Galaxy Distances

Adam Boustani, Guney Coban, Anastasia Marine

King's Certificate

Project Mentor - Lee Stothert

Acknowledgements

Foremost, we would like to give special recognition to Lee Stothert for creating this project and for guiding us throughout our journey. He has helped us with writing the code and understanding the key concepts. His advice has helped shape our article and research.

We would also like to thank Mrs Lawsen who helped us cultivate and expand the necessary skills for this project. Under her guidance, our writing has improved immensely.

Recognition goes to J.Lee, J.Bossom, L.Tattersall, and A.Heeley. Their King's Certificate project on "The Young Tableux" was a great example of how our research should be conducted and we often consulted their article for ideas.

Finally, we would like to thank the King's College London School for creating this opportunity and allowing us to research a topic we are all passionate about, whether it be programming or cosmology.

Abstract

The discovery of redshift in 1929 helped astronomers understand the structure and behaviour of the observable universe. The concept supports ‘Hot Big Bang Theory’ and has been essential in calculating the size and age of the universe. Currently, surveys to find the distance, redshift and magnitude of distant galaxies are very time consuming. They have to be this way to yield accurate results. One such survey is the Sloan Digital Sky Survey, which our group will use when setting out to create an algorithm that could use the redshift of a galaxy to estimate its distance from Earth. This will allow researchers to attempt to carry out rough calculations using the characteristics of a galaxy without having to wait several years to begin their work based on definite values. However, our predictions will still require high levels of accuracy in order for the program to be credible. Therefore, we will test different regression models to see which one will return values that match the results of the current SDSS. Our first approach is to assume that the relationship between redshift and distance is linear, and thus use the linear regression model within Python’s Scikit-Learn package to provide us with initial results. We later discover that at higher redshifts, the program overestimates the distance between the galaxies. Hence, we assume that the relationship between these two variables is not linear but logarithmic. As a result, we use the random forest model instead of the linear regression model as it will be able to pick up on this much more accurately. For our final code, we implement this into the machine learning package and retrieve results which are closer to the existing SDSS data.

Contents

1 Introduction

There are approximately 170 billion galaxies in the observable universe, with this figure rapidly rising towards 200 billion as telescope technology continues to make significant improvements [1]. For centuries, the origin story of these massive systems, with often billions of stars, has been a cause of debate and scientific interest. The most accepted theory currently is the 'Hot Big Bang Theory'. Unlike many other concepts, there is a significant amount of evidence to suggest its occurrence, and it can be seen in effect through studies of redshift, where light waves from galaxies seem to be stretched and hence display longer wavelengths and appear redder in spectrograms. This phenomenon is an example of the Doppler Effect. Surveys of the observable universe have collected data showing the spectrogram for each specific type of galaxy. Comparing these values to the apparent values from Earth proves that a vast majority of galaxies are travelling away from ours, with the most distant ones receding at a greater velocity. This further shows that the space between celestial objects is expanding at an increasing rate as described by the theory.

Redshift can also be used to compare galaxies to one another. This is because it can be used to calculate the distance between galaxies based on the colour that they appear. Historically, these have been measured as part of large sky surveys, of which the Sloan Digital Sky Survey (SDSS) is an example. Whilst these have been calculated by scientists for years, small errors can cause inconsistencies for larger data sets and it may take a long time before patterns begin to emerge. Therefore, by undertaking the project we aim to diminish this problem by developing a model which can take in various inputs and rapidly estimate distances to a very high degree of accuracy. Whilst errors may therefore still arise, they will be minimal and with the data processing power of computers, we will be able to determine patterns much more rapidly. The introduction of computers to these calculations also helps to protect accuracy for the future. As the distance between celestial objects is expanding at an increasing rate, results in these measurements become outdated due to the time taken to collect the data. The use of computers, and their ability to find correlations and carry out large calculations for millions of data points which humans simply cannot do allows us to make estimates for upcoming data and distances, which is even more useful. Importantly, we want to develop a machine learning platform which can be adapted simply to aid anyone seeking to explore beyond our galaxy.

Throughout the project, we will be developing a model capable of accurately approximating the distance to a galaxy based on its colour and brightness. These values will be retrieved from the Sloan Digital Sky Survey and used to train the machine learning algorithm with both linear regression and random forest. This program would give us predictions for the redshift of the galaxies. In addition, we plan on using the survey data to create 3D plots of redshift, brightness and colour which will help us visualise the relationship between the galaxies' different characteristics. Using these plots, we will make an initial estimate on which of the modelling assumptions will suit the relationship between redshift and distance more accurately. Then, we will implement both of the models within a program and analyse the results to come to a conclusion on which one fits our brief better. This will enable us to meet our initial target set when beginning the project, which is to create a very accurate machine learning algorithm.

2 Literature Review

The relationship between redshift and galaxy measurements is becoming increasingly useful in developing our knowledge of the universe. As the project focuses on this correlation, we started our research by reading into relevant topics such as the concept of redshift itself. We found a website which explained the fundamental ideas of cosmological redshift as well as the Doppler effect. In addition, Neta A. Bahcall's article on Hubble's Law and the expanding universe showed us that galaxies are constantly accelerating away from us, making calculating accurate distances between them very challenging. After this, we decided to use Idit Zehavi's article to gain a better understanding of the galaxy survey library that we will be using, including information on its background and its different measurements. Finally, a major part of the project involves creating machine learning code, so we read two programming articles that would develop our understanding of data handling and teach us the basics of the Scikit-Learn model.

