Using Early Student Engagement to Predict Academic Risk

Enrique Hirigoyen

2025-12-17

Contents

title: "Using Early Student Engagement to Predict Academic Risk" author: "Enrique Hirigoyen" date: "2025-12-17" output: html_document: toc: true toc_depth: 2 pdf_document: toc: true

1. Overview

Educational institutions often identify academic problems too late, when students have already failed or withdrawn. This project examines whether **early student engagement** can be used to detect academic risk at the beginning of a course, when intervention is still possible.

Two datasets are analyzed:

- **FNU Analytics**: an internal institutional dataset
- **OULAD**: an external benchmark dataset from the Open University

The project uses SQL Server for data preparation, Excel for data modeling, and R for analysis and visualization. Using multiple tools ensures accuracy, consistency, and reproducibility.

---

2. Problem Definition (Ask)

The main business question of this project is:

> *Can early student engagement during the first weeks of a course be used to predict academic risk and support timely intervention?*

Many students fail or withdraw because they disengage early. If this disengagement can be detected in the first weeks, institutions can act sooner and improve student outcomes.

The goal is not to replace academic judgment, but to **support instructors and advisors with data-driven insights**.

---

3. Data Description (Prepare)

This project uses two datasets: one internal dataset from FNU and one external benchmark dataset (OULAD). Both datasets capture student activity during the early weeks of a course.

3.1 FNU Dataset (Internal)

The FNU dataset represents institutional learning activity. It includes structured tables created in SQL Server and populated with simulated but realistic data.

Key early engagement metrics (Weeks 1–3):

- Number of logins (`logins_w1_3`)
- Time spent in the platform (`minutes_w1_3`)
- Assignments submitted (`assignments_w1_3`)

Academic outcomes:

- Pass indicator (`pass_flag`)
- Withdraw indicator (`withdraw_flag`)

3.2 OULAD Dataset (External Benchmark)

The Open University Learning Analytics Dataset (OULAD) is a public dataset widely used in learning analytics research.

Key early engagement metrics (Days 0–21):

- Total clicks in the virtual learning environment (`clicks_day0_21`)
- Number of active days (`active_days_0_21`)

Academic outcomes:

- Final result (`Pass`, `Distinction`, `Fail`, `Withdrawn`)

Although the metrics differ in scale and structure, both datasets measure the same underlying concept: **early student engagement**.

---

4. Data Engineering and Preparation (Process)

All data preparation was performed in **SQL Server**.

Main steps:

1. Creation of relational databases (`FNU_Analytics` and `OULAD_Analytics`)
2. Definition of primary keys and foreign key relationships
3. Loading of raw CSV files into staging tables
4. Data cleaning and aggregation

5. Creation of analytical views for Excel and R

For example, the OULAD clickstream data contained duplicate records. These were aggregated to ensure one record per student, activity, and day.

Views were created to protect raw data and ensure consistent analysis across tools.

---

5. Analysis in R (Analyze)

The analytical phase was conducted using **R** and **RStudio**, connecting directly to SQL Server views. This approach ensured that the same cleaned and validated data was used consistently across all tools.

The analysis followed these steps:

1. Load analytical views from SQL Server
2. Explore early engagement metrics
3. Visualize engagement vs academic outcome
4. Define an early-risk threshold
5. Build a predictive model

All analysis was performed using reproducible R scripts.

5.1 Loading Data from SQL Server

Data was loaded from the view `vw_OULAD_EarlyRisk`, which already contained aggregated and cleaned early engagement metrics.

```r
library(DBI)
library(odbc)

con <- dbConnect(
  odbc::odbc(),
  Driver = "SQL Server",
  Server = "LAPTOP-EH",
  Database = "OULAD_Analytics",
  Trusted_Connection = "Yes"
)

oulad_early <- dbReadTable(con, "vw_OULAD_EarlyRisk")
```

5.2 Descriptive Analysis

Basic descriptive statistics were calculated to understand engagement patterns during the first three weeks.

Students who passed the course showed substantially higher early activity than students who did not pass.

5.3 Visualization of Early Engagement

A boxplot was created to compare early engagement between students who passed and those who did not.

This visualization clearly showed a separation between the two groups, confirming that early engagement is strongly related to academic success.

5.4 Early-Risk Threshold Definition

An early-risk rule was defined using the **bottom 25%** of early engagement.

Students below this threshold were classified as early risk. This simple rule is easy to implement and interpret.

5.5 Predictive Modeling

A logistic regression model was built to estimate the probability of passing based on early engagement and basic demographic variables.

The model confirmed that early engagement metrics remained significant predictors even after controlling for age, gender, and education level.

---

6. Comparative Analysis: FNU vs OULAD (Share)

This section compares results from the internal **FNU** dataset with the external **OULAD** dataset. The purpose is not to compare platforms, but to evaluate whether early engagement behaves in a similar way across different learning contexts.

