Contents

[List of Figures 2](#_Toc76461143)

[List of Tables 3](#_Toc76461144)

[Motivation 4](#_Toc76461145)

[Introducing the Glicko rating system 5](#_Toc76461146)

[Data 7](#_Toc76461147)

[Modeling 8](#_Toc76461148)

[Baseline model motivation 8](#_Toc76461149)

[Baseline 1 9](#_Toc76461150)

[Baseline 2 9](#_Toc76461151)

[The Model 12](#_Toc76461152)

[Example 14](#_Toc76461153)

[Discussion of the Results 26](#_Toc76461154)

[Glicko-in-Glicko with teams 26](#_Toc76461155)

[Glicko without teams 27](#_Toc76461156)

[Further Extensions 29](#_Toc76461157)

[Works Cited 30](#_Toc76461158)

# List of Figures

# List of Tables

# Motivation

Predicting the outcome of sports has been a tried using many different methods in recent years. There have been papers focused on predicting road cycling results using machine learning approaches (Kholkine, 2020) or time-series regression (Arie-Willem de Leeuw, 2020). This work explores predicting results of road cycling races using a novel method based on glicko model as described in (Glickman, 1995). The main focus is to develop a model for predicting the outcome of cycling races, such as the Tour de France.

Glicko model has been successfully applied to many sports, including but not limited to chess (Glickman, n.d.), e-sports (Pradhan, 2020) or soccer (Boice, 2020). The reason why applying a glicko system isn’t as straightforward in cycling as in other sports fields is the fact that road cycling is both a team sport, as well as an individual sport. The outcome of a race largely dependent on both of the factors in varying amounts.

Goal of this work is to create a model capable of predicting the winner of cycling races, such as the Tour de France.

# Introducing the Glicko rating system

The Glicko model was proposed as an improvement over the existing system of calculating the ranking of players in individual games – the Elo system, developed by a Hungarian-born American scientist Arpad Emrick Elo. For a detailed description of Glicko and Elo system, refer to (Glickman, 1995) and (Elo, 1978), respectively.

Over the course of this work, the version of the Glicko model as presented in (Glickman, 1995) will be used, instead of the second version described in (Glickman, 2013). The first version works with player rating and rating’s deviance, which describes the size of the credible interval, i.e. the uncertainty of the rider’s rating. The second version is adding an additional measure for describing the volatility of the ranking.

When calculating a Glicko rating, first, a rating period has to be defined, which is a collection of games that are deemed to have occurred at the same time. This can be one round of a chess or tennis tournament, for example. Subsequently, rating (r) and rating deviation (RD) are defined. Rating is simply a score that measures the player’s strength. Rating’s deviation (RD) is a measure of uncertainty. High rating’s deviation captures a high degree of uncertainty about a particular rating. The rating and rating deviation can be used to compute credible intervals (CI95):

The most important formulas used in this work are those concerning the calculation of a new ranking, at the end of evaluation period. These are going to be used exactly in the same form as introduced in (Glickman, 1995). RD’ and r’ denote the newly computed ranking deviation and ranking, respectively.

where

# Data

The dataset used is a set of cycling results from years 2010 to 2020, downloaded from the website Procyclingstats.com. The data only includes races of the category Men Elite, counted in ProTour category of the International Cycling Federation.

The results are split into four groups – sprint stages, mountainous stages, general classifications and one-day races. Every result only gets assigned to one category. For every stage in a multi-day race, depending on the profile (using a Profile Score) of the stage, it gets counted as a mountainous or a flat one. The formula for calculating the Profile Score is described in detail in (Profile Score Explained). The website procyclingresults.com introduced a system for ranking riders based on their final standings in the races they took part in. The first rider gets awarded maximal amount of points, in the respective point category.

Using the dataset, it is possible to get cumulative sums of points for every rider in every category, as well as the sum of points for every year and every category. These values will be useful in evaluating the two baseline models, which are going to be used to compare against the novel Glicko2 model. As mentioned above, all the code used throughout the work is available on <https://github.com/ant1code/github_cycling_glicko/>.

