Mobile Device User Behavior Report

Kenny Zhang

2024-12-03

Introduction

This dataset provides a detailed analysis of mobile device usage patterns, focusing on user behavior across 700 samples. It captures key metrics like app usage time, screen-on time, battery drain, and data usage, alongside demographic data such as age and gender. Users are categorized into one of five behavior classes, ranging from light to heavy usage.

All variables:

- User ID: Unique identifier for each user.
- Device Model: Model of the user's smartphone.
- Operating System: The OS of the device (iOS or Android).
- App Usage Time: Daily time spent on mobile applications, measured in minutes.
- Screen On Time: Average hours per day the screen is active.
- Battery Drain: Daily battery consumption in mAh.
- Number of Apps Installed: Total apps available on the device.
- Data Usage: Daily mobile data consumption in megabytes.
- Age: Age of the user.
- Gender: Gender of the user (Male or Female).
- User Behavior Class: Classification of user behavior based on usage patterns (1 to 5).

In the report, comprehensive Exploratory Data Analysis (EDA) and Classification Modeling was conducted to achieve 100% classification accuracy. The report provides helps to understand different usage behaviors, trends, and building predictive models to classify different users.

EDA

First, let's visualize the distribution of the data to have a better understanding.

The dataset is well balanced in terms of device models, user behavior classes, and gender. However, since only one phone model (iPhone 12) uses the iOS operating system, the majority of users are on Android. App usage time, screen-on time, and data usage exhibit right-skewed distributions, whereas the other variables are generally normally distributed.

Understand User Behavior Classes

We aim to visualize and understand the differences among users across various behavior classes. Specifically, we want to explore the relationship between App Usage Time and Battery Drain for users in each behavior class within the dataset.

As observed, there is a clear positive relationship between App Usage Time and Battery Drain—longer app usage leads to greater battery consumption, as expected. Additionally, we can visualize distinct differences among the user behavior classes. Users in class 1 exhibit significantly lower mobile usage compared to those in class 5, indicating a clear separation between light and heavy users.

