Student(s): Long Le 2203558

Title of the work: Practical Work 1 – Mobile Device Usage and User Behaviour Dataset

Predicting Battery Drain Using Machine Learning

Used algorithms: Random Forest Regressor, Logistic regression

Work description and analysis:

Description of the work:

- The dataset, user_behavior_dataset.csv, includes user behaviour and device data.
- The goal is to Predict daily battery drain (in mAh/day) and classify user behaviour into predefined categories based on mobile usage patterns and demographic data.
- The dataset contains features such as app usage time, screen-on time, number of apps installed, data usage, and demographic data. The goal is to develop a predictive model to estimate battery drain and identify the most influential factors contributing to battery consumption.

Data preparation for the training:

Dataset Overview:

The dataset includes:

- Numerical Features: App Usage Time, Screen On Time, Number of Apps Installed, Data Usage, Age.
- Categorical Features: Device Model, Operating System, Gender.
- Target Variables:
    - Battery Drain (Regression task).
    - User Behavior Class (Classification task).

df - DataFrame

Index	User ID	Device Model	erating Syste	age Time (m	On Time (hou	y Drain (mAh	er of Apps In	Usage (MB/	Age	Gender	r Behavior Cl
0	1	Google Pixel 5	Android	393	6.4	1872	67	1122	40	Male	4
1	2	OnePlus 9	Android	268	4.7	1331	42	944	47	Female	3
2	3	Xiaomi Mi 11	Android	154	4	761	32	322	42	Male	2
3	4	Google Pixel 5	Android	239	4.8	1676	56	871	20	Male	3
4	5	iPhone 12	iOS	187	4.3	1367	58	988	31	Female	3
5	6	Google Pixel 5	Android	99	2	940	35	564	31	Male	2
6	7	Samsung Galaxy S21	Android	350	7.3	1802	66	1054	21	Female	4
7	8	OnePlus 9	Android	543	11.4	2956	82	1702	31	Male	5
8	9	Samsung Galaxy S21	Android	340	7.7	2138	75	1053	42	Female	4
9	10	iPhone 12	iOS	424	6.6	1957	75	1301	42	Male	4
10	11	Google Pixel 5	Android	53	1.4	435	17	162	34	Female	1
11	12	OnePlus 9	Android	215	5.5	1690	47	641	24	Male	3
12	13	OnePlus 9	Android	462	6.2	2303	65	1099	57	Female	4
13	14	Xiaomi Mi 11	Android	215	4.9	1662	43	857	43	Male	3
14	15	iPhone 12	iOS	189	5.4	1754	53	779	49	Female	3

Figure 1. First 14 rows with all columns of training data

Preprocessing steps:

1. Dropped Irrelevant Columns: Removed User ID as it has no predictive value.
2. Encoded Categorical Variables: Used mapping and OneHotEncoder for categorical features.
3. Normalized Numerical Features: Applied StandardScaler for consistent scaling.
4. Split Data: Partitioned into training (70%) and testing (30%) sets.Sample Training Data:

App Usage Time	Screen On Time	Number of Apps Installed	Data Usage	Age	Device Model	Operating System	Gender
393	6.4	67	1122	40	Google Pixel 5	Android	Male
268	4.7	42	944	47	OnePlus 9	Android	Female
154	4.0	32	322	42	Xiaomi Mi 11	Android	Male

Relevant metrics for the case(s):

Regression Metrics:

- Mean Absolute Error (MAE): Measures the average magnitude of errors in predictions.
- Mean Squared Error (MSE): Penalizes larger errors more significantly.
- Root Mean Squared Error (RMSE): Square root of MSE, easier to interpret.
- R-squared (R^2): Proportion of variance explained by the model.

Classification Metrics

- Accuracy: Overall correctness of predictions.
- Confusion Matrix: Breakdown of true positives, false positives, etc.
- Precision, Recall, F1-Score: Measure of classification quality, especially for imbalanced data.

Conclusions of the results:

Random Forest Regressor (Battery Drain Prediction):

- **Performance Metrics**:
    - **R^2**: 0.95
    - **MAE**: 148.96 mAh/day
    - **RMSE**: 176.77 mAh/day

```
[15 rows x 11 columns]
Mean Absolute Error (MAE): 148.95990476190477
Mean Squared Error (MSE): 31249.32521238095
Root Mean Squared Error (RMSE): 176.7747866987285
R-squared (R²): 0.9516723527485257
Predicted Battery Drain for new user: [1415.63] mAh/day
```

Figure 2. Regression (Random Forest) metrics result

- **Model Analysis**:
  The Random Forest Regressor achieved outstanding performance, explaining 95% of the variance in battery drain. Its low MAE and RMSE indicate consistent and accurate predictions.
- **Model Usability**:
  The model is practical for real-world applications, such as optimizing app usage to save battery life

Logistic Regression (User Behaviour Classification):

- **Performance Metrics**:
    - **Accuracy**: 73.81%
    - **Confusion Matrix**:
    - **Precision, Recall, F1-Score**: Moderate to high for most classes, with notable strength in Class 1 and Class 5.

```
Accuracy: 0.7380952380952381
Confusion Matrix:
 [[38  8  0  0  0]
 [ 0 27  6  1  0]
 [ 0  8 31 10  1]
 [ 0  0 13 31  3]
 [ 0  0  0  5 28]]
Classification Report:
              precision    recall  f1-score   support

           1       1.00      0.83      0.90        46
           2       0.63      0.79      0.70        34
           3       0.62      0.62      0.62        50
           4       0.66      0.66      0.66        47
           5       0.88      0.85      0.86        33

    accuracy                           0.74       210
   macro avg       0.76      0.75      0.75       210
weighted avg       0.75      0.74      0.74       210
```

Figure 3. Classification (Logistic Regression) result

- **Model Analysis**:
  Logistic Regression effectively distinguishes certain behaviour classes, such as Class 1 (precision: 1.00) and Class 5 (precision: 0.88). However, moderate performance for Classes 2, 3, and 4 indicates room for improvement.

Comparison of Models:

- **Best Performance**:
  For regression tasks, the Random Forest Regressor outperformed Linear Regression with an R^2 of 0.95.
  For classification, Logistic Regression achieved reasonable accuracy but requires enhancement to reduce misclassifications.
- **Model Usability**:
  Both models are suitable for practical applications in their respective tasks, with Random Forest providing more reliable predictions.

Potential Improvements:

1. Add more diverse and recent data to improve generalization.
2. Hyperparameter tuning for Logistic Regression and Random Forest.
3. Test advanced algorithms like XGBoost or Gradient Boosting.