Project Data Science AY 25–26

Geothermal Powerplant: Predictive Seismic Traffic Light System

Final Technical Report of **TEAM 1**

12th December 2025

Programme: Master of Statistics and Data Science – Project Data Science
Case: Geothermal Energy (Balmatt)

GitHub repository:
https://github.com/kiflomhailu/project_datascience_secondYear/tree/main
Interactive dashboard: https://geothermal-dashboard.onrender.com/

Tanjim Hossain
Muhammad Ammad
Laiba Tahir
Thierry Fotabong
Berhe Kiflom
Alain Patrick

Abstract

Geothermal electricity production at the Balmatt site is constrained by the risk of induced seismicity. The current traffic light system (TLS) is reactive: it labels each 15-minute window as green, orange or red *after* events have occurred. Our goal is to design a predictive TLS that uses operational time-series data to forecast seismic risk hours in advance, allowing operators to adjust injection and production parameters before critical events occur.

We combine several machine learning models built on five years of high-frequency operational and seismic data (696,275 records, 30 initial features). Our final architecture consists of (i) a highly calibrated binary classifier for event occurrence, (ii) a regression model for event magnitude conditional on an event and (iii) a three-class classifier that directly predicts TLS levels (green, yellow, red). To address extreme class imbalance 0.14% events we use a mix of class weighting and targeted downsampling of seismically quiet COVID years (2021–2022).

The best binary CatBoost model achieves an AUC of 0.999997 on a strict time-based test split (2018–2025) and at an optimised decision threshold, yields only 2 missed events and 0 false alarms on 101,404 test samples. The magnitude regressor reaches $R^2 = 0.54$ and correlation 0.97 between predicted and true magnitudes, while the 3-class TLS model attains macro F1 = 0.77 with perfect recall for red states and moderate recall for yellow states. Feature importance analyses highlight the dominant role of peak ground velocity, magnitude history and pressure–temperature interactions.

We conclude that the proposed predictive TLS is technically feasible for near real-time deployment and discuss remaining limitations such as covariate shift, magnitude underestimation and dependency on a single site.

Contents

1 Introduction

1.1 Domain context

This report presents a detailed overview of the Seismic Traffic Light Prediction System v2, designed to detect seismic events, estimate their magnitudes and assign a 3-class traffic light status (GREEN, YELLOW, RED). The improved pipeline incorporates refined handling of sentinel values, COVID-period downsampling, time-based splitting and three specialized machine learning models: event occurrence prediction, magnitude estimation and multi-class traffic light classification.

1.2 Objective and research questions

Our overarching goal is to transform the TLS from reactive to predictive. The central research question is:

> **To develop a predictive traffic light system that forecasts seismic risk 48 hours ahead, enabling proactive operational adjustments to prevent damaging earthquakes while maximizing geothermal energy production efficiency.**

This was broken down to specific sub questions:

1 Which operational variables are most strongly associated with induced seismic events?

2 Can we accurately predict the occurrence of a seismic event within a 48-hour horizon?

3 Conditional on an event occurring, can we forecast its magnitude with sufficient precision?

4 Can we map model outputs to a robust three-colour TLS (green, yellow, red) that balances early warning and false alarms?

1.3 Overview of our approach

The project was organised in several sprints. In early sprints we focused on understanding the data, cleaning and building baseline models (CatBoost, LSTM). We then iteratively refined the modelling pipeline, addressed issues like missing values, temporal leakage and class imbalance.
Our final solution is a multi-model CatBoost-based system:

1. A binary classifier for event occurrence
2. A magnitude regressor trained only on event windows
3. A three-class classifier directly predicting traffic light colours.

A time-based train (2018 to 2024) at 80 and test (2024 to 2025) at 20 split ensures proper temporal evaluation for deployment.

2 Data and Exploratory Analysis

2.1 Data sources

We worked with vito operational and seismic dataset containing:

- High-frequency operational data from injection and production wells (pressures, flows, temperatures, energy, cumulative volumes)

- Seismic measurements (magnitude, peak ground velocity, hourly seismicity rate)
- Timestamps for operational phases and seismic occurrences.

The combined dataset contains 696,275 records with 30 base features, sampled at roughly 15-minute intervals between 2018 and 2025.

