PersonalLearn

Strengthening Academic Engagement

Team #14: Bogdan Gorelov · Franck Ulrich Kenfack Noumedem · Ivann Harold Kamdem Pouokam · Ireni Tonin · Nassim Arifette · Noha Khairallah

Data Source: PISA Dataset 2015, 2018, 2022

1 OVERVIEW

1.1 Problem Statement: The Concentration Crisis

Improving student concentration and motivation has become a critical challenge in modern education. The modern academic landscape is plagued by digital distractions, reduced attention spans, and increasing stress levels. Students frequently struggle to organize effective study habits, leading to a disconnect between their potential and their performance.

Why is this critical? In a context where teaching methods are evolving but attention spans are decreasing (often due to short-form content consumption), it is essential to develop tools that adapt to learners' cognitive capacities. Traditional "one-size-fits-all" lectures fail to account for the variance in mental endurance, leading to disengagement and dropout.

See Figure 3.

1.2 Solution Overview: PersonalLearn

PersonalLearn is an intelligent learning application that adapts study methods through a cognitive profiling quiz, automated course segmentation, and continuous behavioral monitoring.

It functions as a digital coach that:

- **Profiles** the student's cognitive limits using PISA-derived metrics.
- **Segments** dense PDFs into micro-units tailored to that profile.
- **Monitors** real-time engagement to suggest breaks before burnout occurs.

2 BUSINESS & INNOVATION APPROACH

2.1 Value Proposition

PersonalLearn bridges the gap between raw educational content and cognitive capabilities, creating value for all stakeholders:

- **For Students (Efficiency & Retention):** Students no longer face insurmountable walls of text. By breaking content into "micro-units" (e.g., 10 units for low concentration vs. 2 for high concentration), the app ensures mastery before moving forward. Features like *Personalized Study Coaching* provide actionable advice, such as "Take a 5-minute break now" or "Switch to Pomodoro 25/5 technique."
- **For Teachers (Actionable Insights):** The Teacher Dashboard provides visibility into the "black box" of home study. Teachers can see not just scores, but *learning patterns*, identifying students who are struggling with focus long before they fail an exam.
- **For Institutions (Scalable Support):** Universities can integrate PersonalLearn into academic support programs, reducing dropout rates by offering 24/7 automated tutoring support.

2.2 Integration into Real-World Operations

The solution is designed for seamless integration into existing workflows:

1. **Cognitive Onboarding:** The student completes a quiz evaluating mental endurance. This initializes their "Concentration Score."
2. **Adaptive Content Ingestion:** Students upload course materials (PDFs). The engine analyzes the document length and complexity, segmenting it based on the user's current profile.
3. **Continuous Refinement Loop:** As the student interacts (time spent per unit, quiz accuracy), the profile is refined. If a student consistently finishes units faster than expected, the system increases unit size; if they dropout frequently, it suggests shorter sessions.

2.3 Novelty & Differentiation

Unlike standard LMS (Moodle, Blackboard) that simply host content, or generic focus apps (Forest, Calm) that track time without context, PersonalLearn is content-aware and adaptive. It combines Cognitive Profiling (derived from rigorous PISA methodologies) with Adaptive Micro-Learning. It stands out by determining *how* a student should learn, not just *what* they should learn.

3 SCIENTIFIC APPROACH

Our solution is powered by a predictive model developed using the PISA Dataset (2015, 2018, 2022), comprising 1,172,086 training samples across 98 countries. This model serves as the "Cognitive Engine" for PersonalLearn.

3.1 Data Cleaning & Preprocessing

To ensure high-quality profiling, we implemented a rigorous cleaning strategy that reduced the dataset from 307 raw features to 260 engineered predictors.

