University of Central Lancashire
UCLan

SROC
STUDENT RECORDS OFFICERS' CONFERENCE

Combining Student Data and Machine Learning: Predicting End of Year Outcomes

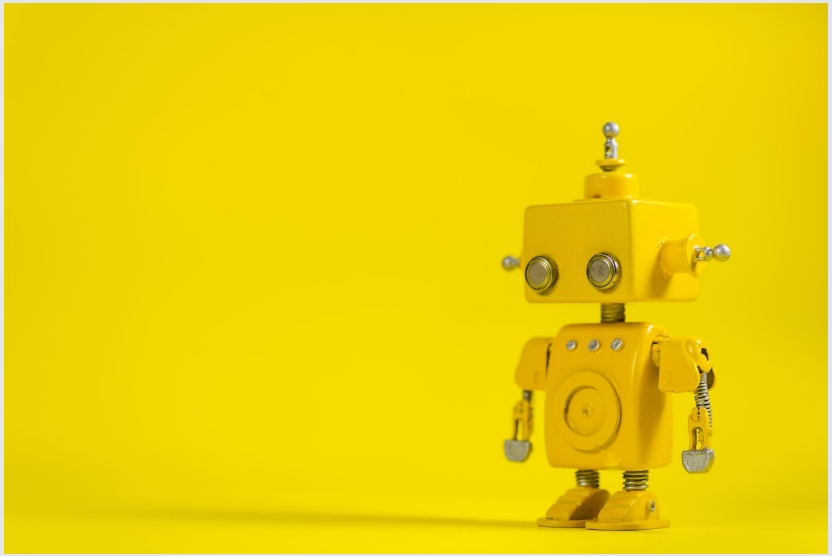
Joe Taylor & Danny Lee

Where opportunity creates success

#sroc24 www.sroc.ac.uk

Intro

- Data and Outcomes in Higher Education
- What Engagement data is available
- Intro to Machine Learning
- Predicting Outcomes with Machine Learning
- Future Work



University of Central Lancashire
UCLan

Student Data in Higher Education

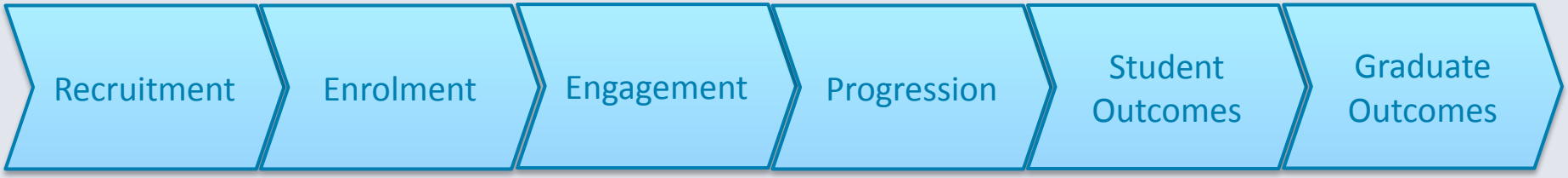
Data Sources in HE

Multiple sources of data across HE:

- HESA
  - Staff, Student & Finance Records
- Internal Banner data
- UCAS data
- OfS data
- University League Table data
- Graduate Outcomes survey
- National Student Survey

Data Across the Student Cycle

University of Central Lancashire
UCLan



- Together, data sources span the full student cycle
- Offers an incredible amount of data
- Our focus for this project is on using Engagement as a predictor for Student Outcomes

Data Across the Student Cycle

University of Central Lancashire
UCLan

- Together, data sources span the full student cycle
- Offers an incredible amount of data
- Our focus for this project is on using Engagement as a predictor for Student Outcomes

University of Central Lancashire
UCLan

Student Outcomes in Higher Education

Higher Education & Outcomes

University of Central Lancashire
UCLan

Many data sources offer different measures of student outcomes in Higher Education:

- OfS: Regulatory measures of Continuation, Completion, and Progression
- Graduate Outcomes survey: Meaningful employment, Highly Skilled / Further Study
- League Tables: Continuation, Graduate Outcomes, Value Added
- Internal: measures of Withdrawals, Non-Returners, Good Honours, etc

