Project B10: KAGGLE - Mobile Device Usage and User Behavior Dataset

Team member: Karina Pinajeva

Business understanding

Background

In today's digital landscape, mobile devices are an integral part of daily life. The data generated by mobile users - such as app usage time, screen-on time, battery consumption, and mobile data usage - holds valuable insights into user behavior. By analyzing this data, we can uncover patterns and trends that help understand how users interact with their devices. These insights are essential for mobile app developers, marketing teams, and telecom companies aiming to optimize user experiences, improve app performance, and create personalized advertising strategies. Understanding how users consume mobile data and interact with applications can directly inform decisions regarding product design, user engagement strategies, and service offerings.

Business Goals

- Analyze User Behavior Patterns: Understand the ways in which users engage with their mobile devices by identifying trends and patterns in app usage, screen time, battery consumption, and mobile data usage.
- Develop a Predictive Model for User Behavior Classification: Use machine learning techniques to develop a model that can predict user behavior based on the collected metrics. This model will allow businesses to classify users into different behavior categories and predict future behaviors.
- Identify Factors Influencing High Mobile Usage: Determine which variables contribute the most to higher mobile device usage. This analysis will provide valuable insights for targeted marketing campaigns, app optimization, and user engagement strategies.

Business Success Criteria

- Model Accuracy: The predictive model should achieve a high accuracy rate (at least 85%) in classifying user behavior.
- Insight Generation: The analysis should identify key factors that significantly influence mobile device usage, particularly those related to app engagement, battery consumption, and data usage.
- Business Application: The insights and models produced should lead to actionable strategies that improve user engagement, help optimize app performance, and enhance marketing efforts. Success will be measured by how well these strategies lead to business growth and user retention.

Inventory of Resources

People: The project team consists of a single individual, who is a student studying Computer Science. As the sole member of the project, I will take on the roles of data scientist, machine learning expert, and business analyst. I will also work independently to analyze mobile usage behavior. Additionally, I may consult resources, such as online guides or experts, to help with specific technical aspects and business insights as needed.

Data: The primary data source is a dataset containing mobile device usage information for 700 users. This dataset includes metrics such as app usage time, screen-on time, battery consumption, and mobile data usage.

Technology: The tools and technologies available for this project include Python, Jupyter Notebooks, and machine learning libraries like scikit-learn for model building. Data visualization tools such as matplotlib will be used to present the analysis results.

Requirements, Assumptions, and Constraints

To successfully complete this project, access to a clean and comprehensive dataset representing user behavior across different mobile platforms is essential. The dataset should include relevant metrics such as app usage, screen time, battery consumption, and mobile data usage to provide a complete picture of user behavior. Additionally, the necessary tools and computational resources are required to process, analyze, and model the data, including software like Python, machine learning libraries, and cloud or local computing resources. Finally, the ability to interpret the results through business insights is crucial for effectively applying the findings and ensuring that the analysis leads to actionable recommendations for improving user engagement, app optimization, and marketing strategies.

It is assumed that the dataset accurately represents a diverse range of mobile users, capturing a wide array of usage behaviors across different demographics and device types. Additionally, it is assumed that the behavior patterns observed in the data can be generalized, allowing for meaningful insights and predictions about user behavior. Furthermore, it is assumed that the collected data is sufficient for building robust and reliable models that can predict user behavior and identify key influencing factors.

Constraints:

- Time limitations: The project needs to be completed within a certain timeframe, which may restrict the depth of analysis.
- Data quality issues: The data may contain missing values, outliers, or inconsistencies that need to be addressed.

Risks and Contingencies

- Data Quality Issues: Missing or inconsistent data could impact the accuracy of the models.
- Privacy Concerns: Ensuring that user data is anonymized and handled securely is crucial. If the data contains sensitive information, this may limit the analyses or models that can be developed.

If data quality issues arise, alternative data imputation techniques or a more refined data-cleaning process can be used to handle missing values.

Terminology

- User Behavior: Refers to the way in which users interact with their mobile devices, including app usage time, screen time, battery usage, and data consumption.
- Mobile Data Usage: The amount of data consumed by users while using mobile apps or services.
- Predictive Model: A machine learning model used to forecast user behavior based on historical data.

Costs and Benefits

- Time and Resources: Significant time will be invested in data preparation, model building, and analysis.
- Computational Costs: Using cloud services or high-performance computing for model training may incur costs.

- Improved User Engagement: By identifying key user behavior patterns, businesses can optimize their apps to increase user engagement and retention.
- Enhanced Marketing Strategies: Insights into user behavior can be used to develop targeted marketing campaigns that lead to higher conversion rates.

- Operational Efficiency: Understanding the factors that drive high mobile usage can help optimize mobile services, reduce battery consumption, and improve user experience.

