From Browsing to Buying: Predicting Online Purchase Intention

Summary

This project uses the Online Shoppers Purchasing Intention Dataset to study whether an online browsing session is likely to end in a purchase. The response variable is Revenue, where TRUE means the session generated a transaction. The project is motivated by a practical business question: can browsing behavior help an online retailer identify visitors with high purchase intent? Since only 15.5% of sessions end in purchase, the analysis will focus on precision, recall, F1 score, ROC AUC, and precision-recall performance rather than accuracy alone.

Introduction and Motivation

Most e-commerce visitors leave without buying anything, but websites collect useful session-level information before that happens: page visits, time spent on pages, traffic source, visitor type, and timing. A prediction model could help retailers decide which users should receive a promotion, recommendation, reminder, or other marketing intervention. In this setting, false positives may waste promotional resources, while false negatives may miss customers who were close to purchasing.

The main research questions are:

- Which browsing features are most associated with purchase completion?
- How do purchase sessions differ from non-purchase sessions?
- Can interpretable models compete with nonlinear models?
- How should performance be evaluated when the minority class is the business outcome of interest?

Data Overview

The dataset contains 12,330 online shopping sessions, 17 predictors, and one binary response variable. There are no missing values. The target distribution is imbalanced: 10,422 sessions did not lead to purchase (84.5%), while 1,908 sessions did (15.5%). The predictors describe page activity, web analytics, timing, traffic source, technical category, and visitor type.

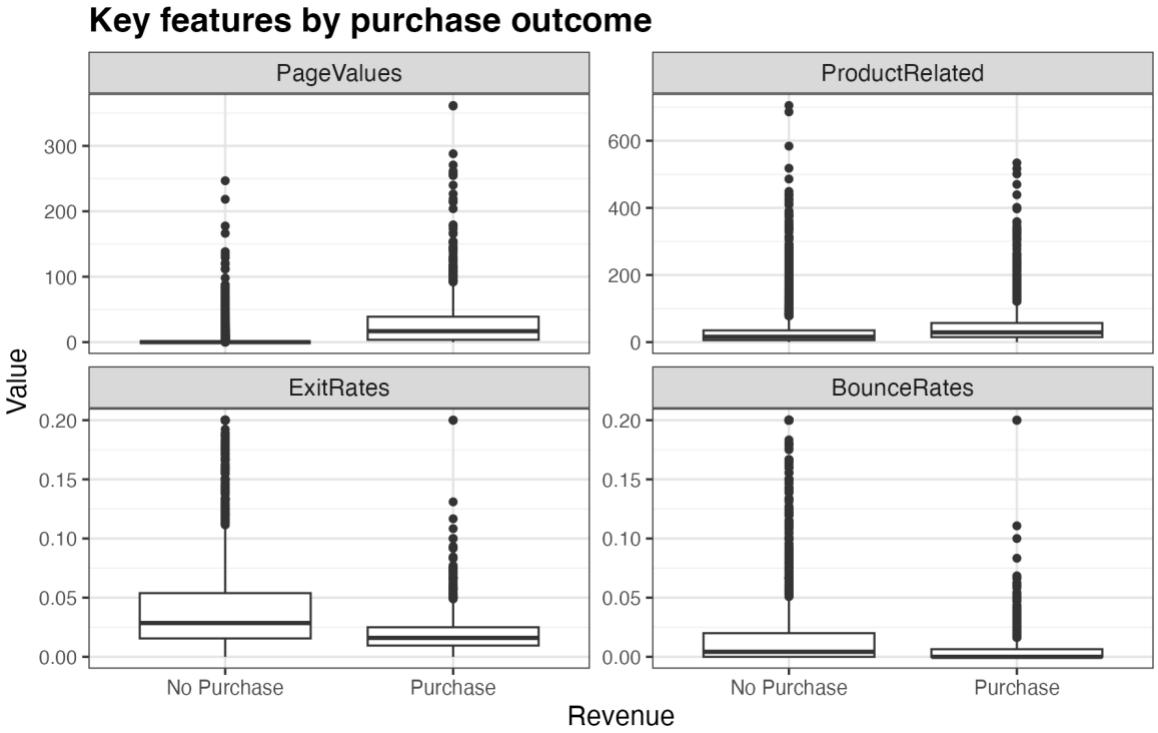
Data Preprocessing and Exploration

Categorical variables such as Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, and Weekend will be treated as factors. Although some are stored as integers, they represent category codes rather than continuous measurements. For regression-based and distance-based models, they will be dummy encoded.

Numerical variables have very different scales. For example, ProductRelated_Duration ranges from 0 to 63,973.52 seconds, while BounceRates and ExitRates range only from 0 to 0.20. Numerical predictors will therefore be standardized for scale-sensitive methods such as Logistic Regression, Lasso, SVM, KNN, and PCA.

Outliers were checked using the 1.5 x IQR rule. Many behavior variables are right-skewed and zero-inflated, so the rule flags many meaningful high-engagement sessions. For example, PageValues has 2,730 flagged observations and Informational has 2,631. These observations will be retained because high page values, long durations, and high page counts are likely useful behavioral signals rather than data errors.

Preliminary exploration shows clear differences between purchase and non-purchase sessions. Purchase sessions have much higher average PageValues (27.26 vs. 1.98), more product-related page views (48.21 vs. 28.71), longer product-page duration (1,876.21 vs. 1,069.99 seconds), lower ExitRates (0.020 vs. 0.047), and lower BounceRates (0.005 vs. 0.025). Purchase rates also vary by timing and visitor type: November has a purchase rate of about 25.4%, and new visitors have a purchase rate of about 24.9%.



Correlation analysis shows two strong numerical relationships: BounceRates and ExitRates are highly correlated (r = 0.913), and ProductRelated is strongly correlated with ProductRelated_Duration (r = 0.861). This motivates regularization or dimension reduction in the modeling stage. Since PageValues appears especially predictive, the final project will also compare models with and without this variable to check whether the conclusions depend too heavily on a single web analytics feature.

Planned Modeling Approaches

The final project will compare Logistic Regression, Lasso Logistic Regression, PCA + Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine with a radial kernel. This model set covers classical modeling, regularization, dimension reduction, ensemble methods, and nonlinear classification. The final evaluation will use cross-validation for model tuning and a held-out test set for performance comparison, with special attention to the minority purchase class.