# Modeling

## Baseline model motivation

To have a relevant comparison for the new model, it is useful to define a baseline model first. That is a model we are going to be comparing the new model against. This baseline model is a naïve approach, which consists of choosing the rider with the most points in the given year as the potential winner. We can also try out another variant of this baseline, such as combining the rider’s point with a measure of team strength (Baseline 1.2).

Another import aspect is that we account for the fact that cycling season runs from March to September. This doesn’t mean that we don’t weight the predictions of the model on these off-high-season races highly, even though we use them for training, these races tend to be not important and more “random”. The reason for that is that most riders might already quit the season by September, or not start racing until March.

## Baseline 1

The very first naïve baseline model is only considering the points awarded during the current season in the relevant discipline. For example, if the Baseline 1.0 was trying to predict a one-day race it would take the current season’s one-day points and simply take the highest-ranked rider as the most likely winner.

As can be seen, the percentage where this model has the correct answer is not very high. If we take the average, the model has a correct prediction in 0.157 races. Note that we don’t need to train any hyperparameters here, because of the way the model is constructed.

## Baseline 2

The second model is doing a very similar thing to the first one, except it is taking a weighted average of the individual’s and team’s current season points. It is then predicting in the same manner, that means, the highest-ranked rider is likely going to win. In the first step, we are going to try out the model with an equal weighting of both of the components. Suddenly, the accuracy of the model improves, and it predicts correctly in 0.286 cases, which is by itself a significant improvement. The accuracy in the last 3 years was 0.264.

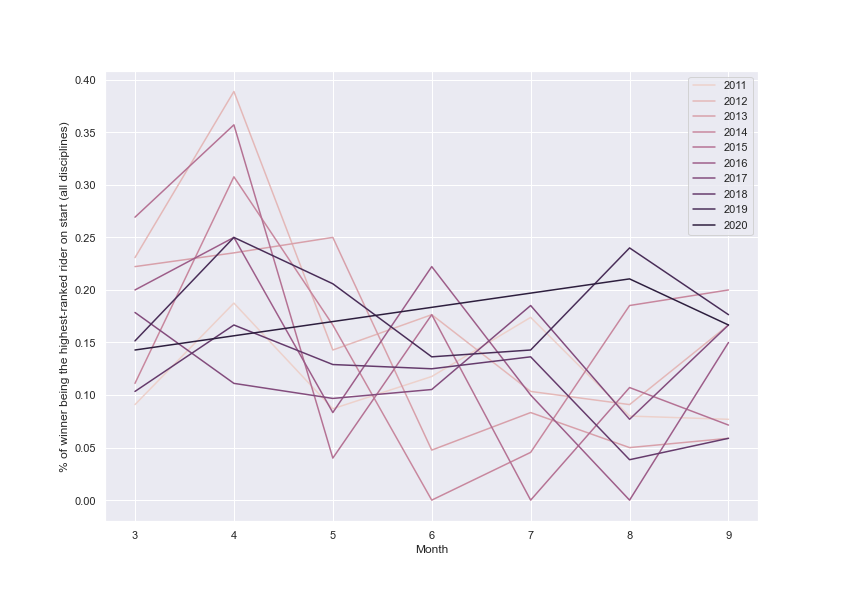


Figure – Percentage of the winner being the highest-ranked rider on the start.

