# Bike2Vec: Vector Embedding Representations of Road Cycling Riders and Races

Ethan Baron[a], Bram Janssens[b,c,d] and Matthias Bogaert[b,c]

[a] University of Toronto / Zelus Analytics, eth.baron@mail.utoronto.ca
[b] Department of Marketing, Innovation and Organisation, Ghent University
[c] FlandersMake@UGent–corelab CVAMO
[d] Research Foundation Flanders

**Abstract**

Vector embeddings have been successfully applied in several domains to obtain effective representations of non-numeric data which can then be used in various downstream tasks. We present a novel application of vector embeddings in professional road cycling by demonstrating a method to learn representations for riders and races based on historical results. We use unsupervised learning techniques to validate that the resultant embeddings capture interesting features of riders and races. These embeddings could be used for downstream prediction tasks such as early talent identification and race outcome prediction.

## 1 Introduction

Professional road cycling offers several interesting challenges in an analytics setting due to its unique properties. Notably, races have a variety of different formats (e.g. one-day races, stage races) and profiles (e.g. flat, hilly, or mountainous), each suiting riders of different characteristics. Although several past works have demonstrated the potential of using machine learning models to predict road cycling race results, these models rely on significant feature engineering efforts and are tailored to predicting specific outcomes, such as rider performance in a specific race.

We present a framework forming the foundation for a generalized prediction algorithm that does not depend on labour-intensive feature engineering efforts. Specifically, we introduce a method to train vector embeddings for riders and races based on historical race results.

In representation learning, vector embeddings are used to capture the key qualities of entities such as words, images, or songs. If trained effectively, these vector embeddings can then be used for a variety of downstream tasks. For example, word embeddings trained using a large corpus of text can be used for emotion recognition or sentence completion.

Likewise, we show that our cycling embeddings capture the key characteristics of riders and races. The embeddings can be used in downstream prediction tasks and eliminate the need for a manual feature engineering process.

## 2   Literature Review

### 2.1   Machine Learning in Road Cycling

There are several prior works which apply machine learning to road cycling.

Multiple works focus on predicting the ProCyclingStats ("PCS") points, a system developed by the website `procyclingstats.com` to assign scores to riders based on the results achieved in certain races. For example, Janssens and Bogaert (2021) construct a random forest regression to predict the total PCS points scored by under-23 prospects in their first two years as professional athletes. They engineer a large set of features, including the riders' performances in particular under-23 races, and compare various methods to impute non-participated race results. This imputation method is used to detect the most promising young athletes (Janssens, Bogaert, and Maton 2022). Similarly, Van Bulck, Vande Weghe, and Goossens (2021) compare linear regression and random forest regression to predict the points scored by under-23 riders in their first three years as professionals. They also hand-craft a number of features summarizing the riders' performance at the youth level.

Other works focus on predicting the outcomes of particular races. De Spiegeleer (2019) develops machine learning models to predict various outcomes of stages from the Tour de France, Giro d'Italia, and Vuelta a Espana, including their average speed, the difference between the average speed of the winner and that of a particular rider, and the head-to-head performance between two riders. The predictions are based on an extensive set of engineered features related to the terrain, weather, rider characteristics, and historical results. Mortirolo (2019) uses Bayesian Additive Regression Trees to simulate races and obtain predictions for the probabilities of specific race outcomes. The simulation uses over one hundred features, including ratings for riders' various abilities, indicators of riders' recent form, historical results from the past three years, and team-level indicators. Kholkine, De Schepper, et al. (2020) apply an XGBoost model to predict the outcomes of the Tour of Flanders using riders' performances in relevant build-up races. They also engineer several advanced features related to past results in similar races, historical weather data and overall team performance. Kholkine, Servotte, et al. (2021) also employ an XGBoost model within a learn-to-rank framework to predict the top ten riders in several one-day races using a suite of engineered features based on historical results and the riders' ages. Finally Demunter (2021) compares linear regression, random forest, XGBoost, and neural networks to predict the rankings of riders in a given race. Again, various features related to the rider's recent and historical results are developed and used as inputs for these models.

To the best of our knowledge, we present the first framework for a generalized prediction algorithm for road cycling which does not rely on a hand-crafted set of features for the particular outcome of interest.

### 2.2   Representation Learning

Representation learning is the field of machine learning concerned with automatically learning meaningful and compact representations of data without requiring features for the data. These representations aim to capture the underlying structure and patterns in the data, enabling more effective performance on a variety of downstream applications. Some primary applications of representation learning include natural language processing, computer vision, and recommendation systems.

