CS-433 Project 1

Abdullah Aydemir, Ernesto Bocini, Elif Kurtay

***Abstract*—In this report, we present our analysis of the Higgs Boson Machine Learning Challenge, aiming to improve the procedure that produces the selection region. Along explanatory data analysis, feature processing methods and several models have been implemented to improve and compare models. The developed model has reached categorical accuracy score of 0.826 and F1 score (model scoring system based on precision and recall) of 0.737.**

I. Introduction[1]

The Higgs boson is an elementary particle which explains why other particles have mass, and it is observed by its "decay signature" among events. Observing the decay signature is particularly challenging due to large background signals of uninteresting events and already know events. The study presented tries to address this challenge by developing a Machine Learning (ML) model that can successfully detect the signal region by separating the background and the signal event.

II. Models and Methods

In pursuance of a good accuracy in predicting whether a given vector of features is representing signal (Higgs boson) or background, we decided to split the challenge into the following steps: *A. Exploratory Data Analysis*, *B. Feature Processing* and *C. Model Selection*.

A. Exploratory Data Analysis

The training set consists of 250'000 events. The file starts with the ID column, then the label column (-1 for "background" and 1 for "signal"), and 30 feature columns. The test set consists of 568'238 observations with the same features of the training set.

- *PRI_jet_num*: All variables are continuous, except *PRI_jet_num* which is categorical. From the challenge documentation, we realized that some features behave different for different number of Jets, for instance *DER deltaeta jet jet*, *DER mass jet jet*, etc. Thus, we decided to split our data into four subsets having *PRI Jet num* respectively equal to 0, 1, 2, 3.
- *Missing Values*: The data set has many missing values (depicted as -999.0). Missing values were dealt with the following approach: First, features with more then 70% missing values were eliminated. Then, features left with missing values were imputed with their median, which is a more robust estimator compared to the mean.
- *Outliers*: In order to deal with the outliers, we decided to perform a quantile study, feature by feature. For each variable, we substituted all the values below and above a certain α-percentile with that threshold. The best value for α depends on the number of Jets and on the learning method implemented, therefore cross validation was used to find it.

[1]This section has been delivered by referring to the challenge documentation from Claire Adam-Bourdarios *et al.* (2014) [1]

B. Feature Processing

- *Feature Distribution Approach*: Empirical distribution of each feature was plotted by separating the two labels. As Jason Brownlee([2]) suggests, these plots are very useful in feature engineering, since they give insights on usefulness, skewness and symmetry of features. With this in mind, we decided to remove the following variables, that didn't provide enough differentiation among the dependent labels: 15,16,18,19,20,21, see Fig. 1 for an example.
  The Feature Distribution Approach has given the best accuracy for our models, however the following methods have also been used sepertely and in combinations:
  1) *Multi-Collinearity Check*: Each feature pair has been regressed on each other and mean squared errors (MSE) have been computed using least square. Using MSE, the R squared value, and ultimately the variable inflation factor (VIF) has been calculated. One feature per feature pair has been dropped for VIF values higher than 4.5.
  2) *Correlation Check*: Features were regressed on y with logistic regression and the pseudo-R square values were calculated using the Tjur approach[3]. Features with very low correlation with y have been dropped.
  3) *Forward feature selection:* Starting from a single feature that performs the best within the model, the feature set can be enlarged by iterative checks of the combinations of the current feature set and features left, by adding the feature that improves the model the best until no improvement is achieved.
- *Polynomial Expansion*: After feature selection through the feature distribution approach, for each *'class Jet'*, we added an intercept (column of ones) to the feature matrix and we performed polynomial expansion with a *degree* that has been accurately selected through cross validation for each data and each ML implementation. In addition, we have also standardized the data. In this way we were able to maximize the accuracy in prediction.

C. Model Selection

Both linear and logistic models have been implemented within this study to improve the accuracy of the final model.

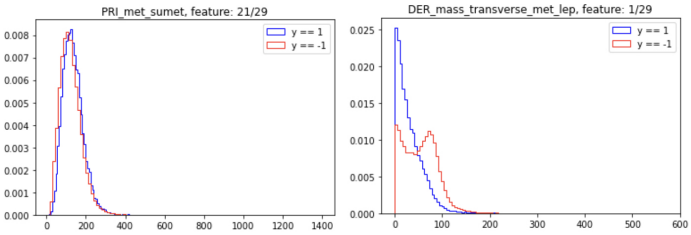

Fig. 1. On the left: feature 21, which doesn't provide enough differentiation among labels. On the right: feature 1, optimal differentiation.

