Cycling predictions using Machine Learning

Clément ANNE

08/03/2022

1. Introduction

With the increasing number of sports statistics publicly available, machine learning has become a popular tool to develop algorithms aiming at predictions outcomes of sports events. While some sports such as basketball, baseball, or soccer has seen a development of advanced statistics and prediction models over time, such developments have not reached road cycling to the same extend.

In fact, cycling races seem more challenging to predict since the prediction output should rank all starting riders instead of predicting a binary or limited-class output such as win-loss-draw.

To fill this void, I relied on web scrapping to extract publicly available cycling data from the https://www.procyclingstats.com. My aim is to develop a machine learning algorithm to predict outcomes from one-day World Tour races over 2014-2021.

World Tour races are the top races in the world cycling calendar, which are expected to bring the highest level of competition. As a first machine learning algorithm developed to cycling race data, I chose to restrict the focus to one-day races to limit the noise in stage results from multi-stage races arising from general classification strategies (e.g., teams not chasing back breakaways having no fearful riders for the general classification in latter stages).

The resulting sample consists of 130 one-day World Tour races over 2014-2021.

One major contribution of this analysis relied in the development of an algorithm cleaning each race results data by grouping riders finishing within similar time gaps from the winner, into categories (fought for the win, fought for the podium, fought for the top 10, was active in the final, fought to finish, did not finish). This approach aims at bringing predictions more relevant than simple rank predictions.

As a first step in cycling predictions, this analysis focuses on predicting riders fighting for the top 10 in those 130 one-day World Tour cycling races over 2014-2021.

Following section will present the methodology applied for this analysis.

2. Methodology

2.1. Data source

This analysis relies exclusively on data scrapped from the https://www.procyclingstats.com website. Extracted data contain:

-**Race results** for the 130 one-day World Tour races over 2014-2021 including the time gap from the winner, the BIB number, and the team.

-**Race level information** (distance, difficulty rating (Profile Score from the https://www.procyclingstats.com website), vertical meters climbed, date in the annual calendar, number of riders).

-**Rider level information** (age, experience, height, weight).

-**Extensive results from each rider's career** at the time of each race start. I limited data extraction to all races prior to the last one-day World Tour race in our dataset.

2.2. Cleaning race results

2.2.1. Grouping riders in race results

One innovation of this analysis relies in the identification of groups of riders inside race results to derive outcomes.

If a time gap below 10 seconds remains between 2 subsequent groups of riders, we group them and assign the time gap of the biggest group. We apply these corrections as long as a gap below 10 seconds remain between subsequent groups. This enables grouping riders which finished close from a main group for various reasons (e.g., A rider not wanting to sprint after helping a teammate).

2.2.2. Identifying race outcomes

After those corrections, we derive some binary outcomes catching the result for those groups:

-**Fought for the win**: In case the winner was in this group

-**Fought for the podium**: In case the better classified rider in this group finished on the podium

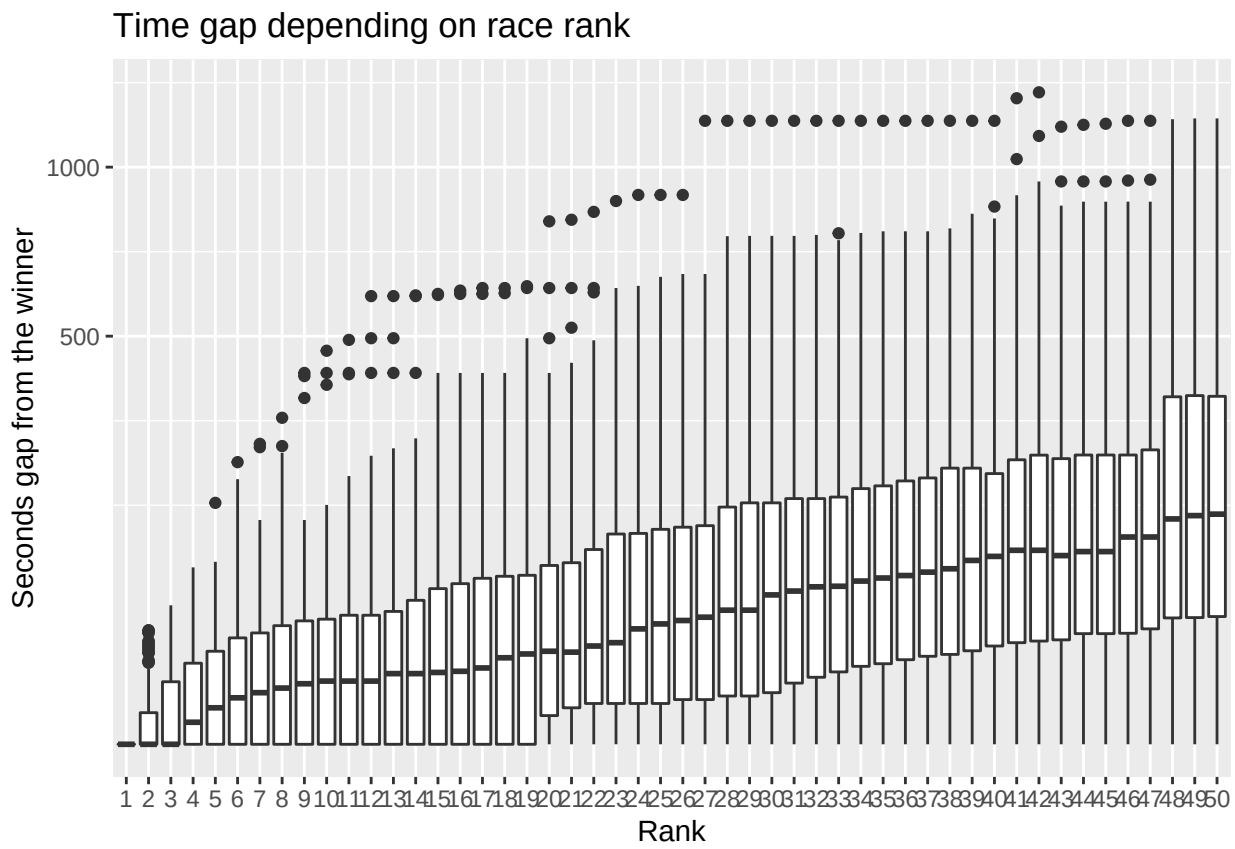
-**Fought for the top 10**: In case the better classified rider in this group finished in the top 10. I chose the top 10 cut-off since it is the standing appearing on TV at the end of the race, thus proving extra motivation to reach the top 10.

