Finding Exoplanets

Ricardo Alexandro Aguilar

11/18/2020

Summary of Methods and Results

Three models were used to classify whether a star has at least one exoplanet in orbit. Logistic regression, random forest (RF), and gradient boosting machines (GBM) were created using the `caret` package in R. Different combinations for tuning parameters were tested to determine the best combination for RF and GBM. All models were created with three separate K-fold cross-validations with $K = 10$. Due to class imbalances, the Synthetic Minority Over-sampling Technique (SMOTE) was used to for RF and GBM.

Due to the nature of the data, the dimensions had to be reduced. This was done by first computing the difference in light intensity from time point t and $t - 1$. The mean, variance, and maximum of these differences were calculated for each star. The maximum of the standardized light intensity for each star was computed as well. The intuition behind this was that the highest intensity of the star might give us some useful information while standardizing it allows the distance of maximum light intensity from the mean for each star to be comparable to that of the other stars. The last variable created is the amount of times the light intensity remains consistent. Let $t \geq 2$ and $i = t - 1$. Then, this can be written as

$$s = \sum_{i=2}^{n} I\left(sign(\mid d_i \mid) \neq sign(\mid d_{i-1} \mid)\right)$$

where n is the total number of differences and d_i is the difference between light intensities at t and $t-1$. The value $\mid d_i \mid$ is zero when no difference exists and a positive number when there is a difference. Thus, the indicator function determines whether there was a difference and then no difference and vice versa. This sum is the amount of times this occurred. For example, the vectors of differences (13, 0, 6, 0) and (12, 2, 36, 0) will result in the sums equal to 3 and 1, respectively. This variable is computed for each star and appears to perfectly dichotomize the data. Stars with sums greater than 223 contain at least one exoplanet while stars with sums less than or equal to 223 contain no exoplanets. It is important to note that cut off of 223 is not a "one true" cutoff since any value s such that $s \in (152, 294)$ would also perfectly dichotomize the data. This is true, for this particular dataset, because $\max(s_{\text{no exoplanet}}) = 152$ while $\min(s_{\text{exoplanet}}) = 294$.

Figure 1 shows the predictive probabilities for each model on the training set. All of the models are able to perfectly predict whether a star has at least one exoplanet. All of the models have an accuracy, sensitivity, and specificity of one. The logistic regression model and GBM produce probabilities of essentially zero and one. The RF model's predictive probabilities appear to vary more compared to the other two models, but is still able to correctly classify with a "large" interval for possible cut points. A variable importance plot, Figure 2, was created to determine the relative importance of the variables in the RF and GBM. We can see from Figure 2 that our new variable is by far the most important one.

Figure 3 shows the predictive probabilities for each model on the testing set. We can see that these results are the same as the ones for the training set. All of the models are able to perfectly classify the test set. Again, the accuracy, sensitivity, and specificity for all three models are equal to one. This provides evidence that the model was not overfit on the training set and can be generalized.

The results suggest that the structure of the model is not important when the variable for consistent light intensity is included. This means that simpler models can be used to predict the outcome for very large datasets without requiring as much computational power. Additionally, the results seem to suggest that stars with more consistent light intensity are more likely to have exoplanets. Of course, a variable that perfectly dichotomizes data might be too good to be true. This variable might not be as important in data that is structured differently or it could also possibly generalize to other data structures. A more thorough investigation is needed to determine the validity of these results. They are, however, promising and could lead to meaningful research.

