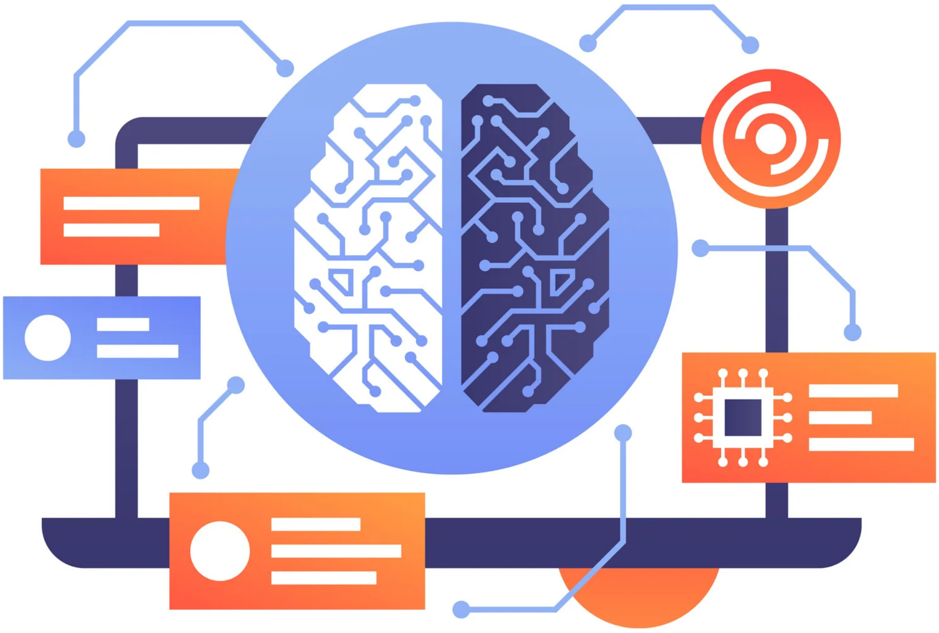
MODEL BUILDING PROCESS



Overview

Goals

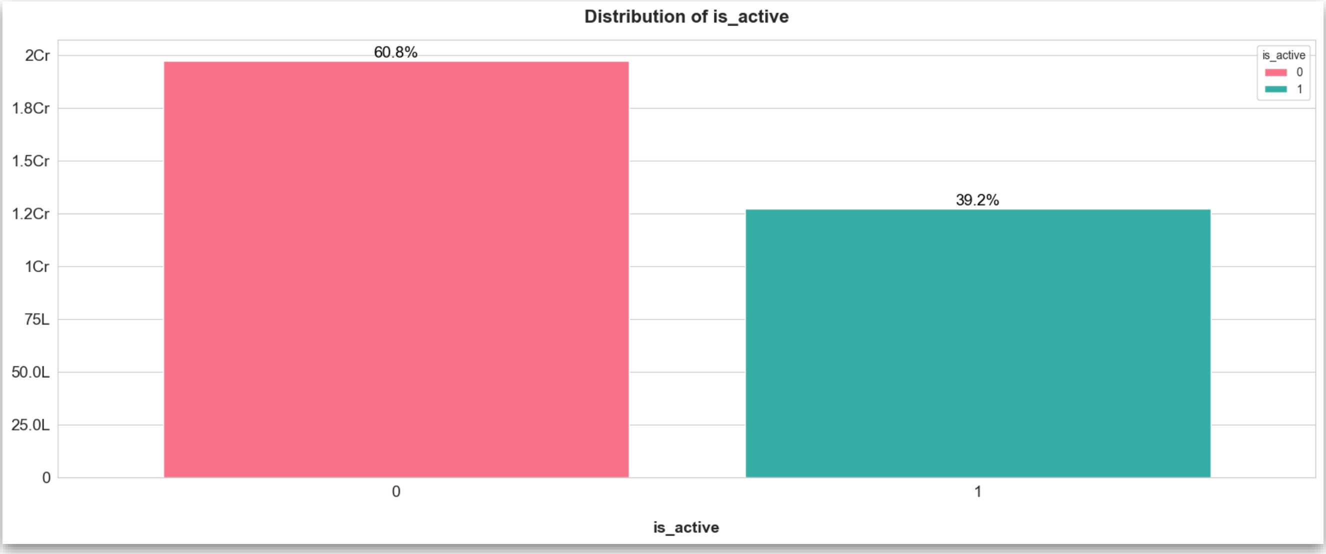
This report provides a detailed walk-through of the model-building process for predicting gender and age using machine learning techniques in a Jupyter Notebook environment. The analysis includes exploratory data analysis (EDA), feature engineering steps, and evaluation metrics for the models.

