Predicting the results of ballroom dancing competitions

Matyáš Mattanelli

Outline

- Introduction
- Data
- Methodology
- Results

Introduction

- A couple (male and female) dancing together
- Standard/Latin
- Czech Dance Sport Federation
- Classes: A, B, C, D, E
- Weekly competitions
- Adjudicators decide the final rankings
- Need 200 points and 5 **“Finals”** to advance to the next class
- **My goal:** Predict the probability of obtaining a “Final”

03.12.2023 - Zimní SPARTA CUP 2023 - Praha

Dospělí-D-LAT (počet párů: 9) [postupová soutěž]

Rank	Number	Couple	Club (country)	Obtained points ("Finals")	Points ("Finals") after competition	New class	
Finále							
1	106	[redacted]	TŠ Easy Dance 2000 Nymburk (CZ)	25(F)	192(F3)		
2	117	Mattanelli Matyáš & Šupíková Dorota	STK Praha (CZ)	19(F)	216(F5)	C	
3	118	[redacted]	STK Praha (CZ)	13	174(F2)		
4	105		LR Cosmetic Dance Team Ostrava (CZ)	7	44(F0)		
5	85		Top Dance Prague Team (CZ)	6	61(F0)		
6	73		TK Sparta Praha (CZ)	5	16(F0)		
Semifinále							
7	75		TK Astra Praha (CZ)	2	174(F5)		
8	82		Taneční klub FIS (CZ)	1	37(F0)		
9	79		Chomutov (CZ)	0	22(F0)		

Introduction

Number of couples in the competition	Ranks that obtain a “Final”
2	-
3 – 5	1
6 – 10	1 – 2
11 – 15	1 – 3
16 – 20	1 – 4
21 – 25	1 – 5
26 and more	All finalists

Data

Scraping

- Obtaining URLs to all competitions
- Each competition on a separate URL

Processing

- Unifying column names
- Concatenating competitions
- Filtering (duplicates, less than 3 participants, etc.)

Feature engineering

- Using data from previous competitions to create features relevant for the current competition

