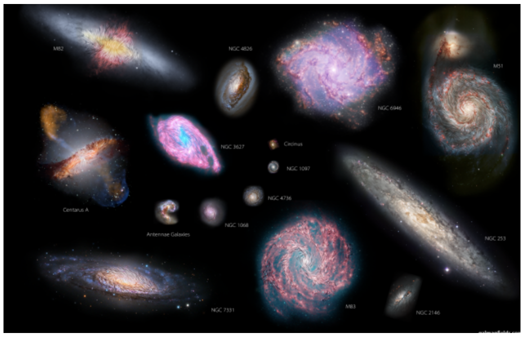

SDS 326E Final Project Report
Spring 2025

Group Members:
Óscar A. Chávez Ortiz
Ethan Abraham
Kyla Ko
Vincent Cheng

April 30, 2025

Contents

1 Project Description and Background

The period known as the Epoch of Reionization (EoR) is an active area of research in astronomy. This period marks the last major phase change the universe has undergone as the universe went from having predominantly neutral Hydrogen to completely ionized Hydrogen. We have evidence of this event occurring due to Quasar (an active black hole) light marking regions in space where the universe was once neutral to ionized. We know this occurred 1 Gyr years after the big bang but getting a complete timeline of this period is challenging due to several factors:

1. Limited amount of tracers that can probe the full timeline of reionizations
2. The high neutral fraction at earlier times makes studying the light emitting from these galaxies very difficult as the light we need to trace reionization gets absorbed and re-emitted in a random direction

We aim to overcome these challenges by using a readily available tracer called Lyman-alpha (Lyα), which is the n = 2 to n = 1 transition in Hydrogen. This emission line is sensitive to neutral Hydrogen and is produced plentifully in star-forming galaxies, which galaxies during the EoR are. The main drawback to using Lyα is that in order for one to use it as a tracer of reionization, one needs to know how much Lyα is being emitted by a galaxy. This is difficult due to point number 2 mentioned above. Since Lyα is sensitive to neutral hydrogen if there is any neutral Hydrogen in the immediate vicinity of a galaxy the Lyα emission will get scattered out of our line of sight and drastically reduce the observability of it. Our work is going to aim to overcome this by studying galaxies much closer to us and in a period where the universe is completely ionized. The reason for this is that there is very little neutral hydrogen across the line of sight from a galaxy to us and so all the Lyα emission that is emitted from a galaxy can traverse unimpeded to our detectors.

For this final project we have a sample of over 11,000 galaxies, each of those galaxies was run through a Bayesian code to ascertain the physical galaxy properties. We also have the strength of the observable which is the Lyα emission strength, as well as another quantity called the equivalent width. The equivalent width normalizes the strength of the Lyα emission by the continuum near the emission line and can be a proxy of the strength of the emission line relative to the continuum level of the galaxy spectrum.

2 Data Set Questions

The main science question we are after is can we generate a model to accurately predict the Lyα predictor (Lyα EW$_r$). Once we have a model that is able to predict the Lyα predictor, we want to see if we are able to interpret the model and see which subset of features is most impactful to accurately predict the rest-frame Lyα equivalent width. Once we know the features that impact the observable, we can begin to understand why the model is able to predict the observable so well.

This interpretability on the features has very important implications to help us understand the physical processes that lead to the observable. Once we have a good physical understanding going on we can apply this model to constrain the neutral Hydrogen fraction at high redshifts since we expect the neutral Hydrogen of the universe to increase as we go further back in time. However, in order for us to compute the neutral Hydrogen we need to have a good way of modeling the expected emerged Lyman-alpha emission from galaxies at those high redshift and distances. In the following sections we outline the data and then outline some of the key science questions we set out to answer. Thus, the two questions we seek to answer in this paper are as follows:

1. **Can machine learning be used to accurately predict Lyman-alpha emission from a galaxy, using its physical properties?**
2. **Which galaxy properties have the greatest impact on Lyman-alpha emission predictions?**

2.1 Data Description

The data was processed by a custom pipeline made by Oscar Chavez Ortiz. Table 1 defines all variables.

Table 1: Definitions of all variables

Variable	Description
Burst	This is a burstiness metric for galaxies; it indicates whether the galaxy is currently undergoing a burst of *star formation*.
Age [Gyr]	The age of the galaxy, in Gyr.
Mass Formed	Total mass formed throughout the history of the galaxy, in units of $\log_{10}(M)$ (with M in $M_\odot$).
Metallicity	The metal content of the galaxy relative to solar, i.e. $Z/Z_\odot$.
τ [Gyr]	The exponential delay time in a delayed-τ star-formation history; the characteristic time when star formation begins to decline.
A_V	Dust attenuation in the V-band, in magnitudes.
$\log_{10}(U)$	Ionization parameter, measuring the intensity (“harshness”) of the radiation field.
Stellar Mass	Current stellar mass of the galaxy (in $M_\odot$).
Formed Mass	Total stellar mass formed over the galaxy’s history (in $M_\odot$).
SFR [$M_\odot$ yr^{-1}]	Star-formation rate; how actively the galaxy is forming stars.
sSFR [yr^{-1}]	Specific SFR, $\log_{10}(\mathrm{SFR}/M_*)$, i.e. SFR per unit stellar mass.
Mass-Weighted Age [Gyr]	Stellar population age weighted by mass.

