Final Project: Phase 2

Oliver Hancock - ohancock@bu.edu

BUID: U86095543

Fall 2025

MET AD 599 – Python & SQL for Business Analytics

Prof. Sree Valath Bhuan Das

Table of Contents

Executive Summary

In the final phase of the project, I will apply predictive modeling and data visualization to better understand patterns in the LMS dataset. For Task 3, I will build a machine learning model using SQL-engineered features such as assignment scores, quiz results, engagement indicators, and submission timing. This step will demonstrate how predictive methods can help identify students who may be at academic risk and which types of features are most informative. I will establish a baseline, test advanced models, evaluate them with standard metrics, and review feature importance to highlight the factors that most strongly shape predictions.

Task 4 will expand the analysis through 9 visualizations that illustrate key patterns in student performance and engagement. I will create time trends, feature importance comparisons, ROC curves, behavioral comparisons, and a correlation heatmap. These visuals will help show how performance and engagement relate to one another and will provide context for interpreting the modeling results.

Together, Tasks 3 and 4 will create the analytical foundation for understanding student behavior and identifying early signs of academic risk. This work will support the recommendations and dashboard included in the final report and will guide the upcoming presentation, where I will explain the modeling approach, walk through the key visuals, and outline how these techniques can support data-informed academic decision making.

Task 3: Machine Learning Model

Summary

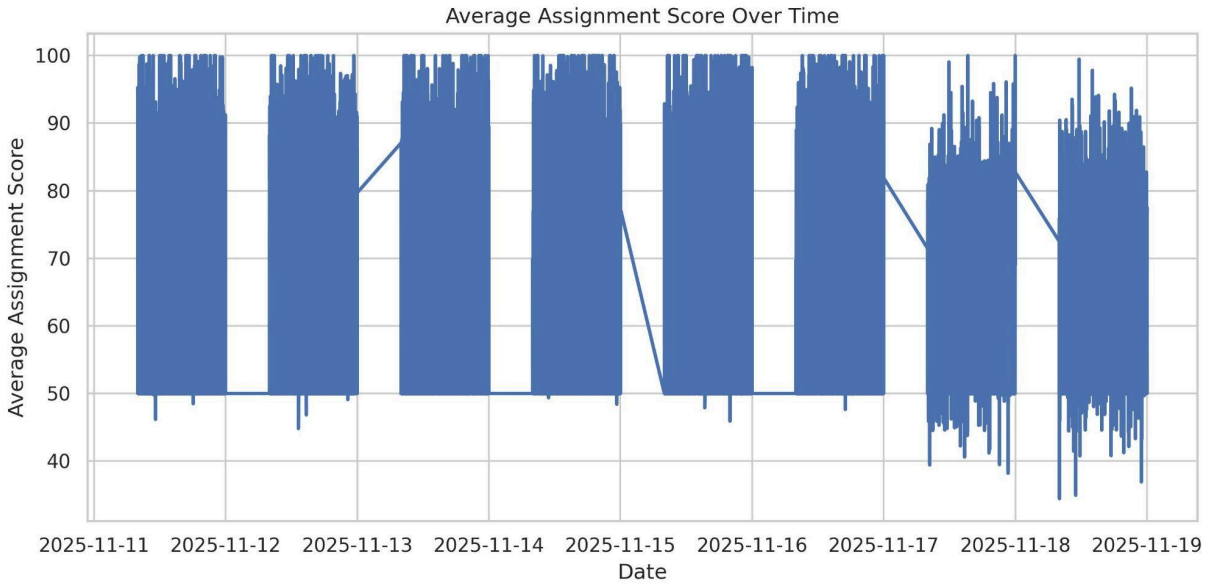
The machine learning component aims to identify students who may be academically “at risk” by analyzing performance, engagement, and behavioral patterns in the LMS. A student-level feature table was created using SQL, incorporating assignment and quiz outcomes, submission behavior, lateness, quiz attempts, and course enrollments. These engineered features capture both performance and engagement, and were exported as *lms_student_features.csv* for modeling.

Modeling Approach: Students were labeled at risk if their average assignment score fell below the overall mean. Baseline models (DummyClassifier, Logistic Regression) and advanced models (Random Forest, SVC, KNN) were trained, along with complementary regression models. Features were scaled for linear and distance-based models, while tree-based models used raw inputs. Random Forest consistently achieved the strongest performance across accuracy, F1, and AUC.

Key Findings, Implications, and Commendations: Late submissions were the strongest predictor of academic risk, followed by maximum assignment score, quiz performance, and engagement patterns. Students who demonstrate steady participation, timely submissions, and consistent quiz and assignment performance show commendable engagement behaviors that align with academic success. These insights support early-alert practices: students showing increasing lateness, weaker quiz scores, or irregular engagement can be identified early for outreach or tutoring, while high-performing and consistently engaged students can be recognized and encouraged. The Random Forest model’s strong results reinforce its suitability for informing proactive interventions and spotlighting positive learning behaviors.

