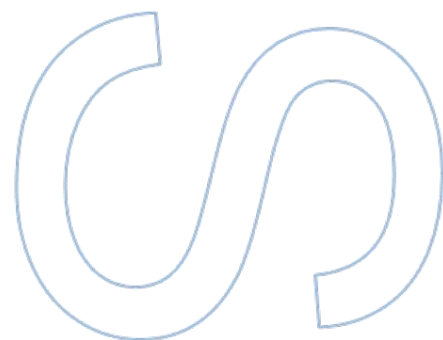
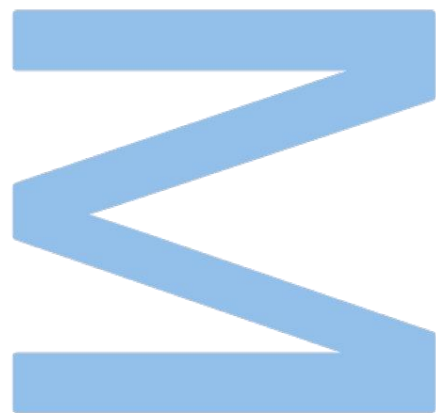


Boosting Lyman- α Identification and Characterization in the Era of Large Surveys

Afonso Vale

Master in Astronomy and Astrophysics
Department of Physics and Astronomy
Faculty of Sciences of the University of Porto
2024



Boosting Lyman- α Identification and Characterization in the Era of Large Surveys

Afonso Vale

Dissertation carried out as part of the Master in Astronomy
and Astrophysics
Department of Physics and Astronomy
2024

Supervisor

Dr. Ana Paulino-Afonso,
Researcher, Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto

Co-supervisor

Dr. José Fonseca,
Researcher, Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto

Acknowledgements

I conclude this work, marking a significant chapter in my academic and personal life. My journey began with a Bachelor's degree in Mathematics, which, despite its challenges, equipped me with the essential tools for developing analytical reasoning and critical thinking. Pursuing a Master's degree in Astronomy and Astrophysics, I had the opportunity to study and do research in a scientific field that has always captivated me. I am grateful to FCUP for the chance to explore these two wonderful areas.

A significant portion of this work was developed under a fellowship (Ref: IA2022-13-BI-M) at the Centro de Astrofísica da Universidade do Porto in the context of the project "FLAEMING", funded by Portuguese funds through FCT - Fundação para a Ciência e a Tecnologia (reference EXPL/FIS-AST/1085/2021) and I am very grateful for this opportunity. Besides the fellowship, I must acknowledge the "FLAEMING" project for giving me the first opportunity to do research, and for all the support provided. I would also like to acknowledge Instituto de Astrofísica e Ciências do Espaço for being the host institution for all the work developed.

To my supervisors, I extend my thanks for your guidance and support during this research project, both during the research fellowship I enjoyed last year and throughout the work for this dissertation. To Ana, for your dedication and for providing opportunities that allowed me to participate in conferences, give oral presentations, and present posters. To José, for your availability and invaluable advice over the time. To Andrew, for being an indispensable help throughout this project. I also extend my gratitude to the entire FLAEMING project team.

I also want to thank my family, father, mother, sisters, and Fátima—for their unwavering support from the beginning of my life to the completion of this dissertation. To Bernardo and Jorge, for all the conversations about everything and nothing, and for their enduring friendship. To Rita, the greatest joy that the Faculty has given me, thank you for walking this path with me, and for making every moment memorable.

This work is based on observations collected at the European Southern Observatory under ESO programme ID 179.A-2005 and on data products produced by CALET and the Cambridge Astronomy Survey Unit on behalf of the UltraVISTA consortium.

HETDEX is led by the University of Texas at Austin McDonald Observatory and Department of Astronomy with participation from the Ludwig-Maximilians-Universität

München, Max-Planck-Institut für Extraterrestrische Physik (MPE), Leibniz-Institut für Astrophysik Potsdam (AIP), Texas AM University, Pennsylvania State University, Institut für Astrophysik Göttingen, The University of Oxford, Max-Planck-Institut für Astrophysik (MPA), The University of Tokyo and Missouri University of Science and Technology.

Observations for HETDEX were obtained with the Hobby-Eberly Telescope (HET), which is a joint project of the University of Texas at Austin, the Pennsylvania State University, Ludwig-Maximilians-Universität München, and Georg-August-Universität Göttingen. The HET is named in honor of its principal benefactors, William P. Hobby and Robert E. Eberly. The Visible Integral-field Replicable Unit Spectrograph (VIRUS) was used for HETDEX observations. VIRUS is a joint project of the University of Texas at Austin, Leibniz-Institut für Astrophysik Potsdam (AIP), Texas AM University, Max-Planck-Institut für Extraterrestrische Physik (MPE), Ludwig-Maximilians-Universität München, Pennsylvania State University, Institut für Astrophysik Göttingen, University of Oxford, and the Max-Planck-Institut für Astrophysik (MPA).

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high performance computing, visualization, and storage resources that have contributed to the research results reported within this paper. URL: <http://www.tacc.utexas.edu>

Funding for HETDEX has been provided by the partner institutions, the National Science Foundation, the State of Texas, the US Air Force, and by generous support from private individuals and foundations.

Finally, I would like also to acknowledge the SC4K ([1]) and DEIMOS ([2]) teams for making publicly available their catalogs, which were fundamental for this work.

UNIVERSIDADE DO PORTO

Abstract

Faculdade de Ciências da Universidade do Porto

Departamento de Física e Astronomia

MSc. Astronomy and Astrophysics

Boosting Lyman- α Identification and Characterization in the Era of Large Surveys

by [Afonso VALE](#)

Identifying Lyman- α emitting galaxies (LAEs) typically involves dedicated surveys with custom narrow-band filters. With the growing data volumes from current and future surveys, AI-driven approaches are essential. We utilize gradient-boosting algorithms trained on balanced data from the COSMOS2020 catalog and the SC4K survey to identify and characterize LAEs. The models achieved F1-scores of 85-90%, while also showing promise for pre-selecting LAEs in large-scale surveys. Some predicted LAEs were confirmed using public spectroscopic catalogs. In addition to selecting new LAE candidates, we trained regression models to estimate redshift, Ly α line luminosity and equivalent width, achieving mean absolute errors below 0.18 for redshift and Ly α luminosity, and below 0.25 for the Ly α equivalent width. We also assessed the models' robustness against minor data perturbations. Our models can help to efficiently identify and characterize LAEs, discovering over 6000 new LAEs in the COSMOS field, and enhancing the study of these galaxies with machine learning.

Keywords: galaxies: high redshift; galaxies: photometry; galaxies surveys and catalogs; machine learning; supervised learning from labeled data.

UNIVERSIDADE DO PORTO

Resumo

Faculdade de Ciências da Universidade do Porto

Departamento de Física e Astronomia

Mestrado em Astronomia e Astrofísica

Potenciando a Identificação e Caracterização de Lyman- α na Era dos Grandes Levantamentos de Dados

por [Afonso VALE](#)

Encontrar galáxias emissoras de Lyman- α (LAEs) geralmente envolve estudos dedicados usando filtros de banda estreita específicos. Com o crescente volume de dados provenientes de pesquisas atuais e futuras, decisões conduzidas por inteligência artificial são cada vez mais necessárias. Usamos algoritmos de *gradient boosting* treinados com dados equilibrados do catálogo COSMOS2020 e do levantamento SC4K para identificar e caracterizar LAEs. Os modelos alcançaram pontuações de F1 entre 85 – 90%, demonstrando também potencial para pré-selecionar LAEs em futuros levantamentos de grande escala. Algumas LAEs previstas foram confirmadas utilizando catálogos espectroscópicos públicos. Além de selecionar novos candidatos a LAE, também treinamos modelos de regressão para obter uma primeira aproximação do desvio para o vermelho, luminosidade da linha de Ly α e largura equivalente da linha de Ly α , alcançando erros médios absolutos menor que 0.18 para o desvio para o vermelho e luminosidade da linha de Ly α e menor que 0.25 para a largura equivalente da linha de Ly α . Também avaliamos a robustez dos modelos contra pequenas perturbações nos dados. Os nossos modelos podem ajudar a identificar e caracterizar mais eficientemente LAEs, descobrindo mais de 6000 novas LAEs no campo COSMOS e melhorando o estudo destas galáxias com aprendizagem automática.

Palavras-chave: galáxias: alto desvio para o vermelho; galáxias: fotometria; levantamentos de galáxias e catálogos; aprendizagem automática; aprendizagem supervisionada usando dados rotulados.

Contents

Acknowledgements	v
Abstract	vii
Resumo	ix
Contents	xi
List of Figures	xiii
List of Tables	xv
1 Introduction	1
1.1 Historical background, definition of LAEs and Ly α physical origin	1
1.2 LAEs general properties	3
1.3 LAEs search techniques	3
1.4 Previous Machine Learning applications to LAEs studies	6
1.5 Structure of this dissertation	7
2 Machine Learning Framework	9
2.1 Ensemble learning	10
2.1.1 Gradient boosting algorithms	11
2.2 Model evaluation	15
2.3 Data splitting	18
2.4 Hyperparameter optimization	20
3 Data Handling and Processing	25
3.1 Observational surveys	25
3.1.1 COSMOS	25
3.1.2 SC4K	26
3.2 Data preparation	27
3.3 Feature selection	29
3.4 Data characterization	31
4 Identifying LAEs using photometric data	37
4.1 Performance metrics and feature importance analysis	37
4.2 Predicting LAE candidates in COSMOS2020 at $z \approx 2 - 6$	44

4.3	Crossmatching predicted LAEs with spectroscopic catalogs	45
4.3.1	HETDEX	45
4.3.2	DEIMOS	47
4.4	Generalization in the era of large surveys	48
4.5	Assessing the robustness of the classification models	51
5	Properties of LAEs candidates: redshift, Lyα luminosity and equivalent width	57
5.1	Revising data preparation	57
5.2	Performance metrics, feature importance, and generalization	59
5.3	Characterizing LAE candidates in the COSMOS2020 survey	63
5.4	Robustness analysis of the regression models	67
6	Conclusion	71
A	Supplementary information - Classification and Regression models	75
	Bibliography	81

List of Figures

1.1	SFR as a function of stellar mass for LAEs at $z \approx 2$ and SFGs at similar redshifts.	4
1.2	Illustration of the narrow-band selection for an LAE at $z = 6.6$	5
1.3	LAE candidate found in VLT/FORS grism data in a blank field.	5
2.1	Summary sketch of Friedman’s gradient boosting algorithm.	14
2.2	Difference of tree growth strategies.	15
2.3	Example of a binary task confusion matrix.	17
3.1	Relative transmission curves for the photometric bands present in COSMOS2020.	26
3.2	The 3D distribution of the SC4K sample in the COSMOS field	27
3.3	Non-LAE and LAE number of sources distribution per redshift bin.	28
3.4	Non-LAE and LAE samples i-band magnitude distribution.	28
3.5	Non-LAE and LAE samples magnitudes distribution and statistics.	32
3.6	Stellar Mass and SFR distributions for LAEs and non-LAEs samples.	33
3.7	Mass-SFR relation for the LAEs and non-LAEs samples at different redshift ranges.	33
3.8	NUVrK color-color diagrams for LAE and non-LAE samples.	34
4.1	Confusion matrices for LightGBM.	39
4.2	Confusion matrices for XGBoost.	39
4.3	Confusion matrices for CatBoost.	40
4.4	Classification: Feature importances of LightGBM.	41
4.5	Classification: Feature importances of XGBoost.	42
4.6	Classification: Feature importances of CatBoost.	42
4.7	Prediction probability distribution for LightGBM.	43
4.8	Prediction probability distribution for XGBoost.	43
4.9	Prediction probability distribution for CatBoost.	43
4.10	Distribution of the number of times a source is predicted as an LAE.	44
4.11	Distribution of the prediction probabilities for each source.	45
4.12	Variation of the F1-Score through perturbations.	52
4.13	Variation of the number of predicted LAEs through perturbations.	53
4.14	Comparison of the perturbed and unperturbed prediction probabilities.	54
4.15	Linear fits to the 2D histograms from Figure 4.14.	55
4.16	Comparison of the perturbations in the prediction data with different feature sets.	55

5.1	Distribution of redshift, Ly α luminosity, and Ly α equivalent width for the LAE sample.	58
5.2	Feature Importance for the redshift regression models.	60
5.3	Feature Importance for the Ly α luminosity regression models.	60
5.4	Feature Importance for the Ly α equivalent width regression models.	61
5.5	Comparison between the properties' real values and the values predicted by the regression models in the test set.	62
5.6	Comparison between the properties' real values and the values predicted by the regression models for the prediction sample.	64
5.7	Comparison between the spectroscopic redshift from HETDEX and DEIMOS with the model predictions.	65
5.8	Comparison between the Ly α luminosity from HETDEX with the model predictions.	65
5.9	Number density of LAEs per luminosity interval.	67
5.10	Variation of MAE through perturbations in the regression models.	68
5.11	Comparison between the properties values predicted by the unperturbed and perturbed regression models.	70
A.1	Classification: Feature importances for LightGBM detailed per sample. . . .	77
A.2	Classification: Feature importances for XGBoost detailed per sample. . . .	78
A.3	Classification: Feature importances for CatBoost detailed per sample. . . .	78

List of Tables

2.1	Summary of the grids used in the hyperparameter optimization.	23
3.1	Fluxes and Magnitudes used to train the models. For both fluxes and magnitudes, 3 arcsec diameter aperture was used.	30
3.2	Information of the broad-bands used to train the models.	30
3.3	Correlation coefficients for the 10 most correlated features with the target label.	35
4.1	Classification evaluation metrics with default settings.	38
4.2	Classification evaluation metrics with optimized settings.	38
4.3	Confusion Matrix for the crossmatch with HETDEX.	46
4.4	Evaluation Metrics for the crossmatch with HETDEX.	46
4.5	Confusion matrix for the crossmatch with HETDEX (threshold of 25%) . . .	47
4.6	Evaluation metrics for the crossmatch with HETDEX (threshold of 25%) . .	47
4.7	Classification evaluation metrics for our work compared with other surveys. .	50
5.1	Redshift model evaluation metrics with optimized settings.	59
5.2	Ly α luminosity model evaluation metrics with optimized settings.	59
5.3	Ly α equivalent width model evaluation metrics with optimized settings. . .	60
5.4	Regression evaluation metrics for our work compared with other surveys. .	63
A.1	Sample 1 classification evaluation metrics	75
A.2	Sample 2 classification evaluation metrics	76
A.3	Sample 3 classification evaluation metrics	76
A.4	Sample 4 classification evaluation metrics	76
A.5	Sample 5 classification evaluation metrics	77
A.6	Redshift models evaluation metrics with default hyperparameters.	78
A.7	Ly α Luminosity models evaluation metrics with default hyperparameters. .	79
A.8	Ly α equivalent width models evaluation metrics with default hyperparameters.	79

Chapter 1

Introduction

In this opening chapter, we will start by providing a brief but concise overview of Lyman- α (hereafter, $\text{Ly}\alpha$) emitters, covering their historical context, the origins of the $\text{Ly}\alpha$ emission line, as well as the main features that distinguish them from other star forming galaxies. The most common search and detection methods of these sources will also be covered and how machine learning can help to overcome the difficulties in assembling larger samples of $\text{Ly}\alpha$ emitters.

Please consult the primary references provided for a detailed outline of $\text{Ly}\alpha$ emitters ([3] and [4]).

1.1 Historical background, definition of LAEs and $\text{Ly}\alpha$ physical origin

The $\text{Ly}\alpha$ emission line stands out as one of the most prominent spectral features emitted by star-forming galaxies. It arises from the abundant presence of hydrogen and involves an atomic electron transition from the $n = 2$ state to $n=1$, the ground state. With its short wavelength of $1216\text{\AA}$, the $\text{Ly}\alpha$ emission proves highly effective for identifying faint, high-redshift objects through optical and near-infrared (NIR) observations. The idea of $\text{Ly}\alpha$ emission serving as an indication of young galaxies was initially proposed by Partridge and Peebles [5] in 1967, where they explored the feasibility of detecting such distant galaxies, particularly those in their early formation stages. Their investigation was based on a specific model of galaxy formation, positing that gravitational instabilities in expanding gas led to galaxy formation, with early stages characterized by high

luminosity, resultant from the presence of luminous and young O-B type stars, that ionize the surrounding gas and their subsequent recombination could generate strong Ly α emission. Despite various attempts to discover such galaxies during the 1970s, significant breakthroughs came only in the early 1990s with the adoption of the narrow-band technique, leading to the first detections of Ly α emitters at high redshifts, in [6], [7], and [8] at redshifts $z = 2.3$, $z = 2.466$, and $z = 3.428$, respectively.

Formally defining Ly α emitters, these are objects emitting Ly α radiation with a rest-frame equivalent width (EW_0) exceeding approximately $20\text{\AA}$. The criteria of Ly α $EW_0 \gtrsim 20\text{\AA}$ is justified as a consequence of the selection limit of narrow-band imaging surveys for Ly α emitting galaxies. With the introduction of the concept of Ly α emitters (hereafter referred to as LAEs), it becomes pertinent to briefly explore the various astrophysical environments where Ly α production occurs and the main differences between LAEs and other star-forming galaxies.

Ly α emission in the interstellar medium (ISM) can originate from two main astrophysical processes, both involving hydrogen recombination. One primary source is interstellar HII regions, where the surrounding gas is ionized by the intense ultraviolet radiation from massive, young stars. In these regions, recombining electrons and protons generate a range of emission lines, including Ly α . This emission is often associated with active star formation, but it can also be triggered by active galactic nuclei (AGN), which produce highly ionized gas in both the broad and narrow-line regions near the galaxy's center.

Ly α emission can also arise from the circumgalactic medium (CGM) and more distant regions, where the gas may be ionized through various mechanisms. These include collisional excitation of neutral hydrogen (HI) atoms due to outflows of gas ([9]), the release of gravitational energy by infalling gas that radiates as Ly α , or fluorescence from HI in the CGM that is photoionized by ultraviolet background radiation, potentially from energetic sources such as quasars ([10]).

Due to the large Ly α scattering cross-section of HI, Ly α photons undergo resonant scattering, which can alter their direction relative to the observer's line of sight, either enhancing or diminishing the observed flux. This scattering also impacts the velocity profile, affecting measurements of redshift and line width ([11]). It may also lead to the formation of neutral gas haloes, where Ly α emission is visible primarily due to scattering effects ([12]). Resonant scattering of Ly α photons encodes key information about the spatial distribution and kinematics of HI gas ([3]). By examining the Doppler shift of Ly α

emission, it is possible to infer gas motions, while mapping the intensity of Ly α can reveal the distribution of HI gas. This allows the study of HI in high-redshift systems, which are otherwise difficult to observe directly ([3]; [13]).

1.2 LAEs general properties

Investigating the physical characteristics of Lyman-alpha emitters (LAEs) sheds light on the differences between these galaxies and other populations, while also offering valuable insights into the physics of Ly α emission. LAEs are typically compact, with an average half-light radius of $r_e \approx 1$ kpc, and are smaller than other star-forming galaxies (SFGs) at all redshifts ([14]), with an inverse correlation observed between the size of a galaxy and its Ly α emission. This compact nature reduces the distance Ly α photons travel, thus lowering the chances of absorption by dust and enhancing the likelihood of their escape. Studies of the stellar populations in LAEs suggest that they are generally low in mass ($M_* \approx 10^{8-9} M_\odot$) and have low metallicities ($Z < 2 \times 10^{-2} Z_\odot$) ([3], indicating minimal chemical enrichment. This is consistent with their classification as star-forming galaxies, characterized by a star formation rate (SFR) of $\approx 1 - 10 M_\odot \text{ yr}^{-1}$ and young stellar ages (about 10 Myr) ([3], [15], and [16]). These findings suggest that LAEs may provide a window into the properties of the earliest galaxies before they grew through mergers or accretion of material.

Additionally, plotting the star formation rate (SFR) against stellar mass (M_*) offers insight into the activity level of galaxies or galaxy populations. For instance, the Milky Way ($M_* \approx 10^{11} M_\odot$, $\text{SFR} \approx 1 M_\odot \text{ yr}^{-1}$) falls near the middle of such a plot, while starburst galaxies show higher SFRs, and quiescent galaxies exhibit lower ones. In Figure 1.1, SFR is displayed as a function of stellar mass for LAEs at $z \approx 2$ ([17]) and for SFGs at comparable redshifts ([18]). LAEs tend to have higher SFRs relative to typical SFGs, likely due to the young ages of their stellar populations.

1.3 LAEs search techniques

The most common method for detecting LAEs is narrow-band (NB) imaging, which involves comparing images taken using a narrow-band filter (100–200 Å wide, see [19] or [20]) with a broad-band image to identify a flux excess in the narrow-band image (see Figure 1.2). The central wavelength of these narrow-band filters is selected based on the

scientific goals of the study, such as the redshift of the targeted LAEs or to avoid contamination from other emission lines. However, this technique has some limitations, such as probing only small redshift ranges and being biased toward galaxies with relatively bright Ly α emission, and the possibility that these NB surveys might be contaminated by galaxies, specially at lower redshifts, that reveal other emission lines, such as CIV emission at 1549Å (e.g., [21]), MgII at 2798Å (e.g., [4]), [OII] at 3727Å (e.g., [22]) or [OIII] at 5007Å (e.g., [23]).

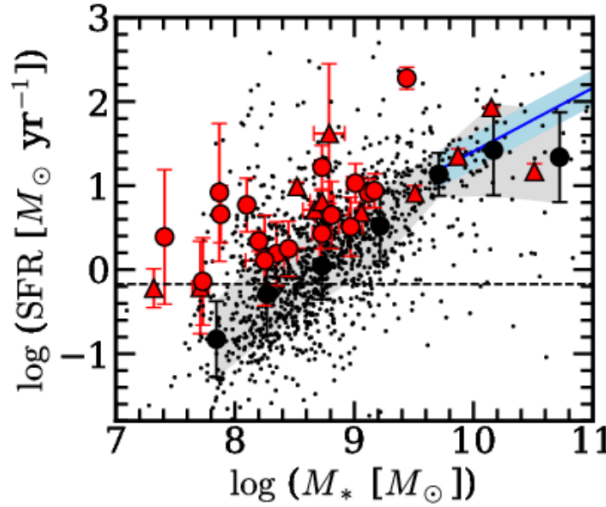


FIGURE 1.1: The star formation rate (SFR) as a function of stellar mass for LAEs at $z \approx 2$ is presented using data from [17]. Red circles and triangles represent LAEs detected in the narrow-band (NB) survey and the HETDEX Pilot Survey, respectively. For comparison, black dots correspond to star-forming galaxies (SFGs) at similar redshifts, with larger black circles indicating their average values ([18]). The blue solid line and grey shaded region illustrate the average SFR and associated uncertainties for SFGs at comparable redshifts, as derived from deep observational studies ([24]). This figure and the information provided in caption was taken from [3].

Another popular technique to search for LAEs is blind spectroscopy, where the LAE candidate is detected as a single-line emitter in a grism image (see Figure 1.3), which is an example of slitless spectroscopy, that is not efficient in assembling large number of LAEs sources. The blind spectroscopy method also includes slit spectroscopy, that has a wider field of view, but leads to lower spectra resolutions and higher background noise.

Additionally, integral field spectrographs (IFS) can also be used, where the spectrograph is equipped with an integral field unit, combining spectrography and imaging techniques. Thus, they can observe spectra of unbiased samples of galaxies, but their limited field of view restricts their effectiveness as wide-area survey probes.

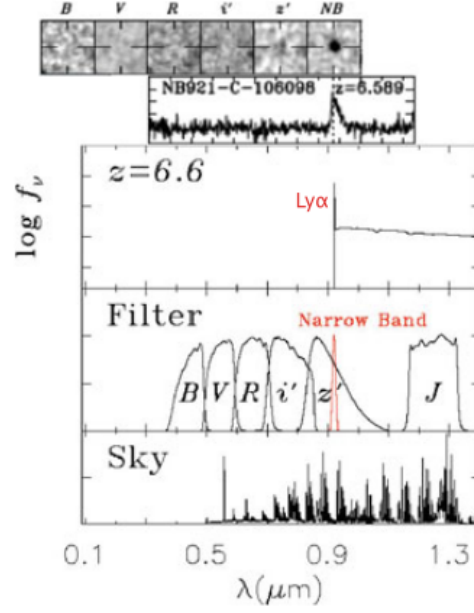


FIGURE 1.2: This figure illustrates the narrow-band selection for an LAE at $z = 6.6$. The top panel displays images of the LAE captured using both broad-band and narrow-band filters, with the narrow-band having a central wavelength of approximately $9200 \text{ \AA}$. The transmission curves for these filters are shown in the second panel from the bottom. The second panel from the top shows the spectrum of the LAE within the wavelength range of $9050\text{--}9275 \text{ \AA}$. The third panel from the top presents the model spectrum of an LAE redshifted to $z \approx 6.6$. The bottom panel shows the atmospheric OH emission lines. This figure and the information provided in caption was taken from [25].

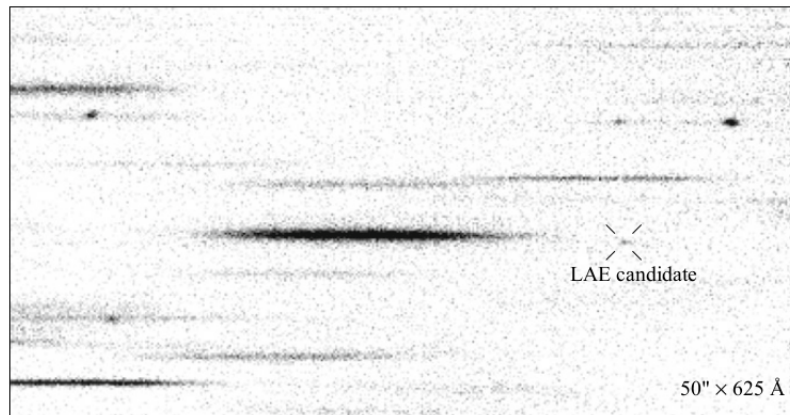


FIGURE 1.3: Spectral image of a blank field obtained with VLT/FORS in slitless spectroscopy mode limited to the atmospheric window at $\lambda \approx 9100 \text{ \AA}$. The $\text{Ly}\alpha$ emission line from a LAE candidate is marked with a cross and is at $z = 6.5$. The $\text{Ly}\alpha$ line detected has a flux of $1.9 \times 10^{-17} \text{ erg cm}^{-2} \text{ s}^{-1}$ and rest frame equivalent width $> 80 \text{ \AA}$. Other emission lines from other objects in the field are also detected, including an $[\text{O III}]$ emitter at $z = 0.8$ at the top right. This image was taken from [25].

In this context, machine learning emerges as a possible solution to overcome the current limitations in assembling larger samples of LAEs, allowing a selection of potential LAE candidates and a first approximation of key properties, such as their redshift, Ly α luminosity, and Ly α equivalent width. Astronomical data is undergoing a swift growth in size and complexity, primarily due the advent of the large-scale survey era. These ongoing and future surveys generate massive datasets that need to be processed and analyzed. The quantity of observed celestial objects and their characteristics is set to increase by over a magnitude (e.g., Euclid and LSST) and so a need for automated tools to efficiently process this data arises. The intersection of astronomy and machine learning has opened new avenues for LAE detection and study.

1.4 Previous Machine Learning applications to LAEs studies

There has not been a vast number of machine learning techniques applied to the study of LAEs but we provide here a brief outline of the main past work regarding this topic.

In [26], convolutional neural network models were built to remove contaminants from photometrically selected LAE candidates by performing image classification based on narrow-band and broad-band filter data. Unsupervised machine learning algorithms were also applied to select LAEs in the HETDEX spectroscopic survey ([27]), employing dimensionality reduction algorithms. In [28], a linear regression was used to predict Ly α luminosity, based upon observed properties, such as broad-band luminosities or other nebular lines and also using derived physical properties, such as the stellar mass or star formation rate.

Finally, in [29], a machine learning model based on Random Forest was developed to select LAEs, achieving $(80 \pm 2)\%$ accuracy and $(73 \pm 4)\%$ precision in the redshift range $z \in [2.5, 4.5]$. At higher redshifts ($z \in [4.5, 6]$), they obtain a 73% accuracy and a 80% precision. This model was built on the basis of the physical (stellar mass, SFR, reddening, metallicity, age) and morphological properties (Sérsic index, half-light radius, projected semi-major axis) and the training sample was based on several samples of spectroscopic confirmed LAEs.

In order to overcome the difficulties in assembling a large sample of LAEs, in this work we intend to train supervised machine learning gradient boosting algorithms, using more than 3000 LAEs from the SC4K survey and "non-LAEs" from COSMOS2020 and building on correlations between LAEs and their fluxes, magnitudes and colors in optical and

NIR broad-bands information, paving the way for future application to larger surveys. The main purpose is to pre-select LAEs and then obtain a first approximation of several key properties, such as their redshift, $\text{Ly}\alpha$ luminosity and equivalent width, gathering information that could be used in the context of target selection for spectroscopic surveys. Overall, with this work we aim to have a tool that is useful to pre-select a large amount of LAEs at $z \in [2, 6]$ using only photometric information. Proven successful, this is a novel method that could also drive spectroscopic follow-ups of the selected LAE candidates.

1.5 Structure of this dissertation

This dissertation is structured as follows. Chapter 2 outlines the algorithmic framework used, detailing model evaluation techniques, data division strategies, and hyperparameter optimization methods. Chapter 3 provides a description and a concise characterization of the data used. In Chapters 4 and 5, the results of the Machine Learning methods are presented for the classification and regression tasks, respectively. Both these chapters include an in-depth analysis of the training performance, and the predictions concerning both tasks are discussed, alongside evaluations of the models' generalization abilities and their robustness to minor data perturbations. Finally, Chapter 6 offers a summary of the key findings and suggests potential paths for future research.

