Stellar Flare Detection Final Report

Isaac Baguisa

2025-04-16

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?

Data Description and Exploratory Data Analysis

The dataset consists of time series data from the Transiting Exoplanet Survey Satellite (TESS) mission. Three stars (TIC 0131799991, TIC 031381302, and TIC 129646813) are used for analysis. The primary task involves detecting spikes in brightness, which indicate flares. Note that these flares are unlabelled, therefore we do not have true indication of when a flare occurs. Missing data occurs due to satellites turning off, and are not missing at random. To account for these gaps, imputation methods will be used. The most important variables for this analysis are time, pdcsap_flux (Pre-Search Data Conditioning Simple Aperture Photometry - PDCSAP Flux), and pdcsap_flux_err, which represent the observation time, corrected photometric flux, and the associated uncertainty respectively.

Visualizing flux over time allows us to detect patterns or irregularities such as sudden spikes that could denote flares. Sudden peaks in these plots may indicate flares, while consistent patterns could suggest underlying stellar processes or cycles. For TIC 031381302, the flux shows consistent variability with occasional sharp spikes, suggesting frequent, moderate fluctuations in brightness which might be indicative of stellar flares. TIC 129646813 exhibits a relatively stable flux level with very distinct, sharp peaks likely pointing to infrequent but significant stellar flares.

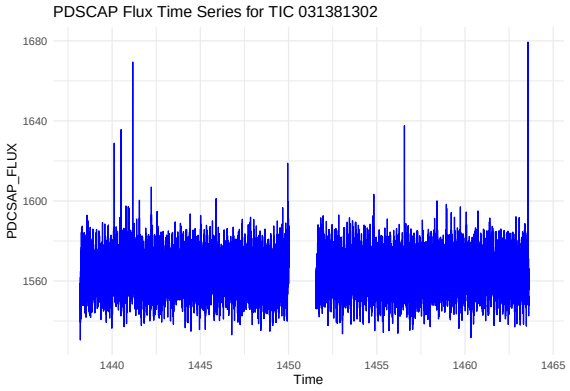

Figure 1: PDCSAP Flux for TIC 031381302

The data contains gaps due to the satellite turning off, therefore imputing missing values and comparing them with actual observations allows us to validate the imputation method and understand its impact on potential flare detection. Effective imputation that closely follows the true data patterns is important for analysis to ensure that potential flares are not missed during gaps in data collection.

Table 1: Number of Missing Values in PDCSAP_FLUX by Star

Star	Available	Missing
TIC 031381302	17033	686
TIC 129646813	18188	91
TIC 0131799991	13034	338

The imputation method used is the Kalman filtering algorithm. The Kalman filter works well 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.

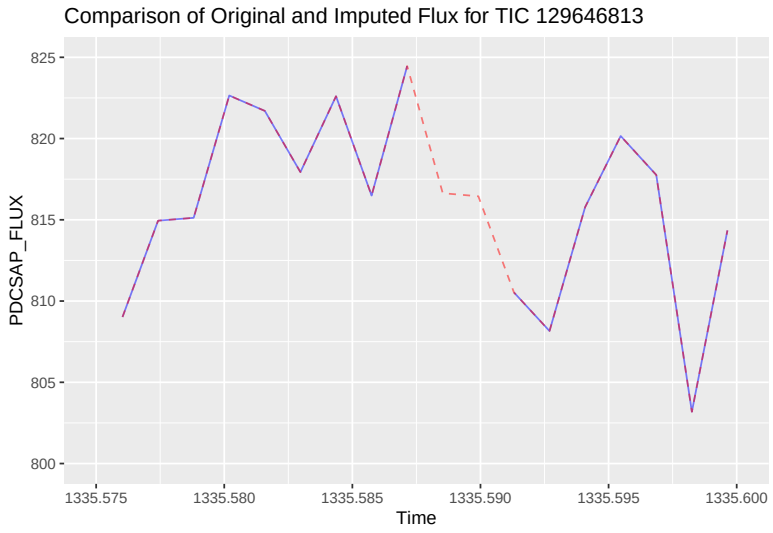

Figure 2: Comparison of Orginal and Imputed Flux for TIC 129646813

General Distribution of PDCSAP Flux

The histogram plots provide a better understanding of the flux distribution, highlighting typical brightness levels. By examining the distribution's skewness and kurtosis, we can infer the frequency and intensity of anomalous brightness events.

