Exoplanets and Machine Learning

Jacob Bratton, Jessica Lynch, Jett Pavlica, Alexander Wigger

May 4, 2023

Abstract

Exoplanets are an upcoming topic of research in the scientific community. Exoplanets are planets that orbit stars like Earth orbits the Sun. Machine learning is a new technology that can be used to assist in determining whether certain stars have exoplanets or not, and it might be able to find planets like Earth. We used a database with stars classified as having exoplanets (a) or not (b). We then used four different machine learning models with three different versions of the data we had (raw, feature-engineered, SMOTE). We used two forms of measurement to rate the model's accuracy and F1 minority score. We used F1 minority score because the data is imbalanced and we wanted to make sure no stars with exoplanets would be classified as not having them because they are much rarer. We studied the results from each of the models and found the random forest model to be the most accurate predictor as well as having the highest F1 minority score with the data having been feature-engineered and SMOTE.

Introduction

On March 7th, 2009, the Kepler space telescope was launched, sparking new excitement about the possibility of finding an Earth-like planet in our galaxy. One of Kepler's objectives was to find exoplanets (any planet outside of our solar system). To achieve this, Kepler would conduct "campaigns", where it would focus its camera on one area of space and observe that region for around 80 days. In order to discover if an exoplanet exists around a given star, Kepler would measure the flux (light intensity) of the star at regular time intervals. If a planet crosses in front of the given star (in an event called "transit"), there would be a small dimming of the flux for an interval of 2 to 16 hours (NASA). Evidence of this transit event would mark this star as a candidate system, to be explored further. Being able to determine whether or not an exoplanet exists around a star informs if there is a planetary system that needs to be explored further, in the constant search for Earth-like habitable planets outside of our solar system, as well as increasing our knowledge of the galaxy as a whole. Our dataset consists of these flux measurements, over 3000 of them for each star, and a label that is either "1", meaning there is no exoplanet found around the star, or "2", meaning there is at least one exoplanet found around the star. We will attempt to train a machine learning model that can accurately predict whether or not an exoplanet orbits a star, given the flux measurements of that star.

Related Works

Our review of related work highlighted several studies that have utilized machine learning and deep learning algorithms to identify exoplanets. One study proposed a method for classifying exoplanet signals using a convolutional neural network, while another presented ASTRONET, a deep learning architecture that searches for habitable exoplanets based on their physical characteristics. We also found a third study that compared the performance of multiple machine learning models on predicting the existence of exoplanets using both supervised and unsupervised tasks. The study concluded that decision trees and neural networks gave the highest accuracies on the Kepler dataset. We plan to build upon these studies to contribute to the exploration of exoplanets and their characteristics by experimenting with additional datasets and machine learning methods than the ones done in these studies.

Approach

Our approach started with finding a data set that contained information on a variety of stars and marked whether or not they had exoplanets. We chose to use two different data sets, one for training the model and the other for testing the model. The training set had 3197 flux measurements for each star. There were 37 stars classified as having exoplanets and 5050 without exoplanets. The test data on the other hand had 5 stars classified as having exoplanets and 565 without exoplanets. We decided to look for patterns within the training set data first to help figure out how to train models and what would be effective. In figure 1 we have two examples of stars with exoplanets and two without that show the raw data and the results as we transform it. This is used to show the difference and help us classify stars more easily.

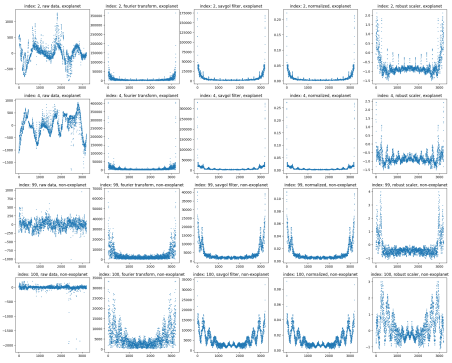
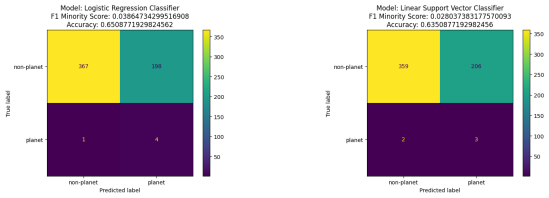


Figure 1: Example star flux scatter plots

Methodology

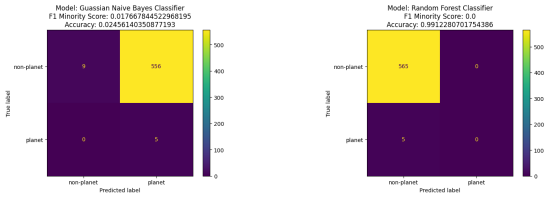
Now that we understand the data that the models will be trained on, we selected a few different machine learning methods to see each of their effectiveness. We selected logistic regression, linear SVClassifier, Gaussian NaiveBayes classifier, and a Random Forest classifier from the sklearn library. We started by using these models from sklearn on the trained data and passed them the test data and scored them with an F1 score and accuracy. We need to use F1 score because there is such a small number of stars with exoplanets (5) in the test data compared to stars without (565). The F1 score allows us to emphasize the score by correctly identifying stars with exoplanets compared to correctly classifying stars as not having exoplanets. This is done to prevent a model scoring well by just always classifying a star as not having an exoplanet.