Student Outcomes: Reputational

University of Central Lancashire
UCLan

- Complete, Guardian, and Times all use some measure of continuation
- Unhelpfully, the methodology differs slightly across each
- Published data does not reproduce overall scores
- A low score in one metric can substantially impact performance
- It can however highlight broad areas where support could be targeted

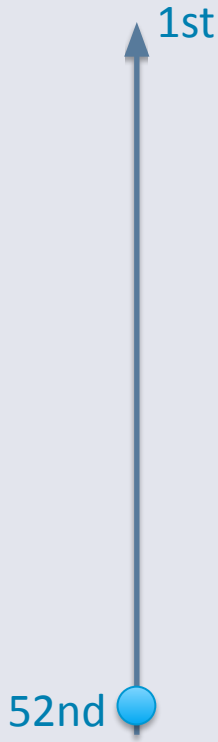
Guardian Subject Ranking Model

University of Central Lancashire
UCLan

- Motivate support by modelling potential reputational impact of improved metric scores
- Use Pandas and Statsmodels packages in Python to approximate overall scores with Ordinary Least Squares regression model
- Chose a recent Guardian subject ranking to model
- Some assumptions needed to substitute missing data; we assume institutions achieve the subject average score where data is missing

Guardian Subject Ranking Model

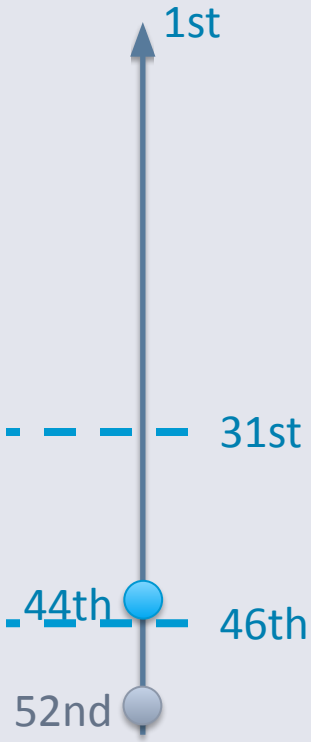
- We can evaluate the model with two statistical measures
- R^2 suggests how well the model fits the data, a score of 1 is a perfect fit, a score of 0 is no fit
  - Our model scores $R^2 = 0.96$
- Root Mean Squared Error (RMSE) measures average difference between predicted and actual scores
  - Our model has RMSE = 3.1
  - Approximate prediction interval of ± 6.2 at 95% level
- Focus on institution with the lowest continuation score in chosen subject which ranked 52nd



Guardian Subject Ranking Model

Scenario 1: Institution achieved subject-average continuation score of 89%

- Model predicts overall score of 50.2 ± 6.2
- Could have achieved a rank of 44th
- Ranges from 46th to 31st at the 95% level



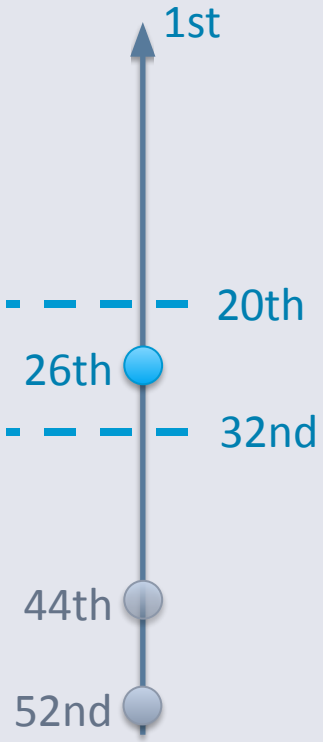
Guardian Subject Ranking Model

University of Central Lancashire
UCLan

Scenario 2: Institution achieved the best continuation score of 100%

- Model predicts an overall score of 61.4 ± 6.2
- Could have achieved a rank of 26th
- Ranges from 32nd to 20th at the 95% level

Summary: Continuation can have significant, meaningful impact on league table rankings



Why is Student Non-continuation important?

