Assignment 2

Fundamentals of Data Science Lab

Session: Jan - May 2025

Programme: BTech. CS - Data Science

Sem: 4
Batch: 5

Submitted By:

Name: Kshitij Chandrakar
SAP ID: 500124827

Submitted To:

Dr. Sachi Chaudhary

1 Dataset Information

Dataset Chosen: APOGEE2 from SDSS
`https://www.sdss4.org/dr17/irspec/spectro_data`

Objective: Identifying Stars from the Halo of Our Galaxy Based on Various Stellar Parameters

Description and Selection Process We choose the dataset as it has good support with multiple Value Added Catalogs, Was released relatively recently so it would have more accurate Information, it is Well Documented and has Features which would help our Objective.
It is a part of The Sloan Digital Sky Survey. They have created the most detailed three-dimensional maps of the Universe ever made, with deep multi-color images of one third of the sky, and spectra for more than three million astronomical objects.

Hypothesis From the current theory, we can Hypothesise that the Halo stars would have the following properties.

- Spatial Distribution with an angled Galactic Latitude
- Metal Poor $[Fe/H] < 1$
- Highly Turbulent Motion
- Highly eccentric and inclined orbits
- alpha-element enhancement (O, Mg, Si)

2 Exploratory Data Analysis

On doing EDA we find various trends, the following are the Observations.

2.1 Data Density and Galactic Coordinates

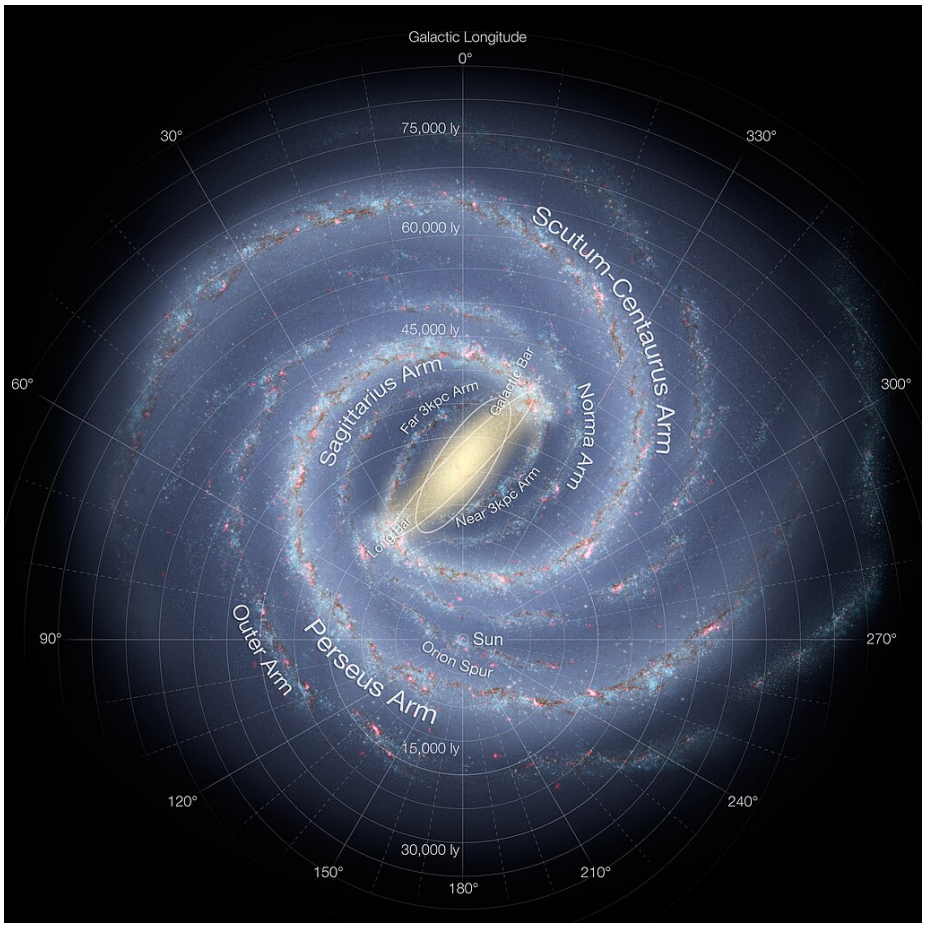

Figure 1: Galactic Coordinate System

The Artist's depiction Above of the Milky Way Galaxy showing the origin and orientation of galactic longitude. The galactic longitude (l) runs from the Sun upwards in the image through the center of the galaxy. The galactic latitude (b) is perpendicular to the image (i.e. coming out of the image) and also centered on the Sun.

