PREDICTING CONSUMER BEHAVIOR: INSIGHTS FROM E-COMMERCE DATA

Marco Capoccia, Viraj Pande, Dheeraj Shetty, Victoria Smith

Business Case

- Online retailers face a major challenge: most website visitors do not convert (2.63% conversion rate)
- Customers Acquisition costs have surged 220% (from $9 to $29)
- Companies collect behavioral data but lack insight into which behaviors drive purchase decisons

By using predictive models, we aim to help e-commerce businesses:

- **Boost conversion rates**
- **Improve ROI**
- **Make smarter data-driven decisions**

Business Question

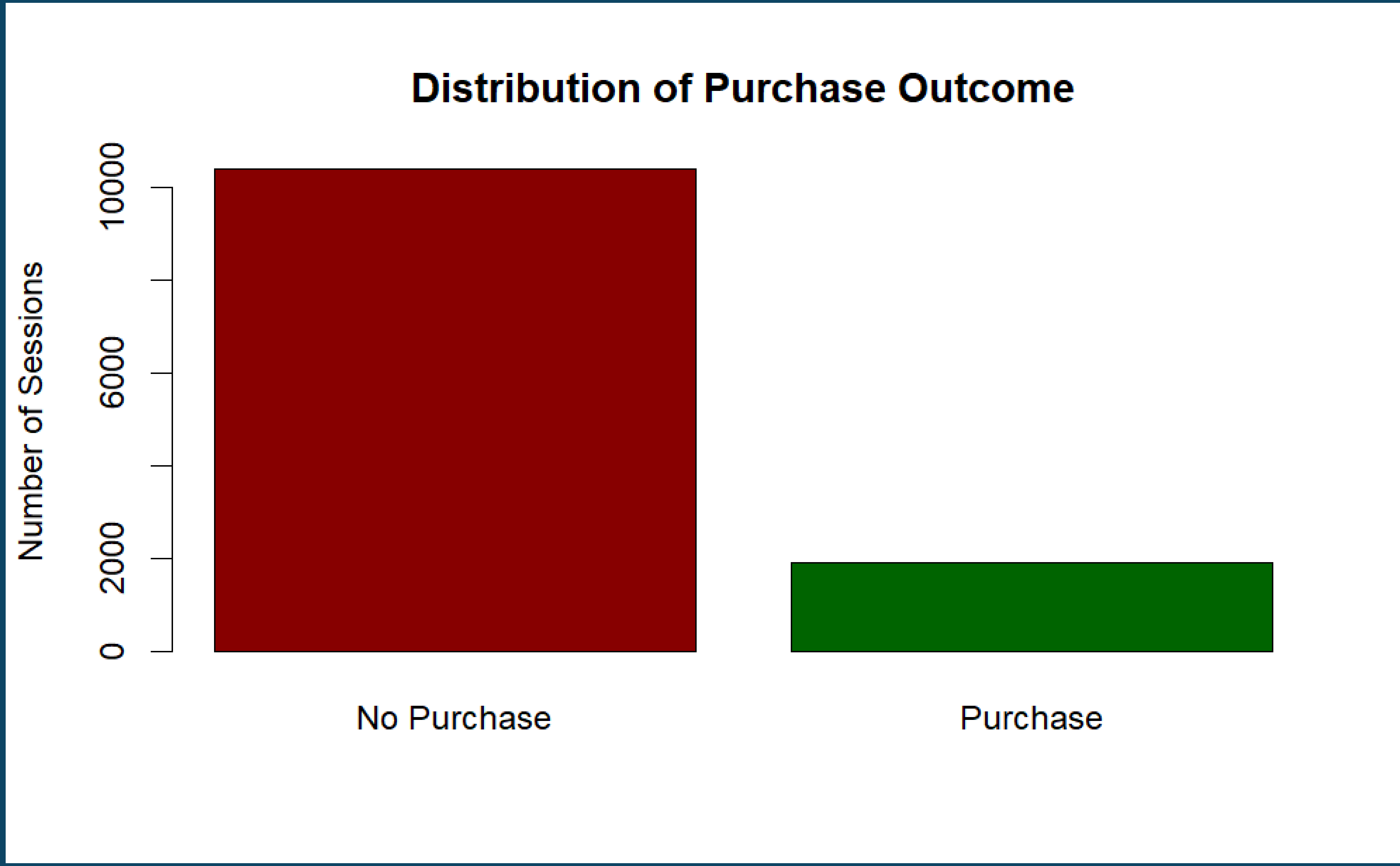
How can an online retailer increase sales by identifying key behavioral factors that drive shoppers to complete a purchase?