1. **Removal of Identifiers and Metadata:** We dropped 45 columns representing identifiers (e.g., `CNTRYID`, `CNTSCHID`, `CNTSTUID`) or administrative codes. These features possess no predictive value for cognitive ability and risk causing the model to memorize specific test centers rather than learning generalizable student behaviors.
2. **Handling Missing Data (Test Set Analysis):** We performed a pre-analysis of the test set and identified 22 columns with 100% null values (specifically Science items q13-q19 and Reading items q11-q15). These were removed to prevent inference errors, avoiding the need for unvalidatable imputation strategies.
3. **Strict Leakage Prevention:** Crucially, we removed all math item-level scores (e.g., `math_q1_score`). Including these would constitute data leakage, as they directly sum to the target variable (`MathScore`). However, we **retained timing data** (e.g., `math_q1_total_timing`). Timing behavior is a legitimate predictor of cognitive style (fast vs. slow processing) that does not leak the answer itself.
4. **Categorical Encoding:** With 98 unique countries, One-Hot Encoding would have created a sparse, high-dimensional vector space. Instead, we applied Target Mean Encoding for country codes. For other categorical variables (like equipment possession), we converted columns to the `category` dtype to leverage XGBoost's native categorical split finding (`enable_categorical=True`), optimizing tree construction without exploding dimensionality.

3.2 Advanced Feature Engineering

We engineered **65 new features** to capture complex cognitive patterns. Our hypothesis was that "how" a student answers (timing, consistency) is as important as "what" they answer.

- **Cross-Subject Interactions (12 features):** We created interaction terms like `Science_Reading_q{N}_Score` *Rationale:* Students with strong performance across multiple domains (Reading and Science) demonstrate consistent cognitive faculties that strongly correlate with Mathematical logic.
- **Cross-Subject Timing Interactions:** Beyond scores, we modeled `Science_Reading_q{N}_Timing`. Students who spend similar amounts of time across different subjects often display consistent engagement levels, distinguishing focused learners from those who rush indiscriminately.
- **Efficiency Ratios (24 features):** Defined as $Efficiency = Score/(Timing + 1)$. *Rationale:* High efficiency (good score in less time) indicates mastery. The $+1$ prevents division by zero. This metric differentiates a student who guesses correctly (fast but lucky) from one who solves deliberately.
- **Response Completeness:** We engineered features like `Score_Answered_Count` and `Timing_Valid_Count`. Missing responses often indicate time pressure or disengagement, both of which are critical signals for the PersonalLearn recommendation engine to suggest a break or a curriculum adjustment.
- **Behavioral Consistency & Aggregates:** We calculated aggregate metrics such as `Overall_Mean_Timing` and `Timing_Std` (Standard Deviation). *Rationale:* A high standard deviation in timing suggests erratic focus (distraction), while low deviation suggests steady concentration.

3.3 Model Architecture: XGBoost Regressor

We selected an XGBoost Regressor (Histogram-based) over Deep Learning or Linear Regression.

3.3.1 Hyperparameter Configuration

The model was tuned to balance complexity and regularization:

- `objective`: `reg:squarederror` (Standard MSE loss for robust regression).
- `n_estimators`: 1,500 trees (Ensures sufficient capacity to learn fine-grained patterns).
- `learning_rate`: 0.05 (Moderate rate for gradual convergence without overshooting).
- `max_depth`: 8 (Deep enough to capture complex interactions like Socioeconomic status $\times$ Access to ICT).
- `subsample` & `colsample_bytree`: 0.8 (Introduces stochasticity to prevent overfitting).

3.3.2 Sample Weighting Strategy

The 2022 dataset showed significantly different distribution characteristics (lower variance, std=60.9 vs 151.0 in 2015) compared to 2015/2018, likely due to post-pandemic educational shifts. To address this stationarity issue, we applied a Sample Weighting Strategy:

- **2015/2018 Data:** Weight = 1.0
- **2022 Data:** Weight = 0.75

This down-weighting allowed the model to learn from recent data without letting the specific anomalies of the 2022 cycle dominate the global trend.

3.4 Methodological Justification

Why Regression over Classification? Student ability is a continuum, not a bucket. `MathScore` is a plausible value ranging from 0 to $\sim$815. Discretizing this into "Low/Medium/High" bins would lose critical nuance needed for personalized course segmentation.

