v                       4.72
## 20 cli                                                         4.65
## 21 maentropy.v                                                 4.57
## 22 entropy                                                     4.57
## 23 RuleCaseRepetition.max_repetition_count.v                   4.50
## 24 RulePredSubjDistance.max_distance                           4.31
## 25 RuleCaseRepetition.max_repetition_frac.v                    4.10
## 26 mattr.v                                                     3.96
## 27 RuleCaseRepetition.max_repetition_frac                      3.92
## 28 RuleTooManyNegations.max_negation_frac.v                    3.89
## 29 RuleTooManyNegations.max_allowable_negations                3.65
## 30 RulePredObjDistance.max_distance                            3.64
## 31 RuleTooManyNegations.max_allowable_negations.v              3.59
## 32 ttr                                                         3.57
## 33 RuleCaseRepetition.max_repetition_count                     3.49
## 34 RulePredSubjDistance.max_distance.v                         3.47
## 35 RuleInfVerbDistance.max_distance.v                          3.38
## 36 RuleTooManyNominalConstructions.max_noun_frac.v             3.29
## 37 RuleInfVerbDistance.max_distance                            3.29
## 38 RuleMultiPartVerbs.max_distance                             3.26
## 39 fre                                                         3.23
## 40 RuleMultiPartVerbs.max_distance.v                           3.18
## 41 RuleDoubleAdpos.max_allowable_distance.v                    3.13
## 42 RulePredObjDistance.max_distance.v                          3.07
## 43 RuleDoubleAdpos.max_allowable_distance                      2.85
## 44 hpoint                                                      2.57
```

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     0.761    10 0.0127  Preprocessor1_Model1
## 2 brier_class binary     0.159    10 0.00404 Preprocessor1_Model1
## 3 roc_auc     binary     0.852    10 0.00930 Preprocessor1_Model1
```


68
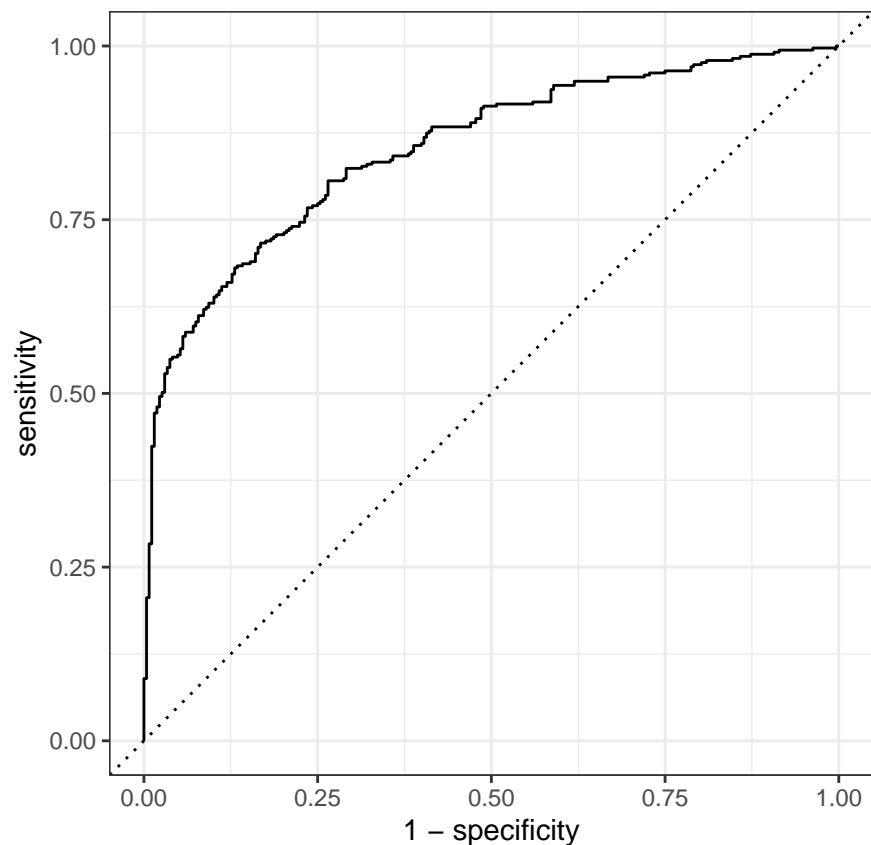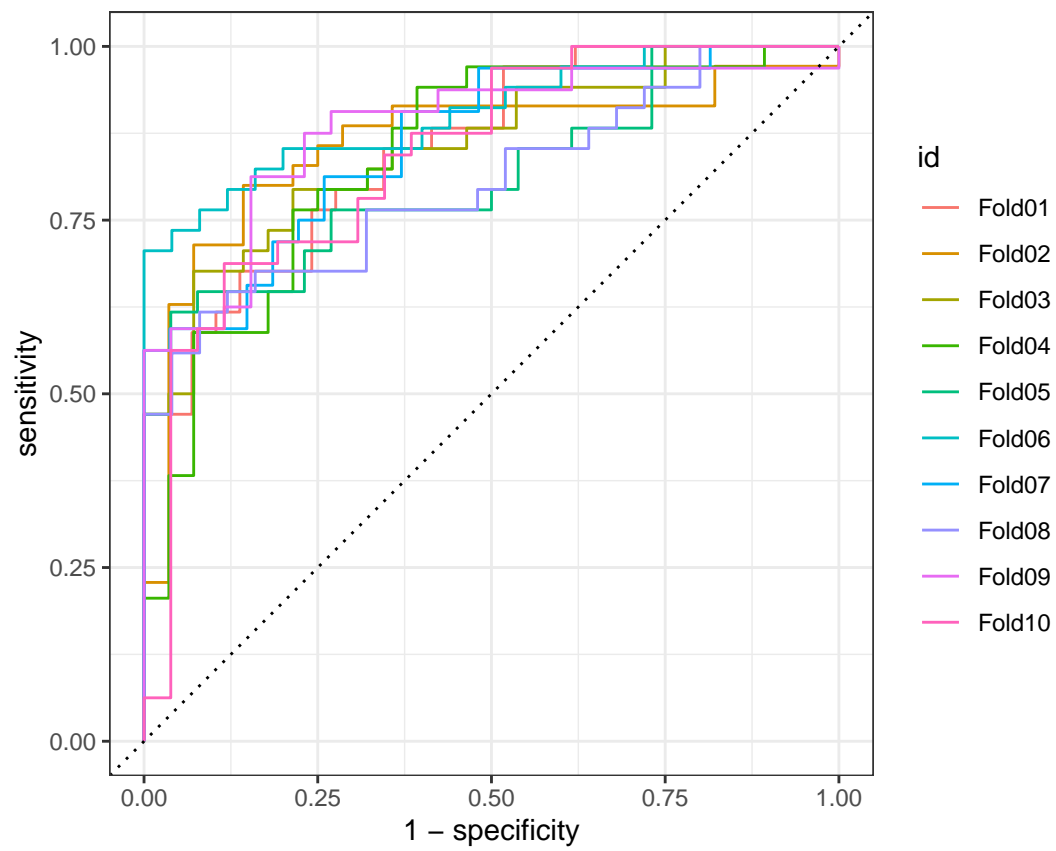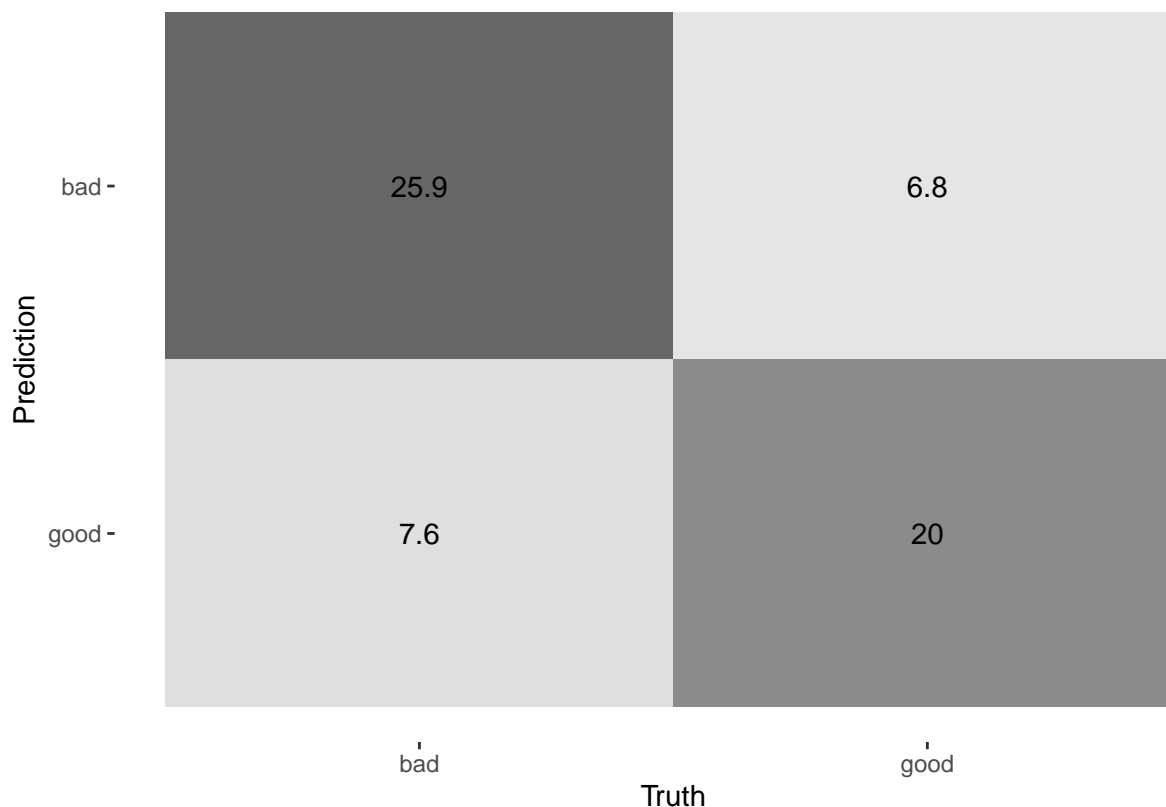
```
## [1] "\n"
## # A tibble: 24 x 2
##    Variable                  Importance
##    <chr>                          <dbl>
##  1 RulePassive                     36.8
##  2 RuleMultiPartVerbs              26.1
##  3 RuleLiteraryStyle               24.3
##  4 RulePredSubjDistance            23.5
##  5 RuleInfVerbDistance             18.6
##  6 RuleVerbalNouns                 13.7
##  7 num_hapax                       11.1
##  8 RuleAbstractNouns               10.4
##  9 RuleTooLongExpressions           9.85
## 10 RulePredObjDistance              9.66
## 11 RuleDoubleAdpos                  9.37
## 12 RuleGPwordorder                  8.94
## 13 RuleAnaphoricReferences          7.89
## 14 RuleWeakMeaningWords             7.56
## 15 RuleReflexivePassWithAnimSubj    6.66
## 16 RuleGPdeverbsubj                 4.56
## 17 RuleGPpatinstr                   4.15
## 18 RuleGPdeverbaddr                 3.55
## 19 RuleGPcoordovs                   3.47
## 20 RuleRelativisticExpressions      2.82
## 21 RuleGPpatbenperson               2.77
## 22 RuleConfirmationExpressions      2.16
## 23 RuleGPadjective                  0.887
## 24 RuleRedundantExpressions         0.506
```

Evaluations

Decision tree

All variables

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   61   18
##       good  19   53
##
##                Accuracy : 0.755
##                  95% CI : (0.6784, 0.8212)
##     No Information Rate : 0.5298
##     P-Value [Acc > NIR] : 1.014e-08
##
##                   Kappa : 0.5086
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.7465
##             Specificity : 0.7625
##          Pos Pred Value : 0.7361
##          Neg Pred Value : 0.7722
##              Prevalence : 0.4702
##          Detection Rate : 0.3510
##    Detection Prevalence : 0.4768
##       Balanced Accuracy : 0.7545
##
##        'Positive' Class : good
##
```

No TL

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   61   18
##       good  19   53
##
##                Accuracy : 0.755
##                  95% CI : (0.6784, 0.8212)
##     No Information Rate : 0.5298
##     P-Value [Acc > NIR] : 1.014e-08
##
##                   Kappa : 0.5086
##
##  Mcnemar's Test P-Value : 1
```

```
##
##              Sensitivity : 0.7465
##              Specificity : 0.7625
##           Pos Pred Value : 0.7361
##           Neg Pred Value : 0.7722
##               Prevalence : 0.4702
##           Detection Rate : 0.3510
##     Detection Prevalence : 0.4768
##        Balanced Accuracy : 0.7545
##
##         'Positive' Class : good
##
```