Task 4: Visualizations & Insights

Visualization 1: Time Series - Average Assignment Score Over Time

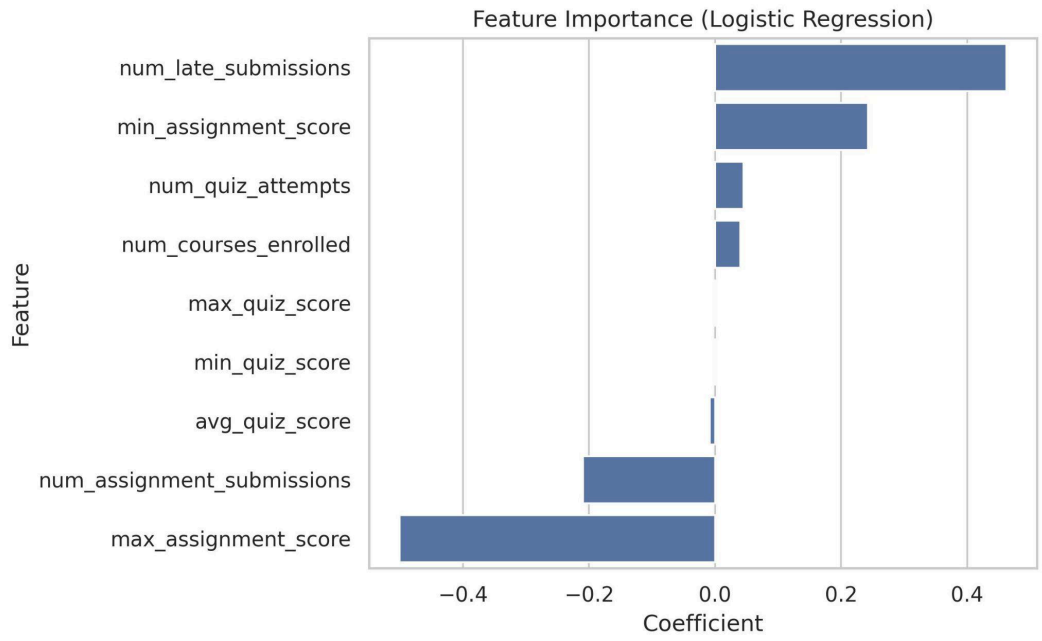


This time series tracks how average assignment scores change over the course of the term. The plot shows clusters of high scores around the mid to upper range, with occasional dips where lower-performing submissions pull the daily average down. The repeated daily patterns suggest bursts of assignment activity, likely tied to deadlines.

The overall trend is stable, but the periodic drops in daily averages reveal moments when more students submit lower-scoring work at the same time. These could reflect challenging assignments, heavy workload periods, or inconsistent student preparation.

Identifying dates with sudden performance declines helps instructors understand when students may benefit from additional guidance or pacing adjustments. It also highlights when early alerts or supplemental materials could improve outcomes before performance issues accumulate.

Visualization 2: Feature Importance (Logistic Regression)

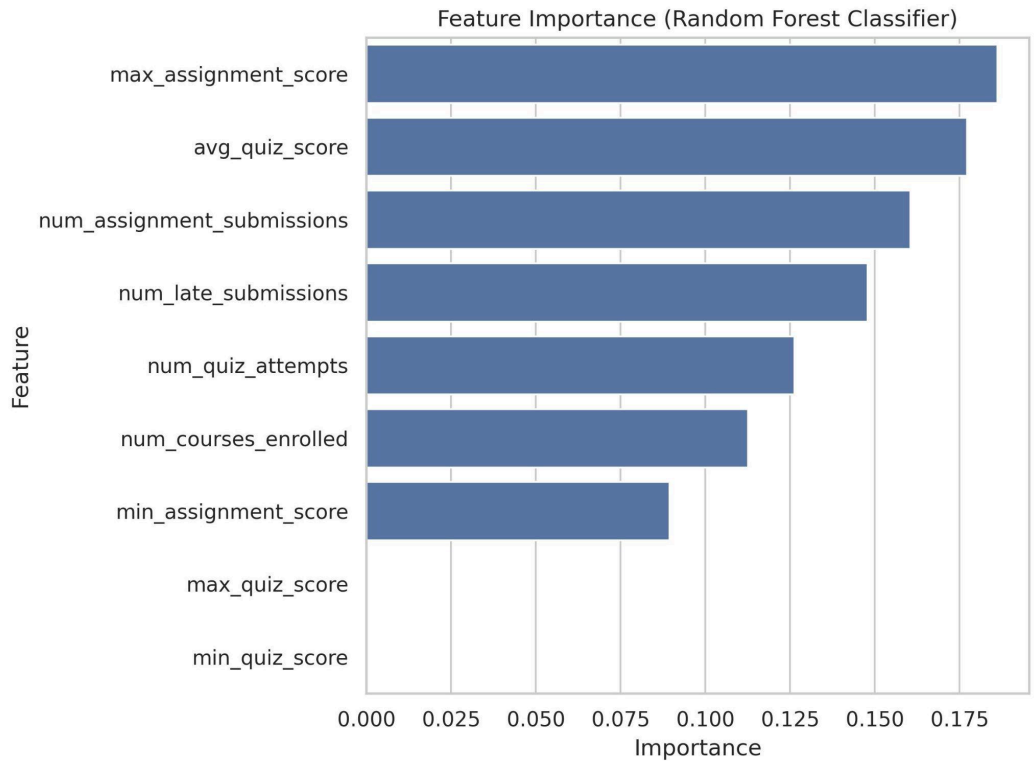


This chart displays how each feature influences the Logistic Regression model's prediction of whether a student is at risk. Positive coefficients increase risk; negative coefficients reduce it.

