Classification Project MVP

Predicting Space Objects (Star, Galaxy, Quasar)

Hayat Aldhahri & Juri Alsayigh 10/26/21

Objective

The aimed output of the prediction model would have a throughout report with expected or predicted classification of the type of space object whether it’s a Star, Galaxy, or Quasar, where this information will be useful for our clients (astrophysicists, researchers, space observers) where they are going to use those findings to focus on actual research rather than wasting time on identifying space objects. This would be achieved by Building classification models then selection and evaluation the best model using proper validation and testing methods on the SDSS data.

MVP Goal

The original data sourced from SDSS containing 18 features. Since some of the columns were not critical and important for EDA and model budling, they were dropped. These include ('objid','rerun','specobjid','fiberid'). We started exploring the data using and the following heatmap illustrates correlation between the features.

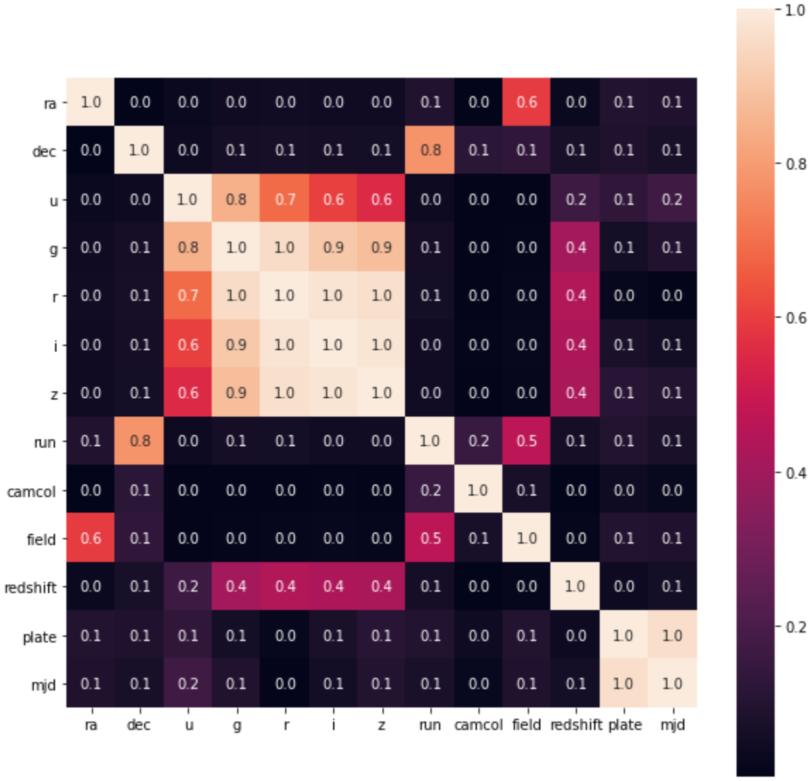

Figure 1: Features Heatmap

The below chart presents the class distribution for the acquired data. It can be clearly seen that the majority of the data are classified as STAR and GALAXY. Quasar, labeled as QSO, is relatively low in data counts. This means that the classes of our target variable were imbalanced. To solve this, we used “smote” over sampling method where we ended up with highly balanced data.

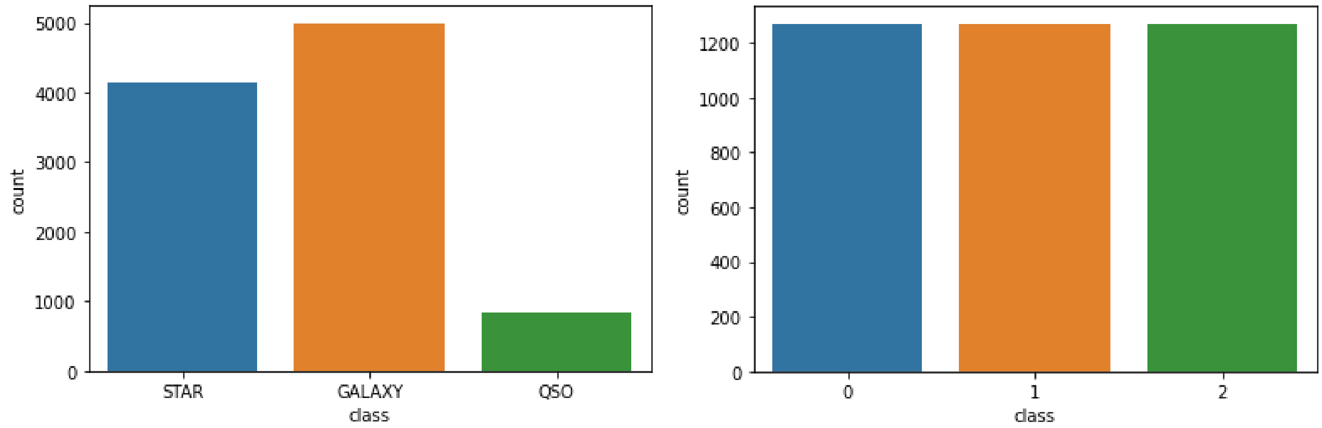

Figure 2: Data Classification Count (Right: Before, Left: After Smote Over Sampling)

KNN is our baseline model where we faced some problems regarding the imbalance, which was fixed with smote over sampling.

Additionally, panda profiling was created to explore and understand the data, which will be included in the final code and report.