Data Analysis Report for MINGAR Fitness Wearables

Research on the characteristics of MINGAR's new line customers and device performance for users with darker skin

Report prepared for MINGAR by Newsim

2022-04-11

Contents

Executive summary

This report aims to provide professional consultation addressing several questions from our client MINGAR. As the Wearable market grows rapidly, MINGAR released "Active" and "Advance" lines to strengthen their competitiveness in the market. In order to inform the marketing teams' strategy in the Canadian market, the first part of our study focuses on analyzing the personnel characteristics of MINGAR's newer products buyers. According to the users' feedback, the devices are performing worse for dark-skin users. Hence the second part of our study aims to investigate the interaction between racial background and device performance particularly with respect to sleep scores.

Key Findings of the study are summarized below:

- *Compares with the traditional products, the newer and more affordable "Active" and "Advance" products are more attractive to older people and people living in regions with lower household median incomes*
- *Customers' age for original products are mainly from 30 to 60, whereas new products attracts customers from a wider age range from 20 to 80, with an average age 1.45 higher than that of original customers*
- *Customers are 1.45 times more likely to buy the "Active" and "Advance" products with every 74 years increase in their age*
- *Average household median income for buyers of "Active" and "Advance" products is 4354.08 Dollars less than that of original customers*
- *For every 153,680 dollars increase in the median household income, customers living within that category area are 0.93 times less likely to buy the "Active" and "Advance" products*
- *No evidence shows other features of customers, such as skin tone and sex, have contribution towards distinguishing buyers of traditional products and newer products*
- *In investigating the issues relating to device performance, particularly with respect to sleep scores, the devices perform poorly on dark-skin users, and the devices performance is not as satisfactory within the younger customers compares to the older customers*
- *In investigating the flags per minute in sleep scores recorded by the devices, compared with other skin-tone emoji users, dark-skin emoji users are the only group with greater than 0.05 flags per minute*
- *The mean of default in sleep scores per minute recorded by the devices of users aged below 45 years old is higher than 0.013, whereas that of users aged over 65 years old is lower than 0.013*
- *No evidence shows device line have a significant effect on the device performance*

Limitations of the study are summarized below:

- *Lacking of data regarding to each customer's household income might lose some level of accuracy when analyzing the plausible features for potential new customers. The median household income in the neighborhood which a customer is living in cannot fully represent the true consumption level of them.*
- *Not including customers that are unwilling to provide gender identification in the study may produce biased estimates that lead to invalid results.*
- *Issues with the viability of using customers' skin tone emoji to represent their races and ethnicities when analyzing whether the MINGAR devices perform poorly for users with darker skin. There are many default skin tone emojis, and one can choose a different skin tone emojis based on one's preference. Hence, using emojis to represent a customer's race and ethnicity may lead to invalid results.*
- *Data for some specified group is too small, so it may cause bias in our analysis.*

Key Visualizations:

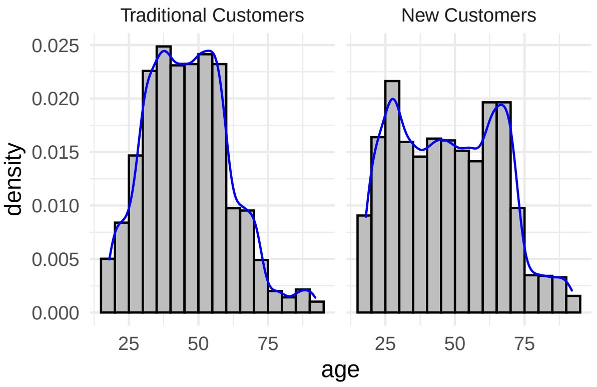

Figure 1.Histogram of the Distribution of Age of New Customers V.S. Traditional Customers

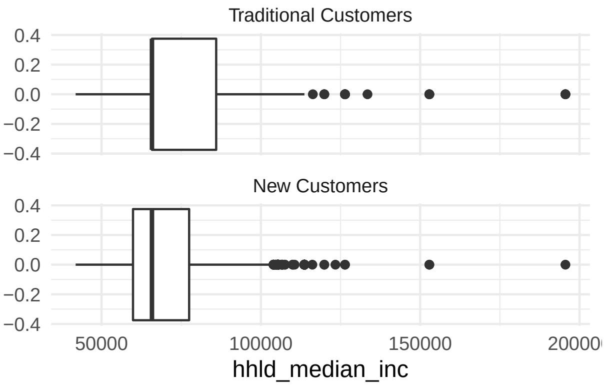

Figure 3. Boxplot Displaying Household Median Incomes of New Customers V.S. Traditional Customers

