Stellar Spectra of Red Giant Stars

By: Kelly Wong, Macy Chen, Wenping Xiao

Abstract:

The SDSS APOGEE Survey contains data on stellar spectrums and their related information. This research investigates the 99,705 observations of stellar spectra from red giant stars and associated measurements. Specifically, we examined the effective temperature, surface gravity and abundance of different elements on the surface of the stars to look for correlations between the various features. Using bootstrap resampling, the results from the first research question suggest that the abundance of iron on most stars is lower compared to the sun. The second investigation — the relationship between the magnesium-to-iron abundance ratio and its effective temperature — shows little to no correlation between the two variables. The third topic, classifying surface gravity based on temperature and oxygen abundance, suggests that surface gravity can be reasonably predicted.

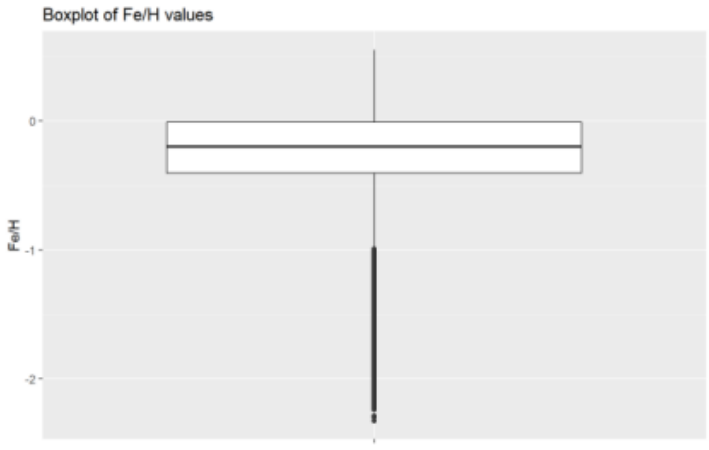
Q1. Abundance of Iron on the Surface of Stars and the Sun

Overview of Data and Methods

This investigation uses the "rhdf5" package in R Studio to read the "STA130_APOGEE.h5" file and obtain data on 99705 stars collected by the SDSS APOGEE Survey. We examined the abundance of iron on the surface of stars and the sun, that is, Fe/H (the decimal logarithmic unit of abundance relative to the same element measured on the surface of the sun).

When Fe/H is greater than 0, it means that the star has more iron on its surface relative to the sun; When Fe/H is less than 0, it means that the star has less iron on its surface relative to the sun; When Fe/H is equal to 0, it means that the star has the same amount of iron on its surface as the sun.

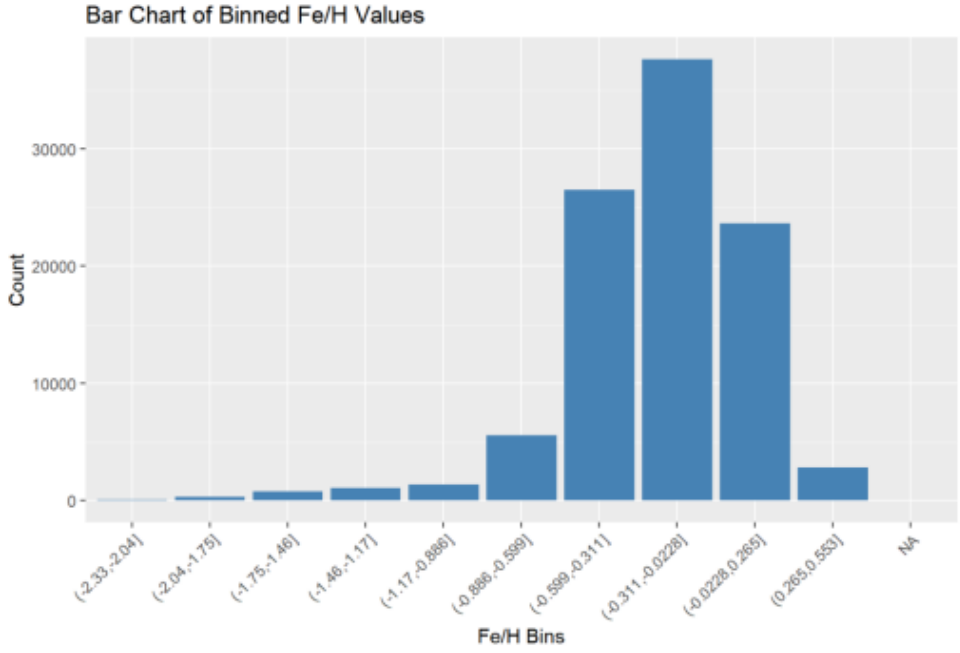
Descriptive Statistics



Indicators	Value
Min	-2.3256
The first Qu.	-0.4029
Median	-0.1994
Mean	-0.2322
The third Qu.	-0.0097
Max	0.5528