Figure 1: Data Distributions

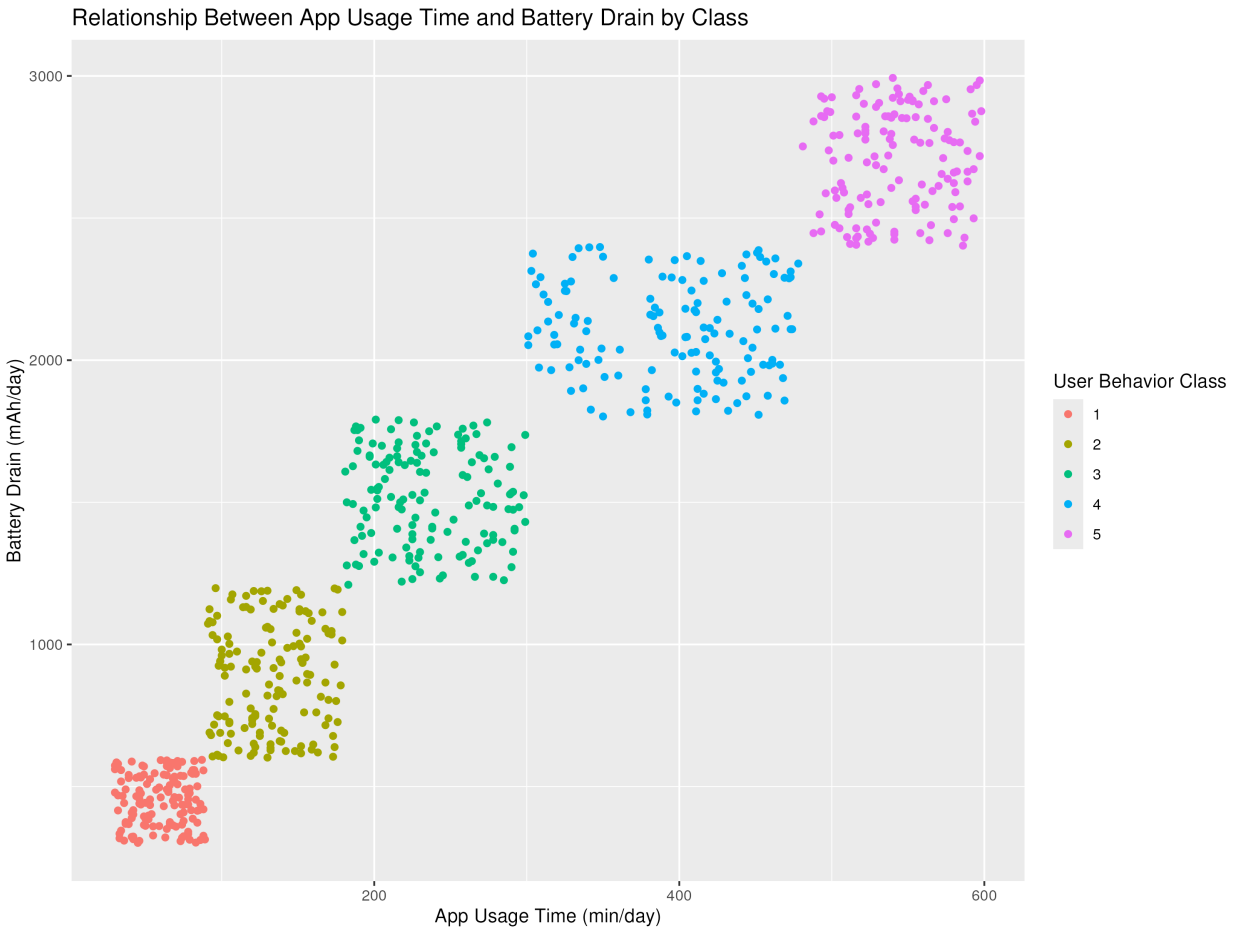

Figure 2: Relationship Between App Usage Time and Battery Drain by Class. The x-axis is daily time spent on mobile applications, measured in minutes, and the y-axis is daily battery consumption in mAh, color coded by user's behavior class.

Effect of age

Age could be an influential factor in determining user behavior. In this section, we explore the relationship between age and the number of apps installed on a device. Specifically, we are interested in whether younger users tend to have more apps installed and use their phones more frequently compared to older users.

