# Decoding Online Purchase Patterns Using Classification

Heidi Rogers

November 2024

## 1 Introduction

Understanding consumer behavior is a critical component for the success of a business and an important issue for any marketing team. The objective of this study is to uncover the factors that drive an internet user to complete purchases online. For this analysis, 12,330 online sessions were observed, each for a different user. Each session has 18 recorded features including types of web pages visited, duration of visits, and time of year, among others. This paper describes the process of building and testing a classification model used to predict whether users ultimately made an online purchase or not. Both logistic regression and random forest models were considered and can be used to predict future consumer behavior and identify effective marketing tactics by uncovering deterministic features of whether a user will generate revenue.

## 2 Data Exploration

### 2.1 Data Cleaning

All data was processed and analyzed using R 4.2.1 (R Core Team 2022). No missing values were present in the original data set and no transformations were necessary. In examining three particular variables, *operating systems*, *browser*, and *traffic type*, many of their categories had few observations. To reduce the complexity of these features, all groups with fewer than 5% of the total number of observations were grouped into a separate ”other” category. Further investigation of variable distribution showed no difference in the proportion of users that generated revenue across browser type or geographical region, thus these variables were removed completely. The original *new visitor* variable contained three categories: new visitor, returning visitor, and other. Since a user can only be new or returning, the small proportion of ”other” users were grouped with the new users as their online behavior was most similar.

Incorporating these categorical variables into the statistical model was accomplished through one-hot encoding. This technique converts the categories into binary variables, allowing them to be used with models that only take numerical inputs. This process increases the dimension of the data set to be 12,234 observations with 39 total variables. Feature elimination is later implemented to reduce the total number of predictors.

### 2.2 Feature Engineering

To enhance analysis and interpret-ability of the data for more effective model inputs, new features were produced to better capture the online behavior in relation to online purchases. One feature, *bounce rate*, refers to the proportion of visitors who enter and leave a website without triggering any requests. Similarly, *exit rate* is calculated as the proportion of all views to that webpage that were the last visit in the user session. With the high correlation of these two variables, the features were combined by dividing exit rate by the bounce rate.

Exploratory data analysis reveals potential varaible relationships with the response. We find that both longer time spent online and more pages visited during a session more often results in a purchase than a non-purchase, as seen in figure 1. In an attempt to account for this, the proportion of each type of page visited out of the total web pages visited in that session was calculated to help quantify the user's attention to each type of content- informational, administrative, or product-related. Since knowing two proportions implies the third, the proportion of product-related page views was removed for simplicity. Additionally, average minutes spent per page for each type of webpage was calculated to capture the level of engagement

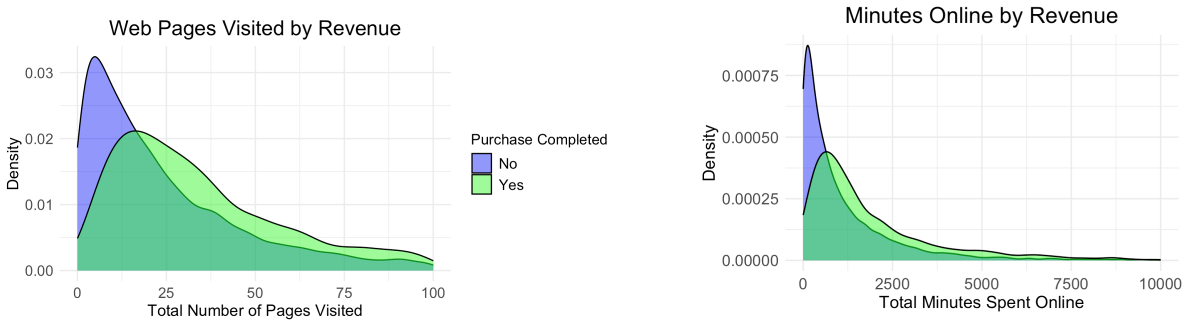

Figure 1: Density plots of total pages visited and total time spent in a single session colored by revenue generation

with different page categories. Following this, redundant and highly correlated variables were removed. The final set of features, along with brief descriptions, are presented in table 1.

| **Variable** | **Description** |
|---|---|
| Revenue* | Whether or not a purchase was completed |
| Pages Value | Avg. value of a webpage visited before completing a purchase, measured by Google Analytics |
| Special Day | Measure of closeness of site visit time to a holiday, accounting for shipping time |
| Month | Month of the year the site was accessed |
| Region | Geographical location of the online user |
| Traffic Type | The method through which the website was accessed |
| Operating System | The type of operating system used by the consumer |
| Weekend | Whether or not the visit was during the weekend |
| Exit-Bounce Rate | For a specific webpage, proportion of page views that were last in the session relative to the proportion of users that entered and left the page without triggering any requests |
| Min Per Page | Avg. minutes spent on each page visited (Product, Administrative, and Informational) |
| Page Proportion | Proportion of visits to each page type out of total visits in that session (Product, Administrative, and Informational). |
| New Visitor | Whether the user was new or returning to the webpage. |

Table 1: Features considered in the modeling process and their descriptions. *Revenue is the response variable.

