MATH 111A Project-
Modeling multi-stage competitive cycling events

Dominic Clark
dmclark@ucsd.edu

December 11, 2020

Abstract

In my project I have modeled competitive road cycling using various factors of the riders. Using modeling as used in modern sports betting as inspiration, I used various factors of the different riders in order to estimate performance in previous and future cycling races, and also determine which of these factors are most correlated with success in cycling events.

1 Introduction

For my project I have looked into road cycling and how we can apply various statistical modeling techniques to competitive road cycling. Overall, my project has two parts. The first section of my project is doing linear regression on various aspects of individual cyclists, such as height, weight, team ranking, and age. I then compare these factors to the recent performance of these cyclists to see which factors are the most correlated with success in elite level cycling.

As a second section, I have built a statistic simulation using a multinomial logistic regression model to see how the factors that I analyzed in the first part of this project can be used in simulations of cycling events in order to predict overall performance. My focus for both sections is general classification riders in multiple stage cycling events. Factors I have considered for these simulations are success in past cycling events, success of the team in past events, as well as the factors I found meaningful in my linear regression portion of the project.

As for my motivation for this project, I am very much interested in the ways that the data in cycling can be used. When I was getting into cycling, I was surprised by just how much data is available to cyclists from devices such as power meters, GPS trackers and heartrate monitors. I believe that we can use this data in many interesting ways beyond the scope of this project as well, but for this project, I have focused on how we can use this data in the context of competitive road cycling events.

In doing my research for this project, I was unable to find regression analysis applied to competitive cycling in the same way that I have used it. However, in the article "Riding against the wind: a review of competition cycling aerodynamics.", they used regression analysis in order to analyze rolling resistance and aerodynamics for cyclists. Thus, the idea of applying linear regression has been used before even in the field of competitive cycling, but not in this way as far as I know.

As mentioned in the article by David Percy, stochastic "models have been applied successfully to select strategies and to predict outcomes in the context of games, tournaments and leagues." Thus, there is a precedent for using similar simulations in analysis of sports, but in doing my research I was unable to find a similar model being applied to competitive cycling. In this article however, they built a stochastic simulation and applied it to modeling a university boat race, so similar approaches have been used in sports analysis to this in the way that I have applied my model to cycling events.

My model draws a lot of inspiration from model outlined in the referenced article by Benter. The article talks about a model that Benter used in the Hong Kong horse races in the 90s to great success. I was very much interested in looking at the parallels between horse racing and road cycling races and seeing how we can apply those parallels to modeling them.

2 Data

For my data, I gathered it all from a website called procyclingstats.com. On this website we have access to lots of the statistics about various pro cyclists, such as their height, weight, age, performance in UCI events, and also a lot of data about the pro cycling teams.

One thing that is worth noting about this data is that some of it changes, so using the procyclingstats data may introduce extra error. For example, over the course of these multi stage races, racers often lose one to two kilograms and their performance changes throughout the races. Also, gathering the data from this website was moderately difficult from a webscraping perspective in order to get the data that I wanted, so I ended up only analyzing the top 17 GC cyclists on the websites' leaderboard at the time I most recently updated my data, which was relatively shortly after the Giro d'Italia ended.

3 Regression analysis on factors of pro cyclists

First, I did regression analysis on various factors of cyclists, such as height, weight and FTP and compared them to the success of the individual cyclists in their recently completed multi stage races.

The factors that I chose are as follows:

- Height
- Weight
- Age
- FTP (I will go into this in a bit)
- FTP/Weight (as a factor)
- current Team ranking in 2020 season

I then compare all of these factors to the average GC finish position in last 5 multi stage races that each of these cyclists competed in. I considered comparing the results to average finish time in races, but there were a few problems with that method. First, not all of these cyclists are competing in the same races, so it is hard to measure time across multiple races with different conditions, courses and stages. Second, the cyclists don't always race for time but rather to try to win, so performance would be better measured by the place these cyclists reached in the races in that case. One flaw of this method is that it rates those who have competed in less competitive races higher as they will have an easier time winning in races that are less competitive.

