Jose Diaz-Rivera
Dominick DeCanio
JR Kargbo

Exoplanet Classification Uncertainty Analysis

Project Description

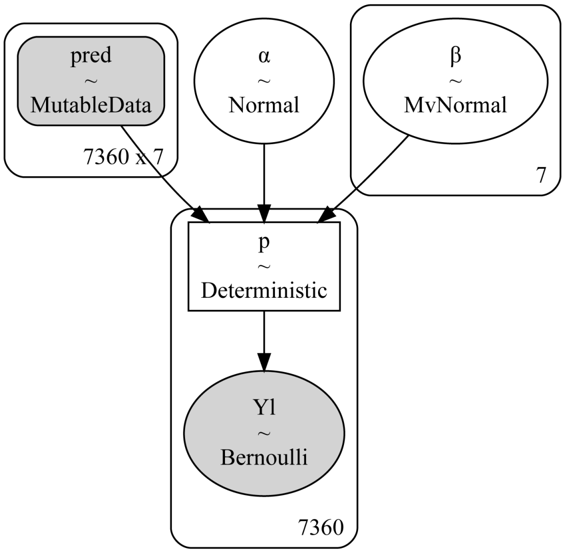
Throughout the 20th century our collective fascination with extraterrestrial life has grown. We travel to distant worlds in science fiction books, movies and shows. Space has cemented itself as the modern stage of drama. Curiosity of the vast universe around us has prompted public programs in space exploration which reverberate to this day. Americans fantasized about space exploration in all of its forms, and the famous missions of the space race made the fantasy of reaching other worlds more tangible. To detect planets in other star systems (exoplanets) manned spacecraft were replaced with unmanned telescopic satellites which could detect planets millions of miles from Earth.

The Kepler Space Observatory is a NASA-built satellite that was launched in 2009. The telescope is dedicated to studying our region of the milky way by searching for exoplanets that may be habitable and determining what proportion of star systems might contain such planets. The nearest exoplanet to Earth is Proxima Centauri, which is 4.2 lightyears away [1]. Because most star systems are significantly further from earth, measuring these bodies tests the limits of Kepler's instruments. Consequently, possible exoplanet observations are classified as either “CANDIDATE” or “FALSE POSITIVE” based on the probability that the observation is caused by an exoplanet. This system allows the classification rule to include different levels of measurement error (resulting from the disparate distances of the planets) within the exoplanet classification rule [2].

Because NASA makes a soft classification of exoplanet observations, our group’s objective is to use the Kepler Space Observatory measurements to evaluate the possible uncertainty of exoplanet classifications using a reduced bayesian logistic regression model.

Probability Model

We sourced our data from a kaggle submission made by NASA [3] and used their published data dictionary [4] to determine the probability model best suited to represent our problem. Because variables included in [4] are measurements regarding a star which might contain an exoplanet, we find it logical to proceed with the conjecture that these measurements might follow a gaussian distribution. Below is an image representing our probability model where we have used seven of the available parameters as predictors for the classification of a possible exoplanet observation. We utilized the PyMC “MutableData” method [5] to create a train-test split of the inputs of this model to allow us to check the predictive ability of the model. The plate on the left hand side of the probability model represents this “MutableData” implementation, whereas the plate on the right represents the seven variable multivariate gaussian distribution we will assume for the posterior distribution, and use in our uncertainty analysis of exoplanet classification.



Approach

In order to classify Kepler Objects of Interest (KOIs) we created a Bayesian logistic regression model using the probability model in the prior section. A logistic regression model is suitable because we are addressing a binary classification problem; we are concerned with whether an observed exoplanet is habitable for human beings or not and predicting that probability. We chose to conduct analysis using this model through two main approaches: Markov Chain Monte Carlo (MCMC) and Variational Inference (VI). We used no u-turn sampling (NUTS) and ADVI to process our MCMC and VI approaches respectively.