Data Analysis and Preprocessing

App Events

- **Observation:** No duplicates or missing values were found.
- **Action Taken:** No preprocessing was required for this dataset.

Distribution of Active Users



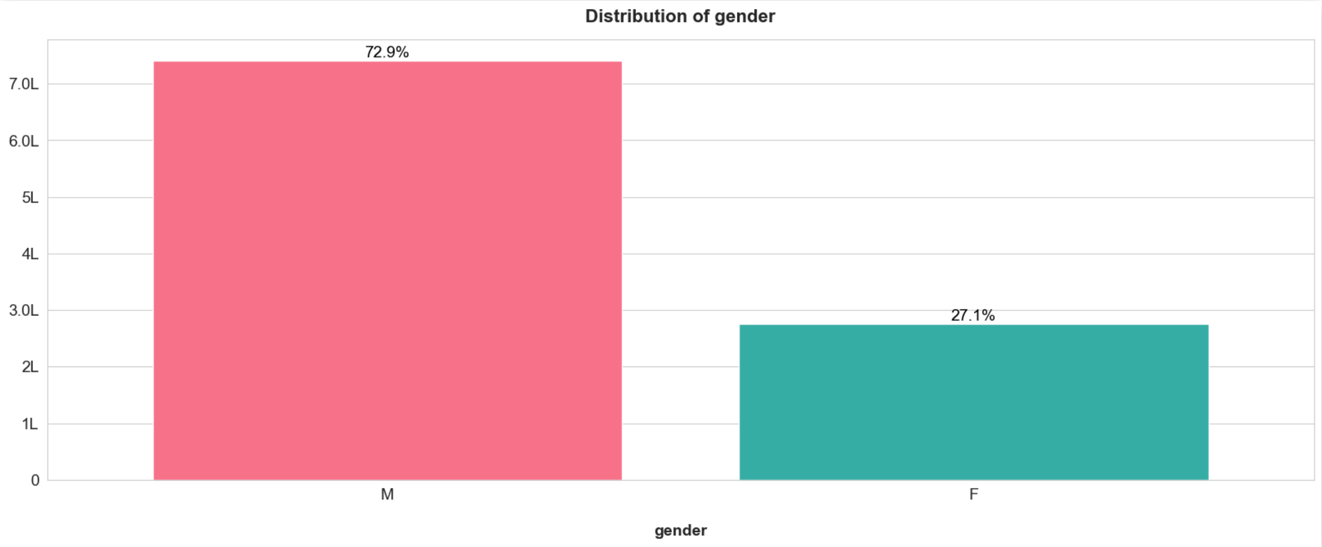
0 -Inactive, 1-Active

Train Event

- **Observation:** Approximately 4% of records had missing data.
- **Action Taken:** A new dataset was created to exclude records with missing data.
- **Additional Preprocessing:**
  - Device IDs with latitude or longitude =0 were identified.

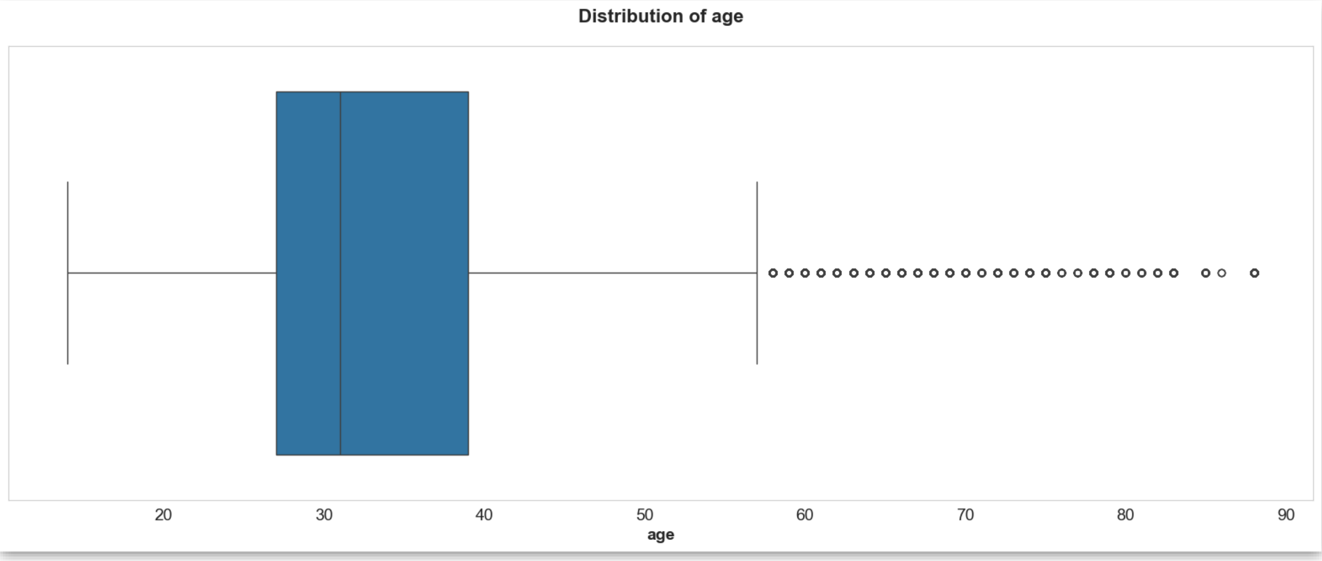
- For Device IDs with valid location data, the `latitude` and `longitude` values were replaced with the mean location of the respective Device ID.
- Records with no valid location data were dropped.