3.1 Approximating FTP

For all of these cyclists, I was unable to find up to date power data. Therefore, instead I approximated the FTP by using a power formula that was outlined in the article "Validation of a Mathematical Model for Road Cycling Power". In that paper, they produced the following model for approximate power:

$$P_{TOT} = P_{AT} + P_{KE} + P_{RR} + P_{WB} + P_{PE}/E_C$$

Where,

$P_{AT} = .5\rho V_a^2 V_g(C_d A + F_w)$
$P_{KE} = .5(m_t + I/r^2)(V_{gf}^2 - V_{gi}^2)/(t_f - t_i)$
$P_{RR} = V_g C_{rr} m_t g * cos(tan^{-1}(G_f))$
$P_{WB} = V_g(.091 + .0087 V_g)$
$P_{PE} = V_g m_t g * sin(tan^{-1}(G_f))$

and
$P_{TOT} = total\ power\ required$
$P_{AT} = power\ required\ to\ overcome\ aerodynamic\ drag$
$P_{KE} = power\ required\ to\ overcome\ kinetic\ energy$
$P_{RR} = power\ required\ to\ overcome\ rolling\ resistance$
$P_{WB} = power\ required\ to\ overcome\ drag\ in\ wheel\ bearings$
$P_{PE} = power\ required\ to\ change\ potential\ energy$
$\rho = air\ density$
$V_a = air\ velocity$
$V_g = ground\ velocity$
$C_d = coefficient\ of\ drag$
$A = frontal\ area\ of\ bike\ and\ rider\ system$
$F_w = wheel\ rotation\ factor$
$m_t = total\ mass\ of\ bike\ rider\ system$
$I = moment\ of\ inertia\ of\ wheels$
$r = outside\ radius\ of\ tire$
$V_{gf} = final\ ground\ velocity$
$V_{gi} = initial\ ground\ velocity$
$t_f = outside\ radius\ of\ tire$
$t_i = initial\ time$
$t_f = final\ time$
$C_{rr} = coefficient\ of\ rolling\ resistance$
$g = acceleration\ due\ to\ gravity$
$G_f = road\ gradient$
$E_C = efficiency\ of\ chain\ drive$
I made the assumption that the weight of the bike is 6.8 kg which is the UCI weight limit for bikes in these races. As a simplification, I made the assumption that for road gradient was the average across the entire course for the course that the cyclist was racing on. To get the approximate FTP, I took the most recent Time Trial result for each of the cyclists in a multi stage cycling race and then applied that time to this formula to approximate power. I also treated several factors I was unsure about the same as the way that the article treated them in the sample calculation that they provided. I then approximated this over an hour period as opposed to the duration of the Time Trial the cyclists competed in to produce approximate FTP.

3.2 Results of Approximating FTP

I'm not going to go into solving the approximate FTP for the cyclists in these events, because it is just a bunch of computations using the assumptions that I mentioned in the section above. In the referenced paper that produced this equation, they did a sample calculation using this equation and I just followed that template when solving for the power of the cyclists that I was analysing. The table below shows the results of approximating the FTP for the 17 cyclists that I chose to analyse for this project using the formula and assumptions from above.

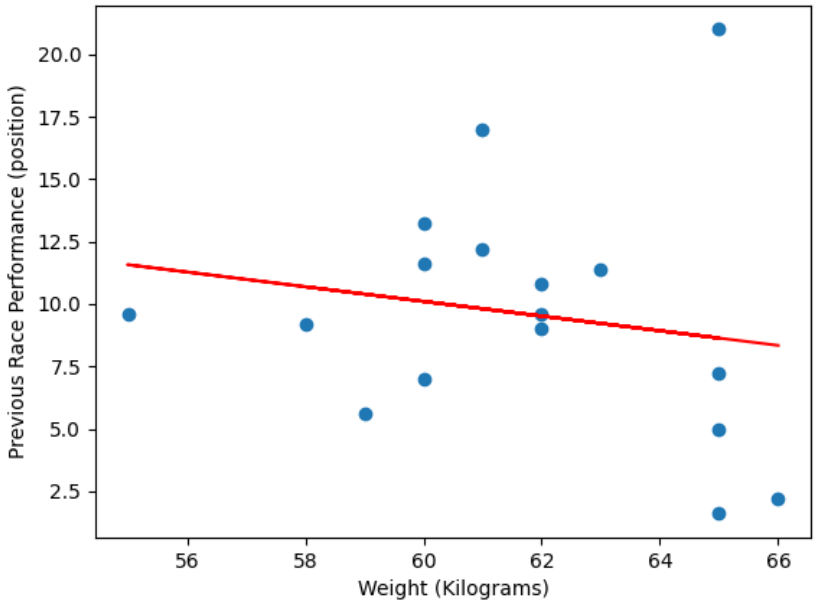
Cyclist name	FTP(Watts)
Primoz Roglic	404
Tadej Pogacar	412
Richie Porte	368
Remco Evenepoel	358
Wilco Kelderman	364
Guillame Martin	305
Jakob Fuglsang	353
Diego Ulissi	334
Nairo Quintana	313
Simon Yates	329
Mikel Landa	335
Rafal Majka	321
Richard Carapaz	354
Jai Hindley	336
Enric mas	341
Tao Geoghegan Hart	366
Ajejandro Valverde	342