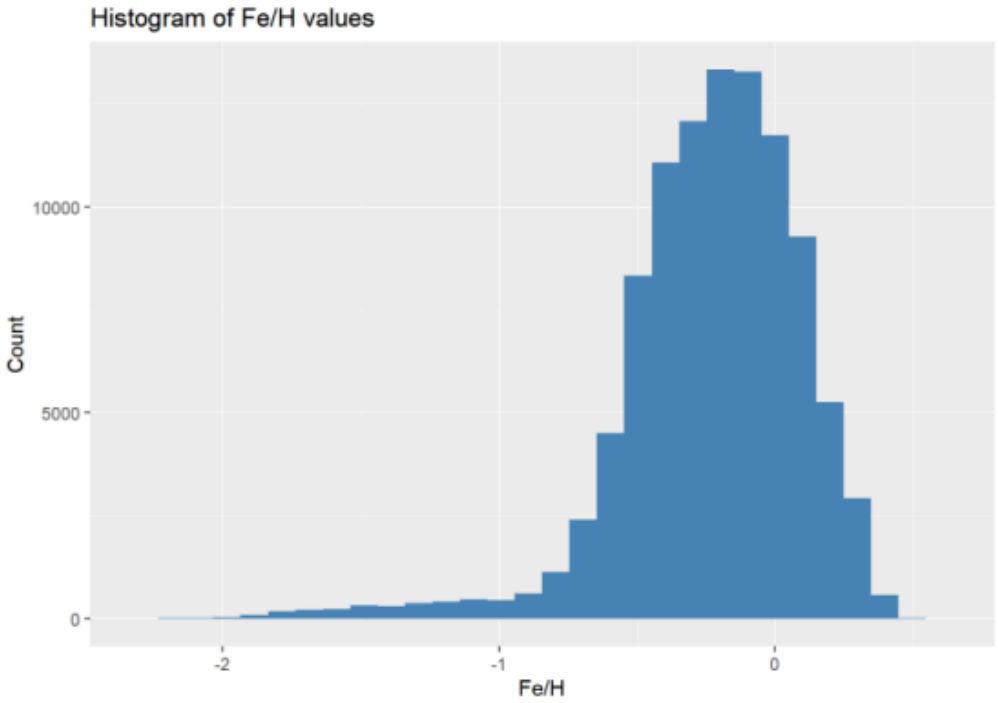
When we look at the overall data statistics, we can see that the lowest abundance of iron indicated by the stars in this data set is -2.32, meaning that the abundance of iron on the surface of this star is much smaller than that of the Sun. The mean value is -0.2322, which means that the mean value of all stars indicating iron abundance is smaller than that of the Sun. The maximum value is 0.5528, which means there are stars in this group with surface iron abundances greater than the Sun.

Data visualization



Bin	Count
(-2.33,-2.04]	52
(-2.04,-1.75]	285
(-1.75,-1.46]	746
(-1.46,-1.17]	1047
(-1.17,-0.886]	1381
(-0.886,-0.599]	5558
(-0.599,-0.311]	26520
(-0.311,-0.0228]	**37632**
(-0.0228,0.265]	23650
(0.265,0.553]	2833

According to the histogram, the abundance of iron on the surface of star 99705 in the data set is close to less than 0, meaning that some of the stars have less iron on their surface than the Sun. Most of them are concentrated at (-0.311,-0.0228]. A small fraction of the stars show iron abundances greater than 0, meaning that some stars have more iron on their surfaces relative to the Sun.



The distribution of the population Fe/H values is approximately normal. In addition, the Fe/H values are mainly concentrated in the range of -0.5 to 0. Most of the stars indicate a lower Fe abundance than the Sun.

