Evaluating the Impact of Equipment Investments on Olympic Medal Probabilities for Australian Professional Cyclists

by

Katherine E. Kostecki

S.B. Computer Science, Economics, and Data Science and Business Analytics, MIT, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2025

Authored by: Katherine E. Kostecki
Department of Electrical Engineering and Computer Science
May 9, 2025

Certified by: Anette Hosoi
Professor of Mechanical Engineering, Thesis Supervisor

Accepted by: Katrina LaCurts
Chair
Master of Engineering Thesis Committee

Evaluating the Impact of Equipment Investments on Olympic Medal Probabilities for Australian Professional Cyclists

by

Katherine E. Kostecki

Submitted to the Department of Electrical Engineering and Computer Science
on May 9, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE, ECONOMICS, AND DATA SCIENCE

ABSTRACT

AusCycling is the National Sporting Organization for cycling in Australia. Their oversight includes all aspects of the sport, including the high-performance program. As AusCycling begins preparations for the 2028 LA and 2032 Brisbane Olympics, they look to invest in new cycling equipment to boost their expected medal counts. This research takes a three-phased approach at selecting an efficient equipment investment portfolio for AusCycling that results in high medal probability. Using machine learning techniques, we first build a CATBoost prediction algorithm that classifies future Olympic performance into the categories of "earn a medal", "close to earning a medal", and "not close to earning a medal". The model predicts Olympic performance from performance at competitions prior to an Olympic Games with an overall accuracy of 96.3%. In the second phase of the research, a methodology is built to compute the percentage in race time an athlete needs to improve between two World Championship competitions in order to meet a probability threshold for earning a medal at the next Olympics. The final phase of this research combines the models and methodology of the first two sections by creating a Mixed-Integer Nonlinear optimization model which selects optimal equipment investments to maximize predicted Olympic performance while minimizing cost. When used on synthetic data similar to that available to AusCycling, the optimization selects an investment portfolio that yields an expected number of medals of 2.15 across men's and women's Team Pursuit and men's Team Sprint for the 2028 Olympics. This methodology may help AusCycling determine which equipment investments to make ahead of future Olympic competitions.

Thesis supervisor: Anette Hosoi
Title: Professor of Mechanical Engineering

Acknowledgments

This research was completed with the unwavering support and guidance of many remarkable individuals who provided invaluable mentorship, encouragement, and insight throughout this journey.

To begin with, I would like to sincerely thank my advisor, Prof. Anette Hosoi, for her unparalleled guidance. The optimism, passion, and insights she brought to each research meeting were a guiding example of how to conduct research while also serving as an inspiring leader and mentor. I am continuously amazed by the breadth of her multidisciplinary knowledge and her immense talent for tackling any obstacle with a clear vision. Her steadfast support of each lab member, including myself, is instrumental to the development of the profound work that comes from the Sports Lab. It has been a privilege to learn from Prof. Hosoi and witness the ways in which she thinks through problems, encourages her students, brings enthusiasm to each new idea, and fosters a supportive, inquisitive, and motivated lab environment.

I feel incredibly fortunate to have had the opportunity to conduct this research with the Sports Lab. I would like to thank all of the students researching with the lab. Their willingness to explore ideas deeply and discuss even the smallest details was not only helpful but consistently inspiring throughout the research process.

Next, I would like to thank Prof. Christina Chase for her extraordinary mentorship and support of my work with the Sports Lab. Her belief in me made a profound impact on this experience and deeply shaped the quality of this research.

This research would not have been possible without the support of AusCycling. It has been a gift to have the opportunity to work closely with an organization as goal-oriented and analytically-driven as AusCycling is. I would particularly like to extend my gratitude to the Executive General Manager, Jesse Korf. Jesse has aided significantly in the shaping of this research from his extraordinary vision for the future of cycling in Australia to his genuine belief in the role of research in achieving many of the organization's most ambitious goals. Beyond the research itself, Jesse has been an outstanding mentor to me throughout this experience. His insights into leadership, organizational culture, and bridging the gap between goals and actions will no doubt leave a lasting impression on me. Our conversations have significantly shaped my development academically, professionally, and personally.

Finally, I would like to thank my family and friends for their unwavering support throughout this journey. Thank you for grounding me through the challenges of the research process, for encouraging my academic pursuits, and for empowering my passion in this field. I am deeply grateful for your support which has been a constant source of motivation throughout this experience.

Contents

List of Figures

List of Tables

Chapter 1

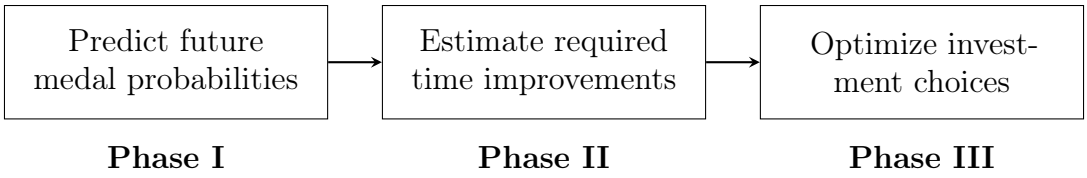
Introduction

Training high-performing professional athletes is a complex task that requires large volumes of resources and attention. Elite sporting organizations must provide coaching, facilities, athletic training, and equipment to athletes in addition to their own focus on research, strategy, and growth of the sport. Hence, efficiently allocating resources is a critical task. Organizations must consider questions regarding which teams should be invested in, which investment opportunities lead to better results, the time horizon for returns on investments, and whether past performance results would have differed had other investment decisions been made.

One such organization that has a particularly large ecosystem under which they face many important investment decisions is the governing body for professional cycling in Australia. Better known as AusCycling, this organization has jurisdiction over all phases of the sport from recreation cycling to elite high-performance athletes spanning all cycling disciplines [1]. Their oversight comprises the work that was formerly carried out by 19 separate governing bodies [2]. It is evident from the size of the organization that efficiently investing resources is critical for the success of AusCycling.

With many possible investment directions, it is important to understand where investing in high-performance cycling fits into AusCycling's priorities. This begins with an understanding of how sports are funded by the Australian government. Since the early 1970s, the government has consistently promoted physical activity through the funding of sporting organizations. After the 1976 Olympics, the intention of sporting investments expanded to include an emphasis on elite international competitions. Since then, recognized National Sporting Organizations (NSOs) have been the beneficiary of funds specifically allocated for high performance. This has resulted in the success of high-performance teams becoming a main priority for many NSOs [3] and therefore implies the critical importance of the high-performance umbrella under AusCycling.

The intention of this research is to provide AusCycling with a methodology for determining efficient investments within their high-performance teams. The research specifically considers efficient investments in cycling equipment and their impact on predicted performance in future Olympic Games. The process for reaching such recommendations is a three phased approach as depicted in the following diagram.



Selecting efficient investments first requires understanding how they contribute to a broader objective. In this case, we establish a relationship between investments and medal probability using a model capable of predicting Olympic performance. This forms the focus of the first phase of the project. In Phase II, we develop a method for estimating the required time improvements necessary to move from a given placement at a prior competition to some new placement at a future competition. The methodology can subsequently be utilized to evaluate of the impact of different equipment investments on future placement based on the unique physical specifications of each investment. The third phase of this project utilizes the methods of the first two phases to construct an optimization model which relates different packages of equipment upgrades to predicted Olympic performance across a portfolio of cycling disciplines. The remainder of this paper discusses the following content as it relates to the process of developing investment recommendations: relevant background information (Chapter 2), the Phase I prediction model (Chapter 3), the Phase II method for calculating required time improvements (Chapter 4), the Phase III optimization model which produces the final recommendation (Chapter 5), and a concluding discussion and opportunities for future work (Chapter 6).

Chapter 2

Background

Sport is one of the few domains with the power to bring together individuals who might otherwise remain divided. The Olympics exemplify this on the world's biggest stage, uniting entire nations in support of their athletes and bringing people across the globe together to follow compelling stories, cheer for underdogs, and witness records being broken. Because of the cultural, social, and economic importance of the Olympics, many countries place a large emphasis on high performance at Olympic competitions. This emphasis manifests itself differently across countries and sports. The work in this thesis focuses exclusively on maximizing performance of Australian athletes across various Olympic cycling disciplines.

2.1 Government funding for sport in Australia

Performance at Olympic competitions is taken quite seriously by the Australian government. This is due to a combination of factors of historical government oversight and a belief that more medals won at global competitions boosts Australian viewership and thus boosts both national pride and activity levels of citizens [4]. In order to achieve such high performance, the Australian government has been a consistent investor in sports through Australia's complex sports ecosystem.

2.1.1 A brief history of the Australian sports ecosystem

Government involvement in sports in Australia began in 1941 with the National Fitness Act [3]. Passed during World War II, the National Fitness Act was created with the intention of improving fitness of young people as it related to preparing them for military service and industry work.

Despite the efforts that came out of this program, government involvement in sports remained limited until 1972 when the Whitlam Government established the Department of Tourism and Recreation and appointed the first Minister for Sport. Out of this department came several reports, study groups, and discussions related to improving sports infrastructure within Australia, establishing sports medicine and coaching pipelines, and improving the general fitness of ever-growing sedentary Australian citizens. The sport portfolio has remained within an Australian Government department since its establishment, although it has moved

several times and currently sits within the Department of Health and Aged Care [3].

After much back-and-forth across several administrations, and ultimately after a poor performance at the 1976 Summer Olympics, the groundwork for the modern Australian sports ecosystem began to develop with several schemes relating to coaching, officiating, and sports governance. In 1981, the Fraser Government opened the Australian Institute of Sport (AIS) as a public company dedicated to facilitating sports across the country and providing a structured system for coaching, sports science, and sports medicine. The AIS would later become a Commonwealth Statutory Authority in 1986, one year after the Australian Sports Commission (ASC) was established as a Commonwealth Statutory Authority to oversee resource allocation and policy formulation for sports in Australia. The AIS and ASC merged in 1989 to become the "principal Australian Government agency responsible for sport" [3].

In 1993, the National Elite Sports Council (NESC) was formed as a communication forum for coordination across the Australian sports industry. The Council had representation from the AIS, Australian Institutes and Academies of Sport, Australian Olympic Committee, Australian Paralympic Committee, and the Australian Commonwealth Games Association. The following year, in 1994, the government transferred nearly all sports responsibilities from the Department of Environment, Sport and Territories to the ASC [3].

Since then, the Australian government, AIS, and ASC have continued to release inquiries, reports, frameworks, and suggestions for the evolving sports industry. While this research does not explore these works, a full history of publications and reports can be found at the Clearing House for Sport's website [3] and a visual representation of the current Australian sports ecosystem can be found in Appendix A.

2.1.2 A brief history of government funding of sports in Australia

As the sports ecosystem has evolved both within and outside of the Australian Government, policies related to government funding of sports have also evolved. The first such policy, "The Australian sports kit - Next Step" was developed to improve sports participation among all Australians and to improve the performance of athletes. This was followed by the 1992 "maintain the momentum" policy program which aligned government funding with the cycle of the Summer Olympics and funded continued growth from gains under the Next Step policy. In 1994, the Olympic Athlete Program: making great Australians was established after Australia won the bid for the 2000 Olympic and Paralympic games, providing $135 million in funding over six years. This was additional funding on top of the funding provided by the "maintain the momentum" policy [3].

Throughout the late 1990s and 2000s, the Australian Government and ASC provided frameworks and policies that addressed dual objectives of increasing participation in physical activity among all Australians and a focus on improving the performance of Australia's elite athletes. In 2010, the Australian Government committed to bridging these two objectives by funding each of them and providing a significant investment to a development pathway to link the two. This was followed by a 2011 policy framework which promoted cross-jurisdictional communication, identified targets for funding and policies moving forward, and led to a collaborative initiative that detailed five target areas for taking Australian performance to the next level [3].

In 2018, the Australian Government released the "National Sport Plan: Sport 2030",

highlighting its commitment to Australian sports. In an effort to foster collaboration and align goals, the ASC co-designed "Australia's High Performance 2032+ Sport Strategy: Win Well" (HP2032+) [5] in coordination with Australia's High Performance Sport System in 2022. This was followed in 2023 by "Australia's Participation Sport Strategy: Play Well" which was developed by the ASC to support the development of sports environments. These two programs have worked to align the many key players in the sports ecosystem in a common vision for the future of sport [3].

The Australian Government's most recent focus includes the "Sport Horizon - National Sport Strategy 2024-2034" and "Australia's Sports Diplomacy Strategy 2032+" which both look to align resources and refresh goals. These strategies complement funding from the Government such as the $249.7 million investment into AIS facility upgrades and the recently announced $283 million investment into elite athletes, coaches, and support staff [6].

While these initiatives are some of the largest-scale programs developed by the ASC and Australian Government, a breakdown of yearly funding since 2020 from the ASC to National Sporting Organizations (NSOs) and National Sporting Organizations for People with Disability (NOSDs) can be found in Table 2.1. A full breakdown of government initiatives and investments can be found on the Parliament of Australia website [7].[1]

Year	High Performance	Sport Participation	Other	Total (2023–2024 adjusted)
National Sporting Organizations				
2020-2021	133,983,147	18,337,500	11,888,783	**191,854,115**
2021-2022	125,157,183	12,950,000	17,199,648	**170,948,360**
2022-2023	127,923,829	15,725,000	18,292,059	**168,118,140**
2023-2024	157,069,691	15,725,000	18,679,859	**191,474,550**
2024-2025	186,694,862	22,932,500	6,691,100	**210,528,917**
Total:	**730,828,712**	**85,670,000**	**72,751,449**	**932,924,082**
National sporting organizations for people with disability				
2020-2021	7,987,808	1,535,000	857,500	**12,127,835**
2021-2022	5,843,148	1,535,000	2,216,255	**10,560,691**
2022-2023	4,645,000	1,535,000	3,324,200	**9,866,739**
2023-2024	6,519,500	1,535,000	5,151,300	**13,205,800**
2024-2025	14,389,370	2,362,500	4,135,000	**20,327,854**
Total:	**39,384,826**	**8,502,500**	**15,684,255**	**66,088,919**

Table 2.1: ASC funding to NSOs and NSODs. The table has been condensed and modified from the Parliament of Australia website to only include the last five years [7].

Funding from the ASC is also directed to specific NSOs. Table 2.2 shows the total funding received by the top 3 highest funded NSOs from 2015-2016 to 2024-2025.[2,3]

As seen in Table 2.2, cycling is the second highest funded NSO, trailing only swimming.

[1] All dollar amounts referenced in this paper are in Australian dollars.

[2] The top 10 highest funded NSOs can be found on the Parliament of Australia website [7].

[3] Funding totals are not adjusted for inflation.

NSO/ NSOD	High Performance	Sport Participation	Other	Total
Swimming	130,976,720	6,987,000	2,324,491	**140,228,211**
Cycling	107,579,565	5,945,698	4,434,206	**117,959,469**
Athletics	103,434,882	4,920,000	1,778,450	**110,133,332**

Table 2.2: NSOs with the highest funding from the ASC since 2015-2016. [7]

This highlights the importance of cycling participation among Australians as well as high-performance programs for Australian cyclists and is a main motivator for this research.

2.2 AusCycling

AusCycling Limited, known by its trading name of AusCycling, is the National Sporting Organization for cycling in Australia and was formed in 2020 through the merging of 19 separate governing bodies. They represent the interests of all 450 cycling clubs and over 50,000 individuals and their oversight includes youth programs, recreation, commuter cycling, junior clubs, and elite and professional cycling in all disciplines [1],[2].

As the second highest funded NSO (both in total funding and in funding for high performance), AusCycling places a large emphasis on strengthening the performance of their elite athletes with ambitious goals for performance on the international stage. The main objective of this research is to assist AusCycling in identifying new equipment investments that will assist Australian athletes in achieving these goals and will be an efficient use of AusCyling resources.

2.3 Cycling disciplines to be considered

There are eight different cycling disciplines as depicted in Figure 2.1 [8]. Given that this work is driven by a desire to earn medals at Olympic competitions, we do not consider the Cyclo-cross or Gravel disciplines as these are not Olympic events. Of the six cycling Olympic disciplines, we only consider BMX Racing, Road, and Track events.

The three Olympic disciplines not being considered are left out of this work due to limited data availability:

i. BMX Freestyle debuted at the Tokyo Olympics in 2021. Prior to its Olympic debut, BMX Freestyle had limited competition standardization and thus the only consistent data for Freestyle competitions is from the last three years [9].

ii. Mountain Biking data is limited and inconsistent due to the lack of standardization across competitions. Each Mountain competition varies drastically and is largely impacted by the geography of the competition location.

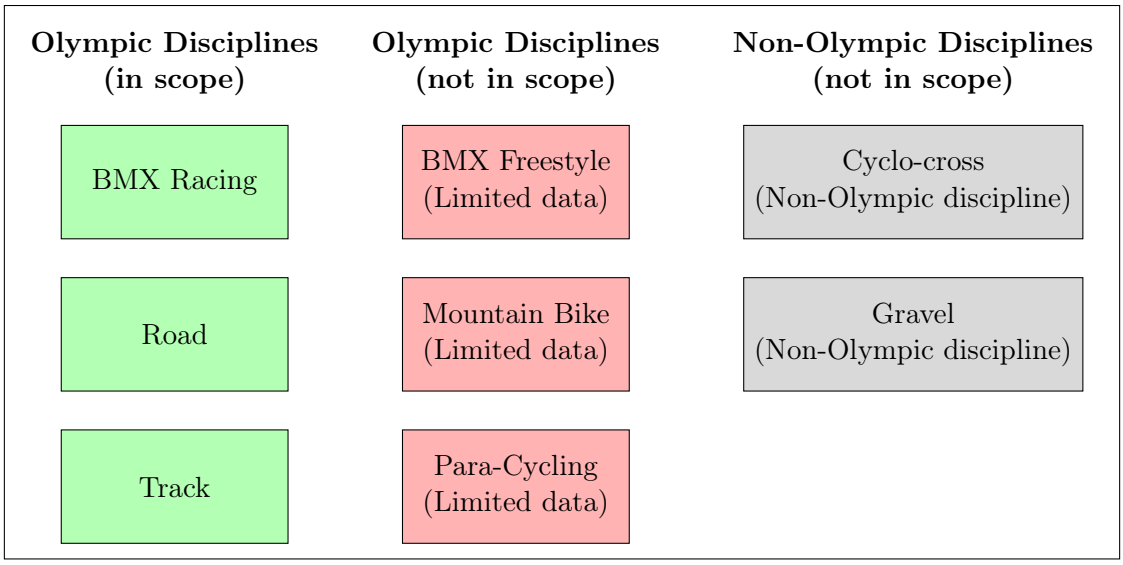


Figure 2.1: Overview of cycling disciplines considered and not considered in this study.

iii. Para-Cycling has limited data due to a small sample size of athletes competing within each classification. Additionally, data tracking for Para competitions is relatively recent and not yet standardized [10].