We chose to use MCMC in this analysis because of this method's ability to accurately approximate the target distribution, allowing better predictive performance at the cost of computational complexity. With this Bayesian sampling approach, we can compute the posterior probability distribution of the model parameters conditionally upon the predictors and the outcome variable. Given that this model is not meant to be run many times, or in fast succession, the tradeoff between predictive performance and computation complexity is appropriate. After conducting posterior estimation with MCMC, we repeated this process using VI to see if the results were similar. We chose to add VI to our analysis because of this method's ability to be much faster than MCMC as it is less computationally expensive, although the disadvantage is that VI is less accurate. Although VI is less useful for finding the globally optimal solution for the target distribution like a sampling approach, because our goal is optimization rather than purely inference, we can learn whether the distribution has converged and the bounds on accuracy. Consequently, if we see similar results when we conduct both approaches, we will be more confident that they are correct.

To evaluate these methods, we utilized a trace plot to gauge the convergence of the methods and a forest plot to assess variable significance. We proceed with the models after ensuring convergence to compare the effectiveness of both approaches in modeling the posterior distribution, and predicting the exoplanet class of an observation.

Results

The body of our analysis rests in the uncertainty analysis of the classification of exoplanets. We will accomplish this analysis of the posterior distributions calculated using NUTS and ADVI using a series of diagnostic plots to understand how these posteriors fit the observed distribution, how they differ, and what they say about the variables that compose them. We will begin the analysis by checking each method for convergence and interpreting their parameter estimations. Then, we will move to a visual comparison of the posterior predictive distributions and trace plots of these methods. After this we will proceed to the predictive segment of the analysis.

For a NUTS approach to succeed, the estimated posterior distribution must converge as a result of the sampling procedure. To check for convergence we reference the trace plot. The trace plot of our posterior distribution as sampled through NUTS [Fig. 1] shows convergence by the normal-distribution shaped density plots of the parameters in the posterior distribution. We can see from the forest plot of the NUTS posterior distribution [Fig. 2] That all of the variables are significant-their credible intervals of their coefficients do not bound zero-except for parameter 4 (Stellar Radius) and alpha, the intercept term.

We check these conditions for the variational inference approach using ADVI, and we observe that the posterior distribution converges because of the normal distribution shape to the parameter density plots [Fig. 3]. In contrast to the forest plot observed from our sampling approach [Fig. 2], the ADVI forest plot [Fig. 4] shows that there are many parameters whose 94% HDI bounds 0: 7 (Estimated Magnitude Kepler-Band), 6 (KIC Declination), 4 (Stellar Radius), 3 (Stellar Surface Gravity), and the intercept term. These results indicate that, with respect to NUTS, the variational analysis using ADVI attributes larger credible intervals to each parameter which indicates that the parameter estimation is less precise. This is exactly what we expected when comparing these methods and we will proceed to treat the credible intervals of our NUTS analysis as the true values of the posterior because of this higher accuracy.

Next we will visualize the difference between the posterior distributions estimated using NUTS and ADVI through posterior predictive distribution plots. A posterior predictive distribution plot creates a visualization which shows the accuracy of the posterior distribution to the observed distribution by creating an overlay of a sample of posterior distributions and the observed distribution. For this analysis, we created two such plots to analyze the accuracy of the NUTS and ADVI approaches. We can see from these two posterior predictive plots [Fig. 5], [Fig. 6] that the posterior estimations from the NUTS approach exhibit dramatically less variation than those generated using the ADVI approach. These plots do not show the shape of the distribution as we saw it in the trace plots [Fig. 1], [Fig. 3], but they allow us to declutter the plots by giving discrete values of the estimate at two locations for each approach. We can see that the variance among samples is the same for both sides of [Fig. 5] and [Fig. 6], and the distributions of both plots are centered around the observed distribution with their posterior predictive means laying precisely on the observed distribution. This reinforces our conclusion that the posteriors of the NUTS and ADVI approaches appropriately converged.

The final comparison of the posterior distributions of the NUTS and ADVI approaches comes from the overlaid forest plots [Fig. 7]. In [Fig. 7], we can see that the forest plot of the NUTS posterior is dramatically narrower than that of the variational approach using ADVI. The ADVI trace plot exhibits so much variance that in areas where the NUTS model would certainly have a probability of exoplanet smaller than 0.5, the ADVI model predicts the opposite. We can see that the NUTS model is preferable because of this increased accuracy.

