Who Shops Where? A Data-Driven Analysis of Consumer Preferences for Online Shopping

Hung Tran

DSAN 5000

December 16, 2024

Executive Summary

This report investigates customer preferences for online vs. in-person shopping, focusing on how demographic and financial information can predict these behaviors. Key findings reveal that higher income and spending levels are associated with online shopping, while demographic factors like race, education, and marital status can influence channel preferences. Leveraging machine learning models, businesses can tailor strategies to optimize customer segmentation, enhance omnichannel experiences, and drive sales. Models like XGBoost and Random Forest emerged as top performers, offering robust predictive capabilities and actionable insights.

Objective

The primary objective of this project is to analyze and predict customer preferences for online vs. in-person shopping using demographic and financial data. The insights aim to help retail businesses optimize marketing strategies, inventory management, and customer engagement across channels in a digital age of ever-increasing competition.

Definitions

Online shopping in this context refers to buying tangible retail goods online to then be delivered or picked up. For example, such goods can be clothing, electronics, furniture and jewelry. Online shoppers refer to those who have made at least one purchase online in 2023, while non-online or “in-person” shoppers have not.

Key Insights

1. **Income and Spending Patterns**:

	Non-online shopper	Online shopper
Mean Annual Income	$118,269	$153,273
Median Annual Income	$105,910	$146,477
Mean Annual Expenses	$103,344	$138,153
Median Annual Expenses	$80,362	$114,650

Figure 1: Table showing mean and median annual income and expenses for non-online and online shoppers.

- Online shoppers have significantly higher annual income and total expenses compared to non-online shoppers. The median annual income for online shoppers is a little more than $40,500 greater than non-online shoppers which is about 38% more. The median total annual expense for online shoppers is about $34,300 more, or around 34% more.

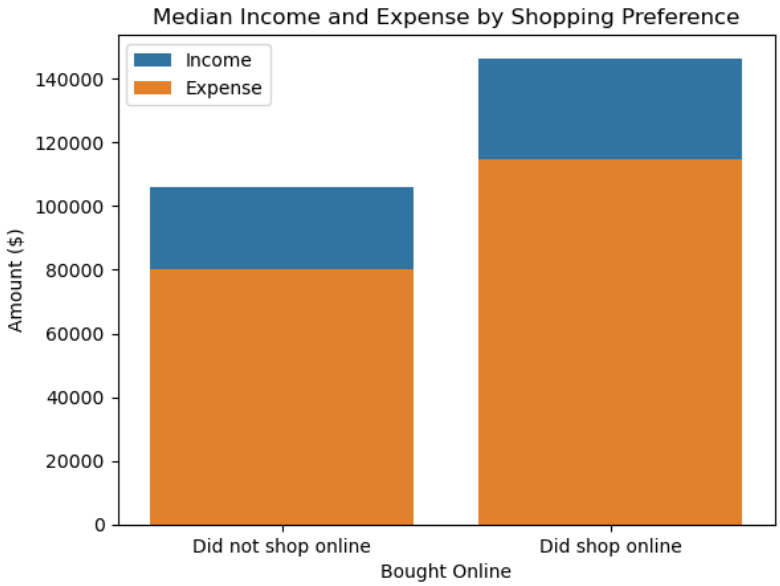

Figure 2: Bar plot of median annual income and expenses by online preference

- Online spending habits also varies significantly across racial groups, with Asian-identifying individuals spending more online compared to other racial categories at $8,170.44 on average. This amount is around 5.7 times more than the next highest average, Black customers, at $1430.52. The fact that the mean online shopping expenditure for Asians was more than the other groups was statistically significant. Conversely, the Native American community on average spent the least in 2023 at $356.67 on average.

2. **Demographic Influences**:

- Factors like race, education level, and marital status are strongly associated with online shopping preferences. After running statistical testing, there was evidence to conclude a statistically significant association between each of the three variables and whether or not they have shopped online in 2023.
    - **Race:** Only 5% of Black survey participants are found to have shopped online which indicate preference for in-person shopping,

and possibly reflect differences in access or trust in online platforms.

