Lab 2: Modeling Stellar Spectra

Andrew Goh

Nov 2022

Abstract

We demonstrate a machine learning method employed by Ness et al. 2015 (*The Cannon*) to model stellar spectra without the direct calculations of synthetic spectra. Using the APOGEE (Apache Point Observatory Galactic Evolution Experiment) spectral data and its corresponding labels from the ASPCAP/APOGEE pipeline, labels being atmospheric parameters of T_{eff}, $log(g)$, $[Fe/H]$, $[Mg/Fe]$, $[Si/Fe]$, we train a generative model to predict a spectrum given these atmospheric parameters. Using a training set and a test set of equal size of mostly red giant stars, the model is able to predict the labels given a spectrum to moderate accuracy.

1 Introduction

At the zeroth order, the spectrum of a star can be approximated as a blackbody, described by the Planck function given an effective surface temperature:

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5}\frac{1}{e^{hc/\lambda kT}} \tag{1}$$

Integrating over all wavelengths, it will give the flux as a function of wavelength. However, in actuality, the spectrum of a star will not be a smooth function, and will include many absorption/emission features due to chemical abundances at the stellar atmosphere. Photons escaping the star will pass through the atmosphere, and predicting the flux of these photons with accuracy involves utilizing a model atmospheres and solving the so-called radiative transfer equation for each wavelength. This method generates synthetic spectra given inputs of effective temperature, surface gravity, and chemical abundances. The ASCAP pipeline [3] uses this procedure which then compares synthetic spectra to the APOGEE spectra [4] using χ^2 minimization to predict atmospheric parameters.

By using the spectral data from APOGEE and their corresponding labels from ASPCAP, we are able to train a generative model to predict stellar spectrum given labels (forward method) and also predict the labels given a stellar spectrum (inverse method) using a basic neural network model. Predicting the atmospheric parameter labels using a generative model is useful as it can work more efficiently than modeling atmospheres and calculating synthetic spectra. This lab closely follows the procedure described in *The Cannon* [5].

2 APOGEE Spectra and ASPCAP labels

The APOGEE spectra is directly available to download via the SDSS (Sloan Digital Sky Survey) website. The command line tool, wget, allows us to streamline the download of specific datasets

from SDSS straight into our local directory. The fields/clusters we are interested are "M15", "N6791", "K2_C4_168-21", and "060+00", all of which are arbitrarily chosen. The spectra are given as *apstar* .fits files, which includes multiple visit spectrum flux, the corresponding wavelength bins, errors, and bitmasks, which will be discussed below. Most stars include multiple visit spectra, which are combined spectra of a star observed multiple times to increase the S/N ratio. Effects from the doppler shift would contribute differently to the spectrum for each visit because Earth's orbit constantly changes the observed radial velocity of the star. To correct for this each spectra has been shifted to the barycentric frame which is the frame of the center of the mass of the solar system. Barycentric correction refers to the correction in which Earth's radial velocity relative to the observed source is subtracted from the observed radial velocity. An example spectra is shown below.

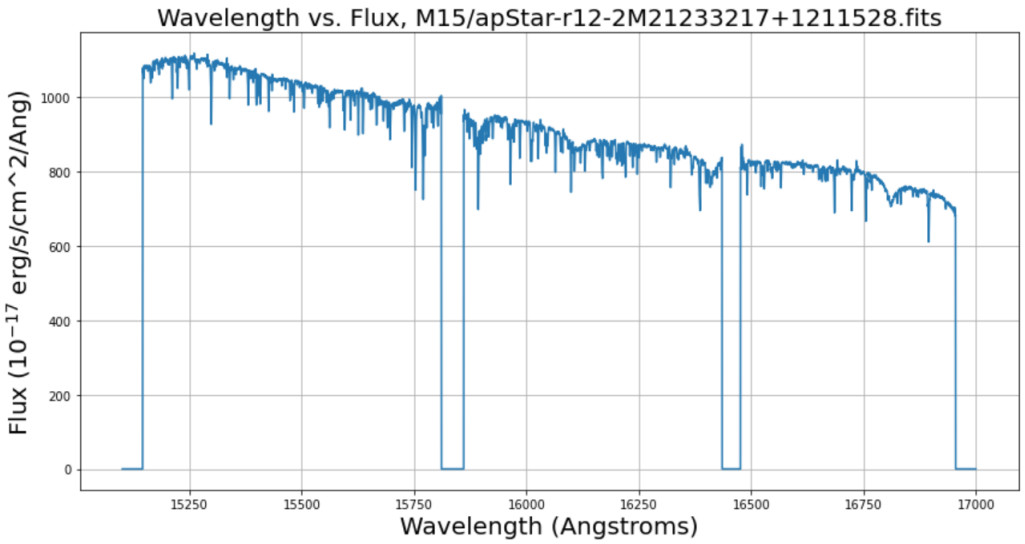

Figure 1: Spectrum of an arbitrary star with 2MASS ID: 2M21233217+1211528 of the M15 cluster. The gaps can be seen between the three detectors, where the flux is unmeasured (zero). Compared to the blackbody planck function, this is essentially just a snippet of whole spectrum, as there are wavelengths of interest in this region that correspond to common absorption/emission features that characterize a star.