To conclude the analysis, we created predictions using both the NUTS and ADVI approaches to evaluate how these models would perform in this predictive setting and evaluated this performance using predictive accuracy, confusion matrices, and plots of the classification threshold versus the true labels of the observations. In these plots 0 corresponds to "CANDIDATE", and 1 corresponds to "FALSE POSITIVE"

The predictive accuracies of our NUTS and ADVI models were .6016 and .6010 respectively. We can also see from the confusion matrices [Fig. 8] and [Fig. 9] that the performance of these models was nearly identical. The ADVI model only misclassified two observations more than the NUTS model. This is somewhat surprising given the extreme variance observed in the ADVI model's trace plot when compared to that of the NUTS model [Fig. 7], and that four (4) of the parameters were bounding zero in the ADVI model forest plot [Fig. 4]. This is more surprising because we would expect the large proportion of "CANDIDATE" labels in data versus "FALSE POSITIVE" (75% vs 25% in training set) to create base-case bias, where the model classifies all observations as the dominant class ("CANDIDATE" in this case) to minimize the loss function. Because we are working with a logistic regression, we are instead maximizing the probability, and thus there must be observations of the "FALSE POSITIVE" class which are strongly influential in the training of the model, which closely align with the characteristics of the "CANDIDATE" class.

To explore the predictive abilities of this model further, we plotted the predicted probability of the observation being classified as "CANDIDATE" versus one of the parameters after standardization (Stellar Surface Gravity). We then overlaid our classification rule: 'if the predicted probability of the observation being classified as "CANDIDATE" was greater than 0.5, it was labeled as "CANDIDATE"' by adding a horizontal line to the plot at the 0.5 level. Each point is an observation from the testing set, and their colors correspond to their true classifications. We can see from these plots [Fig 10], [Fig. 11] that the density of observations is highly concentrated around the mean, which helps explain the similar misclassification struggles evidenced in [Fig. 8] and [Fig. 9].

Conclusions

We have met our objective of evaluating the uncertainty of exoplanet classification by illustrating the significance of seven low level measurements of a KOI on the classification of a KOI as either "CANDIDATE" or "FALSE POSITIVE". Through our predictive performance analysis, we can determine that a combination of predictors are jointly responsible for accurate classification of exoplanets, and we have great confidence in the accuracy of NASA's exoplanet classifications. Future works could include more variables, and those made by NASA through feature engineering, to accurately evaluate the performance of models that align more closely to those used by NASA to make exoplanet classification decisions.

Appendix

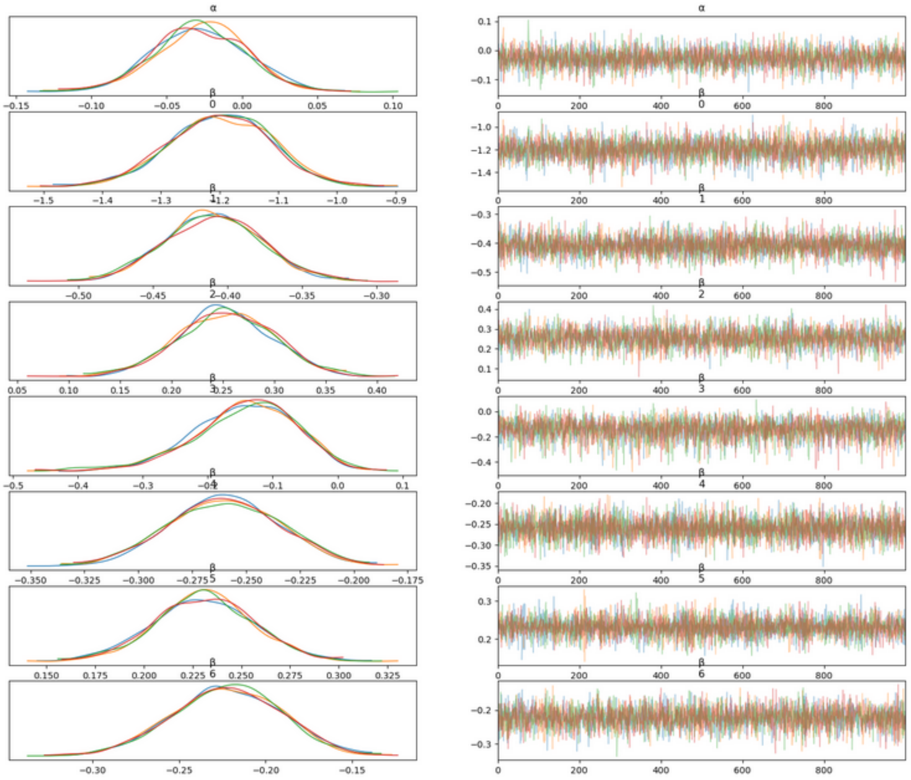


Fig. 1. MCMC Trace Plot.

