Finding the Higgs Boson Machine Learning Project

Can Yilmaz Altinigne, Gunes Yurdakul, Bahar Aydemir
can.altinigne@epfl.ch, gunes.yurdakul@epfl.ch, bahar.aydemir@epfl.ch
Department of Computer Science, EPFL, Switzerland

***Abstract*—Discovering evidence of the Higgs boson is a particular game-changer in the field of particle physics. Besides the conventional methods, machine learning techniques are proven to be effective on complex and high-dimensional datasets. In this project, we have developed and compared several models by various machine learning techniques. Our model's objective is to predict whether given features of a collision event is a result of Higgs boson or a background noise.**

I. Introduction

The Higgs particle has a fundamental role in the Standard model by explaining why subatomic particles have mass. On July 2012, the evidence of the Higgs boson announced as a result of the experiments in CERN's Large Hadron Collider. [1] During the experiments, the decay procedure of proton-proton collisions are observed. The collisions may produce Higgs particle, which is not directly observable due to its very short lifespan, or other subatomic particles. The decay process recorded as a ”decay signature” and it is possible to estimate the likelihood of a event was the outcome of a Higgs particle or the other particles. In this project, the prediction of the Higgs boson in a given event treated as a binary classification problem. We have applied several machine learning methods to solve this problem.

II. Models and Methods

A. Dataset

The dataset provided by the ATLAS experiment at CERN consists of simulated data of the collisions. [2] It consists of 250000 events which are described by 30 feature columns. The label is ’s’ (signal) if the Higgs particle is present. Otherwise, the label is ’b’ which implies the background. There are 85667 signal events and 164333 background events in the dataset. Moreover, the incalculable values are given as -999 and 1.5 million values over a total of 7.5 million values, corresponding to 21%, have invalid values.

B. Exploratory Data Analysis

We have observed that the number of meaningless values are much more greater in some columns. 11 of 30 features have meaningless values whereas the remaining 19 features does not have any. Moreover, we have observed that the meaningless values reside in the same rows. Therefore, we have decided to analyze the features in two categories which includes meaningless values and which does not.

We found that there is a relationship between some features regarding meaningless values. For instance, when PRI_jet_num is 0; $DER\_deltaeta\_jet\_jet$, $DER\_mass\_jet\_jet$, $DER\_prodeta\_jet\_jet$, $DER\_lep\_eta\_centrality$, $PRI\_jetleadings$ and $PRI\_jetsubleadings$ are -999. [3] Also when PRI_jet_num is 1, the same features are -999 also except for $PRI\_jetleadings$.

Values for the features above with respect to the value of PRI_jet_num can be seen in Figure 1.

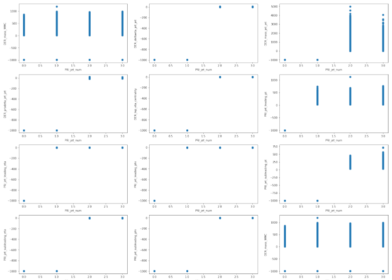

Figure 1. Relationship between features that have meaningless values.

C. Feature Selection and Feature Processing

We used Pearson correlation coefficient for feature selection. We consider the correlation between all features and the label. We have observed that some of the selected features are highly correlated with each other. In such a case, we selected the feature having highest correlation with the label.

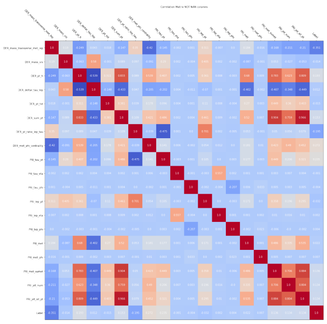

Figure 2. Correlation between non-meaningless features and label).

We choose 14 different features, 7 from features without meaningless values and 6 from features with meaningless values. The last selected feature is PRI_jet_num which will help us to split training set in the prediction phase. The selected features are given in Figure 3.

