Data Analytics Capstone Topic Approval Form

Student Name: Alexa Fisher

Student ID: 000354665

Capstone Project Name: Predictive Modeling via Logistic Regression for Online Purchasing Intention

Project Topic: This project will use Online Shopper's Purchasing Intention data published by UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset). This contained information about consumers who visited an e-commerce site within a one-year period and their various page metrics as noted by Google Analytics.

☒ **This project does not involve human subjects research and is exempt from WGU IRB review.**

Research Question: What attributes within the captured Google Analytics metrics can contribute to the possibility of a customer purchasing an item?

Hypothesis: **Null hypothesis**-. The logistic regression model shows that the predictor variable of "Revenue" and the explanatory variables do not have a statistically significant relationship.
Alternate Hypothesis-The logistic regression model shows a statistically significant relationship between the predictor variable "Revenue" and explanatory variables.

Attribute	Data Type	Brief Description
Administrative	Quantitative	The number of Administrative type pages a user visited.
Administrative_Duration	Quantitative	The amount of time a user spent on an Administrative page.
Informational	Quantitative	The number of Informational type pages a user visited.
Informational_Duration	Quantitative	The amount of time a user spent on an Information page.
ProductRelated	Quantitative	The number of Product Related type pages a user visited.
ProductRelated_Duration	Quantitative	The amount of time a user spent on a Product related page.
BounceRates	Quantitative	The percentage of visitors who enter the website and exit without triggering any other tasks.
ExitRates	Quantitative	The percentage of page views on the website ending on a specific page.
PageValues	Quantitative	The average value of the page that a user visited before completing a transaction
Month	Qualitative	The month the page view occurred.
SpecialDay	Quantitative	The value indicating the closeness of the browsing date to a special day or holiday.
OperatingSystems	Qualitative	The integer value representing the operating system a user was using.
Browser	Qualitative	The integer value representing the browser a user was using.
Region	Qualitative	The integer value representing the region a user was located.
TrafficType	Qualitative	The integer value representing what type of traffic the user is categorized into.
VisitorType	Qualitative	Identification of whether a user is a new visitor, returning visitor, or other.
Weekend	Qualitative	The Boolean value if the session was on a weekend.

Revenue	Qualitative	The Boolean value if the user completed the purchase or not.

Context: Online Purchasing has become a common occurrence in today's society. The utilization of Google Analytics within E-commerce has offered an enormous variety of insights that improve profitability. It is captured by user behavior within the website platform such as the number of times each page is visited, how long a user interacted with the page, or even if a product was purchased. The project can identify if any of these metrics could provide a significant relationship between if an item is purchased or not. Significance is notated as a p-value of less than 0.05.

Data: The data needed to populate the logistic regression model is published by UC Irvine Machine Learning Repository. It has a Creative Commons Attribution 4.0 International license, which allows for the adaptation of the dataset for any purpose. The data set had a total of 12,330 entries and 18 variables. The variables are notated as 10 quantitative or continuous variables, and 8 qualitative or categorical variables. There was 0% sparsity within the dataset. The below chart provides a brief description and data type classification of the attributes within the dataset.

Data Gathering: The data containing the user website behaviors captured by Google Analytics, specific time period, and if revenue was obtained was published by the UC Irvine Machine Learning Repository over a one-year period. It was limited to one year to capture all the annual holidays. The dataset is downloadable in a .CSV format from the UC Irvine website. The data was free of any missing or duplicated values. The categorical variables with high cardinality were omitted from analysis. The dataset was also analyzed to not include any variables with high multicollinearity.

Data Analytics Tools and Techniques: Data cleaning and exploratory data analysis will be performed to explore and visualize the dataset before any analysis is completed. This ensures that the dataset is clean and free of any outliers as well as gain insight on the distributions of the continuous variables. Upon completion of this, the dataset is split into training and testing sets. The testing set consists of 30% and the training set consists of 70% of the observations. The

The project will utilize the following tools: Jupyter Notebook, Python, and Libraries of Pandas, NumPy, MatPlotLib, Seaborn, SciPy, Statsmodels, and SciKit Learn

Justification of Tools/Techniques: Python is a programming language that supports data analysis very well due to its use of various packages and libraries. It also has great readability and ease of use. The usage of Jupyter Notebook allows for visualizations and computations to be performed in a web-based environment. The following libraries were also useful for the analysis:

Pandas allows for the dataset to be imported within the environment. It also allows for the data to be explored, cleaned, and analyzed with the various functions included.

NumPy allows for different mathematical computations and the ability to work with array within the data wrangling and predictive model phases of the analysis.

MatplotLib and Seaborn are both useful in providing visualizations during various stages.

SciKit Learn and Statsmodels allows for statistical computations to be completed. These were shown within their packages of variance inflation factor, logit, confusion matrix, and train test split. Variance inflation is used to check on multicollinearity. Logit is used to perform the logistic regression model to show the coefficients within the dataset and how they compare to the target variable as it is associated with the probability of occurrence. The confusion matrix and train test split packages provide data for visualization of the True or False positive and negative predictions.

Project Outcomes: The project will provide model that is able to provide which variables, if any, that are statistically significant to the target variable of "Revenue". Statistical significance is deemed as a p-value of less than 0.05. The accuracy rate would provide how well the explanatory variables could predict the chance of "Revenue". An accuracy rate of over 80% is ideal.

Projected Project End Date: 07/31/2023

Sources: Sakar,C. and Kastro,Yomi. (2018). Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5F88Q.

Google. (n.d.). How page value is calculated - analytics help. Google. https://support.google.com/analytics/answer/2695658?hl=en

Understanding "traffic sources" in google analytics. Practical Ecommerce. (2022, April 16). https://www.practicalecommerce.com/Understanding-Traffic-Sources-in-Google-Analytics

Kumar, Braveen. (2023, March 6). Google Analytics for Ecommerce in 2023 (Complete Guide). Shopify. https://www.shopify.com/blog/14681601-google-analytics-for-ecommerce-a-beginners-guide

Course Instructor Signature/Date:

☐ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor's Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.