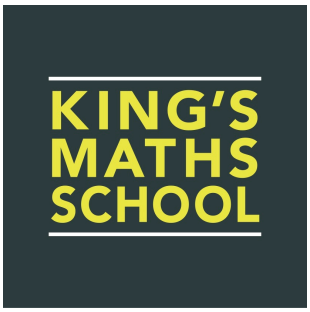
King's College London Mathematics School



King's certificate

Estimating Galaxy Masses from 3D Survey Data

Group P

Thivyaa Rahulan, Shenjun Lu, Buse Ozturk, Aramis Marti Shahandeh

Project Mentor: Lee Stothert

June 25, 2021

Abstract

The aim of this project was to investigate whether a machine learning model would be able to replicate the galaxy mass estimates made by the SDSS. We began this project by deriving an equation to show the link between stellar mass and luminosity. From this, we tested two supervised learning models and then compared the results produced by the better of the two to the existing mass estimates. Our results showed that our model was successful at replicating these estimates, but only when supplied with sufficient information. This was a significant result as it means that machine learning models could be used to produce other important cosmological research.

1 Introduction

Due to the fact that the majority of the information about the universe is unknown, cosmologists tend to rely on theories and models to conduct their research. In estimating the growth of structures over time in the universe, galaxy masses are a metric of interest. However, as these are not directly observable, we require models to estimate these values based on what can be observed.

The first time we had proof of other galaxies existing in our universe was in October 1923, when Hubble saw a Cepheid variable at what he thought at the time was the Andromeda nebula. By timing the 31-day cycle of the star as it flickered, Hubble deduced its distance to be vastly outside the previously estimated bounds of the Milky Way. Therefore, that nebula then became known as the Andromeda Galaxy [1].

Since this revelation, we have discovered many other galaxies that exist in our universe and, with this newfound knowledge, cosmologists have strived to find out more information about them. The Sloan Digital Sky Survey (SDSS) has created the most detailed three-dimensional maps of the universe ever made, with deep multi-colour images of one third of the sky, and spectra for more than three million astronomical objects. It began regular survey operations in 2000, after being designed and constructed for 10 years. There have been several phases, SDSS-I (2000 – 2005), SDSS-II (2005 – 2008), SDSS-III (2008 – 2014), and SDSS-IV (2014 – 2020), with each phase involving multiple surveys that have interlocking goals [2].

Data Release 16 (DR16) is the fourth data release of the fourth phase of the SDSS, SDSS-IV, containing all the SDSS observations through to August 2018 [3]. It includes the latest and final data release of optical spectra from the SDSS component extended Baryon Oscillation Spectroscopic Survey (eBOSS), the most current reprocessed imaging and spectra from the SDSS legacy survey, and the legacy sets of estimates for intrinsic properties of galaxies from Data Release 12 (DR12) [4] which contains three different estimates for the stellar mass and velocity dispersion estimators for galaxies. Therefore, this data release is likely to be the most useful for our project.

This project aims to develop a machine learning model to forecast the mass estimates of a galaxy using its brightness at different wavelengths. We intend

to approach this by applying existing mathematical relations such as the mass-luminosity relation, the initial mass function, Wien's displacement law and the Stefan-Boltzmann law, to determine a very small-scale model for a star. From this, we will then think of a galaxy as a single stellar population to create our first set of results based on a machine learning model and compare them to the existing estimates within the SDSS data. These estimates should hopefully replicate the current mass estimates found from DR16 [3] therefore prove that it is possible to train a model on a large, simulated galaxy catalogue and then apply those results on the real data.

2 Literature Review

The study of the known universe is constantly evolving as technological advancements enable us to discover new insights into the way the world beyond our planet works. There is a long history of research into our neighbouring galaxies, with many papers being written on the theoretically derived formulae which we have used to estimate galaxy masses. In order to understand how we should approach this task, we must first understand the results that we are trying to replicate as well as the mathematical functions and processes previously used to undertake similar projects.

To begin our research, we decided it would be best to begin with investigating and analysing the existing SDSS data in order to get a better understanding of how those stellar mass estimates were gained. The first article we studied used 'star count data' to infer more precise details about the structure of the Milky Way [5]. The authors of the article: "Stellar Population Studies with the SDSS. 1. The Vertical Distribution of Stars in the Milky Way" are Bing Chen et al., all of whom worked at a well-respected university or institution, such as John Hopkins, Fermi National Accelerator Lab and Princeton and most had written several articles related to similar topics. Robert Lupton and Zelko Ivezic in particular have collaborated on multiple occasions and work at the same institutions. Due to this, this source is reliable to use.

This article detailed building a galaxy model in the SDSS photometric system which means making a galaxy model using stars' positions in the colour spectrum and then using this model to parameterise the vertical distribution of stars in the Milky Way. This article is useful for our brief because it is detailing a project which is fundamentally very similar to ours as it is about creating a model to predict a certain characteristic about stars, and thus their stellar populations. The intended audience for this article is people who already well-versed with the concept of stellar populations, are familiar with the SDSS and its data and how to interpret it and its (u'g'r'i'z') photometric system. However, this article does have some limitations: firstly, it only included data from the Milky Way, so we do not necessarily know if this method will be applicable for all galaxies we look at, and this article is quite complex to fully understand for someone who is not experienced in this field of study.