IAC

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   56   19
##       good  24   52
##
##                 Accuracy : 0.7152
##                   95% CI : (0.6362, 0.7856)
##      No Information Rate : 0.5298
##      P-Value [Acc > NIR] : 2.467e-06
##
##                    Kappa : 0.4307
##
##   Mcnemar's Test P-Value : 0.5419
##
##              Sensitivity : 0.7324
##              Specificity : 0.7000
##           Pos Pred Value : 0.6842
##           Neg Pred Value : 0.7467
##               Prevalence : 0.4702
##           Detection Rate : 0.3444
##     Detection Prevalence : 0.5033
##        Balanced Accuracy : 0.7162
##
##         'Positive' Class : good
##
```

Counts

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

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad   56   22
```

```
##       good  24   49
##
##                Accuracy : 0.6954
##                  95% CI : (0.6153, 0.7676)
##     No Information Rate : 0.5298
##     P-Value [Acc > NIR] : 2.505e-05
##
##                   Kappa : 0.3895
##
##  Mcnemar's Test P-Value : 0.8828
##
##             Sensitivity : 0.6901
##             Specificity : 0.7000
##          Pos Pred Value : 0.6712
##          Neg Pred Value : 0.7179
##              Prevalence : 0.4702
##          Detection Rate : 0.3245
##    Detection Prevalence : 0.4834
##       Balanced Accuracy : 0.6951
##
##        '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         0.762 Preprocessor1_Model1
## 2 roc_auc     binary         0.853 Preprocessor1_Model1
## 3 brier_class binary         0.159 Preprocessor1_Model1
```

```
## Variable importance:
## # A tibble: 71 x 3
```

```
##       Variable                                            Importance Sign
##       <chr>                                                     <dbl> <chr>
##  1 activity                                                    0.541  POS
##  2 atl                                                         0.381  POS
##  3 RuleLiteraryStyle                                           0.267  NEG
##  4 smog                                                        0.182  NEG
##  5 RulePassive                                                 0.173  NEG
##  6 maentropy                                                   0.162  NEG
##  7 entropy                                                     0.0937 NEG
##  8 RuleAnaphoricReferences                                     0.0539 POS
##  9 RuleGPcoordovs                                              0      NEG
## 10 RuleGPdeverbaddr                                            0      NEG
## 11 RuleGPpatinstr                                              0      NEG
## 12 RuleGPdeverbsubj                                            0      NEG
## 13 RuleGPadjective                                             0      NEG
## 14 RuleGPpatbenperson                                          0      NEG
## 15 RuleGPwordorder                                             0      NEG
## 16 RuleDoubleAdpos                                             0      NEG
## 17 RuleDoubleAdpos.max_allowable_distance                      0      NEG
## 18 RuleDoubleAdpos.max_allowable_distance.v                    0      NEG
## 19 RuleReflexivePassWithAnimSubj                               0      NEG
## 20 RuleTooFewVerbs.min_verb_frac                               0      NEG
## 21 RuleTooManyNegations.max_negation_frac                      0      NEG
## 22 RuleTooManyNegations.max_negation_frac.v                    0      NEG
## 23 RuleTooManyNegations.max_allowable_negations                0      NEG
## 24 RuleTooManyNegations.max_allowable_negations.v              0      NEG
## 25 RuleTooManyNominalConstructions.max_noun_frac               0      NEG
## 26 RuleTooManyNominalConstructions.max_noun_frac.v             0      NEG
## 27 RuleTooManyNominalConstructions.max_allowable_nouns         0      NEG
## 28 RuleCaseRepetition.max_repetition_count                     0      NEG
## 29 RuleCaseRepetition.max_repetition_count.v                   0      NEG
## 30 RuleCaseRepetition.max_repetition_frac                      0      NEG
## 31 RuleCaseRepetition.max_repetition_frac.v                    0      NEG
## 32 RuleWeakMeaningWords                                        0      NEG
## 33 RuleAbstractNouns                                           0      NEG
## 34 RuleRelativisticExpressions                                 0      NEG
## 35 RuleConfirmationExpressions                                 0      NEG
## 36 RuleRedundantExpressions                                    0      NEG
## 37 RuleTooLongExpressions                                      0      NEG
## 38 RulePredSubjDistance                                        0      NEG
## 39 RulePredSubjDistance.max_distance                           0      NEG
## 40 RulePredSubjDistance.max_distance.v                         0      NEG
## 41 RulePredObjDistance                                         0      NEG
## 42 RulePredObjDistance.max_distance                            0      NEG
## 43 RulePredObjDistance.max_distance.v                          0      NEG
## 44 RuleInfVerbDistance                                         0      NEG
## 45 RuleInfVerbDistance.max_distance                            0      NEG
## 46 RuleInfVerbDistance.max_distance.v                          0      NEG
## 47 RuleMultiPartVerbs                                          0      NEG
## 48 RuleMultiPartVerbs.max_distance                             0      NEG
## 49 RuleMultiPartVerbs.max_distance.v                           0      NEG
## 50 RuleLongSentences.max_length                                0      NEG
## 51 RuleLongSentences.max_length.v                              0      NEG
## 52 RulePredAtClauseBeginning.max_order                         0      NEG
```

```
## 53 RulePredAtClauseBeginning.max_order.v                    0        NEG
## 54 RuleVerbalNouns                                          0        NEG
## 55 sent_count                                               0        NEG
## 56 word_count                                               0        NEG
## 57 syllab_count                                             0        NEG
## 58 char_count                                               0        NEG
## 59 cli                                                      0        NEG
## 60 ari                                                      0        NEG
## 61 num_hapax                                                0        NEG
## 62 ttr                                                      0        NEG
## 63 mattr                                                    0        NEG
## 64 mattr.v                                                  0        NEG
## 65 maentropy.v                                              0        NEG
## 66 mamr                                                     0        NEG
## 67 verb_dist                                                0        NEG
## 68 hpoint                                                   0        NEG
## 69 fre                                                      0        NEG
## 70 fkgl                                                     0        NEG
## 71 gf                                                       0        NEG
```

```r
lasso_all_devs <- lfit_lasso_all %>% get_mismatch_details(data)
```



```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
## .pred_class bad good
##        bad   39    0
```

```
##          good   0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad   3    4
##        good  13   47
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad   11   11
##        good  3    9
##
## , , subcorpus = LiFRLaw
##
##            class
## .pred_class bad good
##        bad   1    0
##        good  0    0
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad   5    0
##        good  5    0
##
##
## Greatest deviations:
## # A tibble: 36 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.363  good        bad   FrBo       orig_Jak uspořádat shromáždění
##  2        0.324  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  3        0.312  bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
##  4        0.275  bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
##  5        0.256  good        bad   FrBo       orig_Jak namítat podjatost_final
##  6        0.249  good        bad   FrBo       orig_Kterých řízení se může váš s~
##  7        0.230  bad         good  KUKY       Reakce_na_dopis_pracovni
##  8        0.227  good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
##  9        0.227  bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 10        0.225  good        bad   KUKY       016_Obcane-EU
## 11        0.218  bad         good  KUKY       33 Cdo 30_2024
## 12        0.194  good        bad   FrBo       64
## 13        0.188  good        bad   FrBo       orig_Jak probíhá správní řízení
## 14        0.174  bad         good  KUKY       11_vizum_pred
## 15        0.169  good        bad   FrBo       149
## 16        0.165  good        bad   FrBo       orig_Jak zajistit měření hluku
## 17        0.155  bad         good  FrBo       red_Certifikáty autorizovaných in~
## 18        0.129  good        bad   FrBo       153
## 19        0.113  good        bad   FrBo       orig_financovani_politickych_stran
```

```
## 20          0.112  bad          good   KUKY        857_2024_VOP
## 21          0.0960 good         bad    FrBo        142
## 22          0.0933 bad          good   KUKY        Reakce_na_dopis_rev
## 23          0.0774 bad          good   KUKY        Mestsky_urad_Vyzva_k_zaplaceni_na~
## 24          0.0741 good         bad    OmbuFlyers  Soudni-poplatky
## 25          0.0664 good         bad    OmbuFlyers  Studny
## 26          0.0581 bad          good   KUKY        6421_2023_VOP
## 27          0.0573 bad          good   FrBo        red_závazná stanoviska_aktualizov~
## 28          0.0482 good         bad    OmbuFlyers  Spolecenstvi-vlastniku
## 29          0.0463 good         bad    FrBo        orig_Pozemkové úpravy_pracovní ve~
## 30          0.0430 good         bad    OmbuFlyers  Sikana-na-pracovisti
## # i 6 more rows
## Highest-deviating documents names:
## [1] "orig_Jak uspořádat shromáždění"
## [2] "orig_Zastupitelstvo_o čem a jak rozhoduje"
## [3] "Mestsky_urad_usneseni_-_slouceni_pred"
## [4] "2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni"
## [5] "orig_Jak namítat podjatost_final"
```

```
# 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         0.762 Preprocessor1_Model1
## 2 roc_auc     binary         0.853 Preprocessor1_Model1
## 3 brier_class binary         0.159 Preprocessor1_Model1
```

```
## Variable importance:
## # A tibble: 67 x 3
```

```
##    Variable                                              Importance Sign
##    <chr>                                                      <dbl> <chr>
## 1 activity                                                   0.541  POS
## 2 atl                                                        0.381  POS
## 3 RuleLiteraryStyle                                          0.267  NEG
## 4 smog                                                       0.182  NEG
## 5 RulePassive                                                0.173  NEG
## 6 maentropy                                                  0.162  NEG
## 7 entropy                                                    0.0937 NEG
## 8 RuleAnaphoricReferences                                    0.0539 POS
## 9 RuleGPcoordovs                                             0      NEG
## 10 RuleGPdeverbaddr                                          0      NEG
## 11 RuleGPpatinstr                                            0      NEG
## 12 RuleGPdeverbsubj                                          0      NEG
## 13 RuleGPadjective                                           0      NEG
## 14 RuleGPpatbenperson                                        0      NEG
## 15 RuleGPwordorder                                           0      NEG
## 16 RuleDoubleAdpos                                           0      NEG
## 17 RuleDoubleAdpos.max_allowable_distance                   0      NEG
## 18 RuleDoubleAdpos.max_allowable_distance.v                 0      NEG
## 19 RuleReflexivePassWithAnimSubj                            0      NEG
## 20 RuleTooFewVerbs.min_verb_frac                            0      NEG
## 21 RuleTooManyNegations.max_negation_frac                   0      NEG
## 22 RuleTooManyNegations.max_negation_frac.v                 0      NEG
## 23 RuleTooManyNegations.max_allowable_negations             0      NEG
## 24 RuleTooManyNegations.max_allowable_negations.v           0      NEG
## 25 RuleTooManyNominalConstructions.max_noun_frac            0      NEG
## 26 RuleTooManyNominalConstructions.max_noun_frac.v          0      NEG
## 27 RuleTooManyNominalConstructions.max_allowable_nouns      0      NEG
## 28 RuleCaseRepetition.max_repetition_count                  0      NEG
## 29 RuleCaseRepetition.max_repetition_count.v                0      NEG
## 30 RuleCaseRepetition.max_repetition_frac                   0      NEG
## 31 RuleCaseRepetition.max_repetition_frac.v                 0      NEG
## 32 RuleWeakMeaningWords                                     0      NEG
## 33 RuleAbstractNouns                                        0      NEG
## 34 RuleRelativisticExpressions                             0      NEG
## 35 RuleConfirmationExpressions                             0      NEG
## 36 RuleRedundantExpressions                                0      NEG
## 37 RuleTooLongExpressions                                  0      NEG
## 38 RulePredSubjDistance                                    0      NEG
## 39 RulePredSubjDistance.max_distance                       0      NEG
## 40 RulePredSubjDistance.max_distance.v                     0      NEG
## 41 RulePredObjDistance                                     0      NEG
## 42 RulePredObjDistance.max_distance                        0      NEG
## 43 RulePredObjDistance.max_distance.v                      0      NEG
## 44 RuleInfVerbDistance                                     0      NEG
## 45 RuleInfVerbDistance.max_distance                        0      NEG
## 46 RuleInfVerbDistance.max_distance.v                      0      NEG
## 47 RuleMultiPartVerbs                                      0      NEG
## 48 RuleMultiPartVerbs.max_distance                         0      NEG
## 49 RuleMultiPartVerbs.max_distance.v                       0      NEG
## 50 RuleLongSentences.max_length                            0      NEG
## 51 RuleLongSentences.max_length.v                          0      NEG
## 52 RulePredAtClauseBeginning.max_order                     0      NEG
```
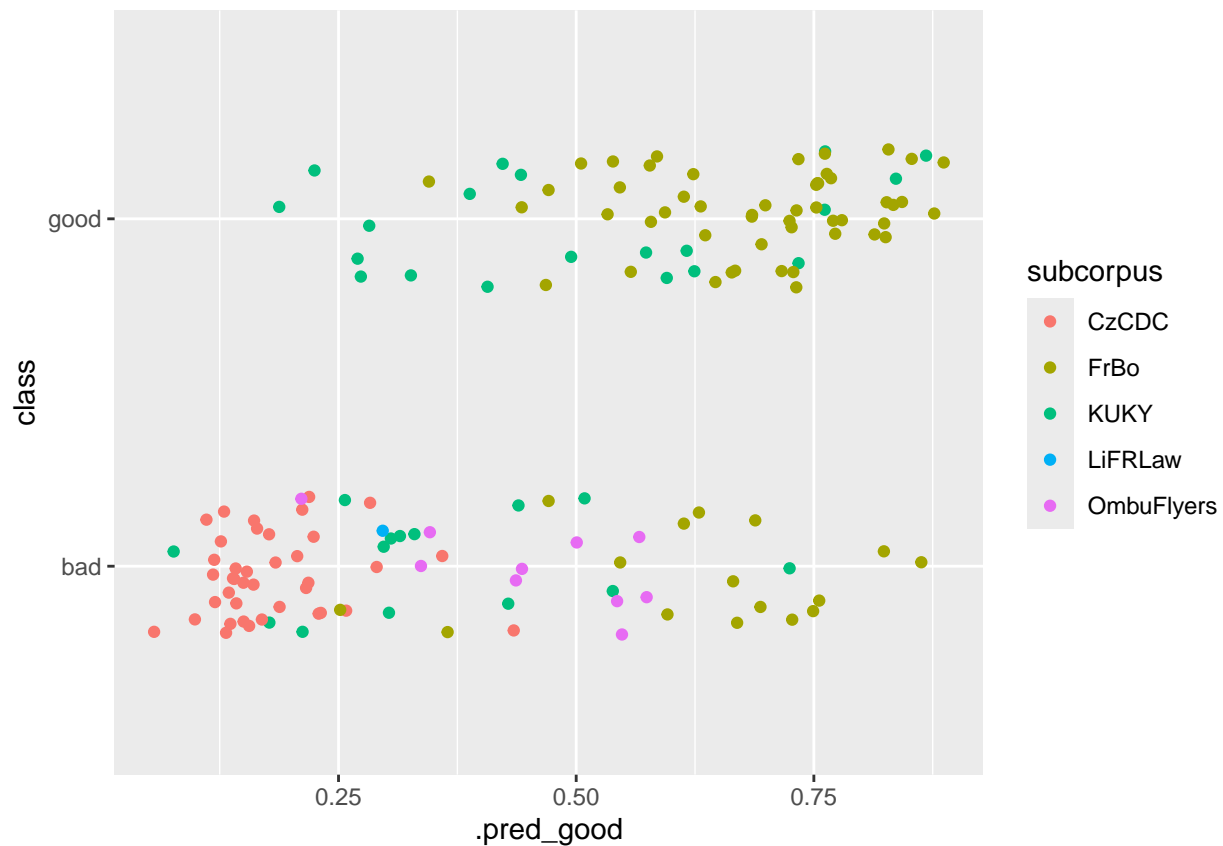
```
## 53 RulePredAtClauseBeginning.max_order.v               0        NEG
## 54 RuleVerbalNouns                                      0        NEG
## 55 cli                                                  0        NEG
## 56 ari                                                  0        NEG
## 57 num_hapax                                            0        NEG
## 58 ttr                                                  0        NEG
## 59 mattr                                                0        NEG
## 60 mattr.v                                              0        NEG
## 61 maentropy.v                                          0        NEG
## 62 mamr                                                 0        NEG
## 63 verb_dist                                            0        NEG
## 64 hpoint                                               0        NEG
## 65 fre                                                  0        NEG
## 66 fkgl                                                 0        NEG
## 67 gf                                                   0        NEG
```

```r
lasso_notl_devs <- lfit_lasso_notl %>% get_mismatch_details(data)
```



```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##           class
## .pred_class bad good
##        bad   39    0
##        good   0    0
##
## , , subcorpus = FrBo
##
```

```
##           class
## .pred_class bad good
##        bad   3    4
##        good 13   47
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##        bad  11   11
##        good  3    9
##
## , , subcorpus = LiFRLaw
##
##           class
## .pred_class bad good
##        bad   1    0
##        good  0    0
##
## , , subcorpus = OmbuFlyers
##
##           class
## .pred_class bad good
##        bad   5    0
##        good  5    0
##
##
## Greatest deviations:
## # A tibble: 36 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
## 1         0.363  good        bad   FrBo       orig_Jak uspořádat shromáždění
## 2         0.324  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
## 3         0.312  bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
## 4         0.275  bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 5         0.256  good        bad   FrBo       orig_Jak namítat podjatost_final
## 6         0.249  good        bad   FrBo       orig_Kterých řízení se může váš s~
## 7         0.230  bad         good  KUKY       Reakce_na_dopis_pracovni
## 8         0.227  good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
## 9         0.227  bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 10        0.225  good        bad   KUKY       016_Obcane-EU
## 11        0.218  bad         good  KUKY       33 Cdo 30_2024
## 12        0.194  good        bad   FrBo       64
## 13        0.188  good        bad   FrBo       orig_Jak probíhá správní řízení
## 14        0.174  bad         good  KUKY       11_vizum_pred
## 15        0.169  good        bad   FrBo       149
## 16        0.165  good        bad   FrBo       orig_Jak zajistit měření hluku
## 17        0.155  bad         good  FrBo       red_Certifikáty autorizovaných in~
## 18        0.129  good        bad   FrBo       153
## 19        0.113  good        bad   FrBo       orig_financovani_politickych_stran
## 20        0.112  bad         good  KUKY       857_2024_VOP
## 21        0.0960 good        bad   FrBo       142
## 22        0.0933 bad         good  KUKY       Reakce_na_dopis_rev
## 23        0.0774 bad         good  KUKY       Mestsky_urad_Vyzva_k_zaplaceni_na~
```
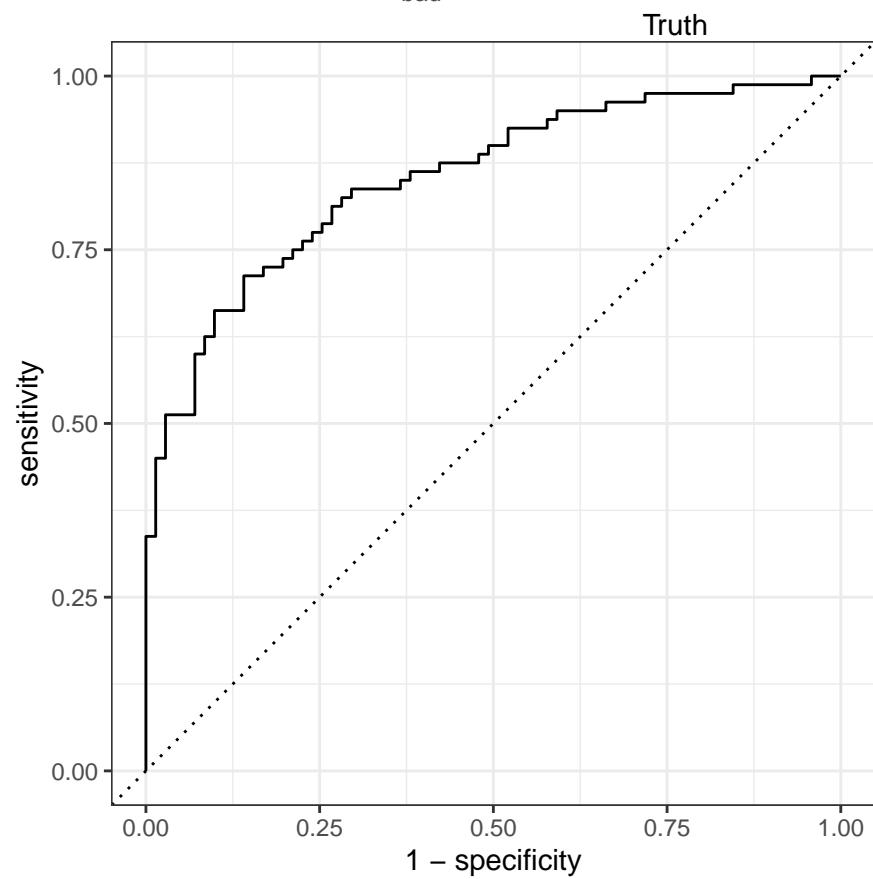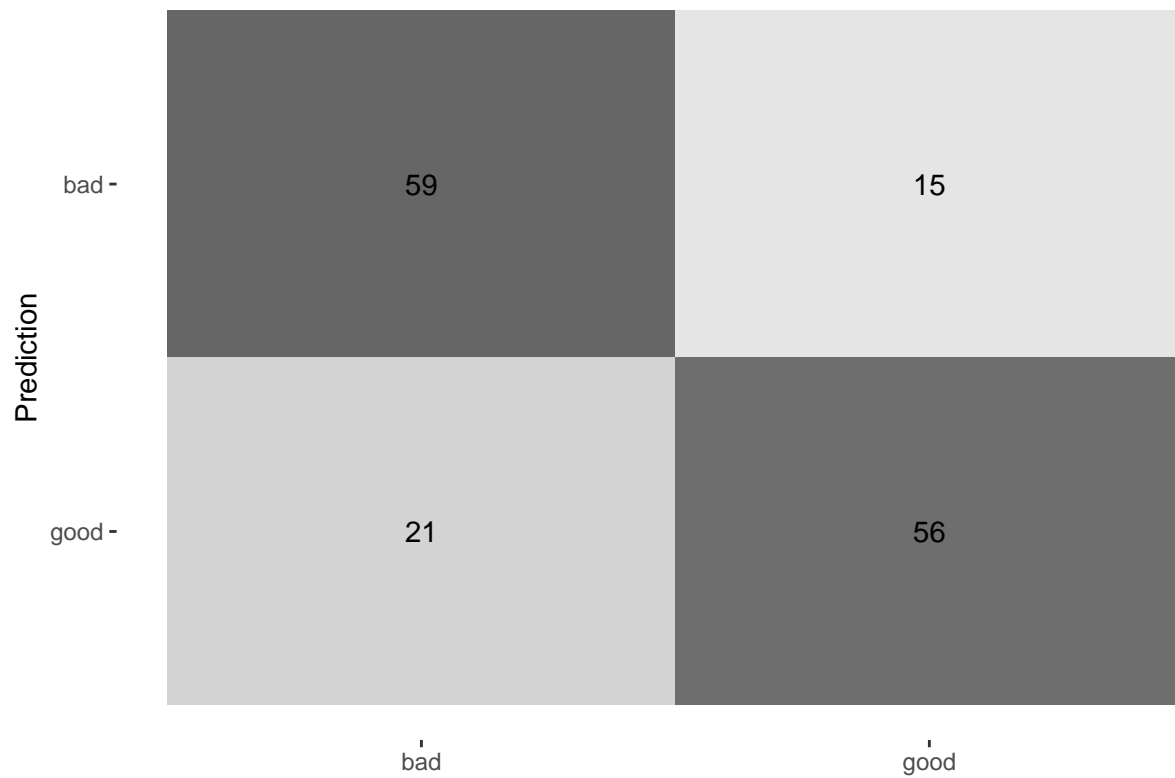
```
## 24          0.0741 good          bad     OmbuFlyers Soudni-poplatky
## 25          0.0664 good          bad     OmbuFlyers Studny
## 26          0.0581 bad           good    KUKY       6421_2023_VOP
## 27          0.0573 bad           good    FrBo       red_závazná stanoviska_aktualizov~
## 28          0.0482 good          bad     OmbuFlyers Spolecenstvi-vlastniku
## 29          0.0463 good          bad     FrBo       orig_Pozemkové úpravy_pracovní ve~
## 30          0.0430 good          bad     OmbuFlyers Sikana-na-pracovisti
## # i 6 more rows
## Highest-deviating documents names:
## [1] "orig_Jak uspořádat shromáždění"
## [2] "orig_Zastupitelstvo_o čem a jak rozhoduje"
## [3] "Mestsky_urad_usneseni_-_slouceni_pred"
## [4] "2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni"
## [5] "orig_Jak namítat podjatost_final"
```

```
# 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         0.748 Preprocessor1_Model1
## 2 roc_auc     binary         0.838 Preprocessor1_Model1
## 3 brier_class binary         0.167 Preprocessor1_Model1
```

```
## Variable importance:
## # A tibble: 44 x 3
```
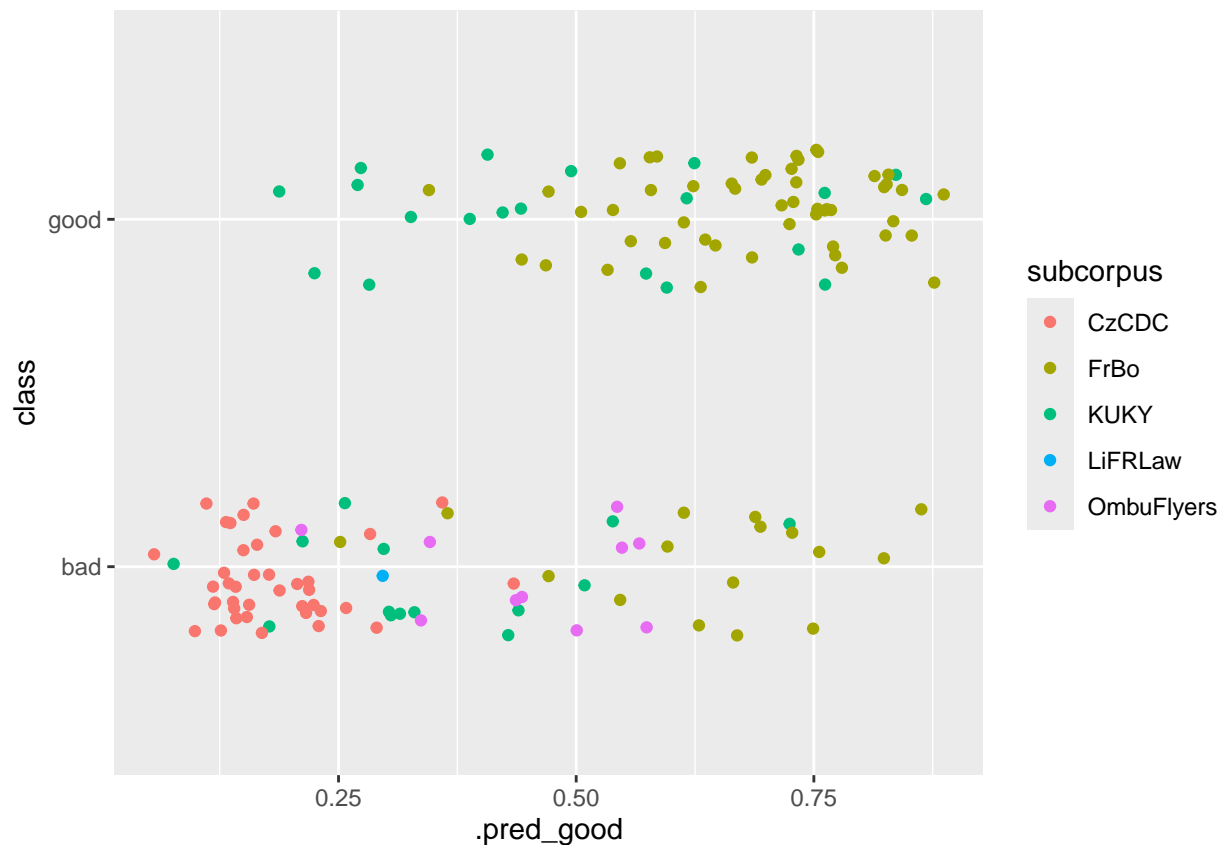
```
##      Variable                                              Importance Sign
##      <chr>                                                      <dbl> <chr>
##  1 activity                                                      5.25 POS
##  2 maentropy                                                     1.41 NEG
##  3 atl                                                           1.03 POS
##  4 RuleTooManyNegations.max_allowable_negations.v               0.345 NEG
##  5 entropy                                                       0.258 NEG
##  6 smog                                                          0.111 NEG
##  7 RuleTooManyNominalConstructions.max_allowable_nouns.v       0.0104 NEG
##  8 gf                                                        0.000136 NEG
##  9 RuleDoubleAdpos.max_allowable_distance                           0 NEG
## 10 RuleDoubleAdpos.max_allowable_distance.v                         0 NEG
## 11 RuleTooFewVerbs.min_verb_frac                                    0 NEG
## 12 RuleTooManyNegations.max_negation_frac                           0 NEG
## 13 RuleTooManyNegations.max_negation_frac.v                         0 NEG
## 14 RuleTooManyNegations.max_allowable_negations                     0 NEG
## 15 RuleTooManyNominalConstructions.max_noun_frac                    0 NEG
## 16 RuleTooManyNominalConstructions.max_noun_frac.v                  0 NEG
## 17 RuleTooManyNominalConstructions.max_allowable_nouns              0 NEG
## 18 RuleCaseRepetition.max_repetition_count                          0 NEG
## 19 RuleCaseRepetition.max_repetition_count.v                        0 NEG
## 20 RuleCaseRepetition.max_repetition_frac                           0 NEG
## 21 RuleCaseRepetition.max_repetition_frac.v                         0 NEG
## 22 RulePredSubjDistance.max_distance                                0 NEG
## 23 RulePredSubjDistance.max_distance.v                              0 NEG
## 24 RulePredObjDistance.max_distance                                 0 NEG
## 25 RulePredObjDistance.max_distance.v                               0 NEG
## 26 RuleInfVerbDistance.max_distance                                 0 NEG
## 27 RuleInfVerbDistance.max_distance.v                               0 NEG
## 28 RuleMultiPartVerbs.max_distance                                  0 NEG
## 29 RuleMultiPartVerbs.max_distance.v                                0 NEG
## 30 RuleLongSentences.max_length                                     0 NEG
## 31 RuleLongSentences.max_length.v                                   0 NEG
## 32 RulePredAtClauseBeginning.max_order                              0 NEG
## 33 RulePredAtClauseBeginning.max_order.v                            0 NEG
## 34 cli                                                             0 NEG
## 35 ari                                                             0 NEG
## 36 ttr                                                             0 NEG
## 37 mattr                                                           0 NEG
## 38 mattr.v                                                         0 NEG
## 39 maentropy.v                                                     0 NEG
## 40 mamr                                                            0 NEG
## 41 verb_dist                                                       0 NEG
## 42 hpoint                                                          0 NEG
## 43 fre                                                             0 NEG
## 44 fkgl                                                            0 NEG

lasso_iac_devs <- lfit_lasso_iac %>% get_mismatch_details(data)
```
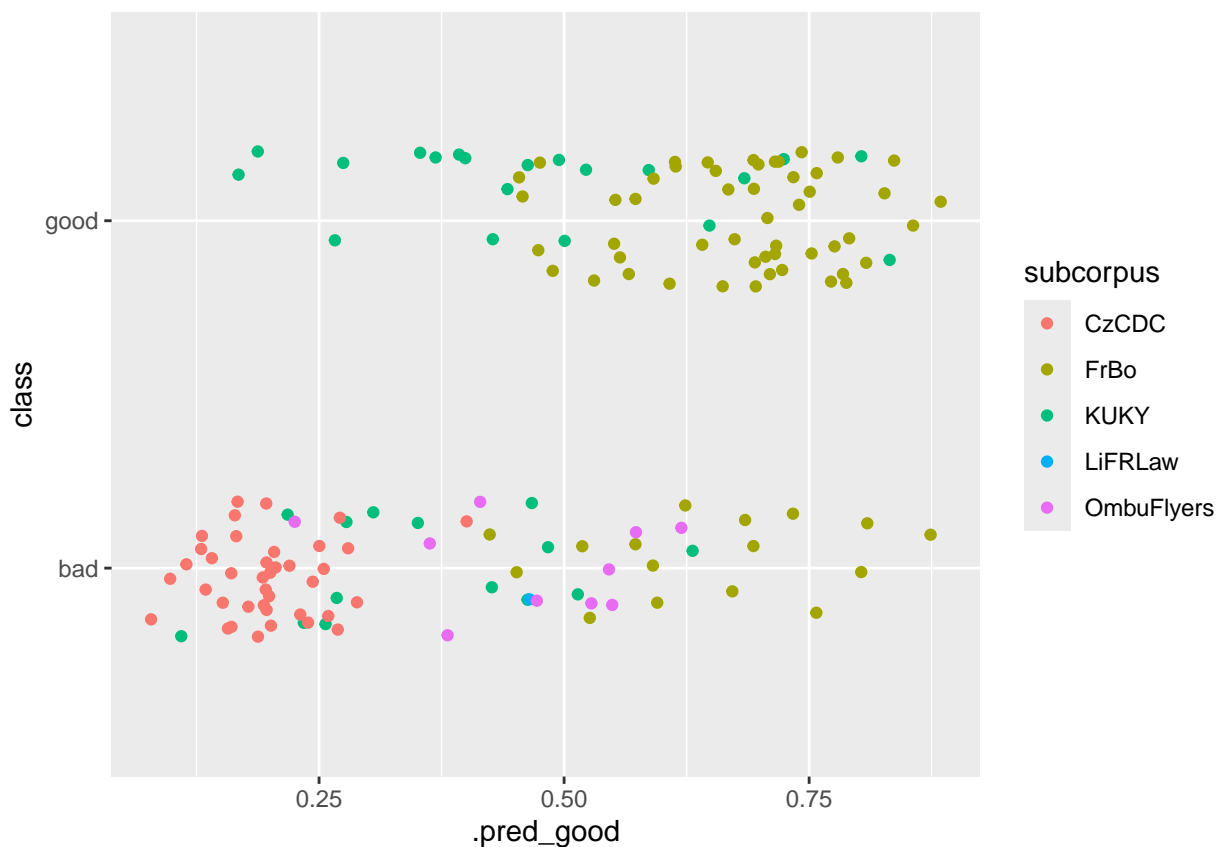
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##           class
## .pred_class bad good
##        bad   39    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##           class
## .pred_class bad good
##        bad    2    5
##        good  14   46
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##        bad   12   12
##        good   2    8
##
## , , subcorpus = LiFRLaw
##
##           class
## .pred_class bad good
##        bad    1    0
```

```
##        good    0    0
## 
## , , subcorpus = OmbuFlyers
## 
##            class
## .pred_class bad good
##        bad   5    0
##        good  5    0
## 
## 
## Greatest deviations:
## # A tibble: 38 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1        0.374  good        bad   FrBo       orig_Jak uspořádat shromáždění
##  2        0.332  bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
##  3        0.313  bad         good  KUKY       Mestsky_urad_Vyzva_k_zaplaceni_na~
##  4        0.309  good        bad   FrBo       orig_Jak namítat podjatost_final
##  5        0.303  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  6        0.257  good        bad   FrBo       orig_Jak probíhá správní řízení
##  7        0.234  bad         good  KUKY       33 Cdo 30_2024
##  8        0.234  good        bad   FrBo       orig_Kterých řízení se může váš s~
##  9        0.225  bad         good  KUKY       11_vizum_pred
## 10        0.193  good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
## 11        0.185  good        bad   FrBo       64
## 12        0.172  good        bad   FrBo       orig_Jak zajistit měření hluku
## 13        0.147  bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 14        0.131  good        bad   KUKY       016_Obcane-EU
## 15        0.131  bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 16        0.124  good        bad   FrBo       153
## 17        0.120  good        bad   OmbuFlyers Soudni-poplatky
## 18        0.107  bad         good  KUKY       Reakce_na_dopis_pracovni
## 19        0.101  bad         good  KUKY       857_2024_VOP
## 20        0.0951 good        bad   FrBo       149
## 21        0.0907 good        bad   FrBo       142
## 22        0.0733 good        bad   OmbuFlyers Spolecenstvi-vlastniku
## 23        0.0728 good        bad   FrBo       orig_Pozemkové úpravy_pracovní ve~
## 24        0.0727 bad         good  KUKY       6421_2023_VOP
## 25        0.0578 bad         good  KUKY       6525_2022_VOP
## 26        0.0491 good        bad   OmbuFlyers Studny
## 27        0.0459 bad         good  FrBo       1
## 28        0.0458 good        bad   OmbuFlyers Sikana-na-pracovisti
## 29        0.0423 bad         good  FrBo       orig_Soustavné obtěžování hlukem ~
## 30        0.0370 bad         good  KUKY       KVOP_19_Stavarska_zprava_JSm
## # i 8 more rows
## Highest-deviating documents names:
## [1] "orig_Jak uspořádat shromáždění"
## [2] "Mestsky_urad_usneseni_-_slouceni_pred"
## [3] "Mestsky_urad_Vyzva_k_zaplaceni_nakladu_rizeni_pred"
## [4] "orig_Jak namítat podjatost_final"
## [5] "orig_Zastupitelstvo_o čem a jak rozhoduje"
## [6] "orig_Jak probíhá správní řízení"
```

```
# 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         0.762 Preprocessor1_Model1
## 2 roc_auc     binary         0.830 Preprocessor1_Model1
## 3 brier_class binary         0.173 Preprocessor1_Model1
```

```
## Variable importance:
## # A tibble: 24 x 3
##    Variable                    Importance Sign
##    <chr>                            <dbl> <chr>
##  1 RuleRelativisticExpressions      140.  NEG
##  2 RulePassive                      122.  NEG
##  3 RuleLiteraryStyle                102.  NEG
##  4 RuleAnaphoricReferences           39.8 POS
##  5 RuleMultiPartVerbs                24.0 POS
##  6 RulePredSubjDistance              14.6 POS
##  7 RuleVerbalNouns                    5.24 POS
##  8 RuleInfVerbDistance                1.02 POS
##  9 RuleGPcoordovs                     0   NEG
## 10 RuleGPdeverbaddr                   0   NEG
## 11 RuleGPpatinstr                     0   NEG
## 12 RuleGPdeverbsubj                   0   NEG
## 13 RuleGPadjective                    0   NEG
## 14 RuleGPpatbenperson                 0   NEG
## 15 RuleGPwordorder                    0   NEG
## 16 RuleDoubleAdpos                    0   NEG
## 17 RuleReflexivePassWithAnimSubj      0   NEG
## 18 RuleWeakMeaningWords               0   NEG
## 19 RuleAbstractNouns                  0   NEG
## 20 RuleConfirmationExpressions        0   NEG
## 21 RuleRedundantExpressions           0   NEG
## 22 RuleTooLongExpressions             0   NEG
## 23 RulePredObjDistance                0   NEG
```

```
## 24 num_hapax                                    0     NEG
```

```
lasso_counts_devs <- lfit_lasso_counts %>% get_mismatch_details(data)
```



```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##           class
## .pred_class bad good
##        bad   38    0
##        good   1    0
##
## , , subcorpus = FrBo
##
##           class
## .pred_class bad good
##        bad    3    6
##        good  13   45
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##        bad   11   11
##        good   3    9
##
## , , subcorpus = LiFRLaw
##
```

```
##            class
## .pred_class bad good
##        bad    1    0
##        good   0    0
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad    8    0
##        good   2    0
##
##
## Greatest deviations:
## # A tibble: 36 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
## 1        0.403   bad         good  KUKY       Reakce_na_dopis_pracovni
## 2        0.298   bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 3        0.280   bad         good  KUKY       Reakce_na_dopis_rev
## 4        0.273   good        bad   FrBo       orig_Kterých řízení se může váš s~
## 5        0.252   bad         good  FrBo       red_Certifikáty autorizovaných in~
## 6        0.251   bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 7        0.248   good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
## 8        0.234   good        bad   FrBo       orig_Jak uspořádat shromáždění
## 9        0.198   good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
## 10       0.188   bad         good  FrBo       red_10 významných práv účastníka ~
## 11       0.180   bad         good  KUKY       857_2024_VOP
## 12       0.180   bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
## 13       0.135   bad         good  KUKY       33 Cdo 30_2024
## 14       0.135   good        bad   FrBo       orig_Jak probíhá správní řízení
## 15       0.125   good        bad   FrBo       orig_Změny v zákoně o EIA
## 16       0.122   bad         good  KUKY       11_vizum_pred
## 17       0.122   good        bad   FrBo       149
## 18       0.121   good        bad   FrBo       orig_Jak namítat podjatost_final
## 19       0.120   bad         good  FrBo       red_závazná stanoviska_aktualizov~
## 20       0.117   bad         good  FrBo       red_Jak podat trestní oznámení
## 21       0.112   good        bad   CzCDC      4-34-13_1
## 22       0.104   good        bad   FrBo       orig_financovani_politickych_stran
## 23       0.101   good        bad   KUKY       016_Obcane-EU
## 24       0.0997  good        bad   FrBo       orig_Jak zajistit měření hluku
## 25       0.0929  good        bad   FrBo       64
## 26       0.0777  good        bad   FrBo       153
## 27       0.0668  bad         good  KUKY       6421_2023_VOP
## 28       0.0622  bad         good  KUKY       důchod-dorovnávací přídavek_1298-~
## 29       0.0620  good        bad   OmbuFlyers Studny
## 30       0.0578  good        bad   FrBo       orig_provokace_korupcniho_jednani
## # i 6 more rows
## Highest-deviating documents names:
## [1] "Reakce_na_dopis_pracovni"
## [2] "2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni"
## [3] "Reakce_na_dopis_rev"
## [4] "orig_Kterých řízení se může váš spolek účastnit_FINAL"
## [5] "red_Certifikáty autorizovaných inspektorů"
```

```
## [6] "1A_dokument_puvodni_ustanoven_zastupce_vyzva_k_doplneni_kast_pouceni"
```

