

# Predviđanje ponašanja klijenata banke

Predaja rješenja u sklopu kolegija *Strojno učenje*

Tim *Petty*

Prirodoslovno-matematički fakultet – Matematički odsjek  
Sveučilište u Zagrebu

Zagreb, lipanj 2019.

# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

- predviđanje eventualnog prijevremenog raskida ugovora o kreditu/depozitu (*RBA*) — **binarna klasifikacija**

# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

- predviđanje eventualnog prijevremenog raskida ugovora o kreditu/depozitu (*RBA*) — **binarna klasifikacija**
- 12 značajki

# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

- predviđanje eventualnog prijevremenog raskida ugovora o kreditu/depozitu (*RBA*) — **binarna klasifikacija**
- 12 značajki
  - 2 identifikacijske značajke

# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

- predviđanje eventualnog prijevremenog raskida ugovora o kreditu/depozitu (*RBA*) — **binarna klasifikacija**
- 12 značajki
  - 2 identifikacijske značajke
  - 5 kategorijskih značajki

# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

- predviđanje eventualnog prijevremenog raskida ugovora o kreditu/depozitu (*RBA*) — **binarna klasifikacija**
- 12 značajki
  - 2 identifikacijske značajke
  - 5 kategorijskih značajki
  - 5 numeričkih odnosno vremenskih značajki

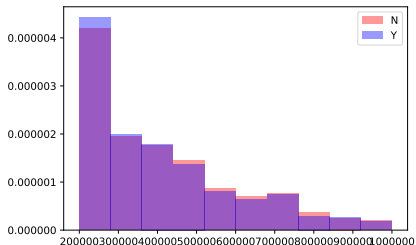
# Problem – rekapitulacija

Natjecanje *Mozgalo* 2019. godine

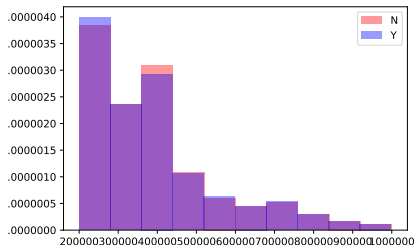
- predviđanje eventualnog prijevremenog raskida ugovora o kreditu/depozitu (*RBA*) — **binarna klasifikacija**
- 12 značajki
  - 2 identifikacijske značajke
  - 5 kategorijskih značajki
  - 5 numeričkih odnosno vremenskih značajki
- $5 \cdot 10^6$  primjera za treniranje **prije spljoštenja**



# Distribucije značajki



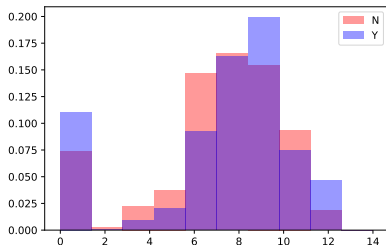
(a) 'A'



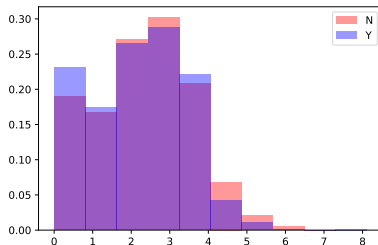
(b) 'L'

Slika: Ugovoreni iznos

# Distribucije značajki



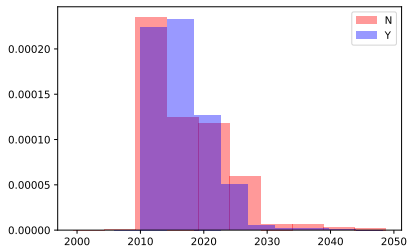
(a) 'A'



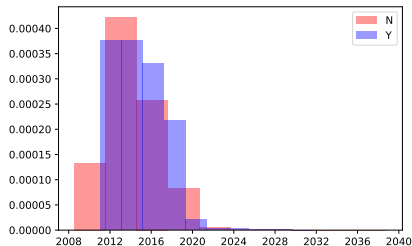
(b) 'L'