Continued on next page

Variable	Description
Mass-Weighted Metallicity	Stellar population metallicity weighted by mass (same units as "Metallicity").
Redshift	Cosmological redshift of the galaxy.
χ^2_{phot}	Chi-squared value from photometric fitting.
SN	Signal-to-noise ratio of the detected emission line (unitless).
EW$_r$ [Å]	Rest-frame equivalent width of the emission line, computed as $\mathrm{EW}/(1+z)$ where z is the Lyα redshift.

2.2 Data Cleaning

Before we can run our non-linear ML models we need to make sure that the input data is of good quality to make for accurate predictions. This involves removing obvious outliers, extreme data and transforming the data to remove any skews in the feature distribution that could bias the data. To generate the data we used to train and test our models we imposed physical and statistical cuts to remove obviously bad data. These cuts included making sure the signal-to-noise ratio (SNR) of the Lyα emission line is of good significance (SNR $>$ 5.3). We also made sure that the galaxy properties can be reliably trusted and this required us to use a photometric cut of $\chi^2_{phot} < 100$. The last cut we made was remove any unphysical Lyα equivalent width and for this cut we use Lyα EW < 500. In Figure 1 we show the feature distribution of all our input features. We can see from the plot that there is a lot of skewness in some of the feature distributions and as such we need to do an additional pre-processing step to make sure that the data is well suited for the non-linear ML models. As such, we used the sklearn package `StandardScalar` to standardize and rescale the data so that no one feature dominates the training in any of our ML algorithms.

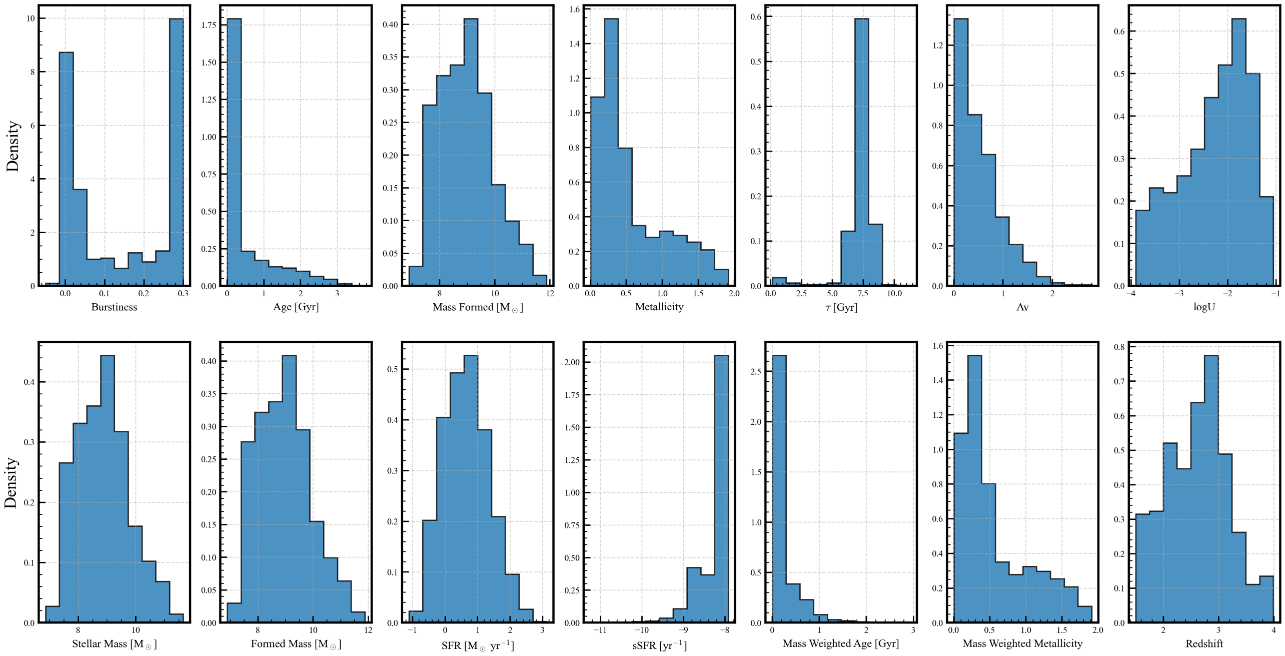

Figure 1: This plot shows the distribution of our features after we applied all the selection cuts to filter out the data. We can see that some features are normally distributed wheras other features show strong skewness. To remove the impact of the skewness in the input features we applied the sklearn function `StandardScalar` to rescale and tranform the data.

