

Решение соревнований по анализу данных

Занятие 1. Введение

Гущин Александр
Осенний семестр 2015 года

Alexander Guschin

Science can do it!



Verified
account

MASTER



?

Highest†
29th

Current†
58th
/377,180



57,081.9 points
Joined a year ago

†Ranking method changed 13 May 2015 (?)

Profile

Results

Scripts

Forum

Account

Activity

\$100

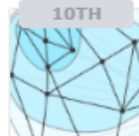
Edit Profile

2ND

otto group

2nd/3514

10TH



10th/203

TOP 10%



12th/1528

TOP 10%



47th/1785

TOP 10%



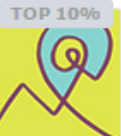
12th/375

TOP 10%



46th/1306

TOP 10%



69th/1049

TOP 25%



226th/1604

11

Competitions

Сайты с соревнованиями по анализу данных: kaggle.com, drivendata.org, datascience.net, ...

Задачи

качество решения
интерпретируемость
скорость работы модели
скорость обучения модели
ограничение на объём RAM
теоретическое обоснование

соревновательные задачи: в основном (качество)

бизнес-задачи: могут включать всё перечисленное, чаще всего
(качество, скорость работы),
(качество, интерпретируемость),
(качество, интерпретируемость, теоретическое обоснование)

Соревнования

сложность
предобработки
данных

извлечь фичи
из сырых данных

“очистить” данные

заполнить NaN

обычное
соревнование
для kaggle

1

...

2

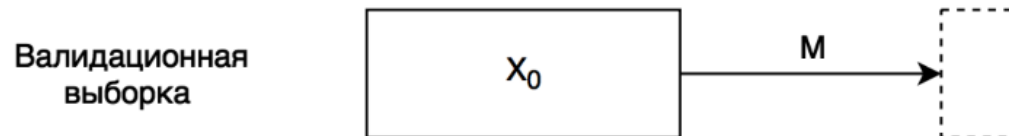
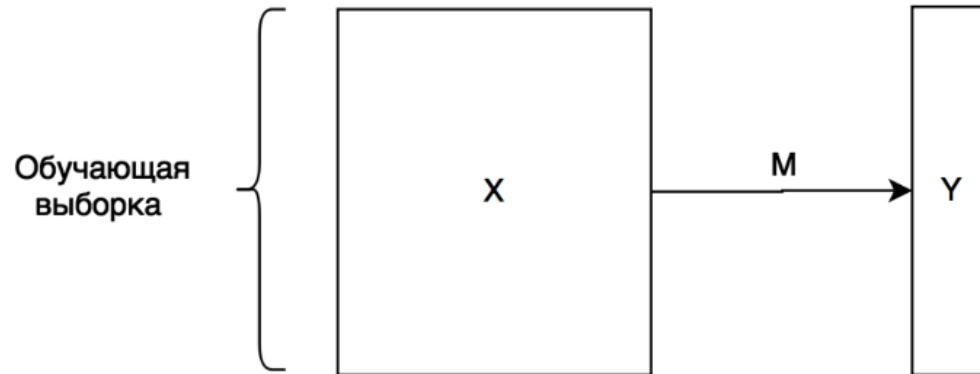
1 - Microsoft Malware
Classification Challenge
(BIG 2015)
2 - **Coupon Purchase
Prediction**

