Advanced labaratory course

Selection of $B_s \to \psi(2S)K_s$ decays using a multivariate analysis

Theodor Zies
theodor.zies@tu-dortmund.de

Can Toraman
can.toraman@tu-dortmund.de

Pre-discussion: 07.07.2024

Hand-in: 28.07.2024

TU Dortmund – Physics department

Contents

1 Goal

The goal of this analysis is the classification of B_s candidates in a data samples from the LHCb experiment. This is achieved by employing simulation samples next to the data samples to train a multivariate classifier.

2 Theory

2.1 $B_s \to \psi(2S)K_S$ Decay

Before analyzing the data sample, looking at the studied decay $B_s \to \psi(2S)K_s$is important to understand it's kinematics. The B_s usually decays weakly as seen in Figure 1. At tree level the W boson creates the K_S (K-Short) and the $\psi(2S)$. The $\psi(2S)$ may decays into $\mu^+\mu^-$. This decay channel leaves a signature inside the LHCb detector which is very well reconstructable. Additionally the K_S's most probable decay channel is into $\pi^+\pi^-$. Next to the K_S the B_s may also decay into the K_L, but the ratio of K_L to K_S is negligible. Another interesting part of the decay are the lifetime of the particles. The K_S decays weakly leading it to have a relatively long lifetime, compared to the fast decaying $\psi(2S)$ which mostly decays strongly. At the LHCb experiment the vertices are reconstructed from the final states down to the *primary vertex* (pp collision). So the data samples consists of candidates where the two pions are reconstructed to the *tertiary vertex* (K_S decay). The K_S and the two muons are then reconstructed to the *secondary vertex* (B_s Decay).

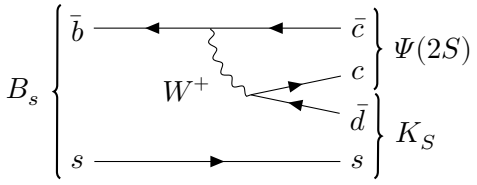

Figure 1: Feynmangraph of the decay $B_s \to \psi(2S)K_s$.

Through different effects inside the detector the data sample also consist of candidates that do not originate from the decay of the B_s or a different decay channel altogether. These *background* candidates may consist of unrelated particles mimicking the signal decay (combinatorial background) or the so-called partially reconstructed background. This secondary type of background is not considered in the analysis.

2.2 Boosted Decision Tree

To distinguish between background and signal candidates a multivariate classifier is applied. The classifier employed in the analysis is a Boosted Decision Tree (BDT). A Decision Tree is a classifier used to obtain a probability for each candidate of being signal or background. Boosting is the process of training multiple Decision Trees each on sub

sample of the training data. Each decision of the tree is then weighted depending on the performance of the tree on the last decision. A final decision is obtained by taking the weighted average of each tree. The process of splitting the data sample into k sub samples and then using $k-1$ for the BDT and one as a validation sample is called k-folding.

A Figure of Merit (FOM) is used during classification to obtain the optimal cut on the determined probability. A FOM is constructed depending on the classification problem at hand. In the employed analysis the FOM used is

$$\mathrm{FOM} = \frac{\epsilon_{\mathrm{sig}}}{\frac{5}{2} - \sqrt{N_{\mathrm{bck}}}}. \tag{1}$$

The idea in this case is to retain most of the signal candidates while reducing most of the background candidates.

To claim a discovery a significance S of 5 is required. S is determined by dividing the observable (in this case the signal counts N_{sig}) by its uncertainty. As the measurement is a poisson process its uncertainty can be approximated by $\sigma = \sqrt{N_{\mathrm{sig}} + N_{\mathrm{bkg}}}$, where N_{bkg} is the number of remaining background candidates. From this the assumed significance m calculates to

$$m = \frac{N_{\mathrm{sig}}}{\sqrt{N_{\mathrm{sig}} + N_{\mathrm{bkg}}}}. \tag{2}$$