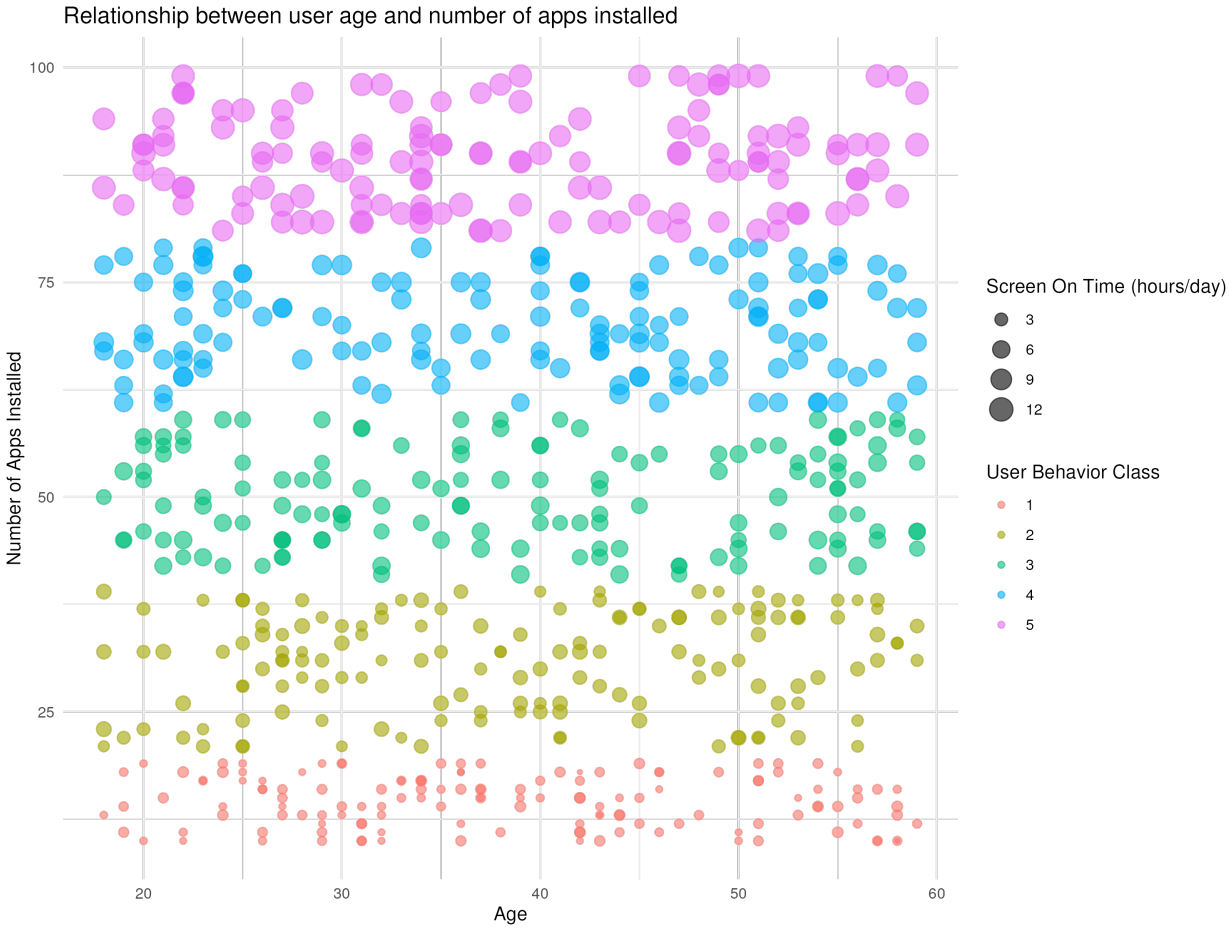

Figure 3: Relationship between user age and number of apps installed. The x-axis is age, the y-axis is the number of apps installed on the device. Each point is colored by their corresponding user behavior class and the size represents the average amount of hours per day the screen is active, the larger the points,the longer the corresponding screen time.

The scatter plots suggest that age does not significantly impact the number of apps installed or screen-on time. These variables appear to be randomly distributed across different age ranges without any clear trends or patterns. Instead, user behavior class serves as a more accurate indicator of how active a user is on their phone. Contrary to our initial expectations, the data suggests that older users are just as active on their phones as younger users.

Exploring difference between genders

Next, we examine whether gender influences user behavior, focusing specifically on App Usage Time as the variable of interest to measure potential differences.

The violin plot indicates that the distribution of daily app usage time is quite similar between genders. However, there is a higher concentration of females around the 100-minute range compared to males. When comparing the average usage times, the difference is not significant, as shown by the closely aligned red

