Data Science Project Review: How to use health fitness data to help users lead more fulfilled lives

Table of Contents

1. Project Summary/Executive Summary

The objective of this work was to predict lifestyle using health fitness tracker data at an accuracy of 95% or better.

The core problem that this work aims to solve is user happiness with respect to exercise.

The Baseline Solution is to use the features mean_workout_minutes and mean_steps per device ID to predict lifestyle. The user's of the health fitness tracker device will benefit from the Baseline Solution, because they will receive exercise product and event information related to their lifestyle group and a similar lifestyle group.

The methodology used to implement the baseline solution, train and test lifestyle prediction models, is classification analysis using logistic regression; the time required is minimal and it does not require additional data or money. The evaluation metric includes prediction accuracy for trained and test datasets.

In summary, the project/business objective was SATISFIED because the baseline solution produced an analysis result showing that lifestyle can be predicted at 100% accuracy for both training and test data. The project can be improved by measuring happiness and monitoring long term changes in predicted lifestyle, to understand whether the user feedback related to lifestyle improves quality of life.

2. Business Objective

The objective of this work is to predict lifestyle using health fitness tracker data at an accuracy of 95% or better.

3. Project Background

The core problem that this work aims to solve is user happiness with respect to exercise.

Happiness and quality of life are strongly linked to lifestyle, and lifestyle is often driven by one's environment [https://worldhappiness.report/ed/2022/]. If one does not have access to information regarding their likes and desires, they may missout on situations/events that could have improved their quality of life and thus their happiness.

In the week 4 project, we had access to a dataset with 3000 user fitness profile's, where each user had registered data everyday for 365 days. Exploratory Data Analysis (EDA) was performed by investigating eight questions regarding the health tracker fitness data. The questions are as follows:

1. Do people workout more or less on weekend, than during the weekday? They workout the same amount on the weekend and weekday (non-statistically, they workout less on the weekday)
2. Are weekend exercisers likely to have a ('Sedentary') lifestyle versus weekday exercisers? No, the probability of Sendentarity is 0.016 for the weekend and weekday
3. Does the number of hours of exercise predict vo2? No
4. What features are most important for predicting vo2 ('natural healthiness')? mean_resting_heartrate' is the best predictor of vo2,
5. Are there more or less types of lifestyle groups? Looking at the Variance between centroids and data, and the Bayesian information criterion (BIC) the best number of clusters are either 3 or 7. So there could probably be 'Sedentary', 'Moderate', 'Active' lifestyle categories.
6. Can we detect what types of exercise people perform (type_of_exercise_group), using steps and workout_minutes? Can we label each of the new groups by looking at mean_workout_minutes and mean_steps? Do they correlate with lifestyle groups? There are 4 types of type_of_exercise_group ('Running', 'Weights_Gym', 'Sedentary', 'Walking'). And the 4 types of exercise groups do not correlate with lifestyle groups (corr=-0.037857).
7. Can the number of minutes of exercise and steps predict the lifestyle? Yes! With the two combined they perfectly predict lifestyle! Mean_steps has a prediction accuracy of 0.35 and mean_workout_minutes has a prediction accuracy of 0.66.
8. Can we predict the type_of_exercise_group using heart rate? We know that the type_of_exercise_group was found by clustering using steps and workout_minutes, but is heart_rate a distinguishing feature to detect these type_of_exercise_group? Using logistic regression we can predict type_of_exercise_group, with resting and active heart rate, with an accuracy of 0.62.

To summarize, the eight questions focused on the number of exercise minutes as a predictor for activity such as; weekday/weekend behavior, vo2 activity, the types of exercise performed, lifestyle. Additionally, I found that heart rate could be used to predict natural healthiness (vo2) and the types of exercise performed.

4. Baseline Solution

The Baseline Solution is to use the features mean_workout_minutes and mean_steps per device ID to predict lifestyle.

5. Project Deliverables: Methodology & Evaluation metrics

5.1 Methodology

The methodology used to implement the baseline solution, train and test lifestyle prediction models, was classification analysis using logistic regression; the time required to perform the analysis was minimal and it did not require additional data or money.

5.2 Evaluation metrics

Classification prediction accuracy was used to evaluate model performance for both train and test datasets; default percentage values of 0.75 and 0.25 were used for the train and test dataset respectively.


```python
X = ht_agg_pandas_df[['mean_workout_minutes', 'mean_steps']]

# y = [int(i) for i in ht_agg_pandas_df['mean_lifestyle_num'].to_numpy()]
# OR
le = LabelEncoder()
lifestyle = ht_agg_pandas_df['lifestyle']
le.fit(lifestyle)
y = le.transform(lifestyle)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y)

lr = LogisticRegression(max_iter=10000)
lr.fit(X_train, y_train)

y_train_predicted = lr.predict(X_train)
y_test_predicted = lr.predict(X_test)

print("training accuracy: ", accuracy_score(y_train, y_train_predicted))
print("test accuracy:     ", accuracy_score(y_test, y_test_predicted))
print("training confusion matrix")
print(confusion_matrix(y_train, y_train_predicted))
print("")
print("test confusion matrix")
print(confusion_matrix(y_test, y_test_predicted))
```

```
training accuracy:  1.0
test accuracy:      1.0
training confusion matrix
[[646   0   0   0]
 [  0 790   0   0]
 [  0   0 235   0]
 [  0   0   0 579]]

test confusion matrix
[[213   0   0   0]
 [  0 274   0   0]
 [  0   0  77   0]
 [  0   0   0 186]]
```

Command took 1.05 seconds -- by j622amilah@gmail.com at 20/01/2023 15:54:15 on temp

The analysis results show that lifestyle can be predicted at 100% for both train and test datasets, with logistic regression classification using mean_workout_minutes and mean_steps as features.

6. Recommendations for improving the project in the future

In summary, the project/business objective was SATISFIED because the baseline solution produced an analysis result showing that lifestyle can be predicted at 100% accuracy for both training and test data.

The project can be improved by measuring happiness and monitoring long term changes in predicted lifestyle, to understand which types of user feedback related to lifestyle improve quality of life. Long-term monitoring includes both: 1. retraining the deployed model to keep the model up-to-date with respect to the user's current information and 2. evaluating the effectiveness of the user feedback to improve happiness. The following diagram shows that the green, red, and orange steps are recommended for future project success for monitoring and predicting exercise lifestyle. Retraining the model every month will satisfy model maintenance and allow for current accurate lifestyle predictions.