Jak dużo mocy (i po co) można wycisnąć z modelu predykcyjnego?

Artur Suchwałko, Tomasz Melcer, QuantUp

Why R?, 28 września 2017

QUANTUP

Wstęp

QUANTUP

1997

Information Value

IV Value	Usefullness
< 0.02	useless for prediction
0.02 - 0.10	weak predictor
0.10 - 0.30	medium predictor
0.30 - 0.50	strong predictor
> 0.50	suspicious or too good to be true

QUANTUP

2017

Information Value

IV Value	Usefullness
< 0.02	useless for prediction
0.02 - 0.10	weak predictor
0.10 - 0.30	medium predictor
0.30 - 0.50	strong predictor
> 0.50	suspicious or too good to be true

QUANTUP

Kod

```r
# "czysta" regresja logistyczna
model.log <- glm(
    loan_status ~ ., data = data_train, family = binomial)

# regresja logistyczna + wybór cech bazujący na IV
library(Information)
(IV <- create_infotables(
    data = data_train, y = "loan_status", parallel = F))

model_log_IV <- glm(
    loan_status ~ int_rate + grade + term + purpose,
    family = binomial, data = data_train)
```

QUANTUP

Jak zrobić lepiej?

- Co można poprawić:
  - Wybór cech — za prosto
  - Model — zbyt sztywny
  - Koszty błędnych decyzji — teraz takie same
  - Liczymy na sztuki: uwzględnienie kwoty (i kosztów) — teraz ignorujemy
  - Kryteria optymalizacji — MSE, Gini, Partial AUC
- Jak to poprawić?
- Po co poprawiać, tzn. jaki jest cel biznesowy?

QUANTUP

Lepszy wybór cech

QUANTUP

Lepszy wybór cech

- Information Value
- Forward-Backward (stepwise) Feature Selection
- Variable Importance Ranking (Random Forest)

QUANTUP

Kod: metoda krokowa

```
formula_big <- loan_status ~ .
formula_small <- loan_status ~ 1

model_start <- glm(
    formula_small, data = data_train, family = binomial)

logit_log_step <- step(
    model_start,
    scope = list(lower = formula_small, upper = formula_big),
    direction = "both")
```

QUANTUP

Kod: random Forest

```
library(randomForest)

model_rf <- randomForest(
    loan_status ~ ., data = data_train, importance = TRUE)

varImpPlot(model_rf, type = 2)

model_log_rf <- glm(
    loan_status ~ revol_util + dti + annual_inc + int_rate +
        emp_length + loan_amnt + total_acc,
    family = binomial("logit"), data = data_train)
```

QUANTUP

Bardziej złożony model

QUANTUP

Bardziej złożony model

- Logistic regression
- XGBoost i dobór hiperparametrów
  - domyślne
  - eksperckie
  - dostrojenie

QUANTUP

Kod: xgboost, zadane hiperparametry

```
model <- xgb.train(
    data=xgb.DMatrix(data_train, label=target),
    eval_metric='auc', nrounds=300, eta=0.03, max_depth=3,
    subsample=0.5, colsample_bytree=0.1,
    objective='binary:logistic')
```

QUANTUP

Kod: xgboost, dostrojenie hiperparametrów (grid search)