```
# 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         0.755 Preprocessor1_Model1
## 2 roc_auc      binary         0.835 Preprocessor1_Model1
## 3 brier_class  binary         0.163 Preprocessor1_Model1
```
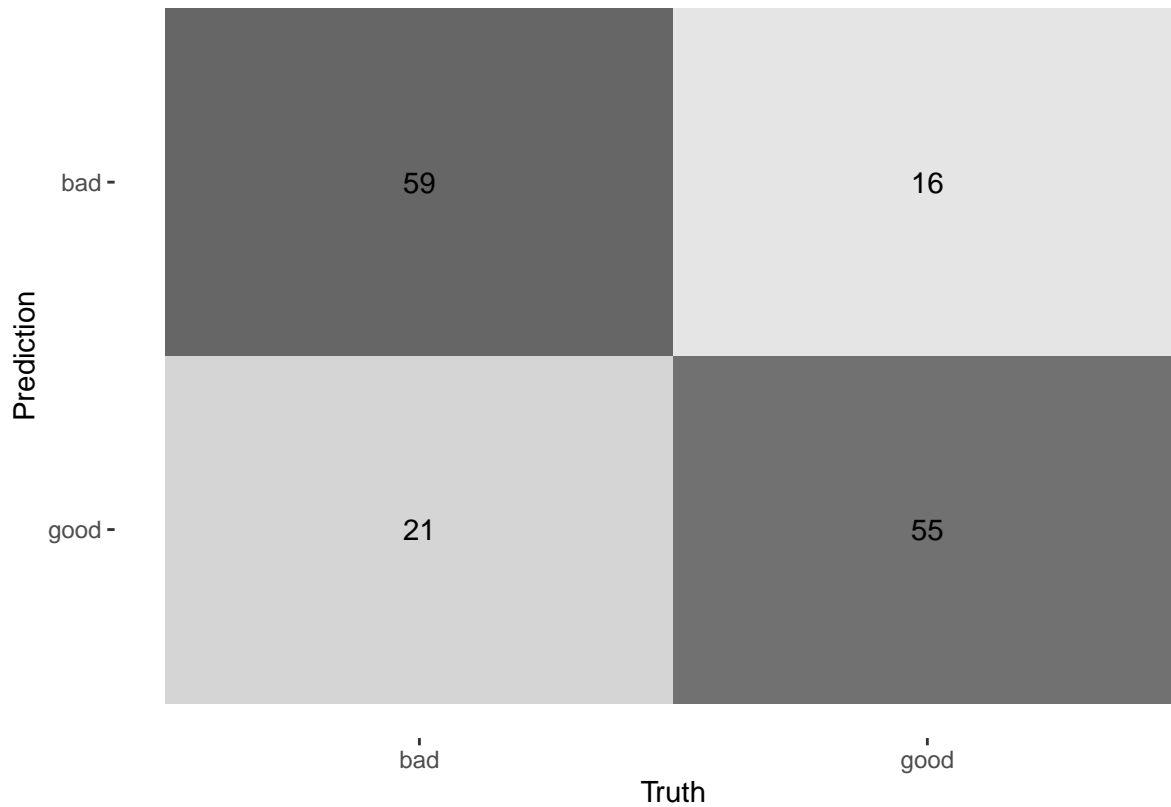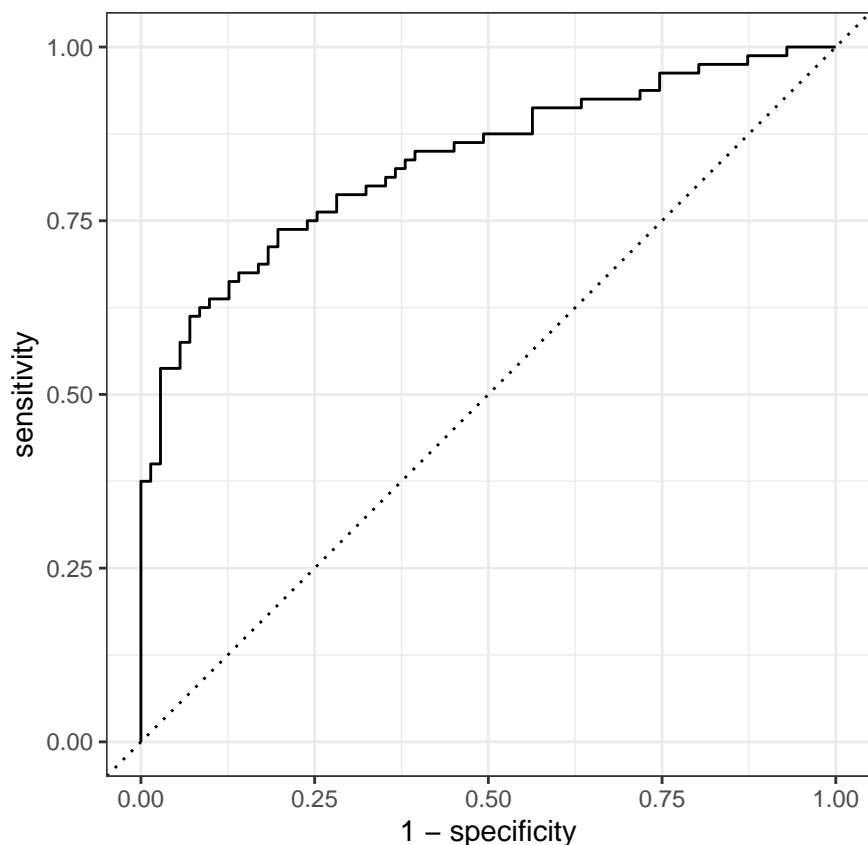
```
## Variable importance:
## # A tibble: 71 x 2
##    Variable                                          Importance
##    <chr>                                                  <dbl>
##  1 activity                                                16.7
##  2 RuleTooFewVerbs.min_verb_frac                           14.4
##  3 verb_dist                                               12.5
##  4 RuleTooManyNominalConstructions.max_allowable_nouns     10.3
##  5 RuleLongSentences.max_length                            10.1
##  6 ari                                                      8.91
##  7 RulePredAtClauseBeginning.max_order                      8.87
##  8 RulePassive                                              8.68
##  9 gf                                                       8.52
## 10 RuleLiteraryStyle                                        6.38
## 11 smog                                                     6.23
## 12 atl                                                      5.84
## 13 maentropy                                                4.71
## 14 fkgl                                                     4.60
## 15 RuleTooManyNegations.max_negation_frac                   4.43
## 16 mamr                                                     4.35
## 17 RulePredAtClauseBeginning.max_order.v                    4.32
## 18 mattr                                                    4.31
## 19 RuleTooLongExpressions                                   4.22
## 20 RuleTooManyNominalConstructions.max_noun_frac            3.92
## 21 RuleMultiPartVerbs                                       3.88
## 22 RuleVerbalNouns                                          3.67
## 23 cli                                                      3.60
```

```
## 24 RulePredSubjDistance                                   3.53
## 25 RuleCaseRepetition.max_repetition_count.v              3.28
## 26 RulePredSubjDistance.max_distance                      3.27
## 27 RuleAnaphoricReferences                                3.23
## 28 maentropy.v                                            3.19
## 29 RuleCaseRepetition.max_repetition_frac                 3.18
## 30 RuleCaseRepetition.max_repetition_frac.v               3.11
## 31 RuleLongSentences.max_length.v                         3.02
## 32 RuleTooManyNegations.max_allowable_negations.v         2.93
## 33 entropy                                                2.80
## 34 RulePredSubjDistance.max_distance.v                    2.78
## 35 mattr.v                                                2.76
## 36 fre                                                    2.75
## 37 RulePredObjDistance.max_distance                       2.72
## 38 RuleCaseRepetition.max_repetition_count                2.64
## 39 RuleDoubleAdpos                                        2.56
## 40 RuleInfVerbDistance                                    2.55
## 41 RuleTooManyNegations.max_negation_frac.v               2.46
## 42 RulePredObjDistance                                    2.42
## 43 RuleInfVerbDistance.max_distance                       2.30
## 44 RuleTooManyNegations.max_allowable_negations           2.30
## 45 RuleTooManyNominalConstructions.max_noun_frac.v        2.29
## 46 RuleMultiPartVerbs.max_distance                        2.26
## 47 RuleAbstractNouns                                      2.18
## 48 num_hapax                                              2.15
## 49 RuleInfVerbDistance.max_distance.v                     2.11
## 50 RuleMultiPartVerbs.max_distance.v                      2.10
## 51 char_count                                             2.04
## 52 RulePredObjDistance.max_distance.v                     2.04
## 53 ttr                                                    2.03
## 54 sent_count                                             2.03
## 55 word_count                                             2.00
## 56 RuleDoubleAdpos.max_allowable_distance.v               1.97
## 57 RuleWeakMeaningWords                                   1.96
## 58 syllab_count                                           1.88
## 59 RuleDoubleAdpos.max_allowable_distance                 1.63
## 60 RuleGPwordorder                                        1.54
## 61 RuleGPcoordovs                                         1.41
## 62 RuleReflexivePassWithAnimSubj                          1.33
## 63 hpoint                                                 1.30
## 64 RuleGPpatinstr                                         1.27
## 65 RuleGPdeverbaddr                                       0.943
## 66 RuleRelativisticExpressions                            0.780
## 67 RuleGPpatbenperson                                     0.728
## 68 RuleGPdeverbsubj                                       0.633
## 69 RuleConfirmationExpressions                            0.514
## 70 RuleGPadjective                                        0.328
## 71 RuleRedundantExpressions                               0.0917
```

```r
rf_all_devs <- lfit_rf_all %>% get_mismatch_details(data)
```

```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##           class
## .pred_class bad good
##        bad   39    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##           class
## .pred_class bad good
##        bad    3    6
##        good  13   45
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##        bad   10   10
##        good   4   10
##
## , , subcorpus = LiFRLaw
##
##           class
## .pred_class bad good
##        bad    1    0
```

```
##        good   0   0
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad   6   0
##        good  4   0
##
##
## Greatest deviations:
## # A tibble: 37 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
## 1         0.402  good        bad   FrBo       orig_Jak uspořádat shromáždění
## 2         0.383  bad         good  KUKY       33 Cdo 30_2024
## 3         0.363  bad         good  KUKY       11_vizum_pred
## 4         0.352  good        bad   FrBo       orig_Kterých řízení se může váš s~
## 5         0.321  good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
## 6         0.302  good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
## 7         0.299  good        bad   FrBo       orig_Jak probíhá správní řízení
## 8         0.287  good        bad   FrBo       orig_Jak namítat podjatost_final
## 9         0.245  good        bad   OmbuFlyers Soudni-poplatky
## 10        0.240  bad         good  FrBo       red_Certifikáty autorizovaných in~
## 11        0.229  good        bad   FrBo       142
## 12        0.225  bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
## 13        0.211  good        bad   OmbuFlyers Studny
## 14        0.203  good        bad   FrBo       64
## 15        0.199  bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 16        0.164  good        bad   FrBo       orig_Jak zajistit měření hluku
## 17        0.144  good        bad   FrBo       orig_provokace_korupcniho_jednani
## 18        0.143  bad         good  KUKY       Reakce_na_dopis_rev
## 19        0.136  bad         good  FrBo       1
## 20        0.122  bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 21        0.113  bad         good  FrBo       red_Jak podat trestní oznámení
## 22        0.107  good        bad   KUKY       016_Obcane-EU
## 23        0.103  good        bad   KUKY       sluzebni_hodnoceni_puvodni
## 24        0.0964 good        bad   OmbuFlyers Socialni-sluzby
## 25        0.0882 good        bad   KUKY       7-Co-1752-2016-Vyber-judikatury
## 26        0.0843 good        bad   KUKY       UOOUOsobniUdajePuvodne
## 27        0.0775 bad         good  KUKY       Mestsky_urad_Vyzva_k_zaplaceni_na~
## 28        0.0693 bad         good  KUKY       důchod-dorovnávací přídavek_1298-~
## 29        0.0656 bad         good  FrBo       red_Les - co smíme a co je zakázá~
## 30        0.0539 bad         good  FrBo       red_závazná stanoviska_aktualizov~
## # i 7 more rows
## Highest-deviating documents names:
## [1] "orig_Jak uspořádat shromáždění"
## [2] "33 Cdo 30_2024"
## [3] "11_vizum_pred"
## [4] "orig_Kterých řízení se může váš spolek účastnit_FINAL"
## [5] "orig_lhuty_v_jednani_s_urady_a_soudy"
## [6] "orig_Zastupitelstvo_o čem a jak rozhoduje"
## [7] "orig_Jak probíhá správní řízení"
## [8] "orig_Jak namítat podjatost_final"
```

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         0.762 Preprocessor1_Model1
## 2 roc_auc      binary         0.843 Preprocessor1_Model1
## 3 brier_class  binary         0.160 Preprocessor1_Model1
```
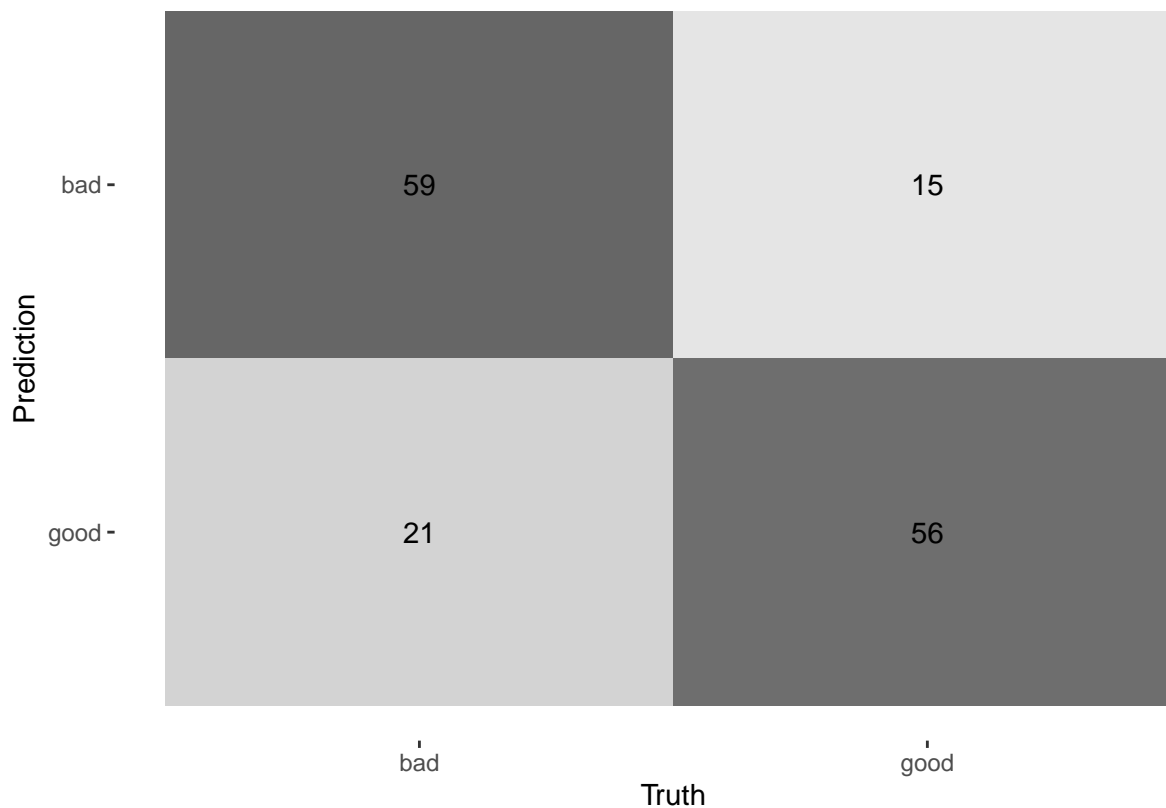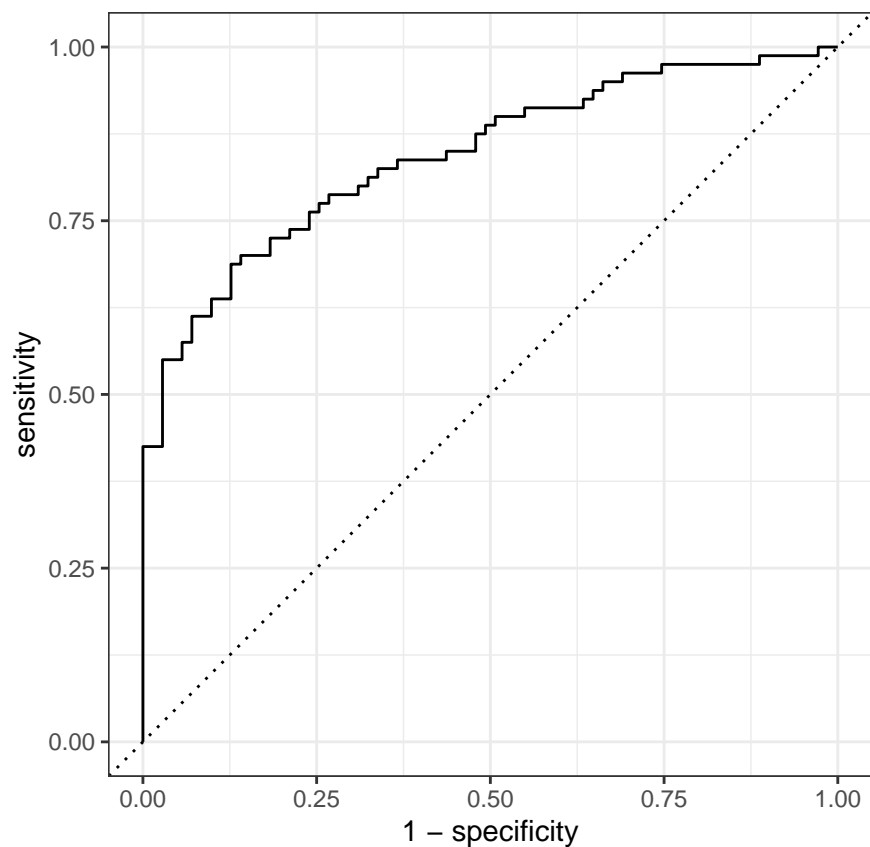
```
## Variable importance:
## # A tibble: 67 x 2
##     Variable                                                Importance
##     <chr>                                                        <dbl>
##  1 activity                                                      15.1
##  2 verb_dist                                                     14.0
##  3 RuleTooFewVerbs.min_verb_frac                                 13.3
##  4 RuleLongSentences.max_length                                  11.6
##  5 RuleTooManyNominalConstructions.max_allowable_nouns           10.3
##  6 RulePredAtClauseBeginning.max_order                            9.69
##  7 ari                                                            9.56
##  8 RulePassive                                                    9.06
##  9 gf                                                             8.31
## 10 RuleLiteraryStyle                                              6.88
## 11 atl                                                            6.28
## 12 smog                                                           6.13
## 13 fkgl                                                           5.56
## 14 RuleTooManyNegations.max_negation_frac                         4.96
## 15 mamr                                                           4.50
## 16 RuleTooLongExpressions                                         4.47
## 17 maentropy                                                      4.36
## 18 RuleVerbalNouns                                                4.22
## 19 RulePredAtClauseBeginning.max_order.v                          4.17
## 20 RuleMultiPartVerbs                                             4.12
## 21 RuleTooManyNominalConstructions.max_noun_frac                  4.04
## 22 mattr                                                          3.76
## 23 RulePredSubjDistance                                           3.69
```

```
## 24 RuleLongSentences.max_length.v                      3.60
## 25 RuleAnaphoricReferences                             3.37
## 26 RulePredSubjDistance.max_distance                   3.34
## 27 cli                                                 3.32
## 28 RuleCaseRepetition.max_repetition_count.v           3.30
## 29 RuleCaseRepetition.max_repetition_frac.v            3.20
## 30 entropy                                             3.17
## 31 maentropy.v                                         3.14
## 32 RuleCaseRepetition.max_repetition_frac              3.08
## 33 mattr.v                                             3.06
## 34 RulePredSubjDistance.max_distance.v                 2.87
## 35 RuleCaseRepetition.max_repetition_count             2.80
## 36 RulePredObjDistance.max_distance                    2.69
## 37 RuleTooManyNegations.max_allowable_negations.v      2.66
## 38 num_hapax                                           2.58
## 39 RuleInfVerbDistance                                 2.57
## 40 RuleTooManyNegations.max_allowable_negations        2.51
## 41 RuleInfVerbDistance.max_distance                    2.49
## 42 RuleInfVerbDistance.max_distance.v                  2.34
## 43 RuleTooManyNegations.max_negation_frac.v            2.34
## 44 RuleWeakMeaningWords                                2.34
## 45 RulePredObjDistance.max_distance.v                  2.33
## 46 RuleDoubleAdpos                                     2.31
## 47 ttr                                                 2.31
## 48 fre                                                 2.27
## 49 RuleTooManyNominalConstructions.max_noun_frac.v     2.25
## 50 RuleMultiPartVerbs.max_distance                     2.25
## 51 RuleDoubleAdpos.max_allowable_distance.v            2.22
## 52 RulePredObjDistance                                 2.21
## 53 RuleAbstractNouns                                   2.09
## 54 RuleMultiPartVerbs.max_distance.v                   2.05
## 55 RuleDoubleAdpos.max_allowable_distance              1.88
## 56 hpoint                                              1.78
## 57 RuleGPcoordovs                                      1.56
## 58 RuleGPwordorder                                     1.42
## 59 RuleReflexivePassWithAnimSubj                       1.25
## 60 RuleGPdeverbaddr                                    1.15
## 61 RuleGPpatinstr                                      1.15
## 62 RuleRelativisticExpressions                         0.945
## 63 RuleGPpatbenperson                                  0.792
## 64 RuleGPdeverbsubj                                    0.768
## 65 RuleConfirmationExpressions                         0.654
## 66 RuleGPadjective                                     0.383
## 67 RuleRedundantExpressions                            0.126
```

```r
rf_notl_devs <- lfit_rf_notl %>% get_mismatch_details(data)
```

```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##           class
## .pred_class bad good
##        bad   39    0
##        good   0    0
##
## , , subcorpus = FrBo
##
##           class
## .pred_class bad good
##        bad    3    5
##        good  13   46
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##        bad   10   10
##        good   4   10
##
## , , subcorpus = LiFRLaw
##
##           class
## .pred_class bad good
##        bad    1    0
```

```
##         good    0    0
##
## , , subcorpus = OmbuFlyers
##
##           class
## .pred_class bad good
##        bad    6    0
##        good    4    0
##
##
## Greatest deviations:
## # A tibble: 36 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
##  1         0.416 good        bad   FrBo       orig_Jak uspořádat shromáždění
##  2         0.376 good        bad   FrBo       orig_Kterých řízení se může váš s~
##  3         0.359 bad         good  KUKY       11_vizum_pred
##  4         0.357 bad         good  KUKY       33 Cdo 30_2024
##  5         0.300 good        bad   FrBo       orig_Jak namítat podjatost_final
##  6         0.285 good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
##  7         0.272 good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
##  8         0.265 good        bad   FrBo       orig_Jak probíhá správní řízení
##  9         0.250 good        bad   FrBo       142
## 10         0.238 bad         good  FrBo       red_Certifikáty autorizovaných in~
## 11         0.216 bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
## 12         0.212 good        bad   OmbuFlyers Studny
## 13         0.201 good        bad   OmbuFlyers Soudni-poplatky
## 14         0.187 bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 15         0.182 good        bad   FrBo       64
## 16         0.172 good        bad   FrBo       orig_provokace_korupcniho_jednani
## 17         0.172 good        bad   FrBo       orig_Jak zajistit měření hluku
## 18         0.161 bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 19         0.157 bad         good  KUKY       Reakce_na_dopis_rev
## 20         0.148 bad         good  FrBo       1
## 21         0.128 good        bad   KUKY       016_Obcane-EU
## 22         0.119 bad         good  FrBo       red_Jak podat trestní oznámení
## 23        0.0949 good        bad   KUKY       sluzebni_hodnoceni_puvodni
## 24        0.0878 good        bad   KUKY       7-Co-1752-2016-Vyber-judikatury
## 25        0.0780 good        bad   OmbuFlyers Socialni-sluzby
## 26        0.0654 bad         good  KUKY       Mestsky_urad_Vyzva_k_zaplaceni_na~
## 27        0.0620 good        bad   FrBo       153
## 28        0.0503 good        bad   KUKY       UOOUOsobniUdajePuvodne
## 29        0.0382 bad         good  FrBo       red_Les - co smíme a co je zakázá~
## 30        0.0380 bad         good  KUKY       důchod-dorovnávací přídavek_1298-~
## # i 6 more rows
## Highest-deviating documents names:
## [1] "orig_Jak uspořádat shromáždění"
## [2] "orig_Kterých řízení se může váš spolek účastnit_FINAL"
## [3] "11_vizum_pred"
## [4] "33 Cdo 30_2024"
## [5] "orig_Jak namítat podjatost_final"
## [6] "orig_lhuty_v_jednani_s_urady_a_soudy"
## [7] "orig_Zastupitelstvo_o čem a jak rozhoduje"
## [8] "orig_Jak probíhá správní řízení"
```

```
## [9] "142"
```

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         0.768 Preprocessor1_Model1
## 2 roc_auc     binary         0.836 Preprocessor1_Model1
## 3 brier_class binary         0.164 Preprocessor1_Model1
```
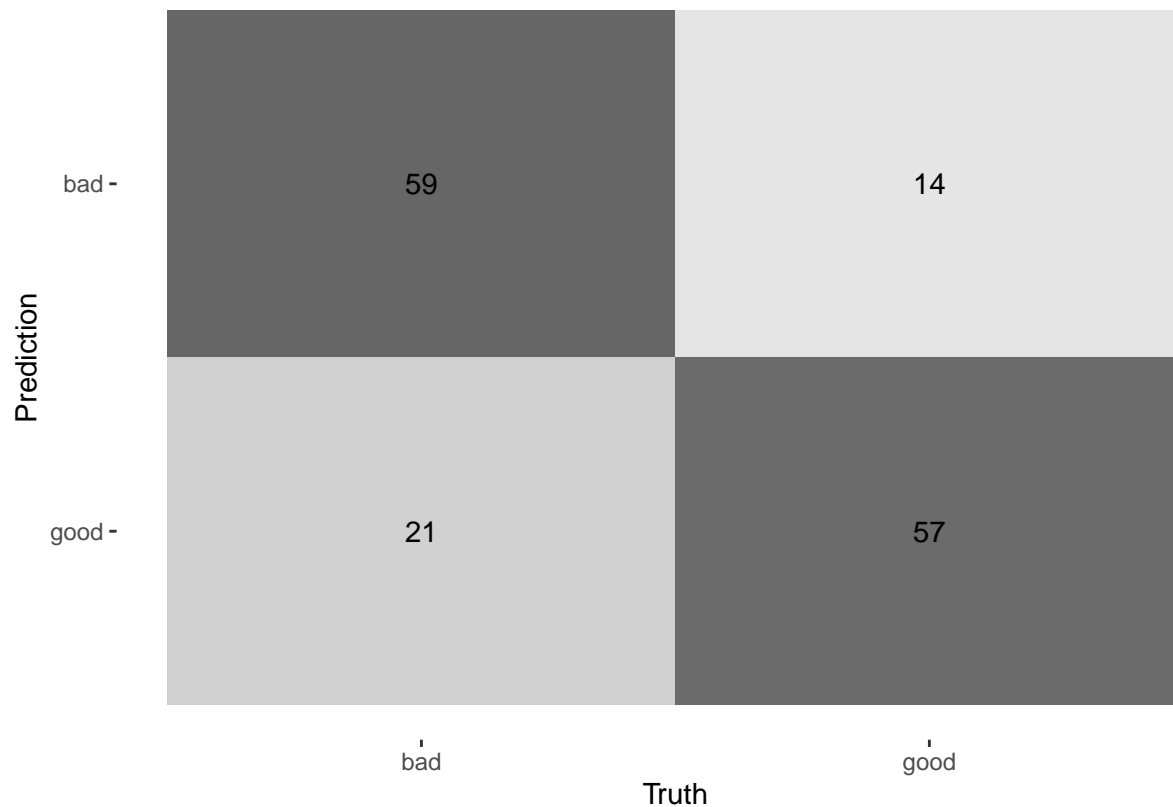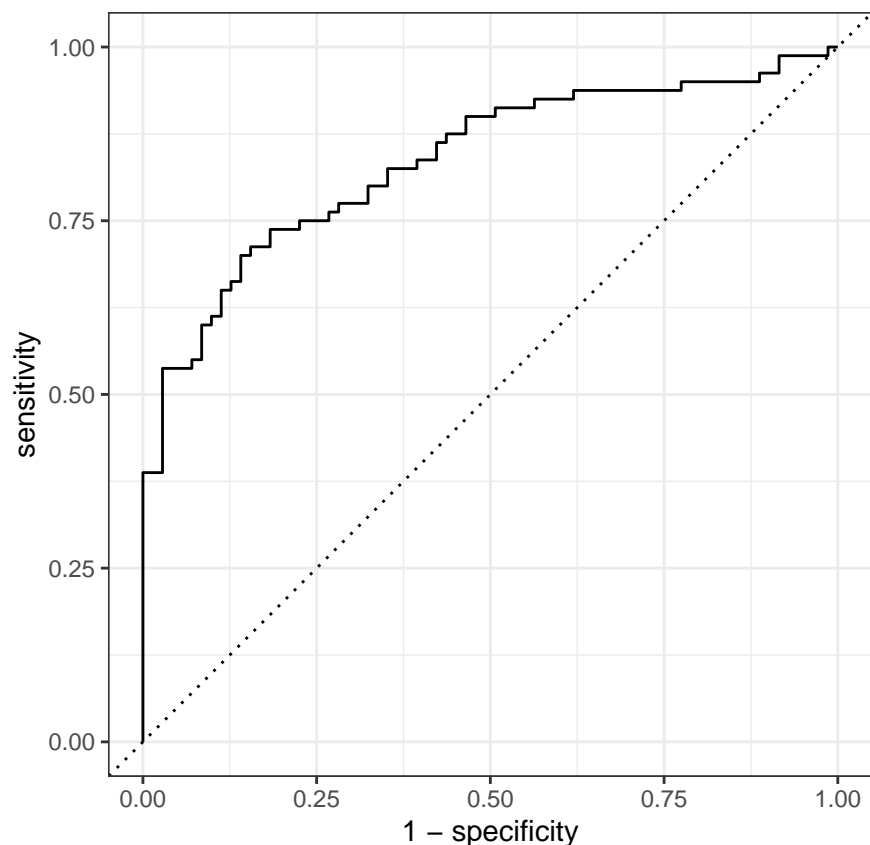
```
## Variable importance:
## # A tibble: 44 x 2
##    Variable                                                  Importance
##    <chr>                                                          <dbl>
##  1 activity                                                        18.2
##  2 verb_dist                                                       16.9
##  3 RuleTooFewVerbs.min_verb_frac                                   15.2
##  4 RuleTooManyNominalConstructions.max_allowable_nouns             13.6
##  5 RuleLongSentences.max_length                                    12.3
##  6 RulePredAtClauseBeginning.max_order                             10.7
##  7 ari                                                             9.64
##  8 gf                                                              9.28
##  9 smog                                                            7.74
## 10 atl                                                             7.20
## 11 RuleTooManyNegations.max_negation_frac                          6.42
## 12 fkgl                                                            6.10
## 13 mamr                                                            5.67
## 14 mattr                                                           5.54
## 15 RuleTooManyNominalConstructions.max_noun_frac                   5.53
## 16 maentropy                                                       5.39
## 17 cli                                                             4.90
## 18 RuleTooManyNominalConstructions.max_allowable_nouns.v           4.84
## 19 maentropy.v                                                     4.80
## 20 RulePredAtClauseBeginning.max_order.v                           4.77
## 21 RuleCaseRepetition.max_repetition_count.v                       4.69
## 22 RuleLongSentences.max_length.v                                  4.61
## 23 RuleCaseRepetition.max_repetition_frac.v                        4.53
```