In natural language processing, representation learning has been utilized to learn word embeddings that capture semantic relationships between words. For example, the word2vec algorithm uses a skip-gram ap-

proach to fit vector embeddings for words. These embeddings yield successful results on downstream tasks such as semantic and syntactic word relationship testing and sentence completion (Mikolov et al. 2013). Another common approach known as GloVe uses co-occurence statistics of words to obtain the vector embeddings. The resulting representations performed strongly on a variety of tasks including word analogies, word similarities, and named entity recognition (Pennington, Socher, and Manning 2014). Finally, pre-trained vector embeddings using bidirectional encoder representations from transformers (BERT) can then be used as inputs into various downstream tasks and have yielded state-of-the-art performance on question answering and language inference (Devlin et al. 2019).

Similarly, representation learning has been successfully applied in the field of computer vision. One notable technique is the use of convolutional neural networks. These convolutional neural networks achieved pioneering performance on image classification tasks, such as handwritten digit recognition (Cireşan et al. 2011) and high-resolution image classification (Krizhevsky, Sutskever, and Hinton 2012). By pre-training large convolutional neural networks on large amounts of data, researchers have then achieved state-of-the-art performance on novel computer vision tasks, including image recognition (Simonyan and Zisserman 2015), object detection (Girshick et al. 2014), scene recognition and domain adaptation (Donahue et al. 2014).

Representation learning advancements in natural language processing and computer vision have exploited observed relationships between words and local patterns in images. In the case of road cycling, we seek to extract representations for both races and riders and exploit historical interactions between these two types of entities. Most relevant to this context are past works on recommender systems surrounding collaborative filtering, which use historical interactions between users and items to recommend new items to a user. One common approach for such problems is to transform both items and users to the same vector space by assigning vector embeddings of the same dimension to both categories. Then, dot products between the vector embeddings of users and items can be used to infer their interaction (Koren 2008).

Although we are not aware of such a representation learning approach being applied in road cycling, similar approaches have recently been tested in other sports, including soccer (Cintia and Pappalardo 2021; Magdaci 2021; Yılmaz and Öğüdücü 2022; Li et al. 2022), basketball (Papalexakis and Pelechrinis 2018), and cricket (Alaka, Sreekumar, and Shalu 2021).

## 3   Methods

We collect historical race results from the 2016-2022 UCI World Tour seasons from `procyclingstats.com`. Specifically, we consider the results of one-day races, and individual stages of stage races (i.e. multi-day races), except for team time trials. We do not consider overall classifications of stage races. Overall, our dataset includes results from 1069 race editions, 118 of which are one-day races.

For each result in our data, we define the *normalized PCS score* as the number of PCS points scored by the rider in the race, divided by the points earned by the winner of that race. For example, if the winner and runner-up of a race earn 500 and 300 PCS points respectively, they are assigned a normalized PCS score of 1 and 0.6 respectively.

We learn vector embeddings of dimension $D$ for individual riders and races by directly optimizing these embeddings' ability to predict historical results. Specifically, we represent a rider's aptitude at a given race by the dot-product between that rider's embedding and that race's embedding. We then pass this dot-product through a sigmoid activation function to predict the normalized PCS score for that rider at that race. The

vector embeddings are trained by minimizing the binary cross-entropy loss between these predictions and the actual normalized PCS scores, according to equation 1.

$$L(R,S,y) = \frac{1}{N} \sum_{i=1}^{N} y_i \log(\sigma(R_{r(i)} \cdot S_{s(i)})) + (1 - y_i) \log(1 - \sigma(R_{r(i)} \cdot S_{s(i)})) \tag{1}$$

Here, $y_i$ records the normalized PCS points scored by rider $r_{(i)}$ in race $s_{(i)}$, $R$ is the matrix of rider vector embeddings, $S$ is the matrix of race vector embeddings, and $N$ is the number of results in our data. $\sigma$ refers to the sigmoid function such that $\sigma(x) = (1 + e^{-x})^{-1}$.

We train a vector embedding for each rider who has scored at least 25 (unnormalized) PCS points in our dataset. We also train a vector embedding for each race edition, except that for one-day races, we use the same embedding across all seasons. We do this since one-day races tend to suit similar riders across years, whereas stages can have very different characteristics across seasons. In total, we train unique embeddings for 973 races and 958 riders.

The results shown below are based on embeddings of dimension $D = 5$ trained using an Adam optimizer with a learning rate of 0.001 for 100 epochs. Reproducible code to implement our methods is available at `https://github.com/baronet2/Bike2Vec`.

