GBM

maycd

Contents

Import train

```r
default_train <- read.csv("default_train.csv", stringsAsFactors = TRUE)
dim(default_train)  # dataset: default_train, response: bad_good
```

```
## [1] 171171     14
```

GBM

```r
tic()
set.seed(123)
default_gbm <- gbm(
  formula = bad_good ~ .,
  data = default_train,
  distribution = "gaussian",  # SSE loss function
  n.trees = 3000,  # start with sufficiently large n.trees
  shrinkage = 0.1,
  interaction.depth = 5,
  n.minobsinnode = 10,
  cv.folds = 10
)
# find index for number trees with minimum CV error
best <- which.min(default_gbm$cv.error)

# get MSE and compute RMSE
sqrt(default_gbm$cv.error[best])
```

```
## [1] 0.0002510817
```

```r
toc()
```

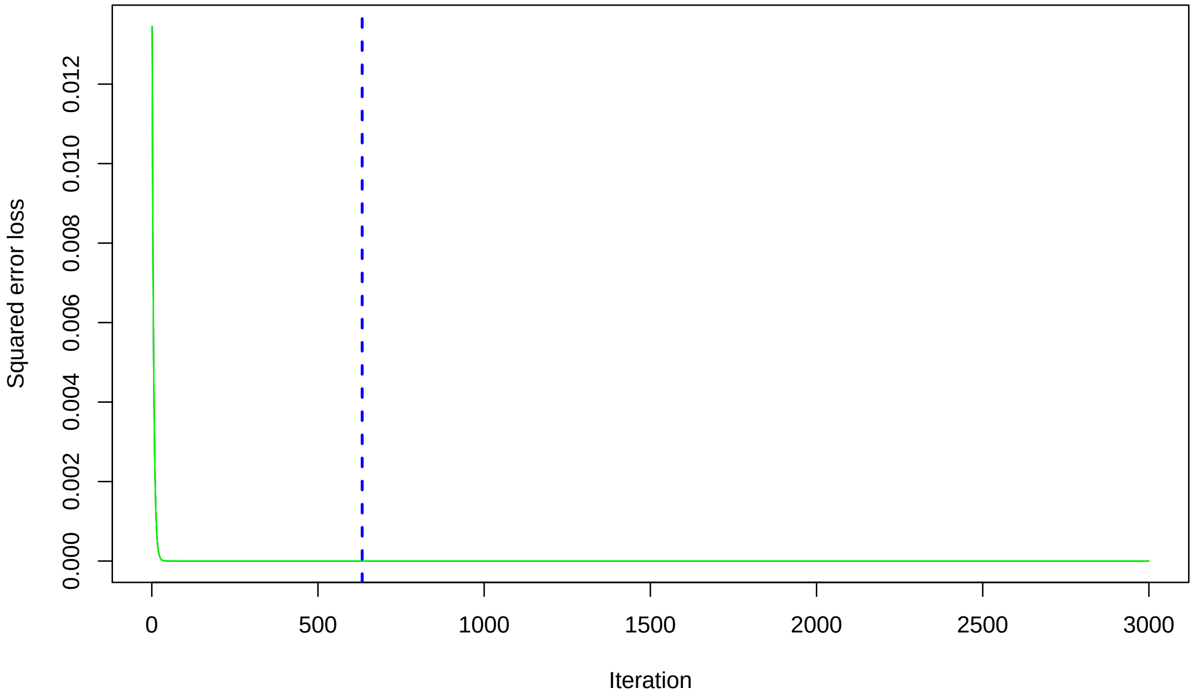
```
## 1549.5 sec elapsed
```

```
best <- which.min(default_gbm$cv.error)
sqrt(default_gbm$cv.error[best])
```

```
## [1] 0.0002510817
```

```
save(default_gbm, file = "default_gbm.rda")
```

```
# plot error curve
gbm.perf(default_gbm, method = "cv")
```