Final data set

- Complete data set prepared for estimation

Data

Each line in the results of a competition represents an observation

Competitions for all classes and all age categories considered

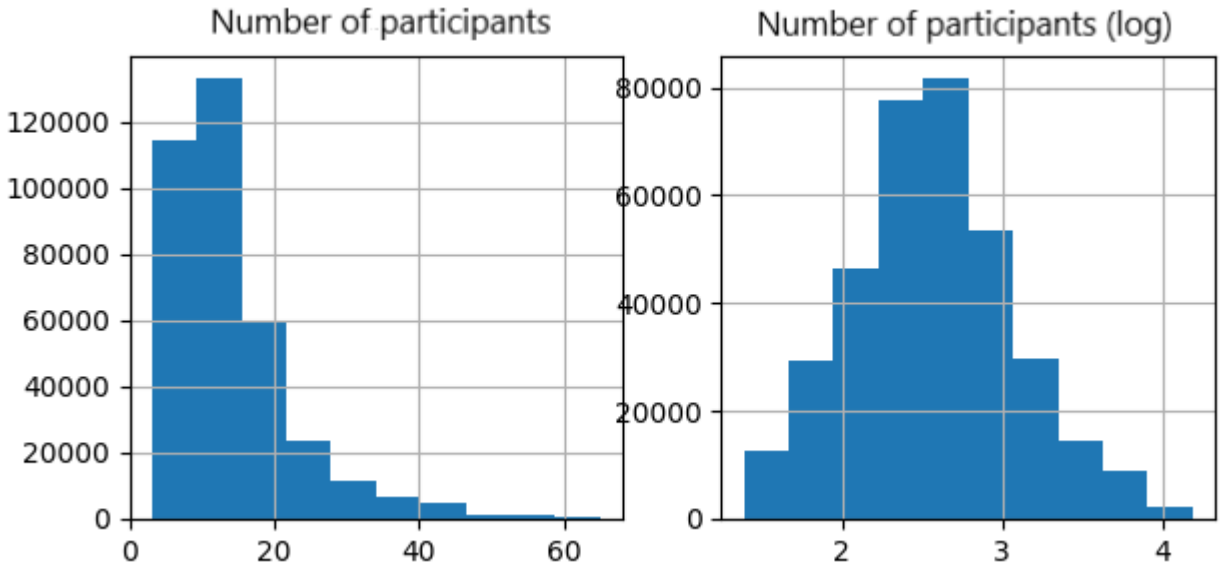
2001 – 2023

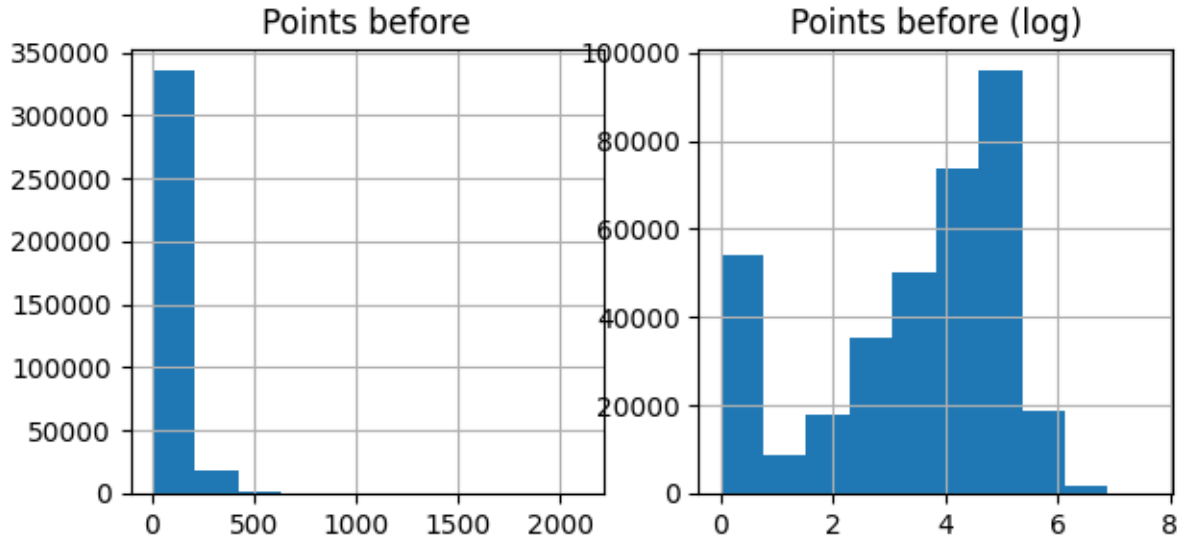
Base features: Club (binary), Country (binary), Number of participants, Number of points (and finals) before competition

Engineered features: Average points (and finals) obtained in given class, Number of participations in a class, Average rank in a class, etc.

