Report

March 21, 2023

1 Project-Develop-an-end-to-end-Machine-Learning-Pipeline

Free project for the Machine Learning Python Labs course # Higgs Boson Event Detection

1.1 Backstory

1.1.1 Particle accelerators

In order to investigate the underlying principles of matter, space, and time, as well as their structure, physicists examine the most basic interactions, such as the collision of subatomic particles, at extremely high energies. Particle accelerators allow scientists to study the fundamental properties of matter by observing the subatomic particles generated during high-energy particle beam collisions. However, the experimental data gathered from these collisions inevitably have limitations in terms of precision. This is where machine learning (ML) plays a role. Researchers typically utilize standard machine learning software packages to analyze the data collected from these experiments. They devote a significant amount of effort to enhance the statistical power by identifying and extracting important features from the raw measurements.

1.1.2 Higgs boson

Higgs Boson. Often referred to as the “God particle” in popular media, the Higgs boson particle is the final component of the Standard Model of particle physics, which defines the fundamental particles and forces that govern the subatomic world. Elementary particles are believed to be massless at extremely high energies, but some can acquire mass at lower energies. The mechanism behind this mass acquisition puzzled theoretical physicists for a long time. In 1964, Peter Higgs and others proposed a mechanism that could theoretically explain the origin of mass for elementary particles. This mechanism involves a field, commonly known as the Higgs field, with which particles can interact to gain mass. The more a particle interacts with the field, the heavier it becomes. Some particles, like the photon, do not interact with the Higgs field at all, remaining massless. The Higgs boson particle is the associated particle of the Higgs field (each fundamental field has one). Essentially, it is the physical manifestation of the Higgs field, which imparts mass to other particles. The discovery of this elusive particle took almost 50 years after its initial theoretical proposal!

The breakthrough. On July 4, 2012, the ATLAS and CMS experiments at CERN’s Large Hadron Collider revealed that both had detected a new particle in the mass region around 125 GeV. This particle aligned with the predicted Higgs boson. This experimental validation led to François Englert and Peter Higgs being awarded the 2013 Nobel Prize in Physics.

The prize was given “for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider."

Giving mass to fermions. The Higgs boson can decay and produce other particles through various processes. These potential transformations during decay are known as channels in physics. The Higgs boson was initially observed to decay in three separate channels, all involving boson pairs. To confirm that the Higgs field is responsible for providing the interaction that imparts mass to fundamental fermions (particles that follow Fermi-Dirac statistics, as opposed to bosons which follow Bose-Einstein statistics), it must be shown that the Higgs boson can decay into fermion pairs via direct decay modes. Consequently, investigating the decay of the Higgs boson into fermion pairs (such as tau leptons or b-quarks) and accurately measuring their properties became a crucial area of research. Among the possible modes, the decay into a pair of tau leptons is the most promising, as it offers a moderate branching ratio along with manageable backgrounds.

The first evidence of $h \to \tau^+\tau^-$ decays was recently reported, based on the full set of proton–proton collision data recorded by the ATLAS experiment at the LHC during $2011 - 2012$. Despite the consistency of the data with $h \to \tau^+\tau^-$ decays, it could not be ensured that the statistical power exceeds the 5σ threshold, which is the required standard for claims of discovery in high-energy physics community.

1.2 LHC at Work

Proton-proton collisions. In particle physics, an event describes the outcome immediately following a fundamental interaction between subatomic particles. This interaction occurs within an extremely brief timeframe and is confined to a specific region in space. At the LHC, clusters of protons are accelerated in both directions along a circular path at extraordinarily high speeds. These clusters are directed to intersect within the ATLAS detector, resulting in hundreds of millions of proton-proton collisions per second. Sensors detect the events that ensue, generating a sparse vector with approximately one hundred thousand dimensions, which is somewhat analogous to images or speech signals in traditional machine learning applications. During the feature construction phase, the type, energy, and 3D direction of each particle are extracted from the raw data. Additionally, the variable-length list of four-tuples is converted into a fixed-length vector of features containing up to tens of real-valued variables.

Background events, signal events and selection region. Some of these variables are initially utilized in a real-time, multi-stage cascade classifier (known as the trigger) to filter out the majority of uninteresting events, referred to as background events. The chosen events, approximately 400 per second, are then stored on disks by a large CPU farm, generating petabytes of data annually. The vast majority of these saved events still represent known processes and are also considered background events. Background events primarily stem from the decay of particles that, while exotic, have been previously identified in earlier experiments.