To begin our research, we needed to first familiarise ourselves with cosmological redshift. The Swinburne University of Technology is one of the most prominent organisations in Australia and has been a trusted resource by many. The university decided to create a website covering astrophysics and supercomputing [2]. This website focuses on the theory behind redshift, briefly touching on the Doppler shift and explaining its effect on wavelengths observed from distant galaxies. In addition, the source covers the difference between redshift and the Doppler effect. The Doppler effect is a result of an object being in motion whereas redshift is caused by the distance between the Earth and the object increasing. This causes the object to appear to be in motion instead of being stationary. The author also explains through one of the known methods of spectroscopy how astronomers can observe cosmological redshift. This is relevant to our project as spectroscopy is a way of obtaining data on galaxies, from which we will have to determine the distance. They explain how it works and how astronomers can calculate redshift from the obtained results using the provided equation. Although introduced, the author does not explore the reason behind cosmological redshift. The main limitation of this source is the lack of depth and detail that the author provides. As our group's project focus is to develop an algorithm to calculate the distances to galaxies using their redshifts, this source is not as useful as other articles and therefore will not form the basis of our research, but rather provide additional information.

Having researched redshift, we decided to investigate its correlation to the distance between galaxies. Neta A. Bahcall is currently the Eugene Higgins Professor of Astronomy at Princeton University and was elected as a member of the National Academy of Sciences in 1997. In her 2015 expository article [3], Bahcall explores redshift's direct proportionality to distance using galaxy survey libraries such as the Sloan Digital Sky Survey which contains data relating to the brightness and colour of galaxies as well as their redshifts. The current redshift appears to suggest that all galaxies are accelerating away from us and the author investigates the relationship between the velocity at which all galaxies recede with their distance from the Earth. Furthermore, the article explains the use of the Hubble Constant to calculate Hubble Time which is significant in approximating the age of the Universe. Despite this, Bahcall does not include any data visualisations from her investigation of the distribution of galaxies as functions of brightness and colour. These would prove useful when highlighting how observed apparent brightness is inversely proportional to distance. The article is aimed at researchers in similar fields of academia as it is quite difficult to understand and is intended to inform readers of previous findings so that research is not repeated. Since the author has previously conducted research related to observational cosmology and the ideas of dark matter and dark energy, it shows her to be a reliable source. Overall, the article will form the basis of our contextual knowledge, however, it will not prove useful when completing initial data analysis using the SDSS galaxy survey or when using linear regression and random forest to train our model.

The observations of distant galaxies and discovering their redshifts and respective distances are a core part of our project. This article [4] explores galaxy clustering specifically in the SDSS. Since we will also be using data from this galaxy survey library, the article will be useful due to its relevance to our project. Idit Zehavi, the chief writer of the article, is an associate professor in the Physics Department at Case Western Reserve University. Zehavi is an astrophysicist whose research interests include the large-scale structure of the universe, galaxy formation and evolution, structure formation, the galaxy-halo connection and cosmic flows. The article dicusses the process of retrieving the data, using a 2.5m telescope and a drift-scanning mosaic CCD camera. The imaging data were photometrically reduced and calibrated after which they selected objects for spectroscopic follow-ups. The wide-range survey of the distant galaxies produced their redshifts measured with a success rate greater than 99% and typical accuracy of 30km/s showing that the collected data is reliable. As this article uses the same survey and data and part of their area of research overlapped with our project, it will be partially useful. However, the author mostly focuses on the luminosity and colour dependence of galaxy clustering which means that its relevance to our project is limited.

Our machine learning algorithm will be created in Python due to the vast array of libraries available in this programming language. They provide several methods for the handling of data, each with their respective benefits and shortcomings. One common module is detailed in Wes McKinney's 2011 article: 'pandas': A Foundational Python Library for Data Analysis and Statistics [5] where he outlines the various approaches to organizing data. As the creator of the open-source package, one may argue the article may have been written to favour the library over other data analysis platforms. However, we had already agreed on the use of 'pandas' due to its ease of use when working with large data sets, as this is crucial for a project where the program relies on a set of statistics for a range of galaxies in the observable universe. Therefore, it was agreed that McKinney's knowledge and description on how to effectively use his own system would surpass any other attempts at tutorials. The article helped us to develop our understandings of Python and give an insight into a new data handling package. Hence, the different sections of the article are laid out in an instructive manner, to ensure that the entire target audience can learn from the concepts being described, compared to an article using more jargon, which would be written for experts in the field. This ensured that our understanding of the article was much better than the previous and therefore allowed us to learn significantly more. Within the paper, McKinney introduces a new concept through the 'DataFrame' which is the foundation of 'pandas'. This class builds on previous data handling packages such as NumPY and R. However, these libraries are not optimised for data analysis, hence why 'pandas' was developed as a bespoke package for handling various sets of data. The article then moves on to highlight the different ways in which it can organise data sets into its unique format. We found this very useful, as it closely ties into the branch of data analysis our group is required to be able to work with to have a practical method of organising the SDSS data within our Python program.