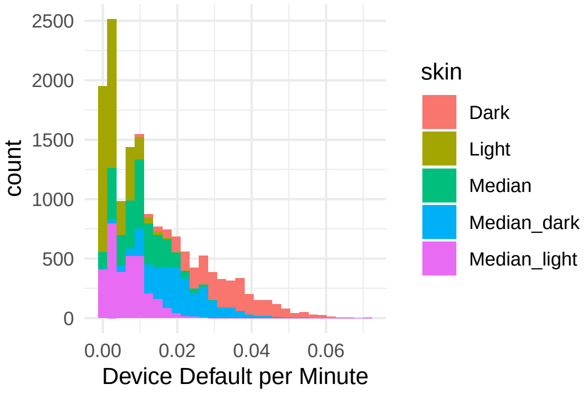

Figure 4. Histogram of Device Default per Minute by Skin

Technical report

Introduction

Wearables are a growing market, and this report provides customer and device analysis to help MINGAR to better understand their customers and to solve their current issues. This report will cover the topic of identifying and distinguishing the buyers of the newer and more affordable products ("Active" and "Advance" line), determining whether the devices' performance can be affected by the user's skin tone and determining whether the sleep score is reliable for users with darker skin. To get a better understanding of the Canadian customer market for MINGAR's newer and more affordable products, the Newsim consultants examine the features of MINGAR's current customers of the newer products, with a focus on the household median income and age, to predict the significant features of their future new customers, and how are these features contributes to the customers' purchasing choice. Additionally, the MINGAR's products aim to be inclusive, therefore this report discusses the differences in device performance across ethnic customers and highlights the potential problems inherent in if MINGAR's devices are performing worse for dark-skin users. A summary of ethical issues and Newsim's Code of Ethical Conduct is included.

Research Questions

- *What are the characteristics associated with the buyers of the MINGAR's newer and more affordable "Active" and "Advance" products, and how are these characteristics distinguish these new customers from MINGAR's traditional customers?*
- *Can the device's performance be effected by the user's skin tone? Is the sleep score reliable for users with darker skin?*

Data Wrangling

For the dataset used in research question 1, we first created a new column in customer dataset that stores the information on whether a customer is using the traditional products or the "Active" and "Advance" products. Then, we created a new binary variable that refers to the type of customers; the value stored for the New Customers is one, and the value stored for the Traditional Customers is 0. The binary information is stored in the new column named "New_Customer." The binary variable allows us to build a model that predicts whether a characteristic of a customer contributes to making them more likely to purchase the newer and more affordable product lines.

To investigate the effect of various variables on sleep scores, we integrated the model with general customer information and sleep data. We eliminated the missing values for validity and operability for this complete dataset. The racial background is an essential parameter in our report. Thus, we excluded the default skin tone input. Our primary concern during the sleep session is the quality flag. However, variations in sleep session time can alter the number of flags, so we add a new variable for flags per minute to our data.

Analysis in MINGAR's Customers of the Newer and More Affordable Lines

Exploratory Data Analyses

Table 1: Summary of Ages of Traditional Customers and New Customers

Customer Type	Observations	Minimum Age	Maximum Age	Mean Age	Median Age
Traditional Customers	8476	18	92	46.50614	46
New Customers	10569	18	92	47.95307	47

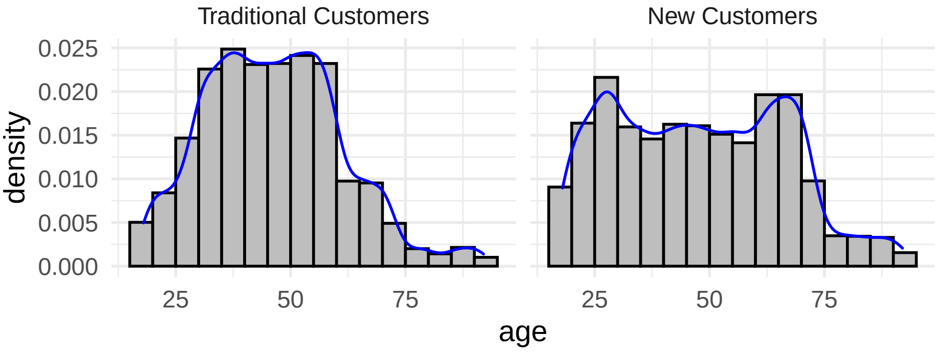

Figure 1: Histogram of the Distribution of Age of New Customers V.S. Traditional Customers

Observed from Table 1, the New Customers are 1.45-year older than the Traditional Customers in average. Fitting a histogram (Figure 1) comparing age between the Traditional Customers and the New Customers, we can also observe that the majority Traditional Customers are between 30 to 60 years old, whereas the age of the New Customers is relatively more evenly spread from 20 to 80 years old, with a greater amount between 25 to 30 years old and between 60 to 70 years. Therefore, we will consider age as a fixed effect in our model to test whether it is significant in predicting customers' purchasing choice.