Analytics Question

What is the impact of session behavior and visitor traits on the likelihood of a purchase? Additionally, how does the month of visit influence purchase probability?

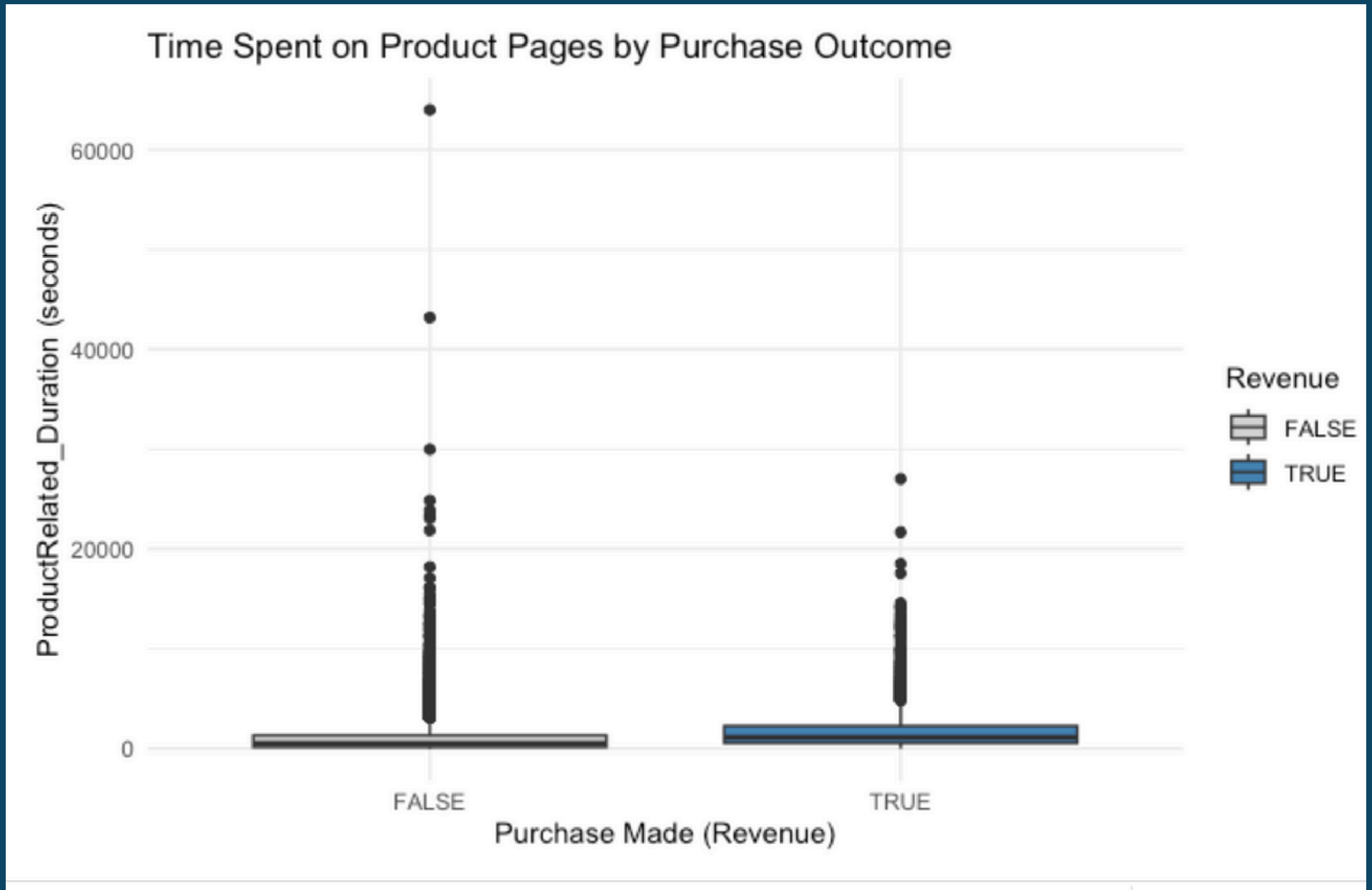
Goal: To identify factors that predict online purchases and provide actionable insights for e-commerce optimization. Using classification models like Logistic Regression and Random Forests, we aim to balance predictive accuracy with interpretability for stakeholders.

