Predicting Workout Quality to Help Coaches Support Sportspeople

Ludovico Boratto
Data Science and Big Data Analytics
EURECAT, Centre Tecnológic de Catalunya
Barcelona, Spain
ludovico.boratto@acm.org

Salvatore Carta
Dip.to di Matematica e Informatica
Università di Cagliari
Cagliari, Italy
salvatore@unica.it

Walid Iguider
Dip.to di Matematica e Informatica
Università di Cagliari
Cagliari, Italy
w.iguider@studenti.unica.it

Fabrizio Mulas
Dip.to di Matematica e Informatica
Università di Cagliari
Cagliari, Italy
fabrizio.mulas@unica.it

Paolo Pilloni
Dip.to di Matematica e Informatica
Università di Cagliari
Cagliari, Italy
paolo.pilloni@unica.it

ABSTRACT

The support of a qualified coach is crucial to keep the motivation of sportspeople high and help them pursuing an active lifestyle. In this paper, we discuss the scenario in which a coach follows sportspeople remotely by means of an eHealth platform, named *u4fit*. Having to deal with several users at the same time, with no direct human contact, means that it is hard for coaches to quickly spot who, among the people she follows, needs a more timely support. To this end, in this paper we present an automated approach that analyzes the adherence of sportspeople to their planned workout routines. The approach is able to suggest to the coach the sportspeople who need earlier support due to a poor performance. Experiments on real data, evaluated through classic accuracy metrics, show the effectiveness of our approach.

CCS CONCEPTS

• **Information systems → Mobile information processing systems**; **Data mining**;

KEYWORDS

Personalized Persuasive Technologies, Health Recommendation, Healthy Lifestyle, eCoaching, Motivation.

ACM Reference Format:
Ludovico Boratto, Salvatore Carta, Walid Iguider, Fabrizio Mulas, and Paolo Pilloni. 2018. Predicting Workout Quality to Help Coaches Support Sportspeople. In *Proceedings of the Third International Workshop on Health Recommender Systems co-located with Twelfth ACM Conference on Recommender Systems (HealthRecSys'18), Vancouver, BC, Canada, October 6, 2018* , 5 pages.

1 INTRODUCTION

A regular physical activity is key to keep a good health [22]. In order to keep motivation high, eHealth persuasive technologies (eHPT) are designed to help people change their habits and to help them overcome their frictions to healthier behaviors [7, 8, 10].

The *u4fit* platform[1] connects users with human coaches, allowing for a tailored exercise experience at a distance [1, 14]. Indeed, users receive tailored workout plans from coaches and, thanks to a mobile application, they are guided to execute the workout correctly. Moreover, coaches receive the results of a workout and can interact with the users via a live chat.

However, a coach usually follows a lot of sportspeople so, after a workout, it is not trivial to understand which sportsperson should be supported first (e.g., who should she chat with). Indeed, a training result is made up of several metrics to be carefully analyzed (e.g., speed and covered distance, just to name a few), so the effectiveness of a workout cannot be easily and quickly estimated.

To face the problem of helping coaches support first the sportspeople that performed a poor workout (since they are, trivially, those who need the most urgent support), in this paper we propose an approach that predicts the quality of a workout result by means of a rating. Based on the features that characterize previous workouts and the ratings assigned to them by the coaches, we train a classifier to predict the rating of the new workouts that the coach has not considered yet. This allows us to recommend to the coach the workouts (and, thus, the sportsperson who performed it), ordered by increasing predicted rating (i.e., those with a low rating are presented first), allowing the coach to take action[2].

Being able to provide effective and timely support to the users who need the most support is a powerful form of motivation that it is crucial for long-term adherence to a training routine [13].

Recommender systems (RS) can help supporting decisions in health environments. As highlighted in [23], when a RS is developed for health professionals (as in our case) they provide information that allows them to address specific cases. Moreover, health RS help providing reliable and trustworthy information to the end users [23].

