Imputation of non-participated race results

Bram Janssens[1] and Matthias Bogaert[2]

[1] Ghent University, Ghent 9000, BE bram.janssens@ugent.be
[2] Ghent University, Ghent 9000, BE matthias.bogaert@ugent.be

Abstract. Most current solutions in cycling analytics focus on one specific race or participant, while a sports-wide system could render huge benefits of scale, by automating certain processes. The development of such a system is, however, heavily inflicted by the large number of non-participations as most riders do not compete in all races. Therefore, value imputation is required. Most popular value imputation techniques are developed for cases where part of the data is fully observed, which is not the case for cycling race results. While some methods are adapted to situations without complete cases, this is not the case for the cross-sectional imputation algorithm suggested by multiple previous studies (i.e., KNN imputation). We therefore suggest an adaptation to the KNN imputation algorithm which uses expert knowledge on race similarity in order to facilitate the deployment of the algorithm in situations without complete cases. The method is shown to be the most performant predictive model and does this within a competitive computation time.

Keywords: Sports Analytics · Scouting Analytics · Missing Value Imputation · Predictive Modeling

1 Introduction

Cycling analytics has grown as a field of study, with researchers exploring a range of possible applications. All these applications are, however, very narrowly scoped with a focus on a specific race's outcome (e.g., [13]) or on the performance of one specific athlete (e.g., [12]). While these solutions are useful, they limit the usability of the systems derived from them. This is a missed opportunity since athlete performance is even more interesting in relation to other athletes' performances and fans and coaches often desire predictions of more than one specific race.

The reason behind these scoped solutions, lays in the nature of the used data. Not all riders compete in all races, which results in a an extreme number of missing values. These missing values are extremely common, as most riders only compete in a selected number out of the hundreds of youth races available on the calendar. This results in a high number of missing values on a year-by-year result basis, but the problem also persists when aggregating results across the youth career, as many riders ride a similar program compared to their previous season. This results into highly specific setting, where generally no complete cases are observed. Most imputation methods are not adapted to this specific situation.

Existing solutions to handle the absence of complete cases, like mean imputation or Multivariate Imputation by Chained Equations [27], result in over-extensive computation times or heavy reduction in data variability. The disadvantages of these solutions, might have dissuaded researchers to come up with a sports-wide system.

We therefore suggest a solution which deploys the K-nearest neighbor (KNN) imputation algorithm [26] on subgroups of races, which due to their similarity in used route attract the same riders. This results in both complete cases being observed and more quality imputations. The method is proven to give the most accurate predictions when combined with a random forest regressor.

The remainder of this study is structured as follows. Section 2 discusses advances made in cycling analytics and discusses current solutions in missing value imputation, followed by the used methodology in Section 3. Section 4 elaborates on the performance and outcome of the various techniques, while we end with a concluding remark and a critical note in Section 5.

2 Literature overview

This section discusses current advances in literature. Section 2.1. describes how the interest in cycling analytics increased in recent years. Nonetheless, this growth in academic interest did not lead to an sports-wide analytical system as the nature of the data, with an abundance of missing values inhibits this. Current solutions on value imputation are therefore discussed in Section 2.2.

2.1 Cycling Analytics

In recent years, there has been an emergence of cycling-related data analytical studies. While some studies use analytical approaches to facilitate recreational and commuter cycling [16], less efforts have been made to harness the power of predictive analytics to boost rider and team performances. Initial introductions towards analytical methods in the field were made by [10]. The authors were able to predict a cyclist's heart rate at various moments in the training ride using a long short-term memory (LSTM) model. The study can be regarded as a proof-of-concept, indicating the feasibility of predictive models in the field of cycling.

This work was quickly followed by a range of studies who focused on practical applications to the cycling community. For example, [13] developed a real-time analytical system to estimate power performance of professional riders at the Tour de France (deployed on 2017 edition) based on GPS and wind sensor data. This would allow fans to have reliable estimates of the performance of athletes during the race. Another interesting study by [3] built a model which predicted the average velocity of a stage, the difference between the average stage velocity and the velocity of a rider and, finally, the head-to-head wins between two riders in a stage, using open data from procyclingstats.com. This popular information-tracking website was also used in other relevant studies. For example, [14] proved

it was feasible to predict race rankings based on previous race rankings scraped from the procyclingstats.com website, while [13] was capable to predict individual rider performance in key mountain stages using a combination of private training data and the open data available on the procyclingstats.com website. These studies, and the wide usage of the website among fans, has clearly established the procyclingstats.com website as the go-to source for open cycling data. The field remains to be developed with recent studies (e.g., [4]) being developed out of collaborations with top-tier teams. Current developments are resulting in-race applications such as race tactics and nutrition schemes [4].

