Early Gameplay Risk Scoring Engine Using Behavioral Data

Malik Ibrahim Ali Khan

March 2, 2026

Executive Summary

This project focuses on predicting whether a gameplay session will result in an incorrect outcome using early behavioral signals. The main idea is to determine whether patterns within the first 20 actions of a session can indicate struggle before the final result occurs.

Using gameplay telemetry data from a Kaggle dataset, I engineered session-level and early-session features that capture timing patterns, pauses, and engagement behavior. Two models were evaluated: Logistic Regression and Random Forest. To prevent inflated performance, a group-based train-test split was applied at the session level to avoid data leakage.

The Random Forest model achieved a ROC AUC of approximately 0.64. While this performance is moderate, the model was effective for ranking risk. By flagging the top 20% highest-risk sessions, the system captured approximately 44% of incorrect outcomes. This shows that early behavioral signals can support targeted and efficient intervention strategies.

Problem Statement and Motivation

The central question of this project is:

Can early gameplay behavior (first 20 actions) predict whether a session will end in an incorrect outcome?

In many learning systems, support is provided only after failure occurs. However, if early signs of struggle can be detected, intervention can happen sooner. The goal is not perfect classification, but rather risk scoring identifying sessions that are more likely to struggle so support can be prioritized.

Data Collection and Preparation Methodology

The dataset comes from the Kaggle competition "Predict Student Performance from Game Play." It includes question-level labels (correct = 1, incorrect = 0), session IDs, level groups (0–4, 5–12, 13–22), timestamped gameplay events, and player interaction logs.

Because the event log file is large, it was processed in chunks, and a 20% sample was used due to computational constraints.

Two sets of features were engineered:

Overall Session Features

These summarize the entire session and include:

- Total number of events
- Session time
- Mean and median time between actions
- Number of long pauses
- Unique event count
- Pause ratio

Early Session Features (First 20 Actions)

These focus only on early behavior and include:

- Early mean timing
- Early timing variability (standard deviation)
- Early long pauses
- Early unique events
- Early pause ratio

Labels were merged with session features using session ID and level group. Because multiple questions belong to the same session, a group-based split was used to prevent data leakage across training and testing sets.

Analysis Methods and Implementation

Two models were trained:

- Logistic Regression (baseline)
- Random Forest (nonlinear ensemble model)

The dataset is imbalanced (approximately 14% incorrect sessions), so class weighting was applied to reduce bias toward predicting correct outcomes.

Model performance was evaluated using ROC curves and Area Under the Curve (AUC). Random Forest achieved a ROC AUC of approximately 0.64, while Logistic Regression achieved approximately 0.61.

Feature importance analysis showed that total events, session time, long pauses, and early timing variability were among the strongest predictors.

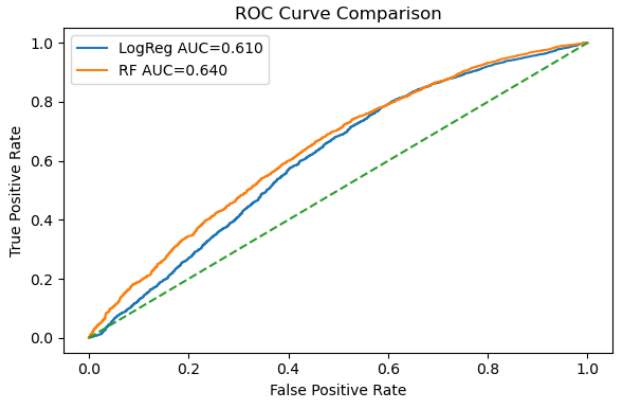

Figure 1: Enter Caption

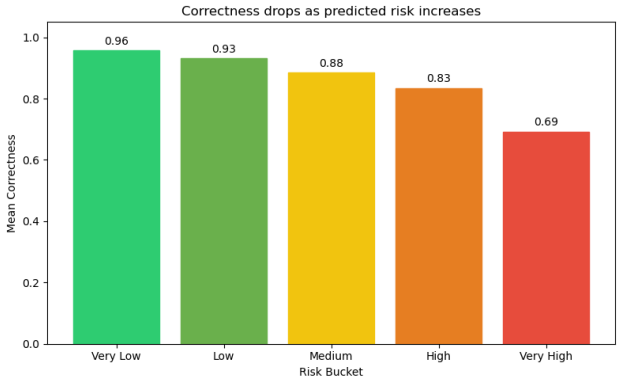

Figure 2: Enter Caption

Results and Interpretation

The ROC curve (fig 1) shows that both models perform better than random guessing, with Random Forest outperforming Logistic Regression.

Risk bucket analysis (fig 2)divided sessions into five groups from Very Low to Very High risk. As predicted risk increased, correctness steadily decreased. The Very High risk group had a correctness rate of about 69%, compared to about 96% in the Very Low risk group. This confirms that the model successfully ranks sessions by difficulty level.

In the intervention simulation:

- Overall correctness rate: approximately 86%
- Top 20% highest-risk sessions were flagged

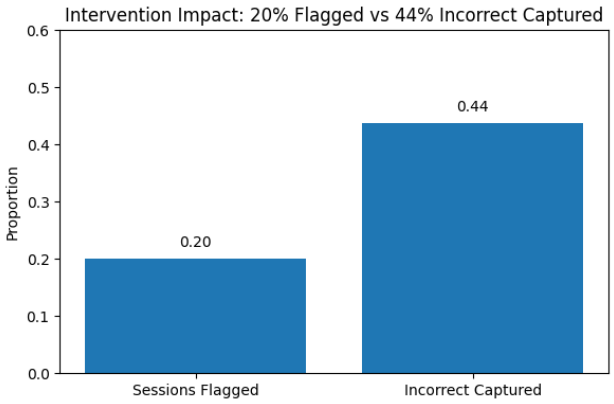

Figure 3: Enter Caption

- Approximately 44% of incorrect sessions were captured

This demonstrates that incorrect outcomes are concentrated in higher-risk groups. Instead of helping every session, intervention can focus on a smaller subset while still covering a large portion of struggling cases.

Key Findings

- Early gameplay behavior contains meaningful predictive signals.
- Timing patterns and pauses are strong indicators of struggle.
- Preventing data leakage reduced inflated model performance and produced more realistic evaluation.
- Even with moderate AUC (0.64), the model is effective for risk ranking.

Conclusion and Future Work

This project demonstrates that early-session behavioral data can be used to estimate risk before the final outcome occurs. Although model performance is moderate, the system effectively prioritizes high-risk sessions for targeted support.

Future improvements may include:

- Using the full dataset instead of sampling
- Adding richer temporal features
- Testing gradient boosting models
- Performing cross-validation across sessions
- Evaluating real-world intervention impact

References

- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1), 3–17.
- Bowers, A. J., Sprott, R., & Taff, S. A. (2013). Do we know who will drop out? A review of predictors of dropping out of high school. *High School Journal, 96*(2), 77–100.
- Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

- Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C, 40*(6), 601–618.

- Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432.

- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics, 7*, 91.