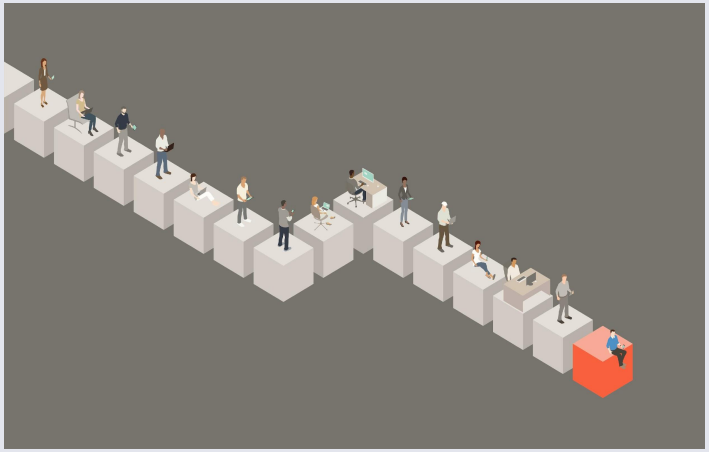
- Non-continuation poses a risk for **most** UK Higher Education providers
- 360 standard students at £9.25k adds up to £10m in revenue
- Non-continuers had **less positive** labour market outcomes than those who **completed** higher education
- Non-continuers also had less positive labour market outcomes than those who had **not entered** higher education

University of Central Lancashire
UCLan

Looking Upstream at Engagement

What data can we use

- The data the students give us at enrolment
- Course / Level / Study Mode
- The engagement data they generate whilst here
  - Attendance
  - Library Visits
  - Interactions with the VLE
  - Graded work
  - Financial data



Existing Tools for Monitoring Engagement

University of Central Lancashire
UCLan

StREAM:

- Student engagement analytics platform
- Allows cohort comparisons with benchmarking

Starfish:

- Case management system
- Lends itself more to operational analytics

Internal Reporting:

- Developed in-house in SSRS / Tableau / PowerBI / Other
- Tailored insights for the institution

Why is early engagement important?

University of Central Lancashire
UCLan

- Results showed that students who obtained the highest end-of-year marks were more likely to be in a higher engagement quintile as early as the first 3–4 weeks

- Students who started in a higher engagement quintile, but where their engagement decreased, were more likely to have higher marks than those that started in a lower quintile and then increased their engagement.

- Early measures of engagement are predictive of future behaviour and of future outcomes. by (Gascoigne & Cole, 2022) and (Summers et al., 2021)

Why is early engagement important?

- Effective intervention with at-risk students can increase course completion rates, retention, academic performance, and overall student success. - (Engage2Serve, 2016)

- The usefulness of the model only becomes realised when combined with human decision making and contact between staff and students.

University of Central Lancashire
UCLan

Machine Learning & Artificial Intelligence

Introduction: What is Machine Learning

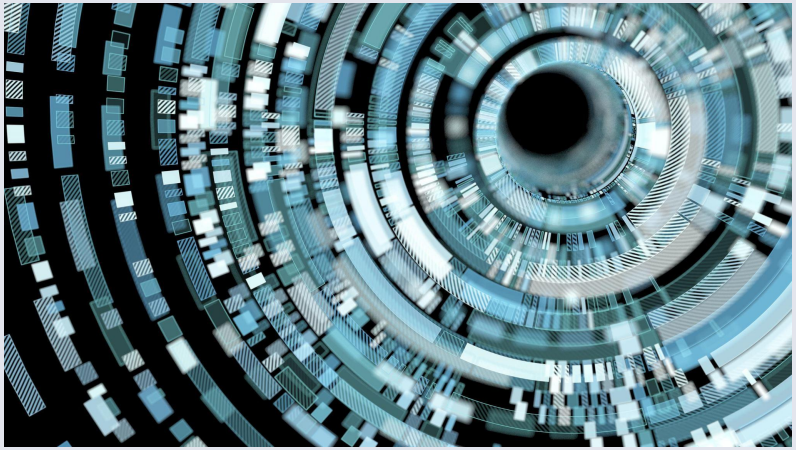
- Machine learning is a branch of artificial intelligence (AI) which focuses on the use of **data and algorithms** to imitate the way that humans learn

- There are **different ways** these algorithms can learn

- We can then **apply** that learning without the need for **human intervention**

Learning Process

University of Central Lancashire
UCLan



- **Supervised Learning.** Trains algorithms using data that has been labelled by the user.
- **Unsupervised Learning.** Finds structure in unlabelled data and makes best guess.
- **Reinforcement Learning.** Trains models through trial and error, a bit of game theory, iterates through multiple generations.