From a the APOGEE Targetting Field Map and Density plot of the data points we can see that the data is slighly biased as we physically cannot observe the stars on the opposite side of our galaxy. However, the stars we can observe are well recorded. with only a few peaks in the interesting regions like the Magellanic Clouds.

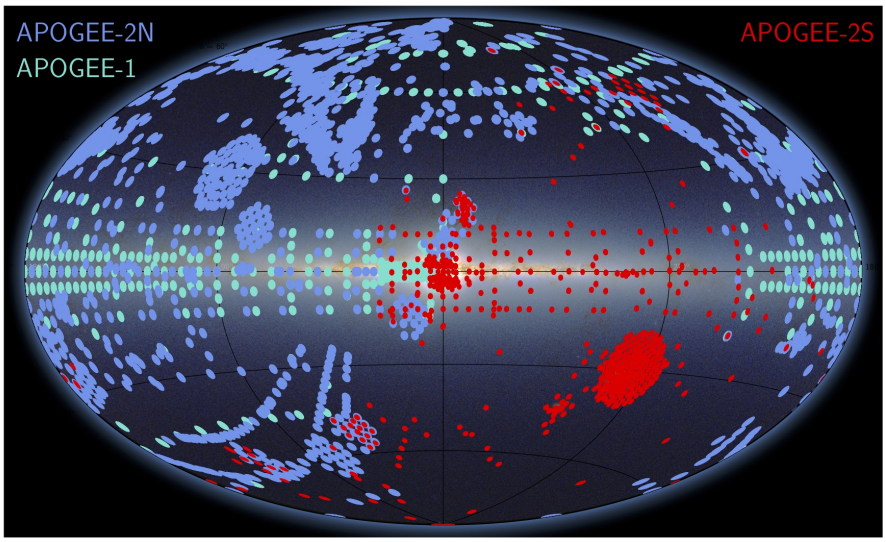

Figure 2: DR17 APOGEE field map where the fields are color-coded by the APOGEE sub-survey. APOGEE-1 in cyan, APOGEE-2N in blue, and APOGEE-2S in red. Figure by C. Hayes. Background image from 2MASS.

