Machine Learning Approach to Predict the Invariant Mass of Two Electrons

**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 detector. The CMS can generate huge amount of data for particle collisions at 0.9 – 13 TeV.

Using the machine learning we can use robust computing system alongside modern algorithms to observe and analyse insights from enormous amount of data quickly and efficiently. In this project, we are not only trying to demonstrate the statistical significance of machine learning models in the field of particle physics by predicting the invariant mass of electrons based on the observation from the CMS detectors but also manipulating the features of the dataset by creating new features such that these features show higher relationship with the target variable, Invariant mass and increase the performance of the ML models.

**Related Works**

Radovic, A, et al., in their research summarized the challenges and opportunities that come with the use of machine learning at the frontiers of particle physics. The paper discusses the challenges of analysing the big data produced by the LHS experiment and how machine learning helps in real time analysis to combat it. It also touches on the use of deep learning such as CNNs and RNNs as the outputs of many particle-physics detectors can be viewed as images.

Richard M. Flores, has done exploratory analysis on the CERN electron dataset and showed the statistical significance in predicting mass in subatomic particle collisions. The model presented by the paper is the CatBoost regression model and only two hyperparameters, “Depth” and “Learning Rate”, were optimized using GridSearch. The model was trained on CERN dataset that was pre-processed and no additional feature engineering was performed.

**Objectives**

* This project aims to expand over the original dataset, taken from CERN, by introducing new features using the original features of the dataset.
* Showing the statistical significance of the new features by comparing the correlations towards the target, Invariant Mass.
* Testing the correlation hypothesis using p-value test and removing any multicollinearity problems that occur due to new features.
* Comparing the performance of 10 regression models for the original dataset and new dataset.
* Choose the best performing regression model and perform hyperparameter tuning.
* Comparing the performance of the chosen model trained on original test dataset and new test dataset.

**Methodology**

Dataset

The dataset used for this research is provided courtesy of the CERN open data portal [8] contains observations of one hundred thousand electrons 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 two electrons in GeV

The dataset is further processed by examining, cleaning, and analyzing the data and its features. As duplicate data are an extreme case of nonrandom sampling, as well as they bias any of the models, leading to overfitting problems.