## 4   Results

In this section, we analyze our learned embeddings to demonstrate that they capture valuable information about the characteristics of riders and races.
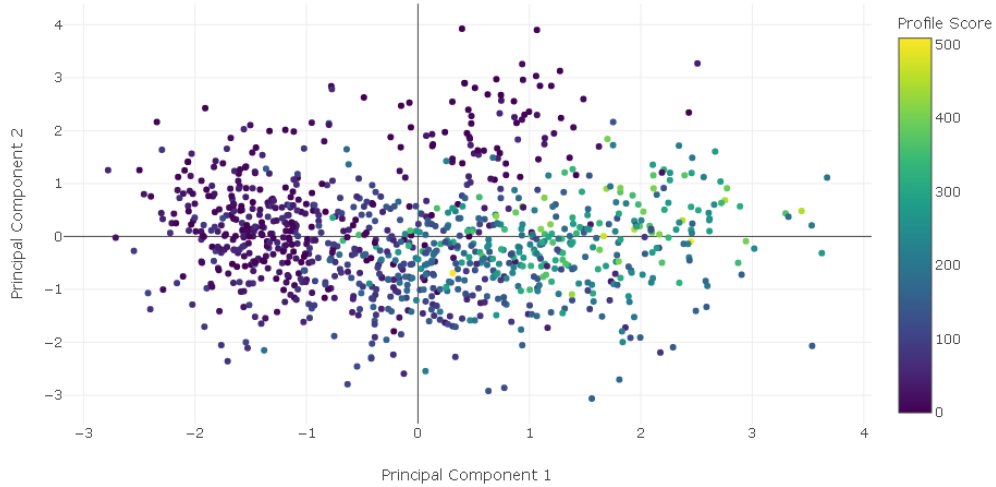


Figure 1:  Visualization of race embeddings coloured by race profile score.

In Figure 1, we plot the embeddings for each race in our dataset, coloured according to the race profile score, a measurement of the amount of climbing in the race developed by PCS. We performed a principal component analysis to reduce the dimensionality of the embeddings to two dimensions for visualization

purposes. Clearly, the primary principal component is capturing significant information about the terrain of a race, with more mountainous races appearing on the right and flat races appearing on the left.
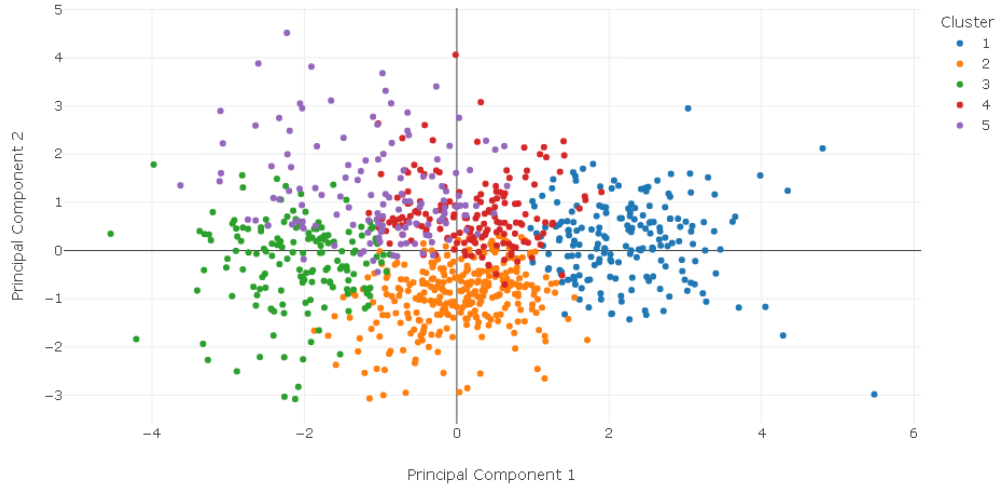


Figure 2:  Visualization of rider embeddings coloured by cluster.

Similarly, in Figure 2, we plot the principal components of the rider embeddings. Unlike races, riders are not labelled by PCS as belonging to a certain category. Therefore, to add interpretability, we perform k-means clustering on the rider embeddings and colour the riders by their assigned cluster.