3.3 Results of Regression Analysis

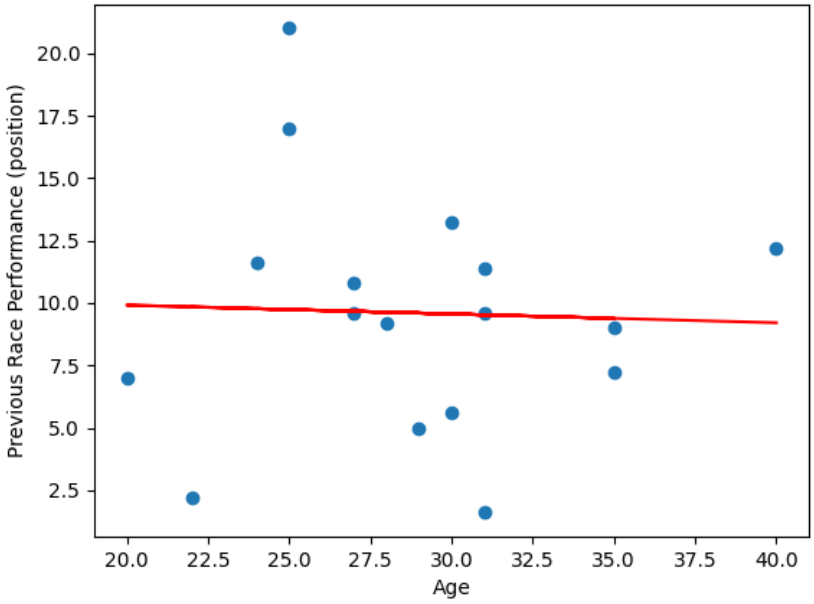
Below is the Python code I used for creating the visualizations for linear regression. I have attached the cyclistData.csv file to this assignment.

```
import numpy as np
# For visualizations
import matplotlib.pyplot as plt
# For reading data
import pandas as pd
from sklearn.linear_model import LinearRegression

# load data set
data = pd.read_csv('tmpCyclistData.csv')
# values converts it into a numpy array, I change this value based on which factor I am co
X = data.iloc[:, 2].values.reshape(-1, 1)
# y axis is previous race performance
Y = data.iloc[:, 5].values.reshape(-1, 1)
linear_regressor = LinearRegression()
# do linear regression
linear_regressor.fit(X, Y)
# make predictions
Y_pred = linear_regressor.predict(X)

plt.scatter(X, Y)
# this will vary based on which factor I am analysing
plt.xlabel("Weight (Kilograms)")
plt.ylabel("Previous Race Performance (position)")
plt.plot(X, Y_pred, color='red')
plt.show()
```

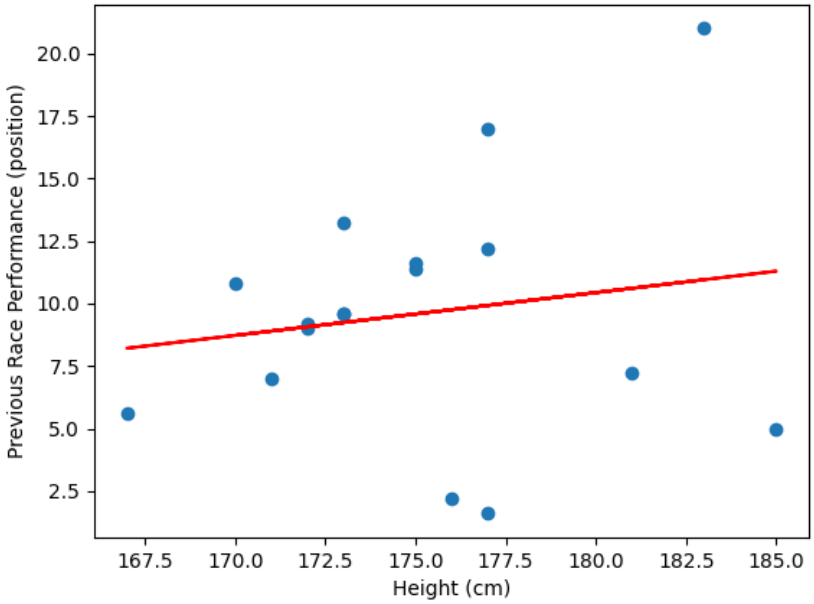
Below are the plots of this Linear Regression comparing these factors:



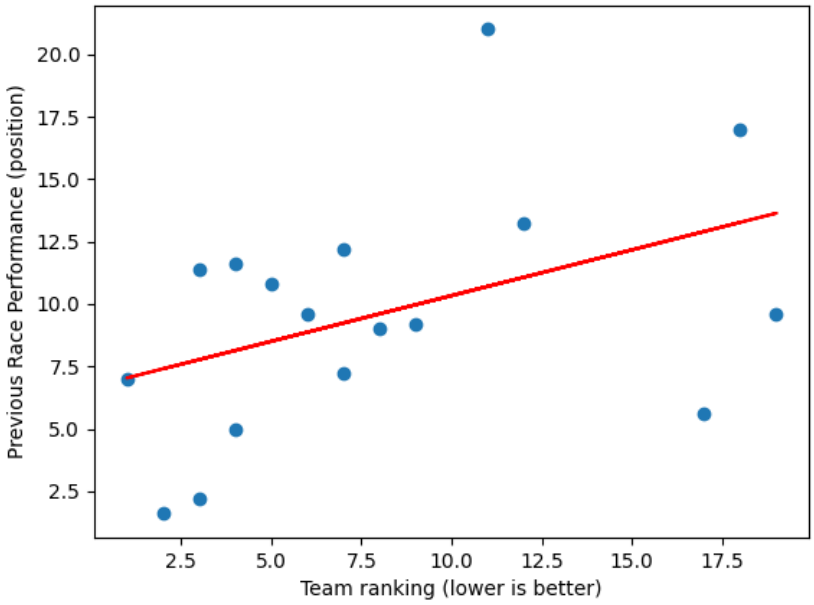
In this data there is a noticeable trend where heavier cyclists perform better. However, I think this is largely related to the two highest performing cyclists, Roglic and Pogacar being 65 and 66 kilograms respectively, causing this relationship, as opposed to actual performance gains from being heavier.



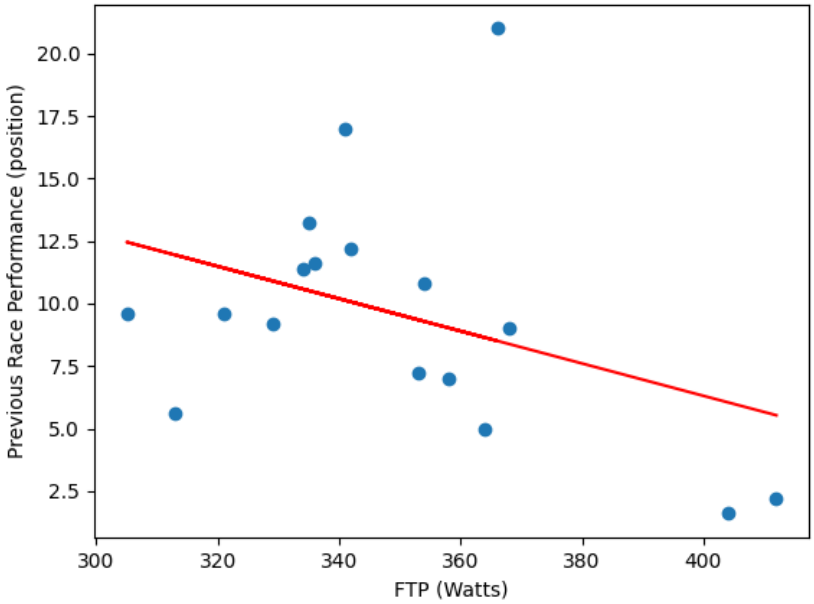
In this data there isn't much of relationship between age and performance.