-**Was active in the final**: In case the better classified rider in this group finished in the top 30. I chose the top 30 cut-off since the number of UCI (alternatively PCS) points earned becomes marginal at this point. Besides, it should catch riders performing well but not enough to fight for accessits or teammates having significantly helped their leader(s).

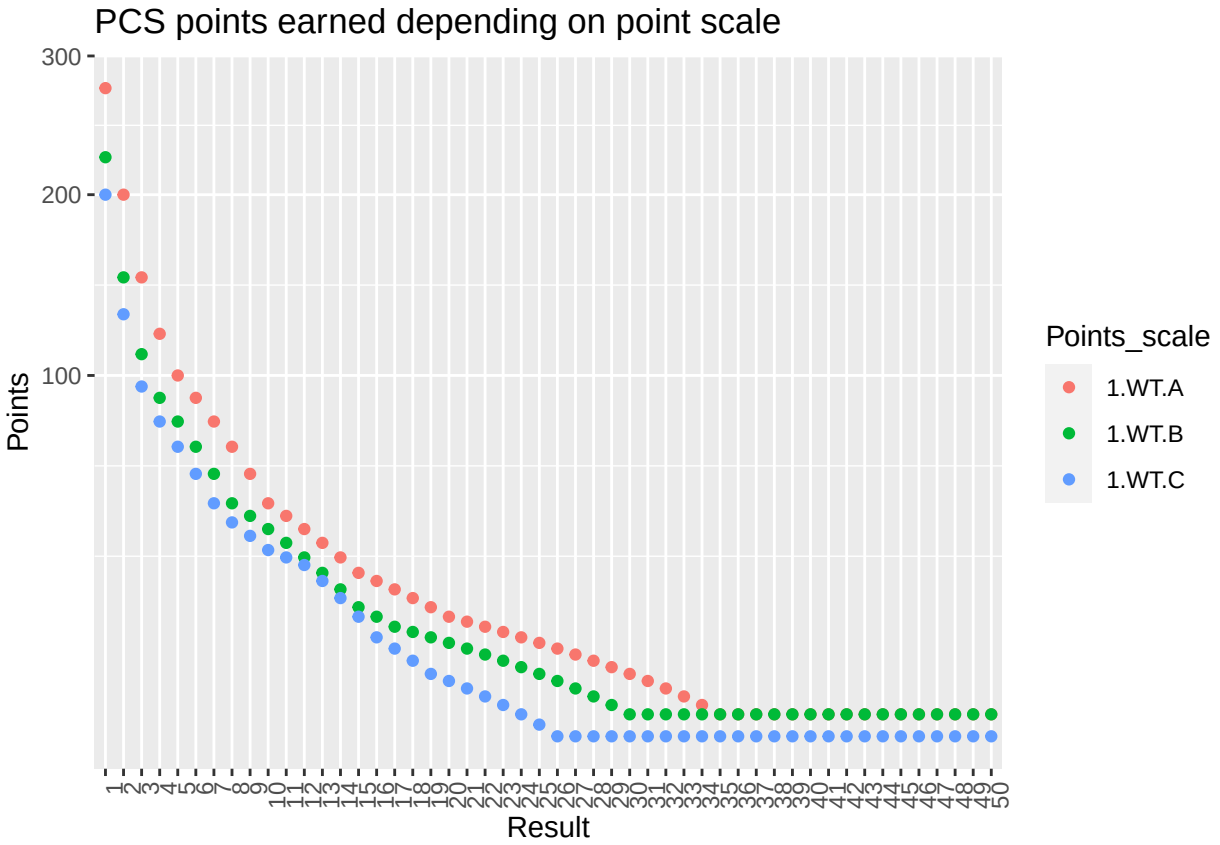
-**Fought to finish the race**: In case the better classified rider in this group finished beyond the top 30. It aims at controlling for riders with a will to finish the race despite having nothing else to fight for. Unlike general classification races, one-day races are all of nothing events so that some riders may be more willing than others to finish it in case they could not play a significant role in the final portion of the race.

-**Did not finished**: In case the rider did not finish the race.

The figure below provides the gap distribution from the winner depending on the final rank. We can evidence that the median values increase by steps which correspond to some of the groups we defined.



The analysis of PCS point depending on point scale highlights that riders beyond the 30th place only gain marginal points. However, focusing on the top 10 seems a more conservative approach for a first analysis since some riders may have different incentives of fighting for the 10-30 ranks.



2.2.3. Example of cleaned race outcome

To illustrate those corrections, table below represents the corrected standing among the top 50 riders from il Giro di Lombardia 2016.