The objective of offline analysis is to pinpoint a region within the feature space (called the selection region) that yields a significantly higher number of events (termed signal events) than can be explained by known background processes. Once this region is established, a statistical test is conducted to assess the significance of the excess. If the likelihood that the excess was generated by background processes falls below a specific threshold, it signifies the discovery of a new particle.

The classification problem. In order to optimize the selection region, multivariate classification techniques are frequently employed. The formal objective function is distinct and somewhat deviates from the classification error or other objectives commonly used in machine learning. However, identifying a pure signal region is roughly analogous to separating background events from signal events, which is a typical classification problem. As a result, established classification methods prove beneficial, as they offer improved discovery sensitivity compared to traditional, manual approaches.

Weighting and normalization. The classifier is trained using simulated background events and signal events. Simulators generate weights for each event to correct for discrepancies between the event's prior probability and the instrumental probability applied by the simulator. These weights are normalized so that within any region, the sum of the event weights falling into that region provides an unbiased estimate of the expected number of events found there for a fixed integrated luminosity. This corresponds to a fixed data collection time for a specific beam intensity. In this case, it is related to the data gathered by the ATLAS experiment in 2012.

Since the probability of a signal event is typically several orders of magnitude lower than the probability of a background event, the signal samples and background samples are usually renormalized to create a balanced classification problem. A real-valued discriminant function is then trained on this reweighted sample to minimize the weighted classification error. The signal region is defined by setting a threshold on the discriminant value, which is optimized using a held-out set to maximize the sensitivity of the statistical test.

The broad goal is to improve the procedure that produces the selection region, i.e. the region (not necessarily connected) in the feature space which produces signal events.

1.3 Enter ML

Shallow neural network. Machine learning is crucial for analyzing data obtained from particle collider experiments. ML classifiers are trained to differentiate between various types of collision events using simulated data from advanced Monte-Carlo programs. Shallow neural networks with a single hidden layer are among the primary techniques employed for this analysis, and standardized implementations are integrated into the widely-used multivariate analysis software tools utilized by physicists. To enhance statistical power, efforts are often concentrated on creating new features to be used in conjunction with existing machine learning classifiers. These high-level features are non-linear functions of the low-level measurements and are derived through an understanding of the underlying physical processes.

Deep neural network. The wealth of labeled simulation training data and the intricate underlying structure make this a perfect application for deep learning, specifically for large, deep neural networks. Deep neural networks have the potential to streamline and enhance the analysis of high-energy physics data by automatically learning high-level features from the data. Notably, they can boost the statistical power of the analysis without relying on manually derived high-level features.

1.4 Data

Source: https://www.kaggle.com/competitions/higgs-boson/data

The simulator. The dataset is constructed using official ATLAS full-detector simulations. The simulator consists of two parts. First, it simulates random proton-proton collisions based on our accumulated knowledge of particle physics, replicating the random microscopic explosions that occur during these collisions. In the second part, the resulting particles are tracked through a virtual model of the detector. This process generates simulated events that closely resemble the statistical properties of real events while providing additional information about what transpired during the collision, prior to the particles being measured in the detector.

Signal sample and background sample. The signal sample includes events where Higgs bosons (with a fixed mass of 125 GeV) were produced. The background sample was created by other known processes that can generate events with at least one electron or muon and a hadronic tau, imitating the signal. For the dataset, only three background processes were retained.

The first process comes from the decay of the Z boson (with a mass of 91.2 GeV) into two taus. This decay creates events with a topology very similar to that of the Higgs decay. The second set of background events consists of those involving a pair of top quarks, which can include a lepton and a hadronic tau among their decay products. The third set of background events is associated with the decay of the W boson, where one electron or muon and a hadronic tau can appear simultaneously only due to imperfections in the particle identification procedure.

Training set and test set. The training set and the test set respectively contains 250000 and 550000 observations. The two sets share 31 common features between them. Additionally, the training set contains **labels** (**signal** or **background**) and **weights**.

1.5 Project Objective

The objective of the project is to classify an event produced in the particle accelerator as background or signal. As described earlier, a **background event** is explained by the existing theories and previous observations. A **signal event**, however, indicates a process that cannot be described by previous observations and leads to the potential discovery of a new particle.

