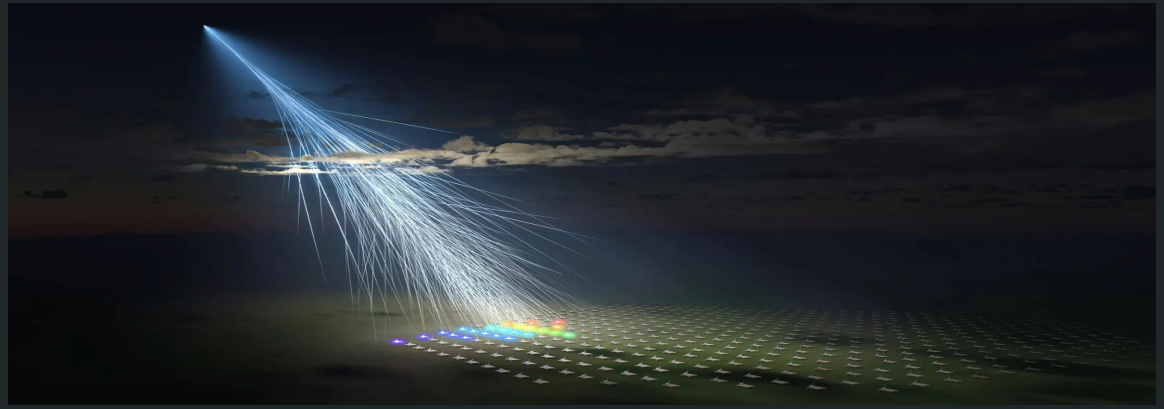
Weather to Cosmic Rays: Machine Learning for Particle Flux

Carter Chapman
Andrei Gogosha

Presentation Outline



Introduction
Overview, Cosmic Rays, Pierre Auger Observatory

Methods
SHAP Feature Importance, Data, Preprocessing, Weather, Model

Results
Model Performance and SHAP Feature Importance

Conclusion
Outlook, Future Improvements

Figure 1: Particle Shower induced by Cosmic Ray [1]

Overview

- Use weather data from the Pierre Auger Observatory to predict cosmic ray count rates
- Understanding this relationship helps pick experiment site location, understand variations in cosmic ray data, and protect sensitive equipment from cosmic ray exposure
- Use SGDRegressor (linear), and a gradient boosted decision tree (non-linear)
- Verify our physics based understanding by investigating feature importance with SHAP (SHapley Additive exPlanations)
- Goal is to have a Mean Absolute Error < 15 counts per second

Cosmic Rays

- Energetic particles from outer space
- Can originate from the Sun, black holes, supernovae, and outside our galaxy
- Interactions with the atmosphere create showers of secondary particles (air showers)
- Secondary Showers are measured on the surface for insight into cosmic ray energy, however are subject to variations in the weather

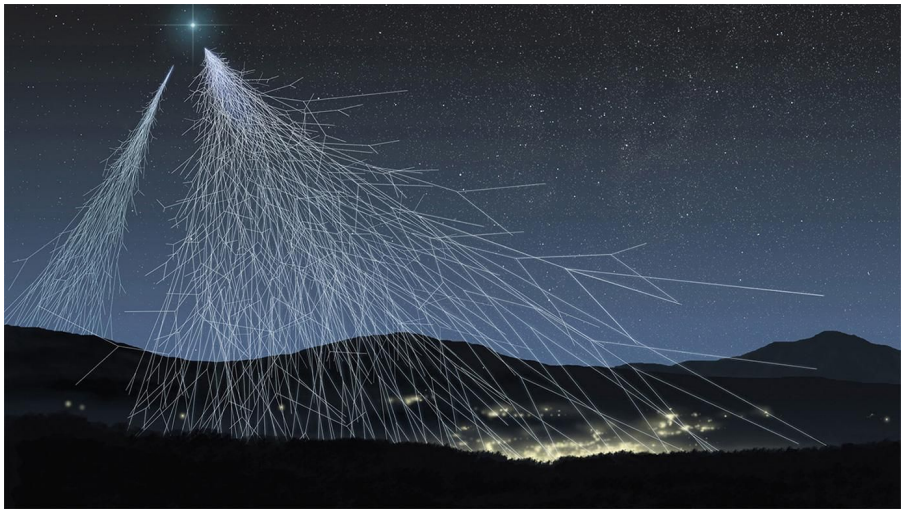


Figure 2: Particle Shower induced by Cosmic Ray [2]