In the following, we adopt the Λ CDM concordance cosmological model ($H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$). The programming language used was Python. The main Python packages used were: Pandas ([30]), Numpy ([31]), Matplotlib ([32]), and Seaborn ([33]).

Chapter 2

Machine Learning Framework

In light of the accelerated growth in size and complexity of astronomical datasets, astronomers are developing automated tools to detect, characterize, and classify objects using these rich and complex datasets (e.g., [34],[35] or [36]). The following overview of machine learning methods applied to astronomy can also be found in [37].

Machine learning algorithms have gained increasing popularity among astronomers and it translates into the use and development of computer systems that can learn and adapt, leveraging algorithms and statistical models to analyze and infer patterns from data. These algorithms are broadly categorized into two groups, depending on whether labeled data is used to predict outcomes:

- **Supervised learning:** Involves training models using labeled datasets, where each input is associated with a corresponding target output. The main goal is to learn a function that effectively maps input features to their correct outputs, enabling the model to make accurate predictions. During training, the algorithm iteratively updates its parameters to reduce the difference between the predicted outputs and the true labels. After training, the model can generalize to new, unseen data, making it an effective predictive tool.

- **Unsupervised learning:** Focuses on training models with unlabeled data, allowing them to uncover patterns and structures without explicit guidance. Key tasks include clustering data points based on similarity, reducing dimensionality for better visualization while preserving important information, identifying anomalies as outliers, and learning useful features automatically. This technique is particularly valuable for exploratory data analysis, as it reveals hidden structures and provides deeper insights into complex datasets.

Throughout this thesis, we will be working with labeled datasets, therefore, our focus will be on supervised learning. Within supervised learning, two learning tasks are commonly focused:

1. **Classification task:** This involves categorizing input data into predefined classes or groups. The main objective is to develop a function that maps input features to discrete labels, resulting in an output that assigns each data point to a specific category. In the context of this thesis, in this task we focus on building a model that can distinguish LAEs from other galaxies, such as those in a photometric catalog.
2. **Regression task:** This task involves predicting a continuous numerical output from given input features. The aim is to learn a mapping from input variables to a continuous target variable, where the result is a numerical value representing the predicted or estimated value of the target. In our case, the goal is to characterize predicted LAE galaxies by estimating key parameters like their redshift, Ly α luminosity, and Ly α equivalent width.

2.1 Ensemble learning

Ensemble learning is a machine learning approach aimed at improving predictive accuracy by combining the outputs of several models. Rather than relying on a single model trained on the dataset, ensemble methods create a collection of base models, such as decision trees, neural networks, or other algorithms. By aggregating the predictions from these diverse models, ensemble learning often results in better accuracy and increased robustness. The three primary types of ensemble techniques ([38]) are bagging, stacking, and boosting, which will be briefly introduced.

Stacking

Stacking is an ensemble learning technique where a model is trained to combine the predictions of multiple learning algorithms. This approach begins by training several base models, called first-level learners, using the available dataset. Then, a meta-model is trained to make final predictions by aggregating the outputs from the first-level models. Typically, these first-level models are derived from different algorithms, forming a heterogeneous ensemble, though homogeneous ensembles can also be constructed using

the same algorithm type. Stacking enhances predictive performance by combining the strengths of diverse models.

Bagging

Bagging, short for Bootstrap Aggregating, is a parallel ensemble technique where base learners are trained independently in parallel, aiming to reduce errors based on their independence. By training each base learner on different subsets of data, the method aims to achieve a more robust model. One challenge is the potential lack of sufficiently large datasets, leading to unrepresentative samples. To address this, bootstrap sampling is applied, where, from a given dataset D , several new datasets D_i are generated by sampling from D , with or without replacement. Each sample is then used to train a base learner, and their outputs are combined, using voting for classification or averaging for regression. Random forest is an example the bagging technique ([39]).

Boosting

Boosting is a sequential ensemble method where base learners are trained one after the other, with each learner building upon the performance of the previous ones. Unlike bagging, boosting explores the dependency between the base learners to improve model accuracy by focusing on minimizing errors. In this work, the algorithms used belong to the boosting family, specifically Gradient Boosting Machines, which will be discussed in further detail in the following section.

2.1.1 Gradient boosting algorithms

Gradient Boosting Machines (GBMs) are a powerful ensemble learning technique that combines predictions from multiple base models, usually decision trees, to form a strong predictive model. This section is mainly based on [40] and [41], and will provide a mathematical overview and brief explanation of how the algorithm operates. Suppose we start with a dataset $(x, y)_{i=1}^N$, where $x = (x_1, \dots, x_d)$ represents the input variables, and y corresponds to the labels of the target variable. The goal is to approximate an unknown function that maps $x \xrightarrow{f} y$, producing an estimate $\hat{f}(x)$. This is achieved by minimizing a chosen loss function $\psi(y, f)$:

$$\hat{f}(x) = y, \quad \hat{f}(x) = \underset{f(x)}{\operatorname{argmin}} \psi(y, f(x)). \quad (2.1)$$

By expressing the equation in terms of expectation, we arrive at the following formulation:

$$\hat{f}(x) = \underset{f(x)}{\operatorname{argmin}} E_x[E_y(\psi[y, f(x)])|x], \quad (2.2)$$

where the goal is to minimize the expected loss over the target variable $E_y(\psi[y, f(x)])$, conditioned on the input data x . The choice of loss function depends on the nature of the response variable y . For a continuous response ($y \in \mathbb{R}$), common loss functions include the Gaussian L_2 loss or Laplace L_1 loss. If the response variable is categorical ($y \in \{0, 1\}$), a binomial loss function is more appropriate.

The next step is to constrain the function search space to a family of parameterized functions $f(x, \theta)$, transforming the function optimization problem into one of parameter estimation:

$$\hat{f}(x) = f(x, \hat{\theta}), \quad \hat{\theta} = \underset{\theta}{\operatorname{argmin}} E_x[E_y(\psi[y, f(x, \theta)])|x], \quad (2.3)$$

which requires an iterative optimization process to estimate $\hat{\theta}$.

Over M iteration steps, the parameter estimates can be expressed in an incremental form:

$$\hat{\theta} = \sum_{i=1}^M \hat{\theta}_i. \quad (2.4)$$

One algorithm used for parameter estimation is the steepest gradient descent. With N data points $(x, y)_{i=1}^N$, the aim is to reduce the loss function $J(\theta)$ over the observed data:

$$J(\theta) = \sum_{i=1}^N \psi(y_i, f(x_i, \theta)). \quad (2.5)$$

In summary, this process relies on iterative improvements along the gradient direction of the loss function $\nabla J(\theta)$. Given the incremental nature of the method, it is useful to distinguish between different notations. The subscript in $\hat{\theta}_t$ refers to the t -th incremental step in estimating $\hat{\theta}$, while the superscript $\hat{\theta}^t$ denotes the cumulative sum of all increments from step 1 to step t . This optimization strategy can be outlined as follows:

1. Initialize the parameter estimates $\hat{\theta}_0$. For each iteration t , repeat:

2. Calculate $\hat{\theta}^t$:

$$\hat{\theta}^t = \sum_{i=0}^{t-1} \hat{\theta}_i; \quad (2.6)$$

3. Evaluate the gradient of the loss function $J(\theta)$:

$$\nabla J(\theta) = \nabla J(\theta_i) = \left[\frac{\partial J(\theta)}{\partial J(\theta_i)} \right]_{\theta=\hat{\theta}^t}; \quad (2.7)$$

4. Obtain the new $\hat{\theta}_t$:

$$\hat{\theta}_t \leftarrow -\nabla J(\theta). \quad (2.8)$$

In these boosting algorithms, due to the fact that the optimization is held in the function space, it is common to parameterize the function estimate $\hat{f}$ in the additive functional form:

$$\hat{f}(x) = \hat{f}^M(x) = \sum_{i_0}^M \hat{f}_i(x), \quad (2.9)$$

where M is the number of iterations, $\hat{f}_0$ is the initial guess and $\hat{f}_{i=1}^M$ are the function increments, called "boosts".

We now introduce the notation of parameterized "base-learner" functions $h(x, \theta)$ (which can have different natures, such as decision trees or splines) to make distinctions from the ensemble function estimates $\hat{f}(x)$. At the t -th iteration, the optimization rule for the estimate is defined as follows:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t) \quad (2.10)$$

$$(\rho_t, \theta_t) = \underset{\rho, \theta}{\operatorname{argmin}} \sum_{i=1}^N \psi(y_i, \hat{f}_{t-1}) + \rho h(x_i, \theta), \quad (2.11)$$

where ρ is the optimal step size at each iteration.

In order to deal with the difficulties of estimating parameters with arbitrary loss functions and base learners, it was proposed to choose a new function $h(x, \theta_t)$ that is the most parallel to the negative gradient $g_t(x)_{i=1}^N$ along the data:

$$g_t(x) = E_y \left[\frac{\partial \psi(y, f(x))}{\partial f(x)} \middle| x \right]_{f(x)=\hat{f}^{t-1}(x)}. \quad (2.12)$$

We can now just choose the new function increment to be the most correlated with $-g_t(x)$, changing this task to a least-squares minimization one:

$$(\rho_t, \theta_t) = \underset{\rho, \theta}{\operatorname{argmin}} \sum_{i=1}^N [-g_t(x_i) + \rho h(x_i, \theta)]^2. \quad (2.13)$$

All of this notation and steps with further explanation can be found in [40] and in Figure 2.1 there is a simplified summary of Friedman’s Gradient Boosting algorithm.

Algorithm 1 Friedman’s Gradient Boost algorithm

Inputs:

- input data $(x, y)_{i=1}^N$
- number of iterations M
- choice of the loss-function $\Psi(y, f)$
- choice of the base-learner model $h(x, \theta)$

Algorithm:

- 1: initialize $\hat{f}_0$ with a constant
- 2: **for** $t = 1$ to M **do**
- 3: compute the negative gradient $g_t(x)$
- 4: fit a new base-learner function $h(x, \theta_t)$
- 5: find the best gradient descent step-size ρ_t :

$$\rho_t = \arg \min_{\rho} \sum_{i=1}^N \Psi[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)]$$
- 6: update the function estimate:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t)$$
- 7: **end for**

FIGURE 2.1: Summary sketch of Friedman’s gradient boosting algorithm from [40].

During this work, three different gradient-boosting decision tree frameworks were used: CatBoost [42], XGBoost [43] and LightGBM [44]. Although they all share a similar algorithm behind, there are some distinctions between them that are worth exploring because they will impact their usage and implementation.

Differences between XGBoost, CatBoost, and LightGBM

The three algorithms utilized in this work exhibit several structural differences (see [45] or [46]), making it essential to highlight the most significant ones:

1. **Tree Symmetry:** CatBoost constructs symmetric trees, meaning that the splitting criteria are uniform across all nodes at the same depth. This consistency ensures that the splitting condition minimizes loss equally across these nodes. In contrast, LightGBM and XGBoost create asymmetric trees, allowing for different splitting conditions at nodes located at the same depth. The advantage of symmetric tree structures includes quicker computation and evaluation, along with improved control over overfitting.

2. **Tree Growth:** LightGBM employs a leaf-wise growth strategy, while XGBoost and CatBoost utilize a level-wise approach (see Figure 2.2). This means LightGBM selectively

expands the tree, resulting in smaller and more efficient models, albeit at a higher risk of overfitting. This tendency can be mitigated by fine-tuning the model's hyperparameters (refer to subsection 2.4).

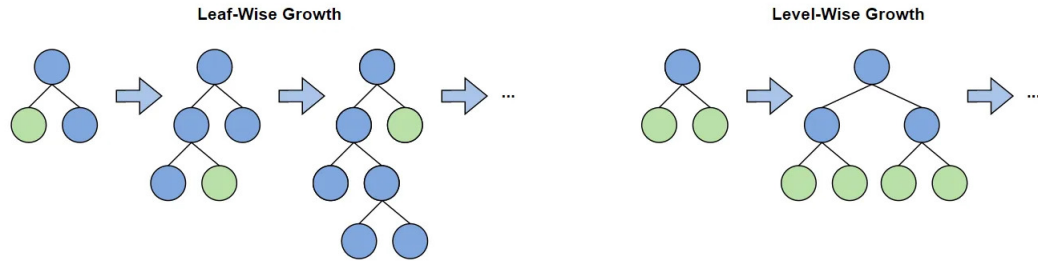


FIGURE 2.2: Sketch of the difference of tree growth strategies: leaf-wise growth (left) and level-wise growth (right). Credit: Kay Jan Wong.

3. Splitting Method: CatBoost uses an algorithm that generates potential feature-split pairs for each leaf node, selecting the split with the least penalty. In comparison, LightGBM implements a technique called Gradient-based One-Side Sampling (GOSS) ([44]), which saves all instances with large gradients and randomly samples from those with small gradients. Instances with larger gradients translate to higher errors and are good indicators of optimal split points. On the other hand, those with smaller gradients contribute to maintaining accuracy in the decision trees. This sampling strategy decreases the number of data instances needed for model training, resulting in quicker training times. Meanwhile, XGBoost employs a pre-sorted, histogram-based method to find the best splits, though it is generally slower than GOSS.

In addition to these three distinctions, there are further differences in how these algorithms manage categorical variables and handle missing values, as well as variations in the processing units employed to execute the algorithms, among other factors.

2.2 Model evaluation

Evaluation metrics are a common tool used to evaluate the performance of a supervised machine learning model ([47]). They can also be used to expand the comprehension of the model's effectiveness and to make quantitative comparisons between different algorithms ([48]). They are also fundamental to optimizing the model hyperparameters (see subsection 2.4). We are interested in evaluating our model's predictive ability, generalization capability, and overall quality and in this subsection, we briefly introduce the

metrics mainly used across this work. For the classification task, where the target variable is discrete, the scores vary between 0 and 1, where 1 is the maximum score.

Precision is the fraction of relevant instances (known as true positives) among the retrieved instances:

$$Precision = \frac{TP}{TP + FP}, \quad (2.14)$$

where TP is the number of true positives in a given class and FP is the number for false positives in a given class.

Recall is the fraction of relevant instances that were retrieved:

$$Recall = \frac{TP}{TP + FN}, \quad (2.15)$$

where FN is the number false negative in a given class.

Given the different nature of these two metrics, it is important to understand how to correctly interpret them and how much importance should be assigned to each. Precision is more important when the cost of false positives is high, i.e. in situations where misclassifying an instance as positive has serious consequences. On the other hand, Recall is more impactful when the cost of false negatives is high, i.e. the consequence of misclassifying an instance as negative is impactful. Due to their different nature and meaning associated, it is useful to combine them both using the F1-Score, which is an harmonic mean of the precision and recall:

$$F1 = 2 \times \frac{Precision \cdot Recall}{Precision + Recall}. \quad (2.16)$$

Additionally, there is also accuracy, which is the fraction of predictions that the model got correctly. Formally it can be defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}, \quad (2.17)$$

and should be used preferably when the number of objects in each class is roughly the same.

All of these four metrics will be calculated for the models obtained, because each one of them attributes different importances to TP, TN, FP and FN and consequently greatest importance should be attributed to the metrics more valuable for the work. In our case,

we are interested in identifying a significant amount of LAE sources while not increasing significantly the number of FP, thus impacting the Recall and Precision, respectively. Therefore, the main metric for the decisions taken from hereon will be F1-Score, which takes both these metrics into consideration.

In order to extract more information about the model performance, a confusion matrix can be used (Figure 2.3). If the model is perfect in their predictions, we could see a confusion matrix purely diagonal, with non-zero elements on the diagonal, and zero elements otherwise, which is rarely achievable.

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

FIGURE 2.3: Example of a binary task confusion matrix.

For regression tasks, where the target variable is continuous, the most common metrics used are the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). Formally, they are defined as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (2.18)$$

where n is the number of objects in the validation or test set, $\hat{y}_i$ is the target variable predicted by the model, and y_i is the target variable provided in the input data. The RMSE, contrary to the mean squared error (MSE), includes the square root, thus contradicting the MSE that has the disadvantage of heavily weighting outliers, since it squares each term in the sum. Additionally, it is also common to use the R^2 , called the coefficient of determination which reflects the portion of the variation in the dependent variable that is predictable from the independent variable. The R^2 is usually defined as:

$$1 - \frac{SS_{red}}{SS_{tot}}, \quad (2.19)$$

where $SS_{red} = \sum_i (y_i - \hat{y}_i)^2$ is the residual sum of squares and $SS_{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares ($\bar{y}$ is the mean of the observed data). An R^2 equal to 1 indicates that the regression predictions perfectly fit the data.

The three metrics presented will all be used when assessing the models performance, but the MAE is the one used to optimized the models hyperparameters (subsection 2.4) and to verify the models robustness to small perturbations (subsection 5.4).

2.3 Data splitting

In machine learning, the input data used to develop the model is typically divided into several datasets, which are used at different stages of the model's creation, specifically during training, validation, and testing ([37]):

- **Training Set:** The initial phase involves training the model on the training set, from where the model and the model parameters are learned. Based on the input data, the model produces an output, which is then compared to the target label for each input. Based on the results of this comparison and the specific learning algorithm employed, the model's parameters are adjusted accordingly.
- **Validation Set:** Once trained, the model is tested on a second dataset known as the validation set. This set provides an impartial evaluation of the model's performance and is crucial for optimizing the model's hyperparameters (see Section 2.4).
- **Test Set:** The primary function of the test set is to provide an unbiased evaluation of the final model, as it remains untouched during the training process. Consequently, this dataset is essential for assessing the model's expected performance on previously unseen data, i.e. a subset of the original data that was not used during the previous stages.

In addition to defining the datasets for this division, it is crucial to determine their sizes and the strategies for partitioning. For all three sets, we opted to implement shuffling and stratification. The stratified split guarantees that all target variables are represented in the same proportions across the different sets. Shuffling disrupts any pre-existing structure within the data, promoting a more unbiased final model. The proportion of data

allocated to each set can also be modified and is dependent on the model's objectives and the characteristics of the data.

A common guideline is to use an 80:20 ratio, meaning that 80% of the data is designated for training (including the validation set) while 20% is reserved for testing. However, there is no definitive guidance on the optimal ratio for a specific dataset. The 80:20 split is justified by the Pareto principle, originally derived from economics but subsequently applied in various domains. In [49], it is suggested that a dataset can be divided in a ratio of $\sqrt{p} : 1$, where p represents the number of parameters to estimate in the "true" model. As we are interested in splitting data in three sets, the suggested division for train, validation and test is $\{(1 - \gamma)^2, \gamma(1 - \gamma), \gamma\}$, where γ is defined as:

$$\gamma = \frac{1}{\sqrt{p} + 1}. \quad (2.20)$$

Considering the number of features used is equal to 53 (see Section 3.3), it implies that $\gamma \approx 0.12$, i.e. 75% for the training set, 10% for validation, and 15% for testing.

For the regression task, however, the dataset used to build the models will be smaller (see Section 5.1), creating a necessity to address this limitation and maximize data utilization. One effective approach to achieve this is by employing k-fold cross-validation for the training and validation sets ([50]). This technique involves randomly dividing the original sample into k equal-sized subsamples, referred to as "folds." From these k subsamples, one is held out as validation data to assess the model (and to optimize hyperparameters), and the remaining $k - 1$ subsamples are used to train. This is repeated k times, so that each subsample serves as validation data. Ultimately, the performance metrics are computed by averaging the results obtained from the k divisions. Additionally, an external test subset is retained, accounting for 15% of the data as previously mentioned. By employing this method, we aim to enhance the utilization of the available data, albeit at the cost of increased computational expense, since training and validating occur k times rather than once. The value of k was established at $k = 3$ throughout the training process, even though it could be increased (typically to 5 or 10). However, doing so would substantially extend the computational time.

2.4 Hyperparameter optimization

In machine learning, hyperparameters are parameters that define various aspects of the learning process. Unlike model parameters, hyperparameters cannot be directly inferred during the fitting process on the training data because the objective function is typically non-differentiable with respect to them. They can significantly influence both the training duration of the algorithm and its overall performance. Therefore, it is essential to identify a set of hyperparameters that results in an optimal model, which minimizes a predefined loss function on specific test data. There are several methods for hyperparameter optimization ([51]), and I will outline three commonly used approaches:

1. **Grid Search:** This method involves an exhaustive search across a manually specified subset of the hyperparameter space for a learning algorithm, defined by certain limits and discrete values. A grid search algorithm relies on a performance metric, which is assessed through cross-validation on the training set or a designated validation set;
2. **Random Search:** Similar to grid search, this method randomly selects a predefined number of combinations from the manually defined hyperparameter space instead of examining all possible combinations as in Grid Search;
3. **Bayesian Optimization:** This technique constructs a probabilistic model that maps hyperparameter values to the objective function evaluated on a validation set. By iteratively testing promising hyperparameter configurations based on the current model and then refining it, Bayesian optimization seeks to gather data that provides valuable insights into this function, particularly regarding the location of the optimum, i.e. the point (set of hyperparameters) that maximizes the performance of the model.

In addition to these methods, there are also more advanced and automated options, such as Hyperopt and Optuna. While grid search is susceptible to the curse of dimensionality, leading to exponential increases in computational time as more hyperparameters are added, its simplicity and reproducibility are significant advantages, especially in this work. Consequently, grid search will be the optimization algorithm employed throughout this work.

This stage of hyperparameter optimization is influenced by the chosen data-splitting method. As discussed in Section 2.3, the data-splitting techniques vary between tasks, resulting in different approaches to hyperparameter optimization. For the classification task, each hyperparameter set is trained on the training set and assessed on the validation set. Finally, an independent test set is utilized to evaluate the performance of the optimal hyperparameter set. This independent assessment is crucial; relying solely on a train-validation split risks overfitting the hyperparameters to the validation set, making the independent test set necessary for a final evaluation of the hyperparameter optimization results. In contrast, for the regression task, the smaller sample size necessitates maximizing its usage. Thus, we opted for 3-fold cross-validation, which increases computational costs since the performance is evaluated across three folds for each hyperparameter set, effectively tripling the running time. However, this approach ensures a more reliable selection of hyperparameters, enhancing the model's suitability when encountering unseen data.

Not all the hyperparameters have the same impact on the performance variation of the algorithms, with tunability being a property that measures how much tuning a certain hyperparameter might have an impact. There is some work in this field, specifically studying the tunability of LightGBM and XGBoost (see, for example, [52] and [53]). Additionally, the documentation pages of the 3 algorithms used have some insight into what are the most impactful hyperparameters for each. Overall, for all the algorithms there are 3 hyperparameters with a highlighted importance:

- **num_estimators**: Specifies the number of boosting iterations (trees). With more iterations, the accuracy of the model increases but it can be at the cost of longer training time and a higher risk of overfitting;
- **learning_rate**: Controls the gradient step size to prevent overfitting, i.e. the rate at which the model weights are updated after working through each batch of training sample;
- **depth**: Controls the maximum depth of each trained tree and setting this value too high can lead to overfitting;
- **subsample**: Specifies the percentage of rows used per iteration to fit each tree, it is used to prevent overfitting.

Additionally, we consider other hyperparameters specifically for each algorithm, starting with LightGBM, the **num_leaves**, which sets the maximum number of leaves each weak learner has, is a parameter that assumes special importance due to LightGBM's trees growing leaf-wise, so the number of leaves is tightly connected with the depth of the tree and so it is necessary to tune the **num_leaves** with the **max_depth**. Connected with the number of leaves, there is also the **min_data_in_leaf**, which controls the minimum amount of data per leaf.

Regarding XGBoost, we also consider the **min_child_weight**, which is an alias of the **min_data_in_leaf** from LightGBM. For XGBoost, the number of leaves does not assume major importance because the trees grow level-wise instead of leaf-wise. Additionally, we also tune the **colsample_by_tree**, which controls the fraction of features that will be used to train each tree and it occurs once for every tree constructed.

Finally, for CatBoost, in addition to the above-mentioned hyperparameters, we consider the **l2_leaf_reg**, which is the coefficient at the L2 regularization term of the cost function and is an effective way of dealing with overfitting.

With this in mind, for the classification task, a first optimization is performed with a broader range of values for the previously mentioned hyperparameters and then a zoom-in is applied on a second round of optimization depending on the first round results. It is important to be mindful of the high degree of computational complexity that hyperparameter optimization requires and that this technique is highly resource-consuming. Having this said, for the regression task, and due to the fact that 3-fold cross-validation is used to tackle the reduced sample size, only a first round of optimization is performed. Overall, we tried to achieve a balance between searching for a reasonable grid and not overextending in computational time. As an example, for the regression task, the optimization time for the LightGBM and XGBoost model was between 15-20h while for CatBoost model it was about 35-40h . The summary of the optimization procedure is shown in Table 2.1, with the default hyperparameters and grids used in the first round of optimization in both tasks for each algorithm, extensively described.

Algorithm	Default hyperparameters	1 st Optimization grid
LightGBM	n_estimators = 100 learning_rate = 0.1 max_depth = -1 subsample = 1 num_leaves = 31 min_data_in_leaf = 20	n_estimators = {10,50,100,500,1000,2000} learning_rate = {0.5,0.25,0.1,0.075,0.05,0.01} max_depth = {-1,3,6,9,12,15} subsample = {0.2,0.4,0.6,0.8,1} num_leaves = {10,31,50,100,200} min_data_in_leaf = {10,20,40,80,100}
XGBoost	n_estimators = 100 learning_rate = 0.3 max_depth = 6 subsample = 1 colsample_by_tree = 1 min_child_weight = 1	n_estimators = {10,50,100,500,1000,2000} learning_rate = {0.5,0.3,0.1,0.075,0.05,0.01} max_depth = {0,2,4,6,8} subsample = {0.2,0.4,0.6,0.8,1} colsample_by_tree = {0.2,0.4,0.6,0.8,1} min_child_weight = {1,10,20,40,80}
CatBoost	n_estimators = 1000 learning_rate = 0.03 max_depth = 6 subsample = 1 l2_leaf_reg = 3	n_estimators = {250,500,750,1000,1250,1500,1750} learning_rate = {0.5,0.3,0.1,0.075,0.05,0.03,0.01} max_depth = {2,4,6,8,10} subsample = {0.2,0.4,0.6,0.8,1} l2_leaf_reg = {0.01,1,3,10,20,30,40,100}

TABLE 2.1: Summary of the grids used in the hyperparameter optimization. The values for max_depth of 0 or -1 result in not limited trees.

Chapter 3

Data Handling and Processing

In this chapter, the data utilized in the MSc thesis will be detailed. It begins with a concise overview of the pertinent observational surveys, followed by an explanation of the data analysis process. Additionally, this chapter outlines the extraction and creation of the final data to be used to train the models, along with other technical specifics. A subsection dedicated to data characterization is included, where the distribution of photometric information and various physical properties are examined to enhance the understanding of the data employed.

3.1 Observational surveys

3.1.1 COSMOS

The Cosmic Evolution Survey (COSMOS) [54] is an astronomical investigation aimed at exploring the formation and evolution of galaxies across different cosmic epochs (redshift) and within various local environments. Spanning a 2 square degree equatorial region, the survey is centered at the coordinates (J2000): RA $+150.11916667^\circ$ ($10^h00^m28.600^s$) and DEC $+2.20583333^\circ$ ($+02^\circ12'21.00''$). It utilizes spectroscopy and imaging across X-ray to radio wavelengths, with observations from major space-based telescopes such as Hubble, Spitzer, GALEX, XMM, Chandra, Herschel, and NuStar, in addition to several large ground-based telescopes including Keck, Subaru, the Very Large Array (VLA), the European Southern Observatory Very Large Telescope (ESO-VLT), and the Canada-France-Hawaii Telescope (CFHT), successfully detecting over 2 million galaxies.

In this study, we utilize the most recent photometric catalog from COSMOS, known as COSMOS2020 [55]. This catalog includes measurements of 1 million sources across more

than 30 photometric bands, spanning from the ultraviolet (UV) to infrared wavelengths. Additionally, it provides photometric redshifts and various physical properties derived using the spectral energy distribution (SED) fitting codes LePhare ([56]; [57]) and Eazy ([58]). A summary of the relative transmission curves for all photometric bands available in COSMOS2020 is presented in Figure 3.1.

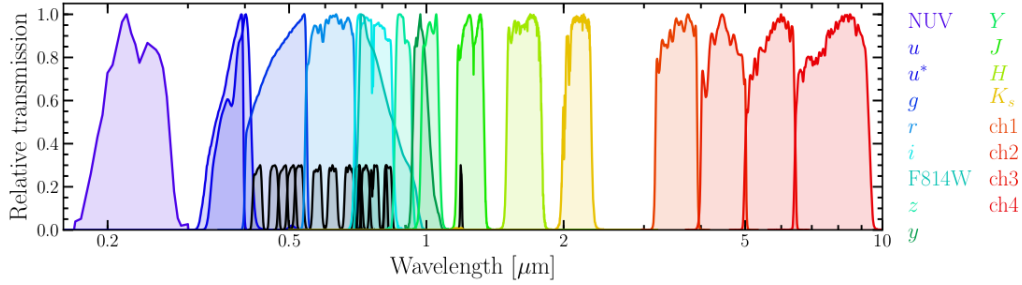


FIGURE 3.1: Relative transmission curves for the photometric bands present in COSMOS2020, where the peaks are normalized to 1 for the broad-bands and to 0.3 for the medium/narrow-bands (represented by black lines). This figure was taken from [55].

3.1.2 SC4K

The SC4K survey [1] is conducted within the COSMOS field and is derived from extensive wide-field observations made with the Subaru and Isaac Newton telescopes. It utilizes 16 (12 medium bands + 4 narrow bands) over an area of 2 deg^2 , encompassing a comoving volume of approximately 10^8 Mpc^3 . This survey has produced a substantial sample of 3,908 sources identified for their elevated $\text{Ly}\alpha$ equivalent width (EW) within the redshift range of approximately $2 < z < 6$. The selection criteria employed include: i) an EW cut of $50 \text{ \AA}$ for medium bands, $25 \text{ \AA}$ for narrow bands, and $5 \text{ \AA}$ for the narrow band at $z = 2.23$; ii) a significant excess of emission in the selected medium or narrow band; iii) a color break occurring blueward of the detected $\text{Ly}\alpha$ emission, indicative of a Lyman Break; iv) exclusion of sources with pronounced red colors, which often signify lower-redshift contaminants where the Balmer break can mimic a Lyman break; and v) a visual inspection of all candidates to eliminate spurious sources and artifacts from stars. For comprehensive details on the sample selection process, please refer to [1]. Figure 3.2 illustrates the “slicing” of the COSMOS field at various redshifts, which is the origin of this sample’s name.

