SCHOOL OF DATA ANALYSIS

University of Zurich UZH

THE UNIVERSITY OF WARWICK

LHCb

”Flavours of Physics” challenge: summary and lessons learned

Tatiana Likhomanneko[1], Marcin Chrząszcz[2], Marc-Olivier Bettler[3], Thomas Blake[4], Francesco Dettori[3], Alexey Rogozhnikov[1] Pavel Serdyukov[1], Andrey Ustyuzhanin[1]

[1]YSDA, [2]University of Zurich, [3]CERN, [4]University of Warwick

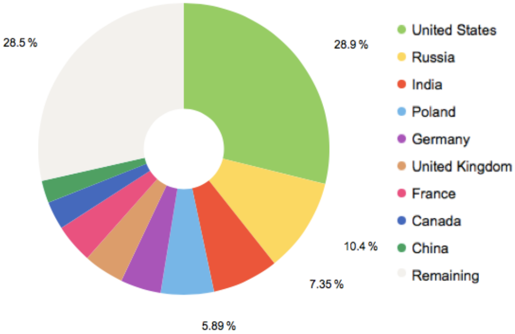
Challenge Summary

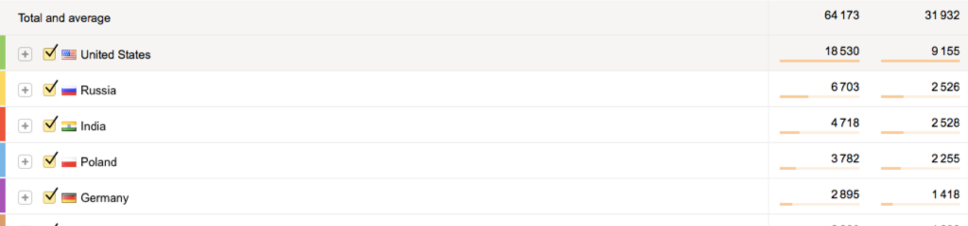
- `https://www.kaggle.com/c/flavours-of-physics`
- Search for new physics
- Goals
  - CERN Public Relations
  - find interesting ML ideas to reuse in HEP data analyses
  - evaluate classifier verification checks
- Real data + MC
- Proxy metric
- Additional restrictions to reflect physics analysis
- Special Prize for Physics

July-
October
2015
https://www.kaggle.com/c/flavours-of-physics
Flavours of Physics Challenge
Organization committee
Marc-Olivier Bettler – CERN
Thomas Blake – University of Warwick
Marcin Chrzaszcz – University of Zurich, IFJ PAN
Francesco Dettori – CERN
Eugene Khartanov – Yandex
Tatiana Likhomanenko – YSDA, HSE, YDF
Pavel Serdyukov – Yandex
Andrey Ustyuzhanin – YSDA, HSE, YDF

Some stats

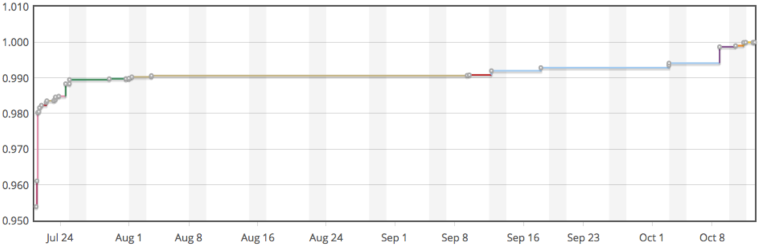
- 3 months
- 673 teams
- 10124 entries
- $15000 USD regular prize
- 2 Physics Prizes

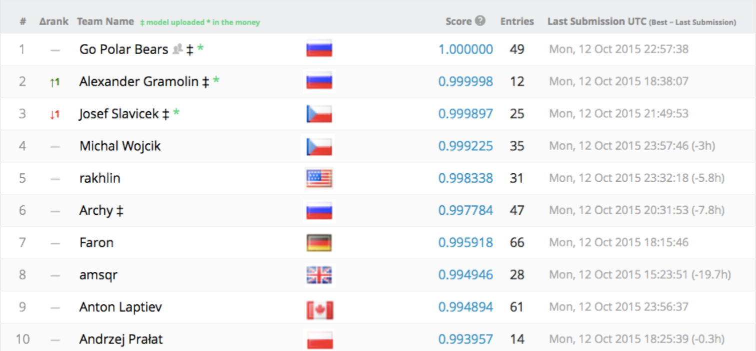




Total and average	64 173	31 932
⊞ ☑ United States	18 530	9 155
⊞ ☑ Russia	6 703	2 526
⊞ ☑ India	4 718	2 528
⊞ ☑ Poland	3 782	2 255
⊞ ☑ Germany	2 895	1 418