While data is freely available, the nature of the data inevitably leads to missing values. Riders do not participate in all races, as they select which races suit their specific skill set the most. This is further complicated by geographical orientation, as non-professional athletes often do not have the means to travel across half the globe. This dispersion has led many studies in cycling analytics to focus on the prediction of a specific race [14], or on a specific rider [13]. This solution enables modelers to select the considered features in such a way that the number of missing values are limited. The usage of such specific scopes, however, limits the applicability of most analytical approaches as no sports-wide system can be developed, such as the detection of young rider talents, or a general race prediction system.

2.2 Missing Value Imputation

Two broad types of imputation methods exist: single imputation and multiple imputation [19]. While single imputation uses the single outcome of a method to impute the value, multiple imputation methods average the outcome across multiple imputed samples, which theoretically ensures a better incorporation of uncertainty about this value. Nevertheless, single imputation methods have been proven capable of outperforming multiple imputation methods [12].

An interesting related field of research, are recommender systems (RS). Whereas value imputation focuses on estimating unknown values in the training, testing, and deployment sets, do RS focus on estimating unknown values in the user-rating matrix. While both fields show a clear overlap, there is still a large distinction. RS solely try to estimate the unknown values, while value imputation methods need a way to transfer the learned practices towards unseen data, used for testing and/or deployment. Accordingly, we observe many popular RS techniques like matrix factorization [24] to be unsuited for missing value imputation in a predictive pipeline.

The three most popular single imputation methods are: mean imputation, regression imputation, and KNN imputation [12]. Mean imputation replaces the non-observed values with the mean of the observed values of the variable and is commonly used due to its simplicity [23, 5]. Regression imputation uses a regression model, which can be any type of regressor, to predict the missing values, by using the complete cases as training set and the missing cases as deployment set. KNN imputation [26] is similar to regression imputation as it also uses the neighbors of the missing case from the complete cases to see which average value

the k nearest neighbors have. The k nearest neighbors are defined by a selected distance measure (in our case the Euclidean distance). It distinguishes itself from regression imputation as no explicit predictive model is fit. Note how each imputation technique can easily be adapted for usage on categorical variables by using the mode instead of the mean or by using classification techniques instead of regression techniques.

An issue with above mentioned methods is that most need to be adjusted when dealing with extreme missingness rates [23]. Mean imputation can be directly implemented, as no complete cases are needed. This ease-of-use might explain the popularity of the method despite the large reduction in variance. Regression imputation and KNN imputation, on the other hand, need complete cases to estimate missing values.

Multivariate Imputation by Chained Equations (MICE [27]) is a solution proposed to handle this issue for regression imputation. The method starts by randomly assigning observed data as imputation of the unobserved data. However, it is stored which values were unobserved. One feature's missing values are then imputed by a regression model which uses the other features' values (actually observed + imputed values) as independent features and uses the observed values as dependent values in the training set. This done for each feature, and imputed values are updated during a number of iterations. The computational time is the largest drawback to the method, as the iterative nature causes the regression imputation method to be very time-consuming.

KNN imputation has no adaptation that handles situations without complete observations. This translates to the situation where KNN imputation is only used in situations where complete cases are observed. This is a missed opportunity for several analytical systems as the algorithm is identified as the best imputation method for predictive modeling [12]. Therefore, we suggest an adaptation to the method which uses expert knowledge to group related races together. By doing so, KNN imputation becomes feasible as no incomplete cases are observed for the grouped races. In this study, we focus on the imputation of youth race results to predict a rider's performance in his professional career. Before explaining this adaptation, the used data and features-to-be-imputed are discussed.

