

# From Browsing to Buying: Predicting Online Purchase Intention

Data source: <https://doi.org/10.24432/C5F88Q>

Group: Jiawei Guo, Duli Lei, Jingyi Guo

## Abstract

Most visitors leave an online store without buying. This project asks whether session-level browsing behavior can reveal which visitors are closer to purchase. Using the UCI Online Shoppers Purchasing Intention Dataset, we model Revenue as an imbalanced binary response, with purchases representing only 15.5% of sessions. The analysis combines exploratory data analysis, preprocessing diagnostics, supervised model comparison, and model interpretation. Purchase sessions show higher PageValues, stronger product engagement, and lower ExitRates and BounceRates. Random Forest and Gradient Boosting provide the strongest held-out test performance, with ROC AUC around 0.92 and F1 around 0.66 after thresholds are selected using training-set cross-validated predictions only. Sensitivity and profile analyses show that PageValues is the dominant predictive signal, but it should be interpreted as an analytics-derived indicator of purchase intent rather than causal evidence.

## 1. Introduction and Motivation

Most e-commerce visits do not convert, but retailers observe rich session-level behavior before the final outcome is known. A purchase-intent model can help decide which visitors might receive recommendations, reminders, or promotional interventions. In this setting, a false positive may waste marketing resources, while a false negative may miss a visitor who was close to buying. The practical goal is therefore not only to classify sessions, but to identify the behavioral signals that separate high-intent visits from ordinary browsing.

- Which browsing behaviors distinguish purchase sessions from non-purchase sessions?
- Can machine learning models identify likely purchasers on unseen sessions?
- Do flexible tree-based models improve performance over interpretable linear models?
- How dependent are the results on PageValues?

## 2. Data and Preprocessing

The dataset and preprocessing decisions are summarized below with emphasis on variable meaning, skewness, and modeling implications.

**1. Dataset structure.** The data contain 12,330 online shopping sessions collected over a one-year period. Each row is a session, not a customer, so the analysis predicts session-level purchase intention rather than long-term customer behavior.

**2. Predictor groups and response.** The response is Revenue, coded as Purchase versus NoPurchase. Predictors fall into five groups: page counts and durations (Administrative, Informational, ProductRelated and their duration variables), web-analytics metrics (BounceRates, ExitRates, PageValues, SpecialDay), timing (Month, Weekend), traffic/technical identifiers (TrafficType, Browser, OperatingSystems, Region), and visitor type.

**3. Class balance and missingness.** Purchase is the minority class, with 1,908 purchase sessions (15.5%) and 10,422 non-purchase sessions (84.5%). The data have no missing values, but the class imbalance makes accuracy alone misleading.

**4. Encoding and scaling.** Categorical variables were treated as factors. Numeric variables were centered and scaled for Logistic Regression, LASSO, PCA, and SVM so that variables measured on large scales, such as duration, would not dominate distance- or penalty-based methods.

**5. Outliers and zero inflation.** Several behavioral variables are highly right-skewed and zero-inflated. For example, PageValues has  $Q1 = Q3 = 0$ , so the IQR rule mechanically flags every positive PageValues observation; this affects 2,730 sessions (22.1%). These values are not treated as data errors because positive PageValues is itself a meaningful purchase-intent signal.

**6. Correlation and redundancy.** Two pairs of numeric predictors are strongly correlated: BounceRates with ExitRates ( $r = 0.913$ ), and ProductRelated with ProductRelated\_Duration ( $r = 0.861$ ). This motivated including LASSO and PCA as checks on redundant predictors, while tree ensembles were used to capture nonlinear behavior.