HealthRecSys'18, October 6, 2018, Vancouver, BC, Canada

[1]www.u4fit.com. Please note that the coaches marketplace is visible only by setting the Italian language on the platform.
[2]In case two users need equally urgent support, different strategies can be carried out, such as supporting first the elder sportperson, or the one who has not received support for a longer amount of time. These decisions on how to rank the equally important cases goes beyond the scope of our paper and are left as future work, when the approach will be implemented in the *u4fit* platform.

Table 1: Samples count for each rating

Rating	Count
1	216
2	723
3	994
4	977
5	683

The goal of health RS is usually to lead to lifestyle changes [20], to support users who are losing motivation when exercizing [15], and to improve the patients' safety [5]. Readers can refer to [3] for a survey on health RS.

To the best of our knowledge, no recommender system can help coaches by suggesting them the sportspeople that need more timely support. This approach can help coaches to provide focused interventions in order to motivate poor performing users. Indeed, coaches can intervene quickly to persuade users change their negative attitude towards physical activity so that to favor a longer-term adherence to their training routines. More specifically, our contributions are the following:

- we provide, for the first time in the literature of health RS, an approach that recommends to a coach the sportspeople she follows who need timely support, considering the workouts they recently performed and that the coach has not considered yet;
- we validated our proposal on a real-world dataset made up of approximately 3 years of data, by comparing different classifiers on standard accuracy metrics;
- our solution can be embedded in real-world persuasive eHealth systems, thus finding practical and effective applications.

We organize the rest of the paper as follows: in Section 2 we introduce the dataset and in Section 3 we present the techniques we employed to preprocess the data. Section 4 presents the classifiers we considered in this study, while in Section 5 we present the experimental framework and results. We conclude the paper in Section 6, with some final remarks and future developments.

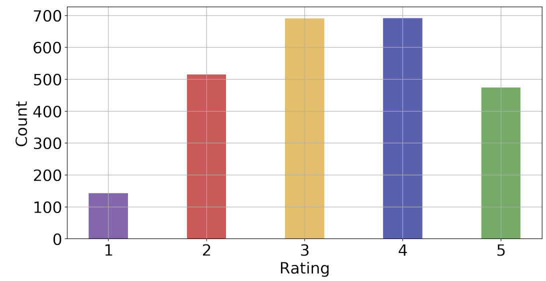
2 DATASET

This research work is based on data collected by means of the u4fit platform. The dataset contains 3593 workouts, which u4fit coaches evaluated by assigning a rating ranging between 1 (poorly performed) and 5 (well performed). Each workout result is represented by the following aggregate statistics:

- Covered distance (in meters);
- Workout duration (in seconds);
- Rest time (in seconds);
- Average speed (in km/h);
- Maximum speed (in km/h);
- User age;
- User gender;
- Burnt calories.

Ratings were distributed as described in Table 1, where "count" indicates the number of samples having the corresponding rating.

Figure 1: Ratings distribution



The workouts we considered are those performed by means of the u4fit mobile app. Indeed, we excluded those performed by means of running watches, since users have to program their workout routines manually and sometimes the workouts do not match painstakingly the workout built by the coach. Instead, users of the mobile application receive their workout plan seamlessly inside the app, so the performed workouts always match those designed by their coaches. This allows the coaches to make a fair evaluation of the workout.

As we are dealing with real-world data, the main issues we encountered were the data imbalance and the small size of the minority classes, as we can clearly notice from Figure 1 that represents graphically the distribution of ratings.

3 PREPROCESSING

Most Machine Learning classifiers get into trouble when dealing with imbalanced data, given that the learning phase of classifiers may be biased towards the instances that are frequently present in the dataset [11, 19].

To deal with imbalanced data, researchers have suggested two main approaches: the first approach consists of adapting the data by performing a sampling, and the other is to tweak the learning algorithm [11]. For the sake of simplicity and due to its effectiveness in our data, we employed the first approach.

Data sampling aims at modifying the data so that all the classes have the same distribution in the training set. There exist two data sampling approaches known as **oversampling** and **under-sampling**.