3 Question 1: Model Selection

Can machine learning be used to accurately predict Lyman-alpha emission from a galaxy, using its physical properties?

3.1 Model 1: Neural Network (Oscar)

My part of the project was looking into how accurate a neural network (NN) model can be to predict Lyα emission. For the purpose of our project I used the Python package `PyTorch` to generate the machine learning model. For each fully connected layer I used the `Linear` function in `PyTorch` and the used a dropout layer between each fully connected layer. I used a dropout percentage of 20% to prevent overfitting of the data as we increased the number of layers.

Because there are a lot of tunable parameters and features in NN models we fixed a couple of them and changed only a few parameters. For the NN, we kept the activation function fixed to the rectified linear unit (reLU), we kept the optimizer to be the Adam optimizer, and the criterion we used to assess the loss of the NN was the MSE-loss function. For the tunable parameters we changed up the number of layers, 2, 3, and 4, but due to time constraints I did not have time to check more layers. We also added into some run the inclusion of K-fold validation to the training set to assess performance of the NN. For the k-fold validation we used another sklearn function called `Kfold.split` which returns back a set of training and testing indexes based off of the K-fold the user specified.

In Figure 2 we see the loss as a function of epoch for a 4 layer model with and without K-fold validation and we can see that the model with K-fold validation has a lower loss through the process. As such for all the subsequent models 2 and 3 layers we kept the K-fold validation for all of them to increase the performance.

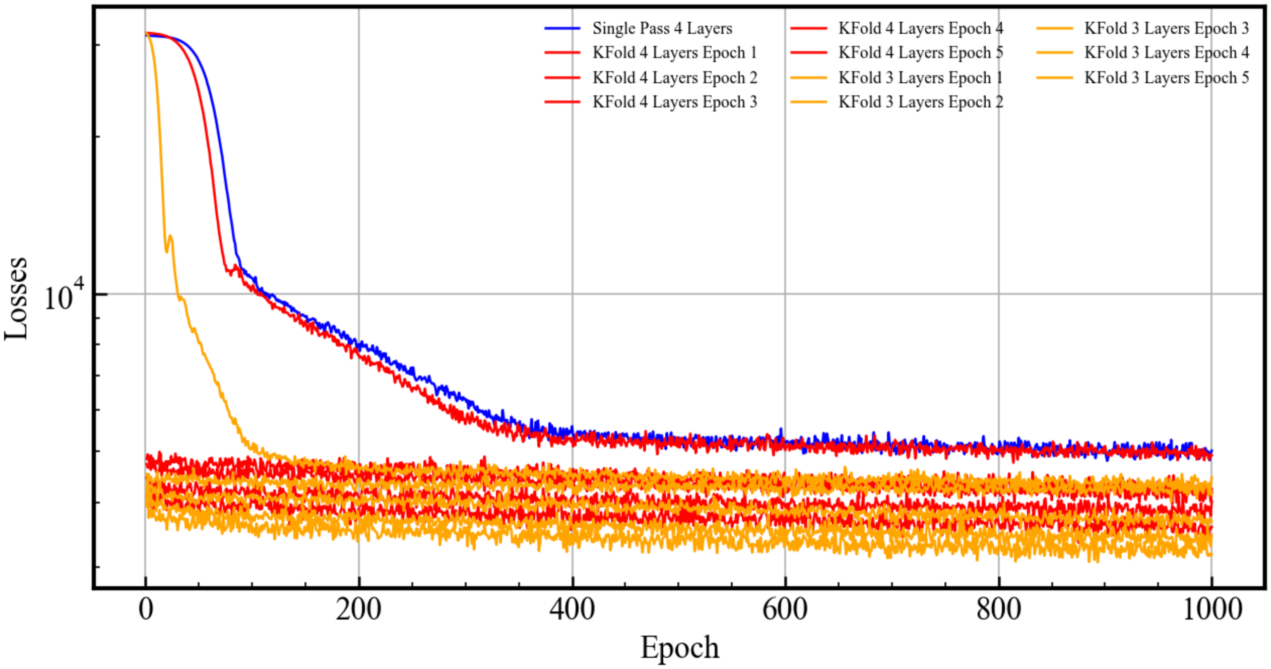

Figure 2: Plot that shows the loss vs Epoch of our NN with 4 layers with and without K-Fold validation. we can see that the NN with only a single pass has a higher loss than the NN that incorporates K-Fold validation. As such for the NN models with 2 and 3 layers we only use the K-fold validation technique as that will drastically reduce the loss in our training.

