Analysis on MINGAR's Fitness Tracker Customers and Their Device Performance

A statistical investigation on new product line customers, and potential wearable device deficiency for MINGAR company

Report prepared for MINGAR by Osar

2022-04-07

Contents

Executive Summary

Introduction

MINGAR company recently intends to focus on developing a new lower price point product line and is interested in the demographics of new customers to target customers more precisely. Moreover, internal concerns about MINGAR's devices potentially have lower performance on darker skin customers than on customers with lighter skin tone. The OSAR Company carries out this report to study the difference between traditional customers and new customers who purchase "Active" and "Advance" line devices and whether device deficiency for sleep scores is based on customers' skin tone.

Key Findings

- As the age of customers increases, the more likely for them to be a customer owning a device from the new line "Active" or "Advance."
- Lower-income individuals are more likely to be customers of the new line products.
- We did not find any evidence supporting that customers' skin tone relates to whether they are new customers or traditional.
- MINGAR's devices have more issues with recording sleep scores for darker skin customers than customers with lighter skin tone. The devices are 5.109 times easier for deep skin tone customers to have performance deficiency. i.e., the device performance needs improvements.
- Elderly customers tend to have more sleep score device issues than younger customers.

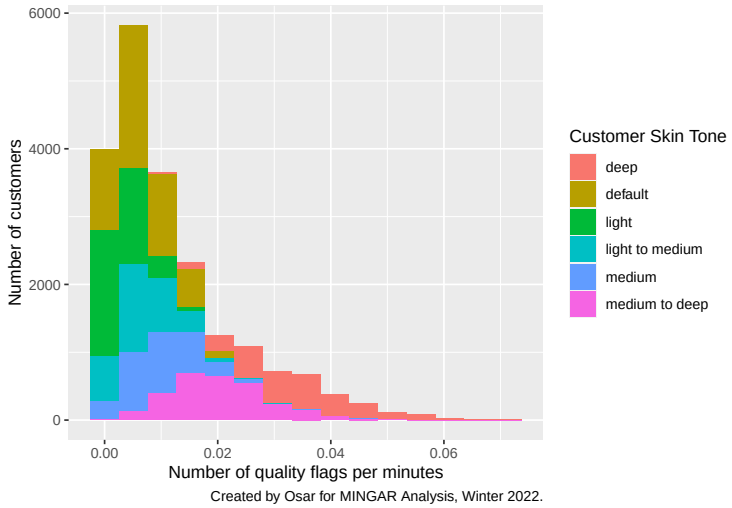

Figure 1: Histogram of number of quality flags per minutes with colour indicating customer skin tones, where default means unclear skin tone.

Figure 1 indicates that only customers with darker skin tones are affected by the device inadequacy problem where the number of quality flags per minute is larger.

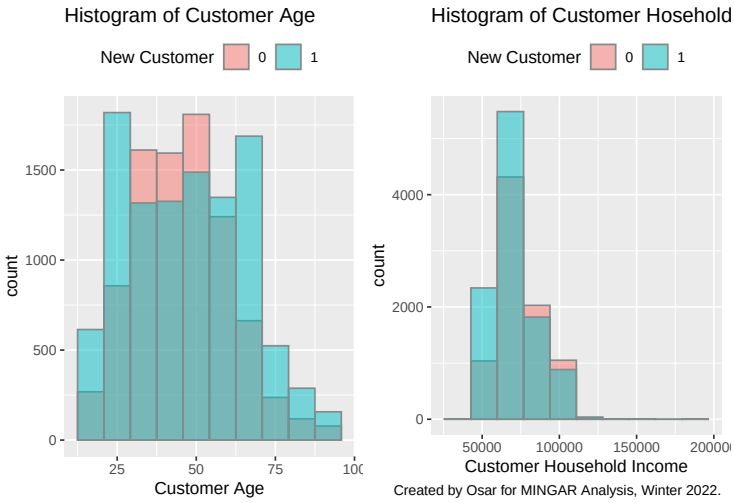

Figure 2: Exploratory Histograms of Customer Demographics

Figure 2 shows the new customers tend to have a wider range of age groups and lower median income than the old customers.