Having now understood how data received from the SDSS can be interpreted

and used to generate concrete results, we decided to narrow our research topic and investigate an article dedicated to the data release whose data is contained within Data Release 16 [3], with that data being what is most relevant to our project – Data Release 12 [4]. In "The Eleventh and Twelfth Data Releases of The Sloan Digital Sky Survey: Final Data from SDSS-III", Shadab Alam et al. analyse the vast amount of new data from the four interlocking surveys of SDSS III (from Data Release 11 to 12): Segue 2, BOSS, APOGEE and MARVELS and it also mentions the future SDSS IV [6]. This is useful for our project as it provides us with the data which we are trying to replicate, as well as a detailed understanding of how the SDSS achieved these results. This article is intended for scholars who want to see data from the SDSS and want to understand what it means and its implications. The limitations for this article are that one needs to be familiar with short abbreviations of commonly used scientific words and that the paper itself contains a lot of raw information to work through. We believe that the authors of this article are reliable as this article was the result of 135 universities (of which the authors were part of), scientific institutes and observatories, and was a large international effort involving well-known, trusted institutes.

With the previous two articles detailing results from the SDSS surveys, exploring how mathematical functions and relations are used to work out stellar masses from their luminosities seemed to be the next logical step. By doing this, we would be able to understand how to develop our mass estimates. The first of those articles uses the initial mass function. The article by Romeel Davé, "The galaxy stellar mass-star formation rate relation: evidence for an evolving stellar initial mass function?", provides a focused view of the understanding of the implications of observed M∗- SFR relation for the theoretical view of how galaxies accumulate stellar mass [7]: the evolution of the galaxy stellar mass-star formation rate relationship (M∗-SFR) provides key constraints on the stellar mass assembly histories of galaxies. It also provides both evidence and evolution of the initial mass function. Davé's article presents varies of results by comparing the relation for simulations using theories and observations, later bringing all the observations together to conclude an evolution of the initial mass function (IMF). This paper is aimed at people who have knowledge of the M∗-SFR relation with a good understanding of the formation of stars. The article provides detailed explanations, such as the implication of the SFR leading to evolving IMF and further showed that the goal of the evolving IMF is to increase the ratio of the SFR of high-mass stars (SFR_{hiM}) to total stellar mass formed, such that the inferred sf(z) is non-evolving. However, it was not our main resources as we only require understanding of the concept of the relationship but not its detailed pieces of evidence. The topic of the paper is too specific, meaning it was only an aid for our research and our understanding. Davé is the Chair of Physics at the University of Edinburgh. He also has previously released several articles around the formation of stars and analyse simulations of galaxies which makes him a reputable source of research.

The next two articles deeply examine the mass-luminosity relation. The first provides a broader overview of how the relation was developed, with a particular

focus on red dwarfs, whereas the second solely investigates the mass-luminosity relation for more massive stars.

"The Mass-Luminosity Relation from End to End" was written by Todd J Henry, who is an American astronomer and Professor of Astronomy at Georgia State University [8]. He is also the founder and director of the Research Consortium of nearby stars, making him a credible resource. In this article, he provides a review of the progress made in mapping this mathematical relation over the course of two decades. Henry places special emphasis on red dwarfs which tend to make up roughly half of a galaxy's stellar mass, despite the fact that they have relatively low stellar masses. This article proves to reliably compare and track the progress of previous articles and studies which have been dedicated to deriving the mass-luminosity relation based on observational results. It also contains the core of our research as well as a detailed analysis of the method that we are using to estimate galaxy masses based on stellar contribution. It is most likely that this paper was intended for university students as it functions as a collection and explanation of previous research so it would serve as a useful catalogue of information for a student looking into this subject. This article has very few limitations as it is well suited to our topic of investigation: however, due to it covering a broad range of studies and mainly focusing on red dwarfs, it lacks the specificity of information about larger stars which is more focused on in the next article: "Mass-Luminosity Relation for Massive Stars" by EA Vitrichenko et al. [9]

Here, the authors collected a catalogue of 73 stars of solar masses between 10M and 50M from previous data collections to try and develop a mass-luminosity relation based on the observable data from massive stars. Both Nadyozhin and Razinkova are researchers for the Institute of Theoretical and Experimental Physics in Moscow, Russia and Vitrichenko has produced 43 publications under the Russian Academy of Sciences, making them very reputable and credible authors. It is most likely that their article is intended for their peers in research as it provides a comparison with the existing theoretical mass-luminosity relation found by Nadyozhin and Razinkova in their 2005 article, "Similarity theory of stellar models and structure of very massive stars" [10]. The research focuses on larger stars to try and discern if there are any differences in the relation for stars with larger masses. This article is useful in our research as, by using different mass-luminosity relations for different populations of stars based on size, we will be able to generate more accurate estimates for galaxy masses based on their luminosities. The main limitation of this article is the fact that it only considers massive stars, rather than all types, so we cannot see how the mass-luminosity relation changes for stars of different sizes. However, when paired with the previous article, we have a very in-depth grasp of the mass-luminosity relation.

The final article that we examined described a method of estimating stellar masses. Having studied the various component parts that will comprise our project, investigating a previous attempt at estimating galaxy masses would provide a viable way in which we could approach our project.

"How well can we really estimate the stellar masses of galaxies from broad-

band photometry?" was written by Peter D Mitchell et al. [11] Mitchell works at the Institute for Computational Cosmology in the Department of Physics at Durham University, making him a credible source. In this article, Mitchell reviews the sufficiency of the information about galaxy brightness and gradient in estimating the stellar masses of galaxies and explores how the mass estimates of the SEDs (spectral energy distributions) can give different results due to the brightness of the galaxies. The research is based on how close the SED fittings are to the established values of the stellar masses of galaxies and whether they can be reproduced at different redshifts. The article reviews the relation between SFH and dust particle masses with an average stellar mass fit. It is useful for our project because it demonstrates a method for estimating galaxy masses as well as the potential inaccuracies and errors that may occur. It will also be a useful guide for laying out our approach during the course of the project. However, as it uses a different method from our project, the data itself is not very relevant for our project.