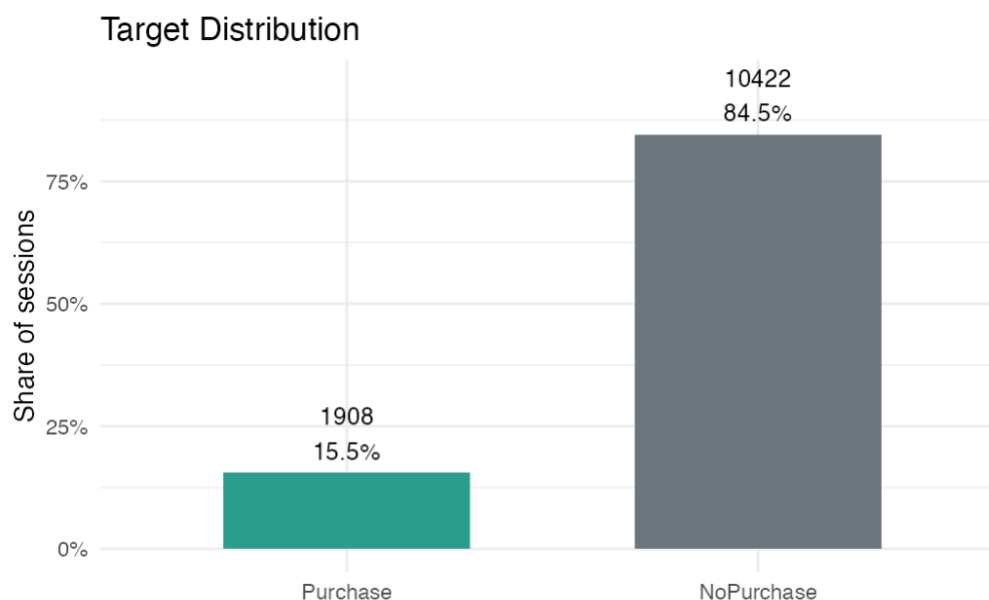


Figure 1. Target distribution shows a strongly imbalanced purchase class.

### 3. Exploratory Findings

Purchase sessions show consistently stronger engagement than non-purchase sessions. Average PageValues is 27.26 for purchase sessions compared with 1.98 for non-purchase sessions. Purchasers also view more product-related pages, spend more time on those pages, and have lower ExitRates and BounceRates. These patterns make business sense: visitors who inspect more products and avoid immediate abandonment are more likely to complete a transaction.

Feature	No purchase mean	Purchase mean	Interpretation
PageValues	1.98	27.26	Much higher page value among purchasers
ProductRelated	28.71	48.21	Purchasers view more product pages
ProductRelated_Duration	1,069.99	1,876.21	Purchasers spend more time on products
ExitRates	0.047	0.020	Purchasers are less likely to exit
BounceRates	0.025	0.005	Purchasers are less likely to bounce