The Elephant in the Room: Generative AI

University of Central Lancashire
UCLan

- Generative AI creates things like images and text
  - Chat GPT
  - DALL-E
  - Midjourney
  - Microsoft's Copilot

- Typically **generates content** in response to a prompt from the user

- In contrast, Machine Learning primarily focusses on making predictions or classifications **based on data**

Difference to Statistics

University of Central Lancashire
UCLan

Ultimately, they're different approaches that could both solve the same problem

Statistical models:

- aim to characterise relationship between the data and our output
- identify the significance of those relationships
- facilitate predictions

Machine learning:

- interested in the performance of predictions
- focus on obtaining a model that can make repeatable predictions
- less focus on the underlying relationship between variables

Use Cases

University of Central Lancashire
UCLan

Machine Learning primarily focusses on making predictions or classifications based on data

- Recognising and classifying handwritten words
- Teaching robots to drive using sensors and cameras
- Suggesting movies or products based on personal preferences (the algorithm)

Student Continuation is a natural extension to this

- Two distinct categories for classification: Continued / Did not Continue
- Predictions target interventions with the intent of improving student prospects and outcomes

What tools enable us to work with ML?

University of Central Lancashire
UCLan

- Not one for Excel
- Tableau Pulse and Copilot for PowerBI offer Generative AI, not ML
- Qlik's Auto ML features, no code needed
  - *Ometis, 2024*
- Largely requires a programming language
- Python probably most accessible though still a bar to entry

Getting Started in Python

University of Central Lancashire
UCLan

- Anaconda Distribution for Python is a user-friendly install experience
  - *https://www.anaconda.com*
- Includes Jupyter Notebooks, an interactive development environment that runs in your web browser
  - *https://jupyter.org*
- Anaconda is a ready-to-use python environment
  - Typically includes Pandas, Numpy, Scikit-Learn, Matplotlib, and many more
  - We use Sweetviz and LazyPredictions in ML model
- Plenty of useful tutorials available on Youtube, LinkedIn Learning, and W3Schools

University of Central Lancashire
UCLan

Machine Learning to Predict Outcomes

How did we decide what data to use

- Exploratory Data Analysis (EDA) – looking at data for completeness
- Examining past work and existing solutions
- Limited by data availability
- Initial Dataset had almost 200 data points per student
- Contained duplicated fields from CRM / HESA data
- Refined to approx. 60 data points with 24 capturing engagement

Approach

University of Central Lancashire
UCLan

- Because we are using historic data to inform the model, this is known as a supervised learning model
- Data points will be taken from UCLan's Data Warehouse
- The project extracted approximately 84000 rows of student data, approximately 43000 individual students across 4 years
- Project data is filtered to only show students who started a course and so completion rate is higher than HESA figures at 83.6%

Approach

University of Central Lancashire
UCLan

- EDA will be performed on these data points to determine **which are suitable** for use in the model
- **Feature selection** will be performed on the dataset to reduce the dimensionality
- Models will be **evaluated** to determine which is the most suitable

Building the classifier

University of Central Lancashire
UCLan

- To build a binary classifier we group end of year outcomes
- This allows the model to choose from a Yes or No value rather than a set of values
- This is a business decision rather than a technical one

End of Year Recommendation	Good Outcome
Awarded	Yes
No Assessment Board Recommendation	Yes
Proceed	Yes
Fail - Recoverable	No
Refer/Defer	No
Withdrawn	No
Fail - Non-Recoverable	No
Exit Award	No
Interruption to Study	No