3 The LHCb detector

The large hadron collider (LHC) located at CERN near Geneva is the largest particle accelerator to date. Opposing proton beams are collided at four different interaction points with center of mass energies up to 13 TeV. Every interaction point houses one experiment, with each of them having a detector that is optimized for a different physics purpose. The four experiments are called ALICE, ATLAS, CMS and LHCb. The data used in this analysis is recorded at the LHCb detector. This detector is optimized for the study of b physics and measurements of CP violation parameters, which is exactly the type of physics present in the B meson decays that are analysed in this lab course.
The LHCb detector is a single-arm forward spectrometer, this particular design was chosen because the particles studied at LHCb are mainly produced with a strong boost in the forward direction. The polar anglular coverage reaches from 15 to 300 mrad [3]. A schematic view of the LHCb detector can be seen in Figure 2 at the status during the first data-taking period. Here, the z-axis is along the beam direction, the negative z-region is referred to as upstream, while the positive z-region is called downstream.

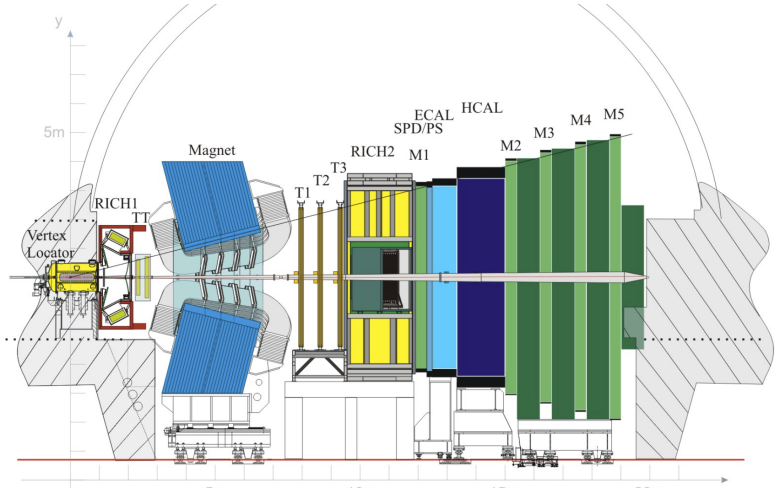

Figure 2: Cross-section of the LHCb detector, the following components are shown from left to right: The vertex locator (VELO), the Cherenkov detectors (RICH1,RICH2), the tracking system (TT and T1,T2,T3) including the magnet, the calorimeters (ECAL, HCAL) and the muon chambers (M1-M5).

The first component of the detector is the vertex locator (VELO), it is located directly at the interaction point. The VELO consits of multiple silicon strip detectors, meant to reconstruct the position of the proton-proton interaction, the primary vertex, and the decay location of the b hadron (secondary vertex). The B mesons typically have a flight path length of a few mm to cm and thus decay directly inside the VELO. The Ring Imaging Cherenkov (RICH) detectors 1 and 2 are used to calculate the velocity of transversing particles. This is archieved by measuring the diameter of light cones emitted by these particles due to the Cherenkov effect.
To obtain information about the particles momentum and charge, they are deflected by a dipole magnet with an integrated field strength of about 4 T m. Charged particles will then travel on a curved track. The tracking stations TT and T1-3 consit of large-area, four-layer silicon strip detectors as well as drift tubes. They measure the curved particle tracks and thus their radius can be determined. This allows it to calculate the momentum of the particle, while its charge is given by the direction of the particles deflection in the magnet. Combining the information about particle momentum, velocity and charge allows for particle identification (PID). Further downstream lies the calorimeter system, it is divided into the electromagnetic (ECAL) and hadronic (HCAL) calorimeters. The main purpose of the calorimeters is the measurement of the energy deposited by photons and electrons in the ECAL and hadrons in the HCAL, respectively. Additionally the calorimeters also contribute to PID. The last main components are the five muon stations located at the end of the detector. Here, muons that traverse all other detector parts without much interaction can be measured.
While the proton proton collisions occur at a rate of 11 MHz, event storage is only possible with a rate of 3 kHz [3]. To reduce the number of events, only physically interesting events are selected via a trigger system consisting of a hardware and software implementation.

