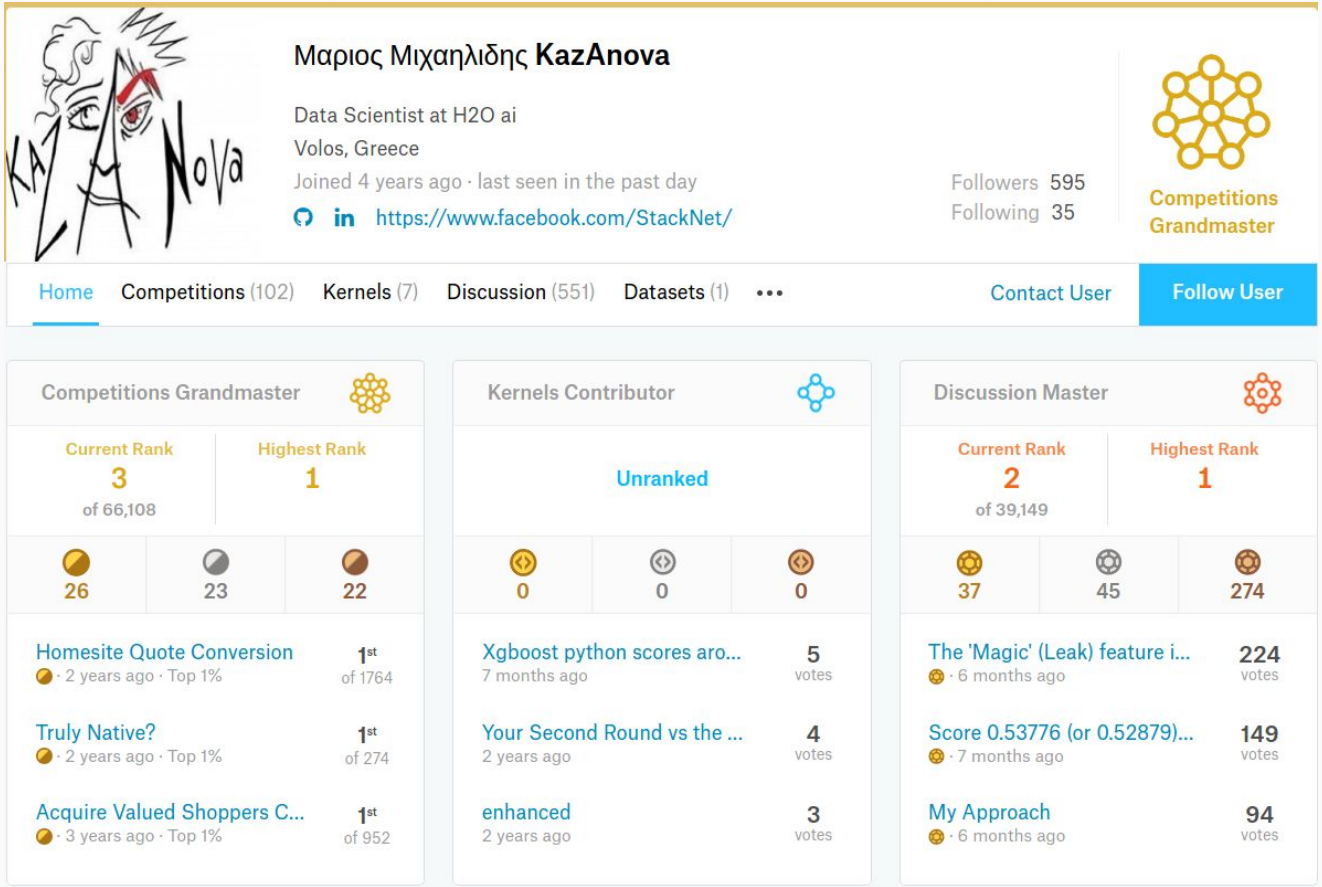
“How to win a data science competition” course

Who we are

Who we are: Marios Michailidis



Who we are: Alexander Guschin

Alexander Guschin

Moscow, Russia

Joined 3 years ago · last seen in the past day

in

Followers 19

Competitions Grandmaster

Home | Competitions (27) | Kernels (3) | Discussion (27) | Followers (19) | Contact User | Follow User

