Nile E-Commerce Review Prediction
Project Summary

1. Introduction

This project was collaboratively developed by Group 28 as part of the Analytics in Practice module in the MSc Business Analytics program at the University of Warwick. We aim to predict which customers of Nile, a fictional South American e-commerce platform, are likely to leave positive reviews. Accurate prediction can help reduce manual reviews, target dissatisfied customers, and optimize loyalty programs.

2. Business Context

Customer reviews are critical for online businesses like Nile. Reviews influence future sales, vendor accountability, and platform trust. Our goal is to identify customers most likely to leave a positive review using machine learning. This enables targeted post-purchase engagement and reduces unnecessary follow-up with already satisfied customers. Understanding the drivers of reviews also supports operational and product-level decisions.

3. Data Engineering

Feature engineering was guided by a combination of domain understanding and exploratory data analysis. Delivery metrics were created from shipping timelines. Review scores were aggregated at multiple levels: seller, category, and customer state. To avoid data leakage, we applied frequency thresholds and used global averages for low-count entities. Review ratings were reclassified into three buckets (1–3, 4, and 5 stars), aligning better with class distribution and industry interpretation.

4. Modelling & Evaluation

Three models were trained: Random Forest, GBDT, and XGBDT. The best-performing model, XGBDT, showed a balance of generalization and precision. Evaluation focused on test set results only, avoiding inflated training scores. We used F1-score, precision, and recall to compare models. XGBDT was chosen due to its superior F1-score and stable performance under class imbalance, which we addressed via class weighting and data balancing.

5. Deployment & Recommendations

The final model can be used to prioritize outreach and improve satisfaction by proactively targeting customers likely to leave negative reviews. We recommend continuous model retraining as more data becomes available, and propose marketing interventions such as coupons or surveys for low-score segments. Class imbalance risks overestimating model reliability; thus, regular monitoring and feedback loops should be established.