“сложность”
метрики

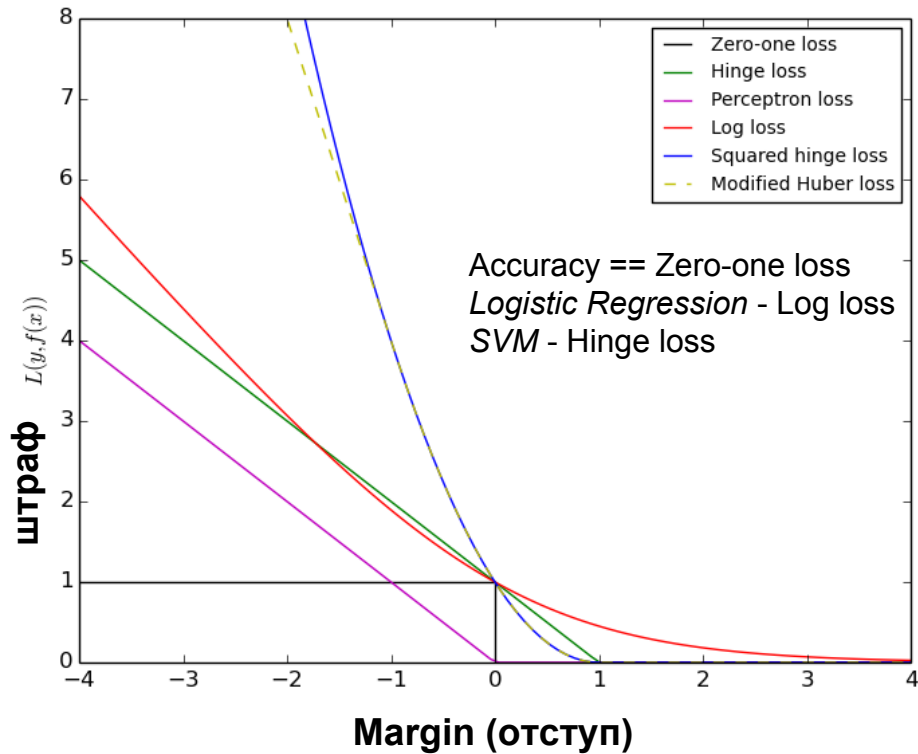
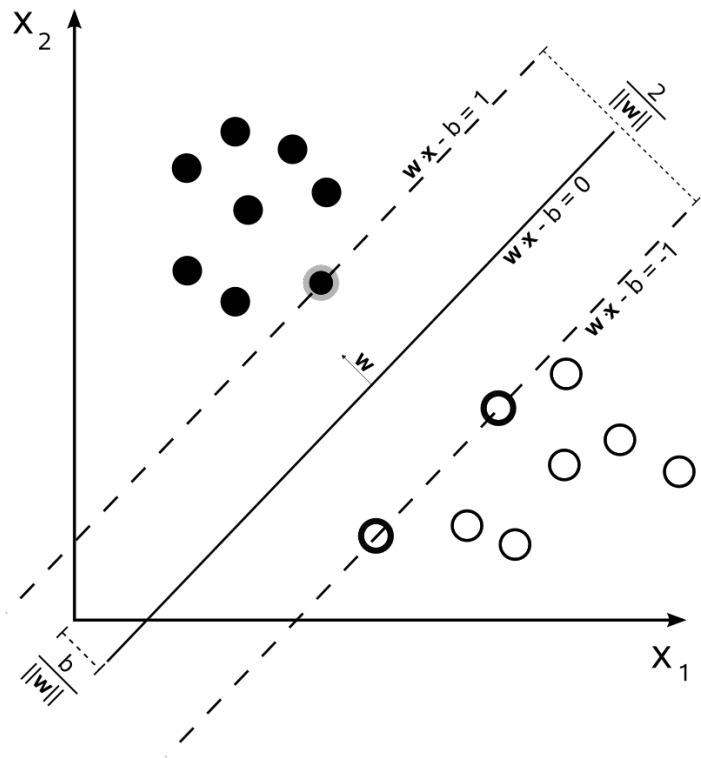
RMSE,
LogLoss

AUC

MAP@10




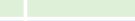



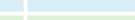


Обучаем классификатор M, “аппроксимирующий” истинную зависимость $X \rightarrow Y$



Изначальная метрика задач классификации - Ассигасу. Но она не дифференцируема

HOSTED BY PLANNED PARENTHOOD FEDERATION OF AMERICA

User or team	Public	Private	Timestamp	Trend	# Entries
 giba	0.2482	0.2539	April 14, 2015, 11:52 p.m.		33
 taguschin	0.2492	0.2543	April 14, 2015, 11:58 p.m.		20
 JYL	0.2497	0.2549	April 14, 2015, 7:49 p.m.		48
 furiouseskimo	0.2513	0.2562	April 11, 2015, 3:51 p.m.		24

otto group

Completed • \$10,000 • 3,514 teams

Otto Group Product Classification Challenge

Tue 17 Mar 2015 – Mon 18 May 2015 (4 months ago)

