Higgs Boson Observation Detection In Experimental Data

Team °Cloud: Farouk Boukil, Albin Vernhes, Juliette Parchet

October 31^{st}*, 2022*

***Abstract*—The Higgs boson is an elementary particle in the Standard Model of physics which explains why other particles have mass. In the field of micro-physics, experimental data is often very noisy because of the complex interactions of multiple phenomena that make it difficult to detect the underlying effect of interest. In this project, we help CERN researchers partially overcome this issue by performing data processing, and implementing and improving a linear classifier, logistic regression, to detect the presence or absence of a Higgs Boson particle based on experimental data, with accuracy** 62%**.**

I. Introduction

We have available to us labeled raw observational data obtained experimentally. The scientific features and data background are explained in details in the scientific description[1].

Given the observations of an experiment, our goal is to detect accurately the presence (denoted HB+) or absence (denoted HB-) of a Higgs Boson. To that end, we use a variety of logistic regression based models with stochastic gradient descent optimization.

Before diving in, data analysis is in order to better grasp the relevance of each feature. Following that, we perform feature selection and engineering to generate a clean data set with limited noise and amplified correlations. Finally, we implement and train our models on a training set, tune the hyperparameters, and validate our model on a validation set. Our best accuracy on logistic regression on AIcrowd is 62%. We try to improve it by combining models to avoid using high order polynomial expansions which are very costly given our resources. Our overall best accuracy on AIcrowd however still is 70% with linear regression.

II. Models and Methods

A. Data Cleaning and Feature Selection

The data set contains 250'000 experiments described by up to 30 observations each. After plotting the feature distribution by label of all features in the data set, and removing outliers using box plots and quantile truncation, we eliminated 19 out of the initial 30 features. A feature is eliminated if:

- Its by label distributions plot does not show a region of significant dominance of either of the labels when the data set is randomly balanced[2]. That is, for most values of this feature F, empirically:
$$\frac{P(HB+|F)}{P(HB-|F)} \approx 1$$
- After removing outliers, the feature is almost orthogonal to the prediction (the covariance is almost zero). An example of such a feature is PRI_lep_eta.
- Its values are undefined, not just missing, for more than 30% of experiments, which is the case for many of the $jet\_(sub)leading$ features group. In fact, we do not wish for our general models to predict based on occasional features that cannot be estimated.

Based on the previous criteria, we remain with the following features: DER_mass_MMC, $DER\_mass\_transverse\_met\_lep$, DER_mass_vis, $DER\_pt\_ratio\_tau\_lep$, $DER\_deltar\_tau\_lep$, PRI_tau_pt, DER_sum_pt, $DER\_met\_phi\_centrality$, PRI_jet_num, PRI_met, PRI_met_sumet

B. Feature Engineering

Our chosen features benefited from further denoising, by clipping the continuous features' by label distributions while keeping at least 95% of their by label density. Figure (1) depicts the result on DER_mass_MMC, that is also an example of a "useful" feature in contrast with the stated elimination criteria (covariance with labels, not mentioned below, is $+0.3687$).

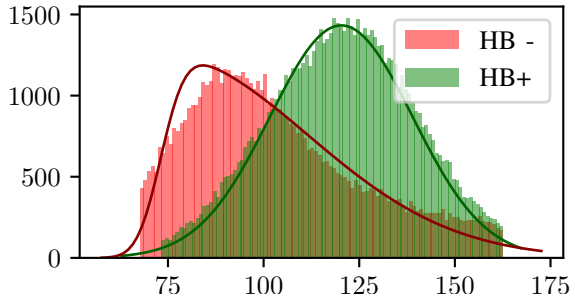

Fig. 1. by label clipped distributions of `DER_mass_MMC`, the data set is balanced to allow comparison.

Apart from $DER\_met\_phi\_centrality$ which was binned at 1.25[3], and PRI_jet_num which is cathegorical, the other features were z-score normalized to avoid gradient explosion and to smoothen the convergence to a local minimum.

We opted for polynomial feature expansion in all of our models prior to normalization.

C. Classifiers

1) Logistic Regression (Logit): With Cross Validation Grid Search (CVGS), for 2000 iterations and SGD batch size of 100, Logit performed best with no regularization (against $10^{-3}, 10^{-2}, 10^{-1}, 1, 10$), 10^{-2} learning rate (against $10^{-3}, 10^{-1}, 1$), null initial weights (against random initialization). Figure (2) depicts the empirical error probability[4] of the Logit under this configuration for increasing degrees of polynomial expansion without any feature interactions. We use the linear (non-expanded) Logit as baseline for comparison.

