**Lincoln Brown**  
**Predicting App Usage Time**  
**Milestone 2**   
**DSC680-T301**  
**Professor Iranitalab**

Business Problem

Predicting app usage time for users is critical for businesses, app developers, and marketers who aim to increase app usage time and user retention by optimizing the user experience for different target demographics (Admin & Admin, 2025). Understanding the factors that influence app usage time can help businesses and developers target high-engagement users, predict app usage time, and improve customer satisfaction and user retention. App developers need insights into user behavior to provide a better user experience that fosters a personalized experience, which can help lead to longer app usage time and better user retention.

This project aims to provide actionable insights by identifying key factors associated with app usage time. Key analysis will focus on the correlation between features such as screen-on time, daily data consumption, age, gender, operating system, and device model to determine which factors are most important for predicting app usage time. Using this information, businesses can develop data-driven strategies to enhance app usage time and user retention.

Background/History

Phone usage has become ubiquitous in society, drastically changing how users interact with companies and conduct business. One important aspect of this change is evidenced by the growing trend of apps. In 2016, roughly 146 billion apps were downloaded from app stores worldwide. By 2023, this number had grown to 257 billion (Statista, 2024). It seems there is an app for everything, ranging from communication and health to education and entertainment. The pervasiveness of apps in everyday life underscores the importance of understanding user behavior, which is essential for developers and businesses looking to retain users and encourage app usage. As the app market is saturating, it is important for businesses to not only attract new users but keep the existing ones they have. This is where predictive analytics demonstrate their value to a company.

Data Explanation

The dataset used in this project is Mobile Device Usage and User Behavior Dataset found on Kaggle, uploaded by user Vala Khorasani. The dataset contains 700 samples of user data including metrics such as app usage time, screen-on time, battery drain, number of apps installed, data usage, and user specific features such as age and gender.

All features were included in the analysis and model creation steps. The data cleaning consisted of dropping 11 rows that had missing ages, scaling numerical features with a standard scaler, and using one-hot encoding on categorical features.

The target feature for the analysis is the ‘App Usage Time (MB/Day)’ feature. There were three models created for this project, a linear regression model, random forest regression model, and an XGBoost regression model.

Methods

**Linear Regression Model**

The linear regression model was used as a baseline for model performance. A linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables (GeeksforGeeks, 2025).

**Random Forest Regression Model**

Random Forests (RF) are an ensemble learning algorithm that use a tree machine learning algorithm that builds multiple decision trees during the training phase. Each decision tree is composed of a random subset of the data set to measure a random subset of features in each partition (GeeksforGeeks, 2024). Utilizing random subsets of features introduces variability among individual trees, which can help reduce the chances of overfitting and improve overall prediction performance. The RF model aggregates the predictions of many decision trees, which provides more reliable and accurate results compared to a single decision tree.

**XGBoost Model**

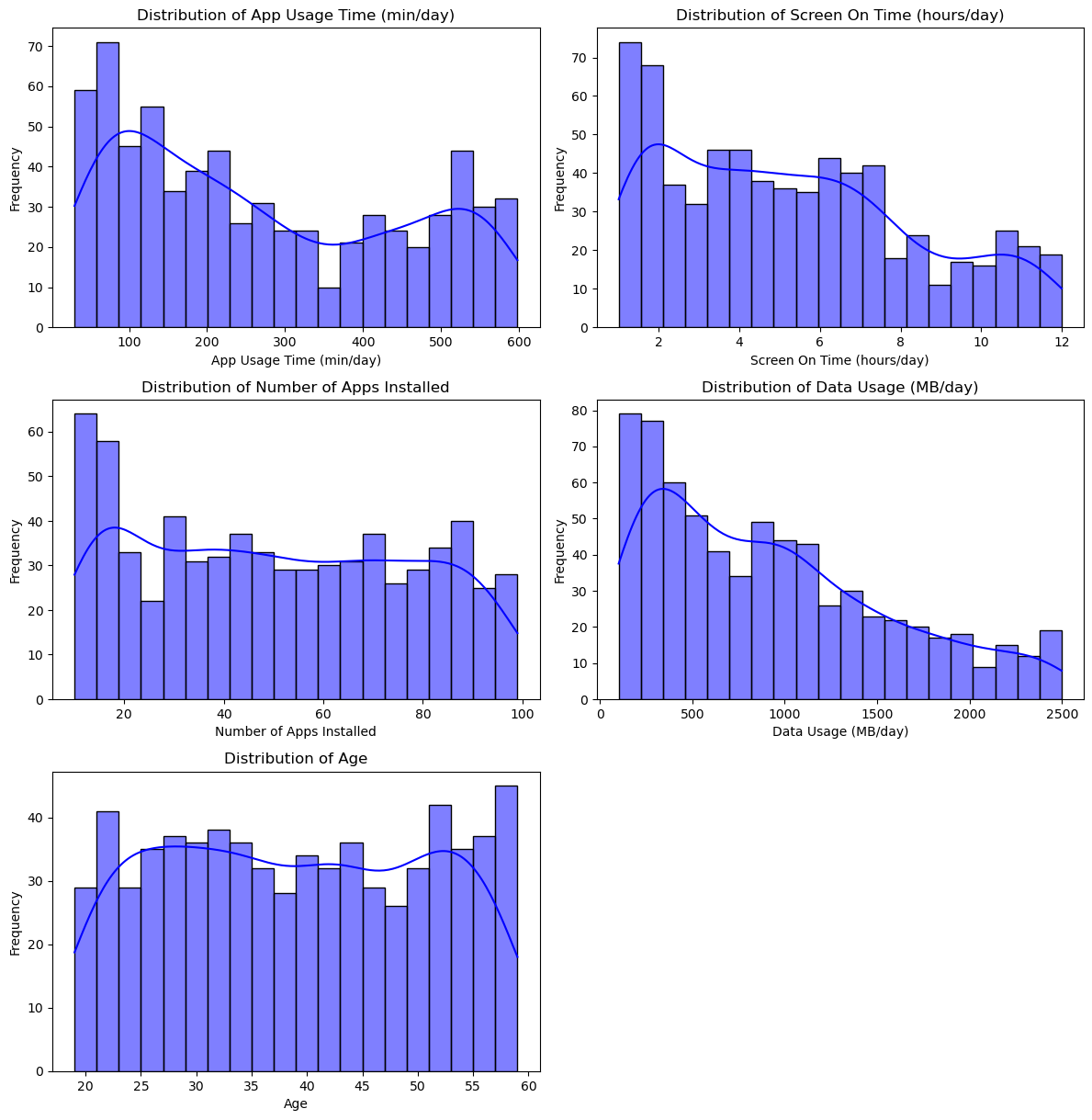
XGBoost or eXtreme Gradient Boosting is a distributed, open-source machine learning library that uses gradient boosted decision trees, a supervised learning boosting algorithm that makes use of gradient descents (Kavlakoglu & Russi, 2024).

