A Machine Learning Approach to Predict the Invariant Mass of Dielectrons

Deepesh Narayan Raul, SK Somaiya

**Abstract**

The elementary particles of the universe and its interactions is said to be concise by the Standard Model of high energy physics or particle physics. Particle collider experiments has the capability to produce enormous and information-rich samples of data. Machine learning techniques can be used to develop how these data samples are interpreted, greatly expanding the discovery of potential present and future experiments. In this paper, data provided by Compact Muon Solenoid built on the Large Hadron Collider at CERN is analyzed to predict the invariant mass of two electrons M using a statistical approach of machine learning.

***Keywords:*** *Machine Learning, CERN, Invariant Mass, Electrons, Regression, Feature Engineering.*

**Introduction**

Mid-1970s was when the Standard Model of particle physics was finalized upon the confirmation of quarks, since then, the evidence of top quark in 1995, tau neutrino in 2000 and the recent Higgs boson in 2012 have solidified the credence of the Standard Model. This theory describes that every observable objects within this universe is made from basic blocks called elementary particles, ruled by the four forces. The invariant mass also known as rest mass or intrinsic mass is the fraction of the total mass of an object or system of objects that is independent of the total motion of the system. A collider experiment is used in particle physics research by colliding pair of particles at very high kinetic energy. The Conseil Europ ́een pour la Recherche Nucl ́eaire or as we call it, CERN houses the world’s largest and highest energy particle collider, the Large Hadron Collider (LHS) and the Compact Muon Solenoid (CMS) a particle physics detectors. The CMS can generate huge amount of data for particle collisions at 0.9 – 13 TeV. Using the Machine Learning we are able to leverage our computing powers alongside modern algorithms to quickly and efficiently observe and analyze insights from enormous amount of data. In this paper, we are not only trying to demonstrate the statistical significance of machine learning models in the field of particle physics by predicting the invarient mass of dielectrons based on the observation from the CMS detectors but also manipulating the features of the dataset by creating new features such that these features increase the performance of the models.

**Data Collection and Processing**

The dataset used for this research is provided courtesy of the CERN open data portal. “Events with 2 electrons from 2010”, McCauley, Thomas, this dataset contains observations of 100K dielectrons events in the invariant mass of 2-110 GeV captured by the Compact Muon Solenoid. This data is organized in a CSV spreadsheet file and include the following observations collected by the CMS:

* **Run:** The run number of the event.
* **Event:** Number of each event
* **E1 and E2:** Total energy of the 2 electrons in GeV
* **px1, py1, pz1, px2, py2, and pz2:** Components of the momentum of the electrons in GeV
* **pt1 and pt2:** Transverse momentum of the electrons in GeV
* **phi1 and phi2:** phi angle of the electrons 1 and 2 in rad
* **eta1 and eta2:** The 2 electrons pseudo-rapidity
* **Q1 and Q2:** The charge of the electrons
* **Invariant Mass M**: The invariant mass of the dielectrons in GeV

As this dataset comes directly from the CERN open portal, it ensured that the observations are reliable, accurate and has been peer reviewed to be scientifically correct. The data released have been thoroughly analyzed and verified its accuracy through simulation events. Any results provided by this paper can be ensured to be true and accurate.

The dataset is further being processed by examining, cleaning, and analyzing the data and its features. First steps taken in data processing is the removal of duplicate data. As duplicate data are an extreme case of nonrandom sampling, as well as they bias any of the fitted models, leading to overfitting problems. In the cases for the CERNs dataset, these duplicates are not real data nor is intentionally oversampled. After analyses of the dataset, the target variable M i.e., the Invariant Mass have no values. As this research deals in predicting the values, the records are removed from the dataset instead of imputing the data which may lead to false results.

**Feature Engineering**

This section provides the explanation of what manipulation of the features of the data was done and show the analysis of these new features. In statistics, and thereby Data Science, correlation analysis is done to calculate the level of relation between one variable to another. In other words, it measures the linear association between 2 variables.

* Where, is the Pearson correlation coefficient between 2 features, and
* is the mean of feature
* is the mean of feature

The values of the correlation coefficient are limited to between +1 and -1. If the coefficient value is close to +1, the 2 variables are perfectly positive interrelated; that is if one variable increase positively, the other also increases. On the other hand, a coefficient value that tends close towards -1, the 2 variables are perfectly negative interrelated: if one variable increases, the other decreases perfectly in the opposite direction. With a coefficient value near 0 there isn’t any interrelation. For the research purpose we are looking for correlation coefficient that tends towards +1 and -1 between each independent features and the target M.

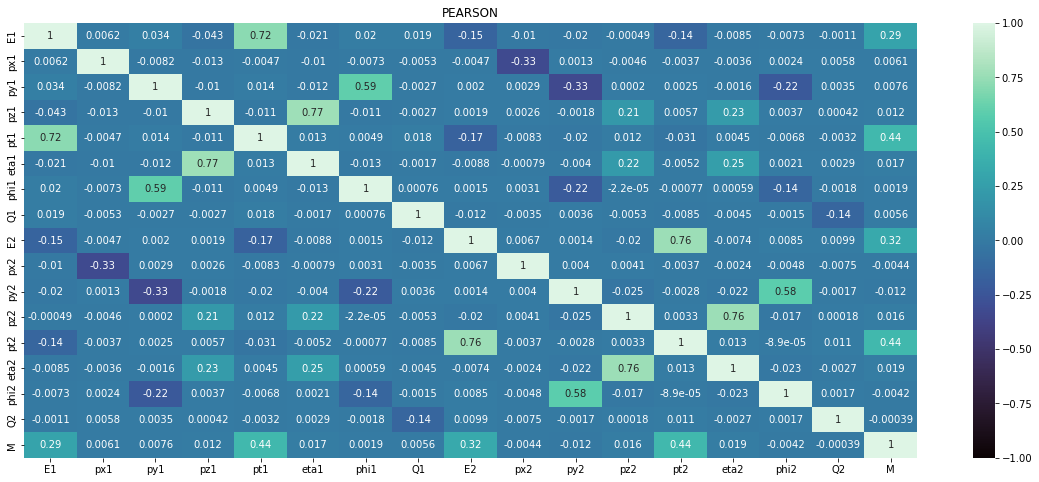


Figure 1 Heatmap for the correlation coefficients

Figure 1 is the heatmap for the Pearson’s correlation coefficient values of the dataset shows us the values between each feature. The major focus here is the coefficient value between the target value M and each independent variable. The heatmap shows that E1, E2, pt1 and pt2 features have high correlation values 0.29, 0.32, 0.44 and 0.44 respectively, with Invariant Mass compared to all the other features. With correlation comes the problem of multicollinearity, where 2 or more independent variables are highly correlated with each other. This problem is detected using variance inflation factor (VIF), it is the measure of how much the standard error of the estimate of the coefficient increased due to multicollinearity.

|  |  |
| --- | --- |
| Features | VIF Value |
| E1 | 2.160088 |
| px1 | 1.120087 |
| py1 | 1.661406 |
| pz1 | 2.486832 |
| pt1 | 2.553705 |