Machine Learning to Extract Predictive Information from the Pollino Seismic Catalog

Mariagiusi Nicodemo (2114171)
A.Y. 2025/2026

Motivation & Context

SAPIENZA
UNIVERSITÀ DI ROMA

1. Earthquake processes are complex and partially stochastic

2. Short-term prediction of individual events is extremely difficult

3. Changes in seismic activity may contain predictive information

4. Machine learning as an exploratory statistical tool

Objectives

- Assess the presence of predictive information in real seismic data
- Identify which features are most informative for activity regimes
- Evaluate intrinsic limitations due to noise and data scarcity

Working Hypotheses

- Multi-scale temporal aggregation enhances predictive signal
- Magnitude-related features (e.g. b-value) provide secondary information
- Variations in seismic rate carry predictive information

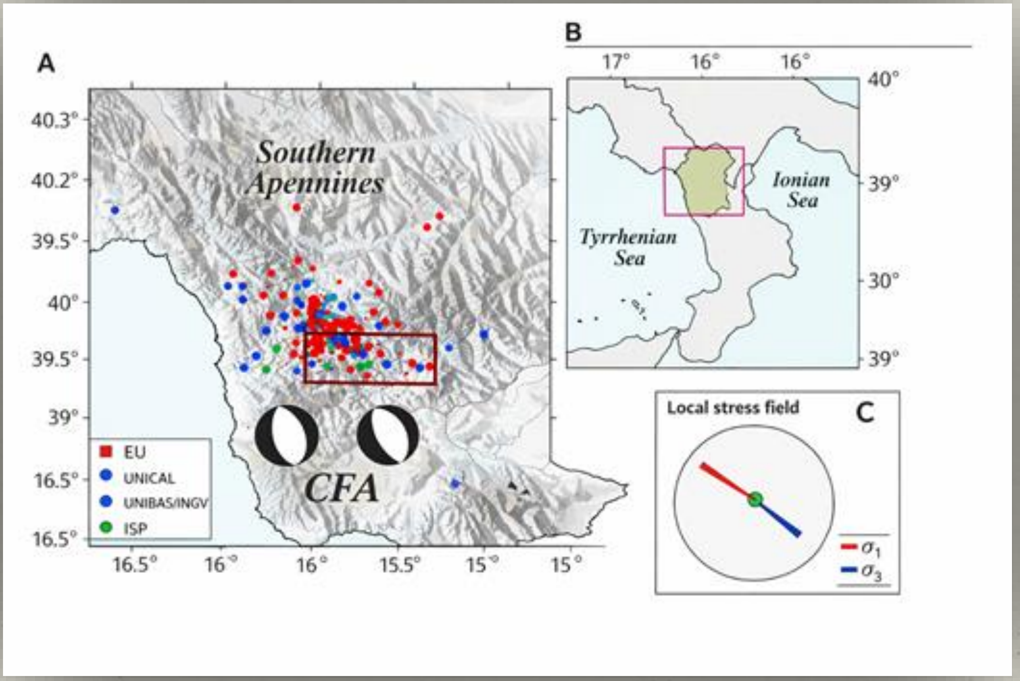
Study area & Data

SAPIENZA
UNIVERSITÀ DI ROMA

- Pollino region, Southern Italy
- Diffuse and swarm-like seismicity
- Homogeneous seismic catalog (late 1990s–present)
- Data source: INGV seismic services

```
client = Client("INGV")
```

