Identifying events involving Higgs Boson in the particle accelerator - applications of Big Data processing in the Cloud and Machine Learning in particle physics

Machine learning model

- Created with Python and XGBoost library using boosted trees technique applied to HIGGS dataset
- Trained on 100000 records of the total 11000000 and 1500 training epochs
- Obtained training score: auc around 0.96, logloss around 2.16
- Obtained evaluation score: auc around 0.81, logloss around 6.48

Thoughts on model performance

- Overfitting occurred despite using regularization, adding more data (the rest of the original dataset) would be likely to fix it
- In order to reduce overfitting gamma and alpha parameters could be increased, also trying other values of sub_sample, max_depth and eta could help
- Training for 3000 epochs instead of 1500 could also slightly improve the obtained scores

Information gathered from the dataset

Besides using it for machine learning purposes, we took a look at the dataset and tried to model the data as information more understandable for human beings.

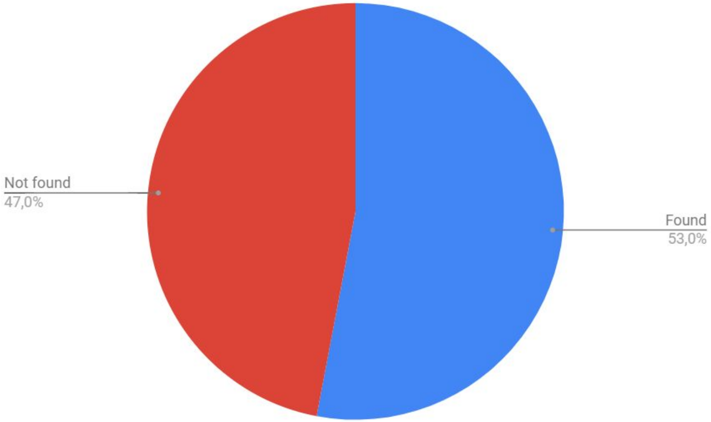
Example:

Number of records where boson was found:
5829123
Number of records where boson was not found:
5170877

The Boson was found in 52.99% of the cases

Information gathered from the dataset

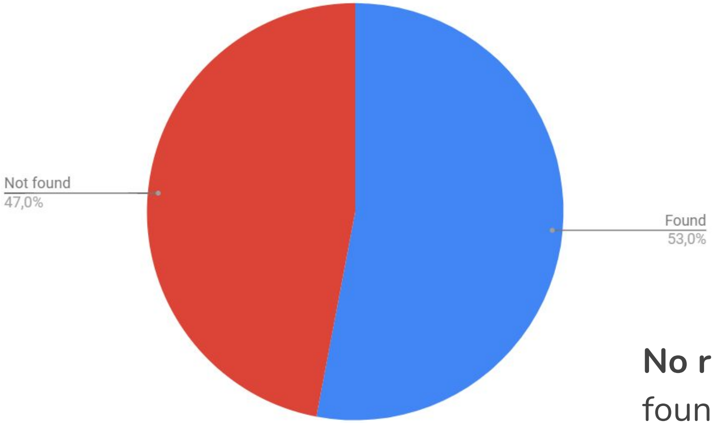
We also tried to study some of the columns in the dataset, to see if there were differences between the cases where boson was found and the cases where it was not.



Positive values for column 2

Boson Found: 2914001

Boson Not Found: 2585459



Negative values for column 2

Boson Found: 2915122

Boson Not Found: 2585418

No relation was found between the sign of column 2 and the presence of the boson