MGT 6203 - Data Analytics for Business

Group Project Final Report

Predicting Reported Thermal Preferences of Smartwatch Users

Team #10

Jack Li

Joanna Stefaniak

Chengxuan Feng

Max Midlash

Silvia Vangelova

Github Repo: https://github.gatech.edu/MGT-6203-Spring-2024-Canvas/Team-10

Content

Introduction

Background & Problem Statement

Cultivating a comfortable environment with ideal temperature conditions is desirable for several reasons. Firstly, it promotes relaxation and well-being by maintaining a pleasant atmosphere. Extreme temperatures can also pose health risks, such as heat stroke or hypothermia, and can exacerbate conditions such as asthma or allergies. By maintaining a comfortable temperature, individuals can safeguard their health and minimize the likelihood of temperature-related ailments. Moreover, a comfortable temperature encourages better sleep quality, as the body can regulate its internal temperature more effectively, leading to more restful and rejuvenating sleep. There are also positive effects regarding cognitive function and productivity. This can be particularly beneficial in work or study environments, where productivity and focus are crucial. Lastly, a comfortable thermal environment enhances social interactions and quality of life. These are all things that are desirable for people individually and for society as a whole.

This background is the motivation behind our team's research. The goal of this project is to predict users' thermal preferences across a diversity of indoor and outdoor spaces within Singapore's urban environment. This allows us to investigate what factors contribute to comfortable or uncomfortable thermal environments and can be useful in many ways. Our team's initial hypothesis is that meteorological data such as temperature and wind speed will be most relevant in predicting a user's thermal preference. Additionally, we suspect that data relating to a user's activity level will also be relevant. By researching what factors contribute to self-reported thermal preferences, more informed decisions can be made in fields such as city-planning to promote thermal comfort.

Business Justification

Addressing the challenge of predicting thermal preferences presents a unique opportunity for the Urban Redevelopment Authority (URA), particularly within a city-state with limited land like Singapore. This project has several significant implications and potential benefits from a socio-economic and operational perspective.

Understanding thermal comfort in disparate areas is instrumental in contributing towards urban planning decisions and policies. It aids in determining the most suitable areas for constructing housing projects and other infrastructure. The process of identifying these areas proactively mitigates potential discomfort for residents, which directly correlates to improved living standards and overall quality of life [1].

There are financial aspects to consider as well. By predicting thermal preferences, the URA can implement cost-effective building designs that require less energy for heating or cooling, thus reducing energy consumption and potentially leading to substantial cost savings in the long run. Furthermore, a suitable thermal environment can attract more businesses into the region and boost property valuations, thereby increasing tax revenues for the city.

In conclusion, understanding what factors contribute to a person's thermal preferences can provide the URA, and by extension, the Singapore government, with invaluable data to assist in their sustainable urban planning efforts, bettering lives while optimizing cost efficiencies.

Literature Review

As coined by the United States Environmental Protection Agency, "Urban heat islands" occur when cities replace natural land cover with dense concentrations of pavement, buildings and other surfaces that absorb and retain heat. This effect increases energy costs (e.g. air conditioning), air pollution levels, and heat-related illness and mortality. Most severely, climate change will likely lead to more frequent, severe, and longer heat waves during summer months (Figure 1).

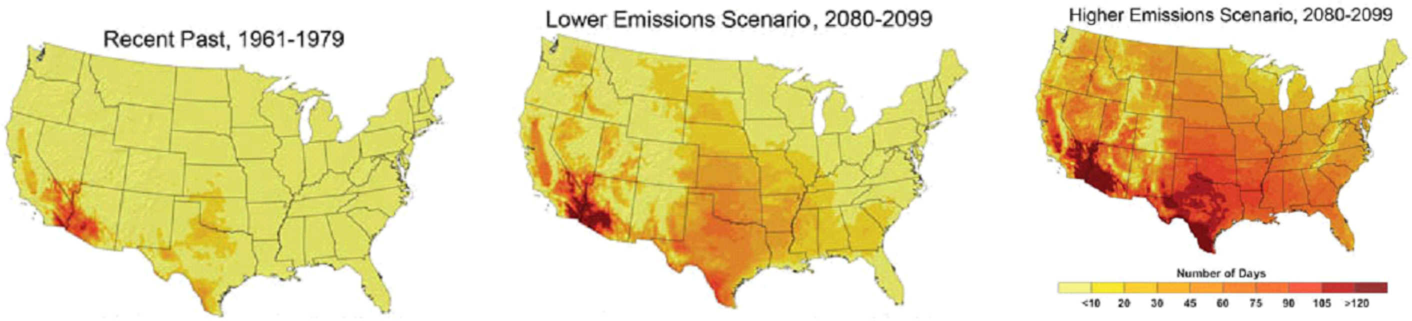

Recent Past, 1961-1979
Lower Emissions Scenario, 2080-2099
Higher Emissions Scenario, 2080-2099
Number of Days
<10
20
30
45
60
75
90
105
>120

Figure 1: Number of days with high heat emissions in USA

Urban places typically experience higher temperatures compared to their rural surroundings [2]. It has been reported that the annual mean air temperature of a city with one million or more people can be as much as 1.8-5.4°F (1-3°C) warmer than its surroundings. Interestingly, on a clear, calm night with negligible winds, this temperature difference can be as much as 22°F (12°C) [2]. Moreover, roof and pavement surface temperatures can be 50-90°F (27-50°C) hotter than the ambient air on a hot sunny summer day, while shaded or moist surfaces in rural surroundings remain close to air temperatures [3].