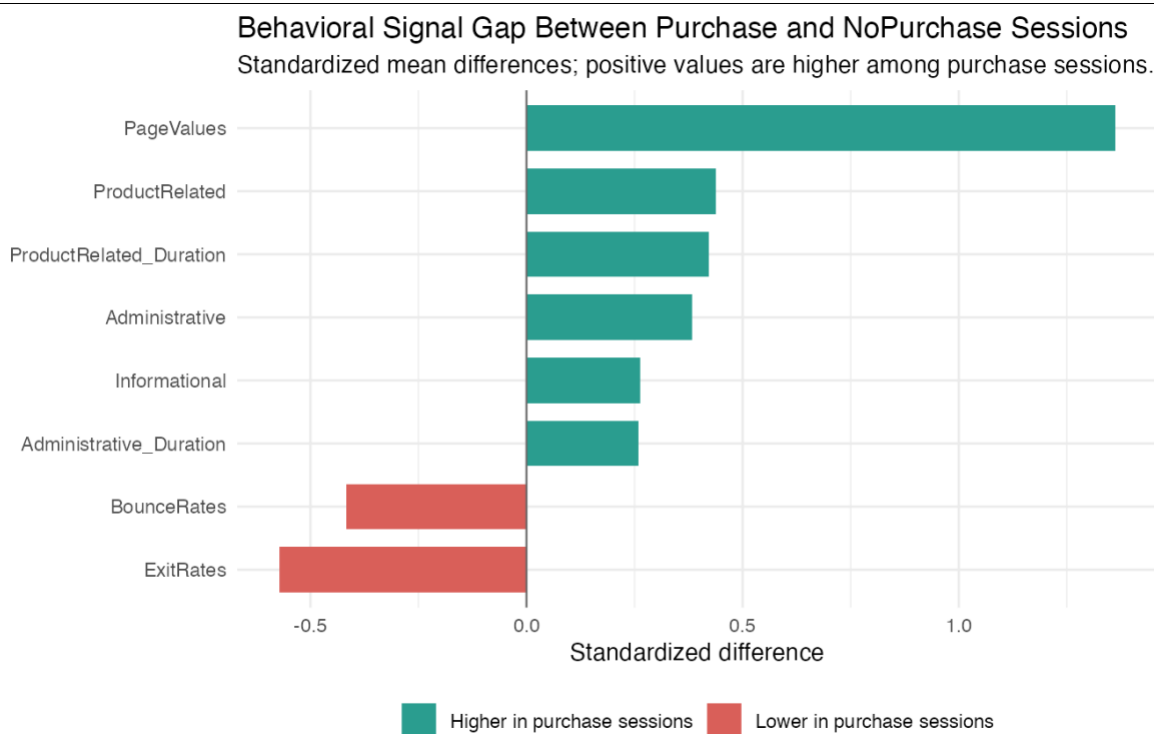


Figure 2. Behavioral signal gap between purchase and non-purchase sessions.

Timing and visitor type also matter. November has the highest major-month purchase rate at 25.4%, while February is only 1.6%. New visitors convert at 24.9%, compared with 13.9% for returning visitors. Weekend sessions have a modestly higher purchase rate than weekday sessions, 17.4% versus 14.9%. Correlation analysis also identifies two highly related predictor pairs: BounceRates with ExitRates ( $r = 0.913$ ), and ProductRelated with ProductRelated\_Duration ( $r = 0.861$ ). These correlations motivate LASSO and PCA as principled responses to redundancy among predictors.

#### 4. Modeling Strategy

The modeling workflow was organized around eight design choices.

**1. Train-test split.** We used a stratified 80/20 split so the purchase rate was preserved in both the training set and the held-out test set.

**2. Model set.** We compared interpretable baselines and more flexible nonlinear methods: Logistic Regression, LASSO Logistic Regression, PCA + Logistic Regression, Random Forest, Gradient Boosting, and Radial SVM. The purpose was not only to maximize accuracy, but also to compare

interpretability, regularization, dimension reduction, nonlinear boundaries, and tree-based interaction modeling.

**3. Parameter sources.** The tuning grids were chosen before looking at the test set and were based on standard course methods and common caret model controls: lambda values for LASSO, mtry values for Random Forest, tree depth and number of trees for Gradient Boosting, and sigma/C values for Radial SVM. This keeps the search broad enough to compare model families without overfitting the held-out test set.

**4. Tuning rule.** Hyperparameters were selected with 5-fold cross-validation on the training set using ROC AUC, which measures ranking quality across possible thresholds. The final selected settings are reported in the appendix.

**5. PCA diagnostic.** PCA was used as a dimension-reduction baseline through the preprocessing pipeline, but a separate diagnostic shows that the data are not well summarized by only a few components: PC1 explains 12.7% of variance, and about 16, 20, and 22 components are needed to explain 80%, 90%, and 95% of variance. This supports the interpretation that PCA compression does not create a clearly superior predictive representation.

Diagnostic	Value	What it means
Encoded predictor dimensions	28	Categorical variables are expanded before scale-sensitive models.
PC1 explained variance	12.7%	No single component dominates the behavioral information.
PCs for 80% variance	16	PCA retains many dimensions before preserving most variation.
PCs for 90% variance	20	Information is spread across many predictors.
PCs for 95% variance	22	Compression is limited, so purchase signal may be diluted.

**6. SVM kernel choice.** A radial kernel was used as the nonlinear SVM baseline because it can capture smooth nonlinear boundaries. A linear SVM would overlap conceptually with the linear baselines, while a polynomial kernel would add degree-specific tuning and is less stable for this mix of scaled numeric and dummy-coded categorical predictors.

**7. Threshold selection.** Because Purchase is the minority class, classification thresholds were selected using training-set cross-validated predictions to improve Purchase-class F1. The held-out test set was used only for final evaluation.

**8. Evaluation metrics.** We report Accuracy, Balanced Accuracy, Precision, Recall, F1, and ROC AUC. Accuracy is included for completeness, but Precision, Recall, F1, and Balanced Accuracy are more informative for this imbalanced classification problem.

