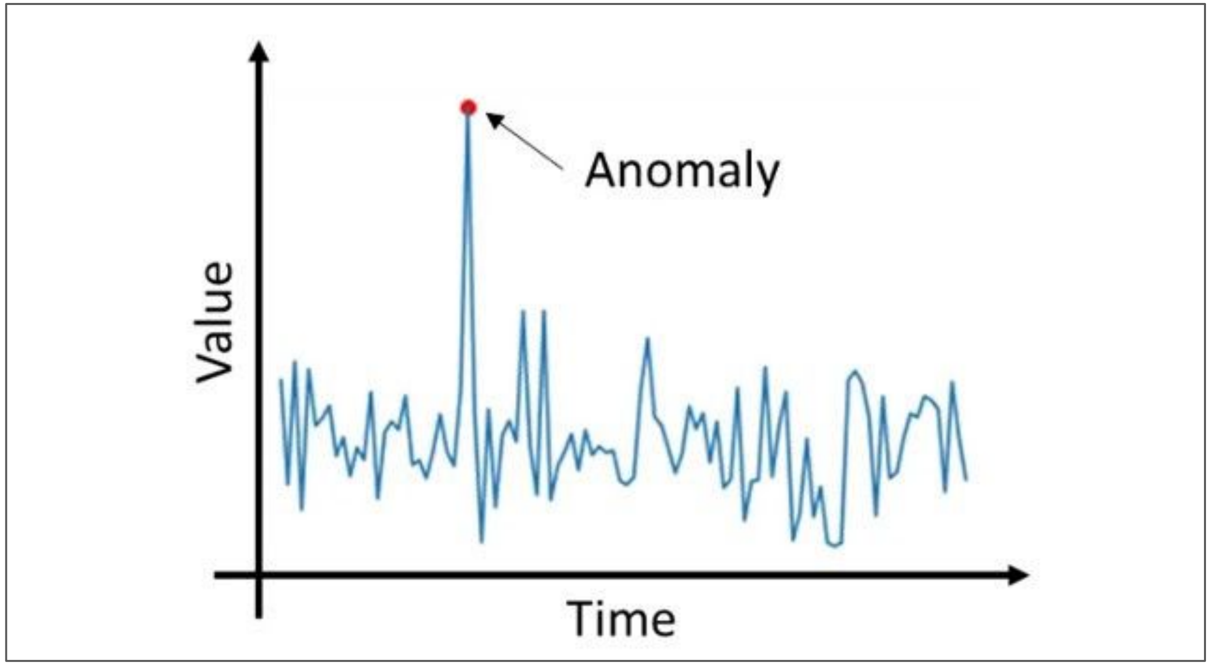
Using Machine Learning Algorithms to detect noise features in ground magnetic data

Exploring the effectiveness of machine learning predictive algorithms on classification tasks.