In conclusion, understanding patterns of heat islands and perceived temperature is crucial for urban planning initiatives. Having predictive models on areas less susceptible to heatwaves allows for more informed decisions in space allocation for different activities. This could lead to more efficient, livable, and sustainable urban environments. Consequently, it is essential that urban planners factor in these elements in neighborhood design, urban regeneration, and new development planning. Considering the implications of activities in specific locations could help mitigate the effects of urban heat islands, contributing to better urban air quality, energy efficiency, and overall resident well-being.

Methodology

Data Preparation

The data used in this analysis comes from a Kaggle competition using smartwatch data from users in Singapore [4]. We have two types of data sets: user response logs combined with their smartwatch readings, and meteorological measurements from specific weather stations in Singapore about humidity, wind, temperature, and rainfall. One important characteristic that both types of datasets have in common is that they are both time-series datasets. Their date-time columns have been converted to time objects. In addition, we have a dataframe with the IDs and locations of all weather stations in the city from which the measurements were taken.

1. User Response Data

The user response dataset consists of two parts - one for training, with 1,149,136 rows, and one for testing, with 996,429 rows. The test to training ratio is 0.88, meaning 46% of the total user response data can be used to compare the quality of different types of models. To do so we further divide the training dataset into two parts - 80% of the data is used for model fitting and 20% is used for model cross-validation. After we have chosen the best of each type of model we can then compare them based on their performance on the remaining test dataset.

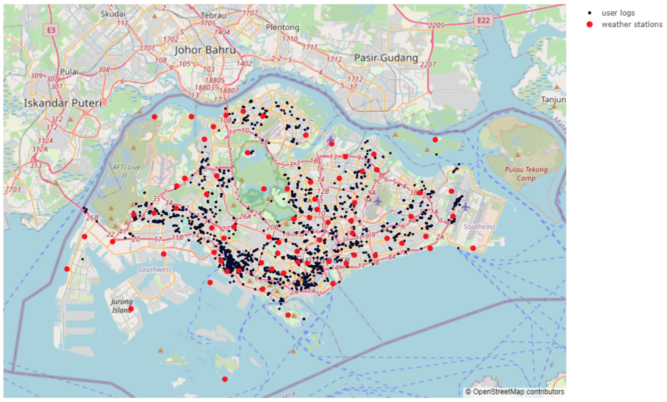
Despite having many observations, most of the data are NaN or empty strings. This is because the dataset is composed of three types of data - physiological measurements (smart watch readings), spatial characteristics, and micro-survey logs. The physiological measurements are frequent at regular intervals. When the user logs a survey response, the physiological measurements stop. Longitude and latitude data are then stored and used to calculate spatial characteristics at that location using the Urbanity framework [5]. So, a row from the user survey response dataset would have either physiological measurements or spatial characteristics and microsurvey logs.

We are interested in the thermal preference response. We hypothesize that some variables that influence a person's perception of temperature are current heart rate and current activity level, such as total steps taken or total distance walked 10 minutes prior to recording the thermal preference response. For each user, we take smartwatch readings 10 minutes prior to each user's survey response. From these data we take average heart rate and total distance walked and add them as variables to our models.

In the training dataset, out of all 996,429 rows, only 4,900 contain survey response logs. 1,943 users wish it were cooler, 304 wish it were warmer and 2,652 wish no change. Most users were located indoors: Indoor - Class: 294 logs, Indoor - Home: 2,200 logs, Indoor - Office: 872 logs, Indoor - Other: 660 logs, Outdoor: 568 logs, Transportation: 306 logs.

2. Meteorological Data

We also have meteorological data about temperature, rainfall, humidity and wind at various different weather stations. Measurements in all datasets start from October 9th 2022 and end July 3rd 2023. Temperature, humidity and wind are measured 1,347 times a day on average, or in other words, about every minute. Rainfall is measured 276 times a day on average, or in other words, about every five minutes. Naturally, the timestamps of weather station measurements differ from the time stamps of user survey response logs. All weather stations have days in which no weather data is reported. Each user’s data reported the latitude and longitude at the time of response. We were able to use this latitude and longitude data to find the weather station located closest to each user. Then, by leveraging the time of response combined with the nearest weather station, we were able to determine the various meteorological data most relevant to the user.