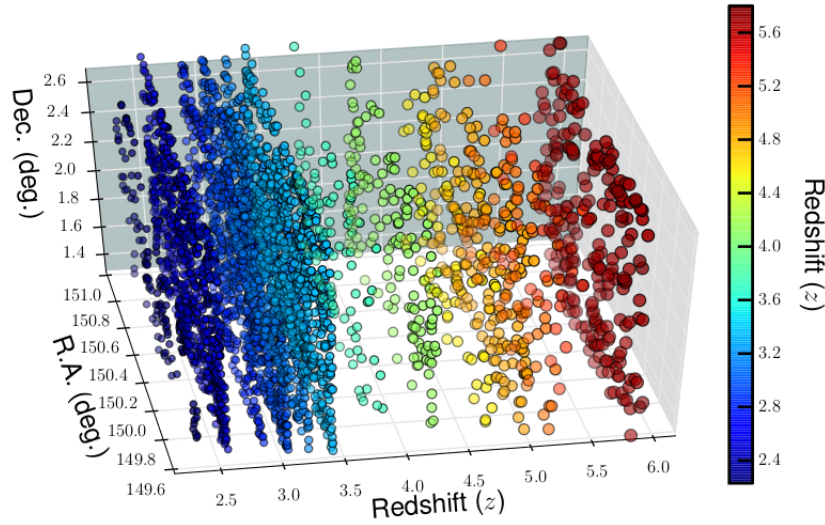


FIGURE 3.2: The 3D distribution of the SC4K sample in the COSMOS field, where the LAEs are divided into 16 different redshift slices with $z \in [2, 6]$. This figure was taken from [1].

3.2 Data preparation

The first step towards building a machine learning model capable of identifying and characterizing objects is to prepare the data to align with this project’s objectives and the algorithms that will be employed for data modeling. This preparatory phase starts with data cleaning, initially focusing on the COSMOS sample, which covers a redshift range from approximately $z \approx 0$ to 10, while the SC4K redshifts span from $z \approx 2$ to 6. Therefore, our first task is to restrict the COSMOS2020 sample to match the same redshift range as SC4K, in order to have comparable samples within the same redshift ranges.

Furthermore, we exclude the SC4K sources from COSMOS2020 during this data pre-processing phase, because we want to create two distinct samples, one with LAE sources and the other with “non-LAE” sources, i.e. sources present in COSMOS2020 but not in SC4K. Following these steps, COSMOS2020 and SC4K end up with 196713 and 3346 sources, respectively. It is crucial to note that we crossmatched SC4K with COSMOS2020 to integrate magnitudes and fluxes information, essential for the machine learning models (see Section 3.3). However, this process resulted in a significant reduction in the number of sources. Hereafter, these altered COSMOS2020 and SC4K datasets will be referred to as non-LAE and LAE samples, respectively.

Before proceeding with further data preparation, it is prudent to compare the number of sources in each redshift bin for both samples (Figure 3.3). Additionally, a comparison

of the distributions in the i-band magnitude is warranted (Figure 3.4), due to the fact that at $z \in [2, 6]$, the i-band captures the rest-frame ultraviolet to visible light continuum emission of the galaxy, and consequently is useful to make estimations of the star formation rate.

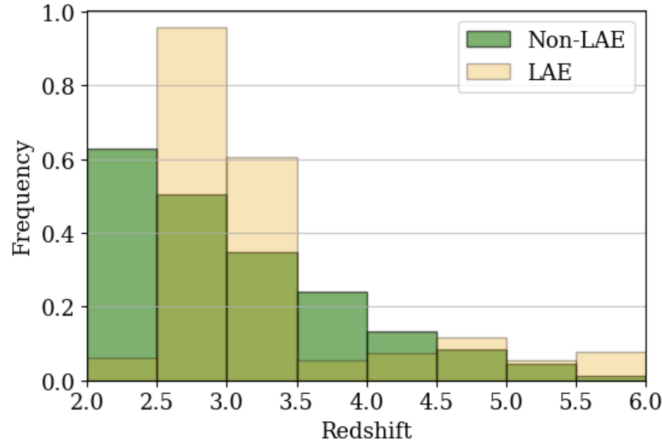


FIGURE 3.3: Non-LAE and LAE number of sources distribution per redshift bin.

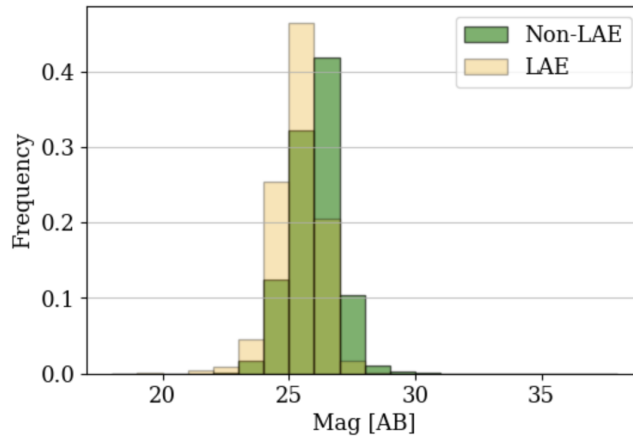


FIGURE 3.4: Non-LAE and LAE samples i-band magnitude distribution.

Our objective is to construct a comparable training sample with LAEs and non-LAEs sources, within which the algorithms can accurately distinguish and identify LAEs. Consequently, it is crucial for the training sample to have similar redshift and i-band magnitude distributions between LAEs and non-LAEs. This similarity allows the algorithms to highlight and utilize more subtle differences between the galaxies' natures for accurate classification. Furthermore, the notable difference in sample sizes between non-LAEs and LAEs introduces another potential challenge, introducing more imbalances in the training sample, that could, in our example, bias the classification model towards the "non-LAE" class, due to the fact that it contains significantly more sources than the LAE sample. This

underscores the necessity for a subsequent processing step: how to attain balanced data that is consistent across redshift and i-band magnitude. To address this issue, the solution entails creating equal bins for these two properties and selecting an equivalent number of sources from both samples for each bin.

Another critical consideration pertains to the uncertainty surrounding the non-LAE sample. While the LAE sample comprises sources obtained through narrow-band surveys, the same level of certainty does not extend to the non-LAE sample nature. We operate under the assumption that sources not detected by SC4K are not LAEs and proceed to train our models based on this assumption. To address this difference in uncertainty between the two samples, instead of relying on a single sample for training and testing, we will use as many samples as possible, ensuring the following conditions are met:

- There is an equal number of sources from each nature;
- The non-LAE sources will be drawn from the non-LAE sample and the LAE sources are essentially the LAE sample previously created;
- The non-LAE sources follow the same i-band magnitude and redshift distribution as the LAE sources.

We are able to achieve a maximum of 5 subsets following the above-mentioned criteria. This approach aims to address the uncertainty surrounding the non-LAE sample and the need for balanced data.

3.3 Feature selection

In machine learning, a feature represents a measurable property or characteristic of a phenomenon, serving as the input for the model to make predictions. Understanding the significance of feature selection is crucial, as it directly impacts the model's performance. A fundamental principle guiding feature selection is Occam's razor, which advocates for constructing explanations with the simplest set of elements possible. In our context, this translates into choosing a robust set of features devoid of unnecessary complexity. The features we have selected primarily consist of broad-band photometric properties, including fluxes, magnitudes, and colors in the optical and near-infrared (NIR) spectra. This choice aims to minimize errors (by choosing a robust number of features) while maintaining simplicity in our inputs, i.e. not using derived physical parameters. Opting for

optical and NIR properties aligns with the overarching goal of creating a model applicable to other large surveys limited to these wavelengths and also due to the fact that the Lyman- α line can be detected in the optical and NIR at $2 < z < 6$. In summary, the set of fluxes and magnitudes used is stated in Table 3.1, and the main properties of the broad-bands associated with these features are summarized in Table 3.2.

Fluxes and Magnitudes
MegaCam/CFHT u*-band
HyperSuprime-Cam g-band
HyperSuprime-Cam r-band
HyperSuprime-Cam i-band
HyperSuprime-Cam z-band
UltraVISTA Y-band
UltraVISTA J-band
UltraVISTA H-band
UltraVISTA K _s -band

TABLE 3.1: Fluxes and Magnitudes used to train the models. For both fluxes and magnitudes, 3 arcsec diameter aperture was used.

Band	Central λ [Å]	Width [Å]	Depth ($3''$) ± 0.1	Error Fact. ($3''$) ± 0.1
u*	3858	598	27.1	1.6
g	4847	1383	27.5	1.8
r	6219	1547	27.3	1.7
i	7699	1471	27.0	1.9
z	8894	766	26.6	1.7
Y	10216	923	26.1	3.1
J	12525	1718	25.9	2.9
H	16466	2905	25.5	2.9
K _s	21557	3074	25.2	2.6

TABLE 3.2: Information of the broad-bands from the features present in 3.1 used to train the models. The central λ is the median of the transmission curve. The width represents the full width of the transmission curve at half maximum. The depth is the 3σ depth computed PSF-homogenized images in empty apertures, and the error fact. is the multiplicative correction factor for photometric flux uncertainties. The information present on this table and caption was taken from [55].

The use of the u*-band instead of the u-band is due to the fact that the u-band does not cover the entire COSMOS field. In the optical, we use the Hyper Suprime-Cam (HSC) data ([59]) instead of the Subaru Suprime-Cam data ([60]) because the HSC data contains deeper broad-band photometric measurements.

The addition of magnitudes to the features list is motivated by the fact that, even though magnitudes are the conversion of fluxes, they are in the form of normalized values, which benefits the algorithms. Besides the features in Table 3.1, the colors created by subtracting the magnitudes are also used as features (important to note that we exclude

redundant colors, i.e. assuming that 'X-Y' is the same as 'Y-X'). Using fluxes, magnitudes and colors was the combination that obtained the best results, when compared to removing any of these features from the feature set.

3.4 Data characterization

One crucial step before delving into machine learning techniques is to characterize our data, examining its features, distribution, and patterns. This characterization is essential for identifying potential data quality issues, including missing values or outliers, and devising strategies to address them. Moreover, it enables us to pinpoint any biases within the data and ensure that our models do not perpetuate or exacerbate them.

With this in mind, we begin by examining the distribution of the magnitudes discussed in the preceding section, along with key statistics such as maximum/minimum values, median, standard deviation, and the count of null values. These statistics are presented separately for both the LAE and non-LAE samples in Figure 3.5.

From Figure 3.5, we see that in the optical bands, specially in the g- and r-bands the LAE sample has brighter sources than the non-LAE samples and, except in the GRIZ bands, there is a significant amount of NaNs values on both samples. However, we infer that their distribution is not uniform, it is dependent on the object's physical characteristics, thus containing valuable physical information about the sources. Additionally, the gradient boosting algorithms that we use in this work handle missing values effectively.

Other relevant distributions to examine are the Star Formation Rate (SFR) and Stellar Mass distributions and compare these between both samples (Figure 3.6).

The samples not containing LAEs exhibit a higher prevalence of galaxies with greater stellar mass and SFR compared to the LAE sample. There is a known correlation between the stellar mass of a galaxy and its SFR in typical galaxies, forming a pattern referred to as the 'Main Sequence' of Star-Forming Galaxies (see, for example, [61] or [62]). Given this, the observed similarities between the two distributions found in Figure 3.6 can be attributed to the SFR-Stellar Mass ($SFR - M_*$) relationship, which itself is also interesting to compute because it can reveal if there are significant differences in the star formation activity between these populations and across different redshift intervals (Figure 3.7). LAEs are typically above the Main Sequence relation, specially for low stellar masses ($M_* < 9.5$), but above this threshold, LAEs tend to follow the Main Sequence line or be below it.

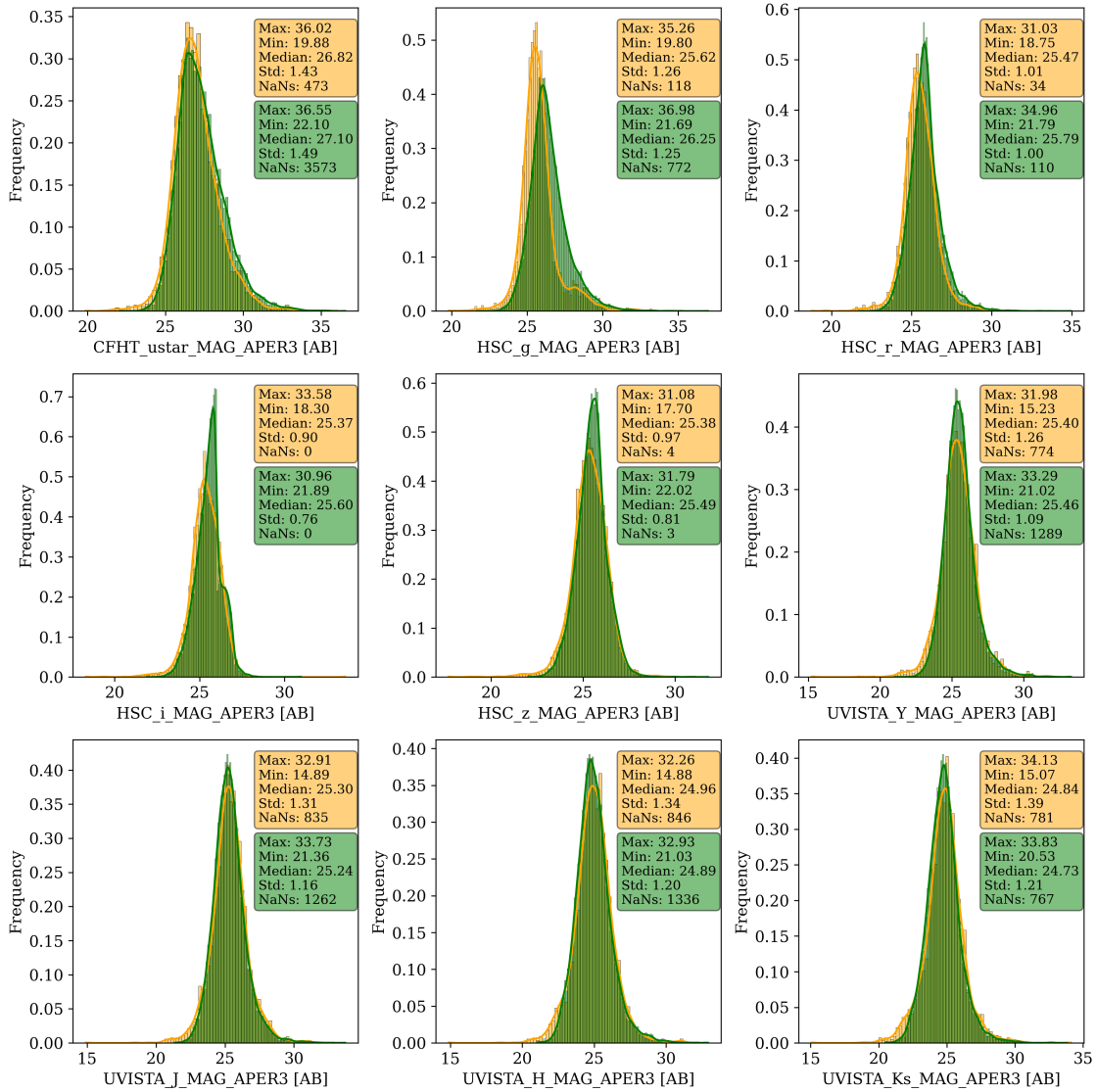


FIGURE 3.5: Non-LAE and LAE samples magnitudes distribution and statistics (maximum, minimum, median, standard deviation and number of NaNs values). The LAE and non-LAE samples are colored orange and green, respectively.

Galaxies undergoing periods of intensive star formation, potentially due to bursty star formation events, are typically found above the Main Sequence. In the non-LAE and LAE samples, there is a noticeable trend of intense star formation in low-mass galaxies across all redshifts and we can also see that the SFR- M_* relation is similar for both samples.

In [64] it is demonstrated that quiescent and star-forming galaxies can be differentiated using the rest-frame $(U - V)$ versus $(V - J)$ color-color diagram, commonly known as the UVJ diagram. Similarly, in [65] it is utilized the $(NUV - r)$ versus $(r - K)$ color-color diagram, referred to as the $NUVrK$ diagram, to enhance the distinction between current star formation activity and dust reddening. NUV light is sensitive to the presence of

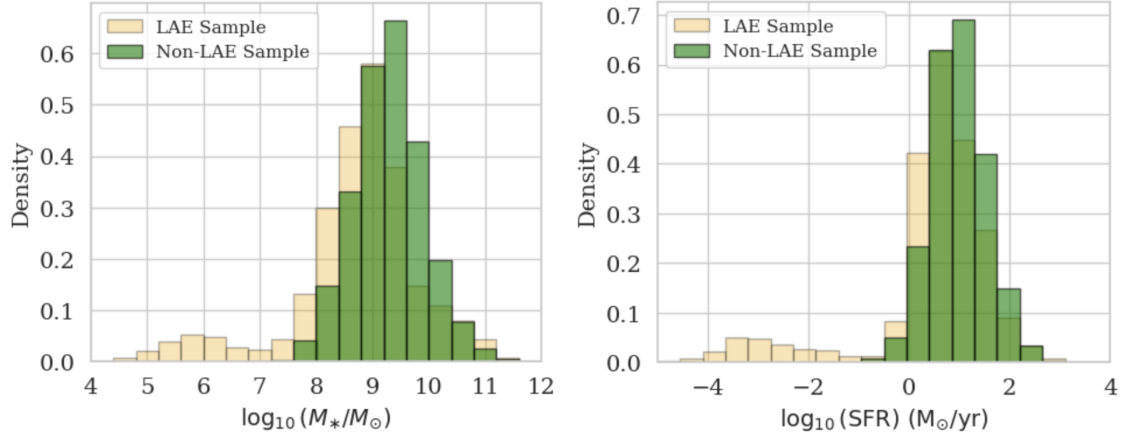


FIGURE 3.6: Stellar Mass (left) and SFR (right) distributions for LAEs and non-LAEs samples (colored orange and green, respectively).

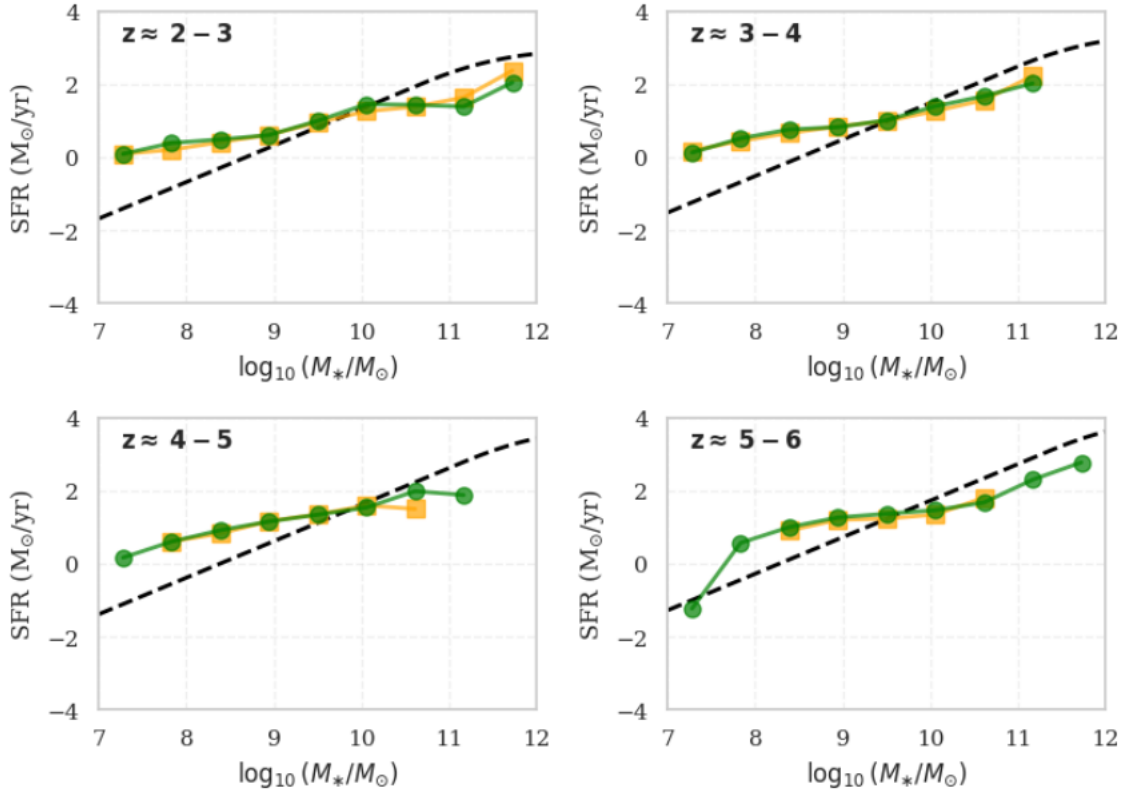


FIGURE 3.7: Mass-SFR relation for both samples at different redshift ranges. The continuous black lines in the last figure are the best-fit relations from [63] computed for the average redshift in each interval. The relations for the LAEs and non-LAEs samples are orange and green, respectively.

hot, massive stars, which are indicative of recent star formation activity in galaxies and the K-band is sensitive to the older stellar population in galaxies and is less affected by dust extinction compared to optical wavelengths while the r-band provides information about the overall stellar content and morphology of galaxies. Therefore, this method has been proven effective in separating passive and star-forming galaxies and consequently, we have constructed these $NUVrK$ diagrams for both the LAE and non-LAE samples as shown in Figure 3.8.

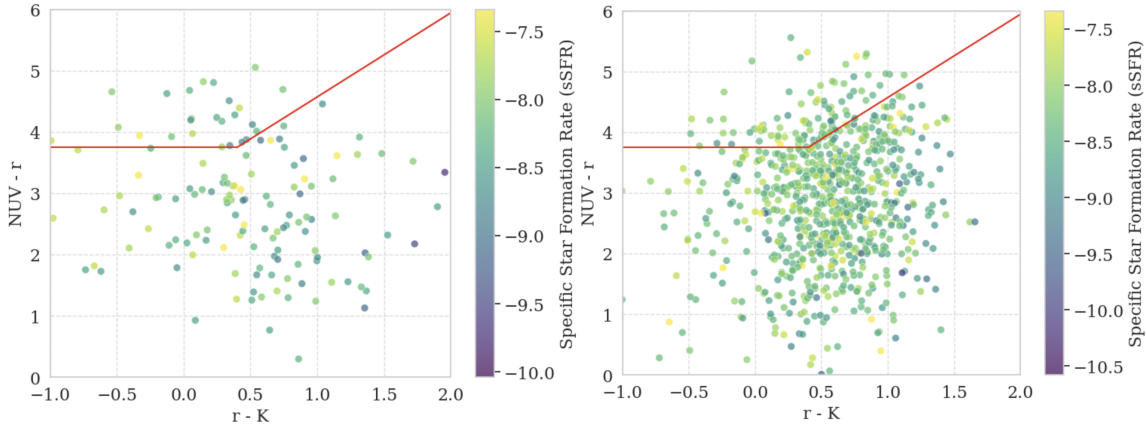


FIGURE 3.8: $NUVrK$ color-color diagrams for the LAE (left) and non-LAE (right) samples. The red line shows the separation between active and quiescent galaxies as in [65].

Another approach to comprehensively understanding data structure is by employing a correlation matrix, which visually represents and examines the degree of correlation between features and the label, specifically whether they belong to the class of LAEs or not. This matrix reveals whether a feature is positively or negatively correlated with the target variable, or if there is no significant relationship at all. Such insights help assess the potential importance of each feature in predicting the nature of the sources. However, as the number of features increases, this visualization can become overwhelming. Therefore, to maintain clarity, we limit the display to the top 10 most correlated features in Table 3.3.

The construction of this table was made by calculating the Pearson correlation coefficient between each pair of features.

Features	Correlation Coefficient
g-K	0.198
HSC_g_MAG_APER3	0.189
g-z	0.188
g-H	0.187
g-J	0.185
g-r	0.181
g-i	0.174
g-Y	0.152
r-K	0.143
HSC_g_FLUX_APER3	0.125

TABLE 3.3: Correlation coefficients for the 10 most correlated features with the target label.

The most correlated feature with the target label is the g- band, which is reasonable since it can be related to the Lyman break. The colors involving the g- band should also have major importance due to the star-forming nature of our sources, i.e. their stellar population should be young in the majority and so it is not expected significant emission detected by the redder bands, resulting in a noticeable slope between the g- band and these.

Interestingly, despite the fact that fluxes and magnitudes encode essentially the same information, it is possible to observe that, for example for the g-band, they do not yield the same correlation coefficient, signalling towards the possibility that feature transformation plays an important role in the models performance. Even though the feature information is the same between fluxes and magnitudes, the way it is provided to the algorithms can have significant consequences.

Chapter 4

Identifying LAEs using photometric data

Due to the exponential increase in astronomical data resulting from large surveys such as Euclid and LSST, there is a growing necessity for efficient techniques to select potential candidates for interesting celestial objects. This necessity serves as the primary motivation for our adoption of a machine learning approach to pre-select LAEs candidates utilizing photometric data. Specifically, we consider broad-band fluxes, magnitudes, and colors across the optical to near-infrared (NIR) wavelength range, leveraging samples derived from COSMOS2020 and SC4K as outlined in Chapter 3. In this chapter, we present the obtained results and engage in a comprehensive discussion, which includes an examination of the robustness of our models to noise and the potential implications of their application to other surveys.

In addition to the Python packages mentioned in Chapter 3, for the classification task, the main machine learning packages used are LightGBM (<https://lightgbm.readthedocs.io>; [44]), XGBoost (<https://xgboost.readthedocs.io>; [43]), and CatBoost (<https://catboost.ai>; [42]).

4.1 Performance metrics and feature importance analysis

The classification evaluation metrics computed for each model, without considering the hyperparameter optimization, are presented in Table 4.1 by averaging the results of the 5 samples.

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM	0.984	0.984	0.993	0.975	0.862	0.864	0.879	0.846	0.873	0.875	0.883	0.864
XGB	1	1	1	1	0.851	0.853	0.861	0.841	0.867	0.835	0.878	0.855
CatBoost	0.960	0.960	0.974	0.946	0.867	0.869	0.878	0.856	0.877	0.879	0.889	0.867
RandomForest	1	1	1	1	0.844	0.848	0.867	0.826	0.853	0.856	0.873	0.834

TABLE 4.1: Average train, validation, and test sets classification evaluation metrics for LightGBM, CatBoost, and XGBoost with default settings.

Overall, CatBoost performed better than the other algorithms, the difference is not that significant compared with LightGBM but XGBoost slightly underperformed in comparison. The three algorithms’ results have in common the noticeable difference in recall when compared to the other metrics and it can translate into a certain conservatism when searching for LAEs, inducing a high number of false negatives which is not ideal as these models are being built considering a possible application to large surveys where the main objective is to pre-select a significant amount of LAEs.

Additionally, we have done an experimental run using the algorithm RandomForest ([39]), which applies the general technique of bagging instead of boosting, introduced in Section 2.1. The performance metrics of the models using RandomForest are worse than the three main boosting algorithms and thus we opt not to perform hyperparameter optimization on it and proceed only with LightGBM, XGBoost, and CatBoost.

As explained in Section 2.4, we perform hyperparameter optimization for all the models using LightGBM, XGBoost, and CatBoost. Having this said, the metrics when running the algorithms with the best hyperparameters set found are also summarized in Table 4.2 to make a comparison with the default settings (Table 4.1), assuming a priori that our performance metrics are, in general terms, robust and it will be difficult to make drastic improvements without risking overfitting.

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM	0.976	0.977	0.986	0.966	0.879	0.881	0.890	0.869	0.875	0.877	0.890	0.861
XGB	0.980	0.980	0.986	0.973	0.882	0.883	0.893	0.871	0.867	0.869	0.882	0.852
CatBoost	0.997	0.997	0.999	0.996	0.883	0.884	0.894	0.872	0.869	0.870	0.878	0.860

TABLE 4.2: Average train, validation, and test set classification evaluation metrics for LightGBM, CatBoost, and XGBoost with hyperparameter optimization.

There is a significant improvement when comparing the validation test scores of the algorithms with default hyperparameter values (Table 4.1) and with the optimization results (Table 4.2), which translates into improvements of 2 – 5% for the validation set, but

these differences become marginal for the test set. It is possible to verify some stability in the results of the test set, meaning that even if the increase in metrics value was not significant, there was also no reduction that could occur due to some overfitting to the validation as a consequence of the extensive grid search performed. The small improvements with hyperparameter optimization are consistent with other works (see, for example, [66] and [67]), which is maybe pointing towards the fact that the default algorithms are already near the maximum performance achievable for these algorithms at this task. From hereon, the results presented for the classification task are obtained considering the algorithms with the optimal hyperparameters found, unless otherwise is stated.

To complement this analysis, the confusion matrices for each algorithm are computed, which also helps to understand the difficulties faced by the algorithms in labeling the sources. The confusion matrices of the five samples for each algorithm were averaged and the results are shown in percentage scale to allow comparisons between the train and test samples. This is summarized in Figures 4.1, 4.2, and 4.3.