For each of the three disciplines that are in scope, we must specify the events within those disciplines which are considered in this research. It is important to understand the Olympic history and format of each of these events.

2.3.1 BMX Racing

BMX racing debuted at the 2008 Olympic Games in Beijing for both men and women and has been at every Olympics since [11].

BMX Racing - Format

The race itself is made up by eight individual cyclists. These athletes race a single lap on a track built with straights, jumps, and turns. Placement is determined by time with the winner being the first athlete to cross the finish line [11].

2.3.2 Road Time Trial

While Road cycling consists of both the road race and an individual time trial, we will not consider the road race in this research. This is due to the fact that the nature of the road race, which is a stage-based event, is significantly different in format and strategy from the in-scope events.

The individual road time trial made its debut on both the men's and women's program at the 1996 Atlanta Games [12].

Time Trial - Format

The Time Trial is an individual event where riders complete a road-based course on their own, spaced at even intervals. While it is possible that riders may catch up to each other, drafting or working together in any other way is not legal. The race distance is typically between 40 and 50 kilometers. Placement is determined by time with the winner being the rider with the fastest split to complete the course [13].

2.3.3 Track Team Disciplines

Track is the largest cycling discipline at the Olympics and is comprised of three team events and three individual events.

Madison

The Madison was on the men's program from 2000-2008 but was removed prior to the 2012 London Games for gender parity reasons. The Madison was not recognized as a women's event by the world governing body for cycling, Union Cycliste International (UCI) [8], until 2016. In 2021, the Madison returned for the men and debuted on the women's program [14].

Madison - Format

Teams consist of two riders who take turns sprinting, as often as they want. Since only one rider is active at a time, the athletes switch off through a hand sling to catapult the resting rider back into action. The men's race is 200 laps (50 km) and the women's race is 120 laps (30 km).

Points are awarded or deducted based on placings in intermediate sprints (every 10 laps) throughout the race, and for laps gained compared to the main bunch. The winning team is the pair with the most points accumulated at the end of the race [14].

Team Pursuit

Team Pursuit has been on the men's program since 1908 and was added as a women's event in London in 2012 [15].

Team Pursuit - Format

Team Pursuit is a head-to-head race with two teams of four riders. Each teams starts on opposite sides of the track and race for 4 km (16 laps). Placement is determined either by time or by one team catching up to the other. The race time is determined as the time in which the front wheel of the third rider of crosses the finish line [1],[8].

While the men's and women's Team Pursuit now have the same format, this was not always the case. Prior to the 2014 season, the women's event was a 3 km race with teams of three. The UCI changed the format in 2014 to create parity with men's Team Pursuit [16].

Team Sprint

The men's Team Sprint made its debut at the 2000 Sydney Games while it appeared on the women's program for the first time in London in 2012 [15].

Team Sprint - Format

The Team Sprint is a race with teams of three riders competing over three laps (750 meters) where each rider leads their team for one lap. Placement is determined by time where the winning team is the one whose last rider crosses the finish line first.

The race format for the Team Sprint is now the same for both men and women after the women's race was changed from a two team, two lap race in 2021 [17].

2.3.4 Track Individual Disciplines

All three Olympic inidivudal track events are in scope for this work.

Keirin

Although a cycling staple since the late 1940s, the Keirin event made its Olympic debut on the men's program at the 2000 Sydney Games and was added for the women at the 2012 London Games [18].

Keirin - Format

In the Keirin, riders follow a pacing motorbike (derny) for 3 laps (750 m) which slowly speeds up. After the third lap, the motorbike leaves the track and the athletes sprint to the finish over another 3 laps (750 m). Placement is determined by the order in which athletes cross the finish line, not by time [18].

Omnium

The Omnium event is the newest event in track cycling and was introduced for both men and women at the 2012 Games in London. It was created to help improve gender parity in track cycling and replaced several individual events. For the first few years of the Omnium, the events that made it up were variable and the current format was set after the 2016 Rio Olympics [19].

Omnium - Format

The omnium is a combined event where athletes compete, in this order, in the Scratch Race, Tempo Race, Elimination Race, and Points Race [1],[8]:

i. **Scratch Race** is a 10 km for men (7.5 km for women) sprint where the first athlete to cross the finish line wins.

ii. **Tempo Race** is also 10 km for men (7.5 km for women) and is similar to an individual sprint version of the Madison where riders can gain points for success in intermediate sprints.

iii. **Elimination Race** is an event where athletes ride at a moderate pace and compete in sprints every 2 laps. At the end of each sprint, the last rider across the finish line is eliminated. This continues until there are two riders left in which case the first rider to finish wins the event.

iv. **Points Race** is a 40 km race for men (25 km for women) that is similar to the Madison. Riders can gain points in intermediate sprints every 10 laps.

In the Omnium, riders gain points in each of the four events. The winning athlete is the rider with the most accumulated points at the end of the fourth event [19].

Sprint

The Sprint (also called Track Sprint or Match Sprint) has been on the men's program at every Olympic Games with the exception of 1904. It was added for the women in 1988 [20].

Sprint - Format

The Sprint is a race between two riders over three laps. The riders start together and employ different tactical strategies before sprinting to the finish. While ultimate placement is determined by finishing rank in these head to head match-ups, the match seeding is determined by time in a 200 m qualifying time trial [20],[21].

Chapter 3

Phase I: Predicting medal probabilities at future Olympics

In order to evaluate which investments are likely to improve medal chances at future Olympics, we must first understand a cyclist's chances of earning a medal based on competition results prior to the Olympics. In this first phase of the project, we develop a model to predict future Olympic placement based on prior competition placement for all in-scope disciplines.

3.1 Data

The raw dataset for this phase of the project is the set of race results for all competitions in the UCI database. While this data was provided by AusCycling, it is pulled from the official records as kept by the UCI. Each entry in the dataset represents an individual athlete's performance in a specific competition. We use all entries in the dataset where the discipline of the competition is one of our in-scope disciplines. A summary of the raw dataset is provided in Table 3.1.

3.1.1 Data by year

The data includes competitions from 1996-2024 with the vast majority of competitions occurring after 2008, as seen in Figure 3.1. This means we only have Olympic results from the 2012, 2016, 2021, and 2024 Olympics and will thus train our prediction model using results from these Olympic years.[1]

Note also that competition results are almost nonexistent prior to 2008, so the relationship between pre-Olympic competitions and Olympic results is limited to a time window of 16 years at a maximum. However, the average athlete in our dataset has much less than 16 years worth of results. Figure 3.2 shows the distribution of the number of years of competition data across all athletes and the average number of years of competition data for athletes in each discipline.

[1]The 2020 Tokyo Olympics were held in 2021 due to COVID-19. This competition will be referred to as the 2021 Olympics throughout this paper.

Data metric	Total number
Competition results	340,912
Countries	177
Athletes	42,395
Australian athletes	2,797
Australian results	15,191
BMX Racing results	72,688
Time Trial results	45,304
Madison results	38,398
Team Pursuit results	24,921
Team Sprint results	23,541
Keirin results	44,008
Omnium results	43,801
Sprint results	48,251

Table 3.1: Summary statistics for raw dataset.

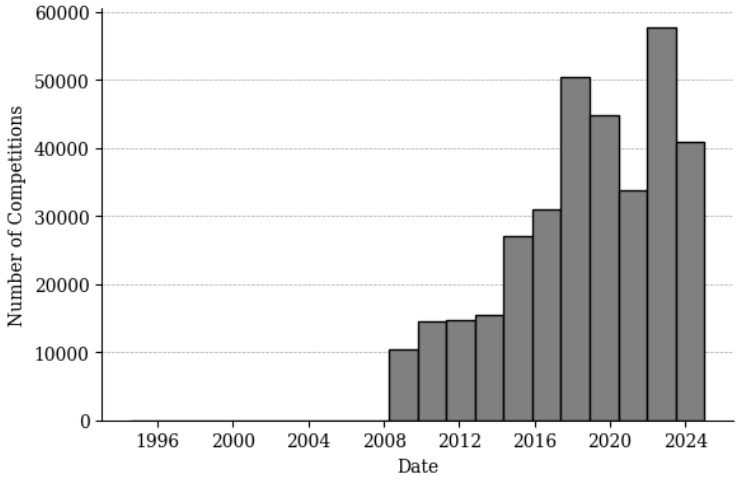
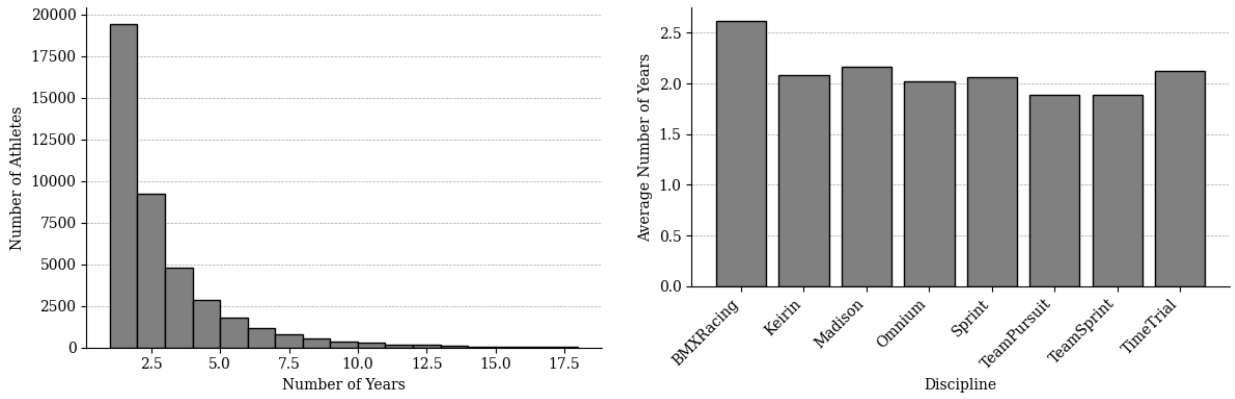


Figure 3.1: Distribution of competitions in the dataset.

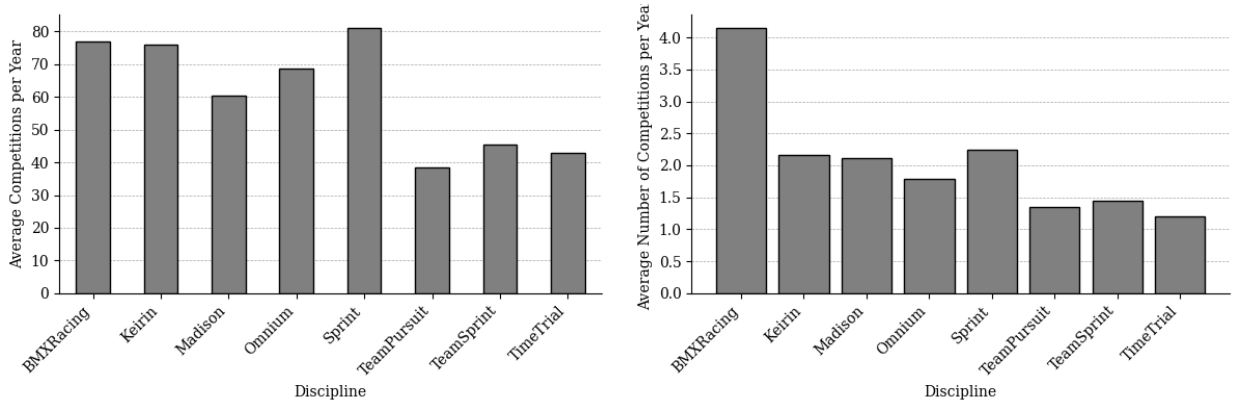
3.1.2 Data by discipline

Since there is a lot of variation across disciplines in terms of Olympic history and race format, we must distinguish between disciplines in our prediction model. Thus, it is important to understand how variable the data is for each target discipline. Figure 3.3 plots the average number of competitions per year for each discipline and the average number of competitions an athlete participates in per year for each discipline. The full distribution of competitions per athlete per year for each discipline can be found in Appendix B.



(a) (b)

Figure 3.2: Number of years of competition data available for athletes: (a) Distribution over all athletes; (b) Average number of years available for athletes in each discipline.



(a) (b)

Figure 3.3: Yearly competition data for each discipline: (a) Average number of yearly competitions with available results per discipline; (b) Average number of yearly competitions per athlete for each discipline.

3.1.3 Data by gender

It is also important to understand the variation in the dataset across gender due to the fact that (i) there is some variability in race format across gender within disciplines, and (ii) the relationship between athletes and race results varies across gender for each discipline. Figure 3.4 depicts the breakdown of male and female competition results for each discipline. As seen in Figure 3.4, there are more results for men than for women for each discipline. This result is likely due to the fact that the women's disciplines were added as official UCI events much later than the men's.

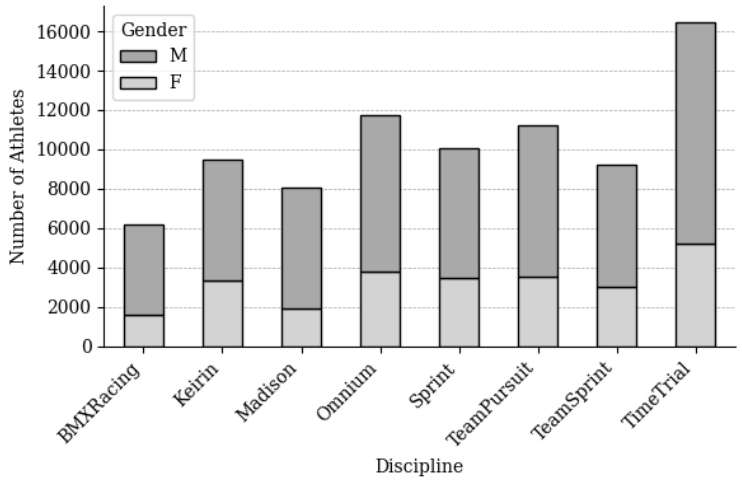


Figure 3.4: Number of male and female competition results per discipline.

3.1.4 Race results: placement vs. time

In the raw dataset there are two variables that track the race result: race value and rank. Race value is the official result of an athlete's race as it is recorded for that discipline (i.e. time for BMX Racing or points for Madison) and rank is the athlete's placement in the race. For the remainder of Phase I, we will use athlete placement to track the result of each race. We use placement instead of race value for several reasons:

i. The data entered under the result value varies drastically for each discipline and the entries with interpretable results are limited. For example, consider the result values entered for BMX Racing. Of the 72,668 entries, 16% of the entries have valid race times while the remaining 84% have the rank as the race result. For this reason, there is not enough consistent result value data to use in our model.

ii. Result values may not always fully reflect performance quality. For instance, there are many competitions, particularly leading up to major competitions like the Olympics, where countries withhold their full-strength packages, choosing not to deploy their best equipment or strategies in the form they plan to use them at the major competition as to not reveal their cards. In such cases, using result value to track the race result can be misleading as a "worse" result value may not necessarily indicate poor performance. Instead, using placement in the competition is a more reliable measure, as it reflects the relative standing of athletes without the influence of performance-specific fluctuations.

3.2 Prediction model motivation – correlated competitions

The relationship between prior competition results and Olympic performance is central to understanding how past events inform future Olympic outcomes and thus plays a critical role in developing our prediction model.

In this section, we examine the correlation between prior competition and Olympic performance to evaluate prior results and motivate our model. We do this by calculating correlations between an athlete's placement in a given competition and their placements in subsequent Olympic competitions.

3.2.1 Correlation methodology

Pairing prior competitions and Olympics

For each competition in the dataset, we form a pair between that competition and all Olympic competitions that occurred after it for athletes who participated in both events. Within these pairs, we track only the distance (in days) between the two competitions. This allows us to abstract the year out of a given Olympics and consider the Olympics themselves as a unified event which we will call $O_{Olympics}$ such that:

$$O_{Olympics} = \{O_{Olympics,i},\ i \in (2012, 2016, 2021, 2024)\}$$

To prevent circularity in our calculations, we do not pair prior Olympic competitions with future Olympic competitions. In other words, we are looking to specifically measure the relationship between non-Olympic prior events and future Olympics.

Correlation calculation

For a given pair of a prior competition and Olympic competition, we compare the placement in both events for athletes that competed in both by calculating the Pearson correlation coefficient between placement in the prior event and placement at the Olympics. The Pearson correlation quantifies the linear relationship between two variables, as given by Equation 3.1

$$r_d = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{3.1}$$

where d is the distance in days between the two competitions, and x_i and y_i are the placement of athlete i in the given prior competition and in the Olympic competition d days afterwards in $O_{Olympics}$, respectively.

Note that r_d cannot be calculated for $n < 2$ and $r_d = 1$ for any $n = 2$. Thus we calculate the correlation only for pairs where at least three athletes participated in both events. We use the Pearson correlation as it captures the linear relationship between athlete placements in prior competitions and subsequent Olympic placements. More specifically, it gives a measure of how similarly athletes rank in both events.