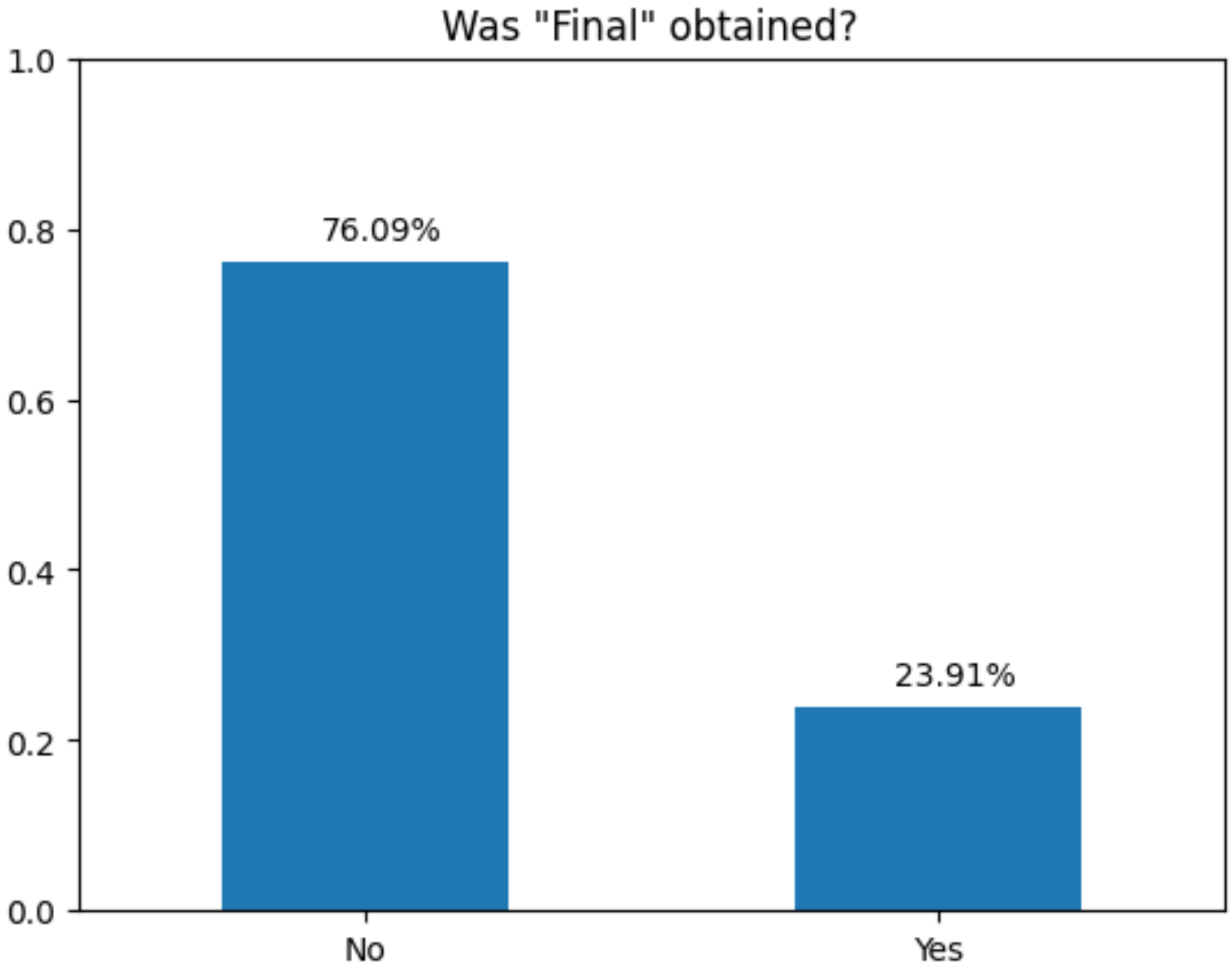
356 124 observations, 14 features (12 numerical, 2 categorical)

Data


Number of participants
120000
100000
80000
60000
40000
20000
0
0
20
40
60
Number of participants (log)
80000
60000
40000
20000
0
2
3
4


Points before
350000
300000
250000
200000
150000
100000
50000
0
0
500
1000
1500
2000
Points before (log)
100000
80000
60000
40000
20000
0
0
2
4
6
8

