Causal Modeling of E-Commerce Purchasing Intention

Foad Assareh
foad.assareh@studio.unibo.it

Master's Degree in Artificial Intelligence, University of Bologna
January 18, 2026

Abstract

This project aims to model the factors behind online purchasing intention using a Causal Bayesian Network. We used the "Online Shoppers Purchasing Intention Dataset" from the UCI Machine Learning Repository [1]. We found that using this data, our model fails to detect buyers due to severe class imbalance. By applying SMOTE oversampling, we improved the Recall for buyers from <20% to 81%, while maintaining an overall accuracy of 87%.

Introduction

Domain

The domain involves modeling the transition from "Visitor" to "Buyer" based on session metrics (e.g., page views) and context (e.g., weekend) to understand user behavior in e-commerce. A key challenge is the prevalence of "window shoppers," resulting in datasets where non-buyers vastly outnumber buyers.

Aim

The purpose of this project is to build a causal model that predicts purchasing intention and explains *why* a purchase fails. Specifically, we aimed to identify the distinct factors—ranging from user behavior to technical context—that influence purchasing probability.

Method

We utilized the `pgmpy` library [2] and the Online Shoppers Purchasing Intention dataset [1] to implement a Causal Bayesian Network. The methodology proceeded in three phases. First, we performed data preprocessing and feature engineering, where continuous numerical variables were discretized into categorical bins to meet the requirements of a discrete network. This step also involved deriving `UserIntent` and `TechFriction` variables, to construct the hidden layer of the network structure.

Second, to address the significant 85:15 class imbalance, we applied random oversampling to the training data only, increasing the representation of the minority (Buyer) class. This step was necessary to allow reliable estimation of CPDs involving the Purchase variable, which would otherwise be biased toward the majority class. We note that oversampling alters the empirical class prior and may affect probability calibration; therefore, model evaluation focused on recall, ROC-AUC, and causal behavior rather than raw probability estimates.

Results

The model achieved an accuracy of **87%** with a Buyer Recall of **81%**, solving the imbalance problem. We also successfully modeled "Explaining Away," where high bounce rates discount the probability of purchase even when page views are high.

Model

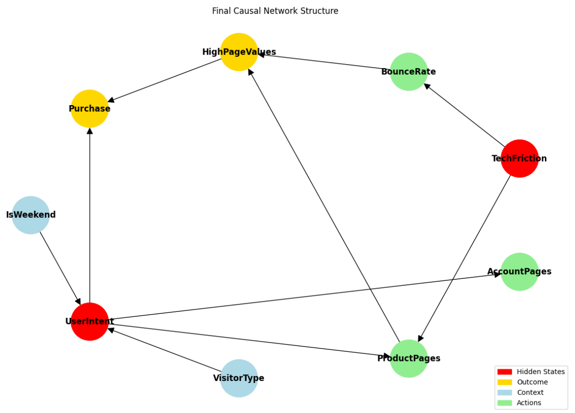

Figure 1: Structure of Bayesian Network

Our model (Figure 1) consists of 9 discrete nodes organized into four causal layers:

- **Context Layer (Root):** Includes `IsWeekend` and `VisitorType`. These define the external environmental context independent of user behavior.
- **State Layer (Hidden):** Includes `UserIntent` and `TechFriction`. These are latent variables representing the user's psychological state and the platform's technical performance, which are not directly observable.
- **Action Layer (Evidence):** Includes `ProductPages`, `AccountPages`, and `BounceRate`. These are the behavioral manifestations caused by the hidden states (e.g., High Friction causes High Bounce Rate).
- **Outcome Layer:** Includes `HighPageValues` and `Purchase`. `HighPageValues` acts as the strongest pre-

dictor (Binary > 0), mediating the link between engagement and the final conversion.

The conditional probability distributions (CPDs) were learned from the balanced dataset using Maximum Likelihood Estimation (MLE), capturing the causal flow from behavioral signals to purchasing intention [2].

Analysis

Experimental Setup

We designed five query scenarios to test specific causal reasoning patterns. We calculated the posterior probability of the target variable `Purchase` given varying sets of evidence E:

1. **Baseline:** No evidence provided ($E = \emptyset$), establishing the prior probability of purchase.
2. **Strongest Predictor:** $E = \{\texttt{HighPageValues} : Yes\}$. Testing the model's sensitivity to the "Ideal User" signal.
3. **Explaining Away:** A comparative test between $E_a = \{\texttt{ProductPages} : High\}$ and $E_b = \{\texttt{ProductPages} : High, \texttt{BounceRate} : High\}$. This tests if the model correctly discounts high engagement when "Friction" is observed.
4. **Context Effect:** $E = \{\texttt{IsWeekend} : True\}$. Testing if temporal context alone shifts user intent.
5. **Ambiguity:** $E = \{\texttt{AccountPages} : Clicked\}$. Testing the impact of a "Soft Signal" (checking account) compared to strong signals.

Results

What did we observe?

- **Scenario 2 (Ideal User):** Observing `HighPageValues=Yes` triggered the largest increase, raising purchase probability to >**85%**, confirming it as the primary causal driver.
- **Scenario 3 (Explaining Away):** While high product views alone increased the probability (Step A), adding `BounceRate=High` (Step B) caused the probability to **drop significantly**. The model successfully "explained away" the views by attributing the bounce to friction or lack of genuine intent.
- **Scenario 4 & 5 (Context/Ambiguity):** The `IsWeekend` evidence resulted in a negligible change, indicating that context plays a minor role compared to behavioral signals like `AccountPages`, which showed a moderate positive lift.

Conclusion

In conclusion, we learned two main lessons. First, we found that **imbalanced data** severely biases the model to ignore rare events (buyers), making balancing techniques essential for accurate predictions. Second, our results showed that **behavioral factors** (like `HighPageValues`) have a much stronger effect on purchasing intention than **contextual factors** (like `IsWeekend`). This proves that observing what a user *does* on the site is far more important than knowing *when* they visit.

Links to external resources

- **Code:** GitHub Repository
- **Dataset:** UCI Online Shoppers Intention `https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset`

References

[1] Sakar, C.O., et al. (2018). *Online Shoppers Purchasing Intention Dataset.* UCI Machine Learning Repository. `https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset`

[2] Ankan, A., Panda, A. (2015). *pgmpy: Probabilistic Graphical Models using Python.* Documentation available at: `https://pgmpy.org/`