Snelling Road Race • Feb 22 2020 • Snelling, CA — Photo by Katie Miu

Predicting winners in cycling races with Machine Learning

With 95% accuracy, we can use performance data to predict which athletes will be on the podium

Bruno Gregory Follow 10 min read · Mar 5, 2021

85 2

A little bit of context

At the beginning of 2020, I decided to go back to amateur cycling racing after more than ten years. But this time, everything was different. It would be my first time racing in The United States and in better shape than when I was a teenager.

As soon as the season started, I found myself checking the list of registered competitors and categories and thinking: How can I increase my chances of having a good result? I didn't know any of the racers and teams. Basically, all I could do was to do a course recon and trust my physical conditioning.

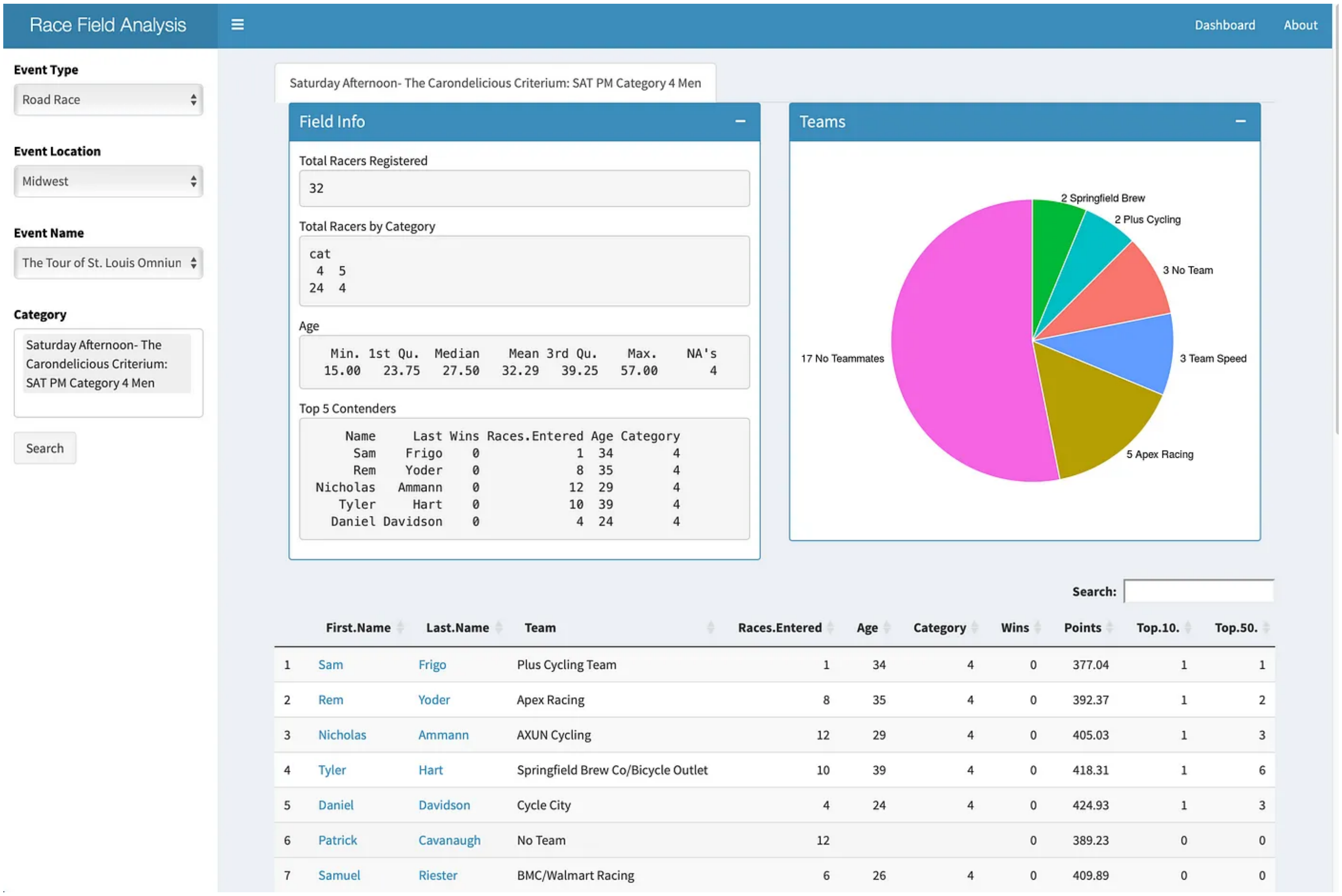
As it is typical of myself, I didn't stop thinking about it, and in no time I was checking each contestant's name and searching for their information. I then discovered that USA Cycling (the American governing body for bicycle racing) has a history of all races and athletes, including amateurs and professionals. Bingo!