Our research showed us that we tend to rely on theoretical formulas that have been previously derived to create estimates such as galaxy masses on account of there being no feasible way to actually measure these values. The Mass-Luminosity Relation and Initial Mass Function will primarily be what we use in our project as they are used to link the luminosity of a star to its mass and the initial distribution of masses for a population of stars. "The Mass-Luminosity Relation from End to End" [8] proved to be a very useful explanation of this relation which will be crucial to the first step of our project. This, supplemented by the paper about the Mass-Luminosity Relation of massive stars [9], will help us to create our first, simple model that will be based on a star. The article by Davé investigated the initial mass function which is what we will be using the second part of our project when we will be looking at stellar populations. This information about stellar populations was improved by the article by Bing Chen et al. [5], but this proved to be more specific and less helpful to our project. However, the article about the 11th and 12th SDSS data releases [6] had analysed the data which we intend to try and replicate and therefore shown us what sort of results we should be expecting from our machine learning model. Finally, the article about broad-band photometry [10] was a very clear explanation and demonstration of how observational results can be used theoretically to estimate values, as well as which errors can occur along the process. Overall, our research has been effective as we have determined which areas we need to focus on and how to approach our machine learning model.

3 Methodology

3.1 Method Overview

This paper aims to replicate the stellar mass estimates of the legacy SDSS, specifically those from the Portsmouth group [12] within DR16 [3], using a machine learning model. Therefore, once we had researched the data which we

would compare our results to and the methods of estimating stellar masses, a detailed enough understanding of our project had been established so that we could move on to practical methods. In order to achieve this, we must first derive the equations that we need to create the necessary parameters for our models. From there, we can use these with an appropriate machine learning model to generate results that are comparable to those from the SDSS.

3.2 Plan

At the beginning of the project, we took steps from the project brief and compiled them with our deadlines and targets in a Gantt chart (see Figure 1). By doing this, we were able to delegate research tasks effectively which resulted in us being able to complete our tasks efficiently with suitable time margins.

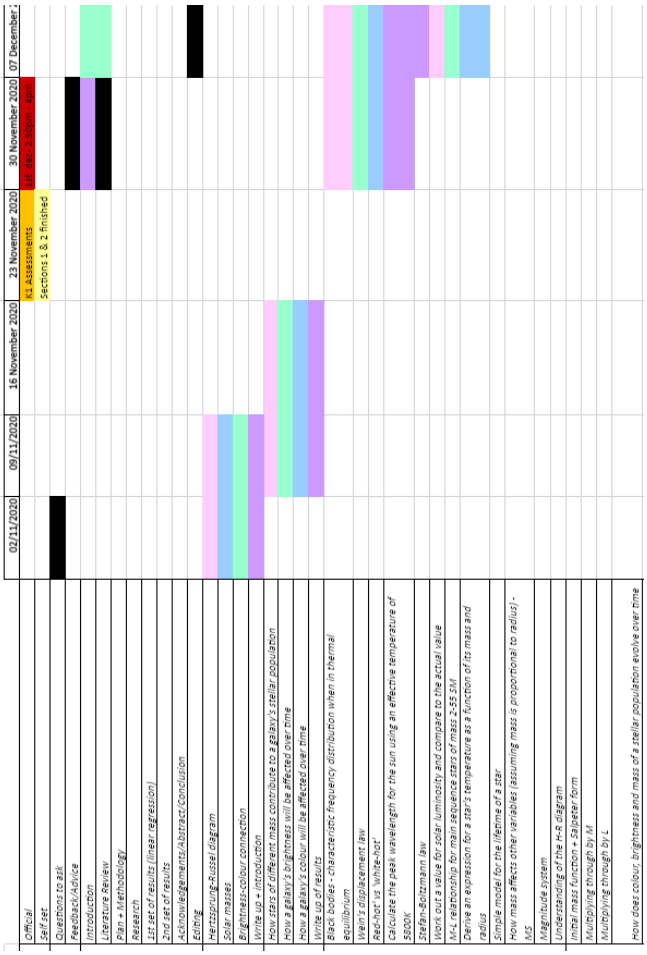


Figure 1: Our Gantt Chart

As shown in Figure 1, the whole group worked on the research together, dividing the sections between us equally. However, when it came to developing a computer model, we decided to allocate the following tasks according to our skill sets – Thivyaa focused on the write up of our article while the others worked on developing the machine learning model, with Shenjun also putting the final document into LaTeX. We also met up once a week to discuss our findings and share our ideas so that everyone understood each part of the research and the process.

3.3 Equations

We started by working of the equations which would be used to show the relationship between a galaxy's mass and its luminosity. This began with Wein's displacement law:

$$\lambda_{peak} = \frac{b}{T} \tag{1}$$

where T is the absolute temperature, λ_{peak} is the wavelength at which the radiance of a surface per unit frequency peaks at, and b is Wien's displacement constant.

This law states that the black-body radiation curve for different temperatures (Figure 2) peaks at wavelengths that are inversely proportional to temperature [13]. This diagram is called the Planck Spectrum.

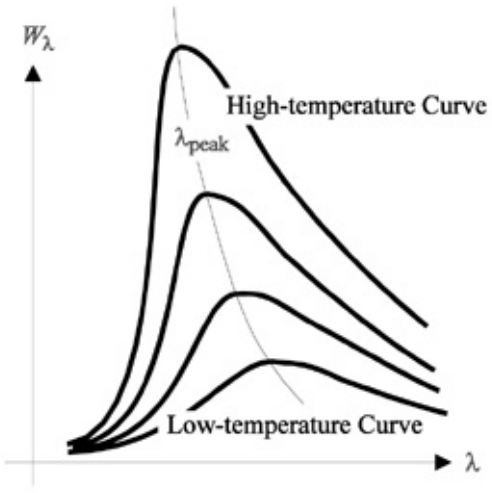


Figure 2: The Planck Spectrum

The Planck Spectrum can demonstrate how the colour of a star corresponds to its temperature by seeing which range of wavelengths (lambda peak) lies within. Based on this information, we can say that if the temperature of a

black body is low, it will appear more red. At high temperatures, it will appear more blue, violet or white the hotter it gets as shorter wavelengths come to predominate the black body's emission. This can then be applied to stars.