Private leader-board



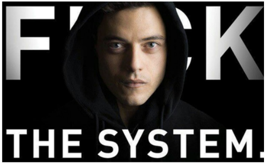


#	Δrank	Team Name ‡ model uploaded * in the money		Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Go Polar Bears ‡ *	🇷🇺	1.000000	49	Mon, 12 Oct 2015 22:57:38
2	↑1	Alexander Gramolin ‡ *	🇷🇺	0.999998	12	Mon, 12 Oct 2015 18:38:07
3	↓1	Josef Slavicek ‡ *	🇨🇿	0.999897	25	Mon, 12 Oct 2015 21:49:53
4	—	Michal Wojcik	🇨🇿	0.999225	35	Mon, 12 Oct 2015 23:57:46 (-3h)
5	—	rakhlin	🇺🇸	0.998338	31	Mon, 12 Oct 2015 23:32:18 (-5.8h)
6	—	Archy ‡	🇷🇺	0.997784	47	Mon, 12 Oct 2015 20:31:53 (-7.8h)
7	—	Faron	🇩🇪	0.995918	66	Mon, 12 Oct 2015 18:15:46
8	—	amsqr	🇬🇧	0.994946	28	Mon, 12 Oct 2015 15:23:51 (-19.7h)
9	—	Anton Laptiev	🇨🇦	0.994894	61	Mon, 12 Oct 2015 23:56:37
10	—	Andrzej Prałat	🇵🇱	0.993957	14	Mon, 12 Oct 2015 18:25:39 (-0.3h)

Platform – Kaggle

- good support from Kaggle
- Kaggle people willing to play with metric to certain extent, support research competitions - “Flavours of Physics” is the 1st competition with additional checks prior to commit implemented
- scripts - allows understand common areas of development, source of rapid bootstrap and knowledge exchange
- lively community
- Caveat emptor!

vs



Data

- Training sample, $\tau \to \mu\mu\mu$
    - Signal - simulated
    - Background - real (taken from regions where signal cannot occur)
    - 40+ features
- Control channel, $D \to \phi\pi$
    - well studied
    - has similar topology to $\tau \to \mu\,\mu\,\mu$
    - Available both MC and real data samples

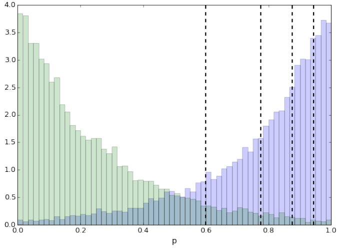
Figure of Merit

- Each event is represented by **x**
- Find classifier $g(\mathbf{x})$ to separate signal from background
- number of background events b is estimated in the selection $G(p)$:

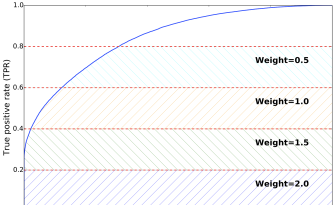
$$G(p) = \{\mathbf{x} : g(\mathbf{x}) \geq p\}$$

- Discovery is made when the number of observed events n in $G(p)$ is significantly higher than b
- **Difficult to estimate!**

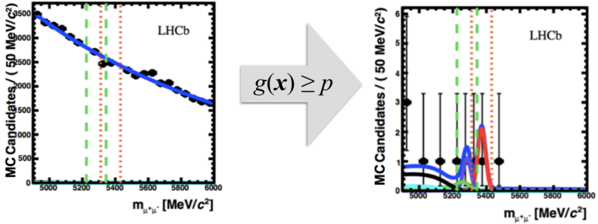
Proxy for significance



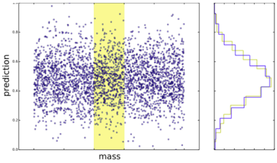
Weighted ROC AUC (for stability sake)

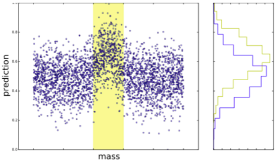


Additional Checks. Mass correlation



$g(\boldsymbol{x}) \geq p$



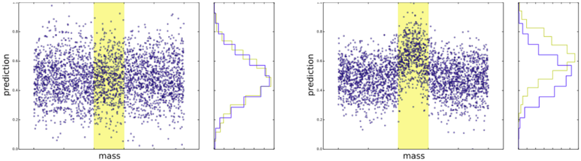


Cramer-von Misés check adaptation

- Mass correlation may sculpt a false peak
- Put all events on 2D plane: mass, prediction
  - Split mass into regions
  - compute Cramer-von Misés distance between CDF of predictions for events in each region (local) and for all events (global)

$$CvM_d = \int (CDF_{\text{global}} - CDF_{\text{local}})^2 dCDF_{\text{global}}$$

