STA2453 Progress Report #1

Isaac Baguisa

2025-03-05

Introduction

This project aims to detect stellar flares, and score the performance of the model to identify these events. Stellar flares are intense bursts of energy, emitted from a star that are thought to be caused by magnetic reconnection. They are usually indicated by a sudden increase in brightness, followed by a slower decay. Detecting and analyzing these flares is crucial for understanding stellar behaviour and its potential effects on its environments. The following questions are aimed to be answered in this project: Can non-parametric models effectively detect stellar flares from brightness time series data? How do these models score based on simulations? How can we validate the quality of simulations for testing the models?

Data

The dataset contains time series measurements of stellar brightness and associated errors from three stars (TIC 0131799991, TIC 031381302, and TIC 129646813) from the Transiting Exoplanet Survey Satellite (TESS). The key variables include: time, the observation timestamp in days (Barycentric Julian Date). PDCSAP_FLUX, the photometric flux after pre-search data conditioning, indicating the star's observed brightness (electrons per second). PDCSAP_FLUX_ERR, the error associated with each flux measurement (electrons per second).

The data contains gaps due to the satellite turning off. The Kalman filter works well for imputation becauses it uses an adaptive process that estimates the state of a linear dynamic system from a series of noisy measurements. It can effectively track the sudden increases in brightness and subsequent decays, even when there are gaps in the data. This makes it robust missing not at random (MNAR) data, as it does not require the missing data to be unrelated to the unobserved data values.

Methods

The Gaussian Process Regression (GPR) model is used to model the underlying trend of a star's light curve, with the goal of detecting stellar flares. The model is trained on one star to predict the flux based on time as the independent variable. The GPR model uses the Radial Basis Function (RBF) kernel, which captures non-linear relationships between time and flux. The model predicts the expected flux values and their associated uncertainty, and flare events are detected by identifying large residuals. These large residuals suggest anomalies in the data, which are flagged as flares.

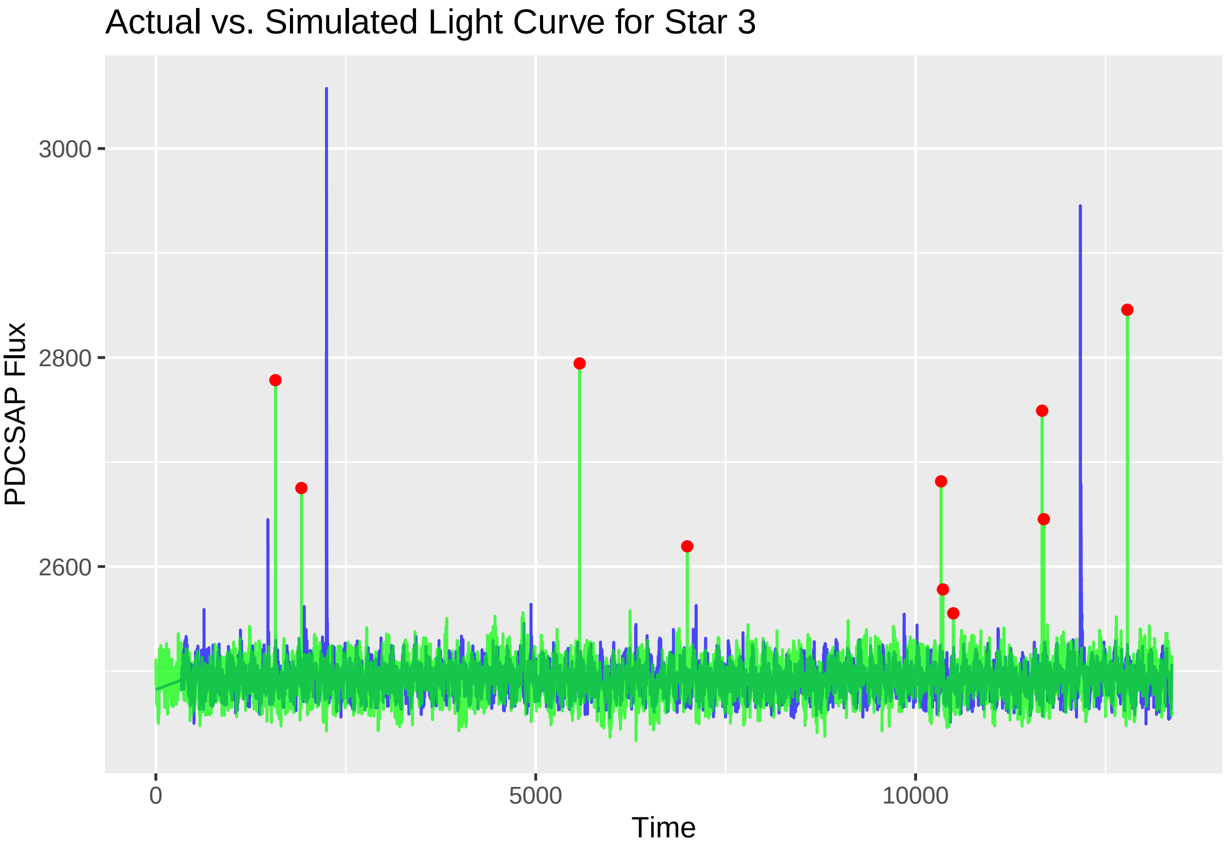
The Poisson Process Regression (PPR) model works by modeling the star's flux as a function of time. The model is trained on one star to predict the flux based on time as the independent variable. This is appropriate for count data where the flux can be treated as event counts over time. The log link function models the logarithm of the expected flux, and the Poisson distribution captures the randomness of flare occurrences. The model detects flares as deviations from the expected flux, with large residuals flagged as potential flares.