Limitations

- We have used emoji customers used in social media and chats to classify their races and skin tone. However, many customers will keep the default yellow colour instead of changing the emoji to their own skin tones. Some customers might modify their emoji skin tone differently from their actual skin tone. Thus, errors in skin tone estimation can lead to inaccurate results.
- We have merged customers' location data with the Canadian Census data to gather the household median income of each customer. However, the household median income might not accurately represent the customer's income. Instead, the median income data we use now only represents the neighbourhood's median income.

Technical report

Introduction

MINGAR and Bitfit are two leading companies of wearable fitness trackers. MINGAR's traditional products are targeting customers with higher income base. To complete with Bitfit, which has products with a lower price point, MINGAR has recently expanded its production by introducing two more affordable lines, "Active" and "Advance."

This report brings key insights into MINGAR's marketing strategy for the new affordable lines in the Canadian market. Specifically, it analyzes the demographic difference between new and traditional customer, including age, sex, skin tones and income base. The report also investigates the complaints on social media where people with darker skin tones tend to have poorer devices performance.

In the following sections, we use the customer and customer device data provided by MINGAR and the census data to further investigate our questions of interest.

Research questions

- Do customers of the two new affordable lines have a significantly lower median income than customers of our traditional product?
  - Is there other demographics of the new affordable lines' customers different from MINGAR's traditional buyers on average?
- On average, does MINGAR's device have significantly more issues with recording sleep scores for customers with darker skin tone than lighter skin tone?
  - What other factors influence the average performance of MINGAR's devices for sleep scores?

Investigation of MINGAR fitness trackers' customer demographics

Data description and wrangling

The customer dataset given by MINGAR consists of 19045 unique customer demographic with their corresponding device line information. We can access customers' age, sex, skin tone, location, population, and median income.

To better investigate our question of interest, we create a new binary variable called new customer, which indicates whether a customer is using an "Active" or "Advance" line device. We also rescale age and household income to a range of [0,1], which has interpretation and prevents R from optimization problems.

Methods

In the dataset, multiple customers might share the same location id. Thus, observations might not be independent of each other. Since our variable of interest: whether a customer is a new customer, is binary, it would make the most sense to use a generalized linear mixed effect model due to the presence of random effect in location.

In exploratory data analysis, we explore the customer data by plotting histograms of rescaled customers' age and household income.

Figure 3: Exploratory Histograms of Customer Demographics

From the customer age histograms above, we notice that numerous new customers are younger

or older individuals than the median age of our traditional buyers. Unlike our traditional line buyers, whose age distribution is unimodal and centred around 0.3, new line device owners' age distribution is comparably more uniform.

Looking at the histogram of customers' household income, we observe the trend that customers with lower household incomes purchase more of the new affordable lines than the traditional ones.