The long tail towards the higher flux values suggests the presence of occasional but significant spikes in brightness, which are characteristics of stellar flares. The skewness could imply that while flares are relatively rare compared to normal stellar activity, they are significant when they occur, as shown by the tail extending to higher values. All three histograms in figure 3 suggest relatively stable stellar brightness with a consistent flux level dominating the observations. The lack of wide variance across all three histograms indicates stable conditions with occasional bursts of higher energy. Additionally, TIC 0131799991 and TIC 129646813 have extreme peaks and shows minimal spread, indicating that most flux measurements are very close to a typical value with very few variations. TIC 031381302 also has an extreme peak, but with a larger variation than the other two stars.

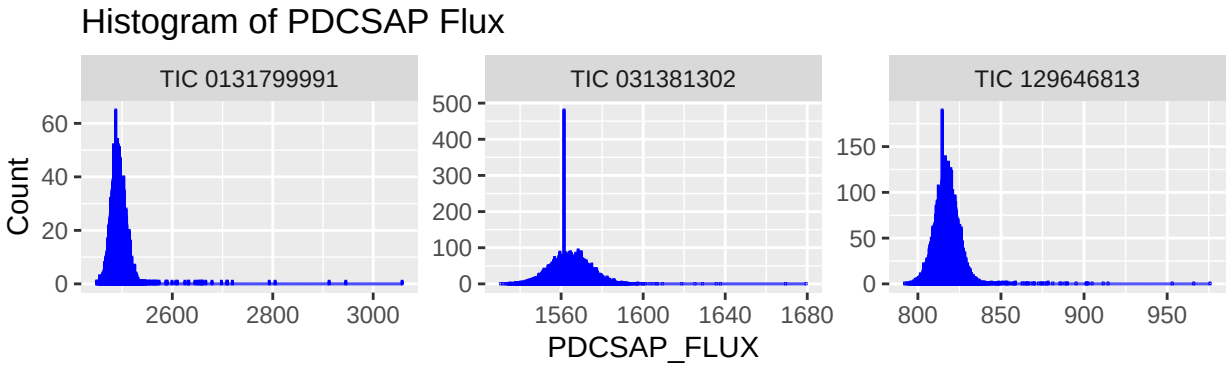

Figure 3: Histogram of Each Stars Brightness

Correlation of Flux Errors

Understanding how measurement errors vary with flux values helps in assessing data reliability. A high correlation between flux and its error could indicate that high readings during a flare are less reliable. A positive correlation might suggest that error increases with flux intensity, which could complicate the detection of true flares from anomalies.

Table 2: PDCSAP Flux and Flux Error Correlation by Star

Star	Correlation
TIC 031381302	0.0831141
TIC 129646813	0.0020326
TIC 0131799991	0.1466368

The correlations between PDCSAP flux and the flux error in these stars are weak, suggesting that errors in the flux measurements are generally not strongly dependent on the flux values themselves. The low correlation suggests that the error associated with flux measurements does not vary significantly with the stellar brightness, indicating stable and consistent data quality across different flux levels.

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.

Additionally, I have explored a Poisson Process Regression (PPR) model, which 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 results are analagous to the GPR therefore it will be omitted in the results.

