Predicting Celestial Bodies Using their Spectral Properties

Souha Kabtni




Who is my stakeholder, and what problem am I solving for them?

The categorization of stars, galaxies, and quasars based on their spectral properties is a crucial concept in the field of Astronomy. By dividing stars into different categories based on factors such as their temperature, luminosity, and chemical composition, we can gain insight into their physical properties and evolutionary stages. Therefore, in this project, I will employ several classification algorithms to help classify observations of space to either stars, galaxies, or quasars based on their observed properties.

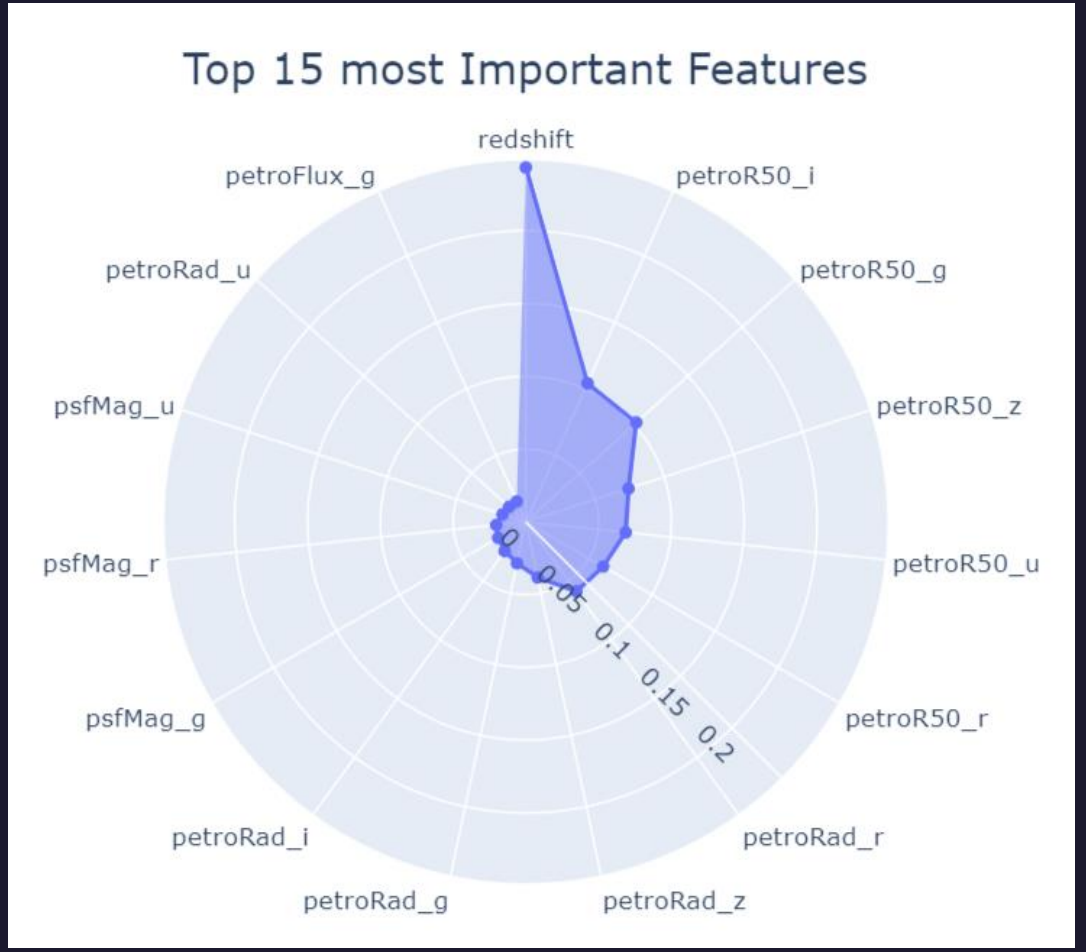
A brief introduction to my data:

- The Sloan Digital Sky Survey (SDSS) is one of the largest, most detailed, and most often cited astronomical data of space observations surveys that has ever existed, with the goal of expanding our understanding of the large-scale evolution and structure of the universe, the formation of celestial bodies.
- Data on Celestial Objects from the Sloan Digital Sky Survey - Data Release 18: Click here to download the dataset.
- For this dataset, there were 100,000 rows and 43 columns.