Method

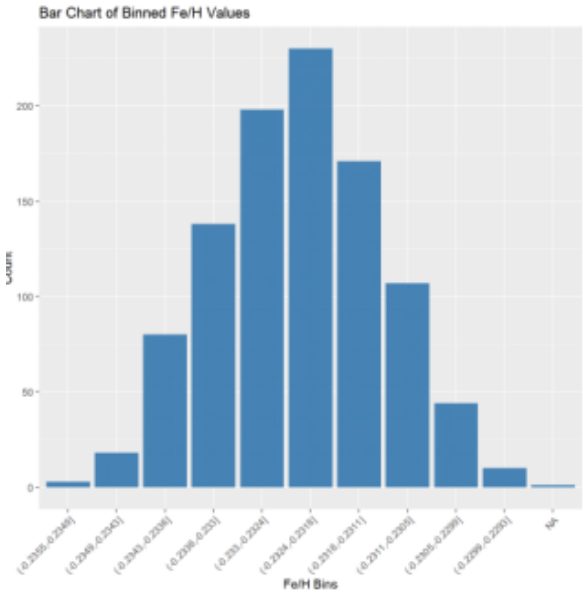
Use bootstrap resampling to get the sampling distribution for the abundance of iron. Visualize sampling distribution using a histogram. Compute the X% confidence interval to estimate the population mean. Compare the population mean to the amount of the sun.

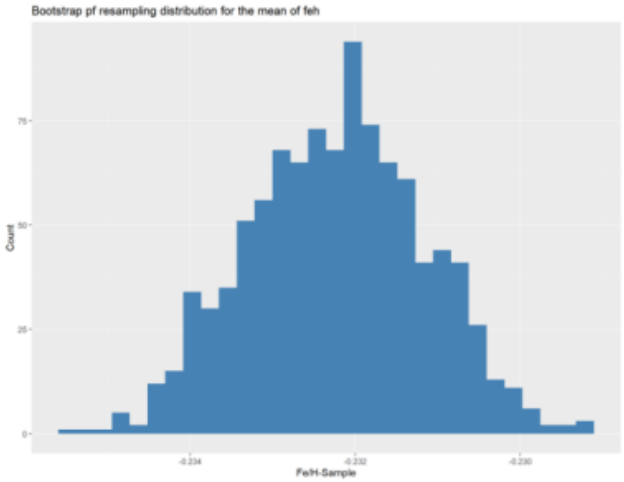
One thousand samples were taken, and the sample set feh_summary was synthesized. The statistics of the sample size were obtained as follows:

Statistic	Sample Value	Population Value
Mean	-0.232	-0.2322
SD	0.00106	0.3294
2.5%	-0.234	-1.0921
97.5%	-0.230	0.2740

Comparing the mean and variance values of the sample and the medium, we found that the

variance of the sample size is smaller and narrower in the range from 2.5% to 97.5%; however, the mean and the total are the same.





Above is the histogram of 1000 samples that are resampled, and it can be seen that the sample distribution shows a normal distribution. The abundance of iron on the surface of most of the sampled samples is lower than that of solar.

Calculate the 95% confidence interval to estimate the population mean

	Upper limit	Mean	Lower limit
Value	-0.2343454	-0.2322343	-0.2301867

Computing the interval and mean of the 1000 samples that are resampled, we know that the confidence interval for this population sample at α=5% is [-0.2343,-2302]. We can be confident that the Fe abundance (Fe/H value) of the sample stars is between -0.2343 and -0.2302. In other words, we are 95% confident that the mean value of the true Fe abundance of the sample stars lies within this interval.

Using the Sun as a reference, we set the abundance of solar-indicative iron to 0 and calculate the mean value of iron abundance on the surface of the sampled star to be -0.2322, which is less than 0. Therefore, we consider the mean value of iron abundance on the surface of the stars in this data set to be less than the abundance of solar-indicative iron.

```
        One Sample t-test

data:  feh$value
t = -222.55, df = 99704, p-value < 2.2e-16
alternative hypothesis: true mean is less than 0
95 percent confidence interval:
       -Inf -0.2305179
sample estimates:
 mean of x
-0.2322343
```

In addition, we used a t-test (alternative hypothesis: the mean value of the sample stars indicating iron abundance is smaller than the Sun). We calculated a t-value of -222.55, with a p-value less than 0.05, which allows us to reject the original hypothesis and conclude that the mean value of the sample stars indicating iron abundance is smaller than the Sun.

Conclusion

In summary, the analysis of a data set encompassing 99,705 stars from across the universe reveals that the mean iron abundance on their surfaces tends to be lower than that of the Sun. This finding, derived from descriptive statistics, data visualization, and bootstrap resampling, demonstrates that most stars have a lower iron content on their surfaces compared to the Sun.

Q2. The relationship between a red giant star's magnesium-to-iron abundance ratio and its effective temperature.

