

# I spot a bot

Projekat iz predmeta Mašinsko učenje

Matematički fakultet

Student: Nemanja Antić 1100/2017

Profesor: dr Mladen Nikolić

Asistent: Anđelka Zečević

- ▶ Analiza
- ▶ Preprocesiranje
- ▶ Učenje modela
- ▶ Rezultati

# Preprocesiranje

- ▶ Podaci prikupljeni iz dataseta: „cresci-2015“ i „cresci-2017“
- ▶ Sadrže podatke o user-ima, njihovim tweet-ovima
- ▶ Niz transformacija:
  - ▶ 10 atributa
  - ▶ Podaci direktno iz dataseta
    - ▶ Followers count
    - ▶ Friends count
    - ▶ Favorites count
  - ▶ Izracunati podaci
    - ▶ Mentions per tweet
    - ▶ Urls per tweet
    - ▶ Hashtags per tweet
    - ▶ Retweets per tweet
    - ▶ Favorites per tweet
    - ▶ Frequency of tweets (in minutes)
    - ▶ Friends to followers ratio

► Ukupan broj user-a

```
In [29]: data.shape
```

```
Out[29]: (14533, 10)
```

► Raspodela klasa

```
print("Bots: {bot}\nHuman: {human}".format(bot = len(y[y == 1]), human = len(y[y == 0])))
```

```
Bots: 11821
```

```
Human: 2712
```

# Učenje modela

- ▶ Logistička regresija
- ▶ Neuronska mreža sa propagacijom unapred

# Logistička regresija

- ▶ Skup je podeljen na trening i test u razmeri 2:1
  - ▶ Trening skup je podeljen na trening i validacioni u razmeri 4:1
- ▶ Podaci su standardizovani pomoću StandardScaler objekta iz sclearn bibl
- ▶ Korišćena je L2 regularizacija
- ▶ Istreniran model:
  - ▶ `LogisticRegression(C=1.1, class_weight='balanced', dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1, penalty='l2', random_state=None, solver='liblinear', tol=0.0001, verbose=0, warm_start=False)`

► Rezultati „cross validation“ metode:

```
score
```

```
array([0.98762887, 0.99312242, 0.95870613, 0.98485891, 0.99242946,  
       0.97040606, 0.87611838, 0.96971783, 0.96971783, 0.68203716])
```

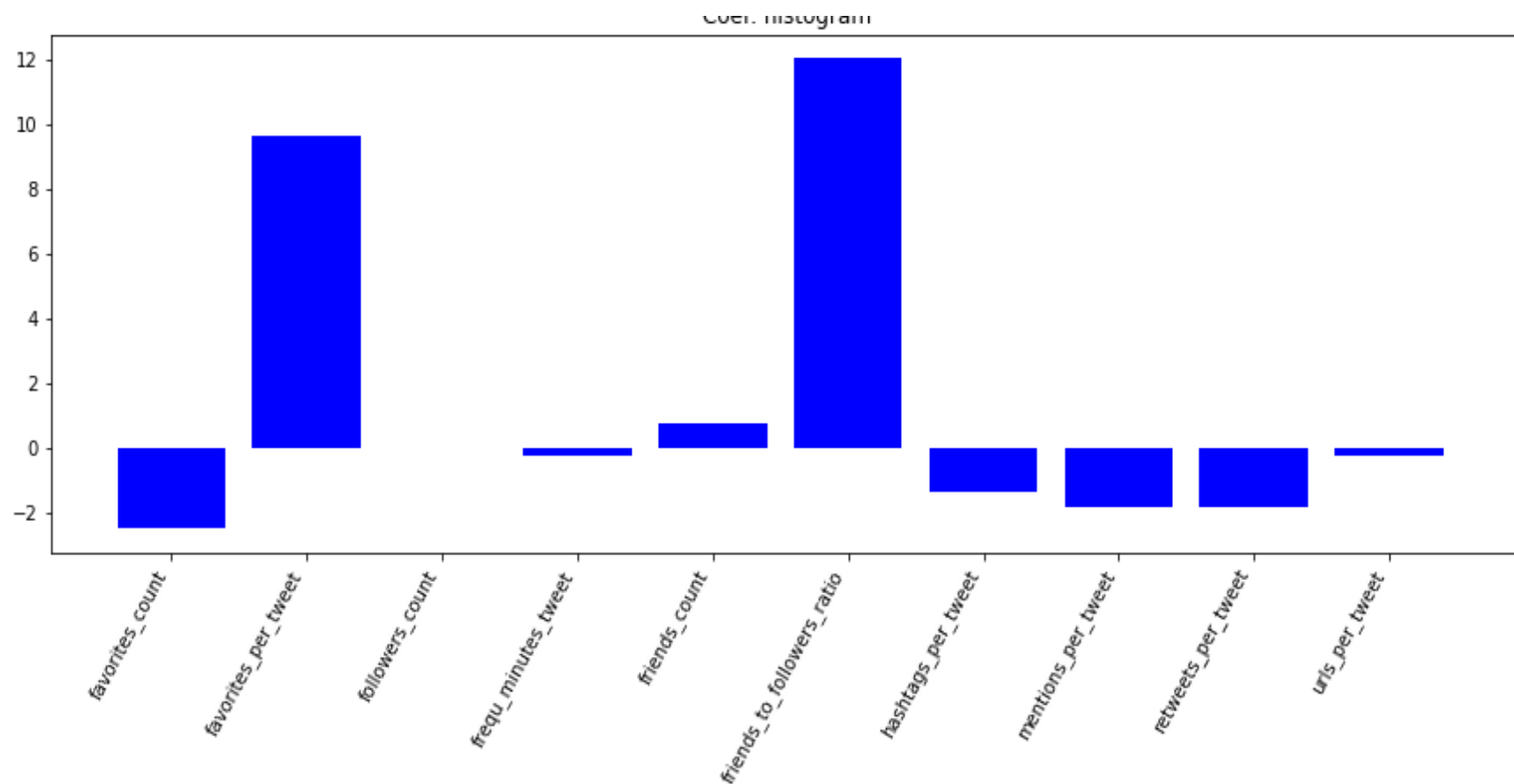
```
score.mean()
```

```
0.9384743028112273
```

```
score.std()
```

```
0.09137550271425189
```