Analysis

The goal of the analysis was to determine the answer to five different questions:

* Which age group has the highest app usage time?
* Are there any differences in app usage time between genders?
* How does daily screen time correlate with app usage time?
* Do heavier app users use significantly more data?
* Which features were the most important for model prediction?

Initial data exploration involved looking at the distribution of the numerical features:

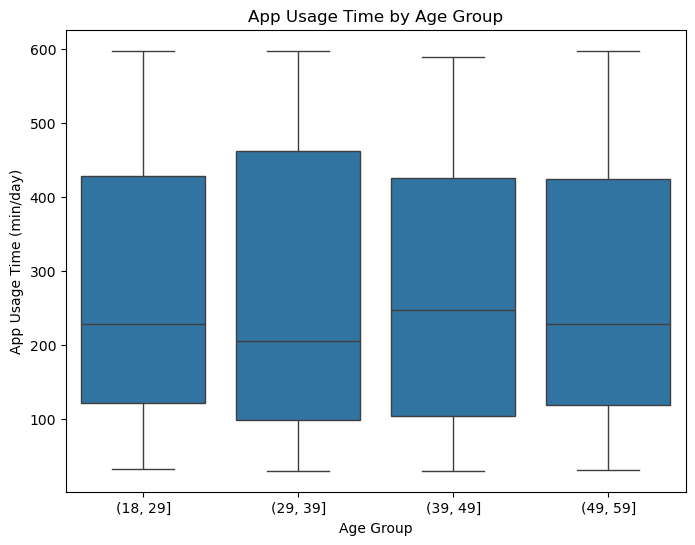


Looking at each of the different features was an important step to see if there were any clear trends in the data. The distribution of age was surprising, as it did not favor younger audiences, which is often expected in app usage studies. Rather, the distribution of age was fairly consistent across all of the age groups, which could imply that app usage is fairly widespread throughout all of the age groups.

The distributions for app usage time, screen on time, and data usage are all left skewed indicating that more users tend to use their devices less.

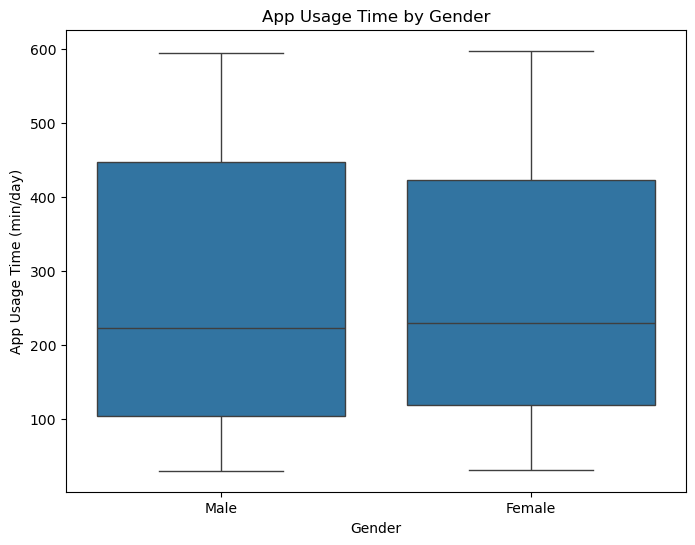
Additionally, the distribution for the number of apps installed indicated that most users have less than 20 apps installed, although the number of apps installed was consistent after the initial spike.