Rnk	Rider	Gap_secs	Gap_secs_fixed	fight_win	fight_podium	fight_top10	fight_active
1	CHAVES Esteban	0	0	TRUE	TRUE	TRUE	TRUE
2	ROSA Diego	0	0	TRUE	TRUE	TRUE	TRUE
3	URÁN Rigoberto	0	0	TRUE	TRUE	TRUE	TRUE
4	BARDET Romain	6	0	TRUE	TRUE	TRUE	TRUE
5	VILLELLA Davide	79	84	FALSE	FALSE	TRUE	TRUE
6	VALVERDE Alejandro	84	84	FALSE	FALSE	TRUE	TRUE
7	GESINK Robert	84	84	FALSE	FALSE	TRUE	TRUE
8	BARGUIL Warren	84	84	FALSE	FALSE	TRUE	TRUE
9	DE MARCHI Alessandro	84	84	FALSE	FALSE	TRUE	TRUE
10	LATOUR Pierre	84	84	FALSE	FALSE	TRUE	TRUE
11	ARU Fabio	86	84	FALSE	FALSE	TRUE	TRUE
12	BRAMBILLA Gianluca	124	124	FALSE	FALSE	FALSE	TRUE
13	TORRES Rodolfo Andrés	125	124	FALSE	FALSE	FALSE	TRUE
14	VISCONTI Giovanni	162	162	FALSE	FALSE	FALSE	TRUE

Rnk	Rider	Gap_secs	Gap_secs_fixed	fight_win	fight_podium	fight_top10	fight_active
15	COSTA Rui	302	302	FALSE	FALSE	FALSE	TRUE
16	MAMYKIN Matvey	302	302	FALSE	FALSE	FALSE	TRUE
17	MOLARD Rudy	302	302	FALSE	FALSE	FALSE	TRUE
18	ATAPUMA Darwin	333	333	FALSE	FALSE	FALSE	TRUE
19	MOLLEMA Bauke	384	384	FALSE	FALSE	FALSE	TRUE
20	FUGLSANG Jakob	396	396	FALSE	FALSE	FALSE	TRUE
21	BAKELANTS Jan	437	437	FALSE	FALSE	FALSE	TRUE
22	ULISSI Diego	490	490	FALSE	FALSE	FALSE	TRUE
23	REICHENBACH Sébastien	490	490	FALSE	FALSE	FALSE	TRUE
24	ELISSONDE Kenny	490	490	FALSE	FALSE	FALSE	TRUE
25	TALIANI Alessio	490	490	FALSE	FALSE	FALSE	TRUE
26	KENNAUGH Peter	490	490	FALSE	FALSE	FALSE	TRUE
27	ANTÓN Igor	490	490	FALSE	FALSE	FALSE	TRUE
28	MORENO Javier	490	490	FALSE	FALSE	FALSE	TRUE
29	SCHLECK Fränk	490	490	FALSE	FALSE	FALSE	TRUE
30	MORENO Daniel	490	490	FALSE	FALSE	FALSE	TRUE
31	WOODS Michael	490	490	FALSE	FALSE	FALSE	TRUE
32	ROSSETTO Stéphane	630	630	FALSE	FALSE	FALSE	FALSE
33	CUNEGO Damiano	630	630	FALSE	FALSE	FALSE	FALSE
34	GILBERT Philippe	630	630	FALSE	FALSE	FALSE	FALSE
35	GOGL Michael	630	630	FALSE	FALSE	FALSE	FALSE

2.3. Assessing rider ability using previous race outputs

Race outcomes may depend significantly on riders abilities. While it is almost impossible to gauge riders abilities, this analysis focuses on past results to derive ability (or strength) related variables.

Those variables are divided into 3 categories: -**Recent form**: They should catch the form of the rider over the most recent races.

-**Status in the peloton**: Those variables aim at capturing the status of a given rider because of his recent results.

-**Career achievements**: They sum up the career stats of the rider.

The 2 latter categories (status and career) are further split into 4 categories:

-**Stats related to the oncoming race**: Statistics related to former participations of the oncoming race (e.g., Paris-Roubaix)

-**Stats related to past 1 day World Tour races**: Statistics related to former participations of 1 day World Tour races. It should capture experience and ability specific to those races. I limited those races to former edition of each of the 21 races appearing at least once in the dataset.

-**Overall stats**: Statistics related to all races

We rely on 2 information of past race achievements for all the abovementioned categories:

-**The number of raced days** (e.g., a multistage race would account for 1 raced day for each stage)

-**The sum of ProCyclingStats points (PCS)** earned

ProCyclingStats points (PCS hereafter) are an alternative to World Tour and/or UCI points earned. In fact World Tour point system changed over time, so that PCS provides a better alternative consistent across time to evaluate riders performance.

It should be noted that PCS points are earned for any race achievements, with more points being earned for better results and on more prestigious races.

The 14 resulting variables are described in the following subsections.

2.3.1. Recent form and the form window

Variables related to the recent form should catch the rider's form when coming to a given race.

In fact, riders are not at their peak level all year long and may target specific objectives in the calendar.

Thus, a rider may perform better when he is coming from more races or at least with better results.

To capture those effect, we compute the variables described previously (number of raced days and sum of PCS points) on a form window.

Form window (fw) is a tuning parameter and we will test values from 2 to 6 weeks.

2.3.2. Status in the peloton and the status window

We aim at capturing the status in the peloton for a given status window sw through the following variables:

-Achievements over 1-day World Tour races during status window sw (number of raced days and sum of PCS points) -Achievements on this oncoming race during status window sw (number of raced days and sum of PCS points) -Achievements over all races during status window sw (number of raced days and sum of PCS points)

Status window is also a tuning parameter and we will test values from 3 to 5 years.

2.3.3. Career achievements

We aim at capturing career achievements through the following variables:

-Career achievements over 1-day World Tour races (number of raced days and sum of PCS points) -Career achievements on this oncoming race (number of raced days and sum of PCS points) -Career achievements over all races (number of raced days and sum of PCS points)

It seems important to distinguish status in the peloton from career achievements as some riders past their prime years would have few top results over the recent years while having great results way back in time.