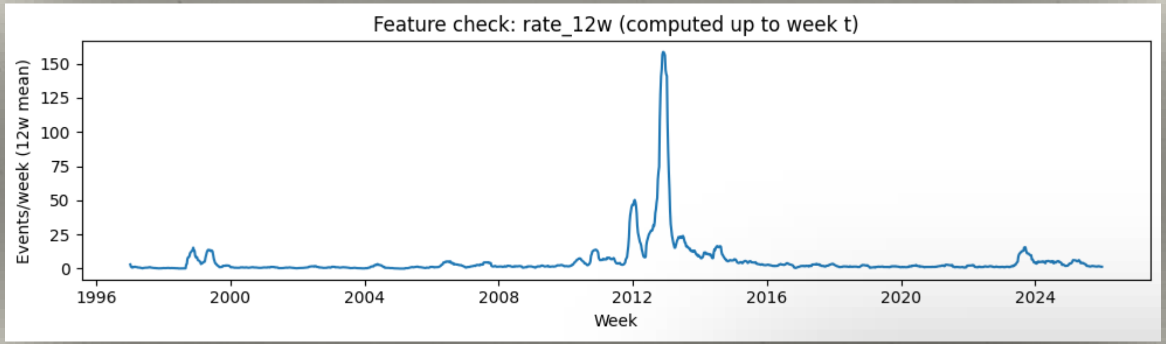
```
# Reference location (near Lauria, Pollino area)
LAURIA_LAT, LAURIA_LON = 40.05, 15.84
RADIUS_KM = 35

# Time slices considered
TIME_SLICES = {
    "1980": ("1980-01-01", "1981-12-31"),
    "1998": ("1997-01-01", "1999-12-31"),
    "modern_2000plus": ("2000-01-01", "2026-12-31"),
}

# Minimum magnitude for catalog retrieval
MINMAG = 1.0
```

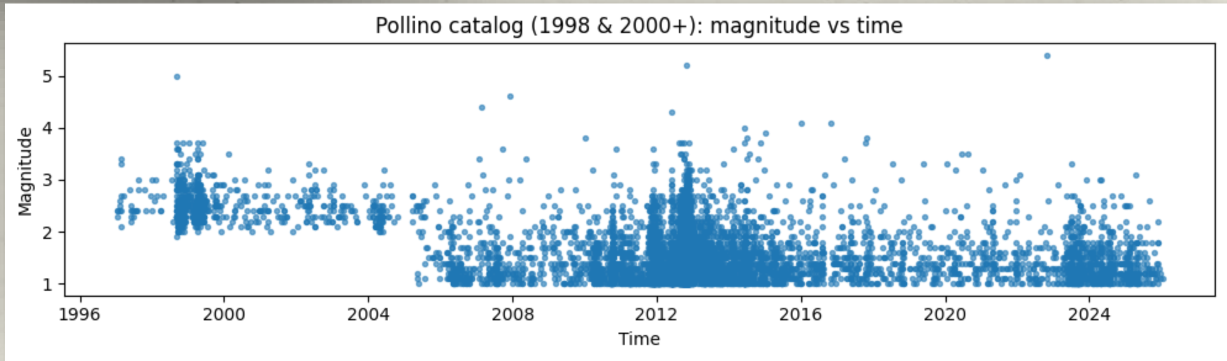


	slice	N_events	t_start	t_end	Mmin	Mmax
0	1980	0	NaT	NaT	NaN	NaN
1	1998	498	1997-01-08 17:01:31.660000+00:00	1999-12-22 11:24:43.930000+00:00	1.9	5.0
2	modern_2000plus	7856	2000-01-08 18:37:56.420000+00:00	2026-01-11 09:30:29.690000+00:00	1.0	5.4

Problem formulation & Temporal aggregation

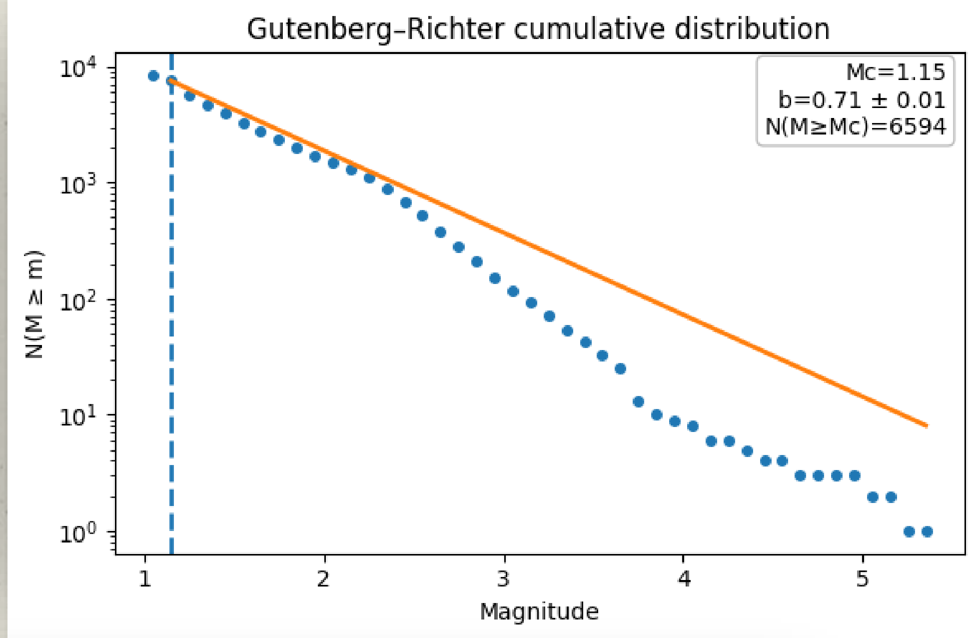
- Event-based seismic catalog aggregated on a weekly basis
- Multi-scale temporal aggregation to capture persistent activity
- Features computed strictly using information up to week *t*