I have always been passionate about data analysis, which among other projects, resulted in Graava — a startup that I founded and where I developed an algorithm to edit videos automatically using data from sensors and images. So when I found myself analyzing the athletes' data and gathering information manually, I immediately came up with the idea of developing a tool to automate this process and analysis.

First version — Race Field Analysis

I did some initial tests and analysis of the athlete's data, and I liked the results. With the first tests performed in R language, I realized that the results could bring me competitive advantages. I mentioned that to some of my cycling teammates, and they all responded the same way: How can I have access to this information as well?

That was how the Race Field Analysis tool was born. An open-source side project, developed in R, and using the Shiny framework. The original idea was to make it possible for my teammates and friends to analyze the dynamics of a race and develop strategic plans to obtain a better result. The tool would provide information such as which team would have more control of the race and which athletes were more experienced.



Event Type: Road Race

Event Location: Midwest

Event Name: The Tour of St. Louis Omniun

Category: Saturday Afternoon- The Carondelicious Criterium: SAT PM Category 4 Men

Saturday Afternoon- The Carondelicious Criterium: SAT PM Category 4 Men

Field Info

Total Racers Registered

```
32
```

Total Racers by Category

```
cat
 4  5
24  4
```

Age

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  15.00   23.75   27.50   32.29   39.25   57.00       4
```

Top 5 Contenders