Familiarising ourselves with the new package allowed us to access different measures of the SDSS data within our program. Using these, we could link individual characterisitcs of galaxies to their relative three dimensional coordinates. In order to utilise this new data within a machine learning algorithm, we explored a package named Scikit-Learn. In their newer journal entry "API design for machine learning software: experiences from the Scikit-Learn project" [6], the writers of the Scikit-Learn project explore the structure of the library in depth. Initially, we believed the article's target audience to be developers who may also be planning to create their own packages, however as stated in the abstract, the article aims to analyse obstacles faced by users and developers of the library. The writers outline the API as consisting of three complementary interfaces: an estimator interface for building and fitting models, a predictor interface for making predictions and a transformer interface for converting data. When beginning a new machine learning algorithm, the article explains how the different estimators are initialised and set up for a new incoming set of data, simply requiring an import and a statement defining the user specific parameters they would like to use as shown [6]:

```
from sklearn.linearmodel import LogisticRegression
clf = LogisticRegression(penalty='11')
clf.fit(x_train, y_train)
```

This initialises a $\ell 1$ regularisation with only the penalty parameter being modified. Depending on requirements more parameters may be changed, and the estimator may also use another form, such as RandomForestRegressor() instead of LogisticRegression(). Therefore, the choice to use this library would significantly reduce the time we have to spend implementing different models as they are already built in to the package. Furthermore, as shown by the code above, only a single line of code is required to change the regression model, which will increase our productivity when comparing results. To ensure the results are not biased, we could use the same program, and only modify a single line when changing between models.

Our literature review has provided us with a strong foundation for the project. By researching cosmological redshift and Hubble's Law, we have also familiarised ourselves with the way in which our project relates to current ongoing research. The article exploring galaxy clustering using SDSS data gave us an example of how the data is used in studies and other projects and solidifies our understanding of the nature of the data. This combined with the programming articles will help structure our project and understand the theory behind it.

3 Methodology and Development

After working with the scientific Python Libraries, a strong understanding of them has been established, which has enabled us to create informative visualisations of the SDSS data. In order to make the most effective use of these, it was necessary to plan a schedule which factored in time to analyse the data for any patterns relating to distance. Throughout this section, the reasoning behind our timings and our analysis of graphical representations will be explored.

3.1 Planning and Time Management

3.1.1 Organisation

At the beginning of our project, we outlined our deadlines and group specific objectives within a Gantt Chart, as shown in Figure 1. Its simple nature gave us a clear way to visualise the seven months we had ahead and this enabled us to complete tasks on time and efficiently. The chart incorporates dates into it which provides a useful time frame that allows the group to see the structure of our project more easily.

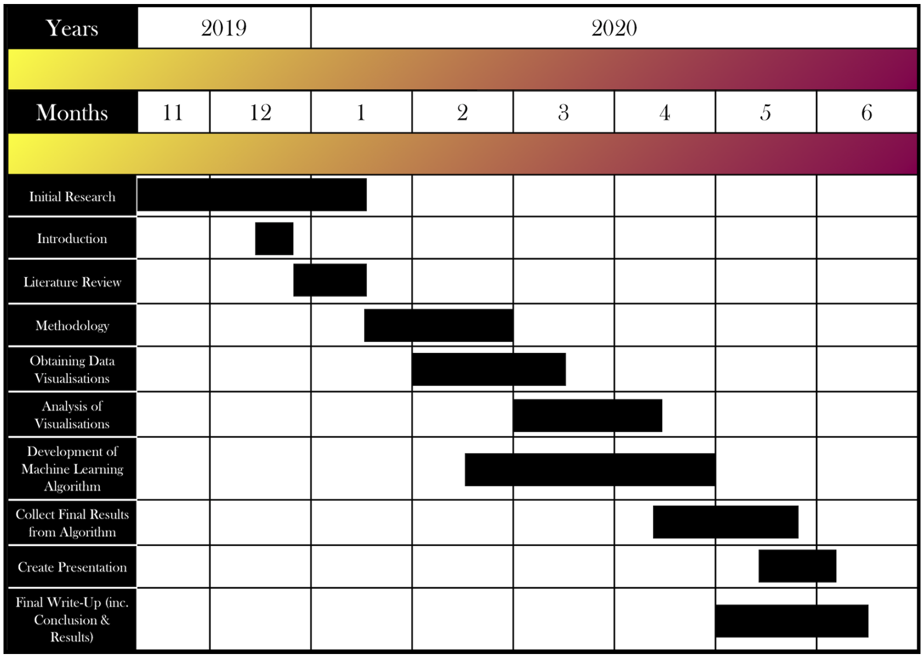

Figure 1: Gantt chart

Our Gantt chart was based upon the set deadlines for the project. A period of one and a half months spanning from February to March was dedicated towards generating and analysing data visualisations. In order to obtain a solution to the project brief, a machine learning algorithm has to be developed. Since this is the main focus of our project, two and a half months were assigned to this task.

3.1.2 Delegation

When beginning the project, our group planned to meet up every Thursday at the end of the school day. In these group sessions we worked on different areas of the project and planned out our next steps. The group also decided on who was going to be in charge of upcoming sections. However, this did not work as well as intended, as some areas required us to work individually, as well as members having other events to attend.