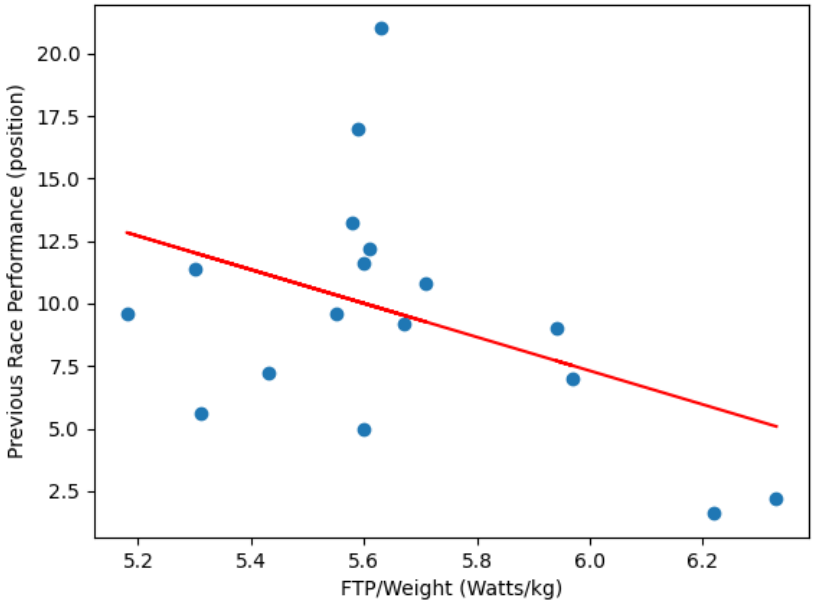
There is a noticeable relationship between height and performance for this dataset, but I think the relationship isn't particularly clear and the dataset is too small to make a reasonable conclusion.



However, there is a big relationship between the team ranking and race performance.



There is also a very clear relationship between FTP and performance.



Similar to FTP, the FTP/Weight factor plays a big role in cycling success as can be seen by the relationship in this graph.

4 stochastic simulations for cycling events

4.1 Overview of the model

For my model, I have produced a Binary Logistic Regression model with 2 outcomes, one where the cyclist is expected to win GC in the next multi stage race they enter, or they are not expected to win GC in that same race. I achieved this by using basic machine learning in Python. For the dependant factors, I have used FTP, FTP/Weight, team ranking and performance over the last 5 races. I trained this model using the results of these cyclists over their past 5 multi stage races, considering whether they won or not.

Here is the Python code I used to produce this result:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import seaborn as sn
import matplotlib.pyplot as plt