Introduction

This investigation explores the relationship between the magnesium-to-iron abundance ratio of red giant stars and its effective temperature. The abundance ratio is defined as the logarithm of the ratio of two individual elements in a star, in which measurements of an element's abundance are acquired relative to its abundance on the sun. This tool can be used to determine its chemical composition, as it quantifies the relative amount of elements in the star. Generally, the range of iron abundance among stars is the widest, so this metallic element was specifically selected with magnesium as the two metallic elements the abundance ratio consists of. The effective temperature is a quantity that depends on the chemical composition of stars, as it is directly impacted by differences in mass or surface gravity resulting from varying chemical compositions. As a result, this leads to the exploration of whether the abundance ratio of this particular combination of elements affects the effective temperature of stars.

Data

The STA130_APOGEE.h5 file is this investigation's main data source, consisting of 99,705 stellar spectra observations from red giant stars collected by Henry Leung. To prepare the data for abundance ratio calculations, the mg_h and fe_h attributes are extracted from the file and then converted to data frames using the h5read and as_tibble functions. This creates 99,705 observations for each metal abundance on the surface of the star. Then, the teff data is processed

using the same procedure to extract 99,705 observations of the effective temperature of stars. The dataset's exact attributes used for data processing and analysis are described below:
mg_h (base-10 logarithm units): The abundance of magnesium on the surface of the star.
fe_h (base-10 logarithm units): The abundance of iron on the surface of the star.
teff (Kelvin): The effective temperature of the star, which indicates approximately how hot it is.

Methods/Analysis
To analyze the relationship between the Mg/Fe abundance ratio and the effective temperature values, the Mg/Fe ratios are determined using the mutate() function to construct a new data frame and compute the ratios by dividing the mg/h and fe/h data columns. In addition, the logarithm of the ratio is taken to obtain an approximated normal distribution of the data, in which the differences are illustrated by the following histograms representing the ratio distributions.

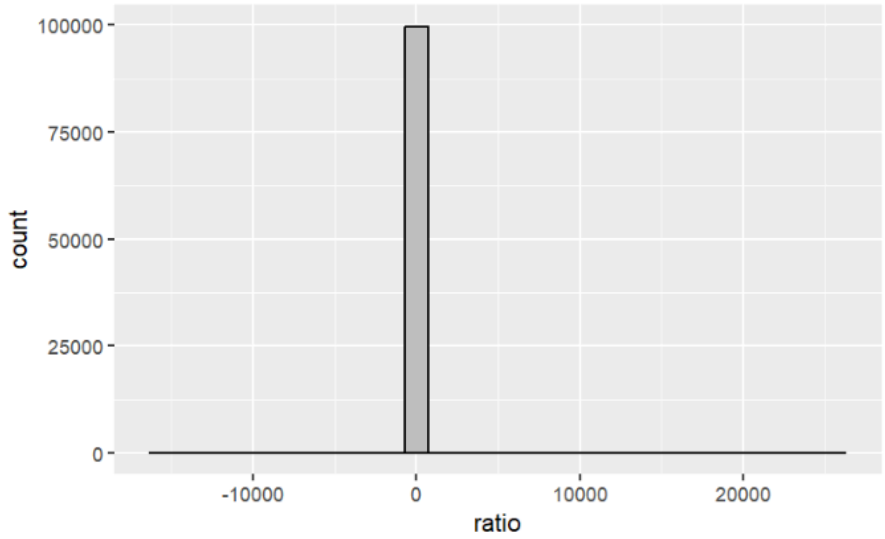


Figure 1.0. The Mg/Fe abundance ratio distribution without taking the logarithm of each ratio.

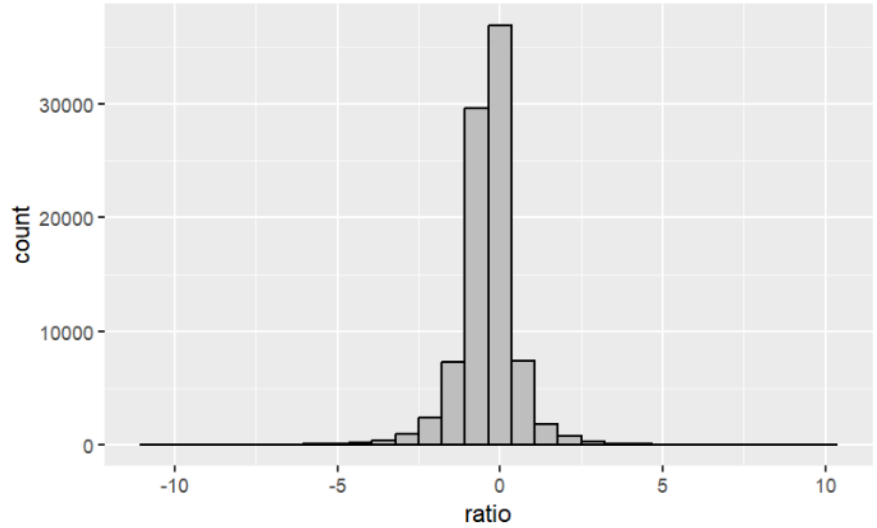


Figure 2.0. The normal distribution of the Mg/Fe abundance ratio after taking the logarithm of each ratio.

