Finding Higgs Boson : Machine Learning

Luis Da Silva, Richie Yat Tsai Wan, Gaëlle Wavre
CS-433 - Ecole Polytechnique Fédérale de Lausanne
October 26, 2020

***Abstract*—This project aims to predict whether a collision at the CERN's Large Hadron Collider (LHC) produces a Higgs boson or not based on decay events. This is done by implementing several machine learning and preprocessing algorithms to 250'000 pre-analysed data. Different methods were compared, such as Ridge Regresion, Regularized Logistic Regression, standardization, and logarithmic transformations. The most efficient model was concluded to be Ridge Regression, achieving a 81.7% accuracy.**

I. Introduction

In March 2013, the Higgs boson, a crucial missing piece in the Standard Model of physics, was discovered at the CERN Large Hadron Collider. The Higgs boson is not directly observable due to its short lifespan, however, particles resulting from the Higgs boson decay can be detected. This discovery was a major step for the scientific community to understand the mass generation mechanism. The aim of the project is to build a model on data to predict, according to its decay signature of the collision, whether or not a Higgs boson was produced in the collision. By first taking a look at the data, followed by feature processing, different learning models were applied to obtain the highest accuracy possible.

II. Methods

The train set is composed of N = 250'000 samples with D = 30 different features each describing different collision decay parameters. Each sample was labelled by the CERN community as 's' for boson signal or 'b' for background noise.

A. Feature Processing

The initial step consisted of doing an exploratory analysis of the data in order to gain a insights on the samples and different features. Then different feature processing steps were made.

1) Separation into clusters: One of the features PRI_JET_NUM, which describes the number of jets in the particle decay, consisted of 4 discrete values. This allowed to separate the data into corresponding clusters. By plotting the distribution of each feature of the 4 clusters different tendencies were observed. Correlation analysis was performed to give a first insight of the clustered data. The results, which are shown on the correlation matrix in figure 1, indicate that some features in specific clusters were all at a meaningless value of -999 or 0. This led to the conclusion that it is possible to drop some features, therefore reducing the complexity of the algorithm, without loosing important information for the predictions.

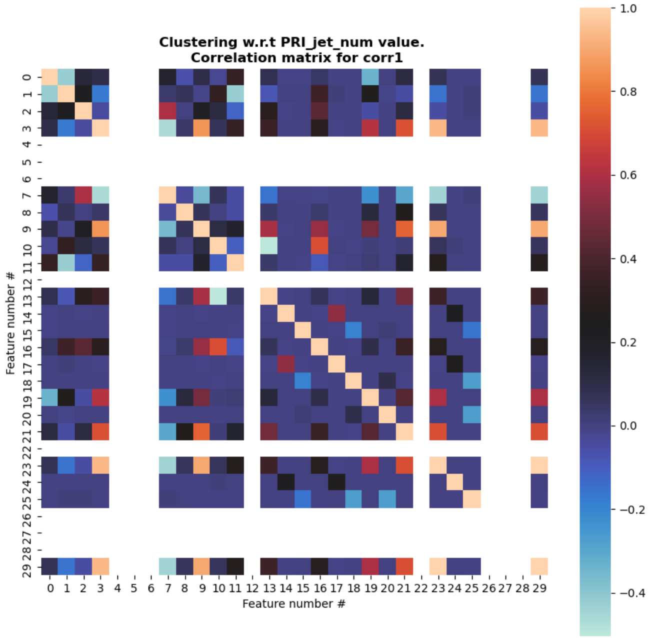

Fig. 1. Heat map of the correlation between features of the dataset cluster n°1. Notice the missing lines which correspond exactly to the features that will be deleted in the preprocessing part due to their missing value (-999 or 0)

2) Non-valid values: After splitting the data in clusters, some of the data nevertheless remained invalid. To reduce the impact of these, their values were replaced by the mean or the median of the other values of the feature.

3) Standardizing: It was observed that there was a large variance in the dataset. An optimization trick was implemented to increase accuracy of the algorithms: standardizing the data to a $\mathcal{N}(0,1)$ distribution.

4) Long tailed distribution features: By plotting the distribution of each feature of the 4 clusters, different tendencies were obtained. Some of the features exposed long-tailed overlapping distributions, making it difficult to distinguish them (cf. figure 2). To reduce the variance of the model, and therefore get rid of the long tail, a logarithmic transformation was applied. In figure 3, it can be seen that this results into a bell-shaped distribution which is less spread (the range of the x-axis is much smaller). However, the separated feature groups are still not distinguishable. To solve this, standardization can be done (cf. figure 4), although in this case, there still is a long tail distribution corresponding to noise. It was then concluded that the best preprocessing results were obtained when combining both standardization and logarithmic transformations as shown in figure 5. Indeed, we can see that in this case, there is a good separation of the clusters and very low variance of the data. This process was applied on the features exposed in table 1.