Rami Abou-Shamalah, MSc, Data Scientist

Contents

01 Background:

Problem overview & the Approach

02 Insights:

1) Exploratory Analysis
2) Model Performance

03 Summary:

Recommendations & Limitations

SGL

01 Background:

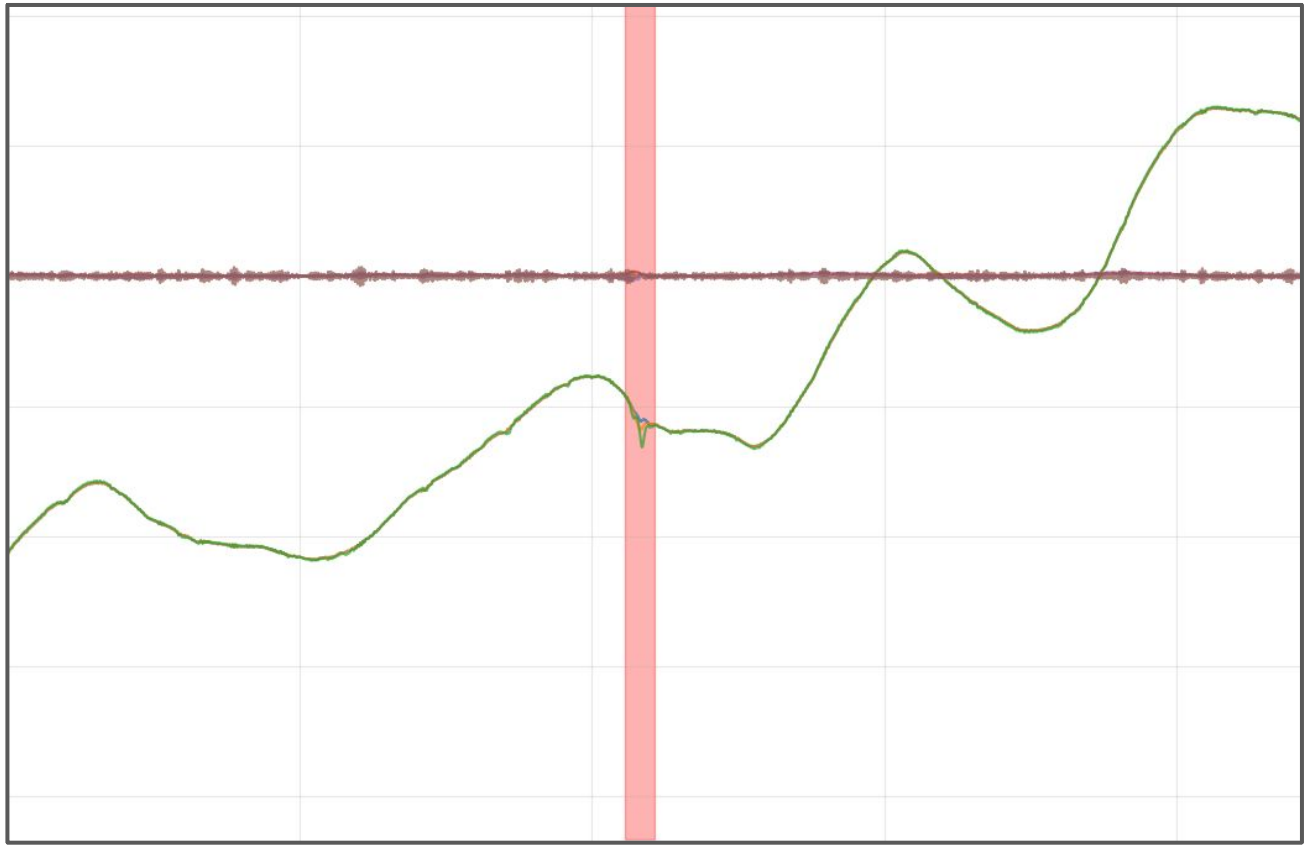
Problem & Approach

Background

- Identifying cultural artifacts has been done manually for decades

- Process of sifting through the time series data and look for sudden changes.

- Humans are prone to error



Approach

Solution

Implement a predictive model to detect cultural artifacts with higher degree accuracy than humans, such that the task can be automated.

How?

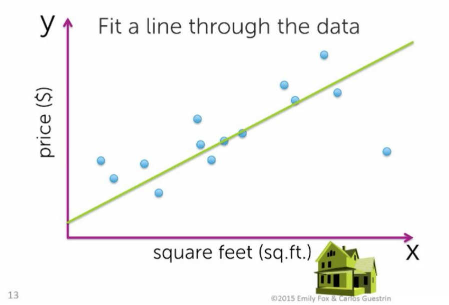
Sanders Geophysics has accumulated more than 5 decades worth of training data. For this particular project, there is 20 years worth of training data.

For the purpose of this project, as a concept of proof, only 2 projects, spanning 2 years, was used as training data.

Modelling Strategy:

Explicit algorithm doesn’t work.

We're employing logistic regression, a basic but powerful statistical tool that excels at binary classification predictions - in our case, whether a time period is cultural or not.



Evaluation Method:

we'll use **F1-Score**, which balances two key metrics to account for errors:

- Type I Error (False Positive): Predicting a noise segment, when it does not exist
- Type II Error (False Negative): Predicting something, when it does not exist

02 Insights:

1) Exploratory Analysis
2) Model Performance

SGL

Exploratory Data Analysis

What characteristics does the anomaly exhibit?