Pierre Auger Observatory

- Cosmic Ray observatory located in Argentina
- 1660 Water tank detectors are spread out over 3000 km^2
- Detect particles on the surface of the earth
- Also have UV detectors that measure the brightness of UV light emitted from interactions in the atmosphere
- Pierre Auger has five different weather stations taking measurements on pressure, density, temperature

Figure 3: Water Detector at Pierre Auger [3]

Weather

- Different atmospheric variables affect the characteristics of an air shower
- While not a huge factor, variables can induce biases in measurements
- Pierre Auger already uses linear correction shown in eq 1, to account for variations in weather
- Our goal is to use ML Models to recreate this relationship leveraging the fact that this relationship exists

$$S(r_{\text{opt}}) = S_0 \left[1 + \alpha_P (P - P_0) + \alpha_\rho (\rho_{\text{d}} - \rho_0) + \beta_\rho (\widetilde{\rho} - \rho_{\text{d}})\right]$$

Equation 1: Correction from Pierre Auger to adjust for weather conditions. S(r_opt) is the detected signal, S_o is the expected signal, alpha and beta are coefficients, P_0 and p_o are the yearly averages for pressure and density, P and p_d are the measured pressure and density, and p_~ is the density two hours ago (to account for changes in density higher up in the atmosphere where the air shower began [4]

SHAP (SHapley Additive exPlanations)

- Method for interpreting machine-learning models by explaining how each feature contributes to a single prediction and to the model overall
- Based on Shapley Game Theory, treats each feature as a player in a game and assigns each feature a share of responsibility in making a decision
- Model Agnostic
- Allows us to directly see which features have the greatest impact on the final predicted cosmic ray flux
- Greater model transparency and illustrates direct relationships between the prediction and feature values [5]

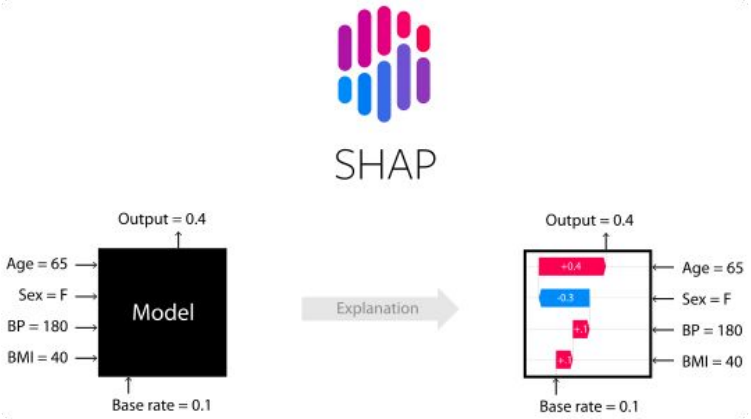


Figure 4: Water Detector at Pierre Auger [7]

Data

- Data is from the Pierre Auger Open-Release Scalar Data set [7]
- Has 10% of all cosmic ray data, and 100% of all weather data collected from March 2005 to December 2020
- Weather data used is a culmination of data collected at every station
- Scalar mode counts the total number of hits across all 1660 detectors within a 1 second coincidence window and is uncorrected for weather effects and operable detectors
- Scalar Data has been used to measure solar activity, and Forbush decreases [8]

Pre-Processing

- Scalar data has average particle count rates over 15 minute intervals
- Atmospheric data has average data collected over 5 or 10 minute periods
- Inspection of Data revealed high dependency on the year the data was taken
- Dependency on year comes from changes in data-taking strategy [9]
- Trained on Data taken post 2012
- Trained models with year as a one hot encoded feature and without it