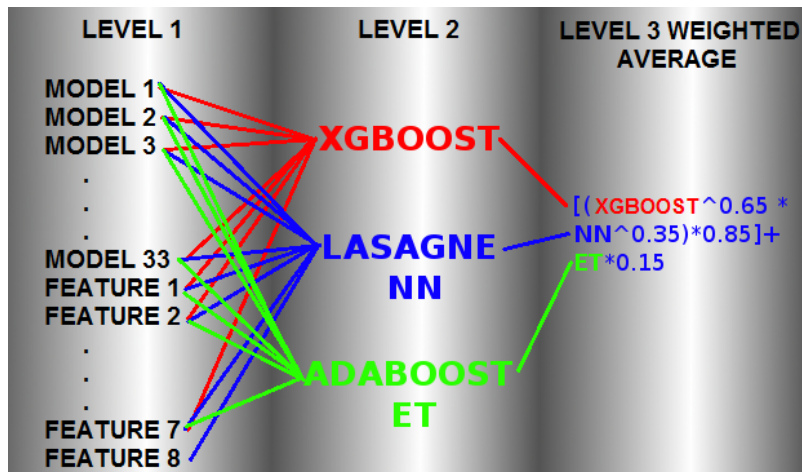
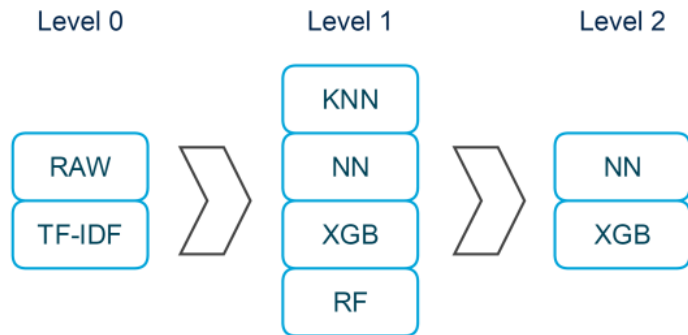
Dashboard

Private Leaderboard - Otto Group Product Classification Challenge

This competition has completed. This leaderboard reflects the final standings.

See someone using multiple accounts?
Let us know.

#	Rank	Team Name	I model uploaded * in the money	Score	Entries	Last Submission UTC (Best - Last Submission)
1	—	Gilberto Titericz & Stanislav Semenov	🏆👑*	0.38243	90	Mon, 18 May 2015 23:42:43 (-6.6h)
2	11	٢\(_٧)_٢👑*		0.38656	26	Mon, 18 May 2015 23:31:27 (-1.5h)
3	1	i dont know	🏆👑*	0.38667	103	Mon, 18 May 2015 23:53:10 (-8.6h)
4	—	👤 Dmitry & Davut	🏆👑	0.39027	168	Mon, 18 May 2015 21:26:53 (-0.2h)
5	—	Mikhail Trofimov		0.39263	27	Mon, 18 May 2015 21:31:14



<обзор кода Health care decisions
01_HealthCareDecisions_2ndPlace.ipynb

цель: дать представление о том, какой сложности
решения бывают + дать материал для
дополнительного изучения>

<туториал по Pandas
01_PandasTutorial.ipynb>



Knowledge • 3,085 teams

Titanic: Machine Learning from Disaster

Fri 28 Sep 2012

Thu 31 Dec 2015 (3 months to go)

Dashboard

Home

Data

Make a submission

Information

Description

Evaluation

Rules

Prizes

Frequently Asked Questio...

Further Reading / Watching

Getting Started With Excel

Getting Started With Python

Getting Started With Pyth...

Getting Started With Rand...

New: Getting Started With R

Submission Instructions

Forum

Scripts

New Script

Leaderboard

Visualization

My Team

GitHub

My Submissions

Competition Details » Get the Data » Make a submission

Predict survival on the Titanic using Excel, Python, R & Random Forests

See [best practice code](#) and [explore visualizations of the Titanic dataset on Kaggle Scripts](#). Submit directly to the competition, no data download or local environment needed!

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

TITANIC sinking



чем мы будем
заниматься в
первую
неделю...

Семестр

планы и домашнее задание

неделя 1: <https://www.kaggle.com/c/titanic/>

анализ данных с помощью Python, Pandas, Numpy, Matplotlib (разбор
туториалов) ~ 5 часов,

“Getting Started with Python/RandomForest” + “Further Reading / Watching”
(на странице соревнования) ~ 3 часа.

неделя 2: построение моделей машинного обучения с помощью
Sklearn, разбор большей части бенчмарков для титаника

неделя 3: начинаем решать более серьёзное соревнование

...