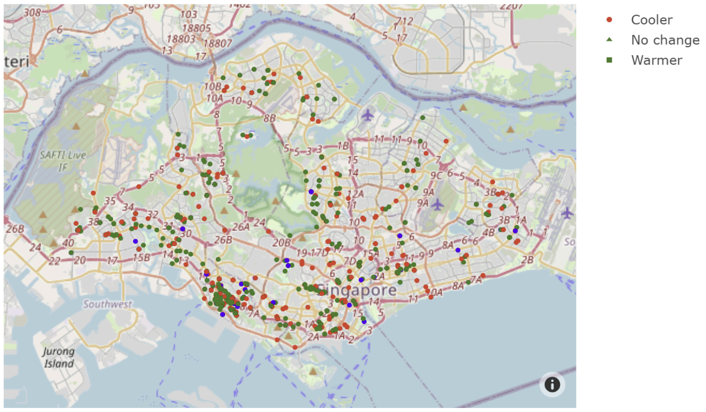
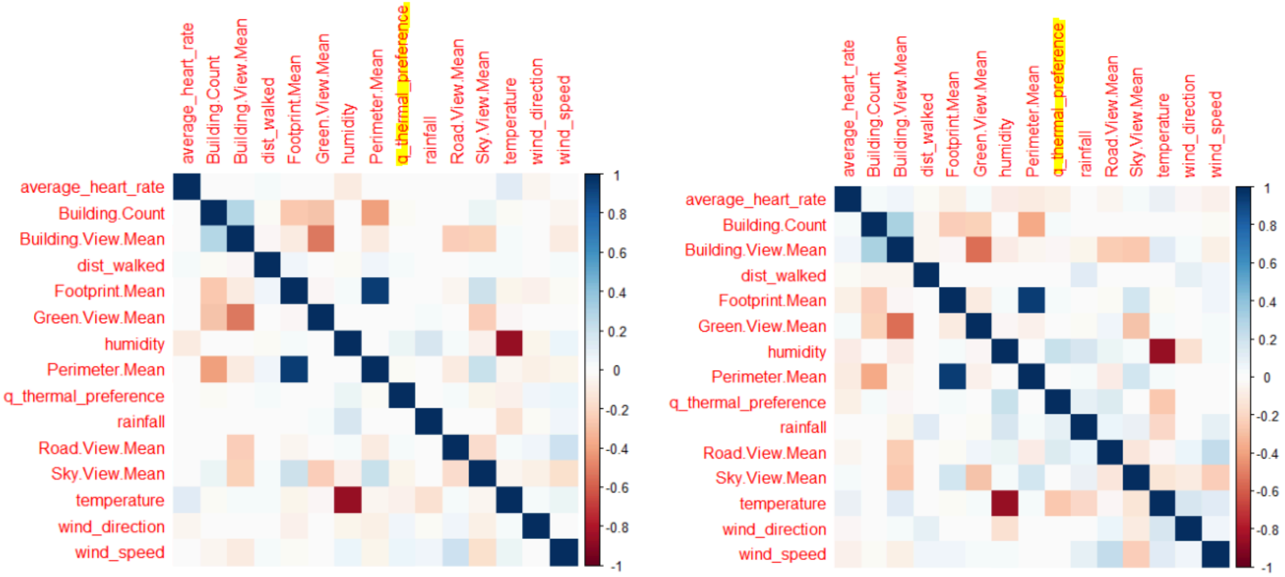


Figure 2: Map of weather stations and user responses   *Figure 3: Map overlay of users’ thermal preference*

Preliminary Analysis

As part of the preliminary analysis, we plotted users’ thermal preferences on top of a map of Singapore to see if there were any patterns that could be visually seen. This plot can be seen in Figure 3 above. There were some pockets of similar user responses clustered together, which we hypothesized could be as a result of localized weather patterns. This further supported our decision to investigate the impact of meteorological factors in the data.

We then plotted variables from the combined dataset as correlation matrices to see if there was any correlation between variables. There were some correlations between variables, such as a strong negative correlation between the humidity and temperature, as well as a negative correlation between Green.View.Mean (density of green space) and Building.View.Mean (density of buildings). These are expected correlations that will not likely affect the approach of our modeling. (See Figure 4 below)



left plot - indoor measurements *right plot - outdoor measurements*

Figure 4: Correlation Matrices

As we continued our exploratory data analysis, we created plots of the input data (Figure 5) to investigate the values that each variable can have, as well as the frequencies of each value.