We show a few examples of riders from each cluster in Table 1. There are clear similarities among the riders in each cluster, indicating that our embeddings are capturing the unique characteristics of each rider. For example, cycling fans would identify that cluster 1 is composed of sprinters and cluster 3 of climbers.

| Cluster | Examples of Riders |
|---------|--------------------|
| 1 | SAGAN Peter, KRISTOFF Alexander, VIVIANI Elia, EWAN Caleb, BENNETT Sam |
| 2 | VAN AVERMAET Greg, COLBRELLI Sonny, NAESEN Oliver, MOHORIČ Matej |
| 3 | ALAPHILIPPE Julian, VALVERDE Alejandro, ROGLIČ Primož, POGAČAR Tadej |
| 4 | VAN AERT Wout, MATTHEWS Michael, STUYVEN Jasper, KWIATKOWSKI Michał |
| 5 | VAN DER POEL Mathieu, GILBERT Philippe, LAMPAERT Yves, ŠTYBAR Zdeněk |

Table 1:  Examples of riders from each cluster.

Furthermore, in Table 2, we show some examples of rider similarities. That is, for each rider on the left-hand-side, we show the name of the other rider with the most similar embedding, according to Euclidean distance. Cycling fans would confirm that these rider pairings strongly reflect these riders' characteristics. For example, Tadej Pogacar and Primoz Roglic are both world-class climbers and time-trialists, Peter Sagan and Sonny Colbrelli are versatile sprinters who also perform well in cobbled or hilly classics, and Julian Alaphilippe and Marc Hirschi are both specialists at climbing short but steep hills.

| Rider 1 | Rider 2 |
|---|---|
| POGAČAR Tadej | ROGLIČ Primož |
| SAGAN Peter | COLBRELLI Sonny |
| ALAPHILIPPE Julian | HIRSCHI Marc |
| YATES Simon | BARDET Romain |
| EVENEPOEL Remco | ALMEIDA João |
| QUINTANA Nairo | ZAKARIN Ilnur |
| VIVIANI Elia | GREIPEL André |
| DENNIS Rohan | CAVAGNA Rémi |

Table 2: Examples of most similar rider embeddings.

Overall, the vector embeddings seem to accurately capture the distinguishing characteristics of both riders and races.

# 5   Conclusion

We present a novel vector embedding approach to represent road cycling riders and races, and implement the approach on seven seasons of data from professional men's road cycling races. We validate the resulting embeddings by showing that they contain useful information about the characteristics of races and riders. These embeddings can form the basis for a variety of downstream prediction tasks, removing the need for extensive manual feature engineering.

Although we have demonstrated that our proposed vector embeddings contain valid information about the riders and races, we have yet to test the inclusion of these embeddings within a downstream prediction task. We leave this as a promising area for future work. Further, augmenting our race embeddings using features about the route, such as the elevation profile, could offer improved race embeddings and enable predictions on new races. Additionally, our current framework assigns a constant embeddings over the span of riders' careers. Future research could seek to incorporate a time-varying element to capture changes in rider skills due to age, physiology, injury, or other effects. Lastly, an interesting avenue for further exploration is extending our framework to women's cycling and comparing the results.

# References

Alaka, Souridas, Rishikesh Sreekumar, and Hrithwik Shalu (Aug. 2021). *Efficient Feature Representations for Cricket Data Analysis using Deep Learning based Multi-Modal Fusion Model*. eprint: 2108.07139. URL: https://arxiv.org/pdf/2108.07139.pdf.

Cintia, Paolo and Luca Pappalardo (June 2021). *Coach2vec: Autoencoding the Playing Style of Soccer Coaches*. Tech. rep. ISTI Research Report. eprint: 2106.15444. URL: https://arxiv.org/ftp/arxiv/papers/2106/2106.15444.pdf.

Cireşan, Dan Claudiu et al. (July 2011). "Flexible, High Performance Convolutional Neural Networks for Image Classification". In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. Vol. 2. Ijcai'11. Barcelona: AAAI Press, pp. 1237–1242. ISBN: 9781577355144. URL: https://people.idsia.ch/~juergen/ijcai2011.pdf.

De Spiegeleer, Emiel (June 2019). "Predicting Cycling Results using Machine Learning". MA thesis. Ghent: Ghent University. URL: https://libstore.ugent.be/fulltxt/RUG01/002/785/834/RUG01-002785834%5C%5F2019%5C%5F0001%5C%5FAC.pdf.

Demunter, Jarne (June 2021). "Predicting Ranking Multientrant Races: Road Cycling". MA thesis. Ghent: Ghent University. URL: https://libstore.ugent.be/fulltxt/RUG01/003/010/353/RUG01-003010353%5C%5F2021%5C%5F0001%5C%5FAC.pdf.

