Emily (Mengqi) Huang

Nov. 2024

**Optimising Student Retention and Educational Outcomes: Early Identification and Tailored Support in Virtual Learning Using Data Science**

Capstone Project Document

Contents

[1. Problem statement 3](#_Toc181568414)

[2. Industry / domain 3](#_Toc181568415)

[3. Stakeholders 4](#_Toc181568416)

[4. Business question 5](#_Toc181568417)

[5. Data question 5](#_Toc181568418)

[6. Data 5](#_Toc181568419)

[7. Data science process 7](#_Toc181568420)

[7.1 Data analysis 7](#_Toc181568421)

[7.2 Modelling 12](#_Toc181568422)

[7.3 Outcomes 18](#_Toc181568423)

[7.4 Implementation 19](#_Toc181568424)

[8. Data answer 19](#_Toc181568425)

[9. Business answer 19](#_Toc181568426)

[10. Response to stakeholders 19](#_Toc181568427)

[11. End-to-end solution 20](#_Toc181568428)

[12. References 20](#_Toc181568429)

# Problem statement

* The problem that this project is investigating and its value
  + As online learning and digital education become more popular globally and in New Zealand, tertiary institutions and online learning providers are seeking innovative ways to improve student retention and success rates. The growing demand for flexible, remote learning—accelerated by the COVID-19 pandemic—has highlighted the need for early identification of at-risk students to provide timely and targeted interventions. This is particularly relevant as institutions face financial constraints, making retention a priority for both student outcomes and institutional viability.
* The current state
  + Many higher education institutions, including those in New Zealand, and especially online learning providers, face challenges in retaining students, which impacts academic success rates and financial stability. Traditional methods may not adequately identify at-risk students early enough, limiting institutions’ ability to provide timely support.
  + New Zealand universities such as Massey University have faced significant financial challenges post-pandemic, which has prompted restructuring and a focus on student retention and support as a strategy to enhance educational outcomes and institutional stability. Similarly, the University of Auckland’s Student Retention Policy emphasises providing personalised academic support to improve engagement and retention. These examples underscore the increasing importance of supporting student success and retention, aligning with the goals of predictive analytics tools like your project. Sources: [Stuff](https://www.stuff.co.nz), [University of Auckland](https://www.auckland.ac.nz).
* The desired state
  + The project aims to leverage predictive analytics to detect students at risk of underperforming early in their courses. By implementing a model that uses data from the initial weeks of a course, the desired outcome is to enable institutions to proactively support these students, improving retention rates and overall educational outcomes.
* Has this problem been addressed by other research projects? What were the outcomes?
  + Previous studies, especially those by the Open University (OU) through their OUAnalyse system, have successfully employed machine learning models for predicting students’ individual assessment performance. However, some models, particularly those on platforms like Kaggle, rely on term-long data, which delays interventions. This project, by focusing on early-stage predictions, aims to fill the gap in providing timely support based on limited initial data.

# Industry / domain

* What is the industry/ domain?
  + Tertiary education and online learning. This includes universities, distance learning institutions, and online training providers like MOOCs.
* Current state of this industry
  + *Global Trends in Online Learning***:** In 2024, the global online education market is projected to generate around USD $185.2 billion, with expected growth at an 8.56% annual rate through 2029. This trend is largely driven by an increased demand for flexible, remote learning solutions that allow individuals to upskill or reskill, which is crucial in today’s evolving job market. Platforms such as MOOCs and online universities cater to a wide audience, including professionals aiming for career advancements or transitions, which illustrates the universal appeal of digital learning for lifelong learning and skill development. Sources: [Purdue University](https://education.purdue.edu), [Statista](https://www.statista.com).
  + *Challenges and Opportunities for Higher Education Providers*: The tertiary education sector is experiencing shifts toward digital transformation, with a focus on improving student outcomes and retention amid financial pressures ([TechnologyOne](https://www.technologyonecorp.co.nz)). In New Zealand, qualification completions saw a minimal increase in 2023, largely from Level 1 courses, while other levels saw declines​ ([Education Counts](https://www.educationcounts.govt.nz/statistics/achievement-and-attainment)). The industry faces challenges in maintaining engagement and outcomes in a growing online learning landscape.
* Industry Value-Chain
  + The value chain includes curriculum development, digital learning platforms, student engagement and support, assessment, and outcome tracking.
* Key concepts
  + Key concepts include predictive analytics, student retention, engagement metrics, personalised learning, and digital transformation.
* Relevance to Other Industries
  + The project’s predictive analytics and data-driven support approach are relevant to industries focused on customer retention, such as online services, corporate training, and employee development programs.