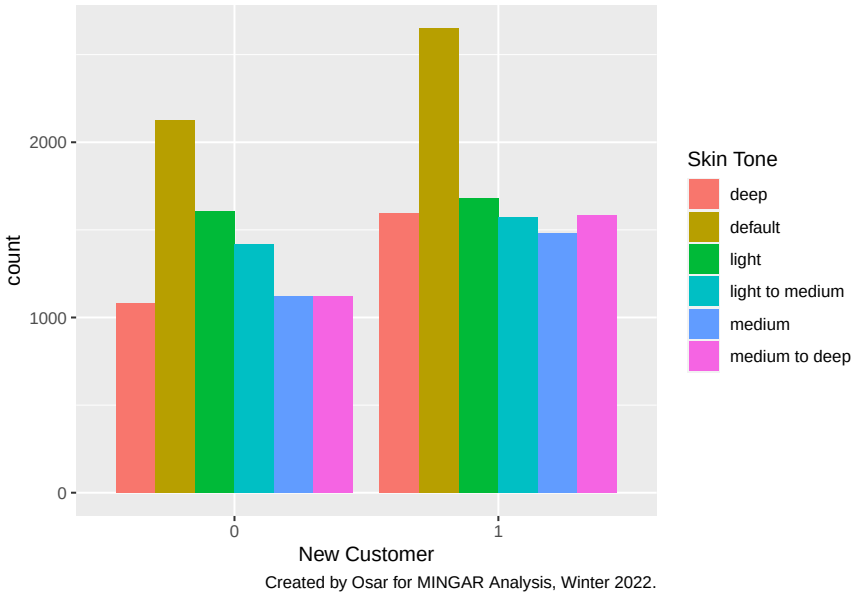

Figure 4: Exploratory Barplots of Customer Skin Tone, where default means unclear.

The above bar plots depict the number of customers with different skin tones, separating traditional customers(left) and new customers(right). We notice that there are more lighter skin tone buyers than deeper skin tone buyers for the traditional lines; While the number of buyers of the new lines is approximately similar for people with all skin tones(excluding default).

Thus, we will consider customer age, household income, and skin tone as predictors of the logistic regression model predicting whether the customer owns a device from the "Advance" or "Active" line.

Results

We build a logistic regression model with **New Customer** being the response variable. The predictors **Customer Age** and **Customer Household Income** are rescaled between 0 to 1. For the predictor **Skin Tone**, we relevel the base level to "default", where customers have no

skin tone information available. We consider "default" as a group of customers that has all possible skin tones. Thus, our final model expression follows

$$Y_{it} \sim Bernoulli(\rho_{ij})$$

$$logit(\rho_{ij}) = \beta_0 + \beta_1 RescaledCustomerAge_{ij} + \beta_2 LightSkinTone_{ij} + \beta_3 LightToMediumSkinTone_{ij} +$$

$$\beta_4 MediumSkinTone_{ij} + \beta_5 MediumToDeepSkinTone_{ij} + \beta_6 DeepSkinTone_{ij} +$$

$$\beta_7 RescaledCustomerHouseholdIncome_{ij} + U_i$$

$$U_i \sim N(0, \sigma^2)$$

where

- Y_{ij} is the response variable, that whether a customer is a new customer for customer i and their location id(CSDuid) j
- U_i is an customer-level random intercept effect based on customer location id, and it follows Normal(0,σ^2)

Considering that multiple customers could have the same location id, they would have the same household income since we joined the customer income from their location's household median income. Therefore, we have set location as an random effect to our model.

Below is the table showing a summary of our model:

Table 1: Summary Statistics of Generalized Linear Mixed Model of Customer Dataset

	Estimated Coefficient	Standard Error	Z Value	P-Value
intercept	0.6581870	0.0957806	6.8718172	0.0000000
rescaled age	0.3800119	0.0658947	5.7669608	0.0000000
deep skin tone	0.0102456	0.0503525	0.2034773	0.8387620
light skin tone	-0.0244955	0.0466346	-0.5252640	0.5993997
light to medium skin tone	0.0209787	0.0478811	0.4381420	0.6612834
medium skin tone	0.0239201	0.0499098	0.4792669	0.6317487
medium to deep skin tone	-0.0311282	0.0500623	-0.6217885	0.5340809
rescaled hoousehold income	-2.6098059	0.3752286	-6.9552424	0.0000000

Table 2: Estimated Odds of Generalized Linear Mixed Model of Customer Dataset

Estimated Odds
1.9312877
1.4623020
1.0102983
0.9758021
1.0212003
1.0242085
0.9693513
0.0735488

Model Interpretation

Looking at the data summary table above, we can see that only the intercept, rescaled age, and rescaled household income are statistically significant predictors of the log odds. In this section, we refer to the estimated odds value for better interpretation.