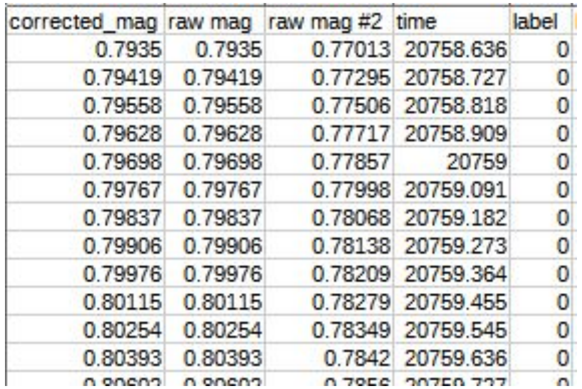
Data

2 Projects were used for the training data:

- Project #1 consisted of 657 files
- Project #2 consisted of 445 files

What is in the ‘file’?

Acquired > Databases > Fetched from Database > Preprocessed



corrected_mag	raw mag	raw mag #2	time	label
0.7935	0.7935	0.77013	20758.636	0
0.79419	0.79419	0.77295	20758.727	0
0.79558	0.79558	0.77506	20758.818	0
0.79628	0.79628	0.77717	20758.909	0
0.79698	0.79698	0.77857	20759	0
0.79767	0.79767	0.77998	20759.091	0
0.79837	0.79837	0.78068	20759.182	0
0.79906	0.79906	0.78138	20759.273	0
0.79976	0.79976	0.78209	20759.364	0
0.80115	0.80115	0.78279	20759.455	0
0.80254	0.80254	0.78349	20759.545	0
0.80393	0.80393	0.7842	20759.636	0
0.80602	0.80602	0.7856	20759.727	0

- On Average 250 seconds for each file,
- At a frequency of 10 data points per second, each file on average is 2500 rows long
- Each file represents a portion of the time that the airplanes were online, so they’ve been properly edited.

Feature Engineering

From: 'raw mag', ‘raw mag #2’:

To: 'Rolling STD', '1st derivative', '2nd derivative', 'mag_difference', 'rolling_mean'

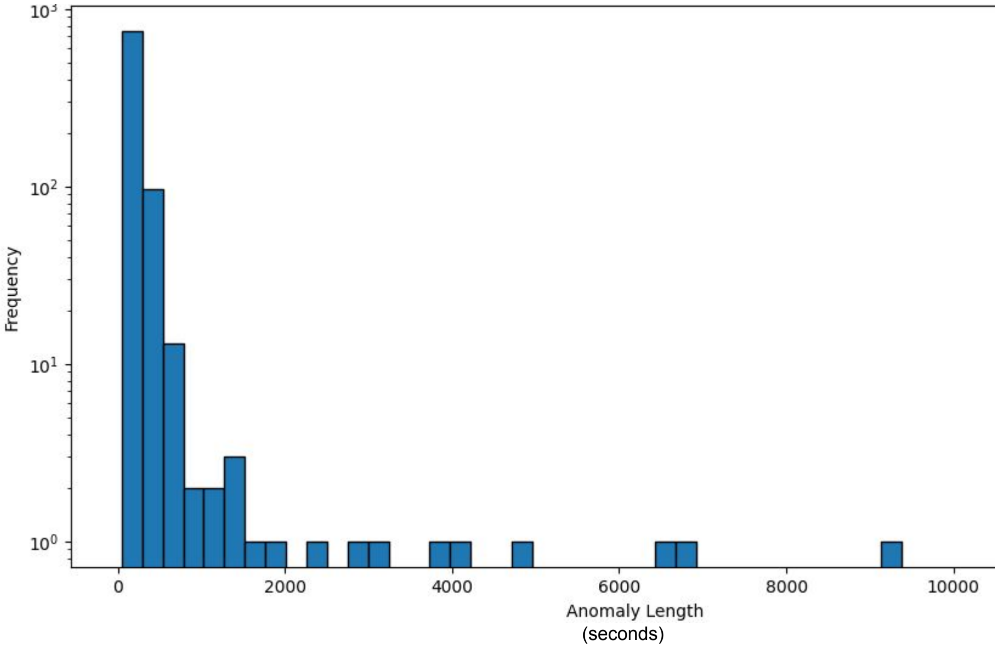
The anomaly

- Imbalanced dataset, the positive class, or 1 label only constitutes 5% of the positive class. Very typical for anomaly classification in time series data
- Length of the anomaly takes a somewhat left-skewed binomial distribution, but dependent on project

