Milestone Report Capstone 1

Maia Paddock

Hubble's law is possibly the most iconic law in all of astrophysics, defining the linear relationship between redshift and the distance of an astronomical object from our location on Earth. Based on the constant speed of light, the redshift also defines the age of the astronomical object at the time of observation. Therefore, it is clear that for astronomical objects outside of the galactic neighborhood of the Milky Way, the redshift is one of the most important physical properties that can be derived. What if applying machine learning methods could drastically reduce the amount of time needed to calculate redshifts for large collections of data and match the accuracy of photometric redshifts?

The most accurate way of determining redshifts is spectroscopic data - measuring how spectral emission and absorption lines shift due to the movement of the object from the expansion of the universe. However, spectroscopic data is difficult to obtain and relatively rare. Photometric data, the measurement of the brightness of astronomical objects through various filters, is available for many objects due to the Hubble Deep Field, Hubble Ultra Deep Field, and many more surveys designed to cover as much of the sky as possible. The magnitudes and brightnesses in different band filters should correlate to the redshift of the galaxy due to the difference in extinction rates for different wavelengths of light as they travel from the object to the observer and the relationship between the age of the galaxy and the wavelengths emitted.

This project is meant to demonstrate to the general public and the wider astrophysics community the potentials of using machine learning in observational astrophysics and cosmology. Training and tests on data from various surveys would most likely be necessary to establish machine learning as a method to quickly and accurately estimate photometric redshifts, but this investigation should act as a proof of concept for the potential usefulness of this method. With the launch of the James Webb Space Telescope in early 2021, large amount of new astronomical data is expected to begin arriving over the next few years. While spectroscopic redshifts would still be needed to confirm many of the more interesting and exciting pieces of data, a quick, reliable machine learning model for estimating the redshifts on large amounts of data could prove valuable to the wider astrophysics community to point out areas of interest and guide further studies and exploration.

The data set that will be used for this investigation is the COMBO-17 object catalogue of the Chandra Deep Field South. The COMBO-17 survey imaged one square degree of sky using 17 optical filters, ranging from 350 to 950nm. In total, 63,501 objects were imaged. Photometric redshifts were calculated for this dataset by Wolf et al. (2004). Below $R<24$, the redshifts are most reliable when compared to the approximately 1.6% of objects with spectroscopic data. Ten-thousand galaxies fall into this category, which will be the objects used for this investigation. The data includes 65 columns of data, 48 of which are magnitudes, photons flux,

and luminosity measurements and errors which will be used to train the model. The remaining data are identifiers for the object within the survey and the redshifts and errors calculated using the methods in the Wolf et al. paper. The mean redshifts calculated for each object will be used as the validation data for the model. A regression model will be used to allow the model to deliver a distribution of predicted redshifts.

At the end of this investigation, I will be able to deliver a regression model which has been trained on my selected portion of the COMBO-17 galaxy dataset and returns predicted redshifts within an acceptable level of accuracy. Additionally, I will write a paper detailing my methods, results, and ideas for further analysis. Finally, the results of the project will be presented in a blog post for easier reading and presentation than a full methods paper.

The COMBO-17 dataset was obtained from https://www.mpia.de/COMBO/combo_CDFSpublic.html, which offers the data in ASCII and FITS format. Initially I downloaded the ASCII file and loaded it into a pandas dataframe but the columns were not aligning properly with the data and there were many trailing instances of NaN at the end of each row. After examining the data as much as possible I could not find a way to properly align all of the data into the appropriate columns. I switched to accessing the data via the FITS file. Flexible Image Transport System (FITS) files are commonly used in astronomy to store data produced by telescopes. It is formatted as n-dimensional arrays and tables. I used astropy.io fits and tables packages to open the file and transfer the data into a pandas dataframe. On a quick inspection, the file held 63501 objects, each of which had 104 variables associated with it. Unlike the ASCII version of the data, all of the columns were properly aligned and contained the correct data.

