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

Jamilah Foucher

July 5, 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 95% or better.

The Baseline Solution is to use at least the features trip_time and rideable_type to predict whether new customers are will be members or casual users; using a kmeans model, the member centroid location was estimated using the cosine similarity distance measure. Thus the distance from the estimated centroid to each feature (trip_time, rideable_type, and birthyear) was calculated and the feature with the largest deviation (ie: outlier) was suggested as a marketing recommendation.

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 the test dataset, and feature distance difference from the member centroid location.

In summary, the project/business objective was SATISFIED because the baseline solution produced an analysis result showing that membership likelihood can be predicted at 98% accuracy for the test data. In addition, the predominate recommendations based on outliers for predicted casual users were in alignment with statistically significant variables for member vs casual, such as trip_time and birthyear.

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 male members. Higher prediction accuracy between members and casual users or better detection of outliers, maybe possible if demographic samples were equally represented.

The code and results for this case study are at https://github.com/j622amilah/automatic_GCP_ingestion. I hope to publish the GCP bigquery statistic library and the GCP bigquery case study library, such that others can use the Google Cloud Platform tools in an automated manner.

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.

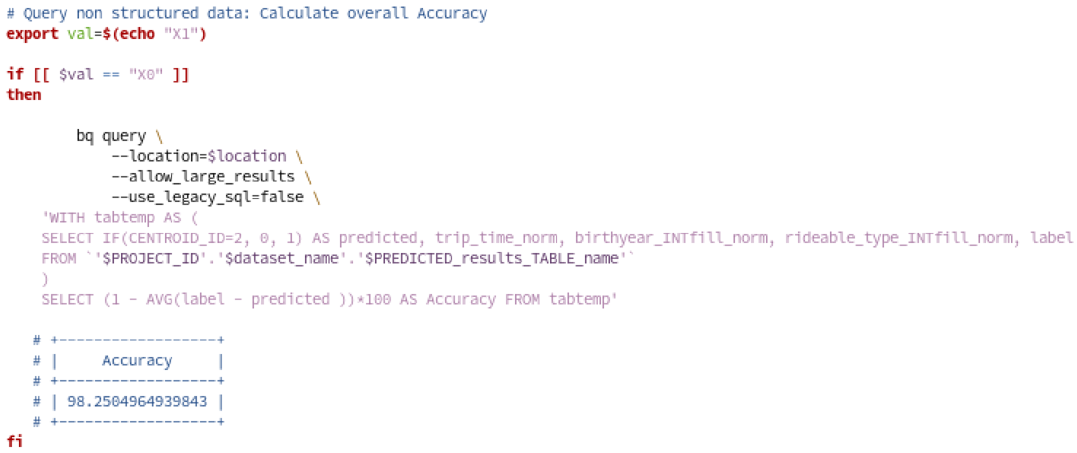
6. A bigquery logistic regression ML model was used to predict the binary label member_casual using the five features.

The categorical probability of occurrence of each feature for member_casual shows that the dataset was statistically bias for male member data ($p < 0.05$); roughly 30 percent of the data consisted of samples from male members.

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.

Figure 1 shows that the prediction accuracy of kmeans for the test dataset was 98%.

Additionally, the equation in the Appendix was implemented and Figure 2 shows the marketing recommendations that were given to casual users; trip_time and birthyear were the most recommended forms of marketing which were in alignment with numerical feature statistical results. This



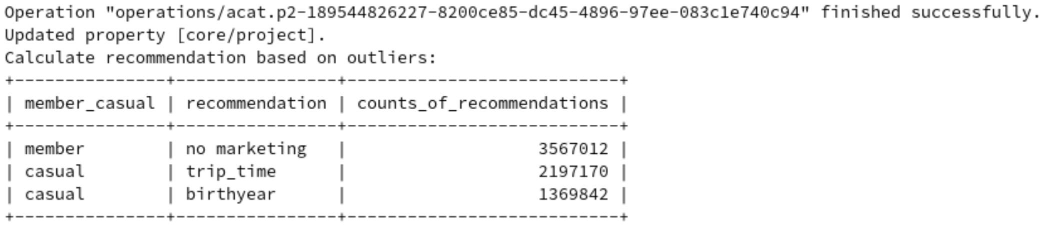
```
# Query non structured data: Calculate overall Accuracy
export val=$(echo "X1")

if [[ $val == "X0" ]]
then

        bq query \
            --location=$location \
            --allow_large_results \
            --use_legacy_sql=false \
    'WITH tabtemp AS (
    SELECT IF(CENTROID_ID=2, 0, 1) AS predicted, trip_time_norm, birthyear_INTfill_norm, rideable_type_INTfill_norm, label
    FROM `'$PROJECT_ID'.'$dataset_name'.'$PREDICTED_results_TABLE_name'`
    )
    SELECT (1 - AVG(label - predicted ))*100 AS Accuracy FROM tabtemp'

   # +------------------+
   # |     Accuracy     |
   # +------------------+
   # | 98.2504964939843 |
   # +------------------+
fi
```