```
      Name    Last Wins Races.Entered Age Category
       Sam   Frigo    0             1  34        4
       Rem   Yoder    0             8  35        4
  Nicholas  Ammann    0            12  29        4
     Tyler    Hart    0            10  39        4
    Daniel Davidson   0             4  24        4
```

Teams

Search:

	First.Name	Last.Name	Team	Races.Entered	Age	Category	Wins	Points	Top.10.	Top.50.
1	Sam	Frigo	Plus Cycling Team	1	34	4	0	377.04	1	1
2	Rem	Yoder	Apex Racing	8	35	4	0	392.37	1	2
3	Nicholas	Ammann	AXUN Cycling	12	29	4	0	405.03	1	3
4	Tyler	Hart	Springfield Brew Co/Bicycle Outlet	10	39	4	0	418.31	1	6
5	Daniel	Davidson	Cycle City	4	24	4	0	424.93	1	3
6	Patrick	Cavanaugh	No Team	12			0	389.23	0	0
7	Samuel	Riester	BMC/Walmart Racing	6	26	4	0	409.89	0	0

Race Field Analysis — Analysis and prediction screen of The Tour of St. Louis race

However, everything changed when I realized that, with the data collected, I could actually make predictions of which athletes would be on the podium. The information of the possible winners could change the whole dynamic of the race. By knowing this information, a contestant could mark the race's potential winners and have a better overall result, including choosing to beat them in the final miles.

The first version of the tool is available at this link. With the tool, it is possible to analyze all the races and categories of cycling in the US, including Road, MTB, Gravel, Track, Cyclocross etc.

As soon as my teammates started using the tool, it was fun to see all kinds of results:

> *“Bruno nailed the 55+ 1/2/3 winner… full field of 60 riders. Consistent with Bruno’s app, 42 old guys had teammates, but the “favorite”, new 55 year old Tom Lyons had no teammates on paper, but lots of help in the form of his old Thirsty Bear teammates.” Larry*

Women Cat 3 Podium • Snelling Road Race

“Thankfully for us, Bruno’s race predictor wasn’t quite as accurate for the women’s 3 :)” Jenn

In all cases, I was happy with the results and how the tool was helping the team. Even with results that were not accurate (like in Jenn’s case), it was fun to see the teammates' comments, in a way saying: "I broke your algorithm, Bruno!" In fact, through machine learning, the algorithm collects more information about the athlete at each race. Consequently, in the next race, this athlete will have a better position in the prediction, and the algorithm will be more precise.

Covid-19

As soon as I released the first beta version of the tool, Covid-19 broke out, and the whole world suffered the consequences of the pandemic and shelter-in-place orders. Along with most group activities, all races have been canceled.

But athletes have not stopped training. Actually, cycling has never grown as much as it did in 2020. In their thirst for competition, cyclists found the solution on virtual training platforms.

Zwift

Image credit: Zwift

Zwift is without a doubt the most famous virtual platform and with the largest number of users today. Basically, Zwift is a multiplayer game where you connect your bike with the computer. As you exercise on the bike, your avatar moves in the virtual race. It is a great evolution of indoor training. What used to be boring and monotonous is now dynamic and super competitive. And guess what? There are races every day with athletes from all over the world. Bingo!

Second version — ZRace