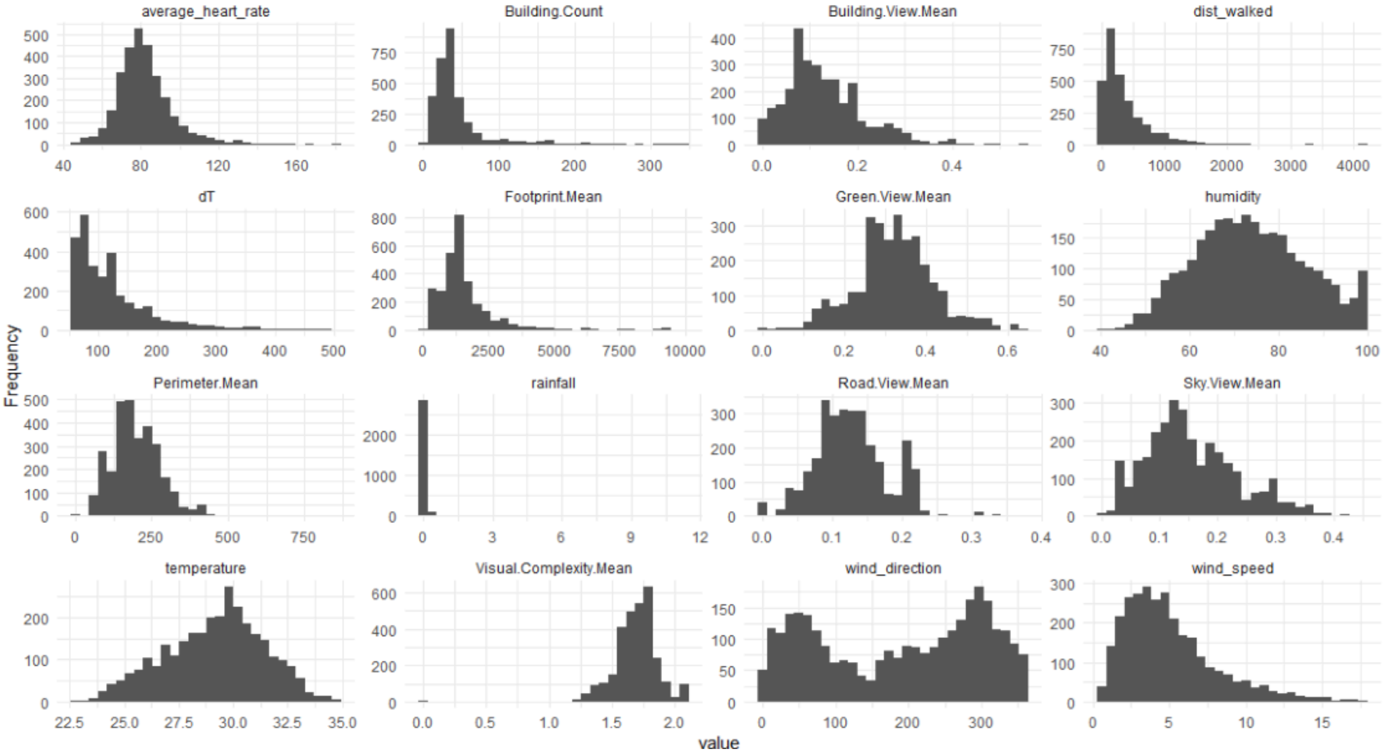


Figure 5: Input Variable Plots

We then plotted the response variable (thermal preference) to examine how many of each response value were in the dataset. As we can see in the figure below on the left, there are a very low number of "warmer" responses compared to the "cooler" and "no change" responses. This relatively small number of responses makes our data biased, which will require care when further analyses are done. In addition, Figure 7 shows the time frame the survey data is collected for reference. This serves as a remark that the data could be biased due to the collection dates being only from Oct-Jan or Apr-June.

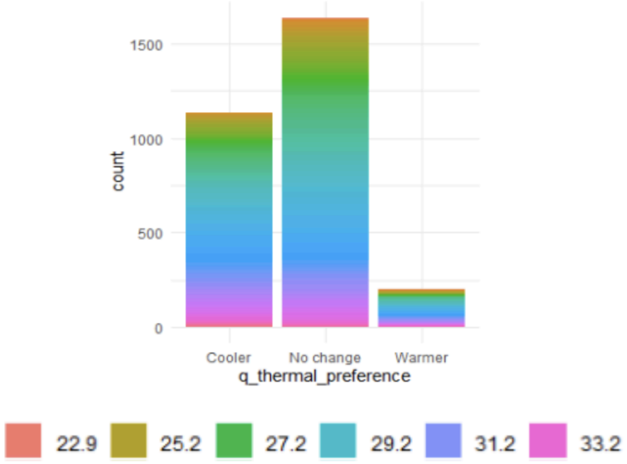


Figure 6: Thermal Preference Frequencies

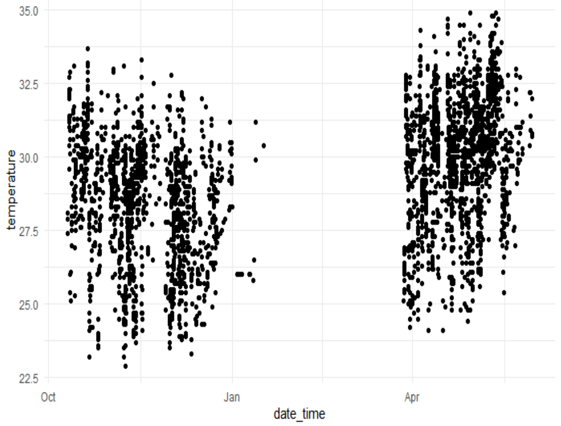


Figure 7: Temperatures in each season of is_winter

Variable Definitions

Before we explore the models used, it is important to clarify what the input variables used represent. The temperature variable represents the air temperature in degrees Celsius, while the humidity variable represents the percent relative humidity. Is_outdoor is a boolean indicator variable created where outdoor responses are true and indoor responses are false. Likewise, is_winter and is_day are also boolean indicator variables representing winter and daytime respectively. Building.View.Mean represents the building view index, which is an index representing the density of buildings of a location. Similarly, Green.View.Mean represents the density of green space, while Sky.View.Mean represents the amount of visible sky from a location. Visual.Complexity.Mean is an index that indicates how varied an environment is in terms of building architecture, height, design, etc. Lastly, Road.View.Mean is an index representing the availability and distribution of roads at a given location. These variables appear in many of our models, so a clear understanding of their meaning is very important. A detailed description of the variables can be found in [5] Table 1.