The One-Class Support Vector Machine (OCSVM) models the normal behavior of the star's flux using a one-class classification approach. It learns the distribution of the normal flux data and identifies outliers (flares). Using a RBF kernel, the OCSVM maps the data into a higher-dimensional space, allowing it to capture non-linear relationships between flux and time. The model defines a boundary around the normal data points, and any points outside this boundary are considered flares.

Simulation using Seasonal ARIMA model

The simulation is created by fitting a SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model to the star's flux data to capture seasonal patterns. It simulates a new light curve based on the fitted model, preserving the underlying seasonal trends. To model flare events, synthetic flares are injected by adding random spikes at 10 randomly chosen time points, which are labeled as 1 in the simulation dataset.

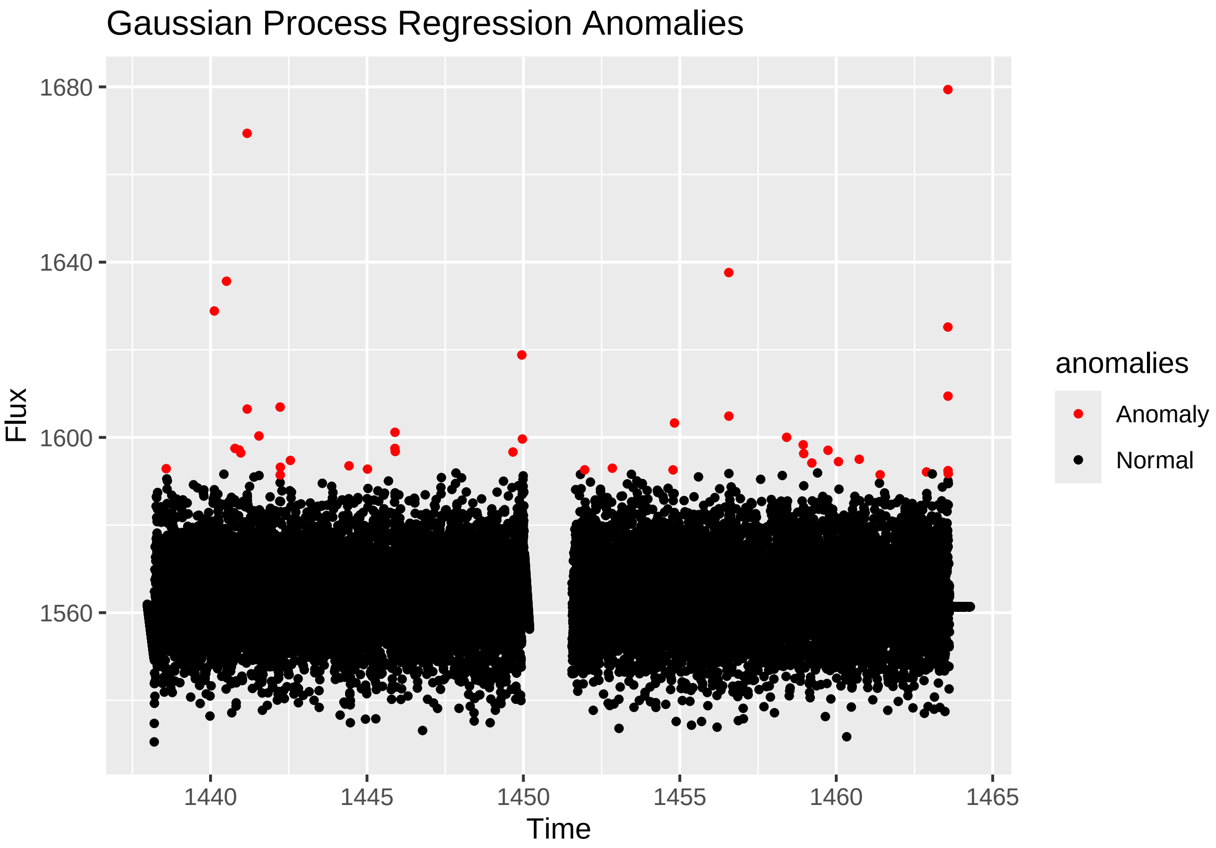
Below is the comparison between the actual flux data and the simulated data, highlighting the flare events in red. The output is a data frame containing the simulated flux values, time, and flare labels. This simulation provides a controlled dataset for testing and validating flare detection models to evaluate their ability to accurately identify flares in the real stellar data.