2.4. Deriving team strength

Cycling is a team sport for which it is not easy to capture team-level features.

Teams differ according to their finances, their material, their material, and their training, which may differ

It is difficult to track teams across times since they often change sponsors, which may bring staff or structure changes (e.g., In case of the arrival of a foreign sponsor).

Thus, I relied on a second-best approach to gauge team strength by summing up the rider strength variables at the team level among riders of the same team at the beginning of each race.

A previous version of this analysis included also the deviation of rider strength variables within his team at the beginning of the race to account for the potential overflow of talent in top teams, which would prevent some talented riders from contending freely for top ranks.

However, it would complicate the model further since teams with few to none experience (e.g., invited teams) may be tricky to extract a team status for each rider. As such, I postponed this inclusion for further developments.

2.5. Machine learning methodology

2.5.1. Data partition

The final dataset consists of 22539 rider-race level observation for the 130 races covered in this analysis. To train a machine learning algorithm I split the data according to their race ID, which randomly assigned 28 races (approximately 20% of races) to a test set and 102 races to the training set.

The performance of the algorithm will only be assessed on the 28 races in the test set after being trained exclusively over the 102 races in the training set.

It seems to mimic better the aim of this algorithm which is to predict race outcomes given observable features at the start of the race. As such, it seems preferable over splitting the whole dataset depending on the race outcome (fighting for the top 10 in our case).

2.5.2. Preprocessing data

The following transformations have been applied:

-**Rider level information** has been scaled among each race participants so that the variables are relative to the start list.

-**Race level variables** have been scaled across all the dataset to catch relative specificities across them.

-**Team strength variables** have been scaled across each race to represent the relative importance of teams inside the peloton for a given race.

2.5.3. Models

Model 1: Knn without matrix factorization

We consider both the status and form windows presented in section 2.3 as tuning parameters since they should catch the rider/team ability at the start of the race.

This analysis being the first in the direction of predicting race outcomes in cycling, it focuses on predicting riders fighting for the top 10.

For this task, I relied on kNN (k nearest neighbors) modelling.

This model applied on binary outcomes computes the average probability of the outcome among the observations with the closest features according to the variables used in the model.

Those closest observation points used depend on the paramt

Model 2: Knn with matrix factorization

This analysis tried to bring extensive data related to rider/team abilities through the indicators presented in the previous sections. One drawback of this approach is to use highly correlated variables.

In fact, we may expect some correlation between the number of races and the sum of PCS points. Likewise, we may observe significant correlation between team strength and rider abilities.

To limit this problem, I relied on matrix factorization through Principal Component Analysis (PCA hereafter) before applying the KNN model.

Principal component disentangle the matrix of predictors into uncorrelated factors of decreasing importance. It aims at preventing high correlation among predictors while also limiting the dimension of predictors.

We select the number of principal components up to representing 95% of the variation from the initial matrix of determinants.

2.5.4. Selection of tunning parameters

I used cross-validation on the training set to tune model parameters. To do so, I relied on 5-fold cross-validation by randomly splitting races from the training set into 5 groups.

For each of those 5 groups, I run estimates on each version of the training set after exclusion of this group before assessing on this excluded group. Then model performance indicators are averaged across those 5 model computations.

This process is repeated for each parameter and those parameters are selected when they maximize the model performance indicators.

Tuning parameters include:

-**Form window**: From 2 to 6 weeks

-**Status window**: From 3 to 5 years

-**Number of neighbors (k)**: For the kNN model.

2.5.5. Measuring model performance

This analysis aims at predicting whether a given rider at the start of a race would finish at different echelons.

In our case, having a low sensitivity would prevent us from predicting enough riders in the corresponding group(e.g., classifying a rider as fighting for the win while he actually did not).

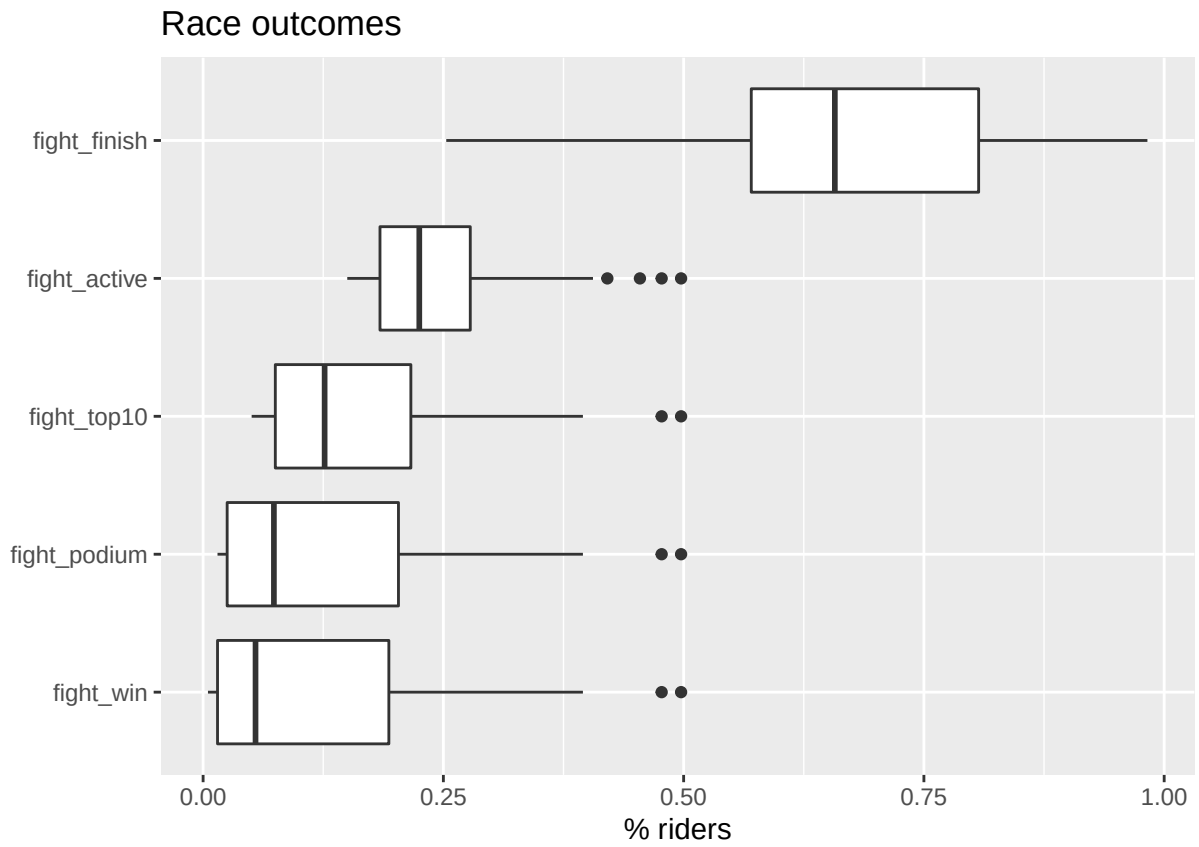
Likewise, having a low specificity would mean the algorithm fails to detect riders finishing in the corresponding group (e.g., classifying a rider as not fight for the win while he actually did).