## 5. Model Evaluation

Random Forest and Gradient Boosting give the strongest overall test-set performance. Random Forest has the highest F1 and balanced accuracy, while Gradient Boosting has the highest ROC AUC. The linear models are slightly weaker but remain useful as interpretable baselines. PCA + Logistic Regression does not improve over the original logistic model, suggesting that unsupervised dimension reduction may remove class-specific signal useful for predicting Revenue.

Model	Threshold	Acc.	Bal. Acc.	Prec.	Recall	F1	ROCAUC
Gradient Boosting	0.25	0.877	0.828	0.577	0.759	0.655	0.922
Random Forest	0.30	0.882	0.836	0.590	0.769	0.667	0.921
Radial SVM	0.14	0.886	0.810	0.614	0.701	0.654	0.894
LASSO Logistic Regression	0.15	0.867	0.808	0.553	0.722	0.626	0.892
PCA + Logistic Regression	0.27	0.880	0.766	0.612	0.601	0.607	0.883
Logistic Regression	0.24	0.871	0.771	0.575	0.627	0.600	0.881

At the selected thresholds, the confusion counts show the tradeoff between capturing purchasers and limiting false alarms. Random Forest identifies the most purchasers among the top models (293 true positives) but also creates 204 false positives; Radial SVM is more conservative, with fewer false positives but more missed purchases.

Model	TN	FP	FN	TP
Gradient Boosting	1,872	212	92	289
Random Forest	1,880	204	88	293
Radial SVM	1,916	168	114	267
LASSO Logistic Regression	1,862	222	106	275
PCA + Logistic Regression	1,939	145	152	229
Logistic Regression	1,907	177	142	239

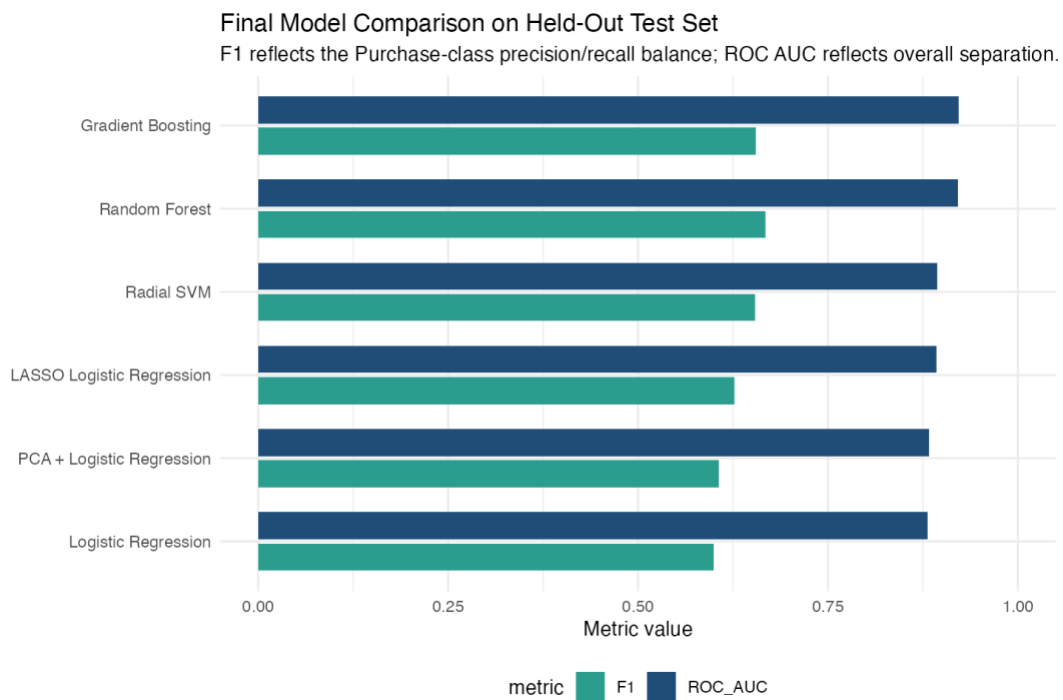


Figure 3. Model ranking by F1 and ROC AUC.

