Early Gameplay Risk Scoring Engine

Using Early Behavioral Signals to Predict Incorrect Outcomes.

Malik Ibrahim Ali Khan

MSDS 692 Practicum

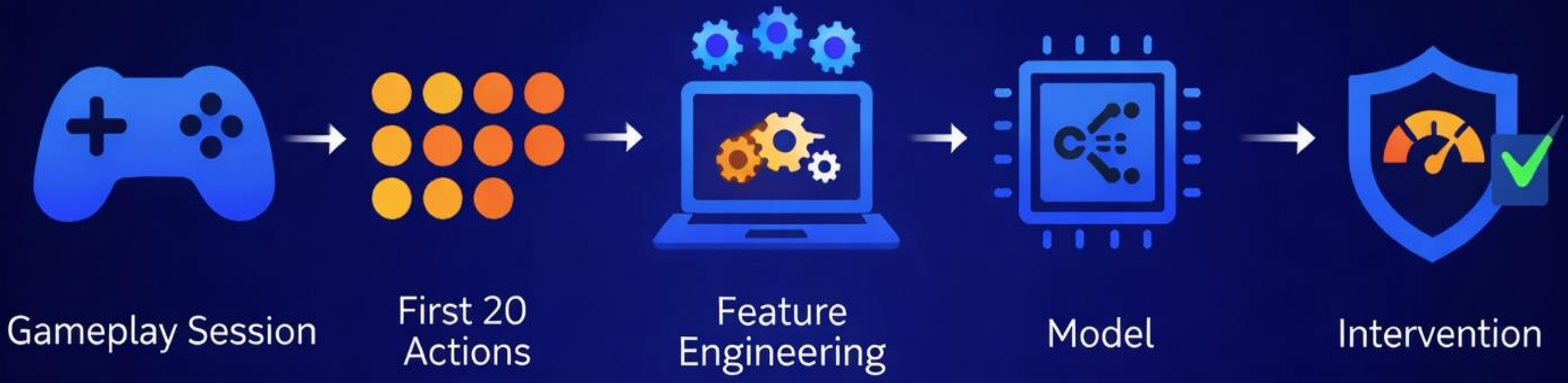
What Does This Project Mean & Why Does It Matter

What it means:

- Analyze early gameplay behavior (first 20 actions)
- Detect signs of hesitation or struggle
- Estimate probability of incorrect outcome
- Generate a “risk score” for each session

Why it matters:

- Struggle is often detected too late
- Early intervention can improve learning outcomes
- Enables targeted support instead of random help
- Scalable solution for large educational systems



What My Project Does

- Extract behavioral features
- Train predictive models
- Generate risk score
- Simulate targeted intervention

Problem Statement & Dataset

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

Source:
Kaggle – *Predict Student Performance from Game Play*

Dataset Overview

- Question-level labels (correct = 0 or 1)
- Session IDs
- Level groups (0–4, 5–12, 13–22)
- Timestamped gameplay events
- Player interaction logs

Methodology

Feature Engineering	Created session-level features from gameplay logs	Built early-session features (first 20 actions only)	Captured timing patterns, pauses, and interaction behavior
Preventing Data Leakage	Used group-based train/test split by session_id	Ensured no session appears in both train and test	Only used early data to simulate real-world prediction