Models

After the extensive data processing and preparation, the first model we created was a logistic regression model predicting thermal preference using our processed variables as features. As logistic regression models return a probability between 0 and 1, we framed the logistic regression model in terms of whether or not a user's thermal preference was "cooler" or not. The negative result in this case could be that they either preferred no change or warmer. Specifically, we used the glm() function in R with the family=binomial(link="logit") parameter. This model was a valuable starting point, as we were able to see initially which variables were statistically significant, and which weren't. For example, some of the statistically significant variables included Sky.View.Mean (the amount of visible sky), temperature, and binary indicator variables we created to indicate if it is outdoors and if it is currently the winter season. These are variables that we initially expected to be relevant. However, the model showed that certain other variables that we expected would be relevant actually were not statistically significant in this model, namely average heart rate, rainfall, and wind speed.

From this initial model, we selected only the variables that were statistically significant to include in a second logistic regression model. Again, this model used the same glm() function in R as the previous model. We then trained this model on the training dataset and iterated through various cutoff values to classify the predictions. The cutoff value that yielded the highest accuracy was a value of 0.6. We then used this cutoff value for our model when evaluating it on the test dataset, which will be discussed further in the results section below.

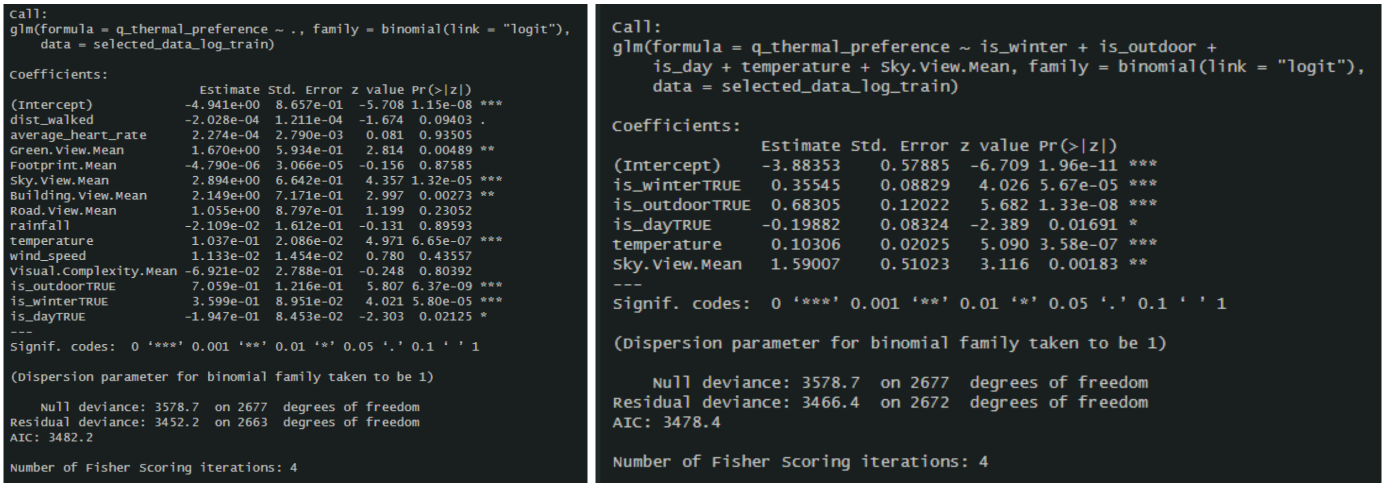
The next models used were tree-based classification models, including decision trees with pruning and ensemble methods such as random forests. To create the tree-based classification model, we used the rpart() function within R to fit this model. Initially we used all variables as inputs into the model to see which variables would actually be useful in classification. We then pruned this tree model to avoid overfitting on our training data.

We then used the randomForest() function from the randomForest library in R to create a new model. As with the other models, we initially used all of our chosen variables as inputs to the model and allowed the function to find the most predictive model. Random forest models use many different classification trees, splitting on different subsets of predictors each time. In our random forest model, we used 100 instances of tree models in our random forest. Further discussion regarding the results of our models will follow in the subsequent results section.