Slika: Visina kamate

# Distribucije značajki



(a) 'A'



(b) 'L'

Slika: Planirani datum zatvaranja

## ① spljoštenje

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP



## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija
- nezaposlenost

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija
- nezaposlenost
- cijena nafte

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija
- nezaposlenost
- cijena nafte

## ③ kombinacije značajki

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija
- nezaposlenost
- cijena nafte

## ③ kombinacije značajki

- trajanje, trajanje u krizi, promjene značajki

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija
- nezaposlenost
- cijena nafte

## ③ kombinacije značajki

- trajanje, trajanje u krizi, promjene značajki
- kamatni račun

## ① spljoštenje

- svi primjeri s istom oznakom partije spljošte se u 1 primjer
- pamte se prva i zadnja varirajuća značajka

## ② makroekonomske značajke

- BDP
- inflacija
- nezaposlenost
- cijena nafte

## ③ kombinacije značajki

- trajanje, trajanje u krizi, promjene značajki
- kamatni račun
- matematički račun nad značajkama baziran na statistici i intuiciji

- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja



- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja
- klasa `CatBoostClassifier`

- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja
- klasa `CatBoostClassifier`
- 3 modela

- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja
- klasa `CatBoostClassifier`
- 3 modela
  - za kredite do 6. listopada 2016.

- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja
- klasa `CatBoostClassifier`
- 3 modela
  - za kredite do 6. listopada 2016.
  - za depozite do 6. listopada 2016.

- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja
- klasa `CatBoostClassifier`
- 3 modela
  - za kredite do 6. listopada 2016.
  - za depozite do 6. listopada 2016.
  - za ostale ugovore

- *Python* biblioteka za gradijentno *jačanje* (eng. *gradient boosting*) na stablima odlučivanja
- klasa `CatBoostClassifier`
- 3 modela
  - za kredite do 6. listopada 2016.
  - za depozite do 6. listopada 2016.
  - za ostale ugovore
- svi su modeli konstruirani simetričnim stablima — **interpretabilnost**

## Kôd 1: Primjer konstrukcije modela

```
model = CatBoostClassifier(  
    iterations = 1000,  
    learning_rate = 0.873,  
    depth = 9,  
    l2_leaf_reg = 743.5,  
    border_count = 168,  
    od_type = 'Iter',  
    leaf_estimation_method = 'Newton',  
    random_seed = 934,  
    random_strength = 1.419,  
    bagging_temperature = 0.415,  
    task_type = 'GPU',  
    sampling_unit = 'Group'  
)
```

# Hiperparametri

## Kôd 2: Primjer treniranja modela

```
train_pool = Pool(train_X, train_y, cat_features =  
    categoricalia)  
test_pool = Pool(test_X, test_y, cat_features =  
    categoricalia)  
  
model.fit(  
    train_pool,  
    eval_set = test_pool,  
    verbose = False,  
    plot = True,  
    early_stopping_rounds = 50  
)  
  
model.save('model')
```



## *Hyperopt*

# Optimizacija hiperparametara

## *Hyperopt*

- *Python* biblioteka za optimizaciju hiperparametara

# Optimizacija hiperparametara

## *Hyperopt*

- *Python* biblioteka za optimizaciju hiperparametara
- pogodna za kompleksne prostore pretraživanja

# Optimizacija hiperparametara

## *Hyperopt*

- *Python* biblioteka za optimizaciju hiperparametara
- pogodna za kompleksne prostore pretraživanja
- realne, diskretne i uvjetne domene

## *Hyperopt*

- *Python* biblioteka za optimizaciju hiperparametara
- pogodna za kompleksne prostore pretraživanja
- realne, diskretne i uvjetne domene
- *Bayesovska* optimizacija

# Rezultati

## Konfuzijske tablice

Tablica: Model A

|            |          | Stvarno     |               |               |
|------------|----------|-------------|---------------|---------------|
|            |          | <i>N</i>    | <i>Y</i>      |               |
| Predikcija | <i>N</i> | <b>7408</b> | 2349          | 9757          |
|            | <i>Y</i> | 2833        | <b>12 143</b> | 14 976        |
|            |          | 10 241      | 14 492        | <b>24 733</b> |