4 Analysis strategy

For this analysis three types of samples are provided. Two simulation samples of the $B_s \to \psi(2S)K_s$ decay used as a signal proxy as well as the so-called control channel, a kinematically similar decay channel $B \to \psi(2S)K_S$ and the actual data sample. This data sample contains candidates from both the signal and control channel as well as the background candidates. For the later training of the classifier the so-called upper sideband (candidates with a invariant mass higher outside the signal region) of the data sample is used as a proxy for background.

As the distributions in the control and signal channel differ from those in the actual data sample it is necessary to align these. This so called *kinematic reweighting* has already been performed. As these generated weights align only some of the distributions in the samples it is important to find those deviating to much. These features are determined using *largest distance between the cumulative probability distributions* F^i

$$\sup_n \left|F_n^1 - F_n^2\right|, \tag{3}$$

where F^i is determined form histogramming the different features.

The considered decays are quite rare leaving the data sample with a lot of background. In order to project the B_0 distribution out of the background the sPlot method is applied. In the sPlot method the so-called sWeights are determined. These also have been provided in the given data sample.

5 Analysis

As a first step, the invariant mass of the simulated $B_s \to \psi(2S)K_S$ decay is visualized in Figure 3.

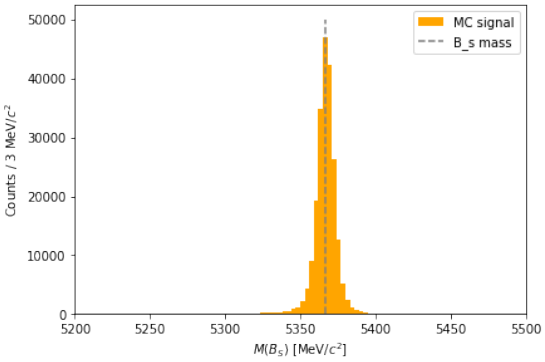

Figure 3: Reconstructed invariant mass of the simulated $B_s \to \psi(2S)K_S$ decay.

A sharp peak is visible around the nominal B_s mass, which is expect as the Monte Carlo simulation of this decay only simulates a signal without any background. Using this

distribution, a signal window is defined as the smallest possible intervall containing 99 % of the events in the simulation sample. The resulting signal interval is

$$I_{\text{sig}} = [5333.4,\ 5394.6]\,\text{MeV/c}^2.$$

The reconstructed mass of real data collected at the LHCb experiment is shown in Figure 4.

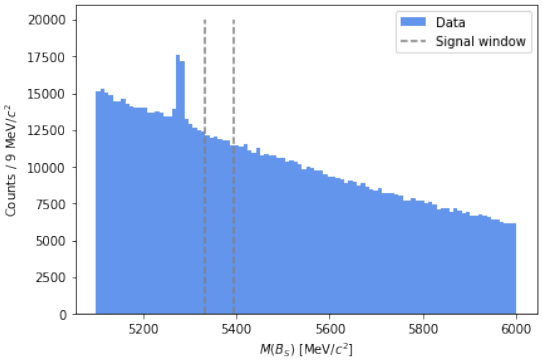

Figure 4: Reconstructed invariant mass of B^0 candidates from LHCb data.

As the data has already been preselected, a peak around the nominal B^0 mass is directly visible without any further selection. However, the B_s peak inside the signal window is still hidden due to the large amount of combinatorial background remaining in the data.

5.1 Feature selection

In order to train a multivariate classifier, suitable features have to be selected that can provide unbiased and meaningful information to the classifier. The first important aspect is that the features have to be well simulated, meaning they have to show good agreement with actual data. To obtain a data sample consisting of only signal events like the simulation, so called sWeights have already been computed for the data. Applying these sWeights to the data will lead to only signal-like events remaining, as it can be seen in Figure 5.