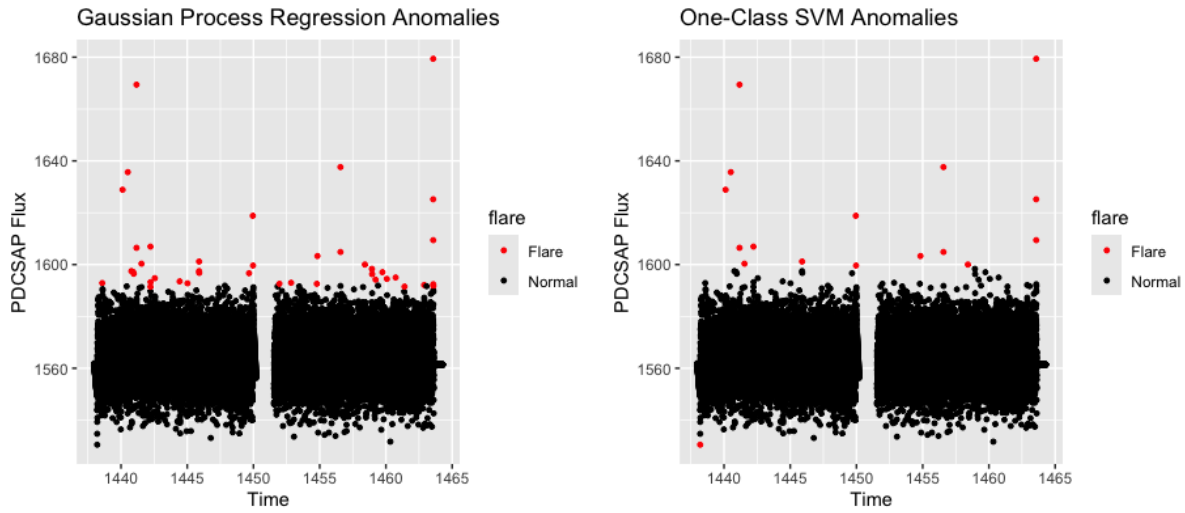

Figure 4: Anomaly Detection using GPR and OCSVM Models on TIC 031381302

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

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 find optimal values of p, d, and q, a stepwise search is performed over combinations from 0 to 5, and the optimal values are chosen based on which fit produces the lowest AIC. For TIC 031381302, TIC 129646813, TIC 0131799991, the values of p, d, and q are (1,0,5), (4,1,1), and (4,0,1) respectively. To model flare events, synthetic flares are injected by adding spikes at randomly chosen time points, which are labeled as 1 in the simulation dataset. 10 flares are injected into the fitted SARIMA simulation. For each selected flare index, the flare amplitude is determined using the Pareto distribution. The parameters shape = 1 and scale = 25 are chosen for generating larger flare amplitudes. The flare is then added to the original flux value at the selected indices to mimic real star data.

Figure 5 is the comparison between the actual flux data from TIC 0131799991 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.

Model Evaluation

Hyperparameter tuning via grid search was applied on the OCSVM model to optimize anomaly detection. The grid search tests various combinations of the ν and γ parameters. ν controls the fraction of anomalies the model can tolerate and the sensitivity to outliers. γ controls the flexibility of the decision boundary, determining how tightly the model fits the data. I loop through a range of values for both parameters (ν from 0.01 to 0.5, γ including 0.01, 0.05, 0.1, 0.5, and 1), training the OCSVM model with each combination. The model is then tested on the simulated star data, and the accuracy of anomaly detection is calculated by comparing predicted anomalies to the true labels. The parameters that produced the highest accuracy were

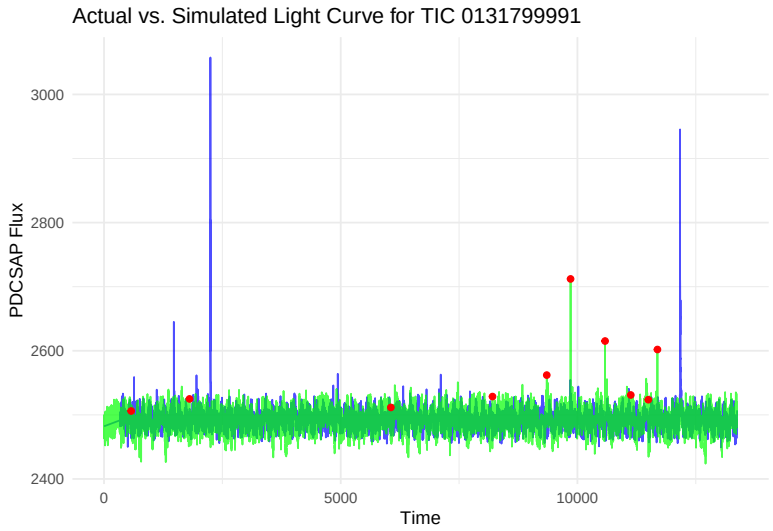

Figure 5: SARIMA Simulation on TIC 0131799991

chosen - in this case were $\nu = 0.01$ and $\gamma = 0.01$.

Results

The models were trained on real star observations from TIC 031381302 and evaluated on simulations from TIC 031381302 and TIC 0131799991. The results show that GPR and OCSVM models successfully identify stellar flares within simulated data, consistently achieving high sensitivity rate of 90%. This performance metric supports each model's ability to detect a majority of synthetic flares accurately, validating the efficacy of SARIMA-based simulations in representing realistic stellar activity.