Correlation results

The correlation calculation in Equation 3.1 assigns each competition a separate correlation for all Olympics occurring after that competition if three or more athletes competed in both that competition and the given future Olympics. Figure 3.5 provides a visualization for the distribution of the correlations of past competitions with the Olympics for all Olympic years in the dataset. The observed correlation distribution reveals multiple insights:

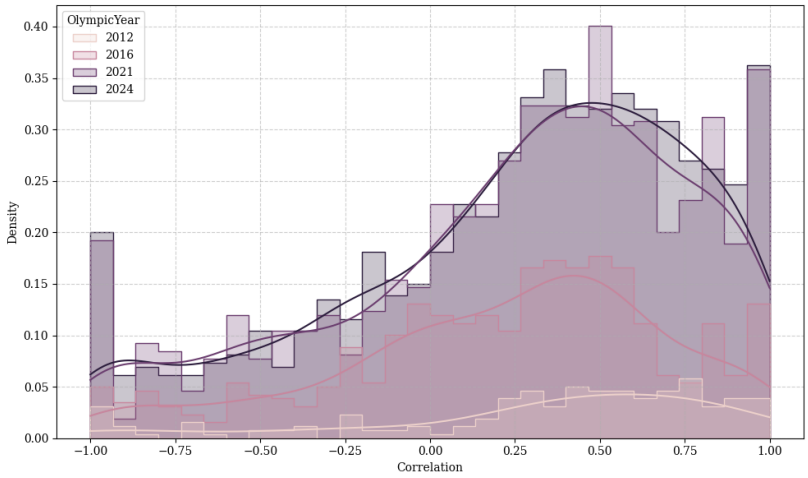


Figure 3.5: Distribution of competition-to-Olympics correlations by Olympic year.

i. While the density increases in later Olympic years as the quantity of data increases, the shape of the distribution of correlation is stable across each Olympic year.

ii. The distribution highlights the variance of competition importance as they relate to Olympic performance.

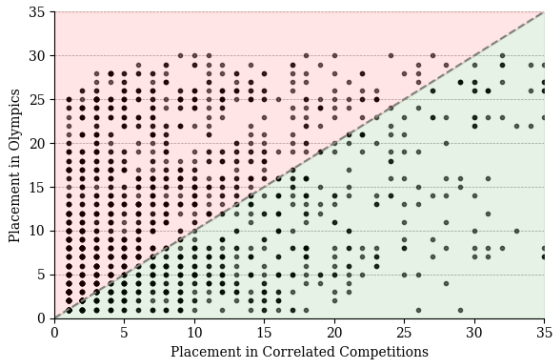
iii. The left-skew of the distributions indicate that the majority of competitions in the dataset have some positive correlation with future Olympic performance.

Given the insights drawn from the distribution of correlation, it is reasonable to consider correlation to be an important feature that may serve as a proxy for competition importance and will therefore be incorporated into our predictive model.

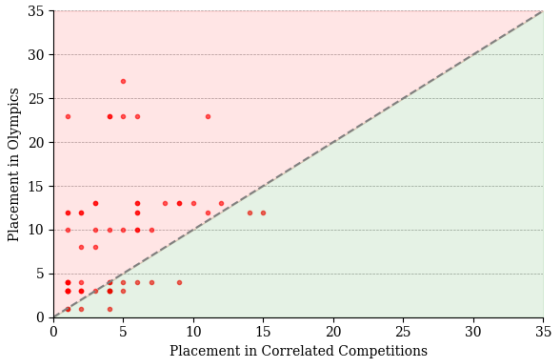
3.2.2 Correlation visualization and performance results

We use the computed Olympic correlation for all valid competitions in the dataset to better understand historical performance in positively correlated events. Figure 3.6 graphs the placement of athletes in prior competitions with their corresponding placement in the Olympics for all prior competitions. We plot this for all $r_d > 0$ for all d such that the prior competition occurred 2 and 4 calendar years prior to the Olympics. We include similar graphs for the other 7 disciplines in Appendix C. Figures 3.6a and 3.6c graph the results for all athletes who participated in both a correlated event and the Olympics while 3.6b and 3.6d graph results only for Australian athletes.

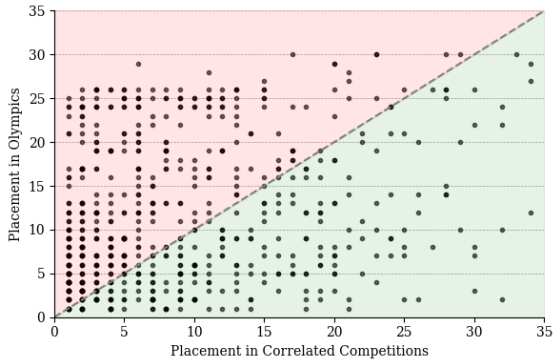
The 45° line is an indicator of perfect parity between performance at the Olympics and prior competitions. We note that the average correlation for competitions represented in this plot will be non-zero and between the x-axis and this line. Any points plotted above the line (in the red region) indicate under-performance at the Olympics relative to theoretically perfect correlation (the line) and under-performance relative to the average correlation for plotted competitions. By symmetry, points below the line (in the green region) indicate over-performance at the Olympics.



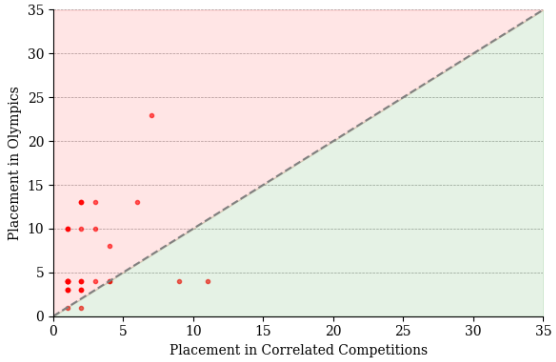
(a) All Sprint results 2 years prior to Olympics



(b) Australian Sprint results 2 years prior to Olympics



(c) All Sprint results 4 years prior to Olympics



(d) Australian Sprint results 4 years prior to Olympics

Figure 3.6: Placement in correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics. The red region indicates under-performance at the Olympics, the green region indicates over-performance at the Olympics, and the 45° line represents perfect parity between Olympic and prior competitions.

In this example, Australian athletes competing in Sprint events at competitions positively correlated with the Olympics 2 and 4 years in advance are under-performing at the Olympics relative to their performance at these competitions. While this result is substantial in its implications, it also solidifies a need to develop a predictive model to evaluate the relationship between a given competition and future Olympic performance as correlation alone does not have complete predictive power.

3.3 Feature Engineering

3.3.1 Average correlation of similar competition

In the previous section, we show that correlation may be an important feature for evaluating the relationship between prior and future Olympic performance. However, we cannot use correlation as defined above as a model feature as that would inject circularity into the model. More specifically, an r_d feature would mean that a component of the input vector was calculated by evaluating the linear relationship between placement in the given competition with placement in the Olympics, which happens to also be the target feature.

The following sections detail the robust methodology used to engineer a correlation feature that captures a value relating to the correlation of similar competitions with the Olympics and removes any circular dependencies.

Clustering competition names

In order to calculate a new correlation feature from similar competitions, we must first cluster competitions to model competition similarity. The intuitive idea behind clustering competitions is that for a given athlete's yearly competition schedule, there is some ordering of importance of the competitions and similar competitions are prepared for in the same way across multiple years.

An initial clustering of competitions based on this intuition may be world competitions, regional competitions, and national competitions. This indicates that the competition name itself may be a feasible clustering dimension. We use the widely-used Natural Language Processing (NLP) feature vectorization method called Term Frequency-Inverse Document Frequency (TF-IDF) to vectorize the text features associated with competition names.

The TF-IDF method evaluates the importance of a given word within a collection of documents, or in this case a collection of competition names. This method quantifies the significance of each word in a name in relation to all other words in the set of competition names by measuring how often a word appears in a given competition name (term frequency) and reducing the weights for often-used common words such as “the" (inverse document frequency). By applying TF-IDF to competition names, we identify patterns in names based on unique aspects of each competition's name.

With competition names in a vectorized form, we apply K-means clustering to segment the competitions. K-means is an unsupervised learning algorithm that iteratively partitions the data by assigning each point to a cluster based on the closest centroid (average position of all points in a given cluster) before updating the centroid. This method helps to uncover

natural patterns in the data by grouping the vectors solely on their form. Table 3.2 displays a summary of the clusters identified through this process.

Cluster Number	Categorization
1	Vier-Bahnen-Tournee (regional European track tour)
2	European continental BMX competitions
3	BMX Supercross World Cups
4	National Track Championships
5	National Road Championships
6	South American national track championships
7	Oceania track championships
8	French-speaking national road championships
9	UCI BMX World Championships and World Cups
10	UCI Track World Cups, Nations Cups, and World Championships
11	Various national championships (Asia/Europe)
12	Miscellaneous international events

Table 3.2: Cluster assignments for competition names.

Computing the correlation feature

Now that we have a defined notion of competition similarity, we create a correlation feature that relates to the correlation of similar competition without circularity. A diagram for this process is defined in Figure 3.7.

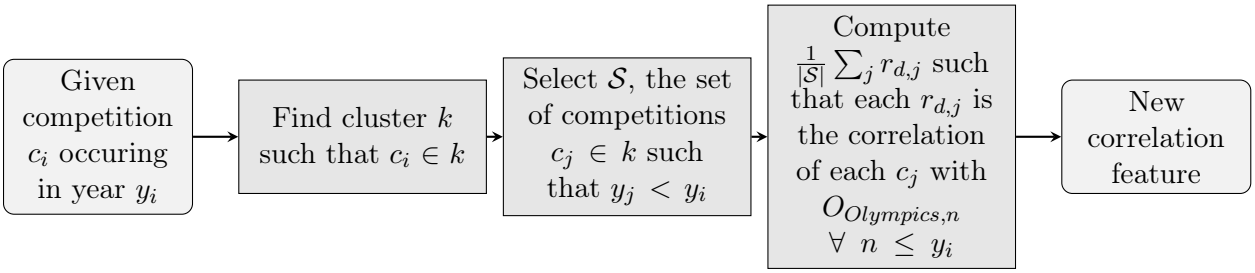


Figure 3.7: Process for creating the correlation feature by using the average correlation of similar competitions from previous years.

The resulting correlation feature is imputed from similar competitions as defined by TF-IDF clustering. The imputation process involves taking the average correlation across the competitions in the cluster that occurred before the given competition, c_i. Note also that we only average across correlations calculated with Olympics occurring before c_i. Therefore, this process removes any circularity but preserves our ability to include some form of correlation as a feature.

3.3.2 Days until Olympics

It is important to capture some notion of the distance between a prior competition and an associated Olympic results. We capture this by creating the *days_until_olympics* variable which is calculated as the difference, in days, between the start of the Olympics and the start of the prior competition. For any competition, we consider only the start date of that competition as the date for all events taking place in the competition. While this may cause a small error in the true number of days until the Olympics, it allows for consistency in evaluating results from the same competition. *days_until_olympics* is positive for all competitions in our final dataset as we are predicting future Olympic performance.

3.3.3 Gender Imputations

The gender column is pre-populated in the raw data for 99% of entries. We perform the following imputation to increase consistency across athletes and limit the number of data points that are thrown out:

i. Many of the competitions are classified as "WE" or "ME" for "Women's Elite" and "Men's Elite", respectively. For any competition result that has one of these classifications as well as a blank gender label, we fill the gender column with "F" for "WE" and "M" for "ME".

ii. Some athletes may have some competition results without a gender label and some with competition results. For any competitions rows missing a gender label, we assign the gender label assigned to that athlete's other competitions.

3.3.4 Handling team-event disciplines

Each entry of the raw dataset corresponds to an individual athlete's performance in a competition. These values can be used directly for individual events, but team events require some additional manipulations. Given that we look to predict future Olympic placement based on prior competition placement, we focus only on the team placement in an event and not on individual performance within a team.

Individuals on a team can be defined as athletes in the raw dataset who are from the same country and have the same data inputs for result value and rank for the same competition in the same discipline. We disregard any instances where the team size identified is larger than the size of the teams for that discipline, as described in Chapter 2.

Team Age

Once the members of a team have been identified, all features in the raw dataset and engineered from the above methods are identical across all athletes on the team except for age. We compute the team age as the average age across athletes on the team. We now have a data point that corresponds to team performance in a competition and we disregard the entries for the individual athletes on the team.

3.3.5 Final model features

Using the engineered features as described above along with some additional information from the raw dataset, we define the final set of features to be used in our prediction model. These features are shown in Table 3.3.

Feature name	Feature type
Age	integer, days
Correlation	float $\in [0, 1]$
Days_until_olympics	integer, days
Rank	integer
Athlete_country	categorical
Competition_country	categorical
Discipline	categorical
Gender	categorical

Table 3.3: Final prediction model features.

3.4 Methodology

As stated throughout this chapter, the goal of this phase is to develop a model to predict outcome probabilities at a future Olympics given prior performance. This is a supervised machine learning problem where we use competition results data labeled with Olympic performance to develop this relationship. The remainder of this section describes the methods used to build the model.

3.4.1 Data format

The format of the original dataset is such that each entry corresponds to an athlete's performance at a competition. We reformat the data such that each entry corresponds to an athlete's performance at a non-Olympic competition as it relates to some future Olympic competition. This expands the size of the dataset. For example, if the original dataset had an entry for an athlete's performance in a 2019 competition, then the new dataset will have an entry for that performance happening before the 2021 Olympics, and an entry for the performance happening before the 2024 Olympics where we also record performance at the corresponding Olympics.

Multi-class classification

AusCycling is looking to predict Olympic performance in terms of the probability of earning a medal at the Olympics (placing no worse than 3rd) and being "close" to earning a medal. We therefore model this as a multi-class classification where the output is a vector of probabilities for each class.

We define the target classes as follows:

- **Class -1:** No Olympic results. Either an athlete did not compete in the Olympics or has no placement result for some other reason such as not finishing the race or a disqualification.
- **Class 0:** Placement at the Olympics earns a medal. The athlete/team placed 1st, 2nd, or 3rd.
- **Class 1:** Placement, p, at the Olympics is "close" to a medal as defined by : $4 \leq p \leq 5$.
- **Class 2:** Placement at the Olympics is not close to a medal, $p > 5$.

For each entry in our dataset, we assign the performance at the corresponding Olympics to one of the four classes.

Input and target features

The target feature for this data is the target class as defined above. Note that we remove any entry from the dataset where the target is -1 as we cannot train a performance prediction model on data without Olympic performance. The input features are all other features as defined in Table 3.3.

3.4.2 Model selection

The model chosen for the Phase I prediction of Olympic performance is a Category Boosting (CATBoost) model. CATBoost is a gradient boosting ensemble method that sequentially builds decision trees in a way that corrects errors made by previous trees. Many of the features of CATBoost make it well-suited for this problem:

i. **Tree-based:** CATBoost makes predictions by building decision trees which is powerful in handling complex, non-linear relationships in the data. This makes CATBoost a reasonable alternative to Logistic Regression which forms predictions by modeling the relationship between features and the target variable as a linear combination and therefore does not capture non-nonlinearities in the relationship.

ii. **Gradient boosting:** CATBoost is a gradient boosting model which means that each successive tree corrects the errors of the prior tree. This process optimizes over prediction accuracy and has been shown to outperform other models such as Random Forests [22],[23].

iii. **Handling of numerical data:** Because CATBoost is a tree-based model, it is not sensitive to the scale of numerical features.

iv. **Handling of categorical data:** CATBoost was developed specifically to provide native handling of categorical features. While categorical features can be used in other predictive models through pre-processing with techniques such as one-hot encoding, label encoding, or target encoding, the encoding of these features may lead to suboptimal handling of the categorical relationships. No other widely used models natively handle categorical variables.

v. **Feature importance:** CATBoost also provides feature importance scores which allows for interpretation of the role of different features in making predictions within the model.

vi. **Overfitting resistance:** CATBoost includes mechanisms that are built to prevent overfitting. These include randomized tree construction and early stopping in training which help to improve generalization. While other tree-based methods are good at preventing over-fitting, CATBoost tends to work better on smaller datasets.

vii. **Comparison to Neural Networks:** While Neural Networks have great success in modeling complex relationships, they are more computationally intensive, less interpretable, and are prone to overfitting.

It is reasonable to hypothesize that the relationship between prior and future performance in cycling is complex and non-linear. Given this hypothesis, and the fact that we have many categorical features of large dimensions (i.e. 177 different competition countries), CATBoost is well suited for our problem.

3.4.3 Train/Test split

For supervised learning tasks, it is necessary to split data into a training and a testing set. The model is be trained on the input features and target value of the training data, and the testing data is used for out-of-sample predictions to measure model performance.

We randomly split 80% of the data into the training set and the remaining 20% into the testing set. This is done such that the split is stratified, meaning that the distribution of the target classes in the dataset is preserved in both the training and test sets. This is essential because the dataset is necessarily imbalanced: the number of Olympic performances where the athlete has not been close to earning a medal across all disciplines and Olympics in our set is significantly smaller than the set of performances where the athlete did earn a medal.

3.4.4 Model parameters

The parameters for the model are selected through hyperparameter fine-tuning using grid search. Grid search is a method that exhaustively searches the the hyperparameter space to find the optimal combinations of hyperparameters for the model. At each iteration during grid search, we evaluate model performance using cross-validation on five splits of our data. This prevents overfitting by training on different subsets and allowing for an unbiased estimate of performance. Each model is trained and evaluated using the **cross-entropy loss function** for multi-class classification. We use grid search with cross-validation to select the parameters as described in Table 3.4.

3.5 Prediction model

After the fine-tuning on the hyperparameters, we train our final CATBoost model on the full training set. As in the models trained during fine-tuning, we use cross-entropy as our loss function.