Feature Engineering

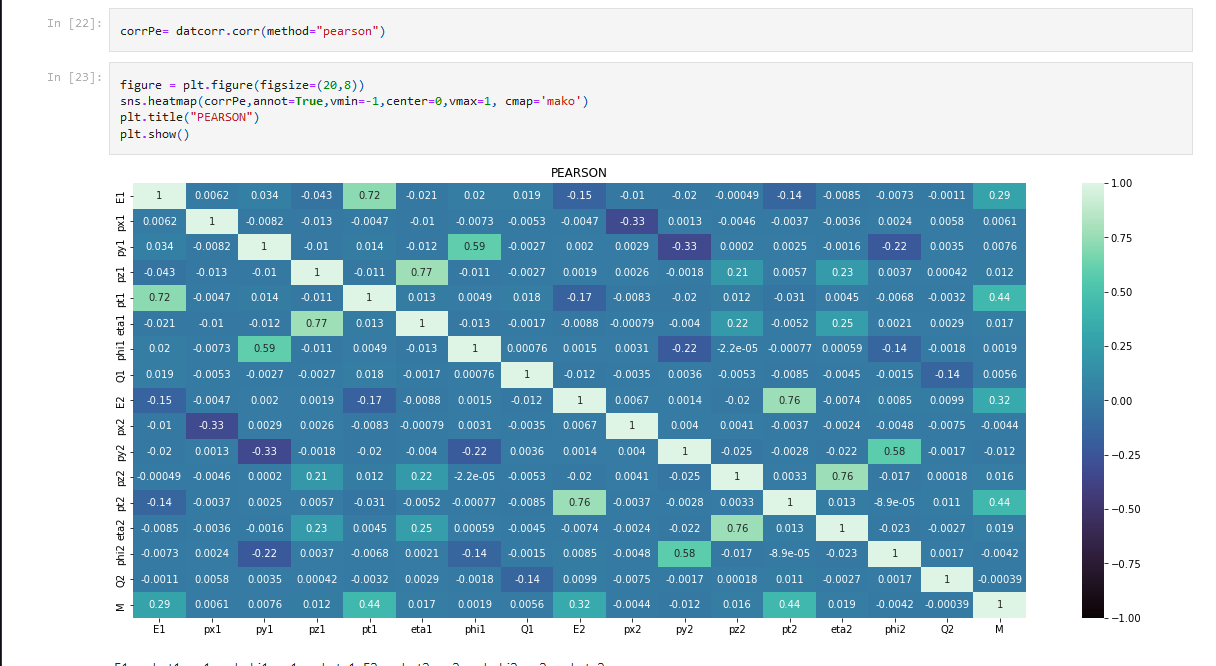


Figure 1 Heatmap for the correlation coefficients of features within the original dataset

Figure 1 is the heatmap for the Pearson’s correlation coefficient values of the dataset. The major focus here is the coefficient value between the target value M and each independent variable. The heatmap reveals that E1, E2, pt1 and pt2 features have high correlation values 0.29, 0.32, 0.44 and 0.44 respectively, with the target M compared to all the other features.

This project presents new features that are created using the original features.

* **E12**: the product of features **E1** and **E2**.
* **pt12**: the product of features **pt1** and **pt2**
* Similarly, **eta12, px12, py12, pz12, phi12**, the products for their respective features.

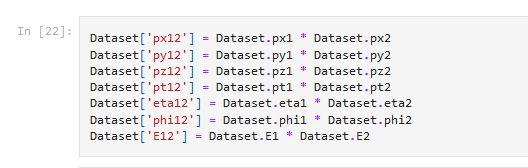


Figure 2 Creating the new features



Figure 3 Correlation Heatmap of the new features

The heatmap from Figure 3 shows the correlation between the new features with the dependent variable M. With the min correlation value to be 0.11, we can say that that the new features have good relation to the dependent variable M.

With the introduction of new features to the dataset, 2 test needs to be done. First, perform the p-value test, to prove the significance of the correlation that is evaluated and secondly deal with the problem of multicollinearity.

The p-values test is a measure of significance to validate a hypothesis against the observed results or data. Here, the null hypothesis is that there is no correlation between the features and the target. The threshold for this test is p<0.05, wherein the null hypothesis is rejected. The p-values for all the features, including the newly created ones, are calculated using Ordinary Least Square Regression model. Every feature has a p-value less than 0.05 and therefore null hypothesis is rejected, proving the correlations found are significant to use.