We can see in Table 2 the breakdown of the RMSE, MAE, R^2 and how they compare between the different models. We can see that the best model is the NN with 4 layers and using K-fold validation for the K we used a value of K = 5 for the cross validation.

We then use the best fitting model to make a prediction on the testing set and we can see in Figure 3 that the model does a good job at predicting the Lyα rest frame equivalent width but fails to reproduce values at high Lyα EW. We think this is due to a sparsity of data at the high EW regime making the NN not being able to predict accurately at those values. We think one thing that could help is getting more data that encompasses the high Lyα EW for better accuracy at this regime. One

Table 2: A metrics table outlining the comparison between the NN models and how they performed based off of MSE, MAE, and R^2 values. The best model is the one with 4 layers and using K-fold validation. Further test is needed to see if adding in 5 layers offers significant improvement as well as the number of nodes in each layer.

Model	**MAE**	**MSE**	R^2
NN 4 Layers (K-Fold)	46.0348	4572.53	0.29142
NN 4 Layers (No K-Fold)	51.77060	5371.919	0.1659
NN 2 Layers (K-Fold)	50.5274	5393.7635	0.0956
NN 3 Layers (K-Fold)	53.768	6154.985	0.1153

thing to also look into is how changing the fixed values we assumed in our NN architecture affects the predictive accuracy. Some of the activation functions we could use instead of reLU can be the `SoftMax` activation function. We can also use different loss functions and asses the change in predictions and some of the other loss functions can be the MAE-loss function or the Huber-loss function.

The final code we used for the 2, 3, 4 Layer NN models can be found in the following github link: GitHub NN Python Script.

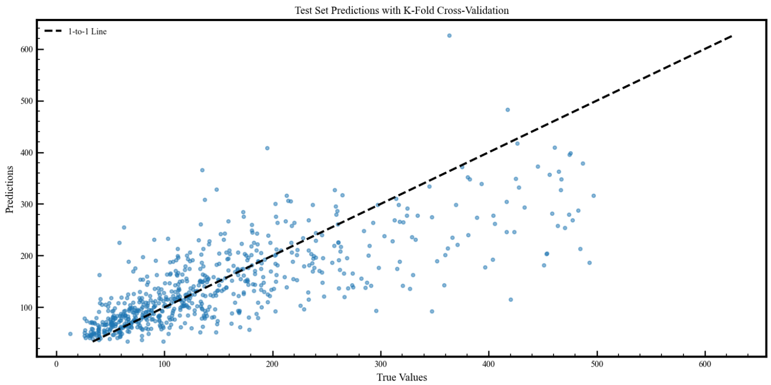

Figure 3: Actual versus predicted rest-frame Lyα equivalent widths for the test dataset. The dashed line indicates perfect agreement.

3.2 Model 2: XGBoost

XGBoost is particularly well-suited for predicting Lyα rest-frame equivalent widths (EW$_r$), as it effectively captures complex non-linear relationships, works well with sparse data, and incorporates built-in regularization techniques such as subsampling, L_1/L_2 penalties, and early stopping. Additionally, XGBoost provides inherent interpretability through feature-importance metrics, enabling clearer identification of key astrophysical predictors compared with neural networks, which typically require extensive tuning and feature engineering to achieve comparable interpretability.

Predictors. Guided by domain knowledge and iterative evaluations aimed at minimizing the Mean Squared Error, the following explanatory variables were selected for the XGBoost model (see definitions in Table 1):

- **dust:** A_V
- **stellar_mass**
- **sfr**
- **mass_weighted_age**
- **redshift**
- **delayed:age**

Data Preparation. The analysis was conducted using Python packages `scikit-learn` and `XGBoost`. The initial dataset of 11,862 galaxies was reduced to 3,393 by removing four entries with missing Lyα rest-frame equivalent widths (EW$_r$) values and applying quality criteria: photometric chi-squared $\chi^2_{\text{phot}} < 100$, Ly$\alpha$ signal-to-noise ratio S/N$_{\text{Ly}\alpha} > 5.3$, and rest-frame Lyα equivalent width $|\text{EW}_r| < 500\,\text{Å}$. These cuts were determined based on subject matter knowledge and evaluating the effect on model performance. The dataset was then split into training (2,714 galaxies) and test (679 galaxies) subsets using an 80/20 stratified split. Finally, all predictors were standardized with `StandardScaler` to ensure zero mean and unit variance, preventing any single variable from dominating the optimization.

Hyperparameter Tuning. Hyperparameter optimization was performed with a 300-trial `RandomizedSearchCV` employing five-fold cross-validation. This approach balances computational efficiency with thorough exploration of the parameter space. Compared to exhaustive `GridSearchCV`, the randomized search offers substantial speed advantages while reliably identifying near-optimal configurations. Key parameters tuned included learning rate, maximum tree depth, subsampling ratios, and regularization strengths (L_1 and L_2). Table 3 lists the optimal hyperparameters.