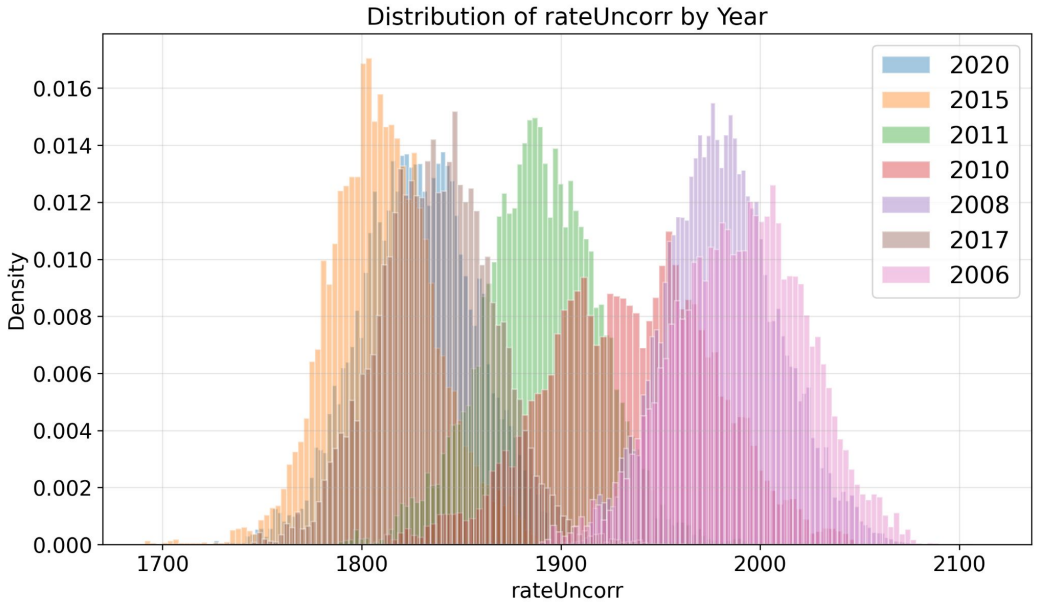


Figure 4: Histogram of counts/s by year. Shows high variance between years when data was taken. Each year appears to have its own gaussian distribution