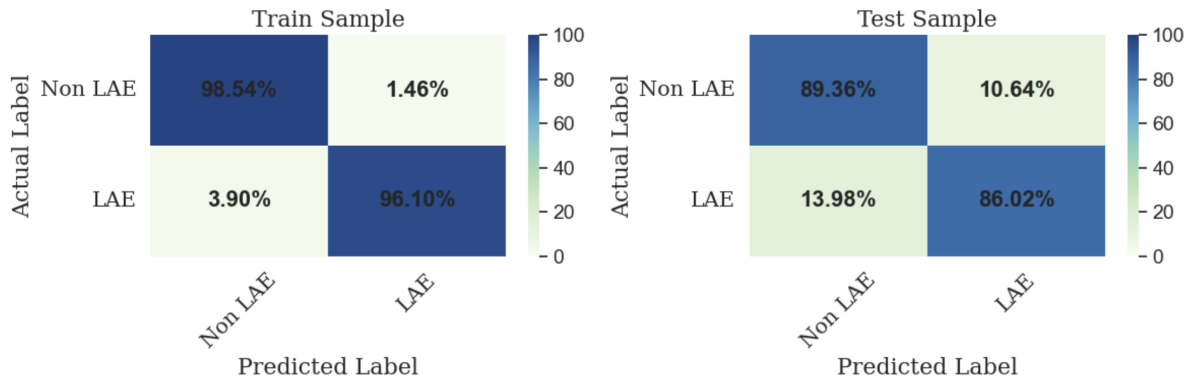


FIGURE 4.1: Average of the 5 samples confusion matrices of LightGBM for the train (left) and test (right) samples.

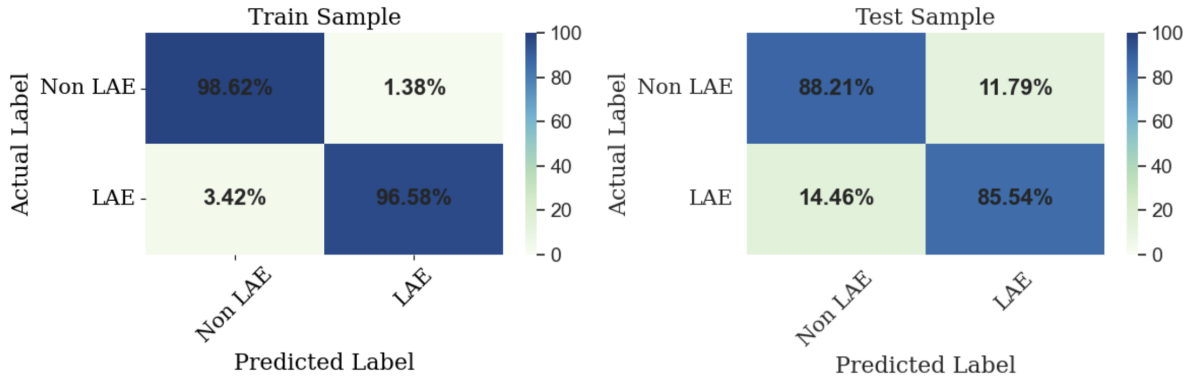


FIGURE 4.2: Average of the 5 samples confusion matrixes of XGBoost for the train (left) and test (right) samples.

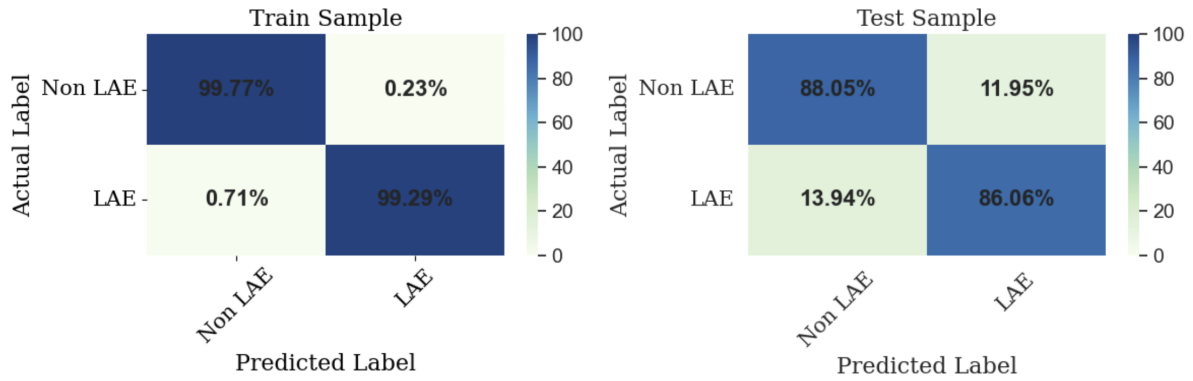


FIGURE 4.3: Average of the 5 samples confusion matrixes of CatBoost for the train (left) and test (right) samples.

The results are relatively consistent across the three algorithms, with both training and testing samples exhibiting more false negatives than false positives. This highlights again the lower recall score when compared to other metrics and the difficulty in discerning subtle differences between LAEs and non-LAEs with the information given. Prioritizing certainty in identifying LAEs often leads to fewer predicted LAEs overall, but with higher confidence in their classification, i.e. a higher percentage of true positives and lower percentage of false positives.

The process by which algorithms assign importance to each feature is termed “feature importance”, allowing the user to obtain insights about the most important features for the models to make the decisions. It is also relevant to compare this with the pre-results correlation analysis summarized in Table 3.3. The importance scale differs depending on the algorithm. For LightGBM and XGBoost a split-based method is used, which attributes the feature importance based on the number of times a feature is used to split the nodes across the trees that constitute the model. Alternatively, another feature importance method called gain can be used, which reflects the difference in error before and after splitting the data on a particular feature. On the other hand, both these methods are not available for CatBoost. The default method for CatBoost is to calculate the feature importance based on how much on average the prediction changes if the feature value changes, so this is the method used.

Figures 4.4, 4.5, and 4.6 display the ten most important features, averaging the five samples, for the LightGBM, XGBoost, and Catboost models, respectively. Notably, the g -band emerges as highly important across all three algorithms. Moreover, there is significant importance attributed to color features involving this band, such as $g - r$ on top of the three lists, but also $g - u$, $g - i$, and $g - z$. Additionally, the relevance of $r - i$ and $i - z$

colors also underscores the overall importance of *griz* broad-bands. An intriguing observation is that, unlike LightGBM and XGBoost, CatBoost assigns greater importance only to fluxes instead of magnitudes, even though the feature importance assignment method differs between the algorithms, raising questions about whether direct comparisons can be made. But this allocation of more importance to fluxes, in the CatBoost case, can be highlighting structural algorithmical differences between the frameworks used, which affects their ability to assimilate the information, resulting in significantly different feature importance rankings.

Overall, comparing these findings with Table 3.3, wherein correlation coefficients of features with the target label were computed, reveals an expected alignment in the importance of the *g* band and its associated colors.

Additionally, Appendix A provides all plots illustrating feature importances separated by sample and algorithm. One comment relevant is that for the XGBoost feature importance results, there are significant differences across the samples, which is not as noticeable for LightGBM and CatBoost. Therefore, the averaged figures should be analyzed together with the feature importances separated by sample.

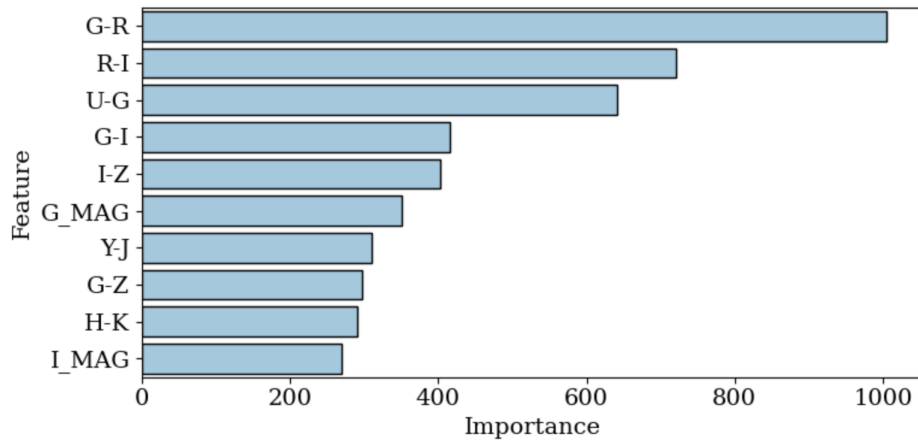


FIGURE 4.4: Average of the feature importances of the 10 most relevant features for LightGBM.

Besides the predicted class, the models also provide the prediction probability, ranging from 0 to 1, which indicates the confidence the model has in its label for each input data point. Analyzing the distribution of prediction probabilities for each algorithm reveals how confident they are regarding their predictions. We expect to observe a distribution that follows a ‘U-shape’, as this shape is expected in a binary classification problem. These distributions are plotted in Figures 4.7, 4.8, and 4.9, separated for each algorithm.

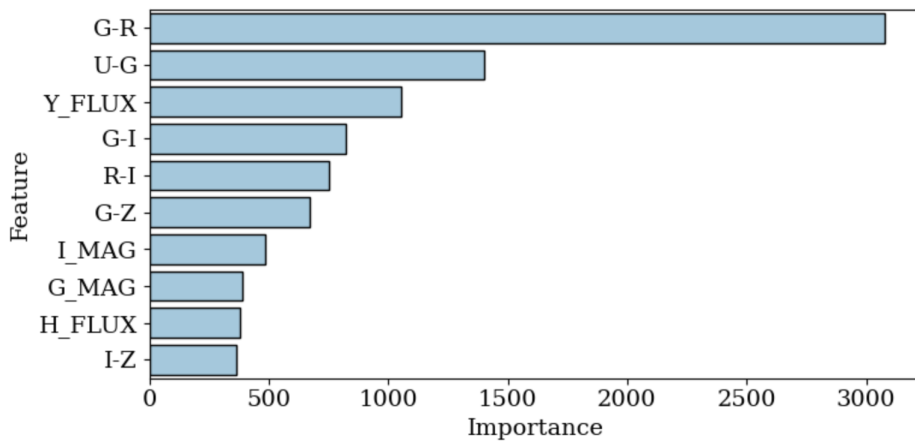


FIGURE 4.5: Average of the feature importances of the 10 most relevant features for XGBoost.

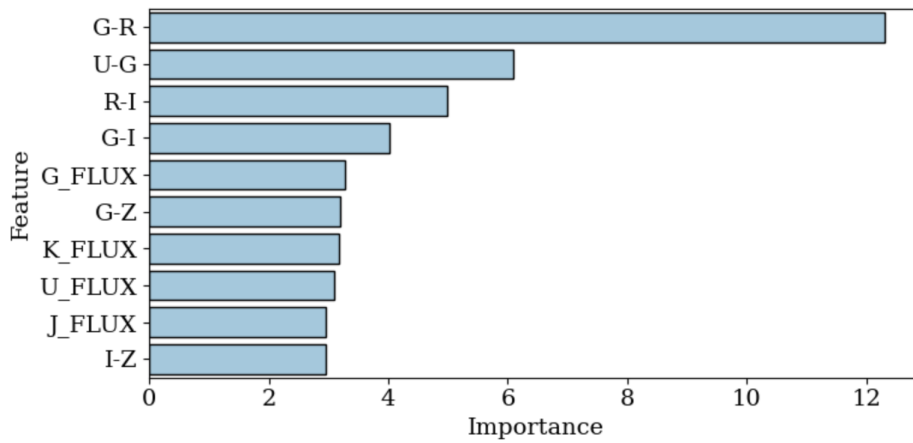


FIGURE 4.6: Average of the feature importances of the 10 most relevant features for CatBoost.

In all three figures, it can be observed the expected 'U-shape', with it being even more pronounced for XGBoost and CatBoost. Additionally, there are no significant differences across the samples for each algorithm.

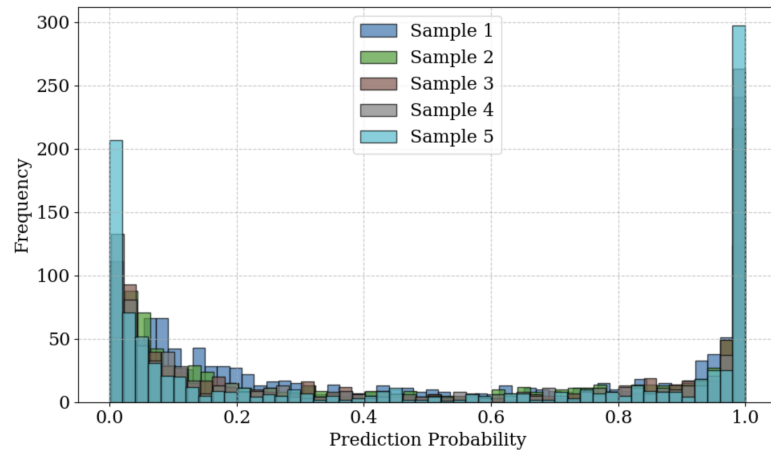


FIGURE 4.7: Distribution of the prediction probability in the test samples for LightGBM.

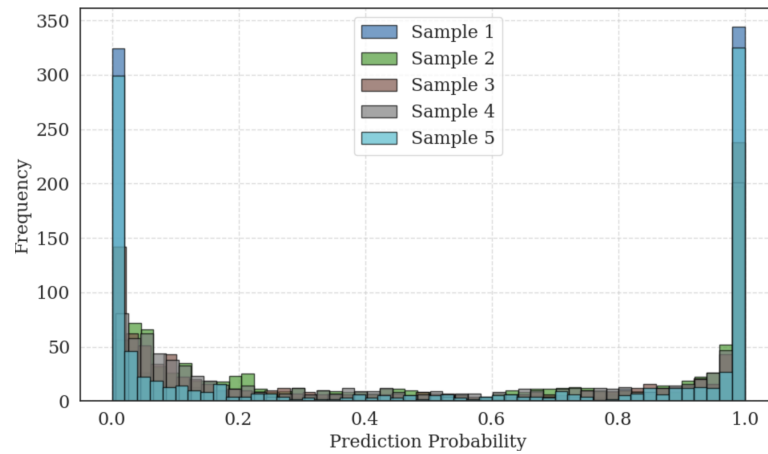


FIGURE 4.8: Distribution of the prediction probability in the test samples for XGBoost.

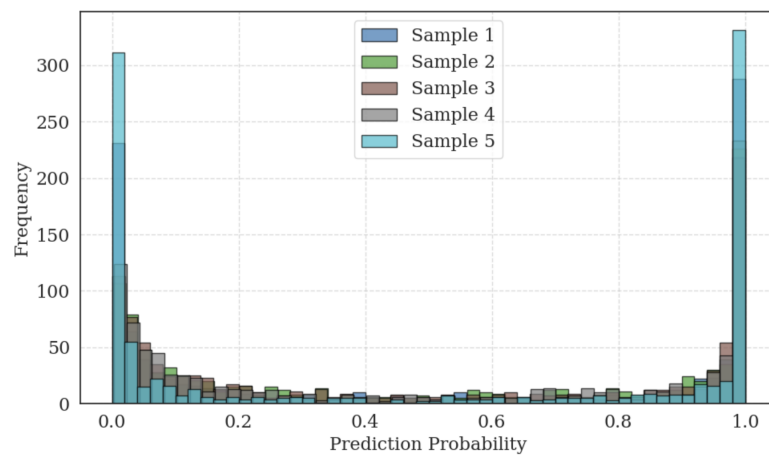


FIGURE 4.9: Distribution of the prediction probability in the test samples for CatBoost.

4.2 Predicting LAE candidates in COSMOS2020 at $z \approx 2 - 6$

The first application for the models created in the previous section will be to predict in the COSMOS2020 dataset with $z \in [2, 6]$, excluding the 5 samples used to construct the models, resulting in 179,469 sources amongst which the models will search for LAEs. With this exercise, we aim to run the models on data not controlled, i.e., not with fixed magnitudes and redshift distributions while still maintaining the redshift range between 2 and 6 due to our a priori knowledge that redshift highly impacts the photometric properties and we should not use the models outside their training boundaries.

In Figure 4.10, the number of sources predicted simultaneously x times as being LAE is shown (where x can vary between 0 and 15). The largest amount of sources (slightly more than 70%) are not predicted as LAEs by any of the models and only less than 4% are predicted by all the 15 models in agreement, resulting in 6667 sources predicted. Confirming the LAE nature of these sources predicted 15 times would more than double the SC4K sample. This can be thought as possible due to the fact that NB/IB only cover limited redshift ranges, thus they risk losing many sources between the “gaps” at $z \in [2, 6]$, that can now be identified.

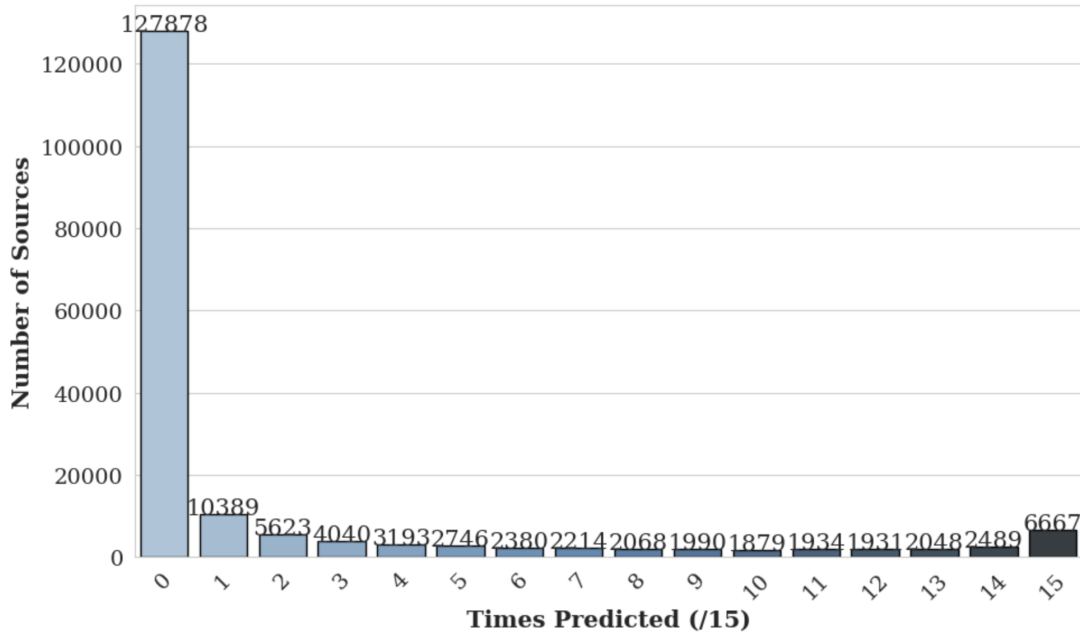


FIGURE 4.10: Distribution of the number of times a source is predicted as an LAE by the 15 models.

The average prediction probabilities for each source were also computed, which is shown in Figure 4.11. The largest amount of sources have prediction probability between

0 and 0.1, and afterward, there is a slow decrease until probability 1. This raises the question of where to draw the threshold in average prediction probability that separates labeling a source as an LAE predicted or not, which will be addressed in the next subsection.

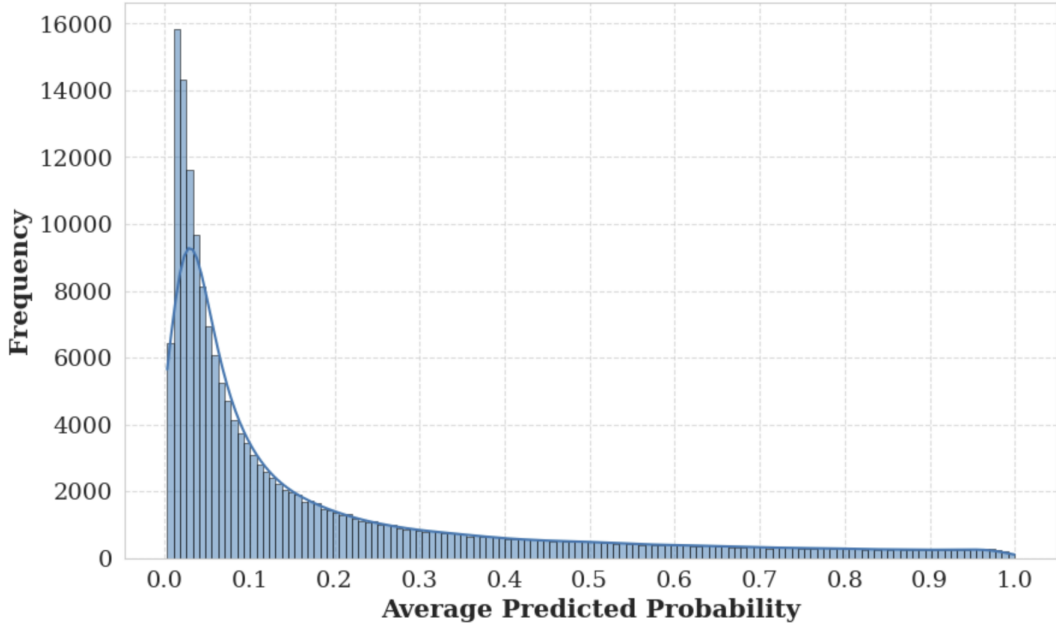


FIGURE 4.11: Distribution of the average prediction probabilities by the 15 models for each source.

4.3 Crossmatching predicted LAEs with spectroscopic catalogs

After predicting inside the COSMOS field with the models built in the previous section, it is important to validate a portion of these predictions with publicly available spectroscopic catalogs available in this region. This confirmation might give us better insights into the decisions made by the models and confirm some sources' nature, ensuring that the model is actually making valid decisions. The procedure followed is to crossmatch the predictions from the previous section with these spectroscopic surveys and analyze the results.

4.3.1 HETDEX

The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) represents a groundbreaking initiative in the quest to understand dark energy's role in the early Universe.

Operating through the Hobby-Eberly Telescope located at McDonald Observatory, alongside advanced spectrographs, HETDEX endeavors to chart the three-dimensional coordinates of over one million galaxies, recording their redshifts. As of the present, the comprehensive HETDEX Source Catalog 1 encompasses more than 220000 sources, prominently featuring over 50000 Lyman- α Emitters, which serve as crucial indicators of cosmic evolution. However, it is noteworthy that HETDEX extends its reach beyond the COSMOS field, encompassing other regions of the sky such as the Extended Groth Strip and GOODS-N.

Please see the survey description [68] and [69] for an extensive explanation of the survey and sample selection/classification methods.

Upon crossmatching this catalog with our predictions in the COSMOS field with the TOPCAT ([70]) software, based on the comparison of the right ascension and declination, and employing the SKY algorithm with a maximum error tolerance of 1 arcsecond for matching, we identify a total of 38 matches. Among these matches, 33 are labeled as LAEs, 3 as stars, and 2 as [OII] emitters. To gauge the efficacy of our predictive models, we establish the criteria whereby a source is classified as an LAE if its average prediction probability, across the ensemble of 15 models, exceeds 50%. The resulting confusion matrix is presented in Table 4.3, alongside corresponding evaluation metrics delineated in Table 4.4.

	True LAE	True Non-LAE
Predicted LAE	16 (TP)	1 (FP)
Predicted Non-LAE	17 (FN)	4 (TN)

TABLE 4.3: Confusion Matrix for the crossmatch with HETDEX.

	F1-Score	Accuracy	Precision	Recall
HETDEX	0.639	0.526	0.941	0.484

TABLE 4.4: Evaluation Metrics for the crossmatch with HETDEX.

The most immediate conclusion from the HEDTEX crossmatch is the high number of false negatives and the consequent low recall. This possibility was foreseen when analyzing the training results, where the recall value was already lower than the remaining metrics. Once again we verify that our model reveals some conservatism i.e. many sources that are spectroscopically confirmed LAEs are not being predicted as such. On

the other hand, it is positive that high precision is achieved, reflected in a high number of true positives and low numbers of false positives.

As referred to in the title of this dissertation, these models are being designed with a possible application to large surveys in mind, and consequently, one of the main goals is to pre-select a large amount of LAE candidates even if sacrificing precision because the low recall score is not desirable as it means losing a significant amount of potential real LAE sources.

The first possible solution to increase the number of sources predicted might be relaxing our imposed restriction that considers a source as a predicted LAE only if the average prediction probability is higher than 50%. This threshold is lowered to 25% and the exercise is repeated. The confusion matrix and evaluation metrics with this new threshold are shown in Tables 4.5 and 4.6, respectively.

	True LAE	True Non-LAE
Predicted LAE	21 (TP)	1 (FP)
Predicted Non-LAE	12 (FN)	4 (TN)

TABLE 4.5: Confusion matrix for the crossmatch with HETDEX assuming a threshold of 25%.

	F1-Score	Accuracy	Precision	Recall
HETDEX	0.764	0.658	0.955	0.636

TABLE 4.6: Evaluation metrics for the crossmatch with HETDEX assuming a threshold of 25%.

The number of false negatives decreased significantly (from 17 to 12), being converted into true positives, while the true negatives and false positives did not change. Naturally, this also affects the metric values, all of them are now higher, but the most important highlight is the increase in the recall score, which increases by approximately 25%. This stresses the importance of taking into consideration the trade-off between precision and recall: as more sources are being predicted as LAEs, increasing recall, inevitably nonrelevant sources will be captured as well, decreasing precision.

4.3.2 DEIMOS

The Deep Imaging Multi-Object Spectrograph (DEIMOS) is a high-resolution, multi-object spectrograph that operates in the optical and near-infrared wavelengths. It gave origin to a spectroscopic catalog containing 10718 objects up to $z = 6$ in the COSMOS

field observed with the Keck II telescope (please see [2] for more detailed information about this catalog).

The crossmatch procedure between our predictions in COSMOS and DEIMOS is similar to what was done with HEDTEX. We consider only the sources with high-quality spectra, i.e. excluding sources with quality flag "Q" signaling unreliable spectroscopic identification (see [2] for more information about the assignment of quality flags to the sources). This resulted in 181 matches (without the spectra quality restriction, there were 827 matches).

This time we will not translate these matches into a confusion matrix and evaluation metrics because we do not have the same confidence level about the sources nature and therefore we can only conclude that we obtained 43 confirmations of galaxies showing a Ly α line and predicted by our models and 62 galaxies that contain a Ly α line but were not predicted by our models. Once again, if we lower our threshold to 25%, we see that the number of correctly predicted sources that contain a Ly α line rises from 43 to 68 and the incorrectly predicted by the model lower from 62 to 36. This crossmatching exercise with both catalogs confirmed that the classification models are actually predicting true LAEs but it unveiled a discussion regarding the significant impact of the criteria used to classify a source as an LAE. This is partly because, although the 15 models are similar in their construction, there are dissimilarities between them due to small changes in the data used to train them which naturally reflect different predictions. Having this said, it is important to thoughtfully reflect on the threshold used, depending on the objective of the task, as it may impact significantly the predictions.

From hereon, a threshold of 50% will be used, unless the opposite is explicitly said.

4.4 Generalization in the era of large surveys

As the name of this thesis suggests it is one major goal to explore the potentialities of this work to be applied to other fields and surveys not included in COSMOS. It served as a motivation, for example, for choosing only broad-band in the optical and NIR regimes as features and also for the precision-recall tradeoff discussion.

This generalization is not as straightforward as one can initially think, it most likely implies the use of encoders-decoders, due to the fact that the filters used in other surveys can be different from the ones present in COSMOS, even though the wavelength coverage is in many cases similar. The analysis and study of encoders and decoders is out of the

scope of this thesis and our simplified approach tries to simulate how these models would behave if not all the bands are available. As mentioned before, our work uses filters from COSMOS that cover the wavelengths from 3560 Å to 23094 Å. So our exercise will be to remove photometric data from the features given to train the models and try to mimic the filters wavelength coverage of potentially interesting surveys and missions where this work can be applied. Please find below a summary of each survey considered for this exercise:

- **Euclid:** The Euclid mission [71] will observe 15 000 deg^2 of extragalactic sky from the Sun–Earth Lagrange Point L2. Among many other data, Euclid’s plans include determining the photometry of at least one billion galaxies. To achieve this goal, the Euclid spacecraft has a visible-wavelength camera (the VISible instrument, VIS) and a near-infrared camera/spectrometer (the Near-Infrared Spectrometer and Photometer, NISP). The VIS instrument covers the wavelength between 5500 – 9000 Å while NISP has 3 broad band filters (Y, J, and H) covering the wavelength ranges of 9000–11920 Å, 11920–15440 Å and 15440–20000 Å, respectively. To mimic the wavelength coverage of the VIS instrument, we consider the r, i, and z bands from Subaru and to mimic the coverage of the NISP instrument the Y, J, and H bands from VISTA.
- **LSST:** The Legacy Survey of Space and Time (LSST) [72] is a survey that will cover about 18,000 deg^2 of the southern sky with 6 filters (u, g, r, i, z, Y) covering a wavelength interval from approximately 3500 to 10600 Å. Consequently, only the homonymous bands available in the COSMOS2020 catalog are used, the u* filter from MegaCam, the griz filters from HSC, and the Y filter from VISTA.
- **DES:** The Dark Energy Survey (DES) [73] is an astronomical survey designed to study hundreds of millions of galaxies and to help in understanding the nature of dark energy. It uses the newly built Dark Energy Camera (DECam) that records images using five filters, g, r, i, z, and Y, spanning wavelengths from 4000 Å to 10800 Å. The filters we are using to mimic the wavelength coverage are g, r, i, z from Subaru and Y from VISTA, which were already presented.
- **ROMAN:** The Nancy Grace Roman Space Telescope [74] is a NASA infrared space telescope created to answer important questions in the areas of dark energy, exoplanets, and infrared astrophysics. It is equipped with a Wide-Field Instrument

(WFI) carrying 8 filters spanning 4800 – 23000 Å. The wavelength range covered is large and only the u^* -band from COSMOS is removed from our exercise.

- **J-PAS:** The Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS) [75] is an astronomical survey with filters that cover the entire visible region of the electromagnetic spectrum (3500 Å to 10000 Å) with 54 narrow-band filters, supplemented by two medium-band and one broad-band filter, but our model only uses broad-band filters to operate. The total coverage is similar to DES but J-PAS upper and lower limits are both lower, therefore we will use the u^*, g, r, i, z bands to mimic the wavelength coverage.