Distribution of Gender



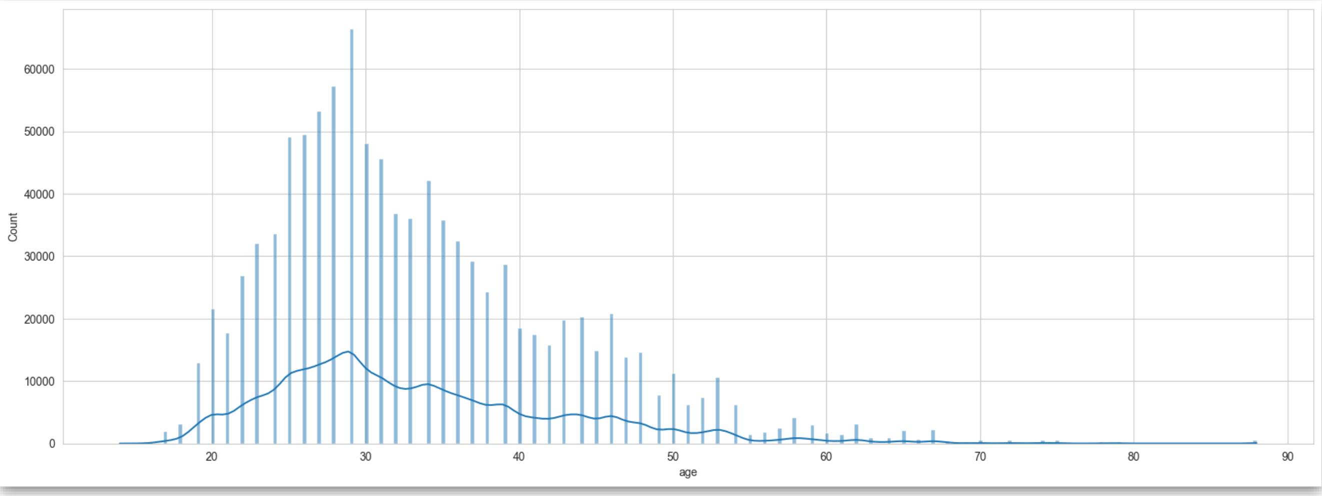
M –Male, F –Female

Distribution of Age



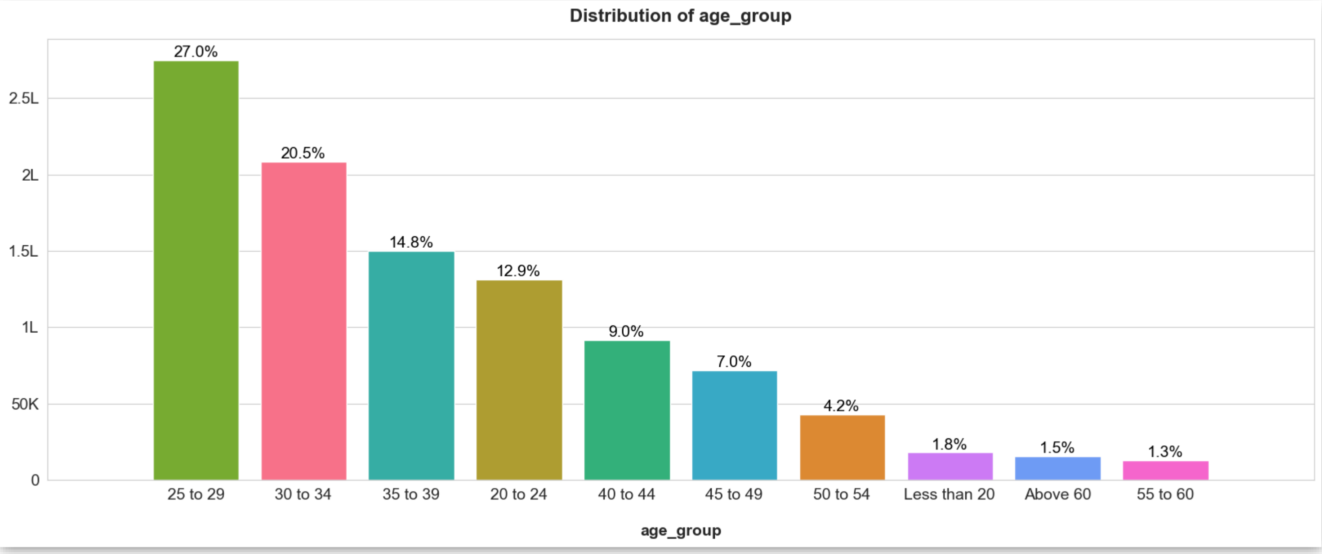
