**Final Project Proposal**

**Background** In an increasingly digital age, companies are looking to find ways to capitalize on the opportunities given when potential customers visit their sites. While many of the same methods to target those customers are still relevant, the digital channel creates new challenges and friction points that are not necessarily present in traditional channels for retail or sales-driven industries. Companies like Adobe and Google have developed some standardized measurements and metrics to help with measuring digital session behavior. Metrics like Bounce Rate and Exit Rate, which are both intended to measure when a consumer or potential consumer would leave a given web page, are prime examples of metrics that can help a business to understand their customer behavior. This information can then be used to address those challenges that customers face that are more particular to the given channel, and not necessarily demographic or economic features that are largely out of a business’s control.

**Motivation**  
 The following paper looks to establish a potential method that an analyst can utilize to identify what might and might not lead to a customer making a purchase during an online session. Some of the questions the analysis will intend to tackle will be as follows:

1. *Is there a significant difference in purchasing decisions between operating systems, browser, region, etc.(categorical/binary)?*Factors such as these would help to explain how customers are affected differently, and that these effects are essentially decided for them prior to them ever accessing the site. If there is a difference in operating system or browser, these may be potential fixes to consider how the site or app performs differently under those circumstances. Similar ideas could be considered with region, in that ISPs or even language may result in different experiences for a customer in one region and a similar one in another.

2. *Is there a significant difference in purchasing decisions based on the number of pages visited, the amount of time spent on different types of pages, bounce rates, exit rates, etc. (continuous/interval)?* Customers who spend more time on some pages, or exit from ones at different rates may be experiencing friction points that an analyst identifying could advise so that resolution can occur.

3. *Can we identify the most important predictive factors for whether someone will make a purchase that visits the website?*Being able to do identify this will allow a business to capitalize on factors that improve the likelihood of a transaction or resolve those that prevent them. The other two questions address this a little more specifically, but building the proper model would allow a solution to be more generalized and potentially quantify the important factors.

**Dataset:** [Online Shoppers Purchasing Intention Dataset](https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset)  
 The dataset consists of 12,330 online sessions in which a person either did or did not make a purchase, which is indicated by the binary response variable. The response is unbalanced, which can be contended with by oversampling the response variable or adjusting the model’s predictions, such as changing the probability threshold for logistic regression. There is an abundance of observations and roughly 17 predictors to be utilized in making predictions. Many of the variables are digital metrics measured by Google Analytics, which are expected to be some of the most important variables, where others are various system, location, and time-based variables to help with explaining some of the variance that the digital metrics would not explain. The observations span a 1-year period, with each observation representing a unique user-session. Curiously, the timeframe looks as though it excludes the month of January and April, and it does not appear there is any indication given in previous literature or documentation as to why that would be. As such I believe it will be easy to find other papers and literature that reference these metrics for similar problems, to include one by the original data provider to the UCI ML database using this very dataset.

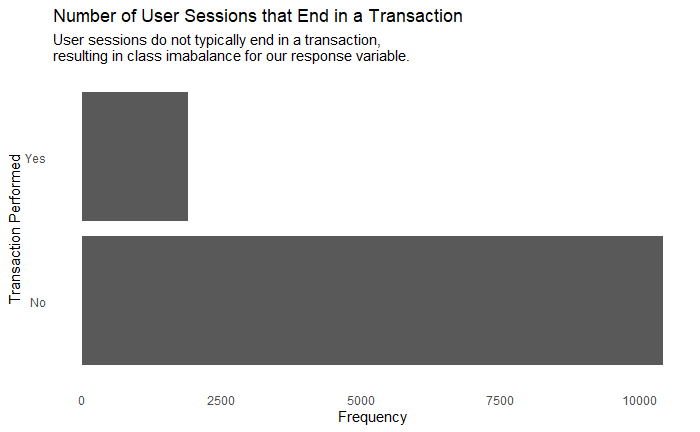
**Proposed Analysis** The analysis will occur in multiple phases, not including things like cleaning and summarizing data. Visual exploration will occur to understand distributions of variables, as well as the relationships between individual predictors. This step of the analysis can potentially be enough to resolve the first two questions; differences in frequency of categorical variables, as well as differences in distribution for the numeric variables, will likely be made apparent from the visualizations. More robust methodologies will also be used, to include tests such as Chi-Square goodness of fit and test of independence for the categorical variables. For the numeric variables, methods such as factorial logistic regression and ANCOVA can be employed to determine the difference in the response variable because of these variables.  
 The tests mentioned above can be a good way to see if the effect on the response variable exists, but a business may not have the resources to address all problems simultaneously. Furthermore, a statistically significant effect could be identified but may mean very little if it cannot be quantified. Creating a proper machine learning model would allow an analyst to properly quantify how much one or multiple factors may be impacting the likelihood of a transaction, giving a sort of priority order to improve a sites performance. Models such as Logistic Regression and Linear Discriminant Analysis should help with quantifying the effect of predictor variables, while other models such as Linear SVM, PLS, MARS, and Random Forest models can at least give an order of importance, if not a quantifying the effect on response. These sorts of methods would be preferred to more black-box methods, because ultimately the focus is on controlling what the business can control and taking advantage of opportunities where possible, and less on the actual prediction itself.