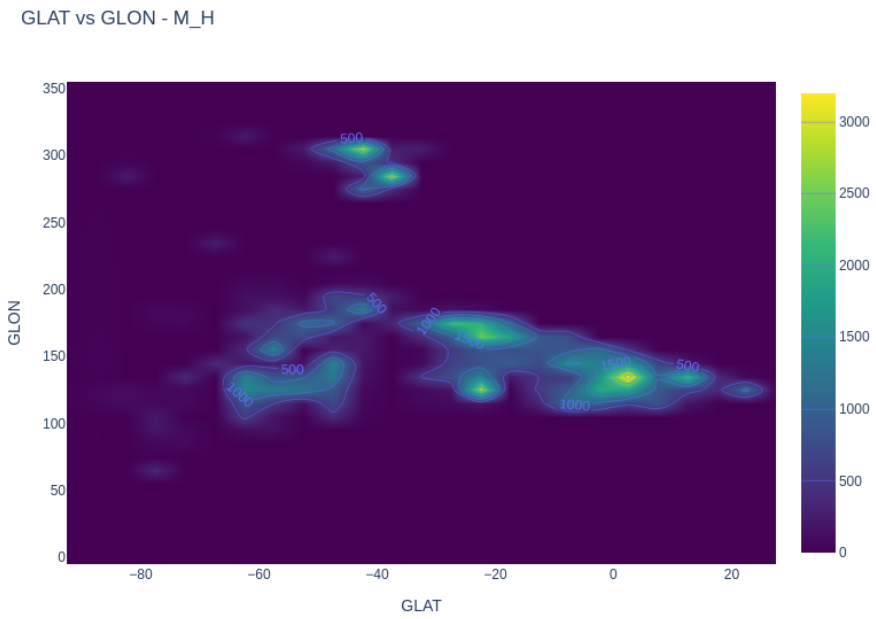

Figure 3: Density Plot

2.2 HR Diagram

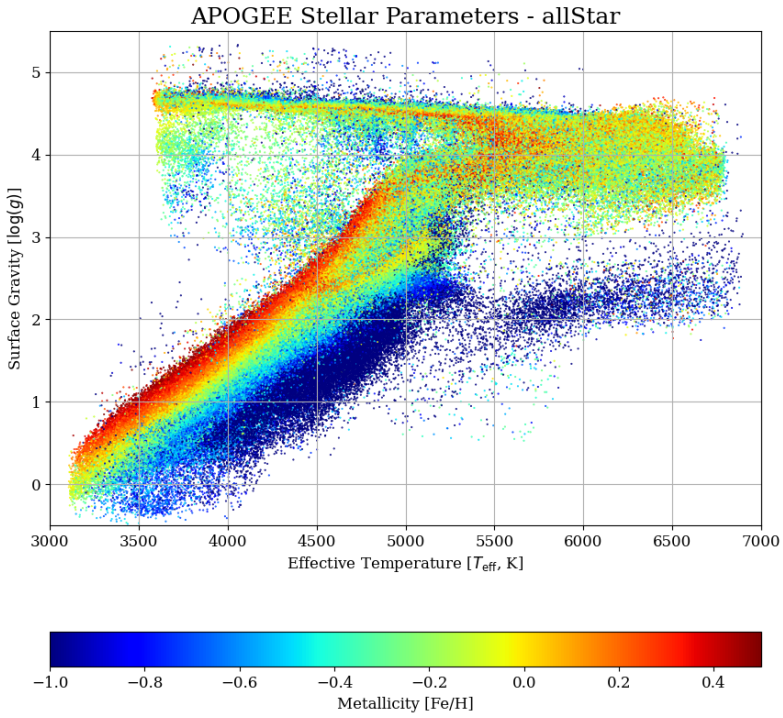

Figure 4: HR - Diagram

From a scatterplot between Effective Temperature at Surface T_{eff} and the Surface Gravity $log(g)$ we can see that:

- the Relationship is mostly linear upto $T_{eff} < 5000K$ and $log(g) < 3$, this is known as the main sequence of stars, which is most of the lifetime of the star. Any given star moves from a lower Metallicity to a Higher Metallicity as it grows Older.
- The star moves above $T_{eff} > 5000K$ and $log(g) > 3$ when its dying, converting to various types of Dead Stars such as Red Giants, Red Supergiants, White Dwarfs, Neutron Stars and Black holes depending on a

mixture of various variables.

2.3 Star Formation Regions

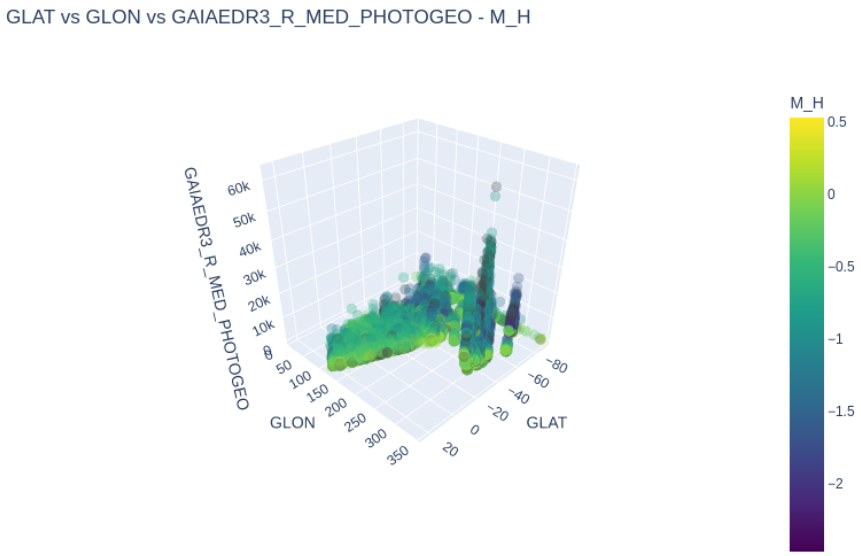

Figure 5: GLAT vs GLON vs Distance Colored by M_H

In the above Scatterplot we can Identify the various regions with a high star formation rate. They can be identified by the clumps of datapoints which are a deep blue.

