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 models the factors behind online purchasing intention using a Causal Bayesian Network. Using the UCI "Online Shoppers Purchasing Intention Dataset" [1], we found that the model fails to detect buyers due to severe class imbalance. Applying Random oversampling improved Buyer Recall from <20% to **81%**, while maintaining overall accuracy of **87%**.

Introduction

Domain

We model the transition from "Visitor" to "Buyer" using session metrics (e.g., page views) and context (e.g., weekend). The dataset contains mostly non-buyers, making predictive modeling challenging.

Aim

The goal is to build a causal model that predicts purchasing intention and explains *why* purchases fail. We aim to identify key factors—from behavioral to technical—that influence buying probability.

Method

We used the `pgmpy` library [2] and the Online Shoppers dataset [1] to implement a Causal Bayesian Network. Methodology included:

- **Preprocessing:** Discretized continuous variables and derived latent variables `UserIntent` and `TechFriction`.
- **Oversampling:** Addressed 85:15 class imbalance using random oversampling on the training set to reliably estimate CPDs for `Purchase`.
- We note that oversampling alters the *empirical class prior* and may affect probability calibration; hence evaluation focused on **Recall**, **ROC-AUC**, and causal reasoning rather than raw probabilities.

Results

Performance: Accuracy = **87%**, Buyer Recall = **81%**. The model captures "Explaining Away": high bounce rates reduce purchase probability even when page views are high.

Model

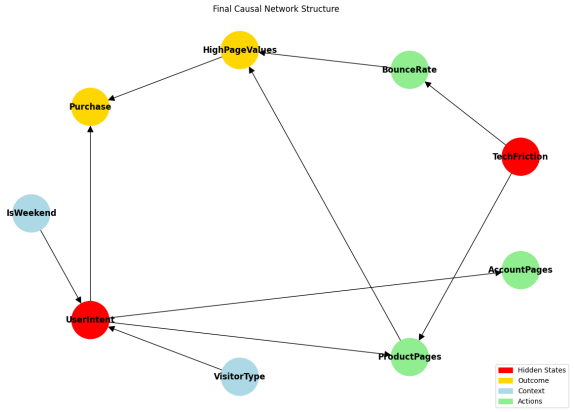

Figure 1: Causal Bayesian Network for predicting purchasing intention.

The network has 9 discrete nodes in four causal layers:

- **Context (Root):** `IsWeekend`, `VisitorType`.
- **State (Hidden):** `UserIntent`, `TechFriction`.
- **Action (Evidence):** `ProductPages`, `AccountPages`, `BounceRate`.
- **Outcome:** `HighPageValues`, `Purchase`.

CPDs were learned from the balanced dataset using Maximum Likelihood Estimation (MLE) [2].

Analysis

Experimental Setup

We tested five query scenarios, computing posterior probabilities for `Purchase` given evidence E:

1. **Baseline:** $E = \emptyset$.
2. **Strongest Predictor:** $E = \{\texttt{HighPageValues} : Yes\}$.
3. **Explaining Away:** Compare $E_a = \{\texttt{ProductPages} : High\}$ vs $E_b = E_a \cup \{\texttt{BounceRate} : High\}$.
4. **Context Effect:** $E = \{\texttt{IsWeekend} : True\}$.
5. **Ambiguity:** $E = \{\texttt{AccountPages} : Clicked\}$.

Results

- **Scenario 2 (Ideal User):** Purchase probability >85%.

- **Scenario 3 (Explaining Away):** Adding high bounce rate reduces probability significantly.
- **Scenario 4 & 5:** Context alone has minor effect; soft signals like account pages provide moderate lift.

Conclusion

- Imbalanced data biases models against rare events; oversampling is essential.
- Behavioral factors (HighPageValues) influence purchase more than contextual factors (IsWeekend).

Links to external resources

- **Code:** GitHub Repository
- **Dataset:** UCI Online Shoppers Intention

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: `https://pgmpy.org/`