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

Feature Selection, Training Models and Cross Validation Results, and Model Selection

Five different sets of features were selected to fit the algorithms to, due to their importances as calculated by SKlearn's ExtraTreesClassifier. Each set respresents the top 10, 25, 50, 75, and 100% of the most important features.

For each model a Sklearn Pipeline object was created to first impute the missing values of the data with the mean using Sklearn's SimpleImputer, then scale the values by replacing them with their z-score using Sklearn's StandardScaler, and finally fit the given model. A parameter space is then defined for each model in the form of a dictionary with important hyper-parameters to be tuned. A RandomizedSearchCV object is then created with the pipeline object, parameter space, scoring metric, and number of folds to implement in k-fold cross validation. The RandomizedSearchCV object is then fit to the training data which performs a randomized search of the best combination of hyper-parameters while applying k-fold cross validation. The average k-fold score is then calculated and used to evaluate the model.

Due to the this particular project being a binary classification problem, and due to the fact that the two classes are imbalanced, the scoring metric chosen for this project was the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC).

Overall, Logistic Regression, Random Forest, and XGBoost performed very well out of the box, with a ROC AUC score above 0.95 using only the top 10% most important features (Appendix, Figures 1-3). Support Vector Machine, was the only model that had significantly poorer perfomance, and also the only one that decreased in ROC AUC score with the increase in number of features used to train the model (Appendix, Figure 4. 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 both Random Forest and XGBoost, ROC AUC score did not increase with more than the top 75% of features.

XGBoost Performance on Test Data

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 was calculated and plotted showing 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 (Appendix, Figure 3). This makes sense as there are nearly twice as many more false positives to train the data on as there are confirmed exoplanets.