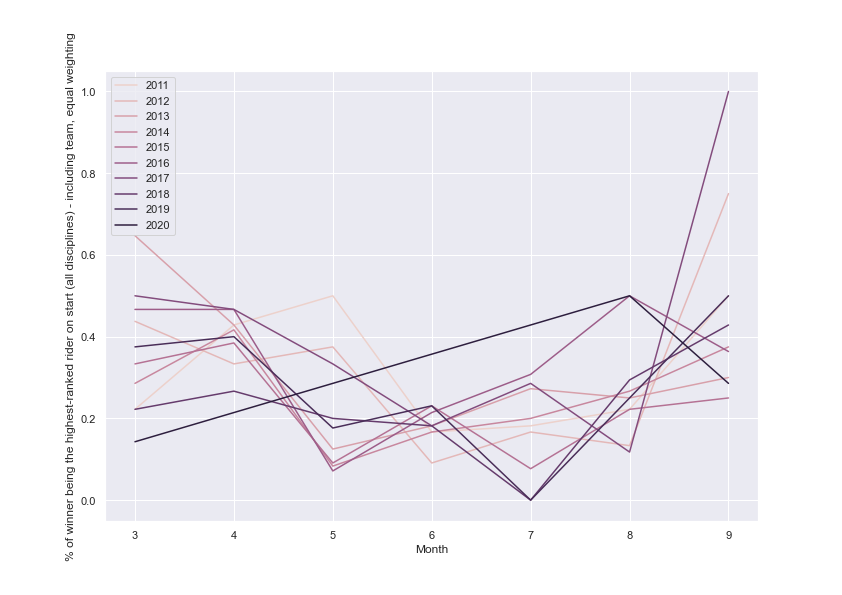


Figure – Probability of winner being one of the three highest-ranked riders at the start for months March until September.

## The Model

In a sense, the construction of the baseline models is unrealistic. Even the better model, the one where we introduce weighted average of team and individual performance spanning the last two years, only works with the current composition of the team. For example, if the team changed, and new riders come in, the calculation behaves as if the team was training and racing together also in the last year, even though every rider was on a different team. This aspect decreases the usefulness of this approach.

This chapter constitutes the main focus of this work, where the main goal is to develop a model that can predict the winner of the race better than just by taking the highest-ranked rider. There are various approaches to incorporate something similar to an elo ranking for cycling. Notably, UCI points, PCR points or others.

For now, the main focus is going to be on modeling multi-day races, because the structure of the race is a bit easier to comprehend than single-day races and additionally, there are fewer of them. Importantly, these races are the most important and prestigious ones with Tour de France being the goal of every professional cyclists. But, the model can and should be extended to one-day races too. It is possible that this would further increase the accuracy and predictive power even for multi-day races.

This chapter is where the whole focus of this work lies, and also the reason why we introduced so many variants to compare against in the previous chapter. The goal is to compare the Glicko2 model against the naïve approaches. Using the Glicko model is appealing for many reasons, for example, the rider who do not directly compete can be compared using their ranking. This is very important – there are two usual variants of races riders take before competing in the Tour de France, and it is vital to be able to compare the relative strength of the riders participating in both of them, even though they haven’t directly competed against each other yet.

As discussed above, a successful model should be able to incorporate several elements.

1. Team strength (importance of which has been growing in the past years)
2. This season’s performance (there is a series of prep-races, who wins those is usually considered to be a favorite to win the whole tour)
3. Past results – it is very rare for a rider to win tdf as his first GC victory, since it requires so much focus, tactics and endurance.

The proposed model is technically a Glicko with an update period of one race, but with a little twist. That twist being that similarly to how a real cycling race works, the element of competition between riders and between teams will be considered.

The proposed model is actually two Glicko models, working side-by-side to achieve better accuracy and predictive power. The first model is focusing on the competition of teams, and the second one is focusing on the competition of team leaders. The logic of the model is the following:

1. Every team has a ranking based on their members’ strength
2. Race is a head-to-head competition of teams, can be imagined   
   as 122 = 144 individual races – with update period of one race
3. Every team’s GC ranking gets updated
4. Additionally, every team has 2 leaders – the same stuff, head-to-head, update ranking and standard deviation
5. Every leader’s ranking gets updated
6. Team rankings and rider ranking get weighted to produce the elo used at the start of the next race

The real beauty of this approach lies in the fact that it captures one of the most important aspects of cycling, which is the fact that it’s both a team sport, as well as an individual sport. Another challenge that this approach is trying to solve is the fact that the exact team and individual component/ranking’s contributions varies, depending on the race type. More specifically, winning a one-day race relies more on individual’s performance than a team strength. On the other hand, winning a three-week race without a strong team is very challenging. By weighting the team and the individual ranking using a parameter, it is possible to adjust the individual and team part, depending on the type of a race.

## Example

Starting off with a very simple example, we have a multiple-day race with three teams. The data is based on the first race of the 2016 season – Santos Tour Down Under in Australia. For the sake of simplicity, suppose there are only three teams, instead of the twelve usual - Orica Greenedge, Team Sky and BMC Racing Team.

Each team is going to have two leaders - as is common in professional cycling nowadays. The reason why teams start with two leaders is that if one of the leaders crashes or experiences technical difficulty, the other can still finish well. This model will assume, that the two leaders to be the two highest-ranked riders of every team before the start of the race.