Appendix

Figure 1: Predictive Probabilities from Training Set

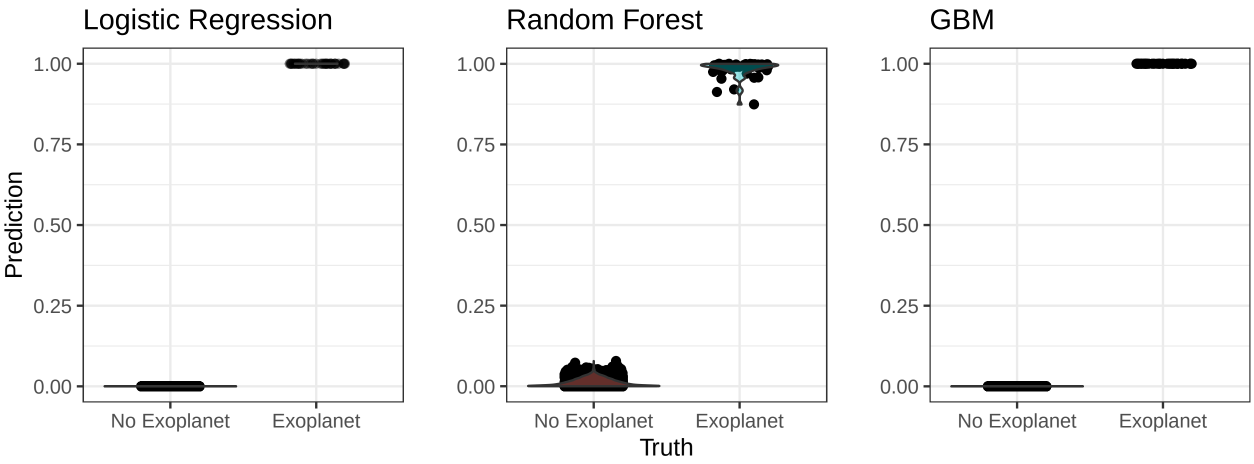


Figure 2: Variable Importance for RF and GBM Models

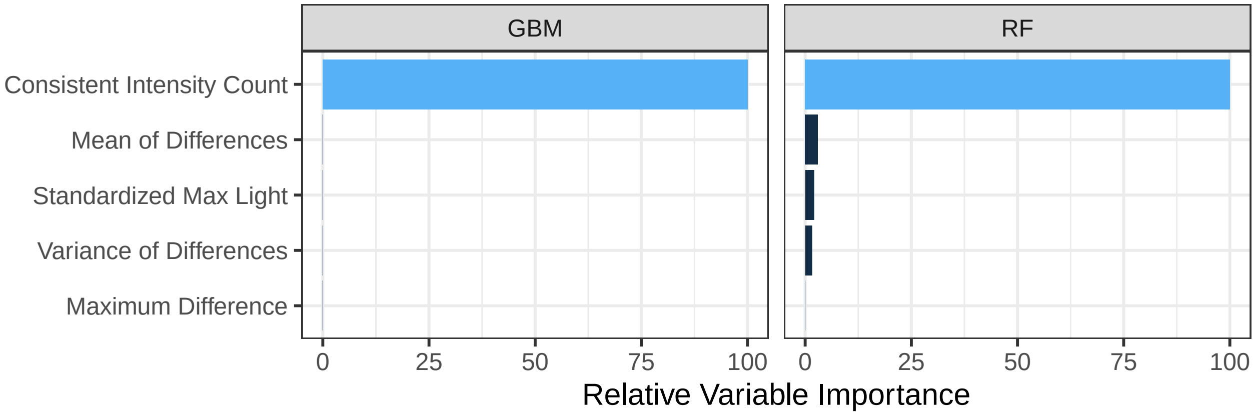
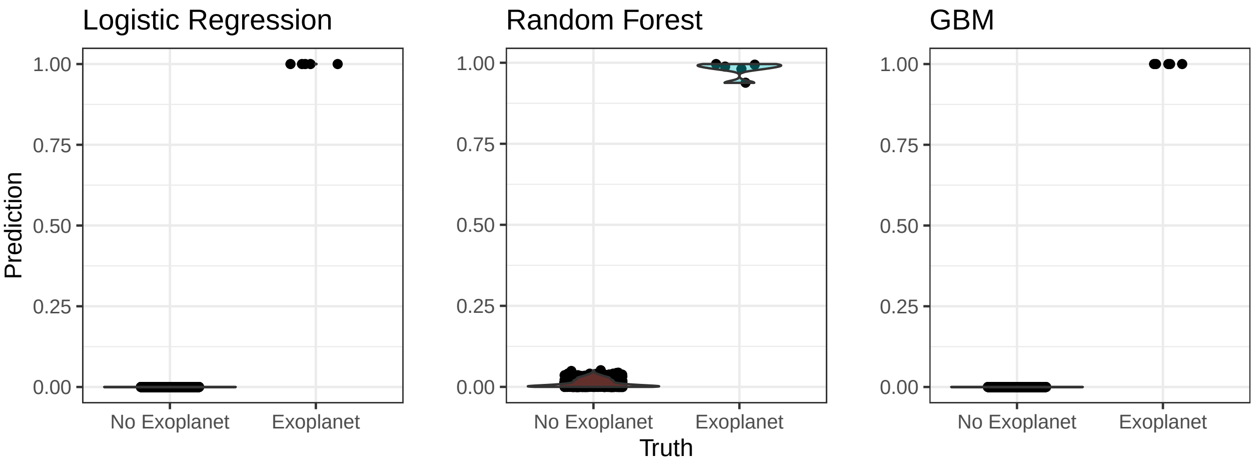


Figure 3: Predictive Probabilities from Test Set


Logistic Regression
Random Forest
GBM
Prediction
No Exoplanet
Exoplanet
Truth