Table 3: Optimal XGBoost hyperparameters

Hyperparameter	**Value**
colsample_bytree	0.92
gamma	0.14
learning_rate	0.06
max_depth	3
n_estimators	291
reg_alpha (α)	2.88
reg_lambda (λ)	3.93
subsample	0.65

Early stopping halted training after 291 boosting rounds when no improvement in validation loss was observed for 50 consecutive iterations.

Model Performance. The final model was evaluated on the test dataset. Performance metrics are summarized in Table 4. The model had an $R^2 = 0.55$, indicating that the model explains 55% of the variance in observed Lyα EW$_r$. Considering the inherent complexity and measurement noise in astronomical datasets, where galaxies lie millions of light-years away, an RMSE of 66 and MAE of 47 demonstrate robust predictive capabilities.

Table 4: XGBoost model performance metrics on the test dataset

Metric	**Value**
R^2	0.55
RMSE	66
MAE	47

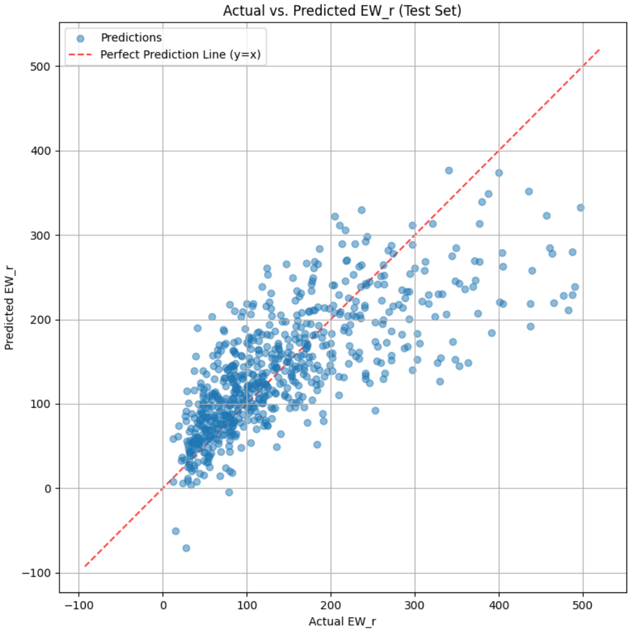

Figure 4: Actual versus predicted rest-frame Lyα equivalent widths for the test dataset. The dashed line indicates perfect agreement.

In summary, XGBoost provided excellent predictive accuracy, efficient handling of complex, sparse astrophysical data, and transparent interpretability, making it ideally suited to probe the physical mechanisms behind Lyα emission and advance our understanding of cosmic reionization. The feature interpretability of the XGBoost model is evaluated below in Section 4.

Code Availability. The XGBoost implementation is available here: XGBoost Model Code.

3.3 Model 3: Random Forest

We implemented a Random Forest regression model to predict the rest-frame Lyman-α equivalent width (EW$_r$) using various physical galaxy properties. Random Forest was chosen for its ability to model complex non-linear interactions and its built-in feature importance measures, which aid physical interpretability.

After applying the quality cuts

- $\chi^2_{\mathrm{phot}} < 40$,
- S/N$_{\mathrm{Ly}\alpha} > 5.5$,
- EW$_r$ $< 500\,\mathrm{Å}$,

we retained 1,965 galaxies, splitting them into 80% training (1,572) and 20% testing (393). We selected seven key features and standardized them with `StandardScaler`.

Table 5: Feature descriptions used in the Random Forest model.

Feature	Description
`burst`	Burstiness metric for star formation
`dust:` A_V	Dust attenuation in V-band (mag)
`nebular:logU`	Ionization parameter (log radiation field intensity)
`stellar_mass`	Current stellar mass ($\log_{10} M_\odot$)
`sfr`	Star-formation rate ($M_\odot\,\mathrm{yr}^{-1}$)
`ssfr`	Specific SFR ($\log_{10}\mathrm{SFR}/M_\odot$)
`mass_weighted_age`	Mass-weighted stellar age (Gyr)

Training and Hyperparameter Tuning

We tuned `n_estimators`, `max_depth`, `min_samples_split`, and `min_samples_leaf` using `GridSearchCV` with 5-fold cross-validation, and employed early stopping (halting training if validation loss failed to improve for 50 consecutive rounds). The final test performance was:

Table 6: Random Forest performance on the test dataset (393 galaxies).

Metric	Value
R^2	0.396
MSE	5,160

Overall, the Random Forest achieves moderate predictive power of ($R^2 = 0.396$) and a ($MSE = 5,160$) of the Lyman-α equivalent width. Due to the noisy underlying astronomy data, this model performs decently well.