It seems that both error types are equally costly, so we can rely on the F1-score to select the best tuning parameters for the algorithm trained on the training set. By weighting both sensitivity and specificity through the harmonic average, it should be a more balanced measure than accuracy.

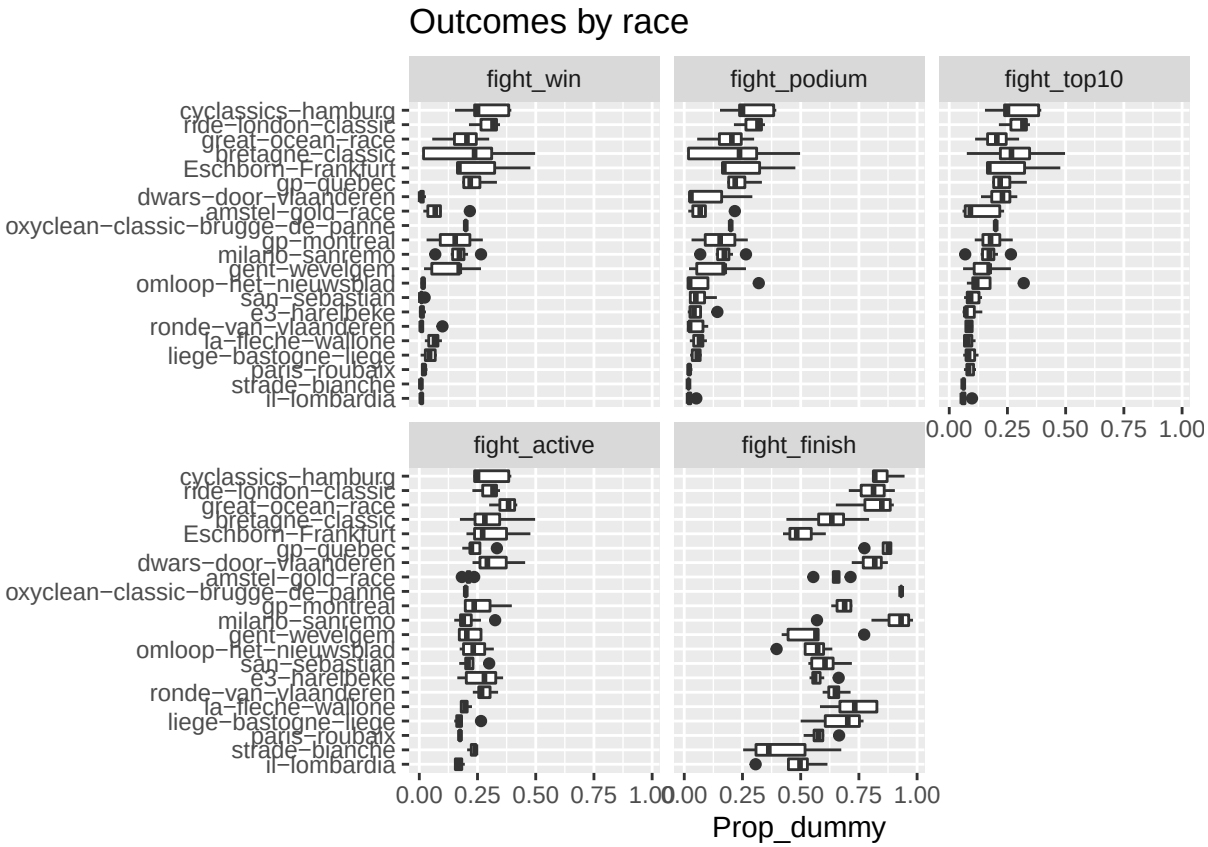
F1-score seems even more relevant that the probability of fighting for the top 10 has a relatively low prevalence so that accuracy could be a misleading indicator.

2.6. Data exploration

The following graph represents the distribution of race outcomes across races in the training set. We can evidence that fighting for the top 10 remains a low prevalence event with the interquartile range being below 25%.



The following graph evidence the same distribution across races in the training set.



One limitation of this current analysis is the mix of races which have different features bringing different scenarii. While some race level determinants have been added in the models (race profile, distance, vertical meters), we cannot rule out that some race-specific features would remain.

