Exploratory Data Analysis:
Online Shoppers Purchasing Intention

Bolai Yin (MSDSA 2024, Silicon Valley Campus) | Contact: yin.bol@northeastern.edu

Introduction:

This project explores online shoppers' purchasing intention using exploratory data analysis. Based on a dataset with 12,330 e-commerce sessions and 18 variables, I examined behavioral patterns distinguishing purchase (15.5%) vs. non-purchase sessions to uncover key factors influencing buying decisions.

Data Summary:

- 12,330 user sessions from UCI Online Shoppers Purchasing Intention Dataset *(https://archive.ics.uci.edu/)*
- 18 features spans Numerical and Categorical variables (e.g. Pages Visited, PageValues, Bounce/Exit Rates, etc.)

Key Features	Mean	Std	Min	Max
PageValues	5.889	18.568	0	361.764
ProductPage_Duration	1194.746	1913.669	0	63973.522
ExitRates	0.043	0.049	0	0.2
BounceRates	0.022	0.049	0	0.2

Tech Stack:

- Python (pandas, seaborn, matplotlib)
- Logistic Regression & KMeans (scikit-learn)
- Visualization: Seaborn, Matplotlib

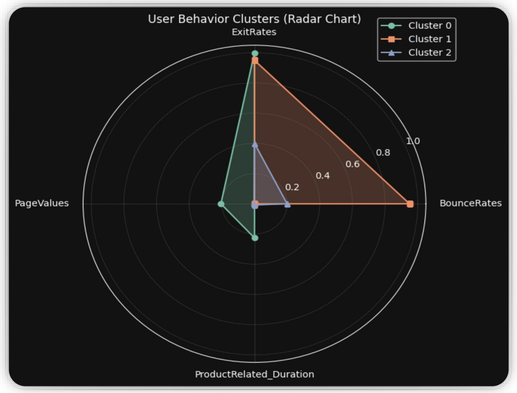

Fig. 1 *User Clustering via KMeans*

Cluster 0 shows high engagement(scaled x 10 for clarity); Cluster 1 has high bounce rates with low PageValues.

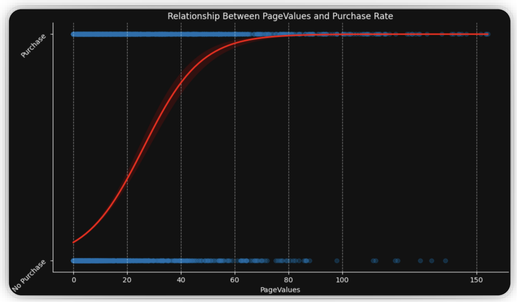

Fig. 2 *Purchase Probability by PageValues (Logistic Fit)*

Probability increases sharply when PageValues exceeds 20; saturates after 50.

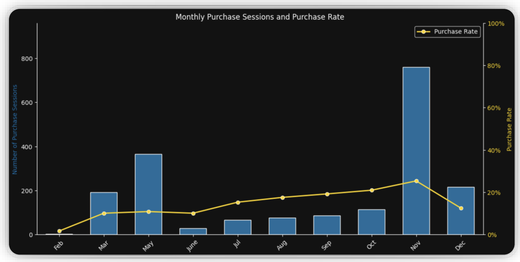

Fig. 3 *Monthly Purchase Sessions and Rate*

Purchase sessions spike in March, May, and Nov, indicating seasonal trends.

Insights & Takeaways:

- Users who spend more time are more likely to purchase
- Cluster analysis reveals clear user segments; Cluster 0 users show highest engagement and conversion (Fig. 1)
- Purchases peak in March, May, Nov(Fig. 2)
- Page Values below 40 are optimal for influencing purchases; logistic model confirms its predictive strength (Fig. 3)
- Behavioral clustering supports audience segmentation for targeted marketing strategies

These findings support data-driven marketing strategies and inform the development of predictive tools..

Future Work:

- Train supervised models (e.g., Random Forest) for conversion prediction.
- Forecast seasonal trends using ARIMA or Prophet.
- Develop an interactive app with Streamlit to visualize user segments and key metrics.