The spectrum covers the wavelength range of 15000 Angstroms to 17000 Angstroms, with 8575 bins. There are wavelength gaps between detector chips, which can be seen in the figure above. However, we are only interested in a subset of spectra from the fields mentioned above. To retrieve the ASPCAP labels (atmospheric parameters) for each of the spectra, the final allstar catalog can be directly downloaded from the the SDSS website. Converting the catalog into a dataframe, rows corresponding to the clusters of interest can be singled out. The columns of interest are effective temperature, log(g) (surface gravity), [Fe/H], [Mg/Fe], [Si/Fe] and their corresponding errors. Many stars in this dataframe have underived parameters, in which they are characterized by a value of -9999.99. Stars that have underived values for any of those atmospheric parameters and that have an SNR ratio < 50 are discarded. Stars with low metallicity are discarded, with [Fe/H] < -1 We also make sure to not include dwarfs by constraining the $log(g) > 4$ or $T_{eff} > 5700K$. The resulting number of stars obtained is 1854.

The apstar .fits file provides a bitmask at each wavelength. The bitmask is encoded in binary, where as each binary digit marks a certain flag in the pixel. If one of the digits are flagged, as in specific binary digit is 1, it could indicates a certain warning on the pixel. For

example, if digits 1 and 2 are flagged, it means the pixel is marked as a cosmic ray detection and as saturated respectively. The documentation for these bitmasks can be found on the SDSS website. In this lab, if the digits 0-7 and 12 are flagged, then the error for that pixel is set to be a high value, 1.0e10.

FILE	APOGEE_ID	FIELD	SNR	TEFF	TEFF_ERR	LOGG	LOGG_ERR
apStar-r12-2M03501997+2458304.fits	2M03501997+2458304	K2_C4_168-21	137.744003	4465.914062	90.295403	1.886413	0.061935
apStar-r12-2M03502656+2445432.fits	2M03502656+2445432	K2_C4_168-21	281.511993	4744.914062	84.443794	2.432934	0.036187
apStar-r12-2M03504772+2514178.fits	2M03504772+2514178	K2_C4_168-21	665.562988	4409.979492	78.179939	1.847668	0.062366
apStar-r12-2M03504852+2433483.fits	2M03504852+2433483	K2_C4_168-21	503.894012	3846.593750	61.748474	1.216197	0.045915
apStar-r12-2M03505216+2442325.fits	2M03505216+2442325	K2_C4_168-21	192.904999	4863.133789	87.082924	2.473944	0.041174
...	...	...	...	...	...	...	...
apStar-r12-2M21351077+1213316.fits	2M21351077+1213316	M15	121.896004	4797.105957	96.422165	3.191941	0.054904
apStar-r12-2M21351671+1221177.fits	2M21351671+1221177	M15	87.771500	4964.708984	117.165840	2.383122	0.051231
apStar-r12-2M21352209+1135485.fits	2M21352209+1135485	M15	58.671902	5067.860840	131.554565	3.642437	0.065028
apStar-r12-2M21352418+1204096.fits	2M21352418+1204096	M15	62.023300	4899.193848	110.572014	3.668981	0.055902
apStar-r12-2M21354701+1209559.fits	2M21354701+1209559	M15	87.247597	4956.580566	113.639565	3.306615	0.061523

Figure 2: A slice of the allstar catalog data frame. The apstar file name is given along with the 2MASS ID. The effective temperature, surface gravity (in log scale), and the chemical abundances are given along with the corresponding measurement errors. By cross matching the apstar file name with our downloaded APOGEE spectra, matching correspondences can be made between spectra respective labels.

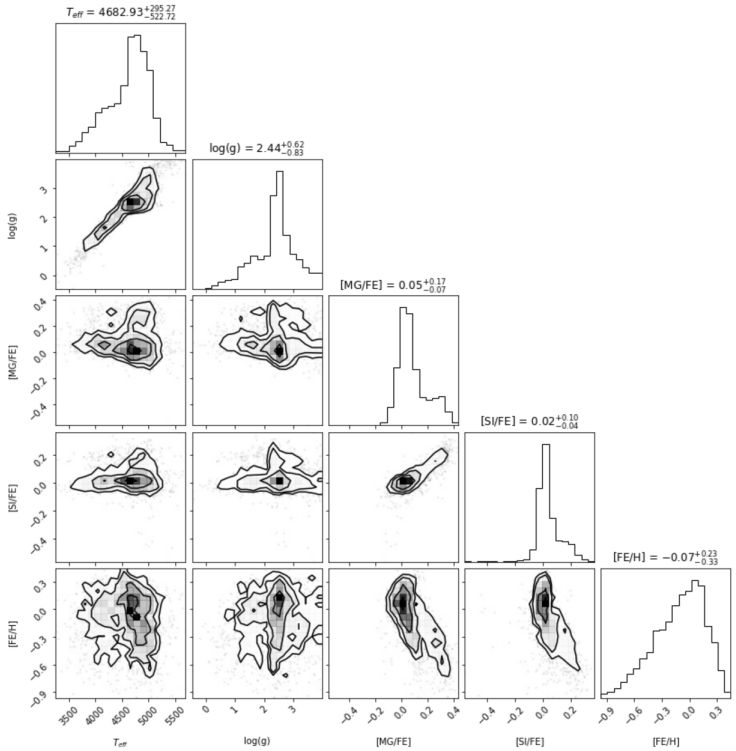

Figure 3: A corner plot to visualize the distribution of atmospheric parameters and their general correlations.