The first step was to filter the data to include only objects positively identified as galaxies and have an R magnitude of less than 24. The R magnitude of less than 24 insures that the photometric redshifts already calculated for each galaxy are accurate. There were five types of objects included in the original file under the column 'MC_class': b'Galaxy', b'Star', b'Quasar', b'Galaxy (Uncl!)', and b'Strange Object'. Approximately 53000 of these were classified as b'Galaxy', 14543 of which also had an R magnitude of less than 24. After these first two filters, I checked all of the columns to remove the ones that were all null or have no meaning for galaxies: MajAxis, MinAxis, PA, phot_flag, var_flag, stellarity, MC_z2, dl, S145Mag, and e_S145Mag. For example, stellarity is a variable calculated to classify an object as a star or other astronomical object. As this sample is all galaxies, this value is unimportant. Additionally, I dropped the columns for UbMag, e_UbMag, BbMag, e_BbMag, VbMag, e_VbMag. According to the dataframe summaries, these columns contained only zeroes for every object.

Filtering out NaN values was relatively simple for this dataset. Each galaxy that had NaN in one magnitude entry had NaN in all magnitude values. Therefore, it was easy to decide to drop these entries from the dataset entirely. There were only approximately 200 galaxies that had missing magnitudes, bringing the total number of samples to 14348. There were 65 objects that had NaN values in one of the remaining magnitude fields, leaving 14283 complete data points in

total. In the final dataset, each object has 62 variables. There are no entries with missing variables or NaNs after data wrangling.

Of the 62 variables, 6 of them are metadata. Four of these reference the object's location in space and in the Chandra imaging. One is the classification of the object, which has already been filtered to contain only galaxies. One of these pieces of data, MU_z_ml, is the redshift that was calculated by the Wolf et al paper using machine learning methods. This can be used at the end of the investigation to compare the results of my model.

The ranges of the values and their correlations are the most important factors to investigate with this data. Additionally, the errors on the magnitudes and photon fluxes are important to see if there is a large range in the reliability of the data. Finally, it is important to see if there are any extra variables than can be calculated from the data that could be more useful in the final model. The cleaned data is contained in the csv file galaxies.csv.

For ease of investigation, the data was split into 7 dataframes, containing each type of data in the dataset: R-band data, magnitudes, magnitude errors, photon fluxes, photon flux errors, redshift data, and metadata. Magnitudes are an logarithmic measurements of the brightness of a galaxy. The magnitudes in this dataset are absolute magnitudes. A galaxy with a magnitude of -15 is 100 times less luminous than a galaxy with a magnitude of -20. The magnitudes in all seven bands form distributions with modes between -20 and -22 magnitudes. All distributions have an elongated tail towards less negative values, suggesting that galaxies are more likely to the less luminous than the mode. The mean value is shifted to the right of the peak of the distribution.

The photon fluxes are the observed brightnesses across different bands. The first 13 fluxes are measured in bands from 420nm to 915nm. The last five fluxes are from measurements in the traditional broad spectral bands, ultraviolet, blue, visible, red, and infrared. As the name broad band suggest, each of these bands covers a wider range of wavelengths than the previous bands. The units are photon flux densities, photons/m^{2}/s/sm. The vast majority of the photon fluxes are close to zero, therefore a log scale for the y-axis is best suited for visualizing the data as a histogram to show the shape of the distributions.

The magnitudes and photon fluxes both have corresponding errors. Histograms of these errors show that overall the errors are very low. There are more outliers for the magnitude errors than for the photon flux errors. This is expected, as magnitude are a calculated value whereas photon fluxes are measured directly by the telescope.

The redshifts for the data set range from 0 to 1.6, corresponding to distances of approximately 0 to 12000 Mpc from Earth. The distribution of redshifts is not normal. This is expected because the distribution of galaxies across the field of view of the Chandra Deep Field South's survey is not expected to be normal. Closer galaxies are easier to see and properly classify, corresponding to more low redshift objects. Galaxy clusters could cause spikes in redshifts around certain values. The tapering off of the distribution towards the high end of the

range of redshifts is expected, as higher redshift objects would be more likely to be obscured or not positively identified as galaxies.