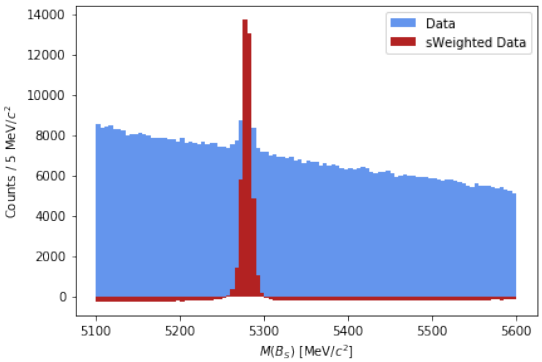

Figure 5: Data with and without applied sWeights.

Now, the distributions of the weighted data and control simulation can be compared. The control simulation is used here because the sWeights were determined in respect to the B^0 decay, as this is a lot more prominent in the data. Additionally, kinematic weights are applied to the control simulation which further improve the accuracy of the distributions. A weighted Kolmogorov Smirnov test is performed using (3), the resulting test statistic is a measure for how similar the distributions are. Subsequently, all features with a test statistic < 0.1 are removed. The same procedure is performed again with the data and signal simulation, however now the remaining features are checked for their discriminating power. This is achieved by using the upper mass sideband of the data (reconstructed mass above signal window), as these candidates only contain combinatorial background. Consequently, only features with with a test statistic > 0.5 are kept, which ensures that the classifier uses features from which it can learn clear differences in the distributions of the signal and background samples. The remaining features are checked for their correlation in respect to the invariant B^0 mass, all attributes with a correlation above 0.1 are discarded, as these could introduce a bias in the classifier. Lastly, the correlation of the leftover features is computed in respect to each other. Highly correlated attributes are removed as they only provide redundant information. In total, 16 features remain after all requirements. A few examples of the distributions of selected features are shown in Figure 6.

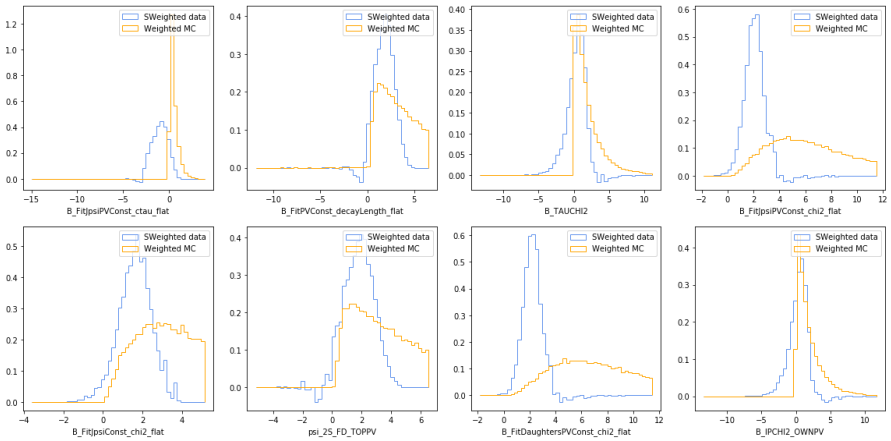

Figure 6: Some examples for the distributions of features selected for the classifier training.

These distributions show clear differences between signal and background, while also ensuring they are correctly modeled. This ensures that the classifier will learn the differences between signal and background and not between data and simulation.

5.2 Classifier training

As explained in the beginning, the classifier used in this analysis is a gradient boosted decision tree. Its implementation is done using the `XGBoost` package [1]. The signal proxy are all events of the signal simulation inside the previously determined signal window. As a background proxy, all events in the data sample with an invariant mass larger than the signal window are selected. Again, kinematic weights are applied to the signal simulation. The background events are also assigned with a weight that ensures that the classifier sees an equal amount of signal and background events. A five-fold cross validation is used during the training, resulting in five individual BDTs. Their performace is nearly identical, verifying this is important to avoid overtraining. The results and scores of the first BDT are shown in the following Figure 7.