In the group sessions we had, we decided that Guney would be the main programmer and work on creating the machine learning algorithm and plots, Anastasia would be the main researcher as the one with the most experience in studying cosmology, and Adam would be in charge of organising our research, writing up our findings, and ensuring that we are following our planned schedule for the project. We believed it would be most effective for all the members of our group to take part in editing our article.

In order to keep track of our progress, we maintained communication through a group chat. Any work that required collaboration was completed during our Tuesday afternoon sessions.

3.2 Main Methods

There are two main types of research in our project:

- Primary research: We constructed our own algorithms to produce graphs comparing galaxy characteristics [7] using Python and analysed the results generated by these. Our knowledge of Python further enabled us to create a basic model which uses a galaxy's redshift to estimate the x-location of that galaxy.

- Secondary research: We used a large a variety of sources to develop our understanding of the general topic surrounding cosmology and galaxy distances.

3.2.1 Computing

Python 3.8 with Anaconda was our choice of programming language. Other options such as C++ are significantly more advanced and would have consumed more of our time when learning the basics of it. Considering that we all had previous experience in Python, we felt no need to learn a more difficult language for our project which required a relatively small program. Furthermore, Python contained several benefits over other programming languages in regards to features. Its pandas library allowed easy modification and access of the SDSS data, which would have remained otherwise unreadable. It also has several extensive machine learning packages, each with complete documentation. Scikit-Learn was the Python package we chose, as it offers a simpler method for creating basic programs, which is sufficient for the task we have.

3.3 Key Concepts

3.3.1 Variables

Our project has required us to use specific variables for our code. These are the redshift values, magnitude values and the distances to the galaxies. The redshift and distance values that we have used for the data plots below are from the SDSS large data set; this is a large collection of data on individual galaxies that have all been measured. These are the values that we will be using to train our machine learning algorithm.

Magnitude and colour are both other variables that we will have to consider. Magnitude is a measure of brightness where the higher the magnitude, the lower the brightness. This means that the greater the magnitude, the greater the distance. Our literature review showed us that the redshift of the celestial object increases as the distance between the object and the Earth increases. Since magnitude also increases with distance, redshift also increases with magnitude.

3.3.2 Data Plots

As mentioned earlier, redshift increases with distance. Figures 2, 3 and 4 below show this relationship in different magnitudes including ultraviolet-infrared (u-i) and green-red (g-r). The plot using the u-i magnitude shows a greater spread across a range of redshifts while the plot using g-r magnitude shows a more concentrated spread.

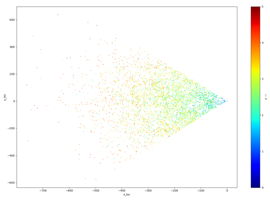

Figure 2: u-i

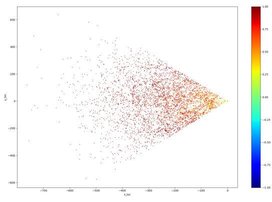

Figure 3: g-r

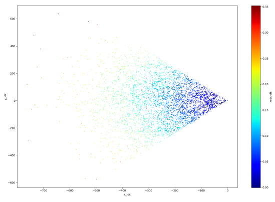

Figure 4: Redshift

In Figures 2 and 3, the data plotted illustrates how the magnitude of the objects increases proportionally with the distance. The x-axis shows the magnitude scale while the y-axis shows the distance. The plots' colour change allows the plots to illustrate how redshift increases with magnitude and distance, thus proving our theory.

Figure 4 depicts redshift against distance. There appears to be a pattern, with a vast number of close galaxies having small redshift values in comparison to the furthest ones which are often above 0.2. The colour in this plot is a visual aid, although the x-axis already represents the redshift. The colour change within the plots gives a visual representation of how the redshift scales with the galaxy's colour and distance.

Redshift values are a measure used to compare the redshift of various objects. Values less than zero are assosciated with objects approaching the observer whereas values greater than zero are assosciated with objects receding from the observer. Hence, we can assume that both redshift and magnitude are dependent on the distance of the object, therefore it is intuitive to infer that these two characteristics have a similar relationship with each other. To test this theory, we created new plots of g-r against redshift and u-i against redshift. These are shown in figures 5 and 6 respectively.

Figures 5 and 6 show directly the relationship between magnitude and redshift with the magnitude on the x-axis and the redshift on the y-axis. The overall trend in both graphs show that as the magnitude increases, so does the redshift.This supports the theory that both redshift and magnitude depend on the distance as magnitude and redshift are also dependent on each other. We can also relate this to the brightness of the galaxy. If the magnitude of a galaxy is high, its brightness is relatively low. As its magnitude decreases, its brightness increases. This means that the lower the brightness, the greater the redshift which further supports our theory as brightness decreases with an increase in distance.