2.2 Target definition

For a predictive TLS we need multiple targets:

- **Event occurrence** (y_{event}): binary indicator of whether a seismic event occurs in the prediction horizon.
- **Event magnitude** (y_{mag}): continuous magnitude for event windows.
- **Traffic light level** (y_{TLS}): three classes based on magnitude:

$$\begin{aligned} \text{GREEN} &: mag < 0.17, \\ \text{YELLOW} &: 0.17 \leq mag < 1.0, \\ \text{RED} &: mag \geq 1.0. \end{aligned}$$

Using these definitions, only 953 of the 696,275 windows are labelled as events 0.14%, while seismic magnitudes are often recorded as a sentinel value of -999 when no event is present.

2.3 Data quality issues and cleaning

Initial exploratory data analysis revealed several issues:

- **NA Values:** NA values in some of the columns like `pgv_max`, `magnitude` and `hourly_seismicity_rate`.
- **Negative magnitudes:** some small-magnitude events have slightly negative values.
- **Timestamp inconsistencies:** operational and seismic timestamps originate from different systems and needed alignment to a unified `recorded_at` index.
- **Extreme class imbalance:** fewer than 1 in 700 records corresponds to a seismic event.

We applied the following cleaning steps:

- Replace the NA in the three columns by 0
- Parse all date-time columns using `pandas.to_datetime`
- Sort chronologically by `recorded_at`

2.4 Exploratory analysis

We conducted descriptive analysis and visualisations like time series, histograms, correlation heatmaps, scatter plots.
Key findings:

- **Operational vs seismic features:** correlation between purely operational variables (pressures, flows, temperature) and seismic magnitude is modest but non-zero. However, the seismic proxy `pgv_max` is highly informative for magnitude.

- **Temporal patterns:** seismic activity clusters in specific episodes linked to injection campaigns. The COVID period (2021–2022) is unusually quiet with many green windows and almost no events.
- **Class imbalance:** the vast majority of windows are green and non-event, implying that naive models can achieve $> 99\%$ accuracy with zero practical usefulness. Metrics such as precision–recall and confusion matrices are therefore crucial.

In an earlier sprint, we trained a first CatBoost classifier and visualised: (i) a confusion matrix at threshold $t = 0.5439$, (ii) a precision–recall curve (average precision ≈ 0.98), (iii) the probability distribution for event and non-event windows and (iv) the top 15 feature importances (dominated by `pgv_max`, `magnitude`, and pressure–flow interactions). While ranking quality was strong, at this threshold the model produced 72 true positives but 138,778 false alarms, which is operationally unacceptable. This motivated the later redesign of the model and thresholding strategy.

3 Methods

3.1 Preprocessing and feature engineering

On top of the cleaned base data, we engineered additional features to capture temporal dynamics and physical interactions.

Temporal features From `recorded_at` we extract:

- Hour of day (`hour`);
- Day of week (`day_of_week`) and weekend indicator (`is_weekend`);
- Month of year (`month`).

Phase-related features Using `phase_started_at` we compute the elapsed time since the start of the current operational phase:

$$\texttt{phase\_duration\_hours} = \frac{\texttt{recorded\_at} - \texttt{phase\_started\_at}}{3600 \text{ seconds}}.$$

Rolling statistics For important operational variables we compute rolling-window statistics with windows of 6, 12 and 24 hours:

- Injection temperature: rolling mean and standard deviation;
- Injection wellhead pressure: rolling mean;
- Production flow: rolling maximum.

These capture short-term trends and variability, which are often more predictive than raw levels.

Rates of change and interactions We add first differences for variables such as injection temperature, injection pressure, cumulative injected energy and production temperature to capture accelerations in stress. We also create physically motivated interaction features:

- $\texttt{pressure\_diff} = \texttt{inj\_whp} - \texttt{prod\_whp}$;
- $\texttt{temp\_diff} = \texttt{inj\_temp} - \texttt{prod\_temp}$;

- `temp_pressure_interaction` = `inj_temp` × `inj_whp`;
- `flow_pressure_interaction` = `inj_flow` × `inj_whp`;
- Normalised cumulative energy: `cum_inj_energy` / (`cum_volume` + 10^{-6}).

After feature engineering we obtain 49 numerical and categorical predictors.

3.2 Train–test split and imputation

To mimic deployment, we use a strict time-based split:

- Training set: first 80% of the time line (2018-11-28 to 2024-09-29), 405,612 samples.
- Test set: last 20% (2024-09-29 to 2025-09-19), 101,404 samples.

We replace infinite values by NaN and impute:

- Median imputation for numerical features (medians computed on the training set and applied to train and test).
- String encoding with "missing" category for any remaining categorical features.

The vector of training medians is stored (`train_medians_v2.pkl`) to ensure consistent preprocessing during inference and in the dashboard.