For the intercept, when a customer has the youngest age and the lowest income among the entire dataset, and with unspecified skin tone, the individual will have estimated odds of 1.9312877. Thus, the probability of this individual being a new customer is $\frac{1.9312877}{1+1.9312877}$, approximately 65.89%.

For rescaled age, if age increases by one unit while holding all other factors constant, the estimated odds will increase by a factor of 1.4623020. Thus, the new probability of an individual being a new customer increases to 73.85%.

And for rescaled household income, if household income increases by one unit while holding all other factors constant, the estimated odds will increase by a factor of 0.0735488. Thus, the new probability of an individual being a new customer decrease to 12.44%.

Accuracy

This section will discuss the coefficients' p-value and their confidence intervals at a 95% confidence level. Also, note that all values in the table below are log odds.

95% Confidence Intervals of Customer Demographics Model Coefficients

	2.5%(lower bound)	97.5%(upper bound)
intercept	0.47261958	0.85115379
rescaled age	0.25093823	0.50924523
deep skin tone	-0.08838393	0.10899544
light skin tone	-0.11588523	0.06692363
light to medium skin tone	-0.07283876	0.11485411
medium skin tone	-0.07384678	0.12179980
medium to deep skin tone	-0.12920353	0.06703963
rescaled household income	-3.35122559	-1.86898603

Observing the confidence intervals above shows that the variable with a confidence interval's range containing 0 is not a significant predictor(since there is no change to log-odds). In our table, skin tone is the only non-significant predictor, which coincides with our results from checking the p-values in the coefficients summary above.

Now, let's look into the device performance analysis by using the customer sleep data.

Device performance analysis

In this section, we want to answer the last two research questions about device performance. Hence, we want to construct the best model that describes device performance patterns during a sleep session and, most importantly, is skin tone of a customer affecting the device performance.

Data description and wrangling

The dataset given by MINGAR consists of 20412 customer sleep records using their wearable fitness device. Each sleep record contains information on quality flags occurrence, the customer id, the duration of sleep time, and the device line this customer possesses. Also, some of the demographics about customers were merged to this dataset from the customer dataset using customer id. Merged demographics include sex, age, location, skin tone and the median income of the customers.

For data wrangling, the age of customers was rescaled from a range of [18, 92] to a range of [0, 1]. Rescaling the age variable allows us to have a more straightforward interpretation of the intercept of the model we build. Otherwise, having an age of 0 as our intercept in the model is not a realistic situation. Moreover, rescaling prevents R from having optimization problems when fitting a generalized linear mixed model, as a model that has large eigenvalues can lead to inaccurate results.

Methods

In the dataset, a subject, which is a customer, might have multiple sleep records. Then, observations within each customer are not independent, but it is sufficient to say each customer is independent. Also, we used quality flags occurrence during the sleep session to access device performance deficiency. The more flags a sleep record has, the worse the corresponding device performed. Quality flags occurrence is a count data, and it can be extended to rates by considering the duration of the sleep session. Therefore, it is reasonable to construct generalized linear mixed models with Poisson response. Flag occurrence is the response variable for our models. We use the log of the duration of the sleep session as the offset to interpret flag occurrence per minute in a sleep session.

In exploratory data analysis, the average number of flags per minute is higher for dark skin tone customers than light skin tone customers. Also, the average number of flags per minute shows a downward trend as the age of customers increases. Customers in some locations seem to have a higher average number of flags per minute. However, it can be due to other confounders since

device performance does not relate to the user's location intuitively. The rest of the variables in the dataset do not seem to relate to flags.

Figure 5: Exploratory histogram and scatter plot of customer locations and age, respectively.