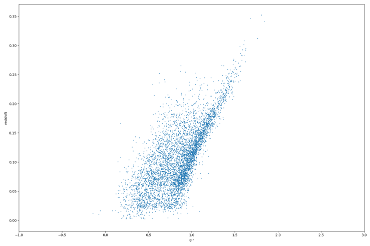

Figure 5: g-r

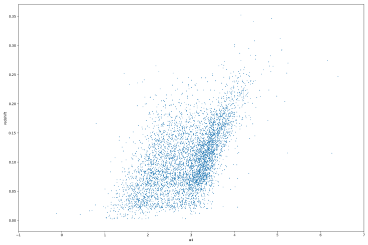

Figure 6: u-i

3.4 Code

After considerations based on the various graphs, we decided that the most clearly visible relationship with distance was redshift. Hence, for the initial development of the machine learning algorithm, we chose to teach the program to interpret this correlation as the learning data set. As seen in Figure 7, we opted for the use of a linear regression model [8] to interpret the correlation as this would provide us with some simple results. Furthermore, the use of a linear regression allows extrapolation of results and helps with anomaly detection as outliers are ignored. This allowed us to have data to compare with true values of galaxies and therefore easily see which aspects of the code to improve on.

```python
import numpy as np
from sklearn.linear_model import LinearRegression

x = np.array(df["redshift"]).reshape(-1, 1)
X = np.array(df["x_loc"]).reshape(-1, 1)

clf = LinearRegression()
clf.fit(x, X)
print(clf.predict([[0.1]]))
```

Figure 7: Code

3.5 Development Strategy

For the project to progress, we needed to produce initial results that we could develop further at a later stage. The code shown in Figure 7 used the x-locations of the galaxies. It then compared the distances between the galaxies and Earth to that of the large data set. These comparisons enable the code to estimate the value of redshift for x-location given. However, the results produced were not precise estimates which are most likely due to the x-location value not being an accurate measure of the distance between Earth and the galaxy. Therefore, we used the calculation

$$\sqrt{((X - Location)^2 + (Y - Location)^2 + (Z - Location)^2)}$$

to calculate the actual displacement between Earth and the respective galaxy. Additionally, we needed to adjust the code to the program to process multiple inputs. The galaxy's distance does not only affect its redshift, but also its magnitude and colour. Figures 2 and 3 illustrated the relationship between redshift and distance and so we utilised the same magnitudes of u-i and g-r and included them as part of the training data for the algorithm. Finally, we must experiment with the vast selection of models provided within the scikit package. Through this testing, we will be able to choose one that best represents the true relationship between the various characteristics and distance. The implementation of these should help us to create a final, successful program.

3.6 Summary

We have familiarised ourselves with the relationship between the redshifts of various colours and distance. Furthermore, we produced data visualisations using Python and interpreted them to help us picture the correlation between these variables. We organised our group by using a Gantt chart, and assigned each other roles based on our strengths and interests.

4 Realisation and Results

4.1 Introduction

Having done substantial research and a deep enough understanding of our next steps, we decided to create plots. By experimenting with different magnitudes we could decide which magnitude filters would be most appropriate to use in our code. We obtained results from our program for various redshift values. We analysed the results and plots which gave us a general direction in which we should go for our end program.

4.2 Data Visualisations

By creating plots, we believed we would gain a good understanding of the theory behind the project. We began by analysing some of the plots of the SDSS data provided by out mentor, Lee Stothert.

In Figure 8, the data is presented in the form of magnitude against redshift. The magnitude is given in the form of colours, which is necessary for our project and for us to understand as our machine learning algorithm must be able to process different types of data, including magnitude and colour. The SDSS measures magnitudes in five colours; this is done by taking the images using five different filters: green (g), red (r), two infrared filters (i and z) and an ultraviolet filter (u). The colour is indicated by subtracting magnitudes of the colours. In Figure 8, the magnitude is given as g-r, where a star with a higher g-r magnitude is more red than a star with a lower g-r magnitude.

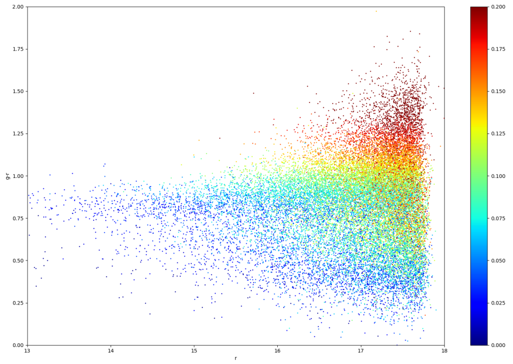

Figure 8: g-r / redshift

In Figure 9, the given magnitude on the axis is u-i which is between the infrared and ultraviolet filters. Generally two stars of the same colour will have the same magnitude in the colour ranges, so two red stars of the same colour will have the same magnitude relative to the filters [9].

