A strategy to tackle the Higgs Boson Challenge

Hugo Witz 284336, Maria Tager 311431, Joana Malvar 315029 *CS-433: Machine Learning, EPFL, Switzerland*

***Abstract*—This project aims at providing an answer to the Higgs Boson Machine Learning Challenge. Scientists are not able to observe a Higgs boson due to its rapid decay, our objective is thus to estimate the likelihood that a given event's signature was the result of a Higgs boson. To tackle this problem, regression methods and feature engineering were applied, achieving binary classification on the provided CERN particle accelerator data.**

I. Introduction

The ATLAS experiment conducted by CERN succeeded at observing a Higgs boson signal as it decayed into two tau particles. However, distinguishing the signal from the background noise was difficult for the physicists. To remediate this, the Higgs Boson Challenge was proposed to data scientists. Given a training set composed of 250 000 events, associated with 30 features and their outcome, the aim was to design a model combining the right machine learning and statistical tools, to perform binary classification on unseen data: "s" for the signal, and "b" for background. In this project, our own model is proposed. It was built after cleaning and processing the data, choosing of an optimal regression method.

II. Methods

A. Data pre-processing

After understanding the structure and the role of each feature in the dataset as described in the CERN documentation, exploratory data analysis followed by pre-processing is primordial in order to prevent errors in the measurements to affect the accuracy of our model. In the following part, the main data cleaning tasks undertaken are described: [1]

1) **Handling missing values:** Some values in the dataset are considered as meaningless or cannot be computed. These are represented with -999 in the raw data. They were thus replaced with the median of the non -999 values of the entire feature column. Since the data is skewed (contains outliers), replacing by the median rather than the mean is more appropriate to handle those missing values.

2) **Handling outliers:** As mentioned previously, the data is skewed, meaning it contains outliers. An outlier is defined as a value that lies an abnormal distance from other values in the same feature column. In order to eliminate the noise they generate, the boundary was set at the 95th percentile and exclude the entries located outside of this range.

3) **Splitting the data depending on the jet number feature:** One categorical feature is present within the data: the number of jets recorded in the detector after collision PRI_jet_num. It can take the discrete values of 0, 1, 2 or 3, possible larger values being capped at 3. Since many measurements in other features are correlated to the number of jets, a decision was made to split the data depending on the jet number and to run the model separately on the different subsets. To maintain balance between the different subsets, the data was split the data into 3 parts: 0 jet, 1 jet, and 2 or more jets.

4) **Correlated features:** The primitives (PRI), come from the raw data measured, whilst the derived (DER), were computed from the primitives. [2] This might cause for dependencies between the attributes. To assess this, a correlation matrix was computed and the features presenting a correlation larger than 0.9 were subsequently removed. After this step, 10 features were eliminated.

5) **Data standardization:** Scaling numerical data prior to modeling is an important pre-processing step to improve the performance of predictive modeling algorithms in machine learning. Standardization was thus performed on the data.

B. Feature expansion:

It is very likely that the theoretical value of the existence of the Higgs boson decaying into tau particles is based on a combination of parameters. Considering this, and with an aim to better fit our date, feature expansion was performed by adding a polynomial basis to the features. This allowed to obtain an extended feature vector. The chosen model is then fit to it, leading to an increased representational power.

C. Model choice:

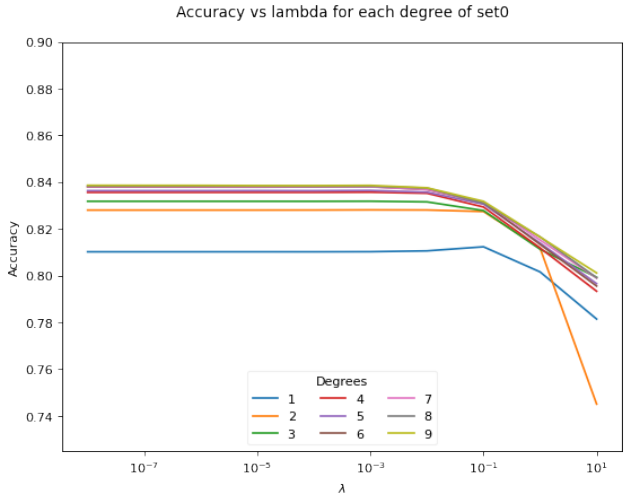
The regular single split of the data into one testing and one training set is not ideal. Instead, K-fold cross-validation is used to obtain an unbiased estimate of the generalization error and variance. Indeed, the model can be trained K times in total, each time training on K-1 groups and testing on the remaining, and outputting the average values. This method allows to have a first intuition of the most appropriate model for our problem. After processing the data, and using 10-fold cross validation, the accuracies were obtained for each tested model, as shown in the Table 1. Ridge regression presented the best result at this stage, so it was selected. Its hyperparameters still had to be optimized, so these were further investigated, as described in the following section of the report.