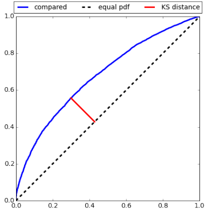
  - Take average over regions



Additional Checks. MC-Real agreement

- Checks that your model behaves similarly on MC and real data
- Uses $D \to \phi\pi$ (proxy channel) that has both simulated and real decay events
- Compute Kolmogorov-Smirnov distance between distributions of your model predictions on real and simulated events only for proxy channel events

$$KS_d = max|CDF_{\text{real}} - CDF_{\text{simulated}}|$$



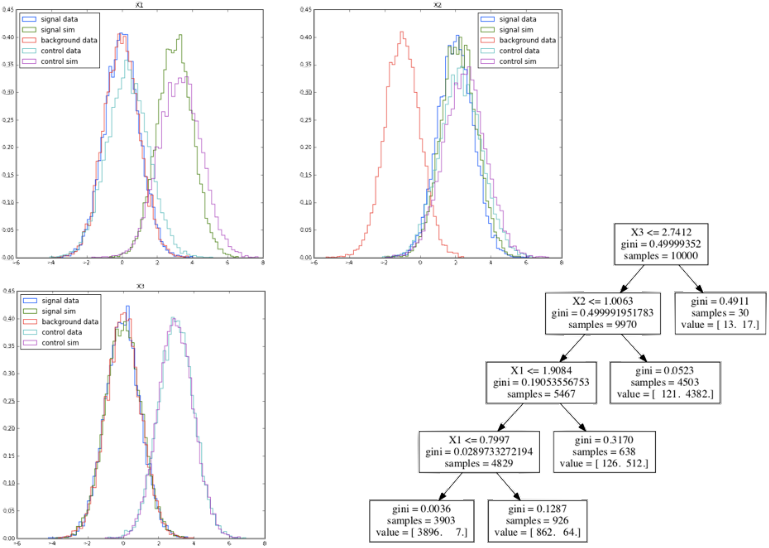
«Data leakage» - constraints overfitting

- Control and signal channels behave quite differently
- `check_agreement.csv` has plenty of events to train classifier to discriminate between control & signal channel
- Exploit idea
  - learn to distinguish between signal and control data,
  - build a classifier on training data, with all the freedom to exploit simulation artefacts,
  - assign predictions to samples predicted as control data, so that the test is passed (e.g., random predictions in the case of the KS statistic), otherwise predict using the classifier found in the previous step
- Tim's and Gilles' *out-of-competition-solution* was the 1st for ~1 month (see next presentation by those guys)

Illustration

- `http://bit.ly/gilles_tim` (by Gilles Louppe, Tim Head)
- artificial classification example between signal and background events, along with some close control datasets. Assume an input space defined on three input variables X1,X2,X3, such that:
  - X1 is irrelevant for discrimination between real data signal and real data background but, because of simulation imperfections, has discriminative power between simulated events and real data events ;
  - X2 is discriminative between signal and background events ;
  - X3 is discriminative between events from the original problem and the control channel, but has otherwise no discriminative power between signal and background events

Illustration, continued



X3 <= 2.7412
gini = 0.49999352
samples = 10000

X2 <= 1.0063
gini = 0.499991951783
samples = 9970

gini = 0.4911
samples = 30
value = [13. 17.]

X1 <= 1.9084
gini = 0.19053556753
samples = 5467

gini = 0.0523
samples = 4503
value = [121. 4382.]

X1 <= 0.7997
gini = 0.0289733272194
samples = 4829

gini = 0.3170
samples = 638
value = [126. 512.]

gini = 0.0036
samples = 3903
value = [3896. 7.]

gini = 0.1287
samples = 926
value = [862. 64.]

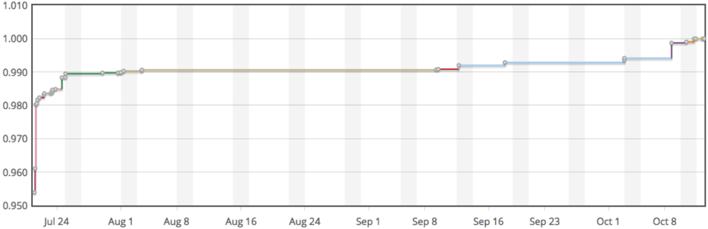
Leakage handling

- ignore
- stop, fix metric and data, restart
  - difficult to specify in mathematically strict way what should be checked in order to keep systematics estimations clear on the later analysis stage, or no new systematics should be introduced by classification model
- inform participants about possible consequences
  - `https://www.kaggle.com/c/flavours-of-physics/forums/t/15735/important-training-dataset-reminder/`
  - `https://www.kaggle.com/c/flavours-of-physics/forums/t/15974/machine-learning-approach-in-high-energy-physics-explanation`