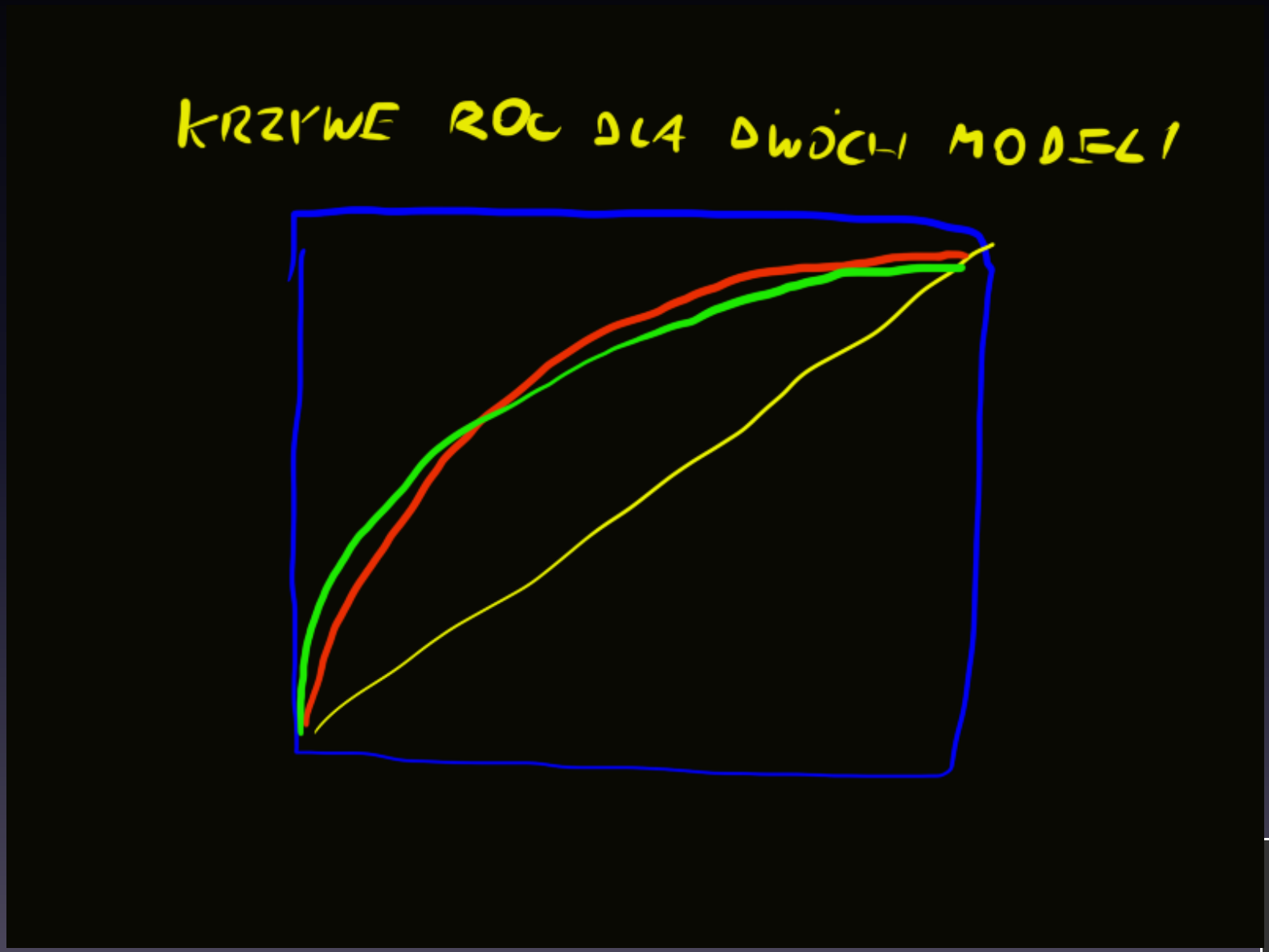
```
hp <- expand.grid(
    max_depth=3:6, gamma=c(0, 1, 2, 3, 5),
    eta=c(0.03, 0.06, 0.1, 0.2))

modele <- lapply(1:nrow(hp), function(idx) {
  model <- xgb.train(
    data=xgb.DMatrix(data_train, label=target),
    eval_metric='auc', nrounds=300, eta=hp$eta[idx],
    max_depth=hp$max_depth[idx], subsample=0.5,
    colsample_bytree=0.1, objective='binary:logistic',
    gamma=hp$eta[idx])
})

najlepszy <- modele[which.max(sapply(modele,
    function(m) auc(
        data_valid$loan_status,
        predict(m, data_valid))))]
```