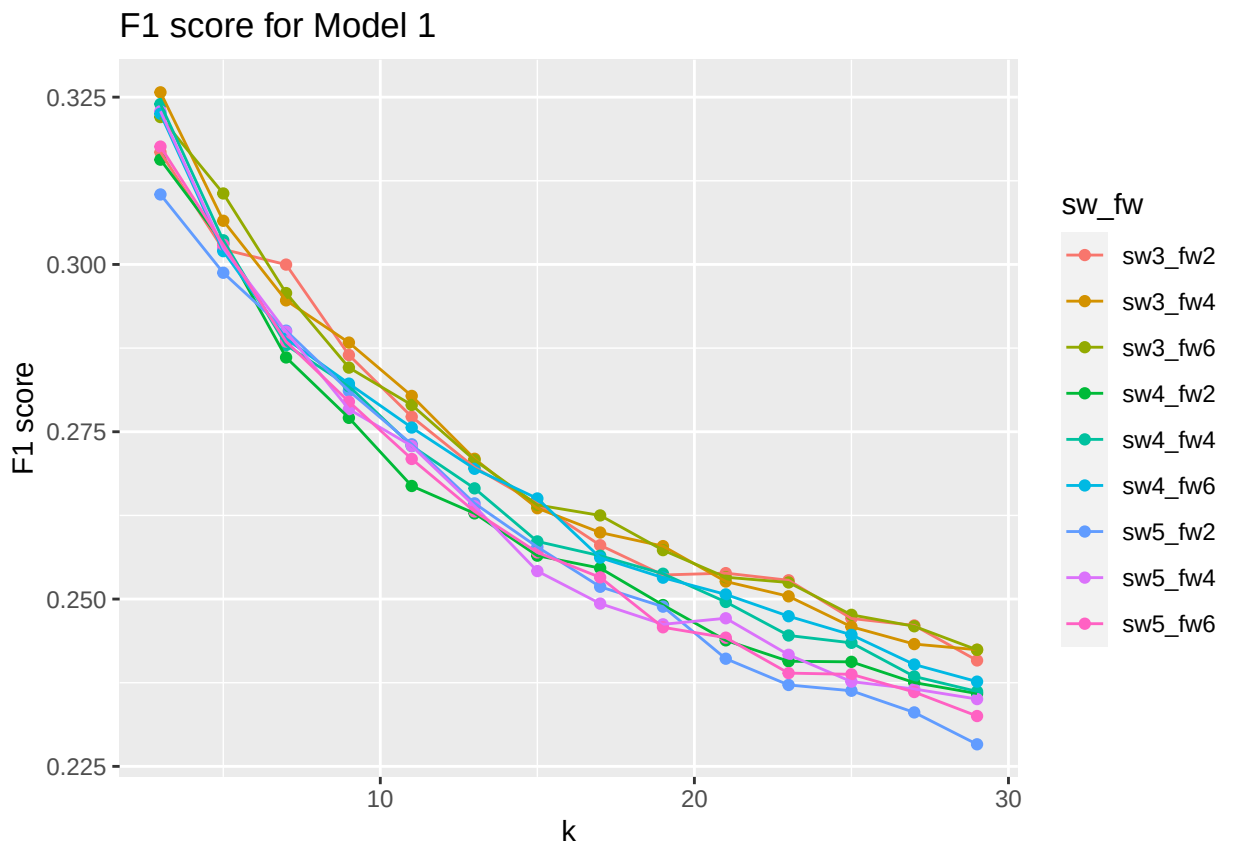
However, focusing on riders fighting for the top 10 seems less prone to this bias compared with riders fighting for the win, since the heterogeneity is slightly lower.

3. Results

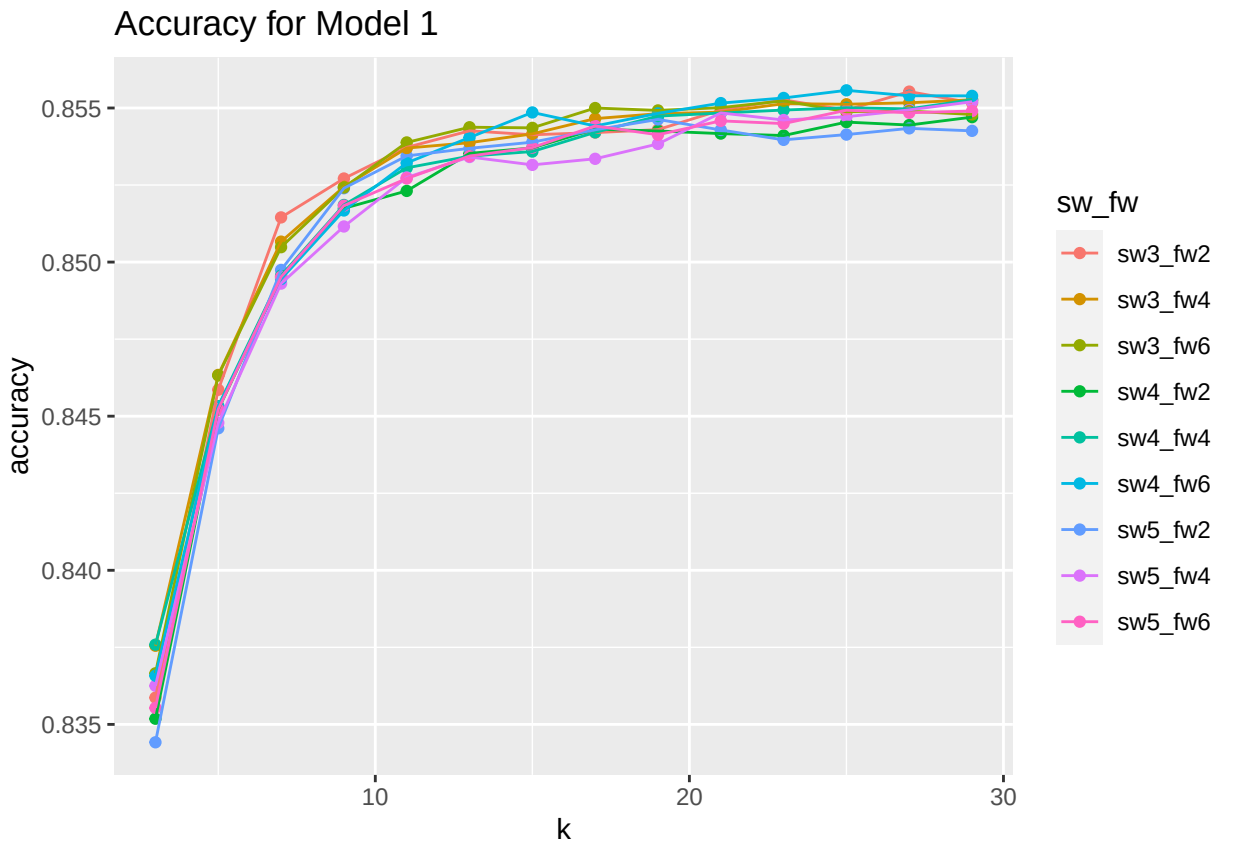
This section presents the results obtained for predicting riders fighting for the top 10 in World Tour 1-day races across 2014-2021.