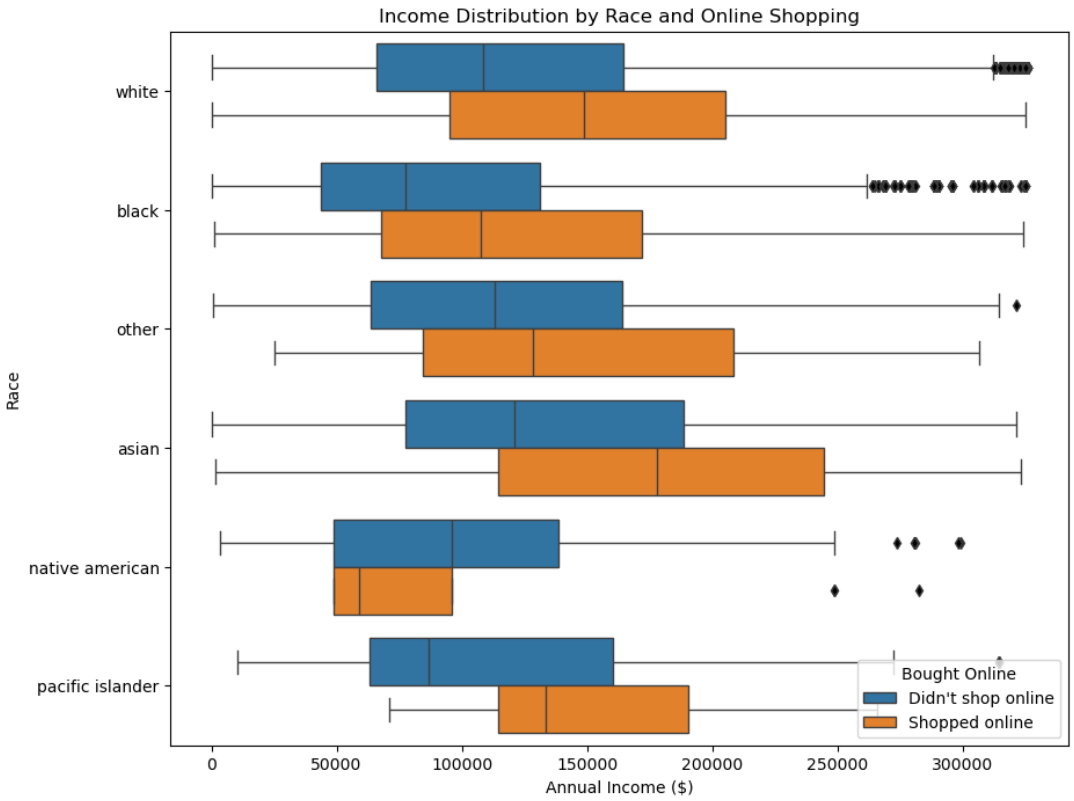

Figure 3: Grouped box plots of annual income organized by race and shopping preference

- **Marital Status**: Married individuals showed the greatest likelihood to shop online while widowed individuals showed the least.
- **Education Level:** In addition, survey participants with a bachelor's or graduate degree are most likely to shop online at 11% and 12%, respectively. This may indicate the possible correlation between education and digital literacy.