Next, we moved on to the Stefan-Boltzmann law which states that the total energy radiated per unit surface area of a black body across all wavelengths per unit time is directly proportional to the fourth power of the black body's thermodynamic temperature [14]. This law is written as:

$$L = 4\pi R^2 \sigma T^4 \tag{2}$$

where σ is the Stefan-Boltzmann constant.

This law can be applied to estimate a value for a star's luminosity if we take it to be a perfect black body.

The equation we looked at next is the second most important equation for our project – the Mass-Luminosity Relation [15]. The link between a star's mass and its luminosity varies depending on the mass range in which the star lies and is written as:

$$\frac{L}{L_\odot} \propto \left(\frac{M}{M_\odot}\right)^a \tag{3}$$

where $L_\odot$ and $M_\odot$ are the luminosity and mass of our Sun respectively, and 1 < a < 6. For main sequence stars which have masses $2M_\odot <$ M $< 55M_\odot$ the equation is:

$$\frac{L}{L_\odot} \approx 1.4 \left(\frac{M}{M_\odot}\right)^{3.5} \qquad (2M_\odot < M < 55M_\odot) \tag{4}$$

This equation, when combined with the Stefan-Boltzmann law can be used to derive an expression for a star's temperature as a function of its mass and radius. By applying the Stefan-Boltzmann law to our Sun and dividing the previously established equation by this new one, we get:

$$\frac{L}{L_\odot} = \left(\frac{R}{R_\odot}\right)^2 \times \left(\frac{T}{T\odot}\right)^4 \tag{5}$$

where $R_\odot$ is the Sun's radius.

We can then replace $\frac{L}{L_\odot}$ with $1.4\left(\frac{M}{M_\odot}\right)^{3.5}$ to give us:

$$\frac{1.4\left(\frac{M}{M_\odot}\right)^{3.5}}{\left(\frac{R}{R_\odot}\right)^2} = \left(\frac{T}{T_\odot}\right)^4 \tag{6}$$

which is the equation we need [16].

Following this, we can make a simple model for the lifetime of a star using the concept that mass is energy as it makes sense to assume that the lifetime of a star is proportional to how fast it uses up its mass of nuclear fuel. A star's

luminosity is a measure of its energy output, which demonstrates how rapidly this occurs. Therefore, the lifetime of a star would be proportional to the mass of fuel available, divided by luminosity, assuming luminosity to be constant. This can be expressed in this form:

$$t = k\frac{M}{L} \tag{7}$$

where k is some constant.

We can then use the mass-luminosity relationship for a main sequence star and the approximation of the lifetime of the sun being 10^{10} to calculate the lifetime of a main sequence star. Using this, stellar lifetime can be expressed as such [17]:

$$\tau \approx 10^{10}\left(\frac{M}{M_{\odot}}\right)^{2.5} yrs \tag{8}$$

Now that we have done this, we can move on to the most important equation for this project: the initial mass function (IMF). The IMF gives the estimated distribution of stellar masses we would expect to find in a new formed stellar population – it is the relative numbers of stars of different masses that emerge, averaged over star formation regions throughout the galaxy [18]. The IMF for stars in the solar neighbourhood was determined by Edwin Salpeter in 1955. We can write it as:

$$\xi(m)dm \sim m^{-2.35}dm \tag{9}$$

We can convert from the number of stars of each mass to the mass contribution at each mass by multiplying through by m:

$$m\xi(m)dm \sim m^{-1.35}dm \tag{10}$$

which still drops rapidly as stellar mass decreases. This tells us that the majority of the mass of a stellar population is found in the large number of low mass stars. We now want to understand which stars contribute the majority of the light of a stellar population. To do this we multiply through the Salpeter IMF equation to give us the light contribution at each mass. Using the Mass-Luminosity relationship of a main sequence star, we can write:

$$L(m)\xi(m)dm \sim m^{1.15}dm \tag{11}$$

This equation now increases with mass instead of decreasing. Therefore, while the majority of the mass of a stellar population comes from the many low mass stars, the majority of the luminosity comes from the very few high mass and extremely bright stars. This is shown by the steep increase of mass with luminosity which is steep enough that the dropoff in the number of high mass stars is overcome.

The equations derived above can be used to show how the mass of a galaxy will vary with brightness as well as colour. We found that brighter galaxies have

more mass and that, at a fixed brightness, redder galaxies would also be more massive. This is because a larger number of low luminance red stars would be required to equal the brightness of a galaxy that had a younger, bluer, or more luminous population. These relationships will be visually represented in the first section of our results.

3.4 Choosing A Machine Learning Model

There are three types of machine learning model algorithm: supervised learning, unsupervised learning and reinforcement learning. Our project aims to forecast the mass estimates of galaxies using brightness at different wavelengths. Given we have a well defined target variable in the masses and well defined inputs in the brightness values we decided that supervised learning was the most appropriate for our problem. Next, we looked through the different supervised learning algorithms and, in the end, we focused on two: linear regression and random forest. Linear regression aims to model the target variable as a linear sum of the input variables, and is arguably the simplest machine learning model. The random forest model fits many decision tree models to random draws of the data to fit to built a robust forecast, and is a good representation of a "non-linear" approach.

3.5 The Pre-existing Data

Each SDSS Data Release includes four types of data: images, optical spectra, infrared spectra and the parameters measured from images and spectra, such as magnitude and redshift. Imaging data provides details on the SDSS imaging pipeline, the calibration process, and which quantities are available in the imaging catalogue data. Optical spectra data includes SDSS two optical spectrographs (SDSS-I and BOSS) and provides details on associated data including target flags, redshifts, and classifications. Infrared spectra data explains what data is available from the SDSS's new infrared spectrograph and provides details on associated data which includes information on the spectra, targets, radial velocities, and determinations of stellar atmospheric parameters [12].