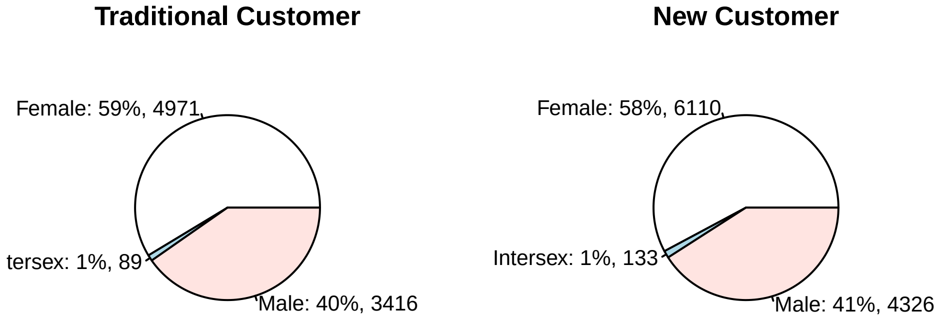

Figure 2: Pie Chart regarding genders for New Customers and Traditional Customers

From Figure 2, observed that the majority buyers of both the Traditional products and the New products are females, resulting in 59% of Traditional Customers are female, and to 58% of New Customers are female. Given the similarity of frequencies of sex in both the Traditional Customers group and the New Customers group, we should be caution when use sex as predictor in our model.

Table 2: Summary of Household Median Income of Traditional Customers and New Customers

Customer Type	Count	Minimum Household Median Income	Maximum Household Median Income	Mean Household Median Income	Median Household Median Income
Traditional Customers	8476	41880	195570	73168.02	65829
New Customers	10569	41880	195570	68813.94	65829

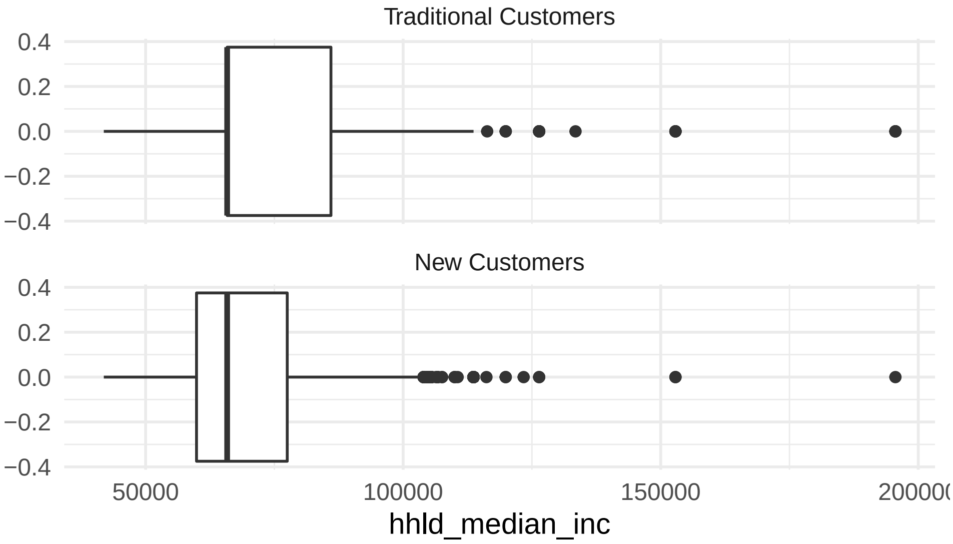

Figure 3: Boxplot Displaying Household Median Incomes of New Customers V.S. Traditional Customers

From Table 2, observed that the mean of household median income of the New Customers is 4354.08 Dollars less than the mean of household median income of the Traditional Customers. Fitting a box plot(Figure 3) comparing household median income between the Traditional Customers and the New Customers, we can also observe that both groups shows similarity in distribution, with the majority household median incomes are between 50k to 100k. The household median income distribution of both groups are tend to be left skewed with some outliers on the right. Also, observed that the household median income of the Traditional Customers has smaller interquartile range, which indicates the Traditional Customers has lower variation in household median income. Therefore, we will consider the household median income as a fixed effect in our model to test whether it is significant in predicting customers' purchasing choice. To distinguish the variation from neighbourhood to neighbourhood, we will use the Census Subdivision Boundary (CSDuid) as our random effect.