Oversampling balances the training set by duplicating instances in the minority class or by generating new synthetic instances using Artificial Intelligence algorithms. **Under-sampling** instead proceeds by removing instances from the majority class.

In our case, we have considered the oversampling approach, since it proved to be more effective for small dimension datasets [21].

More specifically, we opted for *Synthetic Minority Over-sampling Technique* (SMOTE), since it creates completely new samples instead of replicating the already existing ones, which offers more examples to the classifier to learn from [4]. This means that the minority classes are oversampled by introducing synthetic examples of each minority class considering all the k minority class nearest neighbors [4].

4 CLASSIFICATION

In order to identify the classification algorithm most suited for our use case, we compared tree-based and ensemble classifiers, since they perform better than those that are not ensemble or tree-based, when dealing with low dimensionality data [19]. We evaluated and compared the performance of three among the most effective classifiers at state of the art [6].

Gradient Boosting (GB) is an ensemble algorithm that improves the accuracy of a predictive function through incremental minimization of the error term. After the initial base learner (almost always a tree) is grown, each tree in the series is fit to the so-called "pseudo residuals" of the prediction from the earlier trees with the purpose of reducing the error [2].

Random Forest (RF) is a meta-estimator of the family of the ensemble methods. It fits a number of decision tree classifiers, such that each tree depends on the values of a random vector sampled independently and with the same distribution for all the trees in the forest.

Decision Tree (DT) is a non-parametric supervised learning method used for classification and regression. One of the main advantages of decision trees with respect to other classifiers is that they are easy to inspect, interpret, and visualize, given they are less complex than the trees generated by other algorithms addressing non-linear needs [16].

5 EXPERIMENTAL FRAMEWORK

In this section, we will present the experimental setup and strategy, the evaluation metrics, and the obtained results.

5.1 Experimental Setup and Strategy

The experimental framework exploits the Python scikit-learn 0.19.1 library. The experiments were executed on a computer equipped with a 3.1 GHz Intel Core i7 processor and 16 GB of RAM. To balance the data we applied SMOTE, using *imbalanced-learn*, which is a package offering several sampling techniques used in datasets showing strong class imbalance [12]. The classification was performed with 10-fold cross-validation. Both the parameters and the features importance of the classifiers were estimated using Grid Search. The classifier was run with the default parameters, except for the number of boosting stages in Gradient Boosting (*n_estimators* parameter) and the number of nodes in each tree of Gradient Boosting (*max_depth* parameter). This is because a larger number of boosting stages (*n_estimators*) improves the performance of Gradient Boosting and *max_depth* limits the number of nodes of each tree in the boosting stages. The best parameters revealed to be *max_depth* equal to 9 and *n_estimators* equal to 400.

We performed four sets of experiments:

(1) **Classifiers comparison**. We evaluated the classifiers by running them on all the features, then we compared the accuracy metrics they obtained to determine the most effective one.
(2) **Feature sets importance evaluation**. During the feature selection phase, we used the Grid Search algorithm to evaluate the impact of each feature on the result of the classification, for the most effective classifier of the previous experiment.
(3) **Evaluation of the classifier with fewer features**. After choosing the most effective classifier, we took away the least important features one by one, and evaluated the classification accuracy to check how the less relevant features affected the effectiveness of the classifier.
(4) **Features impact on rating values**. In the last set of experiments, we measured the correlation between the value that each feature took in a workout and the rating the workout received. This allows us to evaluate how each feature impacts the quality of a workout.

5.2 Metrics

In order to evaluate the performance of our multi-class model, we had to choose metrics that are most suitable for multi-class datasets. Nevertheless, the majority of the performance measures present in the literature are designed only for two-class problems [9].

Several performance metrics for two-class problems have been adapted to multi-class. Some measures that fit well our needs, give us relevant information about the performance of our classifier, and are successfully applied for multi-class problems are: Accuracy, Recall, Precision, F1-score, Informedness, Cohen's Kappa [9]. In what follows, we present these metrics in detail.