All this data is used to produce different models of stellar populations. Two of these models are assigned to be used for the modelling of galaxies with the observed colours from red to blue, which uses an evolutionary population synthesis model. The output of this model comprises of spectral energy distributions, colours, stellar mass-luminosity rations, bolometric corrections and near-infrared spectral line indices [19].

The training set is the data used to fit a machine learning model and the test set is what we try out our model on to see how well it has worked. In this case, we use the colour bands and mass values from a list of stars to determine a link between the two via the model. The list is split into two subsets using random selection into the training and test sets, so that our data is representative of the original. Our mask for our training set is slightly different than that of the

test set, meaning that the range of redshift that we are examining is slightly different and thus we have a different set of stars.

4 Results and Discussion

4.1 Demonstrating theoretical relationships

In this section we are going to visually represent the conclusion that mass increases steeply with luminosity and that it overcomes the drop-off in the number of stars by mass in the IMF, which we found from our derived equations. This could be achieved on a graph by simply plotting mass against luminosity, but the inputs need to be manipulated so that our plots match our equations as closely as possible.

For each graph, we used a range of redshift for our mask to ensure the bands look at the same wavelengths across galaxies in the sample. The colour bands that we decided to investigate were 'u' which is ultraviolet light, 'g' which is green light, 'r' which is red light and 'i' and 'z' which are in the near-infrared part of the electromagnetic spectrum. We used both the absolute and apparent magnitudes for stellar brightness, where absolute magnitude is the brightness of a star as seen from a distance of 10 parsecs, which is 32.6 light years, and apparent magnitude is the brightness of a star as seen from Earth. In both of these scales, the smaller, or more negative, the value of magnitude, the brighter the star. As we were plotting the properties of individual stars, we used a scatter graph so that we would be able to determine whether the correlation matched our equations or not.

We began our attempts by plotting the log of the stellar masses against the absolute magnitude of the stars within the colour band 'r' and produced Figure 3. This graph used redshift values between 0.08 and 0.12 and it shows brighter galaxies having higher masses that we would expect to see from the Mass-Luminosity relation [15].

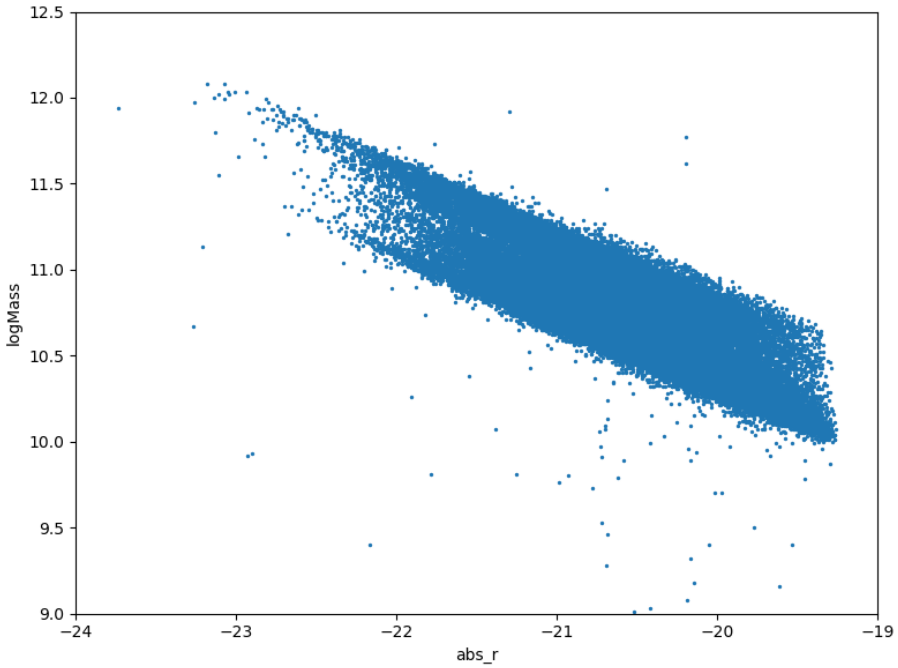


Figure 3: Log(Mass) plotted against the absolute magnitude of the stars in colour band 'r'

In our next graph, we plotted the difference between the colour band filters 'u' and 'r' (the same in both apparent and absolute magnitudes), against the absolute magnitude values within colour band 'r' with a colour spectrum along one side which was labelled with the values of log(mass) (see Figure 4). The same range of redshift values was used.

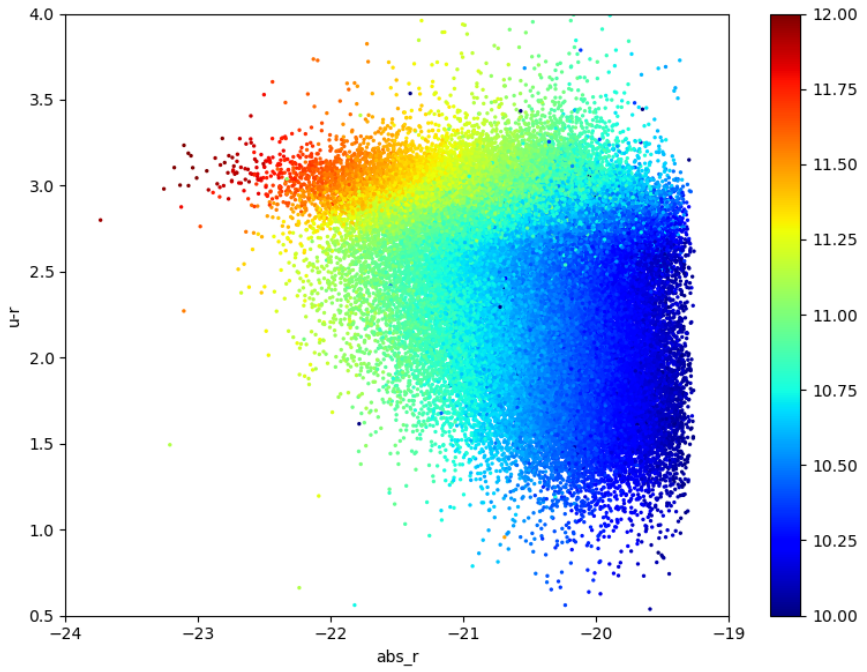


Figure 4: The apparent magnitude of the difference between colour band filters 'u' and 'r' plotted against the absolute magnitude of stars within colour band 'r', showing the colour of each star and the log(mass) corresponding to each colour.