1st visual that demonstrates key findings of interest to my stakeholder:

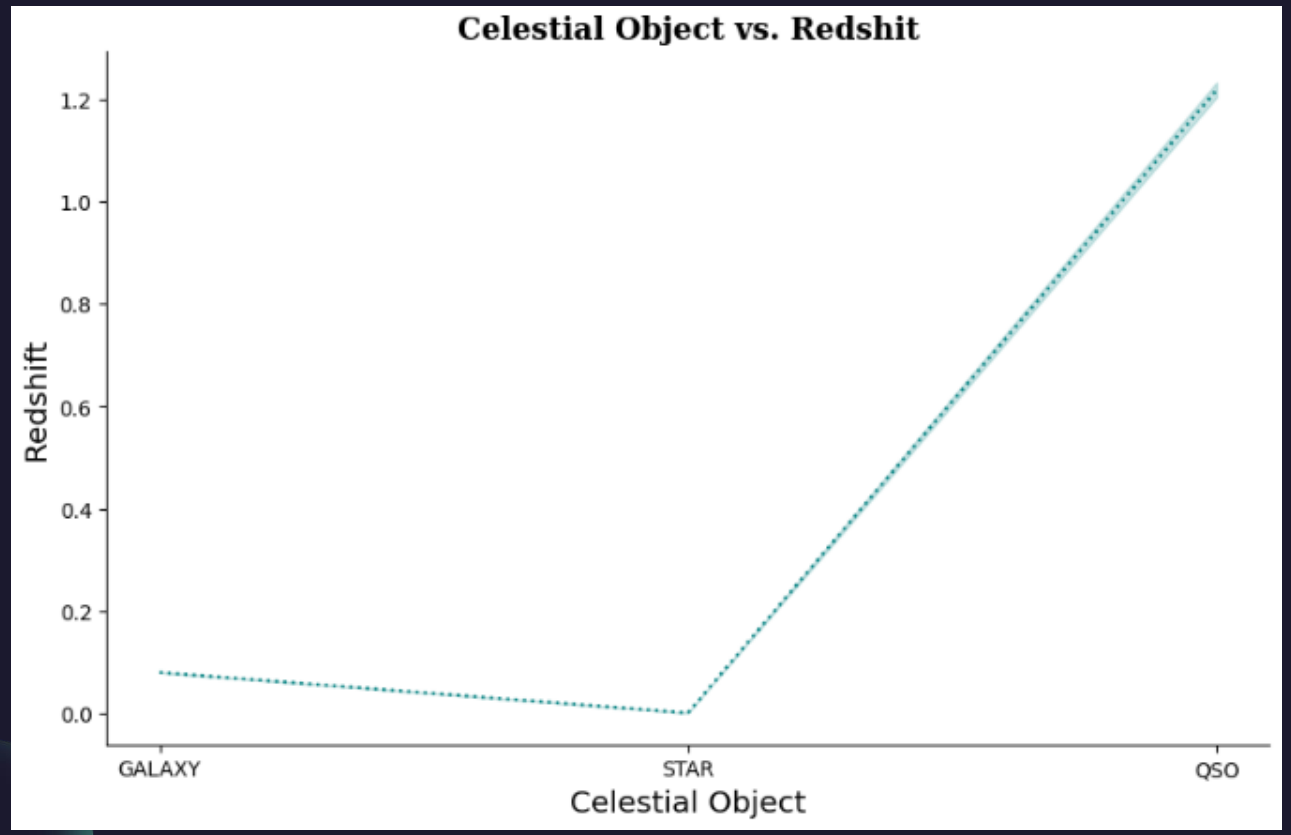
This scatter polar, clearly shows that redshift is the most important feature that helped classify the data. So, this is the best in terms of splitting ability. *PetroR50_i, PetroR50_g, PetroR50_z, PetroR50_u,, and PetroR50_r are the following most important features that impact the distinguishing and hence predicting the 3 celestial bodies.

*PetroR50_(Photometric band): These parameters are important for characterizing the sizes of objects in the SDSS data and how the sizes of the objects vary at different wavelengths.



2nd visual that demonstrates key findings of interest to my stakeholder:

The line plot shows that QSO is by far the most distant celestial object for it is the most redshifted. Both GALAXY and STAR are not that distant from the Earth with close Redshift values.