```
## 24 entropy                                              4.49
## 25 RulePredObjDistance.max_distance                     4.29
## 26 RuleCaseRepetition.max_repetition_frac               4.09
## 27 RulePredSubjDistance.max_distance                    3.99
## 28 mattr.v                                              3.87
## 29 RuleTooManyNegations.max_allowable_negations         3.74
## 30 RuleTooManyNegations.max_allowable_negations.v       3.72
## 31 ttr                                                  3.58
## 32 RuleTooManyNegations.max_negation_frac.v             3.57
## 33 RuleInfVerbDistance.max_distance                     3.50
## 34 RuleTooManyNominalConstructions.max_noun_frac.v      3.38
## 35 RuleCaseRepetition.max_repetition_count              3.38
## 36 RuleInfVerbDistance.max_distance.v                   3.30
## 37 RulePredObjDistance.max_distance.v                   3.28
## 38 fre                                                  3.25
## 39 RulePredSubjDistance.max_distance.v                  3.23
## 40 RuleMultiPartVerbs.max_distance                      3.22
## 41 RuleDoubleAdpos.max_allowable_distance.v             3.06
## 42 RuleMultiPartVerbs.max_distance.v                    3.05
## 43 RuleDoubleAdpos.max_allowable_distance               2.75
## 44 hpoint                                               2.40
```

```
rf_iac_devs <- lfit_rf_iac %>% get_mismatch_details(data)
```
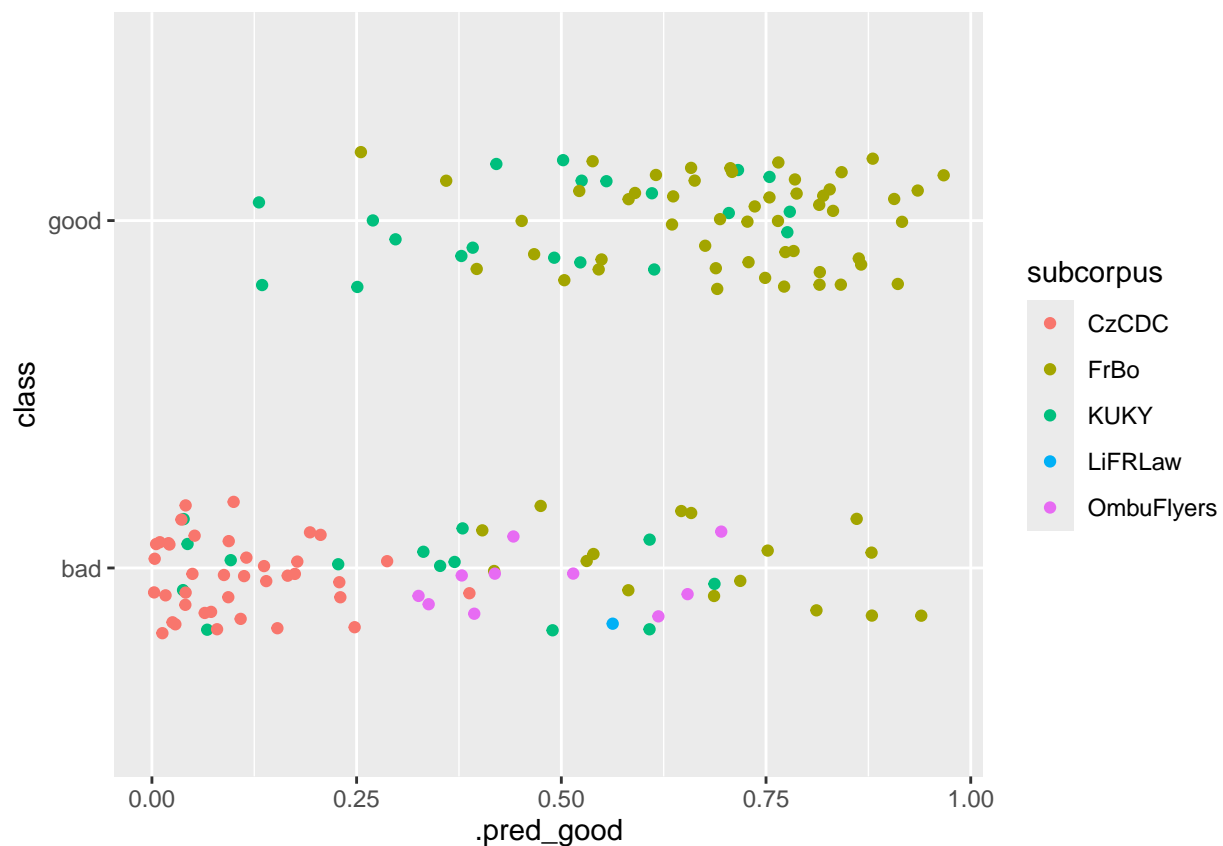


```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##            class
```

```
## .pred_class bad good
##        bad  39    0
##        good  0    0
##
## , , subcorpus = FrBo
##
##            class
## .pred_class bad good
##        bad   3    5
##        good 13   46
##
## , , subcorpus = KUKY
##
##            class
## .pred_class bad good
##        bad  11    9
##        good  3   11
##
## , , subcorpus = LiFRLaw
##
##            class
## .pred_class bad good
##        bad   0    0
##        good  1    0
##
## , , subcorpus = OmbuFlyers
##
##            class
## .pred_class bad good
##        bad   6    0
##        good  4    0
##
##
## Greatest deviations:
## # A tibble: 35 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
## 1          0.440 good        bad   FrBo       orig_Jak uspořádat shromáždění
## 2          0.379 good        bad   FrBo       orig_Kterých řízení se může váš s~
## 3          0.379 good        bad   FrBo       orig_Jak probíhá správní řízení
## 4          0.369 bad         good  KUKY       33 Cdo 30_2024
## 5          0.365 bad         good  KUKY       11_vizum_pred
## 6          0.361 good        bad   FrBo       orig_Jak namítat podjatost_final
## 7          0.312 good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
## 8          0.252 good        bad   FrBo       142
## 9          0.249 bad         good  KUKY       Mestsky_urad_usneseni_-_slouceni_~
## 10         0.245 bad         good  FrBo       red_Certifikáty autorizovaných in~
## 11         0.230 bad         good  KUKY       Mestsky_urad_Vyzva_k_zaplaceni_na~
## 12         0.219 good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
## 13         0.202 bad         good  KUKY       1A_dokument_puvodni_ustanoven_zas~
## 14         0.195 good        bad   OmbuFlyers Studny
## 15         0.187 good        bad   KUKY       sluzebni_hodnoceni_puvodni
## 16         0.187 good        bad   FrBo       orig_provokace_korupcniho_jednani
## 17         0.159 good        bad   FrBo       orig_Jak zajistit měření hluku
```

```
## 18            0.154  good         bad     OmbuFlyers Soudni-poplatky
## 19            0.147  good         bad     FrBo       64
## 20            0.141  bad          good    FrBo       1
## 21            0.122  bad          good    KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 22            0.119  good         bad     OmbuFlyers Bydlení - zajištění bydlení - obe~
## 23            0.108  bad          good    KUKY       Reakce_na_dopis_rev
## 24            0.108  good         bad     KUKY       016_Obcane-EU
## 25            0.108  good         bad     KUKY       7-Co-1752-2016-Vyber-judikatury
## 26            0.103  bad          good    FrBo       red_Les - co smíme a co je zakázá~
## 27            0.0819 good         bad     FrBo       orig_Pozemkové úpravy_pracovní ve~
## 28            0.0793 bad          good    KUKY       Reakce_na_dopis_pracovni
## 29            0.0627 good         bad     LiFRLaw    zastoupeni-3_orig
## 30            0.0484 bad          good    FrBo       red_Jak podat trestní oznámení
## # i 5 more rows
## Highest-deviating documents names:
## [1] "orig_Jak uspořádat shromáždění"
## [2] "orig_Kterých řízení se může váš spolek účastnit_FINAL"
## [3] "orig_Jak probíhá správní řízení"
## [4] "33 Cdo 30_2024"
## [5] "11_vizum_pred"
## [6] "orig_Jak namítat podjatost_final"
## [7] "orig_lhuty_v_jednani_s_urady_a_soudy"
## [8] "142"
```

Counts

```r
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         0.702 Preprocessor1_Model1
## 2 roc_auc     binary         0.812 Preprocessor1_Model1
## 3 brier_class binary         0.176 Preprocessor1_Model1
```
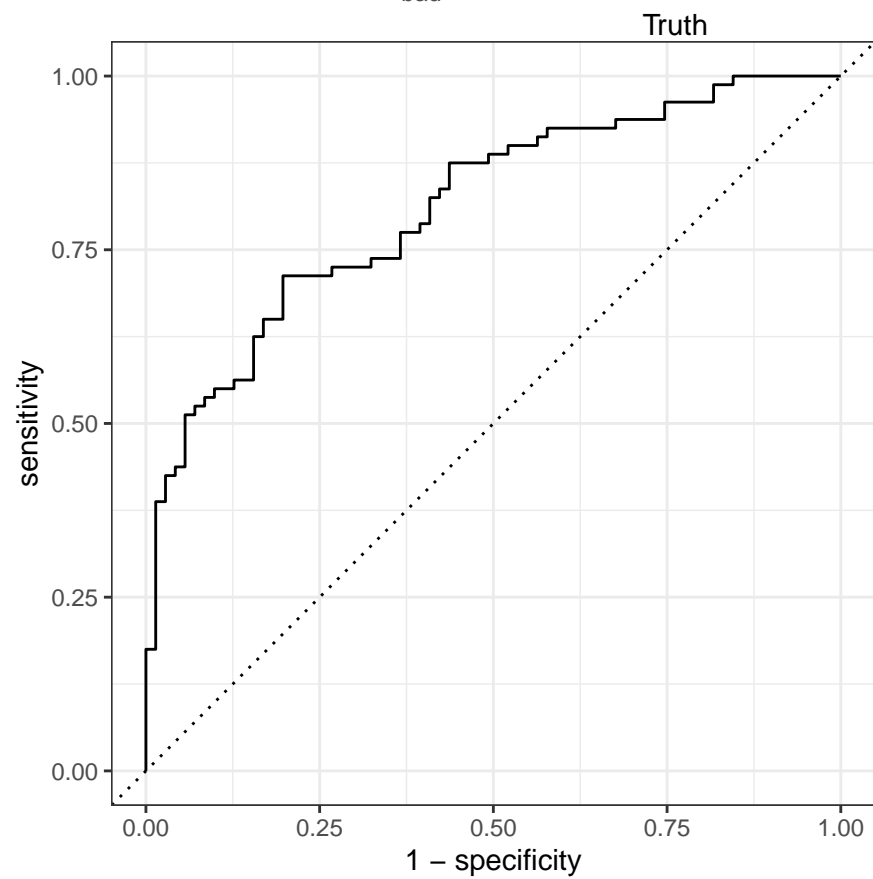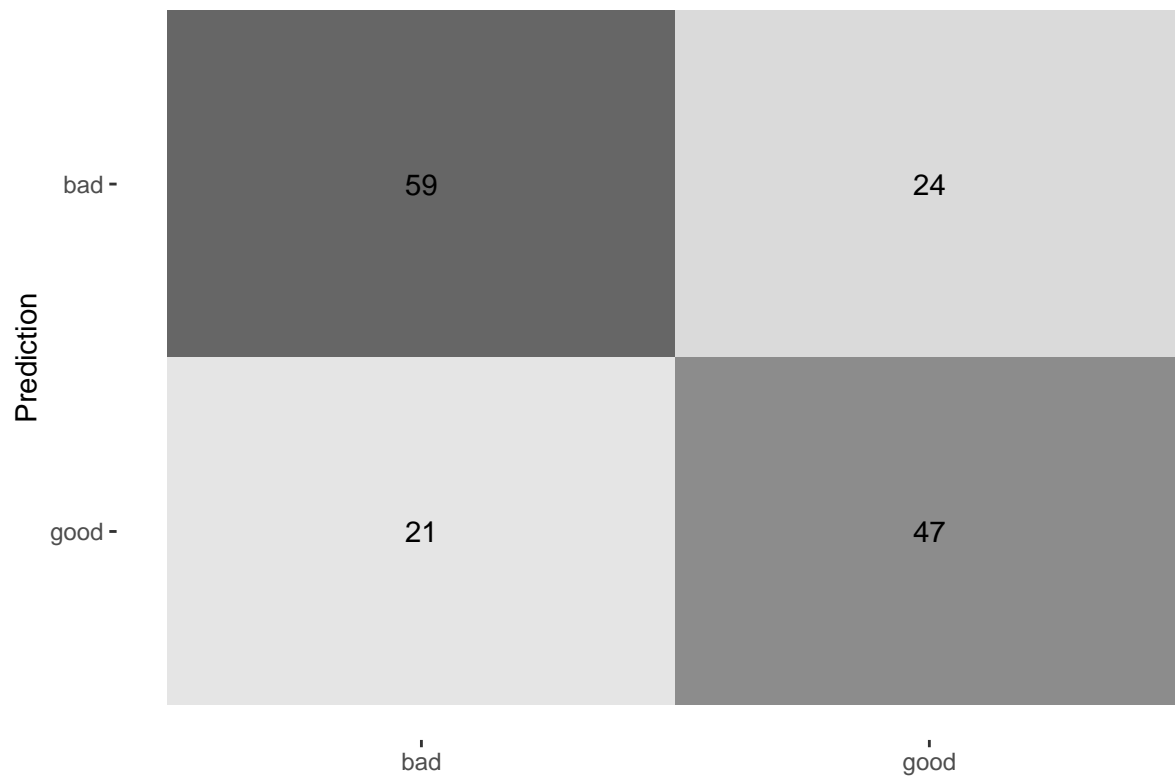
```
## Variable importance:
## # A tibble: 24 x 2
```

```
##    Variable                    Importance
##    <chr>                            <dbl>
##  1 RulePassive                       36.7
##  2 RuleMultiPartVerbs                25.1
##  3 RulePredSubjDistance              24.2
##  4 RuleLiteraryStyle                 23.9
##  5 RuleInfVerbDistance               18.6
##  6 RuleVerbalNouns                   14.0
##  7 num_hapax                         11.4
##  8 RuleAbstractNouns                 10.4
##  9 RulePredObjDistance               10.2
## 10 RuleTooLongExpressions             9.44
## 11 RuleDoubleAdpos                    9.35
## 12 RuleGPwordorder                    8.59
## 13 RuleWeakMeaningWords               7.79
## 14 RuleAnaphoricReferences            7.49
## 15 RuleReflexivePassWithAnimSubj      6.86
## 16 RuleGPdeverbsubj                   4.32
## 17 RuleGPpatinstr                     4.18
## 18 RuleGPdeverbaddr                   3.68
## 19 RuleGPcoordovs                     3.48
## 20 RuleGPpatbenperson                 2.83
## 21 RuleRelativisticExpressions        2.78
## 22 RuleConfirmationExpressions        2.16
## 23 RuleGPadjective                    0.791
## 24 RuleRedundantExpressions           0.394
```

```r
rf_counts_devs <- lfit_rf_counts %>% get_mismatch_details(data)
```
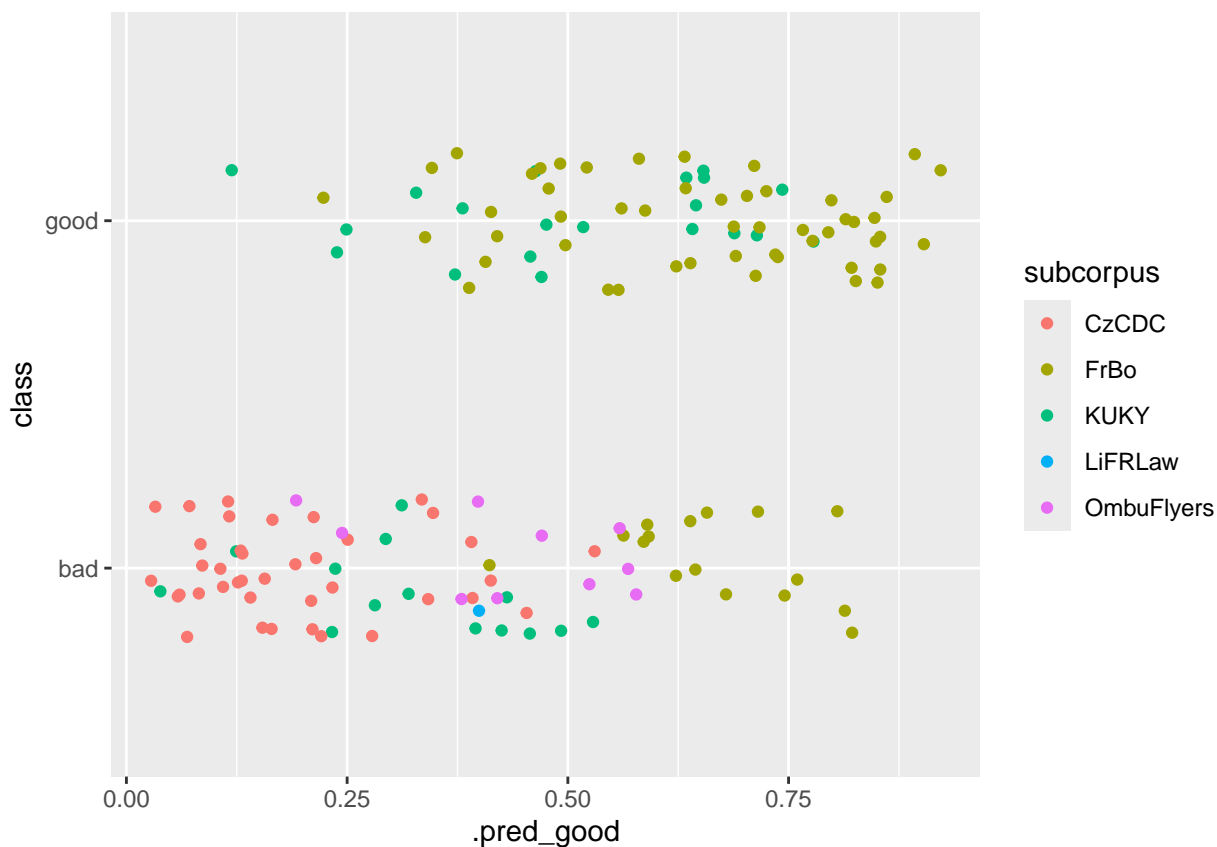
```
## Confusion matrices by subcorpora:
## , , subcorpus = CzCDC
##
##           class
## .pred_class bad good
##       bad  38    0
##       good  1    0
##
## , , subcorpus = FrBo
##
##           class
## .pred_class bad good
##       bad   1   14
##       good 15   37
##
## , , subcorpus = KUKY
##
##           class
## .pred_class bad good
##       bad  13   10
##       good  1   10
##
## , , subcorpus = LiFRLaw
##
##           class
## .pred_class bad good
##       bad   1    0
```

```
##         good    0    0
##
## , , subcorpus = OmbuFlyers
##
##           class
## .pred_class bad good
##         bad    6    0
##         good   4    0
##
##
## Greatest deviations:
## # A tibble: 45 x 5
##    abs_deviation .pred_class class subcorpus  FileName
##            <dbl> <fct>       <fct> <chr>      <chr>
## 1          0.381 bad         good  KUKY       11_vizum_pred
## 2          0.322 good        bad   FrBo       orig_Zastupitelstvo_o čem a jak r~
## 3          0.314 good        bad   FrBo       64
## 4          0.305 good        bad   FrBo       orig_Jak uspořádat shromáždění
## 5          0.277 bad         good  FrBo       red_Certifikáty autorizovaných in~
## 6          0.262 bad         good  KUKY       2A_dokument_puvodni_vyzva_k_zapla~
## 7          0.260 good        bad   FrBo       orig_provokace_korupcniho_jednani
## 8          0.251 bad         good  KUKY       33 Cdo 30_2024
## 9          0.246 good        bad   FrBo       orig_lhuty_v_jednani_s_urady_a_so~
## 10         0.215 good        bad   FrBo       orig_Kterých řízení se může váš s~
## 11         0.179 good        bad   FrBo       orig_Jak probíhá správní řízení
## 12         0.172 bad         good  KUKY       857_2024_VOP
## 13         0.162 bad         good  FrBo       red_závazná stanoviska_aktualizov~
## 14         0.158 good        bad   FrBo       149
## 15         0.154 bad         good  FrBo       red_Jak podat trestní oznámení
## 16         0.144 good        bad   FrBo       orig_Jak namítat podjatost_final
## 17         0.139 good        bad   FrBo       orig_financovani_politickych_stran
## 18         0.128 bad         good  KUKY       6421_2023_VOP
## 19         0.126 bad         good  FrBo       24
## 20         0.122 good        bad   FrBo       orig_Změny v zákoně o EIA
## 21         0.119 bad         good  KUKY       Reakce_na_dopis_rev
## 22         0.112 bad         good  FrBo       red_10 významných práv účastníka ~
## 23         0.0932 bad        good  FrBo       148
## 24         0.0916 good       bad   FrBo       153
## 25         0.0900 good       bad   FrBo       orig_Jak zajistit měření hluku
## 26         0.0870 bad        good  FrBo       orig_Soustavné obtěžování hlukem ~
## 27         0.0859 good       bad   FrBo       orig_Pozemkové úpravy_pracovní ve~
## 28         0.0801 bad        good  FrBo       199
## 29         0.0774 good       bad   OmbuFlyers Spolecenstvi-vlastniku
## 30         0.0683 good       bad   OmbuFlyers Soudni-poplatky
## # i 15 more rows
## Highest-deviating documents names:
## [1] "11_vizum_pred"
## [2] "orig_Zastupitelstvo_o čem a jak rozhoduje"
## [3] "64"
## [4] "orig_Jak uspořádat shromáždění"
## [5] "red_Certifikáty autorizovaných inspektorů"
## [6] "2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni"
## [7] "orig_provokace_korupcniho_jednani"
## [8] "33 Cdo 30_2024"
```

Variable importances

```r
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))) %>%
    mutate(quantile = rank / n()) %>%
    select(rank, quantile, Variable, Importance)
}

importances <- full_join(
  prepare_vi_for_comparison(lfit_lasso_all),
  prepare_vi_for_comparison(lfit_lasso_notl),
  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
## atl                                                                2
## RuleLiteraryStyle                                                  3
## smog                                                               4
## RulePassive                                                        5
## maentropy                                                          6
## entropy                                                            7
## RuleAnaphoricReferences                                            8
## 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
## 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
## ari                                                          NA
## num_hapax                                                    NA
## ttr                                                          NA
## mattr                                                        NA
## mattr.v                                                      NA
## maentropy.v                                                  NA
## mamr                                                         NA
## verb_dist                                                    NA
## hpoint                                                       NA
## fre                                                          NA
## fkgl                                                         NA
## gf                                                           NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v        NA
##                                                    rank.lasso.notl
## activity                                                      1
## atl                                                           2
## RuleLiteraryStyle                                             3
## smog                                                          4
## RulePassive                                                   5
## maentropy                                                     6
## entropy                                                       7
## RuleAnaphoricReferences                                       8
## 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
## 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
## ari                                                            NA
## num_hapax                                                      NA
## ttr                                                            NA
## mattr                                                          NA
## mattr.v                                                        NA
## maentropy.v                                                    NA
## mamr                                                           NA
## verb_dist                                                      NA
## hpoint                                                         NA
## fre                                                            NA
## fkgl                                                           NA
## gf                                                             NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v          NA
##                                                   rank.lasso.iac
## activity                                                        1
## atl                                                             3
## RuleLiteraryStyle                                              NA
## smog                                                            6
```

114

```
## RulePassive                                                      NA
## maentropy                                                         2
## entropy                                                           5
## RuleAnaphoricReferences                                          NA
## 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                   4
## 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
## 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
## ari                                                                NA
## num_hapax                                                          NA
## ttr                                                                NA
## mattr                                                              NA
## mattr.v                                                            NA
## maentropy.v                                                        NA
## mamr                                                               NA
## verb_dist                                                          NA
## hpoint                                                             NA
## fre                                                                NA
## fkgl                                                               NA
## gf                                                                  8
## RuleTooManyNominalConstructions.max_allowable_nouns.v               7
##                                                    rank.lasso.counts
## activity                                                           NA
## atl                                                                NA
## RuleLiteraryStyle                                                   3
## smog                                                               NA
## RulePassive                                                         2
## maentropy                                                          NA
## entropy                                                            NA
## RuleAnaphoricReferences                                             4
## 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                                         1
## RuleConfirmationExpressions                                        NA
## RuleRedundantExpressions                                           NA
## RuleTooLongExpressions                                             NA
## RulePredSubjDistance                                                6
## RulePredSubjDistance.max_distance                                  NA
```

```
## RulePredSubjDistance.max_distance.v                               NA
## RulePredObjDistance                                               NA
## RulePredObjDistance.max_distance                                  NA
## RulePredObjDistance.max_distance.v                                NA
## RuleInfVerbDistance                                                8
## RuleInfVerbDistance.max_distance                                  NA
## RuleInfVerbDistance.max_distance.v                                NA
## RuleMultiPartVerbs                                                 5
## 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                                                        NA
## word_count                                                        NA
## syllab_count                                                      NA
## char_count                                                        NA
## cli                                                               NA
## ari                                                               NA
## num_hapax                                                         NA
## ttr                                                               NA
## mattr                                                             NA
## mattr.v                                                           NA
## maentropy.v                                                       NA
## mamr                                                              NA
## verb_dist                                                         NA
## hpoint                                                            NA
## fre                                                               NA
## fkgl                                                              NA
## gf                                                                NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v             NA
##                                                      rank.rf.all rank.rf.notl
## activity                                                       1            1
## atl                                                           12           11
## RuleLiteraryStyle                                             10           10
## smog                                                          11           12
## RulePassive                                                    8            8
## maentropy                                                     13           17
## entropy                                                       33           30
## RuleAnaphoricReferences                                       27           25
## RuleGPcoordovs                                                61           57
## RuleGPdeverbaddr                                              65           60
## RuleGPpatinstr                                                64           61
## RuleGPdeverbsubj                                              68           64
## RuleGPadjective                                               70           66
## RuleGPpatbenperson                                            67           63
## RuleGPwordorder                                               60           58
## RuleDoubleAdpos                                               39           46
## RuleDoubleAdpos.max_allowable_distance                        59           55
## RuleDoubleAdpos.max_allowable_distance.v                      56           51
## RuleReflexivePassWithAnimSubj                                 62           59
## RuleTooFewVerbs.min_verb_frac                                  2            3
```

