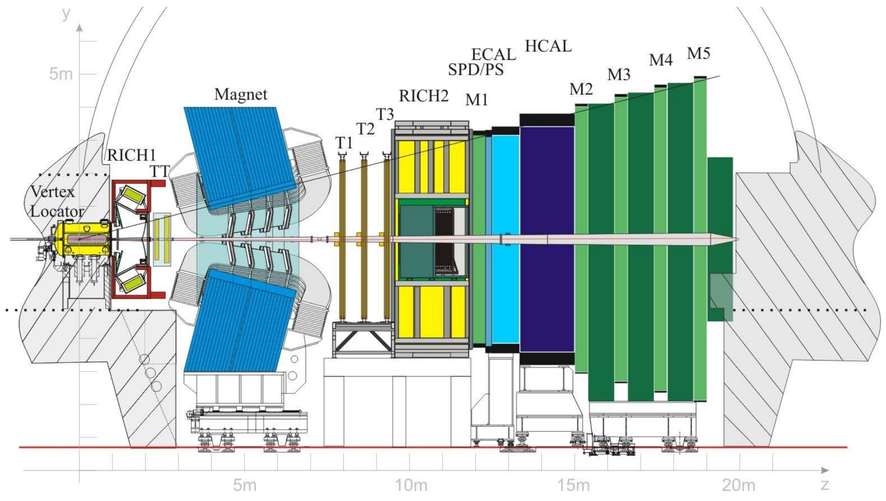
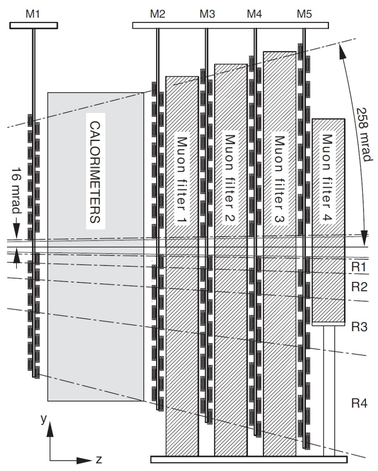
Machine Learning for Muon Identification at LHCb

Nikita Kazeev[1,2,3,4] on behalf o the LHCb collaboration
nikita.kazeev@cern.ch
[1] NRU Higher School of Economics, Moscow, Russia
[2] Università degli Studi di Roma "La Sapienza", Rome, Italy
[3] Yandex School of Data Analysis, Moscow, Russia
[4] INFN - Laboratori Nazionali di Frascati



Particle identification plays a crucial role in the LHCb experiment. Muon subsystem in particular offers a fast way to select events that contain muons in their final state without having to reconstruct the whole event.



The physical idea behind the muon identification is to draw upon the unique penetrating power of muons - if a charged particle was able to pass through the calorimeter and lead shielding, it is highly likely a muon. Muon subdetector consists of 5 sensitive planes (the first one, M1, is not used).
Upgrade will see M1 removed, some additional shielding installed around the beam pipe, and, most importantly, a 5x increase in luminosity that makes it imperative to update the algorithms to cope with the increased occupancy [6].

IsMuon

The first step in muon identification is IsMuon – a very fast yet discriminating algorithm. For Run I it kept muon efficiency in the range of 95-98% and background rejection to the level of 99% [1].

Given a reconstructed track in the LHCb tracking system, hits in the Muon stations are searched around the track extrapolation inside Field of Interest (FOI). FOI is determined from analytical approximation of multiple scattering [2]. If and only if there are hits in enough stations, IsMuon considers the particle a muon. The station requirements are determined by momentum:

momentum, GeV/c	muon stations
$3 < p < p$	M2 && M3
$6 < p < 10$	M2 && M3 && (M4||M5)
$10 < p$	M2 && M3 && M4 && M5

Muon DLL

- N is the number of stations containing hits within the FOI
- $\{x,y\}_{closest}$ are the coordinates of the hit closest to the track extrapolation
- $\{x,y\}_{track}$ are the coordinates of the track extrapolation to the muon stations
- $pad_{\{x,y\}}$ are the muon pad sizes that determine hit coordinates uncertainty

$$D^2 = \frac{1}{N}\sum_{i=0}^{N}\left\{\left(\frac{x^i_{\text{closest}} - x_{\text{track}}}{\text{pad}_x}\right)^2 + \left(\frac{y^i_{\text{closest}} - y_{\text{track}}}{\text{pad}_y}\right)^2\right\}$$

The D^2 values are calibrated and intergrated to obtain the delta log likelihood.

Work on imporoving this approach by taking into account hits correlations is in progress using χ^2 algorithm [3].