DATASET OVERVIEW

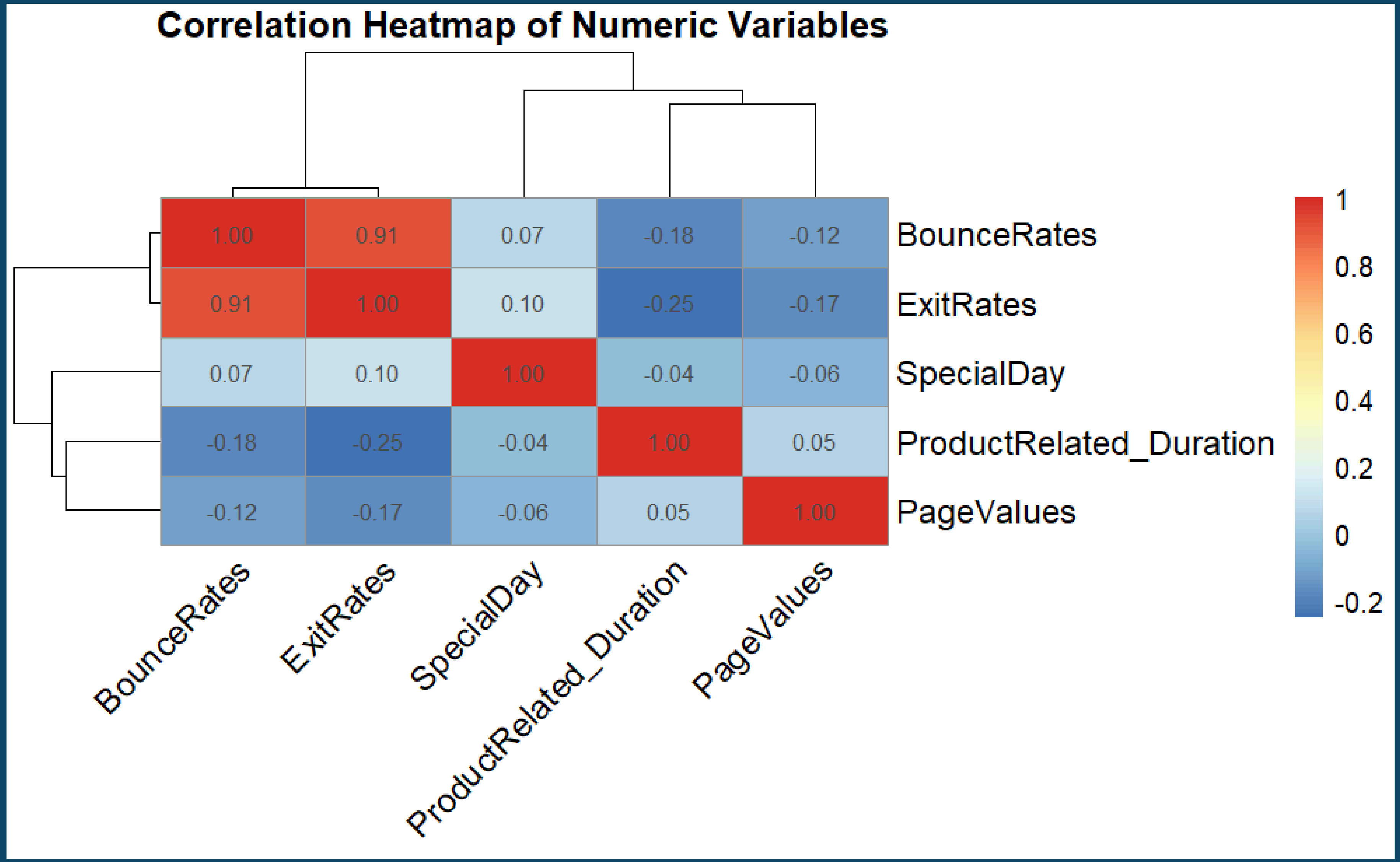


- **12,330 user sessions** from a real e-commerce platform (1 year, unique users)
- **Binary target variable:** Revenue (TRUE = purchase, FALSE = no purchase)
- **Originally 17 predictors** (10 numeric, 7 categorical); we selected 7 key predictors based on business relevance and interpretability
- **Class imbalance:** Only 15.5% of sessions ended in purchases