Hyperparameter	Description	Grid Space	Selection
Learning Rate	Step size at each iteration	{0.01, 0.1, 0.2}	0.2
Depth	Maximum depth of each tree	{4, 6, 8}	8
Iterations	Number of trees to train	{500, 1000}	1000
l2_leaf_reg	Size of L2 regularization penalty	{10, 50, 100}	10

Table 3.4: CATBoost hyperparameters selected via grid search.

3.5.1 Results

The precision, recall, and F1 scores of the CATBoost model are shown in Table 3.5 and the confusion matrix in Figure 3.8 displays the model predictions for the testing data compared with the actual outcome.

	Precision	Recall	F1
Class 0: Olympic medal	0.91	0.92	0.92
Class 1: Close to medal	0.95	0.88	0.91
Class 2: Not close to medal	0.97	0.98	0.98
Accuracy			0.9629
Macro Average	0.95	0.93	0.94
Weighted Average	0.96	0.96	0.96

Table 3.5: Performance statistics for CATBoost Model.

Model Performance on data subsets

While the overall performance of the model on the testing data is important, we are curious about how the model performs for different subsets of the data. Tables 3.6, 3.7, and 3.8 show the performance of the CATBoost model across discipline, gender, and a small sample of countries, including Australia.

3.5.2 Discussion

The CATBoost model performs well in its ability to predict future Olympic performance from performance in a prior competition, with an overall accuracy of 96.3%. The model performs best on the "not close to medal" category which is unsurprising as the vast majority of the data falls into this category. What is particularly of note is that the model has an F1 score of 92% for predicting whether or not a medal is earned at the Olympics. This is impressive performance for an Olympic prediction based on a single performance in some prior competition. The model performs the lowest on the "close to medal" class with an F1 score of 91% resulting largely from the low recall score of 88%. This indicates that while the model performs well when it assigns a "close to medal" label and therefore has minimal false positives for this class, it does produce false negatives in misclassification of some "close" performances. The low recall score means that we may be assigning some instances as "far

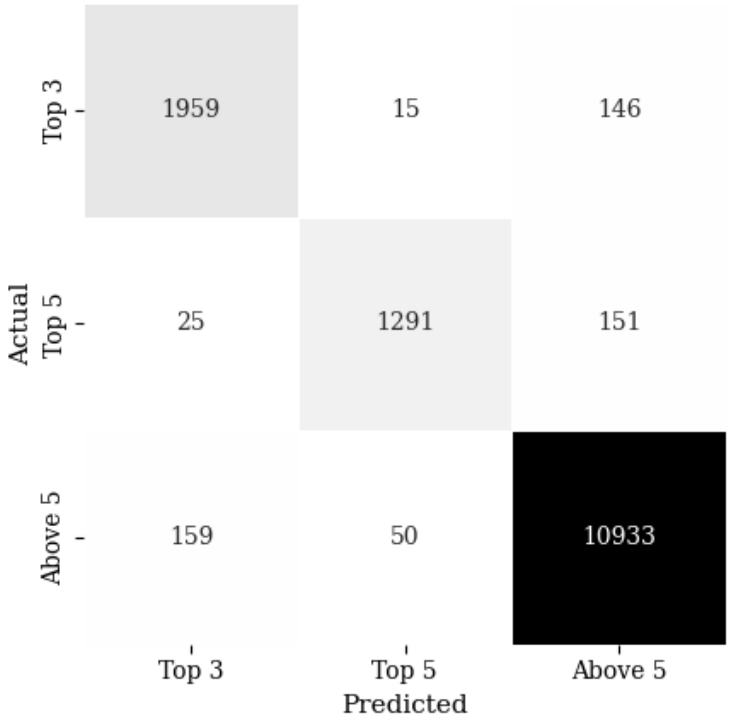


Figure 3.8: Confusion matrix for CATBoost Model.

Discipline	Accuracy
BMX Racing	0.954
Time Trial	0.929
Madison	0.966
Team Pursuit	0.990
Team Sprint	0.990
Keirin	0.969
Omnium	0.964
Sprint	0.974

Table 3.6: Accuracy per discipline

Gender	Accuracy
Male	0.959
Female	0.967

Table 3.7: Accuracy by gender

Country	Accuracy
Australia	0.966
Great Britain	0.923
New Zealand	0.964
United States	0.916

Table 3.8: Accuracy by athlete country

from medal" when they are actually close. However, the model is still useful in its "close" predictions as we are confident that an instance will be labeled as "close" if it is indeed close.

While the high overall accuracy is indicative of model success, it is also important to achieve high accuracy on data subsets. The lowest performance across disciplines is an accuracy of 92.9% for Time Trial. The model is 96.6% accurate for Australian athletes with similar accuracy scores for some of Australia's major competitors. In addition, the model performs well for both male and female competitions with accuracy scores of 95.9% and

96.7%, respectively. High performance across each of these subsets is an indication that this model can be used to make reliable predictions for the types of events in question. In Section 4.3 we discuss the implications of this model for Australian performance at the 2028 Olympic Games in LA.

Chapter 4

Phase II: Estimating required performance improvement

In Phase I, we build a model to predict the probability of medaling at the Olympics, being close to medaling, or being far from medaling. In this phase, we shift the perspective to ask not "what is the chance of high performance at the Olympics" but "how much better do athletes need to be in order to have a high probability of high performance". This chapter details the methodology developed to answer the latter question for Australian athletes.

4.1 Data

The data used for Phase II is a subsection of the raw dataset provided by AusCycling from the UCI database, as described in Chapter 3. While the source data is the same, we use the data differently to meet the unique objectives for this phase.

4.1.1 Race time as target data

Race time is used as the performance metric for this phase in order to quantify performance improvement needed to boost Olympic medal probabilities. This differs from the use of race placement in Phase I. The reason for using time data in this phase is because the continuous nature of time data captures small variations in performance that are not present when looking only at a discrete placement. With race placement, there is no concept of the magnitude of differences between athletes. Athletes who placed 2nd, 3rd, and 4th have the same distance between them when looking at placement, but understanding that the 3rd place athlete was 0.1 seconds behind 2nd and 1.7 seconds ahead of 4th reveals much more about the relative difference between the athletes.

4.1.2 Disciplines

In Chapter 3, we describe the limited number of interpretable time results in the raw dataset. The use of time data for Phase II therefore limits the scope to only those disciplines which have (i) time data as the result value and (ii) sufficient time data.

i. Madison, Keirin, Omnium, and Sprint are all excluded as the final result values in these events are not determined by time.

ii. BMX Racing and Time Trial are also excluded from this phase due to limited time results in the dataset.

After excluding these events, we are left with Team Pursuit and Team Sprint as the two disciplines in-scope for this phase. These disciplines have valid time data for 97.1% and 97.3% of competition results, respectively.

4.1.3 World championships

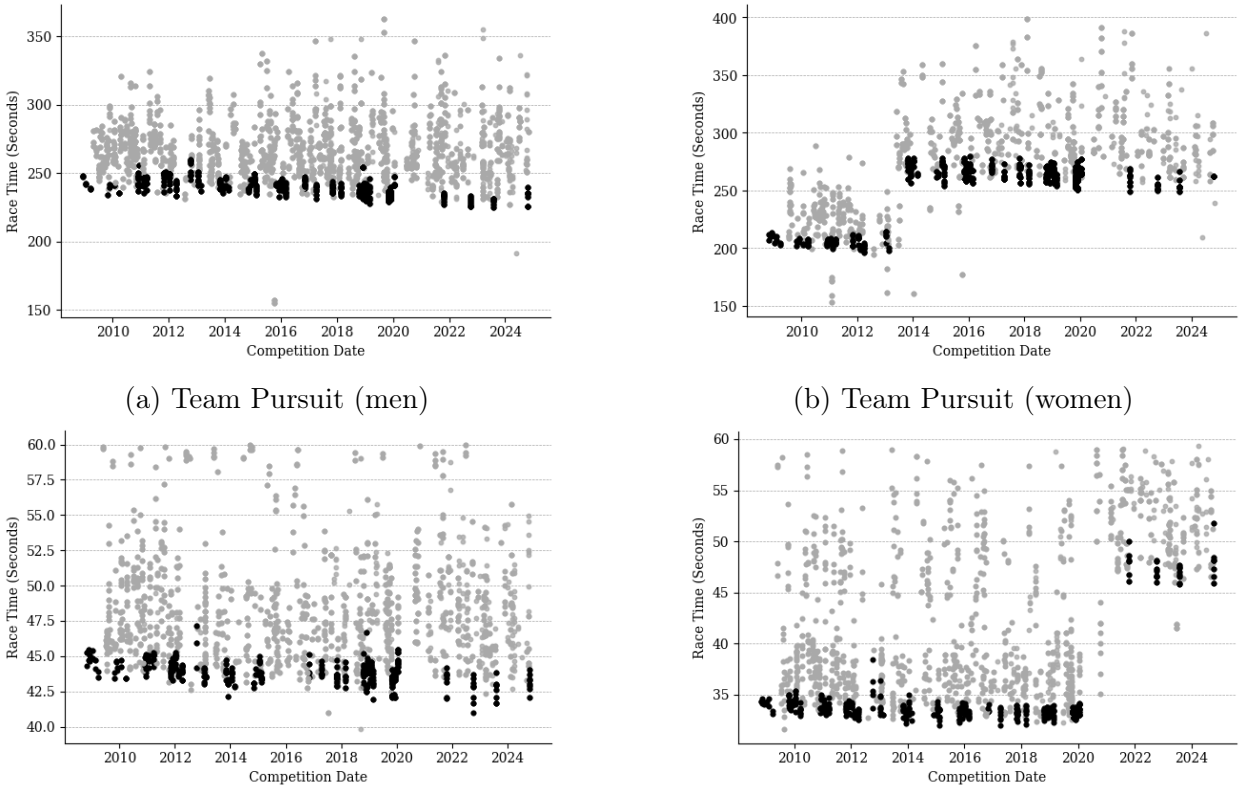
Given the objective to determine the required percentage improvement in time necessary to have a high likelihood of earning a future Olympic medal, the theoretical approach to this phase would be to quantify race times within and across historical Olympic competitions. However, the raw dataset is missing race times for all Olympic competitions after the 2012 London Games.

The first alternative idea may be to quantify race time improvements using all of the time results available in the data. While this is feasible, we have discussed in Chapter 3 that there is variation in the assigned importance and athlete preparation across competitions. Thus the results may be subject to significant noise if we use times from all available competitions. Instead, we look to model the relationship between race time and placement using competitions that have a relatively stable level of importance to most professional cyclist. To find such a set of competitions, it is reasonable to turn to the Olympic qualification system.

Athletes qualify for the Olympics in track disciplines via their place in the UCI Track Olympic Ranking. This ranking is created as a weighted average of performance in the last two Continental Championships, the best two results in the UCI Track Nations Cup (previously known as the Track World Cup) over the two seasons prior to the Olympics, and the UCI Track World Championships from the year prior to the Olympics. The calculation weights results from the Nations Cup and World Championships higher than the Continental Championship [24]. This means that these competitions are consistently highly competitive as they have direct implications on Olympic qualifications. In this phase, we look at result times in World Championships and the Nations Cup over time. For the rest of this paper, we categorize both the UCI Track Nations Cup and the UCI Track World Championships as "World Championship" competitions. We exclude Continental Championships from this process as they are weighted less in the qualification calculation.

To verify that World Championships are a valid competition subset to use, we consider how race times at these competitions compare to all other competitions in the dataset. This can be seen in Figure 4.1 for men's and women's in Team Pursuit and Team Sprint where World Championship results are plotted in black and all other competitions are in gray.

The race times from World Championships are consistently among the fastest results in the available dataset. This confirms the hypothesis that World Championships are consistently held in high regard and can be used to quantify race time relationships to performance. Note also the shift in times for both women's Team Pursuit and Team Sprint. These shifts in time are due to the format changes described in Chapter 2. For this reason, we consider only the data after the format change in both events.



(a) Team Pursuit (men)

(b) Team Pursuit (women)

(c) Team Sprint (men)

(d) Team Sprint (women)

Figure 4.1: Race times at World Championships (black) compared to times at all other competitions in the dataset (gray).

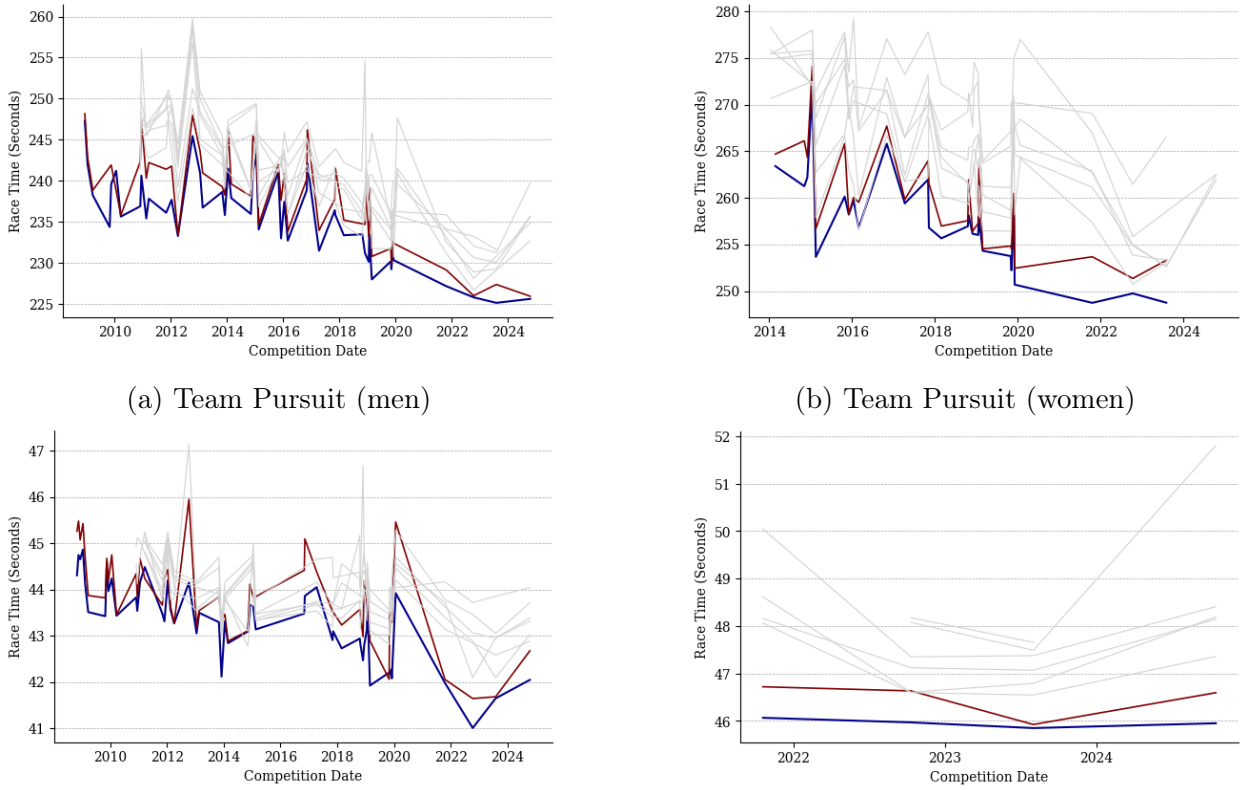
4.2 Methodology

The goal of quantifying time improvement needed for a high likelihood of earning an Olympic medal can be broken down into two key components of (i) natural time improvement over time for a given placement, and (ii) percentage improvement in time needed to improve from one place to another within a competition. This section details the methodology used to compute these two components and combine them to identify total required time improvement.

4.2.1 Modeling natural time improvement

The first component of this process is modeling natural improvement. We define natural improvement as the rate at which result times improve over calendar time when holding placement constant. For instance, we may ask how the time of the World Championship winner has evolved in a recent time window for a given discipline. Figure 4.2 plots race time over time for Team Pursuit and Team Sprint, with the first and second place times highlighted in blue and red, respectively, and all other places 3-8 in gray. Note that due to considering only the World Championships after the change in format for women's Team

Sprint, there are only four data points for this discipline. For this reason, we omit women's Team Sprint for the remainder of this paper.



(a) Team Pursuit (men)

(b) Team Pursuit (women)

(c) Team Sprint (men)

(d) Team Sprint (women)

Figure 4.2: Race times at World Championships for all teams that placed in top 8. Blue lines indicate first place, red lines indicate second place, and gray lines indicate places 3-8.

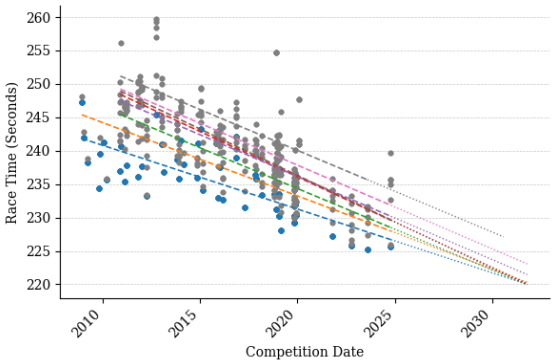
Linearity assumption

Although the time results for each discipline are subject to some variation, there is a stable decline in race time over calendar time. To quantify this relationship, we assume that the decline in race time follows a linear relationship. The assumption of linearity in race times follows from the fact that our dataset considers results within a subset of years. This subset is a small window in the grand scheme of some larger function through which race times evolve. Note that while we do not argue that the larger race time–calendar time relationship is linear, we do assume that we are "zoomed in" on a small enough moment in time of the overall function that it can be linearly approximated in the available subset. Given this assumption, it is sufficient to use a linear regression to describe the natural rate of improvement on the available data and a small window of future time. For any analysis looking further into the future, more detailed research on the evolution of race times in each discipline would need to be conducted, but this is out of scope for this research.