```
## RuleTooManyNegations.max_negation_frac                              15      14
## RuleTooManyNegations.max_negation_frac.v                            41      43
## RuleTooManyNegations.max_allowable_negations                        44      40
## RuleTooManyNegations.max_allowable_negations.v                      32      37
## RuleTooManyNominalConstructions.max_noun_frac                       20      21
## RuleTooManyNominalConstructions.max_noun_frac.v                     45      49
## RuleTooManyNominalConstructions.max_allowable_nouns                  4       5
## RuleCaseRepetition.max_repetition_count                             38      35
## RuleCaseRepetition.max_repetition_count.v                           25      28
## RuleCaseRepetition.max_repetition_frac                              29      32
## RuleCaseRepetition.max_repetition_frac.v                            30      29
## RuleWeakMeaningWords                                                57      44
## RuleAbstractNouns                                                   47      53
## RuleRelativisticExpressions                                         66      62
## RuleConfirmationExpressions                                         69      65
## RuleRedundantExpressions                                            71      67
## RuleTooLongExpressions                                              19      16
## RulePredSubjDistance                                                24      23
## RulePredSubjDistance.max_distance                                   26      26
## RulePredSubjDistance.max_distance.v                                 34      34
## RulePredObjDistance                                                 42      52
## RulePredObjDistance.max_distance                                    37      36
## RulePredObjDistance.max_distance.v                                  52      45
## RuleInfVerbDistance                                                 40      39
## RuleInfVerbDistance.max_distance                                    43      41
## RuleInfVerbDistance.max_distance.v                                  49      42
## RuleMultiPartVerbs                                                  21      20
## RuleMultiPartVerbs.max_distance                                     46      50
## RuleMultiPartVerbs.max_distance.v                                   50      54
## RuleLongSentences.max_length                                         5       4
## RuleLongSentences.max_length.v                                      31      24
## RulePredAtClauseBeginning.max_order                                  7       6
## RulePredAtClauseBeginning.max_order.v                               17      19
## RuleVerbalNouns                                                     22      18
## sent_count                                                          54      NA
## word_count                                                          55      NA
## syllab_count                                                        58      NA
## char_count                                                          51      NA
## cli                                                                 23      27
## ari                                                                  6       7
## num_hapax                                                           48      38
## ttr                                                                 53      47
## mattr                                                               18      22
## mattr.v                                                             35      33
## maentropy.v                                                         28      31
## mamr                                                                16      15
## verb_dist                                                            3       2
## hpoint                                                              63      56
## fre                                                                 36      48
## fkgl                                                                14      13
## gf                                                                   9       9
## RuleTooManyNominalConstructions.max_allowable_nouns.v               NA      NA
##                                                             rank.rf.iac
## activity                                                              1
```

118

```
## atl                                                          10
## RuleLiteraryStyle                                            NA
## smog                                                          9
## RulePassive                                                  NA
## maentropy                                                    16
## entropy                                                      24
## RuleAnaphoricReferences                                      NA
## RuleGPcoordovs                                               NA
## RuleGPdeverbaddr                                             NA
## RuleGPpatinstr                                               NA
## RuleGPdeverbsubj                                             NA
## RuleGPadjective                                              NA
## RuleGPpatbenperson                                           NA
## RuleGPwordorder                                              NA
## RuleDoubleAdpos                                              NA
## RuleDoubleAdpos.max_allowable_distance                       43
## RuleDoubleAdpos.max_allowable_distance.v                     41
## RuleReflexivePassWithAnimSubj                                NA
## RuleTooFewVerbs.min_verb_frac                                 3
## RuleTooManyNegations.max_negation_frac                       11
## RuleTooManyNegations.max_negation_frac.v                     32
## RuleTooManyNegations.max_allowable_negations                 29
## RuleTooManyNegations.max_allowable_negations.v               30
## RuleTooManyNominalConstructions.max_noun_frac                15
## RuleTooManyNominalConstructions.max_noun_frac.v              34
## RuleTooManyNominalConstructions.max_allowable_nouns           4
## RuleCaseRepetition.max_repetition_count                      35
## RuleCaseRepetition.max_repetition_count.v                    21
## RuleCaseRepetition.max_repetition_frac                       26
## RuleCaseRepetition.max_repetition_frac.v                     23
## RuleWeakMeaningWords                                         NA
## RuleAbstractNouns                                            NA
## RuleRelativisticExpressions                                  NA
## RuleConfirmationExpressions                                  NA
## RuleRedundantExpressions                                     NA
## RuleTooLongExpressions                                       NA
## RulePredSubjDistance                                         NA
## RulePredSubjDistance.max_distance                            27
## RulePredSubjDistance.max_distance.v                          39
## RulePredObjDistance                                          NA
## RulePredObjDistance.max_distance                             25
## RulePredObjDistance.max_distance.v                           37
## RuleInfVerbDistance                                          NA
## RuleInfVerbDistance.max_distance                             33
## RuleInfVerbDistance.max_distance.v                           36
## RuleMultiPartVerbs                                           NA
## RuleMultiPartVerbs.max_distance                              40
## RuleMultiPartVerbs.max_distance.v                            42
## RuleLongSentences.max_length                                  5
## RuleLongSentences.max_length.v                               22
## RulePredAtClauseBeginning.max_order                           6
## RulePredAtClauseBeginning.max_order.v                        20
## RuleVerbalNouns                                              NA
## sent_count                                                   NA
```

```
## word_count                                                     NA
## syllab_count                                                   NA
## char_count                                                     NA
## cli                                                            17
## ari                                                             7
## num_hapax                                                      NA
## ttr                                                            31
## mattr                                                          14
## mattr.v                                                        28
## maentropy.v                                                    19
## mamr                                                           13
## verb_dist                                                       2
## hpoint                                                         44
## fre                                                            38
## fkgl                                                           12
## gf                                                              8
## RuleTooManyNominalConstructions.max_allowable_nouns.v          18
##                                                     rank.rf.counts
## activity                                                        NA
## atl                                                             NA
## RuleLiteraryStyle                                                4
## smog                                                            NA
## RulePassive                                                      1
## maentropy                                                       NA
## entropy                                                         NA
## RuleAnaphoricReferences                                         14
## RuleGPcoordovs                                                  19
## RuleGPdeverbaddr                                                18
## RuleGPpatinstr                                                  17
## RuleGPdeverbsubj                                                16
## RuleGPadjective                                                 23
## RuleGPpatbenperson                                              20
## RuleGPwordorder                                                 12
## RuleDoubleAdpos                                                 11
## RuleDoubleAdpos.max_allowable_distance                          NA
## RuleDoubleAdpos.max_allowable_distance.v                        NA
## RuleReflexivePassWithAnimSubj                                   15
## 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                                            13
## RuleAbstractNouns                                                8
## RuleRelativisticExpressions                                     21
## RuleConfirmationExpressions                                     22
## RuleRedundantExpressions                                        24
```

```
## RuleTooLongExpressions                                          10
## RulePredSubjDistance                                             3
## RulePredSubjDistance.max_distance                               NA
## RulePredSubjDistance.max_distance.v                             NA
## RulePredObjDistance                                              9
## RulePredObjDistance.max_distance                                NA
## RulePredObjDistance.max_distance.v                              NA
## RuleInfVerbDistance                                              5
## RuleInfVerbDistance.max_distance                                NA
## RuleInfVerbDistance.max_distance.v                              NA
## RuleMultiPartVerbs                                               2
## 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                                                  6
## sent_count                                                      NA
## word_count                                                      NA
## syllab_count                                                    NA
## char_count                                                      NA
## cli                                                             NA
## ari                                                             NA
## num_hapax                                                        7
## ttr                                                             NA
## mattr                                                           NA
## mattr.v                                                         NA
## maentropy.v                                                     NA
## mamr                                                            NA
## verb_dist                                                       NA
## hpoint                                                          NA
## fre                                                             NA
## fkgl                                                            NA
## gf                                                              NA
## RuleTooManyNominalConstructions.max_allowable_nouns.v           NA
```

```r
importances_ranked <- importances %>%
  mutate(
    mean_rank = rowMeans(
      select(importances, starts_with("rank")),
      na.rm = TRUE
    ),
    mean_quantile = rowMeans(
      select(importances, starts_with("quantile")),
      na.rm = TRUE
    ),
    general_omissions = rowSums(
      select(importances, starts_with("Importance") & (ends_with("all") | ends_with("notl"))) == 0,
      na.rm = TRUE
    ),
    specialized_omissions = rowSums(
      select(importances, starts_with("Importance") & (ends_with("iac") | ends_with("counts"))) == 0,
      na.rm = TRUE
```

```
    ),
    no_of_irrelevance = rowSums(
      select(importances, starts_with("rank")) %>% is.na()
    )
  ) %>%
  mutate(omissions = general_omissions + specialized_omissions)

# working with the means really isn't informative, because:
# - the means don't take predictors omitted by lassos into account
# - the "all" and "no TL" models tend to be the same, thus they essentially get double the weight
importances_ranked %>%
  select(Variable, general_omissions, specialized_omissions) %>%
  arrange(specialized_omissions, general_omissions) %>%
  print(n = 100)
```

```
## # A tibble: 72 x 3
##    Variable                             general_omissions specialized_omissions
##    <chr>                                            <dbl>                 <dbl>
##  1 activity                                             0                     0
##  2 atl                                                  0                     0
##  3 RuleLiteraryStyle                                    0                     0
##  4 smog                                                 0                     0
##  5 RulePassive                                          0                     0
##  6 maentropy                                            0                     0
##  7 entropy                                              0                     0
##  8 RuleAnaphoricReferences                              0                     0
##  9 RuleTooManyNominalConstructions.max_~                0                     0
## 10 sent_count                                           1                     0
## 11 word_count                                           1                     0
## 12 syllab_count                                         1                     0
## 13 char_count                                           1                     0
## 14 RuleTooManyNegations.max_allowable_n~                2                     0
## 15 RuleRelativisticExpressions                          2                     0
## 16 RulePredSubjDistance                                 2                     0
## 17 RuleInfVerbDistance                                  2                     0
## 18 RuleMultiPartVerbs                                   2                     0
## 19 RuleVerbalNouns                                      2                     0
## 20 gf                                                   2                     0
## 21 RuleGPcoordovs                                       2                     1
## 22 RuleGPdeverbaddr                                     2                     1
## 23 RuleGPpatinstr                                       2                     1
## 24 RuleGPdeverbsubj                                     2                     1
## 25 RuleGPadjective                                      2                     1
## 26 RuleGPpatbenperson                                   2                     1
## 27 RuleGPwordorder                                      2                     1
## 28 RuleDoubleAdpos                                      2                     1
## 29 RuleDoubleAdpos.max_allowable_distan~                2                     1
## 30 RuleDoubleAdpos.max_allowable_distan~                2                     1
## 31 RuleReflexivePassWithAnimSubj                        2                     1
## 32 RuleTooFewVerbs.min_verb_frac                        2                     1
## 33 RuleTooManyNegations.max_negation_fr~                2                     1
## 34 RuleTooManyNegations.max_negation_fr~                2                     1
## 35 RuleTooManyNegations.max_allowable_n~                2                     1
## 36 RuleTooManyNominalConstructions.max_~                2                     1
```

```
## 37 RuleTooManyNominalConstructions.max_~               2               1
## 38 RuleTooManyNominalConstructions.max_~               2               1
## 39 RuleCaseRepetition.max_repetition_co~               2               1
## 40 RuleCaseRepetition.max_repetition_co~               2               1
## 41 RuleCaseRepetition.max_repetition_fr~               2               1
## 42 RuleCaseRepetition.max_repetition_fr~               2               1
## 43 RuleWeakMeaningWords                                2               1
## 44 RuleAbstractNouns                                   2               1
## 45 RuleConfirmationExpressions                         2               1
## 46 RuleRedundantExpressions                            2               1
## 47 RuleTooLongExpressions                              2               1
## 48 RulePredSubjDistance.max_distance                   2               1
## 49 RulePredSubjDistance.max_distance.v                 2               1
## 50 RulePredObjDistance                                 2               1
## 51 RulePredObjDistance.max_distance                    2               1
## 52 RulePredObjDistance.max_distance.v                  2               1
## 53 RuleInfVerbDistance.max_distance                    2               1
## 54 RuleInfVerbDistance.max_distance.v                  2               1
## 55 RuleMultiPartVerbs.max_distance                     2               1
## 56 RuleMultiPartVerbs.max_distance.v                   2               1
## 57 RuleLongSentences.max_length                        2               1
## 58 RuleLongSentences.max_length.v                      2               1
## 59 RulePredAtClauseBeginning.max_order                 2               1
## 60 RulePredAtClauseBeginning.max_order.v               2               1
## 61 cli                                                 2               1
## 62 ari                                                 2               1
## 63 num_hapax                                           2               1
## 64 ttr                                                 2               1
## 65 mattr                                               2               1
## 66 mattr.v                                             2               1
## 67 maentropy.v                                         2               1
## 68 mamr                                                2               1
## 69 verb_dist                                           2               1
## 70 hpoint                                              2               1
## 71 fre                                                 2               1
## 72 fkgl                                                2               1
```

```r
importances_ranked %>%
  select(Variable, mean_rank, mean_quantile, omissions) %>%
  arrange(omissions, mean_quantile) %>%
  print(n = 100)
```

```
## # A tibble: 72 x 4
##    Variable                              mean_rank mean_quantile omissions
##    <chr>                                     <dbl>         <dbl>     <dbl>
##  1 activity                                      1        0.0172         0
##  2 RulePassive                                4.83        0.0837         0
##  3 RuleLiteraryStyle                           5.5         0.111         0
##  4 atl                                        6.67         0.114         0
##  5 smog                                       7.67         0.132         0
##  6 maentropy                                    10         0.170         0
##  7 RuleTooManyNominalConstructions.max_allowa~ 12.5        0.284         0
##  8 RuleAnaphoricReferences                     14.3        0.289         0
##  9 entropy                                    17.7         0.296         0
## 10 char_count                                   51         0.718         1
```

```
## 11 sent_count                                        54      0.761      1
## 12 word_count                                        55      0.775      1
## 13 syllab_count                                      58      0.817      1
## 14 gf                                               8.5      0.156      2
## 15 RuleMultiPartVerbs                                12      0.221      2
## 16 RulePredSubjDistance                              14      0.264      2
## 17 RuleVerbalNouns                                 13.2      0.280      2
## 18 RuleInfVerbDistance                               23      0.422      2
## 19 RuleTooManyNegations.max_allowable_negatio~     25.8      0.444      2
## 20 RuleRelativisticExpressions                     37.5      0.693      2
## 21 verb_dist                                       2.33     0.0392      3
## 22 RuleTooFewVerbs.min_verb_frac                   2.67     0.0470      3
## 23 RuleTooManyNominalConstructions.max_allowa~     4.33     0.0740      3
## 24 RuleLongSentences.max_length                    4.67     0.0813      3
## 25 RulePredAtClauseBeginning.max_order             6.33      0.108      3
## 26 ari                                             6.67      0.116      3
## 27 fkgl                                              13      0.221      3
## 28 RuleTooManyNegations.max_negation_frac          13.3      0.223      3
## 29 mamr                                            14.7      0.248      3
## 30 mattr                                             18      0.300      3
## 31 RuleTooLongExpressions                            15      0.308      3
## 32 RuleTooManyNominalConstructions.max_noun_f~     18.7      0.312      3
## 33 RulePredAtClauseBeginning.max_order.v           18.7      0.326      3
## 34 cli                                             22.3      0.371      3
## 35 RuleCaseRepetition.max_repetition_count.v       24.7      0.416      3
## 36 maentropy.v                                       26      0.430      3
## 37 RuleLongSentences.max_length.v                  25.7      0.432      3
## 38 RulePredSubjDistance.max_distance               26.3      0.456      3
## 39 RuleCaseRepetition.max_repetition_frac.v        27.3      0.459      3
## 40 RuleCaseRepetition.max_repetition_frac            29      0.492      3
## 41 num_hapax                                         31      0.512      3
## 42 mattr.v                                           32      0.541      3
## 43 RulePredObjDistance.max_distance                32.7      0.542      3
## 44 RuleDoubleAdpos                                   32      0.565      3
## 45 RulePredObjDistance                             34.3      0.581      3
## 46 RuleAbstractNouns                                 36      0.595      3
## 47 RuleCaseRepetition.max_repetition_count           36      0.618      3
## 48 RulePredSubjDistance.max_distance.v             35.7      0.624      3
## 49 RuleTooManyNegations.max_allowable_negatio~     37.7      0.625      3
## 50 RuleTooManyNegations.max_negation_frac.v        38.7      0.649      3
## 51 RuleInfVerbDistance.max_distance                  39      0.656      3
## 52 RuleWeakMeaningWords                              38      0.667      3
## 53 fre                                             40.7      0.696      3
## 54 RuleInfVerbDistance.max_distance.v              42.3      0.712      3
## 55 RuleTooManyNominalConstructions.max_noun_f~     42.7      0.713      3
## 56 ttr                                             43.7      0.718      3
## 57 RuleGPwordorder                                 43.3      0.737      3
## 58 RulePredObjDistance.max_distance.v              44.7      0.748      3
## 59 RuleMultiPartVerbs.max_distance                 45.3      0.768      3
## 60 RuleReflexivePassWithAnimSubj                   45.3      0.793      3
## 61 RuleMultiPartVerbs.max_distance.v               48.7      0.822      3
## 62 RuleDoubleAdpos.max_allowable_distance.v        49.3      0.827      3
## 63 RuleGPcoordovs                                  45.7      0.834      3
## 64 RuleGPpatinstr                                  47.3      0.840      3
```

```
## 65 RuleGPdeverbaddr                          47.7        0.854         3
## 66 RuleGPdeverbsubj                          49.3        0.860         3
## 67 RuleDoubleAdpos.max_allowable_distance     52.3        0.876         3
## 68 RuleGPpatbenperson                         50          0.906         3
## 69 hpoint                                     54.3        0.908         3
## 70 RuleConfirmationExpressions                52          0.953         3
## 71 RuleGPadjective                            53          0.976         3
## 72 RuleRedundantExpressions                   54          1             3
```

Discussing the variables

We might keep predictors not thrown away by any of the more niche models for the analysis.

Of course, the selection of predictor combinations for the analysis is somewhat arbitrary. We might stick by the characteristics that one group is more focused on more universal properties of the text while the other on more rare of spontaneously-occurring phenomena.

The features not excluded by the model with the richer feature set are the most important ones. The absence of *_counts from the features proves that they are not needed for the recognition of (un)readable texts. This might however be compensated by using entropy for the prediction, as the "most important" features include both regular entropy and the moving average entropy.

Top RF-selected predictors seem not to be omitted completely by the lasso models; the top 20 to 25 ranks seem to overlap somewhat (even if the ordering of the predictors is different). Notable exceptions are:

- `fkgl` (14th for RF.all, but omitted 3 times)
- `cli` (27th for RF.all, but omitted 3 times)
- `mattr.v` (30th for RF.all, but omitted 3 times; maentropy.v omitted only 2 times though)

The RF-selected features start to get omitted more often from rank 38 (`RuleCaseRepetition.max_repetition_count`).

Highly deviating documents

Lasso

FrBo / orig_Jak uspořádat shromáždění
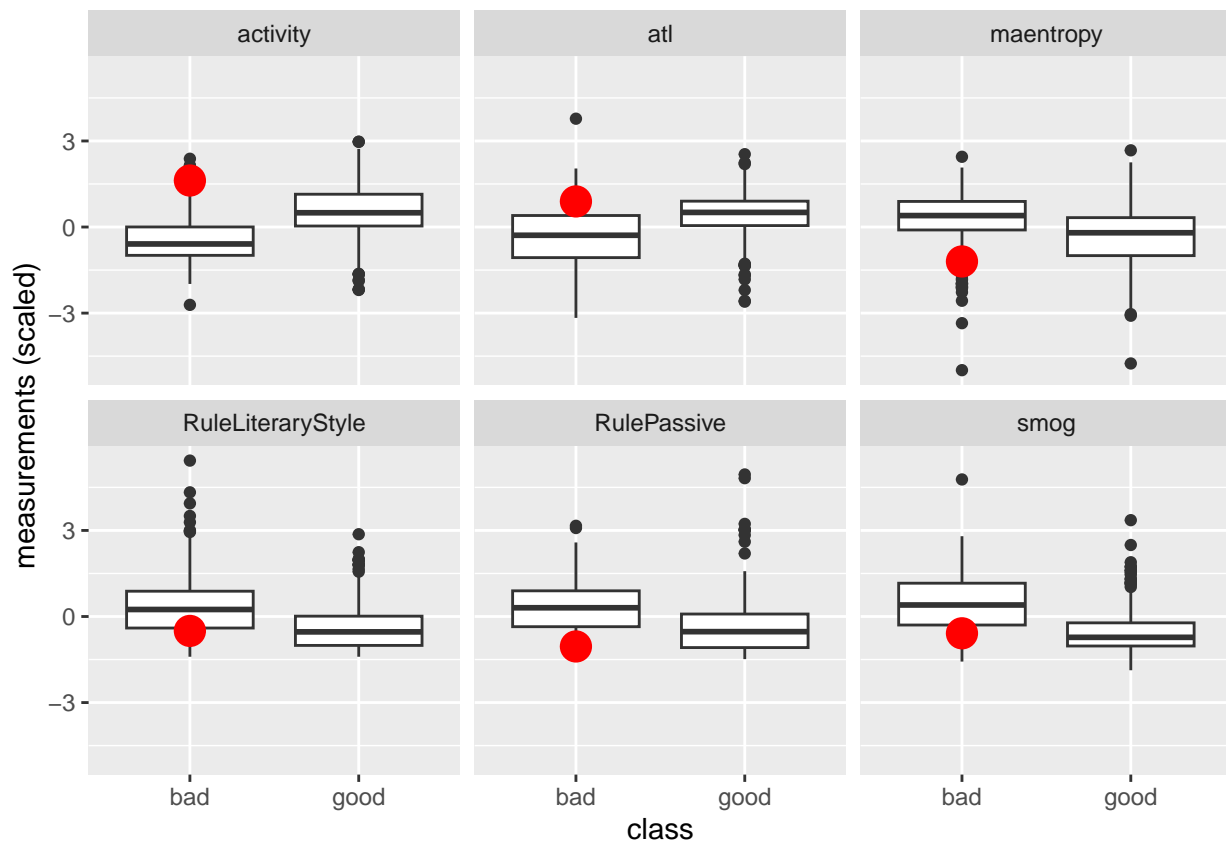
> truth: bad

```
plot_outlier("orig_Jak uspořádat shromáždění", lfit_lasso_all, data_clean)
```
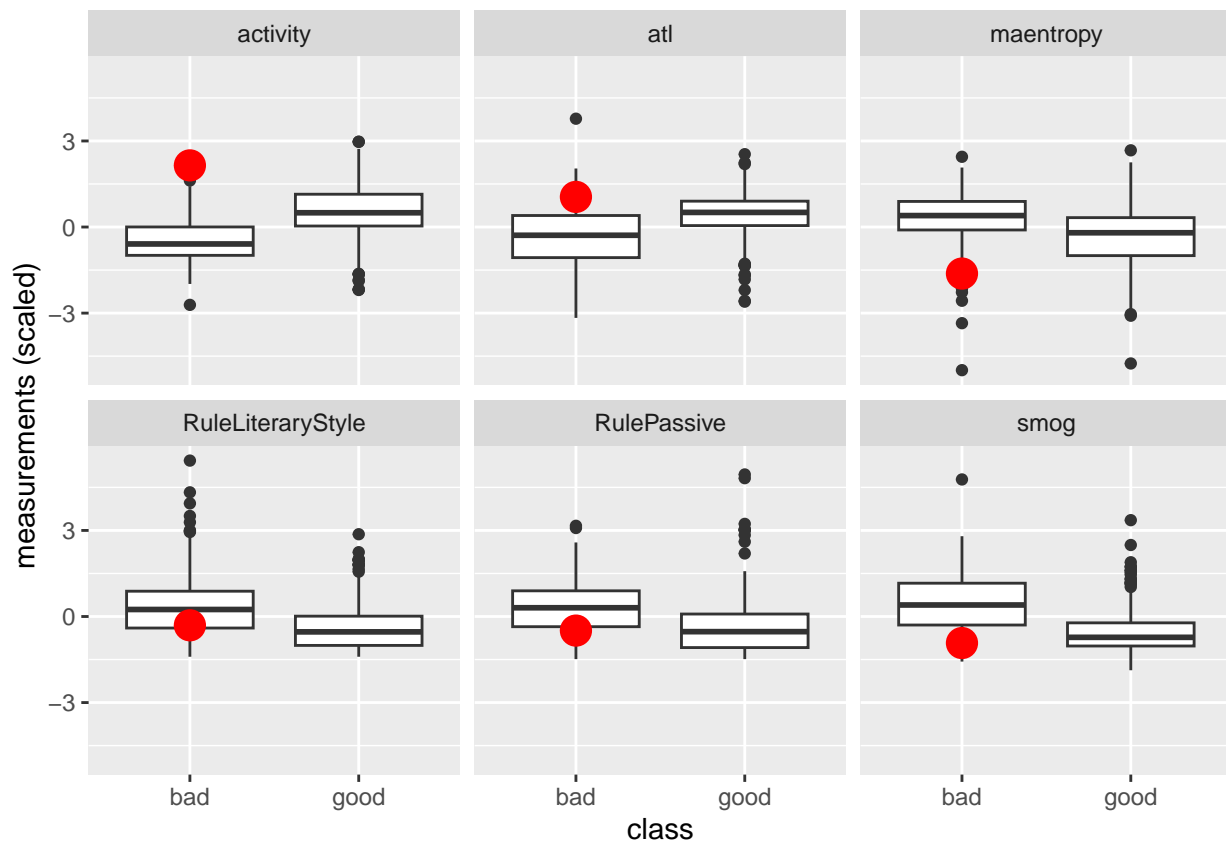
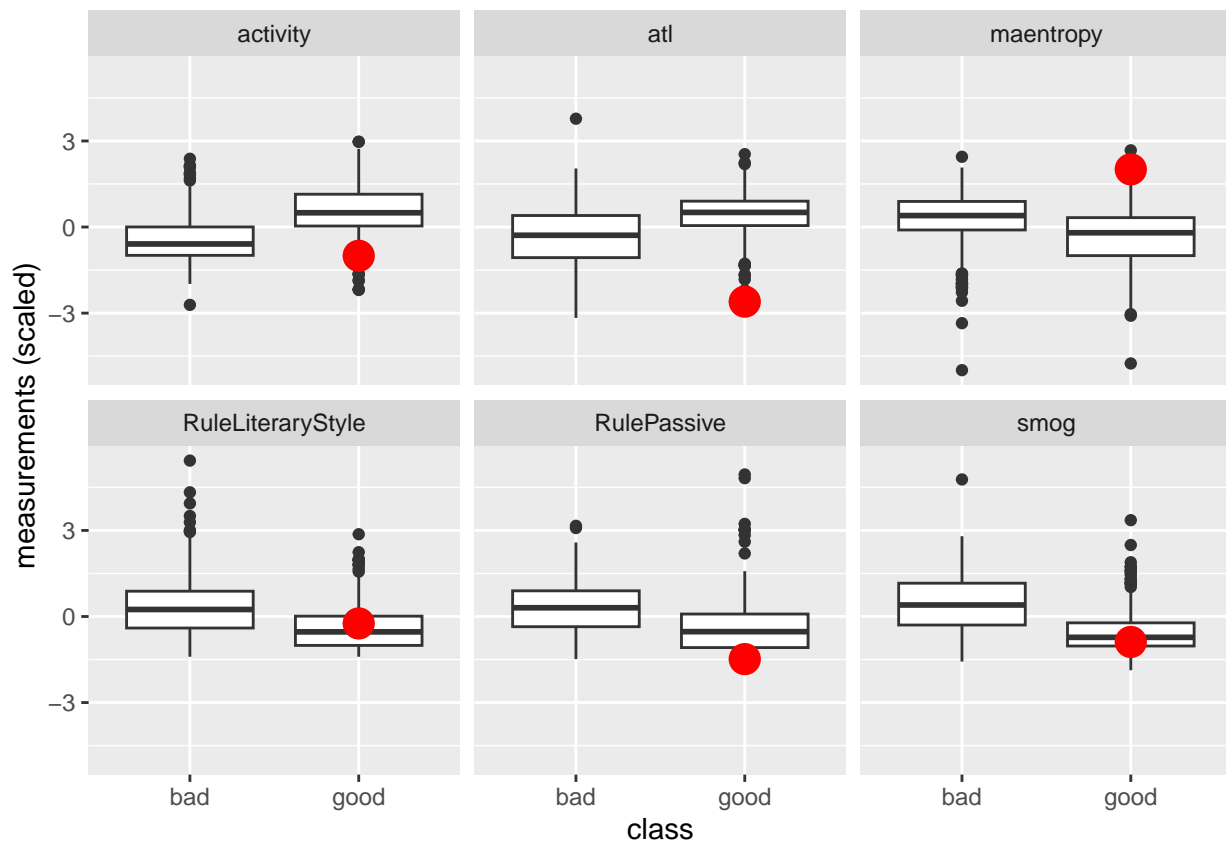**FrBo / orig_Zastupitelstvo_o čem a jak rozhoduje**

truth: bad