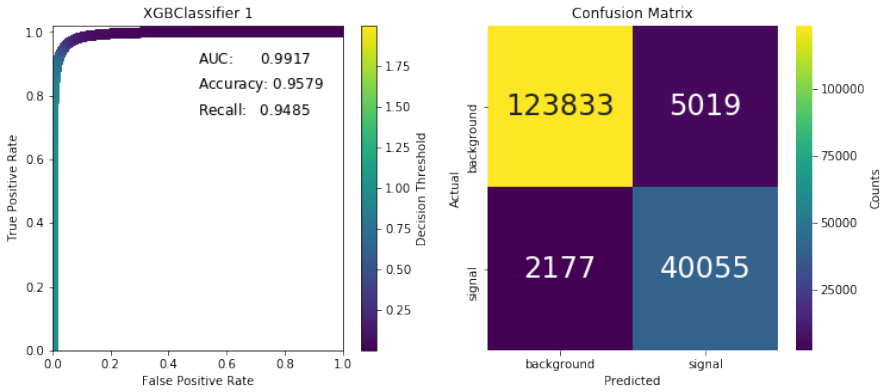

Figure 7: Results and scores of the first BDT.

The area under the ROC curve as well as precision and recall are all high, indicating a successful training. The response of this BDT for training and test data can be seen in Figure 8.

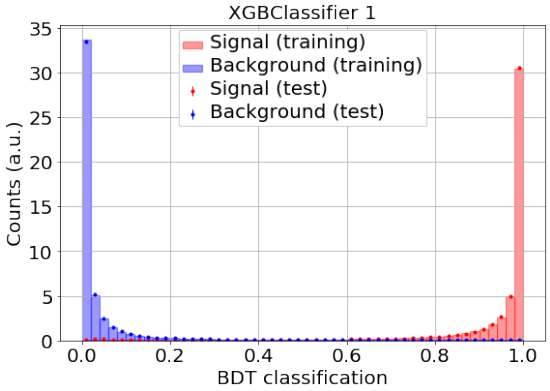

Figure 8: Response of the first BDT to training and testing data.

A very similar response to training and test data is observed, further verifying that no overtraining has occurred. The remaining results of the other four BDTs can be found in the appendix.

5.3 Threshold optimization

To obtain the optimal cut on the prediction of the five BDTs, the Punzi Figure of Merit (FOM) is used (1). For varying threshold values, the signal efficiency and number of background events in the signal window is computed after the respective BDT cut. The signal efficiency simply describes the ratio of signal events before and after the BDT cut. In contrast, the number of background events in the signal window is computed by fitting an exponential background model to the upper sideband and extrapolating the

number of events in the desired signal region. This is done separately for each BDT, the resulting five FOMs are plotted in Figure 9.

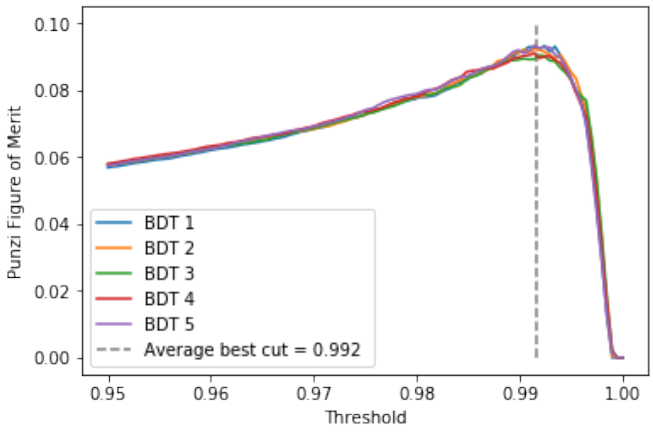

Figure 9: FOM for the five individual BDT including the averaged best cut.

The optimal cut is now determined as the average of the maxima of the five FOMs and also shown in Figure 9.

5.4 Determination of the signal yield

As the best cut has now been determined, the BDT can predict on the data and a corresponding cut is made on the BDT prediction. The following Figure 10 shows the data after this procedure.