Results

Models Predicting "Cooler" Response

The first logistic regression model we created used 2,970 observations total, of which 80% were for training and 20% for testing. This model predicts whether a user's thermal preference was "cooler". The output of this model before and after removing insignificant variables is shown in Figure 8 below. In this model, the variables that were statistically significant in predicting the thermal preference were the indicator variables is_winter, is_outdoor, and is_day. Other significant variables were the temperature and Sky.View.Mean variables. The coefficient of the is_day variable was negative, indicating a negative correlation between it being daytime and a user preferring cooler thermals. The rest of the variables had positive coefficients, meaning that larger values of these inputs increase the odds of a user preferring cooler weather. When a cutoff value of 0.6 is used, the model showed a 62% accuracy in correctly predicting this response. This relatively high accuracy may be misleading however, as this model only predicts if the response was "cooler" or not, as it is unable to predict a warmer response. Additionally, this model had a pseudo-R2 McFadden value of only 0.03205 which indicates a weak explanatory power of the model.



```
Call:
glm(formula = q_thermal_preference ~ ., family = binomial(link = "logit"),
    data = selected_data_log_train)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -4.941e+00  8.657e-01  -5.708 1.15e-08 ***
dist_walked           -2.028e-04  1.211e-04  -1.674  0.09403 .  
average_heart_rate     2.274e-04  2.790e-03   0.081  0.93505    
Green.View.Mean        1.670e+00  5.934e-01   2.814  0.00489 ** 
Footprint.Mean        -4.790e-06  3.066e-05  -0.156  0.87585    
Sky.View.Mean          2.894e+00  6.642e-01   4.357 1.32e-05 ***
Building.View.Mean     2.149e+00  7.171e-01   2.997  0.00273 ** 
Road.View.Mean         1.055e+00  8.797e-01   1.199  0.23052    
rainfall              -2.109e-02  1.612e-01  -0.131  0.89593    
temperature            1.037e-01  2.086e-02   4.971 6.65e-07 ***
wind_speed             1.133e-02  1.454e-02   0.780  0.43557    
Visual.Complexity.Mean -6.921e-02  2.788e-01  -0.248  0.80392    
is_outdoorTRUE         7.059e-01  1.216e-01   5.807 6.37e-09 ***
is_winterTRUE          3.599e-01  8.951e-02   4.021 5.80e-05 ***
is_dayTRUE            -1.947e-01  8.453e-02  -2.303  0.02125 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3578.7  on 2677  degrees of freedom
Residual deviance: 3452.2  on 2663  degrees of freedom
AIC: 3482.2

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = q_thermal_preference ~ is_winter + is_outdoor + 
    is_day + temperature + Sky.View.Mean, family = binomial(link = "logit"), 
    data = selected_data_log_train)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -3.88353    0.57885  -6.709 1.96e-11 ***
is_winterTRUE   0.35545    0.08829   4.026 5.67e-05 ***
is_outdoorTRUE  0.68305    0.12022   5.682 1.33e-08 ***
is_dayTRUE     -0.19882    0.08324  -2.389  0.01691 *  
temperature     0.10306    0.02025   5.090 3.58e-07 ***
Sky.View.Mean   1.59007    0.51023   3.116  0.00183 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3578.7  on 2677  degrees of freedom
Residual deviance: 3466.4  on 2672  degrees of freedom
AIC: 3478.4

Number of Fisher Scoring iterations: 4
```

Figure 8: Logistic Regression Models Predicting "Cooler" response

The next model created was a tree-based classification model that predicts whether the thermal preference is “cooler”. A diagram of this tree model is shown below in Figure 9. In this model, the variables used for classification include the Sky View index, number of buildings, temperature, whether the user is outdoors, and whether it is winter or not. This model had an accuracy of 66% when predicting the “cooler” response.

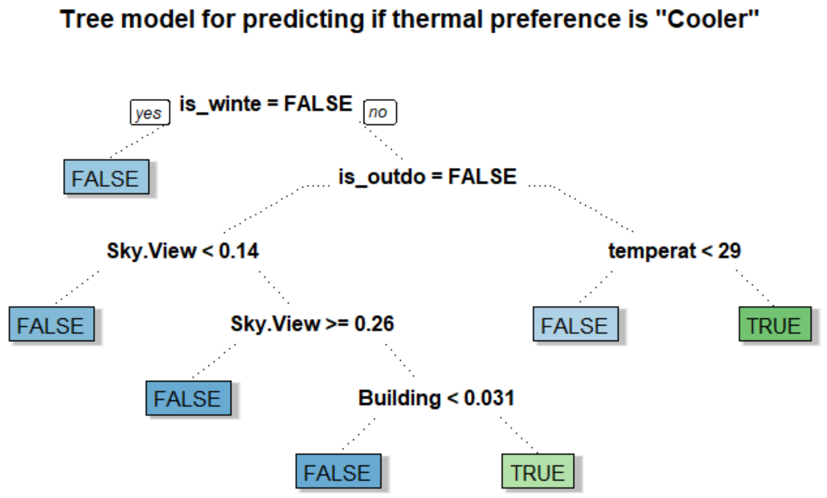
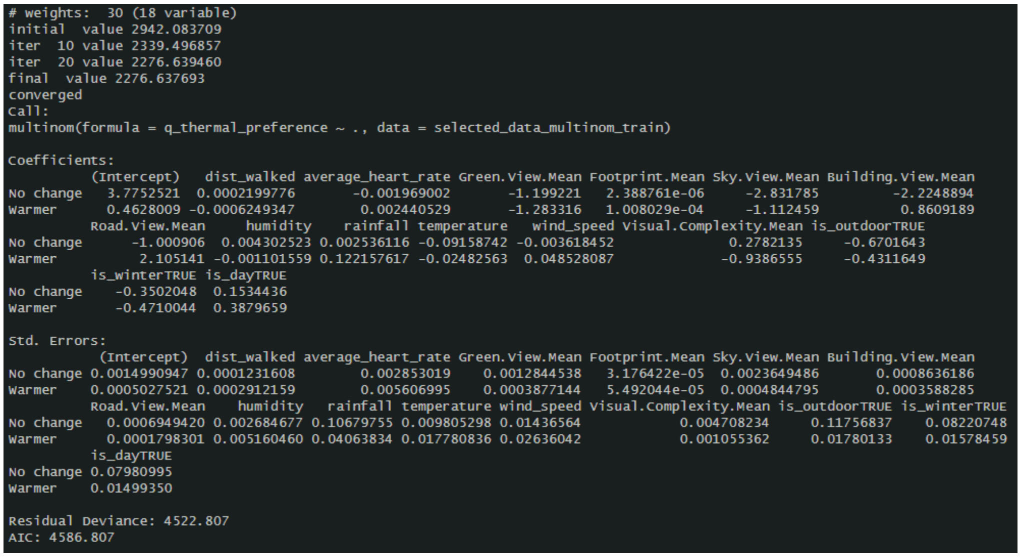