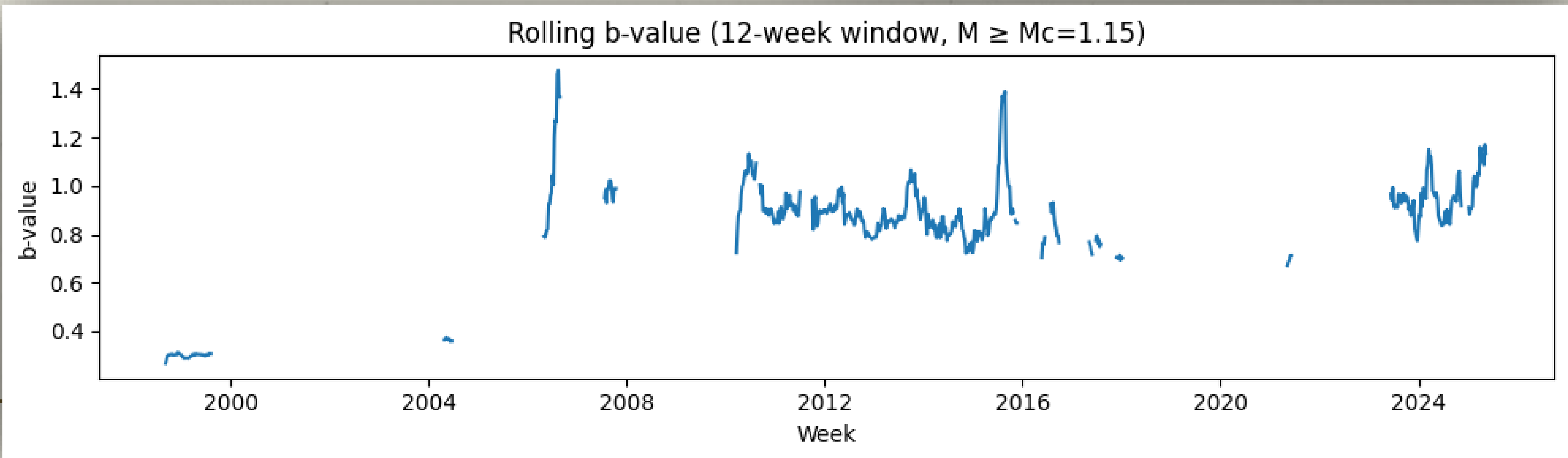




b-value as a physically interpretable feature

The magnitude of completeness Mc is estimated via maximum curvature, and the b-value is computed by maximum likelihood for events above Mc. Deviations at large magnitudes reflect limited statistics, motivating a conservative interpretation of the b-value as a global descriptor.