The bar chart makes the ranking easier to read than the full table: the top models are close, but the two tree ensembles lead on the balance of F1 and ROC AUC. The ROC curves below provide a threshold-free comparison of the same models.

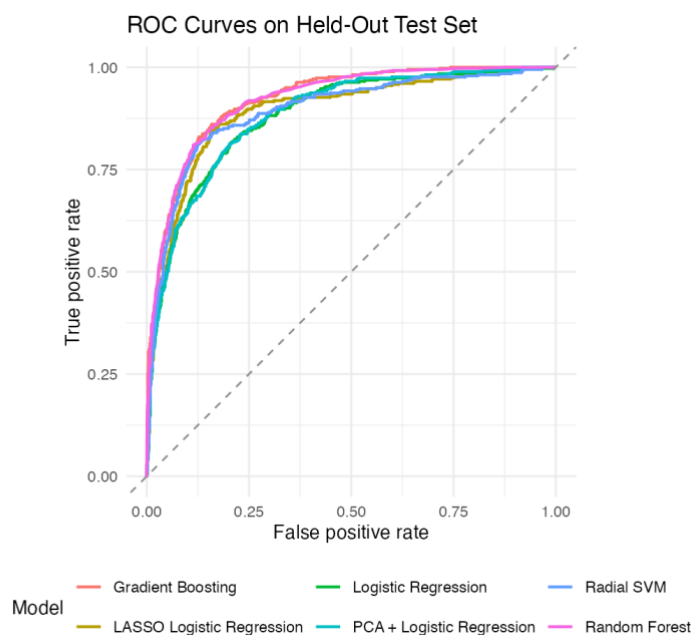


Figure 4. ROC curves on the held-out test set.

## From Prediction to Purchase Intent

To make the prediction results more concrete, we used the Random Forest probabilities on the held-out test set to split sessions into three equal-sized predicted-intent groups. If the model is useful beyond reporting metrics, the high-intent group should contain a much larger share of real purchases than the low-intent group.

Predicted intent group	Sessions	Purchases	Observed purchase rate	Mean predicted probability
Low predicted intent	822	3	0.4%	0.7%
Medium predicted intent	822	34	4.1%	5.0%
High predicted intent	821	344	41.9%	40.7%

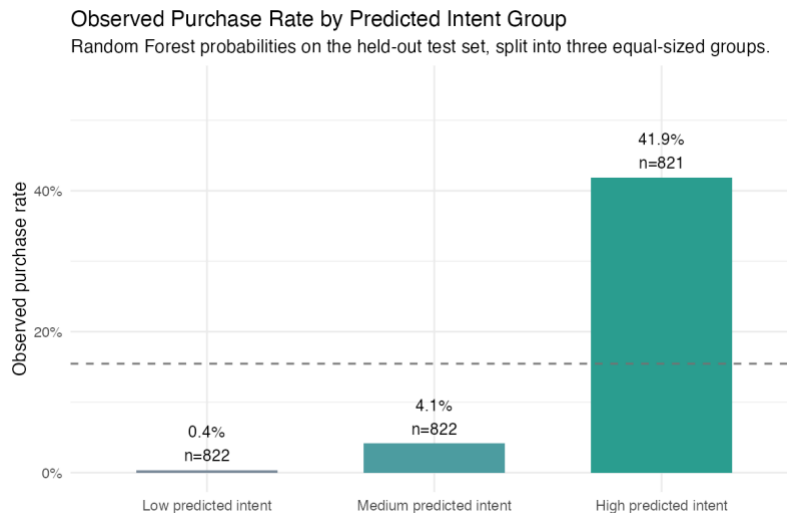


Figure 5. Observed purchase rate by predicted intent group.

## 6. Interpretation and Sensitivity Analysis

### What the Best Models Learned

Page value and product engagement dominate the ensemble models.

