Prediction of online shoppers' purchasing intention using Probabilistic Graphical Models

Chi-Thien Nguyen Thanh-Trung Tran

Abstract

This study addresses the problem of predicting the purchasing intention of users on e-commerce sites by applying Probabilistic Graphical Models (PGM). Using the "Online Shoppers Purchasing Intention Dataset" from UCI, we preprocessed the data by discretizing continuous variables and compared the effectiveness of several Bayesian Network construction methods, primarily the Hill-Climbing (HC) and Tabu Search structure learning algorithms, alongside selective Naive Bayes approaches. The primary contribution of this work is to demonstrate that PGM serves not merely as a predictive "black box", but as a highly interpretable model providing significant business value. The results showed that the Hill-Climbing model achieved the highest performance with an accuracy of 88.70% on the test set, closely followed by Tabu Search at 88.64%. Both models demonstrated the ability to identify positive purchase instances. This study confirms that PGMs, particularly learned with score-based algorithms like HC and Tabu, are effective and interpretable predictive tools, enabling the exploration of dependency relationships between user behavior and purchasing decisions.

1 Introduction

In the explosive growth era of e-commerce, the online market has become increasingly competitive. For businesses, understanding and anticipating customer behavior is no longer a luxury but a critical necessity for survival and success. One of the most pivotal challenges in this domain is the ability to predict a user's purchasing intention in real-time, during their browsing session. This predictive capability is a key that unlocks significant business value. Specifically, by identifying a user with a high probability of making a purchase while they are still on the site, a business can deploy targeted interventions. For instance, it can **personalize the user experience in real-time** by offering a timely discount to a hesitant buyer, triggering a live chat support window to resolve potential queries, or simply ensuring an exceptionally smooth checkout process. These actions can effectively **reduce cart abandonment rates** and significantly boost conversion metrics.

However, predicting purchasing intention is a non-trivial **technical challenge**. Firstly, the problem is characterized by severe **data imbalance**, where the number of sessions ending in a purchase is far outnumbered by those that do not (in our dataset, only approx. 15.6% of sessions are positive). This makes it difficult for standard models to learn the features of the minority class without being biased towards the majority. Secondly, user behavior is inherently **complex and stochastic**; browsing paths are not linear and are influenced by a multitude of interconnected, often subtle, factors. Lastly, the available data is often **session-based**, representing isolated browsing periods rather than a long-term user history, which makes building deep user profiles challenging.

To address these complexities, this paper proposes the use of Probabilistic Graphical Models (PGM), specifically Bayesian Networks. While many "black-box" models like deep neural networks might achieve high predictive accuracy, they fail to provide insights into *why* a prediction was made. The **main contribution** of this paper is to demonstrate that a PGM is not just an effective predictive tool but, more importantly, a **highly interpretable "glass-box" model that delivers tangible business value**. By revealing the probabilistic dependencies between user actions and the final purchase outcome, a Bayesian Network allows stakeholders to understand the key drivers of conversion. This ability to answer "what if" and "why" questions transforms the model from a simple predictor into a strategic decision-support tool, providing actionable insights that can guide business improvements. This study primarily compares two score-based structure learning algorithms, Hill-Climbing (HC) and Tabu Search, and explores simpler structures based on feature importance.

2 Literature Review / Related Work

This section reviews key research that is directly relevant to this paper, focusing on the principles, structure learning, and application of Bayesian Networks for user behavior analysis.

- **Montgomery, A. L., et al. (2004). "Modeling online browsing and path analysis using session-level data."**
  - *How it works:* This study uses Markov models to analyze the sequence of user actions (clickstream) within a session. They model the transition probabilities between states (e.g., from the homepage to a product page) to understand common paths leading to a purchase.

  - *Relevance to this paper:* While not a full Bayesian Network, this work emphasizes the importance of modeling sequential dependencies in session data, which more general PGM models like Bayesian Networks can capture more flexibly.

- **He, D., et al. (2012). "A Bayesian network based approach for churn prediction in e-commerce."**
  - *How it works:* The authors build a Bayesian Network to predict customer churn, integrating both demographic and transactional data to identify key churn factors.
  - *Relevance to this paper:* The problem of churn prediction is similar to purchase intention prediction. This approach demonstrates that Bayesian Networks are effective at integrating different data types and providing interpretable insights, reinforcing our choice of method.

- **A Tutorial on Learning With Bayesian Networks (Heckerman, 1996, revised 2022)**
  - *How it works:* This tutorial provides a foundational overview of Bayesian Networks (BNs) as graphical models representing probabilistic relationships. Heckerman explains that BNs are advantageous as they can handle missing data, learn causal relationships, and combine prior knowledge with data.
  - *Relevance to this paper:* This work establishes the theoretical basis for our choice of BNs. We utilize their ability to model dependencies and handle uncertainty in the online shopper dataset.

- **A survey of Bayesian Network structure learning (Kitson et al., 2023)**
  - *How it works:* This survey reviews 74 different algorithms for learning the structure of Bayesian Networks from data, categorizing and comparing various structure learning methods, including score-based methods like HC and Tabu Search.
  - *Relevance to this paper:* This survey informs our selection of the Hill-Climbing and Tabu Search algorithms. Our experimental results, showing similar performance levels, align with the survey's discussion of score-based search strategies.

- **The Representational Power of Discrete Bayesian Networks (Ling & Zhang, 2002)**
  - *How it works:* This paper explores the relationship between the structural complexity of a BN and the types of functions it can represent. The authors establish that a BN where each node has at most 'k' parents cannot represent functions containing a certain level of complexity (specifically, (k+1)-XORs).
  - *Relevance to this paper:* This research provides theoretical context for our learned networks, which represent a trade-off between fitting the data and avoiding excessive complexity, a concept related to the representational limits discussed by the authors.

- **Giudici, P., & Castelo, R. (2003). "Improving Markov chain model estimation and selection."**
  - *How it works:* This paper proposes using Bayesian Networks as a generalization of Markov chain models for clickstream analysis, allowing for more complex dependency structures beyond just the immediately preceding state.
  - *Relevance to this paper:* Provides a strong argument for using BNs over simpler sequential models, supporting our use of structure learning algorithms like HC to discover potentially longer-range dependencies in user behavior.