```
plot_outlier("orig_Zastupitelstvo_o čem a jak rozhoduje", lfit_lasso_all, data_clean)
```

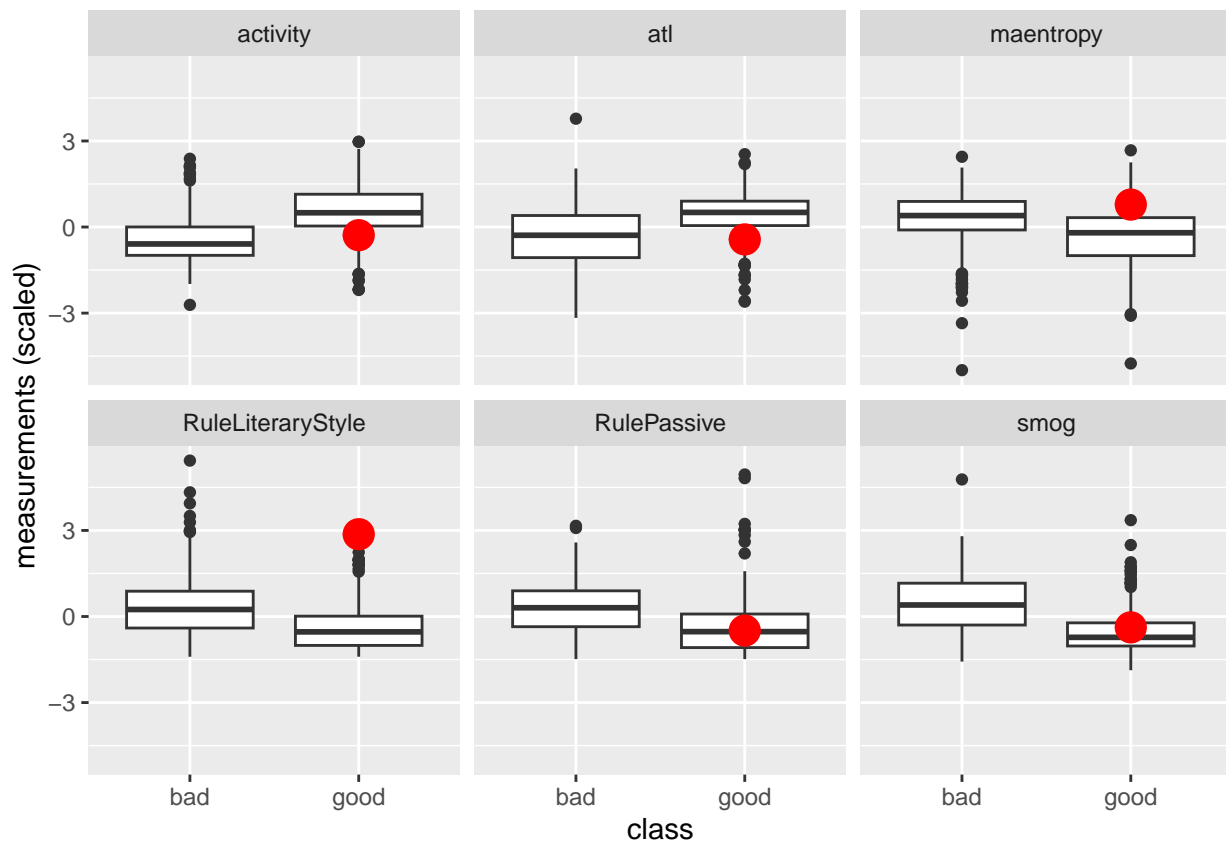**KUKY / Mestsky__urad__usneseni__-__slouceni__pred**

truth: good

```
plot_outlier("Mestsky_urad_usneseni_-_slouceni_pred", lfit_lasso_all, data_clean)
```

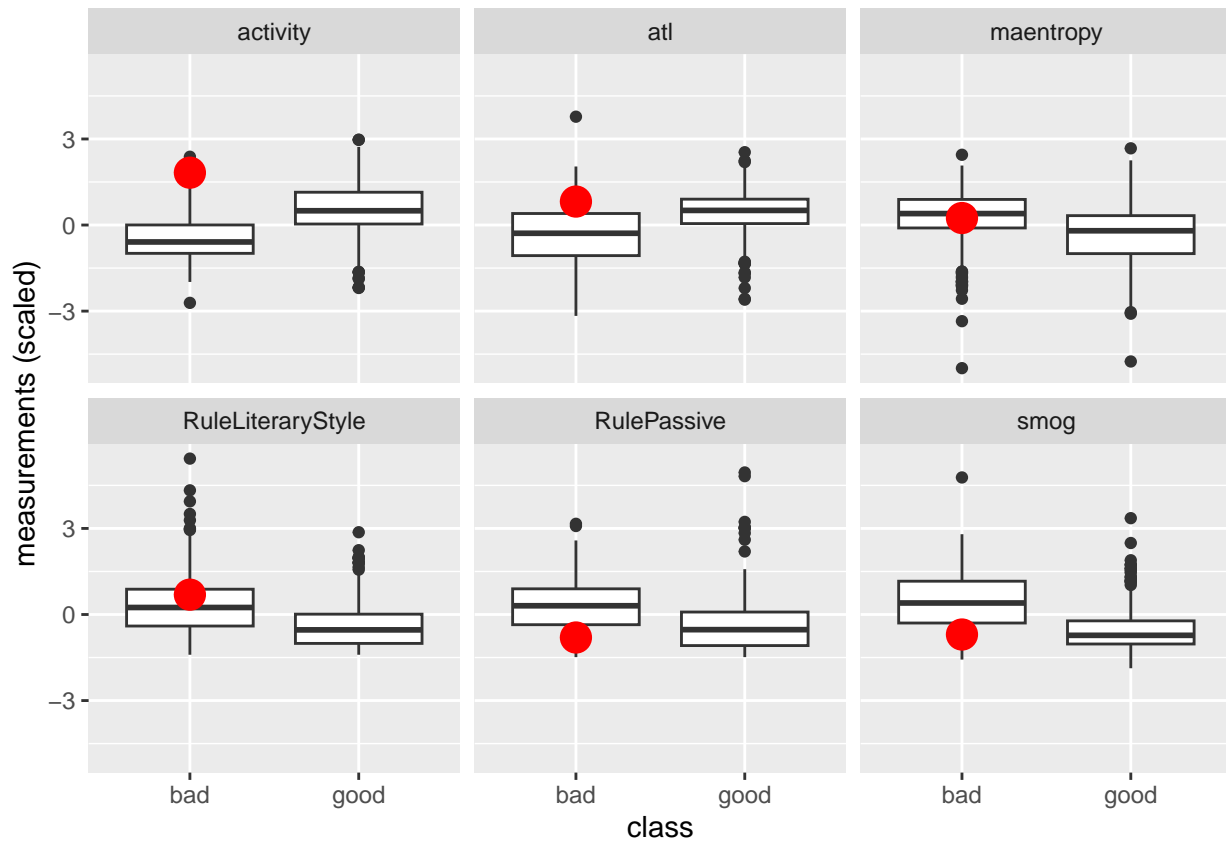**KUKY / 2A__dokument__puvodni__vyzva__k__zaplaceni__SOP__a__k__doplneni__kast__pouceni**

    truth: good

```
plot_outlier("2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni", lfit_lasso_all, dat
```

FrBo / orig_Jak namítat podjatost_final

    truth: bad

```
plot_outlier("orig_Jak namítat podjatost_final", lfit_lasso_all, data_clean)
```

KUKY / Mestsky_urad_Vyzva_k_zaplaceni_nakladu_rizeni_pred

truth: good

IAC-outlier

```
plot_outlier("Mestsky_urad_Vyzva_k_zaplaceni_nakladu_rizeni_pred", lfit_lasso_iac, data_clean)
```
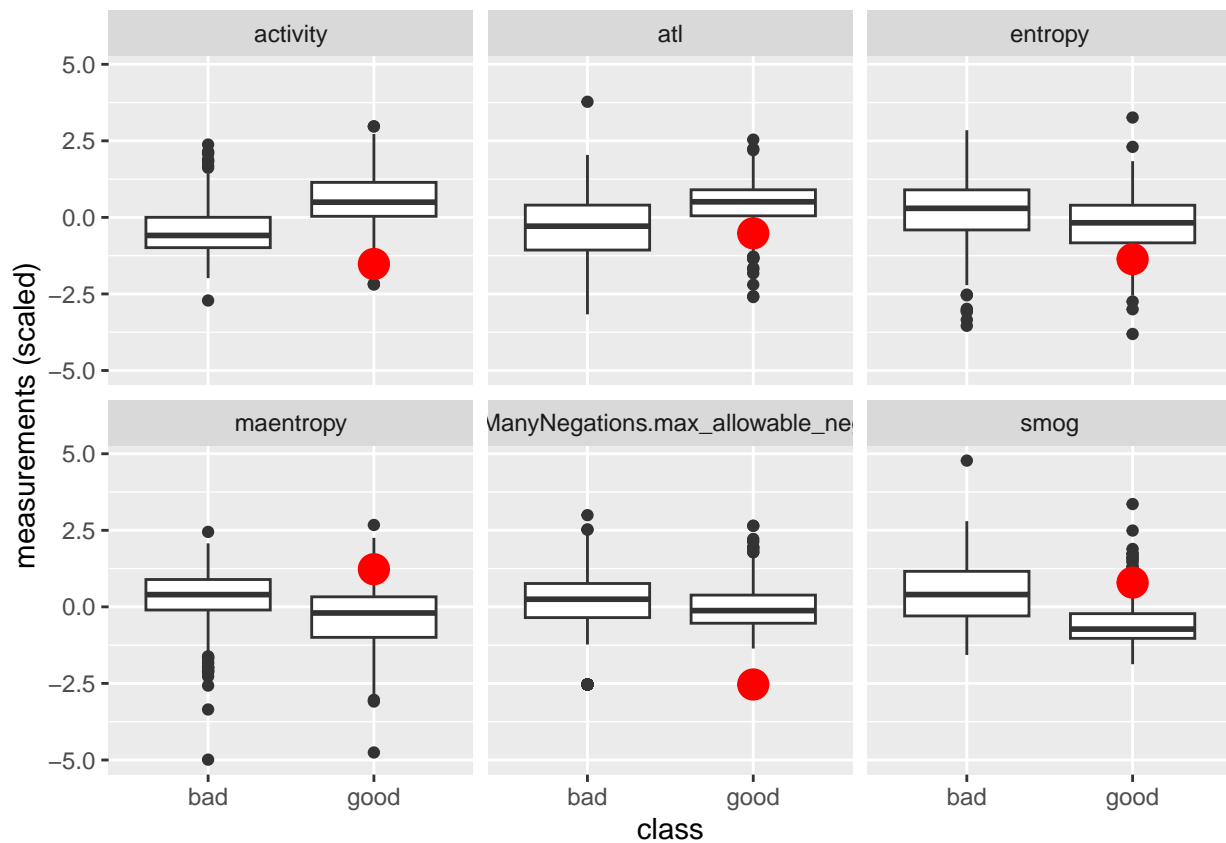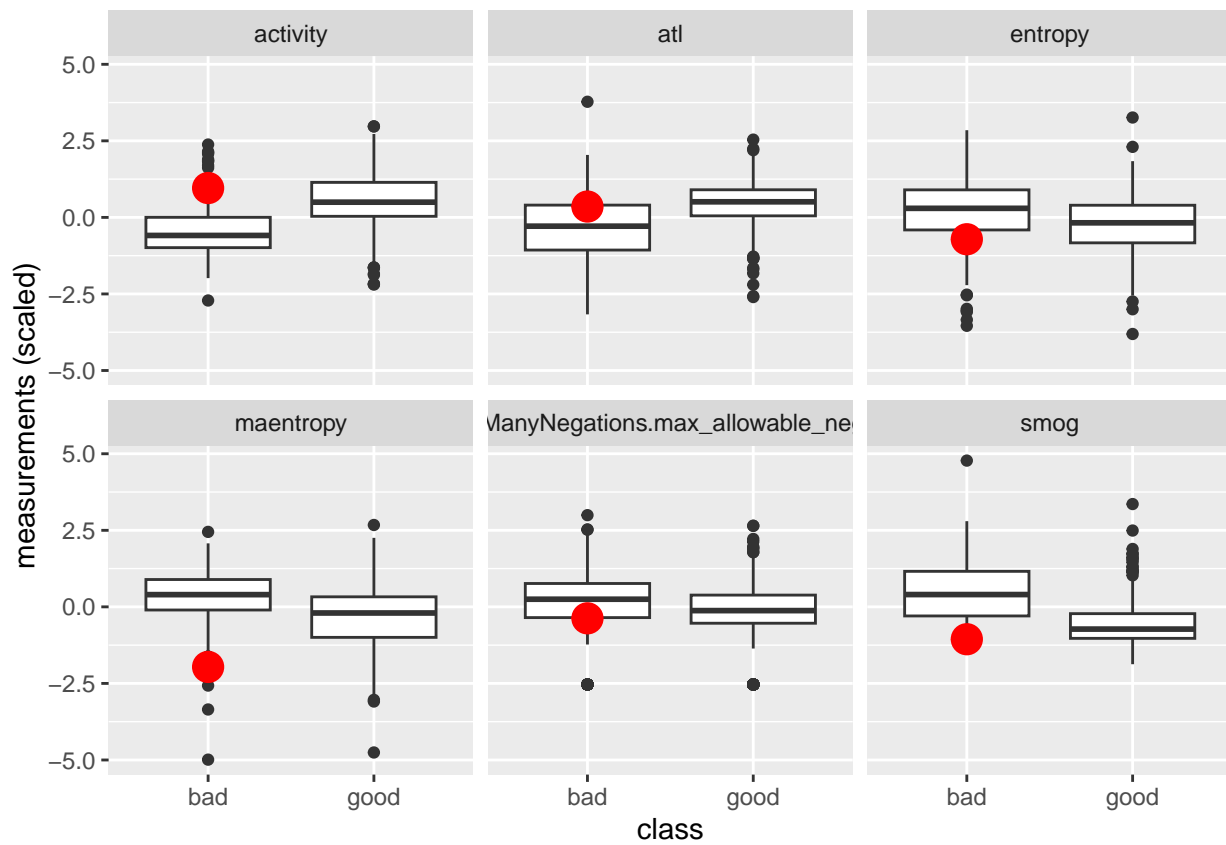
FrBo / orig_Jak probíhá správní řízení

truth: bad

IAC-outlier

```
plot_outlier("orig_Jak probíhá správní řízení", lfit_lasso_iac, data_clean)
```
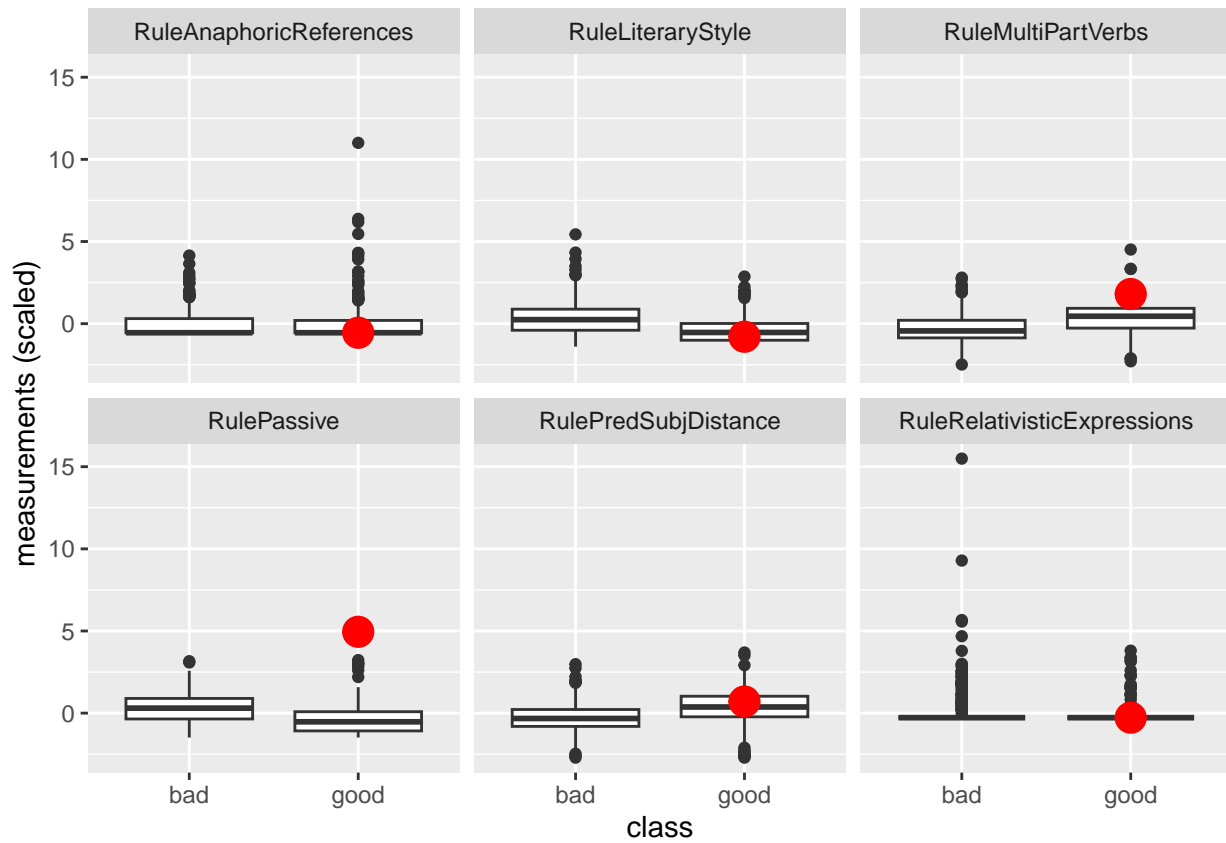
KUKY / Reakce__na__dopis__pracovni

truth: good

counts-outlier

```
plot_outlier("Reakce_na_dopis_pracovni", lfit_lasso_counts, data_clean)
```
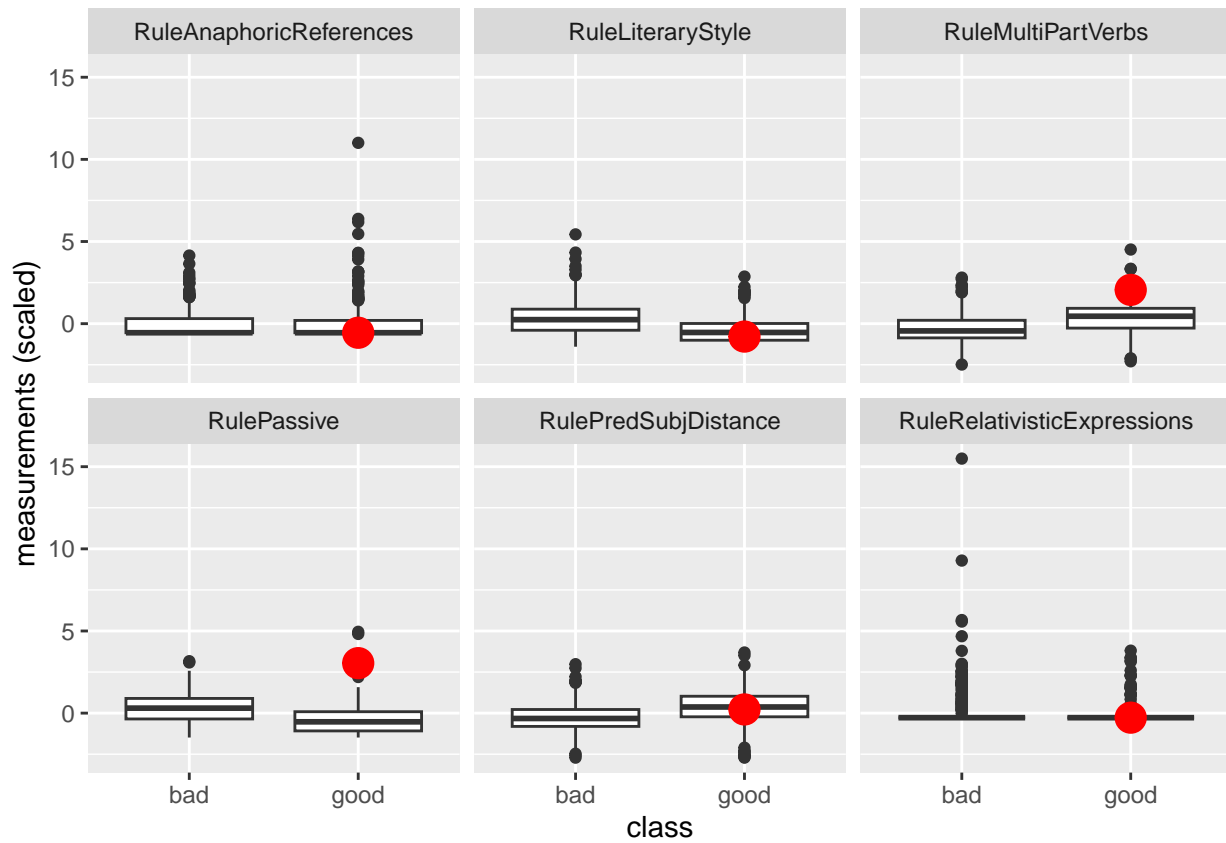
KUKY / Reakce__na__dopis__rev

truth: good

counts-outlier

```
plot_outlier("Reakce_na_dopis_rev", lfit_lasso_counts, data_clean)
```
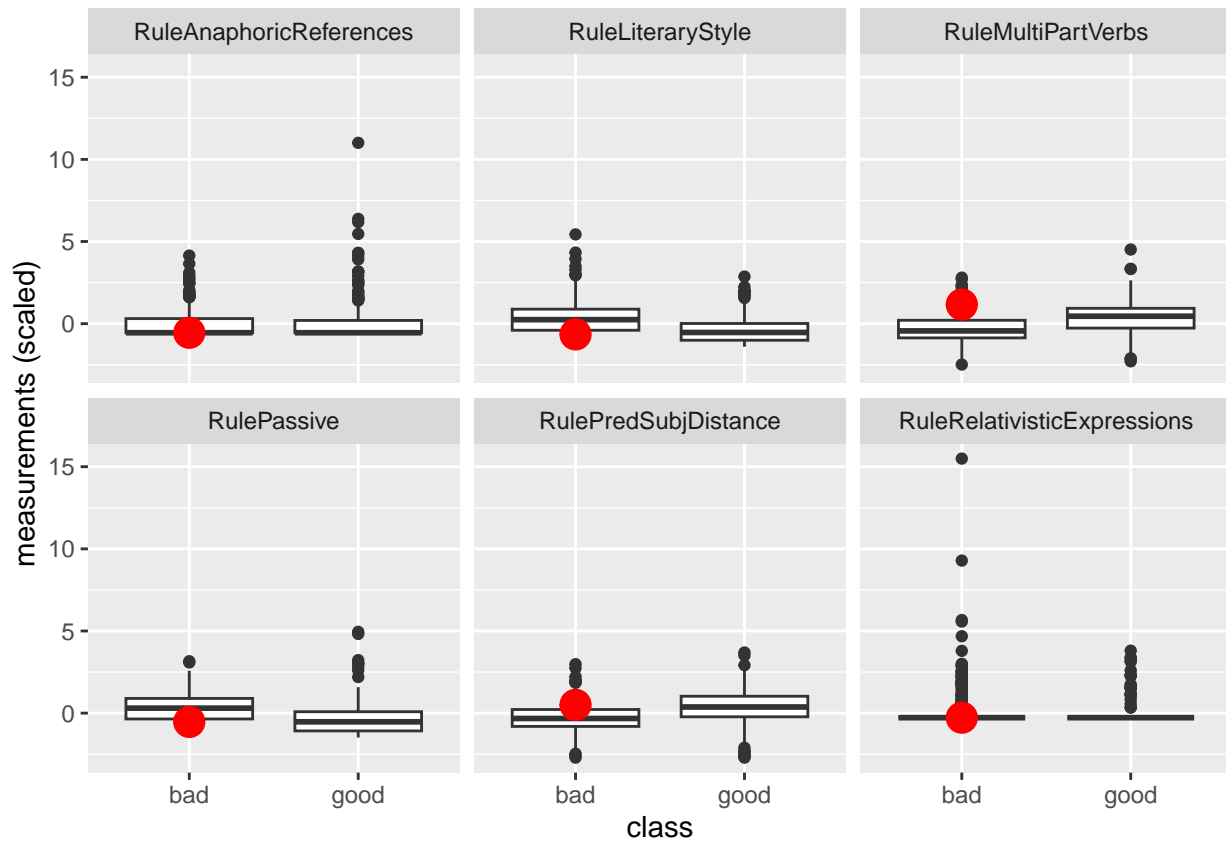
FrBo / orig__Kterých řízení se může váš spolek účastnit__FINAL

truth: bad

counts-outlier

```
plot_outlier("orig_Kterých řízení se může váš spolek účastnit_FINAL", lfit_lasso_counts, data_clean)
```

FrBo / red_Certifikáty autorizovaných inspektorů

truth: good

counts-outlier

```
plot_outlier("red_Certifikáty autorizovaných inspektorů", lfit_lasso_counts, data_clean)
```
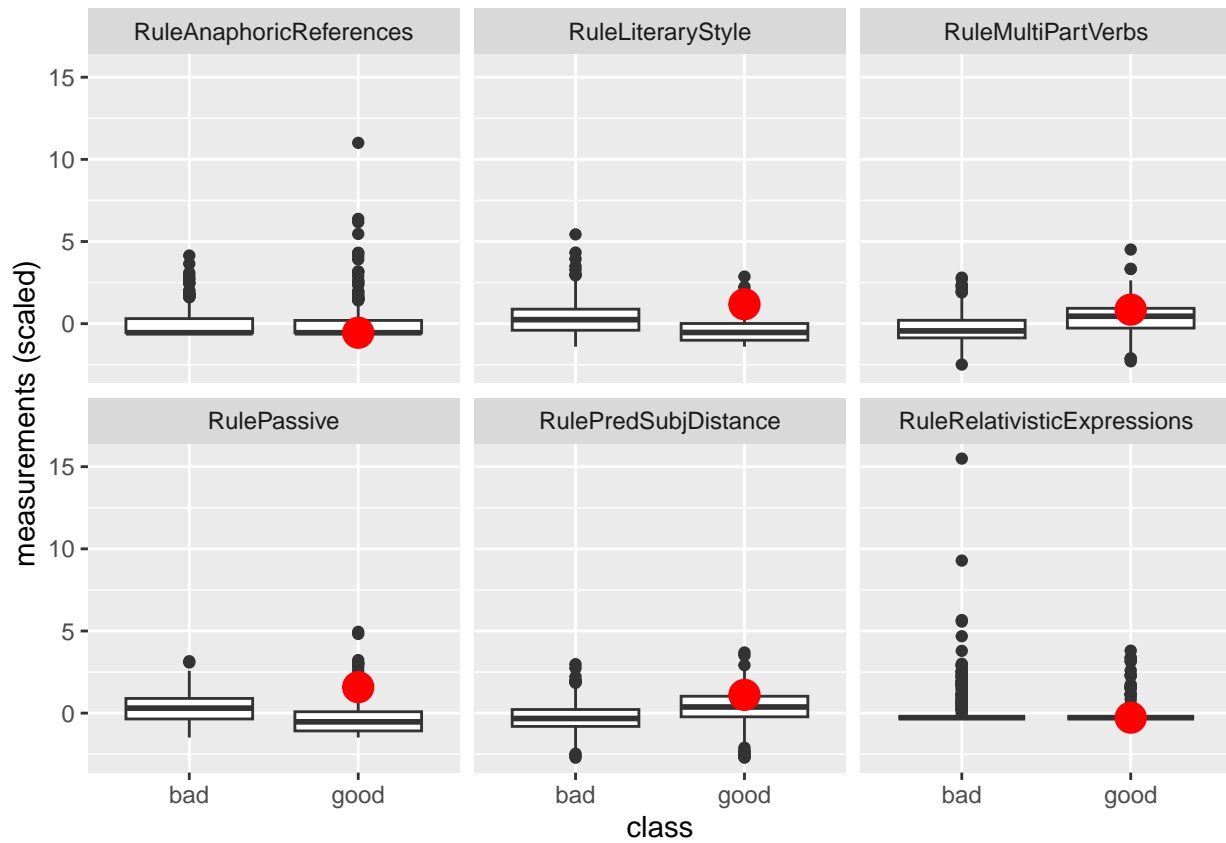
KUKY / 1A__dokument__puvodni__ustanoven__zastupce__vyzva__k__doplneni__kast__pouceni

truth: good

counts-outlier

