LassoModelRaceHispanic.R

hugobaca

2023-07-11

```
# Modelling with Tidymodels
# Example for LASSO
# Predict PHS_race_hispnaic ~ .
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(VIM)
```

```
## Loading required package: colorspace
## Loading required package: grid
## The legacy packages maptools, rgdal, and rgeos, underpinning this package
## will retire shortly. Please refer to R-spatial evolution reports on
## https://r-spatial.org/r/2023/05/15/evolution4.html for details.
## This package is now running under evolution status 0
## VIM is ready to use.
##
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
##
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
# 1. Read Data
data_chrs23 <- readRDS("./data/interim/analytic2023_c1.rds")
desc = readRDS("./data/processed/colDesc_analysis2023.rds")

# We will only focus on raw values
data_core = data_chrs23 %>%
  select(-ends_with("_flag"), -ends_with("cilow"), -ends_with("cihigh"),
         -ends_with("_numerator"), -ends_with("denominator"),
         - c("county_ranked", "statecode", "countycode", "fipscode")) %>%
  filter(state != "US")

data_race_hispanic_only = data_core %>% select(- contains("black"), - contains("white"),
      - contains("asian"), - contains("aian"),-county) %>% drop_na() %>% mutate(across(where(is.characte

# 2. Modeling Pipeline
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.1.0 --
## v broom        1.0.4     v rsample      1.1.1
## v dials        1.2.0     v tune         1.1.1
## v infer        1.0.4     v workflows    1.1.3
## v modeldata    1.1.0     v workflowsets 1.0.1
## v parsnip      1.1.0     v yardstick    1.2.0
## v recipes      1.0.6
## -- 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 recipes::prepare() masks VIM::prepare()
## x yardstick::spec()  masks readr::spec()
## x recipes::step()    masks stats::step()
## * Use suppressPackageStartupMessages() to eliminate package startup messages
```

```
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-7
```

```
# 2.1. Define Receipe specification (saturated model)
target_name = "v005_rawvalue"
recipe_sat =
  recipe(v005_rawvalue~ ., data=data_race_hispanic_only) %>%
  step_naomit(all_predictors()) %>%
  #  step_log(all_numeric_predictors(), offset=1) %>%
  step_dummy(all_nominal_predictors(), one_hot=TRUE)

# 2.2. Define data splits
set.seed(1)
split = initial_split(data_race_hispanic_only, prop=0.7, strata=target_name, breaks=5)
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(target_name)
##
##   # Now:
```

```
##   data %>% select(all_of(target_name))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The number of observations in each quantile is below the recommended threshold of 20.
## * Stratification will use 4 breaks instead.
```

```r
train_data = training(split)
test_data = testing(split)

# Check state imbalance
data.frame(test_ratio = round(test_data$state %>% table / nrow(test_data) * 100, 2),
           train_ratio = round(train_data$state %>% table / nrow(train_data) * 100, 2))
```

```
##    test_ratio.. test_ratio.Freq train_ratio.. train_ratio.Freq
## 1            CA           22.22            CA            28.57
## 2            CO           14.81            CO             5.36
## 3            FL           25.93            FL            19.64
## 4            GA            0.00            GA             5.36
## 5            IA            0.00            IA             1.79
## 6            IN            0.00            IN             5.36
## 7            MI            3.70            MI             3.57
## 8            MN            3.70            MN             1.79
## 9            MO            0.00            MO             1.79
## 10           NE            0.00            NE             1.79
## 11           NJ           11.11            NJ            10.71
## 12           PA            3.70            PA             1.79
## 13           RI            3.70            RI             0.00
## 14           UT            0.00            UT             5.36
## 15           WA           11.11            WA             5.36
## 16           WI            0.00            WI             1.79
```

```r
# 2.3 Define model engine
engine_lasso = linear_reg(penalty = tune(), mixture = 1) %>%
  set_engine("glmnet") %>%
  set_mode("regression") %>%
  translate()

# 2.4. Bind the Workflow
workflow = workflow() %>%
  add_model(engine_lasso) %>%
  add_recipe(recipe_sat)

# 2.5 Hyperparameter tuning with cross-validation
## Create 5-folds
resampling = vfold_cv(train_data, v=5, strata = target_name)
```

```
## Warning: The number of observations in each quantile is below the recommended threshold of 20.
## * Stratification will use 2 breaks instead.
```

```r
## Parallel process to lift computation burden
library(doParallel)
```

```
## Loading required package: foreach
```