Competitions Grandmaster

Current Rank	Highest Rank
51 of 66,108	5

7	7	1

Walmart Recruiting: Trip Ty... · 2 years ago · Top 1%	1st of 1047
Flavours of Physics: Findin... · 2 years ago · Top 1%	1st of 673
Otto Group Product Classifi... · 2 years ago · Top 1%	2nd of 3514

Kernels Contributor

Unranked

0	0	0

Notebook79d6458c53 · 6 months ago	0 votes

Discussion Contributor

Unranked

3	3	10

Features engineering benc... · 3 years ago	29 votes
Beating the 3002 :) · 3 years ago	18 votes
Continuous Ranked Probab... · 3 years ago	13 votes

Who we are: Dmitry Altukhov

utility

Moscow, Russian Federation

Joined 3 years ago · last seen in the past day

Followers 35

Competitions Grandmaster

Home | Competitions (28) | Discussion (55) | Followers (35) | Contact User | Follow User

Competitions Grandmaster

Current Rank	Highest Rank
15 of 66,108	6

13	7	3

Quora Question Pairs · 5 months ago · Top 1%	2nd of 3307
Caterpillar Tube Pricing · 2 years ago · Top 1%	2nd of 1323
Liberty Mutual Group: Prop... · 2 years ago · Top 1%	2nd of 2236

Kernels Contributor

Unranked

0	0	0

No kernel results

Discussion Contributor

Unranked

1	8	23

you were only supposed to ... · a year ago	12 votes
long story of #1 solution · a year ago	10 votes
How to improve the model? · 2 years ago	9 votes

Who we are: Mikhail Trofimov

Mikhail Trofimov

Reseacher
Moscow, Russia
Joined 5 years ago · last seen 2 days ago

Followers 14

Competitions Grandmaster

Home | Competitions (32) | Discussion (95) | Organizations (1) | Followers (14) | Contact User | Follow User

Competitions Grandmaster

Current Rank	Highest Rank
175 of 66,108	21

6	6	2

Crowdflower Search Result... · 2 years ago · Top 1%	2nd of 1326
The Hunt for Prohibited Co... · 3 years ago · Top 1%	2nd of 285
Microsoft Malware Classifi... · 3 years ago · Top 1%	3rd of 377

Kernels Contributor

Unranked

0	0	0

No kernel results

Discussion Expert

Current Rank	Highest Rank
158 of 39,149	109

2	2	52

Share your models! · 2 years ago	15 votes
Brief Description of 3rd Sol... · 3 years ago	11 votes
2nd place solution · 2 years ago	9 votes

Who we are: Dmitry Ulyanov

Dmitry Ulyanov

Moscow, Russian Federation

Joined 4 years ago · last seen in the past day

https://dmitryulyanov.github.io/about

Followers 4
Following 1

Competitions Master

Home | Competitions (19) | Kernels (4) | Discussion (27) | Datasets (0) | •••

Edit Profile

Competitions Master

Current Rank	Highest Rank
219 of 66,108	75

4	5	3

Springleaf Marketing Resp... · 2 years ago · Top 1%	3rd of 2226
Microsoft Malware Classifi... · 3 years ago · Top 1%	3rd of 377
Walmart Recruiting: Trip Ty... · 2 years ago · Top 1%	7th of 1047

Kernels Contributor

Unranked

0	0	0

Simple Exploration Notebo... a year ago	0 votes
[LB 0.158] XGB_handcrafte... 5 months ago	0 votes

Discussion Contributor

Unranked

0	3	9

Code for parallel feature ex... · 3 years ago	10 votes
congratulations to the winn... · 2 years ago	6 votes
Private / Public Split - Chro... · 3 years ago	5 votes

How was it?

21 Ratings

5 out of 5 stars

5 star	95 %
4 star	5 %
3 star	0 %
2 star	0 %
1 star	0 %

All reviews

3 Reviews

To flag an abusive review for removal, please contact partner-support@coursera.org

10 Nov 2017

This course is fantastic. It's chock full of practical information that is presented clearly and concisely. I would like to thank the team for sharing their knowledge so generously.