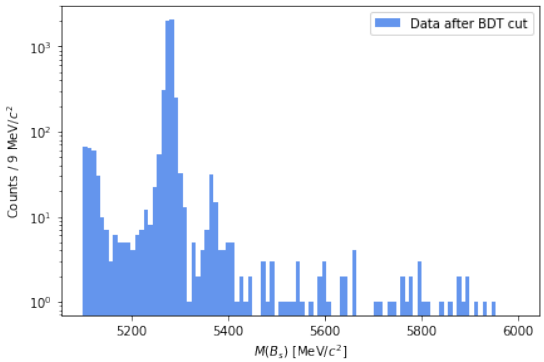

Figure 10: Distribution of the invariant B^0 mass in data after the BDT has been applied.

While the B^0 peak is still clearly visible as before, the second B_s peak can now also be seen. To calculate the yield of B_s decays, a fit has to be performed on the data. A large amount of partially reconstructed events is visible for masses below $5200\,\mathrm{MeV/c^2}$, which

is why the following fits are performed in the interval

$$I_{\text{fit}} = [5200,\, 5800]\,\text{MeV/c}^2.$$

Using the `iminuit` [2] package, an extended unbinned negative log-likelihood fit of the data model is produced. Here, the data model consist of an exponential background function, while the two signal peaks are each independently modeled by the sum of two Gaussian peaks. Before the full fit is performed, the mean and width of both signal peaks is fixed by a fit on the signal and control simulation, respectively. Consequently, the only free parameters of the final fit are the background yield, B^0 (control) yield, B_s (signal) yield and the slope of the exponential background. The fit itself and the results of all yields are shown in Figure 11.

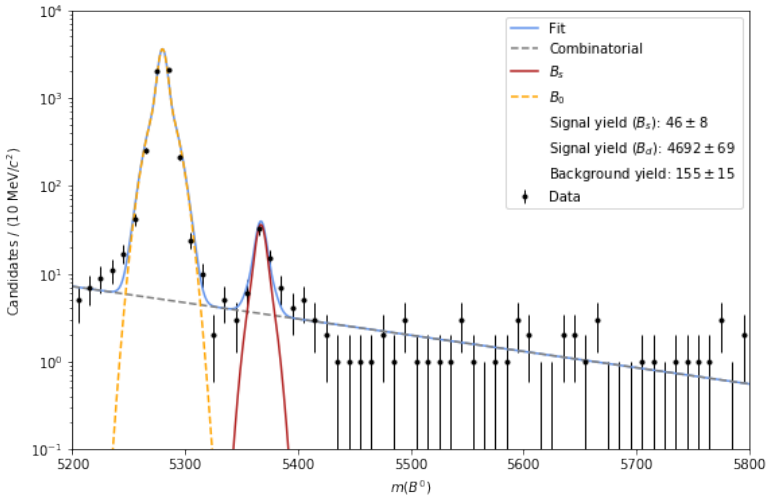

Figure 11: Fit to the invariant mass spectrum of the BDT selected data, consisting of a signal, control and background component.

In total, 46 ± 8 B_s candidates are observed. The significance of this peak is calculated as $S = 5.56\,\sigma$ with the help of (2). The number of background and signal events is obtained by integrating the individual functions of the signal and background component over the signal interval. All of the parameters of these functions are obtained in the previous fit.

6 Discussion

Some remarks on the analysis should be made to put the final result into more context. While the distributions of the chosen variables show good separation, there is still room for optimization. The number of variables was arbitrarily chosen as 16, but can easily be varied when changing the requirements on the Kolmogorov Smirnov test statistics or correlations. Testing the BDT performance with different number of parameters could help to improve its predictive power. The five individual BDTs performed very similar,

indicating that there is not too much overtraining. However, the BDTs have all been initialized with default parameters. Tuning these parameters in a grid search could definitively help improve the BDT performance and thus the final result. The fits on the signal peaks have been modeled by two Gaussian functions, however they cannot describe the asymmetric shape of these peaks perfectly. A more complicated model like a double-sided crystal ball function can improve the fit and thus directly the yield of the signal. The significance was only estimated using (2) and seems rather high being greater than $5\,\sigma$. Including uncertainties should help to mitigate this fact.