Run II

Trained Catboost [7] and Adaboost from TMVA [8] on 2012 calibration samples

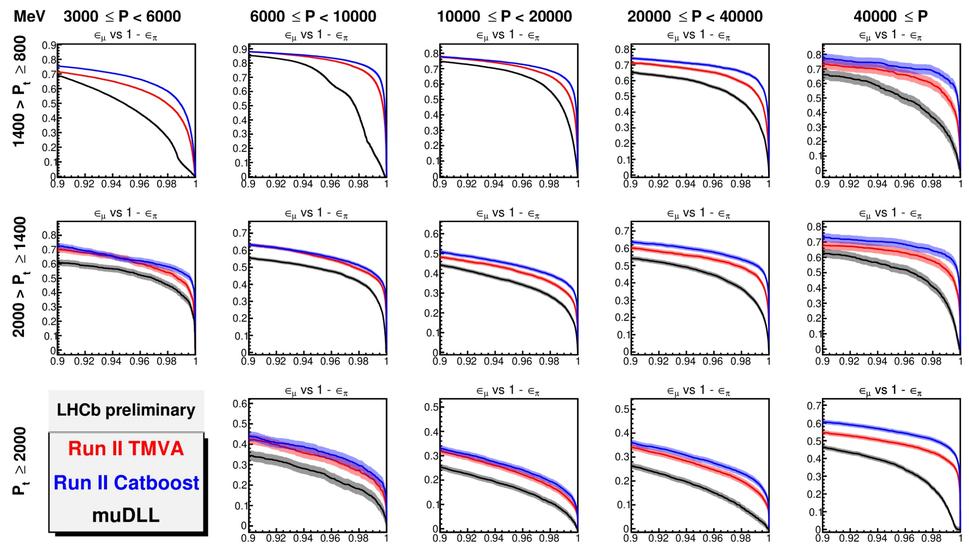
$$\text{muons: } J/\Psi \to \mu\underline{\mu},\ 4\cdot 10^5 \text{ tracks}$$
$$\text{pions: } D^* \to D^0(\to K\underline{\pi})\pi,\ 2\cdot 10^5 \text{ tracks}$$

- sWeights [9] were not used for training
- background was reweighted to match signal (P, P_t) spectrum

As features we used the information about the closest hits:
- hit time
- dT - delta of the hit time read by the vertical and horizontal strips
- whether the hit is crossed
- space residuals in x, y:

$$\frac{x_{\text{closest}} - x_{\text{track}}}{\text{pad}_x/\sqrt{12} + \text{MS}_{\text{error}}}$$

MS_{error} is the track extrapolation error estimated from multiple scattering



Muon efficiency vs pion rejection after applying IsMuon selection, no kinematic reweighting, 2016 calibration data

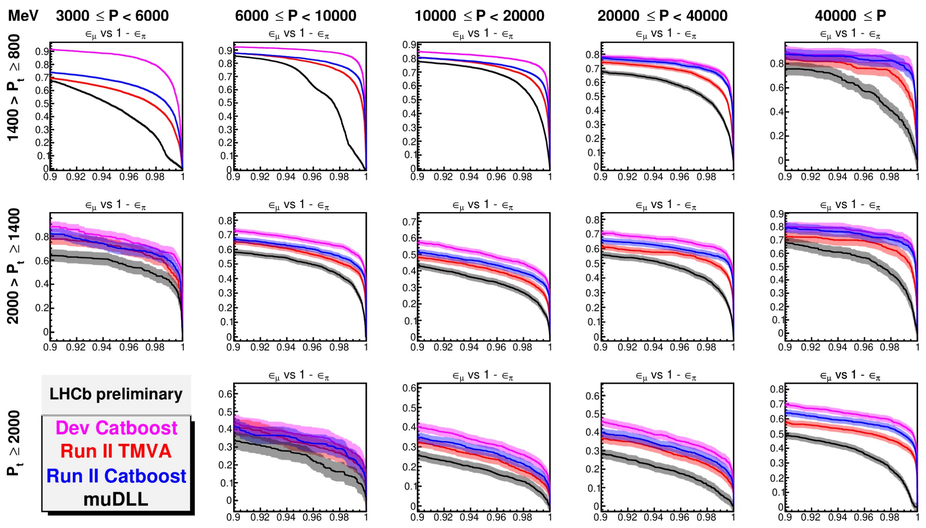
Towards the Run III

Catboost trained on 2016 calibration samples:

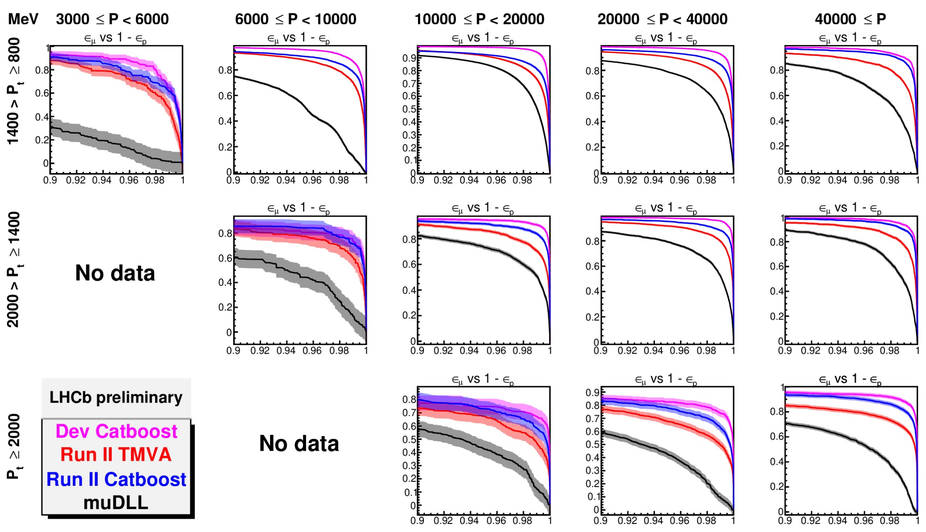
$$\text{muons: } J/\Psi \to \mu\underline{\mu},\ 8\cdot 10^6 \text{ tracks}$$
$$\text{pions: } D^* \to D^0(\to K\underline{\pi})\pi,\ 4\cdot 10^5 \text{ tracks}$$
$$\text{protons: } \Lambda_0 \to \pi\underline{p},\ 3\cdot 10^5 \text{ tracks}$$

- sWeights used in training (see my talk in Track 2 on Wednesday)
- background was reweighted to match signal (P, P_t) spectrum
- Both signal and background were reweighted by the number of primary vertices

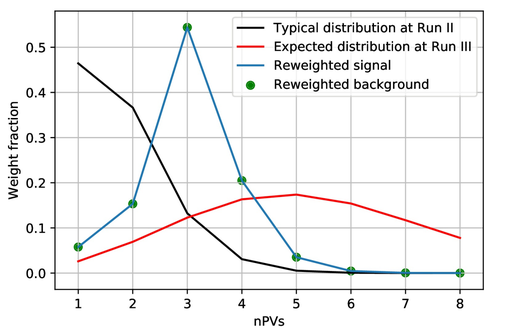
Features:
- Information about the closest hit: space residuals, coordinates, hit time, hit delta time, whether the hit is a crossed one
- Same information about the hits matched by the χ^2 algorithm
- Experimental high-level variables: χ^2, cluster sizes, number of clusters
- Momentum and transverse momentum



Muon efficiency vs pion rejection after applying IsMuon selection, after kinematic reweighting, 2016 calibration data



Muon efficiency vs proton rejection after applying IsMuon selection, after kinematic reweighting, 2016 calibration data

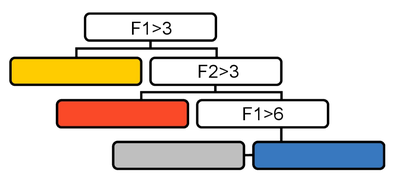


Challenge: there is not enough data with high number of primary vertices to accurately emulate the situation after LHCb Upgrade. We used a reweighting that added more emphasis on high-nPVs events.

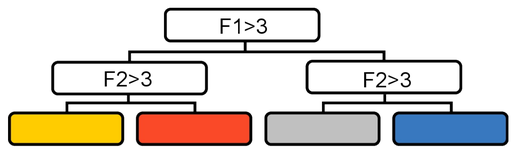
Why we use catboost?

Its oblivious trees [4] allow to gain a factor from 2 to 20 in evaluation speed compared to regular trees [5]. We need that for the trigger.

Decision Tree



Oblivious Trees



Conclusions

- Machine learning allows to gain significant improvement in the problem of matching hits to tracks, background rejection at 90% signal efficiency improves from 45% to 79% compared to muDLL on reweighted data after IsMuon
- For Run II conditions using catboost oblivious trees improves evaluation speed by a factor of 3 compared to TMVA BDT while maintaining higher discriminating power

References

[1] Archilli, F., et al. "Performance of the muon identification at LHCb." Journal of Instrumentation 8.10 (2013): P10020.
[2] Lanfranchi G. et al. The muon identification procedure of the LHCb experiment for the first data. – 2009. – N°. CERN-LHCb-PUB-2009-013.
[3] Cogoni, Violetta. LHCb-Novel Muon Identification Algorithms for the LHCb Upgrade. No. Poster-2016-522. 2016.
[4] Michael Levin - MatrixNet Applications at Yandex
[5] Catboost blog
https://catboost.ai/news/best-in-class-inference-and-a-ton-of-speedups
[6] LHCb Collaboration et al. LHCb PID upgrade technical design report. – 2013. – N°. CERN-LHCC-2013-022.
[7] Prokhorenkova L. et al. CatBoost: unbiased boosting with categorical features //Advances in Neural Information Processing Systems. – 2018. – C. 6639-6649.
[8] Hoecker A. et al. TMVA-Toolkit for multivariate data analysis //arXiv preprint physics/0703039. – 2007.
[9] Pivk, Muriel, and Francois R. Le Diberder. "Plots: A statistical tool to unfold data distributions." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 555.1-2 (2005): 356-369.