This graph clearly shows that the redder galaxies tend to be the brighter ones. When looking at individual values of absolute magnitude, we can see that heavier galaxies at that brightness are usually the ones that emit light that is more red. Therefore, we have also graphically shown the relationship shown by equation 10 where we found that the majority of a galaxy's mass came from the large number of red, low luminance stars, meaning that, at a fixed brightness, the heavier galaxies are those which are redder.

Having demonstrated the links that we found from deriving the equations in section 3.3, we can move on to finding the best machine learning model set up for us to generate a series of stellar mass estimates that are comparable to those found by Data Release 16 of the SDSS [3].

4.2 Choosing the best variables and machine learning model

Having previously decided to use either a linear regression or a random forest machine learning algorithm, now came the time for us to compare the two and decide on the better one to try and replicate the galaxy mass estimates of the SDSS.

In the earlier parts of the research, we have seen that the mass of the galaxy is mainly affected by lower mass stars which are the redder ones. We therefore had a prior belief that the redder filters would produce better results, with the best being the colour band abs_z.

When inputting all the colour bands into our machine learning models, we used more information, so the algorithm would be able to determine the best correlation.

In order to determine the better model, we obtained the root mean square distributions (RMSD) of the data that our models produced against the existing Portsmouth mass estimates. We split the data into training and test sets, the training data was fed in for the model to learn and the test data was data the model had never seen before. The results are displayed in Table 1 below [12].

Linear Regression	Random Forest
$['abs\_u']$ train score: 0.31073433 test score: 0.30951253	$['abs\_u']$ train score: 0.2992366863412904 test score: 0.30869030366130723
$['abs\_g']$ train score: 0.23553129 test score: 0.23410158	$['abs\_g']$ train score: 0.222714061204517 test score: 0.23440753944445897
$['abs\_r']$ train score: 0.1806378 test score: 0.1786006	$['abs\_r']$ train score: 0.16718630858188407 test score: 0.18181389818206428
$['abs\_i']$ train score: 0.16740538 test score: 0.16045268	$['abs\_i']$ train score: 0.15448340463087298 test score: 0.16331865167295995
$['abs\_z']$ train score: 0.14985521 test score: 0.15324342	$['abs\_z']$ train score: 0.13945258754258388 test score: 0.15471695964803753
$['abs\_r', ' u - r']$ train score: 0.14755875 test score: 0.13915081	$['abs\_r', ' u - r']$ train score: 0.0688971330580227 test score: 0.09636822290805444
$['abs\_u', ' abs\_g', ' abs\_r', ' abs\_i', ' abs\_z']$ train score: 0.12728421 test score: 0.1235393	$['abs\_u', ' abs\_g', ' abs\_r', ' abs\_i', ' abs\_z']$ train score: 0.049701354379157806 test score: 0.07497335198599914

Table 1: The RMSDs of our model-produced data for both the linear regression and random forest algorithms for different input parameters

We found that the linear regression model did not work as well because it tried to fit everything to one line, and we saw from Figure 4 that the galaxies could not really be plotted along one straight line. For a single colour band, the results were far more comparable, but the test score was significantly better for the random forest model when using multiple parameters.

If we had a training score that was much lower than the test score, we could conclude that there had been overfitting. We saw train scores that were almost

equal test scores for the linear model, which suggested the model was not flexible enough to capture the relationships in the data. However, if the model had too much flexibility, the fit may have included too much noise in the data and may have not generalised well. Therefore, we judged our machine learning models based on the test score and not on the training score, as a complex model could arbitrarily fit any input data, but we cared about how well it generalised to the test data. In the random forest model we did have some overfitting which was expected in a well-fitting non-linear model, but, in our case, the test scores were still much better than the linear regression model.

4.3 Testing the model

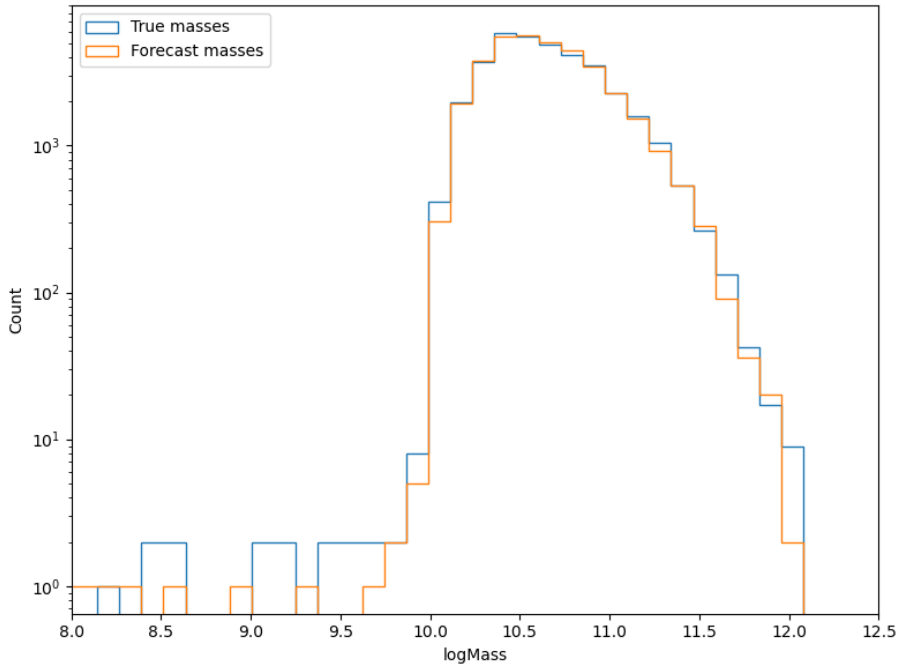


Figure 5: Stellar Mass Function; which shows the number of stars at each log(Mass)