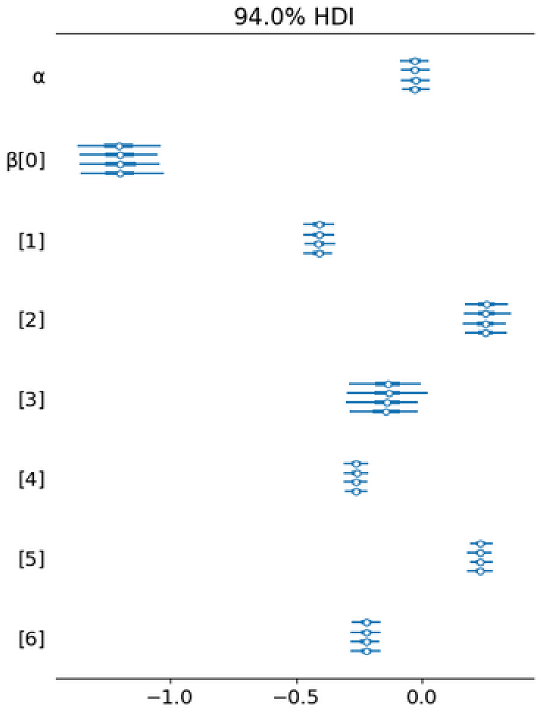


Fig. 2. MCMC Forest Plot.

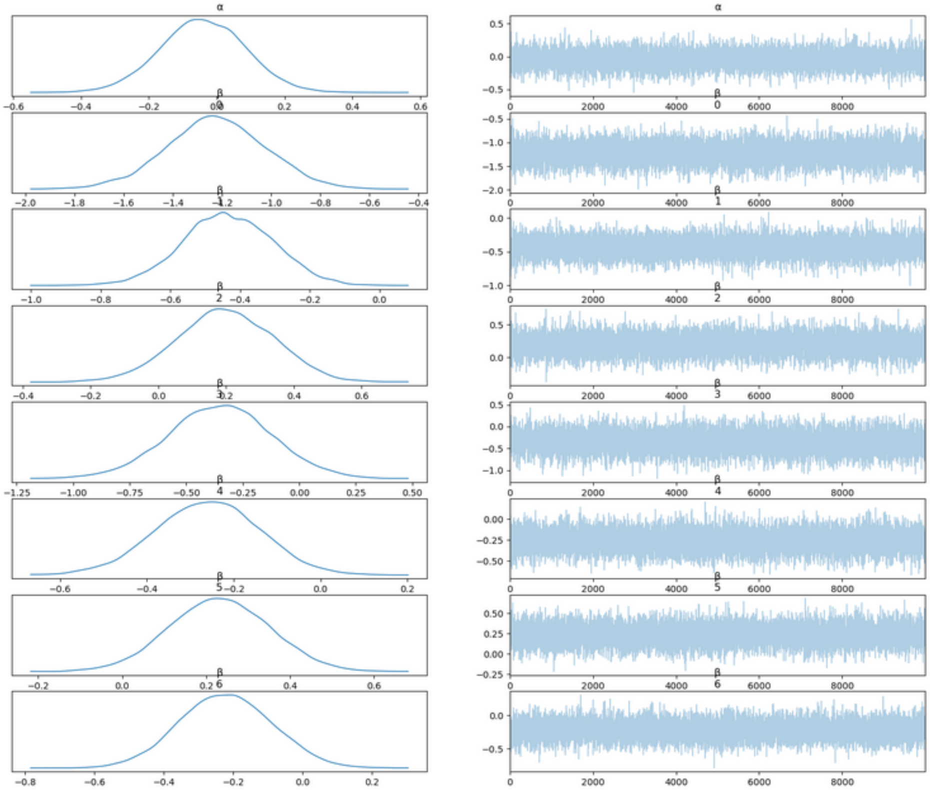


Fig. 3. ADVI Trace Plot.