Figure 9: Tree-based classification model of “cooler” response

The last model used in classifying a “cooler” response was a random forest model. When this model was used in making this classification, it had a 71% accuracy. While this accuracy is higher than the other models above, there is a tradeoff in interpretability. Whereas in logistic regression and single tree classification models, the impact of each variable in the predictions is clear, the impact of variables in a random forest is a bit more complicated to explain. This is a tradeoff often seen with random forests, but this model did show the highest accuracy in prediction.

Models Predicting Actual Response

We then created models to try to predict the actual response. Specifically cooler, warmer, or no change, rather than only predicting a “cooler” response as we did in the models above. To do this, we created a multinomial logistic regression model. This model had a 56% accuracy. The output of this model can be seen below in Figure 10.



```
# weights:  30 (18 variable)
initial  value 2942.083709
iter  10 value 2339.496857
iter  20 value 2276.639460
final  value 2276.637693
converged
Call:
multinom(formula = q_thermal_preference ~ ., data = selected_data_multinom_train)

Coefficients:
          (Intercept)   dist_walked average_heart_rate Green.View.Mean Footprint.Mean Sky.View.Mean Building.View.Mean
No change   3.7752521  0.0002199776       -0.001969002       -1.199221   2.388761e-06     -2.831785         -2.2248894
Warmer      0.4628009 -0.0006249347        0.002440529       -1.283316   1.008029e-04     -1.112459          0.8609189
          Road.View.Mean     humidity    rainfall temperature   wind_speed Visual.Complexity.Mean is_outdoorTRUE
No change      -1.000906  0.004302523 0.002536116 -0.09158742 -0.003618452              0.2782135     -0.6701643
Warmer          2.105141 -0.001101559 0.122157617 -0.02482563  0.048528087             -0.9386555     -0.4311649
          is_winterTRUE is_dayTRUE
No change    -0.3502048  0.1534436
Warmer       -0.4710044  0.3879659

Std. Errors:
           (Intercept)  dist_walked average_heart_rate Green.View.Mean Footprint.Mean Sky.View.Mean Building.View.Mean
No change 0.0014990947 0.0001231608        0.002853019    0.0012844538   3.176422e-05  0.0023649486       0.0008636186
Warmer    0.0005027521 0.0002912159        0.005606995    0.0003877144   5.492044e-05  0.0004844795       0.0003588285
          Road.View.Mean    humidity   rainfall temperature wind_speed Visual.Complexity.Mean is_outdoorTRUE is_winterTRUE
No change   0.0006949420 0.002684677 0.10679755 0.009805298 0.01436564            0.004708234     0.11756837    0.08220748
Warmer      0.0001798301 0.005160460 0.04063834 0.017780836 0.02636042            0.001055362     0.01780133    0.01578459
          is_dayTRUE
No change 0.07980995
Warmer    0.01499350

Residual Deviance: 4522.807
AIC: 4586.807
```

Figure 10: Multinomial Logistic Regression Model

Continuing to this next model, the tree-based model classifies responses as “cooler” or “no change” based on multiple factors. Even though we tried to classify the actual thermal preferences with this model, we were unable to accurately train this model to classify responses as “warmer” because there simply weren’t enough “warmer” responses in the dataset to accurately fit these responses. The variables that were relevant in fitting this model were temperature, is_outdoor, Visual.Complexity.Mean, Sky.View.Mean, Footprint.Mean, humidity, Green.View.Mean, and Road.View.Mean. For further clarification on the meaning of these variables, refer back to the “Variable Definitions” section. This model had a 56% accuracy in predicting the thermal preference response. A visual representation of this model can be seen below in Figure 11.