Then, a linear model is developed for the abundance ratio and effective temperature using the lm() function to determine if there is a correlation between the variables. This method initially assumes a linear relationship between the variables while computing the correlation coefficient.

Results

```
[1] -0.004225479
```

Figure 3.0. Correlation coefficient representing the statistical relationship between the Mg/Fe abundance ratio and the effective temperature.

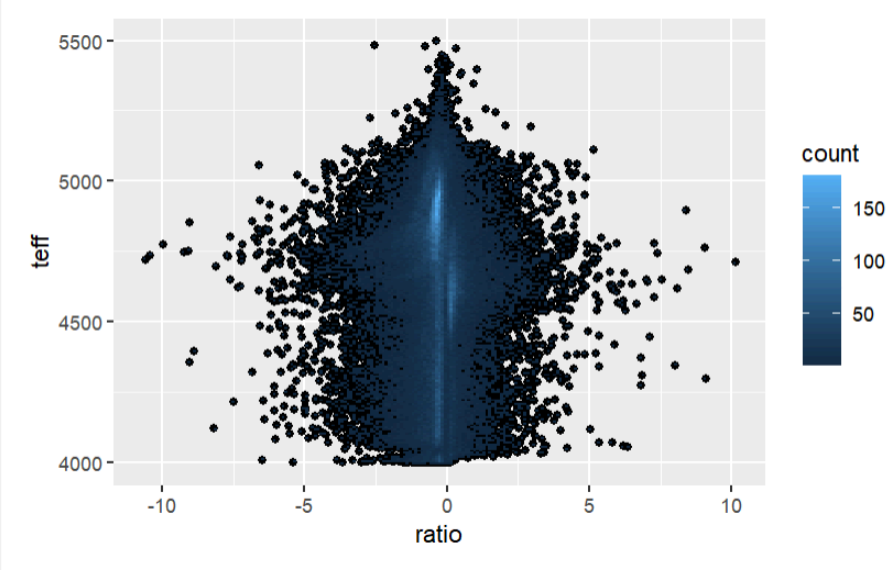
The correlation coefficient is -0.00422549 for the Mg/Fe abundance ratio and effective temperature, which demonstrates a very weak and negative relationship between the variables.

```
Call:
lm(formula = teff ~ ratio, data = abundance_ratio)

Residuals:
    Min      1Q  Median      3Q     Max
-637.88 -208.52   48.69  219.00  865.69

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept) 4632.273      1.027 4508.493   <2e-16 ***
ratio         -1.361      1.083   -1.257    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 283 on 88506 degrees of freedom
  (11197 observations deleted due to missingness)
Multiple R-squared:  1.785e-05, Adjusted R-squared:  6.556e-06
F-statistic:  1.58 on 1 and 88506 DF,  p-value: 0.2087
```

Figure 4.0. The summary of the linear model depicting the abundance ratio as the explanatory variable and the effective temperature as the response variable.

Data Visualizations

The scatter plot graph below illustrates the relationship between the values of the Mg/Fe abundance ratio and the effective temperature.



Discussion

Based on the generated summary of the linear model, we fail to reject the null hypothesis that the coefficient is 0 given a p-value threshold of 0.01, as the probability of observing a value as extreme as 0 for the t-value is sufficiently large ($p > 0.209$). The residual values indicate the difference between the estimated and observed values. However, the sufficiently large residual standard error of 283 agrees with the lack of correlation between the variables, as it communicates that the model cannot be used for accurate predictions with a significant difference in values.

Conclusion

To conclude, there is little to no correlation between the Mg/Fe abundance ratio on the surface of the star and its effective temperature. This result suggests that the ratio of magnesium and iron abundances is insignificant in affecting the effective temperature of the stars measured in this data source.

Instead, an extension to this investigation would be analyzing the relation between the abundance of helium and/or hydrogen on the surface of the star and its effective temperature. The amount of hydrogen equates to more mass and more of every other element included in the data source, while helium affects the star's luminosity and temperature directly.