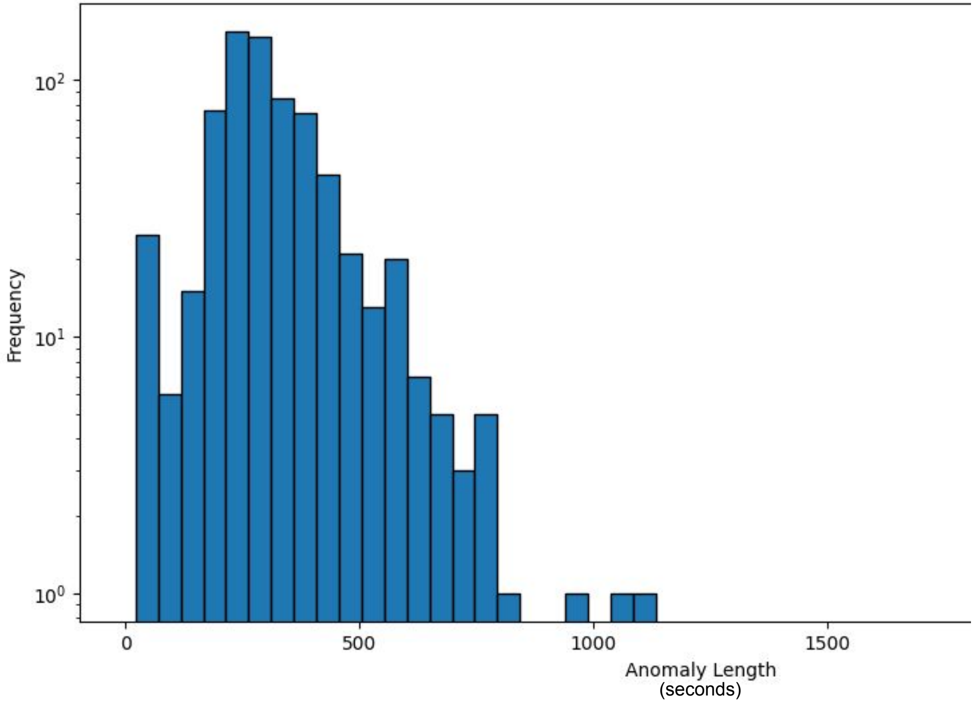
Exploratory Data Analysis

What characteristics does the anomaly exhibit?

Project #1



Project #2



Model Performance

How our models predicts anomalies

Model Performance:

- **Based on a test sample of 20 datasets, our best results:**

True Positives: 37
False Positives: 10
False Negatives: 0

Error Analysis:

- **False Positives:**

Falsely assuming an anomaly occurs when it does not

Occurred 10 times

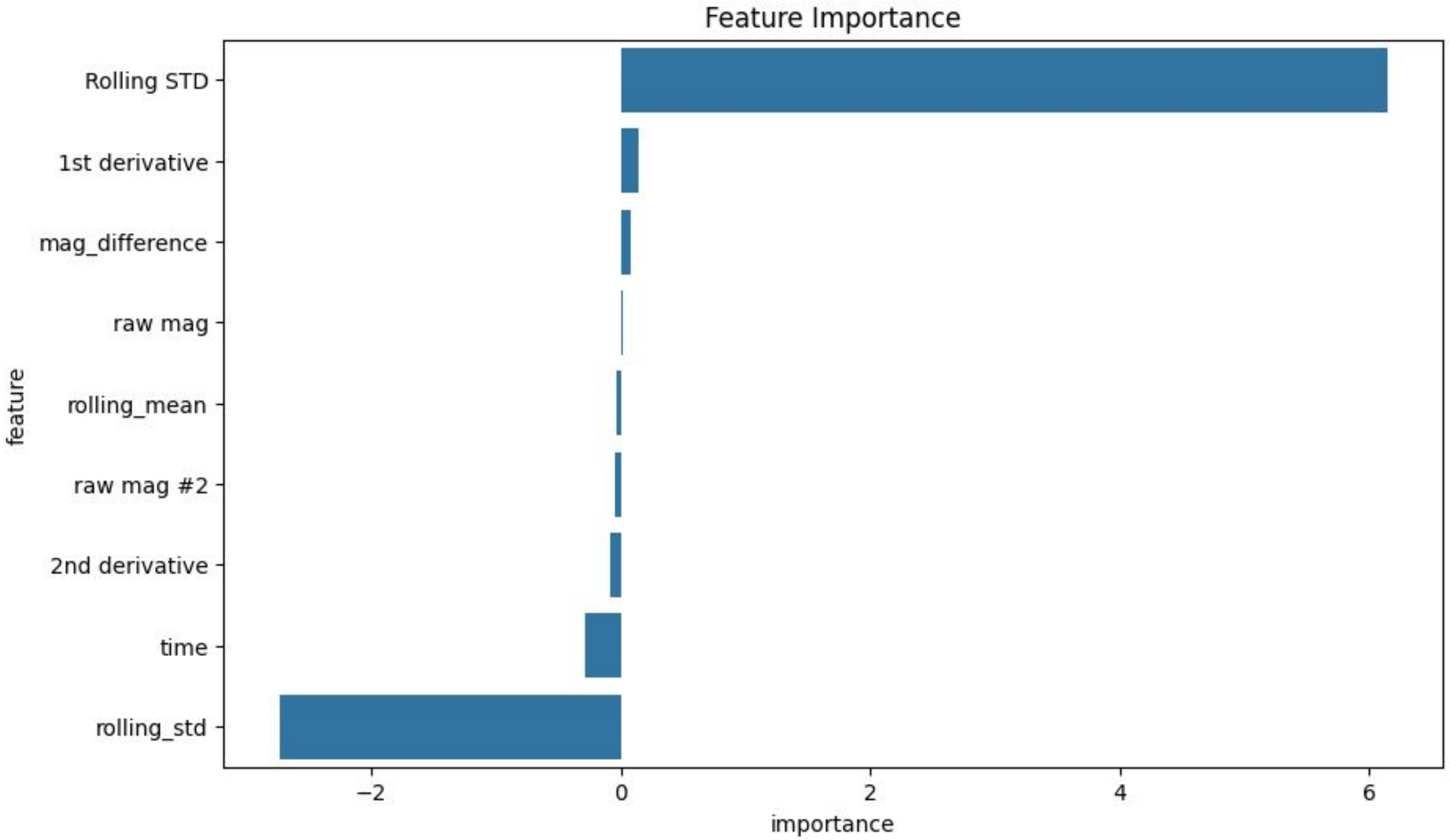
- **False Negative:**

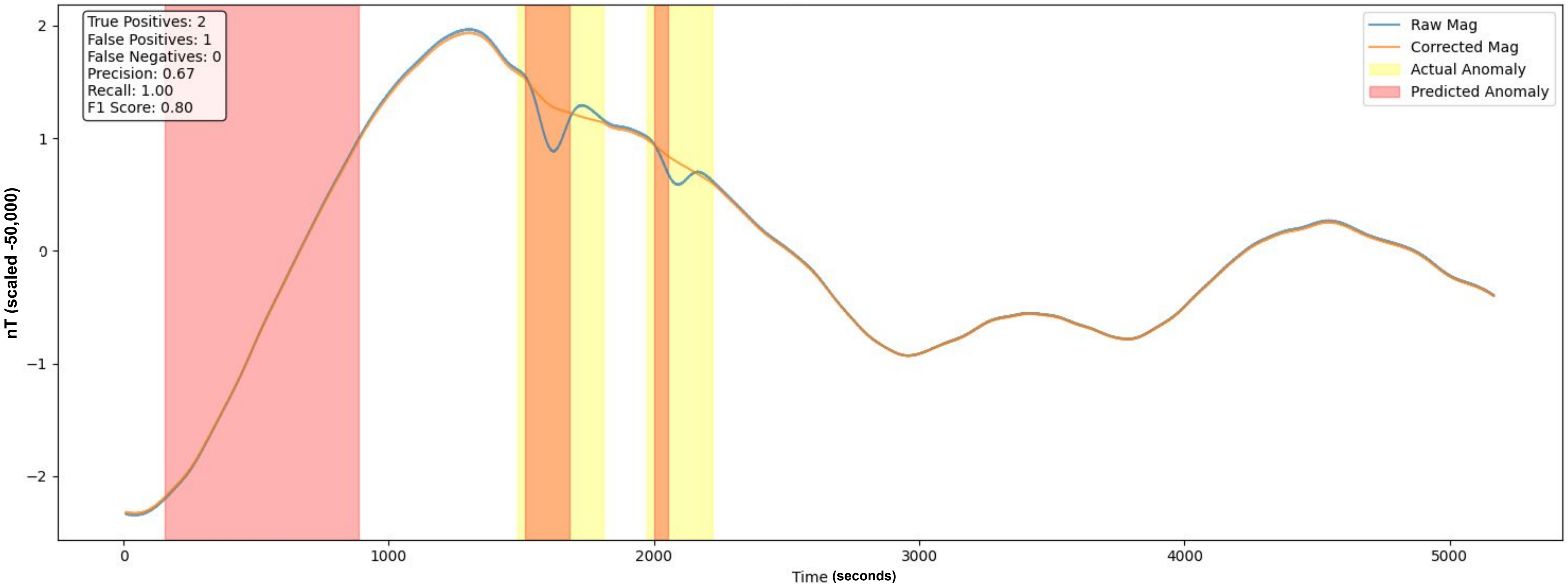
Failing to pick up an anomaly.