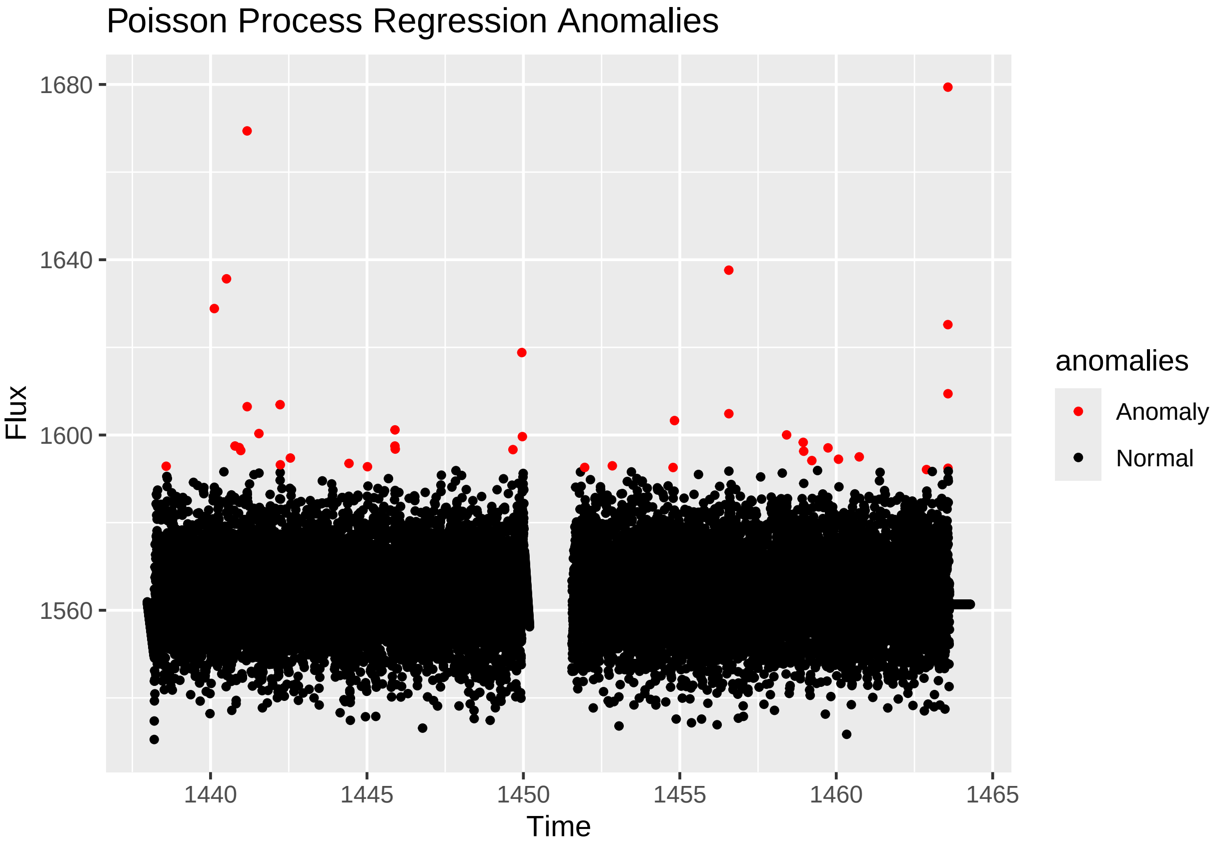


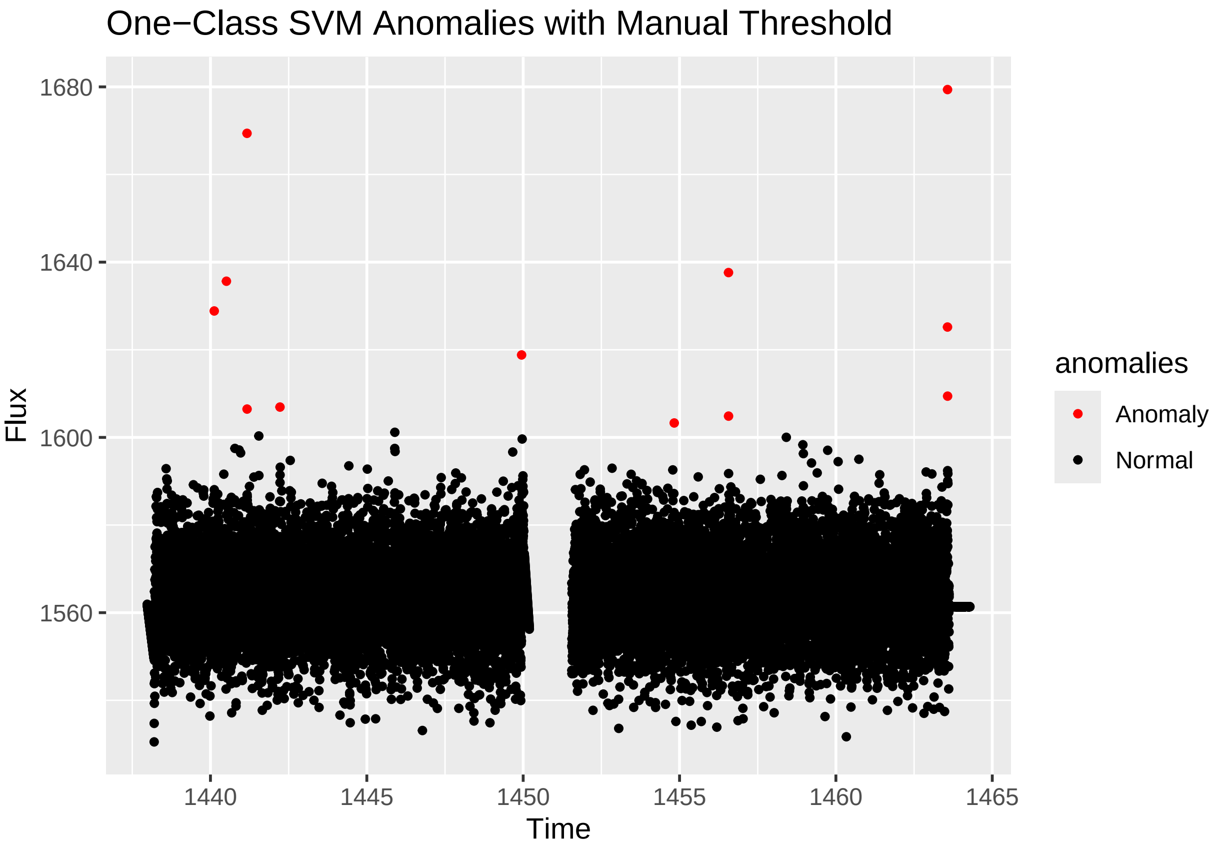
Results and Next Steps

I've tested the GPR and PPR models by predicting flux values from the observed flux values. The anomaly threshold is defined a residual that exceeds 3 standard deviations above the mean residual. Based on this threshold, residuals that are greater than the defined threshold are flagged as an anomaly, otherwise they are labeled as normal.

Testing the GPR, PPR, and One Class SVM models on the simulation I get a predictive accuracy of 0.99+, and detection rate of 0 or 1. The problem with all of my models is that it is predicting "no flare" almost every time. Moving forward I will be tuning my parameters and thresholds for each model and simulation to increase the true positive rate in detecting flares.


Gaussian Process Regression Anomalies
Flux
1680
1640
1600
1560
Time
1440
1445
1450
1455
1460
1465
anomalies
● Anomaly
● Normal


Poisson Process Regression Anomalies
Flux
1680
1640
1600
1560
Time
1440
1445
1450
1455
1460
1465
anomalies
● Anomaly
● Normal



Conclusion

So far I have completed an exploratory data analysis to visualize the trends and distribution of flux, and to better understand the occurrence of flares. In my model classification methods I've identified several non-parametric methods to detect stellar flares, GPR, PPR, and OCSVM. Lastly, to score and compare each model I've started (but still tuning) a SARIMA model to mimic the brightness of a star, with injected with flares.