3.4 Model 4: Decision Tree

We trained a Decision-Tree Regressor because it captures nonlinear relationships and feature interactions. It offers clear interpretability by simple if-then split rules. After data cleaning, we performed grid-search tuning and found the best hyperparameters to be:

Table 7: Optimal Decision Tree hyperparameters

Hyperparameter	**Value**
max_depth	6
min_samples_leaf	0.5
min_samples_split	5
ccp_alpha	0

The best tree achieved:

Table 8: Decision Tree model performance metrics on the test dataset

Metric	**Value**
Cross-validated R^2	0.210
Test-set R^2	0.172
Test RMSE	103.45

In general, the model captures only a fraction of the variance (17%). This suggests the model has substantial residual error.

3.5 Best Model Evaluation

As we can see, as the EW_R increases, our model fails to accurately predict, the outliers here are apparent. So we decided to take a look at the residual plot.

After examining the residual plot we see a downward moving trend as we predict larger EW_r. This suggested some skewness in our original data that cannot be captured by the model. So we decided to move on with another method that could see this. In summary, the decision tree model did poorly in trying to find non-linear relationships between our predictors and the response variable. We needed something like XG-boost to improve on it.

After determining the best model for each of our models the last step was to compare the performance of all of them to the data and see which model is able to perform the best. The results of the comparison can best be seen in the final comparison plot in Figure 7, where we show the RMSE for all the models and we can see that model that performed the best was XGBoost and then the NN with 4 layers came in second.

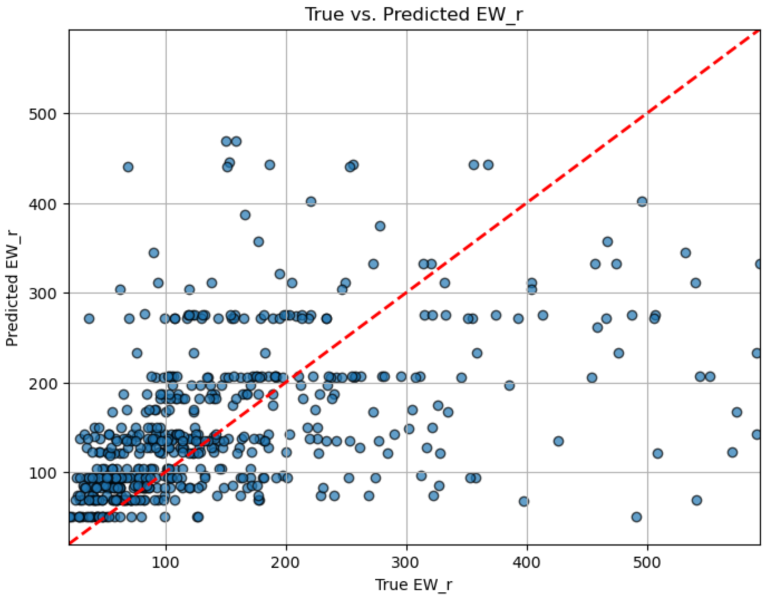

Figure 5: Actual versus predicted rest-frame Lyα equivalent widths for the test dataset. The dashed line indicates perfect agreement.

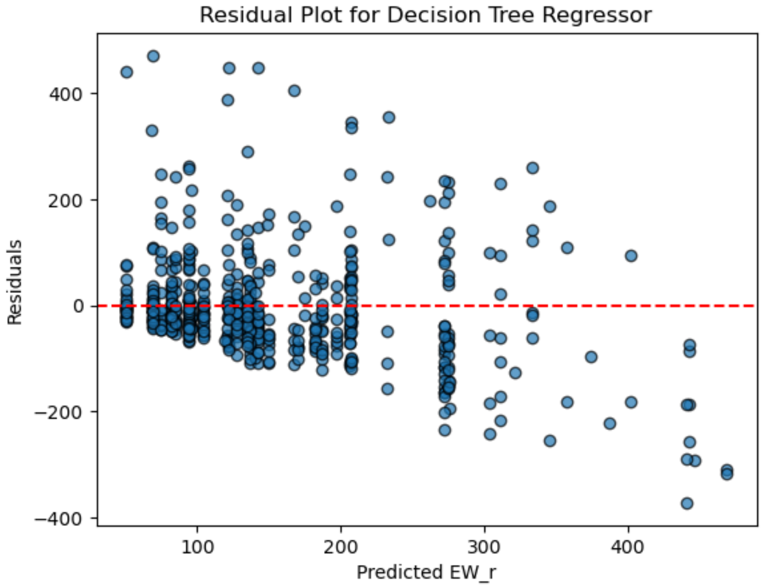

Figure 6: Residual plot of Lyα equivalent widths for the test dataset.

