Machine Learning Project 1
Higgs Boson

Ahmed Ben Haj Yahia, Nour Ghribi, Kushagra Shah
Department of Computer Science, EPFL, Switzerland

***Abstract*—In this project, we used the concepts seen in the lectures and practiced in the labs on a real-world dataset - least squares regression, ridge regression, logistic regression, regularized logistic regression etc. We used exploratory data analysis to understand the dataset and features, cleaned the dataset using feature processing and engineering in order to extract more meaningful information. We implemented machine learning methods on real data, analyzed the model and generated predictions using those methods. In this report, we present our findings.**

I. Introduction

The discovery of Higgs Boson was acknowledged by the 2013 Nobel prize in physics given to Francois Englert and Peter Higgs. The Higgs Boson is an elementary particle in Particle Physics. It is the final ingredient of the Standard Model of particle physics, ruling subatomic particles and forces. [1] This project aims to distinguish between Higgs Boson and background noise. The data-set has 30 features describing a proton collision event. We had to do some data cleaning and feature engineering to remove the noise and implement machine learning methods to solve this binary classification problem.

II. Models and Methods

A. Basic naive prediction

Before diving into data cleaning and feature engineering, we tried 5 basic machine learning algorithms without any optimization and by choosing the parameters by hand. Table I depicts the result we got for each implementation after doing a random 80-20 split on the data.

Methods	Parameters			Pred (%)
	λ	γ	max_iter.	
Least Squares	-	-	-	71.66
Gradient Descent	-	10^{-7}	3000	59.75
Stochastic Gradient Descent	-	4.10^{-7}	2000	59.52
Ridge Regression	10^{-5}	-	-	71.66
Logistic Regression	-	10^{-6}	2000	71.15

Table I
Initial tests of basic ML algorithms and its prediction.

As expected, the initial methods performed quite poorly, especially the gradient descent methods which performed the worst (just a bit better than random flip coin). We also expected that Logistic Regression will perform the best since it's made to deal with this kind of classification, but instead we found that Least Squares and Ridge Regression performed the best, directly followed by Logistic Regression

B. Data Analysis

Next, We decided to get a better sense of the data in order to develop a better model for this classification task. We analysed the data and following are the main observations that grabbed our attention:

1) The feature 22 labeled PRI_jet_num is the only one that takes integer values and are all in $\{0, 1, 2, 3\}$.
2) 11 out of the 30 features presented the existence of NaN and -999 values.
3) The correlation between certain features is shown in Figure 1
4) We plotted the values of each feature and found:
   a) Features that take unique values.
   b) Features that behave as a 'Gaussian' once we apply certain mathematical functions on them.

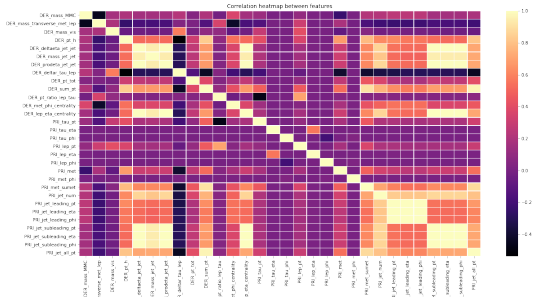

Figure 1. Correlation between different features (yellow means high correlation).

PRI_jet_num represents the number of jet pseudo-particles appearing in the detector. We noticed that the distribution of missing values highly depends on the jet feature, so We decided to partition our Data-set based on the jet_num value leading to 4 different subsets. Once we did that, We noticed that there were entire columns that take the value NaN or a unique value like 0 especially for jet 0 and jet 1 so we omitted them.

After this, we moved on to the feature engineering task where we had to apply some uni-variate transforms to better expose the linear relationship between the inputs and the

output. Here is the list of the applied transformations that worked the best for the different features:

- The Classification by transforming the values that are ranging between two distinct extremes to discrete numbers.
- The Square root.
- Uniform transformation by dividing values of features by their absolute maximum to bring the values between -1 and 1.
- The Logarithm function.
- Replacement of the NaN values by the mean.
- Standardization.