3 Methodology

3.1 Data

As a case study, we will develop a system to predict a young rider's expected future performance. The used data was collected from the procyclingstats.com (PCS) website, which keeps track of all youth results. A list of popular youth competitions was created, and for each of these races all the available results in the period 2005-2020 were scraped, as almost no youth results were available prior to 2005. These scraped results were used as the basis of the independent variables. Given the scarcity of results in earlier years, we decided to only select riders who turned professional in the years 2010-2019. Before 2010, we observed

the riders to have too little observed race results (i.e., less than 40 observations), leading to heavy time-based sample bias. Riders who turned professional in 2020 or 2021 were also not selected, as they did not yet have two full years of observed dependent period. Overall, this resulted in a sample of 1,060 athletes. The goal of our model is to predict the performance during a period in the rider's professional career, based on the results he achieved as youth competitor. This implies that, for instance, when modelling a rider turning professional in 2018, all his results up until 2017 will be used as input of the independent variables, while results from 2018 onwards will be used as input for the dependent variable.

The dependent variable was defined as the PCS points scored in the first two years as a professional athlete. This definition closely follows the regulations of the Union Cycliste Internationale (UCI; international cycling federation), which state that starting professional athletes (defined as competing in one of the top two tier levels) should be awarded contracts of at least two years. By measuring their performance during these two years, we can directly measure the return on investment of the hiring team. This limited time window also filters out potential negative effects of bad talent development. The option for PCS points rather than the official UCI points is inspired by the fact that this points system has remained stable during the entire period 2010-2020, while the current UCI ranking system only dates back to 2016. The ranking is also often used in cycling analytics by other researchers [21, 28].

3.2 Feature Engineering

We created a large set of independent variables (in total 242), which represent both the general rider performance (aggregate features), as well individual rider performance in one particular race. These race-specific features especially lead to large missing rates. Nonetheless, they cannot be disregarded as this gives information about a rider's talent. For example, it could be that a sprinter is more likely to score points straight away compared to climber of cobbled classics specialist. To incorporate individual race results, we included information on best past race result, and best past time difference, as finishing in the same group as the race winner can be regarded as a better result than achieving a largely distanced top-10 placement. When a rider did not finish the race, he received the placing of the last finishing rider plus one. Race participation is included as well, acting as imputation indicator. For stage races, number of stage victories, and best stage result are reported as well. Aggregate features were computed as well, both with a focus on one of the U23 (aged 19-22) and Junior (aged 17-18) categories, or averaged across both youth categories. Fully disclosed information on the set of used features can be obtained when contacting the authors.

3.3 Suggested KNN Adaption

The resulting sample contains a very high number of missing values, with only 49.57% of all the possible feature values observed. The observed rate for the race-based features (besides participation) only ranges between 5 and 40%, indicating

that race-based features (e.g., best result) have 60 to 95% missing values. Note that the overall observed rate of 49.57% is inflated by the fully observed aggregate and participation features. The reason for this high missingness is related to the selection decision of the racer and the coach (i.e., races are chosen that fit with the capacities of the rider and the team), as well as the geographical location. This results in an atypical situation in which no single complete case is observed in the data, this while most analytical models can only be used with complete datasets [15].

As discussed in Section 2.2., current solutions for missing value imputation in situations without complete cases either result in reduced data variability or are extremely time-consuming. Therefore, we suggest an alternative method for value imputation, where we group the races based on domain knowledge into groups that do have complete cases on which KNN imputation can be applied.

In total, eight categories were created. A first category is the Big Tour category, which are races that take place during a period of over a week and over varied terrain. Diverse riders with good recuperation skills excel in the overall classification of this type of race. The importance of the races also attracts riders from quite wide geographical origins and the longitude and importance of the races make it more interesting for some to solely focus on stage victories rather than the overall classification. A related category is the Stage Race Climb category of French stage races over very hilly terrain, attracting many riders from France and neighboring countries. Both categories consist of U23 stage races, Junior stage races are categorized in the Stage Race Junior category, which is more diverse. This due to the fact that Juniors have more limited calendar options. Regarding the one day races, we also followed a similar method, with the One Day Junior races forming one category, and the U23 races divided into Cobbles and Hilly U23. Cobbled races are quite unique as they are the sole type of races which favor more heavy riders, while also being located in and near Belgium. This as opposed to the hilly one day races, which are one a hilly terrain, favoring more light-weight riders, while also being primarily located in Italy. All other races are categorized as Rest.