At this point in the analysis I began making graphs of the different variables to look for interesting trends in the data. Plotting all of the data in the redshift data frame in a pairplot showed that the redshifts are spread between 0 and 1.5 and most fractional errors are low. There is a slight positive correlation between redshift and the error of the redshift which is expected. Histograms were plotted of all of the magnitudes. I calculated the fractional error of the magnitudes and plotted some against their respective magnitudes. All were as expected, with the fractional error staying consistent across the values of the magnitude. Plotting the redshifts vs the errors, chi squared, and fractional errors show that there is very little correlation between the redshifts and the chi squared values or the redshifts and the fractional errors. This is good, given that if there was a positive correlation then the higher redshift values would be less reliable in general. There is a small increase of fractional errors towards values of z=0, which is expected because the denominator would be approaching zero while the numerator stayed roughly constant.

The fractional errors for the magnitudes and fluxes show that almost all of the fractional errors are close to zero with only a few outliers for each magnitude and flux. For the magnitudes, the outliers occur throughout the range of the distribution of magnitudes, centered around where the distribution is most dense. For the fluxes, the outliers of fractional errors all occur close to the flux equaling zero. This occurs for the same reason that there was an upward trend for fractional errors on the redshifts - when the denominator is close to zero the fractional error is more likely to be large.

Histograms of the photon fluxes showed that most of the values were only slightly more than zero. For all fluxes, the vast majority of objects had fluxes less than 0.33 with values rarely reaching higher than 3. Next I looked for correlations between the different magnitudes and fluxes. First I cut down the number of flux columns. Some of the flux columns are data from the same band taken over different runs. These columns were very close to identical, correlated with a Pearson R > 0.99 when plotted against each other. I dropped all but one column from each of these sets of observations from the same flux band. Next I used the Pearson R coefficient to select which pairs of magnitudes and fluxes were highly correlated. Plotting a few again each other it was easy to see that they all follow a similar trend but are not devoid of interesting features.

Next, I plotted the magnitudes and fluxes against the redshift. The magnitudes were all relatively negatively correlated, which aligns with the scientific intuition that further objects should be more dim. The correlations between the magnitudes and redshifts could be vital to building a model to predict the reshifts. All seven of the magnitudes are inversely correlated to the redshifts. Additionally, each graph follows an inverted 'S' curve, a distinct shape showing that the least luminous galaxies could only be seen at low redshifts. There is no correlation between

the redshifts and the errors in the magnitudes. Therefore the magnitude errors may not be useful in the model.

For the next comparison, I plotted the fluxes against the redshift. Each of these showed that most of the high flux values cluster between redshift 0 and 0.2, which will most certainly be a feature of interest in the model. There is not a linear correlation between the fluxes and the redshifts as there was for the magnitudes and redshifts. For each band, the photon flux only ranged above 1 photons/m^{2}/s/sm for redshifts below around 0.2. As expected, the Chandra Deep Field South survey was able to find more photons associated with closer galaxies than farther galaxies. As with the magnitude errors, there is not a meaningful correlation between the flux errors and the redshifts.

The rates of extinction for different wavelengths of light across space leads to the assumption that the subtractions of magnitudes could display a meaningful pattern. When plotted against the redshifts, each subtraction shows a band where the differences in the magnitudes is larger. For example, the subtraction of UjMag-BjMag shows a band from approximately redshift 0.8 to 1.2. On every graph of these differences, there was a range where the difference between the magnitudes grew to approximately plus or minus two. Not all subtractions have the band at the same range of redshifts, suggesting that this subtraction method could be utilized in the model. These graphs all showed a cluster of magnitude subtraction values close to zero across all redshifts. The redshift correlating to this bump in the data differed for each pair of magnitude subtractions, but this bump could be one of the keys in identifying the redshift.

Overall, the plots show a consistent dataset that is well cleaned and ready to be used in a model. The fractional errors on the magnitudes and photon fluxes show that less than 1% of the data has large errors. I believe that this can be taken into account in a model without compromising the predicted reshifts.

