Project 2 Proposal: Supervised Learning (Classification)
Predicting Online Sales

Problem

Critical micro-moments in human decision-making can, if met with the appropriate response or intervention, define the decision-maker's behavior and perception. In the context of shopping, the critical moment is when the shopper is searching for information and deciding whether to purchase a product—at which point the retailer can be there and useful to the shopper. If done well, the retailer's assistance can help increase the likelihood of the shopper making that purchase and the shopper will likely think more favorably of the brand.

The Internet and online shopping has changed the way shoppers interface with retailers, in that web marketing teams have replaced the role of the retail shop worker and now shoppers seek information through web pages. The goal of every web marketing team who are responsible for these pages, then, is to optimize the web page content and design to make the shopper's experience positive and more likely to lead to a purchase.

Through a shopper's website activity, can we predict whether the shopper will make a purchase or abandon the site? With this information, retailers can possibly identify shoppers who are intending to buy something and the critical moments that make or break the sale.

Data

UCI Machine Learning Repository "Online Shoppers Purchasing Intention" Dataset:
http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

Independent Variable	Type	Description	Use for Model?
Administrative	Int	Number of pages visited by the visitor about account management	Y
Administrative Duration	Int	Total amount of time (in seconds) spent by the visitor on account management related pages	Y
Informational	Int	Number of pages visited by the visitor about website, communication, and address information of the shopping site	Y
Informational Duration	Int	Total amount of time (in seconds) spent by the visitor on informational pages	Y
Product Related	Int	Number of pages visited by visitor about product-related pages	Y
Product Related Duration	Int	Total amount of time (in seconds) spent by the visitor on product-related pages	Y
Bounce Rate	Float	Google Analytics Metric - Average bounce rate value of the pages visited by the visitor: percentage of visitors who enter the site from that web page and then leave ("bounce") without triggering any other requests to the analytics server during that session	Y
Exit Rate	Float	Google Analytics Metric - Average exit rate value of the pages visited by the visitor: for all page views to the web page, the percentage that were the last in session	Y
Page Value	Float	Google Analytics Metric - Average page value of the pages visited by the visitor: The average value	

		for a web page that a user visited before completing an e-commerce transaction (and/or landed on the "goal page") - Represents the average "Page Value" for all of the pages visited by the visitor during the session; updated when the visitor moves to another page - Intent: tells which page in the website contributed more to the site's revenue - Ex. if a page wasn't visited during the ecommerce transaction session at all, page value is $0. Interpretation: change your site's content! - Calculated by: $\frac{\text{Ecommerce Revenue + Total Goal Value}}{\text{Number of Unique Pageviews for Given Page}}$ https://support.google.com/analytics/answer/2695658?hl=en More info: https://online-metrics.com/page-value/	
Special day	Float	Closeness of the site visiting time to a special day (e.g., Mother's Day, Valentine's Day): Value is determined by considering the dynamics of e-commerce (e.g., duration between order date and delivery date) - Ex. Valentine's Day: nonzero value Feb. 2-Feb. 12; zero value before and after this date; maximum value of 1 on Feb. 8	Y
OperatingSystems	Categorical int	Operating system of the visitor	Y
Browser	Categorical int	Browser used by the visitor	Y
Region	Categorical int	Geographic region from which the session has ben started by the visitor	Y
TrafficType	Categorical int	Traffic source by which the visitor has arrived at the website (e.g., banner, SMS, direct)	Y
VisitorType	Categorical string/object	Visitor type as "New Visitor", "Returning Visitor", "Other"	Y
Weekend	Categorical (Boolean)	Date of the site visit on weekend (or not)	Y
Month	Categorical string/object	Month value of the visit date	Y
Dependent Variable	**Type**	**Description**	**Use for Model?**
Revenue	Categorical (Boolean)	Class label indicating whether the visit has been finalized with a transaction	Y

Project 2 Proposal: Supervised Learning (Classification)

Predicting Online Sales

Known Unknowns/Barriers

- Operating systems, browser, region, and traffic type are all labeled as categorical integers in the dataset. It would be nice to know their meaning for actual insights, but I have not yet found them
- The 17 features are the only information available.
  - I would also be interested to know what kind of products this dataset is based on. Or if there are differences across products.
  - I would also like to know if these sessions were via mobile devices or computers, if possible, since my outside research on ecommerce trends has noted a higher abandonment rate during mobile searches
  - Some of the features may have covariance or interactions—making the actual number of useful features for my model even smaller
  - While weekend or not is a good starting point to know about timing of web activity, perhaps I would want to know more detail of weekday so I can advise retailers on how to time website content updates or maintenance accordingly
  - While special occasion timing can approximate shopper's intent, I wonder if the "Is this a gift?" checkout option is featured on the website and reveals similar or different "intent"
  - I would also think that whether the shopper added items to their shopping cart may be another measure of the shopper's original intent in visiting the website, but it is not listed among the features in this dataset
  - I think that knowing whether recommender systems were present on the pages correlated with increased number of product related page visits would be interesting to know, to ensure the recommender system is effective—or ideally, whether the shopper clicked into a product page through a recommended item link

Potential Resources

- NumPy, Pandas
- Statsmodels, SK Learn
- Matplotlib, Seaborn
- Shapley or Lime for interpretability
- Possibly other data sources or scraping resources, if time permits/available

MVP

Classification model (likely K-Nearest Neighbors to start) using variables that are "clean" (i.e., consistently represented across all observations in my sample).