Predicting Exoplanet Disposition Using NASA's Kepler Space Observatory Data and Machine Learning

Capstone Project 1: In-Depth Analysis (Machine Learning)

Springboard Data Science Career Track
April 9thth, 2019

Logan Rudd

1 Experimentation

In order to determine what features were most effective in classifying Kepler Objects of Interest as a confirmed exoplanet or false positive, four different sets of features were selected to fit the algorithms to. Each set respresents the top 25, 50, 75, and 100% of the most important features as calculated by SKlearn's ExtraTreesClassifier. For each model we first imputed the missing values of the data with the mean, then we scaled the data by replacing the original values with their respective z-scores. Lastly, we ran a hyperparamter search in conjunction with a 5-fold cross validation fit for each model using the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) as our scoring metric.

All algorithms performed very well overall, with a ROC AUC score above 0.97 using only the top 25% most important features (Appendix, Figure 1). Although Random Forset and XGBoost had nearly identical results and tied in terms of highest ROC AUC score (0.997), XGBoost had a significantly lower runtime, and for that reason XGBoost was selected to predict the test data. For all algorithms, ROC AUC score did not increase with more than the top 75% of features. The following table summarizes ROC AUC score and Total CPU runtime for each of the algorithms used:

Top % of feature importances used:	25%	50%	75%	100%
Support Vector Machine	0.974 (0.003) 52.6s	0.983 (0.002) 1m 23s	0.985 (0.002) 2m 14s	0.985 (0.002) 3m 2s
Logistic Regression	0.983 (0.003) 25.7s	0.989 (0.001) 1m 23s	0.990 (0.002) 1m 23s	0.990 (0.001) 2m 5s
Random Forest	0.994 (0.002) 8m 41s	0.996 (0.002) 14m 28s	0.997 (0.002) 17m 44s	0.997 (0.002) 21m 15s
XGBoost	0.994 (0.002) 7m 4s	0.996 (0.001) 9m 37s	0.997 (0.001) 12m 49s	0.997 (0.001) 16m 41s

2 Results

Prediction of the test data showed an ROC AUC score of 0.996, meaning that there is a 99.6% chance that the model will be able to distinguish between a confirmed exoplanet and a false positive (Appendix, Figure 2). The confusion matrix in Figure 1 below shows that 12 out of 448 'CONFIRMED' exoplanets were predicted as 'FALSE POSITIVES', and 18 out of 981 'FALSE POSITIVES' were predicted as 'CONFIRMED', showing that the model is slightly better at predicting false positives than confirmed exoplanets. This makes sense as there are nearly twice as many more false positives to train the data on as there are confirmed exoplanets.

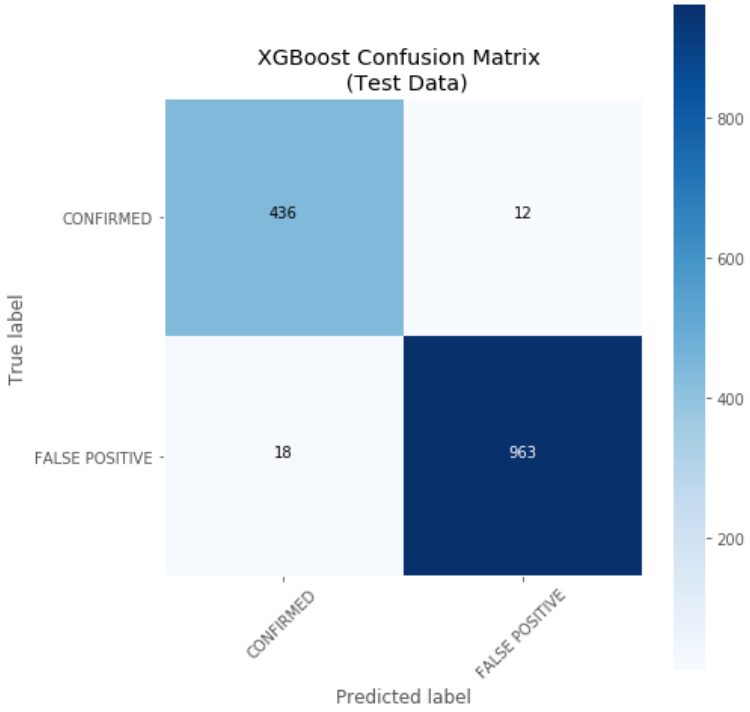

Figure 1: Confusion matrix of XGBoost prediction results on test data

3 Conclusion

A binary classifier was successfully created using data from NASA's Kepler Space Telescope to predict a true exoplanet from a false positive based on features such as transit/stellar properties and instrument characteristic parameters. After analyzing the performance of various machine learning algorithms, XGBoost turned out to be a great machine learning model for this particular binary classification problem.