The data set split according to number of Jets and pre-processed as described in the previous sub-section, was used for cross validation to select the best hyper-parameters for each model. These parameters were then used to compute the accuracy for each jet class. More specifically, for linear regression with gradient and stochastic gradient descent, least squares and logistic regression with gradient descent, the α parameters and the $degree$ able to maximize accuracy for each Jet number was searched. While for methods such as ridge regression and regularized logistic an extra hyper-parameter was introduced for cross-validation, the penalization parameter (λ), which is useful in order to avoid over-fitting. Further detail is given in section *§ III Results* and in figure 2.

III. Results

In order to find the optimal parameters for the selected six Machine Learning models, 3-fold cross validation was performed with a grid search on combinations of different parameters. Then, the selected best parameter of each model was compared to other models to find the most suited model for the data set. However, while observing the accuracy result of our models, we received noticeably low results which was due to the raw initial data. Hence, we processed the data accordingly to the procedures defined in the previous section and then focused only on the data that had a strong influence on the accuracy of the models.

Results of our data analysis explained in detail in the section II. Models and Methods, yielded the most valuable information. After cleaning the data by imputing missing values, finding outliers and standardising; the split into different categories of *PRI_jet_num* increased the accuracy score greatly.

In order to choose the best method, we selected best parameters through a 3-fold cross validation for each method, and then computed the accuracy for each jet category. The results for each method can be observed in Figure 2. The accuracy results are divided per *PRI_jet_num* to provide a better understanding. Due to the distribution of data points in the spatial domain, it was expected to get a lower success rate from MSE-based models and logistic regressions which is exactly what is observed. In between least squares method and the ridge regression, although they provided very close results, we preferred to continue with ridge regression as it extracted better test results in the competition's test data. In conclusion, our final optimizations and more detailed feature

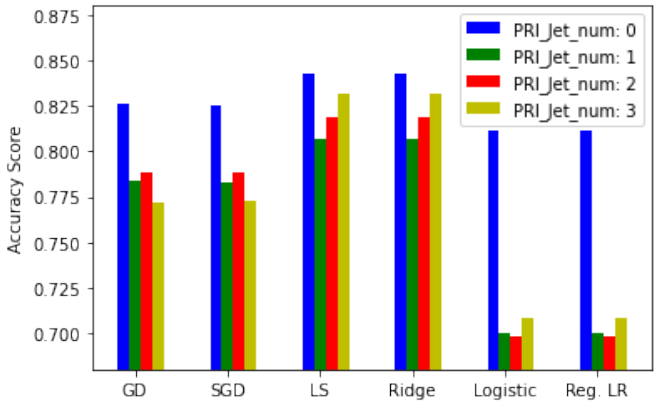

Fig. 2. Accuracy scores of the training data set of all ML methods according to the feature *PRI_jet_num*. "GD": Gradient Descent with MSE, "SGD": Stochastic Gradient Descent with MSE, "LS": Least Squares, "Ridge": Ridge Regression, "Logistic": Logistic Regression, "Reg. LR": Regularized Logistic Regression.

analysis were performed on ridge regression method. Our best result in the competition is obtained by the mentioned feature elimination and data processing in section II and the following best parameters for each jet category [0-3]:

- Degree: [7.0, 7.0, 7.0, 6.0]
- Alpha: [3.0, 3.0, 3.0, 4.0]
- Lambda: [2.5e-05, 1e-06, 2.5e-05, 1e-06]

Using these parameters, we were able to reach an accuracy on the test prediction submitted on AI Crowd of 0.826. Note that this accuracy is by 0.003 smaller than our best one on the competition system. We believe that the loss in the score is caused by cleaning small bugs and code structures.

IV. Discussion

One of the key points that allowed us to succeed in the challenge was creating a model for each different Jet category. In this way we were able to make accurate prediction for each subset using suited parameters for each different class, greatly increasing the final accuracy. In addition to this, the choice of a Ridge Regression based model was best suited for analyzing the features, and this is mainly due to its ability to offset over-fitting via the cross-validated penalization parameter.

V. Summary

The project highlights the importance of the data pre-processing and feature engineer, that, together with cross-validation, let us succeed in the final goal of the challenge.

References

[1] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge (2014). *URL http://higgsml. lal. in2p3. fr/documentation*.

[2] Jason Brownlee. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.

[3] Tue Tjur. Coefficients of determination in logistic regression models–a new proposal: the coefficient of discrimination (2009). *https://doi.org/10.1198/tast.2009.08210*.