Better predictions are obtained while using a narrower status window for assessing the status in the peloton, since the performance improves with a window of 3 years. Results depending on the form window are more blurred.

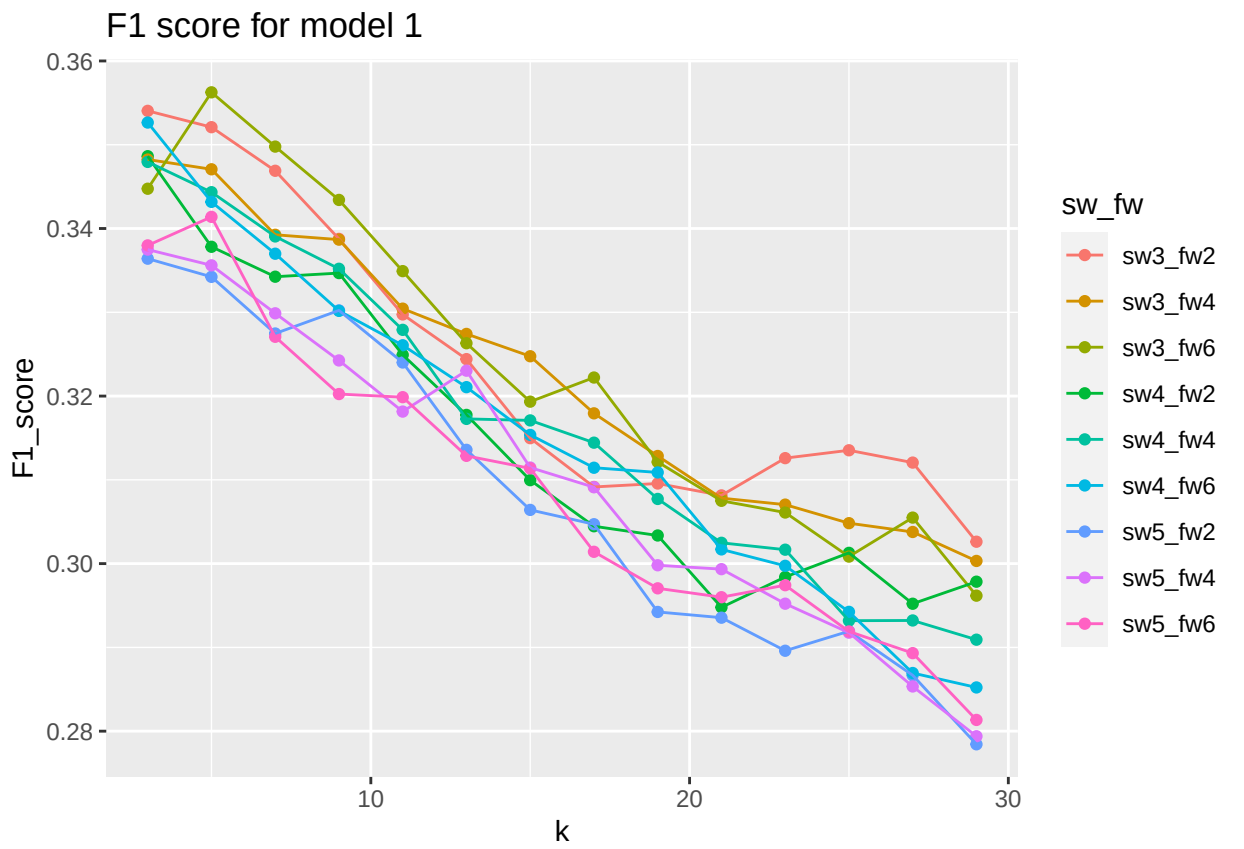
Results related to model bring a relatively low F1 score which is maximized on lower values on k. It may seem counter-intuitive but it seems better to rely on the model using low value of k since it helps.



This pattern could be observed while plotting the same graph for accuracy. Using accuracy we would have prefered a higher number of neighbors k, but it would have been at the cost of a lower sensitivity which motivates our reliance on the F1 score.



Our second model using matrix factorization brings better results with higher F1 scores. The better performing model is obtained for k=5 neighbors over a status window of 3 years and a form window of 6 weeks.



The following graph confirms the importance of relying on F1 score, since the accuracy would have been a misleading performance indicator.