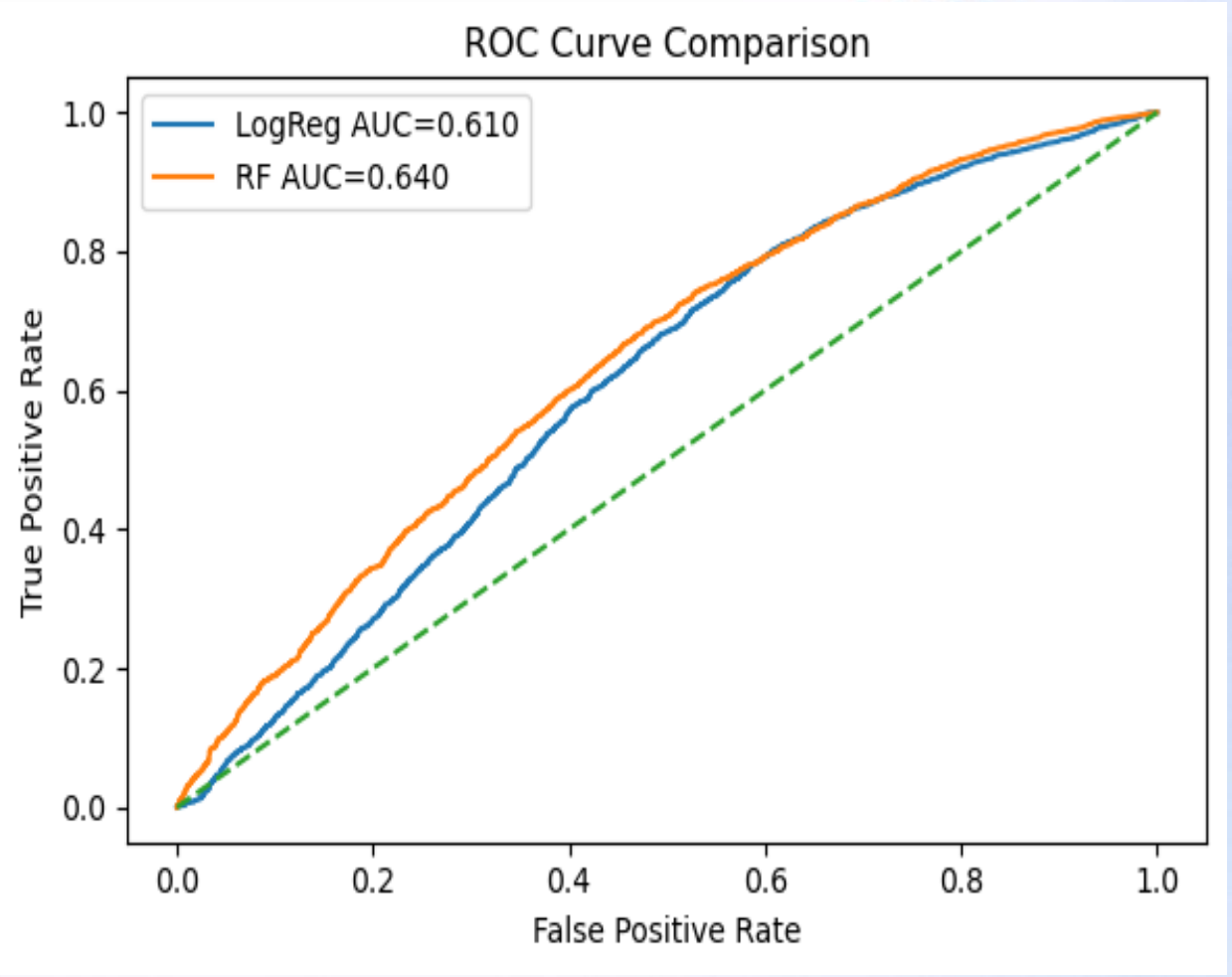
Models Used & ROC Performance

Models

- Logistic Regression
- Random Forest

Performance (ROC AUC)

- Logistic: 0.61
- Random Forest: 0.64
- Random Forest performed better overall.