Accuracy is defined as $(TP + TN)/(P + N)$, where P represents positively labeled instances, whereas N represents negatively labeled ones. TP represents the true positives (i.e., instances of the positive class that are correctly labeled as positive by a classifier), TN represents the true negatives (i.e., instances of the negative class that are correctly labeled as negative by a classifier). It represents the fraction of all instances that are correctly classified.

Recall is defined as TP/P and it measures the completeness of a classifier.

Precision is defined as $TP/(TP+FP)$ and it measures the exactness of a classifier.

F1 score is defined as

$$2 * \frac{TP}{2 * TP + FP + FN} \tag{1}$$

and it is a metric that considers both recall and precision.

None of the metrics presented so far takes into account the true negative rate (defined as TN/N) and this is an issue when dealing with imbalanced datasets [17]. Considered this, we decided to measure *Informedness*, which is the clearest measure of the predictive value of a system [18]. *Informedness* is defined as: *Recall + true_negative_rate* - 1, where *true_negative_rate* is TN/N. It ranges between -1 and 1, where 1 represents a perfect prediction, 0 no better than random prediction, and -1 indicates total disagreement between prediction and observation.

Cohen's Kappa is an alternative measure to Accuracy as it compensates for randomly classified instances. As opposed to Accuracy, Cohen's Kappa evaluates the portion of classified instances that can be attributed to the classifier itself, relative to all the classifications that cannot be attributed only to chance. Its formula is:

$$Kappa = \frac{Accuracy - RandomAccuracy}{1 - RandomAccuracy} \tag{2}$$

where RandomAccuracy is defined as:

$$RandomAccuracy = \frac{(TN + FP) * N + (FN + TP) * P}{(P + N)^2} \tag{3}$$

Table 2: Classifiers comparison table.

Classifier	GB	RF	DT
Accuracy	0.78	0.78	0.76
F1	0.49	0.48	0.44
Recall	0.51	0.50	0.44
Precision	0.48	0.47	0.44
Informedness	0.36	0.36	0.29
Cohen's Kappa	0.35	0.34	0.29

Cohen's Kappa ranges from -1 (total disagreement), through 0 (random classification), to 1 (perfect agreement). This metric is particularly effective for multi-class problems as opposite to the accuracy [9]. Indeed, it scores and aggregates successes independently for each class and thus it is less sensitive to the randomness caused by a different number of instances in each class.

5.3 Experimental Results

5.3.1 Classifiers comparison. Table 2 shows that Gradient Boosting is the classifier that performs better for all the metrics. The accuracy is about 78%, which means that we are correctly predicting the rating of a workout in 78% or more of the cases. This means that, in the vast majority of the cases, the coach would be able to properly support the sportspeople she follows, since she would receive an accurate ranking of those who performed worst in their training.

5.3.2 Feature sets importance evaluation. The feature selection process has shown that the ranking of the features, based on the impact in the classification process (from the most important to the least important), is :

(1) Average speed;
(2) Covered distance;
(3) Burnt calories;
(4) Workout duration;
(5) Maximum speed;
(6) User age;
(7) Rest time;
(8) User gender.

In order to analyze in more detail the relevance of these features, the diagram in Figure 2 shows the importance of each feature, using a scale ranging from 0 (no importance) to 100 (very important); we can see that each feature has an impact on the classification process, since no one has a zero importance rate.

5.3.3 Evaluation of the classifier with fewer features. After evaluating the importance of the features, we removed them one by one, to see how they are affecting the performance of the Gradient Boosting classifier. Table 3 contains the results removing the features in the previous list one by one, starting from the least important one (i.e., setting 1 contains all the features, setting 2 run the classifier without the user gender, setting 3 removed the user gender and the rest time, and so on). As the results show, none of the features is negatively affecting the performance of the classifier, since the best results were obtained when using all the features.

