First, we began with observing the features of data based on the original csv file and we found a large amount of ‘-999’ appeared dispersedly in the data set. Then we found that the column ‘PRI\_jet\_num’ which is consisted only with the integer number from 0 to 3, has a strong connection with the appearance of the value ’-999’. ’-999’ appears when this value is equal to 0 or 1 and disappears when the value is 2 or 3. So, in our first test, we decided to drop all the rows whose ‘PRI\_jet\_num’ is 0. But there are still many error data left as we chose to reserve the data with their ‘PRI\_jet\_num’ equal to 1. We hence plotted the distribution of value for every feature. According to these pictures, we dropped the column which more than half of their data is’-999’. Because they have too little useful data to analyze (ex: DER\_deltaeta\_jet\_jet). We also dropped the features which have uniform distribution and Gaussian distribution for that we couldn’t distinguish two types of labels from their distribution (ex:PRI\_lep\_phi, PRI\_lep\_eta). As for the rest of features, there are some process strong polarization which is quite valuable for the binary classifier. Furthermore, there are some features who peak at one value and have little error data we thus decided to replace ‘-999’ by their median. Then we used python to perform the data processing: first we removed the label ‘s’ and ‘b’ in the original file and replaced them by ‘1’ and ‘-1’ respectively. After that, we trained the logistic regression model with the processed data and using the obtained model to do the prediction on the test data set. The result of the first attempt is 74%.

The second time, we normalized our data using the standard normalization method. This time, we remarked that the compute time loss and the storage loss was been dramatically improved.

The third time we tried to separate our data into 4 small groups based on their value of ‘PRI\_jet\_num’. And for each group, we divided them into two subgroups according to their labels and drew the distribution of the feature. So we obtained, for each feature, 8 pictures. And this time, we tried to study the dependence between two features. We then randomly pick two features and calculate their coefficient of correlation. And we listed all pairs of features who have a coefficient higher than 0.8 which means they have strong association between them. So for each pair, we saved only one feature out of two. And we dropped the other feature. In our case, we left the column’ DER\_mass\_vis’, ’DER\_pt\_h’ and ‘ DER\_sum\_pt’. Furthermore, like the first attempt, we dropped the column which has too many error data and has uniform distribution. But this time, we decided to keep the features that have Gaussian distribution. Then we trained separately the data based on their value of ‘PRI\_jet\_num’ and built four different models using logistic regression. We used these models later to predict the labels of our test data. The final score improved 1% to 75%.☹

So we continued to analyze our features. First we calculated the log of the feature whose value is all positive or calculated the normalized value otherwise. This significantly saved our computed loss and storage loss. After that, we wanted to test the influence of degrees on our final prediction. We used the polynomial feature to process the same features selected during our last test. For each feature, we increased their degrees up to two. And, we randomly picked two features among those who are not strong associated between each other and mutilated them. Then we used the obtained model to do the prediction. This time, the accuracy has a significant improvement to 82%.