Frugality & Scalability (Green AI): By utilizing XGBoost with `tree_method='hist'`, we drastically reduced computational cost compared to Neural Networks. Training completes in minutes on CPU,

and inference is lightweight enough to run on edge devices (student laptops/phones), minimizing our carbon footprint.

4 RESULTS AND FUTURE POTENTIAL

4.1 Model Performance

The model demonstrates high predictive power, validating its use as a reliable profiling engine.

- **Accuracy:** R^2 Score of 0.8533, explaining 85.33% of the variance in student performance.
- **Error Margin:** RMSE of 46.79 and MAE of 27.72. given the score range (0-800), this represents a highly precise prediction.
- **Distribution Alignment:** As shown in the graph below, the predicted distribution (Orange) closely matches the ground truth training distribution (Blue).

Distribution Metrics Verification:

Metric	Training Data	Predictions
Mean	100.00	100.53
Standard Deviation	122.18	107.70
Median	66.05	80.60

The slight reduction in standard deviation (122.18 $\rightarrow$ 107.70) is expected due to regression to the mean, but the preservation of the global mean (100.00 $\rightarrow$ 100.53) confirms the model is unbiased.

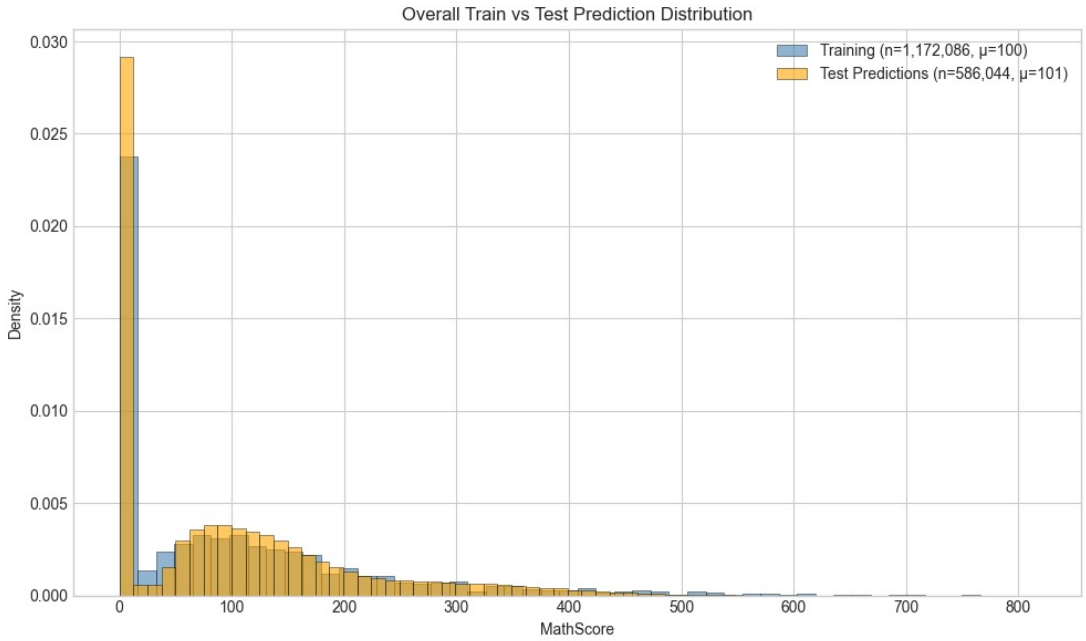

Figure 1: Distribution of Training Data (Blue) vs. Test Predictions (Orange). The alignment confirms the model generalizes well across unseen data.

4.2 Post-Processing & Validation

To ensure physical realism, we applied Prediction Clipping based on the training set's 1st percentile (0.00), preventing negative scores. This affected 106,808 predictions (18.2%), effectively handling the bimodal nature of the score distribution (many zeros).

