# Executive Summary: Predicting Invariant Mass

Richard M. Flores[1]

**Abstract**
The field of Data Science has evolved from statistical calculations to complex algorithms that grant insight into a variety of natural phenomena. Data Science can be accessed to provide information on everything from predicting the future price of corn to the classification of stars in distant galaxies. One application in which Data Science can be utilized is in the analysis of the Petabytes of information generated by the CERN Particle Accelerators in Geneva, Switzerland. In this research report we will analyze the data provided by the Compact Muon Solenoid as the particle accelerator attempts to trigger, reconstruct, and identify events with final state electrons (McCauley 2010).

**Keywords**
CERN — Particle Physics — Invariant Mass

[1] *Master of Science, Data Analytics, Western Governors University, United States*
*__Contact Author__: rflo147@wgu.edu

## Contents

Introduction 1
1 Problem and Hypothesis 1
2 Data Analysis Process 1
3 Outline of Findings 2
4 Limitations of Techniques and Tools 2
5 Proposed Actions 3
6 Expected Benefits of the Study 3
7 References 3

## Introduction

This Executive Summary has been separated into 6 subsections which correlate to the sections listed under the grading rubric for Task 3: Presentation of Findings

## 1. Problem and Hypothesis

This research project seeks to answer if it is possible to predict the mass of an electron by analyzing dielectron events in the mass range 2-110 GeV. Invariant mass is characteristic of a system's total energy and momentum equal in all frames of reference by Lorentz transformations which is defined as a six-parameter group of linear transformations from a coordinate frame to another frame moving at constant velocity relative to the former (Ivanov 2016). The invariant mass is equal to its total mass in the 'rest frame' which is the subject of our project (Wikimedia 2022). We will also explore the idea if a machine learning algorithm will prove to be statistically significant in predicting mass in subatomic particle collisions.

The proposed hypothesis of this research paper is that the observed components of momentum in electron collisions are statistically significant in determining the mass of subatomic particles. A successful outcome from the hypothesis will provide meaningful insight in the application of machine learning in particle physics research and also translates into similar studies in fields like biology, chemistry, and healthcare.

## 2. Data Analysis Process

For this research project, we employed two major analysis techniques for the observation and calculation of the CERN particle physics dataset. The first analysis technique involved visualization of the dataset to gain insights from the raw data. Through the course of exploratory analysis, we created Pearson Correlation and Covariance Heatmaps which demonstrate how two variables are linearly correlated, and the possibility of a direct relationship (DataCamp 2022). We then viewed a histogram visualization where we are able to observe interesting similarities between variables. Further on we visualize Scatterplots between meaningful relationships such as E1 and M. Next we created visualizations of Histplots where we examine the range of variables such as E2, eta2. Then we completed Jointplot visualizations comparing relationships between variables such as E1 and the p-vector. Our next visualizations include Distplots representing the mean of a selected variable as well as the Lesser and Upper bounds which we use as another form of range examination for variables such as E1 (DataCamp 2022). Finally, we examined relationships through the use of 3D plots to examine data along the x, y, and z axis' which allowed us to explore relationships amongst three variables as opposed to the two variables in previous visualizations.