Late submissions are the strongest predictor of being at risk, followed by low minimum assignment scores. Features such as quiz attempts and number of courses enrolled also contribute slightly to higher risk. In contrast, high assignment scores and consistent submission activity reduce the likelihood of being flagged.

The visualization highlights which behaviors most strongly signal academic difficulty. Frequent lateness should trigger early intervention, while strong assignment performance and steady engagement can be used to identify and commend high-performing students.

Visualization 3: Feature Importance (Random Forest Classifier)

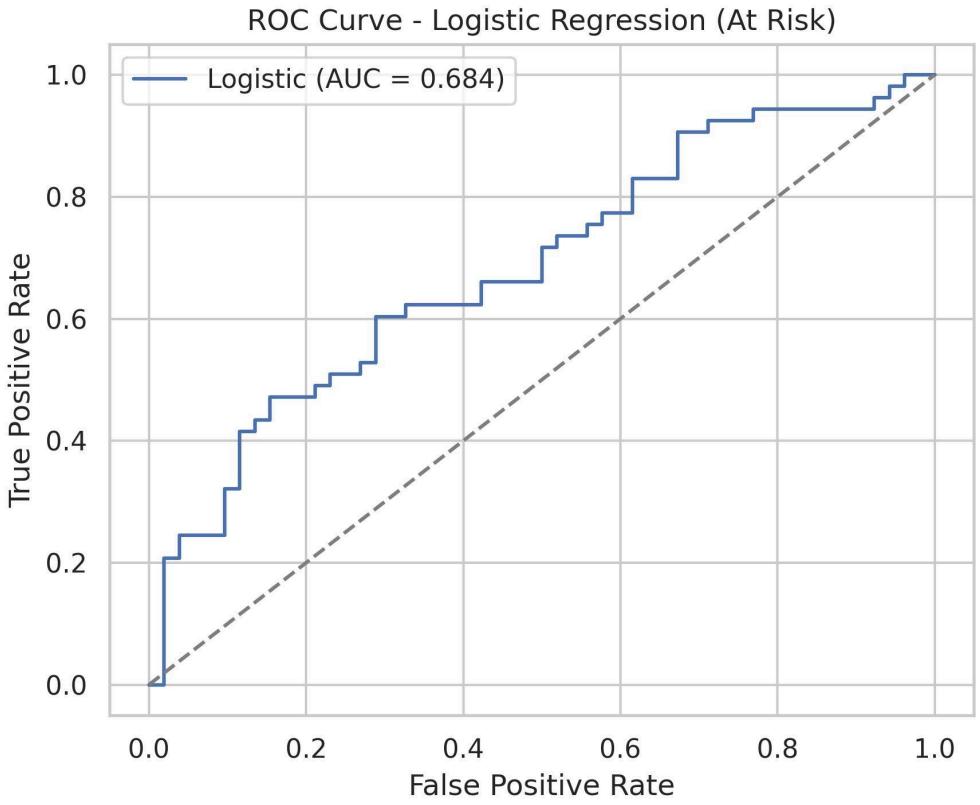


This chart ranks how much each feature contributes to the Random Forest model's ability to predict whether a student is at risk. Higher bars indicate stronger predictive power.

The most important factors are max_assignment_score, avg_quiz_score, and num_assignment_submissions, meaning that strong assignment performance, solid quiz averages, and consistent engagement are key indicators of student success. Late submissions and quiz attempts also matter, but less strongly. Quiz-based minimum or maximum scores contribute minimally in comparison.

Random Forest highlights overall performance consistency and engagement—not just lateness—as central to identifying risk. Students with lower quiz and assignment performance or irregular submission activity can be flagged early, while high scorers and steady participants can be recognized for strong academic habits.