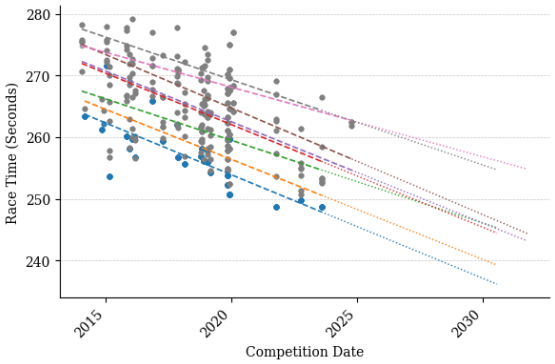
Rate of improvement

Using the assumption of linearity, we can fit a linear model to the race results. We do this separately for each placement position from 1-8. One can think of an example where it is important to model placements separately if, for instance, two or three teams dominate the field over an initial period before the remaining teams begin to "catch up" by closing the gap in race times.

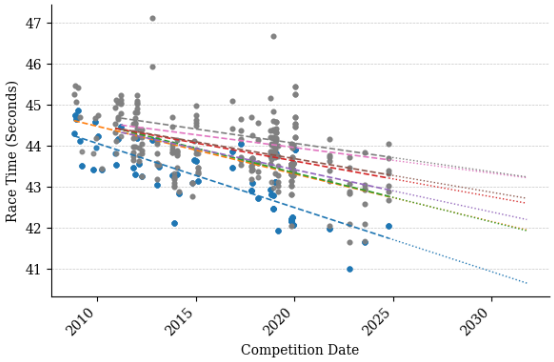
Figure 4.3 plots the results of fitting a linear regression model to the data to predict race times for a given place in a World Championship. We plot this for each discipline that is in-scope for Phase II. Race times are plotted for all teams that placed 1-8 in a given World Championship, with the winning time highlighted in blue. We also graph the result of the regression for each placement with both the fitted line and predictions for the next eight years.



(a) Team Pursuit (men)



(b) Team Pursuit (women)



(c) Team Sprint (men)

Figure 4.3: Linear regression on race times for each placement position in World Championships. Points plotted in blue represent first place finishes at each respective World Championship and points in gray represent places 2-8. Each best fit line represents a linear representation of time improvement for each place 1-8. First place is indicated by blue with the remaining places represented (in place order) by orange, green, red, purple, brown, pink, and gray.

The results of the linear regression model express the relationship between race time and

calendar time for each placement and discipline. We can therefore use the slope of each fitted line as the natural rate of improvement over time at World Championship competitions for a given placement in each discipline. These rates of improvement can be found in Appendix D.

4.2.2 Modeling distance between placements

This section discusses the second component of the Phase II methodology: modeling the distance between placements within a competition. The core concept for this component is exploiting the nature of race time data by computing the percentage difference in race times between any two places. This is calculated using the percent change formula: $\frac{time1-time2}{time2}$. Figure 4.4 shows the percent change in time between a given place and the next-best place for the teams placing 2-8 in men's Team Pursuit in each of the World Championships available in the data. To quantify the distance between places for a general World Championship, we average the percent difference in times between any two places. Table 4.1 shows the specific case where we calculate the average percent change needed to improve to the next-best place in men's Team Pursuit for the top 8 places. The same results for the other disciplines can be found in Appendix E.

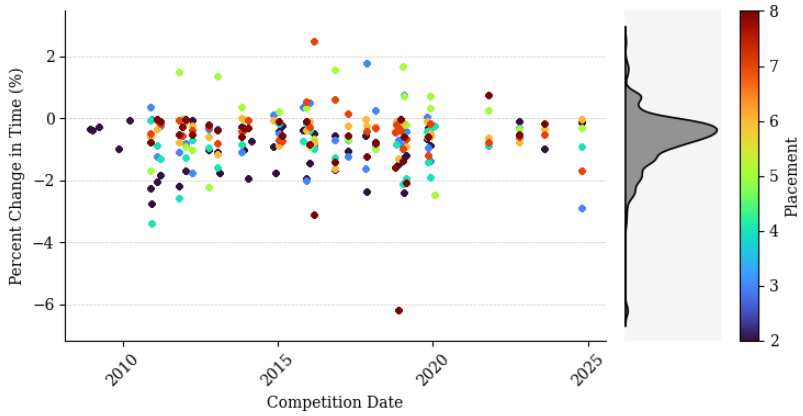


Figure 4.4: Percent change in time from one place to the next-best place in men's Team Pursuit. Each color indicates results for a different place in the range 2-8. The histogram displays the distribution of required change in time across all places in the World Championships dataset.

Note that there are a few instances where the percent change to move up one place may be positive. This typically occurs on the qualification boundary which is defined as the cutoff place for advancement. For both Team Pursuit and Team Sprint the cutoff boundary for the first round is between place 8 and 9, the boundary for the final round is between place 4 and 5, and the boundary for the "Gold medal match" is between place 2 and 3. No team above the boundary can out-place a team below the boundary even if a team below the boundary has a slower time in a later round. This explains occurrences of positive performance gaps, and is the reason why the average percent change to improve one place is lower at the boundaries.

Place	Percent change required
2	-0.999
3	-0.373
4	-0.970
5	-0.329
6	-0.439
7	-0.367
8	-1.090

Table 4.1: Average percent change in time needed to achieve the next-best place within a competition.

4.2.3 Calculating required improvement

We now have a quantity to describe the percent change in time for one place across different competitions and a quantity to describe the percent time improvement necessary to move places within a single competition. These two quantities can be combined to determine the total required time improvement to change placement from one competition to a future competition using the formula for successive percentage improvements given by Equation 4.1 where T is total fractional time improvement for n different fractional time improvements, $t_i(d)$. The fractional time improvements are given by the necessary improvement and the time-delta of that improvement.

$$T = 1 - \prod_{i=1}^{n}(1 - t_i(d)) \text{ where } t_i(d) = a_i \cdot d \tag{4.1}$$

We use Equation 4.1 with $n = 2$ to calculate the percentage change in time necessary to move improve j places at a World Championship d days after a prior World Championship. This example is depicted in Equation 4.2 where a_1 is the time improvement between two World Championships d days apart to remain at place r, and a_2 is the time improvement to move from place r to $r + j$ within a competition ($d = 1$).

$$T_{r,r+j} = 1 - (1 - a_1 \cdot d) * (1 - a_2) \tag{4.2}$$

4.3 Use case – Required performance improvement for Australian athletes

In this section, we use the methodology described above to compute the required performance improvement for Australian athletes from their most recent World Championship performance to the closest World Championships to the 2028 Olympics. While this model can be used for a variety of applications, this example considers the scenario where we compute performance improvements necessary to achieve a threshold Olympic medal probability for each discipline. This requires knowledge of the relationship between placement at the 2027 World Championships and placement at the 2028 Olympics which can be computed using our prediction model from Phase I.

Note that the closest World Championship competition prior to an Olympics is the UCI Track World Championships hosted the year prior to the Olympics. To make predictions for the 2028 Olympics in LA, we consider the 2027 World Championships which will be held starting August 24, 2027 in Haute-Savoie, France.

4.3.1 Using prediction model to identify target placement at future World Championships

We use the prediction model from Phase I to iterate through potential places at the 2027 World Championships and estimate Olympic medal probability for each placement. The target place for the World Championship is selected based on the first place that surpasses the medal probability threshold identified for each discipline. These thresholds are defined based on 2024 Olympic performance:

i. **Team Pursuit (men):** A high threshold of 70% is set due to the team's gold medal results in 2024.

ii. **Team Pursuit (women):** Given the team's seventh place performance in 2024, we set a low threshold of 10%.

iii. **Team Sprint (men):** The threshold is set at 33% based on the team winning a bronze medal in 2024.

Medal probabilities are determined by results from ten independently trained CATBoost models using the parameters identified in Table 3.4 with the following feature inputs:

i. **Age:** We use two different measures for average team age. The first assumes all members of the team remain the same and therefore uses the average team age in 2024 plus 4 years. The second measure assumes that new members are introduced such that the team average age remains the same as the 2024 team. All results for the remainder of the paper use the latter measure.

ii. **Correlation:** Correlation is imputed from the 2027 World Championship competition name as described in Chapter 3.

iii. **Days_until_olympics:** The number of days between the 2027 World Championships and the 2028 Olympics (369 days between the start of both).

iv. **Rank** Iterate through places 1-9 for potential placement r at the World Championship

v. **Athlete_country:** Australia

vi. **Competition_country:** France

vii. **Discipline:** Team Pursuit or Team Sprint

viii. **Gender:** M or F

The resulting mean and standard deviation of 2028 Olympic medal probabilities given placement in the 2027 World Championship are given in Table 4.2. Results are aggregated from calls to ten independently trained prediction models and are reported for places 1-8.

	Medal Probability		
Worlds Place	**Team Pursuit (men)**	**Team Pursuit (women)**	**Team Sprint (men)**
1	0.866 ± 0.038	0.290 ± 0.083	0.433 ± 0.086
2	0.808 ± 0.056	0.195 ± 0.078	0.352 ± 0.088
3	0.737 ± 0.083	0.139 ± 0.057	0.265 ± 0.075
4	0.715 ± 0.087	0.124 ± 0.056	0.249 ± 0.072
5	0.680 ± 0.083	0.111 ± 0.047	0.240 ± 0.059
6	0.673 ± 0.088	0.109 ± 0.046	0.241 ± 0.067
7	0.676 ± 0.076	0.100 ± 0.038	0.250 ± 0.063
8	0.675 ± 0.070	0.097 ± 0.033	0.255 ± 0.068

Table 4.2: CATBoost predictions for 2028 Olympic medal probability in each discipline given potential placement at the 2027 World Championships. The mean and standard deviation for each probability is computed from the results of ten independently trained CATBoost models of the same specifications describe in Chapter 3

Based on our probability thresholds, we select a target placement of 4 for men's Team Pursuit, 6 for women's Team Pursuit, and 2 for men's Team Sprint. Note that the probability thresholds were arbitrarily chosen based on prior performance. This method can be used for future decision-making with any other probability thresholds in place of those used in this example.

4.3.2 Results

Using the target placements selected in the previous section, the Australian team's time in each discipline in the last World Championships appearance prior to 2025, and Equation 4.2, we calculate the total fractional time improvement needed to achieve or surpass the Olympic probability thresholds. Table 4.3 shows the previous Words placement, the target Olympic placement, and the final required change for each discipline.

Note that the Australian men's Team Pursuit team did not have an official time in their most recent World Championship appearance. While they were in the bronze medal match, they were overlapped by their opponent which immediately ends the race. To account for this, we use the time of the fifth place team at that competition as a proxy for the Australian team's time.

4.3.3 Discussion

In Table 4.3, the magnitude of required time improvement for men's Team Sprint is differs from the magnitudes required for time improvement in both Team Pursuit events. This

Discipline	Worlds Place	Target	Required Improvement (%)
Team Pursuit (men)	4	1	-4.87
	4	2	-3.98
	4	3	-3.42
	4	4	-2.43
Team Pursuit (women)	6	1	-7.06
	6	2	-6.02
	6	3	-4.85
	6	4	-3.68
	6	5	-3.59
	6	6	-2.81
Team Sprint (men)	2	1	-2.48
	2	2	-0.79

Table 4.3: Phase II Results: Required percentage change in race time necessary for teams in each discipline to improve from their most recent (pre-2025) Worlds placement to a given target placement at the 2028 Olympics.

result is driven by the difference in how times have improved over the last several years in both disciplines. Figures 4.2 and 4.3 show how the results of Team Pursuit at recent World Championships have been impacted by rapid time improvements across the field. This contrasts with the more stable rate of improvement observed for Team Sprint. Due to the differing trends in World Championship time improvement in each discipline, the regression slopes for both Team Pursuit events are steeper than the slope for Team Sprint. This is the source of the differing magnitudes of required time improvements in each event. Note that these trends are a reflection of how the entire field has evolved in recent years in the two disciplines and we therefore would compute required time improvements of similar magnitudes for any other country in the field.

The methodology developed in this chapter models the relationship between time improvement and placement improvement. Using a linear regression on a small window of World Championship race time results, we can compute results for required time improvements needed to move places between two Worlds competitions. This alone is useful for setting time improvement targets and modeling investments, but the power of the model is amplified when combining it with the prediction model developed in Chapter 3. By combining the two, we have the ability to relate time improvements to Olympic medal probabilities. This relationship provides the tool necessary to set more specific improvement targets in-line with Olympic performance objectives.

Chapter 5

Phase III: Cost optimization and investment selection

In Phase I we build a model to predict future Olympic medal likelihood. In Phase II, we build a method to estimate the required time improvements necessary to improve medal probabilities. Phase III combines the two models, along with an investment horizon, to select the investments which will have the greatest impact on future medal opportunities. This chapter focuses on making these recommendations in the context of equipment upgrades for track cycling. We discuss the physical environment of track cycling, the investment horizon, and the optimization model used to form the final recommendation.

5.1 Physical environment of track cycling

Track disciplines differ from the other disciplines in the scope of this research due to the closed environment in which they take place. All track events occur in a velodrome – an arena with two 180-degree bends, two straights, and steep banking built specifically for track cycling. Velodromes can vary in length, but the official Olympic length is a 250 meter circumference [25],[26]. With the adherence to UCI standards for competitions, velodromes offer a closed environment within which the track disciplines take place. This means that the number of variables influencing track results is much less, and varies much less, when compared to other disciplines that are influenced by the outside environment. For this reason, we focus this section only on track cycling as we can model a direct relationship between track variables and speed which can then be targeted by equipment upgrades. To do this, we must first consider the variables impacting results on the track.

5.1.1 Power needed

Power, one of the key performance indicators for most professional cyclists, is an objective and reliable metric that measures the rate at which riders expend energy [27]. It is therefore important to understand how much power riders must exert in order to ride at a given speed. This allows us to then evaluate physical factors that influence the level of power needed, and how equipment modifications influence them. In the controlled environment of the velodrome,

the total power needed for a rider to sustain a given speed is given by the power needed to overcome aerodynamic drag plus the power needed to overcome rolling resistance as seen in Equation 5.1.

$$P = P_{\text{aero}} + P_{\text{roll}} \tag{5.1}$$

The equations for aerodynamic drag and rolling resistance can be broken down to give the full power needed at the steady state as described in Equation 5.2 where ρ is air density, v is velocity, C_dA is drag coefficient times frontal area, M is the total mass of the system (rider and bike), g is gravitational acceleration, and C_{rr} is the coefficient of rolling resistance [28],[29],[30].

$$P = \frac{1}{2}\rho v^3 C_d A + vMgC_{rr} \tag{5.2}$$

When considering equipment upgrades for track cyclists, it is important to understand their potential impact on Equation 5.2. While g is fixed and ρ is fixed inside the velodrome, the remaining variables can be influenced by equipment:

i. **Aerodynamics:** Research has shown that aerodynamic resistance accounts for nearly 90% of resistance when on a flat surface [30]. Therefore, reducing aerodynamic drag is often the objective of equipment developments or changes. To be precise, the aerodynamic drag can be influenced directly by reducing C_dA, the combined coefficient that reflects the drag coefficient on the frontal area of the system.

ii. **Mass:** The mass of the entire system impacts the power required to accelerate and rolling resistance itself also scales linearly with mass. This includes both the mass of the rider and the mass of their equipment and is therefore a motivator for the development of lighter equipment.

iii. **Tire C_{rr}:** While the surface of the velodrome is flat, rolling resistance still exists and impacts the track environment. Rolling resistance is impacted primarily through the rolling resistance of the tires themselves. Equipment upgrades targeting C_{rr} generally target tire quality and pressure [28].

5.1.2 Physiological power

While Equation 5.2 represents power needed for cyclists to ride at a given velocity, it is important to also account for the physiological power needed from the rider. This quantity is given by Equation 5.3 where P_{needed} is the power needed as in Equation 5.2 and η is mechanical efficiency.

$$P_{physio} = \frac{P_{needed}}{\eta} \tag{5.3}$$

Physiological power represents the output needed to be generated by the rider to ride at a given velocity. If a rider's output capabilities are known as well as the efficiency, aerodynamic qualities, and rolling resistance of the system, then we can solve for the velocity of the rider [27]. In terms of this project, we are concerned with the impact of equipment on speed which can be explained through two variables:

i. **Efficiency:** The efficiency of the system refers to the amount of the rider's output that transitions into mechanical power in the wheels. This is mainly a reflection of drivetrain losses stemming from chains or bearings, rolling resistance energy losses, or slight internal rider losses [31],[32]. In terms of equipment upgrades, efficiency is typically impacted when the drivetrain quality is improved.

ii. **Power (physiological):** The impact of equipment on physiological power relates to how equipment impacts a riders supply of output [33]. In particular, certain equipment upgrades may allow riders to produce a higher output [31]. While the majority of physiological power stems from the abilities of athletes themselves, it is possible for different equipment to enhance or diminish these abilities.

5.2 Data

Since Phase III utilizes the models developed in both Phase I and II, we continue to use the competition results data as describe in Chapter 3 and 4. Note that since we use the model from Phase II, we again only consider Team Pursuit (men and women) and Team Sprint (men). We also introduce new data in this section which outlines the investment horizon to be considered.

The full data for equipment investments will be provided by AusCycling. However, due to timing constraints and data availability, we use synthetically generated data in this discussion. Conversations with the lead performance engineer at AusCycling have influenced the format and magnitudes of all synthetic data used in this model. Magnitudes have also been cross-checked with industry standards [34],[35],[36]. Once AusCycling provides their full investment horizon in the same format as the synthetic data described in this section, the full model in Phase III will output true recommended investments.

5.2.1 Investment horizon

The investment horizon details potential investments being considered for each track discipline. All potential investments are related to equipment upgrades which may include skinsuits, bike components, helmets, booties, or tires. For each investment, we have a low cost, medium cost, and high cost option reflecting the different types of upgrades for a given piece of equipment available to the team. Within each cost option for a given investment is a breakdown of the percentage improvement that investment has on each of the physical variables described above (C_dA, mass, C_{rr}, efficiency, physiological power). Table 5.1 shows the investment categories used in the synthetic data as well as the average impact to each physical variable.