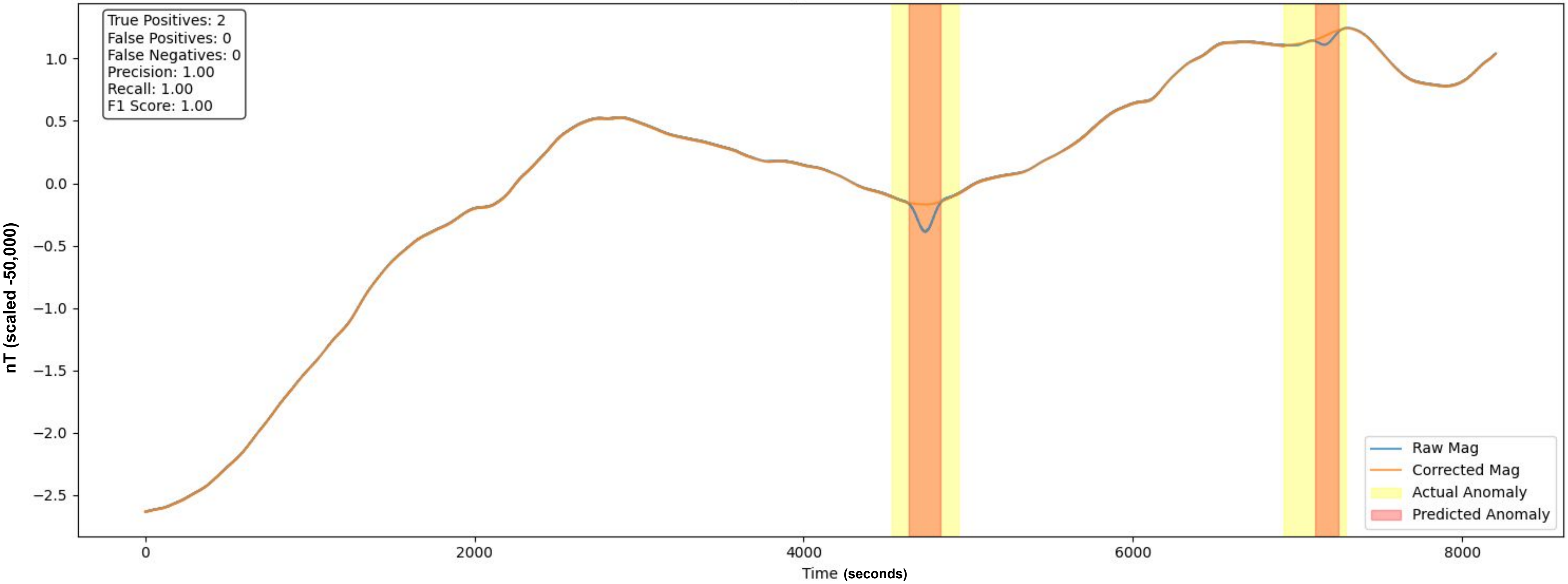
Did not occur.