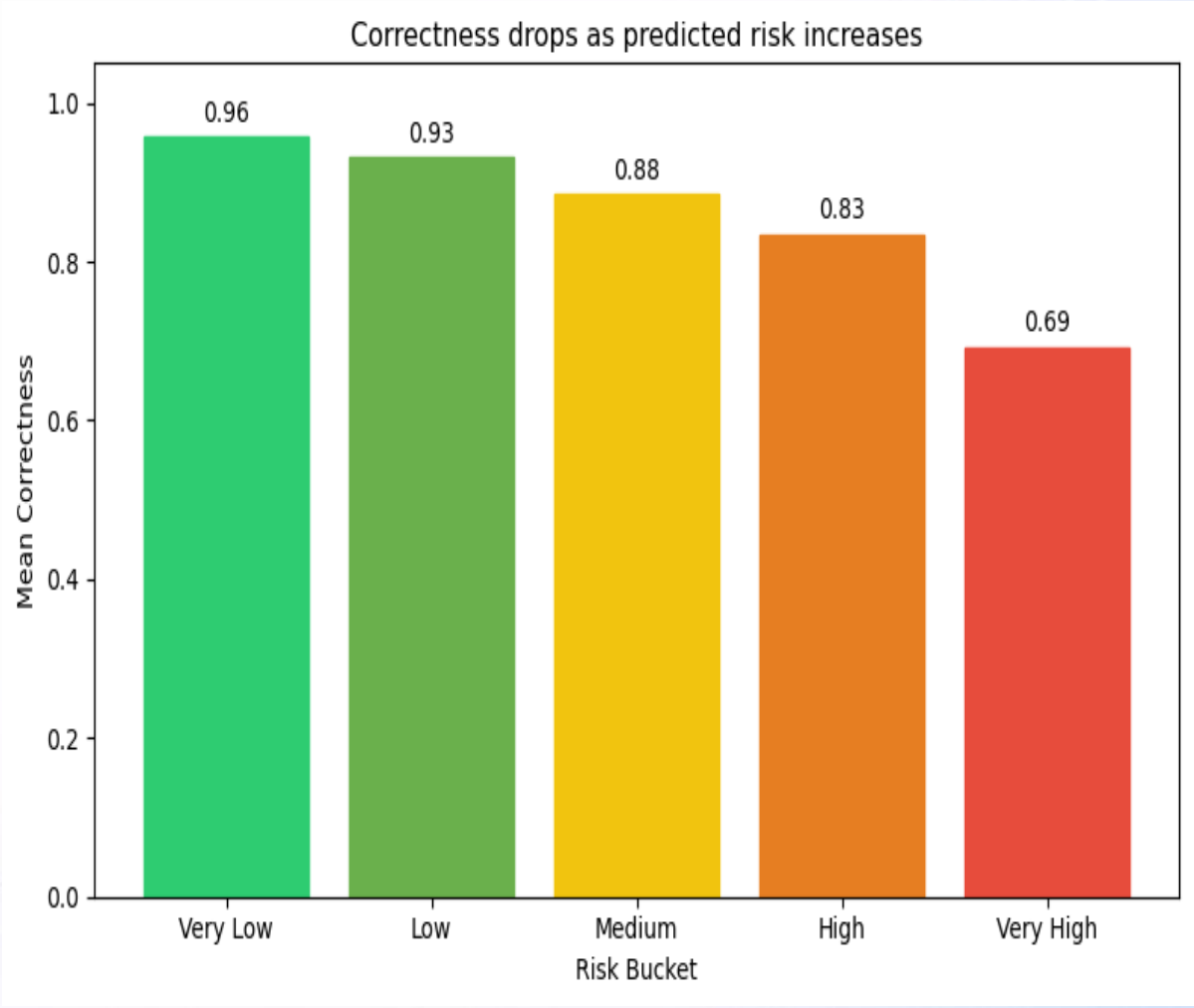
Risk Bucket Analysis

Sessions were grouped into risk levels based on predicted probability:

- Very Low Risk
- Low Risk
- Medium Risk
- High Risk
- Very High Risk

Key Insight

- Accuracy decreases as predicted risk increases.
- The **Very High Risk** group had the lowest correctness rate (~69%), showing the model successfully concentrated incorrect sessions into higher-risk buckets.



Intervention Simulation

Goal:
Simulating what happens if we intervene on the highest-risk sessions.

Strategy

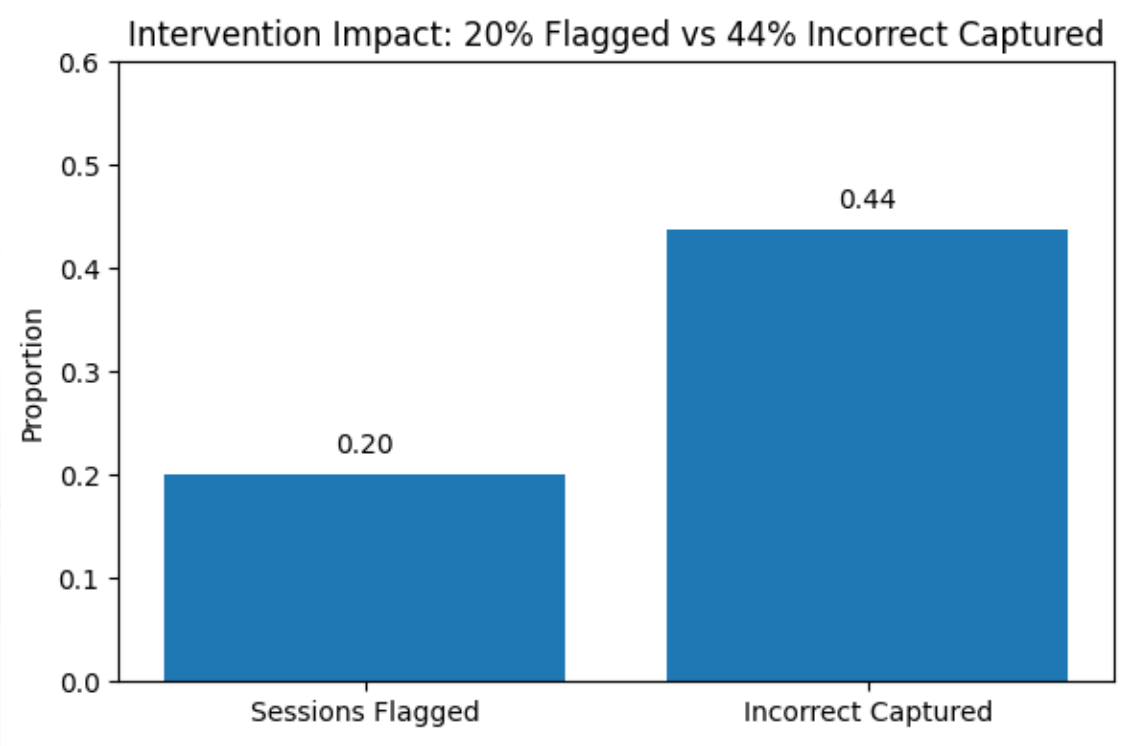
- Flag top 20% highest-risk sessions
- Provide early support or assistance

