An Empirical Study of Race Strategy in College Swimming

Master of Quantitative Economics at the University of California, Los Angeles

Hayden Johnson

Faculty Advisor: Randall Rojas

8th of December 2023

Table of Contents

1. Abstract

This paper offers an empirical analysis of how college swimmers swim the 500-yard freestyle. By Web Scraping an online database, the times and splits for each swimmer are carefully analyzed through statistical analysis. The analysis comprises of charts and tables of sections and subsections of individuals, as well as various regression and machine learning techniques. This paper identifies that the most elite swimmers and the swimmers who improve upon their overall fastest time in a given year race the 500-yard freestyle at a consistent even pace. The less elite swimmers swim their races with larger variations between their splits.

2. Introduction

The four most watched sports in the US–football, basketball, baseball and hockey–struggle to define who the best player, the best team or the best coach truly are. These sports have multiple dimensions to define how good a player, team or coach is; however, a more one-dimensional sport, one that is entirely based on time, better corroborates who the best truly is. Swimming encompasses a single individual in their own lane using every inch of their power in both body and mind to complete their race the fastest. Upon completion of the race, the only thing to show for one's efforts is their time. While this time element forges a steadfast conclusion in deciphering the best from the worst, there are many ways to swim a race. An individual may choose to swim as fast and as hard as they can from the very beginning in hopes of surviving to the very end when they have given so much in the first leg. Others choose to swim at a very even pace, averaging their speed and effort throughout the entire race. One can elect to swim on top of the water for as long as possible, but recent developments in swimming–the past 15 years–have proven that the fastest way to complete a race is under the water. While a swimmer may be defined by their final time, how that time is reached can be very different from swimmer to swimmer and every hundredth counts.

In highly competitive swimming, where athletes have already pushed the limits of what is possible, something as trivial as dropping a tenth of a second is a monumental task. Many swimmers go months or years without seeing their best times achieved, and being a single hundredth faster is celebrated. A hundredth could also be the difference between winning and losing. College swimming houses many of the nation's top athletes, and since swimming is not lucrative nor does it have a large pro scene, many swimmers end their careers after college. This paper utilizes the data from college swimming and its top performers to shine a light on how the elite swimmers of our time swim their race compared to the average swimmer, and how an individual may swim their race differently to improve their time.

3. Background Information of College Swimming

25 yards establishes the length of a U.S. college pool with races ranging from 50 yards to 1,650 yards. For the purposes of this paper, the 500-yard freestyle is the focus. Within each race, splits are taken. A split can be any increment of a race, including one lap, two laps, four laps, eight laps and ten laps, and refers to the time it takes a swimmer to complete the chosen increment. This paper analyzes splits in increments of two laps, or 50 yards, and is referred to both as a "split" and a "50." Accordingly, for the 500-yard freestyle, there are 10 splits (or 10 50s). 10 splits offer a plethora of data of how a swimmer approached a race. Did the swimmer go out fast and spend all their energy on the first third of the race only to see themselves completely out of endurance for the last third? Perhaps the swimmer waited until the last third of their race to expend all their energy to finish. The splits of a race tell the story of the approach for each swimmer. Gathering these splits comes from a website called Swimcloud.com which houses every college swim meet result in the US for free. Every split for each swimmer is recorded and is easily accessed on the site. Overseeing each meet in college swimming is the National Collegiate Sports Association, NCAA for short. Three divisions make up the NCAA–Division I (DI), Division II (DII) and Division III (DIII)–with the Division I athletes generally having the fastest times followed by Division II and finally Division III. The

Men's NCAA Championship swim meets, the swim meets where the fastest swimmers from each division compete at, is analyzed in this paper to compare how the top swimmers in each division swim their race.

4. Theory and Literature Review

Theory of how to swim a race is not empirically studied in the swimming world. Many swimmers have their own varying opinions on how they should swim their race, as well as opinions from each swimmer's coach. For example, the 200-yard freestyle, butterfly, backstroke, or breaststroke is a race consisting of four 50-yard lengths totaling 8 laps in the pool. A majority of coaches will say the third 50 (laps five and six) is the most important 50 of that race. Swimmers often relax their efforts on this 50 as they prepare to give an all-out effort on the final 50 of the race; however, relaxing too much will put a swimmer too far behind in the final 50 and they cannot catch back up. Therefore, setting yourself up well on the third 50, racing hard despite being out of endurance on the final two laps of the race, allows an individual to be in a good position to race others around them, motivating them to swim harder. A similar thought process among coaches for the 500-yard freestyle is common. The sixth and seventh 50s of the race (laps 11 through 14) are considered crucial to the end of the 500. Holding back on these two 50s is often thought to leave swimmers too far behind to catch up because either their physical abilities are not capable of closing the race, or the swimmer gives up mentally due to being too far behind. This may be a general consensus of how these races should be approached, but every swimmer is an individual, requiring their own strategy for each race. Bjorn Seeliger of the University of California-Berkeley's Golden Bear swim team is one of the NCAA's premier sprinters. Specializing in the 50 and 100-yard freestyle, the 500-yard freestyle offers itself as a large endurance challenge. Nonetheless, Bjorn made his 500-yard freestyle debut in 2022 at a meet against the University of the Pacific. He swam the race as sprinters often do all out for as long as possible. His 100-yard (4 laps) splits were as follows: 50.17, 56.12, 59.36, 59.98 and 53.5, with a final time of 4:39.19.[1] These splits are not exactly consistent and there is a significant difference between the first and second splits, the second and third splits, and the fourth and fifth splits. Many swimmers would have trouble finding success with such a variance in their splits. The Golden Bear's Jack Meehan, a swimmer who focuses more specifically on the 500-yard freestyle, swam the race in a way that most endurance athletes do with the 100-yard splits being relatively consistent: 52.13, 54.59, 54.98, 54.43 and 51.89, with a final time of 4:28.02.1 The last 100-yard split, which was faster than the first but within about 2.5 seconds of the middle three splits, would suggest that Meehan saved a little too much energy for the end. Nevertheless, Meehan swam the entire race significantly faster than Bjorn. While it may not be a fair comparison to pit a 500-yard specialist against a non-specialist, both approached the race differently according to the customary methods of racing associated with their specialties.