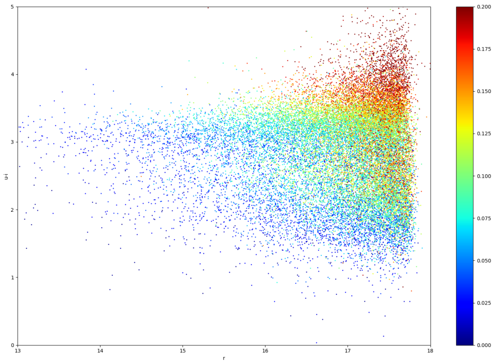

Figure 9: u-i / redshift

4.2.1 Three Dimensions

The first issue that had to be addressed was that our initial attempt at the program only produced an estimate for the x distance of the galaxy. To simplify this process, instead of having the program estimate distances in all three dimensions and then calculate a single distance, we decided to create a new column of data which simplified each galaxy's distance to this new measure and then used this set to train the program. Figure 10 and Table 1 illustrate the code used for this process and its results respectively.

```
import numpy as np
from sklearn.linear_model import LinearRegression

df['Distance3D'] = np.sqrt(np.square(df[['x_loc', 'y_loc',
    'z_loc']]).sum(axis=1))

userRedshift = float(input("Enter the redshift value for which you
    would like to estimate distance: "))

x = np.array(df["redshift"]).reshape(-1, 1)
X = np.array(df["Distance3D"]).reshape(-1, 1)

clf = LinearRegression()
clf.fit(x, X)
print(clf.predict([[userRedshift]]))
```

Figure 10: Linear Regression Algorithm

Redshift	0.1	0.15	0.2	0.25	0.3
Predicted Distance (3.s.f)	291	432	573	714	855

Table 1: Initial Results

4.3 Analysis and Evaluation

The data visualisations have helped us gain a clear understanding of the relationship between magnitudes of galaxies and their respective distances. Furthermore, it has given us an idea of which parameters to use in the machine learning algorithm, as these were influenced the most by the distance to the galaxy.

As we have gained sufficient information from the graphs, we have chosen to change our focus to developing the machine learning algorithm as this will allow us to find a solution to the project brief.

We programmed our machine learning algorithm to output a set of results for distance based on certain values of redshift. For the smaller values of redshift between 0 and 0.2, the distance outputs were more accurate. A reason for this is because a large majority of the galaxies in the data survey library have a redshift which lies in this range, therefore the relationship was more accurately inferred by the program. When redshift above 0.2 were tested, the quantities for distance were less accurate than expected. This was due to there being insufficient data in this range, which meant that a linear regression model was not advanced enough to pick up the trends between redshift and distance as they are non-linear. Hence, this has inspired us to attempt to find a method of doing this more effectively, using a different model.

4.4 Summary

The data plots and our initial data from the machine learning algorithm have allowed us to identify which values of redshift will produce more accurate results. They have given us an insight into the general trend which the distance of a galaxy follows in regards to magnitude and redshift. By examining the data plots, we have also been able to approximate the results we will get from our code. This is very useful since if the code we produce in the end produces incorrect data, we will be able to identify this and make changes accordingly. Also, an initial evaluation of the results from the linear regression model provided us with a starting point around which to base improvements to our code.

5 Final Results

5.1 Introduction

In this section we are going to introduce the new model, the random forest model, and describe its differences from the model in the code used to obtain our first set of results. We will then compare results against a constant measure and come to a conclusion over which model is more suitable for our brief.

5.2 Baseline

When incorporating the new model, we had to use baseline values for each redshift we were estimating, so we could compare and evaluate our results relative to the linear model. For this we decided to use the mean distance at each value as given in the SDSS data. This data is shown in Table 2.

Redshift	0.1	0.15	0.2	0.25	0.3
Baseline Distance (3.s.f)	293	433	571	704	836

Table 2: Baseline Values

5.3 Linear Regression Analysis

As Figure 11 and Table 3 demonstrate, the linear model initially produces predictions close to the baseline values for each redshift. However, as the redshift and distance begin to stop increasing at a near linear rate, the predictions become increasingly incorrect, with there being a 2.27% error at the 0.3 redshift value. This is too large of a margin to be accepted in a project which requires as much accuracy and precision as possible.

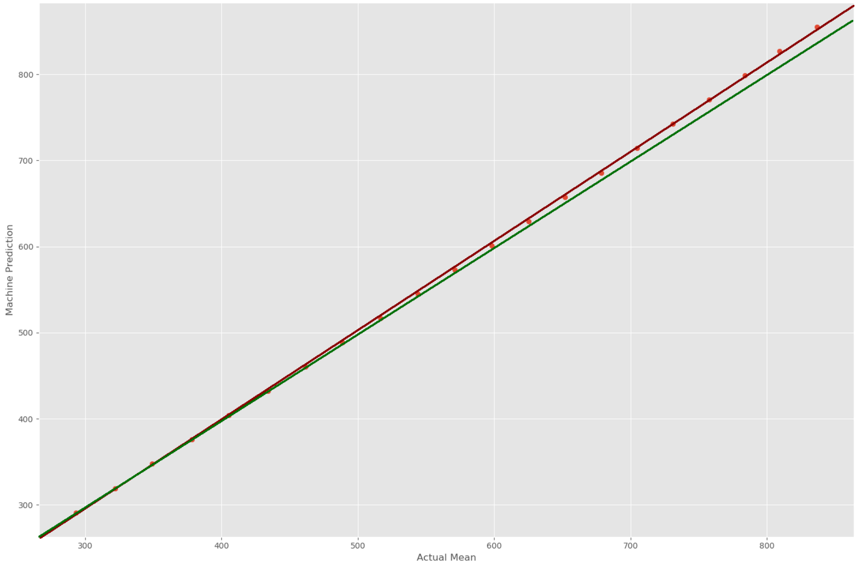

Figure 11: Predictions vs Actual Distances

Redshift	0.1	0.15	0.2	0.25	0.3
Predicted Distance (3.s.f)	291	432	573	714	855
Percentage Error vs Baseline/%	-0.683	-0.231	0.350	1.42	2.27