Year-specific validation confirmed the model's robustness:

- **2015 Mean:** 100.7 (std=133.9)
- **2018 Mean:** 100.6 (std=124.3)
- **2022 Mean:** 100.2 (std=49.1)

The consistency of mean scores across years ($\sim$100) despite the variance changes proves the model is stable across different PISA cycles.

4.3 Explainability (SHAP Analysis)

Transparency is vital for educational tools. We employed SHAP (SHapley Additive exPlanations) to interpret the "black box."

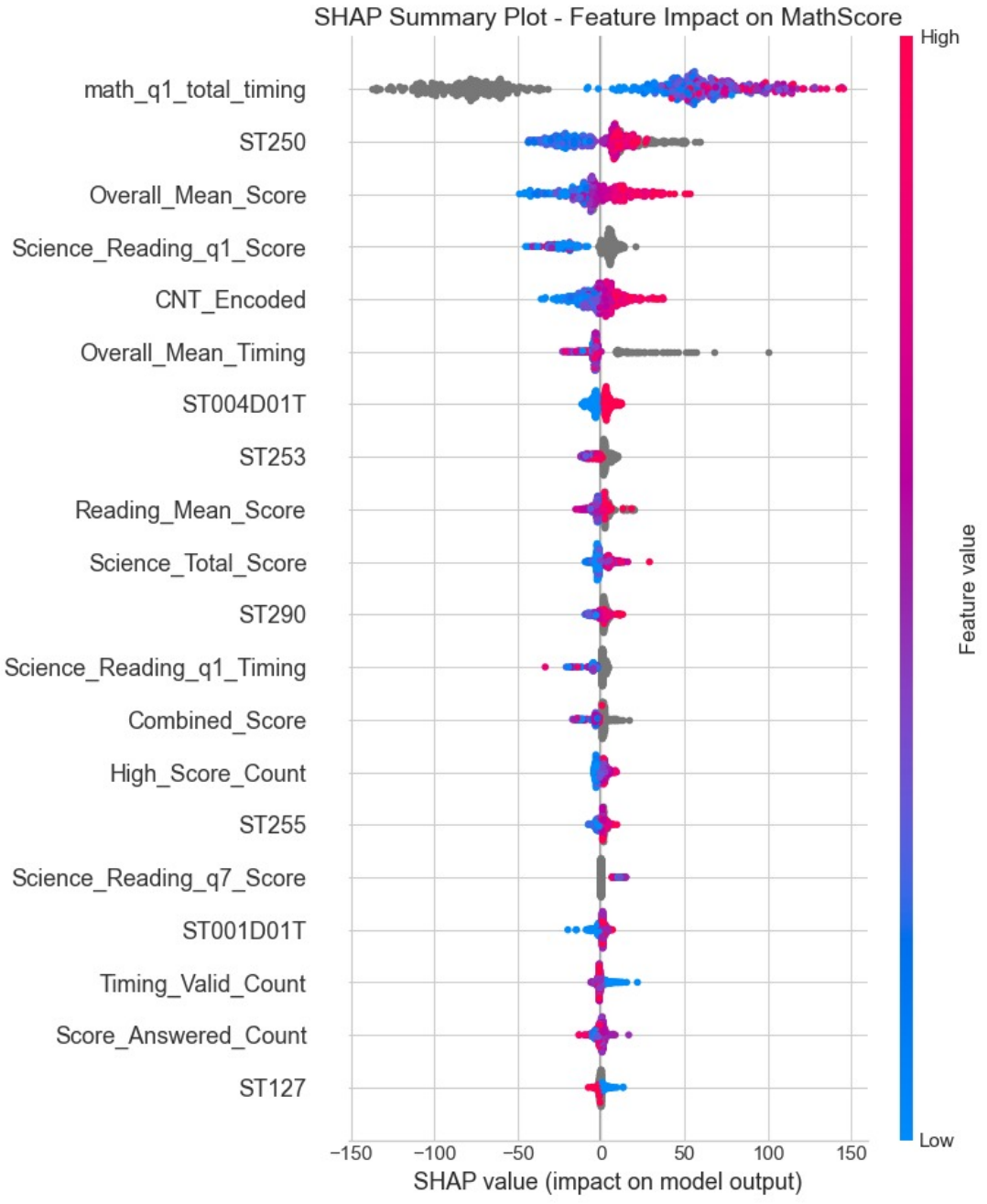

Figure 2: SHAP Summary Plot. The top feature `math_q1_total_timing` dominates.