# Stakeholders

* **Course Tutors and Academic Advisors**  
  Why they care: Tutors and advisors directly support students’ academic success, and effective, data-driven identification of at-risk students enables earlier interventions.  
  *Expectations*: Accurate, actionable predictions on at-risk students and an accessible dashboard that supports proactive student engagement.
* **University / Online Training Administration**  
  Why they care: Administrators focus on improving retention rates and student satisfaction, which impact university funding and reputation.  
  Expectations: A scalable tool that integrates with existing systems, offers data-driven insights, and supports institutional goals around retention and student success.
* **Students**  
  Why they care: Students benefit from personalised support to improve academic outcomes and avoid failure.  
  Expectations: Meaningful feedback, early alerts for struggling students, and guidance on academic improvement strategies.

# Business question

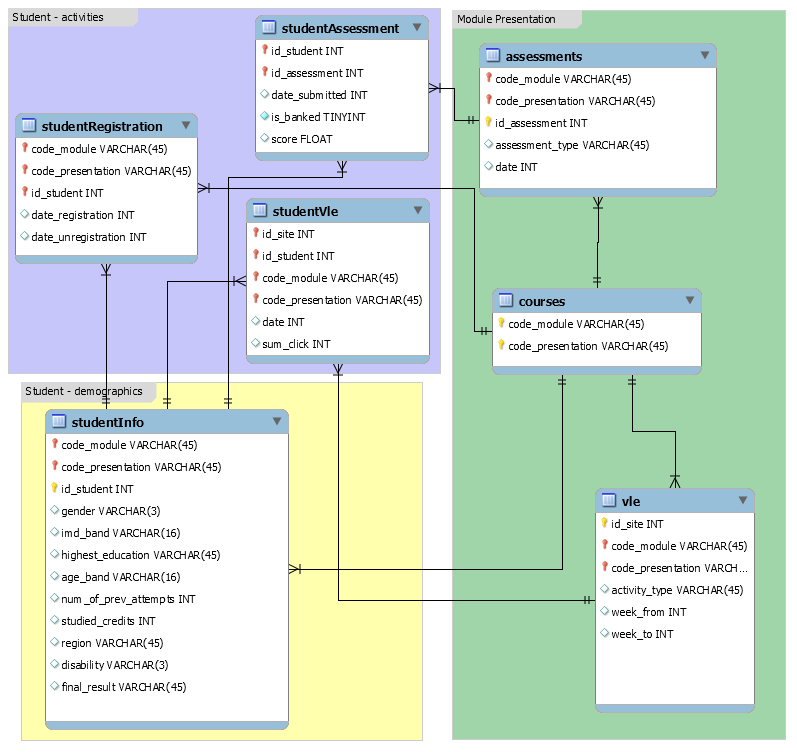
* The main business question that needs to be answered
  + How can we identify students at risk of failing or dropping out early in the term using demographic, engagement, and curriculum-related data, in order to provide timely and tailored support?
* Business value
  + Identifying at-risk students early allows for timely, targeted interventions, potentially reducing attrition and improving academic performance. Assuming a 5% improvement in retention, the university could save substantial costs related to lost revenue per dropout and enhance long-term institutional reputation.
* Required Accuracy and Implications
  + For benchmarking, the target F1 score should align with the OU Analyse project’s reported F1 measure (minimum 50%), aiming to enhance predictive accuracy. A high F1 score is crucial, as it helps balance the identification of true at-risk students (recall) while minimising unnecessary interventions (precision).

# Data question

* The data question that needs to be answered
  + What data is needed to predict which students are at risk of failing or dropping out early in the term, and how can this data be used to improve their learning outcomes?
* Required Data
  + To answer the above question, we need data on student demographics, engagement metrics, and curriculum-related attributes.
  + Specifically, key data points include demographic information (e.g., age, gender, previous academic background), engagement metrics from virtual learning environments within the first 4 weeks of course start date (e.g., frequency of logins, resource clicks), and curriculum characteristics (e.g., assessment weights, module details). This combination enables comprehensive insights for early intervention and personalised support strategies.