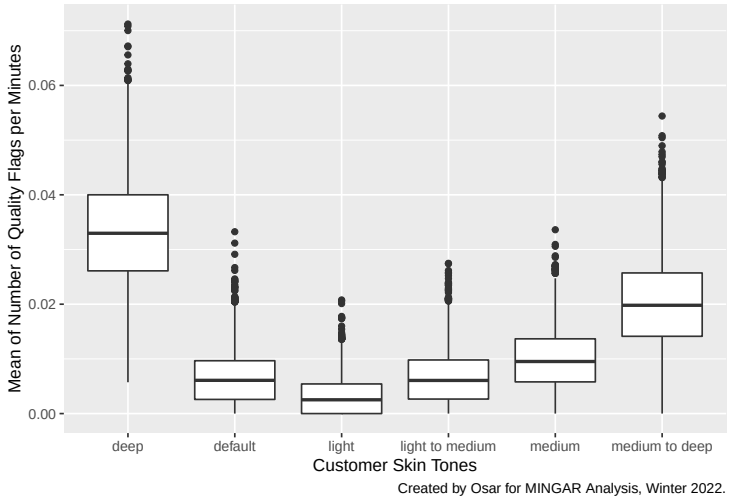

Figure 6: Box plot of customer skin tones, where default means unclear.

Hence, we considered the following variables based on our exploration and the context:
- **Number of quality flags**: Our response variable is the number of quality flags per minute.
- **Customer skin tone**: It is the most important variable we want to investigate regarding the relationship with flags occurrence per minute. They also seem to be related according to our exploration.
- **Product line the device belongs to**: Intuitively, devices produced for different lines might have different performances.
- **Age of the customer**: According to our exploration, the age of a customer seems to relate to flag occurrence.
- **CSDuid**: According to our exploration, customer location might to relate to flag occurrence.
- **Identifier of a customer**: Since each customer has multiple sleep records, the identifier of a customer is a random intercept.

We considered the following models:

- Model 1: The simplest model: A GLMM with customer skin tone as the fixed effect, customer identifier as a random intercept.
- Model 2: A GLMM with customer skin tone, device production line as the fixed effect, customer identifier as a random intercept.
- Model 3: A GLMM with customer skin tone, customer age as the fixed effect, customer identifier as a random intercept.
- Model 4: A GLMM with customer skin tone, customer age and device production line as the fixed effect, customer identifier as a random intercept.
- Model 5: skin tone as the fixed effect, identifier as a random slope depending on skin tone

By comparing models using the Maximum Likelihood test, we have no evidence that the model with customer skin tone and device production line as the fixed effect explains more about the number of flags per minute than the simplest model. We have moderate evidence that the model with customer skin tone and age as the fixed effect explains more about device performance than the simplest model. More complex models are not significantly better than it. Model 5 cannot be constructed due to its complexity and singularity. Considering the complexity and interpretability of Model 5 if we were to resolve the singularity, the model with skin tone and age as fixed effects will be a better choice.

The final model we have concluded utilizes generalized linear mixed model with poisson response and an offset of the duration of sleep record. The response variable is the number of times device performance deficiencies have occured, and the predictors are age, skin tone of customers' emoji as fixed effects, and a random intercept of customer id.

Results

The final model is:

$$Y_{ij} \sim Poisson(\lambda_{ij})$$

$$log(\lambda_{ij}) = \beta_0 + \beta_1 DeepSkinTone_{ij} + \beta_2 LightSkinTone_{ij} + \beta_3 LighttoMediumSkinTone_{ij} +$$

$$\beta_4 MediumSkinTone_{ij} + \beta_5 MediumtoDeepSkinTone_{ij} + \beta_6 Age_{ij} + U_i + log(Duration_{ij})$$

$$U_i \sim N(0, \sigma^2)$$

where

- Y_{ij} is the response variable, that is the number of times there was a quality flag during the sleep session for customer i and sleep record j
- λ_{ij} represents the average number of quality flag in a sleep record j for customer i
- U_i is an customer-level random intercept effect based on customer id, and it follows Normal(0, σ^2)

The offset for this model is the duration of each sleep record j for customer i. Since different sleep records may have different duration of sleep, and the higher the duration, the higher the chance we are going to see a quality flag. Therefore, having offset = duration accounts for the differences in duration time.