BoxPlot

Distribution of Age



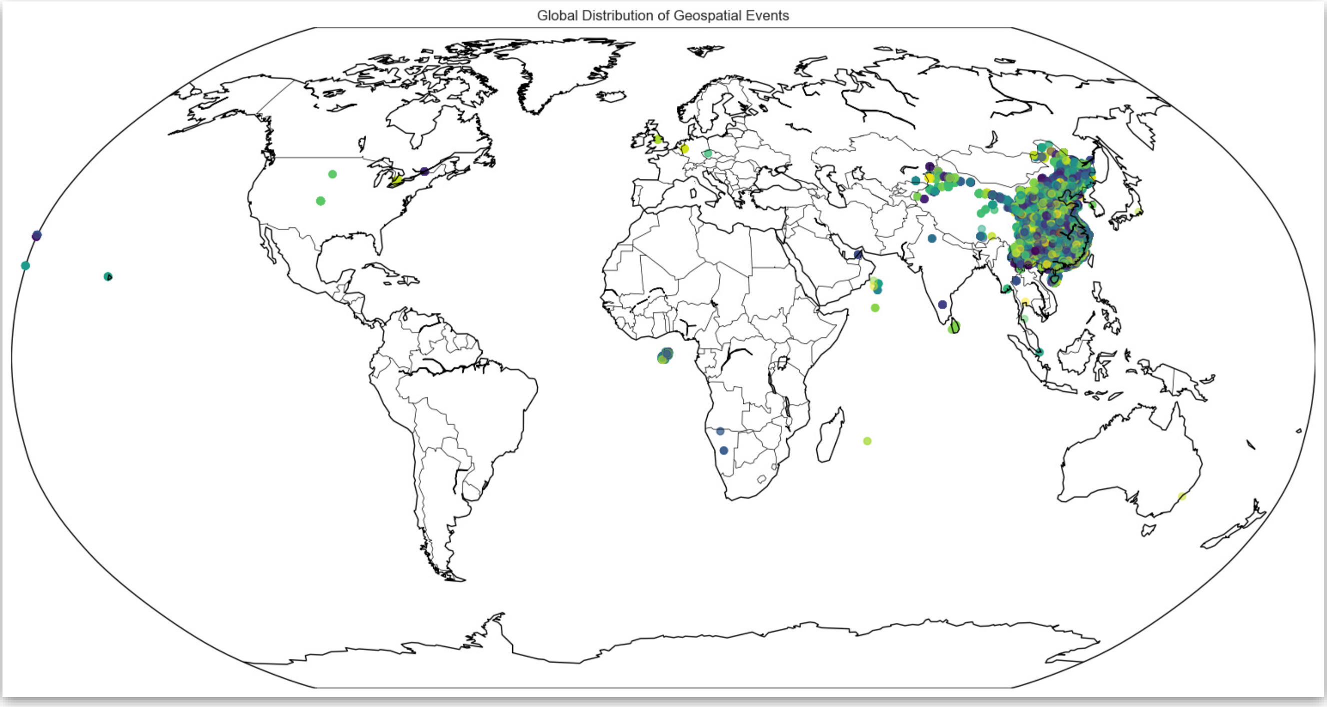
Histplot

Distribution of Age Group

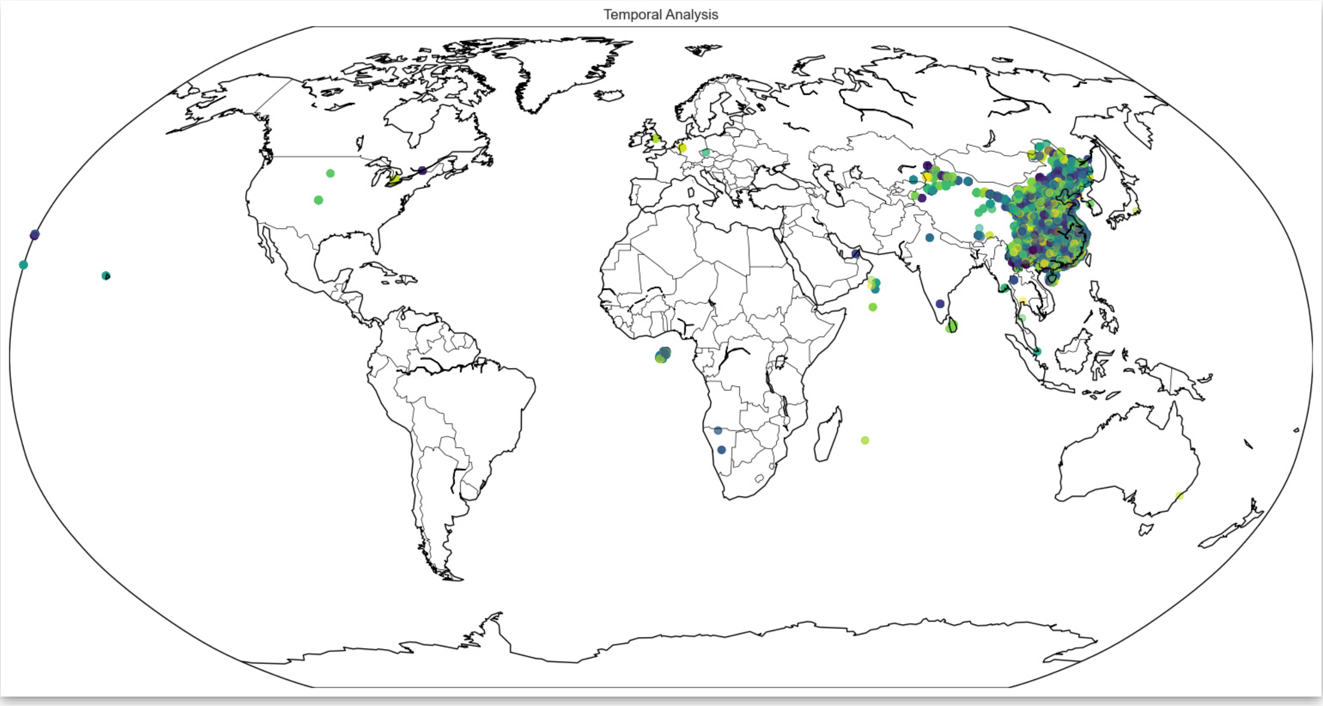


