

Kaggle challenge from LHCb

Thomas Blake¹, Marc-Olivier Bettler², Marcin Chrzęszcz^{3,4},
Francesco Dettori², Andrey Ustyuzhanin^{5,6}, Tatiana
Likhomanenko^{5,6}



University of
Zurich^{UZH}



Yandex



¹ University of Warwick

² CERN, Geneva

³ University of Zurich

⁴ Institute of Nuclear Physics, Krakow

⁵ Yandex School of Data Analysis, Moscow

⁶ NRC "Kurchatov Institute", Moscow

27 February 2015

What is kaggle - The Home of Data Science

- Kaggle is the world's largest portal for the data science community.
- It provides the possibility for data scientists to solve real-world problems across a diverse array of industries including life sciences, financial services, energy, information technology.
- Enables participants use Kaggle to meet, learn, network and collaborate with experts from related fields.
- Usually each contest/challenge has an cash reward (13k\$ in ATLAS case.)

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 to rank solutions.
- For the contests there are allocated prizes.
- When a contest is over (couple of months) the top solutions are made public.
- Current contests:




Active Competitions

All Competitions

17 found, 17 active

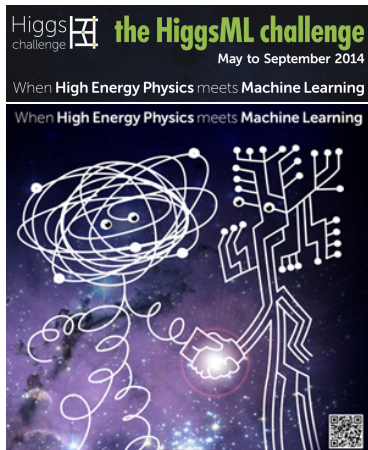
All competitions

Enterable

| 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 |

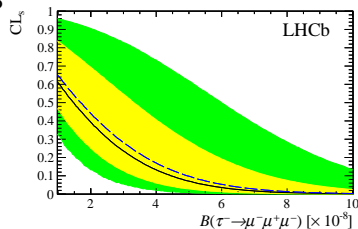
Atlas path to $H \rightarrow \tau\tau$

- Because of its poorer vertex resolution, ATLAS is less sensitive than CMS in modes like $H \rightarrow \tau\tau$ or $H \rightarrow B\bar{B}$
- ATLAS gave their MC samples datasets (for $H \rightarrow \tau\tau$) to train the classifiers that can be used for future analyses.
- After evaluation they gained $\sim 10\%$ on sensitivity!
- Other analysis from ATLAS picked up some open source libraries and are using them.
- During the contest there are discussions between the participants, which are also available for physicist \rightarrow knowledge transfer.
- Over 1.800 teams participated in this contest!



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} ?
- $\frac{5.0 \times 10^{-8}}{\sqrt{\frac{5 \times 2}{3}}} = 2.7 \times 10^{-8}$
- We should aim to do better than Belle (2.1×10^{-8})!
- Help from Kaggle community would be very appreciated for $\tau \rightarrow \mu\mu\mu$, LHCb and HEP community as well.



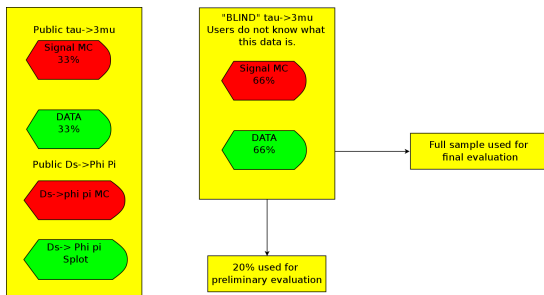
Comparison with Higgs Boson Challenge

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 control channel to access DATA/MC differences.
- 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.
- 6 τ in LHCb comes from five different sources which makes this contest more interesting for machine learning people.