Table 4: Summary Statistics of Generalized Linear Mixed Model of Customer Sleep Dataset

	Estimated Coefficient	Standard Error	Z Value	P-Value
intercept	-5.0151754	0.0114065	-439.678487	0.0000000
deep skin tone	1.6315323	0.0118875	137.247506	0.0000000
light skin tone	-0.7585258	0.0184573	-41.096300	0.0000000
light to medium skin tone	0.0179059	0.0152772	1.172066	0.2411706
medium skin tone	0.4194710	0.0141539	29.636351	0.0000000
medium to deep skin tone	1.1320728	0.0124651	90.819517	0.0000000
rescaled age	-0.0477731	0.0175228	-2.726335	0.0064042

Model Interpretation

The beta 1 - 6 represents the increase in expected log counts of quality flag for a sleep record for every one-unit increase in that predictor holding all other variables constant, including the random constants. Consider taking exponentiation for each coefficients estimate so we refer to the expected counts of quality flag in this section.

For instance, for a one unit increase in Age, a $e_6^{\beta} = e^{-0.477} = 0.9534$ increase in expected counts of quality flag for a sleep record will occur holding all other predictors constant.

For categorical variable skin tones, notice that the base line is default skin tone and notice that the coefficients can also be interpreted as log risk ratios. i.e: $\beta_1 = log(\frac{\lambda_{deep}}{\lambda_{default}})$, and after exponentiation, $e^{\beta_1} = \frac{\lambda_{deep}}{\lambda_{default}}$ is the risk ratio. That is, the percent change in the quality flag for deep skin tone percentage change relative to default skin tone. Therefore, $e^{\beta_1} = e^{1.631} = 5.109$ implies the expected counts of quality flag we get for a deep skin tone customer is 5.109 times than the expected counts of quality flag we get for a default skin tone customer.

The intercept represents the expected counts of quality flags when the customer is aged at 18 with a default skin tone.

The random intercept represents the different intercept for different customers. This means there's going to be multiple responses per customer, and these responses will depend on each customer's baseline level.

Accuracy

In table 1, variables skin tone and Age all are statistically significant(p-value < 0.05) except light to medium skin tone. We have strong evidence to reject the null hypothesis that the coefficient is equal to zero, meaning the predictor we are considering does have a effect in determining

the expected log counts of quality flags. Moreover, the light to medium skin tone predictor is having a p-value of $0.24117 > 0.05$, implying that this predictor is not statistical significant when determining expected log counts of quality flags.

At a confidence level of 95%:

95% Confidence Intervals of Customer Sleep Model Coefficients

	2.5%(lower bound)	97.5%(upper bound)
intercept	-5.0375985	-4.9928532
deep skin tone	1.6082184	1.6548840
light skin tone	-0.7948036	-0.7224384
light to medium skin tone	-0.0120849	0.0478270
medium skin tone	0.3916932	0.4472110
medium to deep skin tone	1.1076218	1.1565376
rescaled age	-0.0821796	-0.0133814

While considering confidence interval in table 2, any interval containing $e^0 = 1$ indicates the failure to reject the null hypothesis. The null hypothesis again is that the corresponding coefficient estimate is zero. (i.e: predictor have no effect in determining the expected log counts of quality flags). From table 2, our result is the same as our p-value. All variables indicates the failure to reject the null hypothesis except light to medium skin tone.

Contexual Interpretation

We observe that having medium to deep skin tone and deep skin tone have larger effect on getting numbers of quality flags during a sleep record. Such deviations across skin tones of the customer confirmed our observations from plots done before and are confirmed by our statistical model chosen. Hence, MINGAR's devices are performing poorly for users with darker skin as quality flags increases as the customer have deeper skin. Furthermore, our model shows an positive relationship between age and the expected counts of quality flags. Therefore, MINGAR's devices also performs poorly for elders than young customers.

