Student engagement model’s implications

1. Project context

This analysis focuses on predicting student engagement. The goal is to support early intervention by identifying students who are at risk of low engagement based on their platform activity.

We used behavioral features such as:

- Number of times logged in
- Lessons accessed
- Assignments submitted
- Forum posts made

These were combined into a single engagement score, and then translated into a binary target:

High engagement (1) vs Low engagement (0) using the median as the threshold.

2. Selected models & evaluation

We trained and compared two supervised classification models:

Model type		**Why used?**
Logistic Regression	Linear, interpretable	Simple and transparent
Random Forest	Non-linear, ensemble	Captures complex interactions

We split the dataset (80/20) and used 5-fold cross-validation for robust evaluation.

Key metrics:

- AUC Logistic Regression: 0.98
- AUC Random Forest: 1.00

Test accuracy: > 95% for both models

Confusion matrices showed:

Very few misclassifications and mostly correct predictions across both classes

3. Visual communication

Visuals used to communicate results:

- ROC Curve Comparison: clear difference in predictive confidence
- Confusion Matrix: shows classification performance visually per model
- Boxplots (earlier step): show how engagement varies by feature or category

4. Interpretation of results

Both models work very well, but Random Forest may be overfitting slightly due to its near-perfect results.

Logistic Regression performs slightly less well but is easier to explain and implement.

Behavioral signals (logins, posts, etc.) are strong predictors of engagement.

5. Limitations

The dataset is balanced and relatively simple. Results may be optimistic.

- No external validation yet (e.g., on a different student cohort).
- Only behavioral data was used, demographic or motivational variables might improve predictions.
- No hyperparameter tuning done yet, future models could be optimized further.

6. Implications & Next Steps

The model could be integrated into an early warning system for student support teams.

Real-time dashboards could use this logic to flag students needing attention.

Future improvements:

- Add more features (e.g., learning pace, survey results)

- Test generalization on other data