QUANTUP

Zmiana kryterium optymalizacji

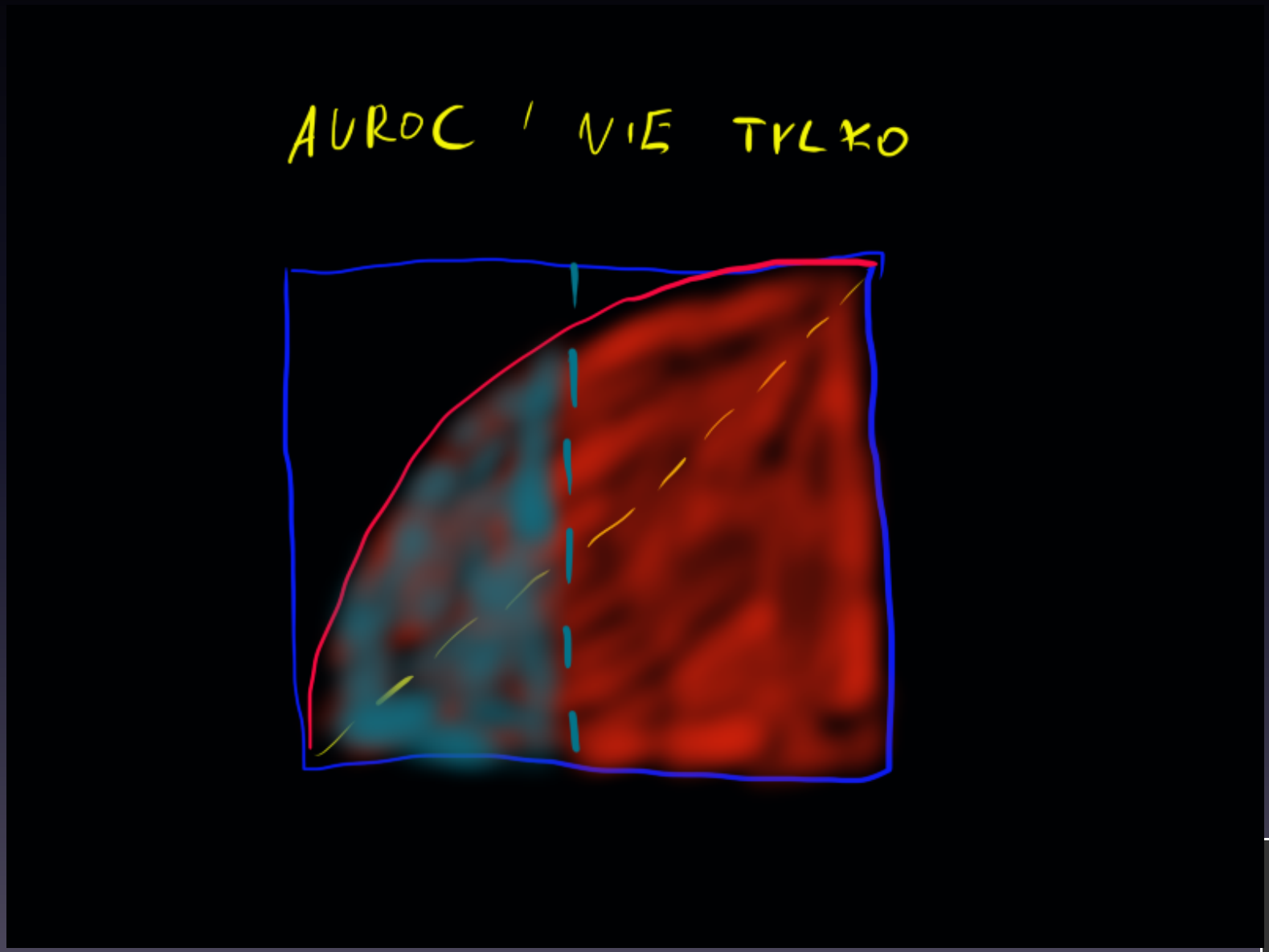
QUANTUP

Krzywe ROC



QUANTUP

Partial AUROC: bliżej celu biznesowego



QUANTUP

Zmiana kryterium optymalizacji w xgboost

- Nie można zaimplementować całkiem dowolnego kryterium
- Implementacja wybranego kryterium w xgboost wymaga dobrego zrozumienia metody
- Implementacja Partial AUC nie jest prosta – zmiana parametru / funkcji `objective`
- Ale sporo można uzyskać stosując wagi:
  - wyrównanie frakcji klas
  - koszt błędnych klasyfikacji

QUANTUP

Kod: zmiana kryterium za pomocą wag

```
wg_constant <- rep(1, length(nrow(LoanStats3a)))
wg_class <- ifelse(target, sum(1 - target), sum(target))
wg_cost <- ifelse(target,
    LoanStats3a$funded_amnt *
        ((1 + LoanStats3a$int_rate / 100) ^
         (LoanStats3a$term / 12) - 1),
    LoanStats3a$funded_amnt)
t_profit <- ifelse(target,
    LoanStats3a$funded_amnt *
        ((1 + LoanStats3a$int_rate / 100) ^
         (LoanStats3a$term / 12) - 1),
    -LoanStats3a$funded_amnt)