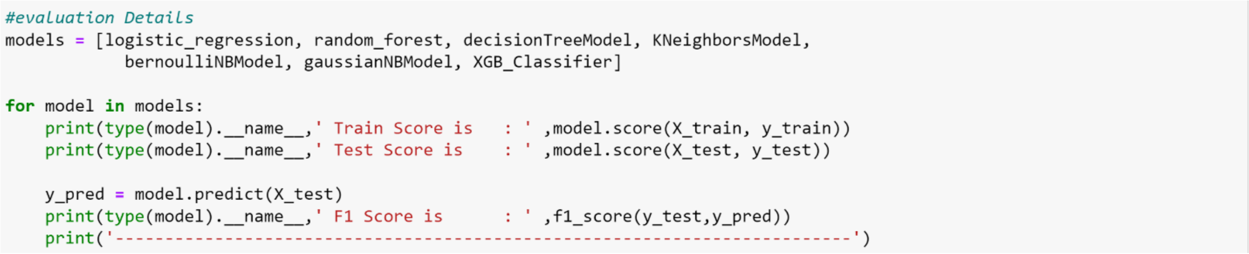
1.6 Evaluation Metric

The **evaluation metric**, used in this project, is the approximate median significance (AMS), given by

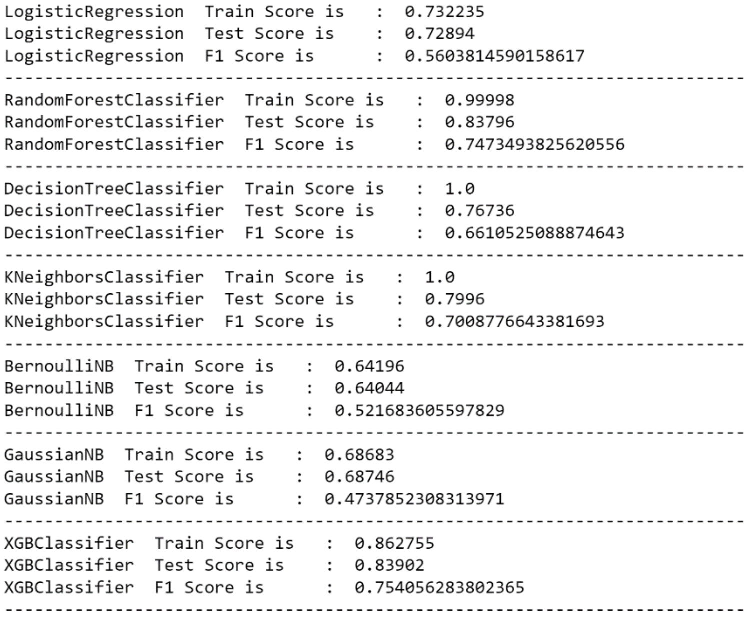
$$AMS := \sqrt{2\left((s+b+b_r)\log\left(1+\frac{s}{b+b_r}\right)-s\right)},$$

where - s : unnormalized **true positive rate**, - b : unnormalized **false positive rate**, - $b_r = 10$: constant regularization term, - $\log$: **natural logarithm**.

[ ]:

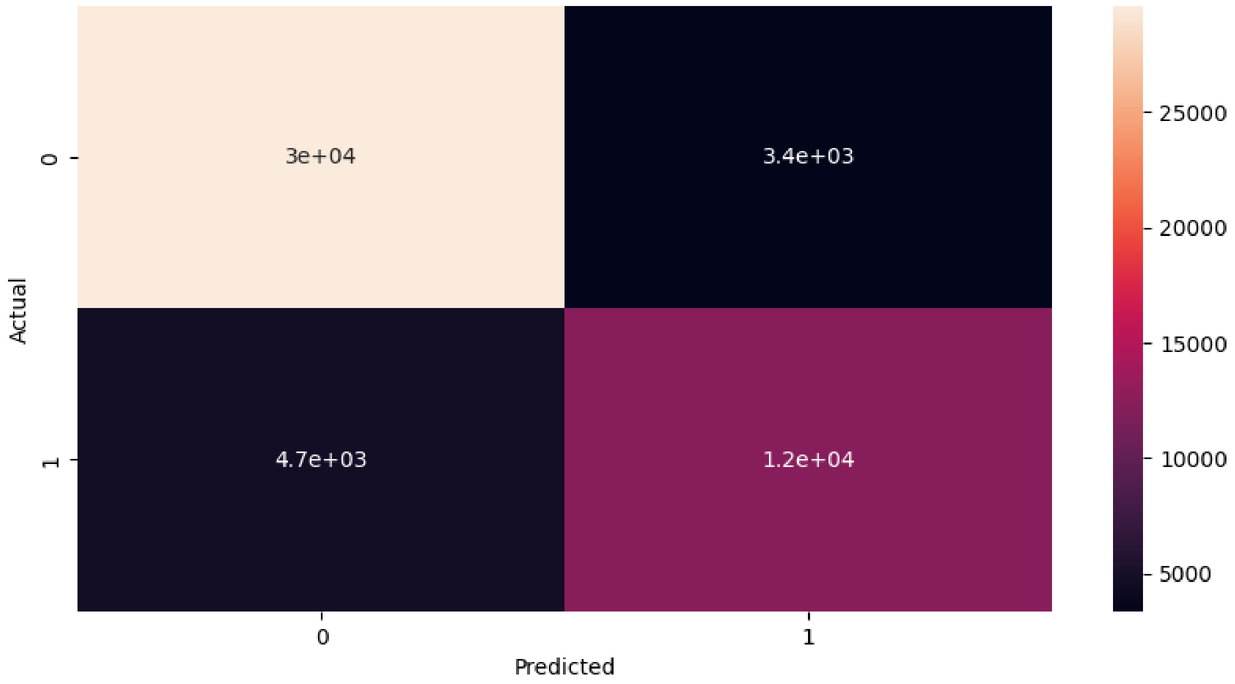


```python
#evaluation Details
models = [logistic_regression, random_forest, decisionTreeModel, KNeighborsModel,
            bernoulliNBModel, gaussianNBModel, XGB_Classifier]

for model in models:
    print(type(model).__name__,' Train Score is   : ' ,model.score(X_train, y_train))
    print(type(model).__name__,' Test Score is    : ' ,model.score(X_test, y_test))

    y_pred = model.predict(X_test)
    print(type(model).__name__,' F1 Score is      : ' ,f1_score(y_test,y_pred))
    print('--------------------------------------------------------------------------')
```



```
LogisticRegression  Train Score is   :  0.732235
LogisticRegression  Test Score is    :  0.72894
LogisticRegression  F1 Score is      :  0.5603814590158617
--------------------------------------------------------------------------
RandomForestClassifier  Train Score is   :  0.99998
RandomForestClassifier  Test Score is    :  0.83796
RandomForestClassifier  F1 Score is      :  0.7473493825620556
--------------------------------------------------------------------------
DecisionTreeClassifier  Train Score is   :  1.0
DecisionTreeClassifier  Test Score is    :  0.76736
DecisionTreeClassifier  F1 Score is      :  0.6610525088874643
--------------------------------------------------------------------------
KNeighborsClassifier  Train Score is   :  1.0
KNeighborsClassifier  Test Score is    :  0.7996
KNeighborsClassifier  F1 Score is      :  0.7008776643381693
--------------------------------------------------------------------------
BernoulliNB  Train Score is   :  0.64196
BernoulliNB  Test Score is    :  0.64044
BernoulliNB  F1 Score is      :  0.521683605597829
--------------------------------------------------------------------------
GaussianNB  Train Score is   :  0.68683
GaussianNB  Test Score is    :  0.68746
GaussianNB  F1 Score is      :  0.4737852308313971
--------------------------------------------------------------------------
XGBClassifier  Train Score is   :  0.862755
XGBClassifier  Test Score is    :  0.83902
XGBClassifier  F1 Score is      :  0.754056283802365
--------------------------------------------------------------------------
```

I will use XGBClassifier Model



```python
from sklearn.metrics import accuracy_score,classification_report

print(accuracy_score(y_test,y_pred).round(4)*100,'\n')

print(pd.crosstab(y_test,y_pred),'\n')

print(classification_report(y_test,y_pred),'\n')
```

```
83.89999999999999

col_0      0      1
Label
0      29612   3359
1       4690  12339

              precision    recall  f1-score   support

           0       0.86      0.90      0.88     32971
           1       0.79      0.72      0.75     17029

    accuracy                           0.84     50000
   macro avg       0.82      0.81      0.82     50000
weighted avg       0.84      0.84      0.84     50000
```