Having defined the changes in the features used to mimic the filter set of each survey presented above, the models are created using the features from the filter set that best matches the wavelength covered by each survey and the results are summarized in Table 4.7, where the performance is evaluated in the test set.

Survey	Algorithm	Test set			
		F1-Score	Accuracy	Precision	Recall
This work	LightGBM	0.875	0.877	0.890	0.861
	XGBoost	0.867	0.869	0.882	0.852
	CatBoost	0.869	0.870	0.878	0.860
Euclid	LightGBM	0.718	0.724	0.735	0.702
	XGBoost	0.707	0.714	0.724	0.692
	CatBoost	0.707	0.714	0.726	0.689
LSST	LightGBM	0.870	0.873	0.887	0.854
	XGBoost	0.863	0.866	0.879	0.849
	CatBoost	0.864	0.866	0.876	0.852
DES	LightGBM	0.808	0.812	0.825	0.792
	XGBoost	0.802	0.806	0.820	0.784
	CatBoost	0.806	0.810	0.824	0.790
ROMAN	LightGBM	0.823	0.825	0.833	0.814
	XGBoost	0.820	0.821	0.825	0.814
	CatBoost	0.828	0.831	0.842	0.815
J-PAS	LightGBM	0.864	0.866	0.881	0.848
	XGBoost	0.854	0.856	0.869	0.840
	CatBoost	0.859	0.861	0.872	0.847

TABLE 4.7: Classification evaluation metrics of the test set for our work compared with other surveys: Euclid, LSST, DES, ROMAN, and J-PAS.

We know a priori the high impact of the g -band and the colors involving this band (from the correlation analysis in Table 3.3 and the feature importance in Figures 4.4, 4.5, and 4.6). Consequently, we expect that removing this band will result in lower metrics. This is demonstrated when comparing the results for Euclid and ROMAN where the only difference is the removal of the u^* - and g -bands or only the u^* -band, respectively. Even

though the difference between removing or not the u^* -band is significant, as reflected in ROMAN results, this difference is further emphasized in Euclid, where the g -band is also removed. Both these bands highly impact our model's performance and therefore are crucial for them to precisely separate LAEs from non-LAES. On the other hand, removing the J -, H -, and K_s -bands does not alter noticeably the performance, as seen by the LSST results.

From what we have already discussed so far, namely related to Table 3.3, this generalization exercise further emphasizes the high importance of the g -band. Regarding the impact of the u^* -band, we have also already discussed the trend of these galaxies to have higher star formation rates, especially at lower masses (see Figure 3.7), consequently, it is expected that they contain a large number of young stars (O, B) that produce an excess in emission at these wavelengths, consequently highlighting the importance of the u^* -band for the correct classification.

4.5 Assessing the robustness of the classification models

Robustness to small perturbations in the input data is an important quality measure for machine learning models, as these perturbations may cause failures in the classification of sources, calling into question the reliability of our models. The main idea behind this verification is that small changes in our input feature data, i.e. fluxes (which indirectly also perturbs the magnitudes and colors) can impact the performance of the classification models, including the predictions obtained when using them. In each training/test sample, every source is perturbed by creating a Gaussian distribution with a mean equal to the flux value and sigma equal to the flux error and then drawing a random value out of this distribution. This is done for every source in the five samples, each 15 times, for every band thus creating five perturbed samples that will then be used to train and test new models and then compare with the unperturbed models previously presented and discussed.

The perturbations can be separated into three different cases:

1. **Perturbations in the input data:** By perturbing the input data, the new models will be trained and tested in these new perturbed samples and therefore we ought to investigate how the perturbations in the input data might affect the performance of the models and the consequences when predicting in unseen data. In Figure 4.12,

there is a variation of the F1-Score across the 15 iterations for the LightGBM, XGBoost, and CatBoost models. There is a clear underperformance of the perturbed models compared to the unperturbed, but this decrease is approximately 5% which is not dramatic. This difference is not surprising because by perturbing the data, we are introducing more noise to the models and therefore a decrease in performance is expected. On the other hand, there is also some stability across the iterations meaning that the different perturbations performed do not differ significantly between them.

2. **Perturbations in the predictions table:** Up until now, we have indirectly perturbed the source labels, when creating five sub-samples with different sources labeled as non-LAEs and directly perturbed the training samples in the previous point. We are now going to also perturbate the sample where the predictions are made, i.e. the COSMOS2020 restricted between $2 < z < 6$ and with the sources from the five subsamples removed, and then apply the unperturbed models to these 15 perturbed predictions tables.
3. **Joint Perturbations:** In this final step, we join the perturbations performed in steps 1) and 2), i.e. applying the perturbed models to the perturbed predictions tables.

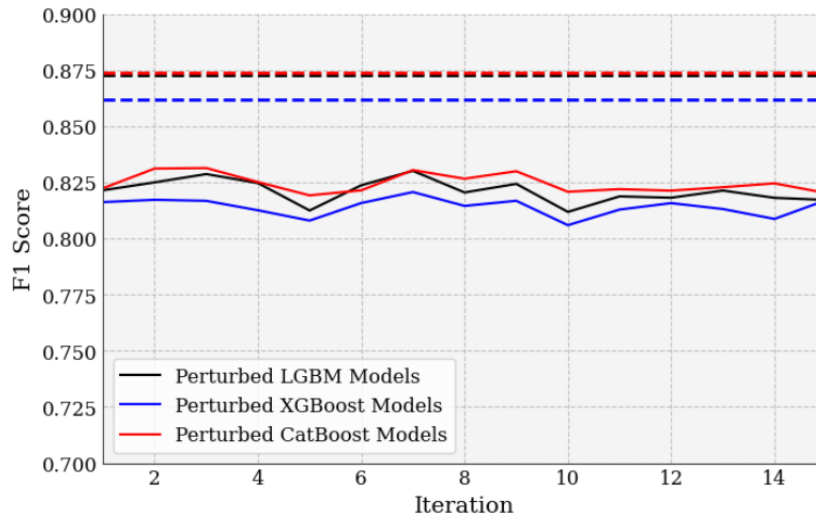


FIGURE 4.12: Variation of the F1-Score metric across the 15 perturbing iterations for LightGBM, XGBoost, and CatBoost. For each algorithm and in each iteration, an average is computed over the 5 samples. The dashed colored lines are the unperturbed F1-Score for each algorithm.

In the following plots, we will compare how the different perturbations have different consequences in the model's outcome. Firstly, the variation in the number of predicted

LAEs with a prediction probability over 90% in the perturbed models for the three different cases is computed in Figure 4.13. There is a big difference in the number of predicted LAEs when considering the input data perturbed or the prediction data perturbed, the latter results in almost three times the sources predicted when comparing to perturbing the input data (and therefore when joining both perturbations, the numbers tend to the unperturbed result). This indicates that there might be a shift in prediction probabilities towards higher values when perturbing the prediction data.

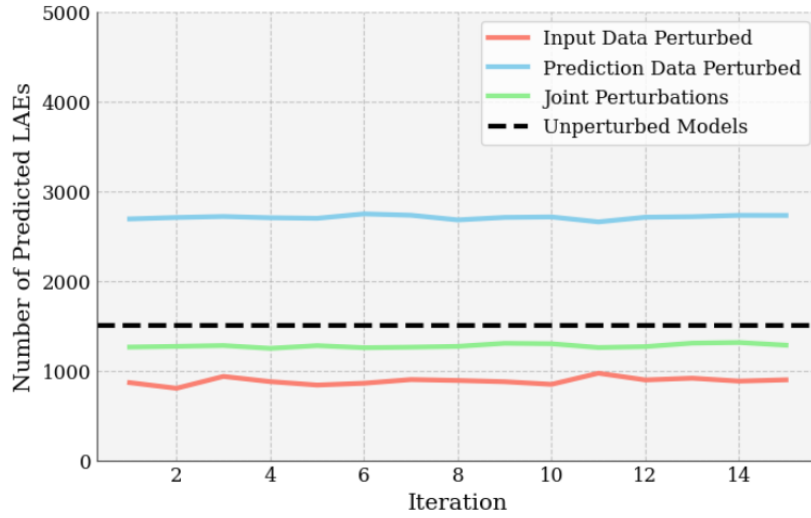


FIGURE 4.13: Variation of the number of predicted LAEs with a prediction probability over 90% across the 15 perturbing iterations for LightGBM, XGBoost, and CatBoost. It is important to recall that an average of prediction probability over all the 15 models (5 samples $\times$ 3 algorithms) is computed in each iteration.

In order to further investigate this anomaly found in the predictions of the perturbations in the predictions data, in Figure 4.14, 2D histograms of the perturbed and unperturbed prediction probabilities are computed for the three perturbation cases above-mentioned. For the first perturbation case, the data points seem to follow a 1:1 relation with some scatter around this line, but there is some trend towards lower perturbed probabilities for sources that have high unperturbed probabilities. The results start being odd in the second case, where besides the trend at the upper right side of the plot, there is also a clear shift of sources with low unperturbed probabilities towards higher perturbed probabilities and this might help to explain the excess in predicted LAES seen in Figure 4.14. Regarding the last case, the excess mentioned diminishes but the trend at the upper right side of the plot aggravates. These three linear fits are isolated in Figure 4.15, where it is again highlighted that in the lower left side of the plot, there is an overestimation of the prediction probabilities by the perturbed case (except when only the input data is

perturbed) and also in the upper right side there is an underestimation of the perturbed prediction probabilities.

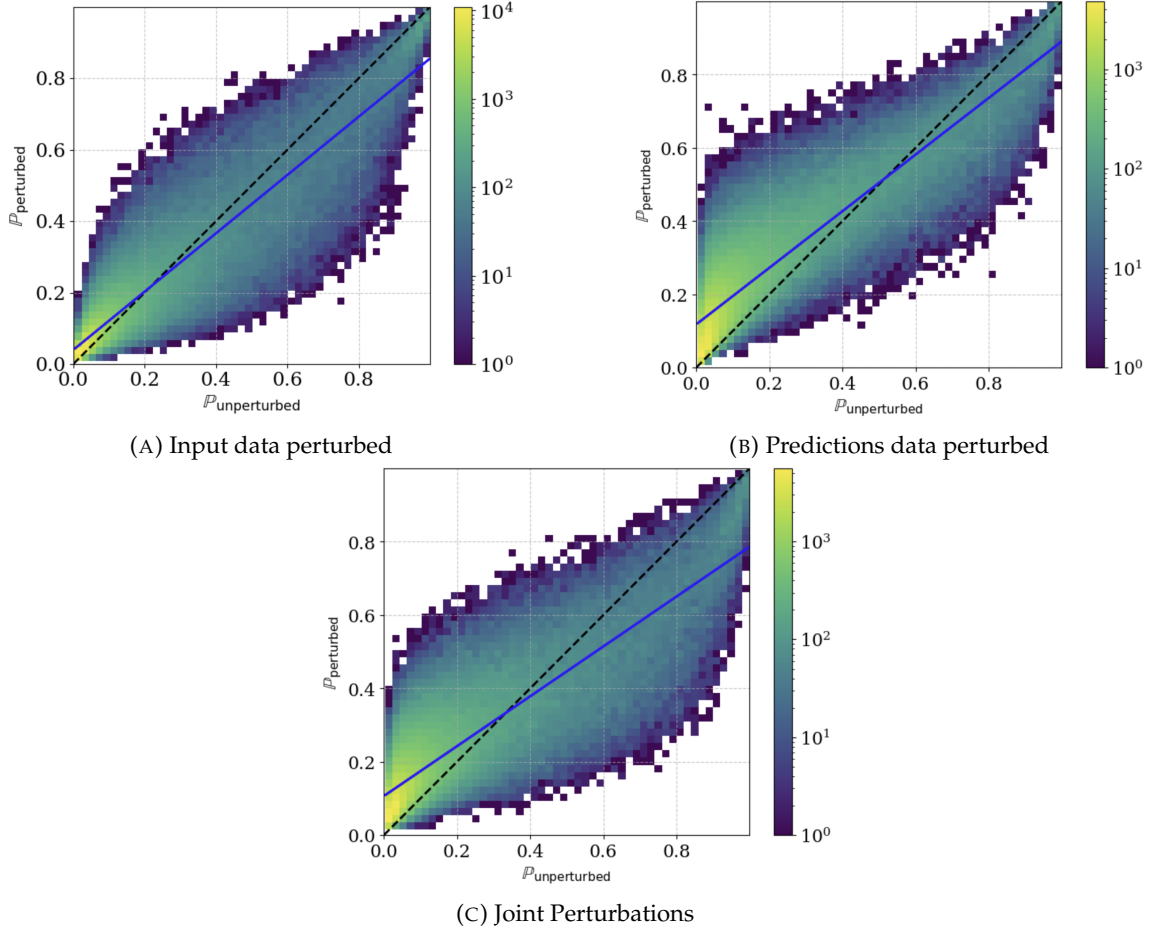


FIGURE 4.14: Comparison of the perturbed prediction probabilities and unperturbed perturbed probabilities for the input data perturbation, prediction table perturbation, and joint perturbation cases. Once again, average prediction probabilities over all the models were computed for each source. The blue lines represent a linear fit to the data points.

In order to verify what might be causing more variation in the prediction data perturbations, this exercise of perturbing the prediction data only was re-did and separated into three cases: (i) only fluxes, (ii) fluxes and magnitudes, and (iii) fluxes, magnitudes, and colors, depending on which features are used. The results are compiled in Figure 4.16.

It is possible to conclude that, as the number of features increases, the scatter around the 1:1 relation between unperturbed and perturbed prediction probabilities also increases, aggravating the “bump” located in the left part of the plots. This raises another trade-off dilemma: even if the model has better metrics performance with more features it also might be simultaneously more prone to noise, making it less robust.

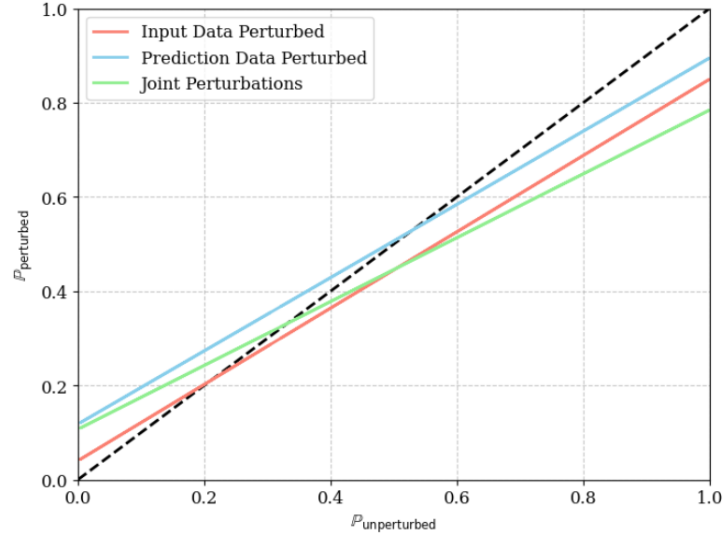


FIGURE 4.15: Linear fits to the 2D histograms from Figure 4.14 for the only input data perturbation, prediction table perturbation, and joint perturbation cases.

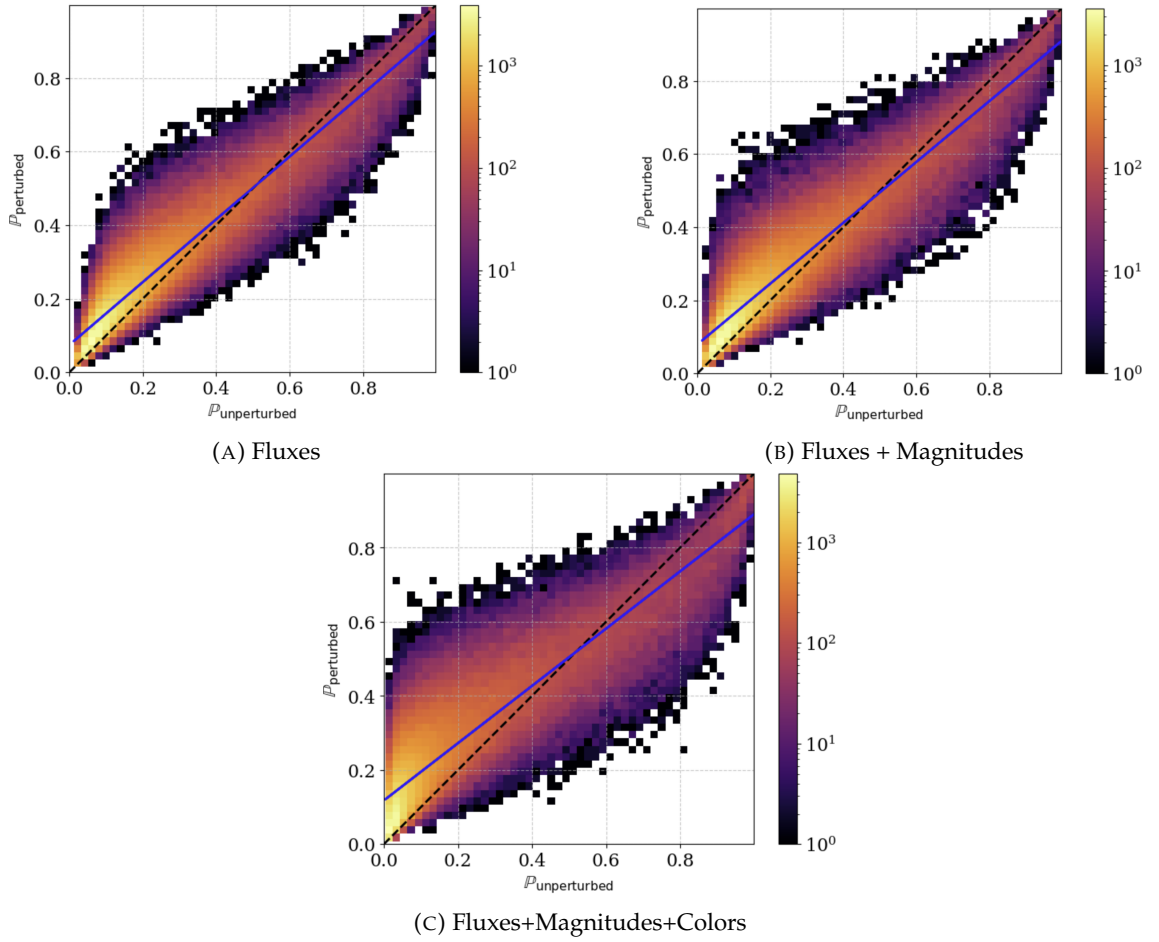


FIGURE 4.16: Comparison of the perturbations in the prediction data results when only using fluxes (a), only using fluxes and magnitudes (b), and using fluxes, magnitudes, and colors (c). The blue lines in each subplot represent a fit to the data.

Chapter 5

Properties of LAEs candidates: redshift, Ly α luminosity and equivalent width

The proposed regression assignment is to use the same photometric information as for the classification task, i.e. broad-band magnitudes, fluxes, and colors, in the optical and NIR, to predict several key properties of LAEs, such as their redshift, Ly α Luminosity, and equivalent width. One of the objectives of this task is to explore whether information on the emission line and redshift is encoded in the broad-band SED, in a way that machine learning can obtain a first approximation of the above-mentioned properties, relevant for target selection in spectroscopic surveys.

Although the input data is the same as in the classification task, there are some differences related to the samples used to train and data division, consequently, it is important to go through it before presenting the results. Afterwards, the performance results and analysis are presented and the generalization and robustness to perturbations are also assessed.

5.1 Revising data preparation

Contrary to the classification task, for the regression only the sample with LAEs (from SC4K) is used because it is, between SC4K and COSMOS2020, the sample that contains information about the above-mentioned properties that we want to study. Therefore, we

only have one sample with 3346 sources (see Section 3.2 for more information about the data processing steps).

The task is to create models that can predict the redshift, Ly α luminosity and equivalent width and it is of major importance to visualize the distributions of these properties in our sample (Figure 5.1) before showing the results.

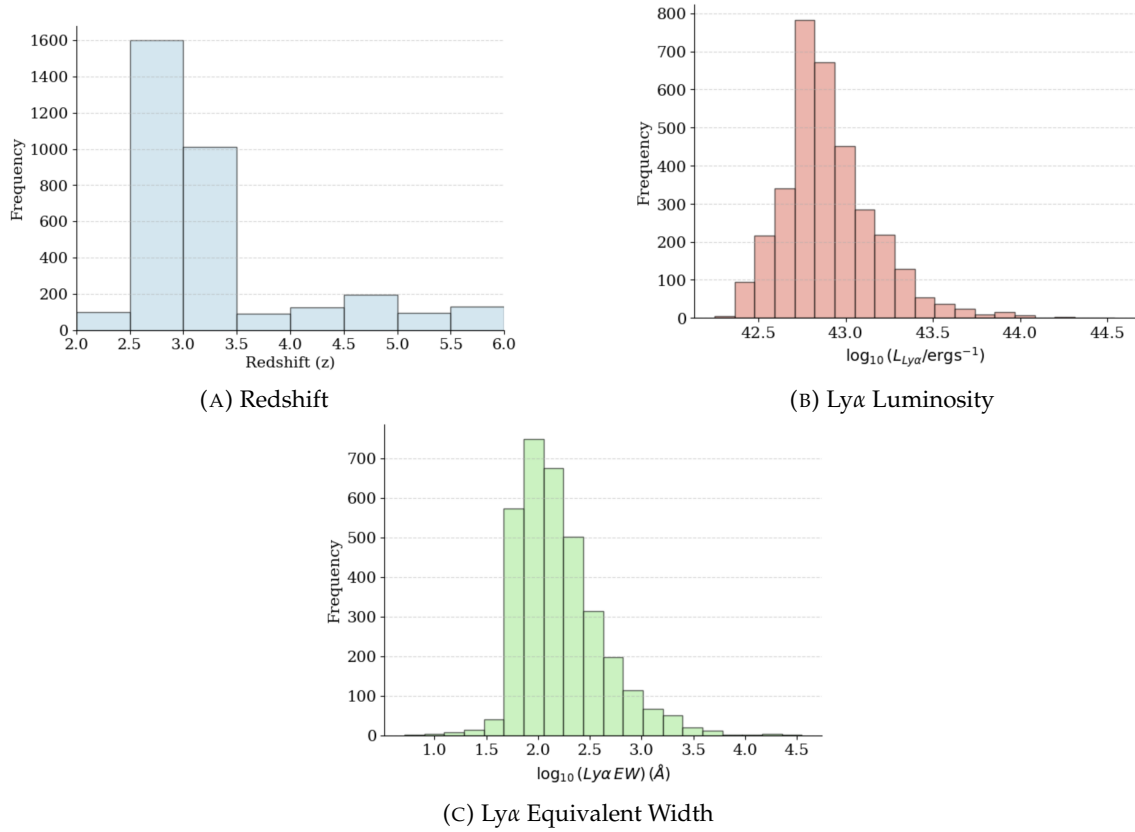


FIGURE 5.1: Distribution of redshift, Ly α luminosity, and Ly α equivalent width for the LAE sample.

The vast majority of sources are at $2.5 < z < 3.5$, which translates to a mean equal to 3.26 and a standard deviation of 0.82. Regarding the Ly α luminosity, the mean logarithmic value is 42.9 ergs^{-1} and the standard deviation is 0.27, with its distribution closely following a normal distribution. For the Ly α equivalent width distribution, the mean logarithmic value is 2.5 and the standard deviation is approximately 3. These statistics are highly influence by the presence of outliers, such as 7 sources with Ly α EW $> 10000 \text{ \AA}$, which can be AGN LAEs ([76]) or also "extreme" Ly α emitters that show emission not only from star-formation neither only AGN activity ([77]). These outliers can impact negatively the training process and the predictive performance of the model. Therefore we use the logarithmic transformed values of the Ly α equivalent width to handle this issue.

This also motivates the different data division strategy used. Contrary to what was done in the classification task, where the data was divided into test-train-validation sets, 3-fold cross-validation will now be used simultaneously to tackle the fact that the training sample is smaller and to deal with the outliers mainly in the $\text{Ly}\alpha$ equivalent widths.

Additionally, the hyperparameter optimization strategy also slightly differs from the classification task, mainly influenced by the usage of 3-fold cross-validation that triples the grid search time and consequently only one round of grid-searching is performed. Please see Sections 2.3 and 2.4 for more details about the data division and hyperparameter optimization strategies.

5.2 Performance metrics, feature importance, and generalization

Having stated the different approaches to the data itself, the data division, and hyperparameter optimization, the models can be trained and tested having this in mind. The metrics obtained for the algorithms with optimized hyperparameters are summarized in Tables 5.1, 5.2 and 5.3 for the redshift, $\text{Ly}\alpha$ luminosity, and equivalent width regressions, respectively. The analogous tables with the performance using the default hyperparameters can be found in Appendix A.

Model	Train			Validation			Test		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
LightGBM	0.090	0.127	0.976	0.155	0.258	0.900	0.184	0.331	0.846
XGB	0.095	0.145	0.969	0.153	0.248	0.908	0.180	0.317	0.859
CatBoost	0.086	0.112	0.981	0.161	0.253	0.905	0.184	0.317	0.858
Voting (soft)	-	-	-	-	-	-	0.180	0.318	0.858

TABLE 5.1: Training, validation, and test sets redshift regression evaluation metrics LightGBM, CatBoost, and XGBoost with hyperparameters optimized.

Model	Train			Validation			Test		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
LightGBM	0.070	0.091	0.884	0.127	0.170	0.592	0.138	0.184	0.544
XGB	0.045	0.060	0.949	0.126	0.169	0.597	0.137	0.183	0.549
CatBoost	0.078	0.100	0.859	0.125	0.169	0.600	0.135	0.179	0.567
Voting (soft)	-	-	-	-	-	-	0.136	0.181	0.560

TABLE 5.2: Training, validation, and test sets $\text{Ly}\alpha$ luminosity regression evaluation metrics LightGBM, CatBoost, and XGBoost with hyperparameters optimized.

For the regression models of all the properties, the performance is similar across the three algorithms, and compared with the default case there is no significant gain with the hyperparameter optimization (similarly to what happened in the classification task), the

Model	Train			Validation			Test		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
LightGBM	0.166	0.234	0.660	0.213	0.298	0.451	0.235	0.331	0.420
XGB	0.184	0.261	0.579	0.210	0.300	0.445	0.232	0.332	0.416
CatBoost	0.149	0.198	0.758	0.211	0.297	0.454	0.230	0.327	0.433
Voting (soft)	-	-	-	-	-	-	0.231	0.328	0.431

TABLE 5.3: Training, validation, and test sets Ly α equivalent width regression evaluation metrics LightGBM, CatBoost, and XGBoost with hyperparameters optimized.

improvements in MAE range from 0% – 7%, depending on the algorithm and property. Another important note is that using a soft voting technique, i.e. an average of the individual regressors predictions, does not significantly improve the performance but it is always better than at least two of the three algorithms individually and so, from hereon the results are obtained using soft voting.

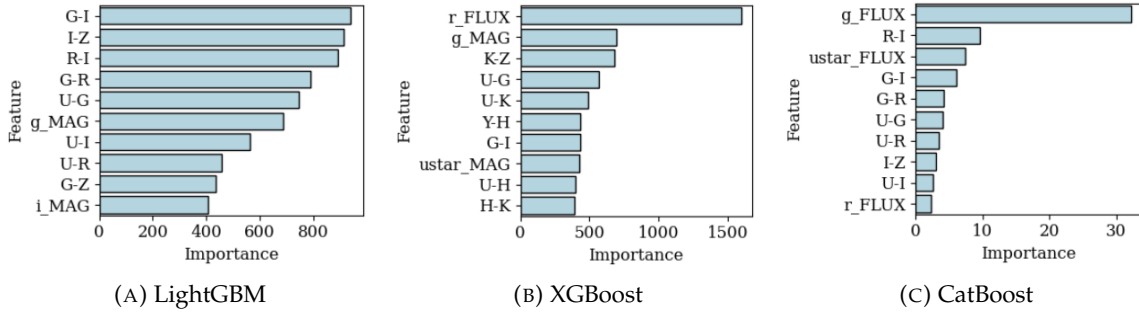


FIGURE 5.2: Feature Importance for the redshift regression models.

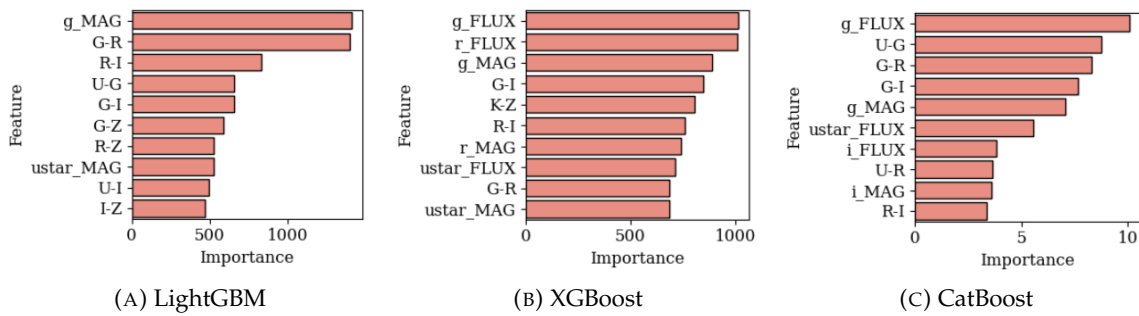
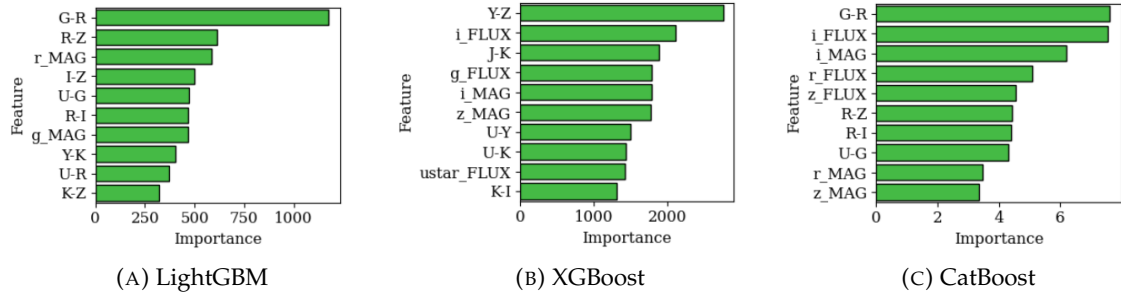


FIGURE 5.3: Feature Importance for the Ly α luminosity regression models.