DESCRIPTIVE STATISTICS



- Buyers spent more time on product pages (mean = 1,876 sec) vs. non-buyers (mean = 1,070 sec).



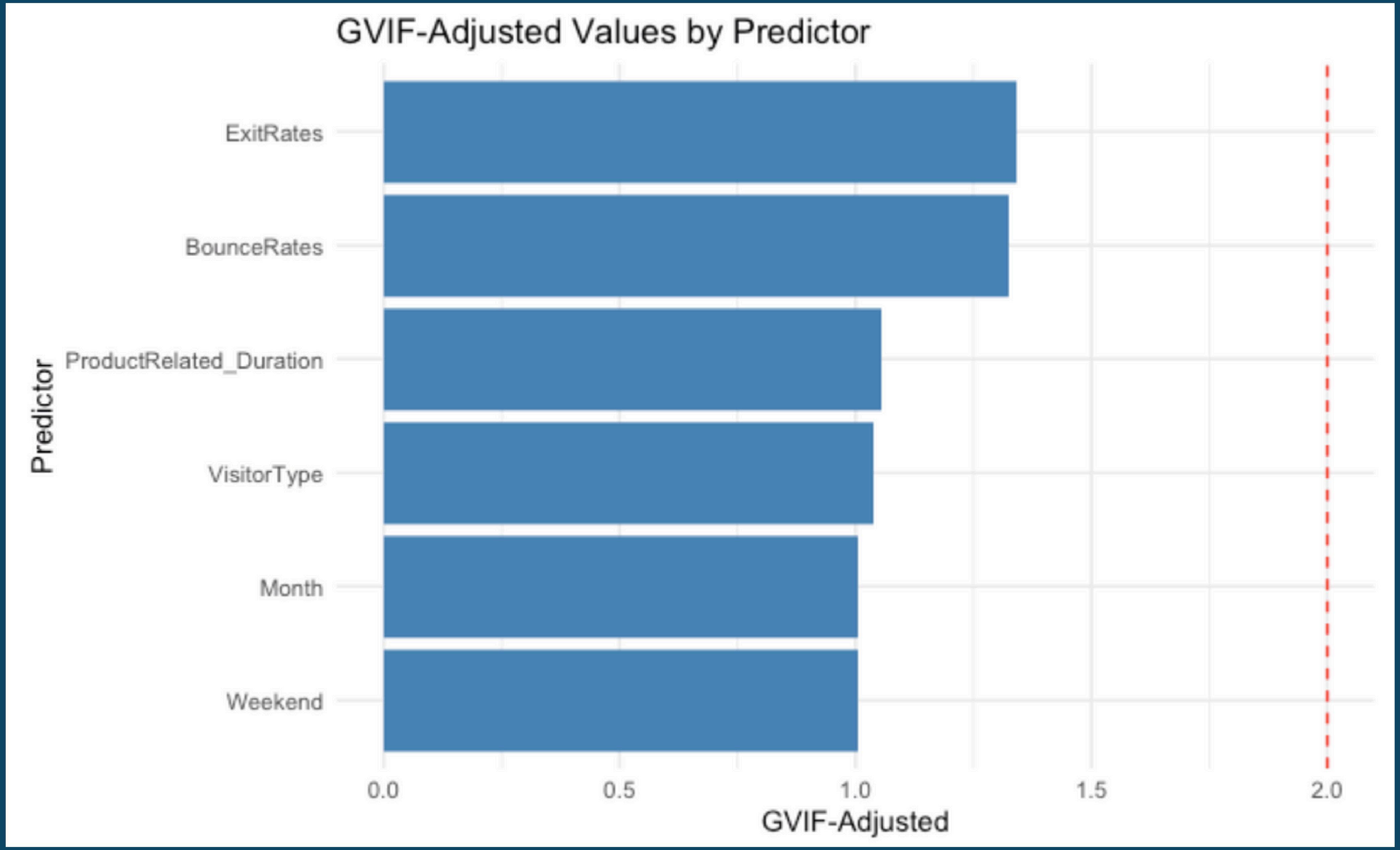
- BounceRates and ExitRates were highly correlated (r = 0.91)