Figure 2: Features' importance

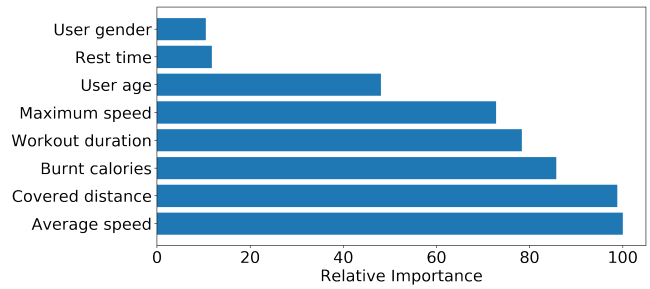


Table 3: Results returned by training Gradient Boosting with different sets of features.

	1	2	3	4	5	6	7	8
Accuracy	0.78	0.78	0.63	0.63	0.63	0.63	0.68	0.68
F1	0.49	0.49	0.03	0.03	0.03	0.03	0.08	0.22
Recall	0.51	0.50	0.20	0.20	0.20	0.20	0.20	0.21
Precision	0.49	0.48	0.04	0.04	0.02	0.04	0.55	0.27
Informedness	0.36	0.36	0.00	0.00	0.00	0.00	0.01	0.01
Cohen's Kappa	0.35	0.35	0.00	0.00	0.00	0.00	0.01	0.03

5.3.4 Features impact on rating values. After analyzing the impact of the features on the rating, we noticed that the workouts with lower ratings are those where the values of the features are low. So, the runners putting more effort during workouts are more likely to have a higher rating. The results of the individual experiments are omitted due to space constraints.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed and validated an approach to identify sportspeople that need immediate coach intervention due to poor quality workouts, so that we could suggest to their coaches to contact them with a higher priority.

Our approach takes into account a set of the workouts performed by a certain user, to which the coach assigned a rating. Then, by exploiting this data, we trained a classifier so that to predict the rating for new workout results.

Thanks to these ratings, we could be able to notify the coach when the algorithm detects that the user is performing poorly. In this way, the coach can intervene quickly to try to overcome this situation.

Experimental results show the effectiveness of our method and, as future work, we will integrate this recommender system in the u4fit platform, to be able to investigate the relationship between workout quality and users motivation. Moreover, we will also analyze the chats between coaches and their users.

ACKNOWLEDGMENTS

The authors would like to thank Marika Cappai, Davide Spano, and Daniela Lai for their contribution in this research work.

This work is partially funded by Regione Sardegna under projects AI4fit (Artificial Intelligence & Human Computer Interaction per l'e-coaching), through AIUTI PER PROGETTI DI RICERCA E SVILUPPO - POR FESR SARDEGNA 2014 - 2020, and NOMAD (Next generation Open Mobile Apps Development), through PIA - Pacchetti Integrati di Agevolazione "Industria Artigianato e Servizi" (annualità 2013).

REFERENCES

[1] Ludovico Boratto, Salvatore Carta, Fabrizio Mulas, and Paolo Pilloni. 2017. An e-Coaching Ecosystem: Design and Effectiveness Analysis of the Engagement of Remote Coaching on Athletes. *Personal Ubiquitous Comput.* 21, 4 (Aug. 2017), 689–704. https://doi.org/10.1007/s00779-017-1026-0

[2] Iain Brown and Christophe Mues. 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* 39, 3 (2012), 3446–3453.

[3] André Calero Valdez, Martina Ziefle, Katrien Verbert, Alexander Felfernig, and Andreas Holzinger. 2016. *Recommender Systems for Health Informatics: State-of-the-Art and Future Perspectives.* Springer International Publishing, Cham, 391–414.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[5] Robert G. Farrell, Catalina M. Danis, Sreeram Ramakrishnan, and Wendy A. Kellogg. 2012. Increasing Patient Safety Using Explanation-driven Personalized Content Recommendation. In *Proceedings of the Workshop on Recommendation Technologies for Lifestyle Change (LIFESTYLE 2012) (CEUR Workshop Proceedings).* CEUR-WS.org, 24–28.