Target definition & Experimental setup

- Binary classification of seismic activity regimes
- High-activity defined via percentile threshold (75%)
- One-week-ahead prediction: features at t, target at $t+1$
- Time-ordered train/test split (no shuffling)

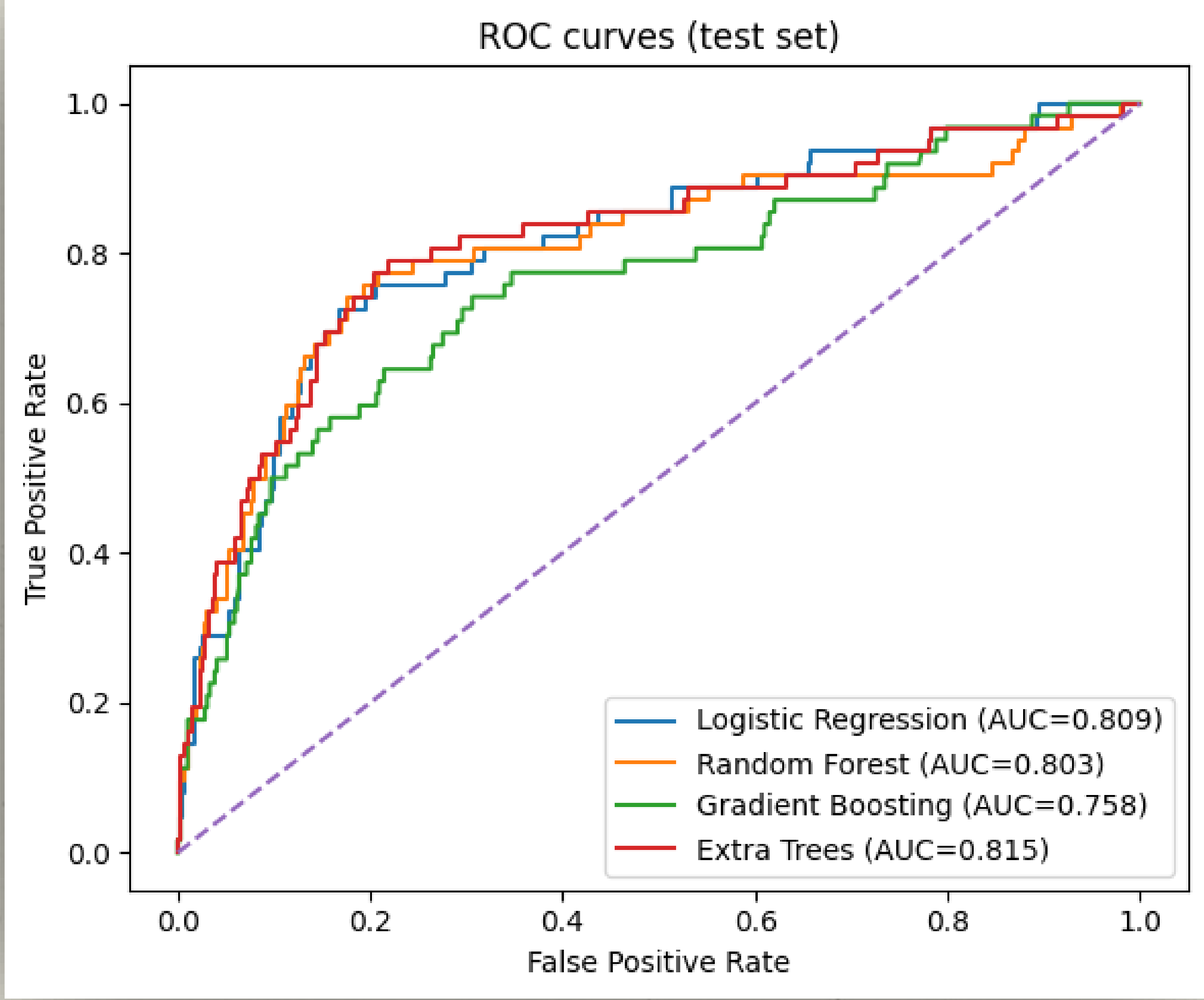
Model evaluation & Metrics

- Multiple models evaluated: Logistic Regression, Random Forest, Gradient Boosting, Extra Trees
- Ensemble methods outperform baseline
- Best-performing model: **Extra Trees**
- Interpretable feature importance

```python
from sklearn.ensemble import ExtraTreesClassifier

et = ExtraTreesClassifier(
    n_estimators=300,
    max_depth=6,
    min_samples_leaf=20,
    class_weight="balanced",
    random_state=42
)

et.fit(X_train, y_train)

y_pred_et = et.predict(X_test)
y_prob_et = et.predict_proba(X_test)[:, 1]

print("Extra Trees")
print(classification_report(y_test, y_pred_et, digits=3))
print("ROC AUC:", roc_auc_score(y_test, y_prob_et))
```

```
Extra Trees
              precision    recall  f1-score   support

           0      0.943     0.847     0.892       392
           1      0.412     0.677     0.512        62

    accuracy                          0.824       454
   macro avg      0.677     0.762     0.702       454
weighted avg      0.871     0.824     0.841       454

ROC AUC: 0.8149687294272547
```