# 3 Methodology

## 3.1 Variable Selection

Variable elimination was implemented using LASSO regression, helping to reduce the dimensionality of the data set. K-fold cross validation with 5 folds was employed, and only variables with non-zero coefficients were maintained. A total of 19 variables were removed, including several browser types, months, and regions, as well as the proportion of pages visited that were informational, and average minutes spent per product-related page.

## 3.2 Class Imbalance

Out of the 12,234 online sessions recorded, only 1,908 of them included an online purchase, creating a class imbalance where approximately 16% of users generated revenue, while 84% did not. The skewed distribution of our response variable presents the issue of over-representation of the majority class and can lead to a biased model. To address this imbalance, various methods such resampling techniques, ensemble modeling, applying class weights, or utilizing different model evaluation metrics are commonly employed.

In this study, the Synthetic Minority Oversampling Technique (SMOTE) was utilized. SMOTE generates synthetic instances of the minority class using existing points and their k-nearest neighbors. Previous studies have demonstrated the efficacy of SMOTE in reducing class imbalance, particularly for large sample data sets with fewer dimensions (Elreedy & Atiya, 2019). Other techniques, such as Adaptive Synthetic Sampling (ADASYN), were considered, though it has not been found to outperform SMOTE when applied jointly with logistic regression (Masruriyah et al., 2023). For each minority class observation, 3 synthetic samples were generated using their 5-nearest neighbors (k=5) giving a more evenly distributed data set of 7,632 revenue-generating sessions and 10,416 session with no purchases.

## 3.3 Model Development

Initially, a multivariate logistic regression model was considered. The classification threshold was optimized using k-fold cross validation with 5 folds. Figure 2 illustrates the impact of threshold on the accuracy, sensitivity, and specificity of the logistic model. A threshold of 0.39 gives the highest accuracy for the model, such that predicted probabilities greater than 39% result in the online user session being classified as a revenue-generating.

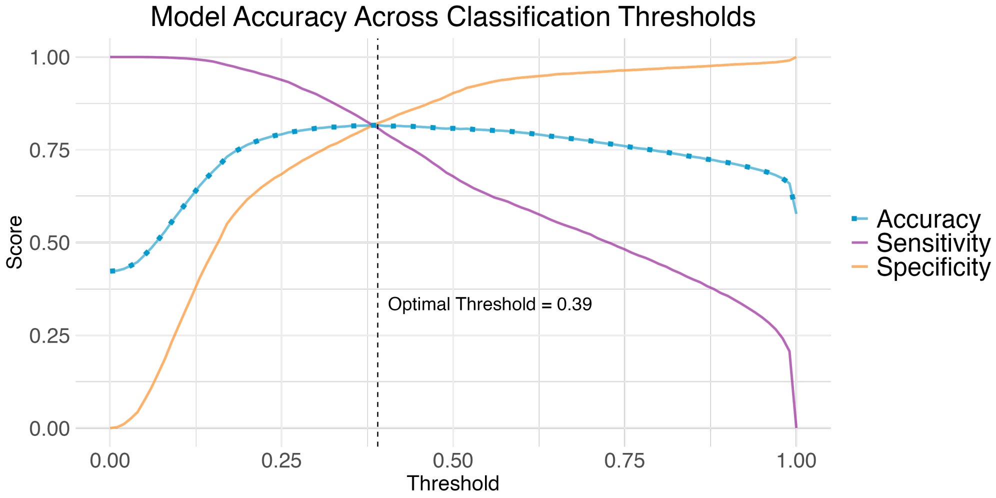

Figure 2: Logistic regression model performance metrics plotted across classification thresholds

A random forest model was also considered to account for potential non-linear relationships with the response. The model was built using 500 trees with 5 variables randomly selected for each splitting node, or the square root of the number of total predictors. The same technique used for logistic regression was used to find the optimal classification threshold for model predictions. By maximizing accuracy, we find that at a threshold of 0.43 is ideal.