Table 3: Model Performance on Simulated Data

Model	Flares_Predicted	Sensitivity	Specificity
GPR on TIC 031381302 Simulation	26	0.9	0.99904
GPR on TIC 0131799991 Simulation	28	0.9	0.99858
OCSVM on TIC 031381302 Simulation	59	0.9	0.99718
OCSVM on TIC 0131799991 Simulation	62	0.9	0.99603

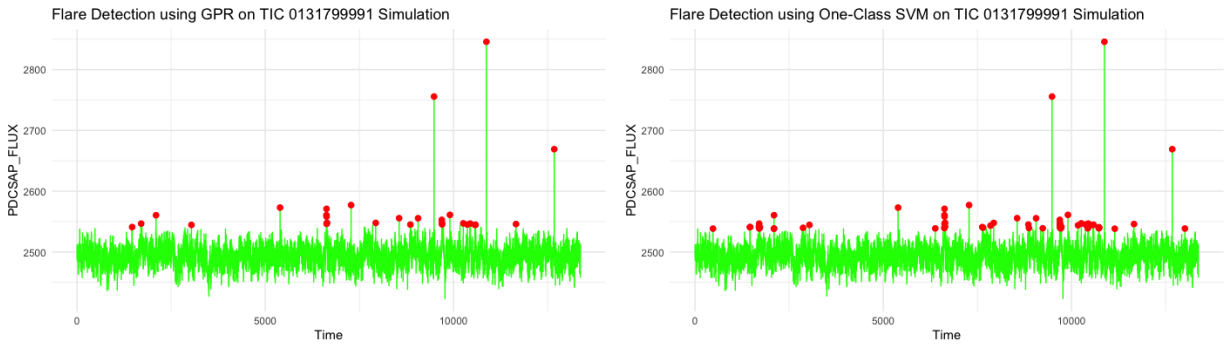

Figure 6: Anomaly Detection using GPR and OCSVM Models on TIC 0131799991 Simulation

Despite both models demonstrating comparable sensitivity, there are notable differences in their specificity rates. Specifically, the GPR model maintained higher specificity (approximately 0.998–0.999) across both simulations compared to the OCSVM, which yielded lower specificity rates (approximately 0.996–0.997). These discrepancies highlight the GPR model's capacity to distinguish effectively between flares normal stellar variability. This advantage is likely from GPR's probabilistic approach, which explicitly incorporates uncertainty estimation, allowing more reliable thresholding (labelling points beyond 3 standard deviations from the mean as flares). Consequently, the GPR model achieves a favorable balance between sensitivity and precision.

Limitations

Imputation assumes a flare does not occur while the satellite is turned off. This assumption might not hold if flares occur during periods of missing data or when the satellite is turned off. This can lead to inaccurate imputation, and biased results. Additionally, simulations were validated visually, comparing each stars flux and simulated flux plotted together. Varying scale parameters were used ranging from 25 to 50. Therefore, robustness of the models implemented are depend on if the simulations accurately resemble real star data. Visually assessing simulations involves subjective judgment, which potentially introduces observer bias and limits reproducibility. The chosen range of 25 to 50 may not fully encompass all realistic scenarios, limiting generalizability and potentially missing optimal scaling values outside of this range.

The performance of the GPR model in anomaly detection heavily depends on the choice of the threshold. If the threshold is too strict, it may fail to detect flares, while if it's too lenient, it may classify too many normal data points as anomalies. A threshold of labelling observations greater than 3 standard deviations from the mean is used as it balances the amount of observations labelled as flares, while maintaining high accuracy. Additionally, the GPR model is computationally inefficient - it has a time complexity of $O(n^3)$ because every pair of training points needs to be compared using the kernel function, which becomes expensive for large n. Lastly, the OCSVM limitations include that even after tuning parameters - we observed overfitting. Many false positives were observed with this model, which may suggest that the RBF kernel is too flexible, which can cause the model to overfit to the noise in the data. Incorporating domain knowledge to tune threshold selection or employing ensemble-based approaches may enhance the robustness and overall accuracy of flare detection.

Conclusion

This analysis demonstrates the effectiveness of GPR and OCSVM methods in detecting stellar flares from simulated stellar brightness data. Both models achieved a high sensitivity rate of 90%, confirming their ability to correctly identify most synthetic flare events. However, GPR consistently outperformed OCSVM by maintaining higher specificity. The GPR model allowed for more robust thresholding, reducing misclassification compared to OCSVM's approach. Several limitations emerged from the analysis, notably the reliance on visual validation of simulations, the subjective selection of scale parameters, and the assumptions made during imputation of missing values. These factors introduce potential biases and constraints on the generalizability of results. The OCSVM method showed overfitting, possibly due to the flexibility of the RBF kernel or threshold setting, suggesting that further parameter tuning or alternative kernels could be beneficial. Future work could focus on addressing these limitations by incorporating objective, quantitative validation methods for simulations, exploring additional kernels or regularization techniques for OCSVM, and employing ensemble-based approaches to improve overall detection robustness. Integrating domain knowledge into threshold selection and model refinement could also significantly enhance model performance.