xgboost Data-Rate Prediction

```r
library(tidyverse)
library(ggplot2)

library(mlr3)
library(mlr3learners)
library(mlr3pipelines)
```

Upload-Rate Prediction

Reading the Data

```r
data_dir = "../datasets/"

dataset_ul = read_csv(
  str_c(data_dir, "dataset_ul.csv"),
  col_types = cols(
    drive_id = col_integer(),
    scenario = col_factor(),
    provider = col_factor(),
    ci = col_factor(),
    enodeb = col_factor()
  )
) %>% select(
  drive_id,
  timestamp,
  scenario,
  provider,
  velocity_mps,
  acceleration_mpss,
  rsrp_dbm,
  rsrq_db,
  rssnr_db,
  cqi,
  ss,
  ta,
  ci,
  enodeb,
  f_mhz,
  payload_mb,
  throughput_mbits
) %>% drop_na() %>% rowid_to_column(var="row_id_original")

glimpse(dataset_ul)
```

```
## Rows: 6,168
## Columns: 18
## $ row_id_original   <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...
## $ drive_id          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ timestamp         <dttm> 2018-12-10 09:08:57, 2018-12-10 09:09:08, 2018-1...
## $ scenario          <fct> campus, campus, campus, campus, campus, campus, c...
## $ provider          <fct> o2, o2, o2, o2, o2, o2, o2, o2, o2, o2, o2, o2, o...
## $ velocity_mps      <dbl> 11.80, 11.49, 7.93, 10.44, 10.92, 12.02, 10.28, 0...
## $ acceleration_mpss <dbl> 0.13, -0.26, 0.23, 0.06, 0.56, 0.09, -1.25, 0.00,...
## $ rsrp_dbm          <dbl> -99, -97, -96, -82, -101, -106, -112, -99, -98, -...
## $ rsrq_db           <dbl> -9, -12, -12, -11, -14, -13, -18, -15, -15, -14, ...
## $ rssnr_db          <dbl> -1, -2, 5, 11, -3, -3, -6, -4, -6, -4, -6, -3, -2...
## $ cqi               <dbl> 8, 9, 5, 15, 6, 6, 3, 4, 7, 4, 4, 5, 6, 5, 1, 4, ...
## $ ss                <dbl> 36, 42, 42, 53, 39, 33, 31, 41, 40, 44, 43, 42, 4...
## $ ta                <dbl> 9, 7, 7, 7, 7, 7, 7, 12, 13, 13, 13, 13, 11, 13, ...
## $ ci                <fct> 13828122, 13416987, 13416987, 13416987, 13416987,...
## $ enodeb            <fct> 54016, 52410, 52410, 52410, 52410, 52410, 52410, ...
## $ f_mhz             <dbl> 1750, 1750, 1750, 1750, 1750, 1750, 1750, 880, 88...
## $ payload_mb        <dbl> 1.0, 6.0, 5.0, 7.0, 5.0, 8.0, 9.0, 7.0, 10.0, 2.0...
## $ throughput_mbits  <dbl> 4.66, 3.97, 6.52, 1.37, 0.80, 1.04, 2.34, 4.09, 2...
```

Create the Prediction Task

```r
task_ul = TaskRegr$new(
  id = "ul_prediction",
  backend = dataset_ul %>% select(-drive_id, -timestamp),
  target = "throughput_mbits"
)

task_ul$col_roles$name = "row_id_original"
task_ul$col_roles$feature = setdiff(task_ul$col_roles$feature, "row_id_original")

task_ul
```

```
## <TaskRegr:ul_prediction> (6168 x 15)
## * Target: throughput_mbits
## * Properties: -
## * Features (14):
##   - dbl (10): acceleration_mpss, cqi, f_mhz, payload_mb, rsrp_dbm,
##     rsrq_db, rssnr_db, ss, ta, velocity_mps
##   - fct (4): ci, enodeb, provider, scenario
```

Create Data Splitting Strategies for Testing and Validation

```r
make_outer_resampling = function(task, drive_ids_train, drive_ids_test) {
  row_ids_train = (
    tibble(task$row_names) %>%
      inner_join(dataset_ul, by=c("row_name"="row_id_original")) %>%
      filter(drive_id %in% drive_ids_train)
  )$row_id

  row_ids_test = (
    tibble(task$row_names) %>%
      inner_join(dataset_ul, by=c("row_name"="row_id_original")) %>%
      filter(drive_id %in% drive_ids_test)
  )$row_id

  result = rsmp("custom")
  result$instantiate(task, train_sets=list(row_ids_train), test_sets=list(row_ids_test))
  return(result)
}
```