After grouping age for better representation

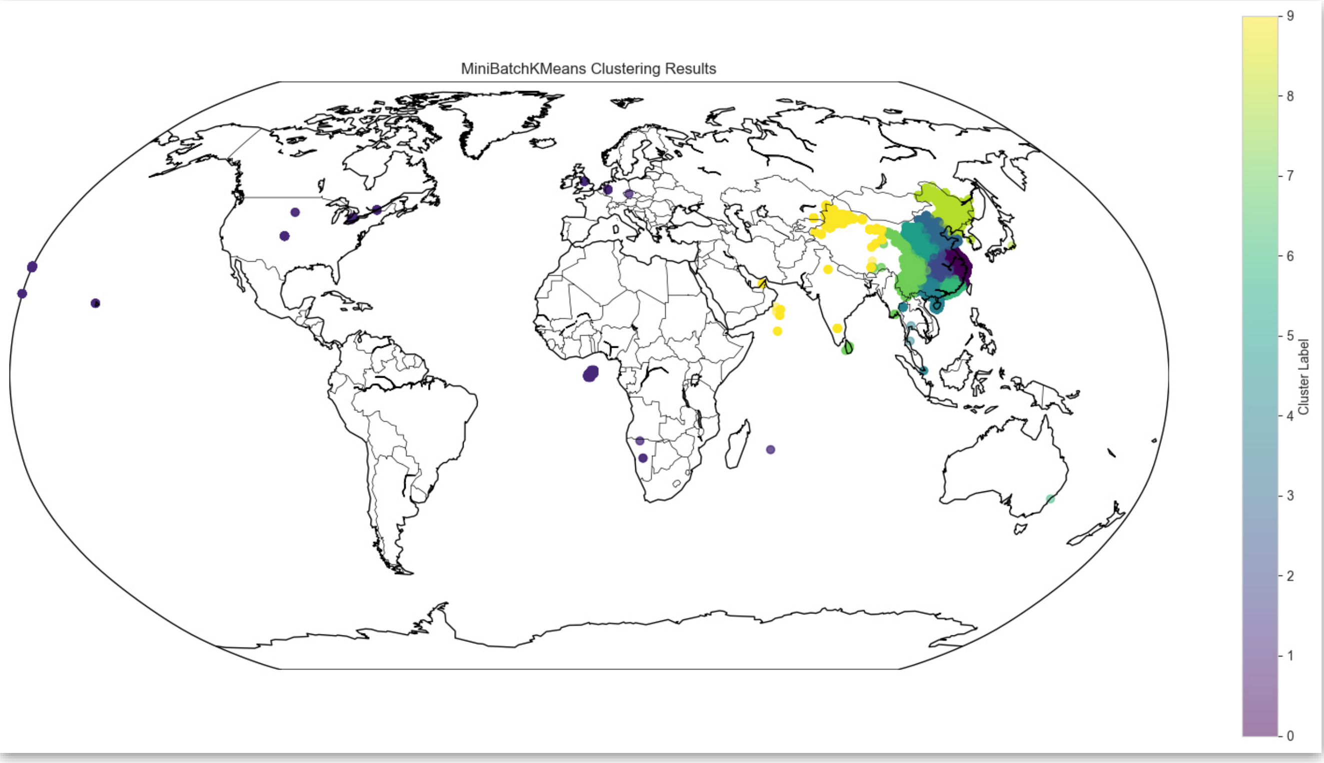
Geospatial DataDistribution



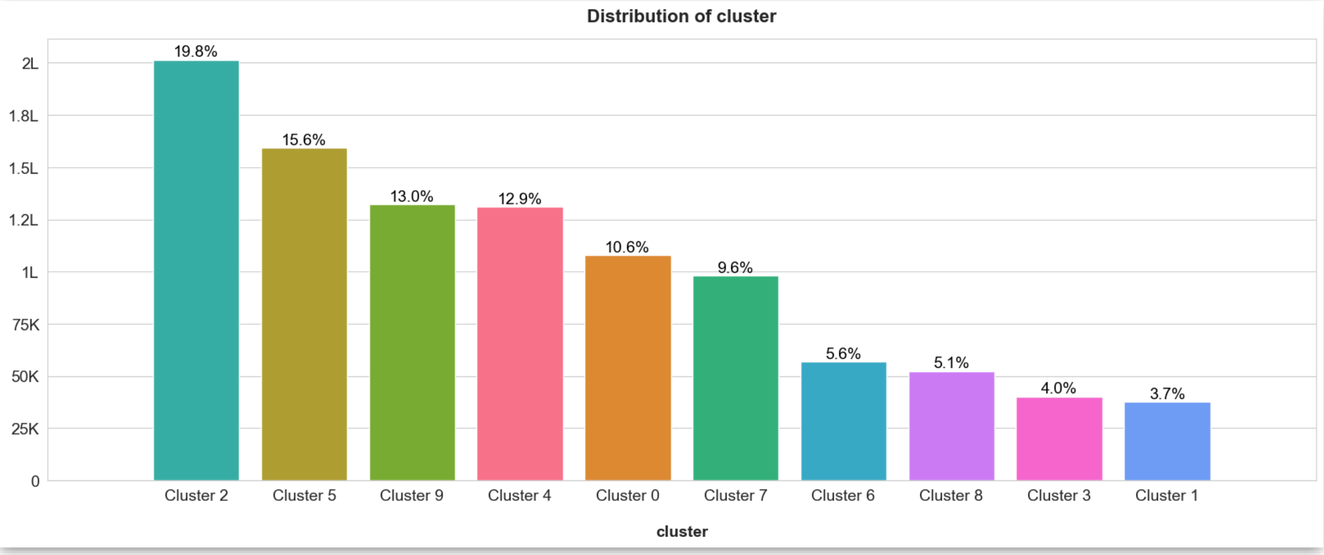