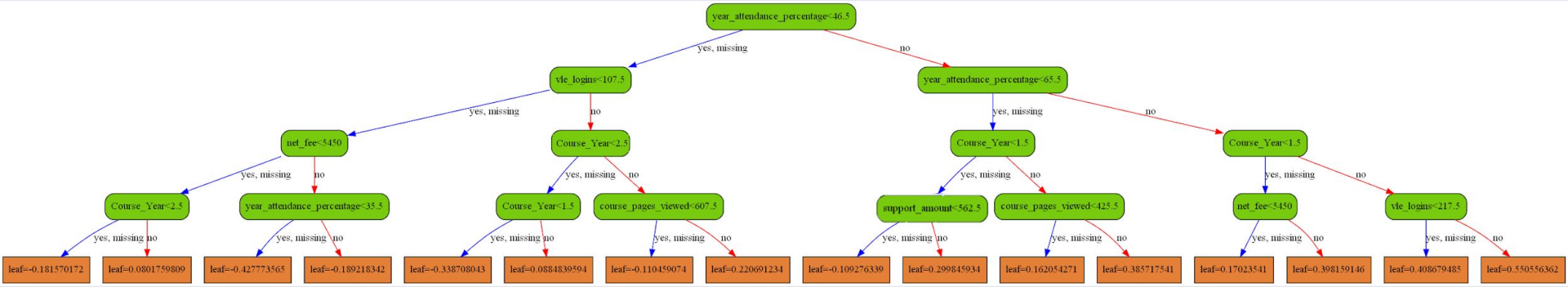
Choosing the model

University of Central Lancashire
UCLan

- 23 classifier models were tested to determine which was the most suitable
- XGBoost was determined as the best for this project
- Individual testing was done to validate results from this package

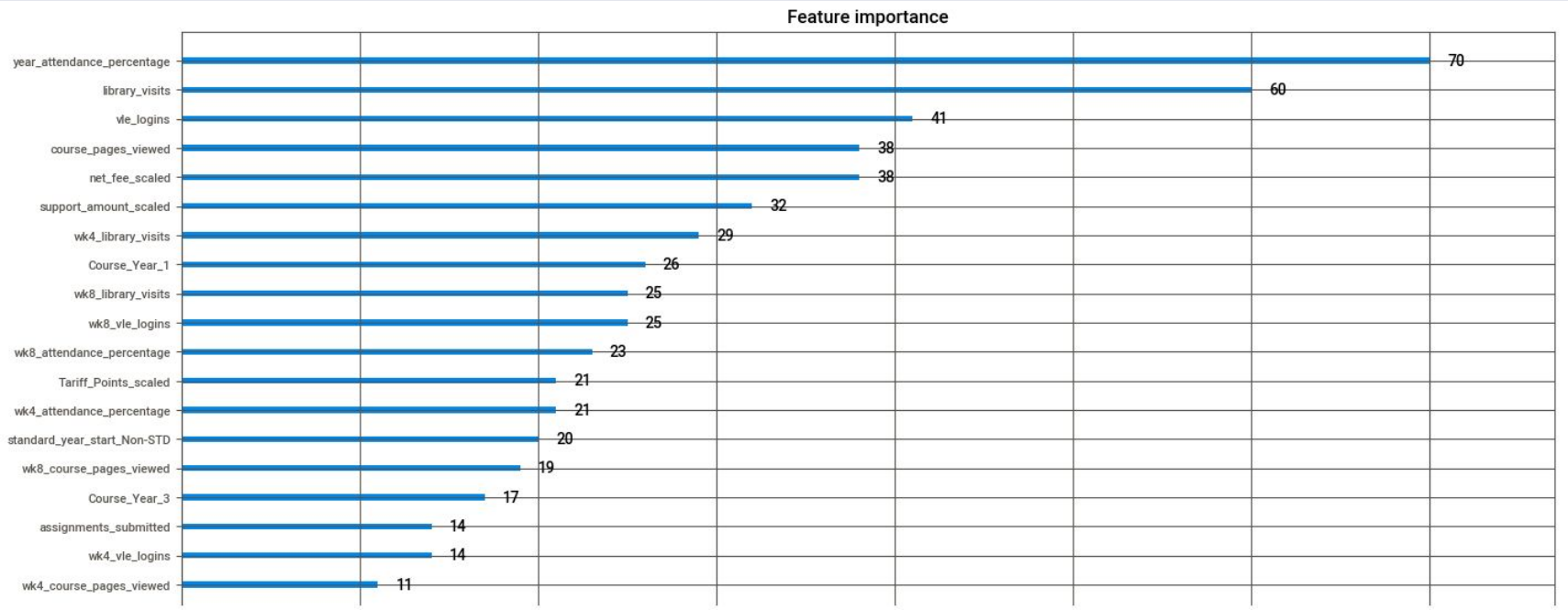
Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
XGBClassifier	0.91	0.74	0.74	0.90	3.81
LGBMClassifier	0.91	0.73	0.73	0.90	1.06
BaggingClassifier	0.89	0.72	0.72	0.89	3.09
NearestCentroid	0.73	0.71	0.71	0.77	0.19
RandomForestClassifier	0.90	0.69	0.69	0.89	6.66
DecisionTreeClassifier	0.85	0.69	0.69	0.85	0.77
BernoulliNB	0.77	0.69	0.69	0.79	0.17
ExtraTreesClassifier	0.90	0.66	0.66	0.88	5.52
AdaBoostClassifier	0.89	0.65	0.65	0.87	2.39
ExtraTreeClassifier	0.83	0.65	0.65	0.83	0.25
SVC	0.89	0.64	0.64	0.87	69.09
GaussianNB	0.46	0.63	0.63	0.52	0.19
KNeighborsClassifier	0.87	0.62	0.62	0.85	46.21
LinearDiscriminantAnalysis	0.87	0.62	0.62	0.85	0.38
LogisticRegression	0.88	0.62	0.62	0.86	0.66
PassiveAggressiveClassifier	0.79	0.61	0.61	0.80	0.24
CalibratedClassifierCV	0.88	0.60	0.60	0.85	41.97
SGDClassifier	0.88	0.60	0.60	0.85	0.76
QuadraticDiscriminantAnalysis	0.34	0.59	0.59	0.37	0.31
LinearSVC	0.88	0.58	0.58	0.85	12.71
Perceptron	0.73	0.57	0.57	0.76	0.22
RidgeClassifier	0.87	0.52	0.52	0.82	0.24
RidgeClassifierCV	0.87	0.52	0.52	0.82	0.33