Mass reconstruction

- active since beginning of September
- `https://www.kaggle.com/c/flavours-of-physics/forums/t/16975/who-else-calculated-the-mother-particle-mass`
- requires a bit of physics background but not much:

$$mass = \frac{p}{speed}$$



Meaningful ideas wrt MC-Real agreement

- feature selection (real-MC inagreement features)
  - train on all
  - remove one by one to minimize KS
- reweighting (e.g. BDT-reweighting, `http://bit.ly/bdt-reweighting`)
- Gradient reversal, cross-domain learning
- training on normalization channel
  - training ($Data^{\tau}_{\text{sideband}}$ vs $Data^{\text{cal}}$)
  - validation ($Data^{\tau}_{\text{sideband}}$ vs $MC^{\tau}_{\text{signal}}$)
  - no mass correlation
  - low-bound estimation
- add KS-term to loss function or splitting criteria
- “Subtraction”

Meaningful ideas wrt mass correlation

- features selection (remove ones that use mass)
- Gradient reversal, cross-domain learning
- alter loss or splitting function to fight correlation with mass (e.g. add CvM-term)
  - uGBoost, http://bit.ly/uGBoost

Useful solutions / reading

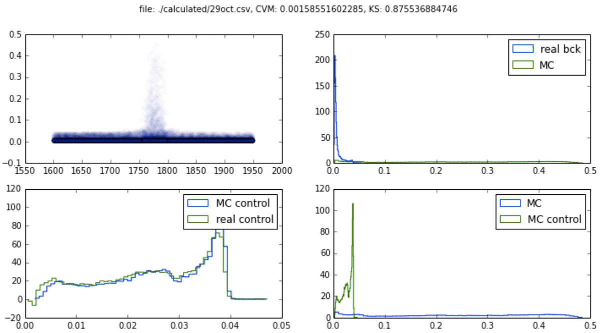
- blog by Hongliang Lui, CMS `https://no2147483647.wordpress.com/author/phunterlau/`
- blog by Vicens Gaitan, ex-researcher at ALEPH, R&D director in the Grupo AIA `http://blog.kaggle.com/2015/10/27/passing-the-tests-strategies-used-in-cerns-flavour`
- notebook by Gilles Louppe & Tim Head `http://bit.ly/gilles_tim`
- `https://www.kaggle.com/c/flavours-of-physics/forums/t/15837/regularized-learning-using-control-channel-data`

Metric tricks

- narrow signal window, i.e. add boolean feature that equals to 1 if event mass is within $[m_\tau - \Delta, m_\tau + \Delta]$ and 0 otherwise and optimize Δ
  - can be fixed by lowering threshold or by changing avg to max in CvM check
- exponential trick (1st and 3rd place)
- adding random noise (see Vicens' blog post)

Metric improvements suggestions

- check distribution difference of the model on MC_{signal} vs MC_{control}
- in general difficult to make it hack-proof



file: ./calculated/29oct.csv, CVM: 0.00158551602285, KS: 0.875536884746

Outcomes

- rule of thumb: if nothing works, use XGboost (present in top3 solutions)
- weighted ROC AUC is stable & understandable
- bunch of ideas to be explored (e.g. BDT-reweighting, `http://bit.ly/bdt-reweighting`)
- more details will follow
  - Physics Prize announcements at NIPS workshop `http://yandexdataschool.github.io/aleph2015/`
  - Zurich workshop, `https://indico.cern.ch/e/HeavyFlavourDM16`

Heavy Flavour Data Mining Workshop

- Focus on applications of ML to HEP problems (advertised in ML community)
- 18-20 Feb 2016
- Zurich University
- Abstract submission deadline: 24th December
- Reports from Physics Prize winners
- Tutorials on popular ML toolkits
- OpenSpace technology to foster collaboration

Check-list for organizers of new challenges

- can you provide unique datasets interesting to the public?
- can you simplify the metric even more? - can you avoid additional checks or embed it in FoM?
- is metric stable? (check with different samplings and random seeds)
- have you tested the metric? have you tested the criteria? did we mention testing?
- Are ‘special’ prizes attractive enough? (at least as big as ranked solutions)
- Is ‘Higgs’ in the title? :)

Conclusion

- some good techniques were discovered
- not all of them has been explored to the full extent
- there is challenge for a good classifier metric (or for evaluation procedure) that evaluates classifier-specific systematics (*“you cannot rely only in the power of machine learning: looking at the data through the glass of a model can expose facts hard to discover by general function approximation”* Vicens Gaitan)
- more extensive reports will follow, NIPS, Zurich (`https://indico.cern.ch/e/HeavyFlavourDM16`)!
- Yandex School of Data Analysis is willing to support such activities in the future

Thank you!