# Data

* Source
  + Seven csv files were sourced from the UK’s Open University (OU) Learning Analytics project, which uses historical data on student demographics, engagement behaviours, and curriculum-related features for predictive analysis.
* Volume and Attributes
  + The dataset contains records from a substantial student population across various courses, capturing attributes such as demographic details, Virtual Learning Environment (VLE) interaction metrics, curriculum characteristics, and assessment performance.
* Reliability and Quality
  + Given OU's structured collection process and extensive records, data reliability is relatively high. The data's continuous updates enable monitoring of student progress over time
  + The raw data is generally of high quality, yet may contain missing values or inconsistencies typical of large-scale educational datasets, especially in VLE engagement tracking.
* Data Generation
  + This data is collected systematically as part of OU’s analytics efforts, tracking both static demographic and academic attributes, and dynamic engagement metrics over the term.
* Ongoing Availability
  + VLE engagement data is continuously updated per login access , enabling ongoing risk analysis for students currently enrolled in courses.
* Database Schema showing the structure and relationship between the seven tables from the Open University Learning Analytics dataset (OULAD)



(Source: <https://analyse.kmi.open.ac.uk/open_dataset#description>)

# Data science process

## 7.1 Data analysis

* The data pipeline is divided into two key parts using Jupyter notebook:
  + **Data Preparation and Feature Engineering**: This part will focus on loading and processing the raw data, including cleaning, exploratory data analysis (EDA), joining multiple tables and feature engineering. After processing, only the selected predictors and target variable will be saved to a single CSV file.
  + **Model Training and Evaluation**: This part will load the preprocessed CSV file, then proceed with model training, hyperparameter tuning, and evaluation, ensuring a clear and modular workflow for future model applications and updates.
* Highlights of the Exploratory Data Analysis (EDA)
  + The stud\_info table includes 11 columns with demographic, educational, and prior attempt data for 32,593 students. Key columns include *gender*, *region*, *highest\_education, imd\_band* (socioeconomic indicator), *num\_of\_prev\_attempts, studied\_credits*, *disability*, and the (oginial) target variable *final\_result* (four categories).

A graph of a distribution of results

Description automatically generated

* + Due to class imbalance and for simplification, a binary target was created Binary Target (*at\_risk*): Categorised as 1 for at-risk (Fail and Withdrawn) and 0 for not at-risk (Pass and Distinction). Students who unregistered within the first 4 weeks were excluded from the dataset, so the adjusted counts are:
    - At-risk (1): 12,168
    - Not at-risk (0): 15,385.

A graph of distribution of status

Description automatically generated

* + Rationale for Binary Target Selection
    - Simplification: A binary target simplifies analysis, focusing on students at risk of failing or withdrawing, which aligns with the goal of early identification and intervention.
    - Project Alignment: Addressing student retention and dropout risk through a binary target makes predictive analytics more actionable for educational support strategies. This choice also supports model accuracy in imbalanced data settings by reducing noise from minor class distinctions.
  + Main new features:
    - Engineered Interaction Features (from the first 4 weeks of data):
      * *vle\_avg\_engagement\_f4w*: Average Virtual Learning Environment (VLE) engagement (active days averaged by the course length
      * *weekly\_avg\_click\_f4w*: Weekly average click counts within the first four weeks
      * *weekly\_avg\_access\_times\_f4w*: Weekly access frequency within the first four weeks
      * *avg\_biweekly\_click\_change*: Average biweekly change in clicks, providing insights into engagement trends
      * *frequent\_activity\_variety\_f4w*: Variety of activities accessed in the first four weeks, measuring diverse engagement
    - Proportion-Based Features for VLE Interactions:
      * Features like *oucontent\_prop, forumng\_prop, homepage\_prop*, and others that represent the proportion of specific VLE activity types out of the total activity. These capture the distribution of interactions with different VLE resources.
    - Course and Assessment Features:
      * *assessment1\_weight* and *assessment1\_weighted\_score*: Key early assessment metrics to capture performance signals
      * *cma\_tma\_weight\_ratio*: Ratio of Computer-Marked Assessments (CMA) to Tutor-Marked Assessments (TMA), relevant for understanding the course structure’s assessment emphasis
    - Demographic and Historical Features:
      * *highest\_education\_grouped* and *imd\_band\_group*: Grouped categories for education level and Index of Multiple Deprivation (IMD) bands (location to indicate the residents’ socio-economic background)
      * *num\_of\_prev\_attempts\_capped*: Limited to 3 attempts to reduce outlier effects, indicating past academic history
      * *cluster* (from K-Means clustering and One-Hot Encoded): Encoded values for student clusters based on demographic and engagement profiles
  + Distribution of key categorical features by At-Risk status