Model Construction and Interpretation

We are attempting to figure out who are the buyers of the "Active" and "Advance" product line, and how are they vary from the customers for the traditional product line. Note that "skin" might also be a factor that will contribute to helping us distinguish the New Customers. As "skin" satisfies the independence assumption, we will treat it as a fix effect in our model. Since we can only access data of the median income of customers from each region, we are considering including a random effect "CSDuid" in our model. And noticing that our response variable is a binary variable representing a individual being a new customer or not, we should use a generalized linear mixed model (GLMM) to fit our data. After checking, the assumptions for GLMM are satisfied, namely the subjects are independent since customer individuals are independent to each other, random effect comes from normal distribution, and appropriate link function is used. Based on the analysis of the data before, the first model we fit uses median income from the customer's region, customer's age, customer's gender and customer's skin tone as fixed effect, and CSDuid as random effect, and due to the significant range difference of median income and age, we rescaled these two variables to be in the range from 0 to 1. Then, originated from the first model, we fitted several reduced models in the lack of certain fixed variables, with models' details presented in table 3. In order to identify the best model, we used likelihood ratio test on several combinations of models, the results of the tests are summarized in table 4.

Table 3: Model Fixed Effects, Random Effect and AIC value

	Model	Fixed effects	Random effect	AIC
AIC	mod1	rescaled median income, sex, skin	CSDuid	25706.1957196162
AIC	mod2	rescaled median income, rescaled age, skin	CSDuid	25674.1774626176
AIC	mod3	rescaled age, sex, skin	CSDuid	25716.1338469602
AIC	mod4	rescaled median income, rescaled age, sex	CSDuid	25666.9880760614
AIC	mod5	rescaled median income, rescaled age	CSDuid	25666.0207011279
AIC	mod6	rescaled median income, rescaled age, sex, skin	CSDuid	25675.1033064465

Table 4: Likelihood Ratio Tests between models

Comparison	p_value	preferred model
mod1 vs mod6	8.78809622414567e-09	mod6
mod2 vs mod6	0.215008420706123	mod2
mod3 vs mod6	5.38919889719943e-11	mod6
mod4 vs mod6	0.864848574107299	mod4
mod5 vs mod2	0.870382822648223	mod5

Comparison	p_value	preferred model
mod5 vs mod4	0.219519868550852	mod5

Table 5: Estimates and 95% Confidence Intervals of Odd Ratios

	Estimate	95% CI
Baseline	1.93	(1.62, 2.33)
Rescaled Customers' Age	1.45	(1.28, 1.65)
Rescaled Household Median Income	0.07	(0.04, 0.15)

According to the test result, our final model will be

$$new\_customer \sim rescaled\_customer\_age + rescaled\_household\_median\_income + (1 \mid CSDuid)$$

Table 5 interpret the output which we are interested in our final model. Observed from the first row of Table 5: the baseline odds is 1.93, which indicates that as we fixed age to be 18 and median income to be 41880, the odd ratio of a customer being a new customer is 1.93:1. The 95% confidence interval is (1.62, 2.33), which indicates that we are 95% confident that the true odd ratio of a customer being a new customer is between 1.62:1 and 2.33:1. Therefore, an 18-years-old customer who lives in a neighborhood with median income of 41880 dollars are 1.93 times more likely to be a new customer.

Then, observed from the second row of Table 5 that the odd of rescaled customers' age is 1.46, which indicates that as we fixed the median income to be 41880, for every unit increased in age, the odd of a customer being a new customer increased by 46%(with 95% confidence interval of (28%, 65%)).

Finally, observed from the third row of Table 5. That the odd of Customer Income Rescaled is 0.07, which means as we fixed the age to be 18, for every unit increased in the median income, the odd of a customer being a new customer decreased by 93%(with 95% confidence interval of (85%, 96%)).

Note that the above age and median incomes are normalized, which means they are rescaled to values from 0 to 1. Therefore, 1 unit of age is not in years but in 74 years, and 1 unit of median income is not in 1 dollar but in 153,680 Canadian Dollars.

To conclude, the youngest customers living in a lowest income neighborhood are 1.93 times more likely to purchase the newer and more affordable products. We can observe that with a minor increase in age, the odds of customers buying the newer products increase by 46%, and with a small increase in income, the odds of customers buying the newer products decreased by 93%. The trend is saying the newer and more affordable products should mainly target older people who live in a region with lower median income.