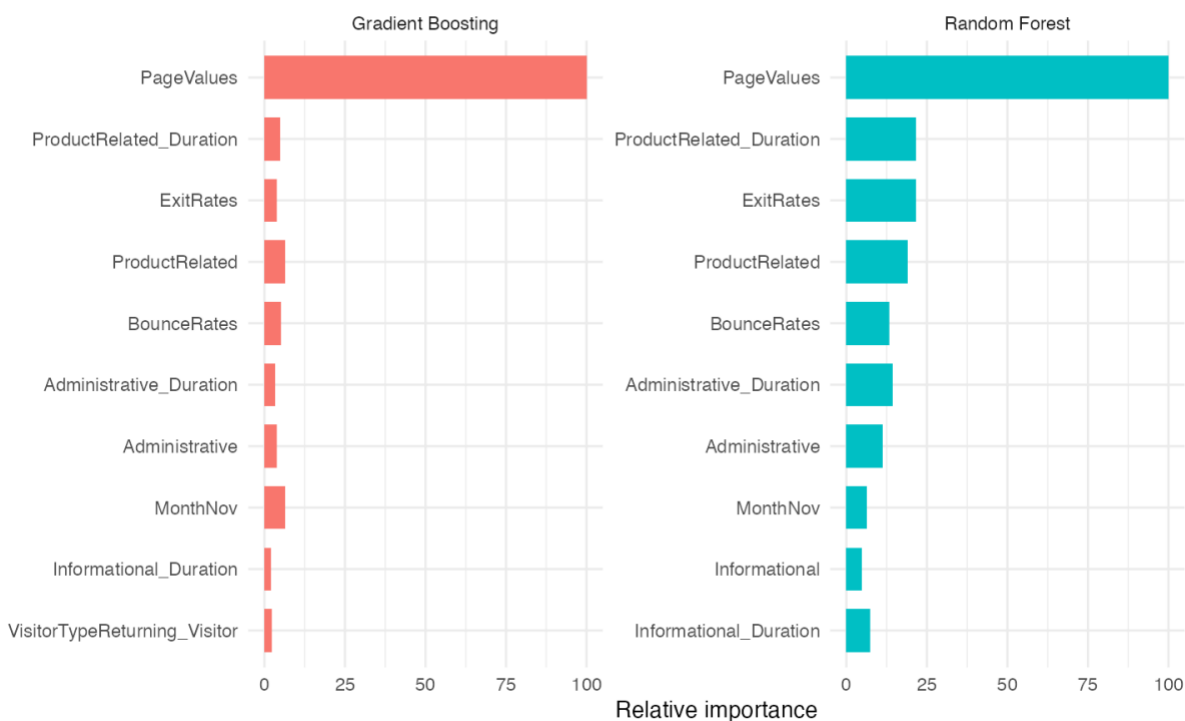


Figure 6. Top variable importance from Random Forest and Gradient Boosting.

Both Random Forest and Gradient Boosting rank PageValues as the dominant predictor. The next tier consists of product-page engagement and abandonment behavior: ProductRelated, ProductRelated\_Duration, ExitRates, BounceRates, and Administrative activity. MonthNov and VisitorType also appear among the more important categorical signals. This reinforces the exploratory

story: conversion is linked to valuable pages, deeper product interaction, lower abandonment, and seasonal shopping behavior.

### PageValues and High-Intent Session Profile

To connect variable importance with user behavior, we profiled sessions by PageValues. The first table shows outcome differences across PageValues groups; the second table adds a more detailed behavioral profile. Together, they suggest that PageValues separates purchase intent into two different forms rather than simply measuring more browsing.

PageValues group	n	Purchase rate	Product pages	Exit rate	Bounce rate	Behavior profile
PageValues = 0	9,600	3.9%	24.6	0.050	0.027	Mostly ordinary browsing
Positive but not top 10%	1,497	43.6%	74.0	0.022	0.007	Deep product browsing
Top 10% PageValues	1,233	71.9%	35.7	0.017	0.004	24.7% new visitors; concentrated in Nov, May, Dec