... <- xgb.train(..., weight=wg_cost, ...)
```

QUANTUP

Jakość kalibracji i zdolność dyskryminacyjna

- Można prognozować wartość (score)
  - możliwie równą p-stwu (`binary:logistic`) – *mierzenie*
  - po uporządkowaniu względem której 0 wyprzedzają 1 (`rank:pairwise`) – *porządkowanie*
- AUC: odsetek par, dla których kolejność wg score jest ok
- Gdy liczba jako target: odsetek par, dla których różnice są w dobrą stronę (czyli kolejność jest ok)
- Jak przejść z AUC jako funkcji celu do funkcji celu na jednym przypadku (niezbędne dla zmiany kryterium w xgboost):
  - Dla tego przypadku losujemy 10 przypadków i dostajemy surogat AUC (interpretacja probab. powyżej)
  - To jest „część AUC powiązana z tym przypadkiem”
- Wyznaczamy kierunek zmiany dotychczasowej predykcji itp.itd. (gradient, hesjan)

QUANTUP

Kod: xgboost, `rank:pairwise`

```
model <- xgb.train(
    data=xgb.DMatrix(data_train, label=t_profit),
    eval_metric='auc', nrounds=300, eta=0.03, max_depth=3,
    subsample=0.5, colsample_bytree=0.1,
    objective='rank:pairwise')
```