3 Pseudo-continuum Normalization

Before training the generative model, it is necessary to normalize the spectra. The model should predict atmospheric parameters based on certain small scale features of the spectra, and should not be trained on the flux values obtained through measurements as these arbitrarily depend on the star's distance from us. Through normalization, we should expect that stars with the same labels should be identical. The normalization procedure as described in Ness et al. 2015 finds pixels/wavelength bins that do not showcase any dependence on the labels by iteratively passing in the spectra in to the generative model. These pixels are then taken as data points to be fit to a Chebyshev polynomial, which then the spectra could be normalized. In this lab, we are preemptively given a set of wavelength bins that is already known to be unaffected by the labels. We take these wavelength bins and the measured flux at these wavelengths to fit 5th order polynomials to the three different detector chips separately using the numpy polyfit module. When performing the fit, the corrected errors according to the bitmasks are also passed in to ensure that bad pixels are not accounted into the fit. This procedure is performed on all 1854 stars in our data set. The normalization procedure is visualized in figure 4.

4 Building the Model

The apstar dataframe contains the APOGEE file names. To create a training and test set of equal size, we can simply create a list of these file names, scramble them, and split them using the sklearn.model_selection module. The training and test set both consist of 927 stars. Creating a list of normalized spectrum as described in the previous section simply involves calling the list of file names of the training or test set.

For a specific wavelength bin, the predicted flux of a single star n is described as a function of a 2nd order polynomial in its corresponding labels.

$$f_{n\lambda} = \theta_\lambda^T \cdot \mathbf{l}_n \tag{2}$$

where $f_{n\lambda}$ is the predicted flux, θ_λ is the coefficient vector to be fitted for, and $\mathbf{l}_n$ is the label vector for star n. For example, a model with 3 labels the label vector is as follows:

$$\mathbf{l}_n = [1, l_1, l_2, l_3, l_1^2, l_1 * l_2, l_1 * l_3, l_2^2, l_2 * l_3, l_3^2] \tag{3}$$

In our model, each element l_n (n = 1,2,3,4,5) corresponds to atmospheric parameters T_{eff}, $log(g)$, [Fe/H], [Mg/Fe], [Si/Fe]. Each of these parameters are scaled to unity by subtracting by the means of the training set labels. The first element of the label vector, 1, allows for an offset fit. The goal is to fit for the coefficient vector θ_λ^T for each wavelength bin, resulting in a matrix of coefficients. For one coefficient vector θ_λ^T at wavelength bin λ, one can solve the matrix equation:

$$\mathbf{X}\theta_\lambda = \mathbf{f}_\lambda \tag{4}$$

$\mathbf{X}$ is a 927x21 matrix, each row consisting of a label vector $\mathbf{l}_n$. $\mathbf{f}_\lambda$ is a vector of length 927, consisting of all the measured flux values at wavelength λ for each star in the training set. Performing this operation for all 8575 wavelength bins will give a matrix of coefficients. However, errors are known for the flux values and some pixels are marked with high errors. So we account for this by solving the problem in weighted least squares. In addition to the