The idea of developing a variation of the Race Field Analysis tool for Zwift came as soon as I started participating in its virtual races. Because it is a virtual race, all athletes are connected, and Zwift uses several sensors from the bike (power meter, cadence, and speed) and the athlete (heart rate) to measure the percentage of effort. That was just perfect. With data such as heart rate, power, weight, age, sex, speed, cadence, historical results, among others, the analysis could be much more accurate and effective.

Thus, ZRace was born: A tool designed for Zwift races where athletes can analyze which race and category fit best with their profile and gain a competitive advantage.

ZRace — analysis and prediction screen of 3R Figure 8 Hilly race

The tool can be accessed at this link: http://zrace.bike

ZRace analyzes all athletes registered in a race and predicts possible winners. It also analyzes each category and presents the average power required for you to have a good result. In addition, athletes with specific profiles are identified, such as climber, sprinter, and time-trialist. This way, depending on the race's course, it is possible to predict who will have a better result or even who you should keep an eye on for a certain part of the race.

Like the first version developed for non-virtual races, ZRace performs analysis and prediction of the winners. However, thanks to the large number of additional information from athletes provided by Zwift, ZRace has become a more powerful tool and with much more accurate predictions.

Under the hood — Statistical models

I have been developing solutions following Agile principles and methodologies for a long time. With this project and the statistical model used, it would not be different. It is very difficult to build a great and complex model from scratch. But it is relatively easy to build a simple model and then iterate, improve the results with new tests. After each iteration, you learn more about the results, importance of each variable and then build the next iteration with the acquired knowledge.

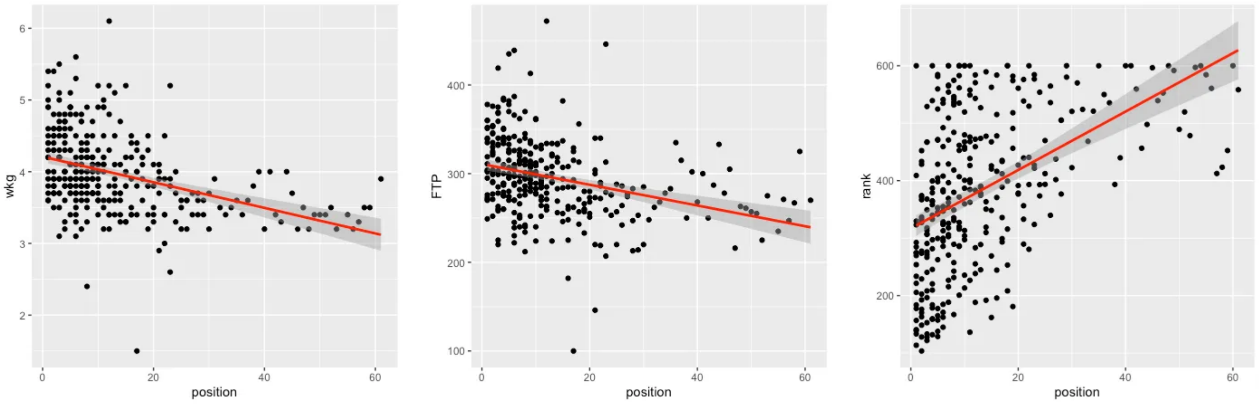
My first iteration idea was to build a simple model that could be used to evaluate and analyze the project end to end. I used an expert model to predict the winners. I looked for the number of victories of each athlete in the last ten races, and the athlete with the highest number of victories would be the winner.

Even with an extremely simple model, this iteration was very important for the whole project. It helped me develop a complete MVP without committing a lot of time and resources to build the actual model. With the results and evaluations of this iteration, I felt confident to go deeper into the project, start collecting and cleaning data, developing the back and front end, and when everything was ready, improve the predictive models.

The predictive model iterations were:

1- Expert model: As mentioned, this model was used as the starting point, and due to the relatively fair result, the project was started.

2- Linear regression: The second iteration of the predictive model was performed using linear models. This was a natural step after an exploratory data analysis and the detection of important variables and their correlations. Specifically, the multiple linear regression model obtained the best results and was used in this iteration.



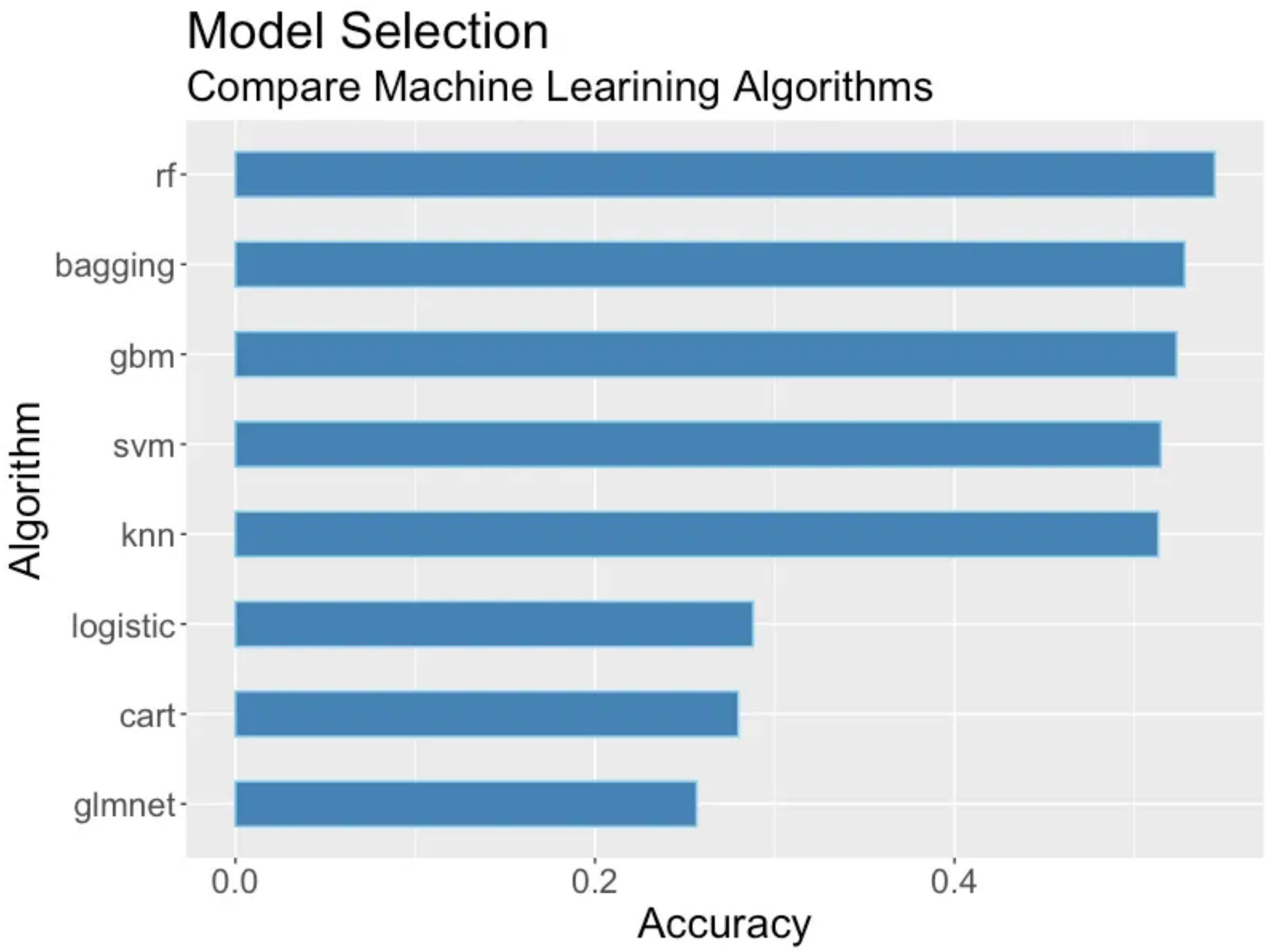