Appendix

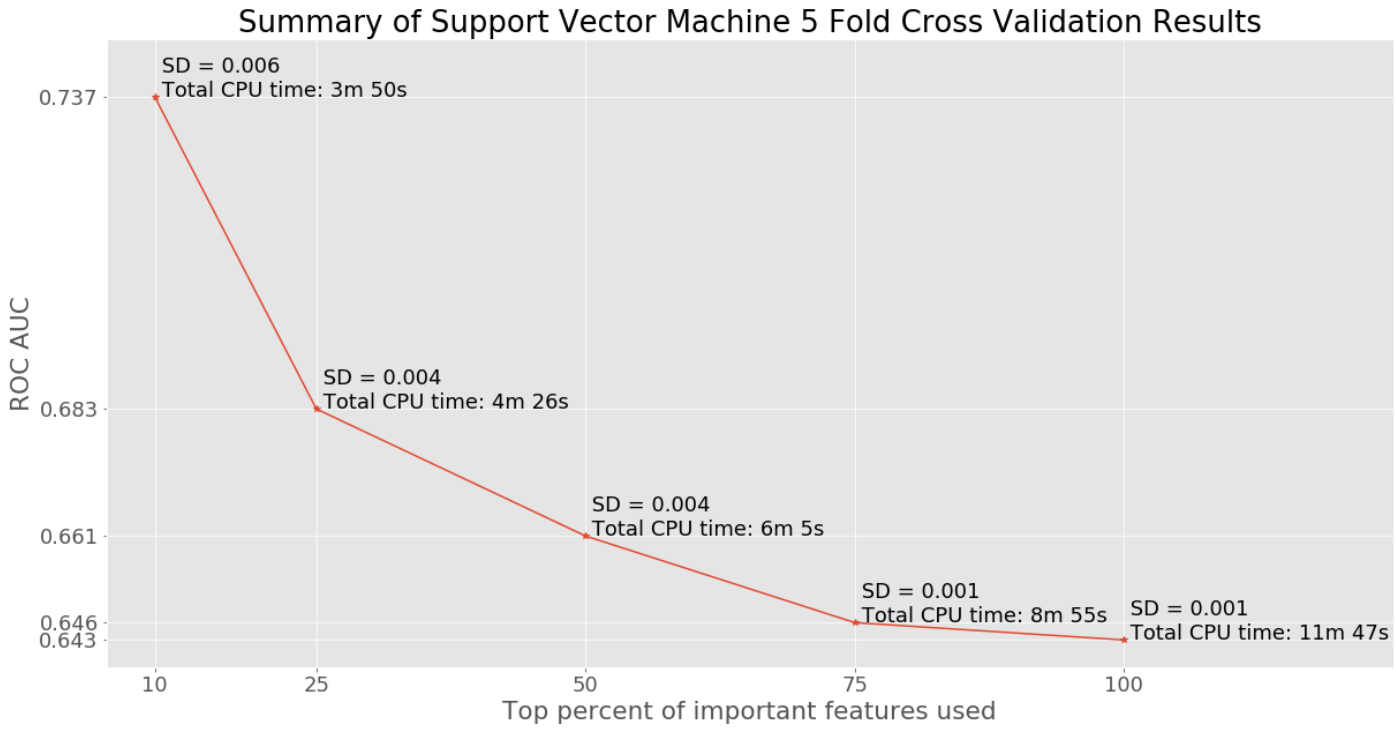

Figure 1: Support Vector Machine results

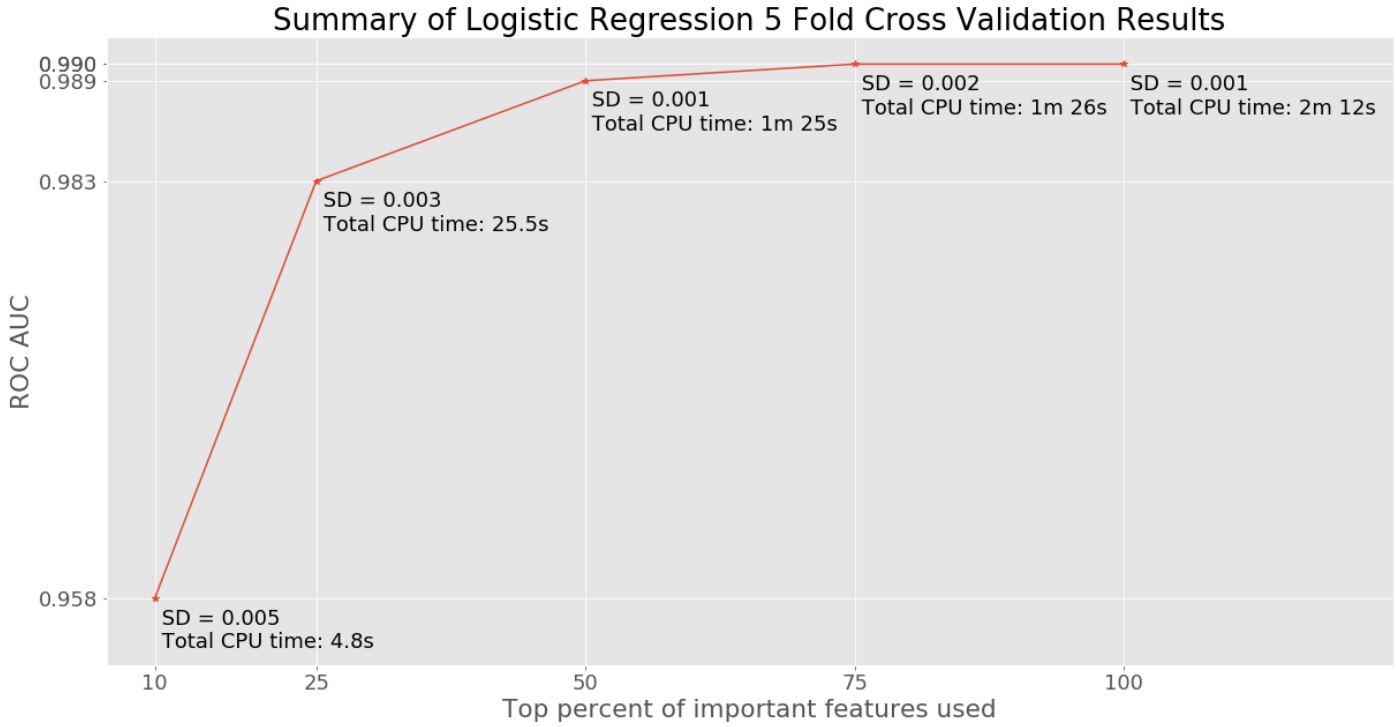