Q3. Predicting Surface Gravity Based on Temperature and Oxygen Abundance

---

Introduction

This investigation examines whether the surface gravity of red giant stars can be classified based on effective temperature and oxygen abundance. The surface gravity is defined as high when it is above the median of the surface gravity observations used in this research and low otherwise. Classification trees are computed using two different functions to form a mapping of binary decisions that leads to a final classification outcome. The Gini impurity function makes

decisions/splits based on choices that maximize the "purity" in each node - the least number of misclassifications expected across all objects if guessing randomly. The entropy function makes decisions/splits based on choices that maximize the amount of "information" retained with each predicted probability. To determine the best-performing model, three types of errors are calculated: precision, recall and classification accuracy. Precision determines if the prediction is “yes” and how likely it is correct. Recall determines how many correct answers are missed. Classification accuracy determines the number of correct predictions. As a result, this leads to the exploration of to what degree the two models can accurately predict whether surface gravity is “high” or “low” using effective temperature and oxygen abundance.

Data

The analysis uses the STA130_APOGEE.h5 file, looking specifically at the surface gravity, effective temperature (eff_temp) and abundance of oxygen (o_h) of the red giant stars collected in the dataset. After cleaning and removing any missing values, 99,705 observations are stored in a data frame and used in the study for data processing and analysis. The variables are defined as follows:

Surface gravity (base-10 logarithm of cgs units): Acceleration due to gravity experienced at the surface of a star.

Effective temperature (Kelvins): How hot the star is.

Abundance of oxygen (base-10 logarithm units): How much oxygen is on the surface of the star compared to the Sun.

To compute a model for predicting the surface gravity, a new variable is created in the datagram using the *mutate()* function. The factor stores the value `High` for stars with a surface gravity greater than the median in Table 1.0 and `Low` otherwise. Due to the sheer amount of data, the dataset is divided into three subsets: training, validation and testing, to prevent overfitting. As our data may get worse when we add too many parameters into the model, the training set used to train the model contains 60% of the total data. The validation set used for deciding which model is best contains 20% of the data, and the testing set that tests the performance of the best-performing model contains the remaining 20%.

Table 1.0: Statistics of the overall sample size (n) and median of surface gravity.

n <int>	median(surface_gravity) <dbl>
99705	2.413821

1 row

Table 2.0: Statistics of the training set, number of observations (n) in each category (sg_category) and the corresponding median.

sg_category <fctr>	n <int>	median <dbl>
High	31315	2.571606
Low	28508	1.892287

Table 3.0: Statistics of the validation set, number of observations (n) in each category (sg_category) and the corresponding median.

sg_category <fctr>	n <int>	median <dbl>
High	10329	2.571122
Low	9612	1.890091

Table 4.0: Statistics of the testing set, number of observations (n) in each category (sg_category) and the corresponding median.

sg_category <fctr>	n <int>	median <dbl>
High	10450	2.565857
Low	9491	1.886261

Methods/Analysis

Before creating subsets and analyzing the data, the distribution of surface gravity is computed to look for any abnormalities in the data values. The distribution is illustrated by the following boxplot:

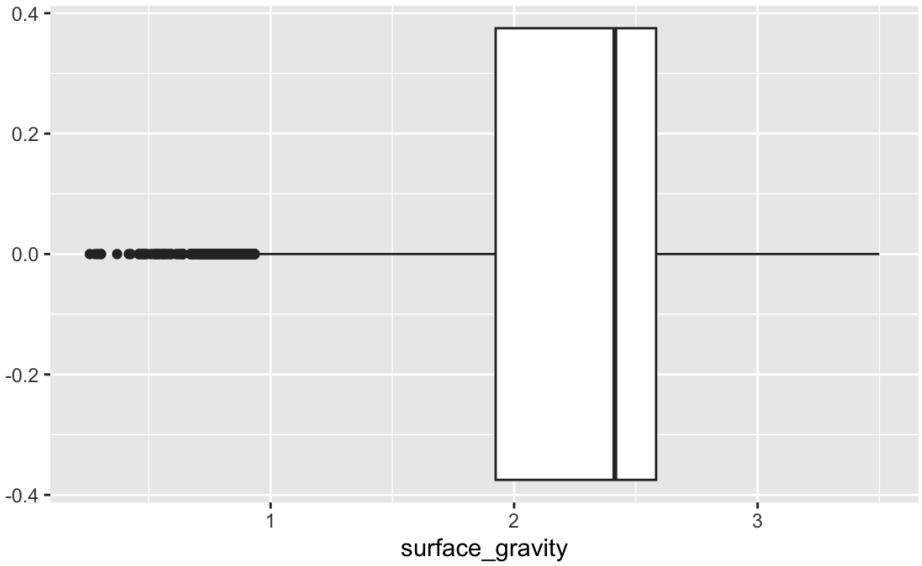


Figure 5.0: The distribution of surface gravity of the 99,705 observations sampled.