The next stage in the project was exploratory data analysis. The main purpose behind this stage of exploratory data analysis with the COMBO-17 data was to solidify the idea that there is a statistically significant relationship between both magnitudes and redshifts and photon flux and redshift. I began by plotting the empirical cumulative distribution function (ECDF) of each magnitude and flux versus the ECDF of a normal distribution with the same mean and standard deviation. This showed that all magnitudes are close to a standard deviation. The tails of the magnitude ECDFs have more points than would be expected by a normal distribution. The fluxes, however, are not a good fit for a normal distribution. The distribution for each of the photon fluxes is heavily skewed towards higher values. However, this should not be an issue for performing bootstrapping on these variables as there are enough data points that the bootstrap replicates should still form a normal distribution by the Central Limit Theorem.

Visually, plotting magnitude versus redshift appears to show a relationship between the two variables. To test if this relationship is statistically significant, I used the null hypothesis that the mean of a magnitude for objects with redshifts between 0 to 0.8 and the mean of the same magnitude measurement for objects with redshifts between 0.8 to 1.6 would be the same. In other

words, the mean of M(0 to 0.8) - M(0.8 to 1.6) equals zero. To test this null hypothesis, I took 10000 bootstrap replicates of the mean for the magnitudes with a low redshift value and the magnitudes of a high redshift value, took the difference, and then calculated the 95% confidence interval and the p-value. For all seven magnitudes, the 95% confidence interval was well above zero and the p-value equaled zero. Therefore, the null hypothesis was easily rejected.

Next, I refined my analysis to see if I could find for which redshift intervals the magnitudes displayed a statistically significant change. To do this, I repeated the bootstrapping method I already tried but on much smaller samples. For this investigation, I took three redshift divisions: 0.05, 0.1, and 0.2. For each of these divisions, I constructed a set of bootstrap replicates comparing the magnitudes from r-x to r and r to r+x, where r is the central redshift of each investigation and x is either 0.05, 0.1, or 0.2. For example, I compared the average of the magnitudes of objects with redshifts between 0.4 to 0.6 to the magnitudes of objects with redshifts between 0.6 to 0.8. The null hypothesis was once again that the difference in the average magnitudes in these redshift bins would be zero.

By doing this investigation for bins of 0.05, 0.1, and 0.2, I could find how closely magnitude and redshift are correlated. The main question I was asking is that at which level are there statistically significant differences in the means of the magnitudes? When looking at redshift bins of 0.2, for each of the seven magnitudes either 0 or 1 of the confidence intervals crossed over zero. Therefore, for redshift intervals of 0.2 the average of the magnitudes is almost always different. For redshift bins of 0.1, approximately 15% of confidence intervals passed through zero. The plot of these confidence intervals also clearly shows that the higher values of redshift are harder to distinguish using magnitudes. Finally, the redshift bins of 0.05, approximately 35% of the confidence intervals crossed over zero. Therefore, magnitudes may not be the best predictors of redshifts at a precision of 0.05 or less.

After finishing this analysis of magnitudes, I repeated the process for the photon fluxes. Plotting photon flux vs. redshift shows that there is a large variation of photon fluxes at redshift values of less than approximately redshift 0.5. Above approximately redshift 0.5, these basic scatter plots do not show a definitive relationship between the photon flux and redshift. Therefore, this investigation of confidence intervals was important to know if photon fluxes would be useful for predicting high redshifts. As with the magnitudes, the confidence intervals showed that fluxes differed for redshift intervals of 0.2. Redshift intervals of 0.1 showed that there was sometimes still enough differentiation for the photon flux to be useful in predicting redshift at this precision. Surprisingly, the lowest redshifts had the largest confidence intervals. Finally, redshift intervals of 0.05 do not appear to produce statistically significant differences in the averages of the photon fluxes. Therefore, the photon flux alone is only enough to predict the redshift to between a precision of 0.1 and 0.2.

Finally, I repeated the analysis for the magnitude subtractions as calculated and plotted in the data story. These plots do not show as clear of patterns as the same analysis for the

magnitudes and fluxes. However, these plots do imply that for redshift intervals of 0.2 the differences between the averages of the magnitudes are mostly not zero.