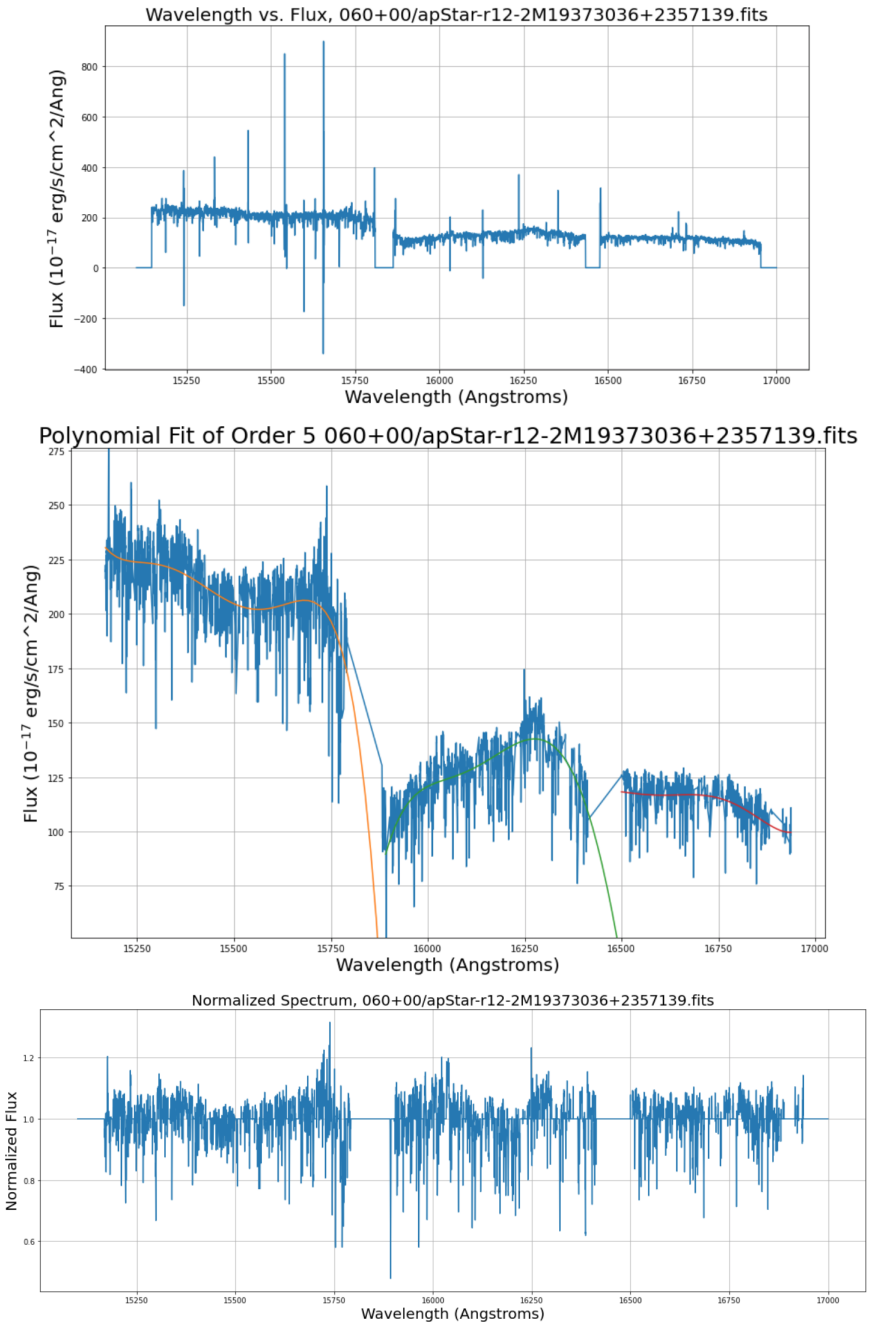

Figure 4: Figures showcasing the normalization procedure of an arbitrary star. The first plot is the raw spectrum obtained from APOGEE. The gaps (denoted by lines at zero flux) between the detector chips can be seen as well as gaps at the beginning and end of the spectrum. There are also unphysical peaks and dips (that reach negative flux) that can be attributed to errors in measurement, for example, a detection of a cosmic ray. In the second figure, these bad pixels have been omitted (marked by an error of 1.0e10) in order to reasonably fit the polynomial. As shown, three polynomials of the 5th order were fit correspondingly to the three detector chips. The third figure is the normalized spectrum that will ultimately be used to train and validate the generative model. For errors that were set to 1.0e10, the flux values at these wavelengths of the normalized spectrum are set to 1.

reported errors, we also include an intrinsic scatter term that accounts for errors not captured by the reported errors. Assuming gaussian errors, maximizing likelihood function in terms of the coefficient vector and intrinsic scatter will give us the coefficient vector.

$$\sigma_\lambda = \sqrt{\sigma^2_{measured,\lambda} + s^2_\lambda} \tag{5}$$

where $\sigma_{measured,\lambda}$ is a vector where each element is the reported error for the 927 stars of the training set, and $\sigma_{i\lambda}$ is the constant intrinsic scatter term at wavelength λ. A weight matrix can then be constructed:

$$W = I \cdot \sigma_\lambda^{-1} \tag{6}$$

where I is the identity matrix.

$$\mathbf{f}_{w\lambda} = \mathbf{f}_\lambda \cdot W \tag{7}$$

$$\mathbf{X}_w = W \cdot \mathbf{X} \tag{8}$$

Then we can solve the following matrix equation using the numpy.linalg module.

$$\mathbf{X}_w \theta_\lambda = \mathbf{f}_{w\lambda} \tag{9}$$

To find the best fit intrinsic scatter and coefficient vector, we solve the following matrix equation and define the log likelihood function as a function of the intrinsic scatter term s_λ:

$$\ln p(f_{n\lambda}|\theta_\lambda^T, \mathbf{l}_{wn}, s^2_\lambda) = -\frac{1}{2}\frac{[f_{n\lambda} - \theta_\lambda^T \cdot \mathbf{l}_{wn}]^2}{s^2_\lambda + \sigma^2_{measured,\lambda}} - \frac{1}{2}\ln(s^2_\lambda + \sigma^2_{measured,\lambda}) \tag{10}$$

Note that $\mathbf{l}_{wn}$ are the rows of $\mathbf{X}_w$ and that θ_λ also depends on s_λ. Summing over each star in the training set gives the total log likelihood. We find the best fit coefficient vector and intrinsic scatter for a wavelength λ by maximizing the total likelihood with respect to those parameters.

$$\theta_\lambda, s_\lambda \leftarrow argmax_{\theta_\lambda, s_\lambda} \sum_1^{927} \ln p(f_{n\lambda}|\theta_\lambda^T, \mathbf{l}_{wn}, s^2_\lambda) \tag{11}$$

5 Results

5.1 Predicting Spectrum as a function of labels

After obtaining the best-fit coefficient vectors for each wavelength, we can now predict flux at certain wavelength bin given a set of labels with equation (2). Solving this equation for each wavelength will give the full predicted spectrum. An example is shown in figure 5.

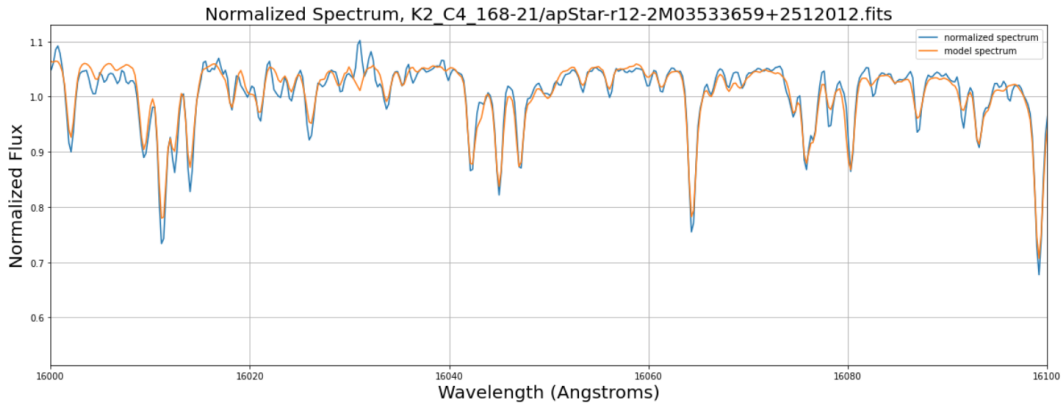

Figure 5: Generated model spectrum overplotted with the original spectrum. The model spectrum is able to capture the general shape as well as the absorption features shown by the characteristic dips in flux. However, there is a consistent underestimation of the magnitude of the dips throughout the spectrum.