- **Padmanabhan, B., Zheng, Z., & Kimbrough, S. O. (2006). "An empirical analysis of the value of recommending links for Web personalization."**
  - *How it works:* This study empirically analyzes web usage data to understand browsing behavior and evaluates the effectiveness of link recommendations for personalization, implicitly dealing with predicting user paths and intentions.
  - *Relevance to this paper:* Relevant as it deals with analyzing web usage patterns, similar to our goal, though using different methods. It underscores the business value of understanding user navigation.

- **Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). "Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks."**
  - *How it works:* This is the source paper for the dataset used in our study. The authors applied MLP and LSTM models to predict purchase intention, providing baseline results using deep learning approaches.

– *Relevance to this paper:* Directly relevant as it uses the same dataset and tackles the same problem. Provides a benchmark against which our PGM results can be implicitly compared, especially regarding the trade-off between accuracy and interpretability.

3 Problem Statement and Related Definitions

Let Y_i denote the random variable representing the purchasing intention for a given user session i. Y_i is a binary variable where: $Y_i = \{1 \text{ (purchase)}, 0 \text{ (no purchase)}\}$.

We consider a set of observed variables $X_i = \{x_i^1, x_i^2, \ldots, x_i^{17}\}$ representing relevant features for session i. These features include behavioral metrics such as `PageValues` and `ExitRates`, and contextual attributes like `VisitorType` and `Month`.

The primary **goal** of this problem is to learn a predictive model, f, that maps the feature set X_i to the purchase intention outcome Y_i. More formally in a probabilistic context, the objective is to estimate the conditional probability distribution of the purchasing intention given the features of the current session:

$$P(Y_i|X_i) = P(Y_i|x_i^1, x_i^2, \ldots, x_i^{17})$$

Based on this distribution, the final **prediction**, $\hat{Y}_i$, is determined by choosing the outcome with the highest posterior probability:

$$\hat{Y}_i = \underset{y \in \{0,1\}}{\text{argmax}} P(Y_i = y|X_i)$$

The computed posterior probability $P(Y_i = 1|X_i)$ also serves as a valuable confidence score for the prediction, allowing for an estimation of uncertainty.

4 Proposed Methods

Based on the analysis of the source code, we implemented and compared four methods for constructing Bayesian Networks. All methods were built on preprocessed data where continuous numerical variables (`Administrative`, `Administrative_Duration`, `Informational`, `Informational_Duration`, `ProductRelated`, `ProductRelated_Duration`, `BounceRates`, `ExitRates`, `PageValues`, `SpecialDay`) were **discretized** into three factor levels ("Zero", "Low", "High") using a custom function based on the median of their non-zero values. Categorical variables (`Month`, `VisitorType`) were converted to numerical factors. The data was split into a 90% training set and a 10% test set using stratified sampling.

4.1 Method 1: Bayesian Network learned with Hill-Climbing (HC) Algorithm

- **Description:** A score-based structure learning method that iteratively searches for the network structure maximizing a score function (e.g., BIC) via local edge modifications.
- **Procedure:** Used `hc()` from `bnlearn` for structure learning on the training set, followed by `bn.fit()` for parameter learning. Inference used `predict()` with `bayes-lw`.

4.2 Method 2: Bayesian Network learned with Tabu Search Algorithm

- **Description:** Another score-based search algorithm that explores the space of network structures. It uses a "tabu list" to avoid revisiting recently explored structures, helping it escape local optima potentially better than simple Hill-Climbing.
- **Procedure:** Used `tabu()` from `bnlearn` for structure learning on the training set, followed by `bn.fit()` for parameters. Inference used `predict()` with `bayes-lw`.

4.3 Method 3: Selective Naive Bayes Network (Top 7 Features)

- **Description:** A simplified structure where the target variable `Revenue` is assumed to be the child of the top 7 most relevant features (based on Cramer's V correlation), and these parent features are assumed to be conditionally independent given `Revenue`. The top 7 features identified were: `PageValues`, `ExitRates`, `ProductRelated_Duration`, `Month`, `TrafficType`, `Administrative`, `BounceRates`.
- **Procedure:** Defined the network structure manually using `model2network()`. Learned parameters using `bn.fit()` on the training set. Inference used `predict()`.

4.4 Method 4: Whitelist Constrained Hill-Climbing Network (Top 3 Features)

- **Description:** This method uses Hill-Climbing but constrains the search space using a whitelist. Only edges directing from the top 3 features (based on Cramer's V: `PageValues`, `ExitRates`, `ProductRelated_Duration`) towards `Revenue` were explicitly allowed in the initial structure considered by the HC algorithm. Parameter learning used Bayesian estimation (`method = "bayes", iss = 10`).
- **Procedure:** Created a whitelist, learned structure using `hc(training_set, whitelist = whitelist_top3)`, learned parameters using `bn.fit(..., method = "bayes", iss = 10)`. Inference used `predict()`.

5 Experiment Design

5.1 Dataset

- **Name:** Online Shoppers Purchasing Intention Dataset.
- **Source:** UCI Machine Learning Repository. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019).
- **Description:** 12,330 sessions, 18 attributes (after reading CSV), highly imbalanced (approx. 84.4% non-purchase, 15.6% purchase in the final used dataset of 12205 samples after minor cleaning/adjustment implied by code).

5.2 Evaluation Criteria

Accuracy is reported as the primary metric from the test set split. Confusion matrices are presented for detailed error analysis. Additionally, Precision, Recall, F1-Score (for the positive class 'TRUE'), and Specificity (for the negative class 'FALSE') were calculated for the Naive Bayes and Whitelist models. 10-fold cross-validation loss (classification error) was calculated for HC and Tabu Search.