What does XGBoost do?



- XGBoost is a gradient boosting algorithm
- It creates many decision tree models – weak learners
- It uses loss functions to build better trees
- It combines a result of these trees

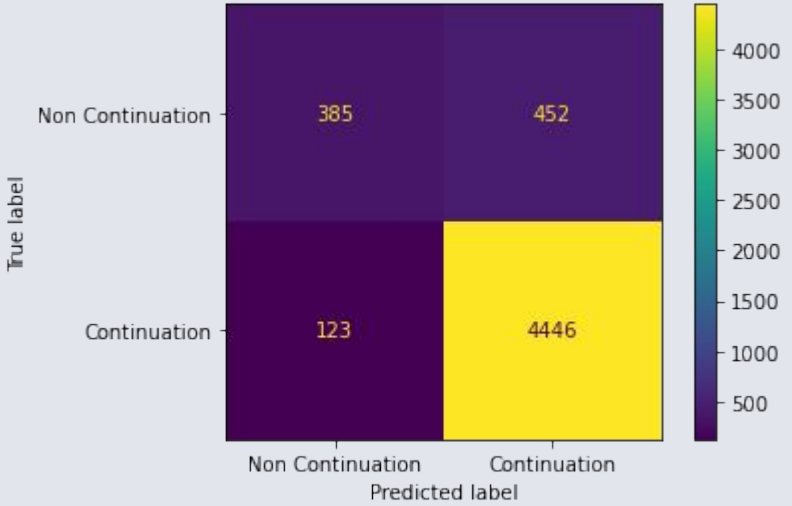
Analysing the important items



Basic Algorithm

- Good outcome correctly classified 97%
- Good outcome incorrectly classified 3%
- Poor outcome correctly classified 46%
- Poor outcome incorrectly classified 54%