Data-Mining Goals

- Analyze User Behavior Patterns: Investigate how variables like app usage, screen time, battery consumption, and data usage interact to reveal trends and typical behavior.
- Develop a Predictive Model for User Behavior Classification: Build a machine learning model to classify users based on their mobile usage patterns and predict future behavior.
- Identify Factors Influencing High Mobile Usage: Identify key factors such as app types, screen time, and data consumption that drive higher mobile usage, aiding in app optimization and user engagement strategies.
- Model Accuracy: Achieve at least 85% classification accuracy in predicting user behavior.
- Pattern Identification: Identify key factors that influence mobile usage, providing actionable business insights.
- Model Validation: Ensure the model generalizes well and provides reliable predictions on new data.

Data understanding

The dataset used in this project contains information about mobile usage behavior for 700 users, which includes various metrics such as app usage time, screen-on time, battery consumption, data usage, and user demographics. The data is structured across multiple columns, each providing insights into user behavior.

To begin with, the data requirements include having a dataset that provides comprehensive user behavior metrics across different mobile platforms, such as app usage time, screen-on time, battery drain, and mobile data consumption. These metrics are essential for analyzing how users engage with their mobile devices. The dataset also includes demographic data, such as age, gender, and device model, which will allow further segmentation of user behavior.

Upon reviewing the data, it was confirmed that the dataset includes columns such as “App Usage Time (min/day),” “Screen On Time (hours/day),” “Battery Drain (mAh/day),” “Data Usage (MB/day),” and demographic information like age, gender, and device model. These variables are relevant for answering the research questions of the project. The “User Behavior Class” column, which categorizes users into different behavioral groups, is also available and will be useful for model classification tasks. The selection criteria for the data are focused on the columns that represent user behavior and the key metrics relevant for analysis: app usage, screen time, battery usage, data usage, and demographic information.

The dataset appears to be comprehensive, containing the necessary features for this analysis. However, during the initial review, some missing or inconsistent values were noticed. For example, there are some cases where data entries for certain metrics, such as screen-on time or data usage, are missing for specific users. These will need to be handled during the data cleaning process. For now, I ensured that all the required data for the analysis was intact, and irrelevant columns were omitted.

When exploring the data, I conducted a basic statistical analysis to understand the distribution of each variable. For instance, the "App Usage Time (min/day)" variable shows a range from 154 minutes to 393 minutes per day, suggesting a variety of usage patterns. Similarly, the "Screen On Time (hours/day)" ranges from 4 to 6 hours per day, with some users spending significantly more time on their devices than others. "Data Usage (MB/day)" ranges from 322 MB to 1122 MB per day, indicating a broad spectrum of data consumption habits. These variations in the data are critical to understanding user behavior and will be used to identify patterns that can predict mobile usage behaviors.

Regarding data quality verification, I observed that while most entries in the dataset are complete, some rows have missing values in specific columns. These missing values may impact the analysis and will need to be addressed by either filling in the missing values or removing the

corresponding rows, depending on the severity of the gaps. For now, initial steps have been taken to identify and document these issues, which will be further handled in the data preparation phase.

In conclusion, the dataset has been gathered and understood, and the necessary data has been identified for analysis. The quality of the data has been initially assessed, and the key fields required for answering the research questions have been confirmed. The next steps will involve data cleaning and preparation before proceeding with further analysis and modeling.

Project Plan

- Data Collection and Initial Inspection (~10 hours)

  Load and inspect the dataset for initial understanding. Identify key variables and check for missing data or inconsistencies. This task will involve exploring the dataset and identifying any immediate data quality issues to address.

- Data Preprocessing and Cleaning (~10 hours)

  Handle missing values, remove duplicates, and standardize the data to prepare it for analysis. This task involves cleaning the data, handling missing entries, and ensuring it is in a format suitable for modeling.

- Exploratory Data Analysis (EDA) (~12 hours)

  Perform statistical analysis and visualizations to identify trends, correlations, and patterns in the data. EDA will help gain insights into user behavior and allow for a better understanding of the data before building models.

- Model Development (~15 hours)

  Develop machine learning models (e.g., decision trees, random forests) to classify users based on their behavior. I will use tools like scikit-learn to build and evaluate predictive models, ensuring accuracy.

- Model Evaluation and Final Reporting (~10 hours)

  Evaluate the performance of the models using appropriate metrics (accuracy, precision, recall) and prepare the final report with insights. This task involves fine-tuning the models and summarizing the results and insights into a cohesive report.

Methods and Tools:

Programming Languages: Python

- Libraries: Pandas, scikit-learn, matplotlib, seaborn
- Tools: Jupyter Notebooks for coding and analysis
- Data Storage: Local storage or cloud (if necessary)

Important Notes:

As a single-member team, I am responsible for all tasks and time management. I will regularly review my progress to ensure timely completion and manage any potential delays in the project timeline.

Github link

https://github.com/kkerychka/Project-B10-KAGGLE---Mobile-Device-Usage-and-User-Behavior-Dataset