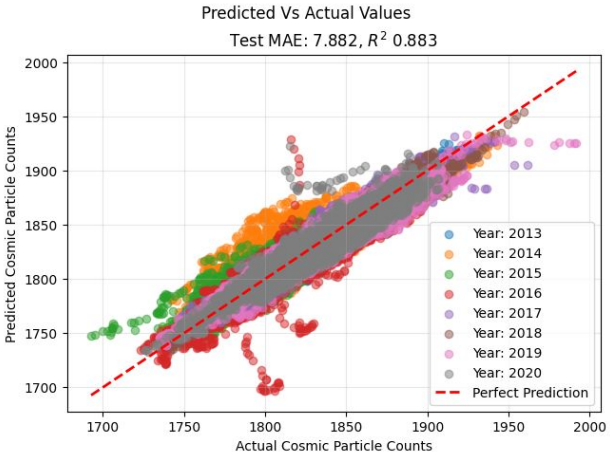
Models

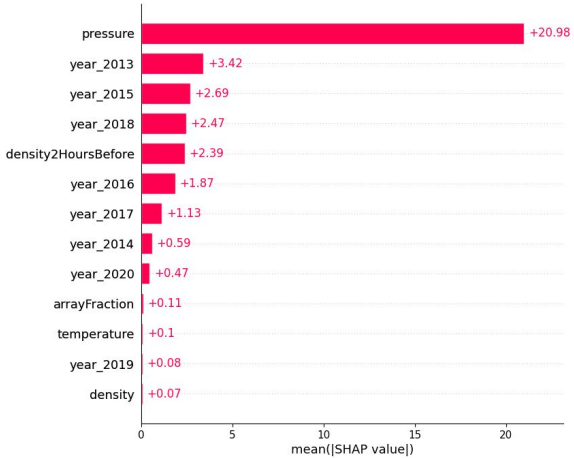
SGDRegressor and HistGradientBoosting

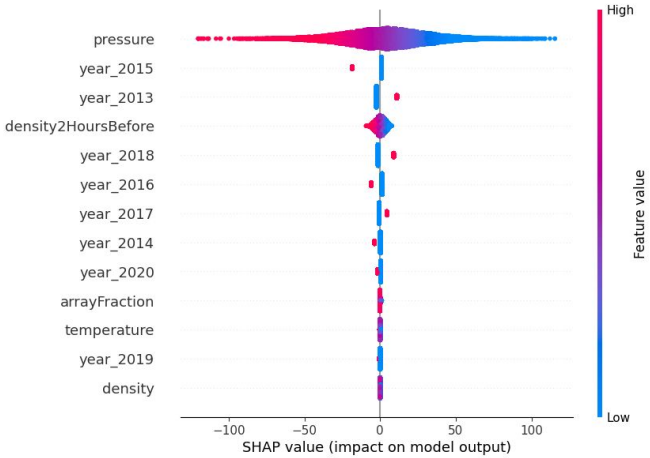
- Performed Nested Cross Validation:
  - 5 Fold Outer CV
  - 3 Fold Inner CV
- Grid Search for Hyper-parameter optimization
- Goal: Minimize Mean Absolute Error
- Trained two models, one with OneHotEncoder Year features, one without

SGDRegressor Performance and Interpretability

- Performance with time features: Nested CV MAE = 7.88 +/- 0.01

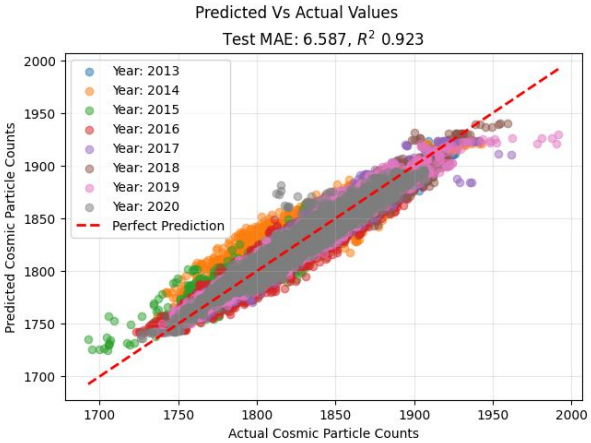


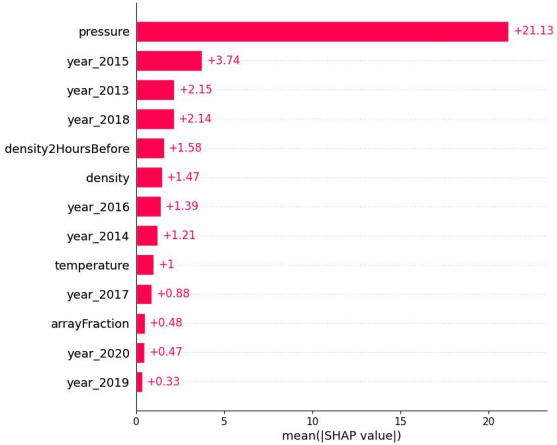


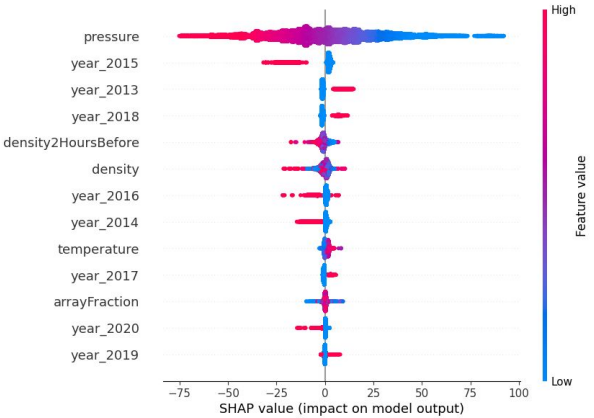


HistGradientBoostingRegressor Performance and Interpretability

- Performance without time features: Nested CV MAE = 6.59 +/- 0.02







Conclusions

Results:

- HistGradient Boost outperformed SGDRegressor, much faster run time and lower MAE
- Both models met out goal of MAE < 15 counts per second with and without year data
- Feature Importance confirmed pressure was the major player in decision making
- Revealed year as an unexpected important player, revealed inconsistencies in data-taking that the model would have been unable to account for otherwise

Future:

- Investigate Weather dependence on particle phenomenology (types of particles, energies, spatial distribution)
- Generalize model to work for different locations