QUANTUP

Inne rozwiązanie

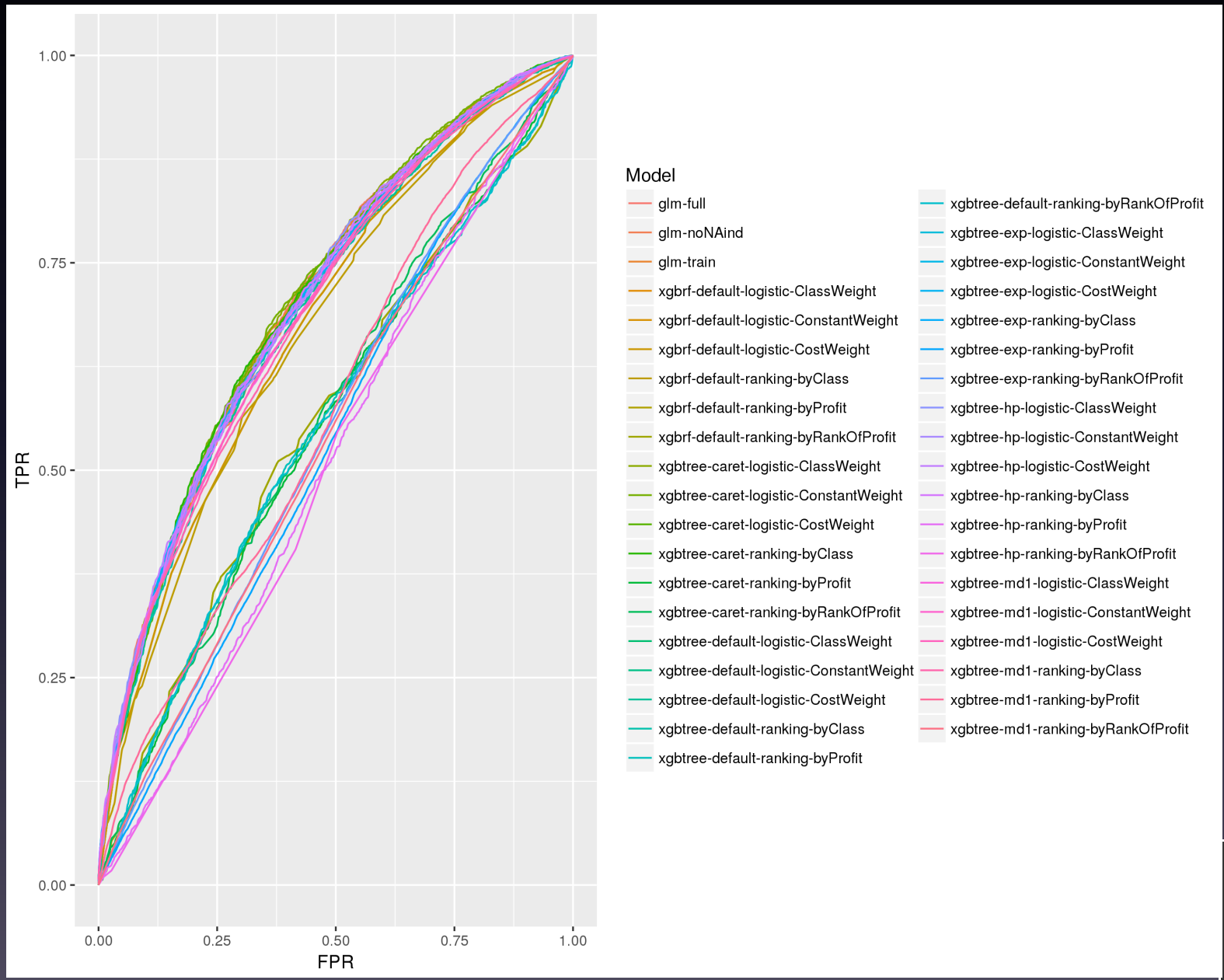
- Sieci neuronowe
- Niekoniecznie głębokie
- Kryteria w miarę dowolne, ale uwaga: nie dla wszystkich tak samo dobrze sieci się uczą
- Kiedyś stosowano modele programowania matematycznego (wprowadzanie ograniczeń biznesowych — czyli specjalne kryteria)
- To poza zakresem prezentacji