[6] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research* 15, 1 (2014), 3133–3181.

[7] Brian J Fogg. 1999. Persuasive technologies. *Commun. ACM* 42, 5 (1999), 27–29.

[8] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 5.

[9] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2011. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition* 44, 8 (2011), 1761–1776.

[10] Wijnand IJsselsteijn, Yvonne de Kort, Cees Midden, Berry Eggen, and Elise van den Hoven. 2006. *Persuasive Technology for Human Well-Being: Setting the Scene.* Springer Berlin Heidelberg, Berlin, Heidelberg, 1–5.

[11] William Klement, Szymon Wilk, Wojtek Michaowski, and Stan Matwin. 2009. Dealing with severely imbalanced data. In *Proc. of the PAKDD Conference.* Citeseer, 14.

[12] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18, 17 (2017), 1–5. http://jmlr.org/papers/v18/16-365

[13] Geneviéve A Mageau and Robert J Vallerand. 2003. The coachâĂŞathlete relationship: a motivational model. *Journal of Sports Sciences* 21, 11 (2003), 883–904. https://doi.org/10.1080/0264041031000140374 arXiv:https://doi.org/10.1080/0264041031000140374 PMID: 14626368.

[14] Fabrizio Mulas, Paolo Pilloni, Matteo Manca, Ludovico Boratto, and Salvatore Carta. 2013. Linking Human-Computer Interaction with the Social Web: A web application to improve motivation in the exercising activity of users. In *Cognitive Infocommunications (CogInfoCom), 2013 IEEE 4th International Conference on.* 351–356. https://doi.org/10.1109/CogInfoCom.2013.6719270

[15] Paolo Pilloni, Luca Piras, Ludovico Boratto, Salvatore Carta, Gianni Fenu, and Fabrizio Mulas. 2017. Recommendation in Persuasive eHealth Systems: an Effective Strategy to Spot Users' Losing Motivation to Exercise. In *Proceedings of the 2nd International Workshop on Health Recommender Systems co-located with the 11th International Conference on Recommender Systems (RecSys 2017), Como, Italy, August 31, 2017. (CEUR Workshop Proceedings)*, David Elsweiler, Santiago Hors-Fraile, Bernd Ludwig, Alan Said, Hanna Schäfer, Christoph Trattner, Helma Torkamaan, and André Calero Valdez (Eds.), Vol. 1953. CEUR-WS.org, 6–9. http://ceur-ws.org/Vol-1953/healthRecSys17_paper_5.pdf

[16] Paolo Pilloni, Luca Piras, Salvatore Carta, Gianni Fenu, Fabrizio Mulas, and Ludovico Boratto. 2018. Recommender System Lets Coaches Identify and Help Athletes Who Begin Losing Motivation. *Computer* 51, 3 (2018), 36–42.

[17] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[18] David MW Powers. 2012. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 345–355.

[19] Santosh S Rathore and Sandeep Kumar. 2017. A decision tree logic based recommendation system to select software fault prediction techniques. *Computing* 99, 3 (2017), 255–285.

[20] Haggai Roitman, Yossi Messika, Yevgenia Tsimerman, and Yonatan Maman. 2010. Increasing Patient Safety Using Explanation-driven Personalized Content Recommendation. In *Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10).* ACM, New York, NY, USA, 430–434.

[21] José A Sáez, Bartosz Krawczyk, and Michał Woźniak. 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition* 57 (2016), 164–178.

[22] Darren E.R. Warburton, Crystal Whitney Nicol, and Shannon S.D. Bredin. 2006. Health benefits of physical activity: the evidence. *CMAJ* 174, 6 (2006), 801–809. https://doi.org/10.1503/cmaj.051351 arXiv:http://www.cmaj.ca/content/174/6/801.full.pdf

[23] Martin Wiesner and Daniel Pfeifer. 2014. Health Recommender Systems: Concepts, Requirements, Technical Basics and Challenges. *International Journal of Environmental Research and Public Health* 11, 3 (Mar 2014), 2580–2607.