Let’s see the starting list and only display the two leaders in every team. Just for the purpose of testing, we will assume the Elo to be 1500 and the RD to be 350. These are the recommended starting values for the Glicko-2 system as per (Glickman, 2013). The starting values of ranking and ranking deviation are shown in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Team | r | RD |
| GERRANS Simon | Orica | 1500 | 350 |
| PORTE Richie | BMC | 1500 | 350 |
| HENAO Sergio | Sky | 1500 | 350 |
| ROHAN Dennis | BMC | 1500 | 350 |
| *GERAINT Thomas* | *Sky* | *1500* | *350* |
| IMPEY Daryl | Orica | 1500 | 350 |

Table Starting values

Then suppose the final results turn out to look like this:

|  |  |  |  |
| --- | --- | --- | --- |
| Rank | Name | Team | Team ranking |
| 1 | GERRANS Simon | Orica | 1 |
| 2 | PORTE Richie | BMC | 2 |
| 3 | HENAO Sergio | Sky | 3 |
| 4 | ROHAN Dennis | BMC | - |
| 5 | GERAINT Thomas | Sky | - |
| 6 | IMPEY Daryl | Orica | - |

Table Finish

Then, we can simply create a matrix of all individual head-to-heads for teams and for the riders.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| J | Name | GERRANS Simon | PORTE Richie | HENAO Sergio | ROHAN Dennis | GERAINT Thomas | IMPEY Daryl | # opponents beaten |
| 1 | GERRANS Simon | 0 | 1 | 1 | 1 | 1 | 1 | 5 |
| 2 | PORTE Richie | 0 | 0 | 1 | 1 | 1 | 1 | 4 |
| 3 | HENAO Sergio | 0 | 0 | 0 | 1 | 1 | 1 | 3 |
| 4 | ROHAN Dennis | 0 | 0 | 0 | 0 | 1 | 1 | 2 |
| 5 | GERAINT Thomas | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 6 | IMPEY Daryl | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table head-to-head, 1 - rider in row won, 0 - rider with name in a column won

It might seem unnecessary to create a table just to show how many opponents has the rider beaten, but that is only because all the riders have equal elo and RD (for now). Once the riders’ ranking start to differ, the computations are going to be more involved. To compute the final elo, we use the formula from the Step 2 of the original Glicko system (glicko). Let’s take rider named Sergio Henao as an example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| J | Rj | Rdj | G(rd)j |  | sj |
| 1 | 1500 | 350 | 1.495 | 0.5 | 0 |
| 2 | 1500 | 350 | 1.495 | 0.5 | 0 |
| 4 | 1500 | 350 | 1.495 | 0.5 | 1 |
| 5 | 1500 | 350 | 1.495 | 0.5 | 1 |
| 6 | 1500 | 350 | 1.495 | 0.5 | 1 |

Table Sergio Henao's result

All the riders start the race with a given Glicko ranking, then all play head-to-head against each other at the same time, and then, after the race, the ranking gets updated. Let’s take an example of Sergio Henao. He finished third, which means he lost against all riders, except for the first two. We want to reflect this in the updating of his rank. Because all the riders have the same ranking, this computation is simple.

His ranking gets updated using the formula shown above:

And calculating the new rating deviation:

Again, this being a very simple example, as the new rating deviation turns out to be the same for all of the riders.

If we repeat this procedure for every rider, we get the following results:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Team | r | RD | r’ | RD’ |
| GERRANS Simon | Orica | 1500 | 350 | 1713 | 99 |
| PORTE Richie | BMC | 1500 | 350 | 1628 | 99 |
| HENAO Sergio | Sky | 1500 | 350 | 1543 | 99 |
| ROHAN Dennis | BMC | 1500 | 350 | 1457 | 99 |
| GERAINT Thomas | Sky | 1500 | 350 | 1372 | 99 |
| IMPEY Daryl | Orica | 1500 | 350 | 1287 | 99 |

