|  |  |
| --- | --- |
| **DATA 430 Technical Report Assignment 1 (a & b): Logistic Regression** | **Sulchan Yoon** |
| **Classification of Stars, Galaxies, and Quasars** | |
| **URL to dataset: https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17** | |

**Assignment 1a (due Week 2):** you should complete the following sections ONLY:

* Overview (Problem Domain)
* Overview (Objective)
* Analysis (Exploratory Analysis)

**Assignment 1b (due Week 3):** all sections of this template should be completed. Modifications of the three sections submitted in Assignment 1a should be made based on feedback from the instructor.

This template should be used in conjunction with the assignment instructions. The size of the text area below will expand to the length of your response; the area should not be interpreted as a required or suggested length of response. Responses within the text area should be single spaced with Times New Roman 12pt font. The body of the document will likely be 6-9 pages, not including the Appendix; length may vary depending on specifics of the analysis and the dataset. As needed, APA format in-text citations should be included, along with a full references list at the end of the document.

|  |
| --- |
| **Overview** |
| **Problem Domain**: give some background and context about the problem domain (application area). For instance, if you are doing the analysis for predicting heart disease, provide some context about the disease and include some interesting statistics about it. Also, discuss how the method is relevant for the chosen problem. |
| Astronomy and Space has been a major topic in recent years with SpaceX bringing exploration much more into the public domain. Observations through a variety of tools and instruments has allowed individuals to classify stars based on their spectral characteristics. With the data gathered through these tools, data scientists, and astronomers have been able to group galaxies into a variety of classifications, allowing us to find those that may be similar to the Milky Way that we are in. Galaxies and star clusters such as the popular Andromeda Galaxy and Pleiades Star Cluster have been those that came about these classification methods.  This is by no means the only method of classification, however it is one set of tools scientists and researchers are able to continue understanding the vast universe in search of potentially a new livable planet. In an Astrophysics sense, we can use these metrics to support how actions and reactions occur as it relates to our known understanding of the Laws of Physics. In Astrometry terms, determining classifications through different filters and wavelengths help map out celestial bodies to better understand things such as a black hole, or dark matter. We also learn that our own galaxy, the Milky Way in its current form is approximately 100,000 light years across, and the closest solar system similar to ours would be Andromeda Galaxy at around 2 Million light years away.  With this information, we want to find a more streamlined method to classify objects and would look to have a predictive machine learning algorithm to support the need. |
| **Objective**: clearly state the objective of the analysis in relation to the kind of algorithm you are employing. Use specific language as to what question(s) you are trying to answer using the specific analysis/modeling type. |
| The purpose of this report is to analyze the given information to find trends of the information as part of exploratory data analysis. We will then go deeper into the data and attempt to produce a viable machine learning algorithm to predict class types. Our first attempt will be through Logistic Regression, then we compare that to other methods such as decision trees, and Bayesian Classifications.  Overall, our purpose is to use these data elements in a way to find high value machine learning techniques to support astronomical research. Our predictive analysis is focused on trying to appropriately classify a celestial body (dependent variable) using the different wavelengths readings, body metrics (i.e. ascension), plate ID information (independent variables). Our first attempt of predictive analysis will be handled via logistic regression. The purpose of logistic regression is mainly designed around classification around a series of independent variables. It measures by estimating probabilities between the dependent variables and independent variables with support through logistic functions. |
| **Analysis** |
| **Exploratory Analysis**: describe the data including the source, the collection method, and variables. Perform exploratory analysis. Also, select few key variables (including the target variable for supervised learning) and study their distributions using plots such as histograms, box plot, bar chart, etc. |
| This data has been provided by the Sloan Digital Sky Survey’s 4th phase (SDSS-IV). The information presented will be based on the Data Release 17 published in 2022 as the final data release. The data has also been peer reviewed and reduced to 100,000 observations for Kaggle use. The main extractions come from the Mapping Nearby Galaxies at Apache Point Observatory survey, bringing in over 10,000 nearby galaxy observations. Data from the DR17 also includes all information from the prior releases and includes information from Apache Point Observatory Galactic Evolution Experiment 2 survey pulling in 650,000 star readings.  Data points captured through this survey have been captured through ground stations. There are a variety of different instruments that have been capturing data points such as the Hubble Space Telescope which gives us a different understanding of the universe. The value of capturing information on ground stations is to see these planets are they relate to Earth, whereas instruments in space can capture with a point of view not equal to Earth.  The total dataset includes 18 total variables, however as we go through and review each of the dataset values, we recognize that some variables are not as valuable due to its variance within each classification. Our initial exploratory data analysis includes descriptive statistics. A lot can be seen with the ID variations and noted that overall IDs are not valuable data points. We remove all ID variables and review the remaining variables in more detail and through visualizations.  We go through and create kernel density estimate (kde) graphs of each variable to find those that may have differences in its distribution. An example graph of the kde plot is shown below. You can find all graphs in the appendix along with the functions used to create the graphs.    Through each of the graphs, we find those that result in variances that we may consider to be unique to each type of Class. This will be used to help create the proper algorithm using the Logarithmic Regression in Python. Our overall result of variables we will use are shown below in the chart of Points of Interest.  Points of Interest   |  |  | | --- | --- | | Variable | Definition | | Alpha | Right Ascension Angle at J2000 epoch | | Delta | Declination Angle at J2000 epoch | | U | Ultraviolet Filter | | G | Green Filter | | R | Red Filter | | I | Near Infrared | | Z | Infrared Filter | | Class | Object Class | | Redshift | Redshift Value on the increase in wavelength | | Plate | Plate ID in SDSS | | MJD | Modified Julian Date | |
| **Preprocessing**: armed with the exploratory analysis, perform the necessary preprocessing, both general and specific types appropriate for the modeling type being employed. |
|  |
| **Model Fitting**: explain the key steps and activities you perform to fit the model. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate ones is the amount of performance tuning performed. |
|  |
| **Results** |
| **Model Properties:** explain the components of the fitted model and their characteristics. Leverage functions to summarize the model properties. Also, leverage visualization as required. |
|  |
| **Output Interpretation**: explain the result and interpret the final model output using terms that reflect the application area and in relation to the stated objective. This is where you check whether or not the stated objective is met. |
|  |
| **Evaluation**: employ appropriate metrics to quantitatively evaluate the performance of the fitted model. For supervised classification, this includes simple accuracy, precision & recall (or sensitivity & specificity), all of which can be generated from a confusion matrix, or ROC. |
|  |
| **Conclusion** |
| **Summary**: highlight the main findings in relation to the stated objective. You don’t need to discuss the details of the analysis and the model such as accuracy here, just focus on the key findings. |
|  |
| **Limitations & Improvement areas**: discuss the limitations of the analysis and identify potential improvement areas for future work. This could be related to the data, algorithm, or a combination of the two. |
|  |

|  |
| --- |
| **Appendix** |
| Functions:    Packages:    Plots: |

**References**

Abdurro’uf et al., The Seventeenth data release of the Sloan Digital Sky Surveys: Complete Release of MaNGA, MaStar and APOGEE-2 DATA (Abdurro’uf et al. 2022 ApJS 259, 35) [arXiv:2112.02026]

fedesoriano. (January 2022). Stellar Classification Dataset - SDSS17. Retrieved June 2023 from https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17.

Yoon. (2023). *Data\_430\_GC Project1* [ipynb]. Retrieved from https://github.com/sulchan/Data\_430\_GC