Table 1. Average results

	Victory ratio U23	Evolution Wins	Omloop der Vlaamse Gewesten best result	Paris-Roubaix U23 best result	Tour de l'Isard best result	Tour d'Alsace best result
Rider 1	0.238	1.608	6	3	**70**	88
Rider 2	0.006	0.884	**6**	8	3	5
Rider 3	0.082	0.589	60	96	1	**5**
Rider 4	0.000	0.000	**60**	77	70	**88**

By using the KNN algorithm on each feature group, rather than across all features, complete cases are observed as the similarity of the race program at-

tracts some riders to complete the fully considered race program. This has also has the advantage that we only use the most relevant features for imputation. Table 1 provides a simplified example of our proposed imputation method. Note that the imputed values are highlighted in bold. From this examples, it is clear that no rider competed in all four races, which would render the base KNN imputation method infeasible as no complete cases are observed. In addition, it also uses a more scoped and better-informed approach. For instance, the imputation-relevant information to predict the result of a rider in Omloop der Vlaamse Gewesten will be mainly situated in the Cobbles group (Paris-Roubaix U23 in this case), while the Stage Race Climb group will contain very limited relevant information. The benefit of our method is nicely reflected for rider 2. He has a highly similar profile to rider 3, which would probably be his nearest neighbor. However, compared to rider 3, he also performs quite well on the cobblestones, as is reflected by his 8th place in Paris-Roubaix U23. This makes rider 1 a better candidate for being rider 2's nearest neighbor in the cobbles group, rendering an imputed value of 6, rather than of 60.

3.4 Experimental Set-up

We will compare our suggested approach to both the mean imputation technique and the Multivariate Imputation by Chained Equations (MICE) technique. For the MICE methodology, we set the number of iterations at 10, acting as a trade-off between computational time and reaching of convergence. As base regressor, we chose random forest [1] due to the algorithms capacity to handle non-linear relationships as well as its good performance without parameter tuning [7]. Previous research [6] also concluded that recursive partitioning methods are recommended over standard applications of MICE. Do note that this implementation is essentially the same as the MissForest implementation by [25] The number of considered neighbors (K) is set to 5 for our KNN adaptation.

The three imputation methods will be compared to each other with regard to both the speed of execution as well as the performance in a predictive modeling pipeline. Besides the imputation step, this pipeline will also include a feature selection step, as a large number of 242 features were considered compared to the maximal sample size of 1,060 athletes, and a regression algorithm. These result in the sequence imputation – feature selection – regression being deployed.

A very popular feature selection method is the Boruta algorithm [17]. The base algorithm used is random forest and the method is based on the idea of 'shadow variables'. These are created by replacing the actual feature values with random permutations of these values. When the shadow variable's variable importance is not significantly different than the actual variable importance, it is decided that the feature in question is not needed and can be excluded from the eventual feature list. Since Gini-based variable importance rankings are unreliable, we use SHAP-based [20] feature importances.

As regression algorithm, we will deploy random forest regression [1]. This algorithm was selected in our experimental set-up as it is very robust and performs well without heavy parameter tuning. The number of trees was set sufficiently

large at 500 and the number of random predictors to select at each tree split was set at the square root of the number of predictors.

As a season follows the subsequent one and riders compete against each other in the same season, rather than act as individualistic competitors, one can safely say that the assumption of independent and identically distributed data is clearly violated. This influences our test design, as a traditional cross-validated approach is not adequate in this situation. Rather, we will follow a rolling window approach where all available information is used up until the moment of prediction [29]. In order to have an unbiased estimation of performance, we use five different periods for testing: starting years 2015-2019. Note that the validation period is only used for hyperparameter tuning and that the combined training and validation period is eventually used for fitting the final model.

Each fold is evaluated against a range of performance measures. The Root-Mean-Squared-Error (RMSE) calculates how exact the method can predict the points scored per participant. This is, however, not the main goal, as teams rather want the best riders to be ranked on top. Therefore, the Spearman rank correlation between actual results and predicted results is calculated as well, indicating how consequent the best riders are ranked on top [9]. Another interesting way of dividing professional athletes is by grouping them into the top 10%, top 25%, or top 50% buckets of all athletes [22]. A good way of measuring the performance of this binned continuous scale, is accuracy within one [8] as this filters out the oversensitivity to misclassifications near the arbitrary cut-off. This adaption to the traditional accuracy measure also accounts ordered predictions as correct if they deviate only one class from the actual class.