7 Appendix

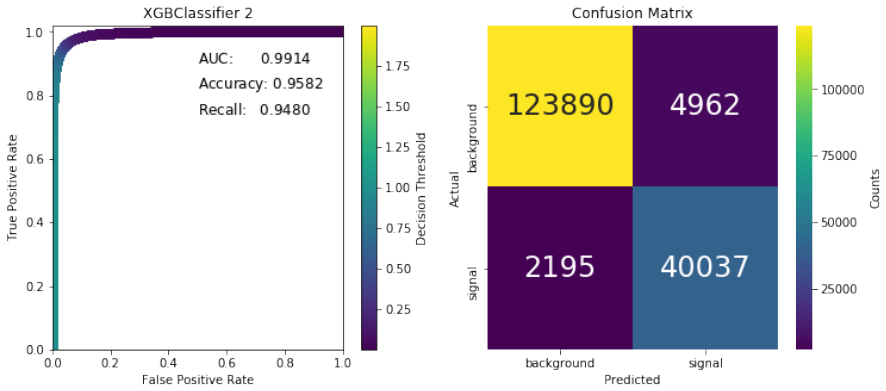

Figure 12: Results and scores of the second BDT.

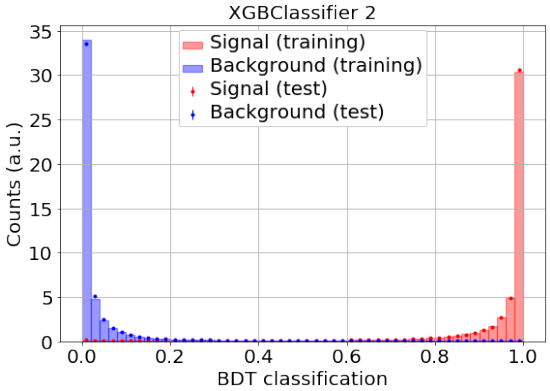

Figure 13: Response of the second BDT to training and testing data.

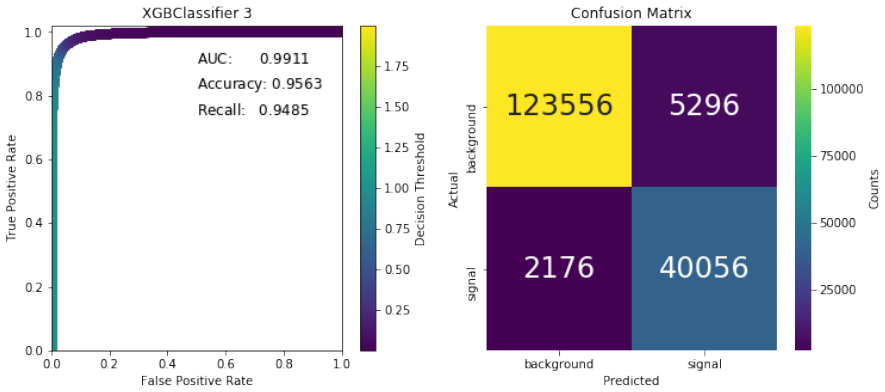

Figure 14: Results and scores of the third BDT.

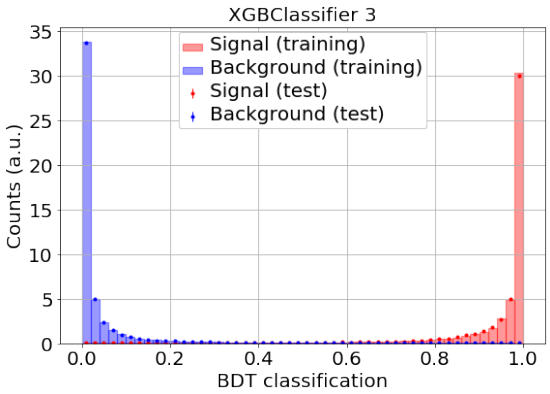

Figure 15: Response of the third BDT to training and testing data.