Figure 1: Accuracy of test dataset of the kmeans model.



```
Operation "operations/acat.p2-189544826227-8200ce85-dc45-4896-97ee-083c1e740c94" finished successfully.
Updated property [core/project].
Calculate recommendation based on outliers:
+---------------+----------------+---------------------------+
| member_casual | recommendation | counts_of_recommendations |
+---------------+----------------+---------------------------+
| member        | no marketing   |                   3567012 |
| casual        | trip_time      |                   2197170 |
| casual        | birthyear      |                   1369842 |
+---------------+----------------+---------------------------+
```

Figure 2: Recommendation result counts for members and casual user for the test dataset.

analysis shows that kmeans is an effective way to recommend marketing using outlier detection of features.

5.2 Evaluation metrics

Evaluation metrics were : 0) probability of occurrence for categorical features, 1) one and two sample z-statistic for numerical features, 2) one sample z-statistic for categorical features, 3) prediction accuracy for kmeans using the test dataset. The train test split percentage was 0.75 and 0.25 respectively.

6 Recommendations for improving the project in the future

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 male members. Higher prediction accuracy between members and casual users or better detection of outliers, maybe possible if demographic samples were equally represented.

7 Appendix

7.1 Calculation of the ”should of, would of, could of” cluster centroid

The ”should of, would of, could of” cluster centroid (SWC_centroid) is an equivalent point in feature space that represents the actual cluster centroid, using the assumption that all the features have the same value.

The goal is to solve for the A variables of the cosine similarity equation. The A variables are the feature values of the desired member cluster.

$$cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=0}^{n-1} A_i B_i}{\sqrt{\sum_{i=0}^{n-1}(A_i)^2}\sqrt{\sum_{i=0}^{n-1}(B_i)^2}} \tag{1}$$

Move the B variables to the left side of the equation, because the B variables are the features and the values are known.

$$\left(\frac{\sqrt{\sum_{i=0}^{n-1}(B_i)^2}}{\sum_{i=0}^{n-1} B_i} cos\theta\right) = \frac{\sum_{i=0}^{n-1} A_i}{\sqrt{\sum_{i=0}^{n-1}(A_i)^2}} \tag{2}$$

Square both sides to remove the square root on the denominator on the right side.

$$\left(\frac{\sqrt{\sum_{i=0}^{n-1}(B_i)^2}}{\sum_{i=0}^{n-1} B_i} cos\theta\right)^2 = \left(\frac{\sum_{i=0}^{n-1} A_i}{\sqrt{\sum_{i=0}^{n-1}(A_i)^2}}\right)^2 \tag{3}$$

$$\frac{\sum_{i=0}^{n-1}(B_i)^2}{(\sum_{i=0}^{n-1} B_i)^2} cos^2\theta = \frac{(\sum_{i=0}^{n-1} A_i)^2}{\sqrt{\sum_{i=0}^{n-1}(A_i)^2}} \tag{4}$$

Let the left portion equal to EQN to simplify the equation. Write the summation out into series expansion to see how to reduce the right side equation.

$$EQN = \frac{(A_0 + A_1 + A_2 + \ldots)^2}{(A_0^2 + A_1^2 + A_2^2 + \ldots)} \tag{5}$$

$$EQN = \frac{(A_0 + A_1 + A_2 + \ldots)(A_0 + A_1 + A_2 + \ldots)}{(A_0^2 + A_1^2 + A_2^2 + \ldots)} \tag{6}$$

$$EQN = \frac{A_0^2 + A_1^2 + A_2^2 + 2A_0A_1 + 2A_0A_2 + 2A_1A_2 + \ldots}{A_0^2 + A_1^2 + A_2^2 + \ldots} \tag{7}$$

$$EQN = 2A_0A_1 + 2A_0A_2 + 2A_1A_2 + \ldots \tag{8}$$

Let the A variables equal each other, so that the estimated centroid distance for one feature does not bias the other features.

Let $q = A_0 = A_1 = A_2 = \ldots$

$$EQN = 2q^2 + 2q^2 + 2q^2 + \ldots = n \cdot 2q^2 \tag{9}$$

Solve for q

$$q = \sqrt{\frac{EQN}{n \cdot 2q^2}} = \sqrt{\frac{\frac{\sum_{i=0}^{n-1}(B_i)^2}{(\sum_{i=0}^{n-1} B_i)^2} cos^2\theta}{n \cdot 2q^2}} \tag{10}$$

In GCP, q will be calculated per row/sample so the SWC_centroid for the member centroid is the average of all the rows when member_casual is equal to member.