Create the Prediction Pipeline

```
make_learner = function() {
  factor_encoding = po(
    "encode",
    method = "one-hot",
    affect_columns = selector_type("factor")
  )
  xgboost = lrn("regr.xgboost", nrounds=100) # default to 100 boosting rounds
  pipe = factor_encoding %>>% PipeOpLearner$new(xgboost)
  learner = GraphLearner$new(pipe)
  return(learner)
}
learner_ul = make_learner()
```

Validation Results

```
resampling_result_ul = resample(
  task = task_ul,
  learner = learner_ul,
  resampling = make_outer_resampling(task_ul, drive_ids_train=1:7, drive_ids_test=8:10),
  store_models = TRUE
)
```

```
## INFO  [11:25:24.207] Applying learner 'encode.regr.xgboost' on task 'ul_prediction' (iter 1/1)
## [11:25:25] WARNING: amalgamation/../src/objective/regression_obj.cu:174: reg:linear is now deprecate
```

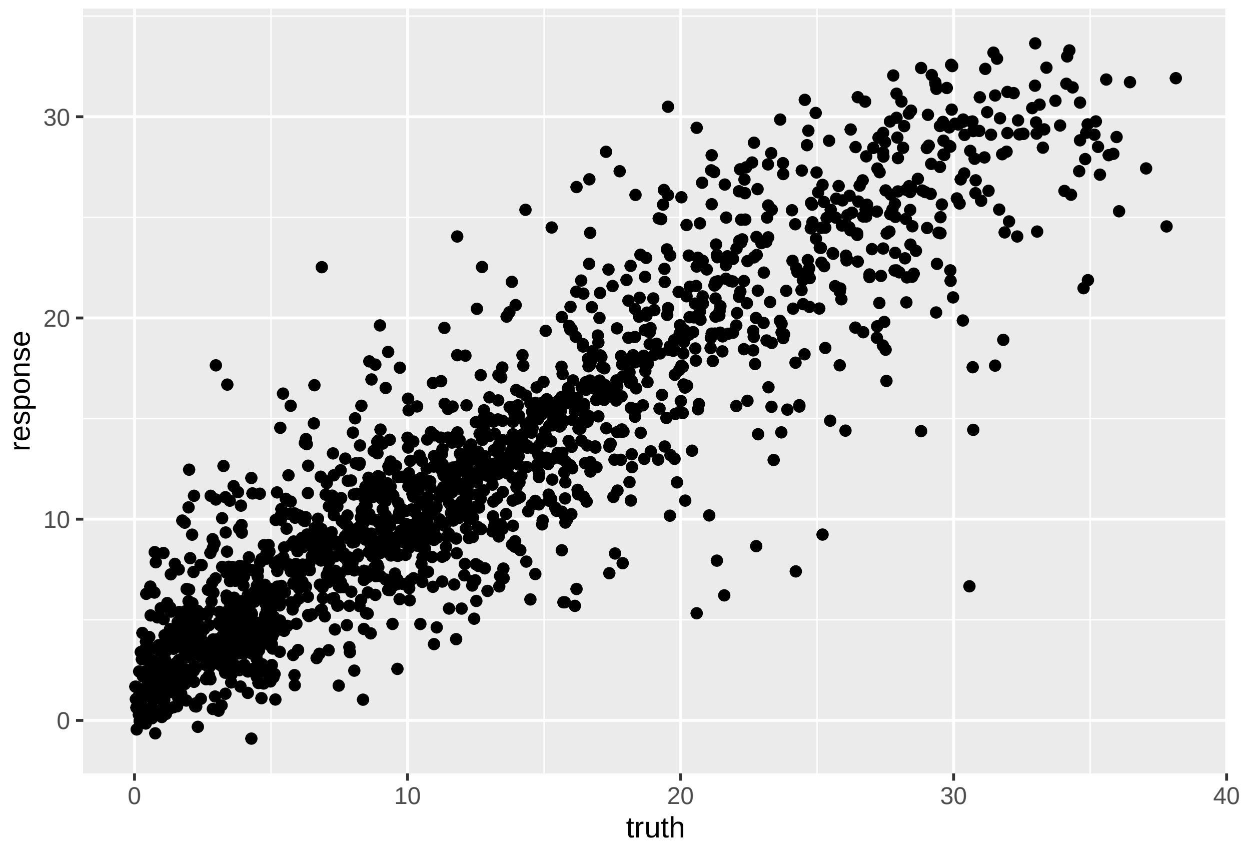
```
resampling_result_ul$aggregate(msr("regr.rsq"))
```

```
##  regr.rsq
## 0.8347534
```

```
resampling_result_ul$aggregate(msr("regr.mae"))
```

```
## regr.mae
## 2.600388
```

```
predictions_ul = as.data.table(resampling_result_ul$prediction())
ggplot(predictions_ul, aes(x=truth, y=response)) +
  geom_point()
```


response
30
20
10
0
0
10
20
30
40
truth