Investigation of the interaction between racial background and sleep scores

Exploratory Data Analyses

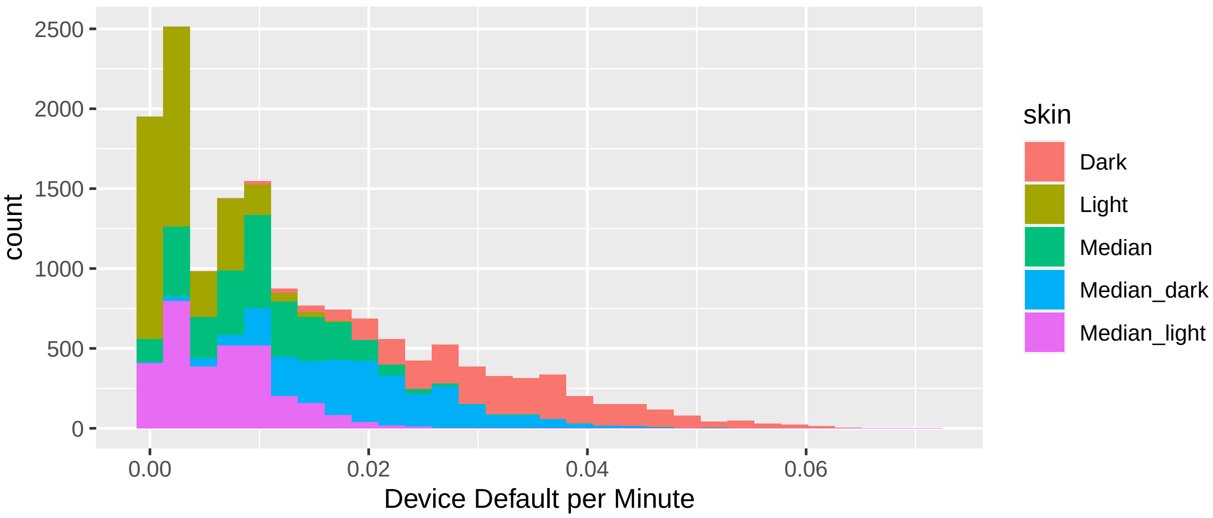

Figure 4: Histogram of Device Default per Minute by Skin

The histogram (Figure 4) of the average device default with varied skin tones of users is shown. The number of flags that occur on the device every minute is displayed. According to observations, the average number of flags for light-skin users ranges from 0 to 0.015, indicating that the equipment outperforms all others. The average device default for both median light and median skin users varies from 0 to 0.025, with the devices performing second only to those of light skin users. The device performance for medium-dark skin and dark skin users is insufficient. Device flags with a frequency greater than 0.025 per minute are almost all theirs. Devices with an average of 0.05 flags per minute are all owned by users with dark skin. This indicates that the devices detect anomalies three times per hour on average. As a result of preliminary analysis, the instruments of darker skin users are not as effective as the devices of lighter skin users. Of course, a follow-up examination is required to confirm a more precise result.

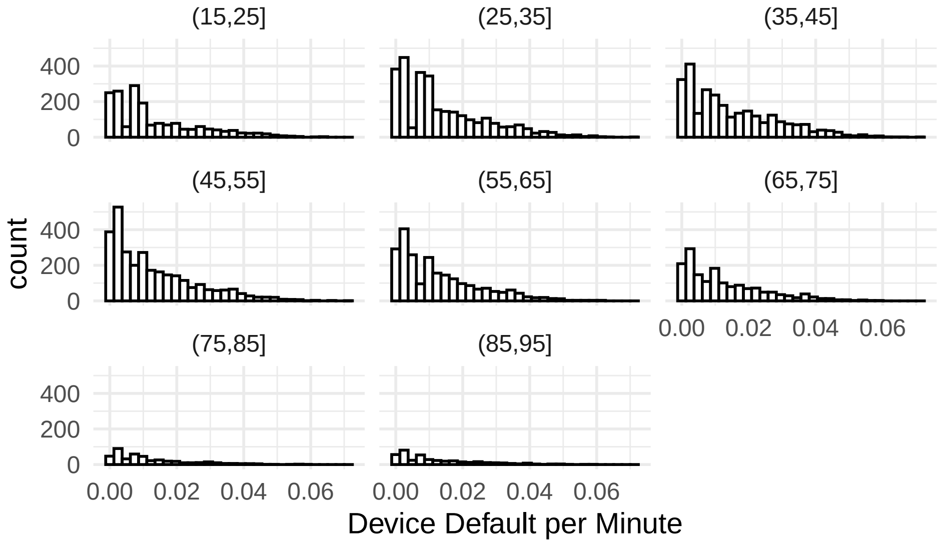

Figure 5: Histogram of Device Default per Minute by Age Group of Users

Table 6: Compare Mean and Variance of Device Default per Minute within Each Age Group

Age Groups	Mean	Variance	n
(15,25]	0.0135249	0.0001773	1773
(25,35]	0.0139924	0.0001700	2886
(35,45]	0.0151769	0.0001783	2764
(45,55]	0.0132827	0.0001598	2976
(55,65]	0.0129979	0.0001451	2346
(65,75]	0.0126473	0.0001391	1641
(75,85]	0.0122802	0.0001383	447
(85,95]	0.0123470	0.0001447	410

Figure 5 reveals a fair amount of variability in groups of age of the users. We observed the most of users in each age group have low average device default, but a few users aged from 15-35, 25-35 and 35-45 years old have high device default per minute (which is higher than 0.05). And the Table 6 of the mean and variance of average device default within each group showed that users aged in 35-45 years old has an outstandingly high mean (0.015 flags per minute) than other age groups. Compare to the senior age group (65-95 years old), the younger age group of 15-35 years old has a higher mean and variance, where younger users have a mean above 0.013 and a variance above 0.00017, but senior users have a mean below 0.013 and a variance below 0.00015. According to the exploratory data analysis of variable age, the device is more effective to the senior users than the younger users. However, the further examination is required to have a precise result about whether age influenced the effectiveness of the device.