To answer our first question, what age group has the highest app usage time, we will look at the distribution of app usage time by age group:



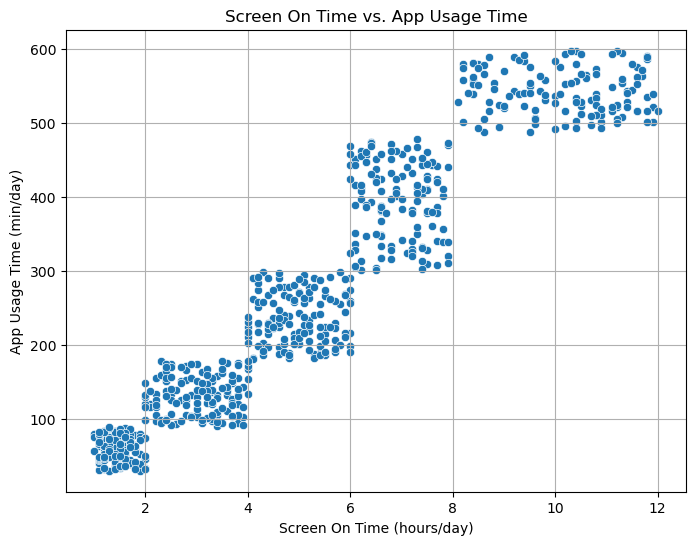
From the graph above, we can see that the 39-49 age group has a slightly higher median app usage time than the other groups. However, the groups are overall very similar.

For our second question, we will investigate whether there is any difference between the app usage time between the genders.



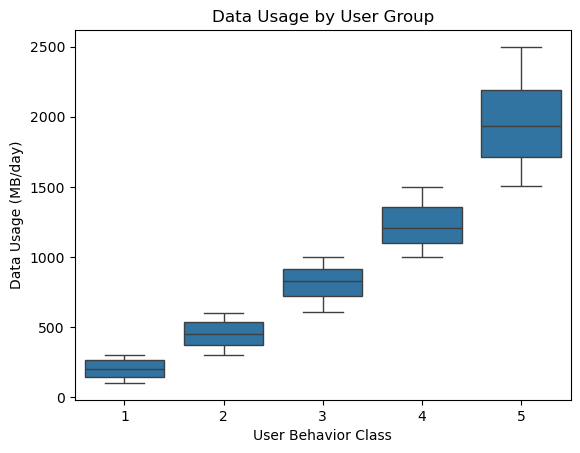
From the graph above, we can see that the app usage time between the two genders is very similar, indicating little to no difference.

Our third question was how does screen time correlate with app usage time? We will investigate this by looking at a scatterplot:



We can clearly see a positive trend in screen time vs app usage time. The longer users have their phone screen turned on, the more likely they are to be using an app.

For our fourth question, we want to see if users assigned to heavier use groups use more data than those that do not:



Again, we see that users in user behavior classes 4 and 5 tend to use more data than those in classes 1 through 3.

Now that we have answered our preliminary questions, we can move into the model building and evaluation section to determine the answer to our final question.

The goal of the analysis is to evaluate and compare the performance and applicability of the Linear Regression (LR), Random Forest (RF), and XGBoost (XBG) models. All three of the models were trained and tested on the dataset and evaluated to determine which is the best candidate for predicting app usage time.

The performance metrics used to evaluate the models were:

Root Mean Squared Error (RMSE) - is a measurement of the average difference between the model’s predicted values and its actual values (Frost, 2023).

Mean Absolute Error (MAE) - is a measurement of the absolute discrepancies between a dataset’s actual values and predicted values (GeeksforGeeks, 2024).

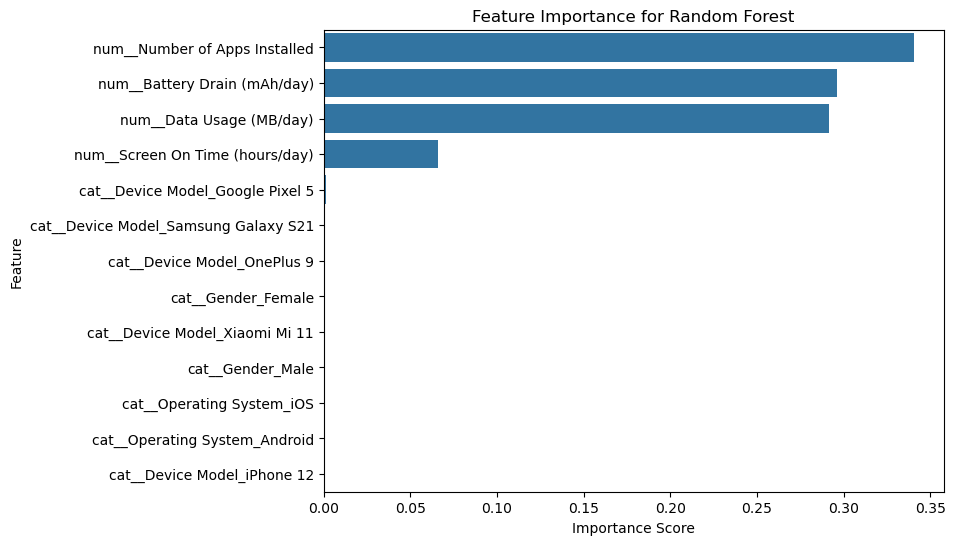
R2 Score – is the percentage of the response variable variation that is explained by a linear model (Frost, 2017).