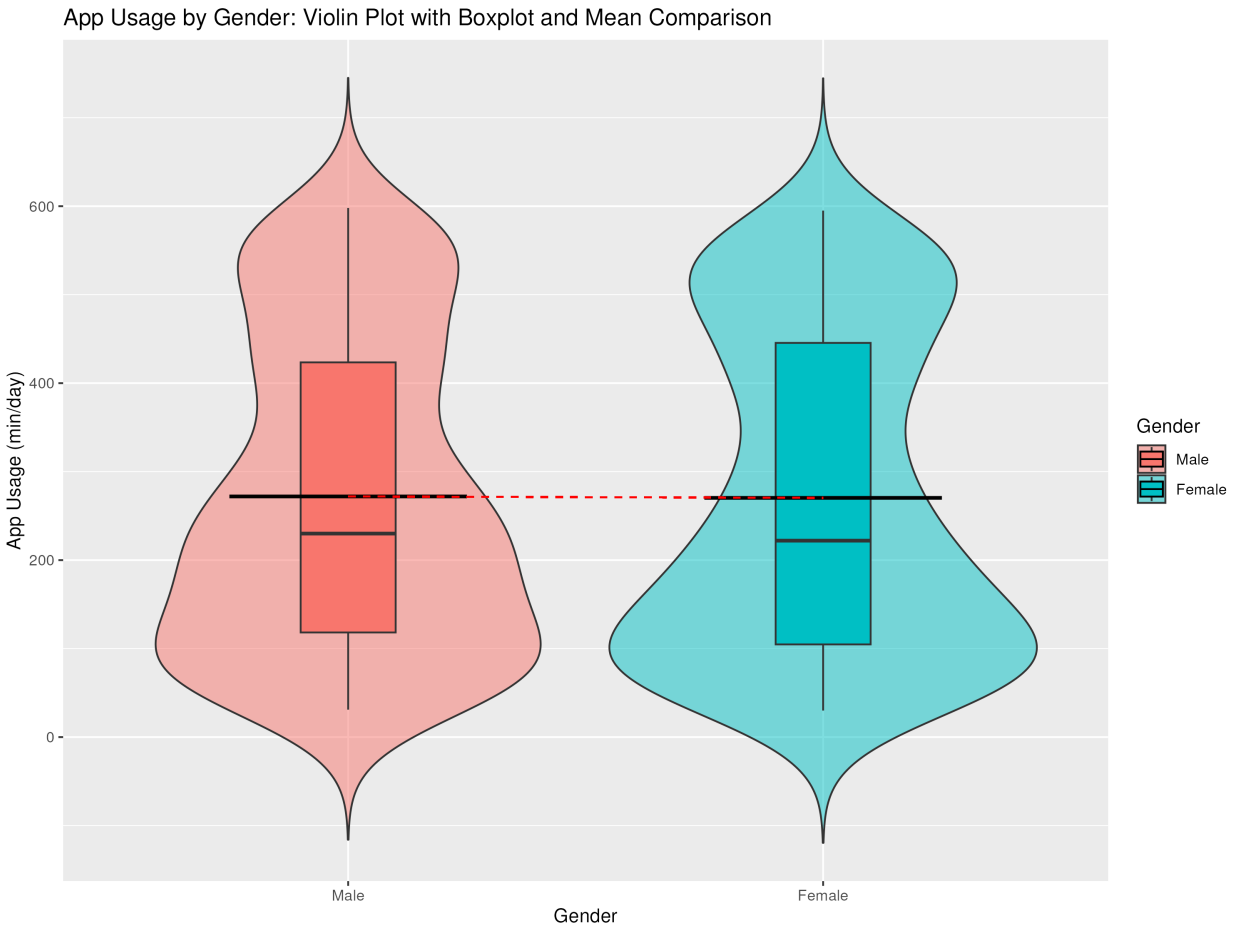

Figure 4: App Usage by Gender: Violin Plot with Boxplot and Mean Comparison. The x-axis is the two genders and the y-axis is daily time spent on mobile applications, measured in minutes. The red-dashed line is used to compare the difference in mean between the two genders.

dashed lines representing the means in the box plots.

Exploring differences among device models with genders

Next, we explore the differences among various phone models, focusing on the distribution of Data Usage across different devices, separated by gender. Our aim is to determine whether certain phone models are associated with lower data consumption and whether gender plays a role in influencing data usage patterns.