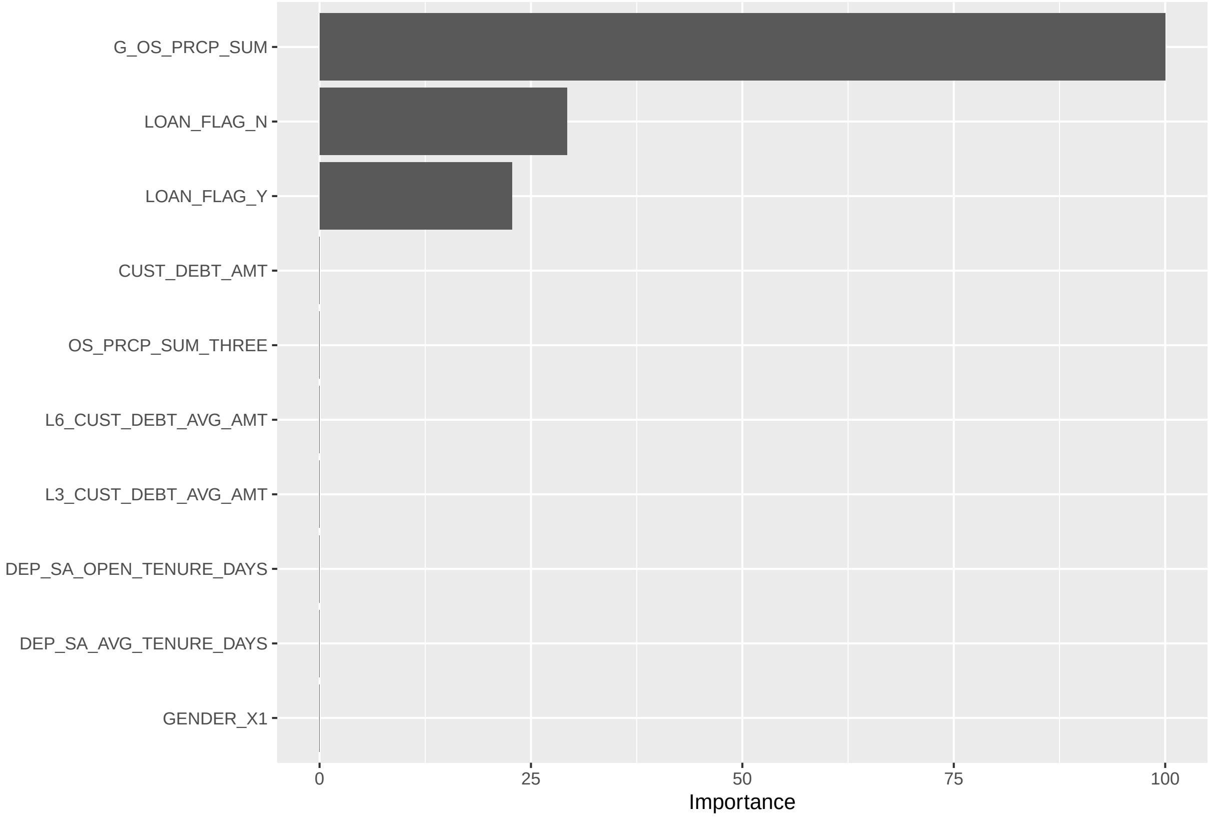
```
## [1] 633
```

Variable importance

```
vi_scores <- vi(default_gbm)
vi_scores
```

```
## # A tibble: 13 x 2
##    Variable               Importance
##    <chr>                       <dbl>
##  1 G_OS_PRCP_SUM             6.58e+ 1
##  2 LOAN_FLAG_N               1.92e+ 1
##  3 LOAN_FLAG_Y               1.50e+ 1
##  4 CUST_DEBT_AMT             8.80e- 4
##  5 OS_PRCP_SUM_THREE         7.25e- 4
##  6 L6_CUST_DEBT_AVG_AMT      2.80e- 4
##  7 L3_CUST_DEBT_AVG_AMT      1.10e- 4
##   8 DEP_SA_OPEN_TENURE_DAYS   5.51e- 5
##   9 DEP_SA_AVG_TENURE_DAYS    4.57e- 5
## 10 GENDER_X1                  1.12e- 6
## 11 GENDER_X2                  9.45e- 8
## 12 GENDER_X                   1.04e-10
## 13 OS_PRCP_SUM_SIX            0
```

```
vip(default_gbm, num_features = 10, scale = TRUE)
```



PDP plots

```
tic()
p1 <- partial(default_gbm, pred.var = vi_scores[[1, 1]], n.trees = 100) %>%
autoplot()
p2 <- partial(default_gbm, pred.var = vi_scores[[2, 1]], n.trees = 100) %>%
autoplot()
p3 <- partial(default_gbm, pred.var = vi_scores[[3, 1]], n.trees = 100) %>%
autoplot()
p4 <- partial(default_gbm, pred.var = vi_scores[[4, 1]], n.trees = 100) %>%
autoplot()
grid.arrange(p1, p2, p3, p4, ncol = 2)
```

```
## Warning: Use of `object[[1L]]` is discouraged. Use `.data[[1L]]` instead.
```

```
## Warning: Use of `object[["yhat"]]` is discouraged. Use `.data[["yhat"]]`
## instead.
## Warning: Use of `object[[1L]]` is discouraged. Use `.data[[1L]]` instead.
```