What would we like to give

- Full MC sample of $D_s \rightarrow \phi\pi$ and $\sim 30\%$ data $D_s \rightarrow \phi\pi$.
- Full MC sample of $\tau \rightarrow \mu\mu\mu$ and full data of $\tau \rightarrow \mu\mu\mu$.¹
- DATA will contain our standard ntuple entries (excluding lumi etc.)
- We want to give as much of those as possible to allow people to construct their own variables (happens often).
- The size of the data sets and the split training/testing 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
- Calculation of Approximate Binned Median Statistics (ABMS), only if above two tests are passed.
- We need to use ABMS because the standard CLs method is computationally too expensive.
- ABMS is just value of statistic and shows how well two hypothesis can be distinguished
- ABMS is similar to the AMS metric used in Higgs competition, but involves all available statistics, making it more meaningful and stable. Details of the 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 could make our data open without any modifications, as performing any analysis without knowledge of the preselection is not possible.
- If collaborations feels strong about protecting our data, we can smear/shift it in a way that physics analysis is not possible, but the training is not distorted.
- Both scenarios are acceptable for us.



- Feb: ask OK from LHCb, allocate prize budget at Yandex (15k \$, more then ATLAS had ;))
 - Mar: prepare website, explanatory materials, refine evaluation procedures, test challenge
 - Mar: propose workshop at KDD/NIPS
 - Apr: announce challenge, start
 - Apr-Jun: run challenge
 - July: announce winners
 - Aug/Sep: run KDD/NIPS workshop, award winners
- 1 KDD/NIPS are very well know maschine learning contests.
 - 2 Plan would be to have a sesion there dedicated to our challange (was the case of ATLAS competition).
 - 3 Afterwards we could organise workshop at CERN as well to hopefulls start a fruitfull collaboration.

Conclusions

- 1 Kaggle constes for LHCb would be very beneficial for us.
- 2 A lot of work has been put in to make the contest as usable for us as possible(correlations check, MC/DATA disagreement).
- 3 All scripts are ready and automatised and tested.
- 4 We look to have feedback from the collaboration.



BACKUPS

Approximate Binned Median Statistics (ABMS)

In real analysis if you are looking for upper limit you compare null hypothesis H_0 (background-only) with a spectrum of other hypotheses representing different branching fractions. Then you choose one with smaller value of branching fraction that gives enough significance. The measure of significance is CLs, that is computed for statistics q equal to ratio of likelihoods of your hypotheses.

For Kaggle to simplify understanding/computation of the metric we do two «tricks»:

- instead of comparing H_0 with spectrum of hypothesis we compare it with hypothesis with specific branching fraction (taken from the paper on $\tau \rightarrow \mu\mu\mu$) and
- instead of estimation of significance we just calculate statistics q (the better it is, the better significance classifier can provide)

Approximate Binned Median Statistics (ABMS)-2

For given classifier g , the number of events n found in a bin (a region of input variable space), is assumed to follow a Poisson distribution with mean $\mu_s + \mu_b$

$$P(n|\mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s + \mu_b)},$$

where μ_s and μ_b are the expected numbers of events from the signal and background, respectively. To establish the existence of the signal process, we test the H_0 of $\mu_s = 0$ against the alternative H_1 with $\mu_s > 0$. We will use several bins, and assume that for each bin we have independent parameter μ_s . Thus the likelihood ratio looks like:

$$Q = \prod_{bin} \frac{P(n_i|0, \mu_b^i)}{P(n_i|\hat{\mu}_s^i, \mu_b^i)} = \prod_{bin} \left(\frac{\mu_b^i}{n_i} \right)^{n_i} e^{n_i - \mu_b^i}, \quad (1)$$

where $\hat{\mu}_s^i$ is the maximum likelihood estimator of μ_s^i given that we observe n_i events in the i -th bin. $\hat{\mu}_s^i = n_i - \mu_b^i$.

Approximate Binned Median Statistics (ABMS)-3

$$q = -2 \ln Q = 2 \sum_{bin} \left(n_i \ln \frac{n_i}{\mu_b^i} - n_i + \mu_b^i \right) \quad (2)$$

For empirical estimations we take $\mu_b^i = b_i$, $n_i = s_i + b_i$, where s_i is estimation of the amount of signal according to the best known upper limit. Then empirical estimation of the statistics is

$$\hat{q} = 2 \sum_{bin} \left((s_i + b_i) \ln \left(1 + \frac{s_i}{b_i} \right) - s_i \right) \quad (3)$$

Adding regularization term we can define

$$ABMS = \sqrt{\hat{q}_{regularization}} = \sqrt{\sum_{bin} AMS_i^2} \quad (4)$$

where AMS_i is calculated for i -th bin in the same way as in HiggsML challenge:

$$AMS = \sqrt{2((s + b + b_{reg}) \ln(1 + s/(b + b_{reg})) - s)} \cong \sqrt{2s^2/(b + b_{reg})}$$