The Deep-Blue Signifies a Low Metallicity which means that the particular star is Young.

3 Feature Selection

Techniques Applied:

- Correlation Matrix
- Random Forest Selection
- Lasso Regression

3.1 Correlation Matrix

We create the Covariance Matrix Heatmap and Remove Highly Correlated Features with $|r| > 0.85$:

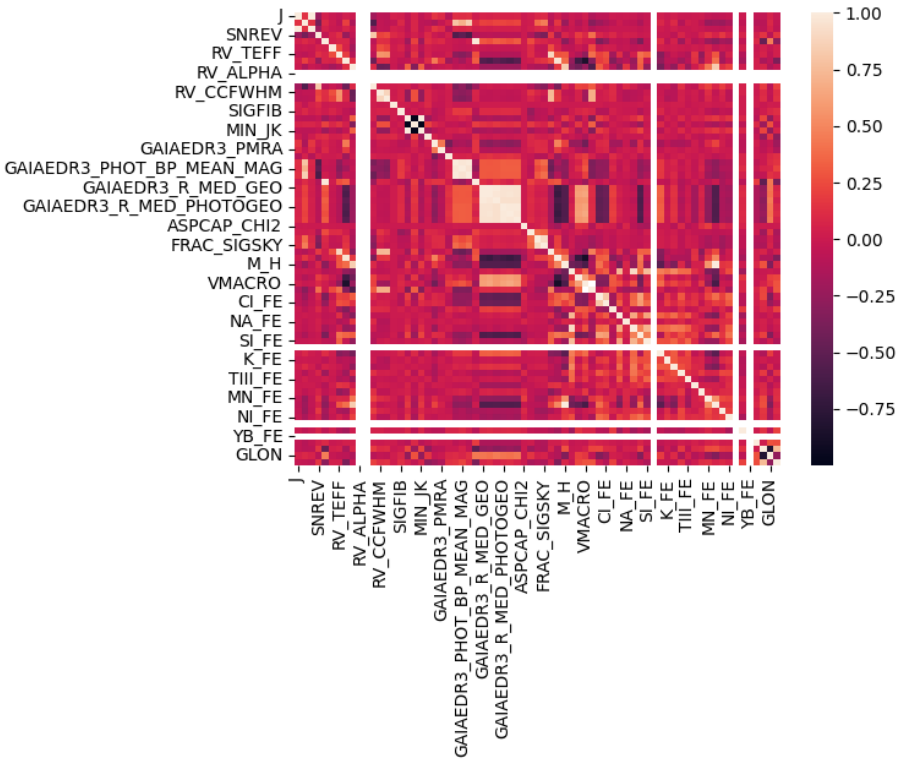

Figure 6: HR - Diagram

After Dropping we get the following columns left:

'J', 'H', 'K', 'SNREV', 'VHELIO_AVG', 'VSCATTER', 'RV_TEFF', 'RV_LOGG', 'RV_FEH', 'RV_ALPHA', 'RV_CARB', 'RV_CHI2', 'RV_CCFWHM', 'RV_AUTOFWHM', 'MEANFIB', 'SIGFIB', 'MIN_H', 'MAX_H', 'MIN_JK', 'MAX_JK', 'GAIAEDR3_PARALLAX', 'GAIAEDR3_PMRA', 'GAIAEDR3_PMDEC', 'GAIAEDR3_PHOT_G_MEAN_MAG', 'GAIAEDR3_PHOT_BP_MEAN_MAG', 'GAIAEDR3_PHOT_RP_MEAN_MAG', 'GAIAEDR3_DR2_RADIAL_VELOCITY', 'GAIAEDR3_R_MED_GEO', 'GAIAEDR3_R_LO_GEO', 'GAIAEDR3_R_HI_GEO', 'GAIAEDR3_R_MED_PHOTOGEO', 'GAIAEDR3_R_LO_PHOTOGEO', 'GAIAEDR3_R_HI_PHOTOGEO', 'ASPCAP_CHI2', 'FRAC_BADPIX', 'FRAC_LOWSNR', 'FRAC_SIGSKY', 'TEFF', 'LOGG', 'M_H', 'ALPHA_M', 'VMICRO', 'VMACRO', 'VSINI', 'C_FE', 'CI_FE', 'N_FE', 'O_FE', 'NA_FE', 'MG_FE', 'AL_FE', 'SI_FE', 'P_FE', 'S_FE', 'K_FE', 'CA_FE', 'TI_FE', 'TIII_FE', 'V_FE', 'CR_FE', 'MN_FE', 'FE_H', 'CO_FE', 'NI_FE', 'CU_FE', 'CE_FE', 'YB_FE', 'RA', 'DEC', 'GLON', 'GLAT'