Of special interest to the professional teams, is the absolute top bin of the top-10% riders. These riders are the ones they want to contact by preference. By considering this bin as the desired class, we can deploy the traditional binary classification performance measures. A popular measure based on the top decile bin, is the lift. By calculating how much more actual top 10% riders there are in the suggested bin, than on average in the dataset, one can derive how much better the model is compared to randomly contacting riders. It is clear that this measure is highly sensitive to the used cut-off of 10% contacted. This is even more worrisome as it is very feasible that the teams won't contact 10% of all riders as their teams simply aren't large enough to contract so many riders. A more complete measure is the average precision, which considers different cut-off rates.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(true_i - predict_i)^2} \tag{1}$$

$$Spearman = \frac{1 - 6\sum_{i=1}^{N} d_i}{N^3 - N} \tag{2}$$

$$Lift = \frac{Precision\ contacted\ top\ decile}{actual\ rate} \tag{3}$$

$$Average\ Precision = \sum_{n}(Recall_n - Recall_{n-1})Precision_n \tag{4}$$

4 Results

Table 2 depicts the average results of the machine learning pipelines across all five folds. KNN scores best on 3 out of 5 performance measures (i.e., RMSE, Spearman correlation, and average precision), with competitive scores on the other 2 measures. This suggests the KNN method as the imputation method which is most useful in the predictive pipeline.

Table 2. Grouped KNN imputation

	RMSE	Spearman	Accuracy Within One	Average Precision	Lift
KNN	**282.89**	**0.5213**	0.8359	**0.3714**	3.2121
Mean	294.44	0.4799	**0.8437**	0.3232	2.9264
MICE	291.36	0.5107	0.8393	0.3681	**3.5996**

A final argument in method selection can be the time required to come up with suggested rider rankings. The computation time of the imputation step per fold is depicted in Table 3. Whereas the grouped KNN and mean imputation methods take only a couple of seconds, does the chained equation regression step take almost 10 hours for the calculation of the largest imputed dataset. This time will probably only further increase with the addition of additional riders to the dataset. As fast imputation allows quick interpretation of new youth race results, this could potentially hinder teams in moving fast with regard of the contacting of a new interesting prospect. Therefore, grouped KNN and mean imputation are suggested above chained equation regression imputation. Overall, our suggested KNN imputation adaptation gives the best results when included in a predictive modelling pipeline, while being highly competitive in terms of computation time.

Table 3. Computation time imputation methods (in seconds)

Fold	Train/val/test size	KNN imputation	Mean imputation	MICE
2015	401/80/112	1.92	0.06	13025.15
2016	481/112/110	1.59	0.03	15035.95
2017	593/110/117	2.46	0.06	20488.19
2018	703/117/131	2.75	0.03	26363.01
2019	820/131/109	3.46	0.03	32997.33

To see whether our approach can be effectively used to detect future star riders early on, we deployed the technique on the riders who turned professional

during the years 2020 and 2021. Interestingly, we observe several prospects in our suggested top-10 who have already showed some good form at the professional level. For instance, Tom Pidcock already finished in the top-5 in the Strade Bianchi and Amstel Gold Race, some of the most important races on the calendar, and won the Brabantse Pijl against a top tier field of participants. Stefan Bisegger also already won a stage in the World Tour Paris-Nice stage race.

5 Conclusion

In this paper we developed a method to impute race results to riders who did not participate. The method leveraged expert knowledge about the similarity between certain youth races, ending up with complete cases for each subgroup, enabling the deployment of the KNN imputation algorithm. The used race groups were Stage Race Junior, One Day Junior, Cobbles, Hilly U23, Big Tour, Stage Race Climb, ITT, and Rest.

The proposed method was shown to yield the best results when included in a predictive modelling pipeline, compared to the traditional mean imputation and MICE solutions. This top performance was achieved within a competitive computation time. We demonstrated that the detection of young cycling talents based on youth race results is feasible despite the tendency of the observed data to have many missing values. The suggested rider rankings have a strong relation to the actually observed rider rankings.

An avenue for future research might be the inclusion of more various regression algorithms. While the adapted KNN is shown to yield the most accurate eventual results, it could be that this is due to a beneficial interplay between the imputation method and the base regressor. The used methodology should therefore be evaluated for other algorithms in the future.

Our method was only deployed onto one specific case, namely the detection of young cycling talents. However, we would like to point out that a similar grouping can be made with regard to professional races, or even amateur races, making predictive analytic systems feasible for a wide range of applications by using the grouped KNN method.