References

[1]astronomers-highest-energy-particles(Picture on Presentation Outline Slide)

[2] what-are-cosmic-rays (Picture on Cosmic Ray Slide)

[3]hybrid-detector (Picture of Water Tank on Pierre Auger Slide)

[4]1ef8bb507e47d8bc4ac3b456b7ff55ef(Weather Correction Pierre Auger)

[5]8a20a8621978632d76c43dfd28b67767-Paper.pdf[SHAP Citation]

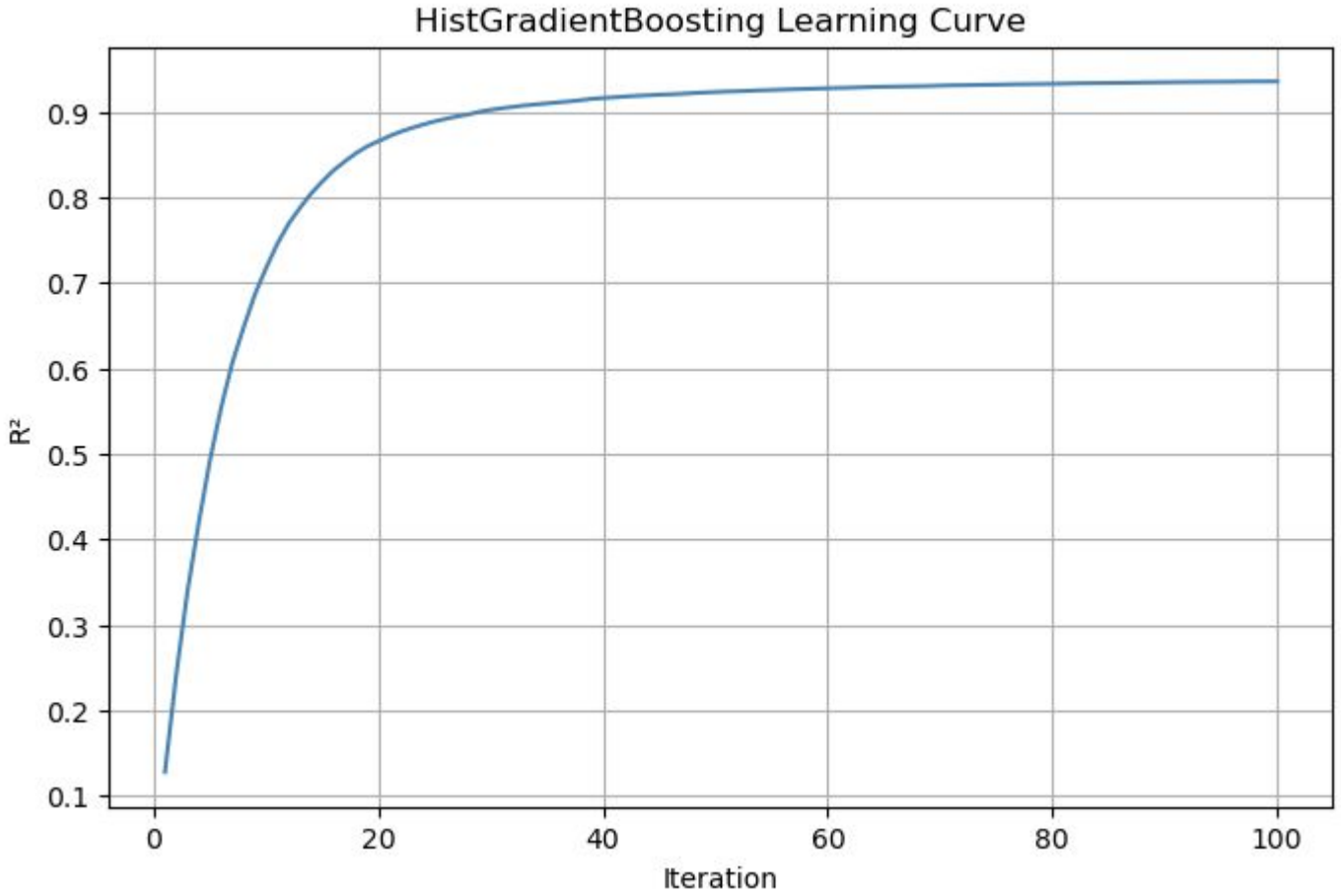
[6]shap (Shap Image and Github)