5.2 Gradient Spectrum and Intrinsic Scatter

To observe which wavelength bins are sensitive to our labels, gradient plots are visualized for each label. The gradient spectrum can simply be obtained by changing one of the label parameters by a small amount and calculating the infinitesimal change.

$$\frac{df_{n\lambda}}{dl_n} = \frac{f_{n\lambda}(l_n + e) - f_{n\lambda}(l_n)}{e} \tag{12}$$

Unfortunately, the plotted gradient spectra seems to be mostly uninformative. There appears to be some correspondence between known absorption lines, however the wavelengths at which the expected gradient occurs is shifted to about 5 angstroms. The fitted intrinsic scatter term is also uninformative as the method used to fit for the term only searched through a small grid of values. There are a few wavelength bins where the intrinsic scatter term becomes large, but it is uncertain whether it is an arbitrary fit or if it corresponds to an absorption feature.

5.3 Predicting Labels given a Spectrum

The inverse problem can now be solved. Given coefficient vectors for each wavelength, we can fit for the labels. However, since the label vector is a 2nd order polynomial, we can not use recast the problem into matrix form as before since the problem is now non-linear. We use the Trust Region Reflective ("trf") method within the scipy.optimize.curve_fit module to solve for the labels.

A one-to-one line is plotted in figure 8 to visualize the spread of values in comparing the best-fit values and measured values. We also calculate the residuals and tabulate the root mean squared error.

5.4 Kiel Diagram comparison with MIST Isochrone

The best fit labels can be tested in correspondence with stellar evolution by plotting a Kiel Diagram, plotting log(g) vs T_{eff} and coloring the points by their metallicity [Fe/H] as shown in figure 9.. This plot can be compared with theoretical isochrones provided by MIST isochrones [2].

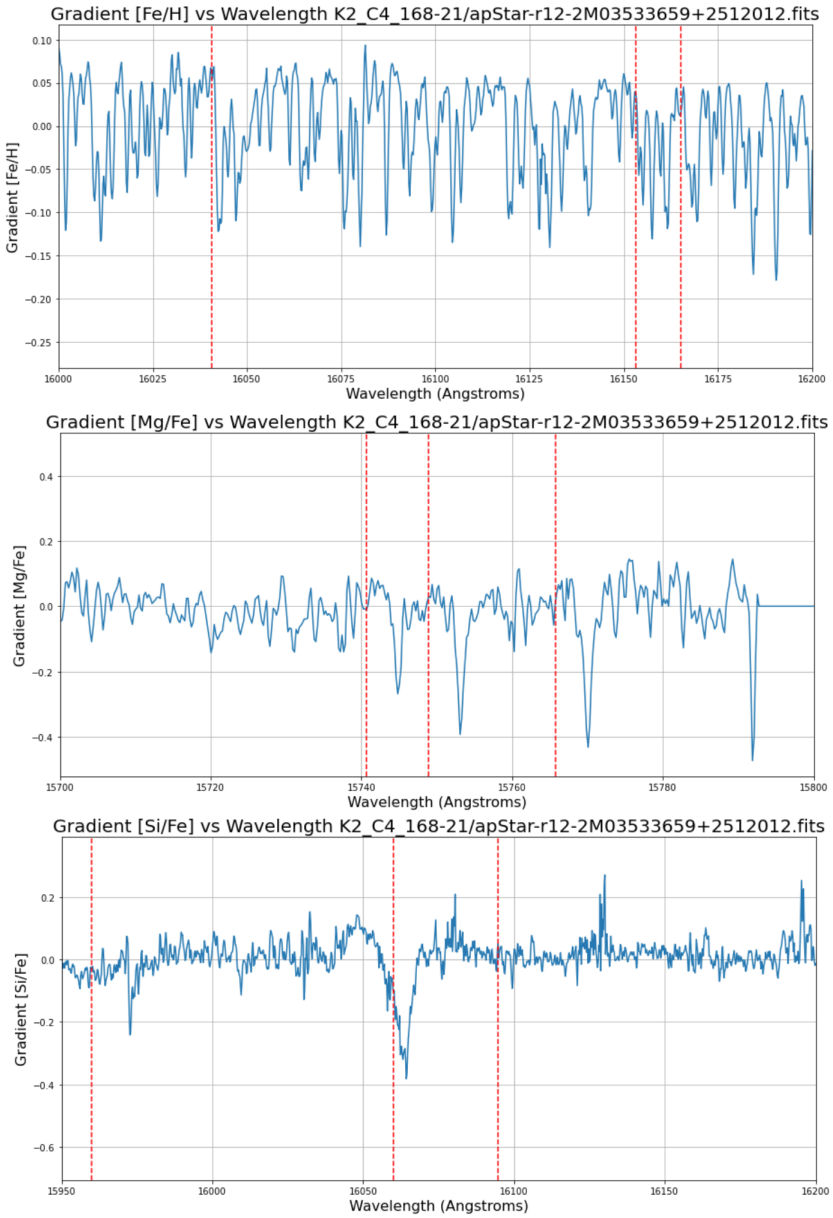

Figure 6: Gradient Spectra for an arbitrary star. The red dashed lines correspond to the known absorption lines. As shown clearly by the second figure, the three characteristic dips are shifted right from the expected wavelength at which the known absorption lines occur.