In this research, we assume that the variable improvement percentages use equipment from the most recent Worlds participation as a baseline. For example, if a new skinsuit improves C_dA by 6%, then this is evaluated as a 6% improvement from the C_dA of the rider at the most recent Worlds competition. The results in Table 5.1 are generated from historic industry standard equipment. However, we expect the physical impact of equipment to be subject to diminishing returns. As advancements in equipment continue to be made, the improvement achieved in the relevant physical variables will decrease. To account for this, we

		Variable improvement (%)				
Investment	**Cost Range**	C_dA	M	C_{rr}	η	P_{physio}
Frames	\$800-\$1,000,000	4	4	0.5	2	2
Wheels	\$3,000-\$200,000	5	3	2	2	3
Tires	\$300-\$5,000	0.5	2	5	2	1
Drivetrain	\$1,000-\$100,000	0.5	2	1	6	4
AeroSuit	\$2,000-\$100,000	6	0.5	0.5	2	0.5

Table 5.1: Investment horizon with synthetic ranges [34],[35],[36].

reduce the impact of each investment type by a uniform factor of 0.8. Note that this is done due to the use of historically-grounded synthetic data. The data provided by AusCycling will be grounded in the current horizon and will therefore not need to be reduced.

Additionally, this analysis is specific to equipment and only examines investments of this kind. We do not consider improvements stemming from other sources such as training, fitness levels, practice facilities, etc. However, the model itself is built to incorporate any potential investment and therefore can be used to identify optimal investments given a larger horizon.

5.2.2 Time savings per variable

In addition to data on potential investments, AusCycling's performance engineer models the relationship between the percent improvement in each physical variable to time savings in each event. While the precise results from the model have not yet been shared, we use the following synthetic data based on preliminary results from the performance engineer as seen in Table 5.2.

	Time savings per variable (%)				
Discipline	C_dA	M	C_{rr}	η	P_{physio}
Team Pursuit	0.29	0.020	0.005	0.020	0.610
Team Sprint	1.460	0.120	0.020	0.010	0.710

Table 5.2: Percentage saving in time per a percent improvement in each physical factor.

5.2.3 Time savings per investment

The investment horizon data, as defined by the investment ranges in Table 5.1, identifies the total improvement to each physical factor provided by each investment. Then, the data for time savings per variable, Table 5.2, reveals the impact of a change in each variable on race time. We combine this information to compute the overall impact on race time for each investment option. This is be done by using the compound improvement formula, Equation 4.1, for each investment and cost level. Equation 5.4 shows the formula applied to this scenario where $t_{i,d,l}$ is the total time improvement for investment i into discipline d at cost level l where we aggregate the improvements over all j variables. $v_{i,d,l,j}$ represents

the percentage impact of investment i into discipline d at cost level l on variable j, and $s_{d,j}$ represents the percentage of time saving in discipline d from a change in variable j.

$$t_{i,d,l} = 1 - \prod_j (1 - \frac{v_{i,d,l,j} \cdot s_{d,j}}{100}) \; \forall i, d, l \tag{5.4}$$

5.3 Methodology

In this section, we detail the methodology used to derive recommendations for investment decisions in order to improve predicted medal probability at future Olympics. This is done via optimization modeling. We formulate the selection of investments from our horizon as binary investment decisions which are made by modeling time improvements and medal probabilities using our nonlinear models from Phase I and II. More specifically, we use the Phase II model to map time improvements resulting from investments to placement at the World Championships in 2027. We then use this placement in our model from Phase I to predict medal probability. The resulting optimization is therefore a Mixed-Integer Nonlinear Program (MINLP). To ensure the recommended investment portfolio is an efficient use of resources, we model this as a scalarized multi-objective optimization. We balance the priorities of maximizing expected medals and minimizing the cost of investments.

We utilize Gurobi Optimizer with the branch-and-bound method for solving the model. Branch-and-bound is a method used for integer or binary optimization problems where the solution is found by splitting the problem into subproblems via fixing variables, solving a relaxed version of the problem with continuous variables, and pruning out suboptimal subproblems.

5.3.1 Smooth approximation of nonlinear functions

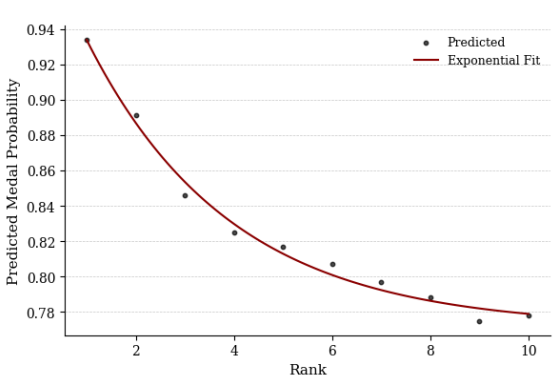
At every node of the branch-and-bound tree, a continuous nonlinear optimization is solved on a relaxed version of the overall model. The subproblem is solved in Gurobi with gradient descent techniques that rely on computing gradients and Hessians where needed. It is therefore important that our optimization model is formulated with smooth, differentiable functions. In this section, we describe smooth approximations for the models described in Chapter 3 and 4.

Medal probability estimation

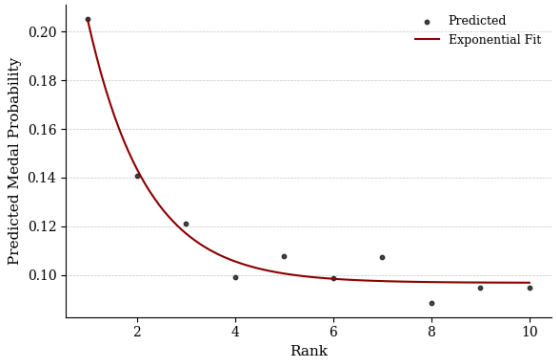
The objective of the optimization is to maximize predicted medals from investments with the intermediate step of predicting 2027 Worlds placement. This indicates that the functional use of the Phase I model is to predict medal probabilities from Worlds placement in 2027 for a given discipline. We compute a smooth approximation for this function by fitting an exponential curve to the relationship between 2027 placement and predicted 2028 Olympic medal probability for each discipline and gender.

As in Chapter 4, we compute medal probabilities with our CATBoost model using the average age of the last Australian team of that discipline and gender to compete in a World Championship, and by predicting the correlation between the two competitions by predicting

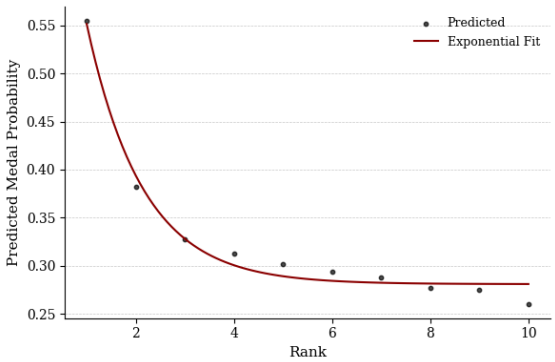
the associated cluster based on the name of the 2027 World Championship. The result of this approximation is shown in Figure 5.2 for each discipline and gender.



(a) Team Pursuit (men)



(b) Team Pursuit (women)

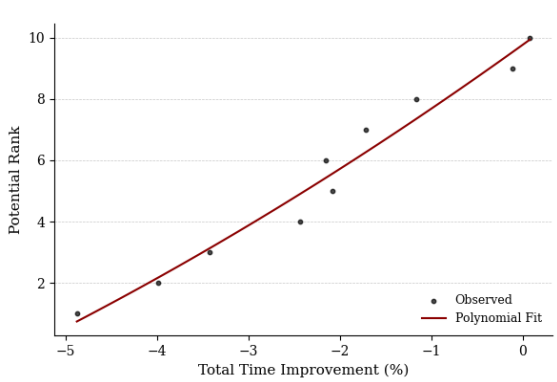


(c) Team Sprint (men)

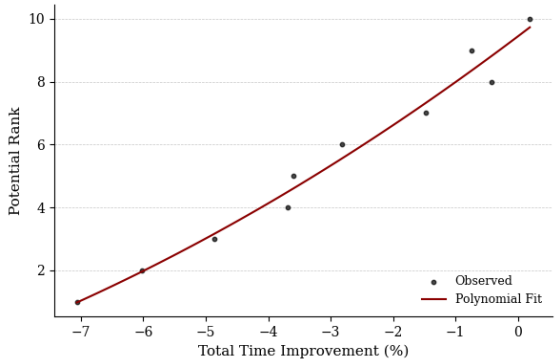
Figure 5.1: Smooth approximation for the function mapping 2027 placement to 2028 Olympic medal probability using CATBoost model.

Predicting placement from time improvement

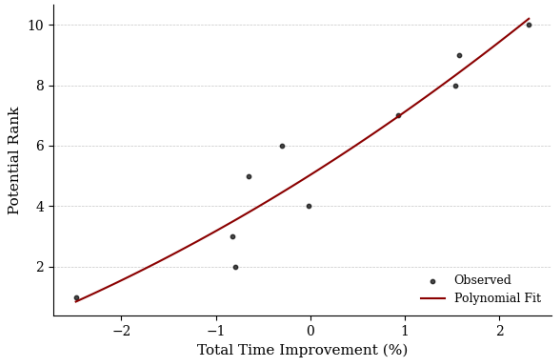
The intermediate step of the optimization flow is to predict placement in the 2027 World Championship based on the total time improvement from a given investment package. To do this, we use the methodology from Phase II. In particular, we assume that the percentage time improvement resulting from a given investment is the time improvement from the race time at the last World Championships competed in by the Australian team of each discipline and gender. Using the prior Worlds time, the Phase II method computes the total time improvement needed to earn any place from 1-10 in 2027. We compute the smooth approximation for this function by fitting a quadratic curve to the relationship between time improvement and placement in 2027. Figure 5.1 shows the results of this approximation.



(a) Team Pursuit (men)



(b) Team Pursuit (women)



(c) Team Sprint (men)

Figure 5.2: Smooth approximation for the function mapping time improvement between the most recent (pre-2025) World Championship and 2027 Worlds to the placement in the 2027 Worlds competition.

5.4 Optimization formulation – Mixed-Integer Nonlinear Programming

This section describes the formulation used for this Mixed-Integer Nonlinear Program in terms of indices and sets, data, decision variables, the mathematical formulation of the problem, and a brief explanation.

5.4.1 Indices and model data

Let:

- $i \in \mathcal{I}$ denote an investment option,
- $d \in \mathcal{D} = \{\text{Team Pursuit}, \text{Team Sprint}\}$ denote a discipline,
- $g \in \mathcal{G} = \{\text{G}, \text{F}\}$ denote gender,
- $l \in \mathcal{L} = \{\text{low}, \text{medium}, \text{high}\}$ denote a cost level

Given the data:

- $cost_{i,d,g,l}$: cost of investment i for discipline d, gender g, and cost level l,
- $t_{i,d,g,l}$: total time improvement from investment option i (as described in 5.4 where total time improvement is the same for both genders for a given investment),
- $a_{d,g}$, $b_{d,g}$, $c_{d,g}$: parameters for the smooth approximation predicting medal probability from rank,
- $cf1_{d,g}$, $cf2_{d,g}$, $int_{d,g}$: parameters for the smooth approximation predicting rank from time improvement,
- Budget $= \$1,000,000$: synthetic budget for AusCycling equipment investments.

5.4.2 Decision Variables

Define:

- $x_{i,d,g,l} \in \{0,1\} = 1$ if we invest in investment i at cost level l for discipline d and gender g, 0 otherwise.

5.4.3 Mathematical Formulation

Objective:

$$max \sum_{d \in \mathcal{D}, g \in \mathcal{G}} \mathbb{E}[P_{d,g}] - \alpha \cdot \sum_{i,d,g,m} c_{i,d,g,m} \cdot x_{i,d,g,m} \tag{5.5}$$

where

Total time improvement:

$$y_{d,g} = 1 - \prod_{i,l} (1 - t_{i,d,g,l} \cdot x_{i,d,g,l}) \ \ \forall \ (d,g) \tag{5.6}$$

Predicted rank:

$$r_{d,g} = cf2_{d,g} \cdot y_{d,g}^2 + cf1_{d,g} \cdot y_{d,g} + int_{d,g} \tag{5.7}$$

Expected medal probability:

$$P_{d,g} = a_{d,g} \cdot e^{-b_{d,g} \cdot r_{d,g}} + c_{d,g} \tag{5.8}$$

$$\mathbb{E}[P_{d,g}] = 1 \cdot P_{d,g} \tag{5.9}$$

s.t.

Constraints:

$$\sum_{(i,d,g,l)} \text{cost}_{i,d,g,l} \cdot x_{i,d,g,l} \leq \text{Budget} \qquad \text{(Budget constraint)} \qquad (5.10)$$

$$\sum_{l \in \mathcal{L}} x_{i,d,g,l} \leq 1 \quad \forall\, (i,d,g) \qquad \text{(Cost level exclusivity)} \qquad (5.11)$$

$$x_{i,d,\text{M},l} = x_{i,d,\text{F},l} \quad \forall\, (i,d,l) \qquad \text{(Gender coupling)} \qquad (5.12)$$

$$r_{d,g} \geq 0 \quad \forall\, (d,g) \qquad \text{(Non-negative rank)} \qquad (5.13)$$

5.4.4 Model Explanation

This model selects the optimal portfolio of investments by solving a MINLP optimization to maximize expected medals at the 2028 Olympics while minimizing the total cost. The objective utilizes α to control the sensitivity of the model to cost. We used $\alpha = 1e-7$ for these results so the impact on the overall objective is minimal. This value can be adjusted in future use to shift the weighting of priorities between expected medal counts and cost.

The total expected medal count is given by the expected medals for each discipline and gender as predicted from the CATBoost model's prediction on the predicted 2027 Worlds placement from the time improvement model for a given investment package. The nonlinear models used to compute expected medals are approximated by smooth and differentiable functions to ensure model feasibility. We use Gurobi Optimizer to solve the optimization.

5.5 Optimization Results

Solving the optimization gives an expected 2028 Olympic medal count of **2.15**. The full results of the optimization on the synthetic investment data are shown in Tables 5.3 and 5.4.

Investment	Discipline	Cost
Graphene Disc Wheels	Team Pursuit	10,104
Low Drag Tires	Team Pursuit	2,682
Aero Tires	Team Pursuit	886
Optimized Drivetrain	Team Pursuit	2,491
Magnetic Drivetrain	Team Pursuit	2,041
Aero Skinsuit	Team Pursuit	2,246
CFD-Mapped Skinsuit	Team Pursuit	2,817
Graphene Tires	Team Sprint	650
Reinforced Drivetrain	Team Sprint	2,643
Carbon-Nanotube Drivetrain	Team Sprint	1,763

Table 5.3: Selected investments from synthetically generated data.

The selected investments make up the most efficient investment package given the budget that maximizes potential medals in the 2028 Olympics. These results are generated from the synthetic data described in Section 5.2 but the tool itself is formulated to make decisions on a real investment frontier with the true cost, variable improvements, and time savings.

Discipline	Gender	Expected Olympic medals
Team Pursuit	M	0.9968
	F	0.4631
Team Sprint	M	0.6892

Table 5.4: Expected 2028 Olympic medal count for each discipline.

5.6 Discussion

In this chapter, we develop an optimization model that uses our models developed in Chapter 3 and 4 to select our investment recommendations. Considering only equipment investments as the source of performance improvement between the most recent World Championship and the 2027 Worlds, we output a solution that expects a 2028 Olympic medal count of 2.15. We have therefore built a model that selects equipment upgrades necessary to overcome the required time improvements needed for high expected Olympic performance. If AusCycling were to invest in all of our recommended equipment upgrades, this would be a lower bound on the true expected number of medals at the 2028 Olympics given the reality that performance improvements can be gained in other areas as well. Future use of this model may include an investment horizon spanning additional investment areas outside of equipment.

While the recommendations shown in this section come from synthetic data, the results enforce the feasibility of using this model for selecting investments. This is particularly true given that the synthetic data is of the same magnitude of the expected data available to AusCycling. Replacing the data used with AusCycling's true investment horizon will therefore give the true recommended investment package ahead of the 2028 Olympics.

Chapter 6

Conclusion

6.1 Discussion

The primary goal of this research is to determine the optimal equipment upgrades for AusCycling to invest in to maximize the expected medal count at the 2028 LA Olympic Games. We use a three-phased approach to determine investment recommendations for Team Pursuit and Team Sprint. The models developed within Phase I and II provide additional utility separate from their contribution to the investment recommendations.

In Phase I, we develop a CATBoost prediction model to estimate the probability of earning a medal, being close to earning a medal, and not being close to earning a medal at a future Olympics given participation in a prior competition. This model predicts outcomes for the eight disciplines described in Chapter 2 and is based on competition placement results for all competitions recorded in the UCI database. The overall accuracy of 96.3% is strengthened by high performance in each of the three Olympic results categories, across all eight disciplines, across gender, and across several well-performing cycling countries. This model may be used both for direct Olympic performance predictions and as an input into the final investment model.

In Phase II, we focus on time improvements required for Australian cyclists to achieve different placements in the 2027 World Championships. Due to the nature of solving for required time improvements and data availability, we focus only on Team Pursuit and Team Sprint and use World Championships as a measure of historical time improvement. We do not build a full formulation for the improvement in race time over calendar times in these disciplines, but instead we "zoom in" on a small number of years such that the time improvement relationship can be linearly approximated. We then model required improvement as the compound improvements of change in race time over calendar time and the percentage time difference between placements at World Championships. Combining this model with that developed in Phase I allows for the computation of total time improvement requirements necessary to achieve certain Olympic performances. This model may also be used as an input into the final investment model.