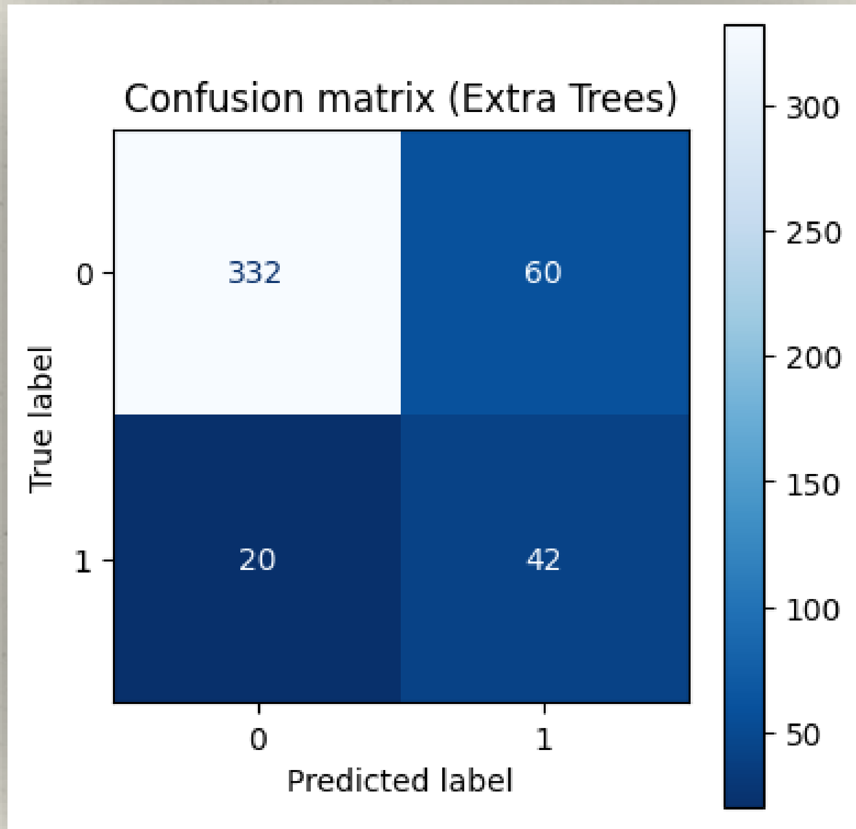
Model Performance: ROC Curves (Test Set)