5.5 Wrapping the model in MCMC

We are provided with a mystery spectrum fits file. Without knowledge of the measured labels, we use MCMC to determine the best fit labels and its corresponding errors. Using the pymc3 module, we are able to wrap our model within model. The priors chosen for all the labels are uniform. For T_{eff}, we chose a uniform distribution from 0 to 10000 K, for log(g), 0 to 10, and for the metallicities, -2 to 2. The likelihood function is a normal distribution centered at the predicted flux values and the observed values to be compared with are the mystery spectrum flux values. By running the pymc3 model, posterior distributions can be retrieved for each label.

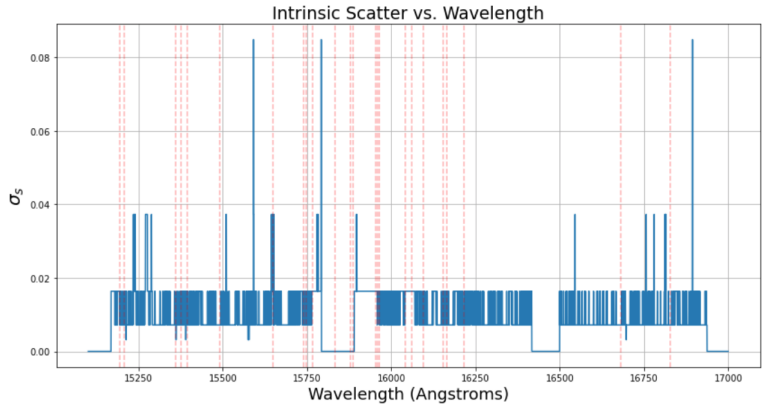

Figure 7: Intrinsic Scatter for each wavelength bin. The 'resolution' of this term is low since only 15 terms were searched in the range of 10^{-5} to 0. This is due to the runtime being extensive since for each scatter term, the coefficient matrix had to be fit for. As shown, it is difficult to attribute the large intrinsic scatter terms to the wavelengths of any of the known absorption lines.

	$T_{eff}(K)$	log(g)	[Fe/H]	[Mg/Fe]	[Si/Fe]
Best-fit	4937	3.4562	-0.15002	0.190667	0.047905
Best-fit error	43	0.0002	2e-5	2e-5	1e-5
Measured	4912	3.39	-0.115	0.197	0.065
Measured error	119	0.06	0.013	0.019	0.019

Table 1: Tabulated values of best fit labels and measured labels of star with APOGEE ID 2M21284724+1326482. The best fit errors obtained from the return values of curve_fit are small compared to the measured error, which could be an indicator of over-fitting. Overall, the best fit labels are within the same order and error range of the measured labels, indicating that our model performs moderately, but it is not accurate in regards to the ASPCAP errors.

	$T_{eff}(K)$	log(g)	[Fe/H]	[Mg/Fe]	[Si/Fe]
RMS error	78.9	0.15	0.056	0.062	.059

Table 2: Tabulated values of root mean squared error between the best fit labels and the ASPCAP labels. The calculated rms errors are larger than most of the ASPCAP measured errors. In theory, our model should be able perform fits within the measured errors, so there is clear room for improvement.

A corner plot visualizing the results of the MCMC is presented in figure1 10.

5.6 Visualizing offset Spectra

Using our model, we can observe how absorption features may change as labels are varied. By varying metallicity [Fe/H], it is expected that absoprtion features become more prominent as it is increased. However, the magnitude our model reacts to these changes are more miniscule than expected. We can also simulate how the spectrum is expected to change as a star evolves across the red giant branch at a fixed chemical composition. This is done by varying the surface gravity and effective temperature. Red giant stars expand while maintaining constant mass, and as a consequence of maintain equilibrium, the surface gravity and effective temperature decreases.