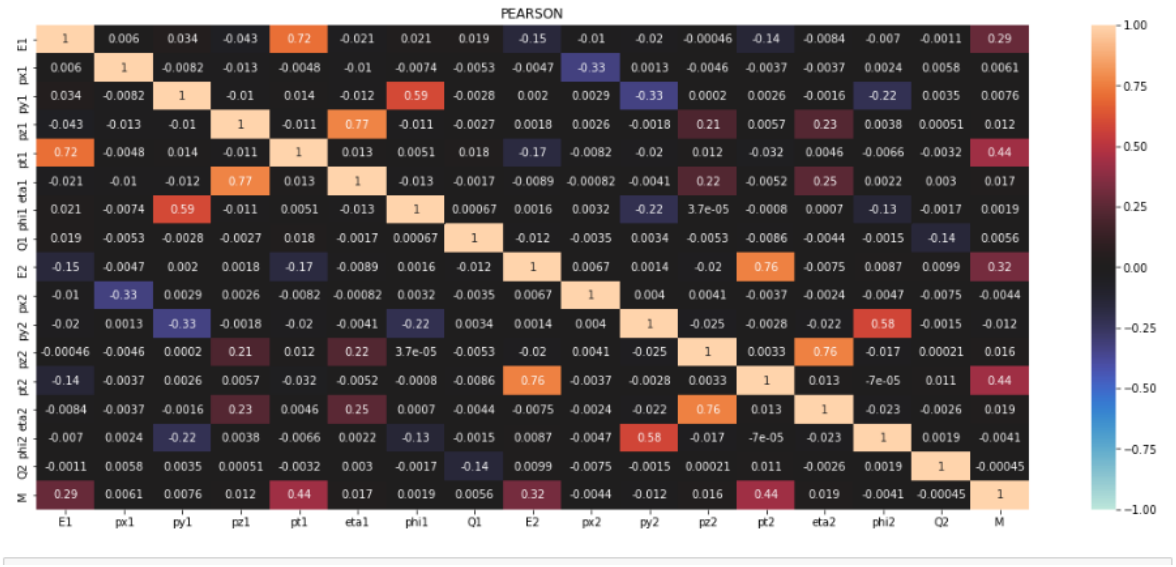


**Figure 1.** Pearson Covariance Heatmap

The second analysis technique will be our CatBoost Machine Learning model to achieve the desired outcome of predicting the invariant mass of an electron through the observed dielectron events. This prediction model was selected for data analysis based on the gradient boosting and decision tree algorithms on which the model is built and excels in analysis of numerical, categorical, and text data (DataAspirant 2021). We have prepared the CatBoost Prediction model by first determining the proper depth and and learning rate for the model. CatBoost was also selected for its optimized algorithm which is able to accurately calculate datasets with relatively less records(Thiesen 2021).

## 3. Outline of Findings

From exploratory analysis techniques we gained critical insights into the CERN dataset and relationships between components of electrons. In the analysis we individually examined each component of electron events including transverse momentum, charge, and phi angle. Our first notable insight appeared during observation of the Pearson covariance heat map. From the visualization we see a direct relationship between the invariant mass of the electrons (M) and the total energy of the electrons (E), we also see a relationship between invariant mass (m) and the transverse momentum of the electron (pt) measured in GeV. We expect to see a relationship between mass and energy as we learned in Einstein's famous equation E = MC2. It is interesting to see an relationship between energy and transverse momentum or the lateral electron momentum distribution which possesses a momentum component perpendicular to the polarization plane.

It was also very interesting, albeit exciting, to see the performance of the open-source CatBoost Machine Learning Prediction Model. Regression based machine learning models are most often evaluated based on their R2 score. An R2 score, also called a coefficient of determination, is an integral metric which measures variance in the predictions of the data (Kharwal 2021). When the value of R2 is equal to 1, this signifies the difference between the samples in the data and the predictions created by the model are in perfect sync of that the model is perfectly trained (Real Python 2021). The CatBoost machine learning model returned an R2 score of 0.98775 which is nearly a perfect score and shows no sign of either under fitting or over fitting. With an R2 score of nearly the ideal target of 1 we ensure the data produced is both reliable and accurate.

## 4. Limitations of Techniques and Tools

The greatest limitation of the research project is the limited availability of records in the Electron Dataset from the CERN organization for analysis. The 100,000 records provided allow for meaningful and insightful results which align with our research goal, however, more data would allow for increased accuracy in the final calculated outcome. It is for this particular reason that the CatBoost Machine Learning Prediction model was selected as the algorithm allows for accurate and efficient calculations on datasets with limited record availability (Affine 2022). CatBoost however was able to overcome this limitation and produce a model that is neither under fit or over fit. Another limitation is the resources available for computation. CatBoost is an extremely efficient algorithm able to calculate complex mathematics quickly. However, even with a sophisticated machine learning model the time to compute the necessary calculations resulted in long and arduous waiting periods. There are many options available for machine learning, but they require the use of complex and expensive equipment. The creation of more sophisticated models and larger datasets will require access to higher performance systems.