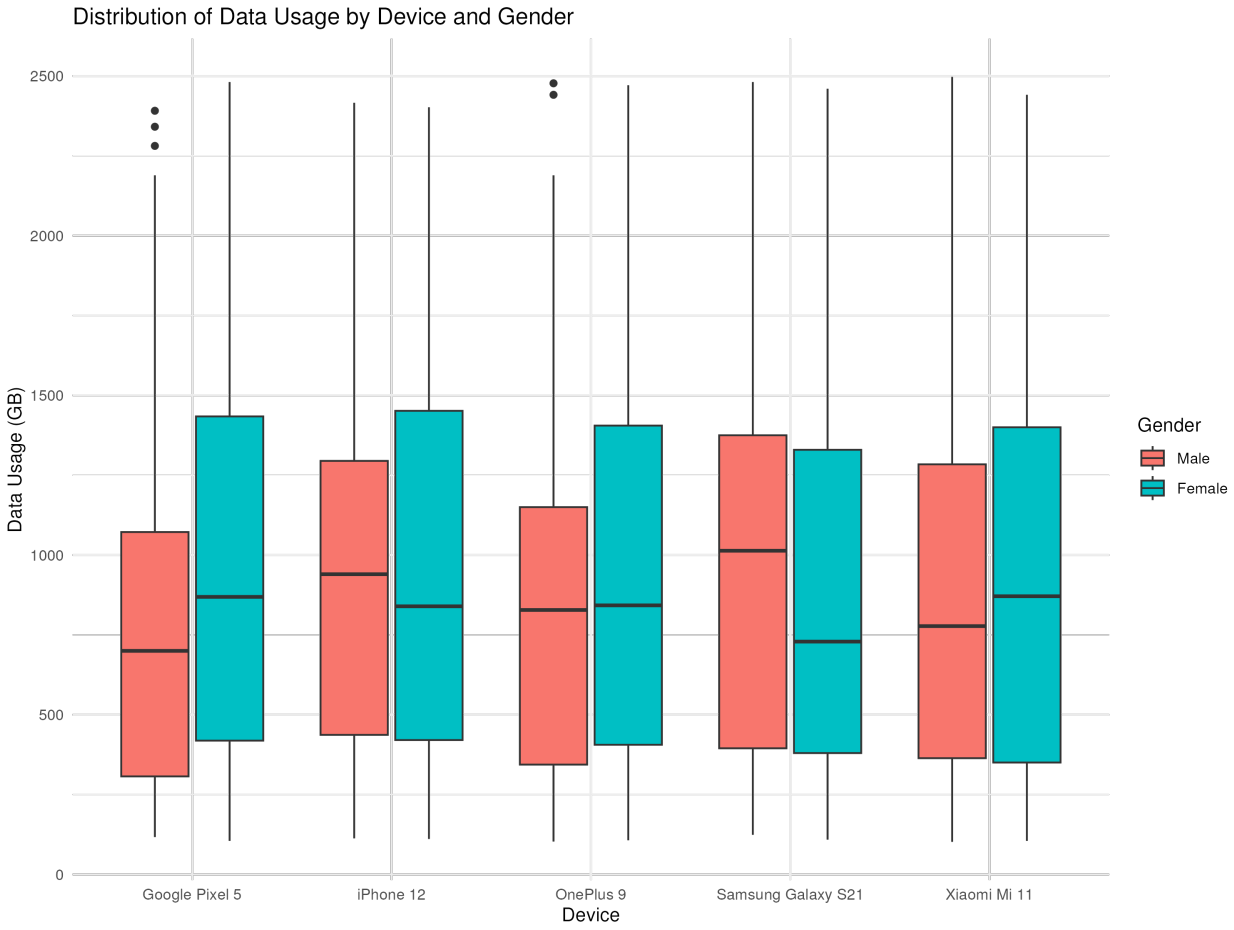

Figure 5: Distribution of Data Usage by Device and Gender. The x-axis is the model of the user's smartphone, the y-axis is the daily mobile data consumption in megabytes. The box plots are also separated by gender.

The box plots indicate that there are no clear differences in data consumption across the various device models, and it is not definitive that one gender consistently uses more data than the other. Instead, data consumption seems to depend on the specific device model. Males tend to use more data with the Google Pixel 5, OnePlus 9, and Xiaomi Mi 11, while females tend to use more data with the iPhone 12 and Samsung Galaxy S21.

Exploring differences in financial efficiency among device models

Different phones exhibit varying levels of battery and data consumption. In this section, we explore the average battery drain and data usage for each device model, aiming to identify which devices are the most efficient in terms of power consumption and data usage, thereby providing more financially favorable options for users.

The average battery and data consumption are quite similar across different devices. However, the iPhone

