# Project Report: Exploiting Higgs Boson Particle

Fereshte Mozafari, Mohammad Tohidi Vahdat, Ehsan Mohammadpour
fereshte.mozafari@epfl.ch, mohammad.vahdat@epfl.ch, ehsan.mohammadpour@epfl.ch,
*EPFL, Switzerland*

***Abstract*—One of the biggest challenges in studying the properties of the Higgs Boson is to extract its signal from the background noise. In this paper, we apply regularized logistic regression to predict the presence of Higgs Boson in the experiments from the Large Hadron Collider at CERN. The proposed model in this paper utilizes the feature engineering techniques such as polynomial expansion and trigonometric functions. The results shows an improvement of $13.89\%$ in the classification accuracy. Finally, submitting our results in aicrowd.com shows $83.7\%$ classification accuracy and $0.734$ F1-score.**

## I. Introduction

The Higgs Boson is an elementary particle in the Standard Model of physics, which explains why other particles have mass. It is a subatomic particle that uses the data from the Large Hadron Collidar at CERN [1], physicists can discover more about this particle. However, the deterministic analysis of data is intensely hard because of the huge volume of the data and the complexity of the physical processes. Since many decay signatures look similar, we can estimate the likelihood that the signature of a given event is the result of a Higgs Boson, namely *signal*, or some other process/particle, namely *background noise*. In practice, this means that there is a matrix of features representing the decay signature of a collision event, and the goal is to predict whether this event is a signal or a background. Hence, we propose a Machine Learning approach for binary classification.

The rest of the paper is organized as follows. In Section II, we explain the data preprocessing in order to remove meaningless data. In Section III, the model selection, the cross validation phase, and feature engineering are discussed. Section IV shows the results of the selected model. Finally we conclude the paper in Section V.

## II. Data preprocessing

This project contains a train dataset consist of $250,000$ sample pairs of features and labels. For each sample, $30$ features have been set to be measured corresponding to a specific label, i.e., signal or background noise. In the primary observation, we noticed that there are some meaningless values equal to $-999$ for some features that is a disguise for *'not available' (NA)*.

Referring to Higgs Boson notes[1], the origins of unavailabilities are two features: number of jets and estimated mass, respectively denoted as 'PRI_jet_num' and 'DER_mass_MMC' in the provided dataset. The former can take a value between 0 to 3; for each number of jets there are some undefined features. The latter feature may or may not be available for some specific cases. This data observation leads to eight sub-datasets: different jet numbers (0 to 3) with or without estimated mass feature. As a result, there is no *NA* value in the created sub-datasets. In the following sections, we train the models after normalizing the sub-datasets and then do the prediction on the test data for each of these sub-datasets independently.

## III. Model training

In this section, first we describe how to select an appropriate model for the test data prediction. Then, we discuss about the cross validation among the training data and the extraction of hyper-parameters, i.e., the regularization parameter ($\lambda$) and the appropriate polynomial degree, to prevent under-fitting and over-fitting of the selected model .

### A. Model selection

To find out the model that provides the best prediction, we have tested six different methods on the original dataset. The testing procedure was done in this order: we split the train dataset into two local train and test subsets; then we train the data on the local train subset using each model and check its accuracy. The size of the local train subset is 0.8 of the size of the provided train dataset and the rest is kept as local test subset. The results of this analysis is shown in Table I.

Table I
Comparison of accuracy over six different method of data training with maximum $500$ iterations.

| Model | Accuracy | $\gamma$ | $\lambda$ |
|---|---|---|---|
| Gradient decent | $71.62\%$ | 0.1 | - |
| Stochastic gradient decent | $56.44\%$ | 0.001 | - |
| Least square | $71.72\%$ | - | - |
| Ridge regression | $71.72\%$ | - | $1.0e{-}5$ |
| Logistic regression | $72.39\%$ | 0.5 | - |
| Regularized logistic regression | $72.39\%$ | 0.5 | $1.0e{-}5$ |

Table I shows that the regularized logistic regression provides the best accuracy among the rest. This also could theoretically be confirmed as the intrinsic of the prediction of a kind of binary classification and the (regularized) logistic regression is a perfect fit for such types of problems.