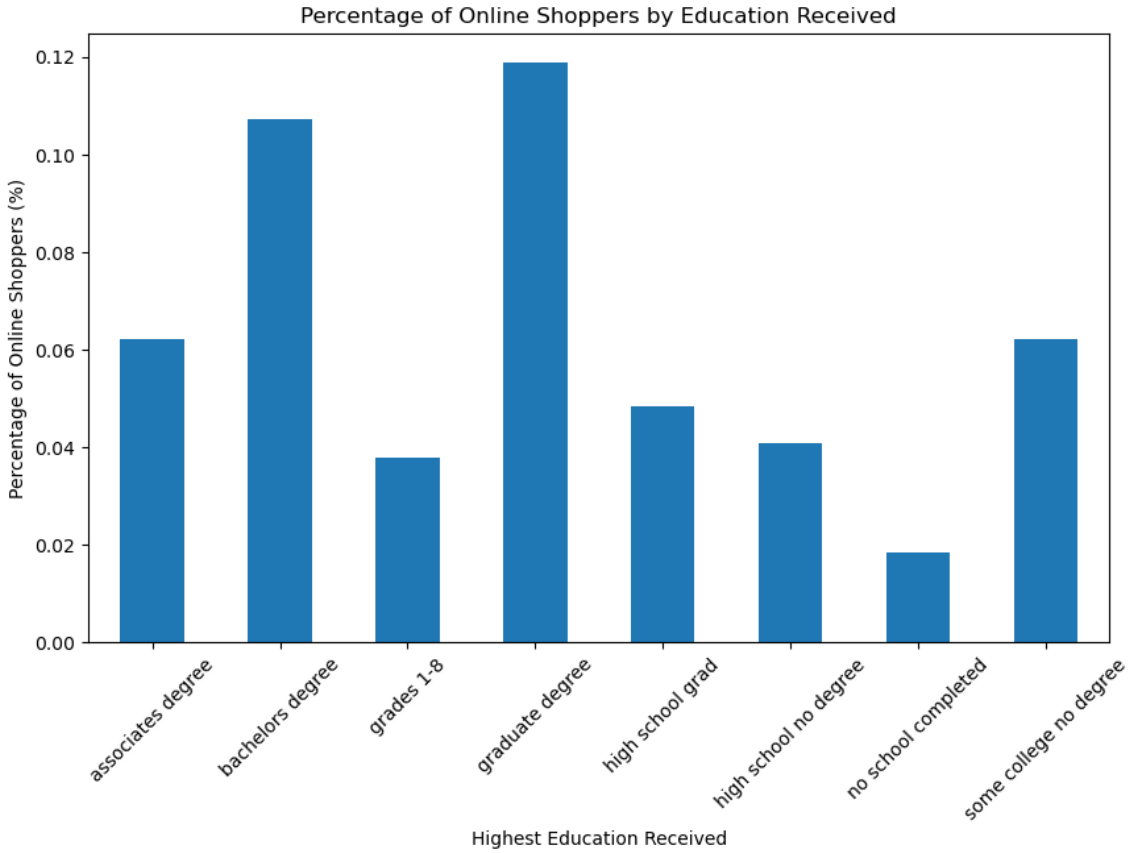

Figure 4: Bar graph of percent of online shoppers by highest education received

- These demographic variables were significant predictors in machine learning models, emphasizing their importance in understanding customer behavior.

3. **Modeling Insights**:

- Logistic Regression, XGBoost, and Random Forest achieved perfect classification metrics, demonstrating the dataset's strong predictive power and the clear separability of online vs. in-person shoppers based on demographic and financial features.

- Linear models like Logistic Regression offer simplicity and interpretability. Ensemble methods like XGBoost and Random Forest excel in handling complex, non-linear relationships, being able to identify nuanced patterns in the data.

- Though these models indicate strong performance on the training data, they need to be tested on validation data to confirm their effectiveness, guarantee good generalization to unseen data, and prevent overfitting.

Business Implications

1. **Customer Segmentation**:
    - Use predictive models to identify high-income customers likely to prefer online shopping.
    - Tailor marketing campaigns to demographics favoring specific channels, such as offering targeted promotions to Asian, married or college-educated customers to drive online channel growth.
2. **Channel Optimization**:
    - Allocate inventory and resources to channels based on predicted preferences. For example, increase online stock at appropriate warehouses and improve delivery options for regions with higher online shopping affinity.
3. **Enhanced Personalization**:
    - Develop personalized recommendations based on spending patterns and demographic profiles. For instance, those who are least likely to shop online can be introduced to this channel through in-store sales staff and be given promotional programs to bring them in.
    - Offer loyalty programs, referral programs, or other incentives that cater to specific channel preferences. This can increase customer lifetime value, reduce churn, and expand customer base.
4. **Reaching Untapped Segments**:
    - Address barriers preventing in-person shoppers from transitioning online, such as improving trust in digital platforms or offering free shipping and returns.

Recommendations

1. **Implement Predictive Analytics**:
    - Adopt XGBoost or Random Forest for customer segmentation prediction. Use these types of models to identify key customer profiles and adjust strategies dynamically.
        - For example, these models can be used to predict *college graduates* and *graduate degree holders*, plus *high earning individuals* to target with marketing campaigns because our research has confirmed that they are most likely to online shop.

- *XGBoost* algorithms offers a combination of robustness, flexibility, and generalizability, decreasing risk overfitting.
- *Random Forest* algorithms can handle large datasets with high dimensionality and also offers interpretability. It also has high accuracy and reduces overfitting risk.

2. **Tailor Marketing Strategies**:
    - Launch personalized campaigns based on income, race, and education level insights. For instance, emphasize convenience and speed for high-income online shoppers and value-driven promotions for in-person shoppers.
3. **Optimize Inventory and Fulfillment**:
    - Use predictions to align inventory allocation with customer preferences, ensuring products are available where and how customers shop. Companies would highly benefit from collecting their own data on shopping preferences. Especially with geographic data, they can stock their stores and warehouses according to regional demands for online and in-store products.
4. **Improve Online Experience**:
    - Simplify the online shopping process for individuals who are less likely to shop online by enhancing user interfaces, providing tutorials, and addressing security concerns.
5. **Enhance In-Store Experiences**:
    - For in-person shoppers, create engaging experiences such as personalized service, special events, or exclusive in-store promotions.

Conclusion

This project highlights the value of leveraging demographic and financial data to predict shopping preferences and vice versa. By adopting predictive models and tailoring strategies from gained insights accordingly, retail companies can enhance customer satisfaction, improve operational efficiency, and increase revenue. Future efforts should focus on validating findings across broader datasets and exploring additional factors influencing shopping behavior. Data is collected at an unprecedented rate so it is time for companies to leverage it for growth and a competitive advantage.