References

1. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
2. Dancey, C. P., Reidy, J. (2007). Statistics without maths for psychology. Pearson education.
3. De Spiegeleer, E (2019). Predicting cycling results using machine learning.
4. de Leeuw, A. W., Heijboer, M., Hofmijster, M., van der Zwaard, S., Knobbe, A. (2020, October). Time Series Regression in Professional Road Cycling. In International Conference on Discovery Science (pp. 689-703). Springer, Cham.
5. Dolatsara, H. A., Chen, Y. J., Evans, C., Gupta, A., Megahed, F. M. (2020). A two-stage machine learning framework to predict heart transplantation survival probabilities over time with a monotonic probability constraint. Decision Support Systems, 137, 113363.

6. Doove, L. L., Van Buuren, S., Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. Computational statistics & data analysis, 72, 92-104.
7. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The journal of machine learning research, 15(1), 3133-3181.
8. Gaudette, L., Japkowicz, N. (2009, May). Evaluation methods for ordinal classification. In Canadian conference on artificial intelligence (pp. 207-210). Springer, Berlin, Heidelberg.
9. Gauthier, T. D. (2001). Detecting trends using Spearman's rank correlation coefficient. Environmental forensics, 2(4), 359-362.
10. Hilmkil, A., Ivarsson, O., Johansson, M., Kuylenstierna, D., van Erp, T. (2018). Towards machine learning on data from professional cyclists. arXiv preprint arXiv:1808.00198.
11. Jadhav, A., Pramod, D., Ramanathan, K. (2019). Comparison of performance of data imputation methods for numeric dataset. Applied Artificial Intelligence, 33(10), 913-933.
12. Karetnikov, A. (2019). Application of Data-Driven Analytics on Sport Data from a Professional Bicycle Racing Team. Eindhoven University of Technology, The Netherlands.
13. Kataoka, Y., Gray, P. (2018, September). Real-time power performance prediction in tour de France. In International Workshop on Machine Learning and Data Mining for Sports Analytics (pp. 121-130). Springer, Cham.)
14. Kholkine, L., De Schepper, T., Verdonck, T., Latré, S. (2020, September). A Machine Learning Approach for Road Cycling Race Performance Prediction. In International Workshop on Machine Learning and Data Mining for Sports Analytics (pp. 103-112). Springer, Cham.
15. Kowarik, A., Templ, M. (2016). Imputation with the R Package VIM. Journal of Statistical Software, 74(7), 1-16.
16. Kumar, A., Nguyen, V. A., Teo, K. M. (2016). Commuter cycling policy in Singapore: a farecard data analytics based approach. Annals of Operations Research, 236(1), 57-73.
17. Kursa, M. B., Rudnicki, W. R. (2010). Feature selection with the Boruta package. J Stat Softw, 36(11), 1-13.
18. Larson, D. J., Maxcy, J. G. (2016). Human capital development in professional cycling. In The Economics of Professional Road Cycling (pp. 129-145). Springer, Cham.
19. Little, R. J., Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.
20. Lundberg, S. M., Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
21. Miller, J., Susa, K. (2018). Comparison of anthropometric characteristics between world tour and professional continental cyclists. Journal of Science and Cycling, 7(3), 3-6.
22. Persson, T. L., Kozlica, H., Carlsson, N., Lambrix, P. (2020, September). Prediction of tiers in the ranking of ice hockey players. In International Workshop on Machine Learning and Data Mining for Sports Analytics (pp. 89-100). Springer, Cham.)
23. Piri, S. (2020). Missing care: A framework to address the issue of frequent missing values; The case of a clinical decision support system for Parkinson's disease. Decision Support Systems, 136, 113339.

24. Ranjbar, M., Moradi, P., Azami, M., Jalili, M. (2015). An imputation-based matrix factorization method for improving accuracy of collaborative filtering systems. Engineering Applications of Artificial Intelligence, 46, 58-66.
25. Stekhoven, D. J., Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.
26. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., ... , Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520-525.
27. Van Buuren, S., Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 1-68.
28. van Erp, T., Sanders, D., Lamberts, R. P. (2021). Maintaining Power Output with Accumulating Levels of Work Done Is a Key Determinant for Success in Professional Cycling. Medicine and Science in Sports and Exercise.
29. Vomfell, L., Härdle, W. K., Lessmann, S. (2018). Improving crime count forecasts using Twitter and taxi data. Decision Support Systems, 113, 73-85.