[1] http://opendata.cern.ch/record/329/files/atlas-higgs-challenge-2014.pdf

Table II
EXPERIMENTAL RESULTS OF USING REGULARIZED LOGISTIC REGRESSION WITH WITH/WITHOUT FEATURE AUGMENTATION FOR EACH SUB-DATASETS.

| # of jets | Est. mass | Test size | Original | Polynomial | (F1) | (F1)+(F2) | (F1)+(F3)+(F4) | (F1)+(F2)+(F3)+(F4) | **Final** | Imp. |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | *NA* | 59263 | 67.9% | 95% | 94.9% | 95.1% | 95.2% | 95.3% | **95.3**% | 27.4% |
| 0 | ✓ | 168195 | 73.1% | 80.5% | 80.6% | 81.2% | 81.5% | 81.5% | **81.5**% | 8.4% |
| 1 | *NA* | 17243 | 69.8% | 92% | 92.5% | 92.7% | 93% | 93.1% | **93.2**% | 23.4% |
| 1 | ✓ | 158095 | 67.3% | 78.5% | 79% | 79.8% | 80% | 80.4% | **80.4**% | 13.1% |
| 2 | *NA* | 6743 | 82.9% | 91.6% | 94.1% | 94.3% | 95.5% | 96% | **97.7**% | 14.8% |
| 2 | ✓ | 107905 | 72.8% | 82.1% | 83.4% | 84% | 84.2% | 84.5% | **84.5**% | 11.7% |
| 3 | *NA* | 3239 | 64.4% | 94.1% | 96.6% | 97.5% | 98.6% | 99.1% | **99.9**% | 35.5% |
| 3 | ✓ | 47555 | 65.4% | 81.9% | 81.8% | 84% | 83.8% | 84.6% | **84.6**% | 19.2% |
| Expected Accuracy | | - | 70.21% | 81.86% | 82.9% | 83.6% | 83.83% | 84.05% | **84.1**% | 13.89% |

### B. Cross validation

A common method to enrich the model is adding polynomial expansion of the features, i.e., addition of $X^n$ terms where $X$ is the original feature matrix. This allows the classifier to model nonlinear relationships among the features. However, it is necessary to find a suitable value of $n$, as if $n$ is a small integer, it causes under-fitting and if $n$ is a large one, it causes over-fitting of the model. Another parameter to control over-fitting and under-fitting is the regularization parameter $\lambda$. The technique to find appropriate value of $n$ and $\lambda$ is cross validation. In this work, we have done 4-folds cross validation to tune these parameters, denoted as hyper-parameters.

### C. Manual feature augmentation

The improvement of the results due to the polynomial expansion of the features encouraged us to utilize other functions that mirror the effect of each feature on the considered model. We have tested several functions to obtain better prediction, e.g. for any feature $x_i$, we added $\sqrt{x_i}$, $\sin\{x_i\}$, $\tan\{x_i\}$, $\log\{x_i\}$, and $\arctan\{x_i\}$. However, neither of them led to any improvement. Then, we focused on the addition of functions based on the *cross terms* $x_i x_j$ for the features $x_i$ and $x_j$. The following is the list of the added features that results in an improvement in the prediction:

(F1) $CT$: cross term of two features $x_i$ and $x_j$, i.e., $x_i x_j$.
(F2) $\sqrt{CT}$: root square of the cross term.
(F3) $\arctan(CT)$: arctangent of the cross term.
(F4) $\sin(CT)$: sine of the cross term.
(F5) $CT^2$: square of the cross term

## IV. RESULTS & DISCUSSION

Table II shows the incremental results of this project using regularized logistic regression with cross validation. The execution of cross validation for the 8 sub-datasets results in finding $n = 2$ and $\lambda = 10^{-20}$ for each one.

Due to space limitation, we did not put some intermediate results and tried to show the effect of manual feature augmentation described in Subsection III-C. The last row shows the expected accuracy of the model over the test data set by calculation of weighted average based on the size of each data-subset.

Our observation shows that adding each of the functions (F1) to (F4), beside the polynomial expansion, leads to an improvement in the predictions. However, addition of (F5) just improves the prediction in two sub-datasets with number of jets equal to 2 and 3 and the estimated mass is *NA*. The column *Final* in Table II shows the final results that utilize (F1) to (F4) for all sub-datasets and also (F5) for the aforementioned sub-datasets. The *Imp.* column shows the amount of improvement in comparison with the original features (no polynomial and manual feature augmentation). We submitted the final predictions in the aicrowd.com with the team *Panda_feat_Ah*, and obtained 83.7% accuracy and F1-score of 0.754.

We have to mention that the same procedure is done using 10-fold ridge regression. Although the local predictions show high accuracy, the result of the prediction over the test dataset was by far worse than the one with regularized logistic regression.

## V. CONCLUSION

In this project, we applied regularized logistic regression to a binary classification problem in order to predict whether a particle is a signal or a background noise, based on some measured features. We also showed that data preprocessing helps to improve the results of a machine learning model by removing meaningless data. We figured out the significant effect of polynomial feature expansion and the cross validation. It is noteworthy to mention that manual addition of some functions to the model substantially increased the model accuracy. To wrap up, the polynomial expansion as well as the feature augmentation led to an average improvement of 13.89% in the local final prediction.

## REFERENCES

[1] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. A. Khalek, A. A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins *et al.*, "Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc," *Physics Letters B*, vol. 716, no. 1, pp. 1–29, 2012.