Table 7: Compare Mean and Variance of Device per Minute with Each Device Line

Device Line	Mean	Variance	n
Active	0.0141850	0.0001460	1499
Advance	0.0141252	0.0001668	6745
iDOL	0.0077066	0.0000334	88
Run	0.0130839	0.0001620	6911

In the table 7, we compare the mean and variance of the average device default with different device lines. It is quite clear that the mean and variance of default per minute of "Active", "Advance" and "Run" device line are close, but the mean of "iDOL" line is much lower than the other 3 device line. This may be because the number of "iDOL" line users is too small. We were unable to clarify whether the different device line would be a factor influenced the device's effectiveness based on the current analysis, hence we need further examination to make the result more precise.

Model Construction and Interpretation

To investigate if the user's skin tone affects device performance and to account for other potential influences, we will construct four models of flags with varying combinations of skin tone, age, and device line. Afterwards, evaluate the goodness of fit of each two statistical models using Likelihood Ratio Tests(LRT). Utilizing LRT reveals whether adding parameters to our model is advantageous or whether we should continue with our simpler model.

During the parameter selection procedure, we constructed generalised linear mixed-effects models. Flags are considered as the response variable for model fitting, while skin, age, and device line are regarded as fixed effects. The customer id appears to be a grouping factor in the data. Since we need to maintain data independence, we interpret it as a random effect. To avoid the possibility of differing durations having an impact on the number of flags, we set offset to log(duration). The model is Poisson distribution since the number of flags is determined per unit time.

$$model_1 : flags \sim (1 \mid cust_{id})$$

$$model_2 : flags \sim skin + (1 \mid cust_{id})$$

We first constructed model 1 with no fixed effects and then model 2 with only skin as a fixed effect. Using LRT, we obtain a p-value of $2.2 \times e^{-16}$, indicating substantial evidence against the null hypothesis; hence, the complicated model should be used. The extremely low p-value also reveals that the user's skin tone considerably impacts the device's performance. The model's skin tone is a crucial consideration.

$$model_3 : flags \sim skin + age + (1 \mid cust_{id})$$

For model 3, the rescaled factor age is added. Age is rescaled by the same method as the previuos section. In addition to skin tone, we expect to determine if age influences the number of flags. After performing LRT on models 2 and 3, we obtain a p-value of 0.007629, considered a relatively low p-value. As a result, there is strong evidence against the null hypothesis, and we select the more complex model3.

$$model_4 : flags \sim skin + age + device + (1 \mid cust_{id})$$

We want to examine if the device line variation influences the number of flags. We added the device line to model 3 to build model 4. When we used LRT to compare models 3 and 4, we received a p-value of 0.3268. As this value is greater than 0.05, we believe there is no evidence against the null hypothesis and choose model3, a simpler model. According to testing, the difference in device level does not affect device performance.

$$flags \sim skin + age + (1 \mid cust_{id})$$

After conducting a series of model comparisons, we eventually established that the frequency of flags is affected by users' varying skin tones and ages. To determine how much different skin tones and age of users affect the number of flags, we apply a summary function to generate estimates for each component.

In the summary, the estimate slope for skin light is -2.39, it indicates that, comparing to skin dark, the number of flags appear on skin light users' device is 2.39 less per minute. The estimate slope for median light skin users is -1.61, indicating that the number of flags occur are 1.61 less than dark skin users per minute. The estimate slope for median skin users is -1.21, shown that the number of flags per minute occurred 1.21 less than dark skinned users. The estimate slope for median dark skin users is -0.50, suggests that the quantity of flags that occurred per minute is 0.5 less than that of dark-skinned users. In addition, the estimate slope for age is -0.05, which specifies that, for each unit increase in age, keeping other variables constant, the number of flags decreases by 0.05 times per minute.

From the above estimates, we can discover that the performance of the devices of the darker-skinned users is generally worse than that of the lighter-skinned users, and thus their sleep score reliability is also worse. The difference in age also has an effect on device performance. The younger the user, the worse the device performance is compared to the older users.