- Shows how accuracy can be misrepresented – Overall Accuracy almost 90%



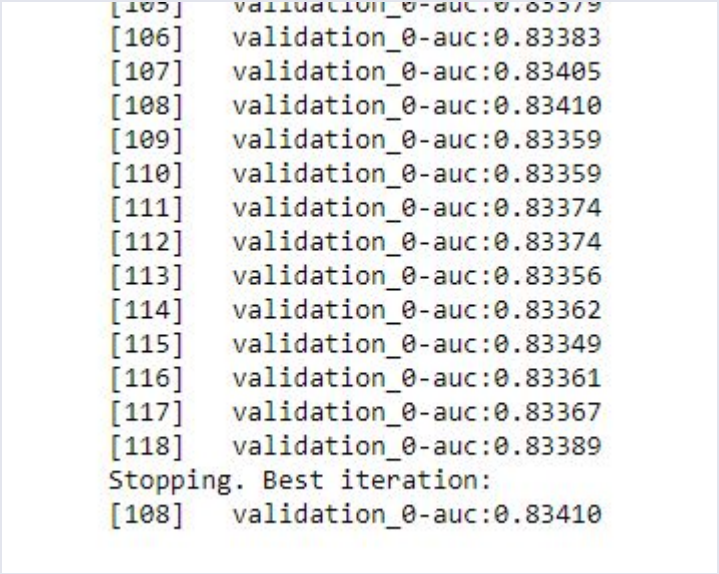
Approach

- The XGBoost model has a number of hyperparameters
- These were trained using a grid search method
- Multiple parameters were tuned over multiple rounds

ght	params	split0_test_score	split1_test_score	mean_test_score	std_test_score	rank_test_score	sp
153	{'gamma': 0.6, 'learning_rate': 0.1, 'max_dept...	0.797715	0.798497	0.798106	0.000391	7	
153	{'gamma': 0.6, 'learning_rate': 0.2, 'max_dept...	0.804290	0.794687	0.799489	0.004802	6	
153	{'gamma': 0.6, 'learning_rate': 0.3, 'max_dept...	0.799619	0.802518	0.801069	0.001450	2	
153	{'gamma': 0.7, 'learning_rate': 0.1, 'max_dept...	0.796944	0.796173	0.796558	0.000385	9	
153	{'gamma': 0.7, 'learning_rate': 0.2, 'max_dept...	0.799929	0.799879	0.799904	0.000025	5	
153	{'gamma': 0.7, 'learning_rate': 0.3, 'max_dept...	0.803009	0.799143	0.801076	0.001933	1	
153	{'gamma': 0.8, 'learning_rate': 0.1, 'max_dept...	0.798445	0.796038	0.797241	0.001203	8	
153	{'gamma': 0.8, 'learning_rate': 0.2, 'max_dept...	0.798691	0.801206	0.799949	0.001257	4	

Approach

- Model trained using these ‘best’ parameters
- Using an early stopping round parameter of 10
- Evaluation method – Area underneath curve

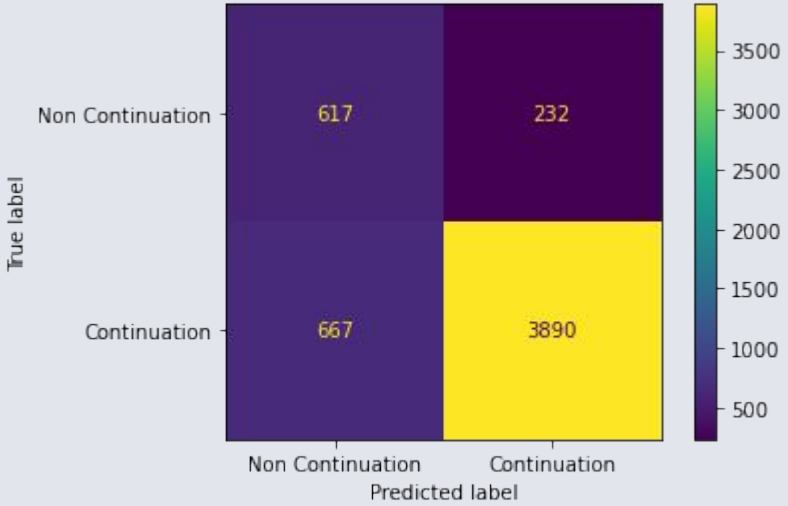


Tuned Algorithm

University of Central Lancashire
UCLan

- Good outcome correctly classified 85%
- Good outcome incorrectly classified 15%
- Poor outcome correctly classified 73%
- Poor outcome incorrectly classified 27%

- Overall Accuracy down to just over 83%
- Balanced Accuracy Score up from 71% to 79%



Algorithm at Week 8

University of Central Lancashire
UCLan

- Good outcome correctly classified 81%
- Good outcome incorrectly classified 19%
- Poor outcome correctly classified 73%
- Poor outcome incorrectly classified 27%