Devlin, Jacob et al. (Oct. 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. eprint: 1810.04805. URL: https://arxiv.org/pdf/1810.04805.pdf.

Donahue, Jeff et al. (June 2014). "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 1. Bejing, China: Pmlr, pp. 647–655. URL: https://proceedings.mlr.press/v32/donahue14.html.

Girshick, Ross et al. (June 2014). "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos: IEEE Computer Society, pp. 580–587. DOI: 10.1109/cvpr.2014.81.

Janssens, Bram and Matthias Bogaert (Sept. 2021). "Imputation of Non-participated Race Results". In: *ECML / PKDD 2021 Workshop, 8th Workshop on Machine Learning and Data Mining for Sports Analytics, Proceedings*. Bilbao, p. 12. URL: https://dtai.cs.kuleuven.be/events/MLSA21/papers/MLSA21%5C%5Fpaper%5C%5Fjanssens.pdf.

Janssens, Bram, Matthias Bogaert, and Mathijs Maton (Jan. 2022). "Predicting the next Pogačar: a data analytical approach to detect young professional cycling talents". In: *Annals of Operations Research*, pp. 1–32. DOI: https://doi.org/10.1007/s10479-021-04476-4.

Kholkine, Leonid, Tom De Schepper, et al. (Sept. 2020). "A Machine Learning Approach for Road Cycling Race Performance Prediction". In: *Machine Learning and Data Mining for Sports Analytics*. Ed. by Ulf Brefeld et al. Communications in Computer and Information Science. Springer International Publishing, pp. 103–112. ISBN: 978-3-030-64912-8. DOI: 10.1007/978-3-030-64912-8\_9.

Kholkine, Leonid, Thomas Servotte, et al. (Oct. 2021). "A Learn-to-Rank Approach for Predicting Road Cycling Race Outcomes". In: *Frontiers in Sports and Active Living* 3, p. 714107. ISSN: 2624-9367. DOI: 10.3389/fspor.2021.714107.

Koren, Yehuda (Aug. 2008). "Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discov-*

*ery and Data Mining*. Kdd '08. Las Vegas: Association for Computing Machinery, pp. 426–434. ISBN: 9781605581934. DOI: `10.1145/1401890.1401944`.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (Dec. 2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc. URL: `https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`.

Li, Yuesen et al. (July 2022). "Characterizing Player's Playing Styles Based on Player Vectors for Each Playing Position in the Chinese Football Super League". In: *Journal of Sports Sciences* 40.14, pp. 1629–1640. DOI: `10.1080/02640414.2022.2096771`.

Magdaci, Ofir (June 2021). *Embedding the Language of Football Using NLP*. URL: `https://towardsdatascience.com/embedding-the-language-of-football-using-nlp-e52dc153afa6`.

Mikolov, Tomas et al. (Jan. 2013). *Efficient Estimation of Word Representations in Vector Space*. eprint: `1301.3781`. URL: `http://arxiv.org/abs/1301.3781`.

Mortirolo (July 2019). *Cycling prediction method*. URL: `https://mortirolo.netlify.app/post/cycling-prediction-method/` (visited on 05/06/2023).

Papalexakis, Evangelos and Konstantinos Pelechrinis (Oct. 2018). "THoops: A Multi-Aspect Analytical Framework for Spatio-Temporal Basketball Data". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Cikm '18. Torino: Association for Computing Machinery, pp. 2223–2232. ISBN: 9781450360142. DOI: `10.1145/3269206.3272002`.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha: Association for Computational Linguistics, pp. 1532–1543. DOI: `10.3115/v1/D14-1162`.

Simonyan, Karen and Andrew Zisserman (Apr. 2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations*. San Diego: Computational and Biological Learning Society, pp. 1–14. eprint: `1409.1556`. URL: `https://arxiv.org/pdf/1409.1556.pdf`.

Van Bulck, David, Arthur Vande Weghe, and Dries Goossens (Oct. 2021). "Result-based talent identification in road cycling: Discovering the next Eddy Merckx". In: *Annals of Operations Research*. DOI: `10.1007/s10479-021-04280-0`.

Yılmaz, Öznur Ilayda and Şule Gündüz Öğüdücü (Apr. 2022). "Learning Football Player Features Using Graph Embeddings for Player Recommendation System". In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. Sac '22. Association for Computing Machinery, pp. 577–584. ISBN: 9781450387132. DOI: `10.1145/3477314.3507257`.