Table Old and new rankings and deviations

Now let’s look at the teams. The exact same procedure is repeated, but the important difference is that when calculating the rank of the team, we choose the highest ranked rider on the team. Otherwise we let teams start with the same default values as the individual riders. Let’s take an example of team A. They have one rider ranked 4th, and another ranked 6th. And, because the team B finished 2nd and 3rd, team A actually finished 3rd in the competition of teams. So we repeat the same as for the riders and arrive at the following results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Team | r | RD | r’ | RD’ |
| Orica | 1500 | 350 | 1585 | 99 |
| BMC | 1500 | 350 | 1500 | 99 |
| Sky | 1500 | 350 | 1415 | 99 |

Table Old and new ratings and deviations for teams

Note that the new ranking deviation RD’ is the same for the teams, as well as for the riders, because the starting values are the same for everyone. This is going to change with time though. Now when we have both the team ranking, as well as the rider’s ranking, it is possible to combine them both. Importantly, for this example we will use an equal weighting of teams and riders, but the parameter gamma is a very interesting one. In particular, it is going to be interesting to see what gamma is the optimal one depending on the category of the race. This way, the model can be easily adjusted to the type of race just by modifying this one parameter.

Let’s get back to our example. Imagine there is another race. Conveniently, we already have results we can use from the past race. By taking a ranking as (team + individual)/2 and applying procedure identical to the last race, we get the following starting table:

Table Starting values in the second race

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Starting individual elo | Starting team elo | Starting  Combined elo | Starting RD | Name | Team | Finished |
| 1713 | 1585 | 1649 | 99 | GERRANS Simon | Orica | 3 |
| 1628 | 1500 | 1564 | 99 | PORTE Richie | BMC | 2 |
| 1543 | 1415 | 1479 | 99 | HENAO Sergio | Sky | 1 |
| 1457 | 1500 | 1478.5 | 99 | ROHAN Dennis | BMC | 4 |
| 1372 | 1415 | 1393.5 | 99 | GERAINT Thomas | Sky | 6 |
| 1287 | 1585 | 1436 | 99 | IMPEY Daryl | Orica | 5 |

When every rider has a different ranking, the computation becomes a little bit more involved. Let’s calculate the ranking for Simon Gerrans after this race. Let’s say he finished 2nd, behind Richie Porte. It is necessary to construct a following table for every rider. In Table, it can be seen that for example the expected value for a head-to-head between Simon Gerrans and Richie Port is 0.62 viewed from the side of Simon Gerrans. This means, that we expect Simon Gerrans to win, on average 62% of the time when racing against Richie Porte based on the ranking we have. But, this time Simon Gerrans lost.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Opponent | Rj | Rdj | G(rd)j |  | Actual outcome |
| PORTE Richie | 1564 | 99 | 1 | 0.619941359 | 0 |
| HENAO Sergio | 1479 | 99 | 1 | 0.726830072 | 1 |
| ROHAN Dennis | 1478.5 | 99 | 1 | 0.727401166 | 1 |
| GERAINT Thomas | 1393.5 | 99 | 1 | 0.81317534 | 1 |
| IMPEY Daryl | 1436 | 99 | 1 | 0.773139913 | 1 |

Table Second race for Simon Gerrans, including team ranking

Now, identically to the previous race we have the same calculations as described above.

Now suppose all the riders finished and by computing the new rankings we arrive at the results shown in Table 9. Note that to compute these results, the combined rankings consisting of 50% individual ranking and 50% team ranking were used.

Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | RD | Name | Team | Finished |  |  |
| 1500 | 1564 | 99 | PORTE Richie | BMC | 1 | 1657 | 1535 |
| 1585 | 1649 | 99 | GERRANS Simon | Orica | 2 | 1664 | 1573 |
| 1415 | 1479 | 99 | HENAO Sergio | Sky | 3 | 1510 | 1393 |
| 1500 | 1478.5 | 99 | ROHAN Dennis | BMC | 4 | 1463 | 1535 |
| 1585 | 1436 | 99 | IMPEY Daryl | Orica | 5 | 1389 | 1573 |
| 1415 | 1393.5 | 99 | GERAINT Thomas | Sky | 6 | 1316 | 1393 |

Because we want to track both the teams, as well as the individuals, the teams’ ranking has to be added too. Once the combined r’ and team r’ are calculated, since:

We can see how the ranking works and sorts the riders. For example, Sergio Henao, who started at 1479, which was more or less the same as Rohan Dennis or Daryl Impey finished at 1510, significantly more than the mentioned riders, since he beat both of them. On the other hand, Geraint Thomas confirmed his role of an underdog, and after two 6th places, has significantly less points than all other riders.