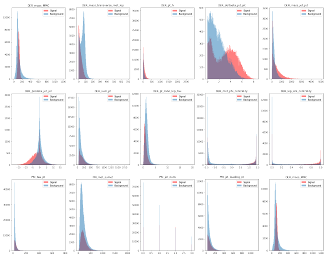

Figure 3. Distribution of selected features

After picking the features for the prediction model, we apply log transform on some features to reduce the skewness. Then, we standardized each feature to have zero mean and unit variance to gain robustness against different ranges and spread of the values. By standardization we prevented the large changes of large values to appear more important. We used the mean and standard deviation values of training set to standardize our testing set as well, since our model is trained using standardized training set. In some models such as Logistic Regression and Linear Regression, we normalize data instead of standardization, since standardized data produces overflow in sigmoid function and error functions. Finally, we impute DER_mass_MMC with its median value, since it has smaller number of meaningless values.

D. Model Comparison and Prediction

We created three separate training sets based on the value of PRI_jet_num feature, which is a categorical feature having 4 categories: 0,1,2 and 3. We use this approach, because we did not want to simply impute a lot of meaningless values, since the features are most likely to be related.

We remove the features that have meaningless values in our three training subsets. Shapes of three subsets are given below.

Subset	
Subset 1	(99913, 8)
Subset 2	(77544, 9)
Subset 3	(72543, 13)

Table I
SUBSET SHAPES

We use Least Squares, Ridge Regression, Logistic Regression and Least Squares with Gradient Descent (Linear Regression) models. We give priority to use Least Squares, since we choose polynomial features that do not cause the rank deficiency problem. We use 5-fold cross validation for Least Squares and Ridge Regression models. For Logistic Regression and Linear Regression, we use 3-fold cross validation, since the computation takes some time.

The most promising results were obtained using Least Squares and Ridge Regression models. We use cross validation to predict the most optimal polynomial degree for Least Squares, the most optimal lambda for Ridge Regression and the most optimal learning rate for Logistic Regression and Linear Regression. We use the same polynomial degrees that we found in Least Squares for Logistic Regression and Linear Regression models, since the cross validation for latter two models takes a lot of time.

We use accuracy to choose the optimal model. Logistic regression uses logarithmic loss and Linear Regression uses mean squared error to calculate gradient and error loss.

III. Results

Optimal hyperparameters and cross validation accuracy for 4 different model is given below with the baseline model accuracy. The validation set accuracies we have obtained for different models are represented in Table - 2. We can see all of the models performed better than the baseline. Moreover, it can be seen that ridge regression performed significantly better than linear regression since it is more robust to overfitting.

Method	Subset 1	Subset 2	Subset 3
Linear Regression (deg, gam)	10, 1e-2	9, 1e-1	10, 1e-1
Least Squares (deg)	10	9	10
Logistic Regression (deg, gam)	10, 1e-4	9, 1e-5	10, 1e-5
Ridge Regression (deg, lam)	12, 1e-14	12, 1e-3	11, 1e-5

Table II

Method	Validation Set Accuracy
Baseline	0.657
Linear Regression GD	0.731
Least Squares	0.810
Logistic Regression	0.751
Ridge Regression	0.813

Table III

We have our highest score in Kaggle with the Ridge Regression model. Accuracy in Kaggle for different models can be seen below.

Method	Test Set Accuracy
Ridge Regression with 3 subsets	0.81476
Least Squares with 3 subsets	0.81179

Table IV

IV. Conclusion

In this project, we have utilized several machine learning algorithms as described above in detail to predict the presence of the Higgs boson in a given collision event. Models produced using Ridge Regression and Least Squares methods have outperformed all the other implemented models. Moreover, using the categorical feature PRI_jet_num to produce different models for each PRI_jet_num category has resulted in a significant increase in accuracy. As future work, we want to use regression methods to impute the missing values rather than imputing with the median of the data.

References

[1] ”The Higgs boson — CERN”, Home.cern, 2018. [Online]. Available: https://home.cern/topics/higgs-boson. [Accessed: 25- Oct- 2018].

[2] C. Adam-Bourdarios, G. Cowan, C. Germain-Renaud, I. Guyon, B. Kgl and D. Rousseau, ”The Higgs Machine Learning Challenge”, Journal of Physics: Conference Series, vol. 664, no. 7, p. 072015, 2015.

[3] ”CERN Open Data Portal”, Opendata.cern.ch, 2018. [Online]. Available: http://opendata.cern.ch/record/328. [Accessed: 26- Oct- 2018].