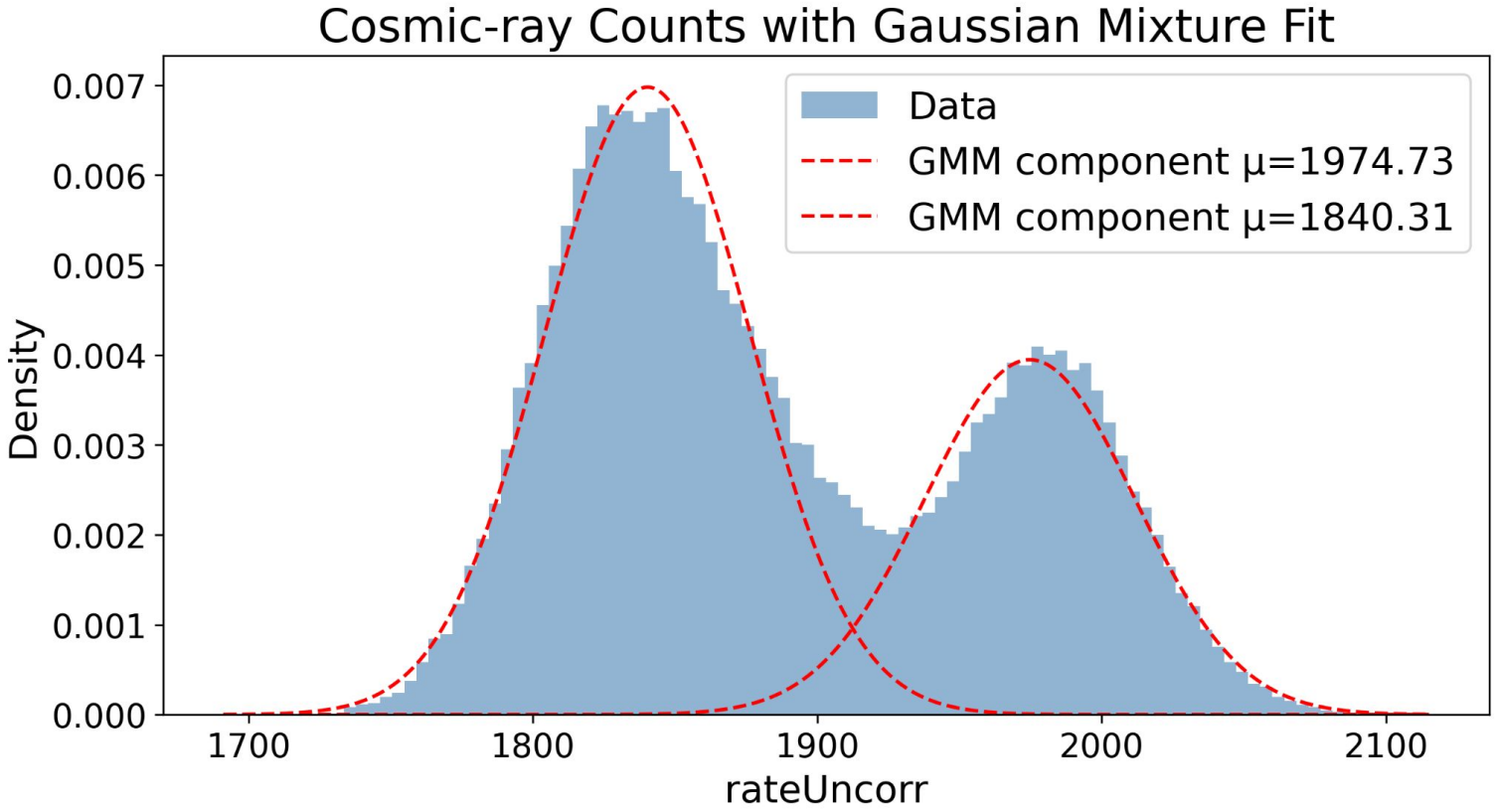
[7]10488964[Open-Data]

