Technical Report — Modeling Early Indicators of Student Performance

1. Problem Statement

The objective is to build a regression model that predicts final course grades in online middle-school science courses using behavioral, linguistic, motivational, and contextual features collected during the learning process. The intent is to evaluate the predictive ceiling of such multimodal educational data and identify structural sources of noise and variance.

2. Data Description

N ≈ 500 students across five online science subjects.

Features domains:

- **Behavioral (LMS traces):** time_spent, engagement_ratio, number_of_posts
- **Motivational:** interest, utility, perceived competence
- **Linguistic:** sentiment, cognitive indicators, social presence tokens
- **Context:** course subject, semester, enrollment_reason (one-hot encoded)

Target: Final numeric course grade (0–100)

3. Preprocessing & Feature Engineering

All feature engineering steps occur on **training folds only** to prevent leakage.

Created higher-level features:

- Composite **motivation_score** from latent motivation dimensions
- Derived **engagement_ratio** (modules_accessed / total_modules)
- Derived **motivation_time = motivation_score × time_spent**
- Created **emotion_balance** = positive_emotion – negative_emotion
- One-hot encoding for subject, semester, enrollment_reason
- Capped extreme engagement ratios at 1st/99th percentiles

Dataset split: **80/20 stratified by final grade distribution**

4. Modeling Framework

Models evaluated:

- **Random Forest** (ranger)
- **Gradient Boosting Machine** (gbm)
- **XGBoost** (standard + tuned)

Cross-validation: 5-fold repeated CV with grid search.

Evaluation metrics: **RMSE / MAE / R^2**

5. Performance Summary

5.1 Predicted vs Actual Distribution

To complement the point-metric evaluation (RMSE/MAE), we visualize the relationship between predicted and actual final grades. Most predictions cluster around the upper–middle grade region (70–90), consistent with the underlying grade distribution. Systematic underprediction appears for a subset of lower-performing students, reflecting the model's difficulty capturing conceptual or non-behavioral learning processes.

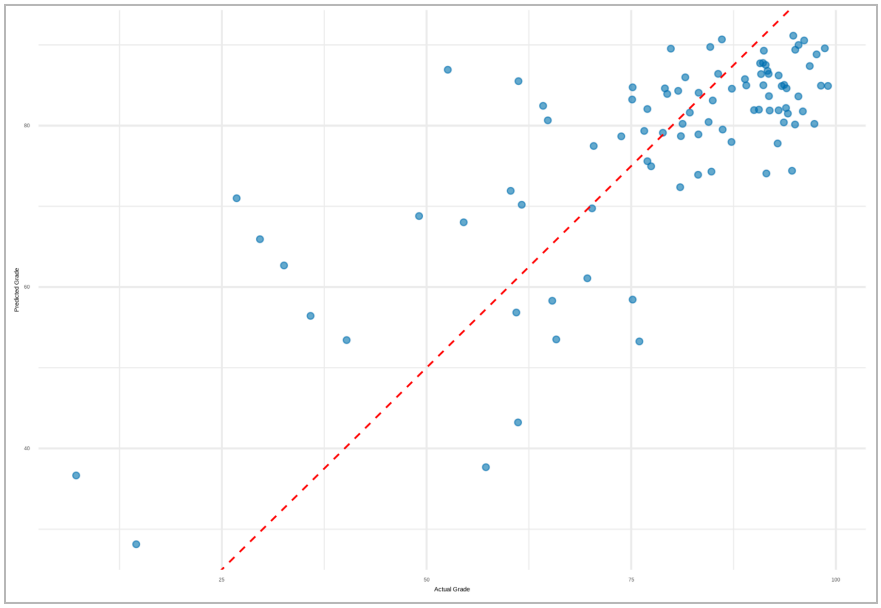

Figure 1. Predicted vs Actual Final Grades
Points close to the dashed line indicate accurate predictions. Deviations highlight systematic challenges with low-performing students and subjects involving high conceptual reasoning (e.g., Physics).

5.2 Model Comparison

Model	RMSE	R^2	MAE
Random Forest	**13.01**	**0.55**	10.08
Gradient Boosting	13.32	0.53	**9.91**
XGBoost (standard)	13.47	0.52	10.15
XGBoost (tuned)	13.84	0.51	10.32

Interpretation:

- Prediction ceiling with this dataset is ~10 points MAE due to noise in linguistic/motivation data.
- Random Forest yields best overall performance → stable under small N and high-variance features.
- GBM has better rank-ordering (lowest MAE).
- XGBoost underperforms, likely due to low sample size + noisy interactions.