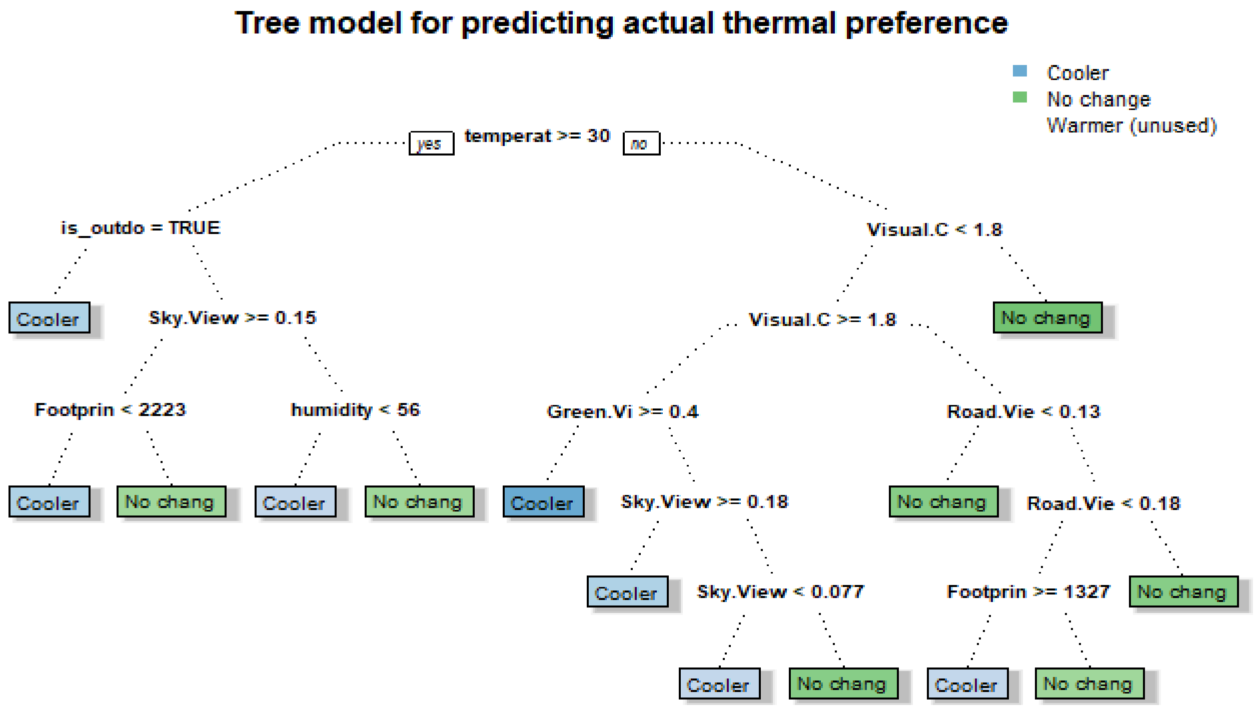


Figure 11: Tree classification model for actual thermal preference

The final model used was another random forest model, but this time we had this model predict the actual thermal preference instead of only the “cooler” response as in the previous random forest model. This model performed with a 65% accuracy. While this model’s accuracy was roughly the same as the accuracy calculated for some of the other models, an important distinction is that this model was able to predict a “warmer” response as well. In this regard, this model was more generalizable than some of the previous models that only worked to detect a “cooler” response.

Discussion

This study aimed to investigate the factors that contribute to a user’s thermal preference, namely when a user reports that they would prefer the environment to be cooler. Many different models were created, each with varying degrees of success. The model that most accurately predicted a thermal preference was the random forest model, though this comes with the tradeoff of being less easily interpretable.

However, regardless of the models used, we saw many of the same variables coming up as significant in predicting the thermal preference. Some of these variables that we saw again and again were temperature, humidity, is_outdoor, Footprint.Mean, Sky.View, Green.View, and Visual.Complexity. At the risk of oversimplification, this outcome suggests that increasing the density of green space and visual complexity of an environment (among other factors) can make people perceive the environment as cooler.

Throughout this research study, many insights were realized regarding what makes the environment feel comfortable. These insights show that the perception of a comfortable environment is complex and relies on many factors. As wearable sensor technology advances, more and more data will be available to researchers which will be valuable in future studies such as this one. Keeping abreast of these developments can further advance the field and improve the effectiveness of urban planning measures. Furthermore, future studies should involve a collaboration with experts in fields such as psychology, sociology, and urban planning to gain a more comprehensive understanding of what influences thermal comfort.

Conclusion

Hopefully, the results of this study will provide actionable insights for stakeholders, ranging from governmental agencies to urban developers. These groups can then address the challenges associated with unpleasant thermal temperatures, such as the urban heat island effect. By leveraging predictive analytics and data-driven approaches, cities can strive towards creating more comfortable, sustainable, and livable environments for their residents. To this end, fostering collaboration between governmental agencies, urban developers, and researchers is essential for implementing effective strategies. These efforts can lead to the creation of urban environments that are not only comfortable but also sustainable in the face of changing climates.

Citations

[1] Bhargava, A., Lakmini, S., & Bhargava, S. (2017). Urban heat island effect: Its relevance in urban planning. Journal of Biodiversity & Endangered Species. DOI: 10.4172/2332-2543.1000187

[2] Oyedepo, S.O. (2012). Noise pollution in urban areas: The neglected dimensions. Environmental Research Journal, 6(4), 259-271. Medwell Journals.

[3] Urban Redevelopment Authority. (n.d.). Heat resilient city. [Website]. Retrieved [insert date, e.g., October 5, 2021], from https://www.ura.gov.sg/Corporate/Get-Involved/Plan-Our-Future-SG/Innovative-Urban-Solutions/Heat-resilient-city

[4] Clayton Miller, Matias Quintana, Mario Frei, Yun Xuan Chua, Chun Fu, Bianca Picchetti, Winston Yap, Adrian Chong, and Filip Biljecki. 2023. Introducing the Cool, Quiet City Competition: Predicting Smartwatch-Reported Heat and Noise with Digital Twin Metrics. In Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '23). Association for Computing Machinery, New York, NY, USA, 298–299. https://doi.org/10.1145/3600100.3626269

[5] Yap, W., Biljecki, F. A Global Feature-Rich Network Dataset of Cities and Dashboard for Comprehensive Urban Analyses. *Sci Data* 10, 667 (2023). https://doi.org/10.1038/s41597-023-02578-1