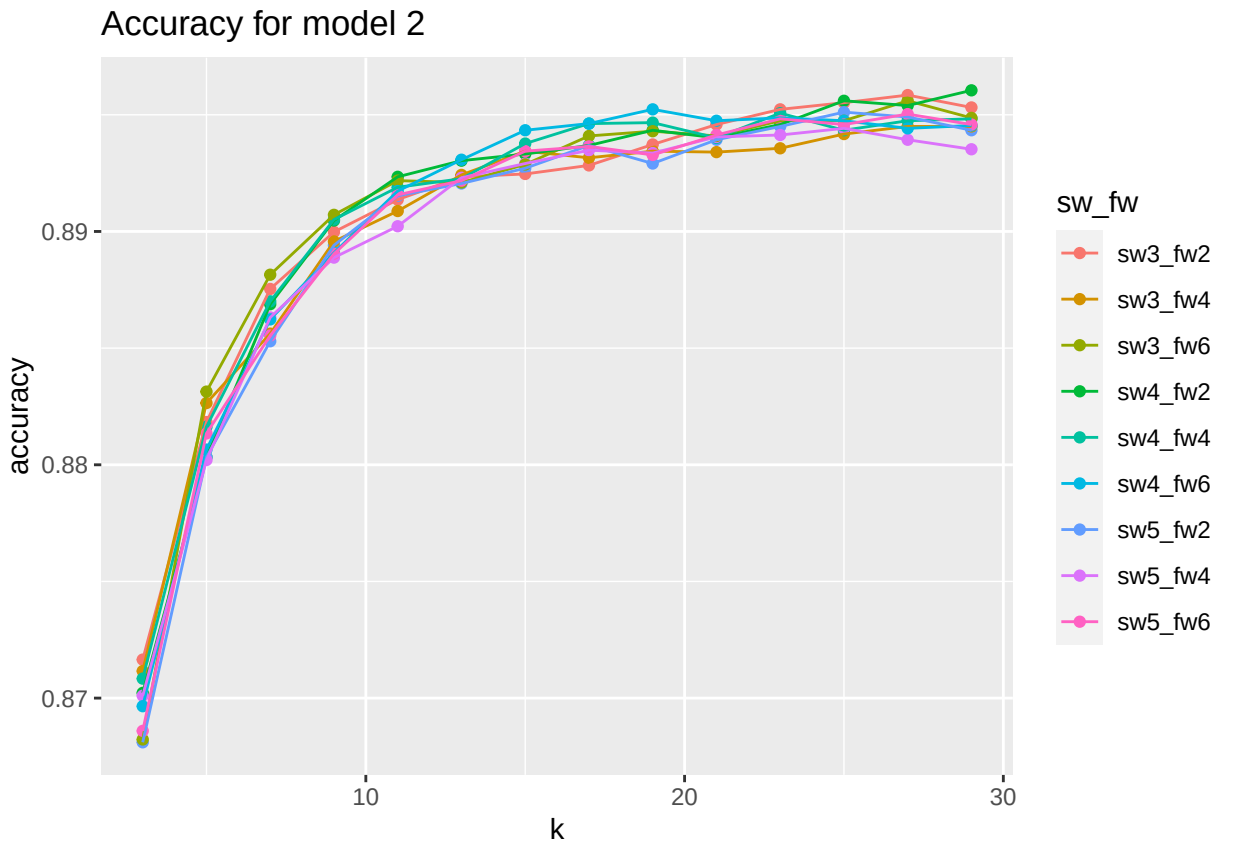


Table below represents the parameters and performance related to the best performing model during the training process on the training set.

Model	k	Form_window	Status_window	F1_score	accuracy	sensitivity	specificity
Knn with PCA	5	6	3	0.356	0.883	0.285	0.96

We use those parameters and apply them on the test set to gauge the real performance of our model. For comparison sake we use those same parameters on model 1 without matrix factorization even though we know it performs worse.

Model	k	Form_window	Status_window	F1_score	accuracy	sensitivity	specificity
Knn without PCA	5	6 weeks	3 years	0.393	0.890	0.313	0.964
Knn with PCA	5	6 weeks	3 years	0.407	0.889	0.335	0.960

The better performing model featuring KNN with matrix factorization provides a F1-score of 0.407. The sensitivity remains a bit low at 0.335 but it improves significantly from the model without matrix factorization.

4. Conclusion

This analysis remains a first step in the development of a machine learning algorithm to predict road cycling race outputs. It provides an innovative way to classify riders' results into group outputs which better represents the race scenario than final ranks.

Using a KNN model with matrix factorization thanks to Principal Component Analysis brings promising results, since it predicts riders fighting for the top 10 with a F1 score of 0.4.

I should acknowledge that the ability indices used for this analysis could be improved using kernels or smoothers to give a higher weight to more recent performances inside the form or status windows. However, the solution may not be straightforward since riders do not race at every point in time, or may even not race at all in case of their season start.

Besides, the reliance on PCS points may blur the real performances which could have been more salient from grouping riders for each race as we did (e.g., overestimating a rider's ability which is steadily in the top 10-20 but will struggle having better results). A more time-consuming alternative bringing the race result correction to all past results may help gauge better rider performance.

Further work seems needed with modeling with other approaches such as the random forest model.

Complementary analyses would focus on predicting other races outcomes (e.g., fighting for the win) as a follow-up to this analysis.

A future development (after developing models for the prediction of each race outcome) would be to develop estimates of the relative sprinting abilities among riders predicted in each group. One way to do this could be to rely on former head-to-head matchups and the group size to predict an order among the groups of riders fighting for at least a top 10. As currently exploited for this analysis, abilities proxied by PCS scores may also overestimate sprinter ability as they tend to have higher PCS scores.