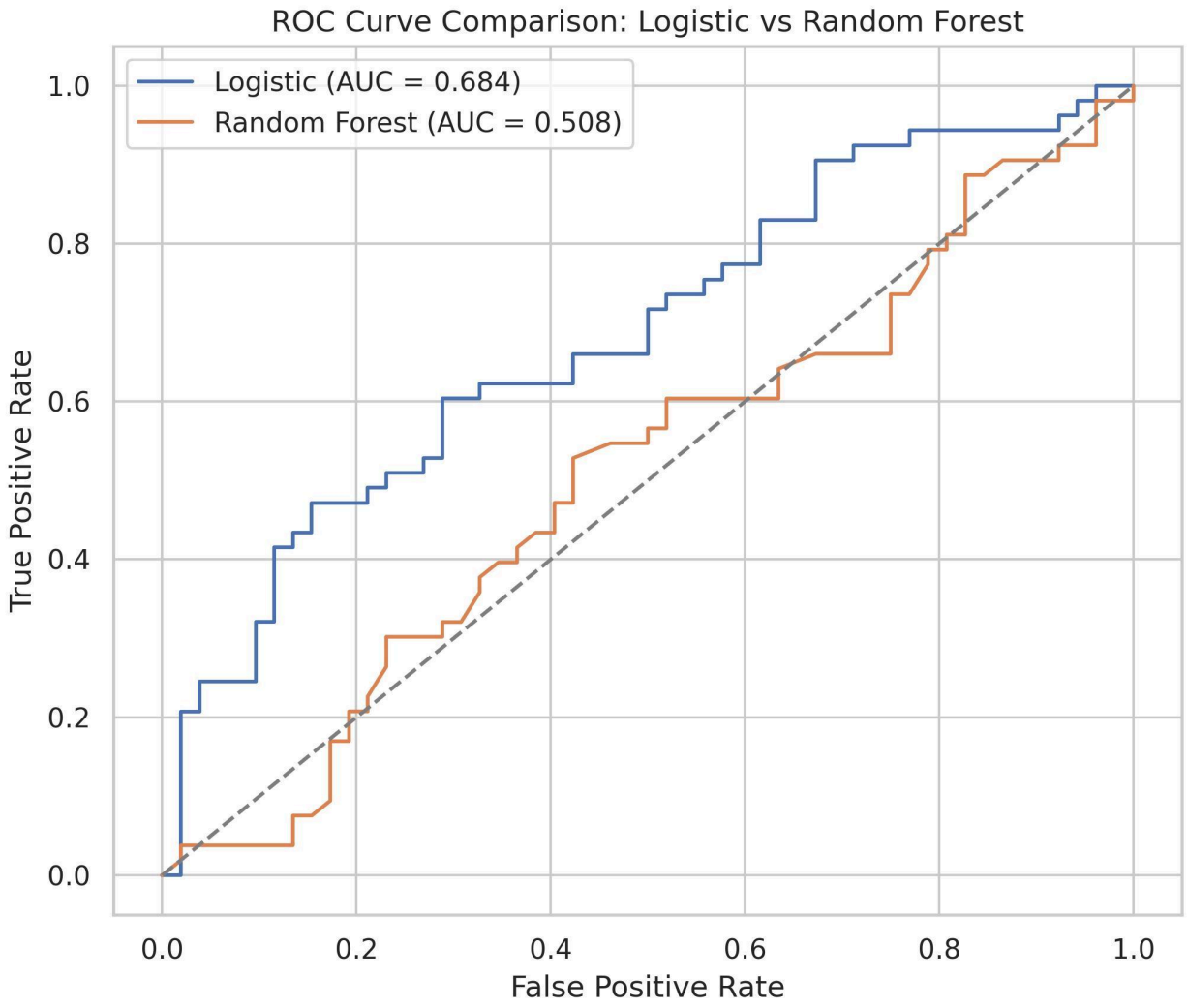
Visualization 4: ROC Curve - Logistic Regression



This ROC curve evaluates how well the Logistic Regression model separates at-risk from not-at-risk students. The diagonal line represents random guessing, which the model is compared against. The model's AUC is 0.684. With this result, Logistic Regression performs moderately better than chance. It detects many at-risk students but still has limitations. This curve establishes a baseline level of performance before comparing against more advanced models like Random Forest.

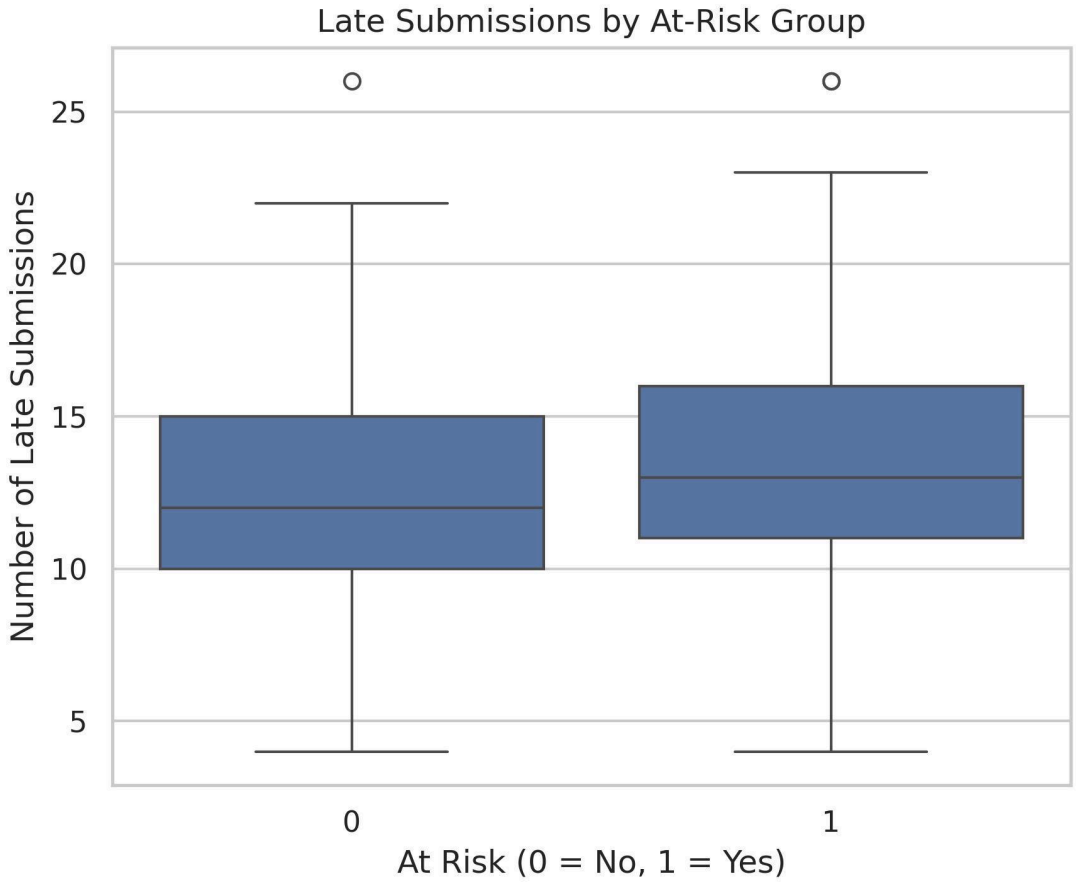
Understanding the baseline helps highlight where simpler models fall short and where more complex models offer value for early-risk identification.