The models achieved these results:

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MAE | RMSE | R2 Score |
| Linear Regression | 34.745836 | 44.74348 | 93.96% |
| Random Forest | 28.735362 | 35.751961 | 96.14% |
| XGBoost | 31.244812 | 38.389111 | 95.56% |

From these results we can see that the Random Forest model performed the best across all of the metrics.

Another key area of analysis was determining which features were the most meaningful for the model’s prediction accuracy:



From the graph above, it is clear that the most important features for predicting app usage time was the number of apps installed, daily battery drain, and daily data usage. This is unsurprising, as these factors directly relate to how often and intensively a user interacts with their phone.

Conclusion

This project aimed to investigate the features that influence app usage time and to determine which predictive model performed the best at predicting app usage time with the dataset Mobile Device Usage and User Behavior and which features were the most important for predicting app usage. Through the comparison of three models Linear Regression, Random Forest, and XGBoost, it was determined that the Random Forest Model had the best predictive performance, with the lowest MAE and RMSE scores and the highest R2 score. The features that were most significant predictors for app usage time were number of apps installed, daily battery drain, and daily data usage. This aligns with expectations, given that these three features directly correlate with how users engage with apps on their device.

In conclusion, this project provides insight into app usage trends and the key factors influencing them. A more robust dataset could provide better information into behavioral trends. Features that would be beneficial in future analysis would include the number of notifications received daily, app types, app names, number of minutes for each app, and time between app usage.

Assumptions

There were several assumptions made in this project. First, that the dataset is accurate, and representative of the population being analyzed. Second, that the dataset provided an equal representation of different user demographics, ensuring that the data is not biased in some manner. Third, that the data was collected from geographically diverse areas to avoid potential bias.

Limitations

Initially, this project was started with another dataset in mind, specifically, the Mobile Apps Screentime Analysis dataset on Kaggle. After determining that this dataset did not have any user information, it was discarded in favor of the Mobile Device Usage and User Behavior dataset. However, a combination of these two datasets would have been preferred, as both seem limited in their applicability to fully analyze how users are interacting with their phones. Future implementations of the models developed in this project would be well served to find more comprehensive datasets to analyze.

Challenges

The dataset used in this project was limited in the amount of features it had. It would have been beneficial to have more information about the specific apps being used, the duration of their use, and the number of notifications that each user received. However, this was out of scope for the dataset and efforts to find more comprehensive datasets were unsuccessful.

Future Uses/Additional Applications

Future implementations of this project are limited due to the vague nature of the dataset itself. The models used in this project are applicable to other datasets and it would be beneficial to investigate more comprehensive datasets in the future for more specific insight into predicting user behavior.

Recommendations

The dataset used in this project was limited in its number of features. It is recommended that future projects find a more comprehensive dataset.

Implementation Plan

This project was purely for academic purposes, and I would not recommend using it for business purposes.

Ethical Assessment

The ethical implications of this project were minimal. The dataset was not sensitive and did not contain any personally identifiable information. Additionally, the model seemed to adequately capture the diversity of age groups and genders represented in the dataset.

References

Admin, & Admin. (2025, January 20). Predictive Analytics in Mobile Apps: Key 2025 trends. *Nordstone - Mobile App Development*. <https://nordstone.co.uk/blog/the-role-of-predictive-analytics-in-mobile-apps-trends-for-2025>

Duarte, F. (2025, January 22). Time spent using smartphones (2024 statistics). *Exploding Topics*. <https://explodingtopics.com/blog/smartphone-usage-stats>

Frost, J. (2023, May 28). *Root Mean square Error (RMSE)*. Statistics by Jim. <https://statisticsbyjim.com/regression/root-mean-square-error-rmse/>

Frost, J. (2017, May 5). *R-Squared - Statistics by Jim*. Statistics by Jim. <https://statisticsbyjim.com/glossary/r-squared/>

GeeksforGeeks. (2025, January 16). *Linear Regression in Machine learning*. GeeksforGeeks. <https://www.geeksforgeeks.org/ml-linear-regression/>

GeeksforGeeks. (2024, December 11). *Random Forest algorithm in machine learning*. GeeksforGeeks. <https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

GeeksforGeeks. (2024, July 29). *Regression metrics*. GeeksforGeeks. <https://www.geeksforgeeks.org/regression-metrics/>

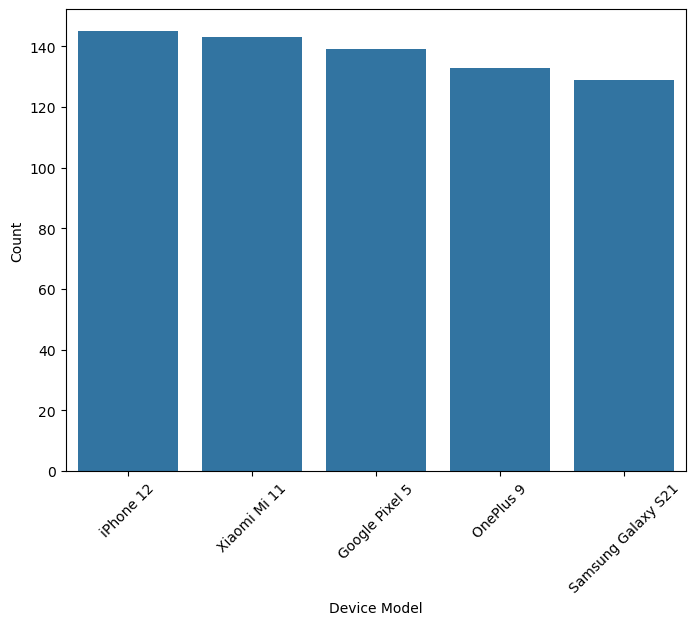
Kavlakoglu, E., & Russi, E. (2024, December 19). *What is XGBoost?*. IBM. <https://www.ibm.com/think/topics/xgboost>