In Phase III, we develop the investment model using Mixed-Integer Nonlinear optimization techniques. The optimization makes use of potential equipment investments, their impact on physical variables that impact power needed and physiological power, and the impact of

those variables on race time. By approximating smooth functions for the models developed in Phase I and Phase II, we form a relationship between the equipment investment horizon and expected medal count across Team Pursuit and Team Sprint in 2028. Due to data availability, we generate results with synthetic data of the same magnitude as the data available to AusCycling. The output of this model generates the recommended equipment investments for maximizing Olympic performance.

Through all three phases, the work in this research achieves its objective of providing AusCycling with a model to efficiently invest in new equipment as a means for improving future Olympic medal probability.

6.2 Future Work

There are several directions where future work can expand on these methods and improve the robustness of the results. We can separate potential future work by phase.

6.2.1 Phase I

i. If race time data were to be made widely available, the model may benefit from the inclusion of the relative difference in the time of a cyclist and the cyclist that finished before them. This may help to clarify current ambiguities the model may face when making decisions solely on placement.

ii. Training the model on a rolling window of performance rather than on a specific competition result may help to capture how performance in a full cycling season impacts predicted Olympic performance. This may reduce sensitivity to outlier performances.

6.2.2 Phase II

i. Developing full formulation for race time improvement, expanding from the current linear approximation, may improve the accuracy of the estimations. This would also allow for estimations to be made over a larger time horizon. The current linearity assumption limits predictions to a small window.

ii. The availability of time data for Olympic performances would allow for the direct computation of improvement in Olympic race times over calendar time. This would eliminate the need for combining predictions with the Phase I model and instead directly relate to Olympic performance.

iii. The availability of consistent race time data for all eight disciplines considered in Phase I would allow for the computation of time requirements for a larger portfolio of disciplines. This would provide AusCycling with a more wholistic understanding across their portfolio of Olympic disciplines.

6.2.3 Phase III

i. The use of the true investment-related data available to AusCycling would allow for valid investment recommendations. Upon analysis of the true data and investment horizon, it may be useful to adjust the level of α used to quantify the sensitivity of the model to cost in the objective function.

ii. The current analysis only considers equipment investments. Future use of this model may include additional domains of investment to expand the use of the tool in decision-making.

iii. To avoid the need to smoothly approximate the Phase I and II models, the optimization model could be solved via a combinatorial optimization approach such as a knapsack problem. This would allow for the selection of discrete placements during solving and could therefore be paired directly with the models of Phase I and II.

Appendix A

Australia's sports ecosystem

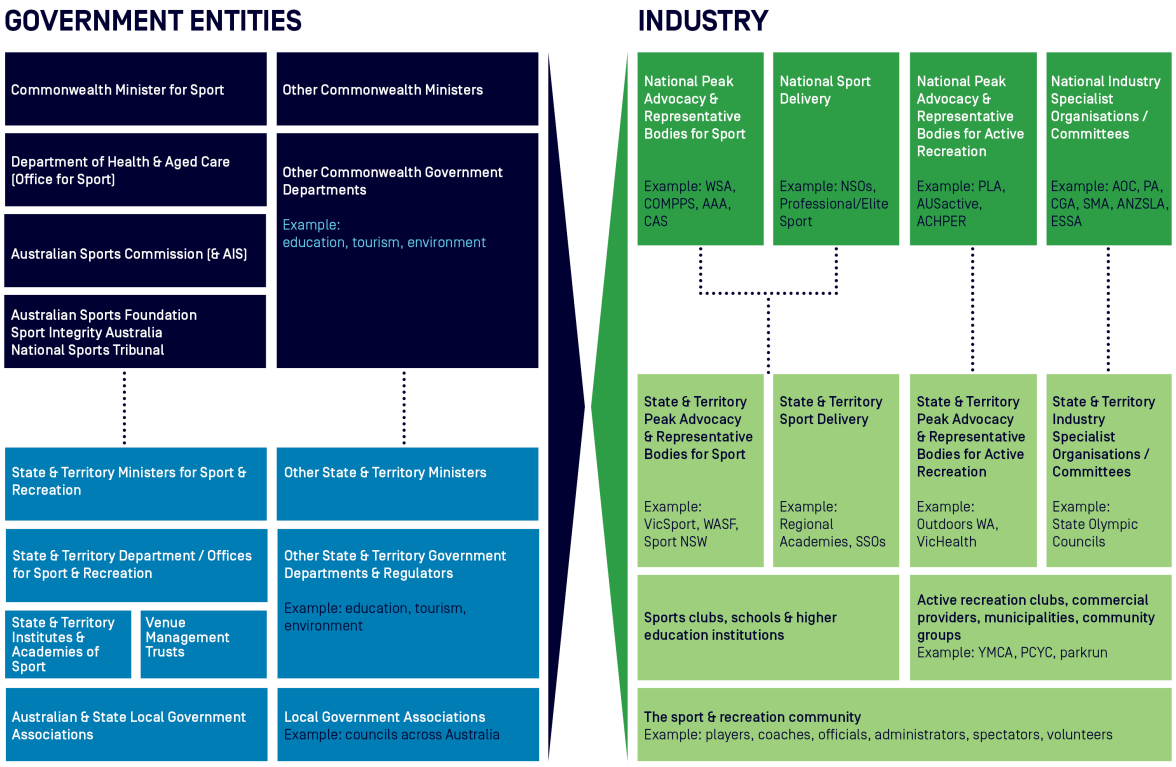
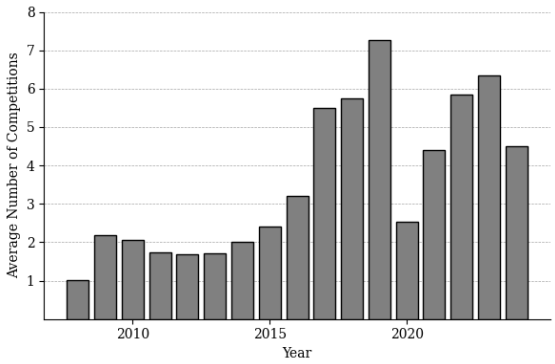


Figure A.1: System map of sport and recreational governance in Australia [7] .

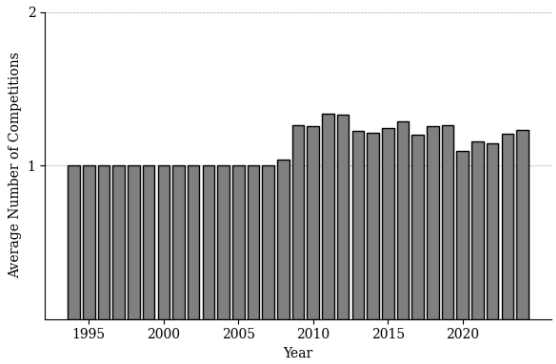
Appendix B

Distribution of competitions by discipline

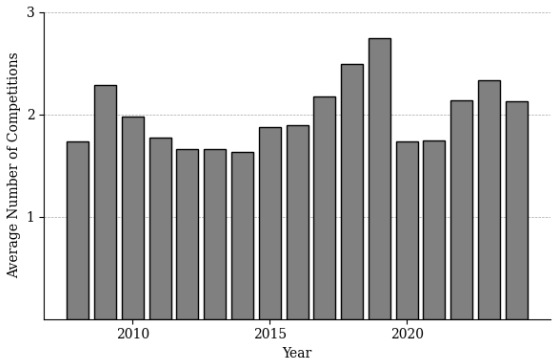
The following figures show a graph of the distribution of competitions per athlete per year for each discipline in the raw dataset.



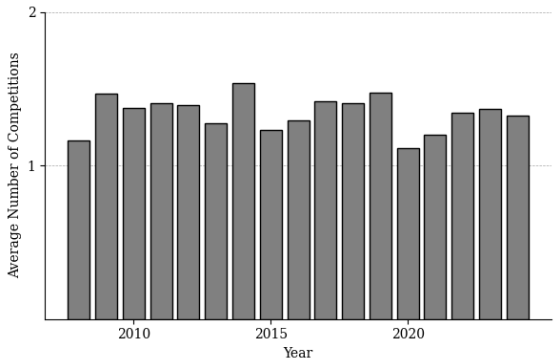
(a) BMX Racing



(b) Time Trial

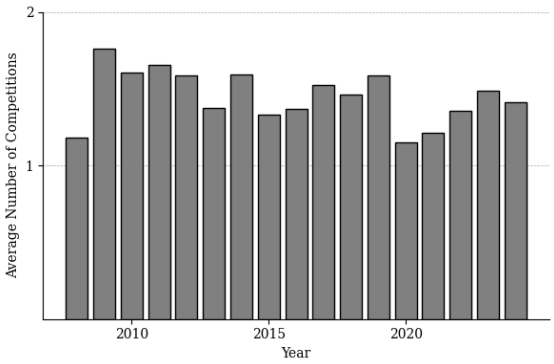


(c) Madison

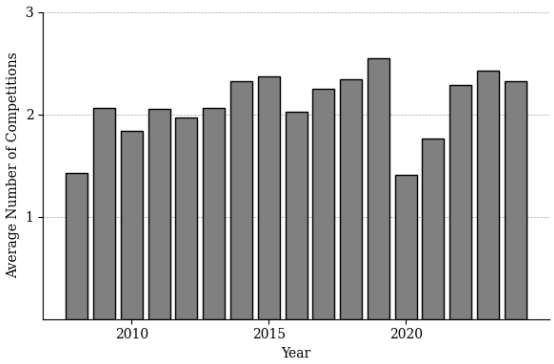


(d) Team Pursuit

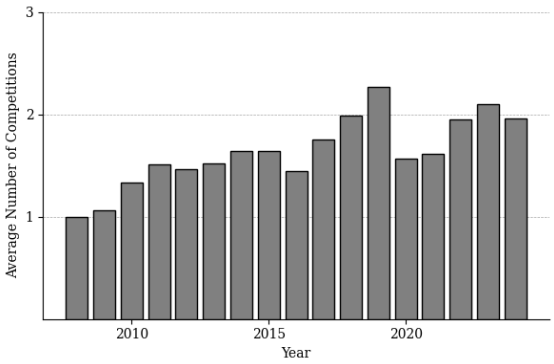
Figure B.1: Distribution of competitions per athlete per year (1/2).



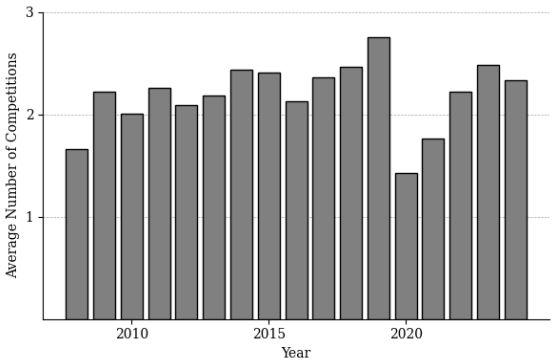
(a) Team Sprint



(b) Keirin



(c) Omnium



(d) Sprint

Figure B.2: Distribution of competitions per athlete per year (2/2).

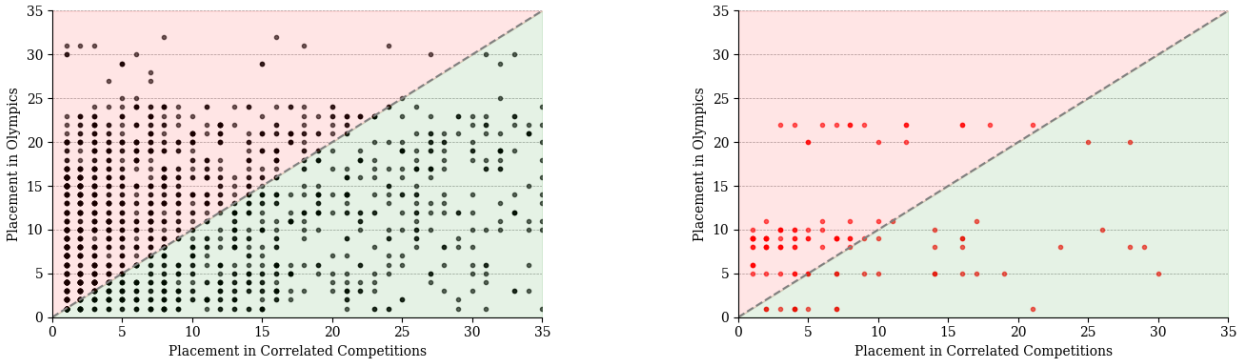
Appendix C

Correlation visualizations by discipline

The following figures show the relationship between placement in Olympic competitions and placement in competitions two years prior. We plot these results only for competitions that are positively correlated with the Olympics for all disciplines not displayed in the main paper. The red region indicates under-performance at the Olympics, the green region indicates over-performance at the Olympics, and the 45° line represents perfect parity between Olympic and prior competitions.

C.1 BMX Racing

The following figures graph the results for BMX Racing.



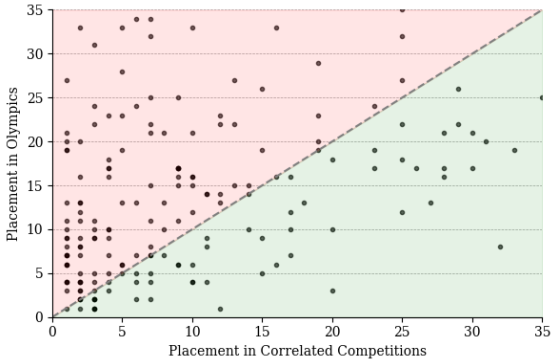
(a) All results 2 years before Olympics

(b) Australian results 2 years before Olympics

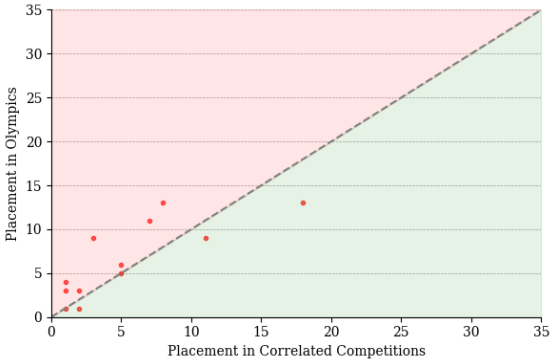
Figure C.1: Placement in BMX Racing events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

C.2 Time Trial

The following figures graph the results for Time Trial.



(a) All results 2 years before Olympics

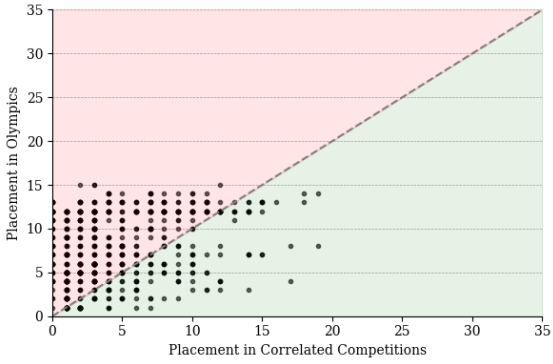


(b) Australian results 2 years before Olympics

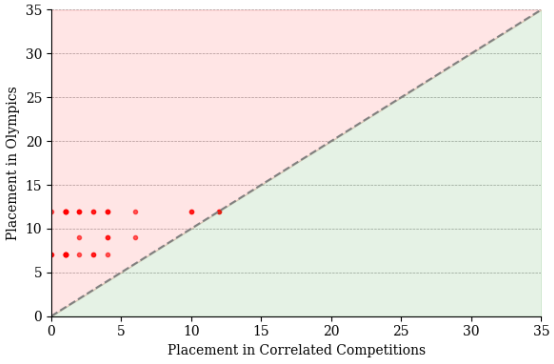
Figure C.2: Placement in Time Trial events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

C.3 Madison

The following figures graph the results for Madison.



(a) All results 2 years before Olympics

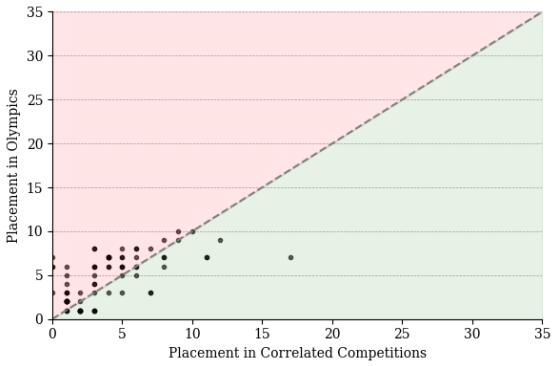


(b) Australian results 2 years before Olympics

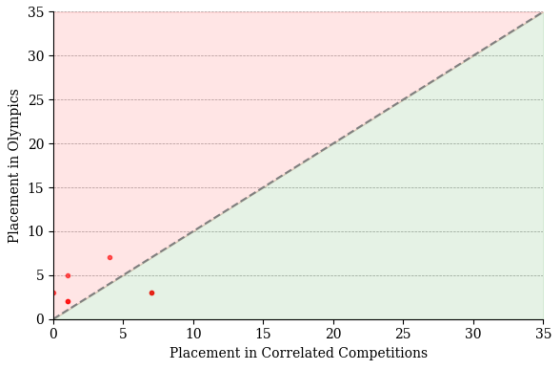
Figure C.3: Placement in Madison events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

C.4 Team Pursuit

The following figures graph the results for Team Pursuit.



(a) All results 2 years before Olympics

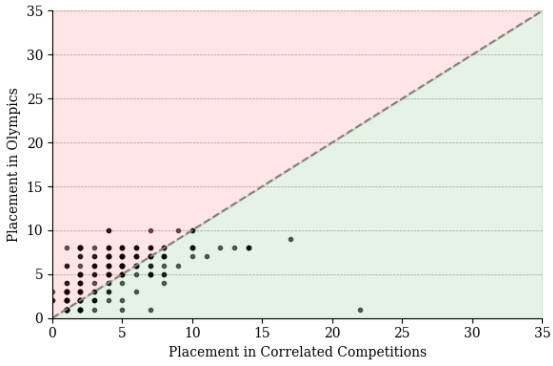


(b) Australian results 2 years before Olympics

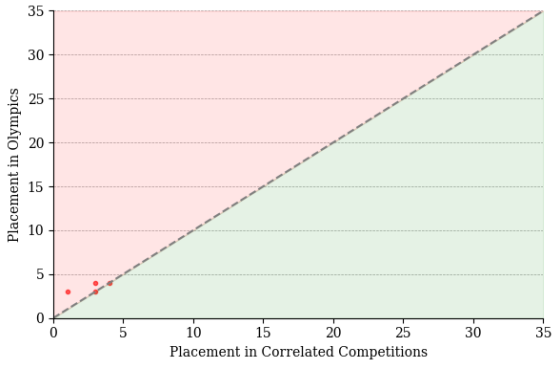
Figure C.4: Placement in Team Pursuit events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

C.5 Team Sprint

The following figures graph the results for Team Sprint.