5.3 Experiment Steps

1. **Preprocessing:** Loaded data, converted specific columns to factors, discretized numerical features, converted `Month` and `VisitorType` to numeric factors, removed unused factor levels, converted to standard data frame.
2. **Data Splitting:** Randomly split the final dataset (12205 samples) into a training set (90%, 10984 samples) and a test set (10%, 1221 samples) using stratified sampling based on `Revenue`.
3. **Model Training:** Trained the four described models (HC, Tabu, Selective Naive Bayes, Whitelist HC) on the 90% training set.
4. **Model Evaluation:** Used trained models for prediction on the 10% test set. Calculated accuracy and confusion matrices. For models 3 & 4, calculated additional metrics. Performed 10-fold cross-validation for models 1 & 2.
5. **Result Comparison:** Compiled tables and figures to compare performance and learned structures.

6 Exploratory Data Analysis (EDA)

Before building the models, we performed exploratory data analysis to understand the data characteristics.

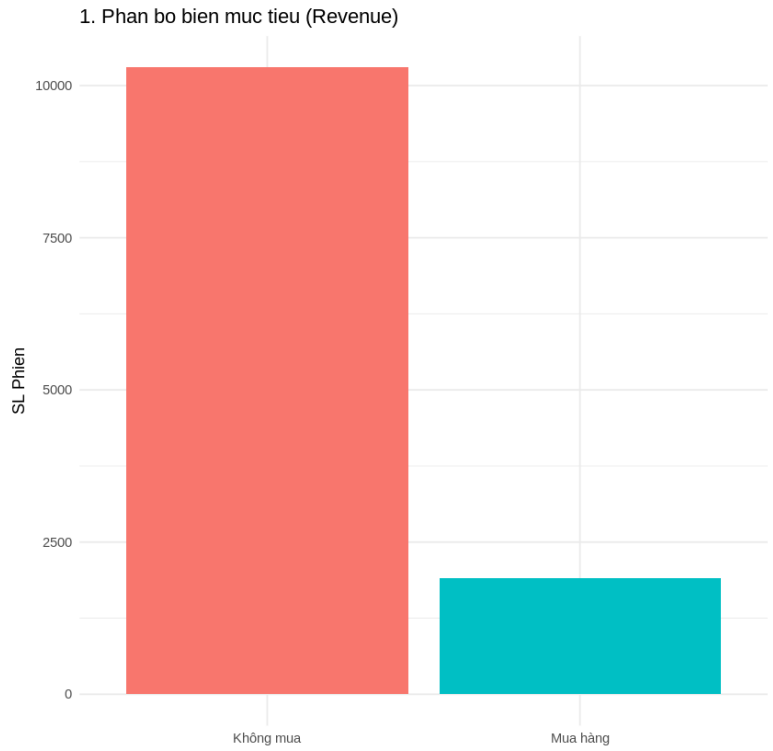

Figure 1: Distribution of the target variable (Revenue), showing class imbalance.

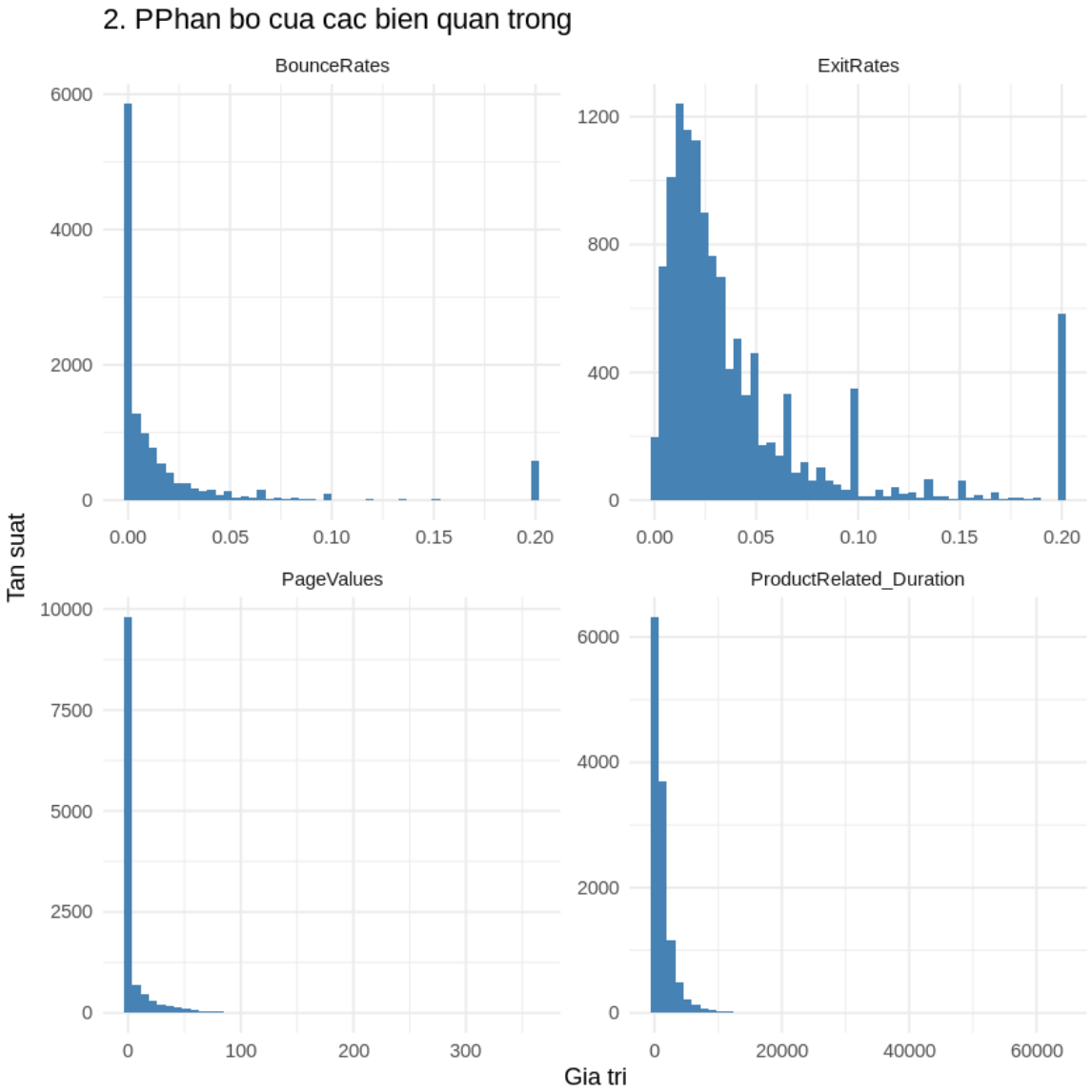

Figure 2: Histograms showing the distributions of key numerical variables. Note the strong left skew.