Trial 1



(a) Logistic Regression (b) LinearSVC

Figure 2: Confusion Matrix



(a) Gaussian Naive-Bayes (b) Random Forest

Figure 3: Confusion Matrix

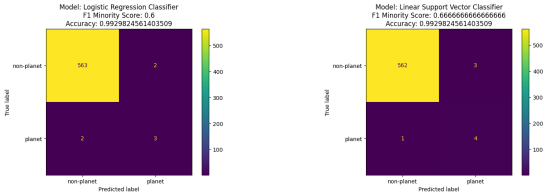
Summary 1

The logistic regression had an accuracy around 63-64% but had a very low F1 score about .03 on the minority class (stars with exoplanets) which indicates it did not do a good job predicting the values we really wanted to be predicted well. The linear SVC classification had very similar results with around a 64% accuracy and a .03 F1 score on the minority class. The Gaussian Naive Bayes method had an accuracy around 2% and F1 score

of .02. It basically attempted to over fix and assume all stars had exoplanets which was very inaccurate. Lastly, the random forests method had a 99% accuracy by classifying all stars as not having exoplanets, but an F1 score of 0 because it never expects exoplanets to exist. None of these results were satisfactory as the accuracy's are either much too low or the F1 score is too low which is the value we really want to predict correctly.

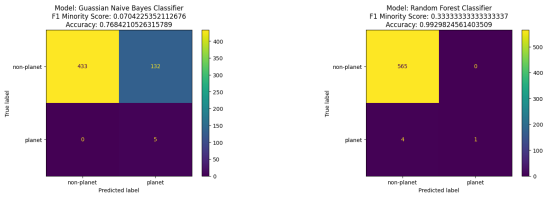
Trial 2

The previous results were unacceptable and therefore we needed to process the data to get better results from our models. We decided to use feature engineering because this is widely known to be useful and help clean the data. We used the sklearn built in functions fourier, savgol, norm, and robust in order to reduce the noise, then normalize the data and use robust to devalue the outliers in the data. After that we ran the same models on the transformed data to see how they fared.



(a) Logistic Regression (b) LinearSVC

Figure 4: Confusion Matrix



(a) Gaussian Naive-Bayes (b) Random Forest

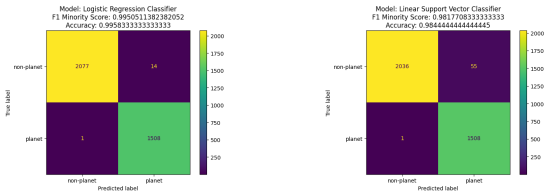
Figure 5: Confusion Matrix

Summary 2

Logistic regression scored a 95% in accuracy and a .08 F1 score. This is a much higher accuracy and a slightly improved F1 score, but it's not good enough. Linear SVC classification scored a 99% accuracy and a .54 F1 score which is a massive improvement. Gaussian Naive Bayes scored a 77% accuracy and a .07 F1 score which is great improvement, but still under performing. Lastly, random forest classifier had a 99% accuracy and a .33 F1 score which was a great improvement. The feature engineering worked, but there are still problems because of how imbalanced the data is when it comes to classifying data because one misclassification of a planet with an exoplanet will plummet the F1 score.

Trial 3

In order to fix the imbalance data problem we used Synthetic Minority Over-sampling Technique(SMOTE). What this does is takes the current minority classifications and finds their nearest neighbors in their subspace and interpolates their data to create more synthetic data that represents the minority class. We used the imblearn SMOTE method to SMOTE the data and then reran our models on the data that has been feature engineered and SMOTE. The results are below in figures 132-14313.

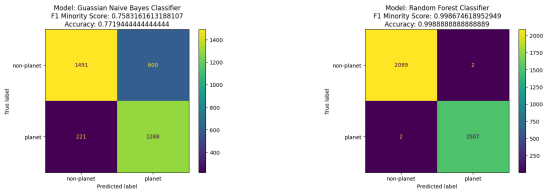


(a) Logistic Regression (b) LinearSVC

Figure 6: Confusion Matrix

Summary 3

Logistic regression scored a 92% in accuracy and a .91 F1 score. This is a slightly lower accuracy