6. Key Predictive Features

Permutation-based feature importance (RF) highlights strong nonlinear behavioral signals.

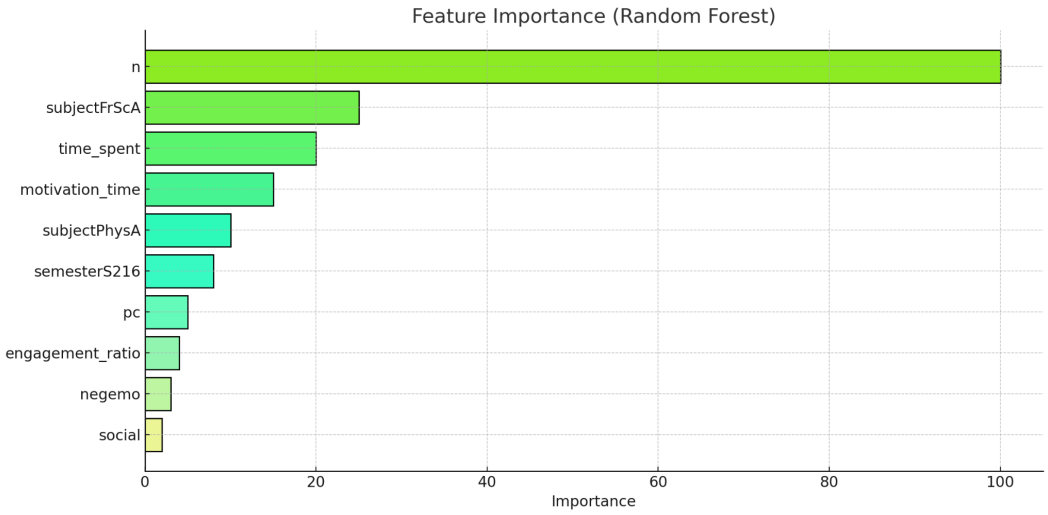

Figure 2. Random Forest Feature Importance
Shows that discussion activity (n) dominates predictive power, followed by subject and time-based features.

Top signals:

1. Discussion activity (n)
2. Course subject
3. Time_spent
4. motivation_time
5. Engagement_ratio
6. Perceived_competence

Note: Linguistic features showed low individual importance — consistent with short forum posts and sparse linguistic signals.

7. Error Analysis

Model errors were examined to identify structural weaknesses:

7.1 High-Error Subjects

- Physics (RMSE ~18.6)
    - Behavioral features insufficient to explain conceptual performance.
    - Indicates subject-level variance not captured by generic LMS signals.

7.2 Enrollment Effects

- Scheduling Conflict (RMSE ~17.6): Behavioral irregularity likely externally driven → high predictive noise.

7.3 Mid-Motivation × Mid-Engagement

- RMSE ~18.9: Stable superficial engagement but weak actual performance → ambiguous signal for ML models.

These high-error zones suggest heterogeneity, measurement noise, or unobserved confounders.

Figure 3. Time Spent vs Final Grade, Colored by Prediction Error
Illustrates struggling high-effort students — high time_spent, low grade, large residuals.

8. Limitations

- Behavioral traces lack sequence/time-order information.
- Motivation self-report introduces bias and ceiling effects.
- Linguistic features low-signal due to brief posts.
- Limited N restricts depth of boosting trees and hyperparameter tuning.
- No explicit modeling of teacher/section effects.

9. Future Directions

Several improvements could meaningfully raise predictive accuracy:

- Incorporate weekly time series instead of aggregated features
- Use mixed-effects models to separate subject/teacher variance
- Add clickstream features (per-module attempts, dwell time)
- Evaluate group-level error fairness (subject, demographic factors)
- Deploy model as lightweight API + dashboard for real-time monitoring

10. Reproducibility

A fully reproducible pipeline is implemented via scripts 00–10:

- 00 load_raw
- 01 clean
- 02 feature_engineer
- 03 split
- 04–07 model training & CV
- 08 evaluation
- 09 error analysis
- 10 visualization

All outputs are version-controlled and seed-locked for reproducibility.

References

- Estrellado, R., Freer, E., Velásquez, I. C., Rosenberg, J. M., & Mostipak, J. (2020). *Data Science in Education Using R*. Dataset: `dataedu::sci_mo_with_text`.