Figure 2: Logistic Regresion results

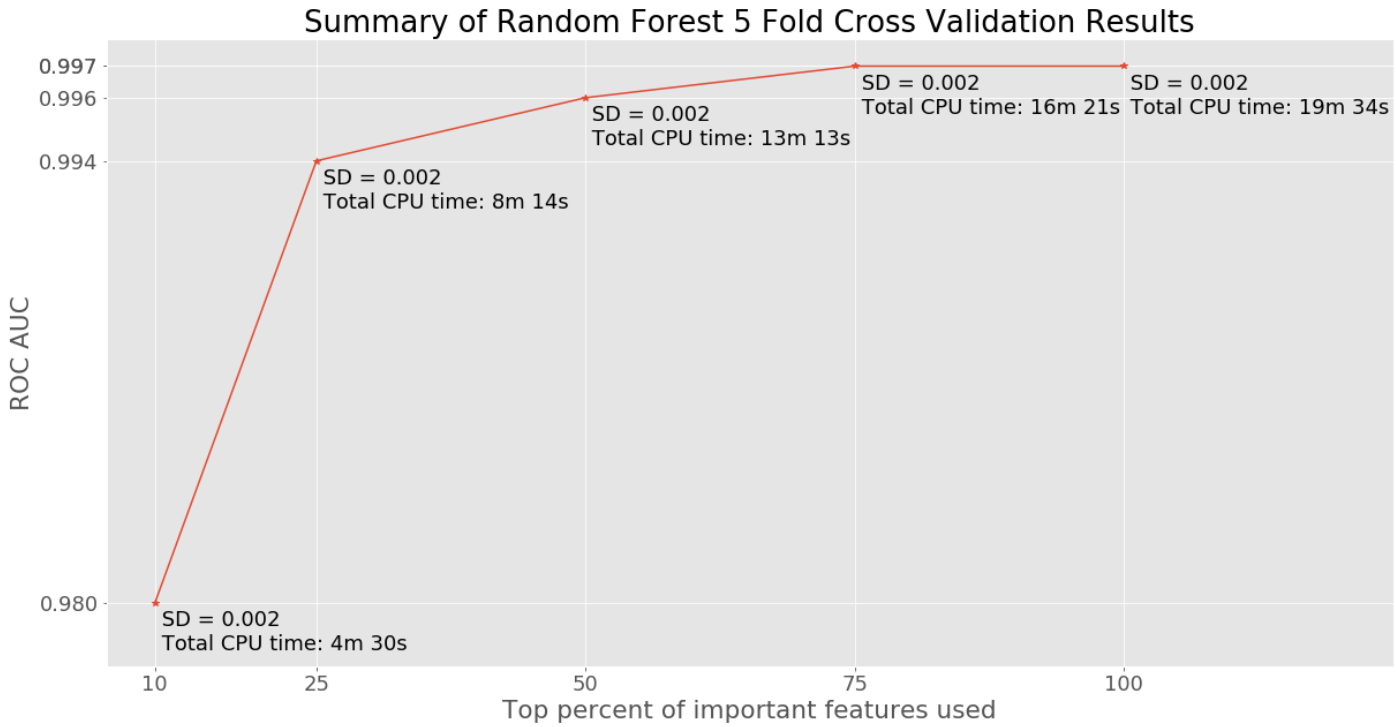

Figure 3: Random Forest results

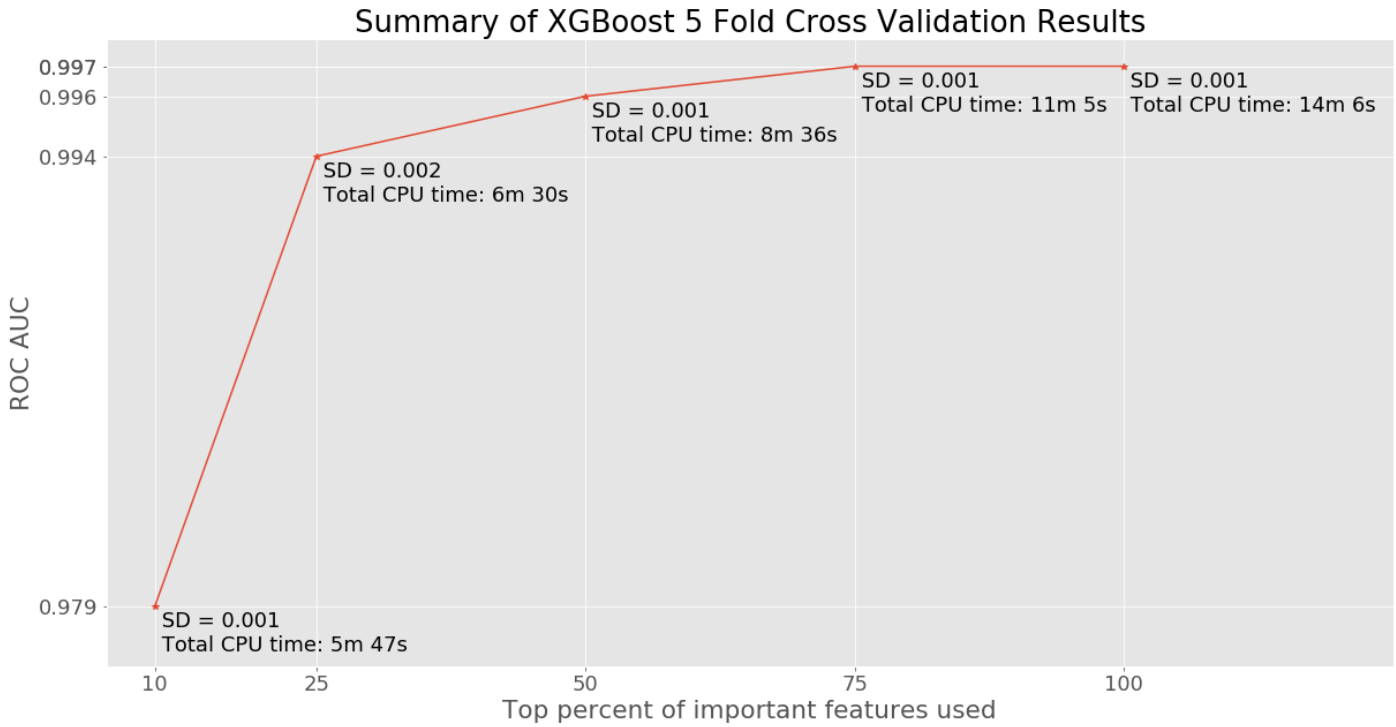