QUANTUP

Wyniki

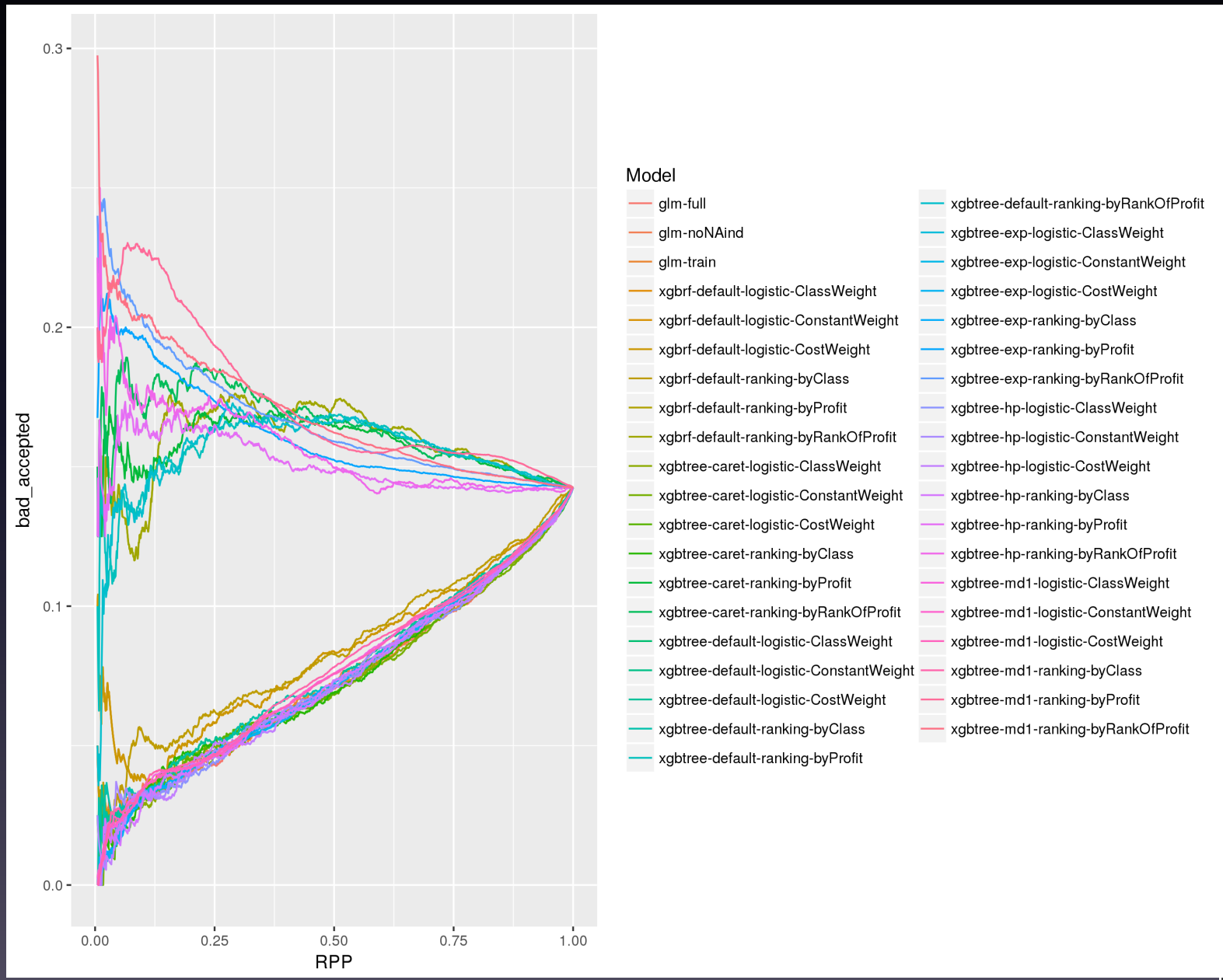
QUANTUP

ROC / AUC



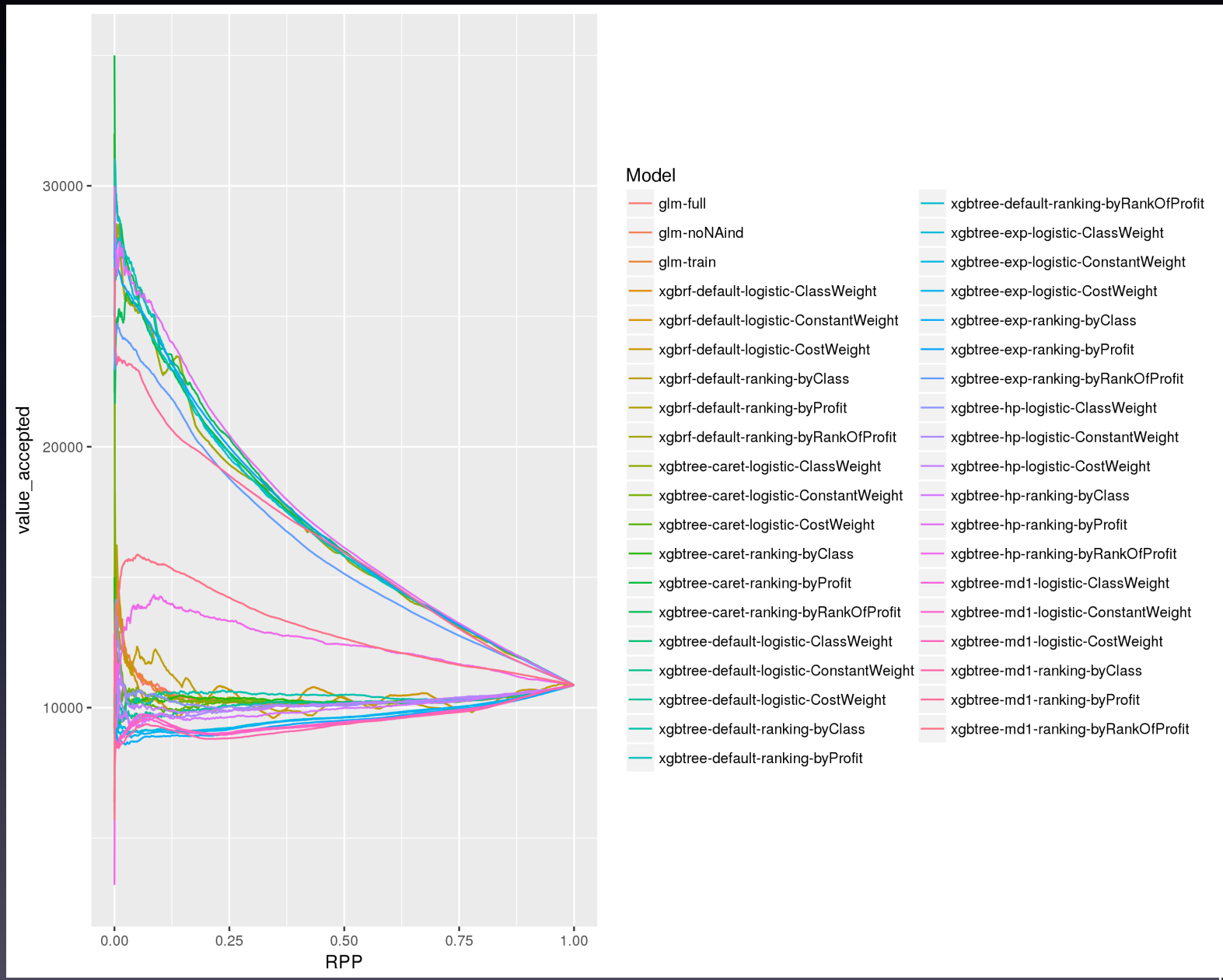
QUANTUP

Ryzyko (w %) dla zaakceptowanych



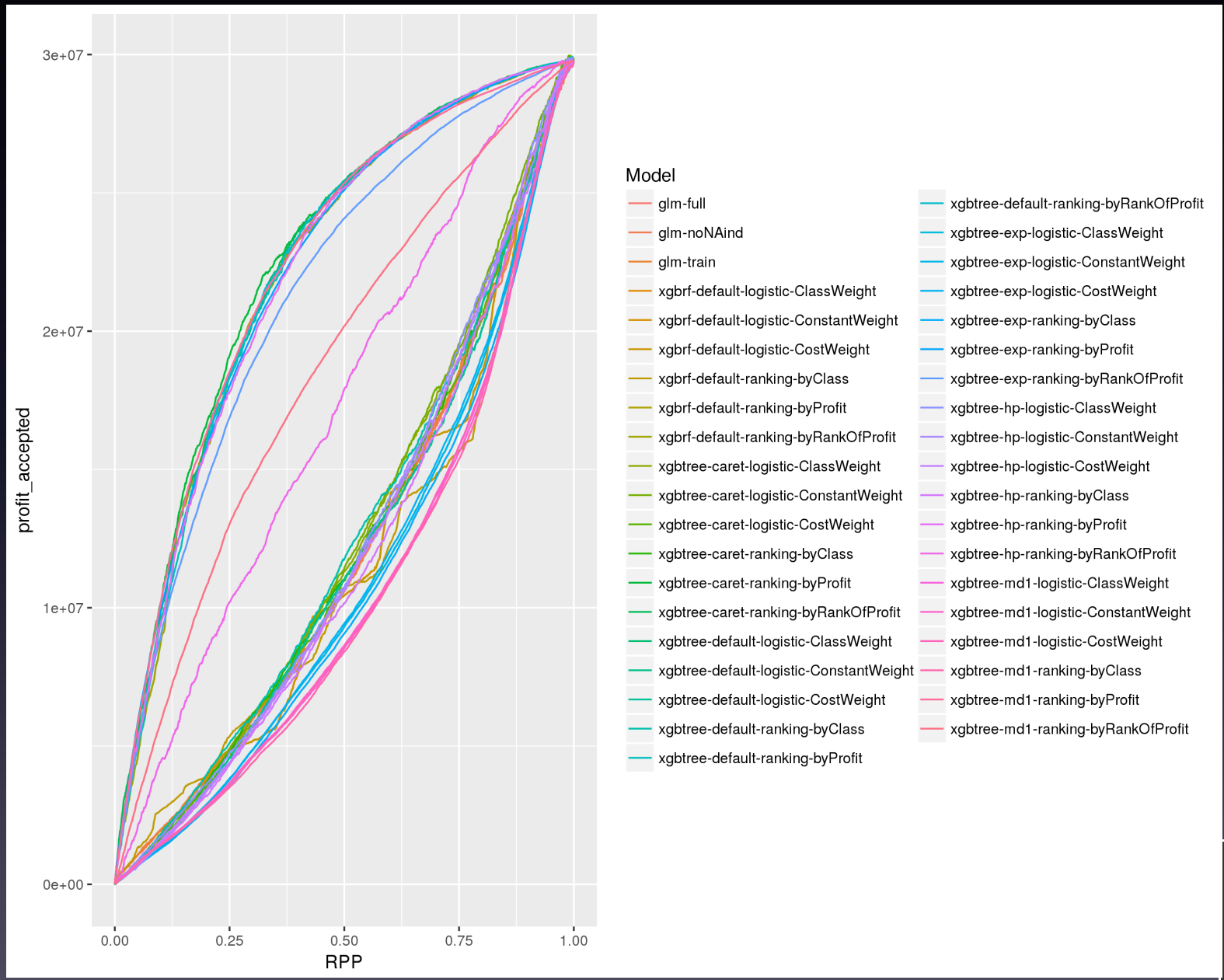
QUANTUP

Kwota (średnia) dla zaakceptowanych



QUANTUP

Zysk dla zaakceptowanych



QUANTUP

Podsumowanie

QUANTUP

Podsumowanie

- Złożoność modelu
- Kryterium optymalizacji, cel biznesowy, odpowiedniość
- Adekwatność metody do celu i zastosowań
- Zgodność kryterium optymalizacji we wszystkich etapach
- Wszystko (łatwo / możliwe) do zrobienia w R!
- Wpływ preprocessingu: kodowanie NA, false predictors, factors
- Inne zagadnienia i dane (churn, fraud) dają inne:
  - kolejność modeli
  - wielkości różnic między skutecznością modeli

QUANTUP

Kontakt

QUANTUP

Kontakt

- Podczas konferencji
- artur *[at]* quantup *[dot]* pl

QUANTUP