The goal of this analysis was to predict appointment attendance failure—defined as both no-shows and cancellations—to help the clinic reduce wasted provider time and improve scheduling efficiency. Three models were evaluated using a unified preprocessing pipeline: Logistic Regression, RandomForest, and GradientBoosting. Performance was measured using accuracy, precision, recall, F1, and AUC under a 0.5 threshold.

The results in the final metrics table show clear behavioral patterns. Logistic Regression performs surprisingly well as a baseline, with accuracy 0.6241, recall 0.5406, F1 0.3647, and the highest AUC (0.6239). RandomForest performs similarly, achieving accuracy 0.6230, recall 0.5129, F1 0.3519, and AUC 0.6026. In contrast, GradientBoosting has the highest accuracy (0.7887) but almost no ability to detect failures (recall 0.0203, F1 0.0369), meaning it is operationally ineffective. These findings indicate that overall accuracy is misleading due to class imbalance: models that appear highly "accurate" may still fail to identify missed appointments. In this project, Logistic Regression and RandomForest provide the most meaningful predictive performance.

Feature importance analysis (from RandomForest) reveals that historical engagement patterns are the strongest drivers of attendance failure. The most influential feature is num_previous_no_shows with importance 0.3307, confirming that patients with repeated no-shows are substantially more likely to miss future appointments. Behavioral and operational factors are stronger predictors than demographics. This implies that risk is closely tied to prior engagement and internal workflow patterns, providing a clear direction for targeted interventions.

Two diagnostic analyses highlight important risks. First, the ID removal experiment shows that eliminating pure identifiers causes performance to drop sharply: with IDs included, RandomForest achieves AUC 0.6026 and F1 0.3519, but without them, performance falls to AUC 0.5325 and F1 0.1014. This indicates that IDs encode hidden structural signals such as provider-specific scheduling patterns or time-slot regularities, but relying on IDs risks reduced generalization. This represents a key modeling limitation and identifies an opportunity for feature engineering.

Second, the threshold tuning experiment demonstrates the trade-off between recall and precision. This reveals an important operational decision point: the clinic can choose higher recall at the cost of more outreach workload, or higher precision with fewer interventions.

Overall, the findings illustrate that historical behavior is the most powerful signal of attendance reliability, while purely demographic variables are weak predictors. Models like Logistic Regression and RandomForest offer practical predictive value when combined with threshold tuning. To strengthen future performance, the clinic could incorporate richer behavioral, temporal, and contextual features; engineer interpretative replacements for ID columns; and adopt cost-sensitive decision thresholds tailored to operational constraints. These insights together provide a foundation for data-driven interventions to reduce missed appointments and improve clinic efficiency.