Figure 4: XGBoost results

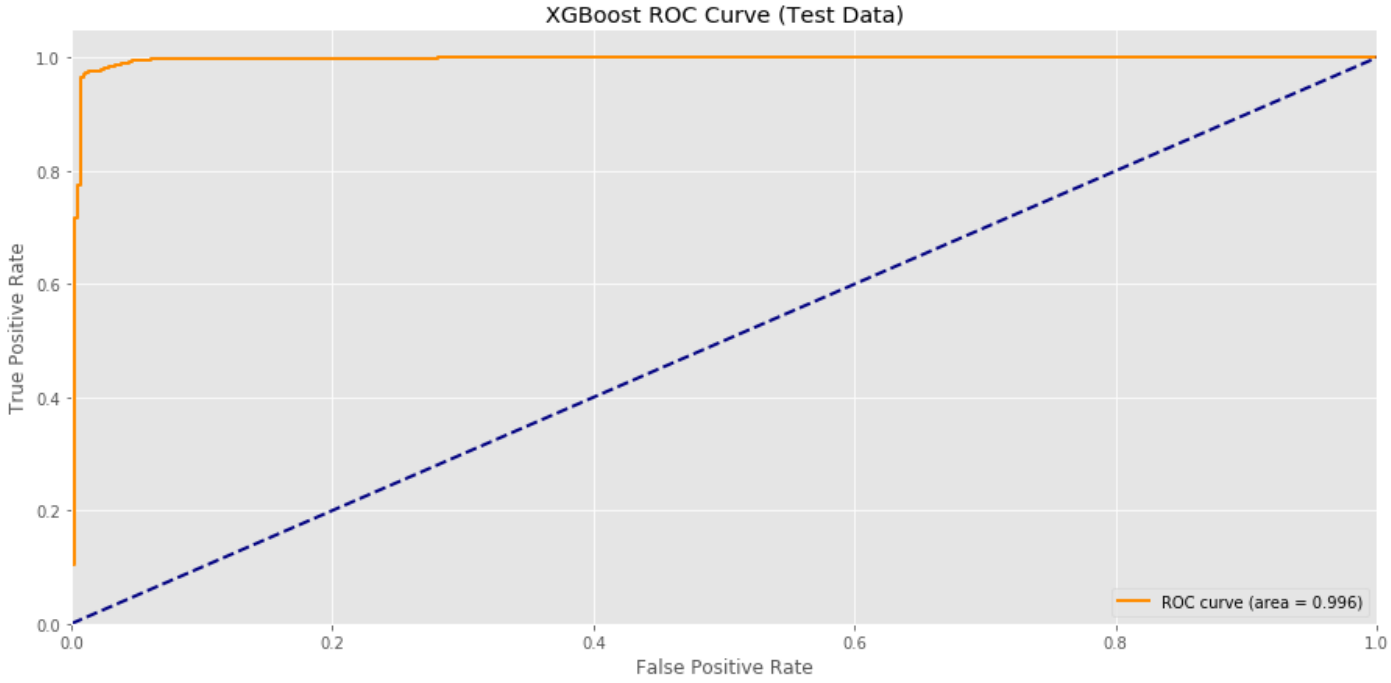

Figure 5: Receiver operating characteristic curve