PRI_jet_num Cluster	Selected features
C0	1, 2, 3, 7, 8, 9, 10, 13, 16, 19, 21
C1	1, 2, 3, 7, 8, 9, 10, 13, 16, 19, 21, 23, 29
C1 & C2	1, 2, 3, 5, 7, 8, 9, 10, 13, 16, 19, 21, 23, 26, 29

TABLE I
FEATURE SELECTION FOR THE DIFFERENT CLUSTERS

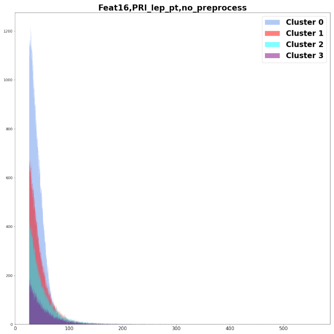

Fig. 2. Distribution for feature n°16 with no processing

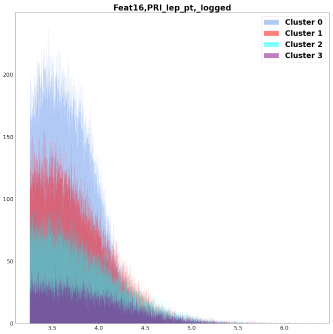

Fig. 3. Distribution for feature n°16 with logarithmic transformation

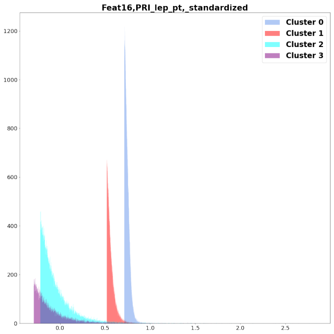

Fig. 4. Distribution for feature n°16 standardized

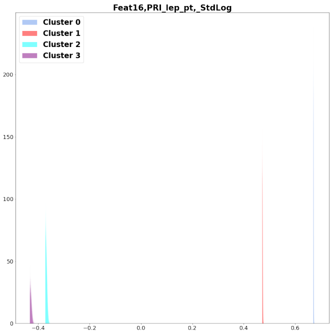

Fig. 5. Distribution for feature n°16 standardized and with logarithmic transformation

B. Learning models

To achieve the greatest accuracy of the predictions possible, different models were implemented and compared. The methods of Least Squares, Gradient Descent and Stochastic Gradient Descent were used. However, the main results of the project were obtained using Ridge Regression and Regularized Logistic Regression (classification model). In the first case, a gridsearch was done to find the optimal hyperparameters L2 regularizer λ (9 values in a logspace from 10^{-12} to 10^{-3}) as well as the polynomial expansion degree (from 1 to 15). Bias was also added for this model. For classification, a Regularized Logistic Regression was implemented, and the regularizer λ and learning rate γ were optimized.

In the two models, 5-fold cross-validation was implemented for testing instead of the standard splitting into train and test datasets.

III. RESULTS

By using the preprocessing methods discussed in section II-A, the accuracy of the predictions was improved. Comparisons were also done between Ridge Regression and Regularized Logistic Regression. The resulting predictions and F-1 scores are exposed in table II.

Method	Prediction (%)	F-1 Score (%)
Ridge Regression	81.7	71.5
Regularized Log Regression	70.0	31.6

TABLE II
ACCURACY RESULTS FOR EACH MODEL

Cluster	Polynomial degree	L2 regularization parameter
0	14	$1e-12$
1	4	$1e-12$
2	14	$1e-12$
3	13	$1e-12$

TABLE III
OPTIMAL HYPER-PARAMETERS FOR EACH CLUSTER

IV. DISCUSSION

The best model was obtained through Ridge Regression algorithm with an accuracy of 81.7%. This method is appropriate to eliminate the problem of multicollinearity, which generally occurs in linear regression problems, as it adds a penalty factor. Unexpectedly, Regularized Logistic Regression did not give satisfying results, with an extremely poor F-1 score of 31.6%, as well as a lower prediction (70.0%). The other implemented algorithms are not suited for categorical data, as despite having a good loss function, they are focused on the loss function minimization and return a poor logistic regression model. As expected, the preprocessing features methods were the key to greatly enhancing the accuracy. By using the log-transformed features, which exhibited initially long-tailed distributions, the variance of the model was reduced. To decrease the wide range of raw values, standardization of the features was a necessary step of our modelization. Moreover, a large part of the improvement in the results came from the clustering of the features according to their number of jets. Indeed, this greatly influenced the model accuracy and allowed to reduce the complexity by removing the meaningless features according to the correlation matrix in figure 1. It was observed that slightly higher accuracies were obtained by replacing the non-valid values by the median rather than the mean. This option was therefore judged the most efficient for the processing. On the contrary, replacing the small number of outliers did not significantly alter the results.

V. CONCLUSION

To conclude, with some feature processing, models such as Ridge Regression and Regularized Logistic Regression able to achieve decent prediction results. Further improvements could be made, such as a more in-depth analysis of the correlation between features, or adding physical meanings to the values.