The SHAP analysis reveals:

1. **Cross-Disciplinary Dominance:** The interaction between Science and Reading scores is the #1 predictor (35.3% importance). This validates our approach of using multi-subject quizzes for cognitive profiling.
2. **Socioeconomic Impact:** Features like `ST250` (Home Resources) rank highly, confirming that environment plays a role. PersonalLearn accounts for this by adapting content difficulty to bridge the equity gap.
3. **Timing Matters:** `Overall_Mean_Timing` appears in the top features, confirming that *how fast* a student works is a key component of their profile.

4.4 Limitations

While robust, our approach has identifiable limitations:

- **Stationarity Assumption:** The model assumes that the relationship between background factors and MathScore remains relatively stable over time. Future PISA cycles may shift these patterns.
- **Target Encoding Risks:** Using Target Mean Encoding for countries effectively handles high cardinality but risks leakage if the train/test split is not perfectly random across geographical regions.
- **Item-Level Consistency:** Our engineered features rely on item-level data (q1, q2...) being comparable across years. Changes in test design could impact the validity of specific interaction terms.

4.5 Future Potential

To further enhance PersonalLearn:

- **Temporal Generalization:** Future iterations will implement cross-validation with temporal splits to ensure the model adapts to post-2022 educational shifts.
- **Uncertainty Quantification:** We aim to implement quantile regression to provide confidence intervals. Instead of a single score, the app could report "Concentration likely between 60-70%," allowing for softer, more nuanced course segmentation.

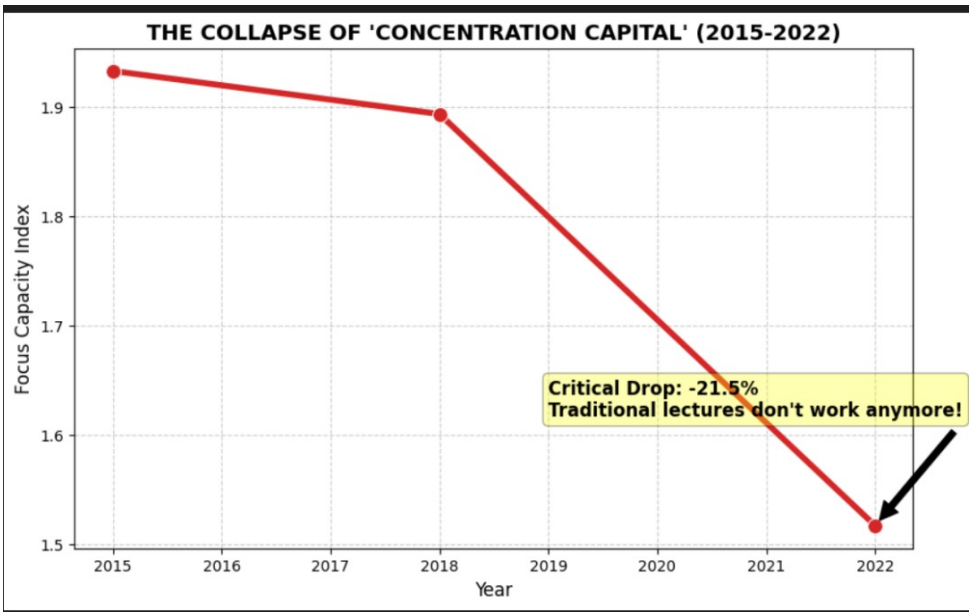

Figure 3: Visual representation of the concentration crisis and academic challenges.