3.3 Handling class imbalance

Even after feature engineering, the event rate in the training data is only 0.21% and in the test set 0.047%. Such extreme imbalance can lead to degenerate models.

We combine two strategies:

1. **COVID-period downsampling** (Step 2.5 in the code): the years 2021–2022 contain many green windows with almost no seismicity. We randomly retain 10% of green windows in this period while keeping all yellow and red windows and all pre/post-COVID data. This reduces the dataset from 696,275 to 507,016 samples while preserving critical events.
2. **Class weighting:** for the binary event classifier we set the CatBoost parameter `scale_pos_weight` to the ratio of negatives to positives (≈ 468). For the 3-class traffic light model we compute inverse-frequency weights per class.

3.4 Model portfolio

Throughout the project we experimented with multiple architectures:

Logistic regression (baseline) Simple linear classifier on a limited set of standardised features. It provided a sanity check but underfit non-linear relationships.

Random forest Ensemble of decision trees with bootstrap sampling. It captured interactions better but struggled with extremely rare events and temporal structure.

LSTM Recurrent neural network applied to short time windows. The model was complex to train, sensitive to hyperparameters and did not outperform tree-based methods given our data volume and engineering time.

CatBoost (final choice) Gradient boosting on decision trees with strong built-in handling of categorical features, class imbalance options and robust performance on tabular data. After hyperparameter tuning, CatBoost delivered the best trade-off between accuracy, interpretability and training time.

The final system is built entirely on CatBoost models.

3.5 Final multi-model architecture

Model 1: Event occurrence classifier A CatBoostClassifier is trained on all samples with targets $y_{\text{event}} \in \{0, 1\}$. Key hyperparameters:

- 3,000 iterations, learning rate 0.02.
- Depth 6, loss function `Logloss`, evaluation metric F1.
- `scale_pos_weight` ≈ 468.
- Early stopping after 200 rounds without improvement.

Model 2: Magnitude regressor Conditioned on event occurrence, we train a CatBoostRegressor to predict y_{mag}. We subset the training and test sets to event windows (865 train, 48 test) and fit with:

- 2,500 iterations, learning rate 0.02
- Depth 7, loss function RMSE, evaluation metric MAE
- Early stopping after 200 rounds.

Model 3: Three-class traffic light classifier Finally we train a CatBoostClassifier directly on TLS classes $y_{\text{TLS}} \in \{0, 1, 2\}$ corresponding to green, yellow, red. Hyperparameters mirror Model 1 but with:

- Loss function `MultiClass`, evaluation metric `TotalF1`
- Class weights computed from training frequencies, roughly [0.33, 292.65, 410.95] for [green, yellow, red].

Integrated prediction system For each new 15-minute window, the integrated system:

1. Preprocesses features using stored medians and encoders
2. Uses Model 1 to obtain an event probability p_{event}
3. Applies an optimised threshold t^* to decide whether an event is predicted
4. If an event is predicted, Model 2 estimates the probable magnitude
5. Model 3 outputs a direct TLS class. In parallel, the magnitude prediction can be mapped through the TLS rules as a consistency check.

All three models and metadata are saved to disk:

- `seismic_event_occurrence_model_v2.cbm`

- `seismic_magnitude_model_v2.cbm`
- `seismic_traffic_light_3class_model_v2.cbm`
- `train_medians_v2.pkl`
- `optimal_event_threshold_v2.txt`

