Classification of Point Sources Using Machine Learning

Jenica Andersen

Abstract

The map of space contains millions of point sources of light that have been catalogued but not classified. This project created models to classify the point sources as either star, galaxy, or quasar. The aim was to identify which model and associated parameters are best for performing this task, and to fit all of the data to the best performing model. A GridSearchCV-tuned Random Forest model outperformed Logistic Regression and k-Nearest Neighbors models. The final model produced an ROC AUC score of 0.9983 on the test data (compared to a baseline Logistic Regression ROC AUC score of 0.77). Application of the final model to the multi class data produced an F1 score of 0.974. Future steps include inspecting misclassifications, tuning the models to the multi class dataset, and trying stacked model for score improvement.

Design

This project aimed to classify celestial point sources as stars, galaxies, or quasars. Ultimately aiming to develop a tuned and highly successful multiclass classification model, the initial stages focused on success with a binary model. The star label was considered the positive class, while galaxy was the negative and quasar was omitted from the initial model development. Three types of models were tuned using RandomizedSearchCV for initial tuning, then further hyper parameter honing using GridSearchCV. The models used were kNN, Logistic Regression and Random Forest.

Data

The was obtained via Kaggle. It contained 100,000 rows, 17 features and 1 target variable (‘class’). Galaxies made up about 60% of the point sources, Stars about 22% and quasars about 18%. No special handling of imbalanced data was applied.

Algorithms

Algorithms include SciKitLearn’s kNN, Logistic Regression, Random Forest. Pipeline was used to organize the models scale the data and tune parameters, using RandomizedSearchCV and then GridSearchCV.

Tools

The tools used include Jupyter Notebook and a collection of relevant python libraries, especially SciKitLearn libraries, matplotlib, pandas, and Seaborn.

Communications

Please see the accompanying slide deck and Jupiter notebook for more results and more information about this project.