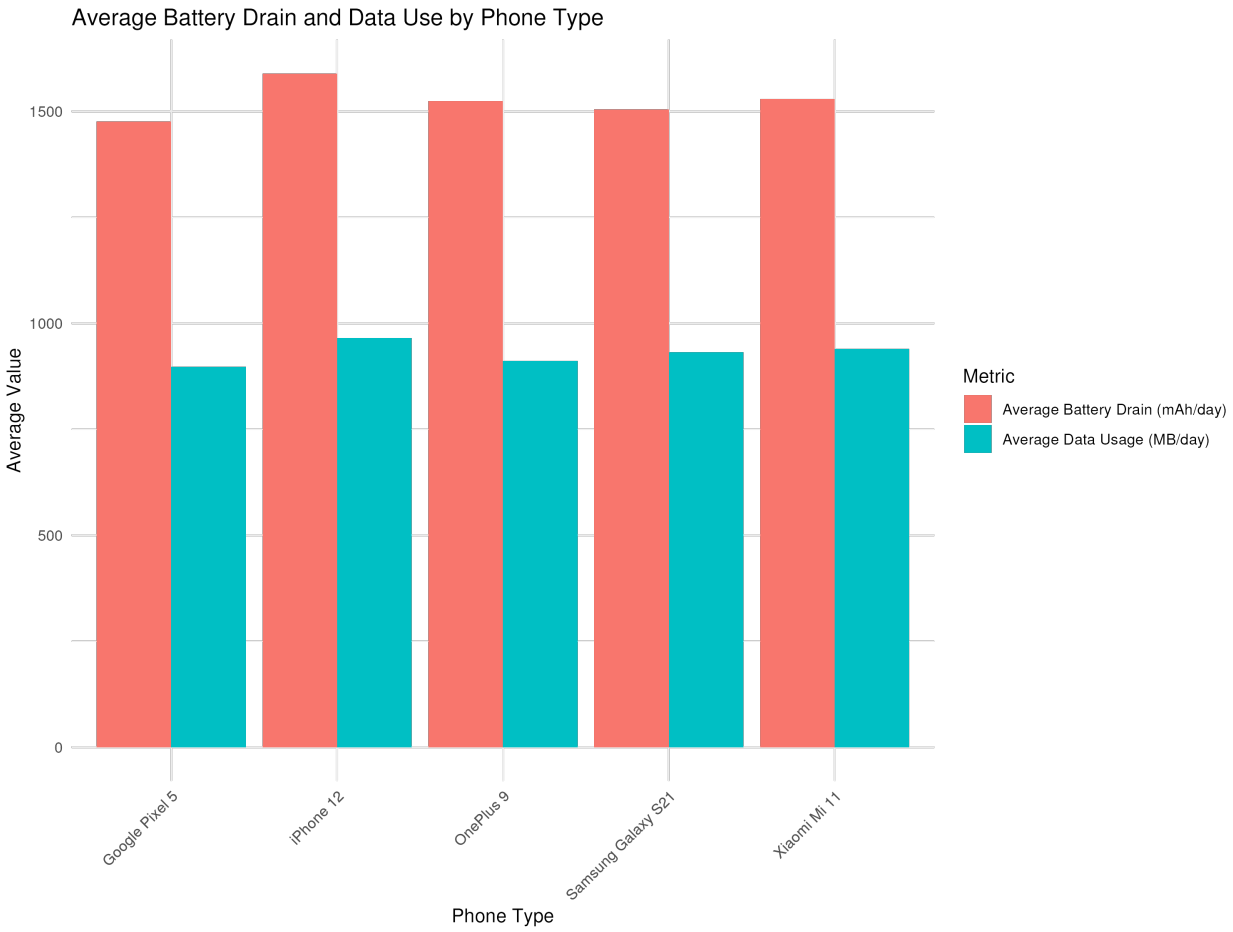

Figure 6: Average Battery Drain and Data Use by Phone Type. The x-axis is the device models, the y-axis is the average values for battery drain and data usage. The representation of different bars are marked in the legend.

12 shows slightly higher battery and data usage compared to other devices. In contrast, the Google Pixel 5 stands out as the most efficient, consuming the least amount of battery and data on average.

To further investigate the differences in battery and data consumption across various phone models, we created an interactive Shiny plot for visualization. This plot allows users to explore the data dynamically by clicking on colored dots in the legend to select or de-select different device models. Additionally, users can utilize the box select and lasso select tools to highlight specific data points and calculate the average battery drain and data usage for the selected groups, providing a more in-depth understanding of device efficiency. This section is only availiable for interactive html.

Finding clusters within the data

Given that the different behavior classes are clearly distinct, it is worthwhile to explore whether clustering can similarly represent these classes. To investigate this, we performed K-Means clustering and Hierarchical Clustering on the first two Principal Components (PCs) of the data. This approach allows us to assess whether the resulting clusters align with the predefined user behavior classes, providing additional insight into how well these behavioral patterns can be captured by unsupervised learning methods.