In Figures 5.2, 5.3 and 5.4, the feature importance for the redshift, Ly α luminosity, and Ly α equivalent width regression models separated for each algorithm are shown. For LightGBM, the results are similar to what was obtained in the classification task, the g-band is highlighted, and also many colors containing it. For XGBoost, the r -band, together with the g -band, have a high importance for predicting the redshift and Ly α luminosity while for the Ly α equivalent width the most important feature is the $Y - Z$ color and

FIGURE 5.4: Feature Importance for the $\text{Ly}\alpha$ equivalent width regression models.

there are also other bands such as the i , u^* , g , and z that have a strong impact. Finally, for CatBoost, we observe again the same pattern from the classification task, that fluxes are more important than magnitudes, and also the g -band and u^* -band are highly important, specially for the redshift, $\text{Ly}\alpha$ luminosity.

Still in the visualization domain, in Figure 5.5, there is a comparison between the properties values from SC4K and the values predicted by the regression model (using soft voting) for the test set.

For the redshift, the scatter plot between the values predicted and the SC4K values must be evaluated cautiously because the redshift values from SC4K are discretely distributed, and consequently, the strange form seen in Figure 5.5 (a). Regarding the $\text{Ly}\alpha$ luminosity, it is possible to observe more scatter for the most extreme $\text{Ly}\alpha$ luminosity values, i.e. $L_{\text{Ly}\alpha} < 42.5$ and $L_{\text{Ly}\alpha} > 43.25$. For sources in the latter case, the predicted values tend to underestimate the real values. Finally, for the $\text{Ly}\alpha$ equivalent width, there is an underestimation by our models for values higher than 10^2 , which may be due to the low presence of high EW sources in the training sample, biasing the predictions towards lower values.

Generalization: Mimicking the filter wavelength coverage from other surveys

Analogously to the classification task, it is relevant to explore the potential suitability of these models for application on other surveys and fields. The method used for this analysis is the same as in Section 4.5, i.e. to mimic the filter wavelength coverage used by several surveys, removing some band information to more or less match the coverage of each survey. The surveys used are the same as in the classification task and please see Section 4.5 for a more detailed description of the methods used for each survey.

In Table 5.4, there is a summary with the metrics values for the test set concerning each survey.

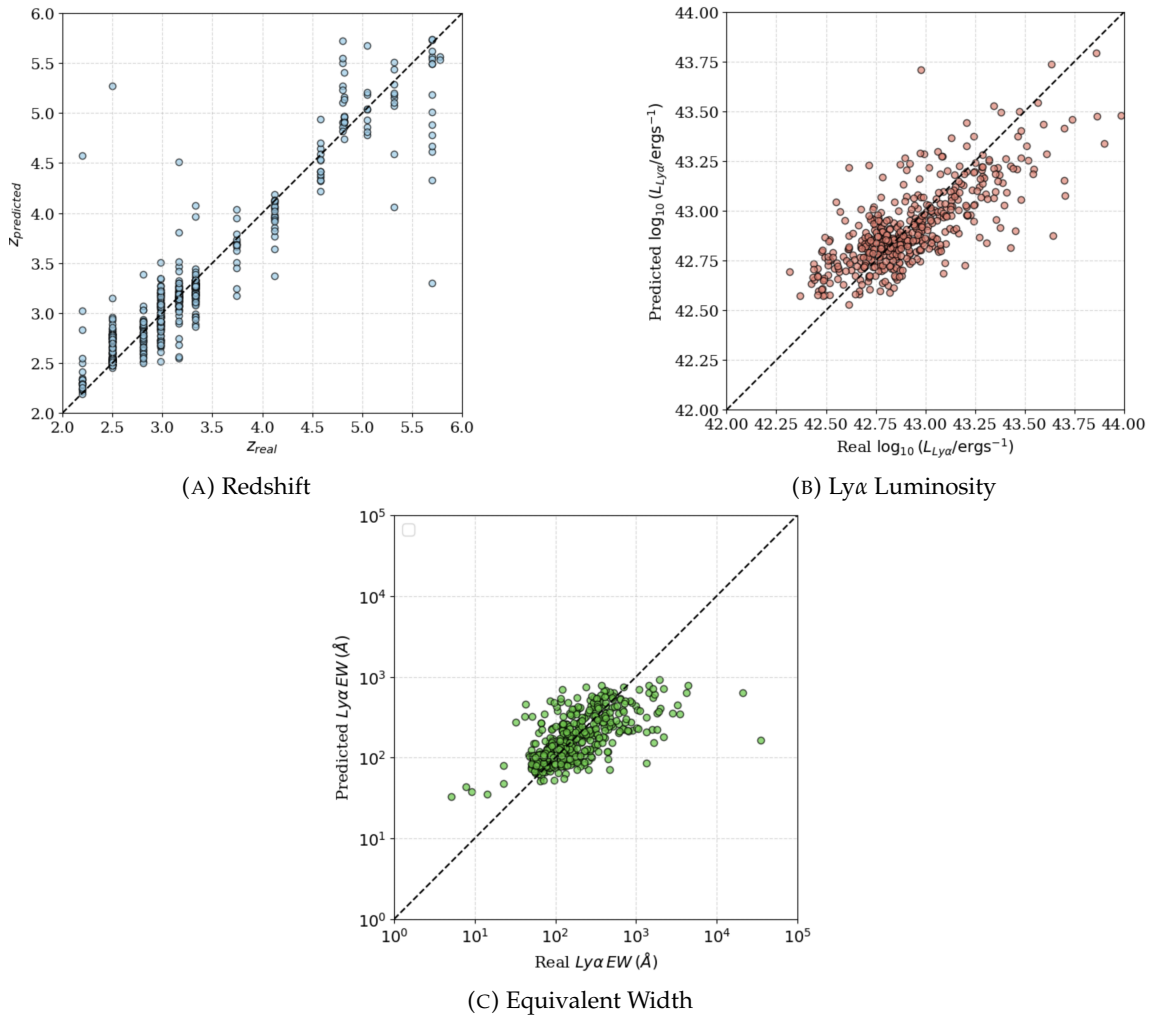


FIGURE 5.5: Comparison between the properties values from SC4K and the values predicted by the regression models (using soft voting) for the test set.

There is again a significant impact of removing the u^* - and g -bands, which is reflected in the worsening of the metrics for Euclid and DES. Specifically, the MAE almost doubles for Euclid (where both u^* - and g -bands are removed) and increases almost 50% for DES (where the u^* -band is not used). The removal of the NIR bands information has again lower impact, see for example the LSST and J-PAS performance.

Survey	Redshift			$\text{Ly}\alpha$ luminosity			$\text{Ly}\alpha$ equivalent width		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
This work	0.180	0.318	0.858	0.136	0.181	0.560	0.231	0.328	0.431
Euclid	0.357	0.504	0.643	0.177	0.232	0.278	0.247	0.346	0.368
LSST	0.190	0.324	0.852	0.136	0.181	0.557	0.238	0.339	0.392
DES	0.255	0.373	0.804	0.166	0.222	0.333	0.246	0.355	0.333
ROMAN	0.217	0.337	0.841	0.154	0.204	0.439	0.238	0.339	0.393
J-PAS	0.187	0.322	0.854	0.137	0.183	0.547	0.240	0.344	0.373

TABLE 5.4: Regression evaluation metrics of the test set for our work compared with other surveys: Euclid, LSST, DES, ROMAN, and J-PAS.

5.3 Characterizing LAE candidates in the COSMOS2020 survey

After having the regression models for the redshift, $\text{Ly}\alpha$ luminosity, and equivalent width optimized and trained, the following objective is to apply them to the LAE candidates found by the classification models in COSMOS2020 with $z \in [2, 6]$. We start by considering the standard threshold for a source to have been successfully identified as an LAE candidate to have an average prediction probability higher than 0.5. Consequently, the target sample to apply the models has 21400 sources and their predicted redshift, $\text{Ly}\alpha$ luminosity, and equivalent width distributions are shown in Figure 5.6 where it is also compared with the distribution of these properties in the SC4K sample, please see section 5.1.

The redshift distribution is similar in both samples, but there is a bigger portion of sources with lower redshift (i.e. $2.5 < z < 3.5$) in the SC4K sample. Regarding the $\text{Ly}\alpha$ luminosity histograms, there is a clear trend for the regression models to find sources with more intermediate luminosity values, i.e. between 42.5 and 43, which might be a consequence of the high capability of the SC4K survey to capture sources with more extreme $\text{Ly}\alpha$ luminosity values, but it tends to allow those with intermediate values “escape” because they might not be so easily differentiable from sources with other nature. For the $\text{Ly}\alpha$ equivalent width distributions, the patterns are more odd because the distributions are not in some degree similar, contrary to what happens in the redshift and $\text{Ly}\alpha$ luminosity cases. This might be partly due to the lower performance of the regression model, causing higher uncertainty in the predicted values.

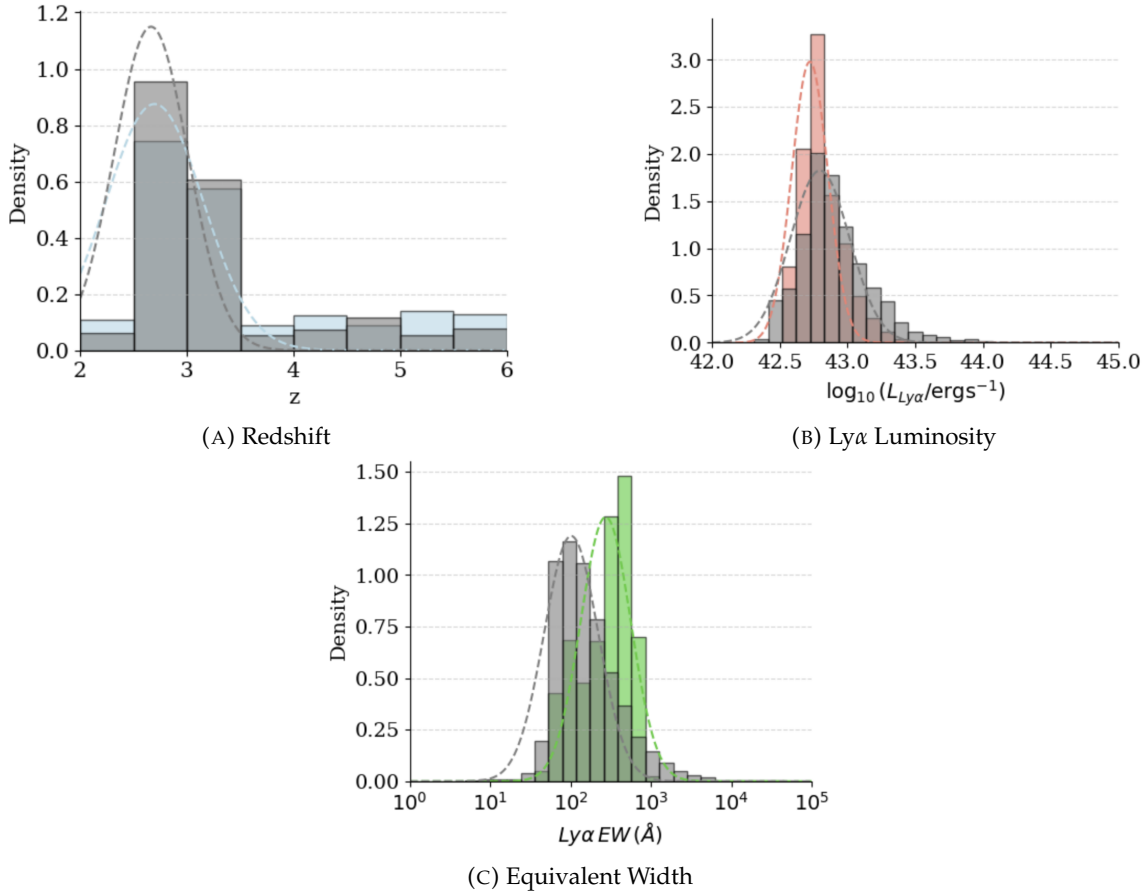


FIGURE 5.6: Comparison between the properties values from SC4K (grey histogram) and the values predicted by the regression models for the predictions sample. The dashed curves represent the fitted gaussians to the histograms. The distributions of redshift, Ly α Luminosity and equivalent width are colored blue, red and green, respectively.

Crossmatching redshift and $L_{Ly\alpha}$ predictions with HETDEX and DEIMOS

In the classification task, the predicted sources were crossmatched with the publicly available HETDEX and DEIMOS catalogs in order to obtain spectroscopic confirmation for some sources (please see Sections 4.3.1 and 4.3.2 for more details about the surveys used to make the crossmatched). Of these two catalogs, both have spectroscopic redshift information but only HETDEX has direct Ly α luminosity information. As mentioned before, there were detected 17 sources in HETDEX by the classification models, and 16 of these were revealed true positives consequently the comparison of their spectroscopic redshift and $L_{Ly\alpha}$ with the model predictions is shown in Figures 5.7 (a) and 5.8, respectively. Similarly for DEIMOS, where 43 sources were true positives and the comparison of their spectroscopic redshift with the predictions is also shown in Figure 5.6. Please note that besides being true positives, for DEIMOS a criteria of quality flag was also applied.

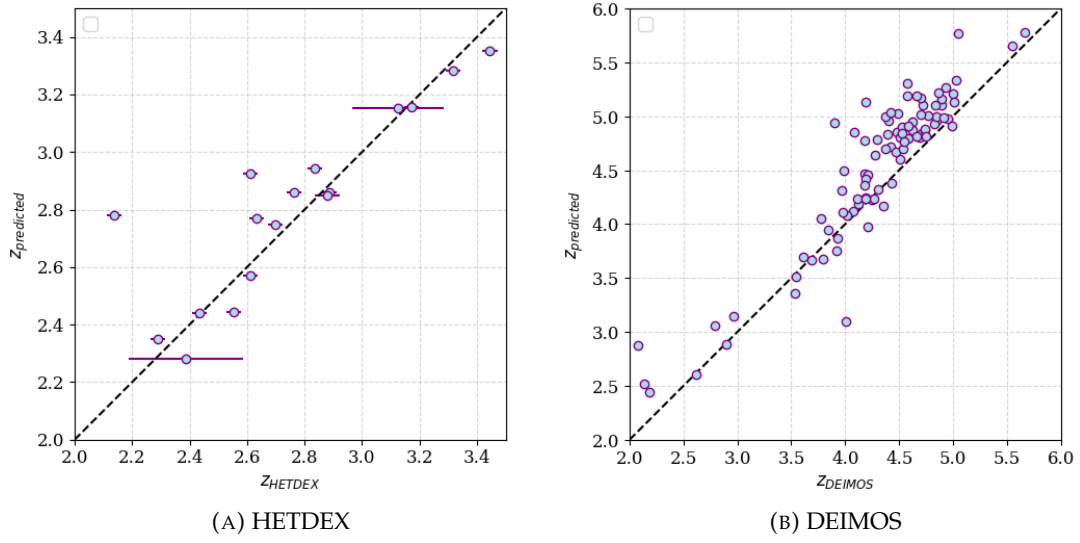


FIGURE 5.7: Comparison between the spectroscopic redshift from HETDEX and DEIMOS with the model predictions.

In Figure 5.7, for both cross-matches, there seems to be a fairly strong agreement between the regression model and the spectroscopic values. However, for DEIMOS, there is a consistent overestimation of redshift for sources with $4 < z_{\text{DEIMOS}} < 5$. But besides the scatter mentioned for DEIMOS, both relations seem to follow the 1:1 relation reasonably.

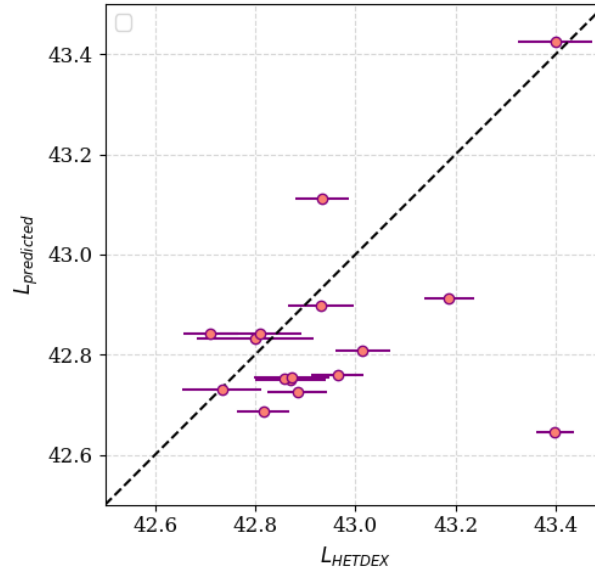


FIGURE 5.8: Comparison between the $\text{Ly}\alpha$ luminosity from HETDEX with the model predictions. There are also present error bars with the $L_{\text{Ly}\alpha}$ errors reported in the HETDEX public catalog.

Regarding the $\text{Ly}\alpha$ luminosity comparison (Figure 5.8), except the source that has $L_{\text{Ly}\alpha} \approx 43.4$, all the sources' deviation from the spectroscopic value is within the mean

absolute error reported (see Table 5.2), also taking into account the $L_{Ly\alpha}$ errors reported in the HETDEX catalog and drawn as error bars in the figure.

Unfortunately, we were unable to crossmatch the $Ly\alpha$ equivalent width predictions with spectroscopic values, due to the fact that neither HETDEX nor DEIMOS contain this information in the publicly available catalogs and computing it is out of the scope of this dissertation.

Number counts of LAEs across the luminosity intervals

The number density of LAEs per luminosity interval, encodes valuable information on the global properties of LAEs and $Ly\alpha$ emission, such as the star formation rate density (SFRD), which depends on the shape of the galaxy luminosity function.

The survey area probed in the interval $z \in [2, 6]$ for the COSMOS2020 sample where the predictions are made is 1.68 deg^2 and we then compute the co-moving volume at redshift 6 and 2. The co-moving volume found between redshifts 2 and 6 is approximately $5.2 \times 10^6 \text{ Mpc}^3$. Moreover, the predicted $Ly\alpha$ luminosities in Section 5.6 are counted and separated into different luminosity bins and the number density is computed for each bin, considering the co-moving volume mentioned. The relation between the number density and luminosity interval is plotted in Figure 5.9, alongside the SC4K data. It is important to note that the number densities provided (and the SC4K's) are not corrected, i.e. completeness, filter profile, or other corrections were not taken into consideration.

Due to the fact that we trained our models with SC4K, it is important to take into consideration our ground truth, i.e. the SC4K survey is more sensitive to bright $Ly\alpha$ emitters, and so it is a poor probe of the number density of fainter LAEs. The data point with $L_{Ly\alpha} \approx 42.5 \text{ erg/s}$ is clearly an outlier compared with the other bins. This is most probably caused by the fact that, as mentioned before, no corrections were applied and, just as what happened in [1], the most affected luminosity bins are, naturally, the fainter luminosity values.

In Figure 5.6, an excess in LAEs was already found at $L_{Ly\alpha} \in [42.5, 42.8]$, which naturally also manifests in Figure 5.9. Even though the redshift interval probed in SC4K is from $z \approx 2$ to $z \approx 6$, the COSMOS field was "sliced" and consequently a discrete sample of LAE sources was obtained in the redshift range mentioned rather than a continuous sample, which arises from the use of NB/IB filters. Additionally, for the last two bins in the bright end, the decrease is significantly steeper than what was found for SC4K and

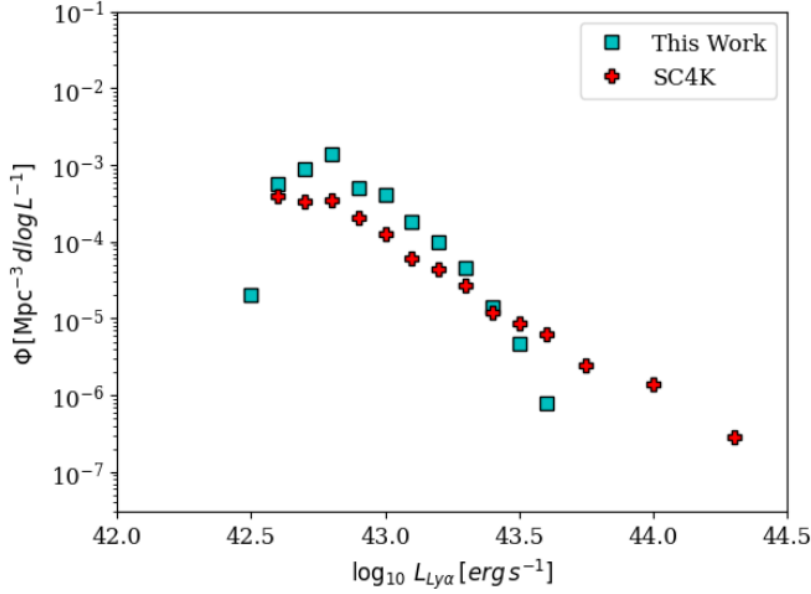


FIGURE 5.9: Number density of LAEs per luminosity interval at $2 < z < 6$ over a co-moving volume of $5.2 \times 10^6 \text{ Mpc}^3$. The data points from this work are the blue squares and this is also compared to data from [1], represented with red plus signals. In both results, corrections were not applied.

also there are no sources with $L_{\text{Ly}\alpha} \gtrsim 43.6$. Both these conclusions are a consequence of the fact the SC4K is more sensitive to brighter LAEs and thus few bright sources of this nature remain in the COSMOS sample to be predicted, leading to few or none in these luminosity intervals. Besides this, our models also contains some false negatives when searching for LAEs which can also impact the results.

Overall, this emphasizes that an overwhelming number of LAE sources were not previously detected, emphasized at $L_{\text{Ly}\alpha} \in [42.5, 42.8]$, and taking them into account can alter our understanding of the LAEs population.

5.4 Robustness analysis of the regression models

The verification of the robustness of the regression models when facing perturbations is of major importance. In the classification task (Section 4.5), the perturbations in the input data and perturbations in the prediction sample were studied by introducing noise into the fluxes, magnitudes, and colors using the errors provided in COSMOS2020. For the regression task, the perturbations will be done in the properties values, using the errors provided in SC4K for redshift, $\text{Ly}\alpha$ luminosity, and $\text{Ly}\alpha$ equivalent width.

For the $\text{Ly}\alpha$ luminosity and $\text{Ly}\alpha$ equivalent width, the 84th and 16th percentiles errors are provided in the SC4K catalog and consequently, our approach is that, for each source,

a random value is withdrawn from the concatenation of two normal distribution centered in the Ly α luminosity/equivalent width value and with standard deviations equal to each percentile error. This concatenation of two different normal distributions is motivated by the fact that the percentile errors are different, resulting in an asymmetric distribution.

For the redshift, the errors are not explicitly provided in the SC4K but they can be deduced following [1], based on the narrow/intermediate band with which a given source was detected with. Additionally, within a given error interval, a source has the same probability of having a redshift value in that interval, which motivates the use of a uniform distribution instead of a normal distribution.

The perturbations are performed throughout 15 iterations and the variation of the MAE can be found in Figure 5.10.

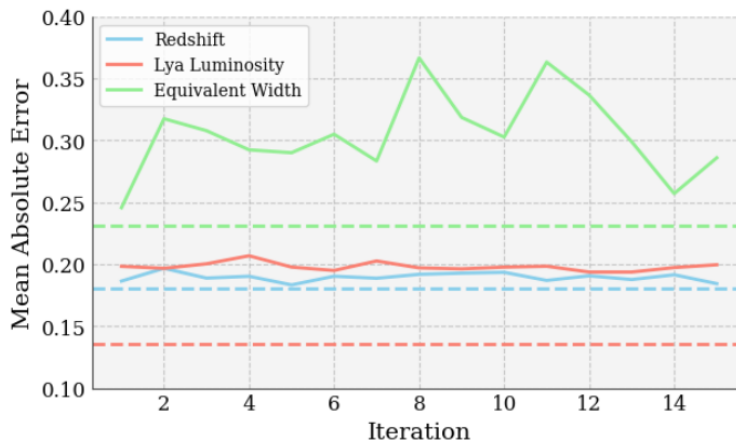


FIGURE 5.10: Variation of MAE through perturbations in the regression models. The dashed colored lines are the unperturbed F1-Score for each algorithm.

Analogously to the classification task, the metrics worsen when data is perturbed but in this case, the equivalent width and Ly α luminosity are clearly more affected by the noise introduced, originating differences of approximately 0.05 between the unperturbed and perturbed cases. The equivalent width perturbations also reveal no stability through iterations which does not happen in the other two cases. The redshift model is revealed to be the most resistant to noise with almost no variation in MAE between iterations and when compared with the unperturbed value.

In Figure 5.11, the values predicted by the unperturbed and perturbed regression models are compared in order to quantify the differences induced by the perturbations.

The redshift and Ly α luminosity comparison plots (Figure 5.11 (a)(b)) reveal a strong agreement between the unperturbed and perturbed values, which evidences that introducing noise to the training values of these properties does not affect significantly the

performance of the models. On the other hand, for the equivalent width, the 1 : 1 correspondence is not as straightforward and there is an overprediction of the equivalent width by the perturbed models mainly for sources with higher values of the unperturbed value, i.e. in the upper right of the plot. Having more scattering for the equivalent width might be a consequence of the worse performance metrics (see Table 5.3) that can lead to a more fragile model and prone to noise. Additionally, the equivalent width also has larger errors which can naturally lead to more variation of the perturbations.

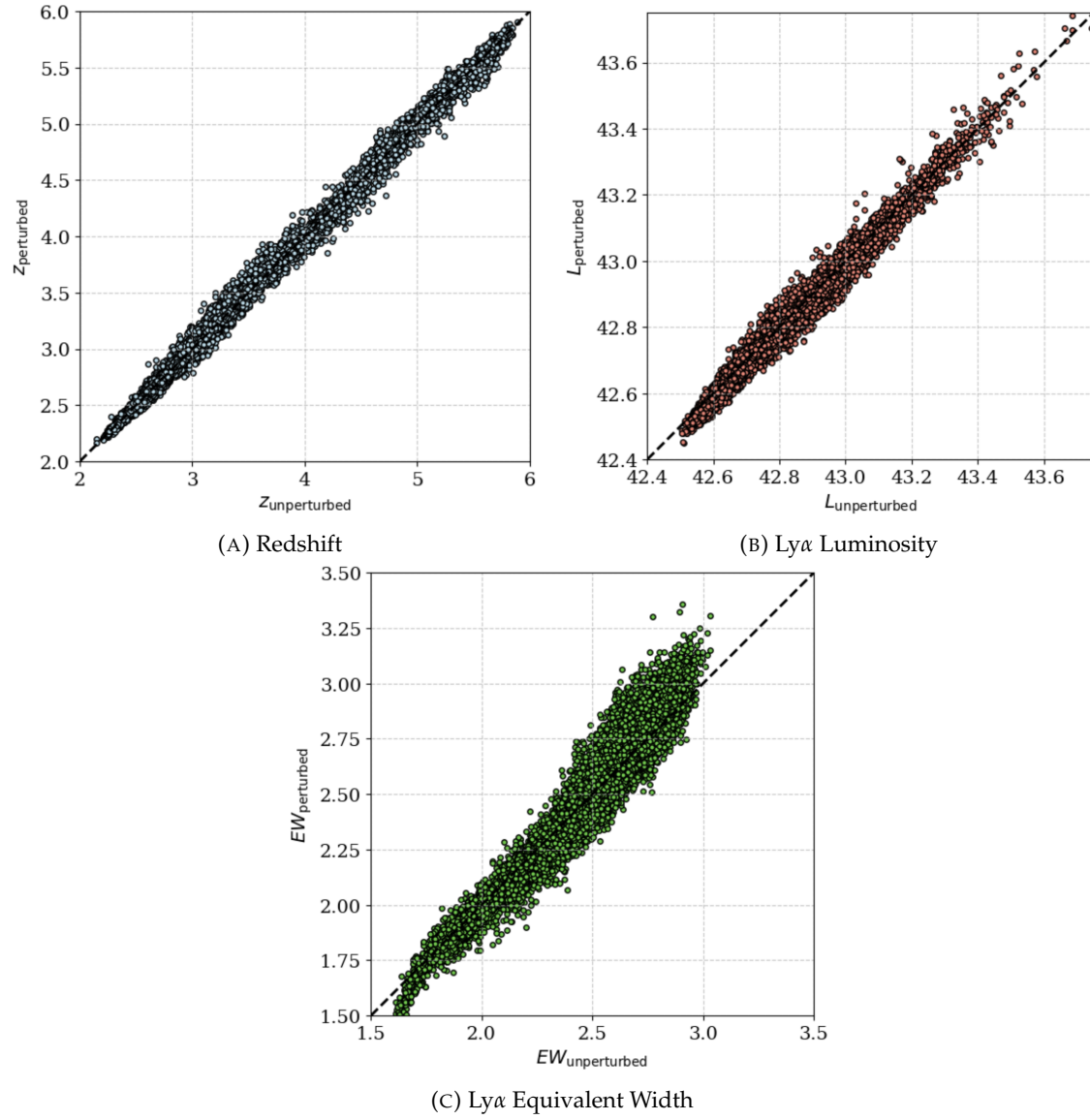


FIGURE 5.11: Comparison between the properties values predicted by the unperturbed and perturbed regression models for the predictions sample. The perturbed values were obtained by averaging the 15 iterations.