Model Performance

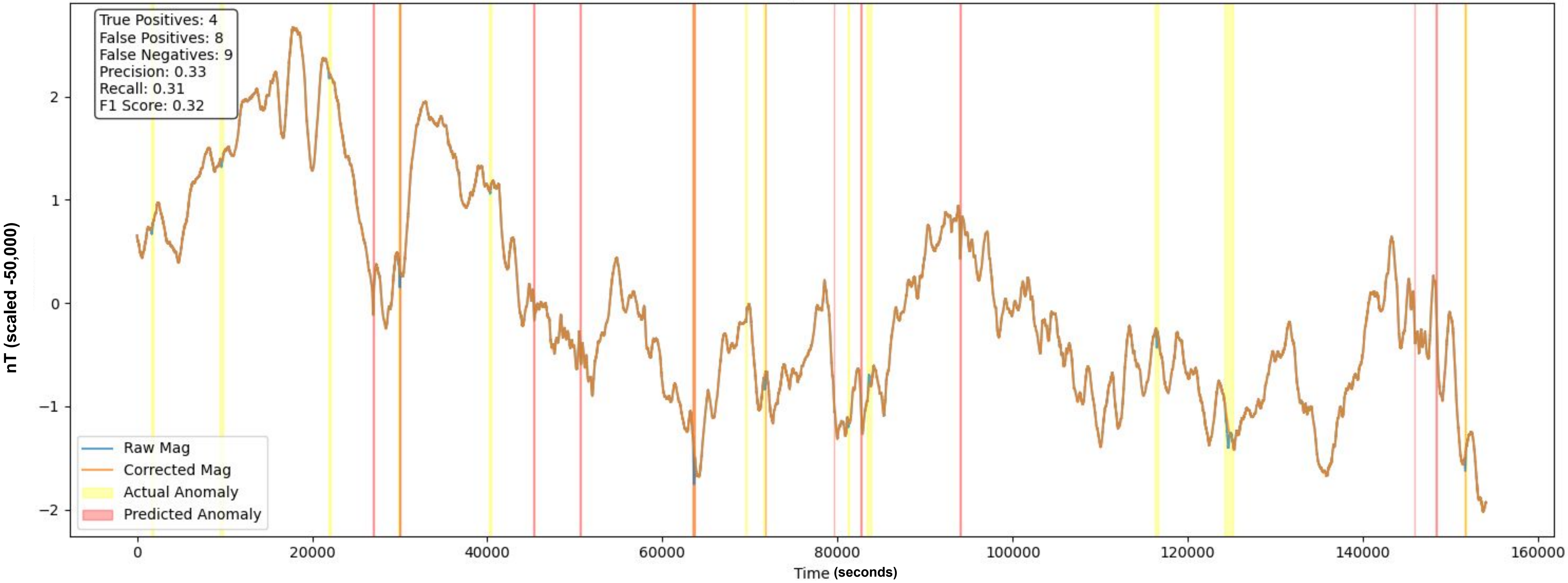
Model performance analysis

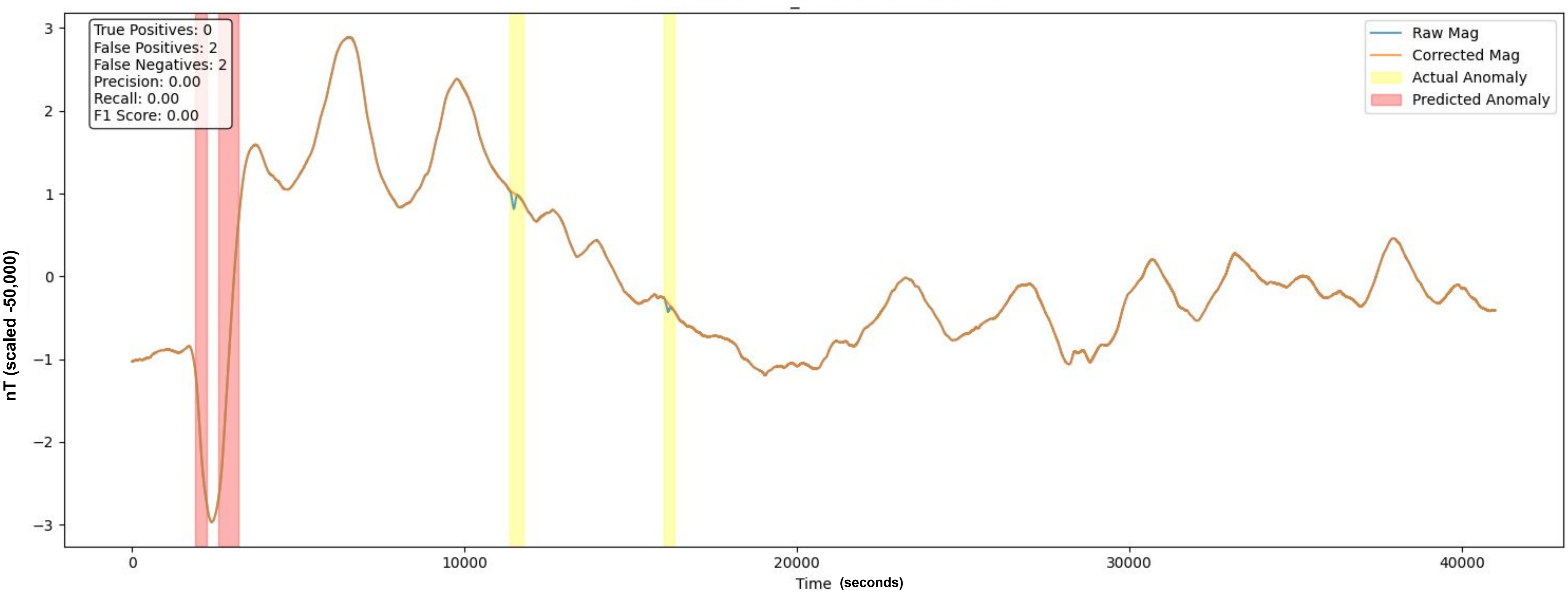



True Positives: 2
False Positives: 1
False Negatives: 0
Precision: 0.67
Recall: 1.00
F1 Score: 0.80
Raw Mag
Corrected Mag
Actual Anomaly
Predicted Anomaly
nT (scaled -50,000)
2
1
0
−1
−2
0
1000
2000
3000
4000
5000
Time (seconds)


True Positives: 2
False Positives: 0
False Negatives: 0
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
nT (scaled -50,000)
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0
2000
4000
6000
8000
Time (seconds)
Raw Mag
Corrected Mag
Actual Anomaly
Predicted Anomaly

SGL


True Positives: 4
False Positives: 8
False Negatives: 9
Precision: 0.33
Recall: 0.31
F1 Score: 0.32
Raw Mag
Corrected Mag
Actual Anomaly
Predicted Anomaly
nT (scaled -50,000)
2
1
0
−1
−2
0
20000
40000
60000
80000
100000
120000
140000
160000
Time (seconds)


True Positives: 0
False Positives: 2
False Negatives: 2
Precision: 0.00
Recall: 0.00
F1 Score: 0.00
Raw Mag
Corrected Mag
Actual Anomaly
Predicted Anomaly
nT (scaled -50,000)
3
2
1
0
−1
−2
−3
0
10000
20000
30000
40000
Time (seconds)

03 Recommendations & Limitations:

Recommendations & Limitations

More data!

The Culprit Revealed:

- Prediction rates increased when using the combined model versus a single model. The obvious conclusion is that, more data will yield better predictions
- per the feature importance graph: ML algorithm is more effective than an explicit algorithm which tries to account for all the nuances

The Impact:

- Automating tasks saves the company time and money by freeing human resources.

Strategic Recommendations:

- Try the program on 4 other more projects and analyze the predictions. If increasing accuracy, collect more data
- Once achieving 95%+ accuracy, we can incorporate the program into the processing stream and allow humans to act as supervisors of the model.
- Include data sets where we have two ground stations in different locations recording simultaneously. We do this now and it is very helpful to distinguish signal from noise.