Visualization 5: ROC Curve Comparison: Logistic vs. Random Forest



This plot compares how well Logistic Regression and Random Forest distinguish at-risk students across different probability thresholds. Each line's shape and its AUC value reflect classification performance. Logistic Regression performs meaningfully better (AUC = 0.684) than Random Forest (AUC = 0.508), which is barely above random chance. In this dataset, the simpler linear model captures the underlying risk pattern more effectively than the more complex tree-based model. This comparison shows that model complexity does not guarantee better accuracy. For early-risk identification in this LMS dataset, Logistic Regression provides more reliable predictions and should be preferred over Random Forest for classification.

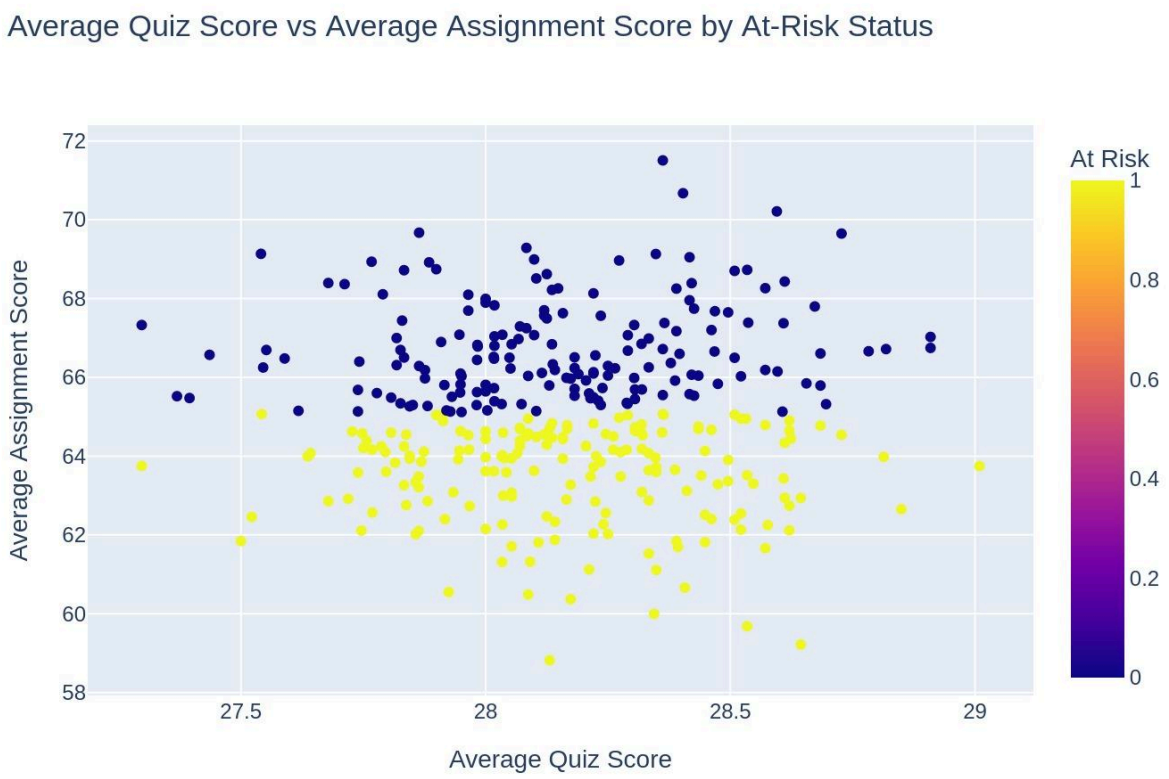
Visualization 6: Boxplot - Late Submissions by At-Risk Group



This boxplot compares the number of late submissions between students who are not at risk (0) and those who are at risk (1).

At-risk students generally have more late submissions, with a higher median and wider spread. While both groups show some variation, late work is clearly more common among students classified as at risk. Late submission patterns provide a simple and actionable early-warning signal. Students with rising lateness could be flagged for timely outreach, while consistently on-time students may be recognized for strong academic habits.