## 5. Proposed Actions

A proposed action for future study is to recommend continued analysis using similar and high-achieving regression algorithms such as CNN and OLS for a more thorough and significant comparison of Prediction Models that best align with our research goal (Katari 2020). There exist a plethora of machine learning algorithms and a study could be continued into the classification of predictive models by ideal R2 scores.

Another proposed action for future study would be to apply the insights and knowledge gained in this research project to observations from lateral CERN research experiments including photon particle physics or to the study of similar topics in the fields of chemistry and biology. The CatBoost Model is not limited to particle physics and significant advances could be made in technology and healthcare by continued study of prediction models.

## 6. Expected Benefits of the Study

Continued study in Machine Learning Prediction Models will serve well to enhance the field of Data Science and concurrently deliver increased insights to the scientific community which can lead to enhancements in countless various fields of knowledge such as economics, bio-sciences, medicine, and improving quality of life. The ability to create accurate prediction models could lead to enormous insights such as the ability to predict commodity prices in economics, or improving the quality of healthcare while decreasing costs. One particular field of interest in which Machine Learning Prediction Models could be applied is the area of inflationary cosmology. This data analysis has proven machine learning to be of value in studying particle physics particularly in areas with high amounts of information in datasets available and similar models could be applied into the study of information generated by simulations into the study of dark matter and dark energy. The prediction and later proof for the existence of dark matter would have profound effects in the world of physics as well as leading to new scientific breakthroughs in macro and micro matter/energy manipulations. It is possible these revelations will lead to a new industrial revolution with an vastly improved energy source.

## 7. References

McCauley, T. (1970, January 1). Events with two electrons from 2010. CERN Open Data Portal. Retrieved May 10, 2022, from https://opendata.cern.ch/record/304

CatBoost – a new game of machine learning. Affine. (n.d.). Retrieved May 10, 2022, from https://affine.ai/catboost-a-new-game-of-machine-learning/

How CatBoost algorithm works in Machine Learning. Dataaspirant. (2021, January 4). Retrieved May 10, 2022, from https://dataaspirant.com/catboost-algorithm/

Thiesen, S. (2021, February 19). CatBoost regression in 6 minutes. Medium. Retrieved May 10, 2022, from https://towardsdatascience.com/catboost-regression-in-6-minutes-3487f3e5b329

Exploratory Data Analysis in python course. DataCamp. (n.d.). Retrieved May 10, 2022, from https://www.datacamp.com/courses/exploratory-data-analysis-in-python

Supervised learning with scikit-learn course. DataCamp. (n.d.). Retrieved May 10, 2022, from https://www.datacamp.com/courses/supervised-learning-with-scikit-learn

Real Python. (2021, March 19). Linear regression in python. Real Python. Retrieved May 10, 2022, from https://realpython.com/linear-regression-in-python/

Katari, K. (2020, October 9). Simple linear regression model using Python: Machine Learning. Medium. Retrieved May 10, 2022, from https://towardsdatascience.com/ simple-linear-regression-model-using-python-machine-learning

Kharwal, A. (2021, June 22). R2 score in machine learning. Data Science — Machine Learning — Python — C++ — Coding — Programming — JavaScript. Retrieved May 11, 2022, from https://thecleverprogrammer.com/2021/06/22/r2-score-in-machine-learning/

Ivanov, I. A. (2016, January 7). Transverse electron momentum distribution in tunneling and over the barrier ionization by laser pulses with varying ellipticity. Nature News. Retrieved May 11, 2022, from https://www.nature.com/articles/srep19002

Wikimedia Foundation. (2022, April 5). Invariant mass. Wikipedia. Retrieved May 11, 2022, from https://en.wikipedia.org/wiki/Invariant$_{m}ass$

Wikimedia Foundation. (2022, April 16). Lorentz transformation. Wikipedia. Retrieved May 11, 2022, from https://en.wikipedia.org/wiki/Lorentz$_{t}ransformation$