3.2 Lasso L1 Regression

We do Lasso Regression with $\alpha = 0.1$ and get the following features with Non-Zero Coefficients. We do not expect this to be accurate as L1 regression is not accurate for non linear data.

Feature	Coefficient
MIN H	-9.317250e-08
MIN JK	-1.959392e-11
GAIAEDR3 R LO PHOTOGEO	3.420404e-09

3.3 Random Forest

We do a Random Forest Feature selection with $n_estimators = 100 and random_state = 42$ and get the following features. We expect this method to give us a more accurate result of what features are the most important as it works well with multivariant and non-linear data.

we can see that the most important features are GLAT, VMACRO, M/H, F/H, N/FE and VSCATTER; which tracks with our objective of mapping Halo stars.

Feature	Importance
Feature	Importance
GLAT	0.063374
VMACRO	0.048827
M H	0.039412
FE H	0.033645

Feature	Importance
Feature	Importance
N FE	0.033473
VSCATTER	0.030167
RV FEH	0.027783
DEC	0.025858
AL FE	0.025106
FRAC LOWSNR	0.023728
NA FE	0.023259
O FE	0.022766
CA FE	0.022734
MN FE	0.02113
MEANFIB	0.021125
ASPCAP CHI2	0.020697
V FE	0.020603
K	0.019685
C FE	0.019623
FRAC SIGSKY	0.019442
GAIAEDR3 R MED PHOTOGEO	0.018614
H	0.018061
RV CCFWHM	0.017715
CO FE	0.01756
SIGFIB	0.017358
SNREV	0.016884
SI FE	0.016617
RV AUTOFWHM	0.015768
GLON	0.014479
RV LOGG	0.013618
GAIAEDR3 PMRA	0.013563
CI FE	0.013318
GAIAEDR3 R LO GEO	0.012231
RV CHI2	0.012108
NI FE	0.011644
MAX H	0.011132
MG FE	0.010685
GAIAEDR3 R HI GEO	0.01064
GAIAEDR3 R LO PHOTOGEO	0.009849
TI FE	0.009843
J	0.009653
GAIAEDR3 PMDEC	0.009627
GAIAEDR3 R HI PHOTOGEO	0.009566
CE FE	0.009226
GAIAEDR3 PHOT G MEAN MAG	0.008478
CR FE	0.008338
MIN H	0.007882

Feature	Importance
Feature	Importance
GAIAEDR3 PHOT BP MEAN MAG	0.007868
MIN JK	0.007806
K FE	0.007662
S FE	0.007464
GAIAEDR3 DR2 RADIAL VELOCITY	0.007266
RA	0.007055
LOGG	0.006129
GAIAEDR3 PARALLAX	0.005646
TEFF	0.005278
GAIAEDR3 R MED GEO	0.004983
ALPHA M	0.00475
TIII FE	0.004284
VMICRO	0.003638
RV TEFF	0.003605
GAIAEDR3 PHOT RP MEAN MAG	0.003507
FRAC BADPIX	0.003299
VHELIO AVG	0.002864
MAX JK	0
RV CARB	0
RV ALPHA	0

4 Conclusion

In Conclusion, We can see that the various stellar parameters selected by the Various Techniques reflect their own Advantages and Disadvantages, with the Random Forest Being the most accurate and suitable for our dataset. It selected the most accurate and applicable Features to fit our Hypothesis.