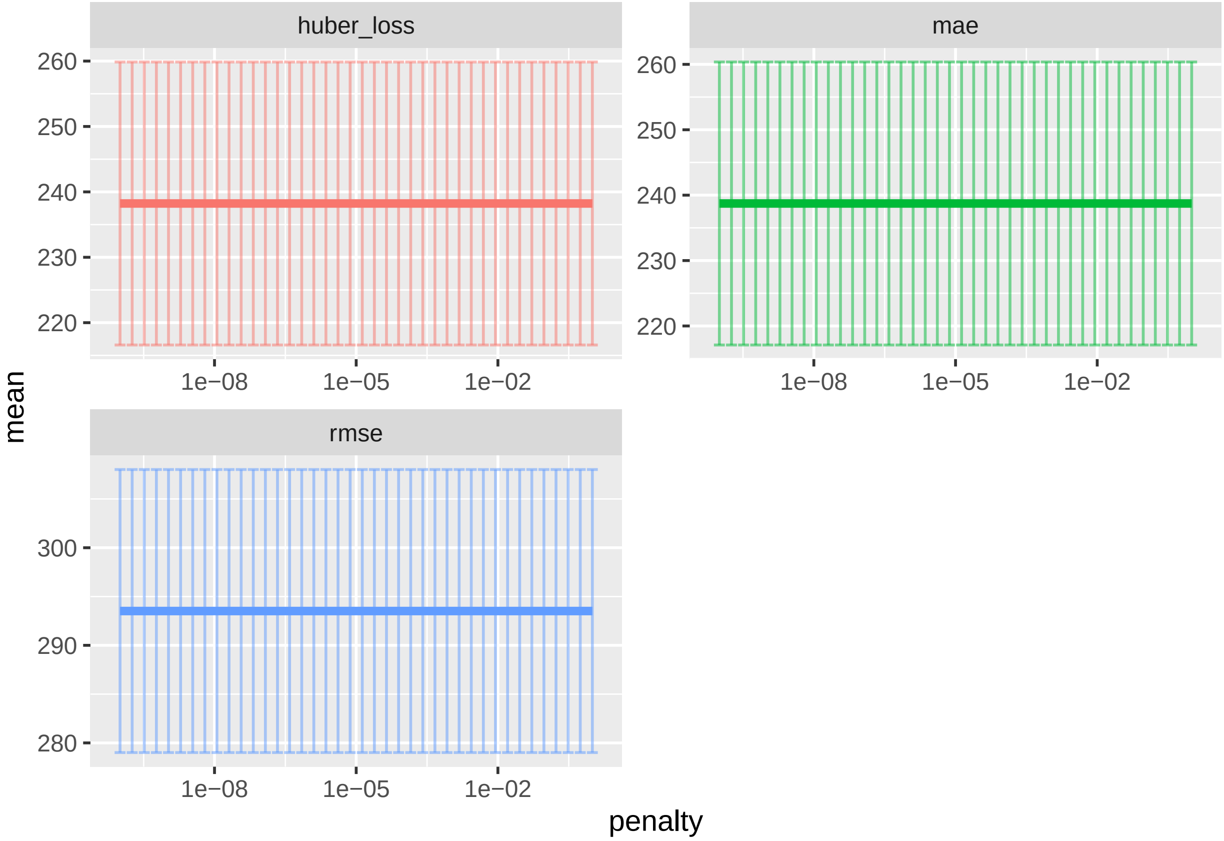
```
##
## Attaching package: 'foreach'
##
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
##
## Loading required package: iterators
## Loading required package: parallel
```

```
cl <- makeCluster(parallel::detectCores())
registerDoParallel(cl)

## Define search grid
param_grid <- grid_regular(
  penalty(),
  levels = 40
)

tuned_model = tune_grid(
  workflow,
  resamples = resampling,
  metrics=metric_set(rmse, mae, huber_loss),
  grid = param_grid,
  control = control_grid(allow_par=TRUE, save_pred=TRUE, parallel_over = "resamples")
)
collect_metrics(tuned_model)
```