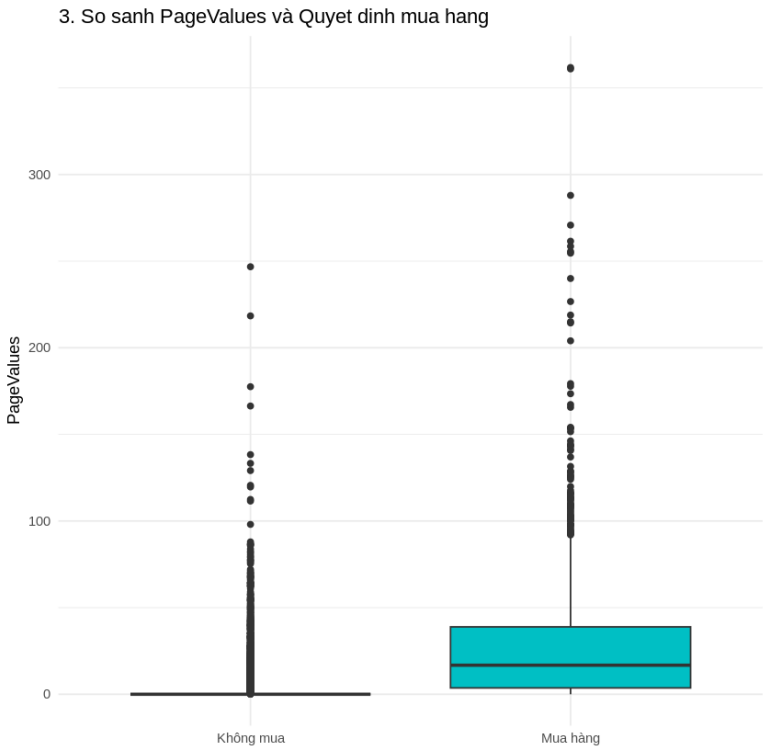

Figure 3: Boxplot comparing PageValues between purchasing and non-purchasing sessions. Higher PageValues strongly indicate purchase intention.

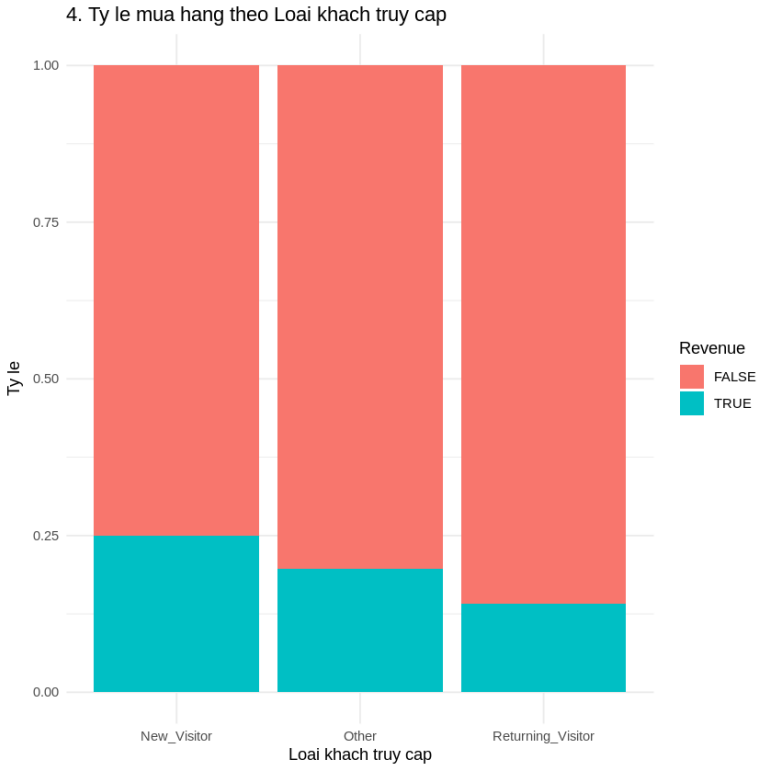

Figure 4: Purchase rate by Visitor Type. New visitors show a higher propensity to purchase compared to returning visitors.

6.1 Correlation Analysis

To understand the association between the features and the target variable (`Revenue`) after discretization, we visualized the correlation using Cramer's V. Figure 5 shows the strength of association between each feature

and `Revenue`.