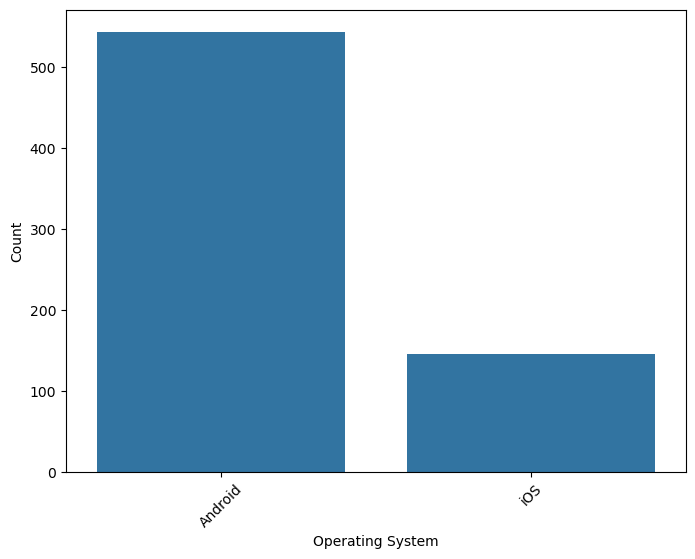
Statista. (2024, April 5). *Annual number of global mobile app downloads 2016-2023*. <https://www.statista.com/statistics/271644/worldwide-free-and-paid-mobile-app-store-downloads/>

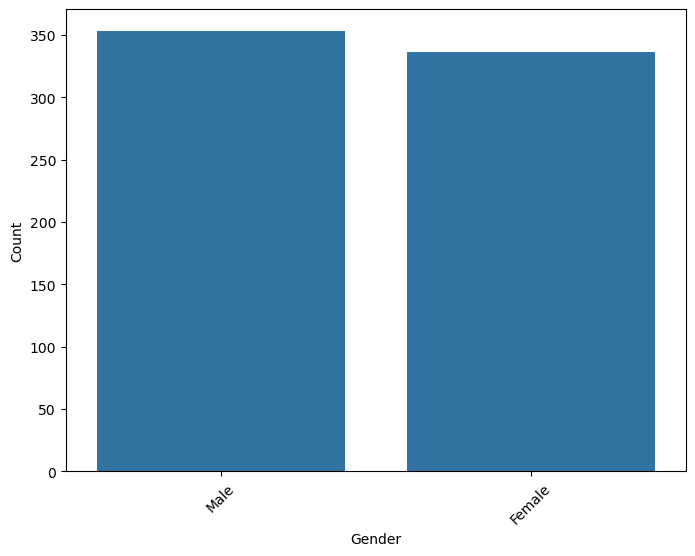
Appendix

Below are some of the EDA visualizations and code snippets that were not included in the white paper.

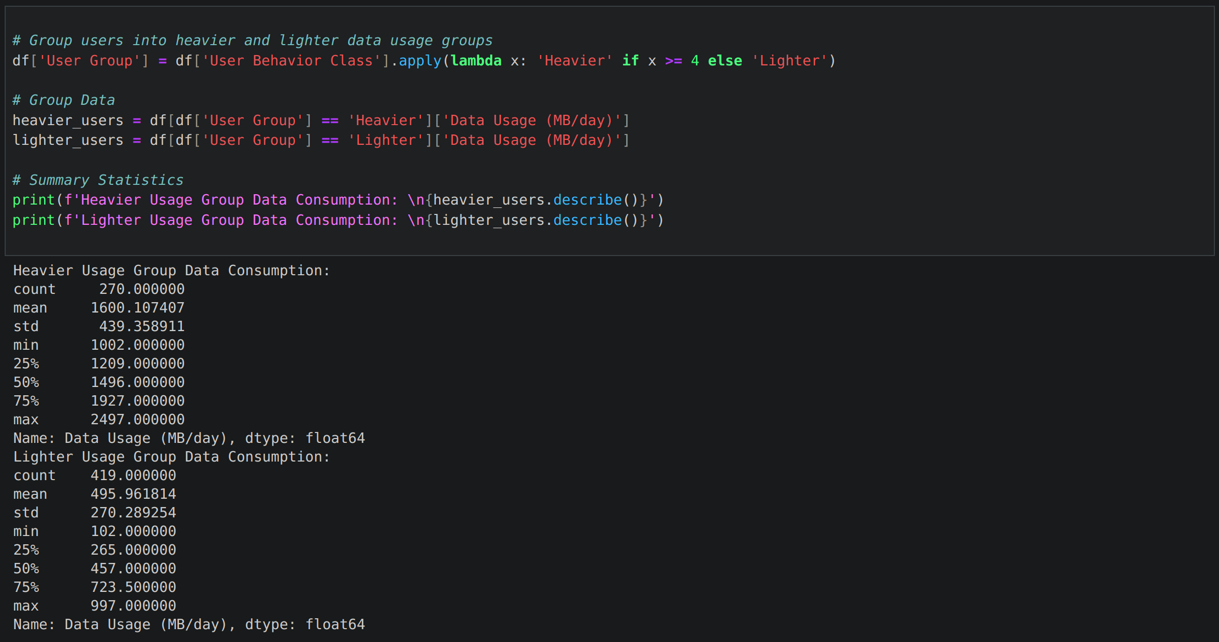
Bar Plots of the categorical features:







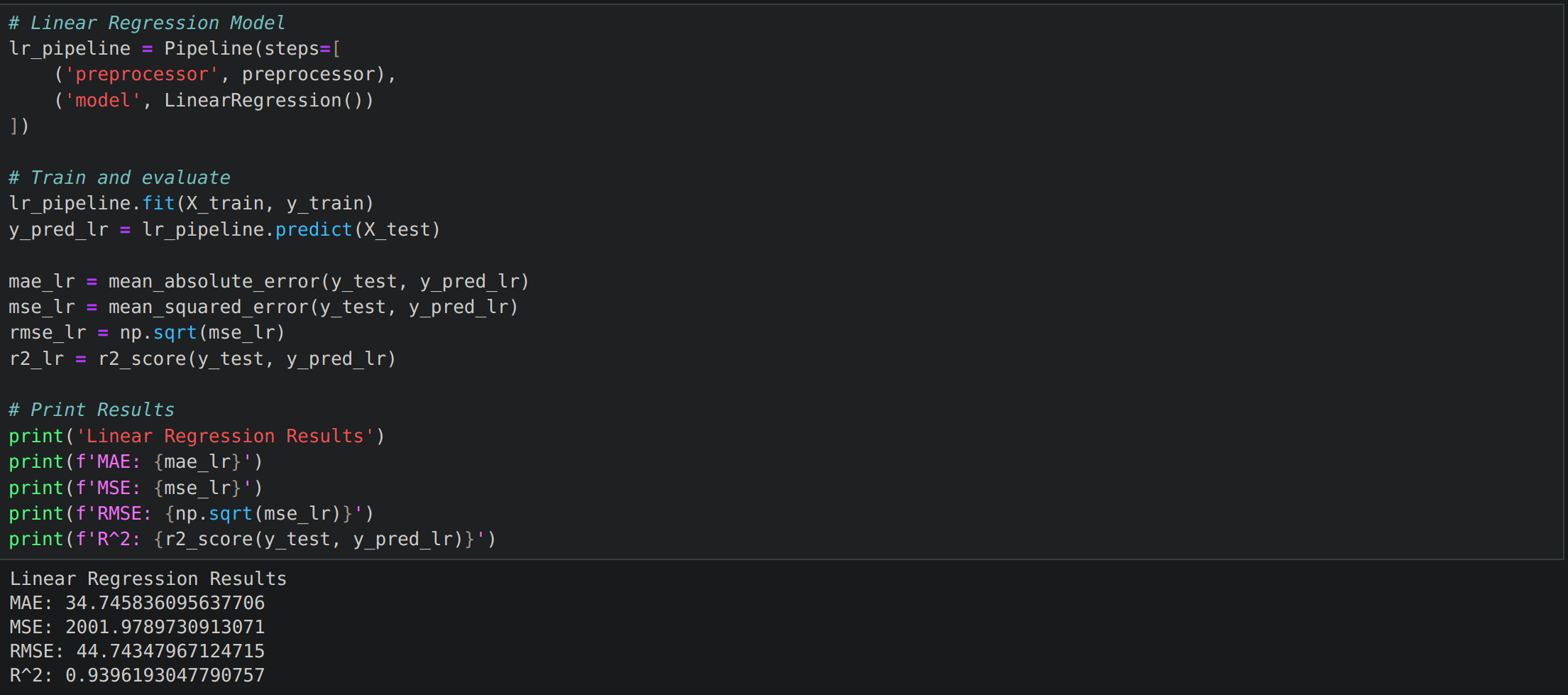
Creation of User Groups for T-Test



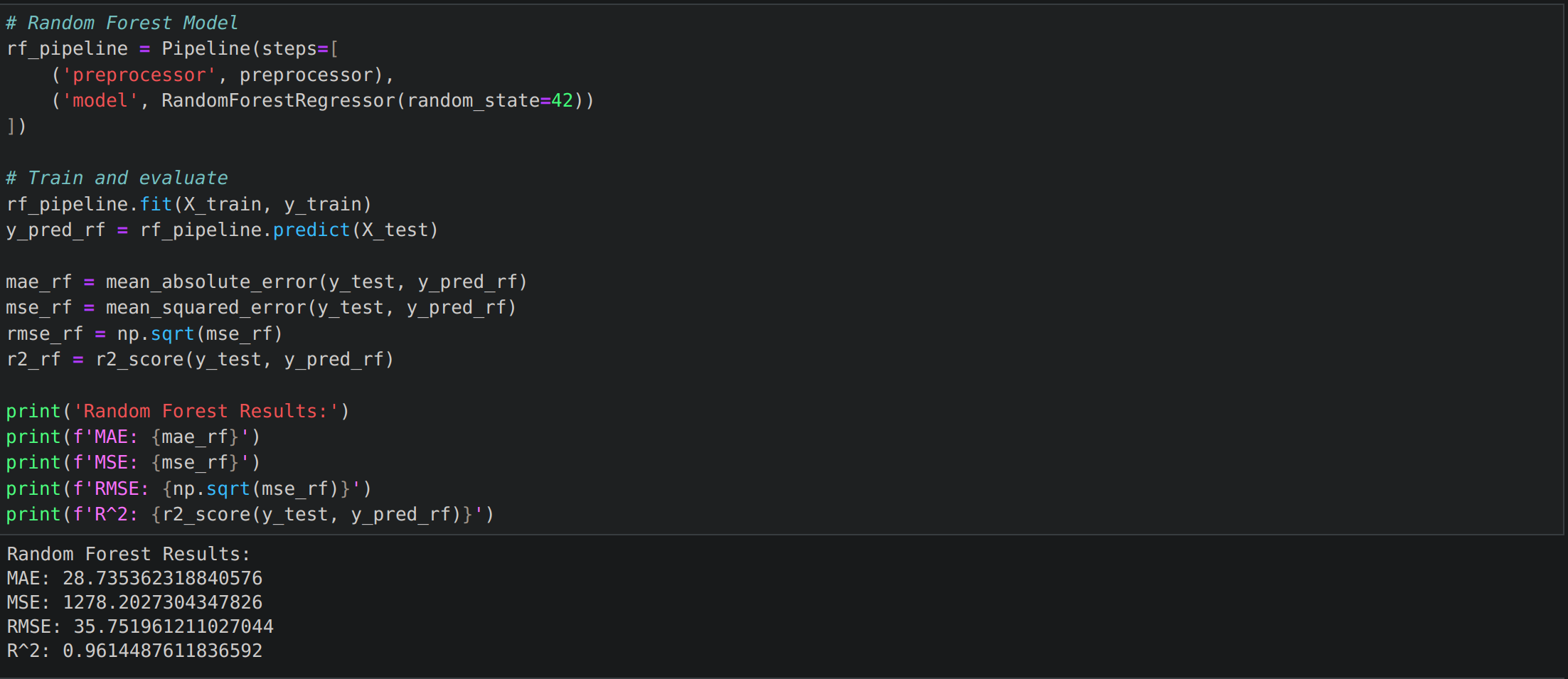
T-Test for statistical significance in data usage between data usage groups.