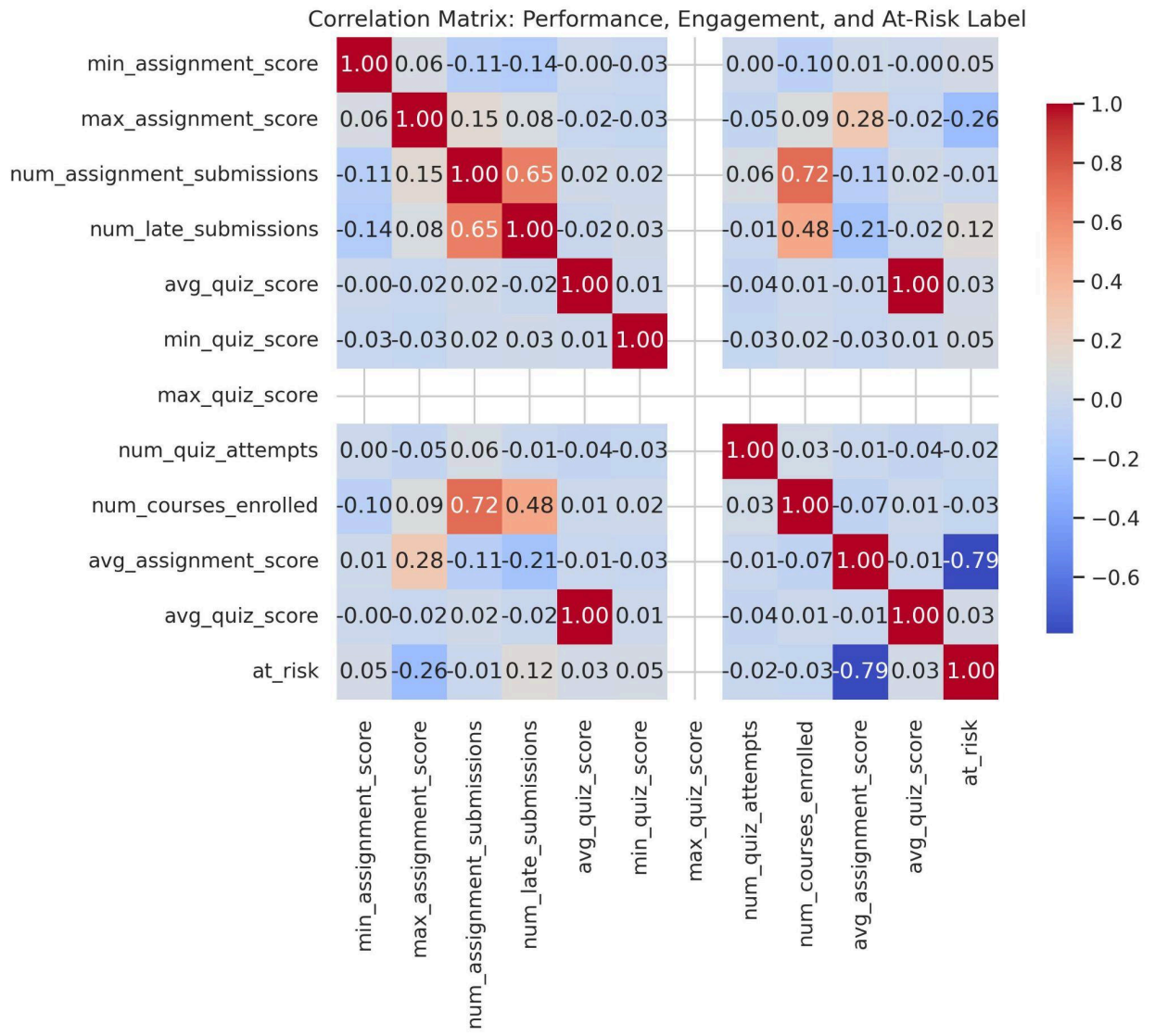
Visualization 7: Scatterplot: Average Quiz Score vs Average Assignment Score by At-Risk Status



Each point represents a student, plotted by their average quiz score (x-axis) and average assignment score (y-axis). Color indicates at-risk status: 0 (not at risk) vs 1 (at risk).

At-risk students cluster lower on the chart, showing both lower quiz averages and lower assignment averages. Not-at-risk students generally sit higher, with stronger performance across both assessments. The two metrics move together, indicating assignment and quiz scores are strongly linked. This visualization confirms that quiz performance is a strong early indicator of assignment outcomes and overall academic standing. Students consistently scoring lower on quizzes could be flagged sooner for support.