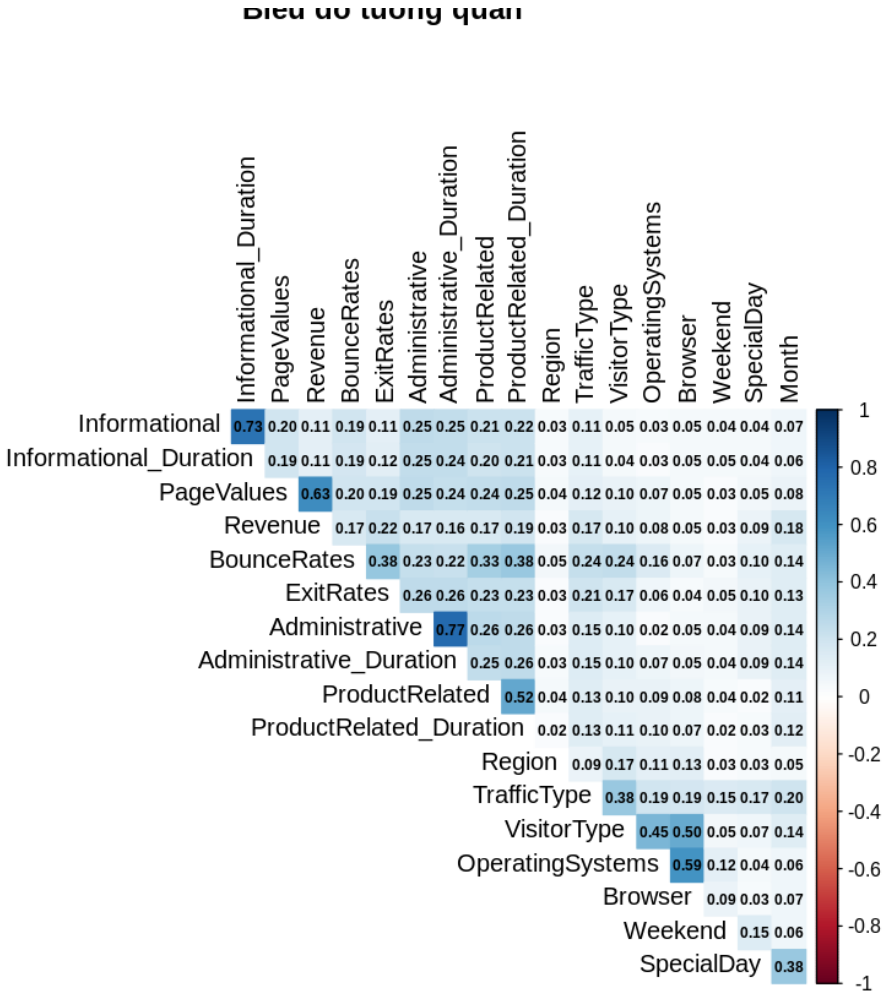

Figure 5: Visualization of Cramer's V association between features and the Revenue target variable.

As depicted in Figure 5 and consistent with the boxplot in Figure 3, `PageValues` exhibits by far the strongest association with `Revenue` (Cramer's V = 0.628). `ExitRates` (0.223) and `ProductRelated_Duration` (0.192) show moderate associations. Other variables like `Month` (0.176), `TrafficType` (0.172), `Administrative` (0.169), and `BounceRates` (0.169) have weaker, yet noticeable, associations. This analysis informed the feature selection for the Selective Naive Bayes (Top 7) and Whitelist (Top 3) models.

7 Experiment Results

The models were trained on 90% of the data (10984 samples) and evaluated on the remaining 10% test set (1221 samples).

7.1 Performance Metrics

Model	Accuracy	Precision	Recall	F1_Score	Specificity
1. Hill Climb	**88.70%**	0.6514	**0.5969**	**0.6230**	0.9408
2. Tabu Search	88.64%	0.6514	**0.5969**	**0.6230**	0.9404
3. Revenue Naive Bayes (Top 7)	85.26%	0.5279	0.5445	0.5361	0.9097
4. Whitelist Top 3 (HC Bayes)	88.40%	**0.7083**	0.4450	0.5466	**0.9658**

Table 1: Comparison of model metrics on the 10% test split. Precision, Recall, F1_Score are for the positive class (Revenue=TRUE). Specificity is for the negative class (Revenue=FALSE).

		Actual	
		False	True
Predicted	False	969	77
	True	61	114

Table 2: Confusion matrix for Hill Climb.

		Actual	
		False	True
Predicted	False	963	77
	True	67	114

Table 3: Confusion matrix for Tabu Search.

7.2 Cross-Validation Results (10-fold on entire dataset)

10-fold cross-validation provided a more robust estimate of generalization error (expected prediction loss/classification error).

- **Hill-Climbing (HC):** Expected Loss = 0.103081 (Equivalent Accuracy ≈ **89.69%**).
- **Tabu Search:** Expected Loss = **0.1029171** (Equivalent Accuracy ≈ 89.71%).

The results show very similar performance between Hill Climbing and Tabu Search, both achieving around 88.6-88.7% accuracy on the test set and estimated 89.7% accuracy via cross-validation. Both successfully identified a significant number of positive cases (114). The Selective Naive Bayes performed worst (85.26% accuracy), while the Whitelist model achieved good accuracy (88.40%) with high precision and specificity but lower recall.