Data



Methodology

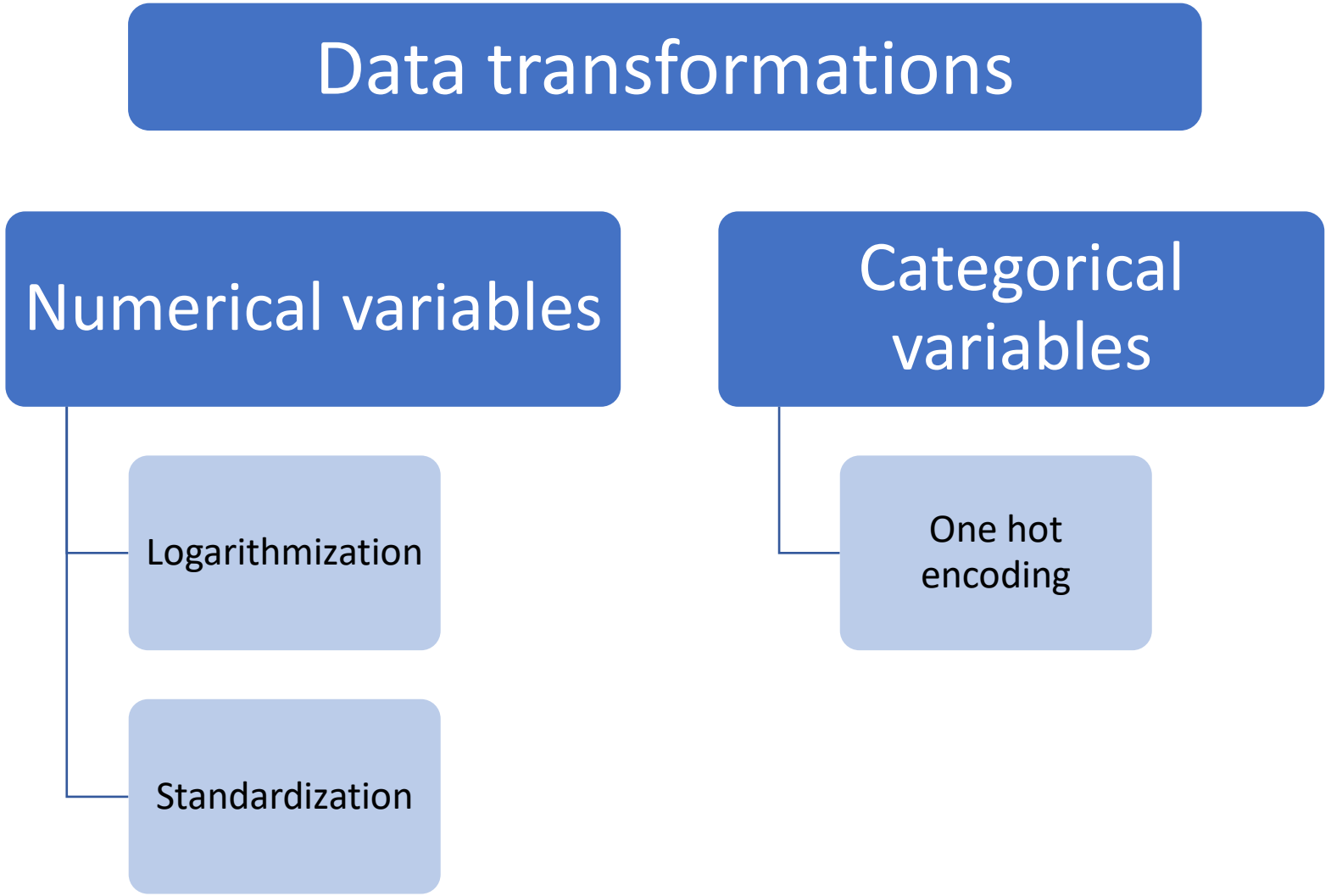
Univariate analysis

- Basic characteristics (mean, min, max, etc.)
- Histogram – normality of the distribution
- Univariate logistic regression

Bivariate analysis

- Pairwise correlation

Methodology



Methodology

Main model

- Feed-forward artificial neural network
- Three layers: input, hidden, output
- Optimization method: Adam
- Hyperparameter optimization through cross-validation

Secondary model

- Logistic regression

Methodology

Hyperparameter optimization

3-fold cross-validation

Evaluation metric: AUC

Parameters

- **Neural network**
  - Hidden layer size
  - Activation function (ReLU, tanh, sigmoid)
  - L2 regularization strength
- **Logistic regression**
  - Regularization type (None, L1, L2)
  - Regularization strength

Methodology

Performance evaluation

Class imbalance => Accuracy not suitable

Area Under the Receiver Operating Characteristic Curve (AUC)

- **ROC:** True Positive Rate against False Positive Rate
- AUC ranges from 0 (worst) to 1 (best)
- Values above 0.7 considered decent
- Not biased by class imbalance
- Evaluates how well the model differentiates between the classes

Data split into training (80%) and testing (20%)

Results

Univariate analysis

	AUC	
	Base	**With squared term**
Average points	0.770	0.782
Average finals	0.741	0.741
Finals before	0.713	0.713
Average rank in current class	0.681	0.712
Points before	0.670	0.700
Previous number of participations	0.555	0.569
Days in previous class	0.551	0.557
Number of participations	0.540	0.555
Number of participants	0.536	0.536
Days since first comp	0.535	0.541
Days in class	0.516	0.531
First comp in class	0.512	