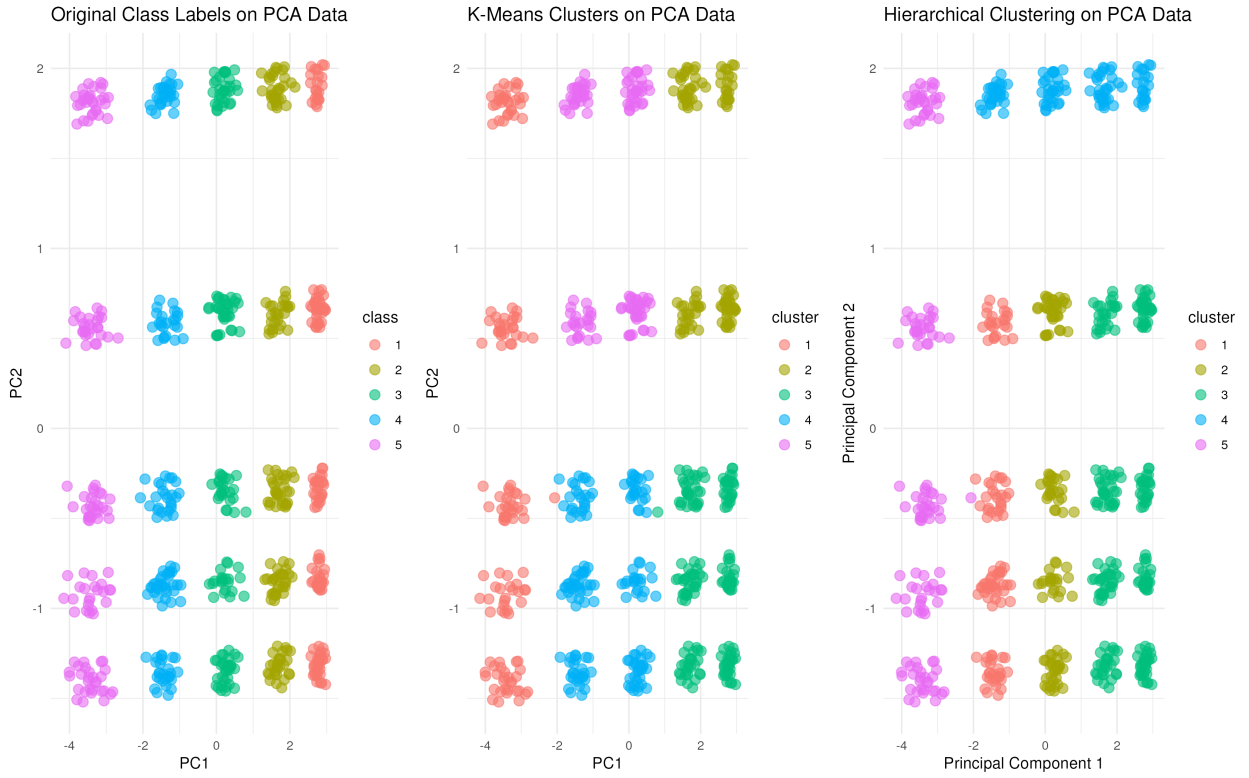

Figure 7: Comparison of Two Clustering Methods with the Actual Classes. The x-axis is the first PC and the y-axis is the second PC. Each class/cluster is labeld with different colors.

The two clustering methods effectively capture the distinction between class 5 and the other classes, but they are less reliable in differentiating among the remaining four classes. Both methods tend to combine similar classes and split classes 1 to 4 into upper and lower segments, indicating the potential presence of subgroups within the data. Therefore, we proceed with classification using machine learning techniques to accurately predict each user's behavior class based on their demographics and phone usage data.

Classification

In this section, we compare three different machine learning methods in terms of their predictive accuracy to determine which model performs best for predicting user behavior classes. The dataset is split into training and test set, with 70% of the data split into training set (489 samples) and 30% as test set (211 samples). All classification models' performances are evaluated on the test set.

Baseline Model - Random Guessing

To determine whether our models learn anything useful, we need a baseline model to compare the performances. The baseline randomly assigns one of the 5 classes to each user, not using any information from the data at all.

Multinomial Logistic Regression

Multinomial logistic regression is a statistical modeling technique used to predict the probability of different outcomes when the response variable is categorical with more than two possible classes. Unlike binary logistic regression, which handles only two categories, multinomial logistic regression can model scenarios where there are three or more classes. It works by estimating the odds ratios of each class relative to a reference category, using a set of predictor variables.

Gradient Boosting Machine

Gradient Boosting Machine (GBM) is an ensemble learning technique that builds a series of weak learners, typically decision trees, to create a strong predictive model. It works by sequentially adding trees, where each new tree aims to correct the errors made by the previous ones, minimizing the overall loss. The "boosting" refers to the process of iteratively improving the model by focusing more on data points that were previously misclassified or had higher residual errors. GBM is known for its flexibility and effectiveness in handling various types of data, often providing high accuracy, making it a popular choice for both regression and classification tasks.

With the baseline model's performance of 20.38% accuracy, both multinomial logistic regression and GBM have large increase in performance. Multinomial logistic regression is able to achieve 99.53% accuracy and GBM achieved an accuracy of 100%. Hence, GBM wins as the most accurate model. Since GBM can already achieve 100% predictive accuracy, there is no need to explore more computational expensive deep learning models such as neural networks.

Conclusion

The analysis of the mobile device usage dataset reveals interesting insights into user behavior, although the relationships between variables like age, gender, and device model with the user behavior classes are not immediately apparent. Despite the absence of clear-cut correlations between these demographic features and behavior classes, we can leverage various machine learning techniques to effectively predict a user's behavior category based on their demographic and usage data. Methods such as multinomial logistic regression and gradient boosting machines have shown promise in capturing complex, non-linear relationships within the data, enabling accurate classification of user behavior.

The ability to predict user behavior classes based on features like app usage time, screen-on time, and battery consumption highlights the predictive power of these models, even in cases where individual variables may not show obvious patterns. By using powerful classification techniques, we can identify meaningful patterns that differentiate light users from extreme users, allowing for better segmentation and targeted analysis.