Chapter 6

Conclusion

In this dissertation, machine learning models were developed to identify and characterize Lyman- α emitters (LAEs). These models effectively perform binary classification and regression tasks using tabular data from astronomical surveys. The algorithmic framework employed includes Gradient Boosting Machines (GBMs) such as LightGBM, XGBoost, and CatBoost.

The training data has a balance of both LAEs and "non-LAEs" (i.e. the same number of sources) and follows the same distribution both in redshift ($2 < z < 6$) and i-band magnitude. The features provided as inputs for the models consist of photometric data from COSMOS2020, including fluxes, magnitudes, and colors in the optical and near-infrared wavelengths.

For the classification task, an F1 Score of approximately 0.87 was achieved across all three algorithms, resulting in a good performance. Feature importance analysis reveals that the g-band and the colors involving it have a significant impact across all the algorithms. These models were then applied to COSMOS2020 at $2 < z < 6$, excluding the training/testing sample (approximately 180,000 sources). The results showed that while the vast majority of sources were confirmed as non-LAEs by all the models, there were 6,667 galaxies identified as LAEs in agreement between all the models (considering the default 0.5 threshold probability to classify as an LAE), potentially more than doubling the SC4K sample. To further assess the validity of the classification models, the predictions made before were cross-matched with HETDEX [68] and DEIMOS [2]. For HETDEX, 38 crossmatches were obtained and for DEIMOS, this number increased to 181. In both cases, precision was higher than recall, prompting a discussion on the precision-recall tradeoff,

if the number of sources being predicted as LAEs increases (the recall increases) and consequently, non-LAEs will also be captured, which decreases precision, which stresses the importance of carefully selecting the classification probability threshold.

Given that the applicability of this work to other surveys was a key motivation, the models' performance on Euclid, LSST, DES, ROMAN, and J-PAS was analyzed by removing some photometric information from the feature list used to train the models, to match the wavelength interval covered by each survey. Once again, the g- and u*-bands emerged as highly impactful, probably due to the importance of capturing the Lyman break and the emission of the young stellar population, respectively.

Finally, the robustness of the classification models was examined by varying the flux values within the provided errors, allowing an assessment of how they behave in the presence of small perturbations to the flux values. Overall, perturbing the input data proved less impactful than perturbing the prediction data. However, increasing the number of features increased the errors, creating less robust models even if they were more accurate. This indicates that with more features the errors propagate, decreasing the models robustness. But this can be further analyzed in order to study the impact of each feature more in depth.

For the regression task, models for predicting redshift, Ly α luminosity, and Ly α equivalent width achieved mean absolute errors of approximately 0.180, 0.136, and 0.231, respectively. Feature importance results again highlighted the significance of the g-band and the colors containing it. Consequently, generalization to other surveys showed similar results, with the removal of the g- and u*-bands leading to significantly poorer outcomes. Crossmatching the predictions with spectroscopic surveys showed promising comparisons between predicted and actual values of redshift and Ly α luminosity.

An analogous perturbation study was conducted for the regression models, where the values of the three properties were perturbed within their provided errors. The results showed that all models were robust to noise in the training data. However, the Ly α equivalent width model was more fragile, possibly due to larger errors associated that lead to more variation in the perturbation results.

The work conducted throughout this dissertation demonstrates that it is both possible and efficient to search for LAEs and predict their key properties using broad-band photometric information through machine learning techniques. Additionally, it revealed that a significant number of LAEs might have been undetected in past surveys, which

could significantly impact previously achieved results, by increasing the size of the LAE population, thus helping their identification and study.

The by-products of this work can be tested and applied, in both the present and near future, to existing datasets, such as the publicly available data from DES. A key step in this process is matching the broad-band filters used in this study with those in the released datasets. Although the wavelength coverage is similar in most cases, the filters used here and in the referenced surveys are not identical. Therefore, a more accurate approach would involve using encoder-decoder models to improve the correspondence between the filter sets. One notable limitation of this work lies in the model optimization, which was primarily based on the F1-Score metric, motivated by the existence of a balance between precision and recall. There is room for improvement by shifting the focus to optimizing the models based on recall. Prioritizing recall could reduce the number of false negatives, resulting in the identification of more LAE candidates, even at the cost of slightly reduced precision. This approach would be particularly beneficial when applying the models to larger surveys, where maximizing the number of LAE candidates is critical. Other aspect of this work that can also benefit from additional research is the perturbations analysis conducted for both tasks, where more perturbations (besides the 15 iterations performed) could be tested in order to achieve more robust results.

Appendix A

Supplementary information - Classification and Regression models

In the present Appendix, the classification evaluation metrics separated by sample are presented (see Tables A.1, A.2, A.3, A.4, and A.5). Additionally, the feature importances for the classification task are also shown separated by sample and for each algorithm (Figures A.1, A.2, and A.3).

The regression evaluation metrics considering the default hyperparameters for redshift, Ly α Luminosity and equivalent width are also presented in Tables A.6, A.7, and A.8, respectively.

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM (D)	0.983	0.983	0.991	0.974	0.851	0.852	0.860	0.842	0.869	0.872	0.888	0.851
LightGBM (O)	0.907	0.910	0.937	0.878	0.868	0.870	0.883	0.854	0.873	0.877	0.906	0.843
XGB (D)	1	1	1	1	0.838	0.839	0.843	0.833	0.871	0.874	0.888	0.855
XGB (O)	1	1	1	1	0.876	0.876	0.880	0.872	0.856	0.861	0.885	0.829
CatBoost (D)	0.958	0.958	0.971	0.945	0.860	0.861	0.867	0.854	0.879	0.880	0.891	0.867
CatBoost (O)	1	1	1	1	0.879	0.881	0.890	0.869	0.870	0.873	0.888	0.853

TABLE A.1: Sample 1 train, validation, and test sets classification evaluation metrics for LightGBM, XGBoost (XGB), and CatBoost, considering the default hyperparameters (D) and the optimized ones (O).

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM (D)	0.984	0.985	0.996	0.974	0.865	0.867	0.882	0.848	0.874	0.875	0.879	0.869
LightGBM (O)	0.982	0.982	0.994	0.970	0.875	0.878	0.892	0.860	0.872	0.873	0.874	0.871
XGB (D)	1	1	1	1	0.839	0.842	0.852	0.827	0.864	0.865	0.866	0.863
XGB (O)	0.963	0.963	0.977	0.949	0.880	0.882	0.895	0.866	0.873	0.873	0.871	0.875
CatBoost (D)	0.962	0.963	0.976	0.949	0.860	0.863	0.876	0.845	0.870	0.871	0.873	0.867
CatBoost (O)	0.996	0.996	0.998	0.994	0.880	0.881	0.888	0.872	0.867	0.868	0.871	0.863

TABLE A.2: Sample 2 train, validation, and test sets classification evaluation metrics for LightGBM, XGBoost (XGB), and CatBoost, considering the default hyperparameters (D) and the optimized ones (O)

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM (D)	0.983	0.983	0.990	0.976	0.863	0.864	0.870	0.857	0.884	0.884	0.891	0.876
LightGBM (O)	0.993	0.993	0.999	0.986	0.878	0.878	0.873	0.884	0.889	0.891	0.909	0.871
XGB (D)	1	1	1	1	0.860	0.858	0.853	0.866	0.870	0.872	0.881	0.859
XGB (O)	0.993	0.993	0.998	0.988	0.879	0.879	0.883	0.875	0.882	0.884	0.899	0.867
CatBoost (D)	0.957	0.957	0.971	0.943	0.867	0.867	0.866	0.869	0.886	0.888	0.905	0.869
CatBoost (O)	0.995	0.995	0.998	0.992	0.874	0.876	0.887	0.863	0.878	0.879	0.891	0.865

TABLE A.3: Sample 3 train, validation, and test sets classification evaluation metrics for LightGBM, XGBoost (XGB), and CatBoost, considering the default hyperparameters (D) and the optimized ones (O).

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM (D)	0.985	0.985	0.994	0.975	0.860	0.864	0.889	0.833	0.866	0.867	0.871	0.861
LightGBM (O)	0.998	0.998	1	0.995	0.880	0.882	0.895	0.866	0.873	0.875	0.881	0.867
XGB (D)	1	1	1	1	0.846	0.849	0.863	0.830	0.859	0.862	0.876	0.843
XGB (O)	0.942	0.943	0.957	0.927	0.882	0.884	0.895	0.869	0.862	0.865	0.881	0.843
CatBoost (D)	0.961	0.962	0.977	0.945	0.868	0.870	0.883	0.854	0.876	0.877	0.889	0.863
CatBoost (O)	0.996	0.996	0.998	0.994	0.886	0.887	0.894	0.878	0.871	0.872	0.875	0.867

TABLE A.4: Sample 4 train, validation, and test sets classification evaluation metrics for LightGBM, XGBoost (XGB), and CatBoost, considering the default hyperparameters (D) and the optimized ones (O)

Model	Train				Validation				Test			
	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec
LightGBM (D)	0.984	0.984	0.994	0.974	0.871	0.875	0.896	0.848	0.873	0.875	0.885	0.861
LightGBM (O)	1	1	1	1	0.894	0.896	0.908	0.881	0.866	0.869	0.882	0.851
XGB (D)	1	1	1	1	0.873	0.876	0.896	0.851	0.869	0.871	0.881	0.857
XGB (O)	1	1	1	1	0.892	0.894	0.910	0.875	0.860	0.863	0.874	0.847
CatBoost (D)	0.960	0.961	0.974	0.947	0.879	0.882	0.9	0.860	0.876	0.877	0.886	0.867
CatBoost (O)	1	1	1	1	0.894	0.896	0.910	0.878	0.857	0.859	0.864	0.851

TABLE A.5: Sample 5 train, validation, and test sets classification evaluation metrics for LightGBM, XGBoost (XGB), and CatBoost, considering the default hyperparameters (D) and the optimized ones (O)

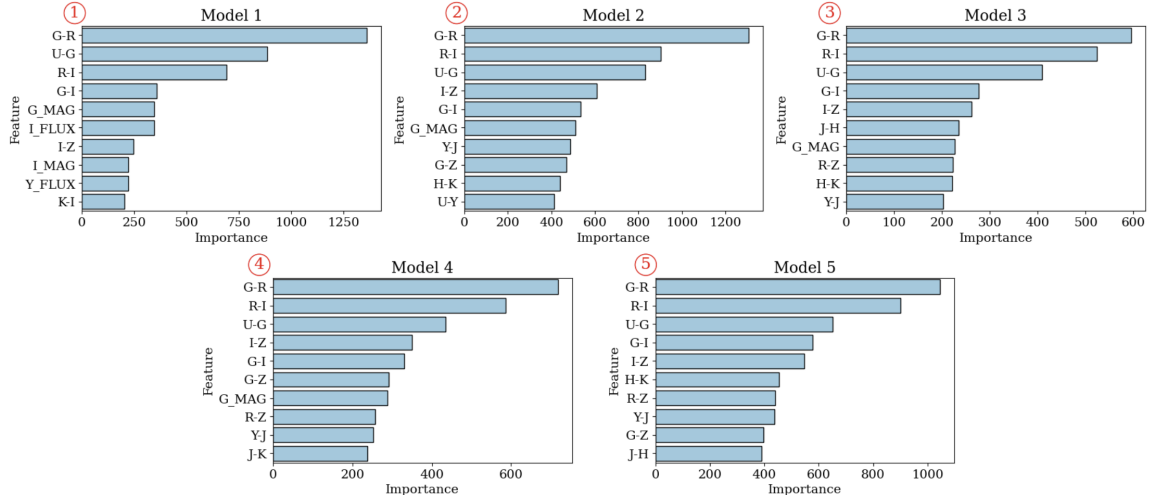


FIGURE A.1: Feature importances of the 10 most relevant features for the 5 samples with LightGBM.

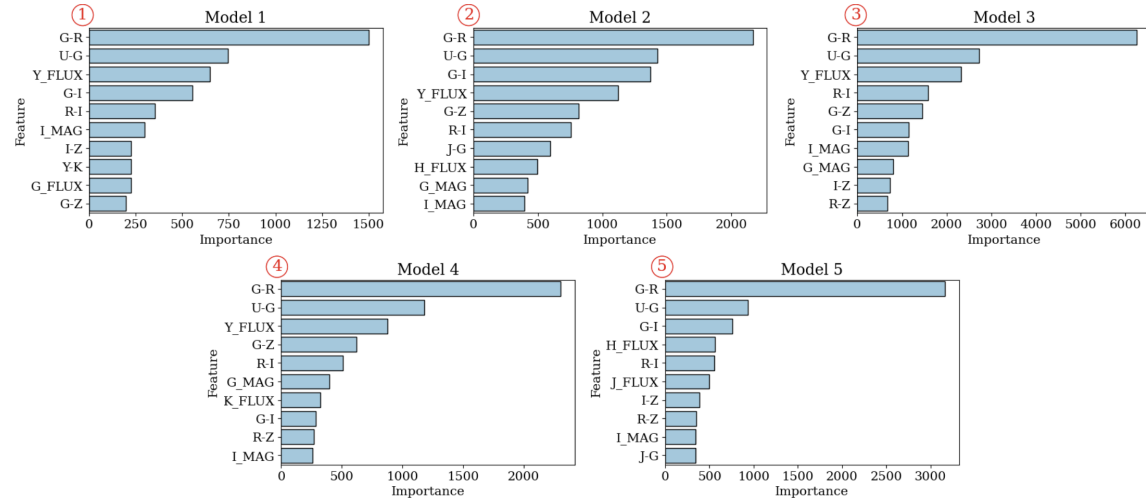


FIGURE A.2: Feature importances of the 10 most relevant features for the 5 samples with XGBoost.

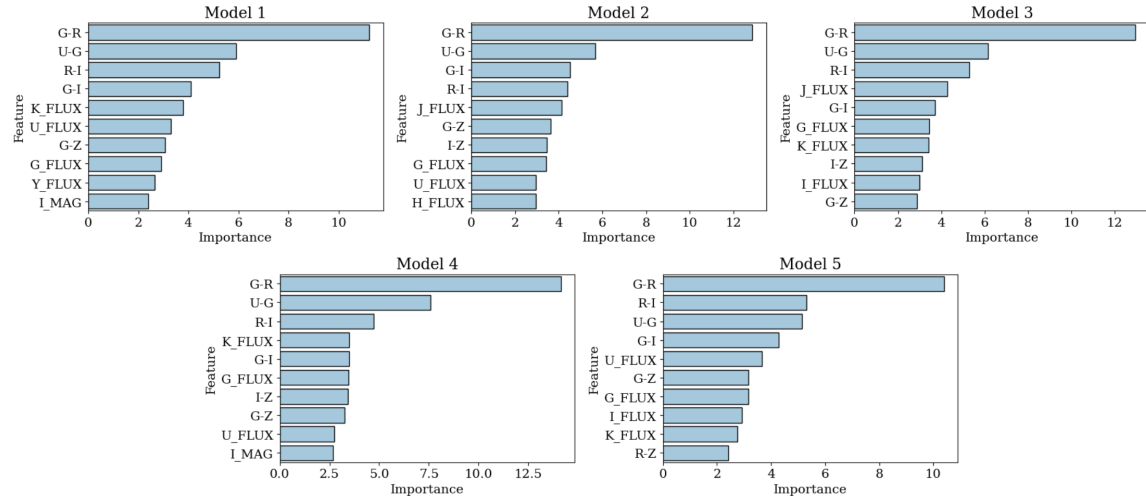


FIGURE A.3: Feature importances of the 10 most relevant features for the 5 samples with CatBoost.

Model	Train			Validation			Test		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
LightGBM	0.058	0.087	0.989	0.163	0.264	0.896	0.184	0.324	0.852
XGB	0.014	0.020	0.999	0.169	0.280	0.883	0.193	0.333	0.844
CatBoost	0.060	0.080	0.990	0.161	0.255	0.903	0.186	0.320	0.856
Voting (soft)							0.181	0.318	0.858

TABLE A.6: This table presents the regression evaluation metrics for redshift prediction using LightGBM (LGBM), XGBoost (XGB), and CatBoost models, all with default hyperparameters. The metrics are reported for the training, validation, and test sets. Soft voting was also applied for the three algorithms.

Model	Train			Validation			Test		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
LightGBM	0.052	0.071	0.929	0.130	0.174	0.572	0.142	0.188	0.522
XGB	0.016	0.023	0.992	0.136	0.183	0.528	0.143	0.191	0.510
CatBoost	0.495	0.064	0.942	0.126	0.169	0.596	0.137	0.181	0.558
Voting (soft)							0.137	0.181	0.556

TABLE A.7: This table presents the regression evaluation metrics for Ly α Luminosity prediction using LightGBM (LGBM), XGBoost (XGB), and CatBoost models, all with default hyperparameters. The metrics are reported for the training, validation, and test sets. Soft voting was also applied for the three algorithms.

Model	Train			Validation			Test		
	MAE	RMSE	R ²	MAE	RMSE	R ²	MAE	RMSE	R ²
LightGBM	0.082	0.111	0.923	0.220	0.308	0.413	0.236	0.331	0.421
XGB	0.022	0.032	0.994	0.230	0.326	0.343	0.247	0.353	0.341
CatBoost	0.083	0.108	0.928	0.216	0.303	0.433	0.233	0.109	0.423
Voting (soft)							0.234	0.331	0.420

TABLE A.8: This table presents the regression evaluation metrics for Ly α equivalent width prediction using LightGBM (LGBM), XGBoost (XGB), and CatBoost models, all with default hyperparameters. The metrics are reported for the training, validation, and test sets. Soft voting was also applied for the three algorithms.

Bibliography

- [1] D. Sobral, S. Santos, J. Matthee, A. Paulino-Afonso, B. Ribeiro, J. Calhau, and A. A. Khostovan, “Slicing COSMOS with SC4K: the evolution of typical Ly α emitters and the Ly α escape fraction from $z \sim 2$ to 6,” , vol. 476, no. 4, pp. 4725–4752, Jun. 2018. [Cited on pages [vi](#), [26](#), [27](#), [66](#), [67](#), and [68](#).]
- [2] G. Hasinger, P. Capak, M. Salvato, A. J. Barger, L. L. Cowie, A. Faisst, S. Hemmati, Y. Kakazu, J. Kartaltepe, D. Masters, B. Mobasher, H. Nayyeri, D. Sanders, N. Z. Scoville, H. Suh, C. Steinhardt, and F. Yang, “The DEIMOS 10K Spectroscopic Survey Catalog of the COSMOS Field,” , vol. 858, no. 2, p. 77, May 2018. [Cited on pages [vi](#), [48](#), and [71](#).]
- [3] M. Ouchi, Y. Ono, and T. Shibuya, “Observations of the Lyman- α Universe,” , vol. 58, pp. 617–659, Aug. 2020. [Cited on pages [1](#), [2](#), [3](#), and [4](#).]
- [4] J. S. Dunlop, “Observing the First Galaxies,” in *The First Galaxies*, ser. Astrophysics and Space Science Library, T. Wiklind, B. Mobasher, and V. Bromm, Eds., vol. 396, Jan. 2013, p. 223. [Cited on pages [1](#) and [4](#).]
- [5] R. B. Partridge and P. J. E. Peebles, “Are Young Galaxies Visible?” , vol. 147, p. 868, Mar. 1967. [Cited on page [1](#).]
- [6] J. D. Lowenthal, C. J. Hogan, R. F. Green, A. Caulet, B. E. Woodgate, L. Brown, and C. B. Foltz, “Discovery of a LY alpha Galaxy near a Damped LY alpha Absorber at $Z = 2.3$,” , vol. 377, p. L73, Aug. 1991. [Cited on page [2](#).]
- [7] P. Maloney, “Low-Redshift Lyman-Alpha Absorption Lines and the Dark Matter Halos of Disk Galaxies,” , vol. 398, p. L89, Oct. 1992. [Cited on page [2](#).]
- [8] D. Valls-Gabaud, “On the Lyman-Alpha Emission of Starburst Galaxies,” , vol. 419, p. 7, Dec. 1993. [Cited on page [2](#).]

- [9] K. Nakajima, M. Ouchi, K. Shimasaku, T. Hashimoto, Y. Ono, and J. C. Lee, “First Spectroscopic Evidence for High Ionization State and Low Oxygen Abundance in Ly α Emitters,” , vol. 769, no. 1, p. 3, May 2013. [Cited on page 2.]
- [10] J. A. Kollmeier, Z. Zheng, R. Davé, A. Gould, N. Katz, J. Miralda-Escudé, and D. H. Weinberg, “Ly α Emission from Cosmic Structure. I. Fluorescence,” , vol. 708, no. 2, pp. 1048–1075, Jan. 2010. [Cited on page 2.]
- [11] C. Behrens and J. Niemeyer, “Effects of Lyman-alpha scattering in the IGM on clustering statistics of Lyman-alpha emitters,” , vol. 556, p. A5, Aug. 2013. [Cited on page 2.]
- [12] I. Roy, C.-W. Shu, and L.-Z. Fang, “Resonant Scattering and Ly α Radiation Emergent from Neutral Hydrogen Halos,” , vol. 716, no. 1, pp. 604–614, Jun. 2010. [Cited on page 2.]
- [13] M. Dijkstra, “Physics of Ly α Radiative Transfer,” *Saas-Fee Advanced Course*, vol. 46, p. 1, Jan. 2019. [Cited on page 3.]
- [14] A. Paulino-Afonso, D. Sobral, B. Ribeiro, J. Matthee, S. Santos, J. Calhau, A. Forshaw, A. Johnson, J. Merrick, S. Pérez, and O. Sheldon, “On the UV compactness and morphologies of typical Lyman α emitters from $z \sim 2$ to $z \sim 6$,” , vol. 476, no. 4, pp. 5479–5501, Jun. 2018. [Cited on page 3.]
- [15] A. Hagen, R. Ciardullo, C. Gronwall, V. Acquaviva, J. Bridge, G. R. Zeimann, G. A. Blanc, N. A. Bond, S. L. Finkelstein, M. Song, E. Gawiser, D. B. Fox, H. Gebhardt, A. I. Malz, D. P. Schneider, N. Drory, K. Gebhardt, and G. J. Hill, “Spectral Energy Distribution Fitting of HETDEX Pilot Survey Ly α Emitters in COSMOS and GOODS-N,” , vol. 786, no. 1, p. 59, May 2014. [Cited on page 3.]
- [16] K. Nakajima, M. Ouchi, K. Shimasaku, Y. Ono, J. C. Lee, S. Foucaud, C. Ly, D. A. Dale, S. Salim, R. Finn, O. Almaini, and S. Okamura, “Average Metallicity and Star Formation Rate of Ly α Emitters Probed by a Triple Narrowband Survey,” , vol. 745, no. 1, p. 12, Jan. 2012. [Cited on page 3.]
- [17] A. Hagen, N. A. Bond, R. Ciardullo, C. Gronwall, E. Gawiser, W. Bowman, J. S. Bridge, H. S. Grasshorn Gebhardt, and D. P. Schneider, “The Rest-Frame Optical

- Morphology of Emission Line Galaxies at $2 < z < 3$: Evidence for Inside-Out Formation in Low-Mass Galaxies,” *arXiv e-prints*, p. arXiv:1610.01163, Oct. 2016. [Cited on pages 3 and 4.]
- [18] P. Santini, A. Fontana, M. Castellano, M. Di Criscienzo, E. Merlin, R. Amorin, F. Cullen, E. Daddi, M. Dickinson, J. S. Dunlop, A. Grazian, A. Lamastra, R. J. McLure, M. J. Michałowski, L. Pentericci, and X. Shu, “The Star Formation Main Sequence in the Hubble Space Telescope Frontier Fields,” , vol. 847, no. 1, p. 76, Sep. 2017. [Cited on pages 3 and 4.]
- [19] M. Ajiki, Y. Taniguchi, S. S. Fujita, Y. Shioya, T. Nagao, T. Murayama, S. Yamada, K. Umeda, and Y. Komiyama, “A Subaru Search for $\text{Ly}\alpha$ Emitters at Redshift 5.7,” , vol. 126, no. 5, pp. 2091–2107, Nov. 2003. [Cited on page 3.]
- [20] L. F. Grove, J. P. U. Fynbo, C. Ledoux, M. Limousin, P. Møller, K. K. Nilsson, and B. Thomsen, “The Building the Bridge survey for $z = 3$ $\text{Ly}\alpha$ emitting galaxies. II. Completion of the survey,” , vol. 497, no. 3, pp. 689–702, Apr. 2009. [Cited on page 3.]
- [21] J. P. U. Fynbo, C. Ledoux, P. Møller, B. Thomsen, and I. Burud, “The Building the Bridge survey for $z = 3$ $\text{Ly}\alpha$ emitting galaxies . I. Method and first results,” , vol. 407, pp. 147–157, Aug. 2003. [Cited on page 4.]
- [22] S. S. Fujita, M. Ajiki, Y. Shioya, T. Nagao, T. Murayama, Y. Taniguchi, K. Umeda, S. Yamada, M. Yagi, S. Okamura, and Y. Komiyama, “The $\text{H}\alpha$ Luminosity Function and Star Formation Rate at $z \sim 0.24$ Based on Subaru Deep Imaging Data,” , vol. 586, no. 2, pp. L115–L118, Apr. 2003. [Cited on page 4.]
- [23] R. Ciardullo, J. J. Feldmeier, K. Krelove, G. H. Jacoby, and C. Gronwall, “A Measurement of the Contamination in [O III] $\lambda 5007$ Surveys of Intracluster Stars and the Surface Density of $z=3.13$ $\text{Ly}\alpha$ Galaxies,” , vol. 566, no. 2, pp. 784–793, Feb. 2002. [Cited on page 4.]
- [24] J. S. Speagle, C. L. Steinhardt, P. L. Capak, and J. D. Silverman, “A Highly Consistent Framework for the Evolution of the Star-Forming “Main Sequence” from $z \sim 0-6$,” , vol. 214, no. 2, p. 15, Oct. 2014. [Cited on page 4.]
- [25] M. Dijkstra, J. X. Prochaska, M. Ouchi, and M. Hayes, “Lyman-alpha as an Astrophysical and Cosmological Tool,” *Saas-Fee Advanced Course*, vol. 46, Jan. 2019. [Cited on page 5.]

- [26] Y. Ono, R. Itoh, T. Shibuya, M. Ouchi, Y. Harikane, S. Yamanaka, A. K. Inoue, T. Amagasa, D. Miura, M. Okura, K. Shimasaku, I. Iwata, Y. Taniguchi, S. Fujimoto, M. Iye, A. T. Jaelani, N. Kashikawa, S. Kikuchihara, S. Kikuta, M. A. R. Kobayashi, H. Kusakabe, C.-H. Lee, Y. Liang, Y. Matsuoka, R. Momose, T. Nagao, K. Nakajima, and K.-i. Tadaki, “SILVERRUSH X: Machine Learning-aided Selection of 9318 LAEs at $z = 2.2, 3.3, 4.9, 5.7, 6.6,$ and 7.0 from the HSC SSP and CHORUS Survey Data,” , vol. 911, no. 2, p. 78, Apr. 2021. [Cited on page 6.]
- [27] K. Shanmugasundararaj, B. Thomas, E. Robinson, K. Gebhardt, E. Mentuch Cooper, and Hetdex Team, “Classification of Lyman-Alpha Emitters in HETDEX Using Un-supervised Machine Learning,” in *American Astronomical Society Meeting Abstracts*, ser. American Astronomical Society Meeting Abstracts, vol. 237, Jan. 2021, p. 135.06. [Cited on page 6.]
- [28] A. Runnholm, M. Hayes, J. Melinder, E. Rivera-Thorsen, G. Östlin, J. Cannon, and D. Kunth, “The Lyman Alpha Reference Sample. X. Predicting Ly α Output from Star-forming Galaxies Using Multivariate Regression,” , vol. 892, no. 1, p. 48, Mar. 2020. [Cited on page 6.]
- [29] L. Napolitano, L. Pentericci, A. Calabrò, P. Santini, M. Castellano, P. Cassata, J. P. U. Fynbo, I. Jung, D. Kashino, S. Mascia, and M. Mignoli, “Identifying Ly α emitter candidates with Random Forest: Learning from galaxies in the CANDELS survey,” , vol. 677, p. A138, Sep. 2023. [Cited on page 6.]
- [30] W. McKinney, “Data structures for statistical computing in python,” in *SciPy*, 2010. [Cited on page 7.]
- [31] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” , vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Cited on page 7.]
- [32] J. D. Hunter, “Matplotlib: A 2d graphics environment,” *Computing in Science and Engg.*, vol. 9, no. 3, p. 90–95, may 2007. [Cited on page 7.]

- [33] M. Waskom, O. Botvinnik, M. Gelbart, J. Ostblom, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. Warmenhoven, J. B. Cole, J. De Ruiter, J. Vanderplas, S. Hoyer, C. Pye, A. Miles, C. Swain, K. Meyer, M. Martin, E. Bachant, Pete Quintero, G. Kunter, S. Villalba, Brian, C. Fitzgerald, C. G. Evans, M. L. Williams, D. O’Kane, T. Yarkoni, and T. Brunner, “seaborn: Statistical data visualization,” *Astrophysics Source Code Library*, record ascl:2012.015, Dec. 2020. [Cited on page 7.]
- [34] P. A. C. Cunha and A. Humphrey, “Photometric redshift-aided classification using ensemble learning,” , vol. 666, p. A87, Oct. 2022. [Cited on page 9.]
- [35] R. Carvajal, I. Matute, J. Afonso, R. P. Norris, K. J. Luken, P. Sánchez-Sáez, P. A. C. Cunha, A. Humphrey, H. Messias, S. Amarantidis, D. Barbosa, H. A. Cruz, H. Miranda, A. Paulino-Afonso, and C. Pappalardo, “Selection of powerful radio galaxies with machine learning,” , vol. 679, p. A101, Nov. 2023. [Cited on page 9.]
- [36] A. Humphrey, P. A. C. Cunha, A. Paulino-Afonso, S. Amarantidis, R. Carvajal, J. M. Gomes, I. Matute, and P. Papaderos, “Improving machine learning-derived photometric redshifts and physical property estimates using unlabelled observations,” , vol. 520, no. 1, pp. 305–313, Mar. 2023. [Cited on page 9.]
- [37] D. Baron, “Machine Learning in Astronomy: a practical overview,” *arXiv e-prints*, p. arXiv:1904.07248, Apr. 2019. [Cited on pages 9 and 18.]
- [38] M. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. Suganthan, “Ensemble deep learning: A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, 10 2022. [Cited on page 10.]
- [39] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, p. 5–32, oct 2001. [Cited on pages 11 and 38.]
- [40] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, vol. 29, pp. 1189–1232, 2000. [Cited on pages 11 and 14.]
- [41] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in Neurorobotics*, vol. 7, 2013. [Online]. Available: <https://api.semanticscholar.org/CorpusID:16864990> [Cited on page 11.]