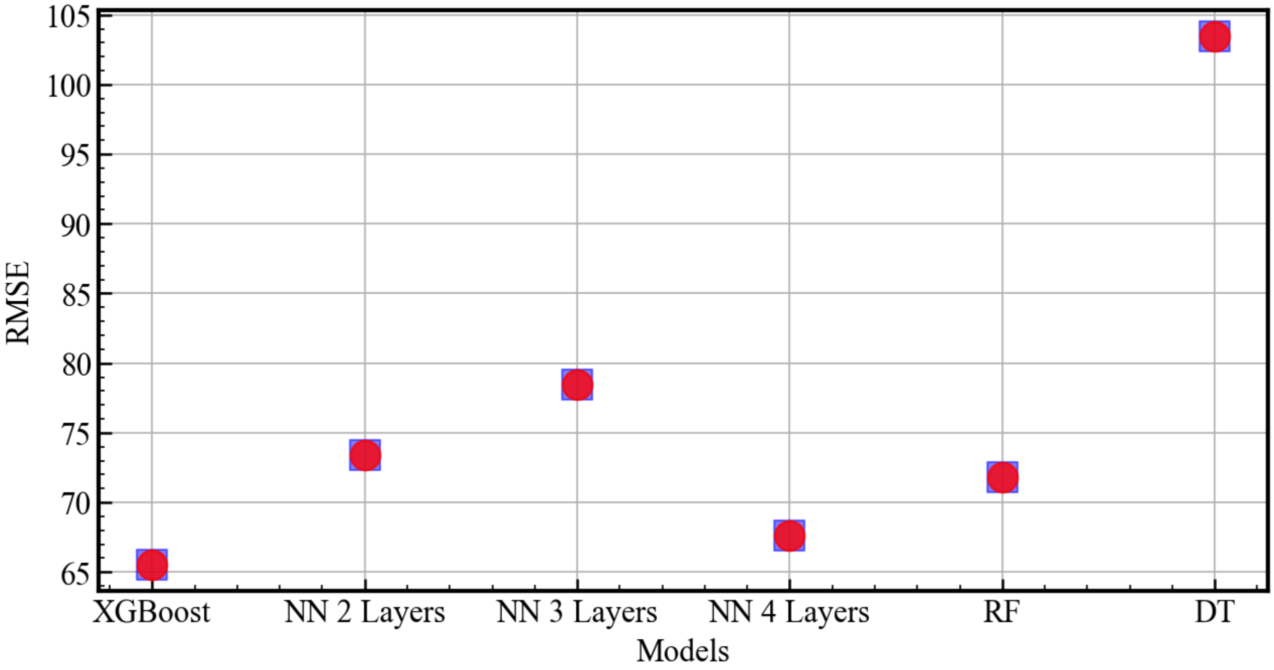

Figure 7: This plot shows a comparison of the RMSE scores between the different models on the respective testing set. We can see that all of the do comparable with an average RMSE of 80, but it is clear that the XGBoost has the lowest error followed by the NN model with 4 layers and K-fold validation. Random Forrest does comparable to some NN architecture and we see that DT is the model with the least predicitve power. RF and DT represent the Random Forrest model and the Decision Tree model respectively.

4 Question 2: Feature Importance

Which galaxy properties have the greatest impact on Lyman-Alpha emission predictions?

Model Interpretability with SHAP. As explained in the section above, the XGBoost Model performed the best. We now wanted to understand why the model makes a given prediction. In order to do this, we computed SHapley Additive exPlanations (SHAP) values for each explanatory variable. Figure 8 shows the full distribution of SHAP values for every galaxy (beeswarm), while Figure 9 ranks features by their mean absolute impact on the output. A quantitative summary is given in Table 9.

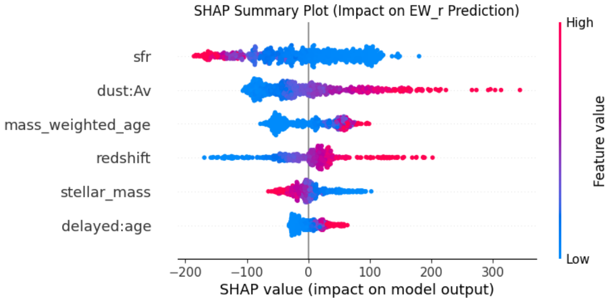

Figure 8: SHAP summary (beeswarm) plot: each dot represents one galaxy. Horizontal position indicates the contribution of that feature to the predicted EW_r; color encodes the raw feature value (blue = low, red = high).