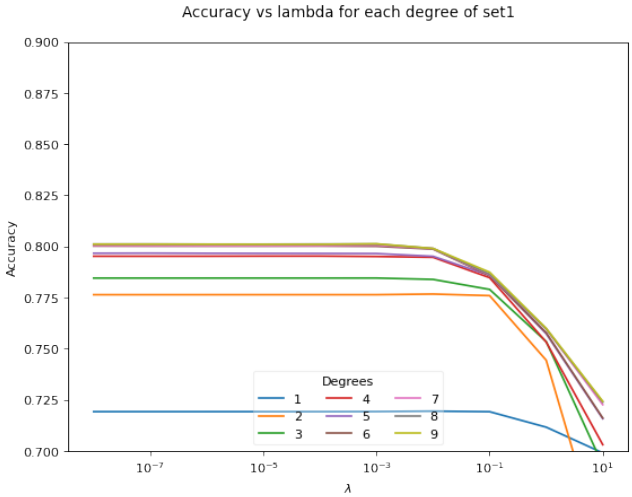
Model	Mean accuracy ± std
Stochastic gradient descent	0.5745 ± 0.0515
Ridge Regression	0.8219 ± 0.0159
Gradient descent	0.6904 ± 0.0387
Logistic regression	0.6168 ± 0.0813

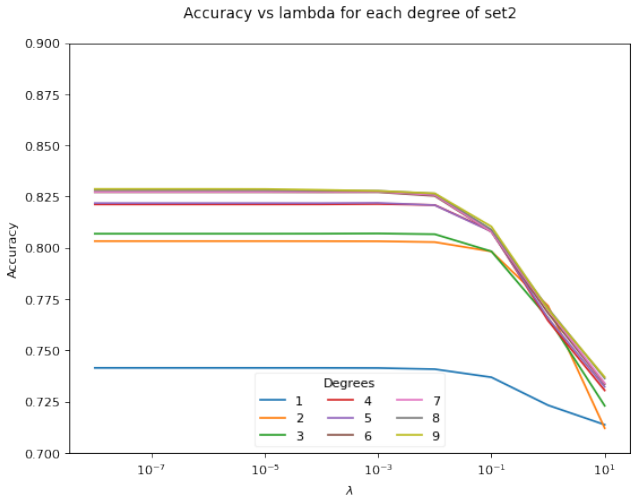
TABLE I
FIRST COMPARISON TO CHOOSE THE MOST APPROPRIATE MODEL USING 7 AS POLYNOMIAL DEGREE

D. Choice of hyperparameters:

Two main parameters are to be investigated and chosen in order to run our model in the best possible way: the tradeoff hyperparameter λ for the ridge regression algorithm, and the degree of the polynomial basis for the feature expansion.
The figures below show the obtained accuracies using different degrees and different tradeoff hyperparameters. Each figure corresponds to one jet number subset. The accuracy is measured using 10-fold cross validation.







The final degree and λ chosen for each jet number subset are presented in the table below. It is clear from the graphs that λ should stay close to 0, to avoid underfitting.

PRI_jet_num	λ	degree
0	1e-5	9
1	1e-5	9
≥ 2	1e-5	9

TABLE II
CHOSEN HYPERPARAMETERS

III. RESULTS

The pipeline described in this report lead to a final accuracy of 0.824 and an F1-score of 0.727 on the provided AIcrowd test dataset.

REFERENCES

[1] CERN. Higgs boson machine-learning challenge, 2014. https://home.web.cern.ch/fr/node/3963.
[2] Kaggle. Higgs boson machine learning challenge, 2014. https://www.kaggle.com/competitions/higgs-boson/data.