Table 3: Linear Regression Model Results

5.4 Random Forest Model

5.4.1 Difference in Methods

After analysing the results from our first algorithm test, we decided to implement a random forest regression into the code. Compared to assuming a linear relationship, the Scikit-Learn algorithm for random forest creates a 'forest' with a number of decision trees. Each tree votes for a class and the forest chooses the tree with the most votes. When performing a regression, the forest takes the average of the outputs by different trees. More decision trees in the forest would produce more accurate results. An advantage of random forest is that it is able to handle large data sets such as the Sloan Digital Sky Survey and it will also maintain the accuracy for missing data. [10]

5.4.2 Training Model

Implementing random forest into the code was very similar to that of linear regression [11], however, this time, there are adjustable parameters. We chose to use 100 trees in our random forest, in order to obtain a relatively high level of accuracy as shown in figure 12.

```
import numpy as np
from sklearn.ensemble import RandomForestRegressor

df['Distance3D'] = np.sqrt(np.square(df[['x_loc', 'y_loc',
    'z_loc']]).sum(axis=1))

userRedshift = float(input("Enter the redshift value for which you
    would like to estimate distance: "))

x = np.array(df["redshift"]).reshape(-1, 1)
X = np.array(df["Distance3D"]).reshape(-1, 1)

rf = RandomForestRegressor(n_classifiers=100)
rf.fit(x, X.ravel())
print(clf.predict([[userRedshift]]))
```

Figure 12: Random Forest Algorithm

After training our algorithm to analyse the data set using our new regression model, we modified the $max\_depth()$ parameter and set it to 3 and only used a single tree, so that we could have a visualisation of some of the trees and decision making processes that the program took. These can be seen in Figure 13.

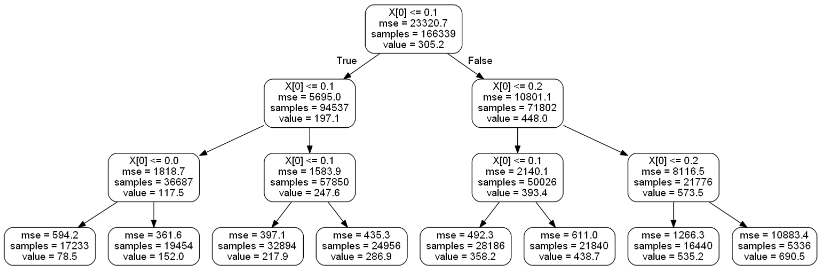

Figure 13: Forest used for $RandomForestRegression(n\_estimators = 1, max\_depth = 3)$

5.4.3 Random Forest Analysis

Similar to the linear model, the random forest also provides accurate predictions for low redshifts. However, it is significantly more accurate at higher redshifts. Whilst the other model was not capable of adapting to the logarithmic relationship, the random forest's algorithm quite easily matches the shape. As seen by the graph in Figure 12, its predictions are almost identical to the baseline values we identified. Furthermore, as seen in Table 4, all error variations are within 0.3% which shows that it is approximately 8x more accurate than the Linear model for redshift above 0.25. Relatively, this is a very small error margin, and due to our rounding of values, it may be even narrower than our calculated percentages.

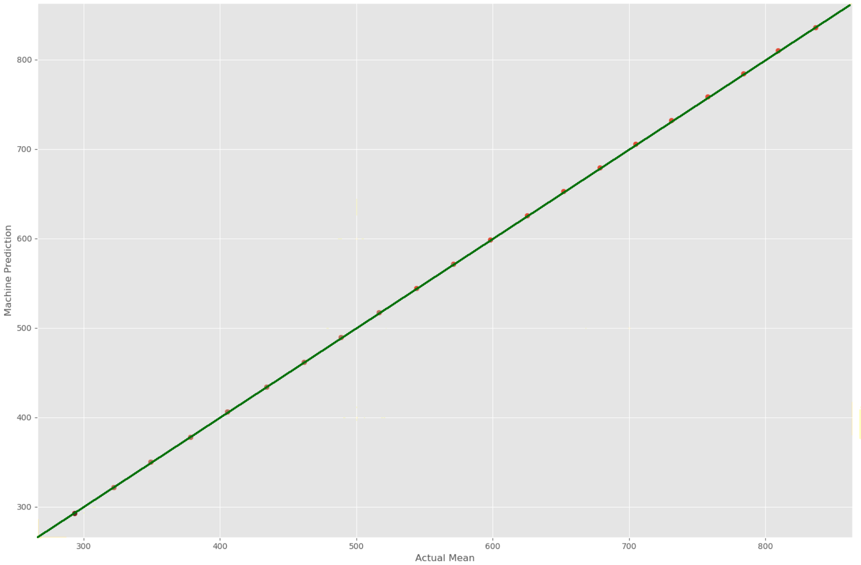

Figure 14: Predictions vs Actual Distances

Redshift	0.1	0.15	0.2	0.25	0.3
Predicted Distance (3.s.f)	293	434	572	706	836
Percentage Error vs Baseline /%	0.00	0.231	0.175	0.284	0.00

Table 4: Linear Regression Model Results

5.5 Summary

Our final results and evaluations of these models solidified our initial assumption that the random forest model would fit the SDSS data more accurately. Relative to the baseline values, the model provided us with data points within a 0.3% margin. Therefore, our initial aim to have a program which was credible for estimating distances has been achieved. However, one problem with this is that outliers in galaxy distances are ignored, and the algorithm can only predict average distances to a high enough level of accuracy.