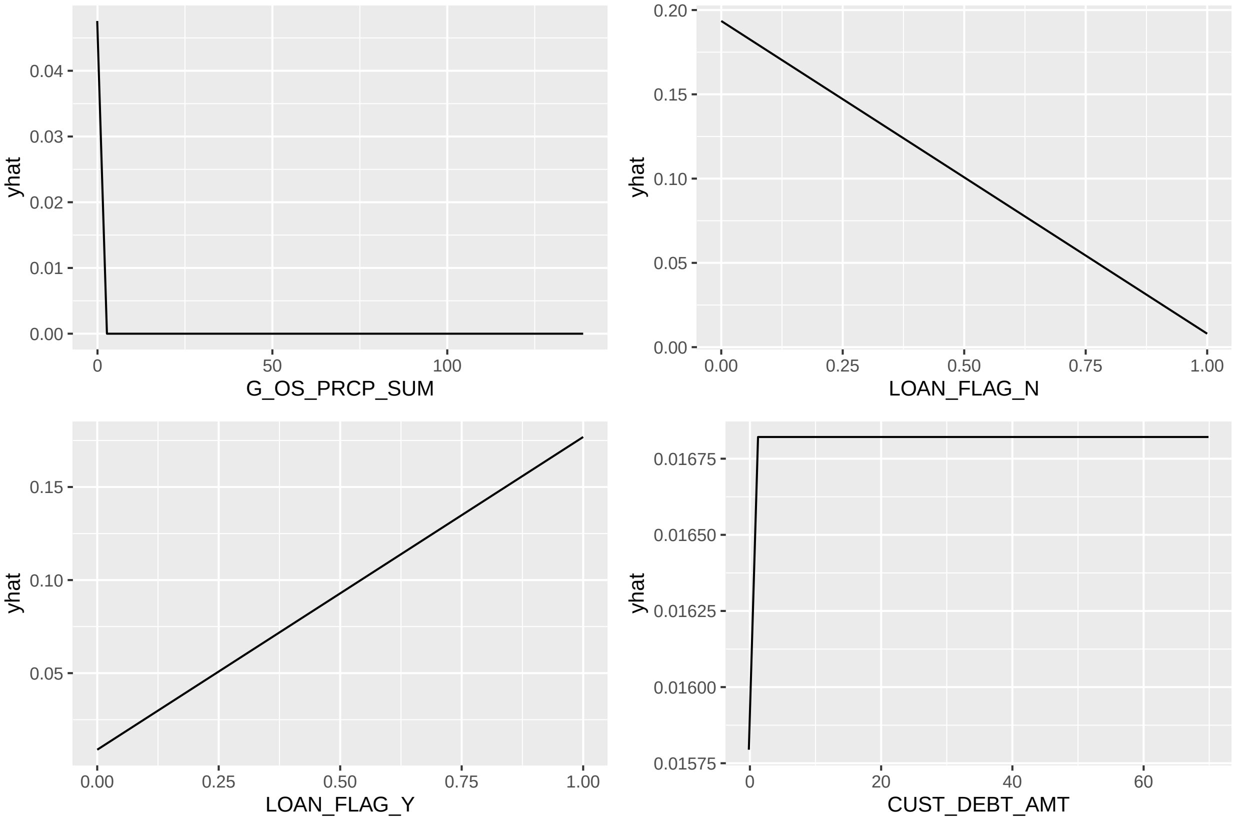
```
## Warning: Use of `object[["yhat"]]` is discouraged. Use `.data[["yhat"]]`
## instead.
```

```
## Warning: Use of `object[[1L]]` is discouraged. Use `.data[[1L]]` instead.
```

```
## Warning: Use of `object[["yhat"]]` is discouraged. Use `.data[["yhat"]]`
## instead.
```

```
## Warning: Use of `object[[1L]]` is discouraged. Use `.data[[1L]]` instead.
```

```
## Warning: Use of `object[["yhat"]]` is discouraged. Use `.data[["yhat"]]`
## instead.
```



```
toc()
```

```
## 0.36 sec elapsed
```