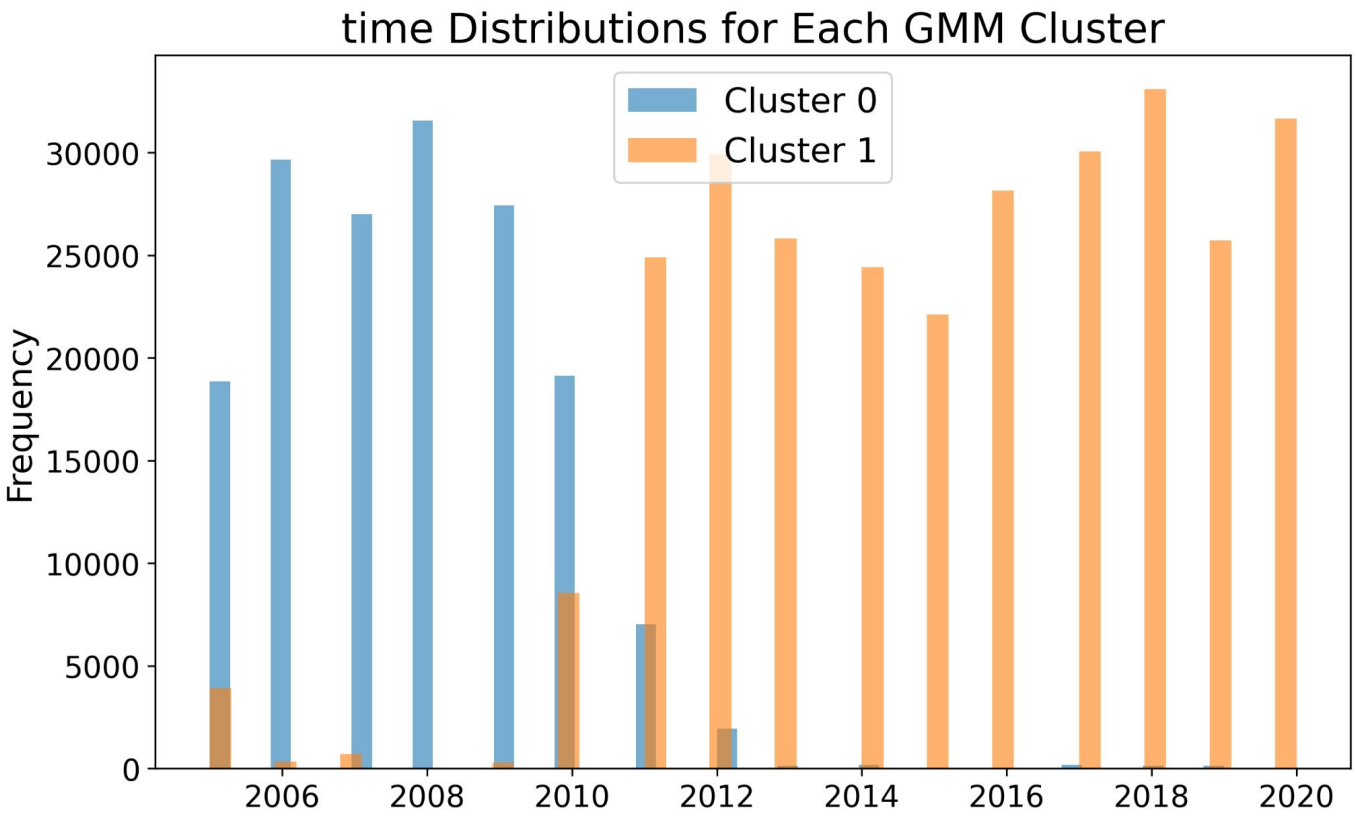
[8] pdf (Pierre Solar Activity Paper)

[9]pdf(Pierre Scalar Data Paper)

Questions?


HistGradientBoosting Learning Curve
R²
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
20
40
60
80
100
Iteration


Cosmic-ray Counts with Gaussian Mixture Fit
Data
GMM component μ=1974.73
GMM component μ=1840.31
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
1700
1800
1900
2000
2100
rateUncorr


time Distributions for Each GMM Cluster
Cluster 0
Cluster 1
Frequency
0
5000
10000
15000
20000
25000
30000
2006
2008
2010
2012
2014
2016
2018
2020