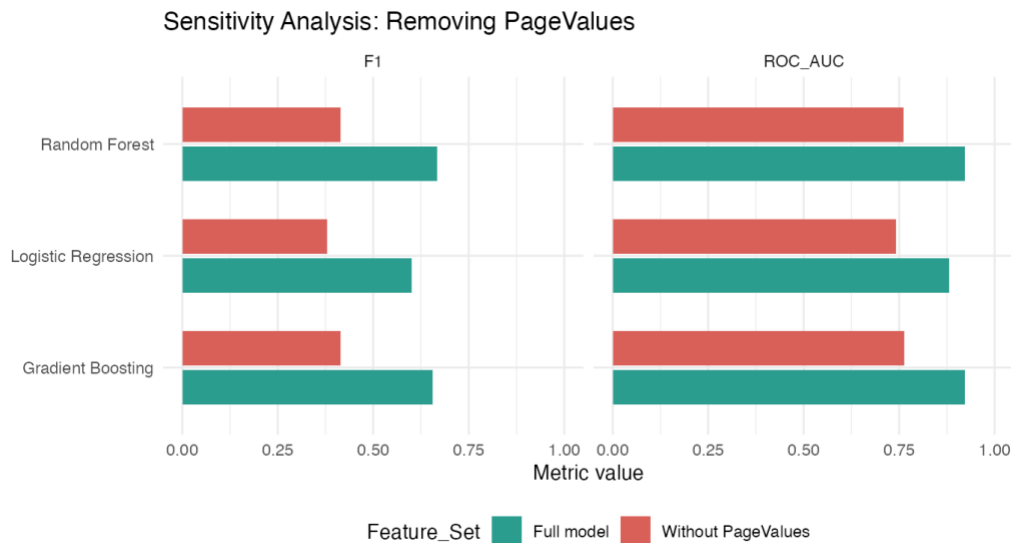
### Detailed behavior profile by PageValues group

PageValues group	Mean PageValues	Product duration	Admin / info pages	New visitors	Weekend	Top months
PageValues = 0	0.0	912 sec	1.7 / 0.35	13.7%	22.8%	May 27.3%, Nov 23.1%, Mar 16.8%
Positive but not top 10%	8.5	2,835 sec	5.5 / 1.32	5.1%	25.9%	Nov 29.1%, May 28.1%, Mar/Dec 10.6%
Top 10% PageValues	48.6	1,408 sec	3.3 / 0.68	24.7%	23.9%	Nov 28.1%, May 26.2%, Dec 14.7%

### The two tables lead to four findings.

- 1. Deep research users.** The positive-but-not-top-10% group has the most product pages (74.0), the longest product duration (2,835 seconds), and a purchase rate of 43.6%. These sessions look like comparison and research behavior: users show clear purchase interest, but they have not reached the highest-conversion state.
- 2. High-conversion intent users.** The top 10% PageValues group views fewer product pages (35.7) and spends less product time (1,408 seconds), but its purchase rate reaches 71.9%, with the lowest ExitRates (0.017) and BounceRates (0.004). This suggests that the strongest purchase intent comes from reaching high-value pages with little abandonment, not simply from browsing more pages.
- 3. New visitors are not low-value users.** The top 10% PageValues group contains 24.7% new visitors, compared with 13.7% in the PageValues = 0 group. High-intent sessions are therefore not only returning-customer sessions; some first-time visitors appear to enter directly into high-value paths.
- 4. Seasonality is linked to high-intent paths.** Top PageValues sessions are concentrated in November, May, and December. This connects the Month pattern to the modeling story: seasonal periods appear to generate more high-value sessions, not only a higher overall purchase rate.

Model	Feature set	ROC AUC	F1	Recall
Gradient Boosting	Full model	0.922	0.655	0.759
Gradient Boosting	Without PageValues	0.764	0.415	0.612
Logistic Regression	Full model	0.881	0.600	0.627
Logistic Regression	Without PageValues	0.742	0.379	0.567
Random Forest	Full model	0.921	0.667	0.769
Random Forest	Without PageValues	0.762	0.414	0.575



*Figure 7. Sensitivity analysis after removing PageValues.*

Removing PageValues sharply reduces performance. For Random Forest, ROC AUC drops from 0.921 to 0.762 and F1 drops from 0.667 to 0.414. Gradient Boosting shows a similar decline. This does not mean the model is invalid; rather, PageValues appears to be a legitimate high-level behavioral signal. However, it should be interpreted carefully because it may summarize prior page-level conversion value. The sensitivity analysis is important because it separates prediction performance from causal explanation.