A graph of a number of bars

Description automatically generated with medium confidence

A chart of a number of bars

Description automatically generated with medium confidence

A chart with different colored bars

Description automatically generated with medium confidence

A graph of a number of red and blue bars

Description automatically generated

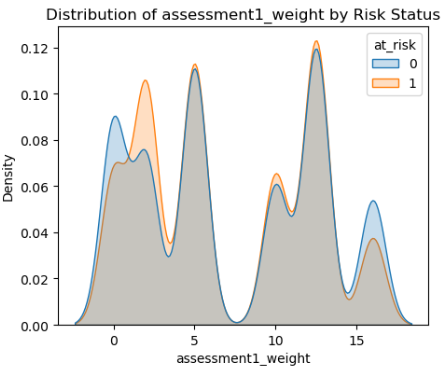
* + Distribution of key numerical features by At-Risk status

A graph of a graph

Description automatically generated A graph of a graph

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a number of people

Description automatically generatedA graph of a box plot

Description automatically generatedA comparison of a graph

Description automatically generated with medium confidence

## 7.2 Modelling

* Feature Selection Process and main Features Used
  + Features were selected initially based on business relevance and validated through trial-and-error to assess impact on model performance. Lasso regression was then applied to identify the top 25 predictive features.
  + The final model used a combination of student demographic information, VLE engagement metrics, and early assessment features.
    - **Course/Assessment Related Features**
      * *assessment1\_weight*
      * *assessment1\_weighted\_score*
      * *cma\_tma\_weight\_ratio*
    - **VLE Engagement Metrics (Online Activities)**
      * *forumng\_prop*
      * *frequent\_activity\_variety\_f4w*
      * *homepage\_prop*
      * *oucollaborate\_prop*
      * *oucontent\_prop*
      * *ouwiki\_prop*
      * *quiz\_prop*
      * *resource\_prop*
      * *subpage\_prop*
      * *url\_prop*
      * *vle\_avg\_engagement\_f4w*
      * *weekly\_avg\_click\_f4w*
    - **Student Demographic Information**
      * *cluster\_1*
      * *cluster\_2*
      * *disability*
      * *gender*
      * *highest\_education\_grouped\_HE Qualification*
      * *highest\_education\_grouped\_Lower Than A Level*
      * *imd\_band\_group\_over 50%*
      * *imd\_band\_group\_up to 50%*
      * *num\_of\_prev\_attempts\_capped*
      * *studied\_credits*

* + Purpose and Value of Clustering (before modelling)
    - K-Means clustering is used to segment students based on demographics, engagement, and academic performance data. The purpose of clustering is to identify patterns among groups of students with similar characteristics and risk profiles. This segmentation enables more tailored, proactive interventions by support teams, allowing them to address the specific needs of each student group more effectively. By understanding these clusters, institutions can offer more personalised support, enhancing student engagement, retention, and overall educational outcomes.
* Feature Interactions - Several interactions between features were insightful:
  + Assessment Scores: *assessment1\_weighted\_score* was calculated by taking the raw *score* from the first assessment, multiplying by its *weight*, and dividing by 100 to standardise its value.
  + *vle\_avg\_engagement\_f4w*: Calculated as the number of active days over the *module\_length*, providing a normalised measure of engagement relative to course duration.
  + Proportion of Specific VLE Activities: For each activity type (e.g., oucontent, forumng), the proportion was calculated by dividing the clicks on each activity by the total clicks within the first four weeks, capturing each activity’s relative interactions, providing predictive value for students' learning habits and early engagement patterns.
* Subset of Key Features for Performance
  + The top features identified by LightGBM’s feature importance ranking proved crucial in the model’s predictive accuracy. Notable features included *assessment1\_weighted\_score*, *vle\_avg\_engagement\_f4w*, various VLE click related features, and demographic related attributes such as studied\_credits and highest education level, which provided a robust predictive basis for identifying at-risk students early in the course.