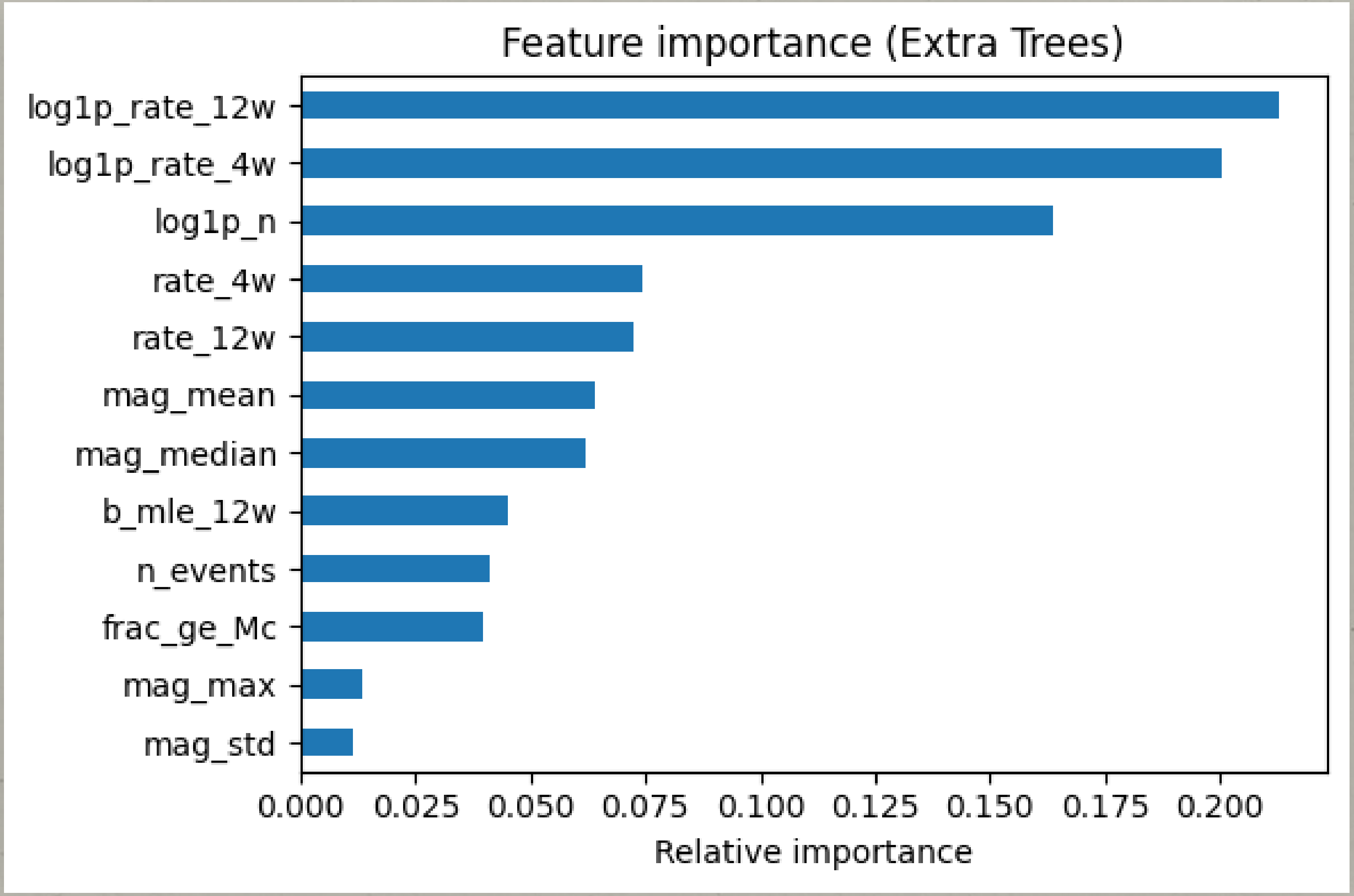
Classification Behavior of the Best Model

SAPIENZA
UNIVERSITÀ DI ROMA

The confusion matrix highlights the ability of the Extra Trees model to identify high-activity weeks while exhibiting an expected trade-off between recall and false positives in an imbalanced setting.



Feature Importance: what drives the predictions?



Robustness checks and Andavanced analysis

Percentile	High-Activity Threshold	Test Class Balance (High Activity)	ROC AUC	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
70th	3.0 events/week	23% (98/438)	**0.760**	0.504	0.684	0.580
75th	4.0 events/week	14% (60/438)	**0.820**	0.409	0.726	0.523
80th	5.0 events/week	10% (45/438)	**0.807**	0.340	0.717	0.462