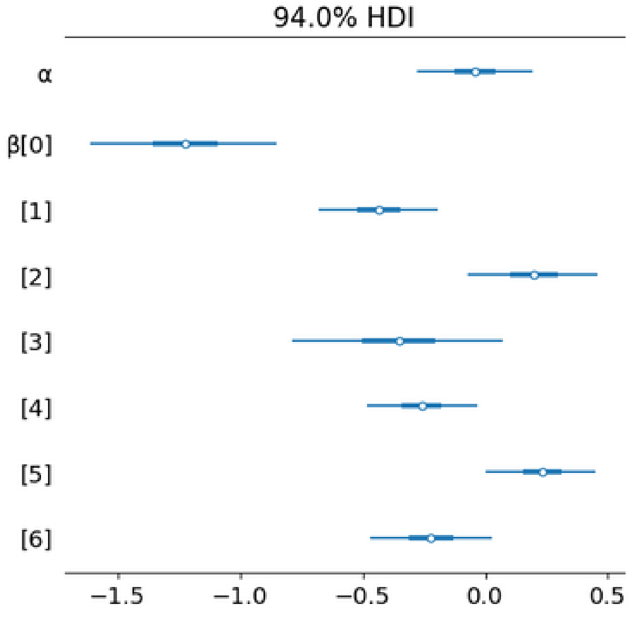


Fig. 4. ADVI Forest Plot.

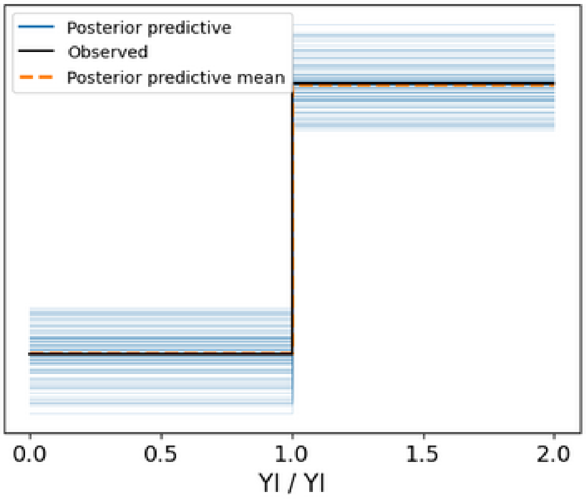


Fig. 5. MCMC Posterior Predictive Distribution.

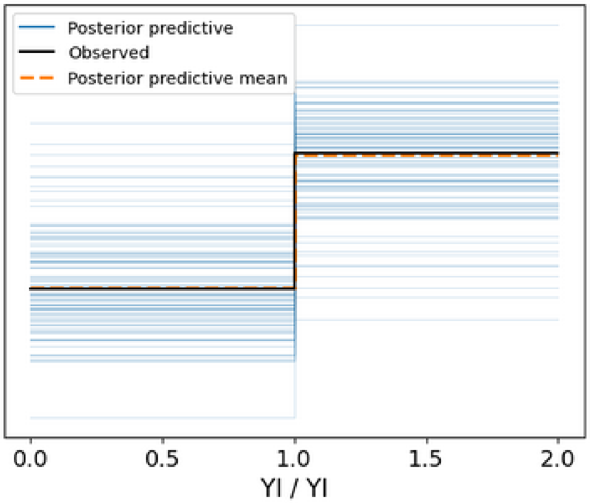


Fig. 6. ADVI Posterior Predictive Distribution.

