Student: Priyanka Gangwal Date: Dec 2022

INFO 656-01 Machine Learning: Final Project Proposal

Introduction: Data science can be used to discover user behavior and needs (Shih, 2020). Designers today are encouraged to leverage from the capabilities of ML to enhance experiences for users (*Machine Learning*, 2017). Investigating Google Analytics (GA) data such as pageviews, bounces, device category, etc. can be used to enhance the UX of a website (Pillai, 2019).

Aim: The project aims to use ML algorithms to investigate the behavior of individuals browsing the Google Merchandise website (https://shop.googlemerchandisestore.com). Google makes their user data available on Google Cloud and can be queried using BigQuery.

Research questions: The project will answer the following research questions:

1. Will a new visitor to the Google Merchandise website make a revenue-generating transaction in the ecommerce store?
2. Which features of the website/user experience are relevant in contributing to making purchases?

Solution:

- Data source: Visitor-wise data for 6 months (1 August 2016 through 31 January 2017) will be collected from Google Merchandise's GA data using BigQuery (*Google Analytics Sample Dataset for BigQuery - Analytics Help*, n.d.). The names of the columns that will be sourced are listed in Appendix A.
- ML algorithms:
    - The project will find the answer to the first predictive question if future visitors of the website make a purchase on the website using 3 learning techniques - logistic regression, neural networks and random forests. Given the independent variable takes binary values, the aforementioned ML techniques are appropriate. The project will use the best model among the 3 for the final prediction of visitors making a purchase.
    - To answer the second research question about identifying important features which impact purchase behavior on the website, the project will calculate the feature importance within the best model.
- Output: The outcome of this project - model predictions and behavior statistics, can be used to make informed UX decisions like prioritizing usability tests of website features or conducting surveys.

Significance: There are very few publicly available ML projects utilizing GA data (Appendix B). The projects reviewed include one on Exploratory Data Analysis of a website's GA data, and three others that analyze Google Merchandise GA data sourced from BigQuery. The latter projects analyze the data by predicting revenues and SEO related behavior. This project will be unique as it will find the best model to predict the probability of a visitor making a transaction and discover UX elements of the website which drive purchases.

Limitations: While Google BigQuery gives a year-long dataset (Aug 1, 2016 - Aug 1, 2017), the project will be extracting data for just 6 months due to Google's data export and storage restrictions. Further, given the data limitations, the exhaustive list of features will not be used - the project will consider only those variables that are expected to be the most impactful.

References:

1. *Google Analytics sample dataset for BigQuery—Analytics Help. (n.d.). Retrieved October 30, 2022, from https://support.google.com/analytics/answer/7586738?hl=en#access-the-dataset&zippy=%2Cin-this-article*

2. *Machine Learning: The Future of UX*. (2017, November 6). Maven Wave. https://www.mavenwave.com/white-papers/machine-learning-future-ux/
3. Pillai, S. R. (2019, June 19). *7 Google Analytics metrics that help in optimizing Website UX for higher conversions*. Medium. https://uxdesign.cc/7-google-analytics-metrics-that-help-in-optimizing-website-ux-for-higher-conversions-7e9309eb6516
4. Shih, B. (2020, February 19). *Applying machine learning to your UX research process*. Medium. https://uxdesign.cc/applying-machine-learning-to-your-ux-research-process-8eb4075ee275

Appendices:

Appendix A: List of dataset columns that will be used in the project -

1. visitNumber,
2. visitId,
3. totals.visits,
4. totals.hits,
5. totals.pageviews,
6. totals.timeOnSite,
7. totals.bounces,
8. totals.transactions,
9. totals.transactionRevenue,
10. totals.newVisits,
11. totals.screenviews,
12. totals.uniqueScreenviews,
13. totals.timeOnScreen,
14. totals.totalTransactionRevenue,
15. trafficSource.source,
16. trafficSource.medium,
17. device.browser,
18. device.operatingSystem,
19. device.isMobile,
20. device.deviceCategory,
21. geoNetwork.continent,
22. geoNetwork.country,
23. geoNetwork.city,
24. hits.hitNumber,
25. hits.time,
26. hits.hour,
27. hits.minute,
28. hits.page.searchKeyword,
29. hits.page.searchCategory,
30. hits.page.pagePath,
31. hits.page.hostname,
32. hits.page.pageTitle,
33. hits.transaction.transactionId,
34. hits.transaction.transactionRevenue,
35. hits.transaction.currencyCode,
36. hits.transaction.transactionCoupon,
37. hits.item.productName,
38. hits.item.productCategory,
39. hits.type,
40. fullVisitorId,

41. userId,
42. clientId

Appendix B: ML projects using GA data

1. https://github.com/GunnarGriese/data-science-blog-post
2. https://github.com/indraninp/Google-Analytics-data-analysis-from-an-e-commerce-store
3. https://www.kaggle.com/code/erickvarela/ga-api-classification-ecommerce-transactions?scriptVersionId=55826491
4. https://pufferr.co.uk/a-machine-learning-study-of-the-google-analytics-metrics-predicting-content-quality/