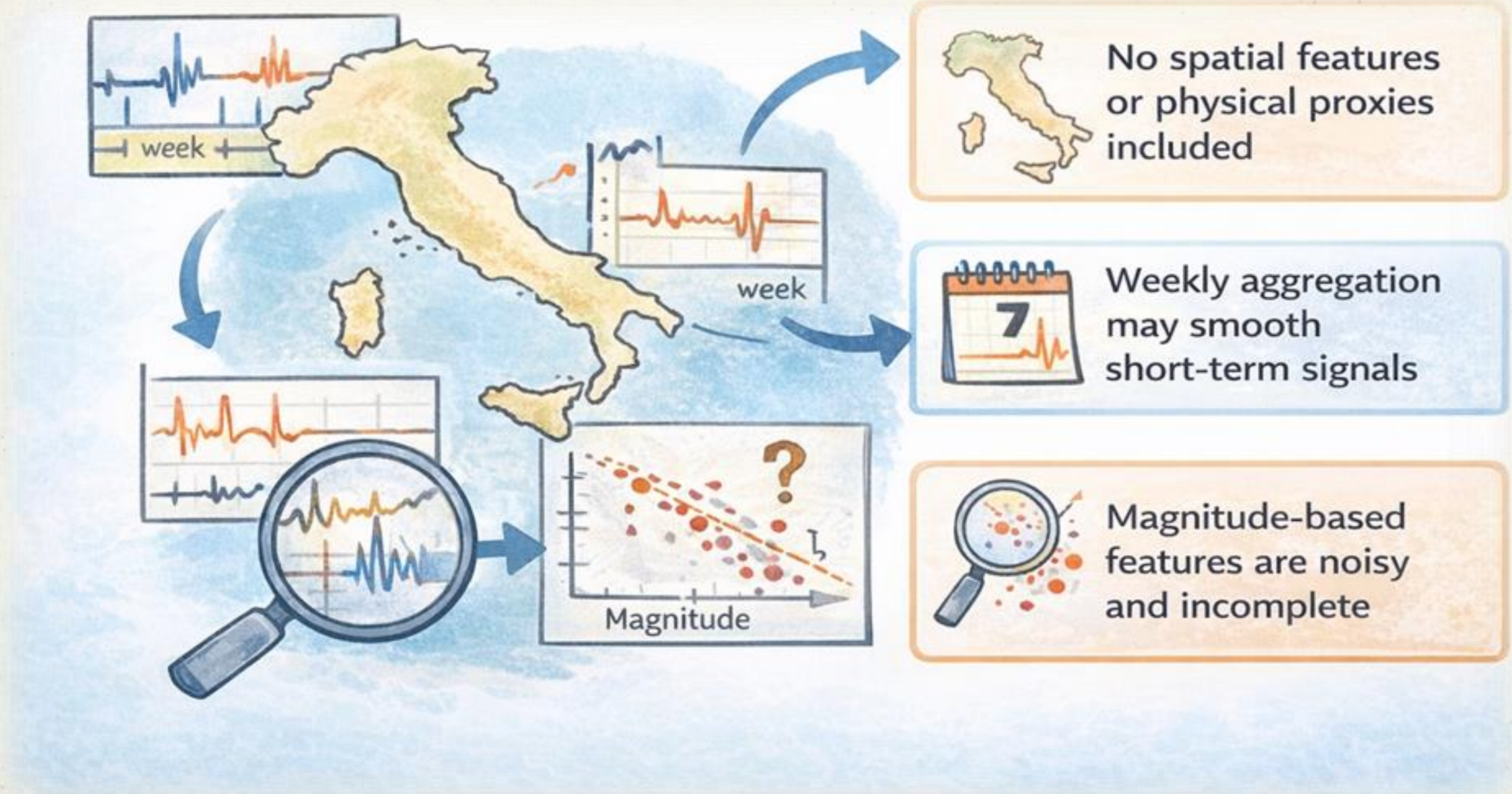
Interpretation

The observed performance plateau suggests that increasing model complexity alone does not substantially improve predictive skill.

The dominant role of rate-based features indicates that the extractable predictive information is primarily related to persistent changes in seismic activity rather than short-term fluctuations.

These findings are consistent with the partially stochastic nature of the seismic process and with intrinsic limitations in the information content of historical catalogs.

Limitations



Conclutions

The analysis is based exclusively on temporally aggregated seismic information and does not incorporate spatial features or additional physical proxies.
The use of weekly time windows may smooth short-lived precursory signals, potentially limiting sensitivity to rapid changes in seismic activity.
Furthermore, magnitude-based features, including b-value estimates, are affected by noise and limited sample size within short temporal windows.

- Extend the proposed framework to different seismic regions and tectonic settings to assess the generality of the results.
- Incorporate additional spatial features and physically motivated proxies, such as stress-related indicators, to enrich the feature set.
- Explore alternative temporal scales and physics-informed machine learning approaches to further improve the interpretability of regime-based predictions.

THANK YOU FOR YOUR ATTENTION