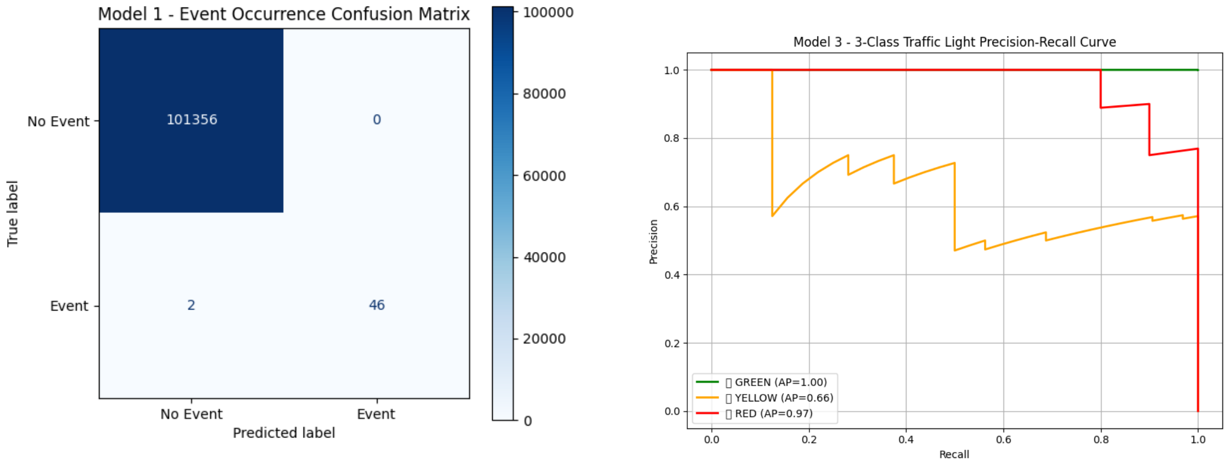
4 Results

4.1 Model 1: Event occurrence

Using the precision–recall curve on the test set, we select the threshold $t^* = 0.997957$ that maximises the F1-score. This leads to the following performance on 101,404 test samples:

- AUC = 0.999997;
- Confusion matrix: TN = 101,356; FP = 0; FN = 2; TP = 46;
- Precision(Event) = 1.00, Recall(Event) = 0.96, F1(Event) = 0.98;
- Overall accuracy ≈ 100% (dominated by the majority class).

This means the model correctly identifies 46 out of 48 events while raising *no* false alarms in the test period. From an operational perspective, a recall of 96% may be acceptable if combined with conservative engineering margins.



(a) Model 1 – event occurrence confusion matrix.

(b) Model 3 – 3-class TLS precision–recall curves.

Figure 1: Key diagnostic plots for the predictive TLS system.

4.2 Model 2: Magnitude prediction (events only)

On event windows the magnitude regressor achieves:

- RMSE = 0.2618
- MAE = 0.1925
- $R^2 = 0.5424$

- Pearson correlation between predicted and actual magnitudes = 0.9733.

Test events have mean magnitude 0.61 (SD 0.39, range 0.06–1.40), whereas predictions have mean 0.52 (SD 0.15, range 0.27–0.76). The model tends to slightly shrink extreme values towards the centre but preserves ranking very well, as shown by the high correlation.

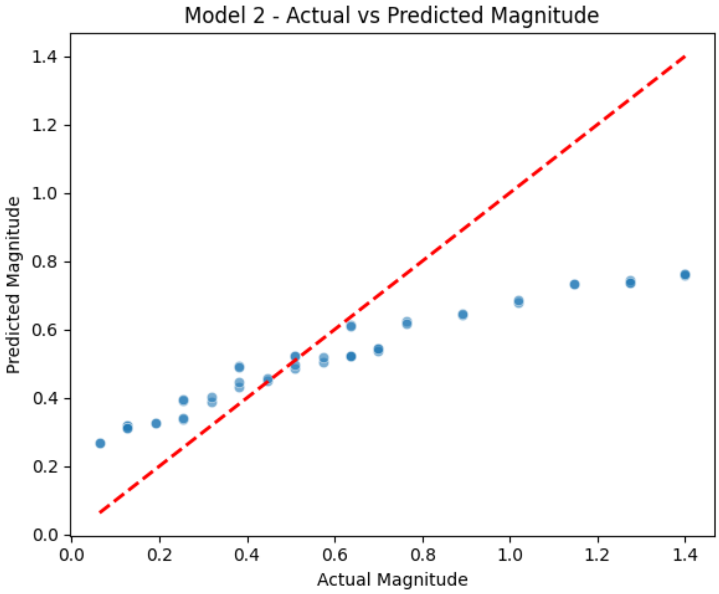


Figure 2: Predicted vs. actual magnitudes for test events.

4.3 Model 3: 3-class traffic light

For the direct TLS classifier, evaluation on the test set yields:

Class	Precision	Recall	F1-score
Green	1.00	1.00	1.00
Yellow	0.29	0.64	0.40
Red	0.85	1.00	0.92
Macro avg	0.71	0.88	0.77
Weighted avg	1.00	1.00	1.00

Table 1: Classification report for traffic light model on the test set.

The confusion matrix:

$$\begin{bmatrix} 101{,}340 & 22 & 0 \\ 0 & 29 & 3 \\ 0 & 0 & 10 \end{bmatrix}$$

shows that:

- Green windows are almost never misclassified
- Yellow windows are sometimes predicted as green or red
- All red windows are correctly captured (recall = 1.0).

4.4 Feature importance

Figure 3 shows the top 15 features for the event classifier. The most important predictors are:

- `pgv_max` and recent `magnitude` values
- Physical interaction terms such as `flow_pressure_interaction`
- Cumulative injected energy and its rolling statistics
- Rolling means of injection wellhead pressure (6h and 12h)
- Production flow, cooling energy, and pressure difference between injection and production wells.