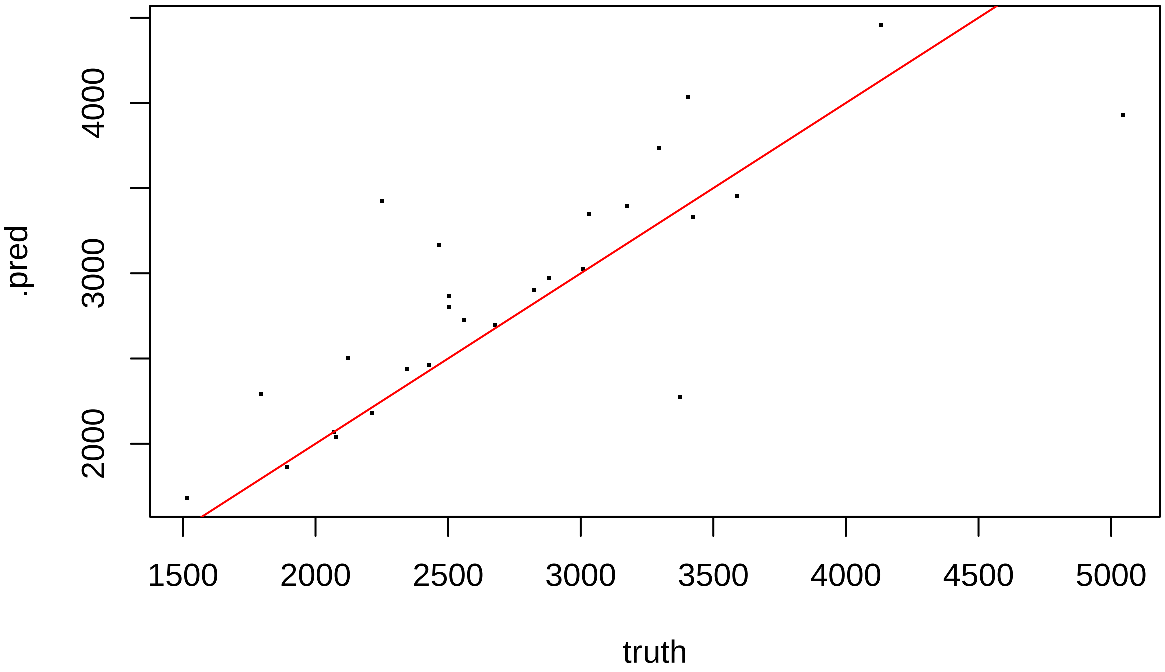
```
## # A tibble: 120 x 7
##     penalty .metric    .estimator  mean     n std_err .config
##       <dbl> <chr>      <chr>      <dbl> <int>   <dbl> <chr>
##  1 1   e-10 huber_loss standard    238.     5    21.6 Preprocessor1_Model01
##  2 1   e-10 mae        standard    239.     5    21.6 Preprocessor1_Model01
##  3 1   e-10 rmse       standard    294.     5    14.5 Preprocessor1_Model01
##  4 1.80e-10 huber_loss standard    238.     5    21.6 Preprocessor1_Model02
##  5 1.80e-10 mae        standard    239.     5    21.6 Preprocessor1_Model02
##  6 1.80e-10 rmse       standard    294.     5    14.5 Preprocessor1_Model02
##  7 3.26e-10 huber_loss standard    238.     5    21.6 Preprocessor1_Model03
##  8 3.26e-10 mae        standard    239.     5    21.6 Preprocessor1_Model03
##  9 3.26e-10 rmse       standard    294.     5    14.5 Preprocessor1_Model03
## 10 5.88e-10 huber_loss standard    238.     5    21.6 Preprocessor1_Model04
## # i 110 more rows
```

```
## Check the search performance
tuned_model %>%
  collect_metrics() %>%
  ggplot(aes(penalty, mean, color = .metric)) +
  geom_errorbar(aes(
    ymin = mean - std_err,
    ymax = mean + std_err
  ),
  alpha = 0.5
  ) +
  geom_line(size = 1.5) +
  facet_wrap(~.metric, scales = "free", nrow = 2) +
  scale_x_log10() +
  theme(legend.position = "none")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



```
# Finalize the model and fit the model
model_fitted = workflow %>%
  finalize_workflow(select_best(tuned_model, metric="huber_loss")) %>%
  fit(train_data)

# Evaluate OOS Performance
test_prediction = predict(model_fitted,
                          new_data = test_data)
test_prediction = test_prediction %>%
  mutate(truth = test_data[[target_name]])

# Check calibration
with(test_prediction,
     plot(truth, .pred, pch=".", cex=2))
abline(a=0, b=1, col="red")
```



```
eval_metric = metric_set(rmse, mae, huber_loss)
eval_metric(
  data = test_prediction,
  truth = truth,

  estimate = .pred
)
```

```
## # A tibble: 3 x 3
##   .metric    .estimator .estimate
##   <chr>      <chr>          <dbl>
## 1 rmse       standard        467.
## 2 mae        standard        317.
## 3 huber_loss standard        317.
```

```
# Assuming 'test_prediction' contains the truth and predicted values
```