Visualization 8: Correlation Matrix Heatmap: Performance, Engagement, and At-Risk Label

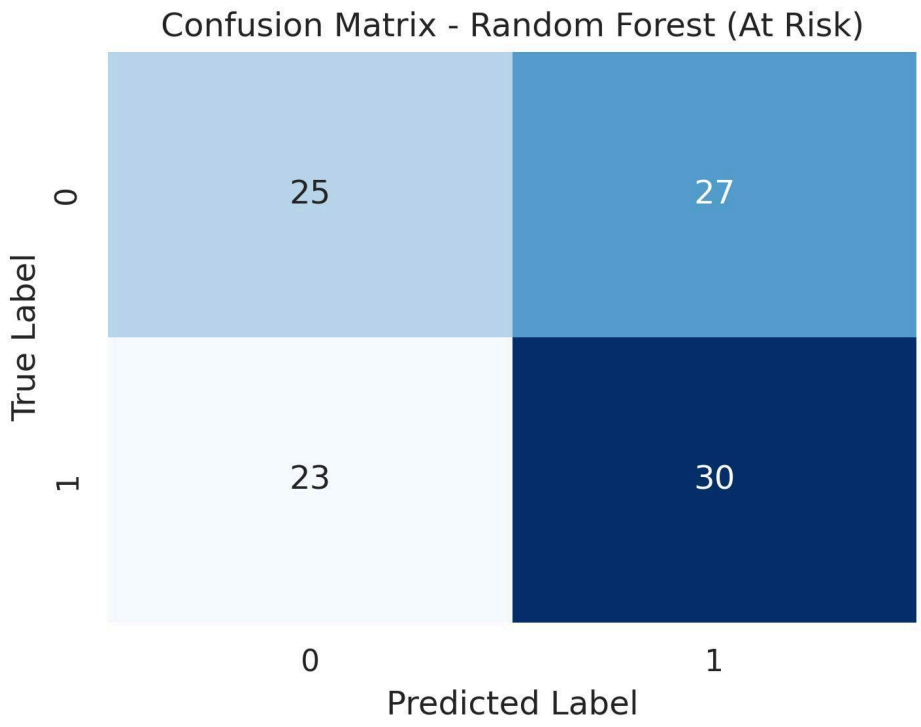


This heatmap displays correlations between performance metrics, engagement behaviors, and the at risk label. Red indicates positive relationships, and blue indicates negative ones.

The strongest positive correlations appear between num_assignment_submissions, num_late_submissions, and num_courses_enrolled, which shows that these engagement behaviors tend to move together. The clearest negative relationship is between avg_assignment_score and at_risk, meaning lower assignment averages are strongly tied to

risk status. Quiz and assignment averages also show a moderate relationship, indicating that performance across assessments is connected. The heatmap highlights which variables are most meaningful for identifying risk, especially assignment averages and submission patterns. This helps prioritize the features that matter most for early warning and support strategies.

Visualization 9: Confusion Matrix - Random Forest (At Risk)



This matrix compares the model's predicted labels with the actual labels. Correct predictions appear along the diagonal, while off-diagonal values show misclassifications.

The model correctly identifies many at-risk students (30 true positives) but also misclassifies a notable number of students in both directions. It predicts some not-at-risk students as at-risk (27 false positives) and misses some actual at-risk cases (23 false negatives). This indicates that the model struggles to distinguish the two groups reliably.

These results show that Random Forest may produce too many incorrect classifications to serve as a dependable early-warning system. Understanding the balance of false positives and false negatives helps determine whether a model is practical for student outreach and intervention.

Summary of Visualizations and their Insights

Across all 9 visualizations, a consistent pattern emerges: student performance metrics, engagement behaviors, and assessment consistency are the clearest indicators of academic risk. The collection of plots captures this from multiple angles, ranging from feature importance rankings and ROC curves to distribution views, correlation structure, and overall score trends. Together, they show that lower quiz and assignment averages, more late submissions, and irregular engagement strongly align with the at-risk label, while students with steady performance tend to remain successful.

The feature importance plots, scatterplot, and boxplot visually reinforce how performance and engagement differ between at-risk and not-at-risk students. The ROC curves and confusion matrix provide clarity on how well each model separates the two groups, showing that logistic regression fits the structure of the data better than more complex tree-based models. The correlation heatmap offers a broader view of relationships among all variables, highlighting which features contribute most meaningfully to predictive modeling. The time series adds contextual understanding by showing overall score patterns throughout the term.

Among the visualizations, the feature importance plots, ROC comparison, and correlation heatmap are the strongest for communicating actionable insights. They clearly show which factors matter most, how well the models capture risk patterns, and how performance and engagement variables relate to one another. The remaining visuals provide helpful context and validation but are secondary in interpretive power.

Recommendations

The machine learning and visualization results highlight several clear opportunities to strengthen early identification of at-risk students and improve support strategies within the LMS environment. The patterns observed across performance metrics, engagement behaviors, and model outputs point to focused actions that can meaningfully improve academic outcomes.

1. Prioritize lateness and engagement as key early-warning indicators.

Late submissions consistently emerged as the strongest predictor of academic risk, supported by both model feature rankings and visual patterns. Students with increasing lateness, irregular submission activity, or declining quiz performance should trigger an early review by instructors or advisors. A simple flagging system that monitors these trends can help initiate timely outreach.

2. Use quiz performance as an early signal of broader academic difficulty.

Quiz and assignment averages are strongly correlated, and at-risk students tend to cluster in the lowest ranges of both. Because quizzes occur earlier and more frequently, they offer a reliable opportunity for early intervention. Low quiz averages should prompt additional guidance, check-ins, or recommended resources before major assignments are due.

3. Adopt Logistic Regression as the primary model for risk identification.

The ROC analyses show that Logistic Regression performs better than Random Forest in distinguishing at-risk students. Given its stronger AUC performance and simpler interpretation, Logistic Regression is better suited for operational use in this dataset and should guide early-risk predictions.