Figure 2 is a perfect example of how one can transform a certain feature.

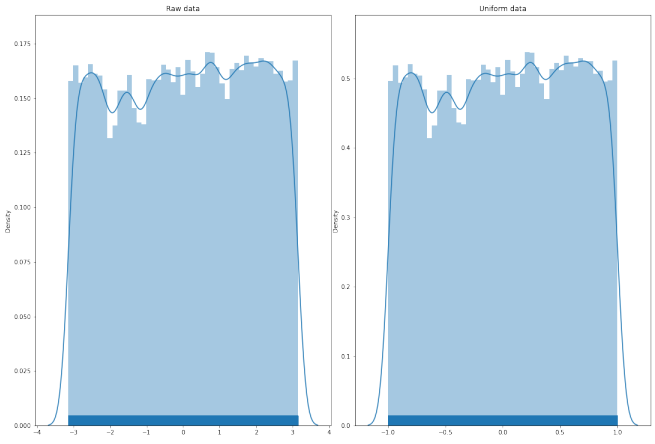

Figure 2. An example of a Uniform feature transformation.

C. Cross validation

Having found some problems with Logistic Regression due to very long training process, we tried to fit the Ridge Regression method. We used cross validation to find the best lambdas and degrees for each subset of the data-set. In Figure 3 you can see the boxplot of RMSEs (Root Mean Square Errors) for different values of lambdas.

D. Feature Expansion

Having found satisfying results, we tried to improve our model further so we tried to expand our features using some other functions: we tried the cosine and sine functions. This operation showed improved results. This is because there is enough features for our algorithm to find a linear relationship between the input and the output.

E. Training

Cross validation returns the degrees and the lambdas that gives the best loss in training without over-fitting. Having these values, we transform our data using all the operations mentioned above and we fit the Ridge Regression model with the best hyper-parameter lambda to find the best values for W.

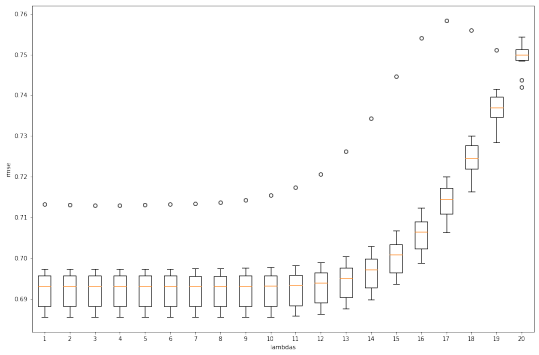

Figure 3. RMSE for different values of lambda for jet 0.

F. Testing

Once we have the best W's we test each one on the different subsets corresponding to their "jet" numbers.

	Subset by jet number				
	0	1	2	3	Overall
Accuracy (%)	84.3	77.7	72.68	80.69	78,84

Table II
RESULTS OF RIDGE REGRESSION.

The table II show the results of our tests done locally by splitting the data into train and test set using a 0.7 split ratio. However we got a score of 83% accuracy on AiCrowd and an F1-Score of 0.739.

III. CONCLUSION

Through this project we implemented the models we have seen in the Machine Learning class [2] and were able to dive more into the problem of prediction and understand the complexity involved in feature engieneering and data preprocessing. [3] We learned how to explore and analyze the data and what adequate transformations to use for each case.

REFERENCES

[1] P. Onyisi, *Higgs boson FAQ*. University of Texas ATLAS group, 2012.

[2] "CS433 epfl, machine learning course," https://mlo.epfl.ch/page-146520-en-html/, accessed: 2018-10-21.

[3] "Machine Learning Mastery logistic regression for machine learning," https://machinelearningmastery.com/logistic-regression-for-machine-learning/, accessed: 2018-10-23.