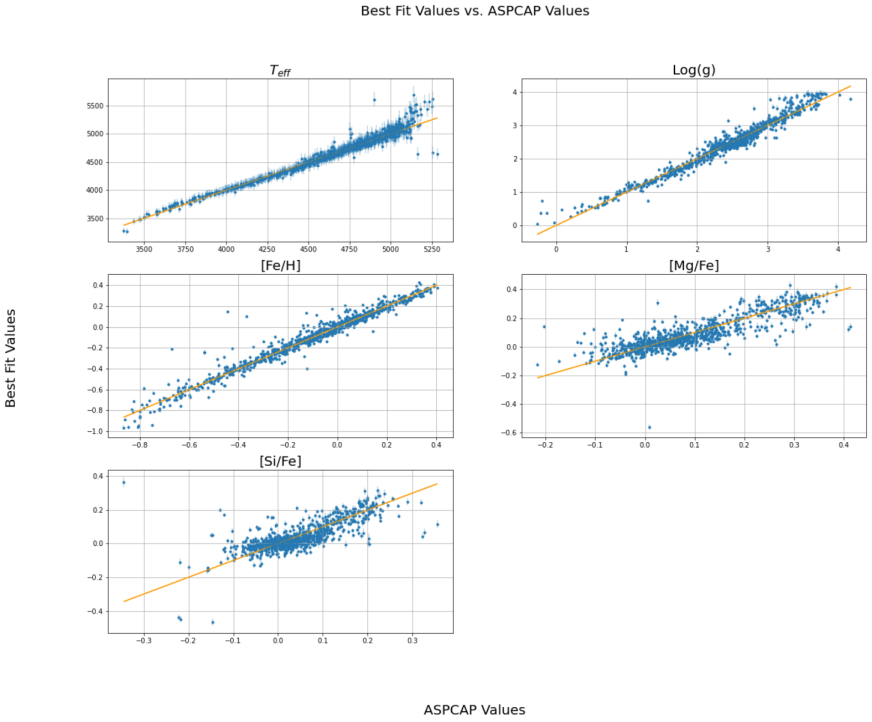

Figure 8: One to one line of Best fit values vs. ASPCAP values with errorbars. The orange line is a line of slope one for comparison. The errorbars are plotted for the reported ASPCAP values. As shown, the spread between the values are larger than the errorbars, indicating that our model is not as accurate as reported values. The effective temperature becomes harder to predict at higher temperatures. There seems to be more spread overall towards edge values for each the labels, which might correspond to insufficient training at those values.

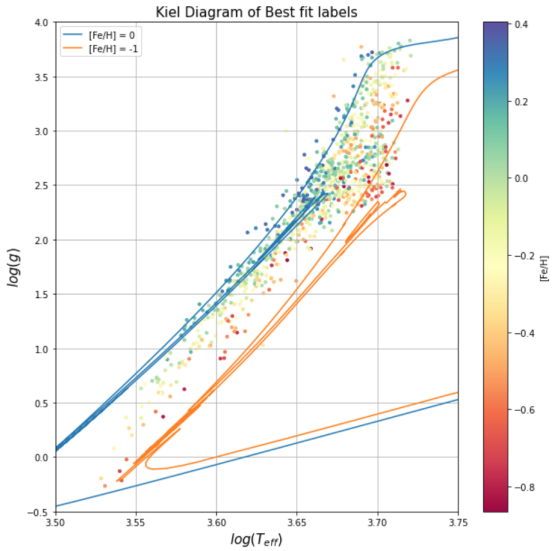

Figure 9: Kiel Diagram of the best fit values overplotted with MIST isochrones of metallicity [Fe/H] = 0 (solar metallicity) and [Fe/H] = -1. As expected from stars that evolve towards the red giant branch, as they become bigger, the surface gravity and temperature both decrease. As shown on the graph the relationship is positively correlated, appearing to be a linear in log space. Stars with higher metallicities have a higher surface gravity than those with lower metallicities as predicted by the best fit values and the MIST isochrones.

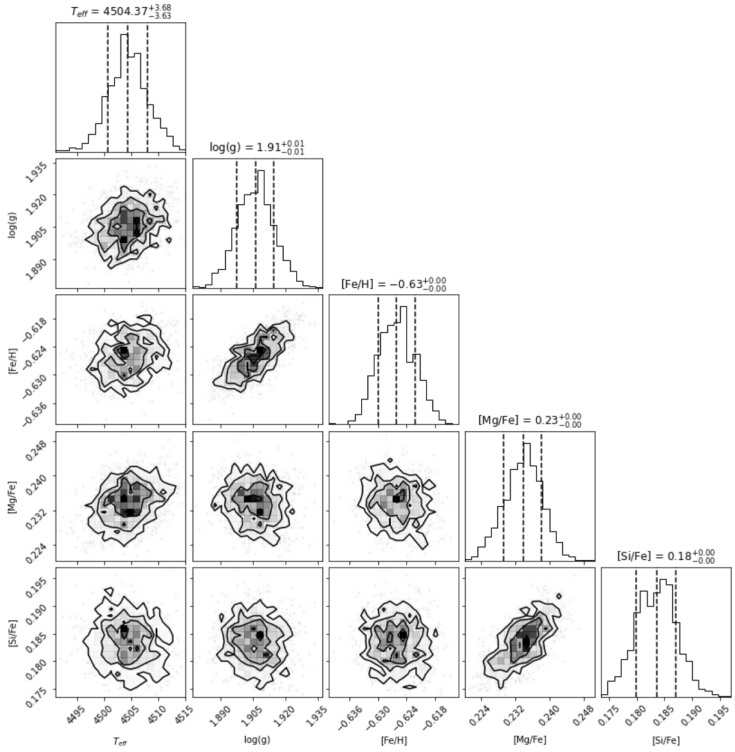

Figure 10: Corner plot of the constraints of the five labels for the mystery spectrum produced with the MCMC method. The errors determined by the MCMC method appear to be too small compared to the ASPCAP errors.