3- Model selection: The third iteration aimed to evaluate and identify the definitive predictive model. Once identified, the goal would be to tune the model until it presented the most precise result possible. For this iteration, more complex models and machine learning were essential to get to the final result.

The goal was to use the most-accurate model for the dataset. This is the essence of predictive modeling. Basically, no one knows what is the best algorithm to use for a given problem or dataset. If you already have experience with a given problem, you probably would not need to use machine learning due to your deep knowledge in the field.

That said, my plan was to use the full potential of machine learning to evaluate and identify the best and most-accurate algorithm for this prediction.

First step: A much larger sample was gathered so that the training of the models would be performed and thus present a better result. 1,700 observations (races) are being used in this test. The number of observations is not ideal, but I am gathering more over time.

Second step: It is essential to have a good set of algorithm representations (lines, trees, instances, etc.), as well as algorithms for learning those representations. At the end of this process, more than 10 algorithms were tested, including: Linear Discriminant Analysis, Logistic Regression, GLMNET, SVM Radial, kNN, Naive Bayes, CART, Bagged CART, Random Forest, Stochastic Gradient Boosting, among others.



After training the models and running tests, 3 algorithms performed well: Bootstrap Aggregating, Generalized Boosted Regression Modeling, and Random Forest. The latter of these had the best performance, so it was the model chosen to go ahead and start fine-tuning.

Random Forest (RF) is a robust machine-learning algorithm that creates hundreds of decision trees based on the dataset variables. With the combination of the trees' decisions, the model combines the predictions of the trees to produce a prediction that is more accurate.

Results

After tuning the model, the accuracy rate in predicting race results reached 58%! This is a very good result given that the model predicts each athlete's **exact** position in the race result. On average, the races have 25 athletes per category, but there are cases of 50, 80, or even more participants.

In order to tune the model and improve accuracy from 54.4% to 58%, some parameters were changed, and the model trained again until the ideal set of parameters was reached.

The ideal model uses 500 decision trees and 10 variables. The graph below shows the importance of each variable for the model, and it is possible to understand why the model is more accurate with 10 variables instead of 16.

Random Forest Feature Importance

But here comes the most exciting part. The problem we are trying to solve here and why the development of this project started was to identify which athletes will be on the podium. That means we want to identify the Top 5. With this information, you can mark the athletes in the race and thus have a better result.

That said, my model has a 75% accuracy rate in the races' predictions that in the Top 5 (on the podium), 4 athletes will be present regardless of the order. And a 95% accuracy rate that 3 predicted athletes will be in the Top 5!

ZRace — Result vs Prediction

This result made me extremely happy with the work done. And I can tell you with 95% accuracy if you are thinking about racing on Zwift now you know who you need to win. :)

Conclusion