Interestingly, team Orica is still the highest-ranked team, even though they lost to BMC and lost some points. BMC came close to Orica in terms of ranking, but not enough to beat them.

One might ask what purpose the Ranking deviation serves, and it is true, in this example it serves no purpose at all, other than complicating things even further. But in the actual season, not every rider is competing in the same number of races, therefore the ranking deviation is going to be different for every rider. Additionally, suppose the rider was sick for two months and couldn’t compete. We want to somehow take this fact into account, in this case by increasing the ranking deviation, because the ability of this rider is suddenly becoming less “certain”.

# Discussion of the Results

Let’s look at the results in more detail and for the sake of clarity, let’s only look at multi-day races, such as Tour de France or Giro d’Italia. Because of the more predictable nature of these races (as opposed to one-day races), it is expected that the model is going to be able to predict the winner with higher accuracy. Additionally, the team plays a much bigger role in the multi-day, long and exhausting races. Some team members spare the resources of the leader by riding in front of him most of the race, so that he can stay covered from the wind and save energy for the deciding parts of the race.

## Glicko-in-Glicko with teams

This is the exact model as described in the last chapter. It can be seen that the performance of the model is getting slightly better over time, but not overly so. In general, the performance of the model is quite disappointing.

To infer the best weighting parameter, Monte Carlo simulation was used and surprisingly, the best predicting setup is achieved by setting the weighting parameter to 0, i.e. disregarding the team component altogether and only predicting based on the individual riders’ past performance, which bri ngs us to the second iteration of the model.

## Glicko without teams

As can be seen, the model where teams are disregarded is consistently outperforming the model with teams. But the performance is still only mediocre, when compared to the baseline model. The baseline model 2 achieved average accuracy of 0.286 on all data, including every one-day race and every stage of every multi-day race. When calculating only multi-day races, its performance was hovering around 0.4, which is almost double that of “Glicko without teams”.

Note that this only applies to Multi-day Races. When modeling one-day races or individual stages, the performance of the Glicko model would still be likely even worse.

After applying the Glicko-in-Glicko model exactly as described in the last chapter on the dataset of the last 10 years of competition cycling results, there are two interesting results to be seen:

1. The model is performing worse than both of the baseline models, by a significant margin.
2. The Glicko-in-Glicko model is performing better in case the team relative performance is not accounted for.

# Further Extensions

In this chapter, I will discuss possible extensions of the model and round up the text.

# Works Cited

Arie-Willem de Leeuw, M. H. (2020). Time Series Regression in Professional Road Cycling. *Discovery Science*, (pp. 689-703).

Boice, J. (2020, 6 2). *How Our Club Soccer Predictions Work*. Retrieved from Fivethirtyeight: https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/

Elo, A. E. (1978). *The rating of chessplayers, past & present.* New York, NY: Arco Publishing.

Glickman, M. E. (1995). The glicko system. *Boston University*(16), pp. 16-17.

Glickman, M. E. (2013, 8 30). *Example of the Glicko-2 system*. Retrieved 4 26, 2021, from Example of the Glicko-2 system: http://glicko.net/glicko/glicko2.pdf

Glickman, M. E. (n.d.). *GLICKO-BOOST Deloitte/FIDE Chess Rating Challenge*. Retrieved 4 28, 2021, from GLICKO-BOOST Deloitte/FIDE Chess Rating Challenge: http://www.glicko.net/glicko/glicko-boost.pdf

Kholkine, L. e. (2020). A Machine Learning Approach for Road Cycling Race Performance Prediction. *International Workshop on Machine Learning and Data Mining for Sports Analytics.*

Pradhan, S. A. (2020). “Power ranking” professional circuit eSports teams using multi-criteria decision-making (MCDM). *Journal of Sports Analytics, 6*(1), 61-73. doi:10.3233/JSA-190420

*Profile Score Explained*. (n.d.). Retrieved 4 28, 2021, from procyclingstats.com: https://www.procyclingstats.com/info/profile-score-explained