of XGBoost on test data

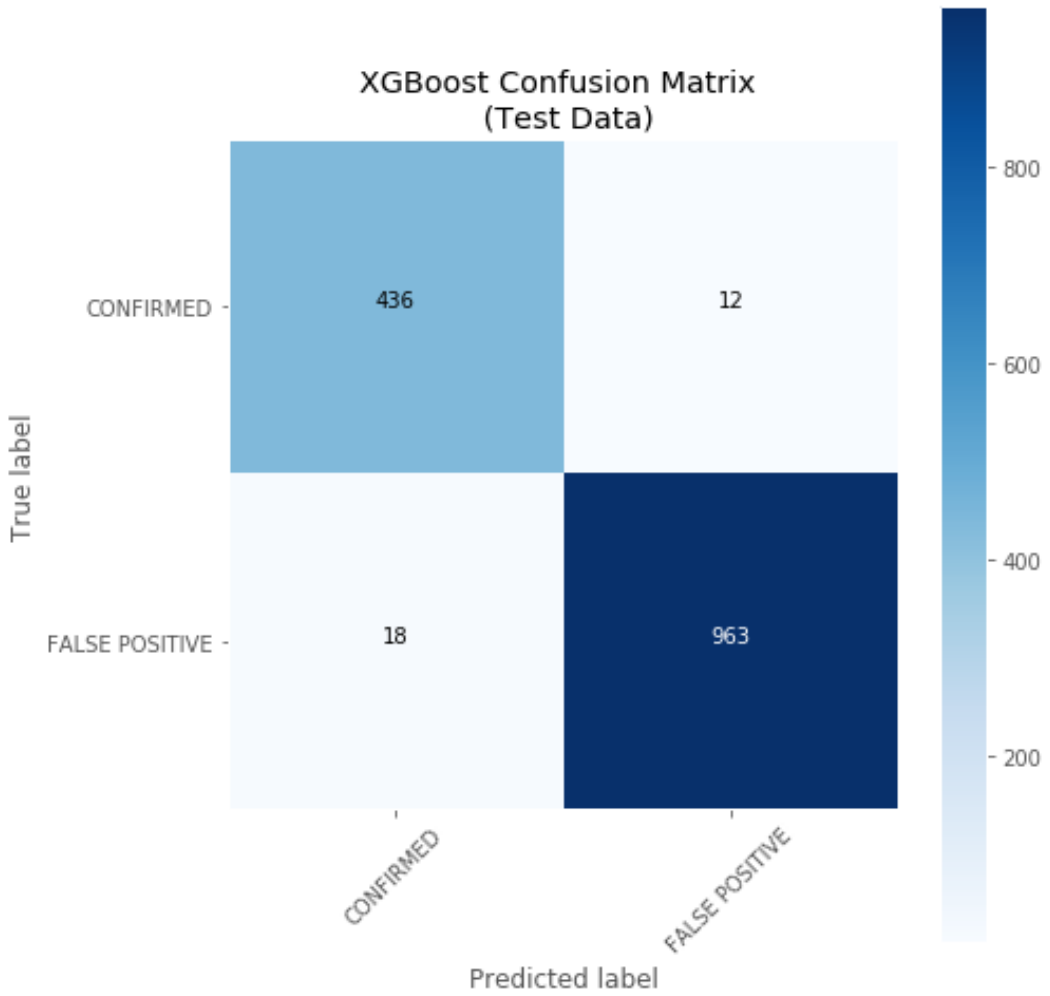

Figure 6: Confusion matrix of XGBoost prediction results on test data