- Working on this project was fun. Comparing the predictions with the races' actual results shows an immediate and real benefit, which was very motivating.
- As much as I am pleased with the current results, I will continue to gather data from the races to reach the magic number of 10,000 observations (races) and then perform another training session for the models. I should test new algorithms also including Neural Network.
- The first version of the project developed for non-virtual racing was an important learning step for developing the second version with a focus on racing on Zwift. But today, it is outdated. I have plans to revisit it as soon as the races resume and apply all the knowledge acquired at ZRace.
- While analyzing the importance plot of the RF model, I noticed an intriguing variable. The weight of the athlete has a high factor in the model's decisions. Theoretically, we know the importance of W/kg and FTP in cycling, but weight alone is new information. Could I explore the data to reach the ideal weight of a super cyclist? In case I come up with an answer, this may be a good topic for a future article.
- So far, all feedback on the tool and benefits for athletes has been positive. And this is rewarding. Please feel free to comment, suggest

changes and features. Any feedback is welcome.

Machine Learning Cycling Zwift Sports Data Science

Written by Bruno Gregory

27 followers · 40 following

Follow

Entrepreneur, Product Manager, and Engineering Leader. Enthusiast of Data-Driven Products and Data Science. From Brazil to Silicon Valley

Responses (2)

George Ogden

What are your thoughts?

Nick Keat
Mar 27, 2021

An interesting idea, and I took a look at a couple of results. One thing that seems obviously missing from the model is the route being raced - is it a hilly vs flat route, is it going to come down to a sprint from a large proportion of the starters… more

1 1 reply Reply

Gina Durante
Apr 25, 2021

Thanks for publishing this! I just learned about your blog and website today. How are you pulling data into the model? Will you be posting to github any time soon?

1 reply Reply

More from Bruno Gregory

Bruno Gregory

Prevendo ganhadores em corridas de ciclismo com Machine Learning

Com 95% de precisão, pode-se utilizar dados históricos de desempenho para prever quais atletas estarão no pódio

Mar 5, 2021

See all from Bruno Gregory

Recommended from Medium

In Long. Sweet. Valuable. by Ossai Chinedum

I'll Instantly Know You Used Chat Gpt If I See This

Trust me you're not as slick as you think

May 16 16.3K 925

In Predict by iswarya writes

GPT-5 Is Coming in July 2025—And Everything Will Change

"It's wild watching people use ChatGPT… knowing what's coming." —OpenAI insider

Jul 7 4.2K 167

Jordan Gibbs

ChatGPT Is Poisoning Your Brain…

Here's How to Stop It Before It's Too Late.

Apr 29 22K 1118

In Coding Beauty by Tari Ibaba

This new IDE from Google is an absolute game changer

This new IDE from Google is seriously revolutionary.

Mar 11 6.1K 366

In The Medium Blog by Zulie @ Medium

Medium versus Substack: Six reasons writers pick Medium

Why Medium is the best platform for most writers

5d ago 8.8K 287

Sohail Saifi

Kubernetes Is Dead: Why Tech Giants Are Secretly Moving to…

I still remember that strange silence in the meeting room. Our CTO had just announce…

✦ Jun 7 3.3K 131

See more recommendations

Help Status About Careers Press Blog Privacy Rules Terms Text to speech