```
plot_outlier("1A_dokument_puvodni_ustanoven_zastupce_vyzva_k_doplneni_kast_pouceni", lfit_lasso_counts,
```
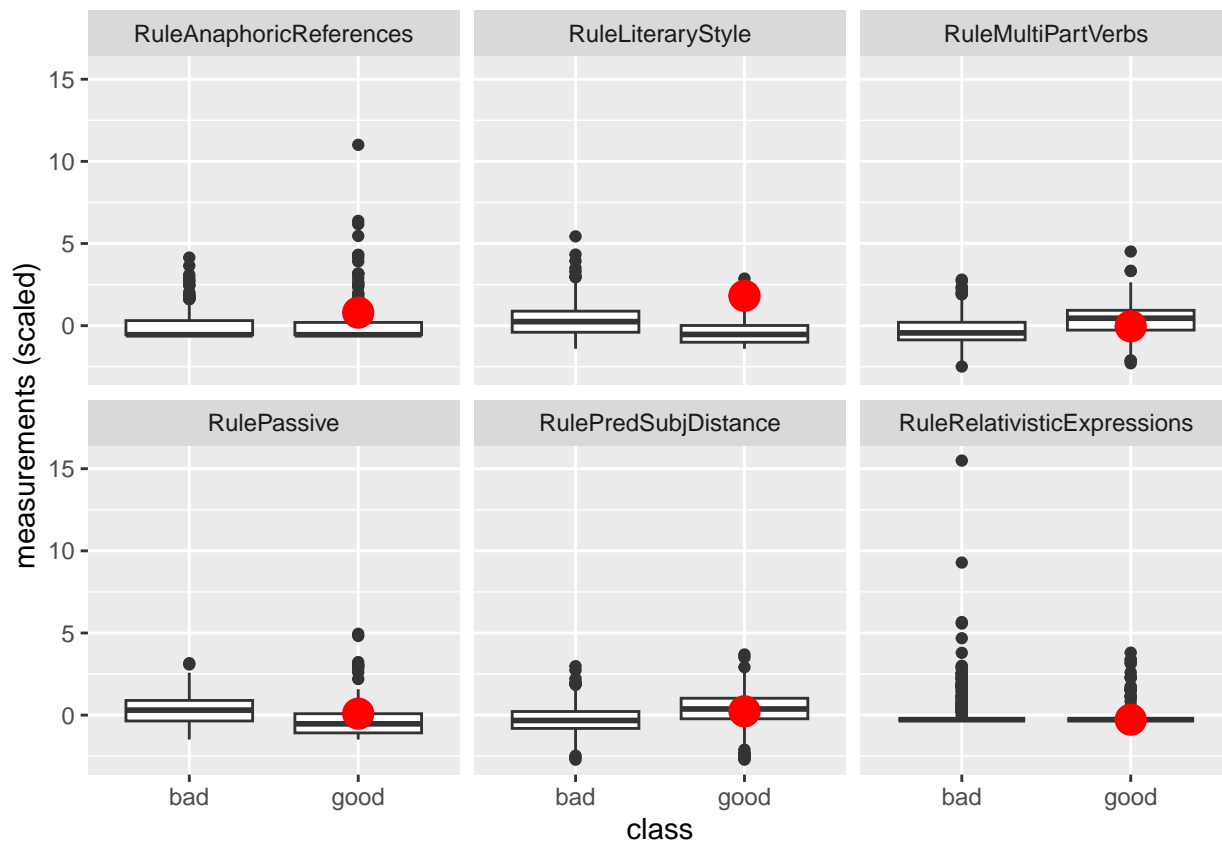
RF

FrBo / orig_Jak uspořádat shromáždění

> truth: bad

```
plot_outlier("orig_Jak uspořádat shromáždění", lfit_rf_all, data_clean)
```
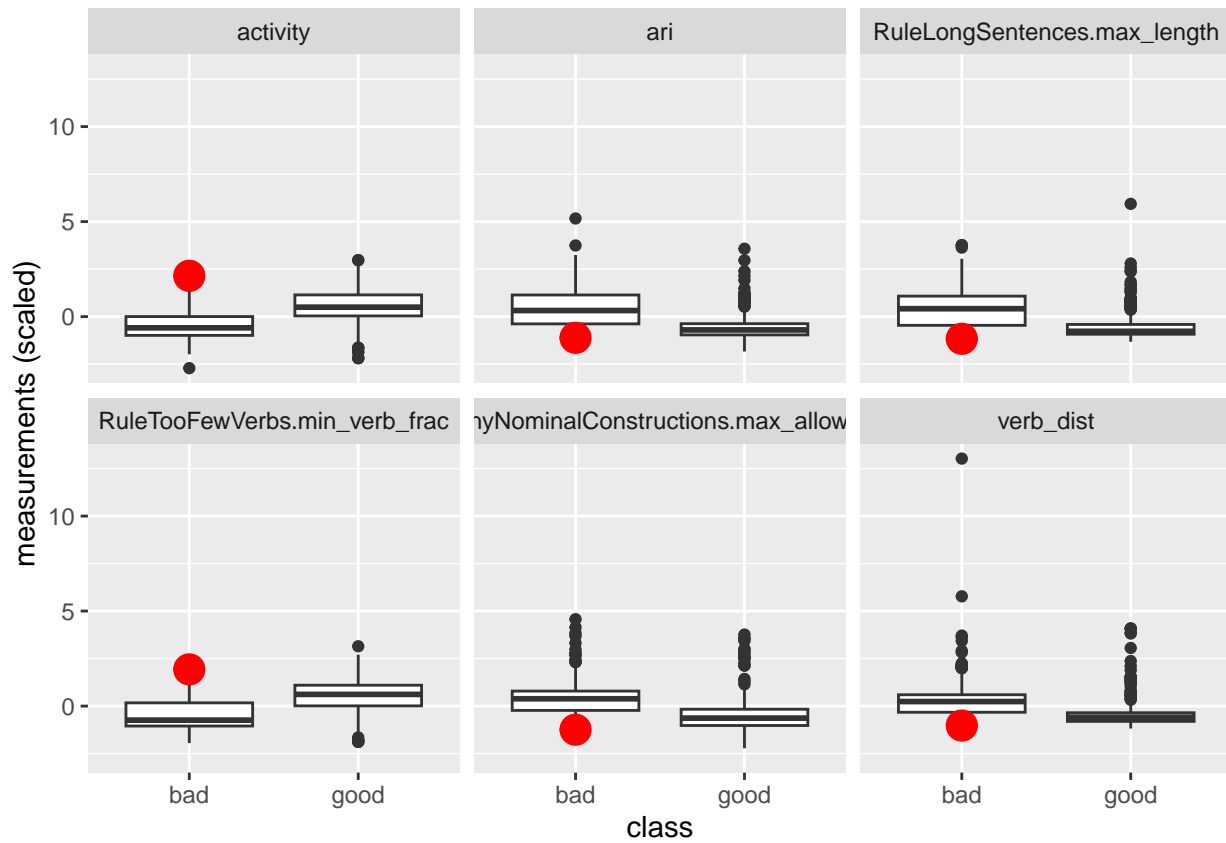
KUKY / 33 Cdo 30_2024

> truth: good

```
plot_outlier("33 Cdo 30_2024", lfit_rf_all, data_clean)
```
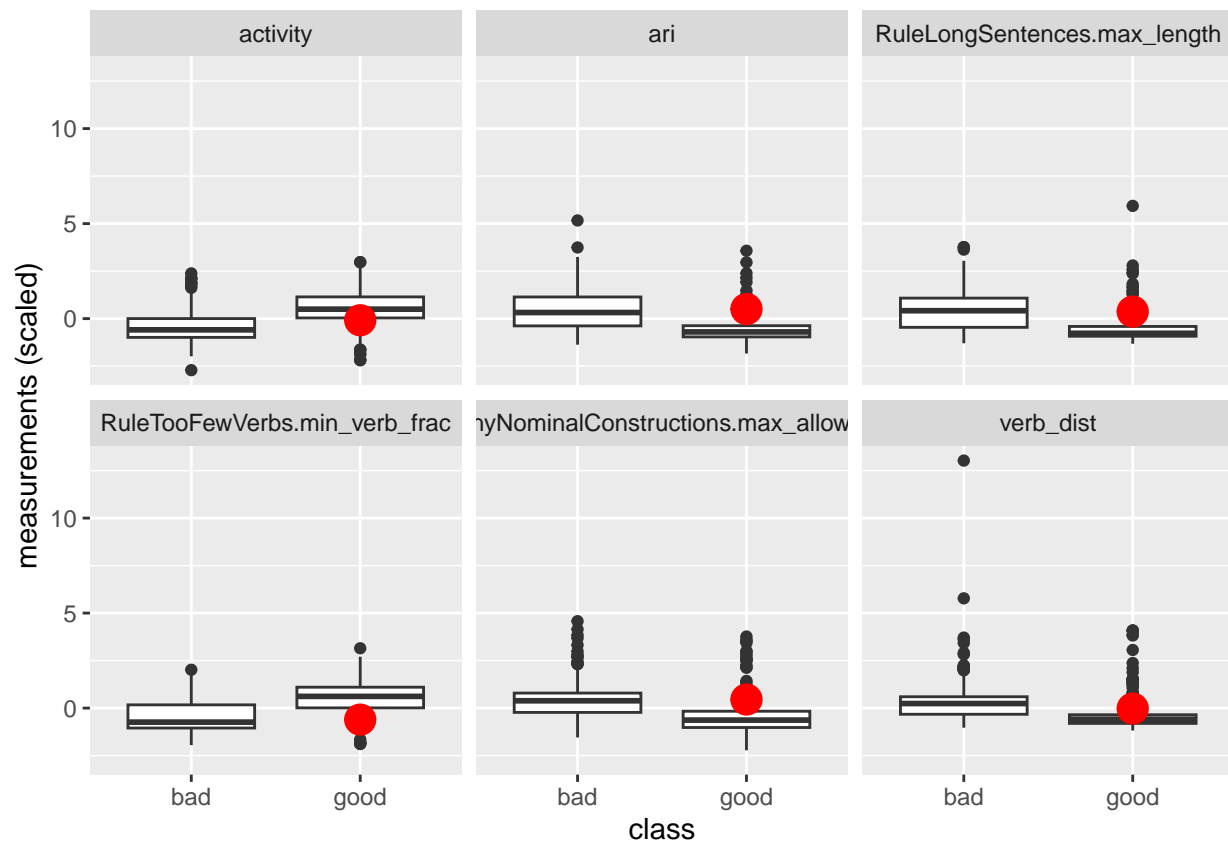
KUKY / 11_vizum_pred

truth: good

```
plot_outlier("11_vizum_pred", lfit_rf_all, data_clean)
```
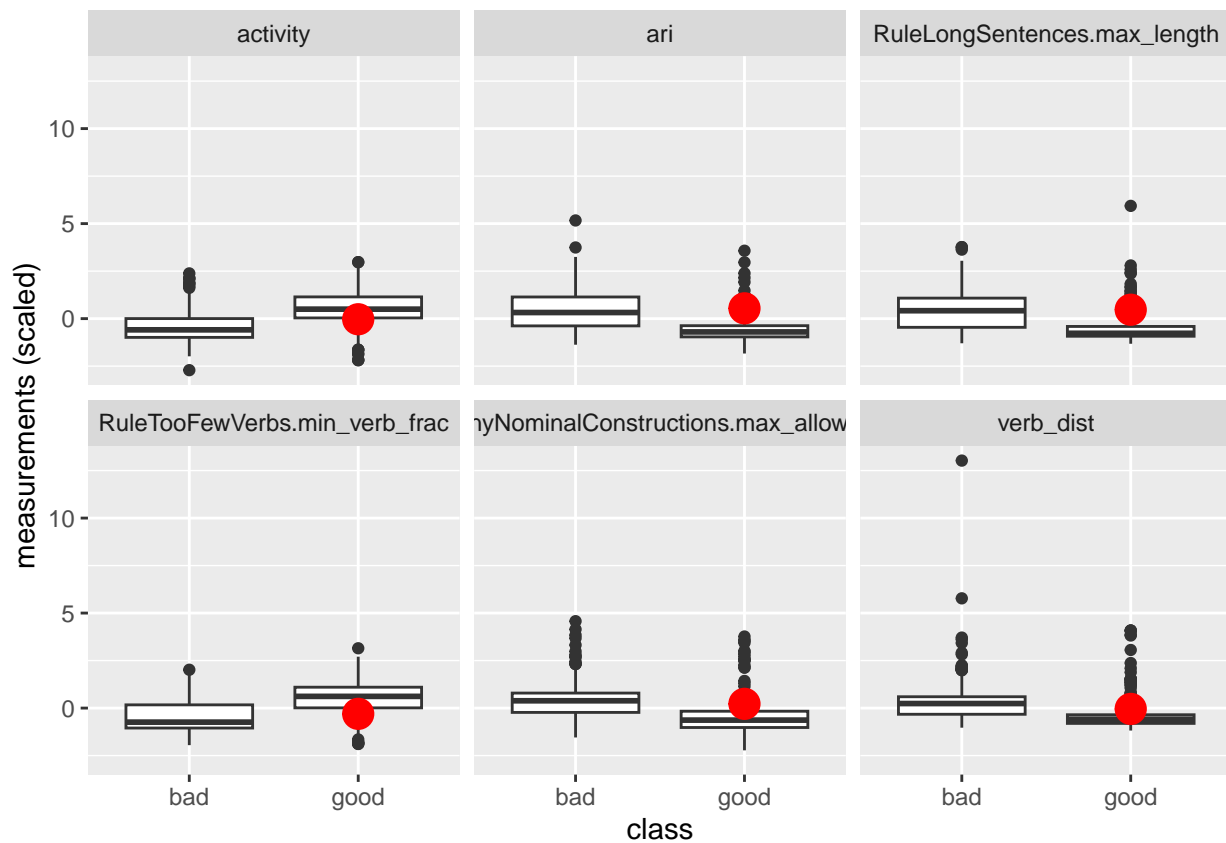
FrBo / orig__Kterých řízení se může váš spolek účastnit__FINAL

truth: bad

```
plot_outlier("orig_Kterých řízení se může váš spolek účastnit_FINAL", lfit_rf_all, data_clean)
```
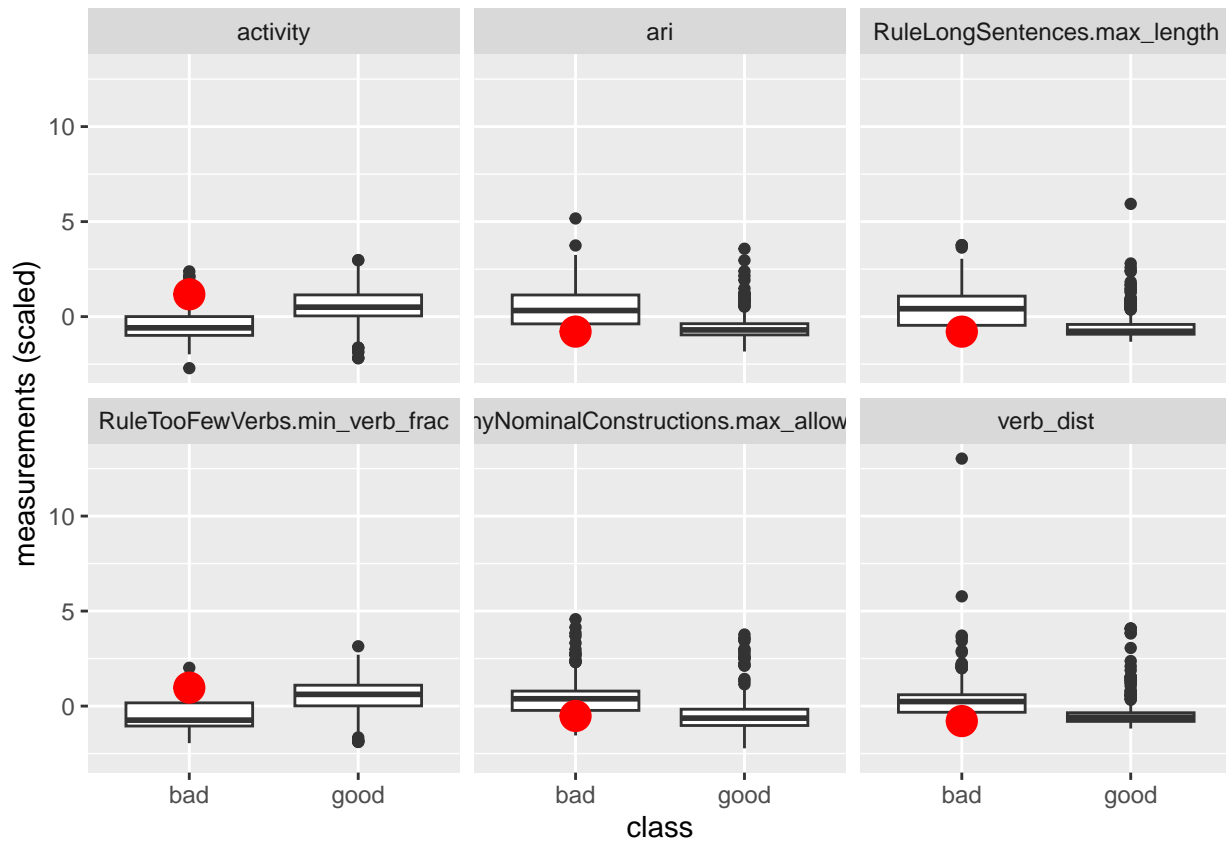
**FrBo / orig_lhuty_v_jednani_s_urady_a_soudy**

truth: bad

```
plot_outlier("orig_lhuty_v_jednani_s_urady_a_soudy", lfit_rf_all, data_clean)
```
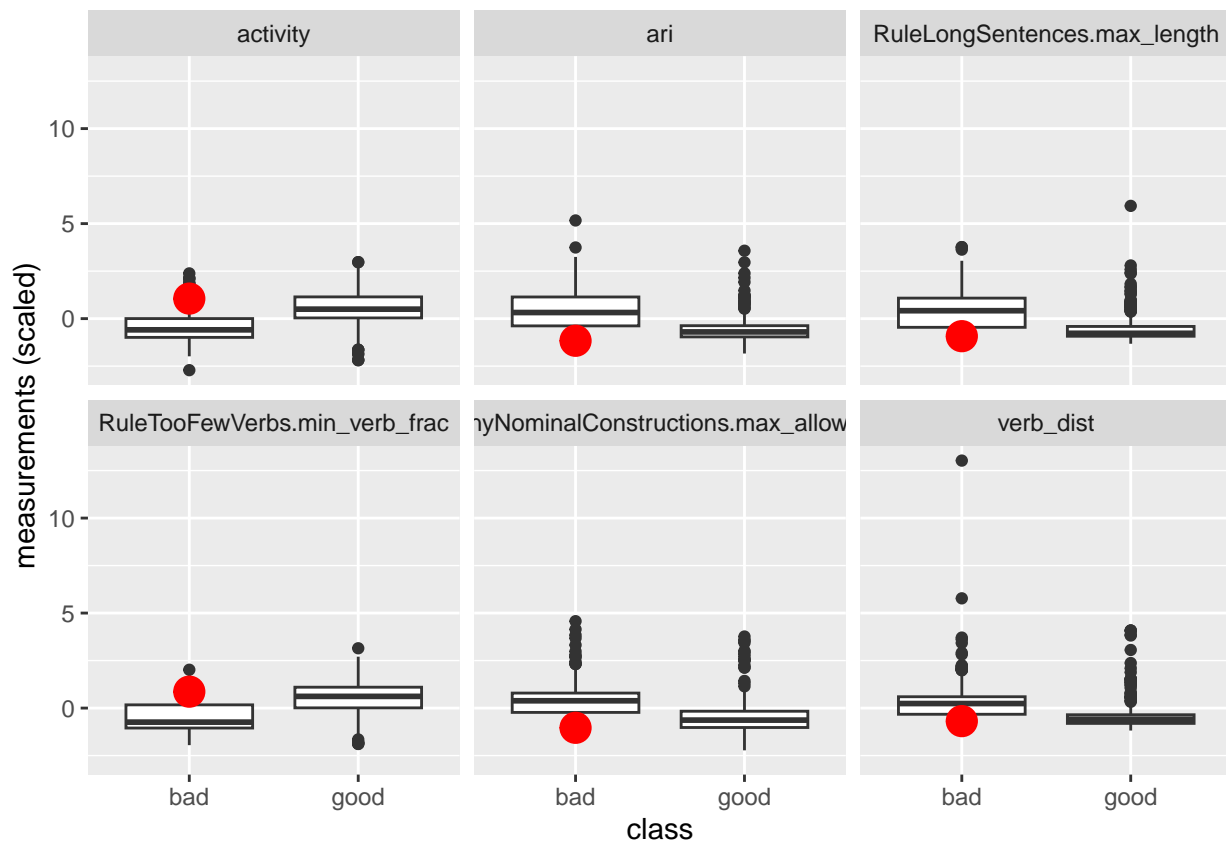
FrBo / orig_Zastupitelstvo_o čem a jak rozhoduje

truth: bad

```
plot_outlier("orig_Zastupitelstvo_o čem a jak rozhoduje", lfit_rf_all, data_clean)
```
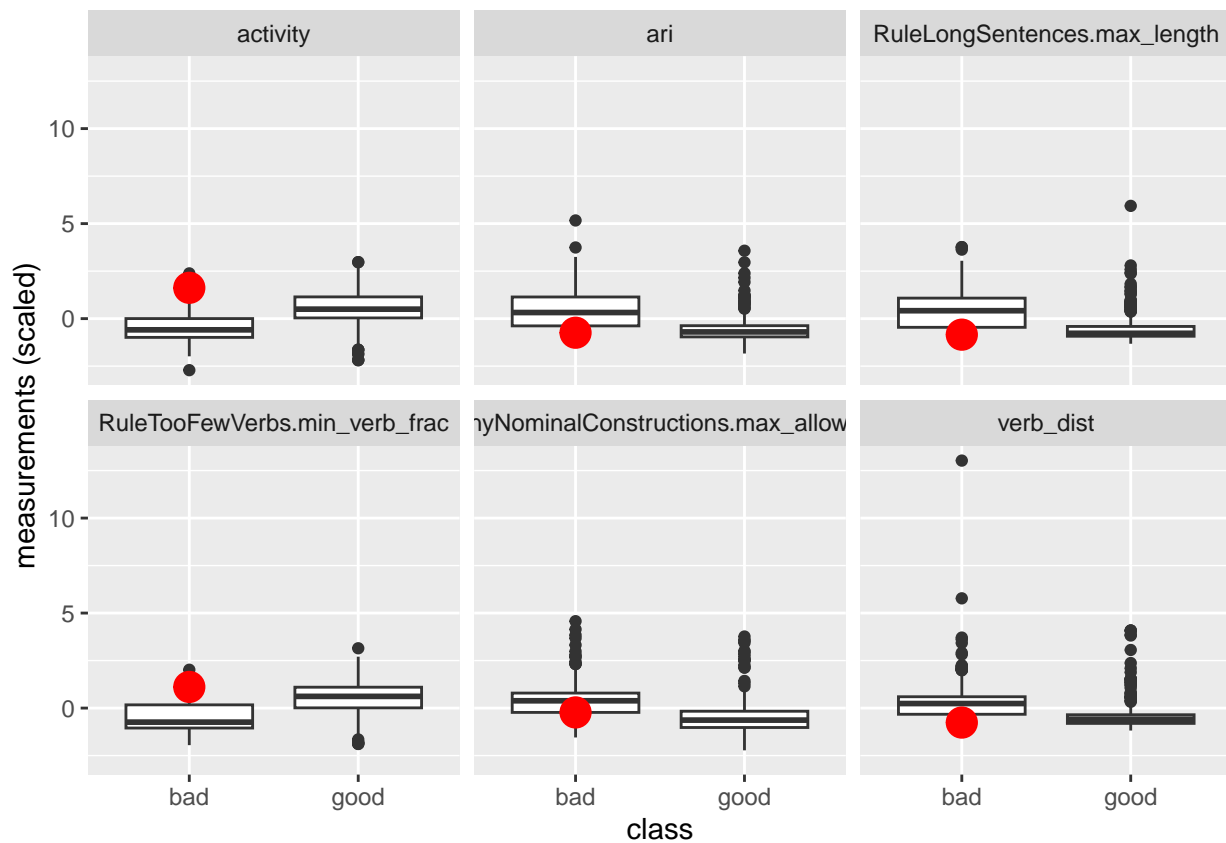
FrBo / orig__Jak probíhá správní řízení

     truth: bad

```
plot_outlier("orig_Jak probíhá správní řízení", lfit_rf_all, data_clean)
```
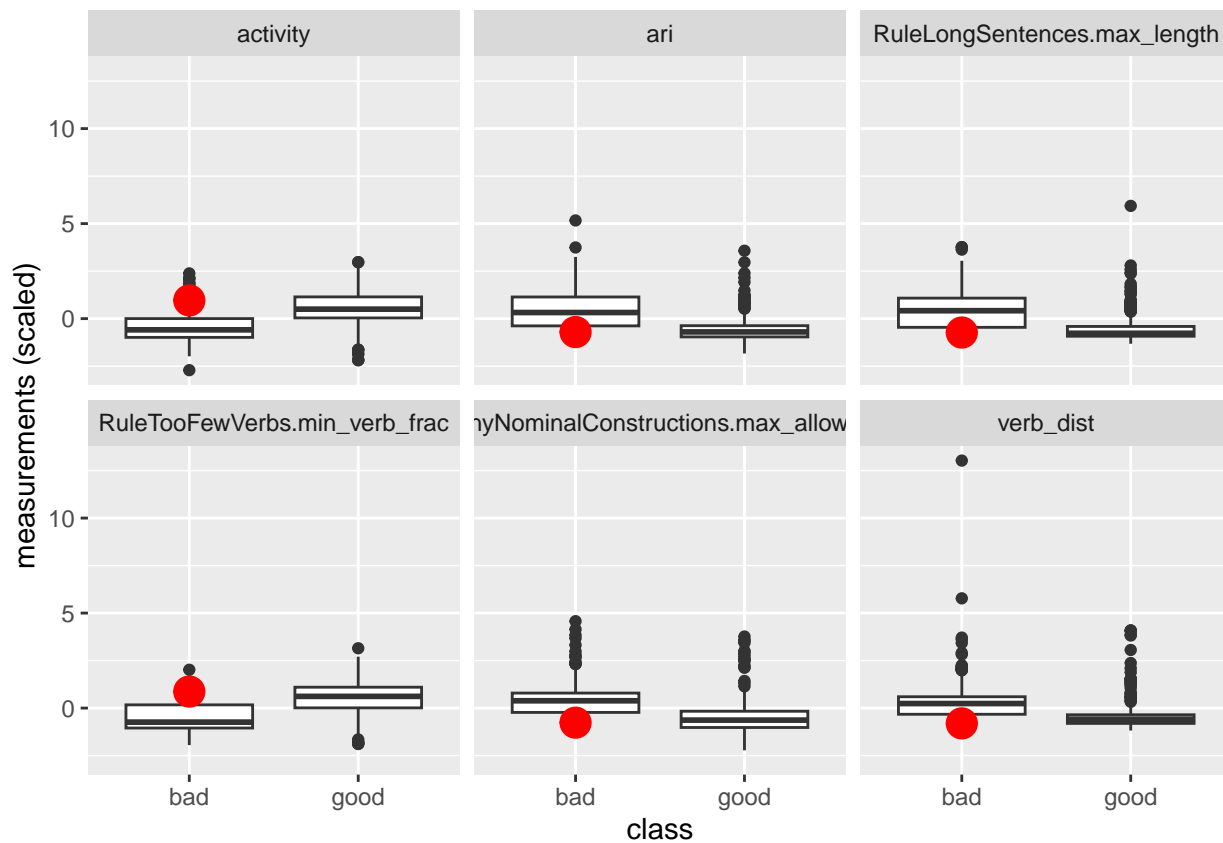
FrBo / orig_Jak namítat podjatost_final

     truth: bad

