**A README FILE**

**BY**

**ANDREW IGNATIUS BROWN**

**MASTERSCHOOL**

**ON**

**5TH OCTOBER 2023.**

**TOPIC: CUSTOMER SEGMENTATION OF CUSTOMERS AT TRAVELTIDE.**

1. **OVERVIEW.**

The project focussed on assigning the users of TravelTide perks which the users could find enticing. This is performed to enable users to rely on the services of TravelTide and that the users always return to web address for more services offered by organisation.

In the project, users which have used the services of the organisation from January 4th, 2023, and have had more than seven sessions in the duration were eligible for categorisation for a suitable perk. This criterion allowed for the analysis of the behaviour of customers within a fixed frame of time and are in recent contact or use of the business.

The perks for this project are free hotel meal, free checked bags, no cancellation fees, exclusive discount, and one-night free hotel with meal. This analysis for perks is to the fact that every of the mentioned perk has a special effect on a user: every user has a favourite perk. There are several attributes in the TravelTide database which most likely suit a specific perk, thus, are used in categorising users to the perks.

In this project, attributes gained from the sessions and the users table are original but those earned from the flights, and the hotels table are made from random numbers. The sessions table had users with numerous trip identification numbers, likewise, the trip identification numbers had different users for same numbers.

**2.0 DATA PROCESSING.**

Here, segmenting users to different perk/rewards was performed. First, the relationship between some numerical value attributes (in the dataset) to each other was performed using the correlation analysis. Through the determination of the mentioned relationship, attributes were considered as relating positively or negatively with each other.

In the project, the median was considered as the cut-off point for which users reached qualification for a perk. The exception to this is the Free Cancellation Perk, where an algorithm is used for determining if users were likely to cancel a booking.

The values gained from the calculations performed on the attributes for each perk are represented on a box and on histogram plots. However, not in all circumstances. The mentioned plots are functional in providing statistical description of values; importantly, showing the median, the minimum/maximum values, and the region above the median (the qualification region).

Henceforth in the README report, the considered users would be referred to as Cohorts.

**2.0.1 RELATIONSHIP AMONGST SOME NUMERIC ATTRIBUTES.**

The relationship (correlation process) tends from –1 to 1, where 1 suggests a very strong relationship. This strong relationship states that, as an attribute increase, the other attribute increases too. A relationship value of zero suggest that there is no relationship amongst the attributes. A vale of negative one (-1) says that as strongly, as one attribute increases, the other decreases. Meanwhile, values of greater than or equal to 0.7 are considered strong and those below 0.5 suggest a weak relationship.

A diagram of numbers and symbols

Description automatically generated with medium confidence

**2.1 FREE HOTEL MEAL.**

For this perk, the attributes considered are the hotel amount per room, number of page clicks, the total session time, nights spent, and number of rooms booked by the cohort. From figure 1, there is a relationship between page clicks and total time of session. Therefore, the use of session time is eliminated in choosing cohorts to assign to this perk.

The attributes night spent, hotel amount per room, rooms booked are chosen, as cohorts who have spent much on booked trip(s) would love to earn some treats for their specialization in the company. Session time, and page clicks are considered as attributes for this perk as frequent activities on the webpage could suggest a search for an offer/bargain by the cohort.

Figure 2 below represents a dashboard made such that cohorts which have scores above or below the median could be visualized or adjusted should more or less number of cohorts be required for the perk in future decisions.

A screenshot of a graph

Description automatically generated

**2.2 FREE CHECKED BAGS.**

For this perk, the attributes to consider were the number of checked bags, page clicks, session time, and flight cost (base fare). The qualified attributes were added together to produce a score. The data was a bit bias as it likely exaggerated the amount spent by a cohort on flight. Therefore, debiasing the base fare amount; the cost is divided by the distance of the journey. The result of the mentioned result is hence considered in this report as “real\_cost\_of\_flight”.

The number of checked bags was considered as cohorts whose bags have been checked would be happy to be offered free checks in subsequent flights. While, flight cost was considered as cohorts who’s spent much on fares would need a bargain.

A graph of a bar graph

Description automatically generated with medium confidence

**2.3 NO CANCELLATION FEES.**

Here, attributes were checked for relationships using the pearson’s correlation coefficient method; details in diagram below. Attributes considered included the number of page clicks, total discount received on hotel bills (rate), rate of flight booked total discount received on flight bills, total number of sessions, cost of flight, amount spent on hotels, and total number of cancellation of sessions by cohorts. The values gotten from the attributes would be studied by an algorithm (K-Nearest Neighbor machine learning algorithm) to predict the likelihood of a cohort cancelling a trip.

The number of page\_clicks and total session time by a cohort could suggest a cancellation by a cohort. Similarly, cost of flight, amount spent on hotels, number of sessions, total discount received (rate) on hotel and flight bookings are attributes which could aid the algorithm learn possibilities of cancellation.

A screenshot of a computer

Description automatically generated

The algorithm prediction was at an accuracy of 70%, with an error of just 0.03. Thus, cohorts predicted by algorithm to cancel sessions (greater than or equal to 0.5) would be considered for this perk. Figure 5 below shows the predicted result compared to the actual cancellation result. Figure 6 show a dashboard distribution of predicted cancellation of cohorts with options to toggle and visualize size of cohort based on possibilities of cancellation.

