Case Study: How Does a Bike-Share Navigate Speedy Success?

Jamilah Foucher

July 4, 2023

Contents

1 Project Summary/Executive Summary

The objective of this work was to understand how annual members and casual riders use Cyclistic bikes differently, and then predict membership using key features from the Cyclistic data at an accuracy of 80% or better.

The Baseline Solution is to use the features trip_time and rideable_type to predict whether new customers are will be members or casual users; using a kmeans model, features (trip_time, rideable_type) can be perturbed systematically to determine with of the two features should be changed to minimize distance to the member centroid such that a predicted casual user would become a member.

The methodology used to implement the baseline solution, train and test lifestyle prediction models, is classification analysis using kmeans. The evaluation metric includes prediction accuracy for trained and test datasets, and minimum feature/s change needed to make a casual user into a member if the model predicts causal user for a new customer.

In summary, the project/business objective was SATISFIED because the baseline solution produced an analysis result showing that lifestyle can be predicted at 80% accuracy for both training and test data.

The project can be improved by collecting more data from casual users and women; the dataset was statistically significantly (one sample z-test: p¡0.05) biased to men members. Higher prediction accuracy between members and casual users, maybe possible if demographic samples were equally represented.

2 Business Objective

The objective of this work is to answer three questions using past data via statistical analysis Hypothesis testing, such that future marketing decisions can be made. The three questions include:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Using the statistical analysis Hypothesis testing results, the final objective of this work is to predict membership using key features from the Cyclistic data at an accuracy of 95% or better.

3 Project Background

Cyclistic is a bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8 percent of riders use the

assistive options. Cyclistic users are more likely to ride for leisure, but about 30 percent use them to commute to work each day.

Future success of Cyclistic depends on maximizing the number of annual memberships, thus it is necessary to design a new marketing strategy to convert casual riders into annual members. Casual riders are customers who purchase single-ride or full-day passes, and Cyclistic members are customers who purchase annual memberships.

There are three teams that operate Cyclistic :

- Manager: responsible for promoting the program (social, media, email, etc)
- Marketing analytics team : responsible for doing the data analysis
- Executive team: responsible for approving the proposal

4 Baseline Solution

The Baseline Solution (top three recommendations) is to :

1. use trip_time, rideable_type, gender, and birthyear_INT to predict member_casual.
2. recommend bike usage type: if the model predicts that a new customer is a casual user, marketing should encourage usage of member bike preferences such as classic or electric bikes (rideable_type)
3. recommend bike routes: if the model predicts that a new customer is a casual user, marketing should encourage usage of the bikes like a member such as recommending short trip time routes that are beautiful/interesting (trip_time).

5 Project Deliverables: Methodology & Evaluation metrics

5.1 Methodology

I wrote an automatic GCP ingestion program in bash (https://github.com/j622amilah/automatic_GCP_ingestion) that:

1. Downloaded the data from the public bucket using the AWS SDK,
2. organized/unzipped the zip files into three folders (csvdata, remaining_files, zipdata),
3. evaluated the header of each csv file using the first header file as a main reference of comparison for all the other files; similar words with respect to the main reference header were replaced in each csv file. Three folders were created such that identical csv files could be grouped together: exact_match_header, no_match_header, similar_match_header. The evaluation algorithm was re-run on the files in the similar_match_header folder to find all unique table types, in a recursive decision tree like fashion.

4. files in each of the exact_match_header folders were uploaded to GCP and a UNION operation was performed to append all the csv files; for this analysis two table types were found. The two types of datasets were: A) 39 datasets targeting latitude & longitude, member type, and bike preference (rideable_type), B) 25 datasets targeting trip duration, member type (usertype), gender, and birthyear. The two large datasets were joined using a FULL JOIN on the ride_id and trip_id primary key. The SQL table was reduced to 12 columns: fin_trip_ID, rideable_type, trip_time, fin_stsname, fin_stsID, fin_esname, fin_esID, trip_distance, member_casual, bikeid_INT, gender, birthyear_INT.

5. An automatic bigquery statistical analysis program was written such that five main features/columns were compared with the member_casual column; two categorical columns [rideable_type, gender] and three numerical columns [trip_distance, trip_time, birthyear_INT]. The categorical columns were evaluated by calculation the probability of occurrence, and the numerical columns were evaluated using the one and two sample z-statistic with respect to the population and individual sample means respectively.

6. A bigquery logistic regression ML model was used to predict the binary label member_casual using the five features.

The categorical probability of occurrence of each feature for member_casual show that

The numerical z-statistic results show that member statistically have shorter trip_time than casual users, also members are statistically older than casual users by 6 years. In terms of occurrence, men are more likely to be members than women because more men use bikes. Classic and electric bikes tend to be used more by members than casual members. Based on these statistics, casual riders might buy annual membership if they grow older, have an age similar to the average age of membership. Similarly, casual riders might buy membership if they start to desire to do short trip time sessions, or have a preference for classic or electric bikes . Digital media about a short organized trip routes, usage of classic or electric bikes, and marketing for young adults might help casual users to become members; members like short trip time sessions and classic or electric bikes. Also, young adults are less likely to be members so special marketing to non-likely member candidates may encourage them to join thus gaining more money for Cyclistic; older adults are already motivated to be members so they need little to no marketing.

5.2 Evaluation metrics

Evaluation metrics were probability of occurrence for categorical features, one and two sample z-statistic for numerical features, accuracy of logistic regression where default percentage values of 0.75 and 0.25 were used for the train and test dataset respectively.

6 Recommendations for improving the project in the future

Using the machine learning model, a prediction of whether a new customer will be a member can be made for each new person who fills out a Cyclistic form. If the model says that the new person will be a casual member, appropriate marketing based on the Hypothesis testing results can be given to the person such that they are more likely to become a member.