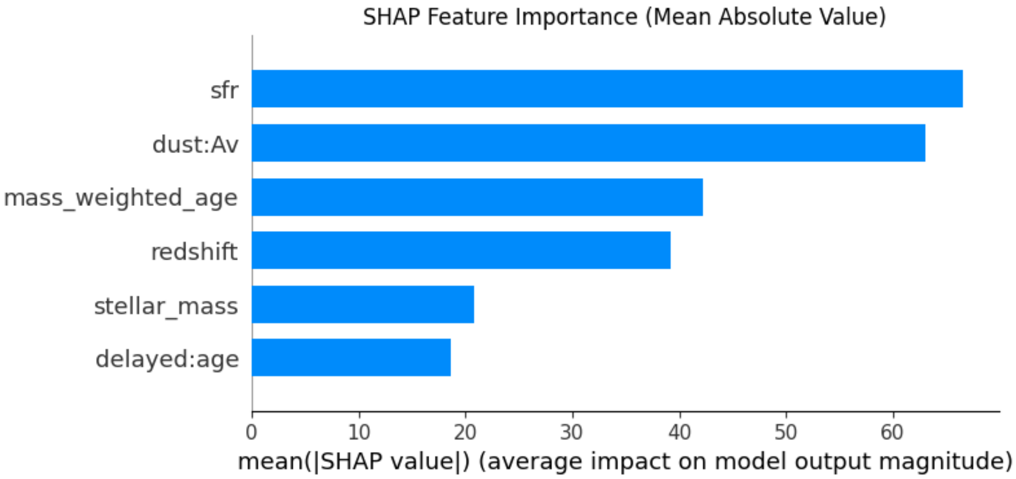

Figure 9: Mean absolute SHAP value for each predictor. Longer bars denote larger average influence on the model output.

The mean-absolute SHAP ranking shows that the predicted rest-frame Lyα equivalent width is primarily driven by changes in star-formation rate and dust attenuation in the V-band, with decreasing sensitivity to the stellar mass-weighted age, the galaxy's redshift, its stellar mass, and finally the delayed-τ star-formation timescale.

Table 9: Mean absolute SHAP values (impact on EW$_r$ prediction)

Feature	$\langle|\text{SHAP}|\rangle$ [Å]
sfr	66.6
dust: A_V	63.1
mass_weighted_age	42.2
redshift	39.2
stellar_mass	20.8
delayed:age	18.6

5 Conclusion and Future Work

5.1 Conclusion

For us the end goal is to come up with a model that can predict Lyα EW$_r$ and the XGBoost model is on par and better than the NN models. The advantage of using XGBoost over a typical NN model is that XGBoost is able to be interpretable. The XGBoost model can produce SHAP plots which show the relative feature importance to the prediction of Lyα which has real physical impact on understanding what galaxy properties drive Lyα observability. We can use physics to understand the deeper connection between the galaxy properties and Lyα observability and to hone in on why this trend exists.

In summary, our goal was to come up with a ML model that can predict Lyα equivalent width and our conclusions are the following:

1. ML approach can be used to predict the Lyα equivalent widths to within an EW of 60-80
2. After comparing multiple non-linear machine learning models we see that the best model is XGBoost
3. Using SHAP analysis on the XGBoost model we can see that the features that impact the predictions the most are Star Formation Rate (SFR) and dust (dust:A_V) which can be followed up on for further analysis to determine their physical connections

5.2 Future Work

1. **Broaden the training set**
   The original dataset contained 11,862 galaxies and after data cleaning, we only used 3,393 of the galaxies. Thus, our analysis was mainly on galaxies with clean data (photometric chi-squared $\chi^2_{\text{phot}} < 100$, Ly$\alpha$ signal-to-noise ratio S/N$_{\text{Ly}\alpha} > 5.3$, and rest-frame Lyα equivalent width $|\text{EW}_r| < 500$ Å). Future research could include analyzing galaxies where the Lyα equivalent widths data is less clean.

2. **Refine model architectures**
   Explore deeper neural networks and alternative activations and loss functions to determine whether current hyperparameter choices limit performance. Further refine the hyperparameter selection for XGBoost, Random Forest, and Decision Trees.

3. **Improve underlying data-quality**
   There were a substantial number of rows with unreasonably high levels of Lyα in the underlying data. Future work could include improving the underlying code used to analyze and collect the galaxy physical properties.

6 Contributions

- **Óscar A. Chávez Ortiz**: Developed the NN model and completed Project Description and Background, contributed to section 2 and the conclusion.
- **Ethan Abraham**: Developed the XGBoost model and completed the XGBoost, feature importance, and future work sections; contributed to data description, optimizing data filtering cutoffs, and formatting document.
- **Kyla Ko**: Developed and implemented the Random Forest model, including feature selection, data cleaning with photometric and emission line quality cuts, and hyperparameter tuning using GridSearchCV. Conducted SHAP-based feature importance analysis and contributed to comparative model evaluation.
- **Vincent Cheng**: Developed Decision Tree model, including feature selection and data cleaning.