A bar graph with text

Description automatically generated

* Feature engineering techniques
  + Joining table: Merging tables by primary and foreign keys while ensuring correct granularity is maintained at the level of module-presentation-student
  + Aggregation of VLE interactions over the first 4 weeks (*vle\_avg\_engagement\_f4w, weekly\_avg\_click\_f4w*)
  + Proportion-based features for activity types (e.g., *oucontent\_prop, forumng\_prop*)
  + Normalisation of continuous features and capping outliers (e.g., *num\_of\_prev\_attempts\_capped*)
  + One-hot encoding for categorical features such as highest education levels and imd band.
* Four models were selected for their unique advantages in binary classification and interpretability
  + **Logistic Regression**: Baseline interpretable model, providing a straightforward benchmark for predictive accuracy.
  + **Decision Tree**: Allows for intuitive interpretation of decision paths, with fast training and reasonable performance.
  + **LightGBM**: Offers high accuracy and handles imbalanced datasets well, making it suitable for nuanced binary classifications.
  + **Neural Network** (with 2 hidden layers): Capable of capturing complex patterns within data, with added flexibility in model architecture and tuning.
  + **Note:** Given the slight imbalance in the target classes, the "class\_weight" parameter is applied in all models to ensure that both classes are represented fairly in training, improving the model’s sensitivity to the minority (at-risk) class.
* Model training time
  + Training times varied across models, with Logistic Regression completing in about 0.1 seconds, Decision Tree in 0.2 seconds, LightGBM around 1 second, and Neural Network at approximately 3.5 seconds.
* What are the tools used?
  + The project was implemented using **Python** in Jupyter Notebook, with libraries including **pandas** and **numpy** for data handling, **matplotlib** and **seaborn** for visualisation, **scikit-learn** for modelling, and **TensorFlow** for building the neural network model.
* What are the model performance metrics?
  + Performance was evaluated using F1 Score, Recall, Precision, Accuracy, ROC AUC, Training time, together with the model’s interpretability and scalability. Given the importance of minimising both false negatives and false positives, F1 Score was prioritised.
  + A screenshot of a graph

    Description automatically generated
* Model selection
  + Based on our priority criteria—F1 score, training time, interpretability, model transparency, and scalability, **LightGBM** and **Logistic Regression** are recommended as primary models, with LightGBM being the top choice if performance is paramount and Logistic Regression as a strong, interpretable backup.
    - 1. LightGBM: This model consistently demonstrates strong performance in terms of F1 score and stability with increasing data, as seen in the learning and scalability curves. Although it shows some fluctuations in performance with fit times, LightGBM’s overall high F1 score and generalisation make it a strong choice for production. However, it is less interpretable compared to Logistic Regression but generally more efficient than Neural Networks.
    - 2. Logistic Regression: Logistic Regression is highly interpretable and offers fast training times, which can be crucial for quick deployments. It maintains a steady performance improvement, though it achieves a slightly lower F1 score compared to LightGBM. Logistic Regression can be a practical alternative for stakeholders who prioritise simplicity, transparency, and ease of interpretation.

A close-up of a test

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

A close-up of a test

Description automatically generated

A graph with a purple line

Description automatically generated with medium confidence

A close-up of a test

Description automatically generated

A graph with a purple line

Description automatically generated with medium confidence

A close-up of a test

Description automatically generated

A graph of a curve

Description automatically generated

A graph of different colored bars

Description automatically generated A graph with text overlay

Description automatically generated

A group of graphs showing different types of curves

Description automatically generated with medium confidence

## 7.3 Outcomes

* Main findings and conclusions of the data science process
  + The data science process achieved a satisfactory F1 score of 68%, favourably comparing to the original benchmark of 50%, using data limited to the first 4 weeks of the course. This balance between moderate predictive accuracy and early intervention highlights a trade-off: while higher accuracy might be possible with more data, timely predictions enable crucial, proactive support.
  + Key aspects that contributed to this success include effective feature engineering (e.g., calculating engagement proportions and early assessment scores), which enabled the models to leverage high-impact features.
  + Clustering further added value by revealing different student profiles, supporting personalised interventions.
  + LightGBM emerged as the optimal model, balancing predictive performance with scalability.

## 7.4 Implementation

* Considerations for implementing the model in production
  + **Data Pipeline**: Automate weekly data updates for VLE activity, demographics, and assessment scores, keeping predictions current.
  + **Dashboard Integration**:
    - Develop a dashboard to visualise at-risk predictions and highlight critical features.
    - Integrate clustering results into the dashboard, providing advisors with easy access to profiles and common characteristics of each cluster. This enables personalised support strategies.
  + **Model Monitoring**: Track model performance, retraining and validating as necessary to ensure accuracy over time.
  + **Data Security**: Implement secure handling for sensitive data to comply with regulations (e.g., GDPR).

# Data answer

* Was the data question answered satisfactorily?
  + Yes, the data question—to identify features that help predict at-risk students based on early-term demographic and engagement data—was answered effectively.
* What is the confidence level in the data answer?
  + The confidence level is high due to the model's consistent performance across multiple validation folds and the clear importance of certain features (e.g., first assessment scores, VLE engagement metrics) aligning with domain knowledge in education.