While outliers can be seen in the lower extreme of data, it is expected that the surface gravity of some stars may be outside the interquartile range. Therefore, the observations of these stars are not dropped as these values represent legitimate data and capture important information on the subject.

Classification trees are built using the training data to predict which stars' surface gravity category is `High` or `Low` based on the input features: effective temperature and abundance of oxygen. The feature importance in percentage is then calculated. The validation data is used to calculate the confusion matrix: the performance and errors of the classification algorithms. The performance of the best-performing model is then tested by computing the confusion matrix and the same metrics over the test data.

Results

	High	Low
Predict No	754	7817
Predict Yes	9575	1795

Precision: 84.2%
Recall: 92.7%
Classification accuracy: 87.2%

Figure 5.0. Summary of the performance of the Gini impurity classification algorithm using Gini impurity loss function.

	High	Low
Predict No	549	7514
Predict Yes	9780	2098

Precision: 82.3%
Recall: 94.7%
Classification accuracy: 86.7%

Figure 6.0. Summary of the performance of the classification algorithm using entropy loss function.

As can be seen from Figures 5.0 and 6.0, the two algorithms produce slightly different results. The classification tree that splits using gini impurity has slightly better overall performance compared to the classification tree using entropy. The Gini impurity tree was tested again using the testing data to confirm the performance of the model and whether it performed as expected.

	High	Low
Predict No	588	7664
Predict Yes	9862	1827

Precision: 84.4%
Recall: 94.4%
Classification accuracy: 87.9%

Figure 7.0. Summary of the performance of the classification algorithm using the Gini impurity loss function.

Based on the calculations of errors in Figure 7.0, the performance from testing matches up well based on the results from the validation data. The quantities for each metric with the test data are similar to the ones obtained previously from the validation set.

The surface gravity of a star is "high" or "low" and can be predicted relatively accurately from its respective effective temperature and abundance of oxygen using a classification tree. For this investigation, using the Gini impurity method to build the tree led to the best performance. Looking at the feature importance in Figures 9.0 and 11.0, the effective temperature of a star matters significantly more than its abundance of oxygen when determining the surface gravity

category.

Data Visualizations

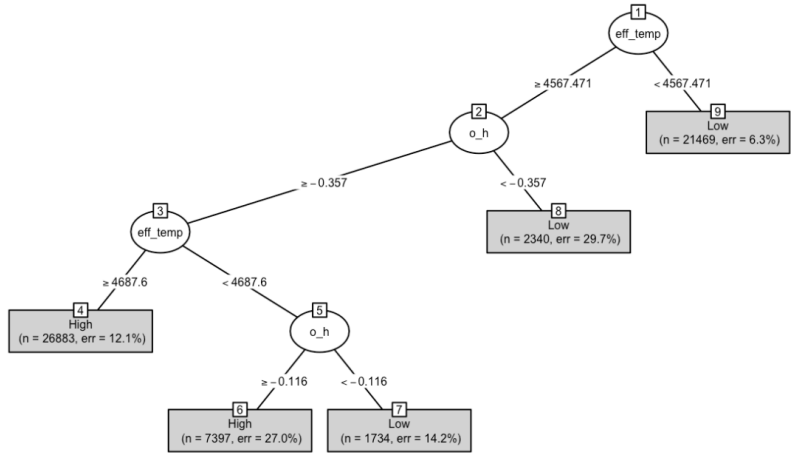


Figure 8.0. Classification tree of surface gravity using the Gini impurity loss function.

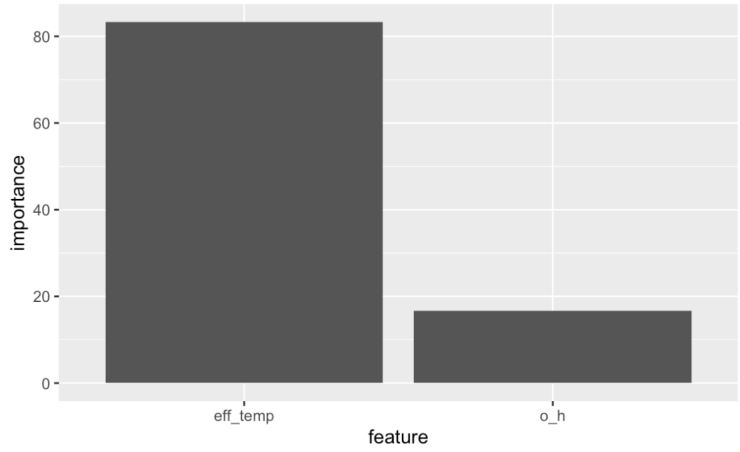


Figure 9.0. Feature importance of variables in the classification tree of surface gravity using the Gini impurity loss function.