Results

Optimal parameters

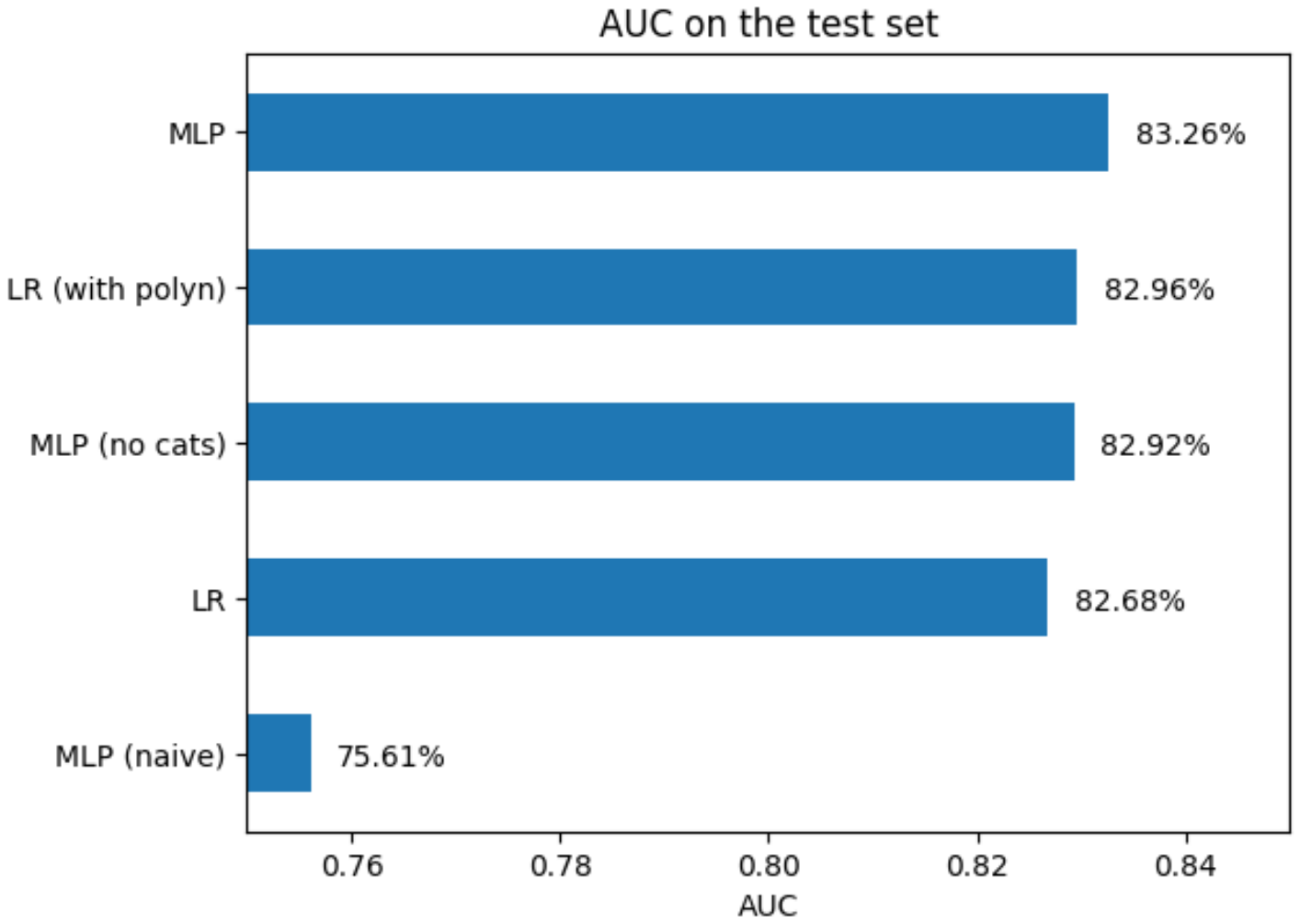
Neural network

- **Hidden layer size**: 100
- **Activation function**: tanh
- **L2 strength**: 0.01

Logistic regression

- **Regularization**: L2
- **Strength**: 10

Results


AUC on the test set
MLP
LR (with polyn)
MLP (no cats)
LR
MLP (naive)
83.26%
82.96%
82.92%
82.68%
75.61%
0.76
0.78
0.80
0.82
0.84
AUC

Results

Variable significance

	Coefficient	Significant at 5%
Average points	1.47	✓
Average rank in current class	-1.14	✓
Finals before	0.46	✓
Number of participants	-0.4	✓
Days in class	0.18	✓
Number of participations	0.16	✓
Country_CZ	-0.87	✓
Country_SVK	-0.27	✗

Results

Example

Rank	Couple	Club	Obtained points	Obtained final	Prediction
1	-	TŠ Easy Dance 2000 Nymburk	25	1	0.3
2	Mattanelli Matyáš & Šupíková Dorota	STK Praha	19	1	0.37
3	-	STK Praha	13	0	0.4
4	-	LR Cosmetic Dance Team Ostrava	7	0	0.1
5	-	Top Dance Prague Team	6	0	0.08
6	-	TK Sparta Praha	5	0	0.05
7	-	TK Astra Praha	2	0	0.51
8	-	Taneční klub FIS	1	0	0.08
9	-	Chomutov	0	0	0.04

Thank You for Your attention