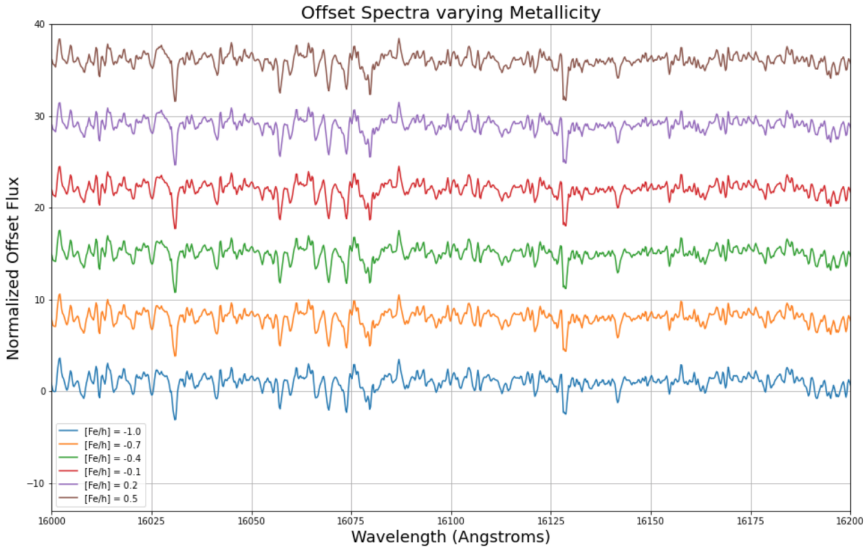

Figure 11: Offset spectra of varying metallicities from [Fe/H] = -1 to 0.5. It is very slight, but [Fe/H] is increased, the absorption features become more prominent, as expected since the higher the chemical abundance in the atmosphere, the higher probability photons of a specific wavelength will be absorbed by Fe atoms, resulting in a lower flux at that specific wavelength.

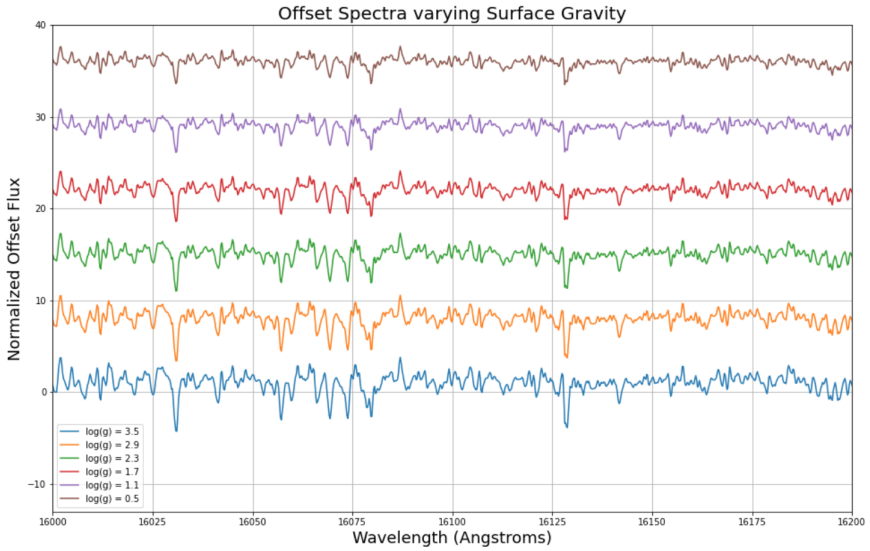

Figure 12: Offset spectra of varying surface gravity from log(g) = 3.5 to 0.5 while fixing [Fe/H], [Mg/Fe], [Si/Fe] at 0, corresponding to a solar metallicity star. As the surface gravity is decreased, the absorption features become less prominent.

6 Conclusion

In theory, our model could be trained to achieve the same accuracy as the reported ASPCAP values. The inaccuracy of our model is likely due to randomly choosing stars in the training and test set, where picking stars with high SNR for the training set would be more ideal. An additional cause of error can be attributed to the model not accounting for potential binaries, which measured spectrum could actually be a sum of two spectrum. In terms of training the model, a consequence would be an overestimation of chemical abundances, as the spectrum will include more prominent absorption features then a single star actually has. El-Badry et al. 2018 [1] accounts for this by generating multiple spectra models and single spectra models for a single isochrone and fitting the data to both models, seeing which of the models are a better fit.

References

[1] Kareem El-Badry et al. "Discovery and characterization of 3000 main-sequence binaries from APOGEE spectra". In: *Monthly Notices of the Royal Astronomical Society* 476.1 (Jan. 2018), pp. 528–553. DOI: 10.1093/mnras/sty240. URL: https://doi.org/10.1093%2Fmnras%2Fsty240.

[2] Jieun Choi et al. "Mesa Isochrones and Stellar Tracks (MIST). I. Solar-scaled Models". In: 823.2, 102 (June 2016), p. 102. DOI: 10.3847/0004-637X/823/2/102. arXiv: 1604.08592 [astro-ph.SR].

[3] Ana E. Garcıa Pérez et al. "ASPCAP: The APOGEE Stellar Parameter and Chemical Abundances Pipeline". In: 151.6, 144 (June 2016), p. 144. DOI: 10.3847/0004-6256/151/6/144. arXiv: 1510.07635 [astro-ph.SR].

[4] Steven R. Majewski et al. "The Apache Point Observatory Galactic Evolution Experiment (APOGEE)". In: *The Astronomical Journal* 154.3 (Aug. 2017), p. 94. DOI: 10.3847/1538-3881/aa784d. URL: https://doi.org/10.3847%2F1538-3881%2Faa784d.

[5] M. Ness et al. "iTHE CANNON/i: A DATA-DRIVEN APPROACH TO STELLAR LABEL DETERMINATION". In: *The Astrophysical Journal* 808.1 (July 2015), p. 16. DOI: 10.1088/0004-637x/808/1/16. URL: https://doi.org/10.1088%2F0004-637x%2F808%2F1%2F16.