Temporal Analysis



Clustering



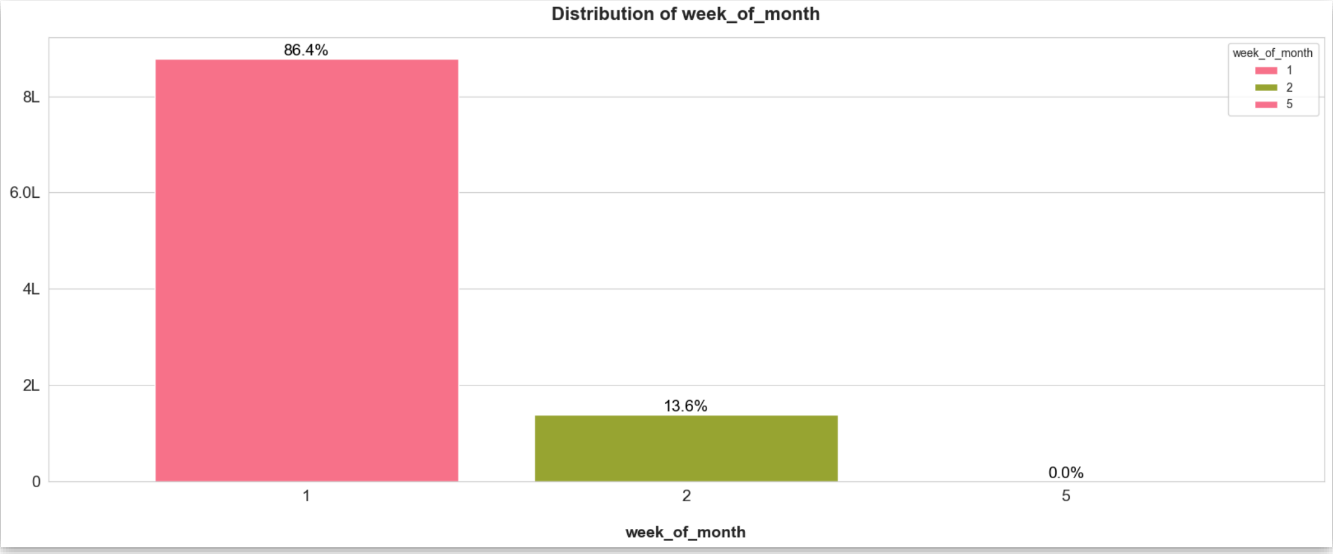
Clustering Distribution



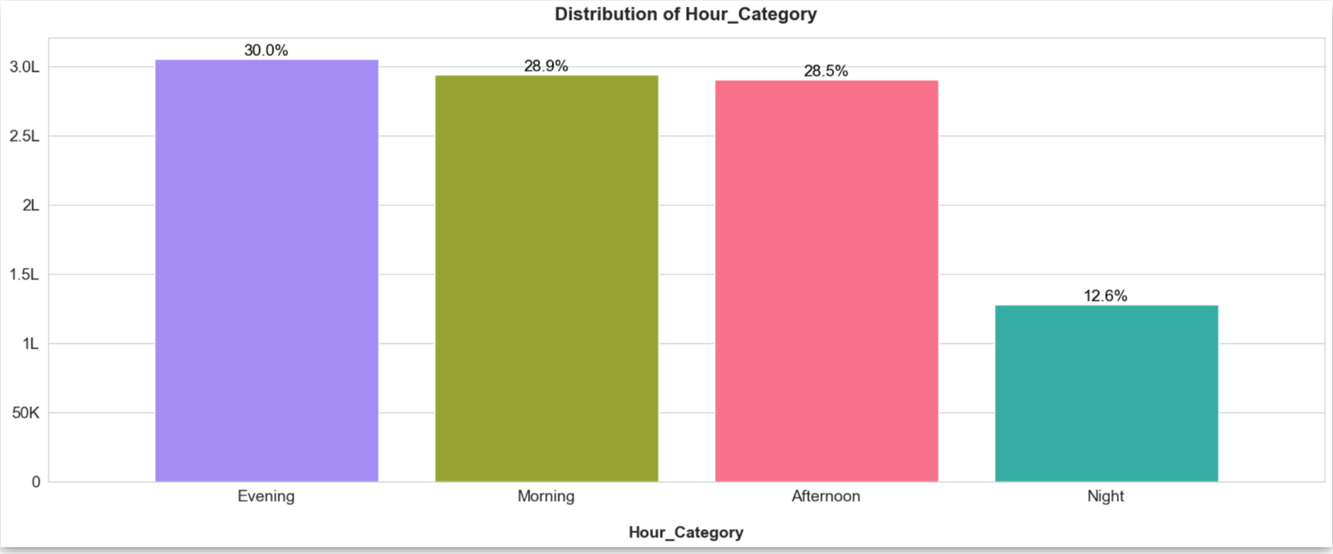
App Events Metadata

- **Observation:** No missing data or duplicates.
- **Action Taken:** Duplicates were removed to ensure data consistency.

