

Kaggle challenge from LHCb

Marc-Olivier Bettler¹, Marcin Chrzęszcz^{2,3},
Andrey Ustyuzhanin^{4,5}, Tatiana Likhomanenko^{4,5}



University of
Zurich^{UZH}



Yandex



¹ CERN, Geneva

² University of Zurich

³ Institute of Nuclear Physics, Krakow

⁴ Yandex School of Data Analysis, Moscow

⁵ NRC "Kurchatov Institute", Moscow

27 February 2015

What is kaggle - The Home of Data Science

- Kaggle is the world's largest community of data scientists.
- It provides the possibility to data scientists to solve real-world problems across a diverse array of industries including life sciences, financial services, energy, information technology.
- In addition to the prize money and data, participants use Kaggle to meet, learn, network and collaborate with experts from related fields.

Who has already used Kaggle:



facebook.

and many many others including:



How does Kaggle work?

- One (ex. LHCb) defines a data analysis problem, provides data sets and rules about the evaluation.
- For the contests there are allocated prizes.
- After the contest is over (couple of months) the solutions are made public.
- Current contests:




Active Competitions

All Competitions

17 found, 17 active

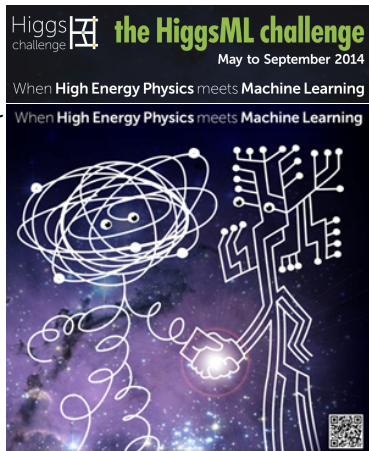
All competitions

Entenable

Competition Name	Reward	Teams	Deadline
 National Data Science Bowl Predict ocean health, one plankton at a time	\$175,000	467	52 days
 Driver Telematics Analysis Use telematic data to identify a driver signature	\$30,000	728	52 days
 Click-Through Rate Prediction Predict whether a mobile ad will be clicked	\$15,000	1560	17 days

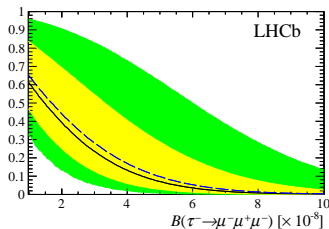
Atlas path to $H \rightarrow \tau\tau$

- As we know ATLAS due to poor vertex detector is worse than CMS in modes like $H \rightarrow \tau\tau$ or $H \rightarrow B\bar{B}$
- ATLAS gave their datasets (for $H \rightarrow \tau\tau$) training the classifiers that can be used for future analysis.
- After evaluation they gained $\sim 10\%$ on sensitivity!
- This would be great for $\tau \rightarrow \mu\mu\mu$.
- This would be great for LHCb and HEP community as well.
- Also after the challenge they organized a workshop \rightarrow other LHCb analysis could benefit from this kind of collaborations, we almost always use MVAs.



Future of $\tau \rightarrow \mu\mu\mu$ in LHCb

- In 3 fb^{-1} we had expected limit of 5.0×10^{-8} .
- What can we expect after another 5 fb^{-1} ? \bar{B}^0
- $\frac{5.0 \times 10^{-8}}{\sqrt{\frac{5 \times 2}{3}}} = 2.7 \times 10^{-8}$
- We should aim to do better than Belle (2.1×10^{-8})!
- Help from Kaggle community would be very appreciated!



Comparison with Higgs Boson Challenge

Proposed challenge is connected with Higgs Boson challenge, but differs from it, being more realistic (closer to real physics analysis):

- 1 training dataset includes not only simulated data but also real data (signal-like events and background-like events have different nature).
- 2 test dataset includes also normalization channel to do calibration of data.
- 3 the submission must pass additional checks (classifier must not be correlated with mass and behave similarly on real and simulated data).
- 4 our quality metric uses predicted probability in all bins as we do in real analysis.
- 5 all scripts for testing mass correlation, DATA/MC agreement and limit evaluation are provided by us.

What would we like to give

- MC and DATA¹ after the preselection.
- DATA will contain our standard ntuple entries (excluding variables, like trigger decisions, BKGCAT, etc.).
- We want to give as much of those as possible as people can then construct their own variables (happens often).
- The size of data that is used for testing and training is up to us.

¹I will comment on protecting our data in couple of slides

Evaluation

- Check correlation between mass and model predictions on all test $\tau \rightarrow \mu\mu\mu$ sidebands using Cramer-von Mises measure (ex. arXiv:1410.4140)
- Check agreement between MC and data on $D_s \rightarrow \phi\pi$ (test MC and test data) using Kolmogorov-Smirnov distance
- Calculate of Approximate Binned Median Statistics (ABMS), only if above two tests are passed.
- We need to use ABMS because the standard CL_s method is computational expensive.
- ABMS is just value of statistic and shows how well two hypothesis can be distinguished
- ABMS metric is similar AMS used in Higgs competition, but involves all available statistics, making it more meaningful and stable. Details of this metric can be found in backups.
- Participants chooses number of bins and bins thresholds themselves (i.e. splitting classifier output)

Evaluation Examples

Ada Boost, Gradient Boost were trained (as participants can do). Also Ada Boost and Uniform Gradient Boost were trained using mass as input for classifier. (CVM - Cramer-von Mises metric, KS - Kolmogorov-Smirnov metric),

	ada	ada(mass)	gb	ugb(mass)
CVM metric	0.005674	0.061837	0.005642	0.005714
CVM p-value	0.918667	1.000000	0.850000	0.970000
KS distance	0.028815	0.018353	0.027854	0.025621
ABMS public	1.557205	1.790282	1.564490	1.545545
ABMS private	1.549253	1.785880	1.562412	1.542357

- We tested already with standard classifiers that this metric work and reject cases where people will try to do something strange (like add mass for training).

Protecting our data

- We should be of course worried that we are giving to public our data set.
- However those data are after preselection, without any PID calibration, only selected variables, etc.
- Plan is to smear the Data, so we can make a statment such as: "Data was prepared in a way that it can't be used for any physical analysis".
- Other tughts are welcome!



- Jan/Feb: confirmation from LHCb, allocate prize budget at Yandex
 - Feb: prepare website, explanatory materials, evaluation procedures, test challenge
 - Feb: propose workshop at KDD
 - Mar: announce challenge, start
 - Mar-May: run challenge
 - Jun: announce winners
 - Aug: run KDD workshop, award winners
- 1 KDD - Conference on Knowledge discovery and Data Mining: 2015 is one of the most important conferences for data scientists.
 - 2 Plan would be to have a session there dedicated to our challenge.
 - 3 Afterwards we could organise workshop at CERN as well to hopefully fruitful collaboration.