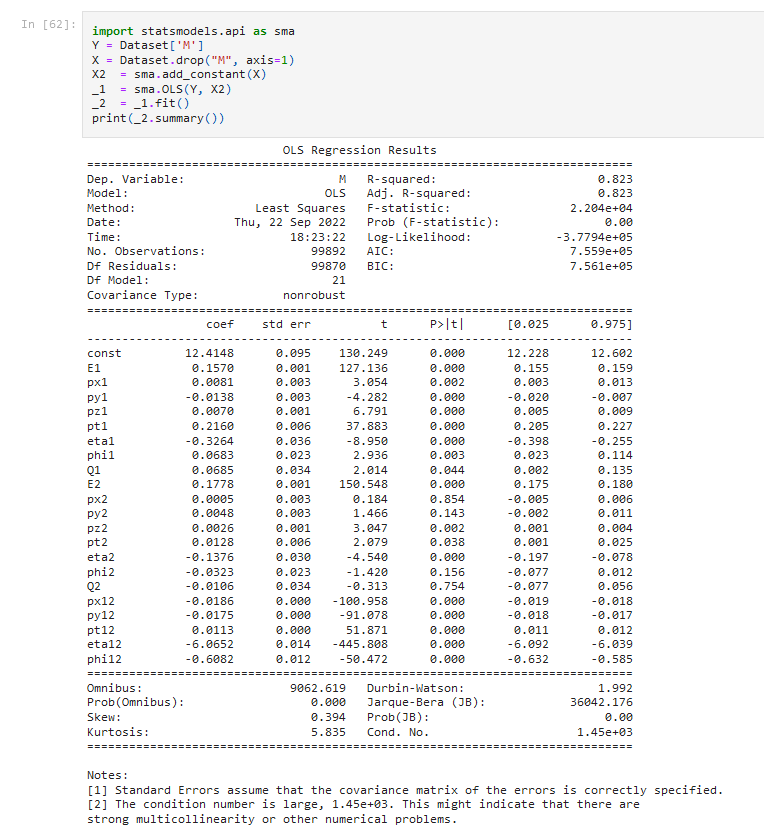


Figure 4 p-value testing using OLS

As new features are created, the problem of multicollinearity needs to be solved, where 2 or more independent variables are highly correlated with each other. Multicollinearity is detected by calculating the variance inflation factor (VIF) of each feature in the dataset, it is the measure of how much the standard error of the estimate of the coefficient increased due to multicollinearity. The threshold, accepted, for the VIF value is less than 10. Above which the feature is removed.

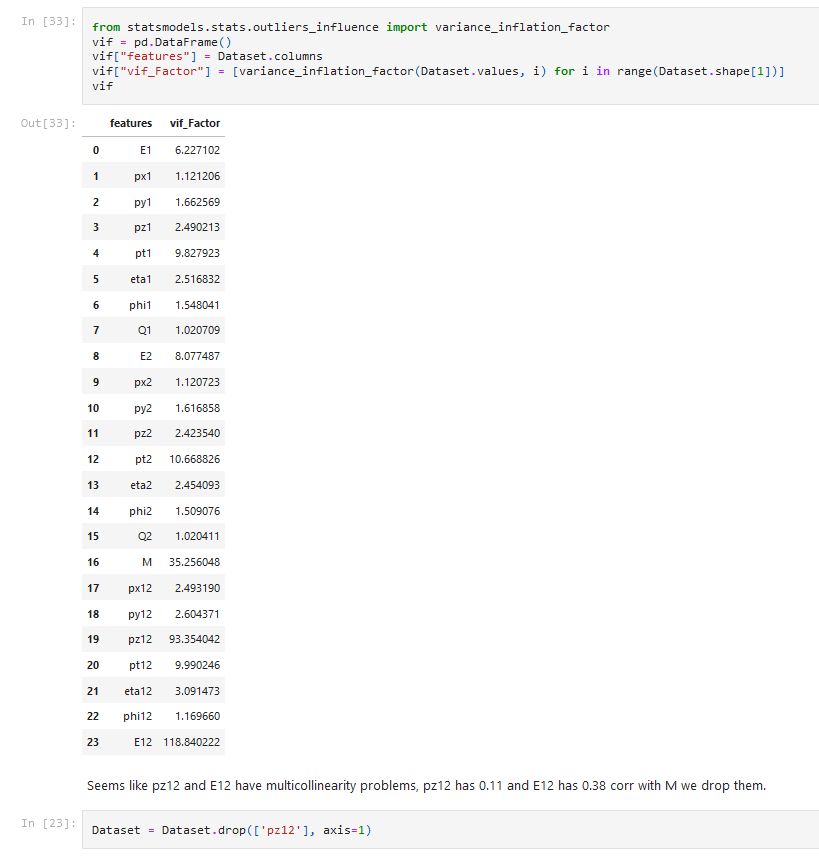
 Figure 5 VIF calculation for multicollinearity test.