# Rezultati

## Konfuzijske tablice

Tablica: Model L

|            |          | Stvarno     |               |
|------------|----------|-------------|---------------|
|            |          | <i>N</i>    | <i>Y</i>      |
| Predikcija | <i>N</i> | <b>8066</b> | 1510          |
|            | <i>Y</i> | 2430        | <b>14 800</b> |
|            |          | 10 496      | 16 310        |
|            |          |             | <b>26 806</b> |

# Rezultati

## Konfuzijske tablice

Tablica: Ostali

|            |          | Stvarno       |             |               |
|------------|----------|---------------|-------------|---------------|
|            |          | <i>N</i>      | <i>Y</i>    |               |
| Predikcija | <i>N</i> | <b>10 748</b> | 3700        | 14 448        |
|            | <i>Y</i> | 2598          | <b>7154</b> | 9752          |
|            |          | 13 346        | 10 854      | <b>24 200</b> |



# Rezultati

## Konfuzijske tablice

Tablica: Ukupno

|            |          | Stvarno       |               |               |
|------------|----------|---------------|---------------|---------------|
|            |          | <i>N</i>      | <i>Y</i>      |               |
| Predikcija | <i>N</i> | <b>26 222</b> | 7559          | 33 781        |
|            | <i>Y</i> | 7861          | <b>34 097</b> | 41 958        |
|            |          | 34 083        | 41 656        | <b>75 739</b> |

# Rezultati

Vlastiti validacijski *dataset*

Tablica: Evaluacijske mjere modela

| Model         | Točnost | Preciznost | Odziv  | $F_1$  |
|---------------|---------|------------|--------|--------|
| <b>A</b>      | 85,3 %  | 85,9 %     | 90,7 % | 88,3 % |
| <b>L</b>      | 74,0 %  | 73,4 %     | 66,0 % | 69,4 % |
| <b>Ostali</b> | 79,0 %  | 81,1 %     | 83,8 % | 82,4 % |
| <b>Ukupno</b> | 79,6 %  | 81,3 %     | 81,9 % | 81,6 % |

# Rezultati

Mozgalo 2019. – evaluacijski i validacijski *dataset*

Tablica: Rezultati na natjecanju

| ◆          | Točnost | $F_1$ | Ostvareni bodovi |
|------------|---------|-------|------------------|
| Evaluacija | 71 %    | 77 %  | 14/15            |
| Validacija | 69 %    | 70 %  | 17/20            |
|            |         |       | <b>31/35</b>     |

# Ukupni plasman

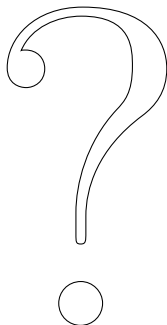
Nismo ušli u finale — bili smo 8., a 6 je finalista

# Ukupni plasman

Nismo ušli u finale — bili smo 8., a 6 je finalista



# Komentari i pitanja



## Implementacija ✈️

-  Tomislav Šmuc, Tomislav Lipić i Matija Piškorec. *Materijali za strojno učenje*. 2019. URL: <http://web.math.pmf.unizg.hr/nastava/su/materijali/> (pogledano 9.6.2019).
-  Trevor Hastie, Robert Tibshirani i Jerome Harold Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.
-  Hrvatska narodna banka. *Statistics – HNB*. 2019. URL: <http://www.hnb.hr/statistika> (pogledano 9.6.2019).
-  International Monetary Fund. *IMF Data*. 2019. URL: <http://www.imf.org/en/Data> (pogledano 9.6.2019).





Andrew Ng. *Machine Learning*. 2019. URL:  
<http://www.coursera.org/learn/machine-learning>  
(pogledano 9.6.2019).



Republika Hrvatska. *Državni zavod za statistiku*. 2019. URL:  
<http://www.dzs.hr/> (pogledano 9.6.2019).