Appendix

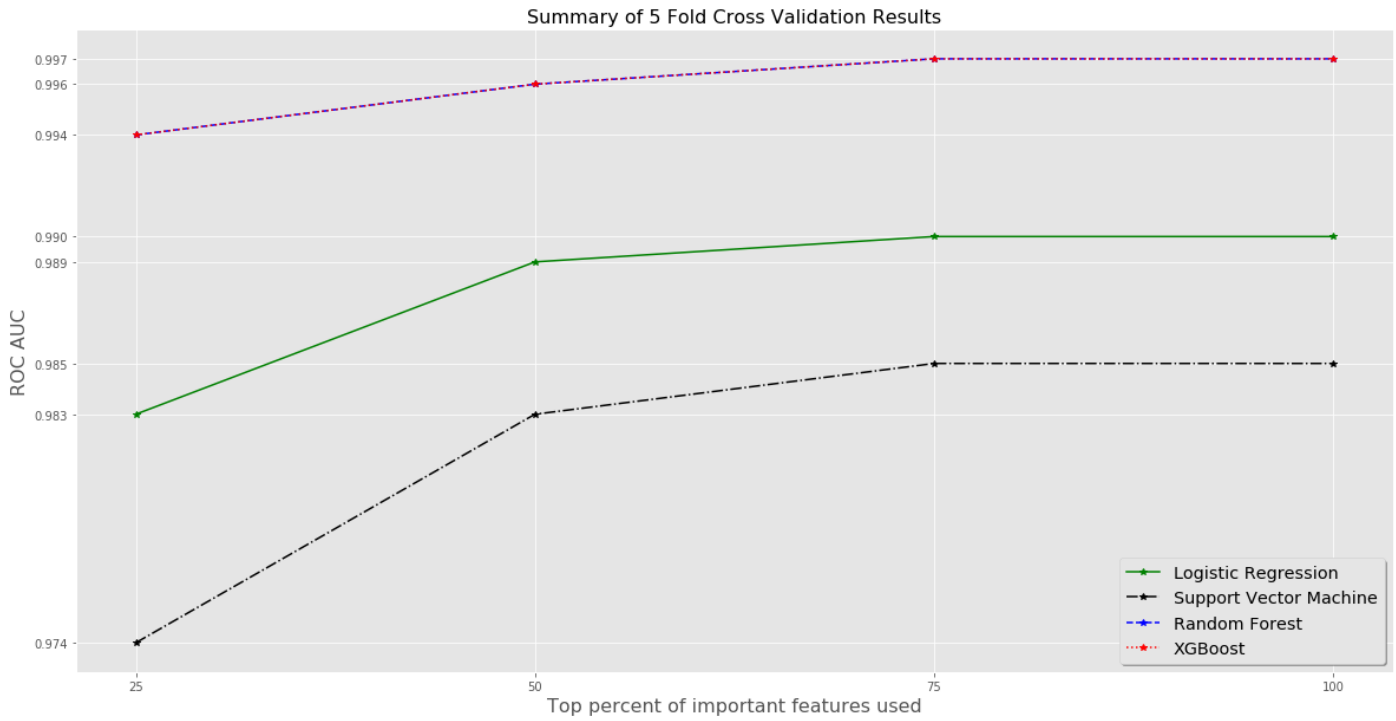

Figure 2: Summary of 5-fold cross validation results

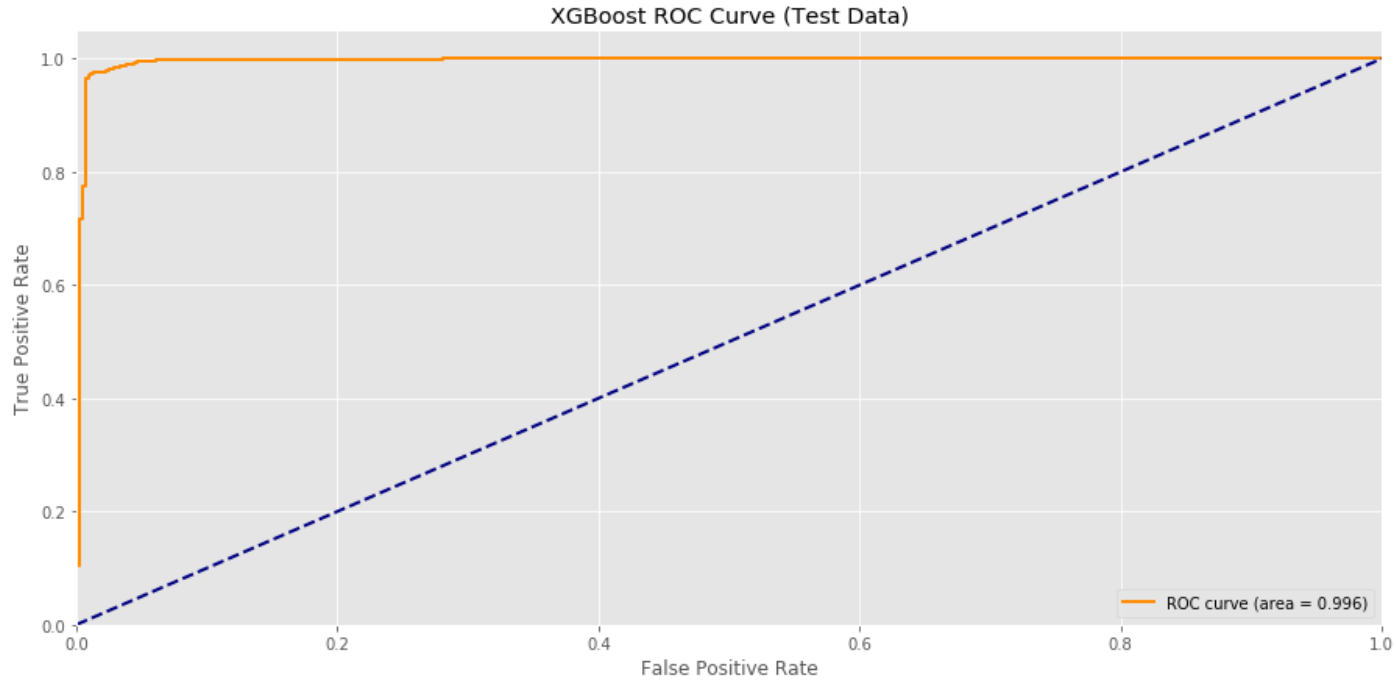

Figure 3: Receiver operating characteristic curve of XGBoost on test data