A graph with blue dots

Description automatically generated

A graph with numbers and a bar

Description automatically generated

**2.4 EXCLUSIVE DISCOUNT.**

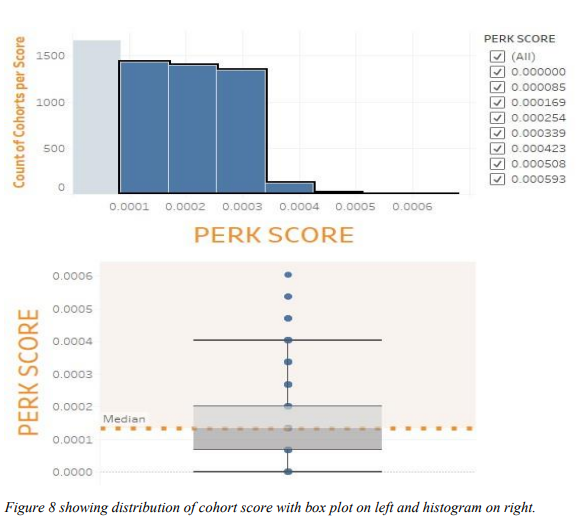
For this perk, the aim was to calculate and ensure that cohorts who may have spent much are considered for discounts. Also, cohorts who may have had huge discounts could be willing to earn more; moreover, it is a marketing strategy to retain these cohorts. The attributes were added up. Attributes used include the real\_cost\_of\_flight, hotel/flight discount amount (in currency), and count of discount received on hotel and flight bookings.

A graph and a diagram

Description automatically generated with medium confidence

**2.5 ONE FREE NIGHT WITH MEAL.**

In this perk, due to correlation of attributes, only the rate of flight booked was the measuring factor.



**2.6 K-MEANS CLUSTERING.**

The k-means algorithm was used to segment the cohorts to perks which better suited. The cohorts were segmented using the metrics calculated in section 2.1 - section 2.5 above.

A diagram of a cluster of dots

Description automatically generated

*Figure 9 showing the distribution of the segmentation done by the algorithm.*

**3.0 SEGMENTATION RESULT, RECOMMENDATIONS, AND FURTHER WORK.**

An option in segmenting the cohorts to perks is assigning cohorts to perks in which they are highly ranked. However, perks like free cancellation requires less cohorts as cancellations are less likely. Therefore, assigning cohorts to perks which they are highly ranked is unsuitable for this project. The recommendation is that cohorts be advised to choose perks for which they are qualified.

The table below present the number of cohorts qualified for each perk.

|  |  |
| --- | --- |
| Perk | Number of Qualified Cohorts |
| Free Hotel Meal | 3013 |
| Free Checked Bags | 3014 |
| No Cancellation Fee | 439 |
| Exclusive Discount | 2982 |
| One Free Night with Meal | 2889 |

Finally, to determine the type of email on which to send to a cohort specifying the qualified perks, further segmentation is required to accurately group the cohorts based on the number and types of perks qualified.

**APPENDIX**

**SQL CODE USED IN THE GENERATION OF THE DATASET**

/\*this code is used to retrieve the necessary attributes from the sessions and users table on which the the date began 4th January 2023, and of which the user had more than seven sessions. The code also perform bit calculations such as scaling, summing attributes and replacing null values.

\*/

WITH ak AS (

select session\_id, user\_id, flight\_booked, hotel\_booked, session\_end, cancellation, session\_start, COALESCE (hotel\_discount\_amount, 0) AS discount\_hotel\_amount, COALESCE (flight\_discount\_amount, 0) AS discount\_flight\_amount, page\_clicks, flight\_discount, hotel\_discount

from sessions

where session\_start >= '2023-01-04 00:00:00' -- choosing the minimum date for the project

),

ak2 AS ( -- summing up attributes required per user whom have had more than 7 sessions from the minimum date mentioned above

SELECT COUNT(session\_id) AS session\_total, user\_id, SUM(session\_end - session\_start) as session\_time, SUM(discount\_hotel\_amount) AS hotel\_discount, SUM(discount\_flight\_amount) AS flight\_discount, SUM(page\_clicks) AS page\_click\_sum, SUM(CASE WHEN flight\_discount THEN 1 ELSE 0 END) AS discount\_flight\_tf, SUM (CASE WHEN hotel\_discount THEN 1 ELSE 0 END) AS discount\_hotel\_tf, SUM(CASE WHEN cancellation THEN 1 ELSE 0 END) AS cancel\_session, SUM(CASE WHEN flight\_booked THEN 1 ELSE 0 END) AS booked\_flight, SUM(CASE WHEN hotel\_booked THEN 1 ELSE 0 END) AS booked\_hotel

FROM ak

GROUP BY user\_id

having count(session\_id) > 7

),

ak3 AS ( -- converting the sessions time spent per selected user for realignment and easy working in mathematics

SELECT

MIN(EXTRACT(epoch FROM session\_time)) AS min\_session\_time,

MAX(EXTRACT(epoch FROM session\_time)) AS max\_session\_time

FROM ak2

),

ak4 AS (

SELECT

user\_id,

session\_total,

cancel\_session,

booked\_hotel,

booked\_flight,

(((EXTRACT(epoch FROM session\_time) - min\_session\_time) / (max\_session\_time - min\_session\_time)) + 1) AS session\_time\_scaled,

hotel\_discount,

flight\_discount,

page\_click\_sum,

discount\_flight\_tf,

discount\_hotel\_tf

FROM ak2

CROSS JOIN ak3

)

SELECT --- specifying every attributes and the pre-calculations performed above to generate the required dataset

h.user\_id,

h.session\_total,

h.cancel\_session,

h.session\_time\_scaled,

h.booked\_flight,

h.booked\_hotel,

h.hotel\_discount,

h.flight\_discount,

h.page\_click\_sum,

h.discount\_flight\_tf,

h.discount\_hotel\_tf,

u.home\_airport\_lat,

u.home\_airport\_lon,

u.gender,

u.home\_airport

FROM ak4 AS h

LEFT JOIN users AS u

USING (user\_id)