- [42] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “Catboost: unbiased boosting with categorical features,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS’18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 6639–6649. [Cited on pages 14 and 37.]
- [43] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’16. New York, NY, USA: Association for Computing Machinery, 2016, p. 785–794. [Cited on pages 14 and 37.]
- [44] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: a highly efficient gradient boosting decision tree,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157. [Cited on pages 14, 15, and 37.]
- [45] “Lgbm vs xgboost,” <https://neptune.ai/blog/xgboost-vs-lightgbm>, accessed: 2024-10-02. [Cited on page 14.]
- [46] “Catboost over xgboost and lgbm,” <https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm>, accessed: 2024-10-02. [Cited on page 14.]
- [47] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Scientific Reports*, vol. 14, p. 6086, Mar. 2024. [Cited on page 15.]
- [48] “Evaluation metrics,” <https://www.aiacceleratorinstitute.com/evaluating-machine-learning-models-metrics-and-techniques/>, accessed: 2024-10-02. [Cited on page 15.]
- [49] V. R. Joseph, “Optimal ratio for data splitting,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, 2019. [Cited on page 19.]
- [50] J. M. Gorriz, F. Segovia, J. Ramirez, A. Ortiz, and J. Suckling, “Is K-fold cross validation the best model selection method for Machine Learning?” *arXiv e-prints*, p. arXiv:2401.16407, Jan. 2024. [Cited on page 19.]
- [51] T. Yu and H. Zhu, “Hyper-parameter optimization: A review of algorithms and applications,” *ArXiv*, vol. abs/2003.05689, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:212675087> [Cited on page 20.]

- [52] M. Sipper, “High per parameter: A large-scale study of hyperparameter tuning for machine learning algorithms,” *ArXiv*, vol. abs/2207.06028, 2022. [Cited on page 21.]
- [53] C. Bentéjac, A. Csörgo, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, pp. 1937 – 1967, 2019. [Cited on page 21.]
- [54] N. Scoville, H. Aussel, M. Brusa, P. Capak, C. M. Carollo, M. Elvis, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, J. P. Kneib, O. LeFevre, S. J. Lilly, B. Mobasher, A. Renzini, R. M. Rich, D. B. Sanders, E. Schinnerer, D. Schminovich, P. Shopbell, Y. Taniguchi, and N. D. Tyson, “The Cosmic Evolution Survey (COSMOS): Overview,” , vol. 172, no. 1, pp. 1–8, Sep. 2007. [Cited on page 25.]
- [55] J. R. Weaver, O. B. Kauffmann, O. Ilbert, H. J. McCracken, A. Moneti, S. Toft, G. Brammer, M. Shuntov, I. Davidzon, B. C. Hsieh, C. Laigle, A. Anastasiou, C. K. Jespersen, J. Vinther, P. Capak, C. M. Casey, C. J. R. McPartland, B. Milvang-Jensen, B. Mobasher, D. B. Sanders, L. Zalesky, S. Arnouts, H. Aussel, J. S. Dunlop, A. Faisst, M. Franx, L. J. Furtak, J. P. U. Fynbo, K. M. L. Gould, T. R. Greve, S. Gwyn, J. S. Kartaltepe, D. Kashino, A. M. Koekemoer, V. Kokorev, O. Le Fèvre, S. Lilly, D. Masters, G. Magdis, V. Mehta, Y. Peng, D. A. Riechers, M. Salvato, M. Sawicki, C. Scarlata, N. Scoville, R. Shirley, J. D. Silverman, A. Sneppen, V. Smolčić, C. Steinhart, D. Stern, M. Tanaka, Y. Taniguchi, H. I. Teplitz, M. Vaccari, W. H. Wang, and G. Zamorani, “COSMOS2020: A Panchromatic View of the Universe to $z \sim 10$ from Two Complementary Catalogs,” , vol. 258, no. 1, p. 11, Jan. 2022. [Cited on pages 25, 26, and 30.]
- [56] S. Arnouts, L. Moscardini, E. Vanzella, S. Colombi, S. Cristiani, A. Fontana, E. Giallongo, S. Matarrese, and P. Saracco, “Measuring the redshift evolution of clustering: the Hubble Deep Field South,” , vol. 329, no. 2, pp. 355–366, Jan. 2002. [Cited on page 26.]
- [57] O. Ilbert, S. Arnouts, H. J. McCracken, M. Bolzonella, E. Bertin, O. Le Fèvre, Y. Mellier, G. Zamorani, R. Pellò, A. Iovino, L. Tresse, V. Le Brun, D. Bottini, B. Garilli, D. Maccagni, J. P. Picat, R. Scaramella, M. Scodeggio, G. Vettolani, A. Zanichelli, C. Adami, S. Bardelli, A. Cappi, S. Charlot, P. Ciliegi, T. Contini, O. Cucciati, S. Foucaud, P. Franzetti, I. Gavignaud, L. Guzzo, B. Marano, C. Marinoni, A. Mazure,

- B. Meneux, R. Merighi, S. Paltani, A. Pollo, L. Pozzetti, M. Radovich, E. Zucca, M. Bondi, A. Bongiorno, G. Busarello, S. de La Torre, L. Gregorini, F. Lamareille, G. Mathez, P. Merluzzi, V. Ripepi, D. Rizzo, and D. Vergani, “Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey,” , vol. 457, no. 3, pp. 841–856, Oct. 2006. [Cited on page 26.]
- [58] G. B. Brammer, P. G. van Dokkum, and P. Coppi, “EAZY: A Fast, Public Photometric Redshift Code,” , vol. 686, no. 2, pp. 1503–1513, Oct. 2008. [Cited on page 26.]
- [59] S. Miyazaki, M. Oguri, T. Hamana, M. Shirasaki, M. Koike, Y. Komiyama, K. Umetsu, Y. Utsumi, N. Okabe, S. More, E. Medezinski, Y.-T. Lin, H. Miyatake, H. Murayama, N. Ota, and I. Mitsuishi, “A large sample of shear-selected clusters from the Hyper Suprime-Cam Subaru Strategic Program S16A Wide field mass maps,” , vol. 70, p. S27, Jan. 2018. [Cited on page 30.]
- [60] Y. Taniguchi, N. Scoville, T. Murayama, D. B. Sanders, B. Mobasher, H. Aussel, P. Capak, M. Ajiki, S. Miyazaki, Y. Komiyama, Y. Shioya, T. Nagao, S. S. Sasaki, J. Koda, C. Carilli, M. Giavalisco, L. Guzzo, G. Hasinger, C. Impey, O. LeFevre, S. Lilly, A. Renzini, M. Rich, E. Schinnerer, P. Shopbell, N. Kaifu, H. Karoji, N. Arimoto, S. Okamura, and K. Ohta, “The Cosmic Evolution Survey (COSMOS): Subaru Observations of the HST Cosmos Field,” , vol. 172, no. 1, pp. 9–28, Sep. 2007. [Cited on page 30.]
- [61] J. Brinchmann, S. Charlot, S. D. M. White, C. Tremonti, G. Kauffmann, T. Heckman, and J. Brinkmann, “The physical properties of star-forming galaxies in the low-redshift Universe,” , vol. 351, no. 4, pp. 1151–1179, Jul. 2004. [Cited on page 31.]
- [62] K. G. Noeske, B. J. Weiner, S. M. Faber, C. Papovich, D. C. Koo, R. S. Somerville, K. Bundy, C. J. Conselice, J. A. Newman, D. Schiminovich, E. Le Floch, A. L. Coil, G. H. Rieke, J. M. Lotz, J. R. Primack, P. Barmby, M. C. Cooper, M. Davis, R. S. Ellis, G. G. Fazio, P. Guhathakurta, J. Huang, S. A. Kassin, D. C. Martin, A. C. Phillips, R. M. Rich, T. A. Small, C. N. A. Willmer, and G. Wilson, “Star Formation in AEGIS Field Galaxies since $z=1.1$: The Dominance of Gradually Declining Star Formation, and the Main Sequence of Star-forming Galaxies,” , vol. 660, no. 1, pp. L43–L46, May 2007. [Cited on page 31.]

- [63] C. Schreiber, M. Pannella, D. Elbaz, M. Béthermin, H. Inami, M. Dickinson, B. Mag-nelli, T. Wang, H. Aussel, E. Daddi, S. Juneau, X. Shu, M. T. Sargent, V. Buat, S. M. Faber, H. C. Ferguson, M. Giavalisco, A. M. Koekemoer, G. Magdis, G. E. Morrison, C. Papovich, P. Santini, and D. Scott, “The Herschel view of the dominant mode of galaxy growth from $z = 4$ to the present day,” , vol. 575, p. A74, Mar. 2015. [Cited on page 33.]
- [64] R. J. Williams, R. F. Quadri, M. Franx, P. van Dokkum, and I. Labbé, “Detection of Quiescent Galaxies in a Bicolor Sequence from $Z = 0-2$,” , vol. 691, no. 2, pp. 1879–1895, Feb. 2009. [Cited on page 32.]
- [65] S. Arnouts, E. Le Floch, J. Chevallard, B. D. Johnson, O. Ilbert, M. Treyer, H. Aussel, P. Capak, D. B. Sanders, N. Scoville, H. J. McCracken, B. Milliard, L. Pozzetti, and M. Salvato, “Encoding of the infrared excess in the NUVrK color diagram for star-forming galaxies,” , vol. 558, p. A67, Oct. 2013. [Cited on pages 32 and 34.]
- [66] M. Bahmani, R. E. Shawi, N. Potikyan, and S. Sakr, “To tune or not to tune? an approach for recommending important hyperparameters,” *ArXiv*, vol. abs/2108.13066, 2021. [Cited on page 39.]
- [67] P. Probst, A. Boulesteix, and B. Bischl, “Tunability: Importance of hyperparameters of machine learning algorithms,” *J. Mach. Learn. Res.*, vol. 20, pp. 53:1–53:32, 2018. [Cited on page 39.]
- [68] K. Gebhardt, E. Mentuch Cooper, R. Ciardullo, V. Acquaviva, R. Bender, W. P. Bowman, B. G. Castanheira, G. Dalton, D. Davis, R. S. de Jong, D. L. DePoy, Y. Devarakonda, S. Dongsheng, N. Drory, M. Fabricius, D. J. Farrow, J. Feldmeier, S. L. Finkelstein, C. S. Froning, E. Gawiser, C. Gronwall, L. Herold, G. J. Hill, U. Hopp, L. R. House, S. Janowiecki, M. Jarvis, D. Jeong, S. Jogee, R. Kakuma, A. Kelz, W. Kol-latschny, E. Komatsu, M. Krumpe, M. Landriau, C. Liu, M. L. Niemeyer, P. MacQueen, J. Marshall, K. Mawatari, E. M. McLinden, S. Mukae, G. Nagaraj, Y. Ono, M. Ouchi, C. Papovich, N. Sakai, S. Saito, D. P. Schneider, A. Schulze, K. Shan-mugasundararaj, M. Shetrone, C. Sneden, J. Snigula, M. Steinmetz, B. P. Thomas, B. Thomas, S. Tuttle, T. Urrutia, L. Wisotzki, I. Wold, G. Zeimann, and Y. Zhang, “The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX) Survey Design, Reductions, and Detections,” , vol. 923, no. 2, p. 217, Dec. 2021. [Cited on pages 46 and 71.]

- [69] The HETDEX collaboration, E. Mentuch Cooper, K. Gebhardt, D. Davis, D. Farrow, C. Liu, G. Zeimann, R. Ciardullo, J. Feldmeier, N. Drory, D. Jeong, B. Benda, W. Bowman, M. Boylan-Kolchin, Ó. Chávez Ortiz, M. Debski, M. Dentler, M. Fabricius, R. Farooq, S. Finkelstein, E. Gawiser, C. Gronwall, G. Hill, U. Hopp, L. House, S. Janowiecki, H. Khoraminezhad, W. Kollatschny, E. Komatsu, M. Landriau, M. Niemeyer, H. Lee, P. MacQueen, K. Mawatari, B. McKay, M. Ouchi, J. Poppe, S. Saito, D. Schneider, J. Snigula, B. Thomas, S. Tuttle, T. Urrutia, L. Weiss, L. Wisotzki, and Y. Zhang, “Hetdex public source catalog 1: 220 k sources including over 50 k ly emitters from an untargeted wide-area spectroscopic survey,” *Astrophysical Journal*, vol. 943, no. 2, Feb. 2023, publisher Copyright: © 2023. The Author(s). Published by the American Astronomical Society. [Cited on page 46.]
- [70] M. Taylor, “TOPCAT: Desktop Exploration of Tabular Data for Astronomy and Beyond,” *arXiv e-prints*, p. arXiv:1707.02160, Jul. 2017. [Cited on page 46.]
- [71] Euclid Collaboration, R. Scaramella, J. Amiaux, Y. Mellier, C. Burigana, C. S. Carvalho, J. C. Cuillandre, A. Da Silva, A. Derosa, J. Dinis, E. Maiorano, M. Maris, I. Tereno, R. Laureijs, T. Boenke, G. Buenadicha, X. Dupac, L. M. Gaspar Venancio, P. Gómez-Álvarez, J. Hoar, J. Lorenzo Alvarez, G. D. Racca, G. Saavedra-Criado, J. Schwartz, R. Vavrek, M. Schirmer, H. Aussel, R. Azzollini, V. F. Cardone, M. Cropper, A. Ealet, B. Garilli, W. Gillard, B. R. Granett, L. Guzzo, H. Hoekstra, K. Jahnke, T. Kitching, T. Maciaszek, M. Meneghetti, L. Miller, R. Nakajima, S. M. Niemi, F. Pasian, W. J. Percival, S. Pottinger, M. Sauvage, M. Scodeggio, S. Wachter, A. Zachei, N. Aghanim, A. Amara, T. Auphan, N. Auricchio, S. Awan, A. Balestra, R. Bender, C. Bodendorf, D. Bonino, E. Branchini, S. Brau-Nogue, M. Brescia, G. P. Candini, V. Capobianco, C. Carbone, R. G. Carlberg, J. Carretero, R. Casas, F. J. Castander, M. Castellano, S. Cavioti, A. Cimatti, R. Cledassou, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, A. Costille, F. Courbin, H. Degaudenzi, M. Douspis, F. Dubath, C. A. J. Duncan, S. Dusini, S. Farrens, S. Ferriol, P. Fosalba, N. Fourmanoit, M. Frailis, E. Franceschi, P. Franzetti, M. Fumana, B. Gillis, C. Giocoli, A. Grazian, F. Grupp, S. V. H. Haugan, W. Holmes, F. Hormuth, P. Hudelot, S. Kermiche, A. Kiessling, M. Kilbinger, R. Kohley, B. Kubik, M. Kümmel, M. Kunz, H. Kurki-Suonio, O. Lahav, S. Ligi, P. B. Lilje, I. Lloro, O. Mansutti, O. Marggraf, K. Markovic, F. Marulli, R. Massey, S. Maurogordato, M. Melchior, E. Merlin, G. Meylan, J. J. Mohr, M. Moresco, B. Morin, L. Moscardini, E. Munari, R. C. Nichol,

C. Padilla, S. Paltani, J. Peacock, K. Pedersen, V. Pettorino, S. Pires, M. Poncet, L. Popa, L. Pozzetti, F. Raison, R. Rebolo, J. Rhodes, H. W. Rix, M. Roncarelli, E. Rossetti, R. Saglia, P. Schneider, T. Schrabback, A. Secroun, G. Seidel, S. Serrano, C. Sirignano, G. Sirri, J. Skottfelt, L. Stanco, J. L. Starck, P. Tallada-Crespí, D. Tavagnacco, A. N. Taylor, H. I. Teplitz, R. Toledo-Moreo, F. Torradeflot, M. Trifoglio, E. A. Valentijn, L. Valenziano, G. A. Verdoes Kleijn, Y. Wang, N. Welikala, J. Weller, M. Wetzstein, G. Zamorani, J. Zoubian, S. Andreon, M. Baldi, S. Bardelli, A. Boucaud, S. Camera, D. Di Ferdinando, G. Fabbian, R. Farinelli, S. Galeotta, J. Graciá-Carpio, D. Maino, E. Medinaceli, S. Mei, C. Neissner, G. Polenta, A. Renzi, E. Romelli, C. Rosset, F. Sureau, M. Tenti, T. Vassallo, E. Zucca, C. Baccigalupi, A. Balaguera-Antolínez, P. Battaglia, A. Biviano, S. Borgani, E. Bozzo, R. Cabanac, A. Cappi, S. Casas, G. Castignani, C. Colodro-Conde, J. Coupon, H. M. Courtois, J. Cuby, S. de la Torre, S. Desai, H. Dole, M. Fabricius, M. Farina, P. G. Ferreira, F. Finelli, P. Flose-Reimberg, S. Fotopoulou, K. Ganga, G. Gozaliasl, I. M. Hook, E. Keihanen, C. C. Kirkpatrick, P. Liebing, V. Lindholm, G. Mainetti, M. Martinelli, N. Martinet, M. Maturi, H. J. McCracken, R. B. Metcalf, G. Morgante, J. Nightingale, A. Nucita, L. Patrizii, D. Potter, G. Riccio, A. G. Sánchez, D. Sapone, J. A. Schewtschenko, M. Schultheis, V. Scottez, R. Teyssier, I. Tutusaus, J. Valiviita, M. Viel, W. Vriend, and L. Whittaker, “Euclid preparation. I. The Euclid Wide Survey,” , vol. 662, p. A112, Jun. 2022. [Cited on page 49.]

- [72] Ž. Ivezić, S. M. Kahn, J. A. Tyson, B. Abel, E. Acosta, R. Allsman, D. Alonso, Y. AlSayyad, S. F. Anderson, J. Andrew, J. R. P. Angel, G. Z. Angeli, R. Ansari, P. Antilogus, C. Araujo, R. Armstrong, K. T. Arndt, P. Astier, É. Aubourg, N. Auza, T. S. Axelrod, D. J. Bard, J. D. Barr, A. Barrau, J. G. Bartlett, A. E. Bauer, B. J. Bauman, S. Baumont, E. Bechtol, K. Bechtol, A. C. Becker, J. Becla, C. Beldica, S. Bellavia, F. B. Bianco, R. Biswas, G. Blanc, J. Blazek, R. D. Blandford, J. S. Bloom, J. Bogart, T. W. Bond, M. T. Booth, A. W. Borgland, K. Borne, J. F. Bosch, D. Boutigny, C. A. Brackett, A. Bradshaw, W. N. Brandt, M. E. Brown, J. S. Bullock, P. Burchat, D. L. Burke, G. Cagnoli, D. Calabrese, S. Callahan, A. L. Callen, J. L. Carlin, E. L. Carlson, S. Chandrasekharan, G. Charles-Emerson, S. Chesley, E. C. Cheu, H.-F. Chiang, J. Chiang, C. Chirino, D. Chow, D. R. Ciardi, C. F. Claver, J. Cohen-Tanugi, J. J. Cockrum, R. Coles, A. J. Connolly, K. H. Cook, A. Cooray, K. R. Covey, C. Cribbs, W. Cui, R. Cutri, P. N.

Daly, S. F. Daniel, F. Daruich, G. Daubard, G. Daues, W. Dawson, F. Delgado, A. Delapenna, R. de Peyster, M. de Val-Borro, S. W. Digel, P. Doherty, R. Dubois, G. P. Dubois-Felsmann, J. Durech, F. Economou, T. Eifler, M. Eracleous, B. L. Emmons, A. Fausti Neto, H. Ferguson, E. Figueroa, M. Fisher-Levine, W. Focke, M. D. Foss, J. Frank, M. D. Freemon, E. Gangler, E. Gawiser, J. C. Geary, P. Gee, M. Geha, C. J. B. Gessner, R. R. Gibson, D. K. Gilmore, T. Glanzman, W. Glick, T. Goldina, D. A. Goldstein, I. Goodenow, M. L. Graham, W. J. Gressler, P. Gris, L. P. Guy, A. Guyonnet, G. Haller, R. Harris, P. A. Hascall, J. Haupt, F. Hernandez, S. Herrmann, E. Hileman, J. Hoblitt, J. A. Hodgson, C. Hogan, J. D. Howard, D. Huang, M. E. Huffer, P. Ingraham, W. R. Innes, S. H. Jacoby, B. Jain, F. Jammes, M. J. Jee, T. Jenness, G. Jernigan, D. Jevremović, K. Johns, A. S. Johnson, M. W. G. Johnson, R. L. Jones, C. Juramy-Gilles, M. Jurić, J. S. Kalirai, N. J. Kallivayalil, B. Kalmbach, J. P. Kantor, P. Karst, M. M. Kasliwal, H. Kelly, R. Kessler, V. Kinnison, D. Kirkby, L. Knox, I. V. Kotov, V. L. Krabbendam, K. S. Krughoff, P. Kubánek, J. Kuczewski, S. Kulkarni, J. Ku, N. R. Kurita, C. S. Lage, R. Lambert, T. Lange, J. B. Langton, L. Le Guillou, D. Levine, M. Liang, K.-T. Lim, C. J. Lintott, K. E. Long, M. Lopez, P. J. Lotz, R. H. Lupton, N. B. Lust, L. A. MacArthur, A. Mahabal, R. Mandelbaum, T. W. Markiewicz, D. S. Marsh, P. J. Marshall, S. Marshall, M. May, R. McKercher, M. McQueen, J. Meyers, M. Migliore, M. Miller, D. J. Mills, C. Miraval, J. Moeyens, F. E. Moolekamp, D. G. Monet, M. Moniez, S. Monkewitz, C. Montgomery, C. B. Morrison, F. Mueller, G. P. Muller, F. Muñoz Arancibia, D. R. Neill, S. P. Newbry, J.-Y. Nief, A. Nomerotski, M. Nordby, P. O'Connor, J. Oliver, S. S. Olivier, K. Olsen, W. O'Mullane, S. Ortiz, S. Osier, R. E. Owen, R. Pain, P. E. Palecek, J. K. Parejko, J. B. Parsons, N. M. Pease, J. M. Peterson, J. R. Peterson, D. L. Petravick, M. E. Libby Petrick, C. E. Petry, F. Pierfederici, S. Pietrowicz, R. Pike, P. A. Pinto, R. Plante, S. Plate, J. P. Plutchak, P. A. Price, M. Prouza, V. Radeka, J. Rajagopal, A. P. Rasmussen, N. Regnault, K. A. Reil, D. J. Reiss, M. A. Reuter, S. T. Ridgway, V. J. Riot, S. Ritz, S. Robinson, W. Roby, A. Roodman, W. Rosing, C. Roucelle, M. R. Rumore, S. Russo, A. Saha, B. Sassolas, T. L. Schalk, P. Schellart, R. H. Schindler, S. Schmidt, D. P. Schneider, M. D. Schneider, W. Schoening, G. Schumacher, M. E. Schwamb, J. Sebag, B. Selvy, G. H. Sembroski, L. G. Seppala, A. Serio, E. Serrano, R. A. Shaw, I. Shipsey, J. Sick, N. Silvestri, C. T. Slater, J. A. Smith, R. C. Smith, S. Sobhani, C. Soldahl, L. Storrie-Lombardi, E. Stover, M. A. Strauss, R. A. Street, C. W. Stubbs, I. S. Sullivan, D. Sweeney, J. D. Swinbank,

- A. Szalay, P. Takacs, S. A. Tether, J. J. Thaler, J. G. Thayer, S. Thomas, A. J. Thornton, V. Thukral, J. Tice, D. E. Trilling, M. Turri, R. Van Berg, D. Vanden Berk, K. Vetter, F. Virieux, T. Vucina, W. Wahl, L. Walkowicz, B. Walsh, C. W. Walter, D. L. Wang, S.-Y. Wang, M. Warner, O. Wiecha, B. Willman, S. E. Winters, D. Wittman, S. C. Wolff, W. M. Wood-Vasey, X. Wu, B. Xin, P. Yoachim, and H. Zhan, “LSST: From Science Drivers to Reference Design and Anticipated Data Products,” , vol. 873, no. 2, p. 111, Mar. 2019. [Cited on page 49.]
- [73] Dark Energy Survey Collaboration, T. Abbott, F. B. Abdalla, J. Aleksić, S. Allam, A. Amara, D. Bacon, E. Balbinot, M. Banerji, K. Bechtol, A. Benoit-Lévy, G. M. Bernstein, E. Bertin, J. Blazek, C. Bonnett, S. Bridle, D. Brooks, R. J. Brunner, E. Buckley-Geer, D. L. Burke, G. B. Caminha, D. Capozzi, J. Carlsen, A. Carnero-Rosell, M. Carollo, M. Carrasco-Kind, J. Carretero, F. J. Castander, L. Clerkin, T. Collett, C. Conselice, M. Crocce, C. E. Cunha, C. B. D’Andrea, L. N. da Costa, T. M. Davis, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, J. Estrada, J. Etherington, A. E. Evrard, J. Fabbri, D. A. Finley, B. Flaugher, R. J. Foley, P. Fosalba, J. Frieman, J. García-Bellido, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, D. A. Goldstein, D. Gruen, R. A. Gruendl, P. Guarnieri, G. Gutierrez, W. Hartley, K. Honscheid, B. Jain, D. J. James, T. Jeltema, S. Jouvel, R. Kessler, A. King, D. Kirk, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, M. Lima, H. Lin, M. A. G. Maia, M. Makler, M. Manera, C. Maraston, J. L. Marshall, P. Martini, R. G. McMahon, P. Melchior, A. Merson, C. J. Miller, R. Miquel, J. J. Mohr, X. Morice-Atkinson, K. Naidoo, E. Neilsen, R. C. Nichol, B. Nord, R. Ogando, F. Ostrovski, A. Palmese, A. Papadopoulos, H. V. Peiris, J. Peoples, W. J. Percival, A. A. Plazas, S. L. Reed, A. Refregier, A. K. Romer, A. Roodman, A. Ross, E. Roza, E. S. Rykoff, I. Sadeh, M. Sako, C. Sánchez, E. Sanchez, B. Santiago, V. Scarpine, M. Schubnell, I. Sevilla-Noarbe, E. Sheldon, M. Smith, R. C. Smith, M. Soares-Santos, F. Sobreira, M. Soumagnac, E. Suchyta, M. Sullivan, M. Swanson, G. Tarle, J. Thaler, D. Thomas, R. C. Thomas, D. Tucker, J. D. Vieira, V. Vikram, A. R. Walker, R. H. Wechsler, J. Weller, W. Wester, L. Whiteway, H. Wilcox, B. Yanny, Y. Zhang, and J. Zuntz, “The Dark Energy Survey: more than dark energy - an overview,” , vol. 460, no. 2, pp. 1270–1299, Aug. 2016. [Cited on page 49.]
- [74] D. Spergel, N. Gehrels, C. Baltay, D. Bennett, J. Breckinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin,

- B. Macintosh, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, D. Benford, M. Hudson, W. S. Jeong, Y. Mellier, W. Traub, T. Yamada, P. Capak, J. Colbert, D. Masters, M. Penny, D. Savransky, D. Stern, N. Zimmerman, R. Barry, L. Bartusek, K. Carpenter, E. Cheng, D. Content, F. Dekens, R. Demers, K. Grady, C. Jackson, G. Kuan, J. Kruk, M. Melton, B. Nemati, B. Parvin, I. Poberezhskiy, C. Peddie, J. Ruffa, J. K. Wallace, A. Whipple, E. Wollack, and F. Zhao, “Wide-Field Infrared Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report,” *arXiv e-prints*, p. arXiv:1503.03757, Mar. 2015. [Cited on page 49.]
- [75] N. Benítez, R. Dupke, M. Moles, L. Sodré, A. J. Cenarro, A. Marín Franch, K. Taylor, D. Cristóbal, A. Fernández-Soto, C. Mendes de Oliveira, J. Cepa-Nogué, L. R. Abramo, J. S. Alcaniz, R. Overzier, C. Hernández-Monteagudo, E. J. Alfaro, A. Kanaan, M. Carvano, R. R. R. Reis, and J-PAS Collaboration, “J-PAS: The Javalambre-Physics of the Accelerated Universe Astrophysical Survey,” in *Highlights of Spanish Astrophysics VIII*, May 2015, pp. 148–153. [Cited on page 50.]
- [76] J. Calhau, D. Sobral, S. Santos, J. Matthee, A. Paulino-Afonso, A. Stroe, B. Simmons, C. Barlow-Hall, and B. Adams, “The X-ray and radio activity of typical and luminous Ly α emitters from $z \sim 2$ to $z \sim 6$: evidence for a diverse, evolving population,” vol. 493, no. 3, pp. 3341–3362, Apr. 2020. [Cited on page 58.]
- [77] S. Santos, D. Sobral, J. Matthee, J. Calhau, E. da Cunha, B. Ribeiro, A. Paulino-Afonso, P. Arrabal Haro, and J. Butterworth, “The evolution of rest-frame UV properties, Ly α EWs, and the SFR-stellar mass relation at $z \sim 2$ -6 for SC4K LAEs,” vol. 493, no. 1, pp. 141–160, Mar. 2020. [Cited on page 58.]