Both models were trained and tested 5 times via k-fold cross validation. Performance metrics were calculated for each fold, and averaged across all folds, including accuracy, sensitivity, specificity, and area under the Receiver Operating Characteristic (ROC) curve.

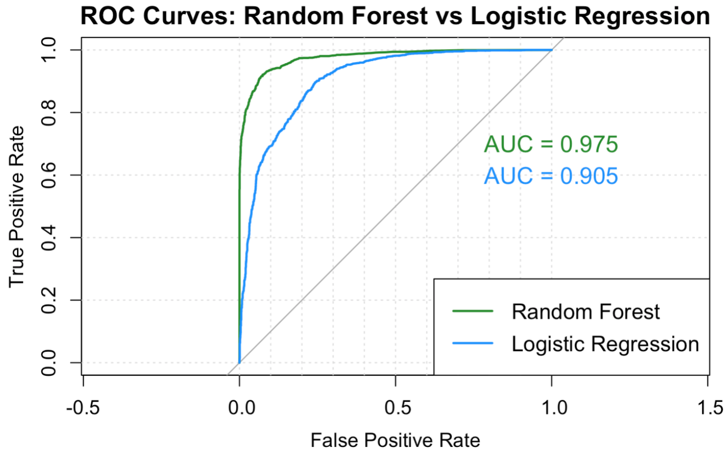

Figure 3: Comparison of logistic regression and random forest model ROC curves for fold 1 of cross validation

# 4 Results

The random forest model outperforms logistic regression, producing a higher accuracy, sensitivity, specificity, and area under the curve (AUC) as displayed in table 2, with slightly different classification thresholds. Figure 3 illustrates the comparison of their respective ROC curves for a single fold of cross validation. While the random forest yields greater accuracy, both models are comparable and logistic regression offers a more straightforward interpretation in terms of odds.

| **Model** | **Threshold** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
|---|---|---|---|---|---|
| Logistic Regression | 0.39 | 0.813 | 0.805 | 0.819 | 0.905 |
| Random Forest | 0.43 | 0.919 | 0.905 | 0.930 | 0.975 |

Table 2: Comparison of average performance metrics across 5-fold cross validation for logistic regression and random forest classification models

Several variables were significant in the final logistic regression model. To understand their impact on revenue, we look at the odds ratios of each one, displayed in table 3. Generally, an odds ratio greater than 1 indicates that revenue generation increases as the predictor value increases, and vise versa for a ratio less than 1.

| **Variable** | **Odds Ratio** |
|---|---|
| Page Value | 1.13 |
| Exit-Bounce Rate | 0.54 |
| Min Per Admin Page | 1.00 |
| Min Per Info Page | 1.00 |
| Admin Page Prop | 2.77 |
| Month: Dec | 0.47 |
| Month: Feb | 0.09 |
| Month: Mar | 0.42 |
| Month: May | 0.38 |
| Month: Nov | 2.15 |
| Operating System 2 | 1.19 |
| Region 5 | 0.51 |
| Region 9 | 0.67 |
| Traffic Type 13 | 0.44 |
| Traffic Type 2 | 1.41 |
| Traffic Type 3 | 0.74 |
| Traffic Type Other | 1.35 |

Table 3: Odds Ratios for Significant Predictors of Revenue

Proportion of pages visited that are administrative and the month of November have the largest odds ratios of 2.77 and 2.15. Given this, we know that the odds of a user generating revenue are 2.77 times higher for every unit increase in proportion spent on an administrative page. Similarly, the odds of revenue are 2.15 times greater if the user is browsing in the month of November versus not. Conversely, many of the odds ratios are less than 1. For instance, the odds of revenue are reduced by a factor of 0.09 if the month if February, or reduced by 46% for every unit increase in the exit-bounce rate.

Overall, the odds ratios give insight into which factors are influential in the outcome of an online shoppers internet session, yet, the random forest model had a significantly high accuracy, and thus is reliable for making future predictions about consumer behavior.

# References

[1] Elreedy, D., & Atiya, A. F. (2019). A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance. Information Sciences, 505, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

[2] Masruriyah, A. F., Novita, H. Y., Sukmawati, C. E., Fauzi, A., Wahiddin, D., & Handayani, H. H. (2023). Thorough evaluation of the effectiveness of smote and ADASYN oversampling methods in enhancing supervised learning performance for Imbalanced Heart Disease Datasets. 2023 Eighth International Conference on Informatics and Computing (ICIC), 1–7. https://doi.org/10.1109/icic60109.2023.10382105

[3] R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.