```
plot_outlier("orig_Jak namítat podjatost_final", lfit_rf_all, data_clean)
```
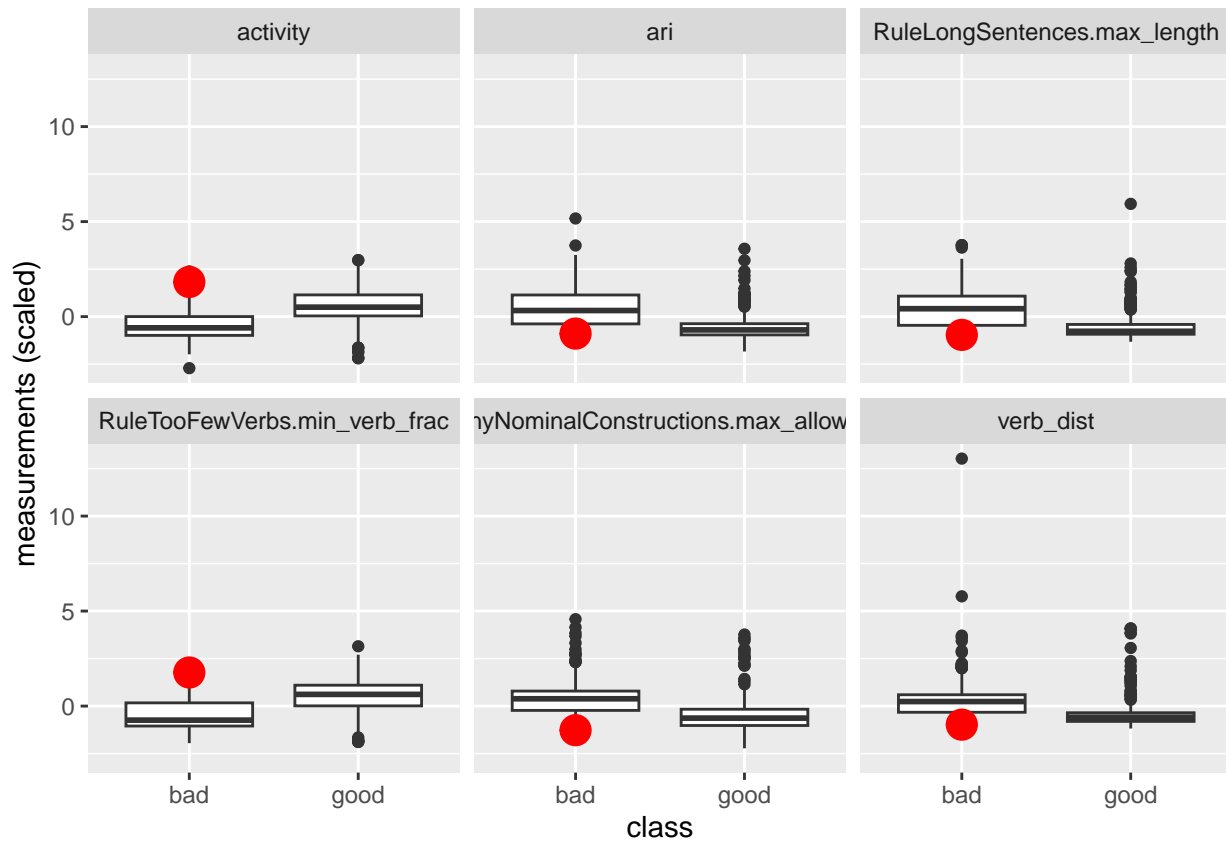
FrBo / 142

     truth: bad

     IAC-outlier

```
plot_outlier("142", lfit_rf_iac, data_clean)
```
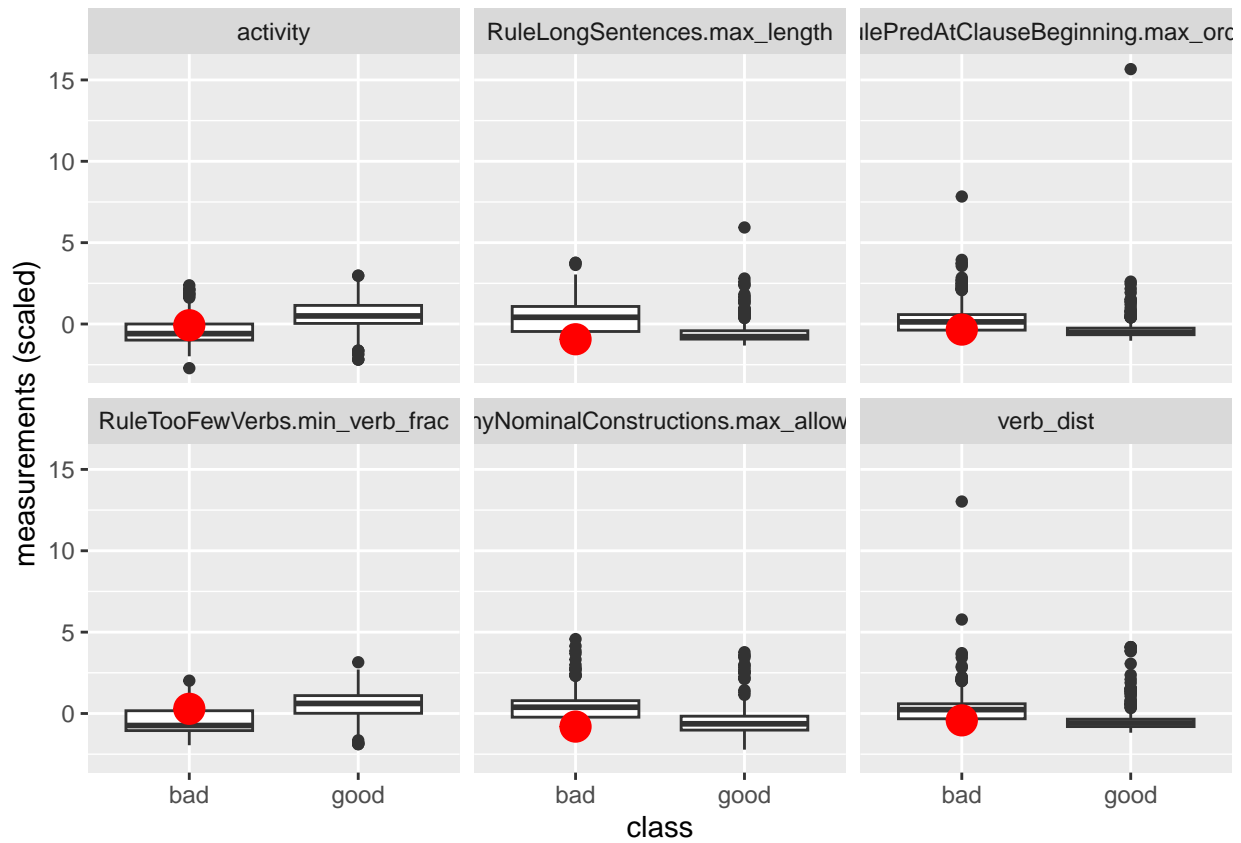
FrBo / 64

truth: bad

counts-outlier

```
plot_outlier("64", lfit_rf_counts, data_clean)
```
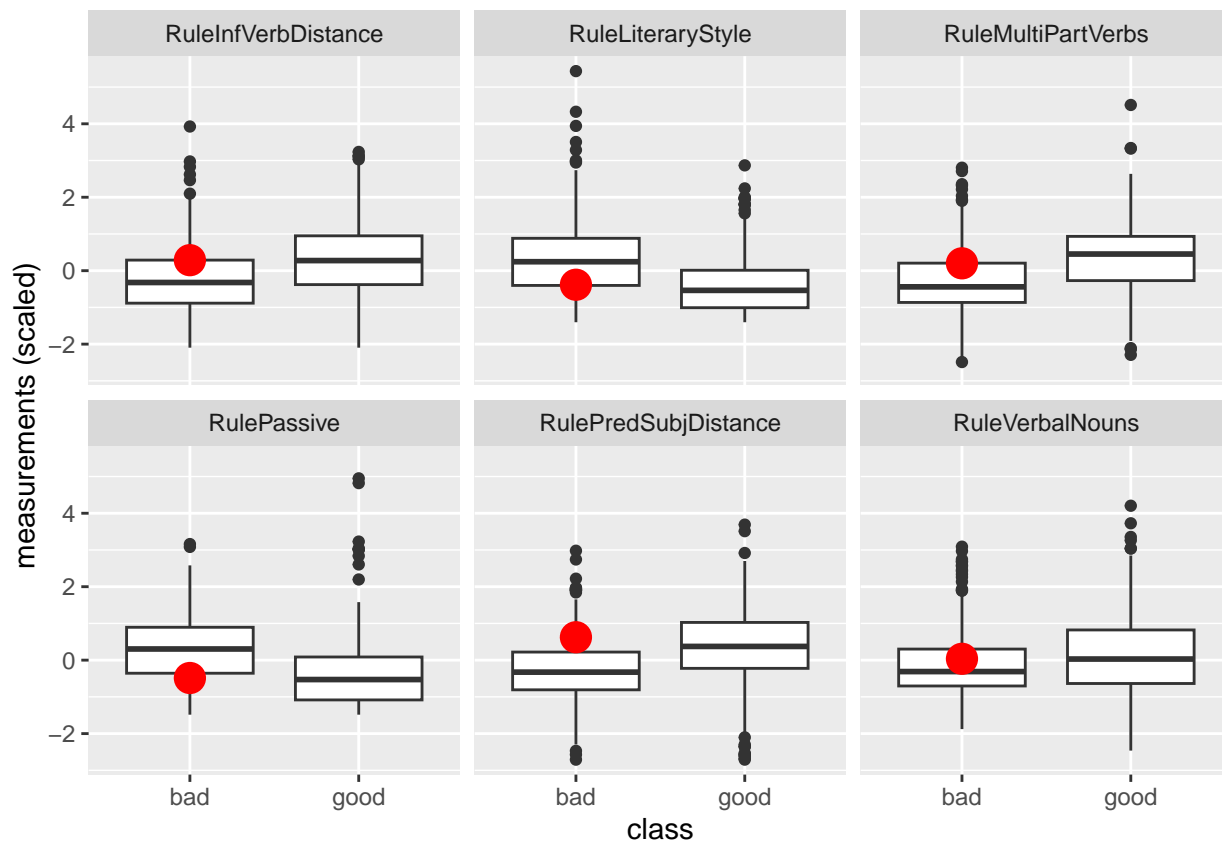
FrBo / red_Certifikáty autorizovaných inspektorů

truth: good

counts-outlier

```
plot_outlier("red_Certifikáty autorizovaných inspektorů", lfit_rf_counts, data_clean)
```

KUKY / 2A__dokument__puvodni__vyzva__k__zaplaceni__SOP__a__k__doplneni__kast__pouceni

truth: good

counts-outlier

```
plot_outlier("2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni", lfit_rf_counts, data
```

FrBo / orig__provokace__korupcniho__jednani

truth: bad

counts-outlier

```
plot_outlier("orig_provokace_korupcniho_jednani", lfit_rf_counts, data_clean)
```

Summary

All outliers (*notl* models excluded since they're the same as *all*):

```
table(c(
  lasso_all_devs$highest_deviations,
  lasso_iac_devs$highest_deviations,
  lasso_counts_devs$highest_deviations,
  rf_all_devs$highest_deviations,
  rf_iac_devs$highest_deviations,
  rf_counts_devs$highest_deviations
)) %>% sort(decreasing = TRUE)
```

```
## 
##                              orig_Jak uspořádat shromáždění 
##                                                          5 
##                              orig_Jak namítat podjatost_final 
##                                                          4 
##                   orig_Zastupitelstvo_o čem a jak rozhoduje 
##                                                          4 
##                                              11_vizum_pred 
##                                                          3 
##   2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni 
##                                                          3 
##                                            33 Cdo 30_2024 
##                                                          3 
##                            orig_Jak probíhá správní řízení 
##                                                          3 
```

```
##                orig_Kterých řízení se může váš spolek účastnit_FINAL
##                                                                    3
##                             Mestsky_urad_usneseni_-_slouceni_pred
##                                                                    2
##                                  orig_lhuty_v_jednani_s_urady_a_soudy
##                                                                    2
##                             red_Certifikáty autorizovaných inspektorů
##                                                                    2
##                                                                  142
##                                                                    1
## 1A_dokument_puvodni_ustanoven_zastupce_vyzva_k_doplneni_kast_pouceni
##                                                                    1
##                                                                   64
##                                                                    1
##                       Mestsky_urad_Vyzva_k_zaplaceni_nakladu_rizeni_pred
##                                                                    1
##                                     orig_provokace_korupcniho_jednani
##                                                                    1
##                                          Reakce_na_dopis_pracovni
##                                                                    1
##                                              Reakce_na_dopis_rev
##                                                                    1
```

Lasso outliers:

```r
table(c(
  lasso_all_devs$highest_deviations,
  lasso_iac_devs$highest_deviations,
  lasso_counts_devs$highest_deviations
)) %>% sort(decreasing = TRUE)
```

```
##
##  2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni
##                                                                    2
##                             Mestsky_urad_usneseni_-_slouceni_pred
##                                                                    2
##                                      orig_Jak namítat podjatost_final
##                                                                    2
##                                       orig_Jak uspořádat shromáždění
##                                                                    2
##                           orig_Zastupitelstvo_o čem a jak rozhoduje
##                                                                    2
## 1A_dokument_puvodni_ustanoven_zastupce_vyzva_k_doplneni_kast_pouceni
##                                                                    1
##                       Mestsky_urad_Vyzva_k_zaplaceni_nakladu_rizeni_pred
##                                                                    1
##                                       orig_Jak probíhá správní řízení
##                                                                    1
##                orig_Kterých řízení se může váš spolek účastnit_FINAL
##                                                                    1
##                                          Reakce_na_dopis_pracovni
##                                                                    1
##                                              Reakce_na_dopis_rev
##                                                                    1
##                             red_Certifikáty autorizovaných inspektorů
```
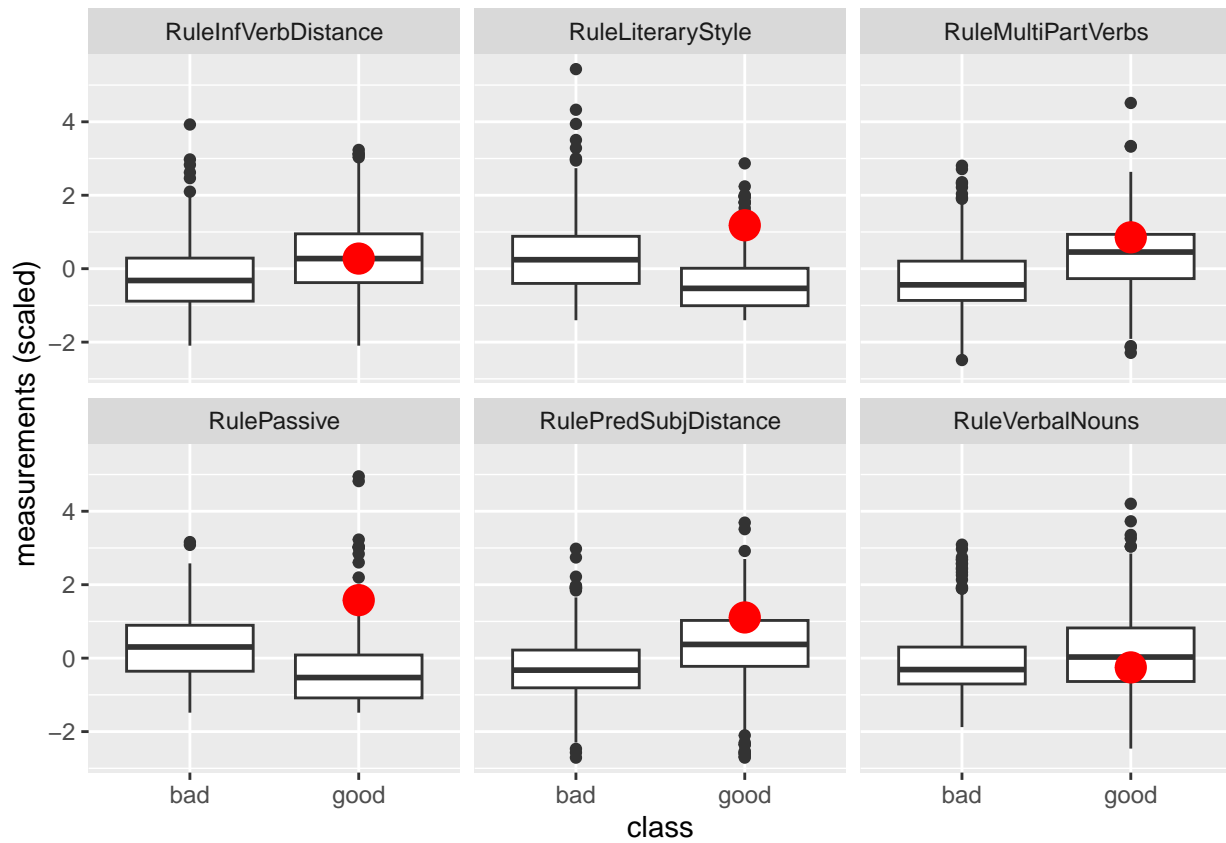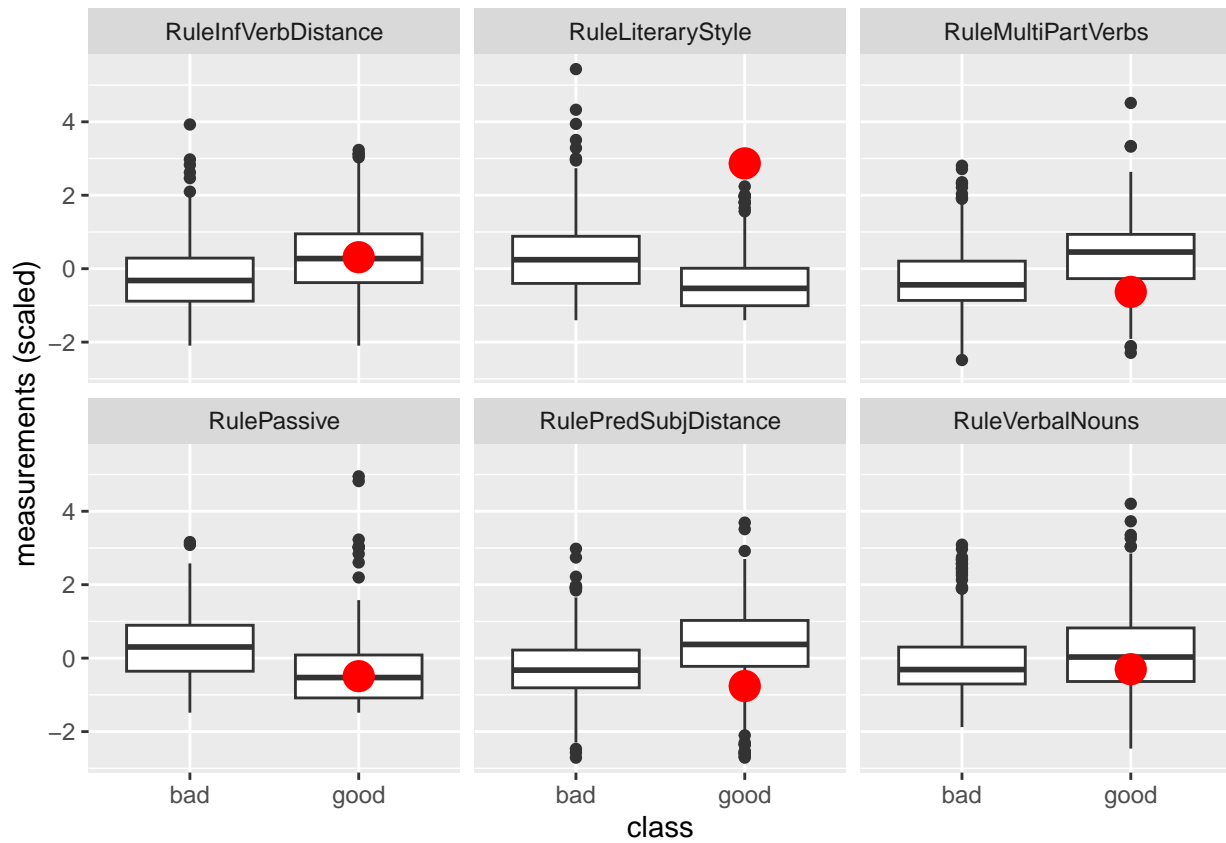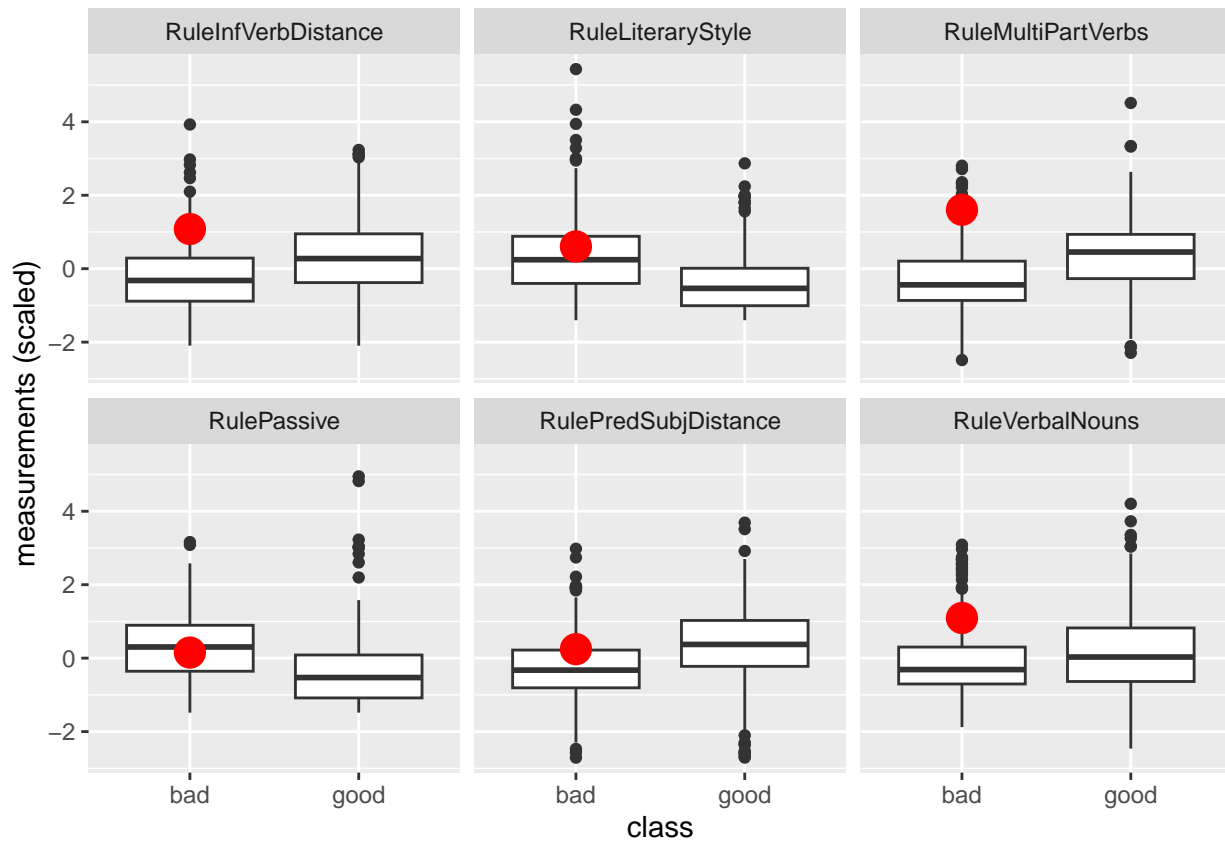
RF outliers:

```
table(c(
  rf_all_devs$highest_deviations,
  rf_iac_devs$highest_deviations,
  rf_counts_devs$highest_deviations
)) %>% sort(decreasing = TRUE)
```

```
## 
##                                                         11_vizum_pred
##                                                                     3
##                                                         33 Cdo 30_2024
##                                                                     3
##                                        orig_Jak uspořádat shromáždění
##                                                                     3
##                                     orig_Jak namítat podjatost_final
##                                                                     2
##                                 orig_Jak probíhá správní řízení
##                                                                     2
##           orig_Kterých řízení se může váš spolek účastnit_FINAL
##                                                                     2
##                                 orig_lhuty_v_jednani_s_urady_a_soudy
##                                                                     2
##                           orig_Zastupitelstvo_o čem a jak rozhoduje
##                                                                     2
##                                                                   142
##                                                                     1
## 2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni
##                                                                     1
##                                                                    64
##                                                                     1
##                                       orig_provokace_korupcniho_jednani
##                                                                     1
##                             red_Certifikáty autorizovaných inspektorů
##                                                                     1
```

all-models outliers:

```
table(c(
  lasso_all_devs$highest_deviations,
  rf_all_devs$highest_deviations
)) %>% sort(decreasing = TRUE)
```

```
## 
##                                     orig_Jak namítat podjatost_final
##                                                                     2
##                                        orig_Jak uspořádat shromáždění
##                                                                     2
##                           orig_Zastupitelstvo_o čem a jak rozhoduje
##                                                                     2
##                                                         11_vizum_pred
##                                                                     1
## 2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni
##                                                                     1
```

```
##                                            33 Cdo 30_2024
##                                                         1
##                          Mestsky_urad_usneseni_-_slouceni_pred
##                                                         1
##                             orig_Jak probíhá správní řízení
##                                                         1
##               orig_Kterých řízení se může váš spolek účastnit_FINAL
##                                                         1
##                          orig_lhuty_v_jednani_s_urady_a_soudy
##                                                         1
```

iac-models outliers:

```r
table(c(
  lasso_iac_devs$highest_deviations,
  rf_iac_devs$highest_deviations
)) %>% sort(decreasing = TRUE)
```

```
##
##                          orig_Jak namítat podjatost_final
##                                                         2
##                             orig_Jak probíhá správní řízení
##                                                         2
##                            orig_Jak uspořádat shromáždění
##                                                         2
##                                              11_vizum_pred
##                                                         1
##                                                       142
##                                                         1
##                                            33 Cdo 30_2024
##                                                         1
##                    Mestsky_urad_usneseni_-_slouceni_pred
##                                                         1
##    Mestsky_urad_Vyzva_k_zaplaceni_nakladu_rizeni_pred
##                                                         1
## orig_Kterých řízení se může váš spolek účastnit_FINAL
##                                                         1
##                          orig_lhuty_v_jednani_s_urady_a_soudy
##                                                         1
##              orig_Zastupitelstvo_o čem a jak rozhoduje
##                                                         1
```

counts-models outliers:

```r
table(c(
  lasso_counts_devs$highest_deviations,
  rf_counts_devs$highest_deviations
)) %>% sort(decreasing = TRUE)
```

```
##
##  2A_dokument_puvodni_vyzva_k_zaplaceni_SOP_a_k_doplneni_kast_pouceni
##                                                         2
##                          red_Certifikáty autorizovaných inspektorů
##                                                         2
##                                              11_vizum_pred
##                                                         1
```

```
## 1A_dokument_puvodni_ustanoven_zastupce_vyzva_k_doplneni_kast_pouceni
##                                                                     1
##                                                          33 Cdo 30_2024
##                                                                     1
##                                                                    64
##                                                                     1
##                                          orig_Jak uspořádat shromáždění
##                                                                     1
##                      orig_Kterých řízení se může váš spolek účastnit_FINAL
##                                                                     1
##                                     orig_provokace_korupcniho_jednani
##                                                                     1
##                            orig_Zastupitelstvo_o čem a jak rozhoduje
##                                                                     1
##                                            Reakce_na_dopis_pracovni
##                                                                     1
##                                                 Reakce_na_dopis_rev
##                                                                     1
```