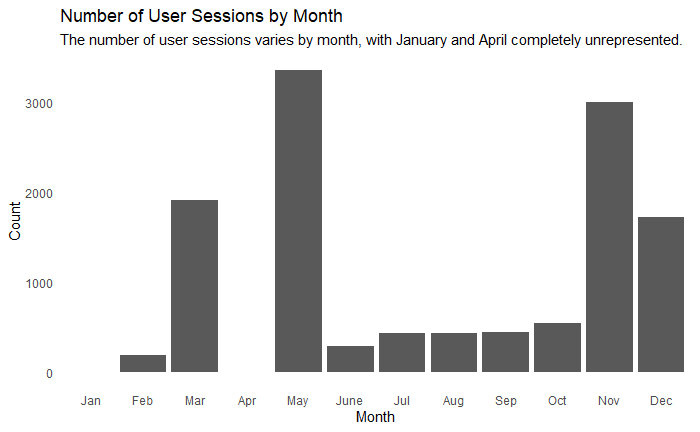
**References**

“Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks” C. Okan Sakar, S. Olcay Polat, Mete Katircioglu & Yomi Kastro, Article from Neural Computing and Applications, available on springer through UTSA Libraries, <https://link.springer.com/article/10.1007/s00521-018-3523-0>

“[GA4] Analytics dimensions and metrics” Google Analytics Help, <https://support.google.com/analytics/table/13948007?sjid=9224210117807248745-NC&visit_id=638650376835318674-629778858&rd=2>

**Appendix**

****

****

**Variable Descriptions**Thefollowing variable descriptions are taken from the article “Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks,” which is the original use case for the dataset.

Numeric Variables **Administrative:** Number of pages visited by the visitor about account management  
**Administrative duration:** Total amount of time (in seconds) spent by the visitor on account management related pages  
**Informational:** Number of pages visited by the visitor about Web site, communication and address information of the shopping site  
**Informational duration:** Total amount of time (in seconds) spent by the visitor on informational pages  
**Product related:** Number of pages visited by visitor about product related pages  
**Product related duration:** Total amount of time (in seconds) spent by the visitor on product related pages  
**Bounce rate:** Average bounce rate value of the pages visited by the visitor  
**Exit rate**: Average exit rate value of the pages visited by the visitor  
**Page value:** Average page value of the pages visited by the visitor  
**Special day:** Closeness of the site visiting time to a special day

Categorical Variables  
**Operating Systems:** Operating system of the visitor  
**Browser:** Browser of the visitor  
**Region:** Geographic region from which the session has been started by the visitor  
**Traffic Type:** Traffic source by which the visitor has arrived at the Web site (e.g., banner, SMS, direct)  
**VisitorType:** Visitor type as ‘‘New Visitor,’’ ‘‘Returning Visitor,’’ and ‘‘Other’’   
**Weekend:** Boolean value indicating whether the date of the visit is weekend   
**Month:** Month value of the visit date  
**Revenue:** Class label indicating whether the visit has been finalized with a transaction