(a) All results 2 years before Olympics

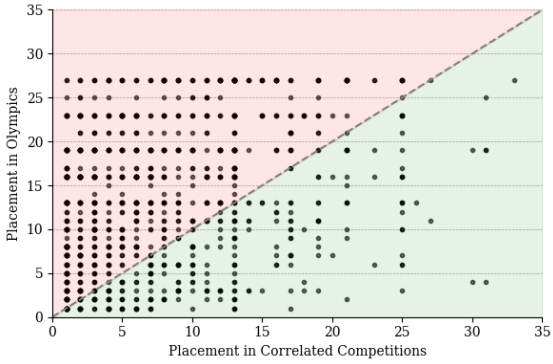


(b) Australian results 2 years before Olympics

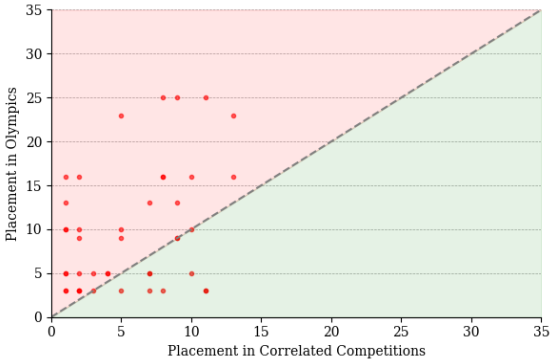
Figure C.5: Placement in Team Sprint events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

C.6 Keirin

The following figures graph the results for Keirin.



(a) All results 2 years before Olympics

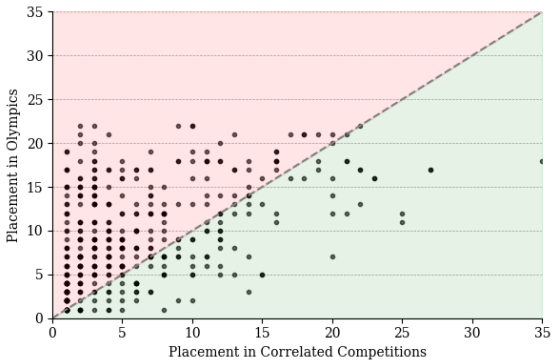


(b) Australian results 2 years before Olympics

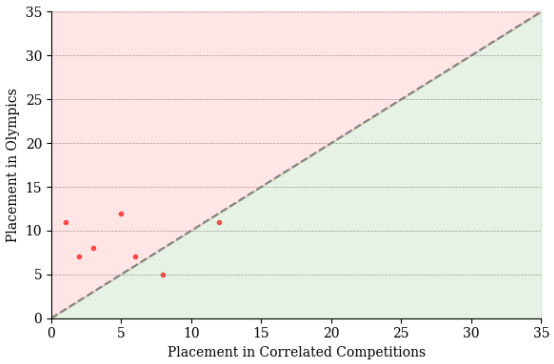
Figure C.6: Placement in Keirin events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

C.7 Omnium

The following figures graph the results for Omnium.



(a) All results 2 years before Olympics



(b) Australian results 2 years before Olympics

Figure C.7: Placement in Omnium events for correlated competitions and Olympic competitions for all competitions that have a positive correlation with the Olympics two years prior to the Olympics.

Appendix D

Natural time improvement coefficients

These tables represent the intercepts and slope for a linear regression fit to World Championship race result times for places 1-8 for men's Team Pursuit and Team Sprint and women's Team Pursuit.

D.1 Team Pursuit (men)

The following table shows the intercept and slope for a linear regression fit to men's Team Pursuit.

Place	Intercept	Slope
1	241.82	-0.0026
2	245.37	-0.0030
3	245.20	-0.0033
4	248.41	-0.0037
5	247.49	-0.0034
6	248.93	-0.0038
7	249.23	-0.0034
8	251.16	-0.0034

Table D.1: Regression results for Team Pursuit (men) race times in World Championships.

D.2 Team Sprint (men)

The following table shows the intercept and slope for a linear regression fit to men's Team Sprint.

Place	Intercept	Slope
1	44.247	-0.00043
2	44.618	-0.00032
3	44.436	-0.00033
4	44.430	-0.00024
5	44.332	-0.00028
6	44.413	-0.00023
7	44.510	-0.00017
8	44.679	-0.00019

Table D.2: Regression results for Team Sprint (men) race times in World Championships.

D.3 Team Pursuit (women)

The following table shows the intercept and slope for a linear regression fit to women's Team Pursuit.

Place	Intercept	Slope
1	263.72	-0.0046
2	265.87	-0.0044
3	267.48	-0.0037
4	271.96	-0.0046
5	272.28	-0.0045
6	275.04	-0.0047
7	274.83	-0.0031
8	277.54	-0.0034

Table D.3: Regression results for Team Pursuit (women) race times in World Championships.

Appendix E

Percentage distance between places

These results show graphs of the percent change in time required to advance from one place to the next-best place as well as a table of the average percent change required for men's Team Sprint and women's Team Pursuit.

E.1 Team Sprint (men)

The following figure and table display percentage change required to move from one place to the next-best over time (Figure E.1) and on average (Table E.1) for men's Team Sprint.

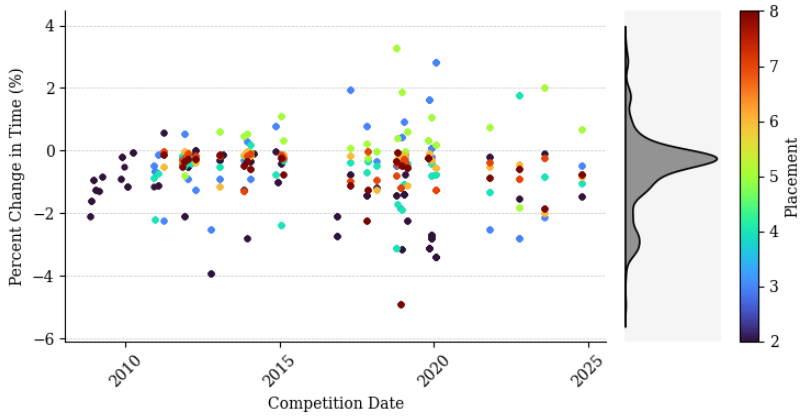


Figure E.1: Percent change in time from one place to the next-best place in men's Team Sprint. Each color indicates results for a different place in the range 2-8. The histogram displays the distribution of required change in time across all places in the World Championships dataset.

Place	Percent change required
2	-1.676
3	0.054
4	-0.706
5	0.425
6	-0.361
7	-0.553
8	-0.516

Table E.1: Average percent change in time needed to reach the next-best place.

E.2 Team Pursuit (women)

The following figure and table display percentage change required to move from one place to the next-best over time (Figure E.2) and on average (Table E.2) for women's Team Pursuit.

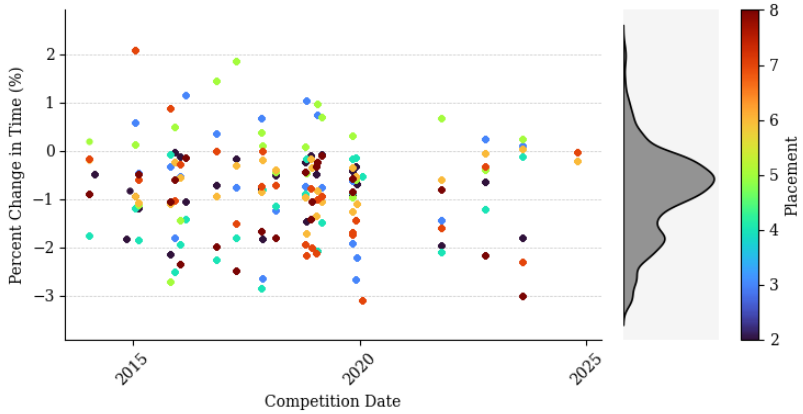


Figure E.2: Percent change in time from one place to the next-best place in women's Team Pursuit.Each color indicates results for a different place in the range 2-8. The histogram displays the distribution of required change in time across all places in the World Championships dataset.

Place	Percent change required
2	-0.797
3	-1.031
4	-0.979
5	-0.147
6	-0.755
7	-1.276
8	-1.078

Table E.2: Average percent change in time needed to reach the next-best place.

References

[1] *About Us | AusCycling | AusCycling*. en. URL: https://auscycling.org.au/about (visited on 04/18/2025).

[2] *AusCycling: Inside Australia's biggest sporting merger - AICD*. en-US. URL: https://www.aicd.com.au/good-governance/organisational-strategy/mergers-acquisitions-takeover-strategies/auscycling-inside-australias-biggest-sporting-merger.html (visited on 04/30/2025).

[3] A. S. C. j. o. A. c. S. Commission. *Australian Sport Policy*. English. URL: https://www.clearinghouseforsport.gov.au/kb/australian-sport-policy (visited on 04/18/2025).

[4] *Why does the Australian Government invest in sport ? One answer is national pride.* en-US. Aug. 2023. URL: https://australiansportreflections.com/2023/08/14/why-does-the-australian-government-invest-in-sport-one-answer-is-national-pride/ (visited on 05/02/2025).

[5] *2032 Strategy | HP2032*. URL: https://www.winwell2032.au/2032-strategy (visited on 05/02/2025).

[6] *Gold medal investment in our elite athletes | Prime Minister of Australia*. en. June 2024. URL: https://www.pm.gov.au/media/gold-medal-investment-our-elite-athletes (visited on 04/18/2025).

[7] *Funding for major sporting codes: a quick guide – Parliament of Australia*. URL: https://www.aph.gov.au/About_Parliament/Parliamentary_departments/Parliamentary_Library/Research/Quick_Guides/2024-25/Funding_for_major_sporting_codes_a_quick__guide (visited on 04/18/2025).

[8] *The UCI*. en-gb. URL: https://www.uci.org/the-uci/2QDMyzhoBXyWbVrsxge0EH (visited on 04/30/2025).

[9] *Olympic Qualifier Series 2024: The evolution of BMX Freestyle cycling*. en. URL: https://www.olympics.com/en/news/olympic-qualifier-series-2024-evolution-bmx-freestyle-cycling (visited on 05/02/2025).

[10] *Para cycling: A brief history of the sport*. en. URL: https://www.paralympic.org/news/para-cycling-brief-history-sport (visited on 05/02/2025).

[11] *Cycling BMX Racing: Olympic history, rules, latest updates and upcoming events for the Olympic sport*. en-us. URL: https://www.olympics.com/en/sports/cycling-bmx-racing/ (visited on 04/18/2025).

[12] *Cycling Road: Olympic history, rules, latest updates and upcoming events for the Olympic sport*. URL: https://www.olympics.com/en/sports/cycling-road/ (visited on 04/18/2025).

[13] *Olympic Road Cycling format: Overview, rules, competition info, road race, time trial, how it works | NBC Olympics*. en. URL: https://www.nbcolympics.com/news/cycling-101-road-cycling-olympic-rules-and-competition-format (visited on 04/18/2025).

[14] M. Mitchell. *Paris Olympics 2024 Track Cycling: Madison Explained*. en-GB. Aug. 2024. URL: https://procyclinguk.com/paris-olympics-2024-track-cycling-madison-explained/ (visited on 04/18/2025).

[15] "CYCLING: History of Cycling Track at the Olympic Games". en. In.

[16] *In pursuit of gold: How the U.S. rebuilt its track team - Velo*. URL: https://velo.outsideonline.com/road/road-racing/in-pursuit-of-gold-how-the-u-s-rebuilt-its-track-team/ (visited on 04/18/2025).

[17] *Olympic Track Cycling format: Overview, rules, competition info, track cycling events, omnium races and scoring, how it works | NBC Olympics*. en. URL: https://www.nbcolympics.com/news/cycling-101-track-cycling-olympic-rules-and-competition-format (visited on 04/18/2025).

[18] *Cycling at the Tokyo Olympic Games - keirin*. en. URL: https://www.britishcycling.org.uk/track/article/20160616-gb-cyclingteam-Cycling-at-the-Olympic-Games---keirin-0 (visited on 04/18/2025).

[19] L. W. l. updated. *Track cycling: What is the Omnium?* en. July 2021. URL: https://www.cyclingnews.com/features/tokyo-olympics-what-is-the-omnium/ (visited on 04/18/2025).

[20] R. W. published. *Track sprint: How the 2024 Paris track sprint works*. en. Aug. 2024. URL: https://www.cyclingweekly.com/news/track-sprint-how-the-tokyo-2020-olympic-track-sprint-works (visited on 04/18/2025).

[21] *Emma Bianchi's Guide to Olympic Track Cycling*. en. Aug. 2024. URL: https://writebikerepeat.com/emma-bianchis-guide-to-olympic-track-cycling/ (visited on 04/29/2025).

[22] S. B. Nadkarni, G. S. Vijay, and R. C. Kamath. "Comparative Study of Random Forest and Gradient Boosting Algorithms to Predict Airfoil Self-Noise". en. In: *Engineering Proceedings* 59.1 (2023). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 24. ISSN: 2673-4591. DOI: 10.3390/engproc2023059024. URL: https://www.mdpi.com/2673-4591/59/1/24 (visited on 05/02/2025).

[23] M. Imani, A. Beikmohammadi, and H. R. Arabnia. "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels". en. In: *Technologies* 13.3 (Mar. 2025). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 88. ISSN: 2227-7080. DOI: 10.3390/technologies13030088. URL: https://www.mdpi.com/2227-7080/13/3/88 (visited on 05/02/2025).

[24] *How to qualify for track cycling at Paris 2024. The Olympics qualification system explained*. en. URL: https://www.olympics.com/en/news/pathway-to-paris-2024-track-cycling-qualification-system-explained (visited on 04/23/2025).

[25] *Track Cycling Guides - Velodrome from Velodrome Shop*. URL: https://www.velodrome.shop/velodrome?srsltid=AfmBOoo22OcmeDRvM6FjHrAlBOd4341ZCceiyk32R9Mc-2M39_bNGoGu (visited on 04/28/2025).

[26] *Track Cycling 101: What to Know*. en-US. URL: https://www.bikeexchange.com/blog/track-cycling-guide-everything-to-know (visited on 04/29/2025).

[27] P. Leo, J. Spragg, T. Podlogar, J. S. Lawley, and I. Mujika. "Power profiling and the power-duration relationship in cycling: a narrative review". en. In: *European Journal of Applied Physiology* 122.2 (Feb. 2022), pp. 301–316. ISSN: 1439-6327. DOI: 10.1007/s00421-021-04833-y. URL: https://doi.org/10.1007/s00421-021-04833-y (visited on 05/02/2025).

[28] J. Thornham. *Understanding Rolling Resistance & Impedance For Cyclists - FLO Cycling*. en-US. Section: Aero Wheel Science. Feb. 2020. URL: https://blog.flocycling.com/aero-wheels/understanding-rolling-resistance-impedance-for-cyclists/ (visited on 05/02/2025).

[29] *The Science of Speed: Aerodynamic Drag in Cycling*. en. URL: https://aerosensor.tech/pages/the-science-of-speed-aerodynamic-drag-in-cycling (visited on 05/02/2025).

[30] S. H. Faulkner, P. Jobling, K. E. Griggs, and P. Siegkas. "Individual aerodynamic and physiological data are critical to optimise cycling time trial performance: one size does not fit all". en. In: *Sports Engineering* 27.1 (Jan. 2024), p. 4. ISSN: 1460-2687. DOI: 10.1007/s12283-023-00446-0. URL: https://doi.org/10.1007/s12283-023-00446-0 (visited on 05/02/2025).

[31] L. Moseley and A. E. Jeukendrup. "The reliability of cycling efficiency:" en. In: *Medicine and Science in Sports and Exercise* (Apr. 2001), pp. 621–627. ISSN: 0195-9131. DOI: 10.1097/00005768-200104000-00017. URL: http://journals.lww.com/00005768-200104000-00017 (visited on 05/02/2025).

[32] c. club creative. *Mechanical efficiency*. en. Dec. 2023. URL: https://www.c2cyclingcoach.com/eficiencia-mecanica/ (visited on 05/02/2025).

[33] A. K. Dunst, C. Hesse, O. Ueberschär, and H.-C. Holmberg. "A Novel Approach to the Determination of Time- and Fatigue-Dependent Efficiency during Maximal Cycling Sprints". en. In: *Sports* 11.2 (Feb. 2023). Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, p. 29. ISSN: 2075-4663. DOI: 10.3390/sports11020029. URL: https://www.mdpi.com/2075-4663/11/2/29 (visited on 05/02/2025).

[34] R. A. Lukes, S. B. Chin, and S. J. Haake. "The understanding and development of cycling aerodynamics". en. In: *Sports Engineering* 8.2 (Dec. 2005), pp. 59–74. ISSN: 1460-2687. DOI: 10.1007/BF02844004. URL: https://doi.org/10.1007/BF02844004 (visited on 05/02/2025).

[35] L. Underwood. "Aerodynamics of Track Cycling". en. In.

[36] *Revealed / Japan's €126k Olympic track bike and the spiralling cost of equipment for Paris 2024*. en. July 2024. URL: https://www.bikeradar.com/news/2024-olympic-track-bike-prices (visited on 05/02/2025).