App usage across weeks of amonth



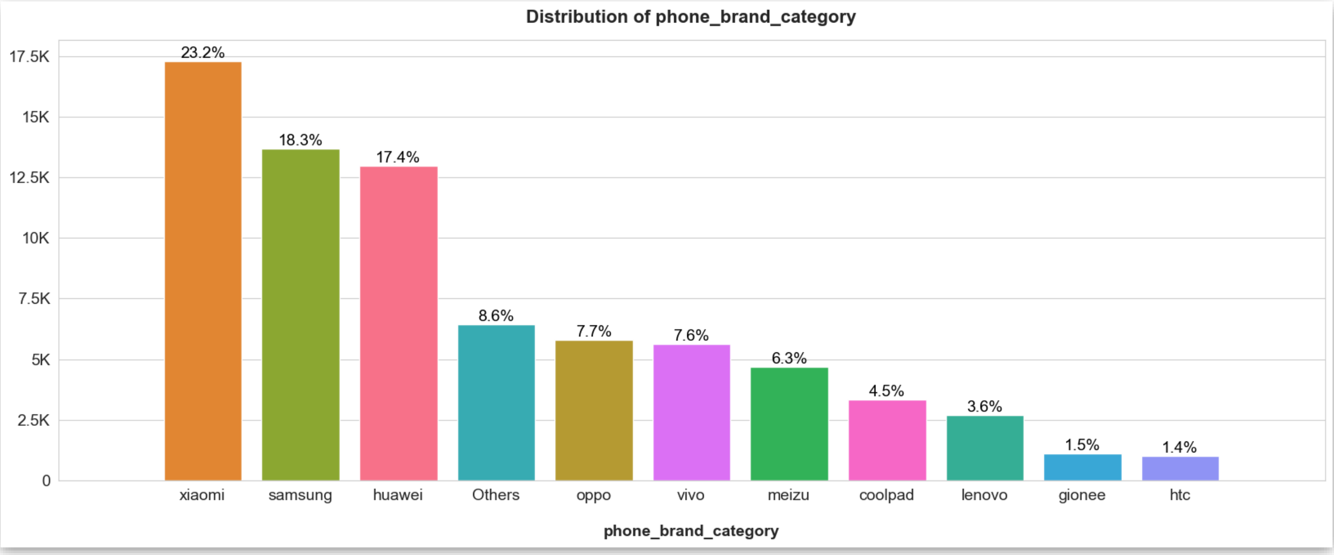
App usage across periods of aday



Train Mobile Brand

- **Observation:** No missing data or duplicates.
- **Action Taken:** Duplicates were removed.

Mobile Phone Distribution



Feature Engineering and Model Building

The feature engineering steps included:

1. Handling categorical variables (e.g., mobile brand) using appropriate encoding techniques.
2. Normalizing numerical features using `StandardScaler` to ensure all features contribute equally to the model performance.
3. Splitting the dataset into training and testing sets.

Gender Prediction Models

- **XGBoost:** Achieved an accuracy of 70.73% in Scenario 1,with a precision of 72.57%, recall of 67.04%, and F1score of 69.69%.
- **Logistic Regression:** Performed poorly compared to XGBoost, achieving only 57.64%accuracy in Scenario 2.
- **Random Forest Classification:** Also performed less effectively than XGBoost for gender prediction.

Age Prediction Models

- **Random Forest Classification:** Achieved near-perfect performance with an accuracy of 95.88%, precision of 95.86%, recall of 95.88%, and F1score of 95.86%.
- **XGBoost:** Performed moderately well, achieving an accuracy of 49.84% in Scenario 2.

Source code

The code is published on GitHub at,

https://github.com/nimunik/Ad-Campaign-Predictor.git

Model Evaluation

Gender Prediction

Model	Accuracy	Precision	Recall	F1Score
XGBoost (Scenario 1)	70.73	72.57	67.04	69.69
XGBoost (Scenario 2)	82.31	84.83	78.66	81.63
Logistic Regression	57.64	56.55	65.54	60.72

Age Prediction

Model	Accuracy	Precision	Recall	F1Score
Random Forest	95.88	95.86	95.88	95.86
XGBoost	49.84	50.51	49.84	47.88

Final Model Selection

Gender Prediction

- **Best Model:** XGBoost (Scenario 2) achieved the highest performance metrics, including accuracy (82.31%),precision (84.83%), recall (78.66%), and F1score (81.63%).
- **Reasoning:** XGBoost consistently outperformed other models like Logistic Regression and Random Forest for gender prediction.

Age Prediction

The trained models were exported to file format using `joblib.dump()`for future use:

- **Gender Prediction:** `age_compressed.joblib`with compression level set to 9.
- **Age Prediction:** `gender.joblib`.

Conclusion

This analysis demonstrates that XGBoost and Random Forest are the most effective models for gender and age prediction, respectively. The choice of model was based on rigorous evaluation of performance metrics, ensuring optimal results for each task.