[1] more about that in this report *Learning to discover: the Higgs boson machine learning challenge*

[2] randomly balancing the data set - having as many positives as negatives - preserves the by label distribution of the features making them comparable

[3] All values smaller than 1.25 were mapped to 0, whereas the rest was mapped to 1, because the by label distribution was overlapping below that value and showed dominance of HB+ above it

[4] 1 - accuracy

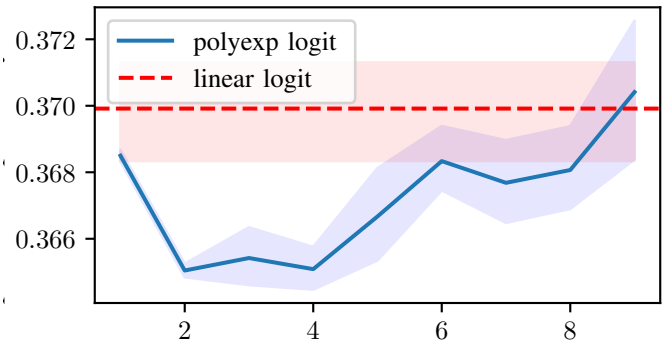

Fig. 2. Logit with polynomial expansion against linear Logit.

We pick therefore polynomial expansion logit with degree 4 without interaction terms as the best model.

On the other hand, the optimal accuracy of this Logit model on AIcrowd is 62%.

2) Logistic Regression Ensemble (LRE): LRE is an ensemble of an odd number of Logit models trained on disjoint partitions of the training set. At prediction time, the most frequent prediction between these models is selected. The idea is to have multiple models that best represent fractions of the data and with little correlation. The Logits are copies of the optimal Logit found in the previous step. Figure (3) shows the cross validation empirical error probability for an increasing number of underlying Logit models.

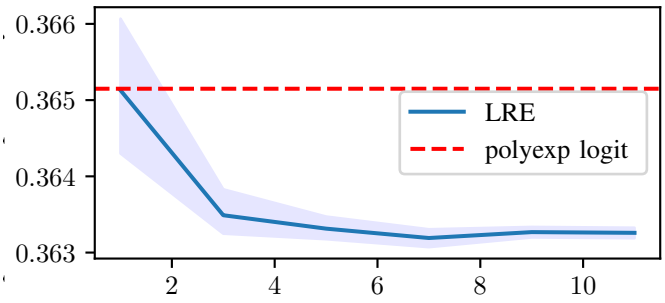

Fig. 3. by label clipped distribution of `DER_mass_MMC`.

Since we expect models at finer granularity to perform better when doing cross validation, we will go with the "elbow" method and pick the optimal number of models to be 5 that does not overfit the data.

On the other hand, the accuracy of this LRE model on AIcrowd is 61%.

3) Logistic Regression Ensemble with Prior Likelihood (LREPL): LREPL is an ensemble of two LRE models. The first is trained on the training data and used to compute the "prior" likelihoods of HB+ that are used to augment the training data. The second LRE model is trained on the augmented data. Prediction is done by predicting the priors with the first LRE, using them to augment the samples, then predicting the class with the second LRE. The idea behind this combination is to push the model towards compensating for outliers around the decision boundary rather than computing the likelihood from scratch. Both LREs are copies of the optimal LRE found previously.

On the other hand, the accuracy of this LREPL model on AIcrowd is 61%.

III. Results

The table below summarizes the accuracies on AIcrowd we obtained with the three previous models on clipped data, which our models had a better performance on compared to unclipped data. Additional baseline models are added for comparison purposes.

Model	Clipped Data
Linear Regression	70%
PolyExp Logit	62%
Logit Ensemble (LRE)	61%
Logit Ensemble with Prior Likelihood (LREPL)	61%

From this table, we see that logistic regression based models did not perform well compared to linear regression. We suspect that our polynomial expansion was not good enough, and that interaction terms of higher degrees may be necessary. In fact, small degree interaction terms only worsen the performance by around 10%.

IV. Summary

After doing the data cleaning, feature selection and feature engineering we have obtained a better dataset to train our models on. Then we have tried different classifiers with linear regression, logistic regression, logit ensemble and logit ensemble with prior likelihood. We have also optimised hyperparameters like degree expansion this way we have obtain an accuracy of 70% with linear regression in predicting the presence (HB+) or absence (HB-) of a Higgs Boson.

V. Conclusion

Linear regression give us the best results with 70% . This is unexpected and one reason might be that we need to try higher order polynomial expansions with complex interactions which we do not have the resources for.