df = pd.read_csv("trainingData.csv")
#trained based on these factors
X = df[['Team Ranking','Performance', 'FTP','FTP/Weight']]
#compared to whether or not they won
y = df['Won']
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=.1,random_state=0)
logistic_regression= LogisticRegression()
logistic_regression.fit(X_train,y_train)
dfTest = pd.read_csv("testData.csv")
#produce the probability of winning and losing for each cyclist from the test data set
newY_pred=logistic_regression.predict_proba(dfTest)
print(newY_pred)
```

4.2 Results of the model

Running my model on this dataset, I got these results:

Cyclist name	Odds to win next GC race
Primoz Roglic	0.50902474
Tadej Pogacar	0.46467739
Richie Porte	0.16873041
Remco Evenepoel	0.34813152
Wilco Kelderman	0.32172863
Guillame Martin	0.08810448
Jakob Fuglsang	0.20309376
Diego Ulissi	0.14624161
Nairo Quintana	0.15800888
Simon Yates	0.16998518
Mikel Landa	0.0701288
Rafal Majka	0.16629388
Richard Carapaz	0.15581682
Jai Hindley	0.15649953
Enric mas	0.02334183
Tao Geoghegan Hart	0.01913797
Ajejandro Valverde	0.113529

Clearly, there are some flaws in this model. You would obviously not expect for there to be more than a 100% chance for these people to win the events, but you get that amount across jut the top three cyclists

from this. However,given the dataset that the model was trained on, these results make sense. For example, in the training dataset I supplied, Primoz won 3 of the 5 events he entered, and that would make since if he is 51% to win any individual race. However, by having chosen the top 17 cyclists as the ones to train this model on, it will lead to significantly higher expected performance for the data across the board.

In terms of ways to improve this, having a measure beyond just winning could be an improvement. In the current model, winning the Tour de France is the same as winning a local or junior multi-stage race, even though the performances are clearly not the same. Additionally, just having much, much more training data and a lot more non winning training data would certainly make it more accurate.

5 Constraints and Further improvements

5.1 Lack of good data set availability

The data , particularly with regards to heart rate and power for the top tier cyclists, was generally inconsistent or unavailable, so I had to use approximate power numbers and avoided using heart rate altogether in my model. If I had access to heart rate or true power data, my model would have both been more robust and likely more accurate.

As mentioned in the the referenced article on the FTP of pro cyclists, "FTPs are often closely guarded secrets for top riders". The article also gives approximate FTPs for many pro cyclists, which are roughly in line with the approximate FTPs I produced. However, by using approximate FTP I have introduced another place where there is going to be errors in my model. As for the accuracy of my FTP estimate, the only recent estimate for FTP for any of these cyclists I could find was on the referenced cyclingtips.com article, which suggested that Tadej Pogacar's FTP is 403 watts as opposed to the 412 my approximation produced. However, Pogacar's performance in the Time Trial portion of this Tour suggests that his FTP was likely higher than this article would suggest, given that he finished over a minute faster than the rest of the field. So it is reasonable to believe that my FTP approximations are at least somewhat accurate.

Additionally, with the data being somewhat hard to gather and compile, I only used a very small group of the top cyclists in the data which I analyzed, which overall causes some errors in my analysis, particularly in the linear regression portion. For example, there is a correlation between being heavier and cycling success in my linear regression analysis, but I think it has more to do with the top two cyclists in this method weighting heavier than much of the rest of the cyclists, which causes this relationship, rather than an actual relation between being heavy and succeeding in cycling. In fact, you would think the opposite is true. Also in this measurement, being "heavy" is 65-66 kilograms, which is very light relative to the general population.

5.2 Some factors of cycling are hard to measure

Many factors of cycling are hard to measure from a mathematical perspective, such as racing strategy, racing skill, aerodynamics, or ability of domestiques for a given team. Additionally, there is a lot of randomness with the human body for a given day, and injuries and crashes all are quite hard to model beyond just treating them as a random variable. Additionally, classifying the role of racers in a given race may be difficult. In a given race, the individual racers are not necessarily aiming for the same goal, and even individuals may have different goals for different races they are entering in, or even for different stages in the same race. I focused on GC competitors, but it isn't even always clear if the cyclists are focusing on winning the race or have other goals in mind.

5.3 Cycling races are very complicated

Relative to how simple this model is, there are many different aspects of competitive cycling that I don't fully take into account. Things like team dynamics and injuries all don't get taken into account by this model. Really only using 4 factors (2 of which are essentially the same) in order to predict something this complicated is much, much to simple. I suspect that if this model were to be fully developed, there would be at least 10 to 20 factors that are being used in the regression. However, this reintroduces the errors I mentioned earlier with lack of clear and available data, so I'm not sure how to model and fix this in that regard.

5.4 Many oversimplifications were made

In building this model and doing the logistic regression, I made many simplifications, particularly with my data. One thing I want to mention about the way I applied the approximate FTP formula is that since I only used one example time trial. For example take Rafal Majka. He got 68th place in the last time trial he did (a stage in the Giro d'Italia), so his FTP estimate is relatively low compared to what it could be if I chose another time trial that he competed in.