(a) Gaussian Naive-Bayes

(b) Random Forest

Figure 7: Confusion Matrix

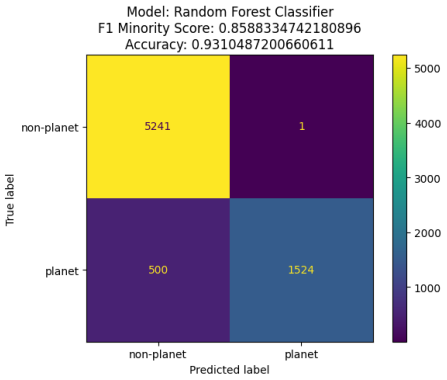


Figure 8: Example star flux scatter plots

and a massively improved F1 score overall. It's much better, but not the best method. Linear SVC classification scored a 98% accuracy and a .98 F1 score which is a very good predictor from both accuracy and F1 score. Gaussian Naive Bayes scored a 77% accuracy and a .76 F1 score which is a great improvement in F1 score, but obviously not the best method. Lastly, random forest classifiers had a 99.9% accuracy and a .99 F1 score which is by far the best method combining accuracy and F1 score. After the data has been SMOTE and feature engineered the models have a much better accuracy and F1 score in general, but the random forest classifier does the best job of these 4 models.

Random Forest Test

After finding that the random forest performed the best by our F1 minority standard and accuracy we decided to use another new test data set. We acquired this data set by selecting individual star's flux records from NASA's public archive and compiling an entirely new csv data-set. We had to manually clean the data by interpolating it so each star had the same number of flux values. After that we feature engineered it and SMOTE it due to the imbalance of data once again.

Summary 4

The model did not perform as well as it did on the previous data-set, but it performed considerably well in both accuracy and F1 minority score for having been tested on a completely unrelated data set of stars. With an accuracy of 93% and F1 minority score of approximately .86 it performed better than any of the other methods and could be considered a very useful model.

Conclusion

Overall, we can conclude that machine learning models can be used as a very good predictor of whether a star has exoplanets or not. We found that using a random forest classifier on processed data is extremely accurate and provides the best results in cases where data may be imbalanced. For further research we would attempt random forest on non-labeled data and with research we could be able to classify unclassified stars. This study has also helped make clear that the majority of stars do not have solar systems / exoplanets like our sun, so using accuracy as the only factor of a prediction model is risky due to the natural imbalance of stars with exoplanets versus those without. To those who wish to repeat or perform similar experiments we'd recommend using feature engineering and SMOTE to fix your data if results on your models are poor due to noise, outliers, or imbalanced data.

Acknowledgments

A big thanks to sklearn, imblab, and kaggle for the machine learning methods and guides. These were extremely helpful in completing this project. Thanks to NASA for researching exoplanets and piloting this field.

References

HARI, AKSHAY. "Model Comparison with Feature Engineering - 99.9%." Kaggle, Kaggle, 19 Jan. 2021, https://www.kaggle.com/code/akshayh007/model-comparison-with-feature-engineering-99-9.

Balmukhanov, Assylkhan. "Easy Way to Achieve 99.1% Accuracy." Kaggle, Kaggle, 5 Apr. 2023, https://www.kaggle.com/code/ assylkhanbalmukhanov/easy-way-toachieve99-1-accuracy.

WDelta,"Exoplanet Hunting in Deep Space." Kaggle, https://www.kaggle.com/datasets/keplersmachines/kepler-labelled-time-seriesdata.

Johnson, Michele. "Mission Overview." NASA, https://www.nasa.gov/ mission_pages/kepler/overview/index.html.

Dattilo, A., Vanderburg, A., Shallue, C. J., Mayo, A. W., Berlind, P., Bieryla, A., ... Yu, L. (2019). Identifying Exoplanets with Deep Learning. II. Two New Super-Earths Uncovered by a Neural Network in K2 Data. The Astronomical Journal, 157(5), 169. doi:10.3847/1538-3881/ab0e12

Jin, Yucheng, et al. "Identifying Exoplanets with Machine Learning Methods: A Preliminary Study." ArXiv.org, 1 Apr. 2022, https://doi.org/10.48550/arXiv.2204.00721.

Shivani Dere, Maziya Fatima, Rutuja Jagtap, Unzela Inamdar, Nikhilkumar Shardoor, "Anomalous Behavior Detection in Galaxies and Exoplanets using ML DL Techniques", 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), pp.938-947, 2021

Dattilo, Anne, et al. "Identifying Exoplanets with Deep Learning. II. Two New Super-Earths ..." The Astronomical Journal, 9 Apr. 2019, https://iopscience.iop.org/article/10.3847/1538-3881/ab0e12/meta.

Jagtap, Rutuja, et al. "Habitability of Exoplanets Using Deep Learning — Ieee Conference ..." IEE Xplore, 14 May 2021, https://ieeexplore.ieee.org/abstract/document/9422571/.

A, Barbara. Mikulski Archive for Space Telescopes. https://archive.stsci.edu/kepler/data$_search/search.php$.