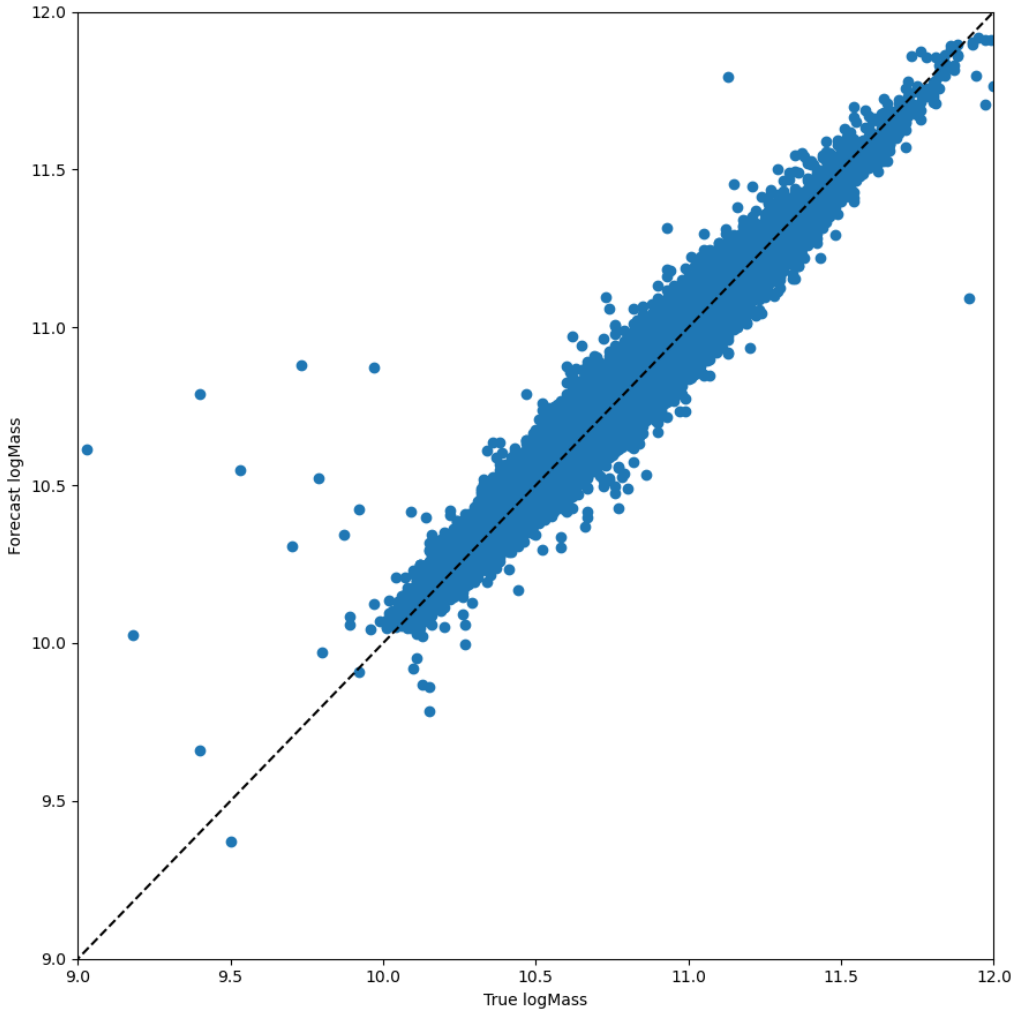


Figure 6: Forecast Log(Mass) plotted against True Log(Mass)

From Figure 5 we can see that the random forest model fits very well for values of Log(Mass) above 10, but below that, our results are insufficient as these lower mass stars weren't luminous enough to be seen by the SDSS creating these anomalous results.

Figure 6 also supports the accuracy of the random forest model, looking at the graph we can see that the majority of results are very close to, if not on the 'y=x' line which is the line that would represent when the Forecasted Log(Mass) is the same as the True Log(Mass). The majority of anomalous results are also in the same region as the Count x Log(mass) graph and the reason for these anomalous results is also the same as those for the Count x Log(mass) graph.

These graphs show proof of concept that galaxy masses estimated using machine learning methods can be used to perform important cosmological measurements.

5 Conclusion

The aim of our project was to test whether a machine learning model would be able to replicate the stellar mass estimates made by Data Release 16 of the Sloane Digital Sky Survey and, therefore, whether it would be possible to train a model on a large, simulated galaxy catalogue and then apply those results to real data.

We began by conducting the necessary research into analysis of previous data collected by the SDSS which included studies on stellar populations as well as on information contained within Data Release 16 [3], which was the focus of our paper. Following that, we moved on to study the existing equations and relationships used to calculate stellar masses based on luminosity such as the Initial Mass Function [18] and the Mass-Luminosity Relation [8]. These equations were built around the properties of individual stars and our plan was to derive a link between stellar mass and luminosity based on them which we would scale up to represent stellar populations and then galaxies.