# Business answer

* Was the business question answered satisfactorily?
  + Yes, the project effectively answers the business question by identifying at-risk students early enough for timely intervention. The use of demographic, engagement, and curriculum-related data provides actionable insights for improving student retention.
* What is the confidence level in the business answer?
  + The confidence level is moderate to high, as the model has been validated to perform well, but real-world application will require testing in a production environment to confirm the expected impact on retention outcomes.

# Response to stakeholders

* Overall message:
  + Early Intervention: The model allows student advisors to intervene early, potentially reducing dropout rates and improving academic success.
  + Key Predictors: Attention should be given to student performance in the first assessment and engagement in the VLE during the first few weeks.
  + Tailored Support: Clustering reveals different student engagement patterns, suggesting that a personalised approach to student support would be beneficial.
* Recommendations:
  + Implement the model into existing dashboards for real-time access by educators and students.
  + Train student advisors on using the insights from clustering to personalise interventions.
  + Regularly retrain and validate the model to ensure ongoing accuracy and relevance.

# End-to-end solution

* **Ingestion and Preprocessing**: Collect data on student demographics, VLE activity, assessments, and courses, cleaning and transforming it as necessary.
* **Model Scoring and Updates**: Use the model to score students weekly, flagging those at risk and updating predictions as new data arrives.
* **Dashboard Deployment**: Feed at-risk predictions and cluster insights into a user-friendly dashboard, enabling student support teams to view risk status and personalised recommendations for each cluster.
* **Feedback**: Monitor model performance and accuracy, collecting feedback from stakeholders to improve model output quality.
* **Regular Model Maintenance**: Retrain and fine-tune the model periodically to adapt to changes in student engagement and demographics, maintaining a positive impact on retention goals.

# References

* Datasets:
  + Kuzilek J., Hlosta M., Zdrahal Z. [Open University Learning Analytics dataset](https://www.nature.com/articles/sdata2017171)Sci. Data 4:170171 doi: 10.1038/sdata.2017.171 (2017). (Accessed in October, 2024, URL: <https://analyse.kmi.open.ac.uk/open_dataset#data> )
* Python libraries:
  + **Data Processing**:
    - pandas, numpy: For data manipulation and handling.
    - scikit-learn.preprocessing: For feature scaling (StandardScaler) and encoding categorical variables.
    - scikit-learn.model\_selection: For cross-validation (StratifiedKFold) and train-test splits.
  + **Visualisation**: matplotlib, seaborn
  + **Machine Learning**:
    - scikit-learn (Logistic Regression, Decision Tree), lightgbm (LightGBM model)
    - TensorFlow and Keras for the neural network model
    - scikit-learn.metrics: For performance metrics such as accuracy\_score, precision\_score, recall\_score, f1\_score, and roc\_auc\_score.
* Other online resources / publications (Accessed in October, 2024):
  + <https://analyse.kmi.open.ac.uk/project_info#publications>
  + Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z. and Wolff, A. [**OU Analyse: Analysing At-Risk Students at The Open University.**](http://oro.open.ac.uk/42529/)Learning Analytics Review, no. LAK15-1, March 2015, ISSN: 2057-7494.
  + Wolff, Annika; Zdrahal, Zdenek; Nikolov, Andriy and Pantucek, Michal (2013). Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment. In: Third Conference on Learning Analytics and Knowledge (LAK 2013), 8-12 Apr 2013, Leuven, Belgium (forthcoming) (URL: <https://oro.open.ac.uk/36936/>)
  + <https://www.lexjansen.com/sesug/2016/EPO-271_Final_PDF.pdf>
  + Kaggle models: <https://www.kaggle.com/datasets/rocki37/open-university-learning-analytics-dataset/code?datasetId=70176&sortBy=voteCount>
  + <https://www.educationcounts.govt.nz/statistics/achievement-and-attainment>
  + <https://www.stuff.co.nz/nz-news/350347935/progress-massey-university-negotiates-through-financial-difficulties>
  + <https://www.auckland.ac.nz/en/about-us/about-the-university/policy-hub/education-student-experience/admissions-enrolment/student-retention-policy.html>
  + <https://education.purdue.edu/2024/01/global-trends-distance-learning/>
  + <https://www.technologyonecorp.co.nz/resources/articles/tertiaryeducation-institutions-are-feeling-the-pressure-to-accelerate-their-digital-transformation>