7.3 Learned Network Structures

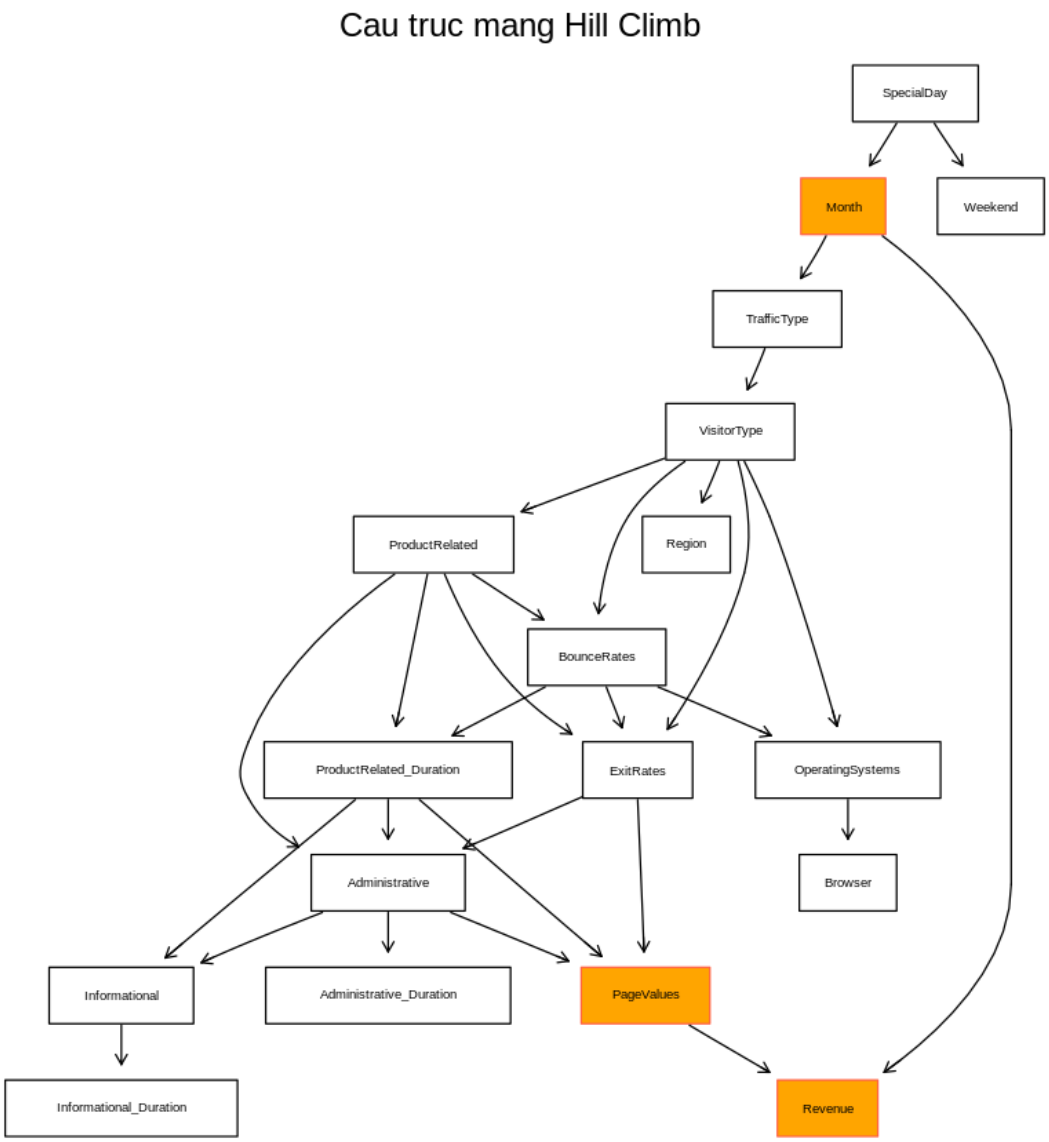

Figure 6: Bayesian Network structure learned using the Hill-Climbing algorithm.

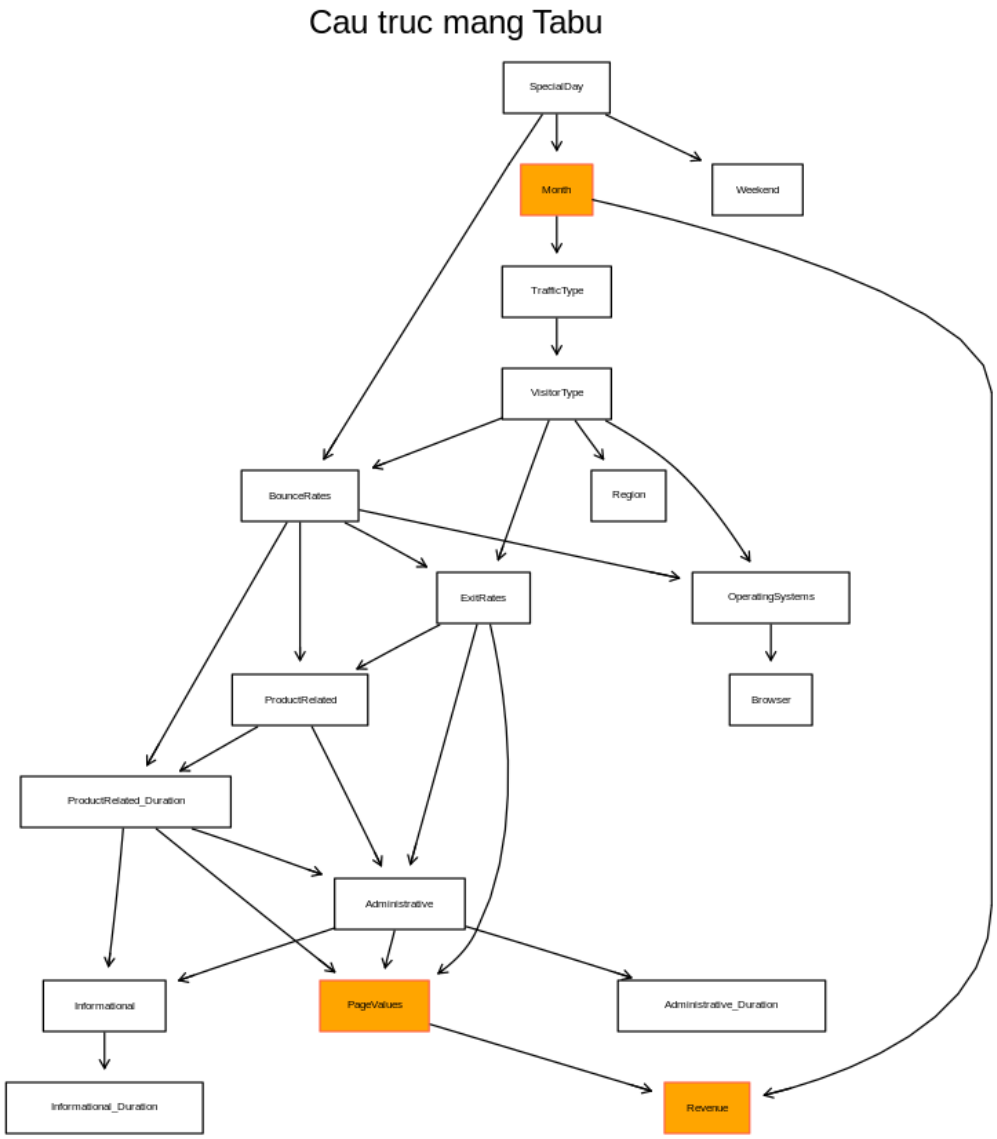

Figure 7: Bayesian Network structure learned using the Tabu Search algorithm.

