Bayesian Prediction of Online Shoppers' Purchasing Intention

Liang-Jen Huang [1]

Abstract

This project uses Bayesian logistic regression to predict whether an online shopping session results in a purchase, based on the Online Shoppers Purchasing Intention Dataset. By modeling both behavioral and contextual variables, the Bayesian framework provides calibrated probabilities and quantifies uncertainty in predictions. The approach offers insights into variation across user groups and visit patterns, supporting a better understanding of online shopping behavior and informing more effective marketing strategies in e-commerce.

1. Introduction

E-commerce companies are constantly trying to understand how customers behave so they can improve marketing, use resources more effectively, and increase sales. A key question is whether a visit to an online store will actually lead to a purchase. Being able to predict this helps businesses focus on the users who are most likely to buy, tailor recommendations, and run more effective campaigns. It also improves the shopping experience while avoiding unnecessary marketing costs.

Many machine learning methods have been used for this type of prediction, but they usually concentrate on accuracy and provide only single-point outcomes. These models often struggle when the data are imbalanced and do not show how uncertain their predictions are. Bayesian modeling offers an alternative: it provides a full probability distribution rather than just a yes/no answer. This means the predictions are better calibrated, and the uncertainty around them can be clearly expressed, giving a deeper understanding of customer behavior.

In this project, a Bayesian logistic regression model is applied to analyze the Online Shoppers Purchasing Intention Dataset, with the goal of identifying key factors that influence purchase decisions and evaluating the uncertainty around them.

2. Data

The analysis is based on the Online Shoppers Purchasing Intention Dataset (Sakar & Kastro, 2018). The dataset contains 12,330 sessions collected over a one-year period, each representing the complete browsing activity of a unique user. The target variable is *Revenue*, a binary indicator of whether the session ended in a purchase. In total, the dataset provides 17 features, including both behavioral information (such as the types and durations of pages visited, bounce and exit rates, and page value) and contextual information (such as month, weekend indicator, visitor type, operating system, browser, and traffic source).

The data are well structured, contain both numerical and categorical variables, and do not include missing values. Among all sessions, 84.5% did not end with a purchase and 15.5% did, resulting in an imbalanced distribution of the target variable. The variables and their corresponding types are summarized in Table 1.

Before conducting the Bayesian analysis, the dataset was preprocessed as follows: categorical variables — including Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend, and Revenue — were converted into factor types so that the model could correctly interpret them as categorical predictors. All remaining variables were kept as numeric. To eliminate potential scale differences among numeric features, z-score standardization was applied, transforming each numeric variable to have a mean of zero and a standard deviation of one. Finally, the dataset was randomly divided into a training set (80

3. Models and Methods

3.1. Bayesian logistic regression

A Bayesian logistic regression model was applied to predict the target variable *Revenue*. Let y_i denote the purchase outcome for session i, where $y_i = 1$ if the session ended in a purchase and $y_i = 0$ otherwise:

$$y_i \sim \text{Bernoulli}(\pi_i), \quad \text{logit}(\pi_i) = \alpha + \mathbf{x}_i^\top \boldsymbol{\beta},$$

where π_i is the purchase probability, α is the intercept, and $\boldsymbol{\beta}$ are coefficients for predictors $\mathbf{x}_i$. Weakly informative

Table 1. Online Shoppers Purchasing Intention Dataset

Variable	Type
Administrative	Integer
Administrative Duration	Continuous
Informational	Integer
Informational Duration	Continuous
ProductRelated	Integer
ProductRelated Duration	Continuous
BounceRates	Continuous
ExitRates	Continuous
PageValues	Continuous
SpecialDay	Continuous
Month	Categorical
OperatingSystems	Integer
Browser	Integer
Region	Integer
TrafficType	Integer
VisitorType	Categorical
Weekend	Binary
Revenue (Target)	Binary

priors were assigned to regularize the estimation:

$$\alpha \sim \mathcal{N}(0, 5^2), \quad \beta_j \sim \mathcal{N}(0, 2^2).$$

The model was implemented in Stan through the RStan interface and estimated using the No-U-Turn Sampler (NUTS). Four Markov chains were run, each with 2,000 iterations including 1,000 warm-up steps.

4. Results

The posterior mean of the intercept was around –2.0 (95% CI: [–2.47, –1.55]), which corresponds to a baseline purchase probability of roughly 12% when all predictors are at their reference levels. Variables related to product engagement—such as PageValues, ProductRelatedDuration, and ProductRelated—appeared to have positive effects, suggesting that longer browsing time and higher page values might be associated with a greater likelihood of purchase. In contrast, BounceRates and ExitRates tended to have negative effects, indicating that early exits or low engagement could lower the chance of conversion.

According to the convergence diagnostics, the model seemed to have converged reasonably well. As shown in Figure 1, the chains mixed consistently without visible divergence, and the $\hat{R}$ values were close to 1. These results suggest that the sampling process was likely stable.

Predictions on the test set were computed using the posterior mean of the parameters. When evaluated on the test data,

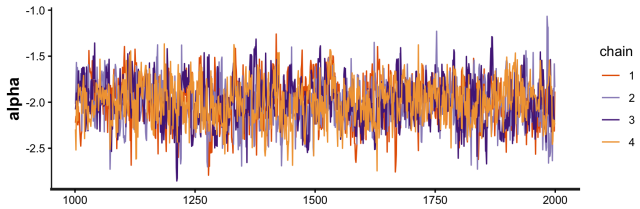

Figure 1. Trace plot for the intercept parameter

the model achieved an accuracy of approximately 87.8%, suggesting that it may generalize well to unseen observations. However, the current dataset was used as-is without addressing potential class imbalance, and both the training and testing samples were randomly split rather than stratified.

5. Conclusions

References

Sakar, C. and Kastro, Y. Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5F88Q.

Appendix A. Stan Code

```
data {
  int<lower=0> N;
  int<lower=0> K;
  matrix[N, K] X;
  int<lower=0,upper=1> y[N];
}
parameters {
  real alpha;
  vector[K] beta;
}
model {
  alpha ~ normal(0, 5);
  beta ~ normal(0, 2);
  y ~ bernoulli_logit(alpha + X * beta);
}
generated quantities {
  int y_pred[N];
  for (i in 1:N)
    y_pred[i] = bernoulli_logit_rng(alpha + X[i] * beta);
}
```

Listing 1. Stan code for Bayesian logistic regression