Assumption Check

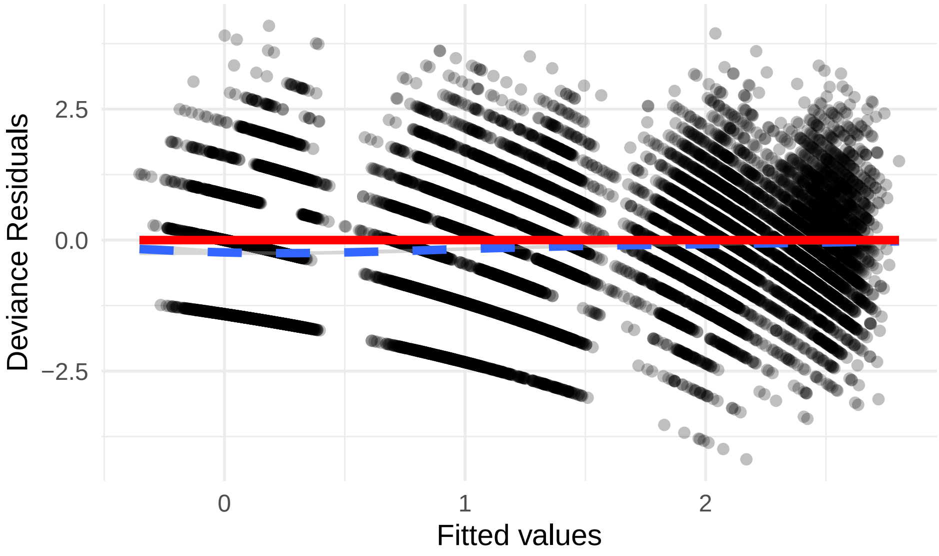

Figure 6: Deviance Residual Plot of Possion Model of Device Default by the skin and ages

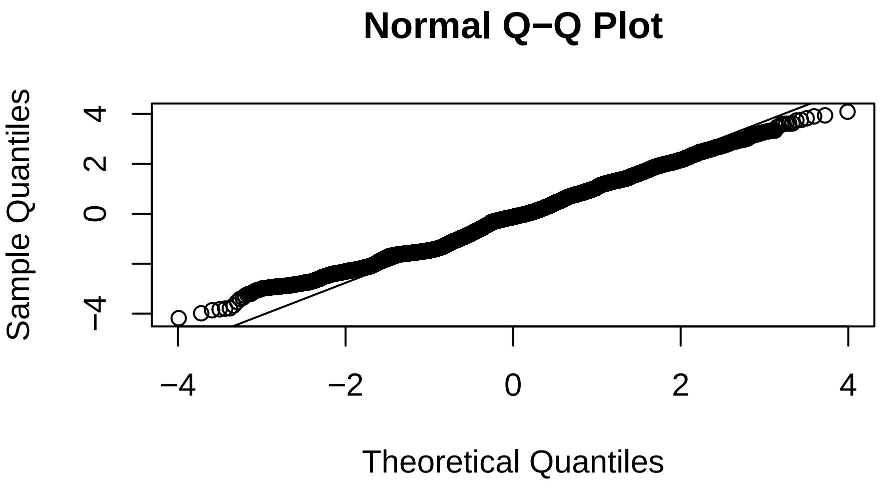

Figure 7: QQ Plot of Possion Model of Device Default by the skin and ages

Plot (Figure 6) of the deviance residuals versus predicted responses for the final model shows that the deviance residuals are most likely centered about the zero residual line. The QQ plot (Figure 7) shows that the residuals only deviate slightly and all points of quantiles lies closely to straight line, hence the regression is fairly robust to departures from normality. Both of plots appear the model to be good for a poisson fit.

Discussion

Our analysis above concluded that the youngest customers living in the lowest-income neighbourhoods are 1.93 times more likely to purchase the newer and more affordable products. Fixed the household median income to 41880, the odds of customers buying the "Active" and "Advance" products are increased by 46% with every 74 years increase in their age. Additionally, fixed the age to 18 years old, for every 153,680 dollars increase in the median household income, the odds of a customer living within that category area buying the "Active" and "Advance" products are decreased by 93%. There is no evidence that other customers' features, such as skin tone and sex, have contributed to distinguishing buyers of traditional products and newer products. Hence, the analysis suggests that newer and more affordable "Active" and "Advance" products should mainly target the older people who live in a region with a relatively lower median income.

To investigate the interaction between racial background and device performance, we found that the number of defaults that appear on light-skin-emoji users' devices is 2.39 less per minute compared to the dark-skin-emoji users' devices. The number of flags per minute recorded on the median-light-skin emoji users are 1.61 less than the dark-skin-emoji users, and the number of device defaults per minute recorded on the median skin users is 1.21 less than the dark-skin-emoji user. The number of flags per minute recorded on the median-dark-skin-emoji users is 0.5 less than the dark-skin-emoji users. In addition, for each unit increase in users' age, keeping the other variables constant, the number of flags per minute decreased by 0.05. No evidence shows that device lines significantly affect the device's performance. Hence, the analysis suggests that the devices perform, particularly concerning sleep scores, poorly on dark-skin users. Also, the device's performance is not as satisfactory among the younger customers compared to the older customers.