Another part of my model that is flawed is that the way I classify people is too dependant on recent consistency. For example, take Tao Geoghegan Hart who is comparatively ranked quite poorly by my model. He won the Giro d'Italia, one of the major Grand Tours, but isn't rated well by my model due to the way that I measure performance. That is, the model says he has less than 2% chance because he did average in most of the other recent he competed in except for the Giro. Also, Egan Bernal is considered fourth most likely to win the Tour de France according to the Bovada sports book betting lines which I have referenced, but by the way I have gathered data he isn't even considered because of his poor performance this season. Thus, there are few ways that my model could be improved in terms of data and performance classification.

5.5 The dataset for the simulations portion was particularly bad

As I mentioned earlier, the dataset that I used for training my Logistic Regression model was way too small. Generally with these types of machine learning models you would want much larger and more diverse datasets. I would very much be interested in the performance of this model if I were to use a larger dataset that is more representative of the general field in these races. Also, I am interested in trying running this model or a similar model on a more specific subset of races to see how it would work on the more competitive races, such as the Grand Tours or the World Championships.

6 Conclusion

As an overall topic of study, this was a very interesting project and I think that there is significant room for improvement in my model, and there is a lot of interesting ways that we can use the data in cycling in mathematical models.

However, I think that my model is overly simplistic for how truly complex these events are from a mathematical perspective, as I had to make far too many assumptions and simplifications for the model to be truly accurate. Frankly, I think that even if one of the more well funded cycling teams hired several full time analysts for building a similar model in order to influence their training, strategy and hiring, they would also struggle and have to make some concessions in their model (but certainly would be able to do better than me).

I also want to bring up horse betting example I mentioned in my introduction and how it compares to my model. As I mentioned, there are many interesting parallels between cycling and horse racing and ways that we can model them. However, I think that although there are parallels, cycling overall is much more complex and harder to model. The team dynamics, different stages, and different overall races make it quite hard to model in a fashion similar to the way that Benter did for horse racing. Also, Benter's model was significantly more complicated than mine even though he was using a otherwise less complicated sport to model. Such as comparing his model's odds to existing betting odds and using more factors in the multinomial logistic regression model, his model was more advanced and more suited for the sport that was being analysed.

Also, unlike in horse betting the only cycling event I could find betting odds for was the Tour de France, so even if we were to build a successful model that performed better than the betting odds, we would never be able to nearly as successful as Benter was in his betting performance. I've always thought the Benter story was interesting, but I'm not sure that we will ever have such a confluence of factors that allow for such a successful model to be applied to the field of sports betting again, and almost certainly not to cycling in the way that it currently exists.

References

[1] Edward A. Bender. *An Introduction to Mathematical Modeling (Dover Books on Computer Science)*. General reference. 1977.

[2] William Benter. *Computer Based Horse Race Handicappingand Wagering Systems: A Report*. URL: https://www.gwern.net/docs/statistics/decision/1994-benter.pdf.

[3] Stephan van der Zwan Bert Lip. *procyclingstats.com*. 2013. URL: procyclingstats.com (visited on 11/30/2020).

[4] Giancarlo Bianchi. *TDF Power Analysis: Pogacar's Peyresourde attacks and new climbing record*. Sept. 6, 2020. URL: https://cyclingtips.com/2020/09/tdf-power-analysis-pogacars-peyresourde-attacks-and-new-climbing-record/.

[5] T.N. Crouch et al. *Riding against the wind: a review of competition cycling aerodynamics*. URL: https://doi.org/10.1007/s12283-017-0234-1.

[6] John E. Cobb James Martin Douglas Milliken and et al. *Validation of a Mathematical Model for Road Cycling Power*. 1998. URL: https://www.researchgate.net/publication/279937184_Validation_of_a_Mathematical_Model_for_Road_Cycling_Power (visited on 12/05/2020).

[7] Mathew Mitchell. *Pro Cyclist FTPs*. https://www.procyclinguk.com/pro-cyclist-ftp/. [Online; accessed 21-October-2020]. 2020.

[8] David Percy. *Strategy selection and outcome prediction in sport using dynamic learning for stochastic processes*. 2015. URL: https://link.springer.com/article/10.1057/jors.2014.137.

[9] Bovada Sportsbook. *Cycling odds, lines, spread and props — Bovada Sportsbook*. 2011. URL: https://www.bovada.lv/sports/cycling (visited on 12/06/2020).

Glossary

domestique In road racing, a domestique is a rider who rides solely for the benefit of their team or other team members, rather than trying to win the race themselves. 9

FTP FTP stands for functional threshold power, and it is essentially a measure of the best average power a cyclist could output for 1 hour in a time-trial setting. A common sentiment in cycling is that a person's FTP with respect to their weight (or watts per kilo) is the most important thing for being a faster cyclist . 2, 9

general classification General Classification, or GC, in a multi stage bicycle race is the racer with the overall fasted time across all the stages of the race. There is some additional complexity with bonuses for high performance in individual stages or time punishments if the rules are broken, but the vast majority of time is based on cumulative time across all stages . 1

Grand Tour In road bicycle racing, a Grand Tour is one of the three major European professional cycling stage races: Giro d'Italia, Tour de France and Vuelta a España. 10

Time Trial Individual Time Trials, or ITTs, are events in which the cyclists race individually and are racing to try to get the fasted time. 3

UCI The UCI, or Union Cycliste Internationale is the world governing body for sports cycling and oversees international competitive cycling events. 3