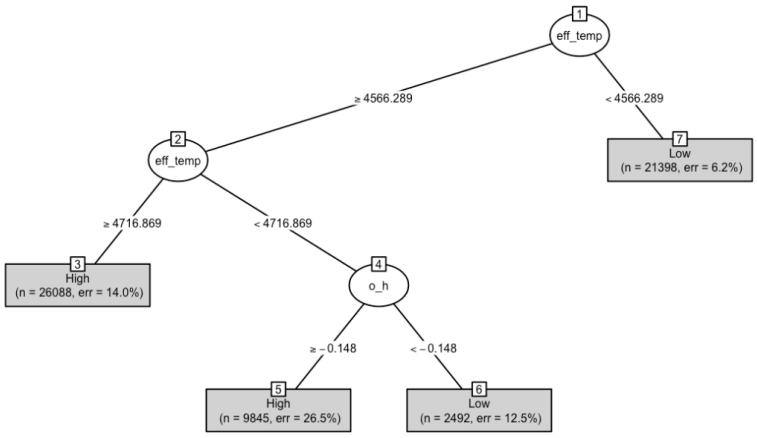


Figure 10.0. Classification tree of surface gravity using the entropy loss function.

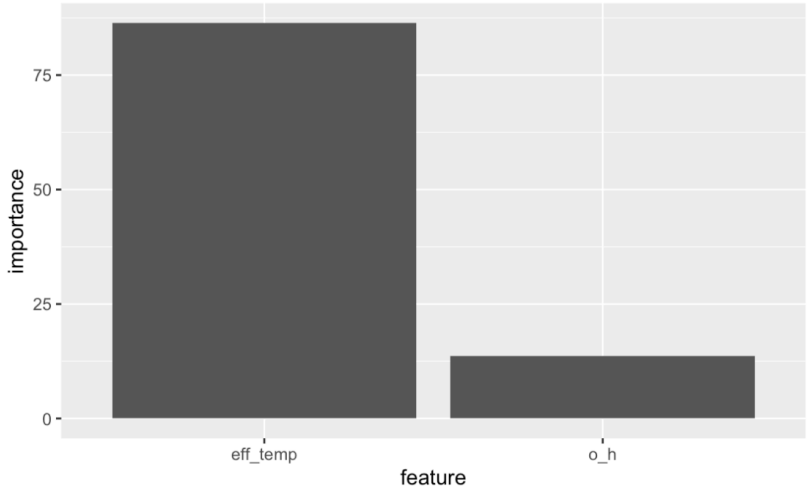


Figure 11.0. Feature importance of variables in the classification tree of surface gravity using the entropy loss function.

Discussion

Based on the generated confusion matrix of the classification models, the high precision, recall and classification accuracy values indicate that the model can be used for largely accurate predictions of whether the surface gravity of a star is above the median (classified as “High”) or below (classified as “Low”) using its effective temperature and abundance of oxygen. For this investigation, using the Gini impurity method to build the tree led to the best performance. The results imply that there is an unlying relationship between the surface gravity and two input features, especially effective temperature, because of its high feature importance. Examining the trees, the relationship may be direct; the higher the effective temperature, the more likely the surface gravity is “High” and vice versa. However, additional tests and research would need to be conducted in future work to confirm this hypothesis.

Conclusion

Using data from the SDSS APOGEE Survey, we analyzed the effective temperature, surface gravity and abundance of different elements of 99,705 red giant stars and the relationship between the variables. Through bootstrap resampling, we concluded that the abundance of iron on most stars is lower compared to the sun. The data from linear regression of magnesium to iron abundance ratio and its effective temperature suggests little to no relationship between the two features. By computing classification trees, the results show that, to a large extent, the surface gravity of a star can be classified based on its temperature and oxygen abundance.

References

Abundance ratio: Cosmos. Abundance Ratio | COSMOS. (n.d.). Retrieved April 11, 2023, from https://astronomy.swin.edu.au/cosmos/A/Abundance+Ratio

Element abundance analyses of stars. Ebrary. (n.d.). Retrieved April 11, 2023, from https://ebrary.net/55407/sociology/element_abundance_analyses_stars