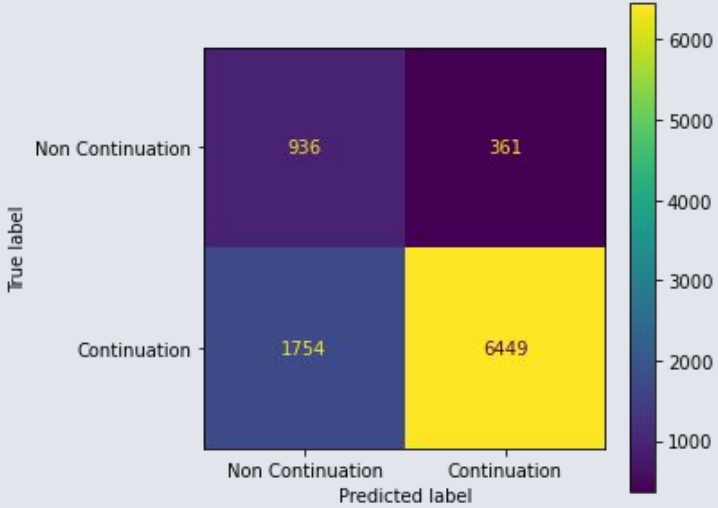
- Overall Accuracy down to 80%
- Balanced Accuracy Score down to 77%

Algorithm at Week 4

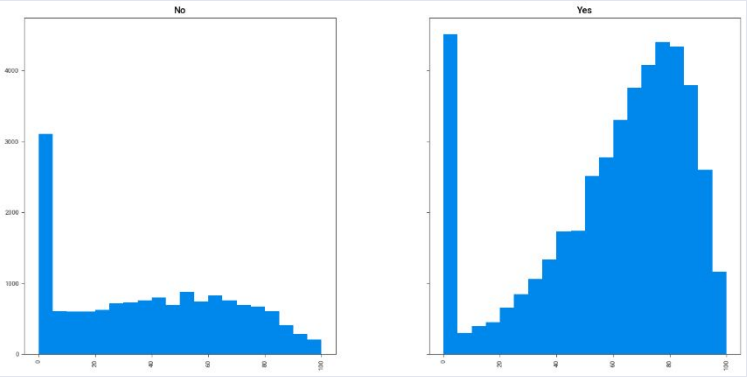
University of Central Lancashire
UCLan

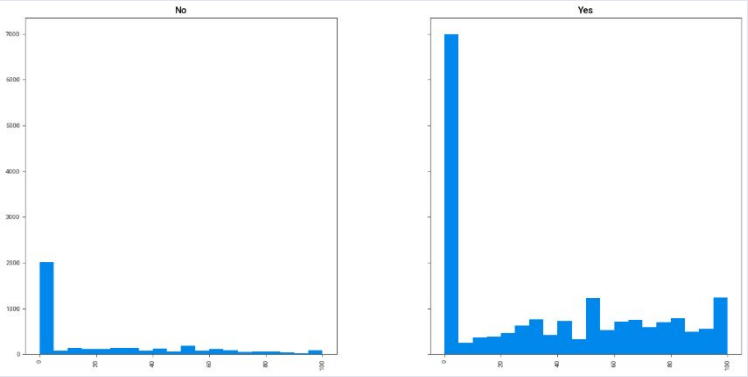
- Good outcome correctly classified 79%
- Good outcome incorrectly classified 21%
- Poor outcome correctly classified 72%
- Poor outcome incorrectly classified 28%

- Overall Accuracy down to 77%
- Balanced Accuracy Score down to 75%



Impact of Covid – absence of attendance measures





University of Central Lancashire
UCLan

Further Development

Further Development & What to Expect

University of Central Lancashire
UCLan

- Addition of more data sets – e.g. financial data
- This model running as a service, updating a student's scores nightly
- Automated alerts sent to an ‘engagement’ team
- The ability to monitor the effect of interventions

University of Central Lancashire
UCLan

Any Questions?

University of Central Lancashire
UCLan

References

References

- Engage2Serve, 2016, https://www.engage2serve.com/uk/blog/student-retention-strategies/
- Gascoigne & Cole, 2022, https://www.solutionpath.co.uk/case-study/aston-university-finds-early-measures-of-engagement-predictive-of-future-outcomes/
- Ometis, 2024, https://www.youtube.com/watch?v=z_8vzw4OYFI&t=912s
- Summers et al., 2021, https://doi.org/10.1080/02602938.2020.1822282