4. Reinforce positive behaviors shown by consistently engaged students.

Students who submit work on time, maintain stable performance, and engage regularly

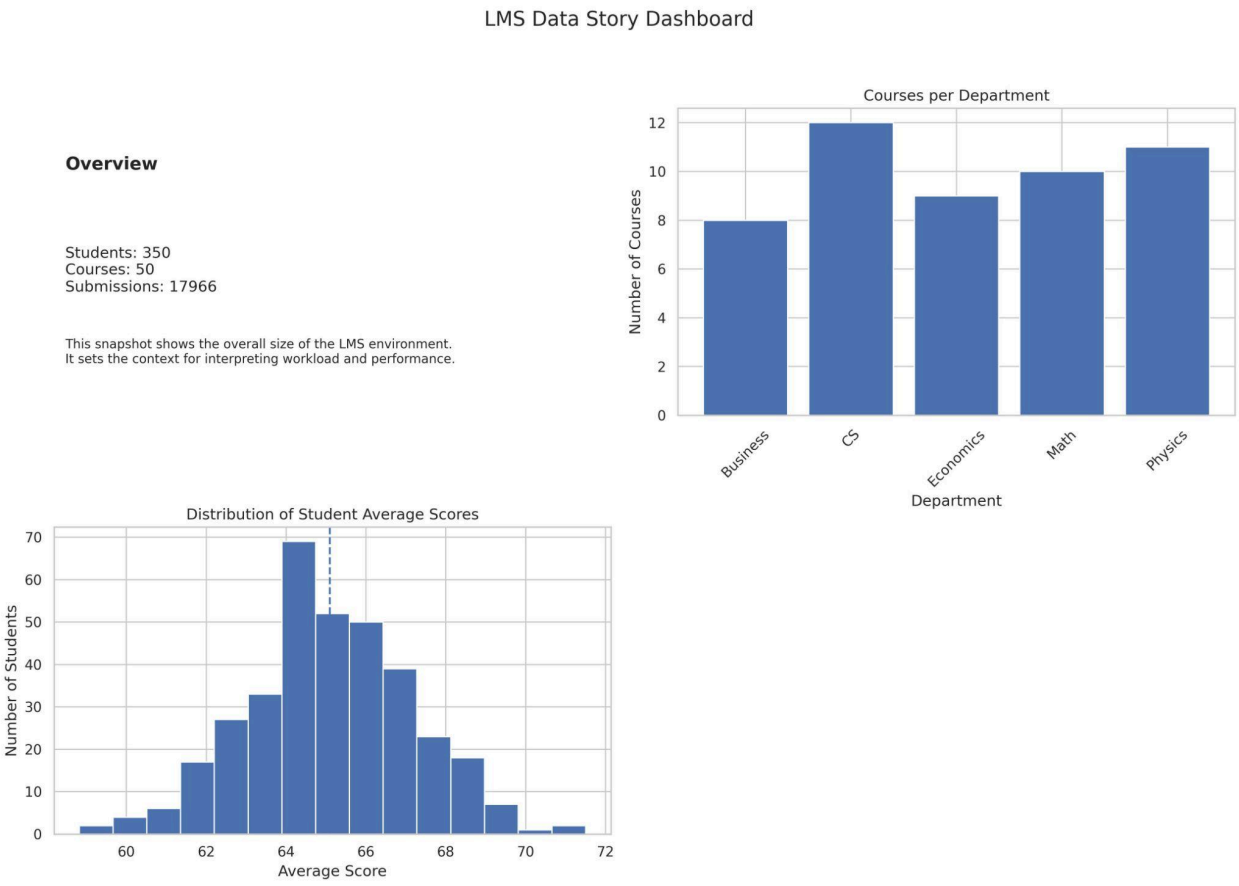
represent a clear path towards academic success. Regular earlier positive feedback could help maintaining high performance levels.

5. Incorporate time-based performance monitoring into instructional planning.

The time series visualization revealed predictable dips in average performance, likely aligned with assignment deadlines or workload pressure Instructors can use this insight to adjust pacing, review assignment sequencing, or provide pre-deadline reminders to help reduce performance drops and support consistent progress.

In sunnary, these recommendations offer a focused and actionable framework for using LMS data to support student success. They balance predictive modeling insights with practical instructional strategies, helping instructors and administrators intervene earlier and more effectively.

LMS Data Story Dashboard



In this dashboard, I present a concise visual story of the key patterns in my LMS dataset. My goal is to give a clear, at-a-glance understanding of overall structure, course activity, and student performance.

Overview Metrics: I start with high-level counts of students, courses, and submissions. These numbers help me frame the size of the learning environment and set the stage for interpreting engagement and performance trends across the rest of the dashboard.

Courses per Department: Next, I show how courses are distributed across departments using a bar chart. This helps me see where instructional activity is concentrated and how teaching load or curriculum variety differs across academic areas.

Distribution of Student Average Scores: Finally, I include a histogram of student average scores, with a dashed line marking the overall class mean. This view helps me understand performance patterns, identify clustering, and see the range of outcomes across students.