## 7. Conclusion and Limitations

The analysis answers the four research questions directly. First, purchase sessions differ from non-purchase sessions through higher PageValues, deeper product-page engagement, and lower ExitRates and BounceRates. Second, machine learning models can predict purchase intention on unseen sessions: Random Forest and Gradient Boosting reach ROC AUC near 0.92 on the held-out test set, and the predicted high-intent group converts at 41.9%. Third, nonlinear tree ensembles outperform the interpretable linear baselines, although Logistic Regression and LASSO remain useful for understanding directionally important signals. Fourth, the results depend heavily on PageValues; removing it sharply lowers ROC AUC and F1, so the strongest conclusion is predictive rather than causal.

The main limitation is that the dataset is session-level and does not include customer history, prices, product categories, cart activity, promotion exposure, or intervention costs. PageValues is also analytics-derived and may summarize historical conversion value, so it should be used as a high-level intent signal rather than treated as a direct cause of purchasing. In practice, a retailer should choose thresholds according to business costs: inexpensive reminders may favor recall, while costly discounts may require higher precision.

## References

Sakar, C. O., & Kastro, Y. (2018). Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository. <https://doi.org/10.24432/C5F88Q>

Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer. <https://www.statlearning.com/>

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

### **Appendix: Reproducibility and AI Disclosure**

The full analysis script is `online_shoppers_analysis.R`. The script sets seed 474, creates a stratified 80/20 train-test split, performs 5-fold cross-validation on the training set, selects final classification thresholds from training-set cross-validated predictions, and evaluates all final models on the held-out test set. It also produces the EDA summaries, PCA diagnostic table, model benchmark table, confusion-count table, variable-importance results, predicted-intent group analysis, PageValues profile, and PageValues sensitivity analysis used in this report.

R packages used: tidyverse, caret, pROC, glmnet, randomForest, gbm, kernlab, forcats, and scales

<b>Model</b>	<b>Selected tuning setting</b>
Logistic Regression	default glm after preprocessing
LASSO Logistic Regression	alpha = 1, lambda = 0.0403
PCA + Logistic Regression	default glm after preprocessing
Random Forest	mtry = 7
Gradient Boosting	trees = 300, depth = 3, shrinkage = 0.05
Radial SVM	sigma = 0.005, C = 0.5

#### **AI/LLM disclosure:**

AI tools were used as limited support during this project. We used AI to help draft and polish some report wording, improve grammar and clarity, and assist with R code writing and debugging. All data analysis, model fitting, numerical results, tables, figures, and final interpretations were checked and finalized by the group.

This project's research question, modeling choices, statistical results, and conclusions are based on our course materials, dataset documentation, R code, and group review

I am working on a machine learning final project using the UCI Online Shoppers Purchasing Intention Dataset. The response variable is Revenue. How can I turn this project from a simple model comparison into a coherent research story?

I'll frame this as a project-design answer rather than just "try models." The goal is to make Revenue prediction answer a business/statistical question, with hypotheses, diagnostics, and interpretable findings.

已思考 7s >

You need a **central question**, not just "which model has the highest accuracy?"

A strong research story for this dataset could be:

**Can early-session browsing behavior and contextual factors predict whether an online shopper will purchase, and what does that reveal about conversion behavior?**

That turns the project into a study of **online purchase intention**, not just a leaderboard of algorithms.

### 1. Start with the business/statistical problem

Bad version:

We compare logistic regression, random forest, SVM, and XGBoost to predict Revenue.

In my Model Evaluation section, I have a performance table, an F1/ROC AUC bar chart, and ROC curves. How can I arrange these so the section is clear and not visually crowded?

Arrange it as a **small narrative sequence**, not as three visuals dumped together.

## Recommended order

### 1. Start with a short evaluation setup paragraph

Before any figure/table, write 3–5 sentences:

Because the target variable `Revenue` is imbalanced, accuracy alone is not sufficient. We evaluate models using F1 score, ROC AUC, balanced accuracy, precision, and recall. F1 emphasizes the tradeoff between precision and recall for the purchase class, while ROC AUC measures the model's overall ranking ability across thresholds.

This tells the reader **how to read the section**.