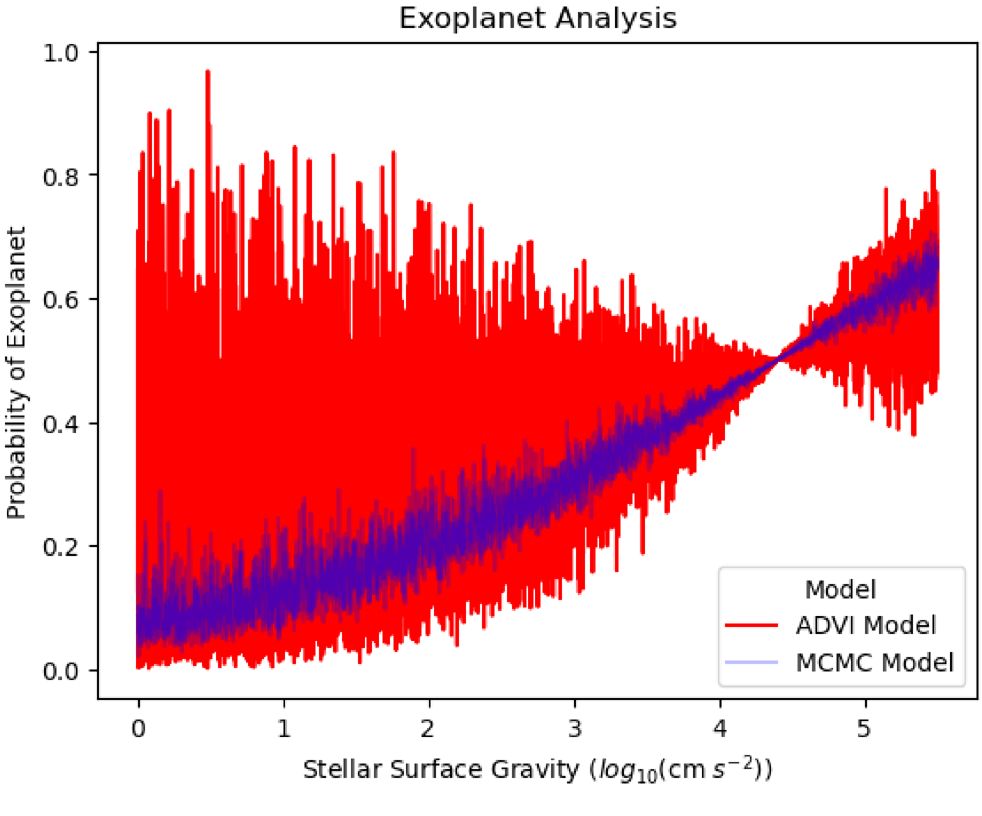


Fig. 7. MCMC vs. ADVI Forest Plots.

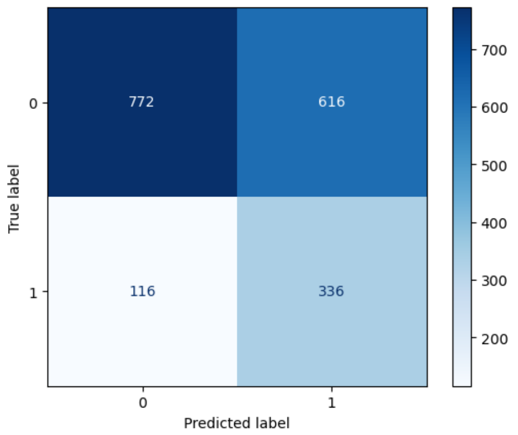


Fig. 8. MCMC Confusion Matrix.

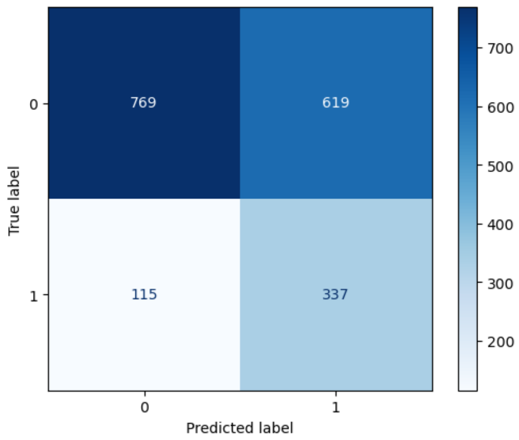


Fig. 9. ADVI Confusion Matrix.