Discussion and Conclusion

Overall, we have found that the new line product targets customers typically younger or older than the average traditional line users. We also found that increasing an individual's income negatively impacts the probability of this individual is an "Active" or "Advance" line tracker owner. However, we could not find any evidence supporting a statistically significant relationship between the skin tone of customers and whether or not they are new customers.

Investing on customer sleep data, this technical report has also explored another potential question: whether MINGAR's wearable devices perform poorly with customers having darker skin tones. We have proposed several statistical models to analyze the data provided by the company thoroughly, and these predictors include sex, age, location, and skin tone of the customer. Ultimately, we have found that only the age and skin tone of the customer have an association with the device deficiencies in sleep score. All skin tones and ages of customers except light to medium skin tone have been shown to be significant predictors the number of quality flags with sleep scores. This suggests MINGAR should consider improving the product and fixing the device issues. Or else, this device issue can be viewed as a racist problem by the customers. We have also found devices have more deficiencies for older customers than younger customers.

Strengths and limitations

In this report, we take the random effect of grouping into account when we construct our models. By doing so, we can handle the dependence between each observations so that our model would not be biased.

For our first model, we used the census data to obtain the household income of a customer based on their neighbourhood's median income. This estimation is not always accurate since customer's income could be an outlier in their neighbourhood.

For the second model, the fundamental assumption of the Poisson regression model is that the mean and variance of our response are equal. The equal mean and variance assumption has been violated in our model. This leads to a limitation of the analysis where the result might be overdispersion if the variance is greater than the mean. Consequently, our standard errors might be falsely small, meaning the model probably has falsely small p-values, leading us to choose a more complicated model.

Then, when we explore the sleep data, there are 2654 observations with quality flags of 0. We should consider the zero-inflated Poisson model to deal with the frequent zero observations for the next steps.

Lastly, the dataset contains noticeable missing values for customers' emoji modifiers. Each

missing value is treated as the default skin tone, which may lead to inaccuracy in determining the actual skin tone of the customers, increasing the uncertainty of our conclusion for the sleep scores relationship.

In regards to suggestions for future considerations, the Osar Group would like to have more explore more into device-related data such as device price, and device functionalities. With the goal of increasing Canadian market share with competitor Bitfits, comparison data between MINGAR's to Bitfits' products would be informative to explore and might come up with a more reliable decision with more device-related data being considered. Additionally , In line with the idea of analyzing the data with a model that has a good balance in accuracy and easy interpretation, the Osar group continuously improves the interpretability of our models with statistically accurate models values at a more significant level, as well as makes the model clear, understandable by more potential models and consider more random or fixed effects, and finally make more comparisons between models with reasoning to improve transparency in statistical practice.

Consultant information

Consultant profiles

Wendy Yusi Cheng. Wendy is a junior consultant with Osar Group. She specializes in model interpretation and reproducible analysis. Wendy is experienced in identifying and executing new and potential products and services. Wendy earned her Bachelor of Science, specializing in Statistics and majoring in Computer Science in 2024.

Haoze Huang. Haoze is a junior consultant with Osar Group. He specializes in data visualization and machine learning. Haoze has experience collecting and analyzing data on established and prospective customers and competitors. Haoze earned his Bachelor of Science, specializing in Computer Science and majoring in Statistics in 2024.

Qingyi Liu. Qingyi is a senior data scientist with Osar Group. She specializes in data cleaning, preparing and report generation. Qingyi has experience working with large data analytics companies such as Tableau and Oracle and converting data into actionable insights by modelling future outcomes. Qingyi earned her Bachelor of Science, specializing in Statistics and majoring in Computer Science in 2020.

Yuchen Zeng. Yuchen is a junior consultant with Osar Group. She specializes in data analysis and statistical communication skills. Yuchen has experience in SQL analytics using Databricks in her internship. Yuchen earned her Bachelor of Science, specializing in Computer Science and majoring in Statistics in 2024.