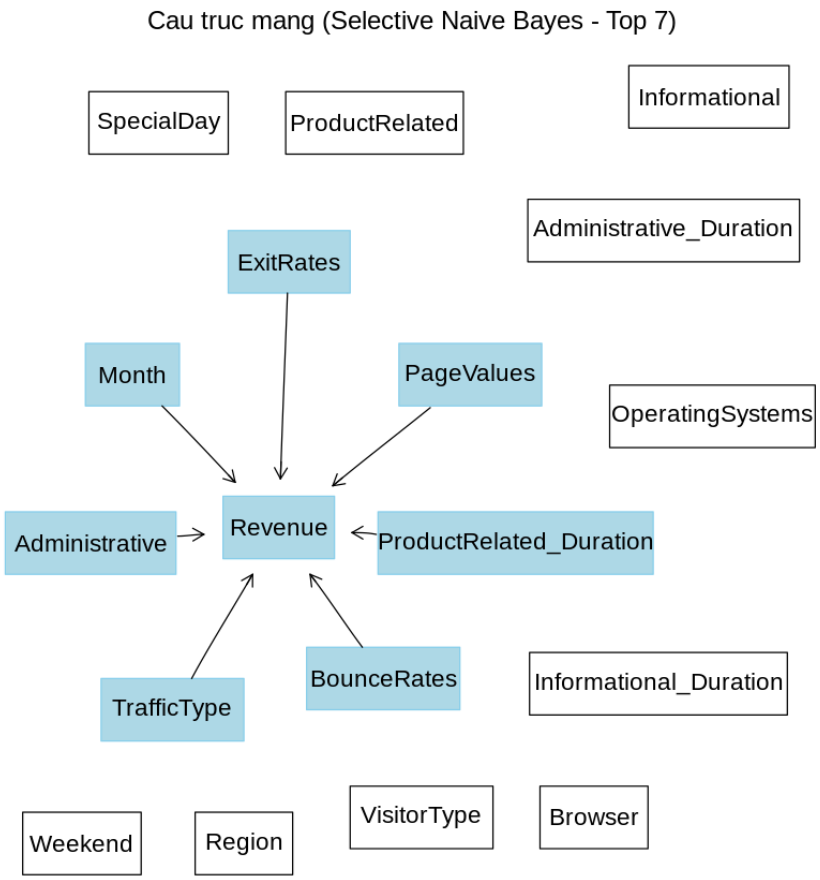

Figure 8: Structure of the Selective Naive Bayes (Top 7) model.

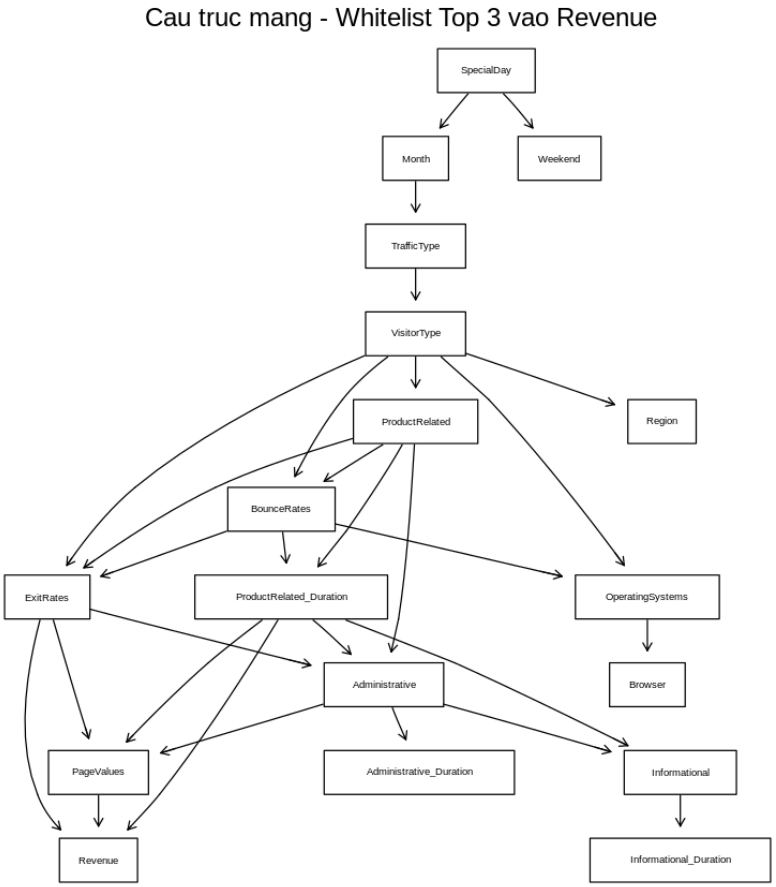

Figure 9: Structure learned using Hill-Climbing constrained by the Top 3 feature whitelist.

Figures 6 and 7 show the complex structures learned automatically by the score-based algorithms. Figures 8 and 9 depict the simpler, constrained structures.