6 Conclusion

Our aspiration for this project was to train a machine learning algorithm to approximate distances to galaxies using brightness and colour as inputs. The Sloan Digital Sky Survey provided us with data relating to five colours, and known redshifts, and this data gave us the means by which we would create our algorithm. Firstly, we conducted research to gain a better understanding of how redshift relates to galaxy distances and learnt that as distance increases, so does the redshift. This relationship is logarithmic. Furthermore, we researched how redshift is actually measured. We found that the light from the specific galaxy is obtained in the form of a spectrum. This spectrum is then compared to the known spectrum of the known materials within the galaxy such as hydrogen. The wavelengths shown on the obtained spectrum and the already known spectrum are used in the equation below.

$$z = \frac{\lambda observed - \lambda known}{\lambda known}$$

After our background research was complete, we began generating 3D plots of the survey data using code provided to us by our mentor, Lee Stothert. These data visualisations illustrated the relationship between magnitude and distance and analysing them helped us to guess the results that would be predicted by our machine learning algorithm. Our initial program took in various redshifts as inputs and output corresponding values for the predicted x-location of a galaxy. This was useful because it gave us a base code that we could expand from. The errors that we could identify highlighted the areas that we needed to work on and gave us a clearer idea of what our next steps should be.

Our initial results used a linear regression and proved to be quite inaccurate for large redshifts. Therefore, we used random forest for our final model. This proved to be more advantageous as it does not assume a linear relationship, and can adapt to the logarithmic nature of the redshift/distance relationship unlike a linear model. Furthermore, the visualisations arising from this model provided us with knowledge on the system behind the code, helping us to understand where the values came from.

There were several limitations throughout our project that prevented us from obtaining completely accurate values. No matter how well designed our model could be, the results produced would still be approximations and this limited the usefulness of our findings. Furthermore, COVID-19 has had a major impact on our project as it has prevented our group from being able to discuss our findings as a collective. Another difficulty our group has encountered as a result of this pandemic has been sharing code and tracking our progress throughout the project.

We have met our brief for the project and now have a machine learning program that can predict to a good degree of accuracy the distance of a galaxy. For further research, we could create a program that would calculate redshift, if given distance as an input. This would require more research in the form of analysing data plots and using different models to experiment. The code would give the user freedom in which input they would like to give, and which output they would like to receive as the result. The inputs and outputs can be magnitude, colour, redshift or distance. Another idea that could build on this would be to create a program that could take an image of a galaxy and analyse the colour of that image to approximate the redshift and distance of that galaxy. This would require a different set of data for the model to be trained with. The model will need a variety of images that it can use alongside with the corresponding values of redshift, magnitude, colour and distance.

References

[1] Howell E. How Many Galaxies Are There?;. Available from: https://www.space.com/25303-how-many-galaxies-are-in-the-universe.html. (Accessed: 12.01.2020).

[2] Swinburne University of Technology. Cosmological Redshift;. Available from: http://astronomy.swin.edu.au/cosmos/C/Cosmological+Redshift. (Accessed: 07.01.2020).

[3] Bahcall NA. Hubble's Law and the expanding universe. Proceedings of the National Academy of Sciences. 2015;112(11):3173–3175.

[4] Zehavi I, et al. GALAXY CLUSTERING IN THE COMPLETED SDSS REDSHIFT SURVEY: THE DEPENDENCE ON COLOR AND LUMINOSITY. The Astrophysical Journal. 2011;736(1).

[5] McKinney W. pandas: A Foundational Python Library for Data Analysis and Statistics. 2011;(Accessed 11.01.2020). Available from: https://www.dlr.de/sc/portaldata/15/resources/dokumente/pyhpc2011/submissions/pyhpc2011_submission_9.pdf.

[6] Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, et al. API design for machine learning software: experiences from the scikit-learn project. ArXiv. 2013;abs/1309.0238.

[7] Stothert L. scatterPlot;. (Accessed: 02.02.20).

[8] scikit-learn developers. Linear Models;. (Accessed: 15.03.2020). Available from: https://scikit-learn.org/stable/modules/linear_model.html.

[9] The Sloan Digital Sky Survey. The Definition of Color in Astronomy;. Available from: https://skyserver.sdss.org/dr1/en/proj/advanced/color/definition.asp. (Accessed: 08.03.2020).

[10] Startups A. Random Forest - Fun and Easy Machine Learning; 2017. (Accessed: 26.06.20). Available from: https://www.youtube.com/watch?v=D_2LkhMJcfY.

[11] Koehrsen W. Random Forest in Python;. (Accessed: 21.06.20). Available from: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0.