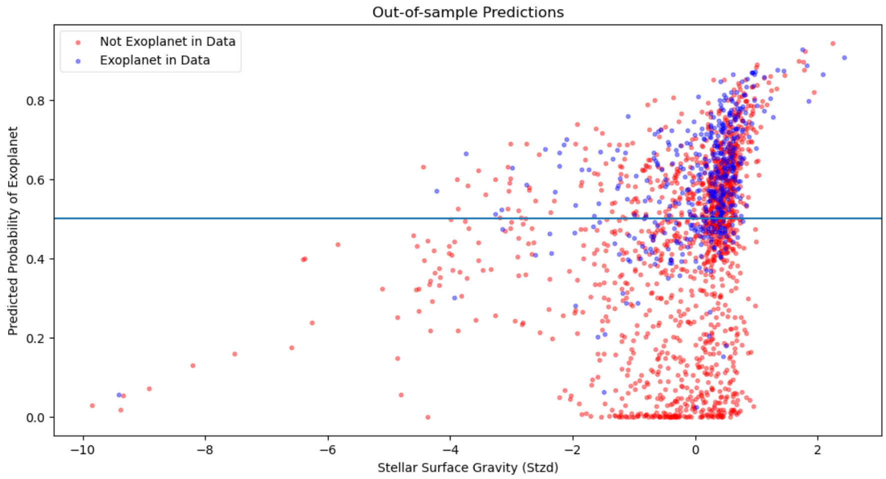


Fig. 10. MCMC, Classification Threshold vs. Truth.

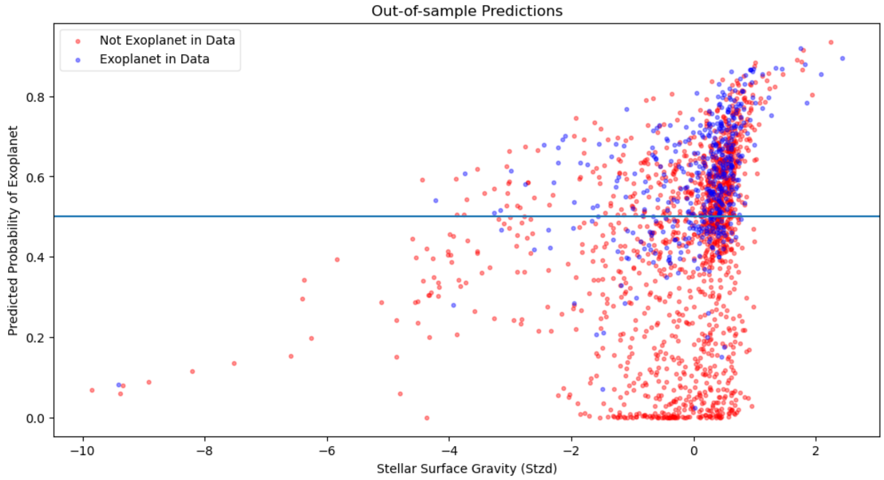


Fig. 11. ADVI, Classification Threshold vs. Truth.

References

[1] K. Haynes, “Proxima Centauri: The closest exoplanet to Earth,” *Astronomy.com*, 20-Feb-2020. [Online]. Available: https://astronomy.com/news/2020/02/proxima-centauri-the-closest-exoplanet-to-earth. [Accessed: 08-Dec-2022].

[2] M. Johnson, “Mission overview,” *NASA*, 14-Apr-2015. [Online]. Available: https://www.nasa.gov/mission_pages/kepler/overview/index.html. [Accessed: 08-Dec-2022].

[3] Nasa, “Kepler exoplanet search results,” *Kaggle*, 10-Oct-2017. [Online]. Available: https://www.kaggle.com/datasets/nasa/kepler-exoplanet-search-results. [Accessed: 08-Dec-2022].

[4] NASA Exoplanet Science Institute, “Data columns in Kepler objects of interest table,” *Data columns in Kepler Objects of Interest Table*. [Online]. Available: https://exoplanetarchive.ipac.caltech.edu/docs/API_kepcandidate_columns.html. [Accessed: 08-Dec-2022]. DOI 10.26133/NEA4

[5] “Prior and Posterior Predictive Checks.” *Prior and Posterior Predictive Checks - PyMC 4.4.0 Documentation*, https://www.pymc.io/projects/docs/en/stable/learn/core_notebooks/posterior_predictive.html.