7.4 Performance Comparison Visualization

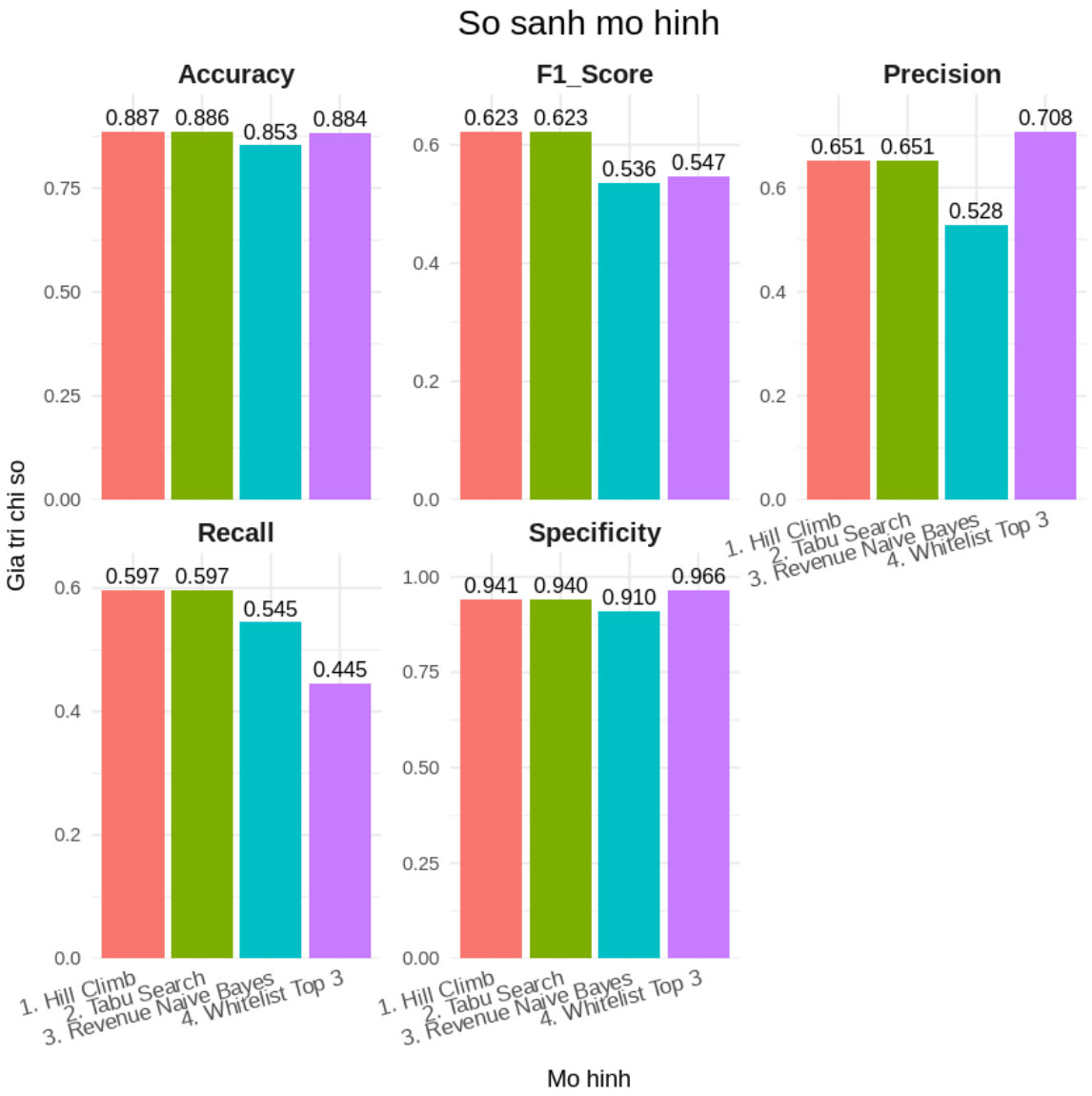

Figure 10: Bar chart comparing key performance metrics across the four models.

Figure 10 visually confirms the similar high accuracy and F1-scores of HC and Tabu, the lower performance of Naive Bayes, and the trade-offs (high precision/specificity, low recall/F1) of the Whitelist model.

8 Conclusions

8.1 Evaluation based on Results

The experimental results demonstrate that score-based structure learning algorithms, specifically **Hill-Climbing (HC)** and **Tabu Search**, are effective for predicting online shopper purchase intention using this dataset. Both models achieved similar high accuracy (around 88.7% on the test set, 89.7% estimated by CV) and successfully identified a substantial portion of the positive (purchase) class, addressing the challenge of data imbalance to a reasonable extent.

The simpler structures, Selective Naive Bayes (Top 7) and Whitelist Constrained HC (Top 3), yielded lower overall performance (especially Naive Bayes) or specific trade-offs (Whitelist HC had high precision but poor recall). This suggests that allowing the algorithms (HC/Tabu) to learn complex dependencies across most features captured the underlying patterns better than relying solely on pre-selected features in a simplified structure for this problem.

The successful application of HC and Tabu Search, combined with the inherent interpretability of the learned Bayesian Network structures (Figures 6, 7), validates this paper's central thesis. We have shown that PGM

can function as a powerful "glass-box" tool for this business problem, providing relatively accurate predictions alongside actionable insights into the probabilistic factors (like `PageValues`, `ExitRates`, and `Month`) that directly influence customer conversion, offering significant value beyond simple prediction.

8.2 Future Work

Despite the promising results, this study has limitations that open avenues for future research:

- **Addressing Data Imbalance:** Explicitly applying advanced balancing techniques (e.g., SMOTE, cost-sensitive learning) during training and evaluating with metrics robust to imbalance (e.g., F1-Score, AUC-PR) remains a key area for improvement.
- **Advanced Discretization:** Investigating supervised discretization methods (e.g., entropy-based binning like MDLP) could potentially enhance model performance by creating more informative feature intervals.
- **Benchmarking:** To provide a comprehensive performance context, the PGM models should be benchmarked against other state-of-the-art machine learning algorithms, including tree-based ensembles (XGBoost, Random Forest) and potentially deep learning models designed for tabular data.
- **Algorithm Tuning:** The parameters for HC and Tabu Search (e.g., scoring function, tabu list length) were used with defaults; tuning these could potentially yield further gains.

References

References

[1] Giudici, P., & Castelo, R. (2003). Improving Markov chain model estimation and selection. *Applied Stochastic Models in Business and Industry*, *19*(1), 15-30.

[2] He, D., Zhao, H., & Bao, F. (2012). A Bayesian network based approach for churn prediction in e-commerce. *Procedia Engineering*, *29*, 1305-1309.

[3] Heckerman, D. (2022). *A Tutorial on Learning With Bayesian Networks* (arXiv:2002.00269v3 [cs.LG]). arXiv. `https://arxiv.org/abs/2002.00269`

[4] Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., & Chobtham, K. (2023). A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, *56*(8), 8607-8687. `https://doi.org/10.1007/s10462-023-10378-5`

[5] Ling, C. X., & Zhang, H. (2002). The Representational Power of Discrete Bayesian Networks. *Journal of Machine Learning Research*, *2*, 1-13.

[6] Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using session-level data. *Marketing Science*, *23*(4), 579-595.

[7] Padmanabhan, B., Zheng, Z., & Kimbrough, S. O. (2006). An empirical analysis of the value of recommending links for Web personalization. *Decision Support Systems*, *42*(2), 407-422.

[8] Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, *31*(10), 6893-6908. `https://doi.org/10.1007/s00521-018-3523-0`