ASSUMPTION TEST: MULTICOLLINEARITY

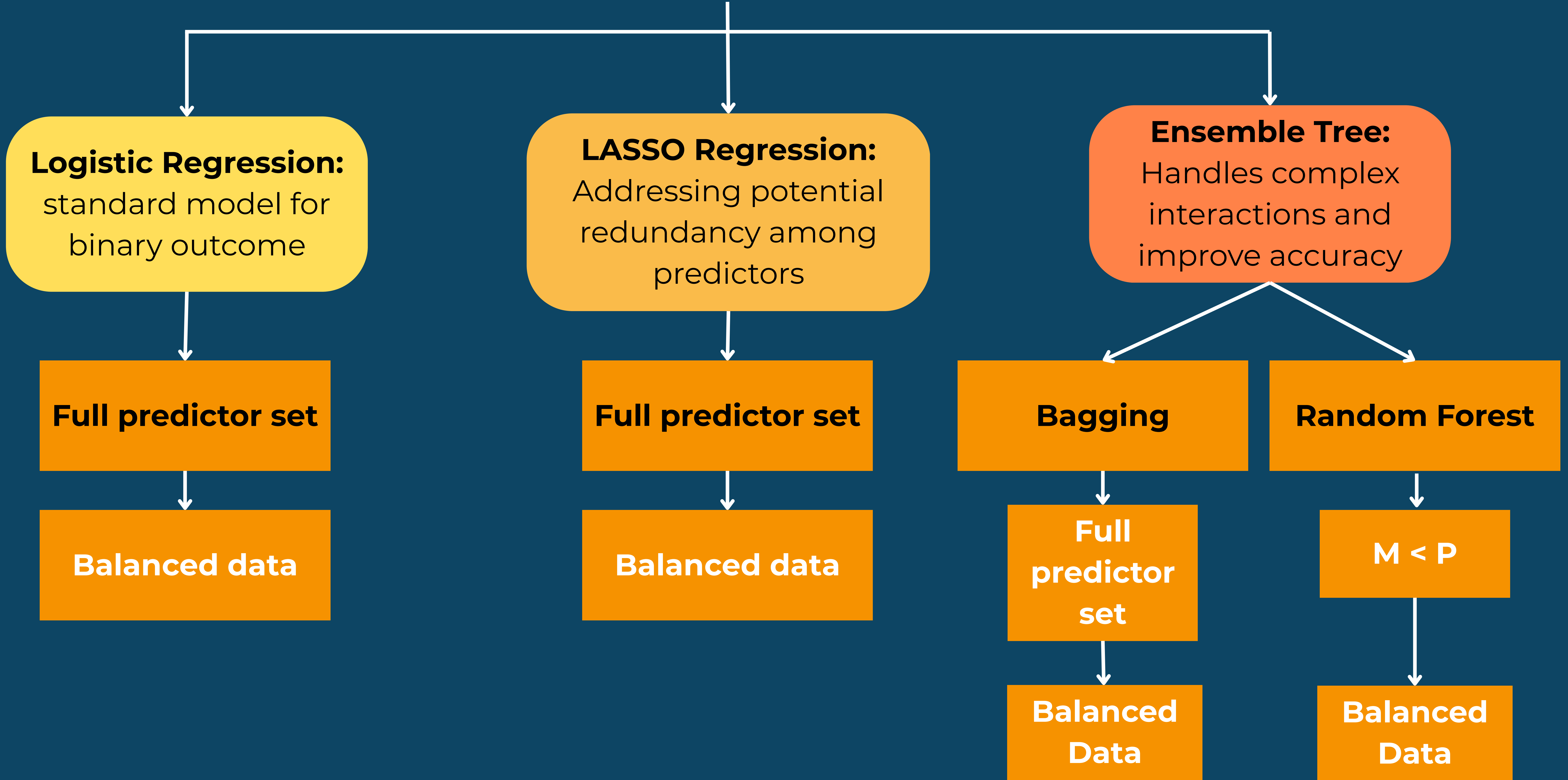
Variance Inflation Factors for Logistic Regression

- **Multicollinearity:** VIF values for all predictors were below 2, indicating no multicollinearity concerns
  - BounceRates & ExitRates were highly correlated (r = 0.91), but VIF = 1.75 & 1.79 confirms both can remain in the model.