Code of ethical conduct

Osar Group will follow the ethical guidelines for statistical practice. We will commit to have professional integrity of data and methods for the analysis. We take responsibility to protect customer privacy by assigning unique id to distinguish customers instead of customer name, age instead of exact birth date, and subdivision id instead of address postal codes. We promise to rigorously investigate the data and make all the assumption and methods transparent to reach the conclusion. We will keep our promises and honor our commitments on the below ethical statements.

1. Responsibility to society: maintain objectivity and make rigorous statistical conclusion avoiding personal bias, and provide assistance in misleading information.
2. Responsibility to Data and Methods: report details for statistical methods used and any additional information for decision making to ensure reproducibility of the analysis

3. Responsibility to clients confidentiality and non-disclosure agreement: all the customer information will be protected and all confidential information (non public) and analysis conclusion may not be referenced without client's permission.

References

[1] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[2] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[3] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48, https://doi:10.18637/jss.v067.i01

[4] Hadley Wickham (2021). rvest: Easily Harvest (Scrape) Web Pages. R package version 1.0.2. https://CRAN.R-project.org/package=rvest

[5] Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=polite

[6] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

[7] Wood, S.N. (2017) Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC. https://cran.r-project.org/web/packages/mgcv/citation

[8] Fitness tracker info hub. (n.d.). https://fitnesstrackerinfohub.netlify.app/

[9] Postal code conversion file. (2016). Toronto. https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file

[10] Code of ethical statistical practice. https://ssc.ca/sites/default/files/data/Members/public/Accreditation/ethics_e.pdf

[11] von Bergmann, J., Dmitry Shkolnik, and Aaron Jacobs (2021). cancensus: R package to access, retrieve, and work with Canadian Census data and geography. v0.4.2. https://cran.r-project.org/web/packages/cancensus/vignettes/cancensus.html

[12] Hadley Wickham and Evan Miller (2021). haven: Import and Export "SPSS", "Stata" and "SAS" Files. R package version 2.4.3. https://CRAN.R-project.org/package=haven

[13] Jared E. Knowles (2020). eeptools: Convenience Functions for Education Data. R package version 1.2.4. https://CRAN.R-project.org/package=eeptools

[14] Dmytro Perepolkin (2019). polite: Be Nice on the Web. R package version 0.1.1. https://CRAN.R-project.org/package=polite

[15] Hadley Wickham and Evan Miller (2021). haven: Import and Export "SPSS", "Stata" and "SAS" Files. R package version 2.4.3. https://CRAN.R-project.org/package=haven

[16] Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

[17] Hao Zhu (2021). kableExtra: Construct Complex Table with "kable" and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra

Appendix

Web scraping industry data on fitness tracker devices

To collect data for MINGAR's wearable fitness trackers, we used web scrapping to extract data for each tracker. The data is provided by https://fitnesstrackerinfohub.netlify.app/. Permissions for web scrapping were granted in the robots.txt. The following is the web scrapping R code we used to scrape the data:

```
url <- "https://fitnesstrackerinfohub.netlify.app/"
target <- bow(url,
        user_agent = "qingyi.liu@utoronto.ca for STA303/1002 project, representing the
        ↪  Osar group",
        force = TRUE)
html <- scrape(target)
device_data <- html %>%
  html_elements("table") %>%
  html_table() %>%
  pluck(1)
```

The above code uses R libraries *tidyverse*, *polite*, and *rvest*. All of them are referenced under the reference section. As displayed in the code, we provided a user agent string that clarifies our intentions and provides our contact email. We requested data at a reasonable rate only when necessary for this analysis.

Accessing Census data on median household income

Census Mapper API provides median household income data. The information is publicly available on Canadian census website. The Osar Group only retrieved the data using the API key for this analysis and adequately referenced in the report.

Accessing postcode conversion files

The University of Toronto portal provides postal code conversion data. As members of the Osar Group are students from the University of Toronto, we retrieved the data with legal access credentials. This data is not available to the public. We only used the information we needed for the analysis, which are postal codes and corresponding ids. Also, the data source link is referenced in the report in addition to the licensing agreement.