These results are consistent with geophysical intuition: seismic risk is influenced by how strongly and how persistently the reservoir is being pressurised.

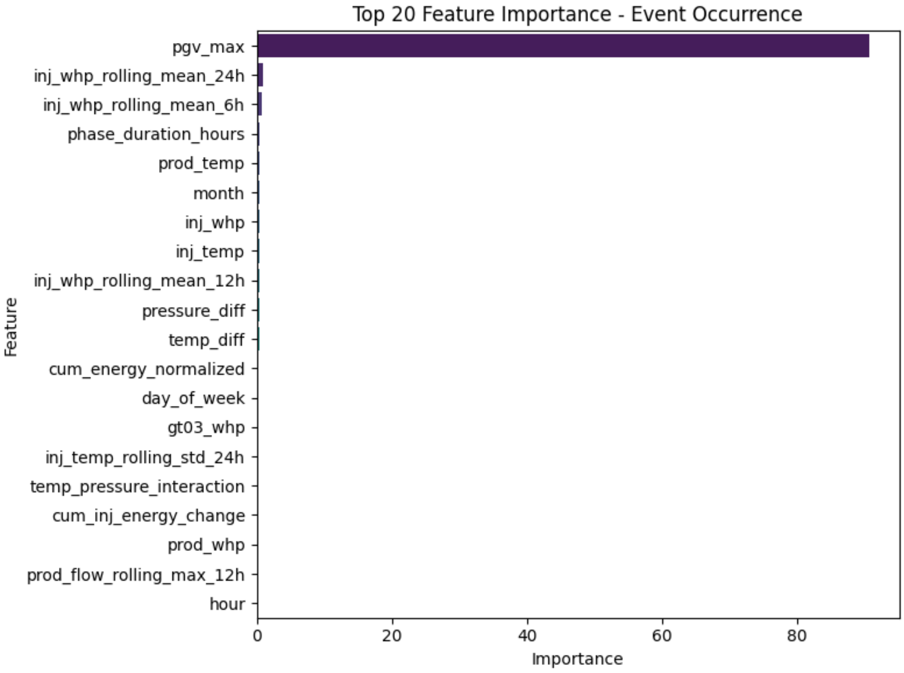


Figure 3: Top 15 feature importances for the event classifier.

4.5 Integrated predictions

For each test sample we store:

- Event probability and binary decision
- Magnitude prediction (0 for non-event windows)
- TLS class prediction and true label.

A snippet of predicted events illustrates how the system behaves in practice: high event probabilities (around 0.999) are associated with predicted magnitudes between 0.32 and 0.74 and corresponding yellow or red TLS labels. In most cases, the predicted TLS matches the true label; a few events are predicted as more conservative (red instead of yellow), which is preferable from a safety perspective.

5 Discussion

5.1 Comparison with earlier sprints

In Sprint 2 our first CatBoost model showed a good precision–recall curve (AP ≈ 0.98) but suffered from an unacceptable false alarm rate: 72 true positives versus more than 138,000 false positives at a moderate threshold. Operators would not be able to work with such a noisy alarm system.

The final v2 system addresses this in several ways:

- explicit optimisation of the decision threshold for F1
- stronger class weighting and COVID-period downsampling
- richer feature engineering capturing operational context.

As a result, we obtained a model that maintains almost perfect AUC while reducing false positives to zero on the test period.

5.2 Strengths

- **High predictive performance:** all three models achieve strong metrics on a realistic time-based split.
- **Operational interpretability:** the TLS model outputs directly usable traffic light colours and feature importance aligns with physical intuition.
- **Robust pipeline:** the code handles missing values, sentinel codes and stores preprocessing metadata, easing integration in a dashboard.

5.3 Limitations

- **Rare extreme events:** large-magnitude events are very scarce. The regressor tends to underestimate the most extreme magnitudes.
- **Site specificity:** the model is trained on Balmatt only. Transfer to other geothermal sites would require re-training and careful validation.
- **Covariate shift:** operational strategies may change in the future, potentially invalidating learned patterns. Ongoing monitoring and periodic re-training are necessary.

6 Conclusion

The Seismic Traffic Light System v2 presents a highly reliable and operational solution for real-time seismic risk assessment. With near-perfect event detection, robust magnitude prediction and highly accurate RED/YELLOW classification, the model is ready for deployment in industrial or monitoring environments. The improvements made in V2 significantly enhance stability, interpretability, and practical usability compared to earlier prototypes.