- All other predictors had GVIF values near 1



MODELING METHODS AND SPECIFICATIONS



ANALYSIS OF RESULTS

- **Final model:** Logisitc Regression with unbalanced data (chosen for interpretability and predictive accuracy)
- **10FCV error rate:** 11.82%- strong performance

Top predictors:

- **ProductRelated_Duration:** ↑ Time = ↑ Purchase Odds (+0.0085% per second)
- **ExitRates:** Strong negative predictor (-100%+ odds with high values)
- **Returning_Visitor:** 30% ↓ purchase odds vs. new visitors

Seasonal Effects:

- **November** → +60% odds (holiday season)
- **February, March, May** → lower purchase odds

```
Call:
glm(formula = Revenue ~ ProductRelated_Duration + BounceRates +
    ExitRates + VisitorType + Weekend + Month, family = binomial,
    data = Shopping)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5551  -0.6364  -0.4621  -0.1462   3.7417

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -5.906e-01  1.442e-01  -4.095 4.23e-05 ***
ProductRelated_Duration      8.474e-05  1.227e-05   6.906 4.98e-12 ***
BounceRates                  8.528e-01  3.281e+00   0.260 0.794943
ExitRates                   -3.172e+01  2.173e+00 -14.597  < 2e-16 ***
VisitorTypeOther             2.953e-01  3.111e-01   0.949 0.342439
VisitorTypeReturning_Visitor -3.525e-01  6.929e-02  -5.088 3.62e-07 ***
WeekendTRUE                  4.624e-02  6.033e-02   0.766 0.443434
MonthDec                    -3.746e-01  1.520e-01  -2.464 0.013729 *
MonthFeb                    -2.034e+00  6.012e-01  -3.383 0.000718 ***
MonthJul                     6.023e-03  1.915e-01   0.031 0.974911
MonthJune                   -3.482e-01  2.427e-01  -1.435 0.151416
MonthMar                    -5.523e-01  1.532e-01  -3.606 0.000310 ***
MonthMay                    -3.585e-01  1.435e-01  -2.499 0.012459 *
MonthNov                     4.693e-01  1.390e-01   3.377 0.000732 ***
MonthOct                     1.170e-01  1.704e-01   0.686 0.492472
MonthSep                     1.667e-02  1.806e-01   0.092 0.926457
---
```

CONCLUSION

- Product engagement and exit behavior are the strongest drivers of purchases
    - Longer ProductRelated_Duration → Higher purchase likelihood
    - Higher ExitRates → Lower purchase likelihood
- Model performance varies by business goal
    - Logistic Regression offers strong interpretability (11.8% error on unbalanced data)
    - Bagging delivers highest accuracy (6.6% error on balanced data)

RECOMMENDATION

- Focus on maximizing product engagement time (User experience improvements, targeted recommendations)
- Monitor and reduce exit rates through session tracking and real-time interventions
- Re-engage returning visitors with incentives or reminders to improve conversions
- Prioritize seasonal campaigns, especially in November

CHALLENGES FACED

Challenge:

- Significant class **imbalance**: only ~15% of sessions resulted in purchases
- High **correlation** between BounceRates and ExitRates: potential reddundancy in predictors
- Same **predictors** yielded different results across models: made interpretation and model selection complex
- Ensuring **model validation** was accurate and not overfit
- Balancing **interpretability** **vs.** **accuracy** for stakeholder recommendations

Solution:

- Random **oversampling** for a balanced dataset
- Evaluated **multicollinearity** (VIF<2)
- Used **LASSO** to isolate the most informative predictors
- Compared **internal errors** from tree models with **10FCV** to confirm reliable results
- Selected **logistic regression** for intepretability, while noting **bagging** as the best for predictive accuracy alone.

THANK YOU

Questions?