Results

- Captured ~44% of incorrect sessions
- High-risk group size: ~15,878 sessions

Key Insight

- By targeting only 20% of sessions,
  we can potentially address nearly half of incorrect outcomes.



Risk-Based Intervention Strategy

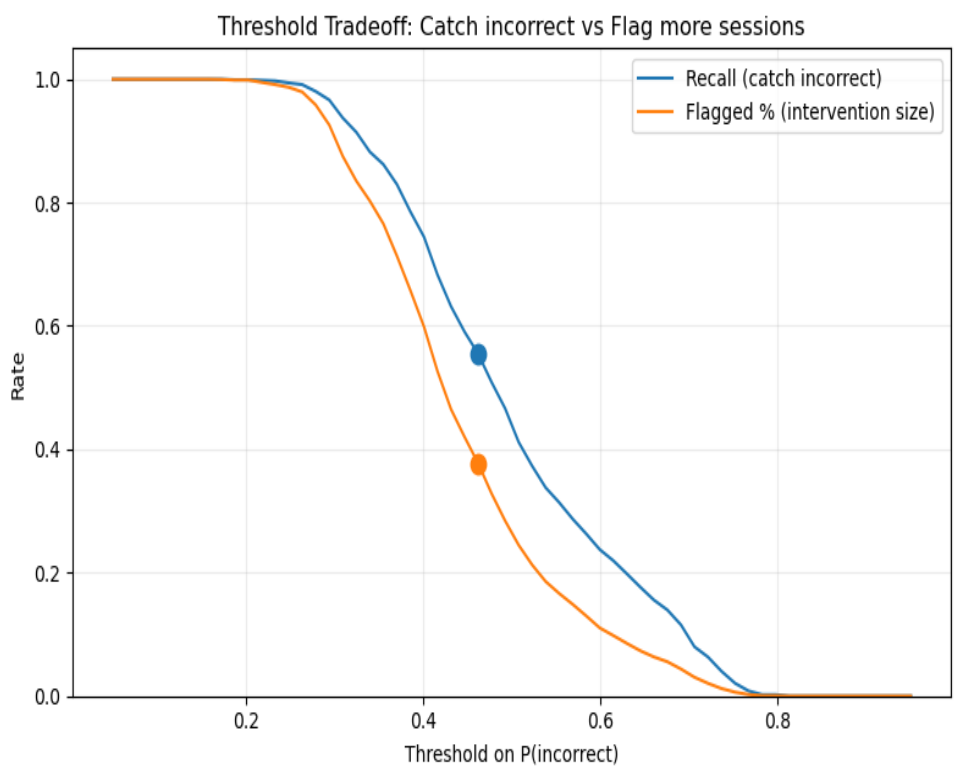
Goal:
Find the best probability threshold for classifying high-risk sessions.

Instead of using the default 0.5 threshold,
we evaluated different cutoff values to balance:

- Recall (capture more incorrect sessions)
- Intervention size (number of sessions flagged)

Trade-Off

- Lower threshold → More sessions flagged, higher recall
- Higher threshold → Fewer sessions flagged, lower recall



Conclusion

- Early gameplay behavior contains meaningful predictive signals
- Random Forest achieved ROC AUC ≈ 0.64
- Top 20% high-risk sessions captured ~44% of incorrect outcomes
- Risk scoring enables targeted, efficient intervention

Main Takeaway:
We can identify struggling sessions early and act before failure occurs.

Even with moderate model performance (AUC ≈ 0.64), early behavioral data is structured enough to meaningfully concentrate risk into a smaller intervention group.

References

- 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.
- 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.
- 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.
- 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.
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- 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.