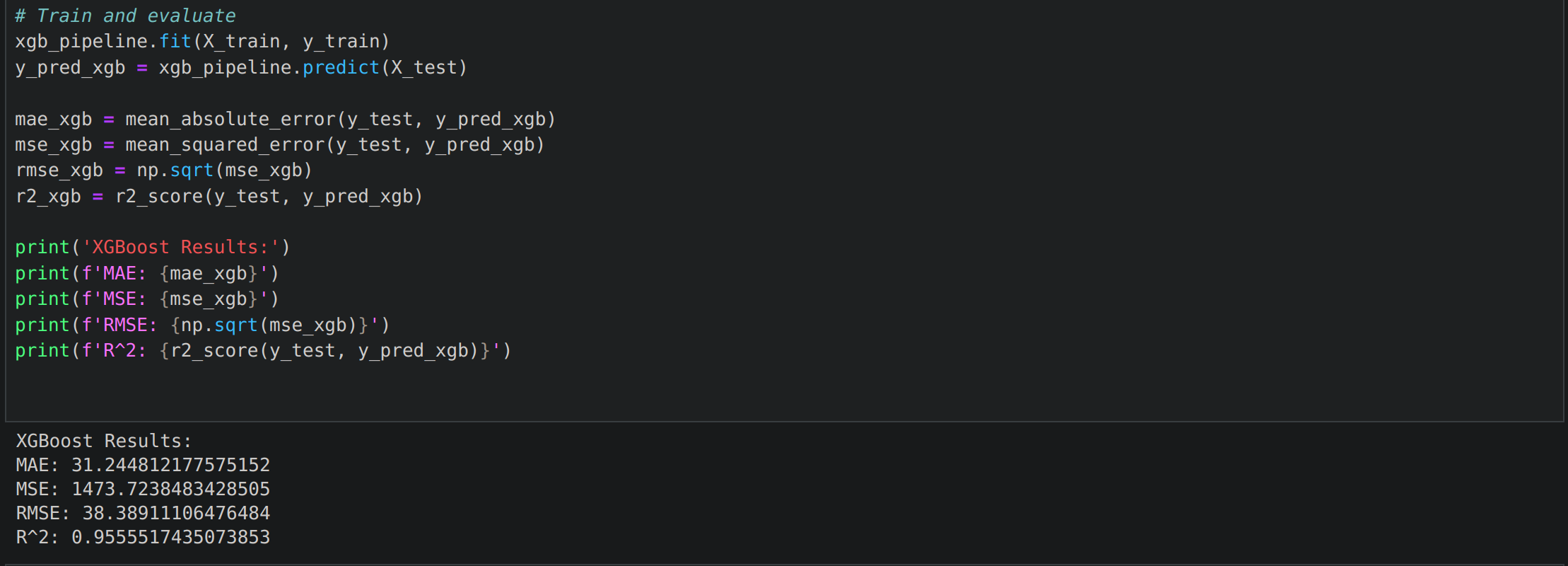
Linear Regression Model Creation

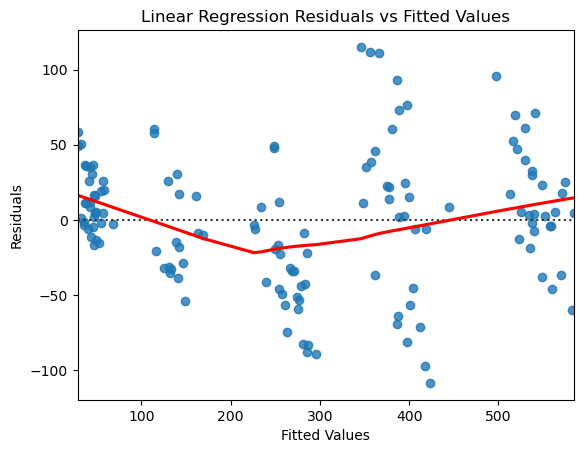


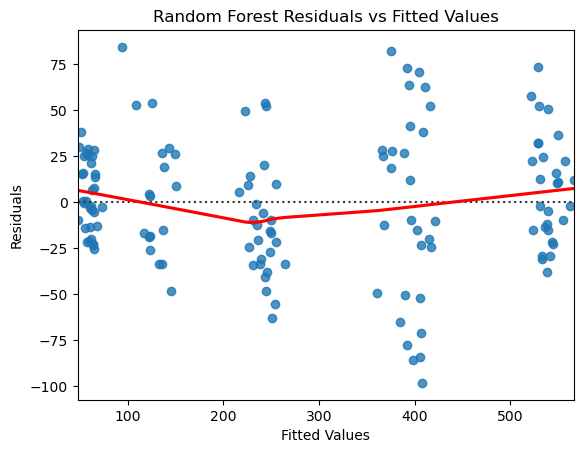
Random Forest Model Creation

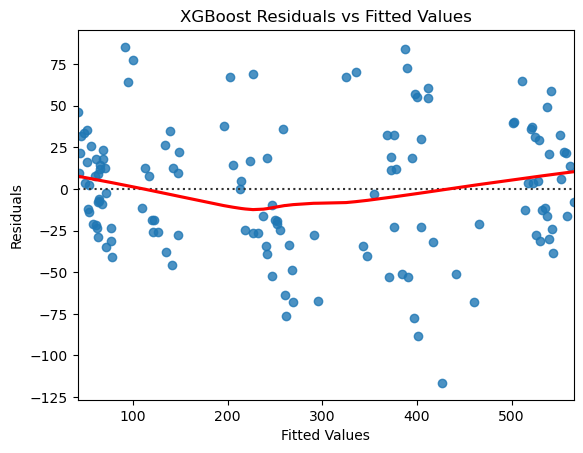


XGBoost Model Creation

Residual Plots for the models:







Ten Questions an audience would ask:

1. Why did you choose the Mobile Device Usage and User Behavior dataset?

Initially, I was looking at another dataset, but after I realized it did not have specific user information, I opted for this dataset instead. However, I would have liked to see more features included with this dataset.

1. How did you handle any incomplete or missing data?  
   A total of 11 records were removed because they were missing the Age of the user. Alternatively, I could have imputed the mean for the age in this fields, but I chose to just remove them to keep the data as realistic as possible.
2. What made the Random Forest model outperform the Linear Regression and XGBoost models?  
   The Random Forest model performed better than the Linear Regression because the ensemble of decision trees allows it to identify nonlinear relationships better. As for the XGBoost model, the random forest model probably performed better because I did not go through extensive hyperparameter tuning which can help the XGBoost model perform better. The Random Forest model was able to generalize better without the need for fine tuning.
3. Were there any features that surprised you by being less significant predictors of app usage?  
   Not after I performed the exploratory data analysis and discovered that there were not any significant differences in age groups. Initially, I suspected that there were would be considerable differences in app usage time between different age groups, but that was not the case with this dataset.
4. Did you find any unexpected trends or patterns in the data?  
   Not especially, as I said previously, I expected there to be a larger differential for app usage between age groups. However, this was not the case.
5. Do you think that the data might have potential bias?  
   No, I do not think that the dataset was biased.
6. What specific industries or organizations could benefit from the models developed in this project?  
   This project was developed as an educational exercise, but these types of models could be put to use by any business using an app that wants to understand its customer base better.
7. If you had a more comprehensive dataset, which new research questions would you want to explore?  
   I would want to see the types of apps people use the most, the time of day where most of the app usage occurs, and the impact that notifications have on app usage time.
8. Could your predictive model be integrated into mobile app development to optimize user engagement?  
   I think that a model similar to mine that was trained on the appropriate data could be used in this scenario, but not this model specifically.
9. What was the average app usage time?  
   The average app usage time for the entire dataset was 271 minutes or about 4 and a half hours a day.