Reply

How was it?

- It was hard.

What the course is about?

- **Week1**
  - Intro to competitions & Recap
  - Feature preprocessing & extraction
- **Week2**
  - EDA
  - Validation
  - Data leaks
- **Week3**
  - Metrics
  - Mean-encodings
- **Week4**
  - Advanced features
  - Hyperparameter optimization
  - Ensembles
- **Week5**
  - Final project
  - Winning solutions

Several words about the process

Вопрос про ГБМ

- Что будет, если из обученной GBDT модели (например XGboost) выкинуть первое дерево?

  a. Все сломается к чертям (почти рандом)
  b. Качество упадет, но не сильно
  c. Качество не изменится
  d. Качество улучшится, но не сильно
  e. Качество станет 146

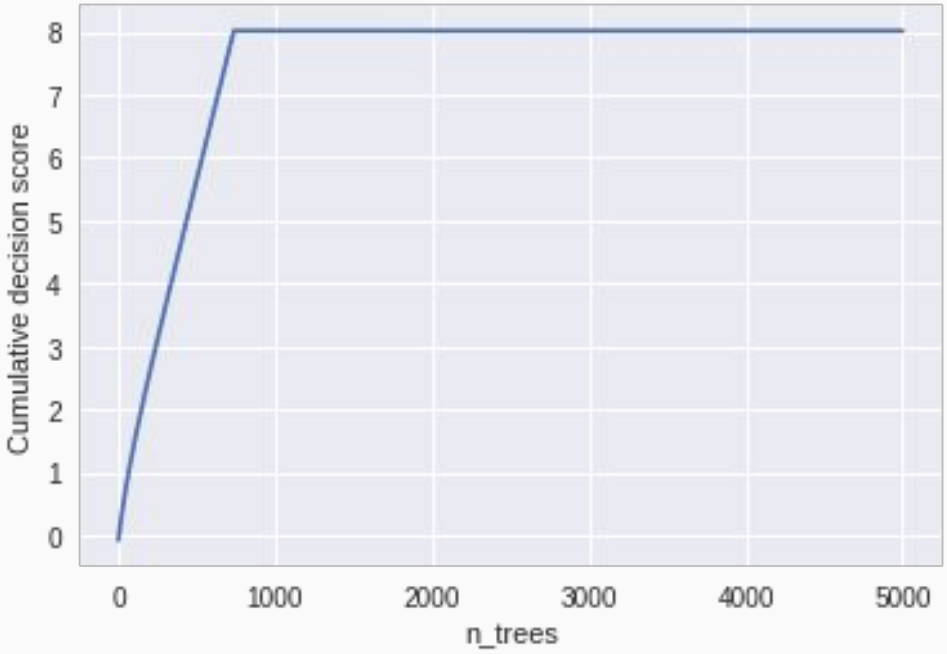
Вопрос про ГБМ

```
X_all = np.random.randn(5000, 1)
y_all = (X_all[:, 0] > 0)*2 - 1
```

```
clf = GradientBoostingClassifier(n_estimators=5000, learning_rate=0.01, max_depth=3,
clf.fit(X_train, y_train)
```

```
Logloss using all trees:          0.00031358024844254 86
Logloss using all trees but last:  0.00031358024844265755
Logloss using all trees but first: 0.00032053682522239753
```

Вопрос про ГБМ



Вопрос про ГБМ

```
clf = GradientBoostingClassifier(n_estimators=5000, learning_rate=8, max_depth=3,
clf.fit(X_train, y_train)
```

```
Logloss using all trees:          3.03310165292726e-06
Logloss using all trees but last:  2.84620992927020 4e-06
Logloss using all trees but first: 2.3463091271266125
```

Вопрос про ГБМ

$$F(x) = const + \sum_{i=1}^{n} \gamma_i h_i(x)$$

Еще вопрос

2. What of these methods can be used to preprocess texts?

- [ ] Levenshteining
- [ ] Stopwords removal
- [ ] Plumping
- [ ] Lemmatization
- [ ] Lowercase transformation
- [ ] Stemming
- [ ] Plumbing