6.1 Alignment of Metrics

The two datasets measure engagement differently:

- **FNU** uses:
  - Logins during weeks 1–3
  - Minutes spent during weeks 1–3
  - Assignments submitted during weeks 1–3
- **OULAD** uses:
  - Total clicks during days 0–21
  - Number of active days during days 0–21

Although the scales are different, both datasets represent the same concept: **student activity during the early part of a course**.

For this reason, comparisons focus on **outcomes** (pass and withdrawal rates) rather than raw engagement counts.

6.2 Outcome Comparison

A summary table was created to compare academic outcomes between FNU and OULAD.

The results showed:

- Very similar pass rates across both datasets
- A higher withdrawal rate in the FNU dataset

This suggests that early engagement is consistently related to success, while institutional factors may influence withdrawal behavior.

6.3 Visualization of Outcomes

A bar chart was created to visualize pass and withdrawal rates side by side for both datasets.

The visualization highlights that, despite differences in learning environments, overall academic performance patterns are comparable.

6.4 Interpretation

The comparison indicates that early engagement is a **robust indicator** of academic success.

Even when engagement is measured differently, students who engage early are more likely to pass and less likely to withdraw. This supports the use of early engagement analytics as a general decision-support tool, not limited to a single institution.

---

7. Recommendations (Act)

This section translates the analysis into clear and practical actions that can be applied by instructors, advisors, and academic support teams.

7.1 Monitor Early Engagement

Institutions should monitor student activity during the first three weeks of each course. Simple indicators such as logins, time spent in the platform, clicks, or assignment submissions are sufficient.

Early monitoring allows institutions to detect disengagement before academic failure occurs.

7.2 Identify At-Risk Students Early

Students with very low engagement during the first weeks should be automatically flagged as early risk. A data-driven threshold, such as the bottom 25% of early activity or a low predicted probability of passing, can be used.

This approach ensures consistent and objective identification of at-risk students.

7.3 Apply Targeted Interventions

Once students are identified as early risk, institutions should apply targeted and timely interventions. These may include academic advising, reminders, tutoring support, or personalized communication.

Early interventions are more effective and less costly than late corrective actions.

7.4 Focus on Withdrawal Prevention

The analysis showed higher withdrawal rates in the FNU dataset. Institutions should prioritize interventions aimed at preventing early withdrawal, especially for students who disengage during the first weeks.

Reducing withdrawals improves student retention and overall academic outcomes.

7.5 Integrate Analytics into Regular Practice

Early-risk analytics should be integrated into regular academic workflows. This includes dashboards, advising reports, and periodic reviews during the term.

Embedding analytics into daily practice ensures sustainability and long-term impact.

---

8. Conclusion

This project demonstrates that early student engagement is a strong and reliable indicator of academic success.

By combining SQL Server for data preparation, Excel for data modeling, and R for analysis and visualization, the project shows a complete and reproducible analytics workflow.

The comparison between the internal FNU dataset and the external OULAD dataset confirms that early engagement patterns are consistent across different educational contexts.

Most importantly, the project highlights how data analytics can support timely intervention, improve retention, and enhance student success when applied early in the academic term.

---

9. Appendix

This appendix provides additional technical information to support transparency and reproducibility. Detailed scripts are summarized here, while full versions are provided as separate files.

Appendix A: Summary of Academic Outcomes

Dataset	Pass Rate	Withdraw Rate
FNU	48.0%	41.5%
OULAD	47.2%	31.2%

These values are derived from aggregated analytical views and were used to create the comparative visualization in Section 6.3.

Appendix B: R Scripts

The analytical work for this project was performed using R and RStudio. The primary script developed for the OULAD analysis is:

- `OULAD_EarlyRisk_Analysis.R`

This script includes the following steps:

- Connecting to SQL Server databases
- Loading analytical views
- Creating early engagement indicators
- Generating visualizations
- Building a logistic regression model to predict pass probability

The example below illustrates how early engagement was visualized using R.

```
library(ggplot2)

ggplot(oulad_early, aes(x = factor(pass_flag), y = clicks_day0_21)) +
  geom_boxplot(outlier.alpha = 0.05) +
  scale_x_discrete(labels = c("0" = "Did Not Pass", "1" = "Passed")) +
  labs(
    title = "Early Engagement vs Final Outcome (OULAD)",
    x = "Final Outcome",
    y = "Total Clicks (Days 0-21)"
  ) +
  theme_minimal()
```

Appendix C: SQL Scripts

SQL Server was used for all data engineering and preparation tasks. SQL scripts were developed to:

- Create relational databases and tables
- Load CSV files into staging tables
- Clean and aggregate raw activity data
- Create analytical views for Excel and R

Key analytical views created for this project include:

- `vw_FNU_Analytics`
- `vw_FNU_EarlyRisk`
- `vw_OULAD_Analytics`

- vw_OULAD_EarlyRisk

The simplified SQL example below illustrates how early engagement was aggregated in the OULAD dataset.

```sql
CREATE VIEW vw_OULAD_EarlyRisk AS
SELECT
    id_student,
    SUM(sum_click) AS clicks_day0_21
FROM StudentVLE
WHERE date BETWEEN 0 AND 21
GROUP BY id_student;
```