Strengths and limitations

The Newsim Team has an advantage in interpreting statistical analysis, data visualizations and Model development. All models and visualizations are interpreted clearly and readable by different audiences, and all models and visualizations are easy to reproduce based on these interpretations within our report. The strength of our model construction is that we consider the possible random effects to avoid possible grouping effects, which gives us a more appropriate model and more accurate coefficient estimates. All the results are interpreted within the report. We succeeded in delivering information to help MINGAR adjust its marketing strategies and solve the problems related to device performance.

Newsim team also faced a few limitations throughout the analysis. Due to the lack of data regarding the real income of each customer, Newsim team decided to estimate it using the median household income in the region where the customer lives, which may fail to reflect the actual features of the potential new customers. Also, we excluded the customers that are unwilling to provide gender identification; this may produce biased estimates that lead to invalid results. Moreover, customers' preference in using skin tone emoji cannot fully represent their races and ethnicities because some users may not choose the same skin tone emoji as themselves. This would cause an issue of finding valid evidence to determine whether the devices' performance can be affected by the user's skin tone and whether the sleep score is reliable for users with darker skin. Lastly, the data for some specified groups, like the "iDOL" line device user, is too small, so the analysis of that variable may produce biased estimates that lead to invalid results.

To address these constraints, MINGAR might send out an optional customer survey in which customers can volunteer to provide information, such as their racial backgrounds, to understand its customers better and better analyze the correlation between the device performance and customers' skin tone. MINGAR should disclose all information and data collected through the survey, and customers must be able to withdraw from the survey at any time.

Consultant information

Consultant profiles

Yutong Chen. Yutong is a junior consultant with Newsim Analytics. She specializes in data visualization and statistical communication. Yutong earned their Bachelor of Science, Majoring in Mathematics and Economics, Minoring in Statistics from the University of Toronto in 2022.

Kexin Li. Kexin is a junior consultant with Newsim Analytics. She specializes in data visualization and reproducible analysis. Kexin earned their Bachelor of Commerce, Specialist in Finance & Economics, Majoring in Economics, Minoring in Statistics from the University of Toronto in 2022.

Jiamei Shu. Jiamei is a junior consultant with Newsim Analytics. She specializes in data processing, model building and text editing. Jiamei earned her Bachelor of Science, Majoring in Mathematics and Statistics, Minoring in Economics from the University of Toronto in 2022.

Xuanzhong Zhao. Xuanzhong is a junior consultant with Newsim Analytics. He specializes in data processing and software development. Xuanzhong earned his Bachelor of Science, Specialist in Computer Science, from the University of Toronto in 2022.

Code of ethical conduct

- Throughout the study by Newsim, we did not use any summary of the data that could be misleading, and we dedicated to make sure that all the assumptions and limitations relating to our study are addressed, as well as presenting all of the information we gathered to our client.
- Throughout the study by Newsim, we only used methods and data that are valid, relevant and appropriate. We did not impose our prejudice upon the methods and data. We also included comments and justifications for each important step, closely abode the criterion to produce valid, interpretable, and reproducible results.
- In our report, we presented the sources of our data used, and fully disclosed the data processing and transformation procedures. We also included the methodology we used regarding data handling.

References

> [1] R Core Team (2021). 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. 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. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36

> [8] Hadley Wickham and Dana Seidel (2020). scales: Scale Functions for Visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

> [9] Fitness tracker info hub. (2022). Retrieved 11 April 2022, from https://fitnesstrackerinfohub.netlify.app/

> [10] Population Density. (2022). Retrieved 11 April 2022, from https://censusmapper.ca/

> [11] Postal code conversion file | Map and Data Library. (2022). Retrieved 11 April 2022, from https://mdl.library.utoronto.ca/collections/numeric-data/census-canada/postal-code-conversion-file

> [12] Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr

> [13] Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Appendix

Web scraping industry data on fitness tracker devices

To see if we are allowed to scrape industry data on fitness tracker devices, we first utilize the bow function to provide a User-Agent string that includes our email address and purpose of using the data. This makes our intentions clear, and the website host can contact us conveniently. According to the response, the path is scrapable for this user agent. Then we made a table with the device data, applied the necessary component to our study, and cited the website in the reference section.

Accessing Census data on median household income

To get income data from the Canadian census, we first signed up for the website to get a public API that provides the data. Then we install the package of cancensus and then fill our API key into the option function to display the result. Since we already have a public API, we should use that instead of scraping. We appreciate any content we maintain and never claim it as our own, and we always give credit to the source in the reference section.

Accessing postcode conversion files

Since the zip code data we want to use is not the same as the one we usually know starts with a capital letter, we need the conversion of zip codes to match the data in our income data. The median household income data we downloaded in the previous part is for 2016. To keep the data consistent, we chose the 2016 census data to match median income data. On the website, we logged in with UTORid and accepted a license agreement to access the data and download data. We then placed it in our data raw folder instead of directly onto our GitHub. Eventually, we applied a portion of the data in our final submission.