Figure 5 reveals that 3 features have strong multicollinearity problems. As E12 and pt2 have high correlation with M compared to pz12, seen from figure 2, this feature is dropped and the new recalculated VIF values for all remaining features are within the threshold.

Model Selection and Model Training

Ten regression models are selected and fitted to the original dataset and the new added feature dataset. Following are the regression models used:

* Decision Tree Regressor
* Gradient Boosting Regressor
* LASSO aka Least Absolute Shrinkage and Selection Operator
* Ridge Regressor
* Random Forest Regressor
* PLS Regressor aka Partial Least Squares Regressor
* Elastic Net
* XGBoost Regressor aka as eXtreme Gradient Boosting
* LightGBM aka Light Gradient Boosting Machine
* CatBoost Regressor

These models are fitted with the 2-dataset using default hyperparameters.

TABLE I  
Evaluation metrics using original dataset

|  |  |  |
| --- | --- | --- |
| **Model** | **R2 Score** | **RMSE** |
| DecisionTreeRegressor | 0.66747 | 14.55148 |
| CatBoostRegressor | 0.99037 | 2.47511 |
| ElasticNet | 0.39832 | 19.56982 |
| GradientBoostingRegressor | 0.74486 | 12.74279 |
| Lasso | 0.39835 | 19.56938 |
| LGBMRegressor | 0.95843 | 5.14299 |
| PLSRegression | 0.39183 | 19.67583 |
| RandomForestRegressor | 0.90155 | 7.91384 |
| Ridge | 0.39804 | 19.57434 |
| XGBRegressor | 0.96420 | 4.77334 |

TABLE II  
Evaluation metrics using new dataset

|  |  |  |
| --- | --- | --- |
| **Model** | **R2 Score** | **RMSE** |
| DecisionTreeRegressor | 0.97820 | 3.72371 |
| CatBoostRegressor | 0.99699 | 1.38363 |
| ElasticNet | 0.81897 | 10.73765 |
| GradientBoostingRegressor | 0.98590 | 2.99343 |
| Lasso | 0.82125 | 10.66939 |
| LGBMRegressor | 0.99401 | 1.95138 |
| PLSRegression | 0.79007 | 11.56486 |
| RandomForestRegressor | 0.99092 | 2.40208 |
| Ridge | 0.82179 | 10.65290 |
| XGBRegressor | 0.99333 | 2.05787 |

Tables I and II are the evaluation metrics of all the regression models that are fitted on the original dataset and the new dataset respectively. These metrics are calculated by taking the means of the R2 score and RMSE values evaluated across 10 – fold cross-validation. Keep in mind these models are fitted on their default hyperparameters. The four gradient boosting regression models, GBR, LGBM, XGBoost, and CatBoost stand out the most out of all the models.

As the CatBoost Model has the best performance out of all models for both dataset, this model is selected for the project.

Hyperparameter Tuning

In the previous section, it is mentioned that the models were evaluated on their default hyperparameters. The software framework Optuna automates the steps involved in hyperparameter optimization. A sampler using TPE algorithm provided by Optuna is used, to find the optimal hyperparameters for the CatBoost Model.



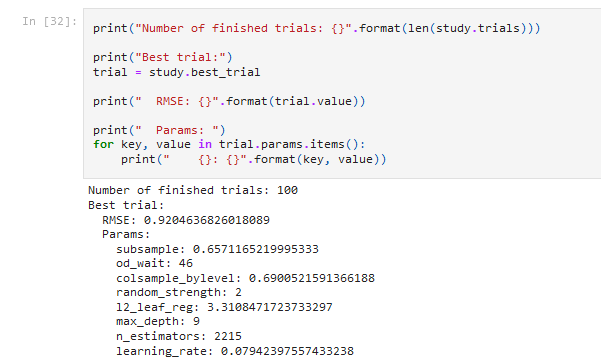


Figure 6 Optuna with TPE sampler for hyperparameter tuning.