In section 3.3, we used Wein's Displacement law, the Stefan-Boltzmann law, the Mass-Luminosity Relation, and the Initial Mass function to derive the equation that would show the link between a star's mass and luminosity. From this, we discovered that brighter galaxies would be those with higher masses and that, at a fixed level of luminosity, the redder galaxies would be heavier ones. This was because a galaxy made up of a larger number of dull, red stars would be needed to have the same luminosity as a galaxy containing younger, bluer, or brighter stars. In section 4.1, we presented this link visually in both Figure 3 and Figure 4.

From this, we were able to choose and test a machine learning model by seeing whether it produced data points based on the training set that fit the relationships that we had previously determined. We compared the results given by a random forest model and a linear regression model in section 4.2 and found that the random forest model's values had less variation from the mass estimates given by the Portsmouth group [12], meaning that it was the better of the two.

We then went on to analyse the model's success at replicating the existing galaxy mass estimates against the original estimates in Figures 5 and 6. In this comparison, we found that our model's accuracy was lower for lower mass stars as they do not tend to be bright enough to be picked up by the SDSS, resulting in anomalies. However, for higher mass stars, the random forest model replicated the results from DR16 [3] very accurately. Our results showed that using a machine learning model would be a viable method of estimating galaxy masses and that, therefore, they could be applied used for other important cosmological measurements. However, the use of a machine learning model is limited by the fact that the model requires a set of training data. Therefore, while we have

proven that models can reproduce existing forecasts, another application would require data from elsewhere, such as galaxy simulations.

Acknowledgements

Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, Center for Astrophysics | Harvard & Smithsonian (CfA), the Chilean Participation Group, the French Participation Group, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, the Korean Participation Group, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New York University, University of Notre Dame, Observatório Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

References

[1] Powell CS. *January 1, 1925: The Day We Discovered the Universe.* Available from: https://www.discovermagazine.com/the-sciences/january-1-1925-the-day-we-discovered-the-universe [Accessed 18th November 2020].

[2] SDSS. *The Sloan Digital Sky Survey: Mapping the Universe.* Available from: https://www.sdss.org/ [Accessed 2nd November 2020].

[3] SDSS. *This is Data Release 16.* Available from: https://www.sdss.org/dr16/ [Accessed 2nd November 2020].

[4] SDSS. *This is Data Release 12.* Available from: https://www.sdss.org/dr12/ [Accessed 2nd November 2020].

[5] Chen B, et al. Stellar Population Studies with the SDSS. I. The Vertical Distribution of Stars in the Milky Way. *The Astrophysical Journal*. 2001;553(1):184-197. Available from: https://iopscience.iop.org/article/10.1086/320647 [Accessed 1st December 2020].

[6] Alam S, et al. The Eleventh and Twelfth Data Releases of The Sloan Digital Sky Survey: Final Data from SDSS-III. *The Astrophysical Journal Supplement Series*. 2015;219(1):12. Available from: https://iopscience.iop.org/article/10.1088/0067-0049/219/1/12 [Accessed 1st December 2020].

[7] Davé R. The galaxy stellar mass-star formation rate relation: evidence for an evolving stellar initial mass function? *Monthly Notices of the Royal Astronomical Society*. 2008;385(1):147-160. Available from: https://doi.org/10.1111/j.1365-2966.2008.12866.x [Accessed 4th December 2020].

[8] Henry TJ. The Mass-Luminosity Relation from End to End. *ASP Conference Series*. 2004;318(159):159–165. Available from: http://articles.adsabs.harvard.edu/pdf/2004ASPC..318..159H [Accessed 8th December 2020].

[9] Vitrichenko EA, Nadyozhin DK, Razinkova TL. Mass-Luminosity Relation for Massive Stars. *Astronomy Letters*. 2007;33(4):251–258. Available from: doi:10.1134/S1063773707040044.

[10] Nadyozhin DK, Razinkova TL. Similarity theory of stellar models and thes tructure of very massive stars. *Astronomy Letters*. 2005;31(10):695–705. Available from: doi:10.1134/1.2075312.

[11] Mitchell PD, Lacey CG, Baugh CM, Cole S. How well can we really estimate the stellar masses of galaxies from broad-band photometry? *Monthly Journals of The Royal Astronomical Society*. 2013;435(1):87–114. Available from: doi:10.1093/mnras/stt1280.

[12] SDSS. *Galaxy Properties from the Portsmouth Group*. Available from: https://www.sdss.org/dr16/spectro/galaxy_portsmouth/ [Accessed 2nd January 2021].

[13] Riedl MJ. *Optical Design Fundamentals for Infrared Systems*. 2nd ed. Washington: SPIE Press; 2001. Available from: https://spie.org/publications/tt48_153_wiens_displacement_law?SSO=1 [Accessed 7th 2020].

[14] Nave R. *Stefan-Boltzmann Law*. Available from: http://hyperphysics.phy-astr.gsu.edu/hbase/thermo/stefan.html [Accessed 17th December 2020].

[15] Kuiper GP. The Empirical Mass-Luminosity Relation. *Astrophysical Journal*. 1938;88(1): 472-507. Available from: doi:10.1086/143999.

[16] Omnicalculator *Luminosity*. Available from: https://www.omnicalculator.com/physics/luminosity [Accessed 16th February 2021].

[17] Nave R. *Stellar Lifetimes*. Available from: http://hyperphysics.phy-astr.gsu.edu/hbase/Astro/startime.html [Accessed 22nd December 2020].

[18] Salpeter E. The Luminosity Function and Stellar Evolution. *Astrophysical Journal*. 1955;121(1): 161-167. Available from: doi:10.1086/145971.

[19] SDSS. *Optical Spectra: Galaxy Properties*. Available from: https://www.sdss.org/dr16/spectro/galaxy/#PortsmouthSED-fitStellarMasses [Accessed 9 March 2021].