Particle Identification

Adam Erler

The Purpose of this project is to test if we can use the values of the detectors to identify particles that pass through them. The overall outcome is that we absolutely can achieve this. The clearest indicator of this is the hypothesis test of our outcome. There was a lot of work to get to that moment so I will outline the steps taken and how I have interpreted the information.

The data had a Label feature with 6 categories. To do EDA on a set of strings can become a problem though so I created 6 dummy variables with 1 or 0 to be indicators of the particles identification. Then I ran each detector's measurements against the set of each dummy variable, all in all 24 tests. Each was correlative with a p value much less than our goal of 0.01. In all instances we accepted the null hypothesis. The tables and their generating code can be found in the slides and the code.

By no means is this analysis complete though. This was a primary set of EDA to see if we were pursuing a situation that was worth our time. There are literally 46 other features to consider and help create the model. While some may be logically removed we need to consider the impact of all of the data. For example TrackP, particle momentum could have a huge impact on identifying particles since each of these have very different masses and the momentum could be impacted by this. This leads me to question my assumption of the first 4 that I chose for my EDA. These may not be the best indicators and I need to explore the data and do more background research to understand their context fully.

There were many challenges in this project, one being trying to analyze string data. This was solved with dummy variables. Another situation that was hard to understand and clarify was the PMF. These values have several thousand unique variables and finding bins to represent them well was time consuming. I need to better understand PDF creation and matching to better analyze this data that is at some points basically continuous.

To summarize though we achieved success in addressing our hypothesis that these variables can be used to model the particle labels. This needs us to take the next step to EDA the rest of the data and begin to create a machine learning model to help the computer better create a model to predict labels.