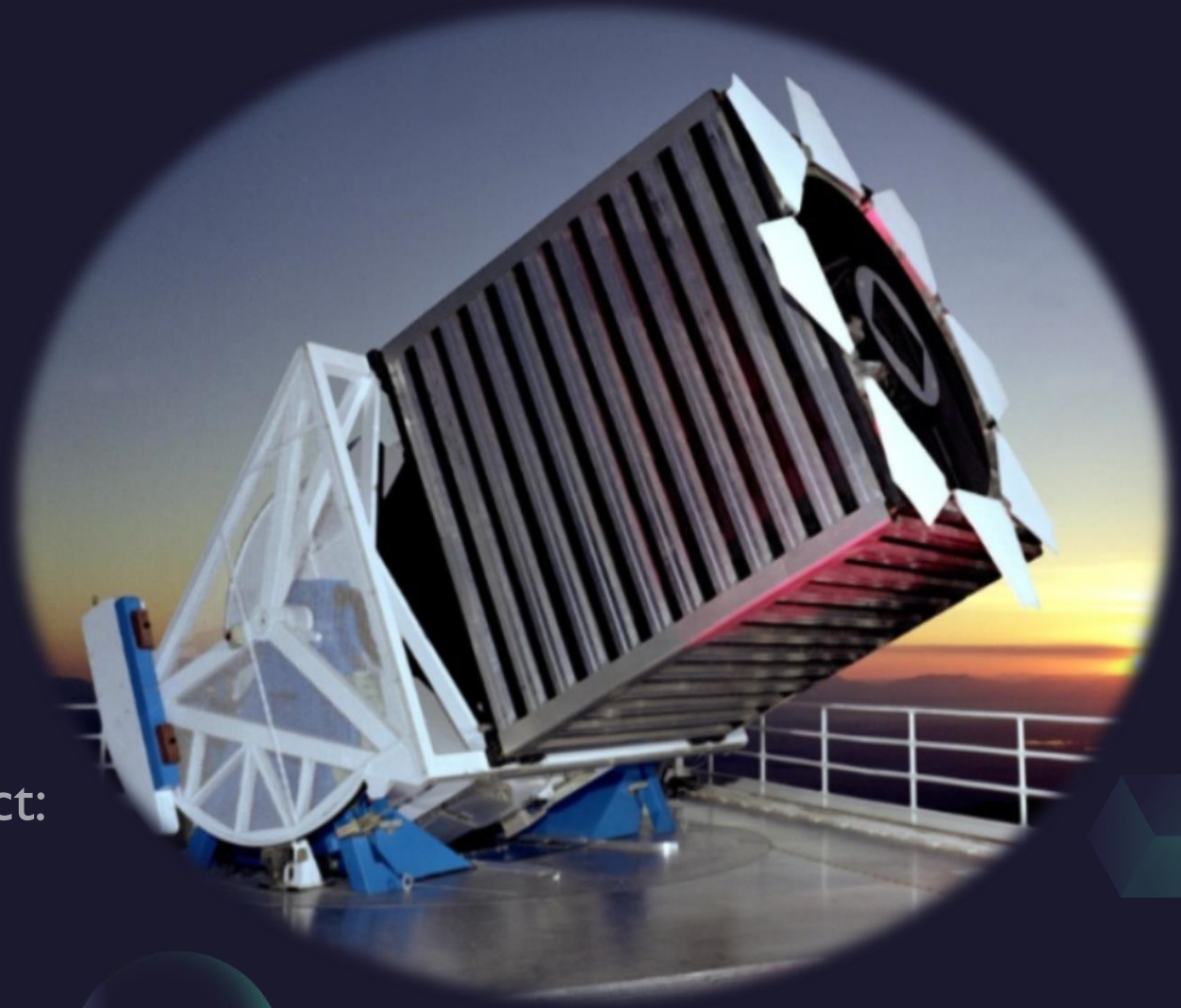
A brief description of the strengths and limitations of my model for my stakeholder:

- **Strengths:** Since no cleaning was involved (no missing values, no duplicates, no inconsistent data, …), all the time was dedicated to the Modelling phase. Besides, all created models are accurate enough to classify these celestial objects.
- **Limitations:** The ratio of the data is drastic: 52343 rows (52.343%) are classified as galaxies, 37232 rows (37.232%) are classified as stars, and only 10425 rows (10.425%) as quasars. Although SMOTE was used to handle the imbalanced classes, it still "fail[s] to reach high levels of recall while creating undue complexity for the machine-learning pipeline." (From Medium Article entitled: Stop Using SMOTE to Treat Class Imbalance). So, other techniques might be used to better address class imbalance.

Final recommendations based on my analysis:

- All created models can be used in the future to classify these celestial objects with great accuracy;
- Although delivered the best results, XGBoost was not tuned. There is still potential for XGBoost to deliver more accurate predictions;
- Deep Learning models were not 'heavily' tuned, and if so, better results might be expected;
- As there is data about the camera as well, we can perform analysis on them as well;
- Some feature engineering might be involved.

THANK YOU FOR YOUR ATTENTION



For any additional questions, please contact:

souha.kabtni.data@gmail.com