## Koeficijenti modela

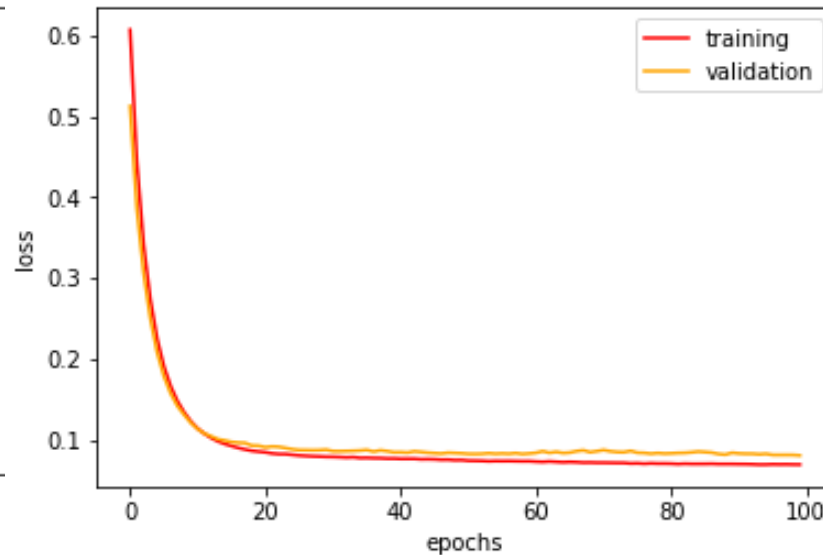
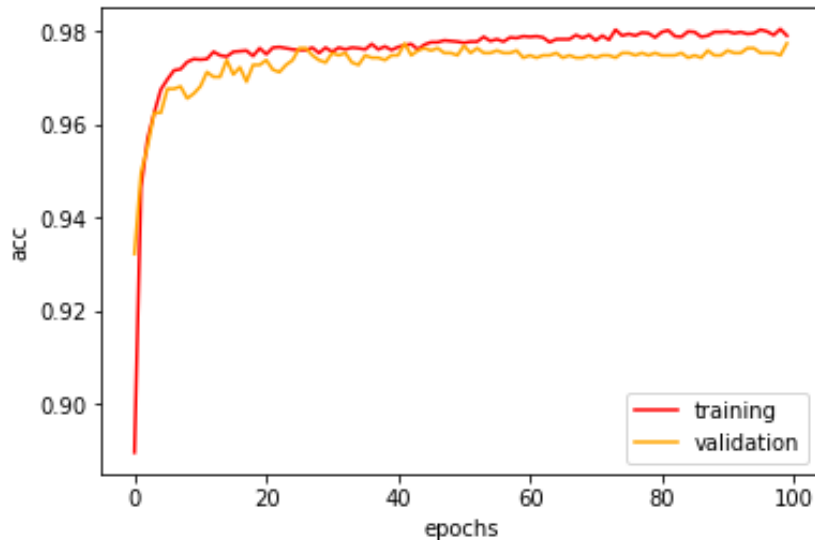




# Neuronska mreža sa propagacijom unapred

- ▶ 2 skrivena sloja:
  - ▶ Sa 4 čvora
  - ▶ Sa 3 čvora
- ▶ Test loss: ~0.077
- ▶ Test accuracy: ~0.97

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 4)	44
dense_2 (Dense)	(None, 3)	15
dense_3 (Dense)	(None, 1)	4
Total params: 63		
Trainable params: 63		
Non-trainable params: 0		



► Projekat:

<https://github.com/Cubaka14/-I-Spot-a-Bot---Building-a-Binary-Classifer-to-Detect-Bots-on-Twitter>

► Podaci:

<https://botometer.iuni.iu.edu/bot-repository/datasets.html>

► Literatura:

<http://cs229.stanford.edu/proj2017/final-reports/5240610.pdf>

<https://developer.twitter.com/en/docs/tutorials>

Hvala na paznji!