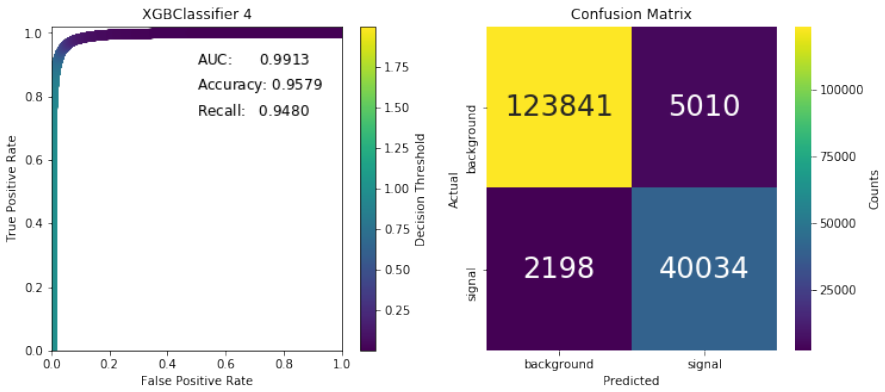

Figure 16: Results and scores of the fourth BDT.

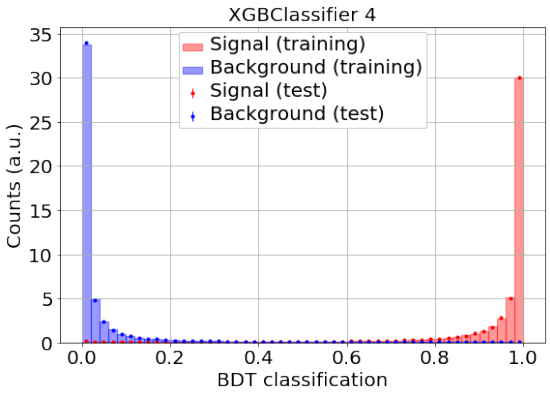

Figure 17: Response of the fourth BDT to training and testing data.

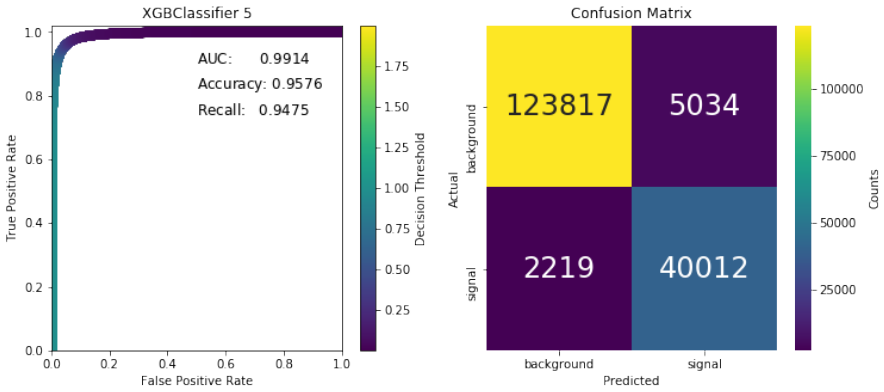

Figure 18: Results and scores of the fifth BDT.

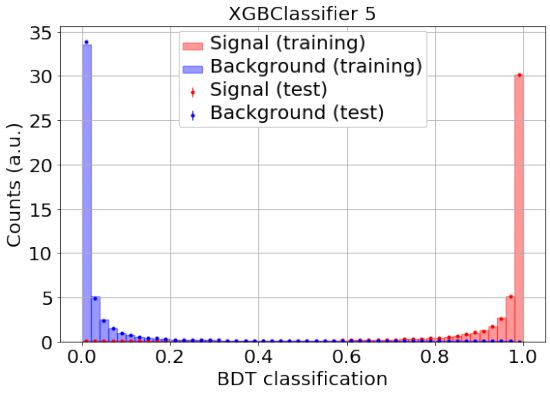

Figure 19: Response of the fifth BDT to training and testing data.

References

[1] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[2] H. Dembinski and P. Ongmongkolkul et al. "scikit-hep/iminuit". In: (Dec. 2020). DOI: 10.5281/zenodo.3949207.

[3] *Selection of Bs0 → (2S)KS0 events*. TU Dortmund, Department of physics.