Results

The original dataset and the dataset with new features are trained on CatBoost model with the hyperparameters found using Optuna in the previous section. First, we will analyze the impact of the new features on training the model compared to features of the original dataset. This is evaluated using SHAP [20] values, SHapley Additive exPlanations, which is used to make machine learning models more transparent and understandable.

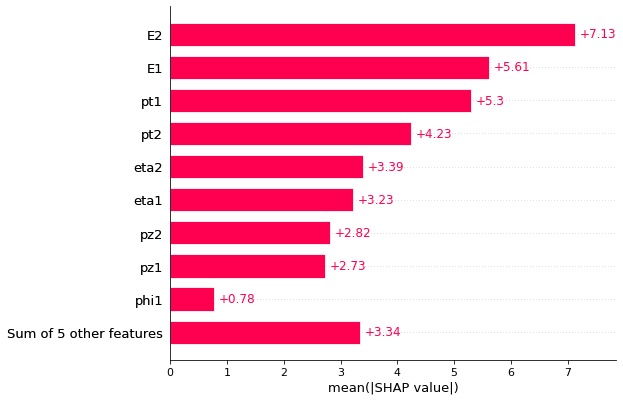
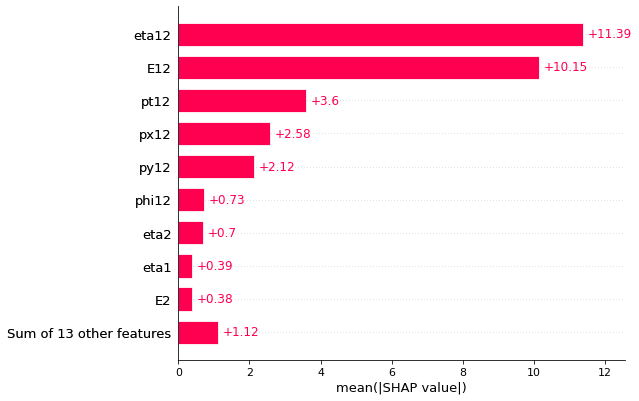
 

Figure 7 Bar graph of the mean SHAP values calculated for each feature of the original (left) & new (right) dataset.

As seen in figure 7, the features are ordered from the highest to the lowest effect on the prediction. This bar plot considers the absolute SHAP value, as such it does not show the positive or negative impact of the features.

The features E1, E2, pt1, pt2, eta1 and eta2 have high mean SHAP values for the left graph. This shows these features have higher impact on the prediction then the other features. From the right figure, it can be inferred that the new features, have higher mean SHAP values. These new features, eta12, E12, pt12, px12, py12, phi12, that are introduced by this paper have a higher impact on the predictions compared to the original features in the dataset.

TABLE III  
Performance of CatBoost Regression model (10-CV) for Test data of original dataset and new dataset

|  |  |  |
| --- | --- | --- |
| **Evaluation Metrics** | **Original Dataset** | **New Dataset** |
| R2 Score | 0.99405 | 0.99710 |
| RMSE | 1.94644 | 1.35979 |

The original dataset and new dataset are divided into training and testing set with 80/20 split. CatBoost with the optimized hyperparameters is trained with both these training sets. Table III shows the mean evaluation metrics, across 10-fold cross validation of the test data for both the datasets. It is clear that the new features not only have more influence but also increase the performance of the model.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 8 Joint graph with regression (scatter) plot and histogram of actual values vs predicted value for both new data (left) and original data (right)

Figure 8 Joint graph with regression (scatter) plot and histogram of actual values vs predicted value for both new data (left) and original data (right)

There is a very small difference between the R2 score of original dataset and new dataset, as such the joint graph [21] in figure 5 is very similar. As the R2 score of new dataset, 0.99710, is slightly greater than the original dataset, 0.99405, the scatter plot is bit more compressed.

The model is now ready to predict the invariant mass for di-electron.