This capability is especially valuable in domains such as marketing and personalized services, where understanding different user groups can lead to more tailored offerings and an improved user experience. Furthermore, clustering analyses provide additional insights into potential subgroups that may not align with the predefined behavior classes, pointing towards more nuanced behaviors within the dataset.

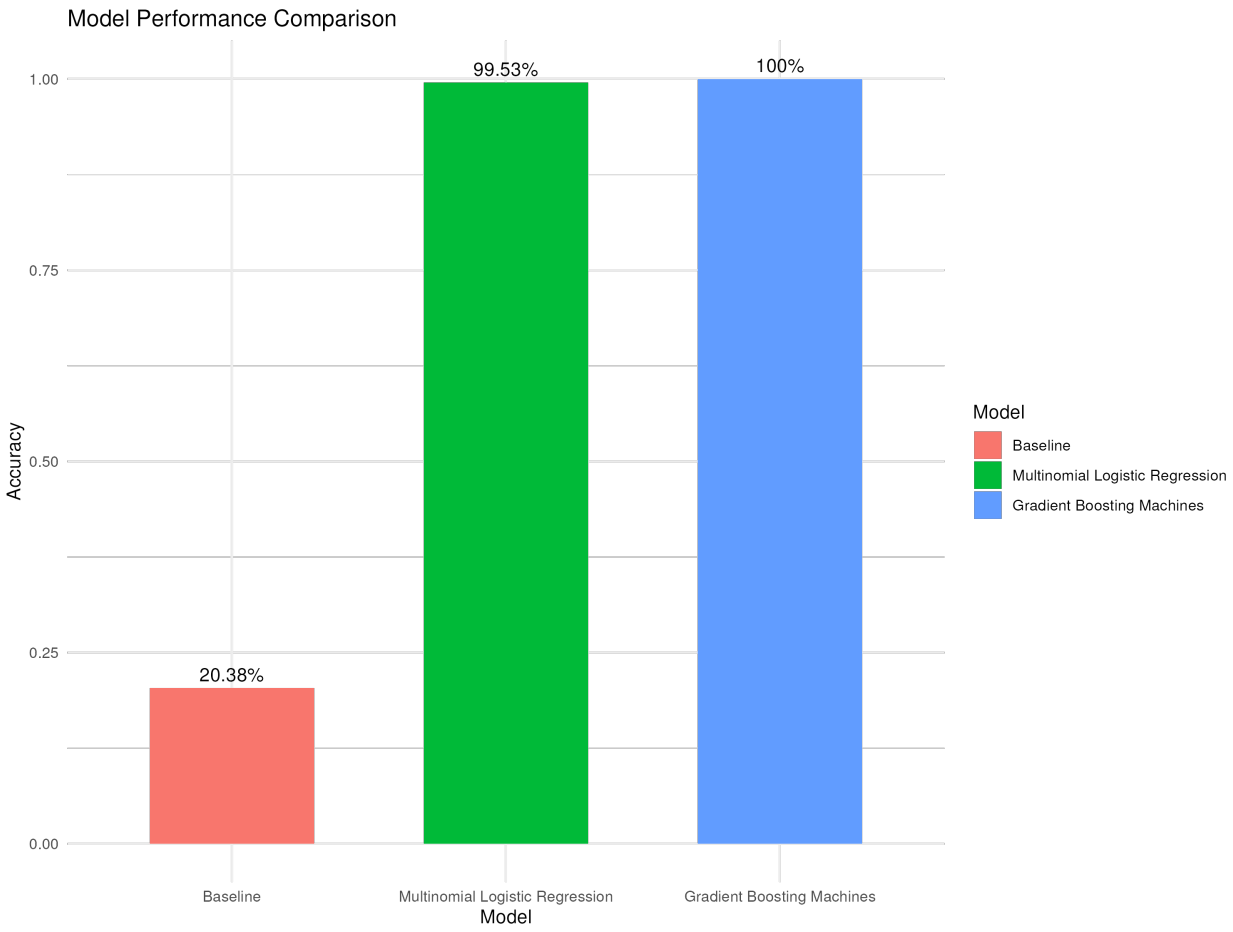

Figure 8: Model Performance Comparison. The x-axis is the classification models and the y-axis is the prediction accuracy.

Overall, while simple relationships between demographic and usage features may be difficult to discern, advanced predictive modeling enables us to build effective classifiers that can help understand and predict user behavior on a broader level. This understanding can drive more personalized recommendations, improve mobile app development, and contribute to more targeted energy-efficient features for mobile devices, ultimately enhancing both user experience and device performance.