Studying the approach for a race has not been empirically explored through data, but the preparation for a meet has. All swimmers do not expect to get a best time in every race. There are usually two to three meets a year that a swimmer focuses on to race faster than they ever have before. All other meets are just for practice. When one of the major meets comes around, swimmers enter a transitory period known as taper. Taper lessens the volume of yardage a swimmer may do at practice and increases the recovery time.

[1] Keith; https://swimswam.com/what-happens-when-you-put-an-18-2-sprinter-in-the-500-free/

Muscular fibers need to be repaired after constant use and stress from training arduously every day. In addition, the VO_2 max of an athlete increases when the workload decreases.[2] The VO_2 max is the maximal oxygen uptake your body can utilize while enduring intense exercise, and increasing this measurement allows for the swimmer to perform at high speeds for longer because the muscles in the body are receiving more oxygen than usual.[2] Recovery time during taper also allows the anaerobic threshold to increase.[2] Purging the body of lactic acid is important for muscle recovery and preventing soreness. Breaking down more lactic acid with a higher anaerobic threshold gives the swimmer faster recovery between swims to compete at a high level between the preliminaries and finals. A body prepared to perform at the highest level is needed in order to seek improvement in one's own time.

The majority of analytics for swimming focuses on stroke and turn efficiency. The 500-yard freestyle comprises of 20 laps and a total of 19 transitions from one lap to another, called turns. If a swimmer can improve their turn technique to be 0.10 seconds faster, a 1.90 second drop in overall time can be achieved– a great victory for any swimmer. A University of Virginia Mathematics Professor, Then Ono, puts trackers on swimmers to study their stroke technique and turns; he then tracks and corrects the inefficiencies in a swimmer's stroke, turn or dive.[3] His approach has been successful, seeing time improvement as great as 6 seconds.[3] Ono's analysis is helping swimmers to be better, faster and more efficient. This paper aims to do the same with the analysis of splits.

5. Data and Analysis

The dataset analyzed comes from Swimcloud.com–a website that records every race and split from every college meet in the US. The NCAA Championship meets for each division are recorded on the website from years past. Web Scraping is used to gather the information needed for each swimmer from the Division I, II and III Championships for the 2018-2019, 2021-2022 and 2022-2023 seasons. The 2019-2020 and 2020-2021 seasons are not chosen due to the COVID-19 Pandemic; moreover, the NCAA Championship was not held in 2019-2020 and many teams did not compete in the 2020-2021 season. In total, 140 swimmers from Division I, 105 from Division II and 109 from Division III are Web Scraped. The Web Scraping tool gives access to many elements: the swimmer's name, school, time improvement from their season best as a percentage, if the swim was a personal best or season best and the place they finished. The races analyzed are selected from the preliminaries (referred to in shorthand as "prelims," which is a portion of the meet where all competitors regardless of number can compete) of the 500-yard freestyle. The top 16 times from the preliminaries swim again at night in the A final–top eight in prelims–or the B final, where 9^{th} through 16^{th} still score but with significantly fewer points based on final position in the race. Results from finals are not chosen to maintain consistent conditions in both setting and swimmers. A full faith effort in the preliminary event is to be expected, for each swimmer attempts to finish prelims with at least the eighth fastest time. The top eight score the most points in a race regardless of how they finish in the final. The splits for each swimmer are gathered by hand. Web Scraping the splits proves to be much too

[2] Mortenson; https://www.swimmingworldmagazine.com/news/what-happens-to-your-body-during-taper/
[3] Hausman; https://www.npr.org/2022/03/12/1085542427/uva-professor-swimmer-math-faster.

difficult. An excel document containing the splits of every swimmer is created and the splits are verified as the accurate splits for each swimmer.

4.2 Summary Statistics

Gathered and ready to be used, the data is imported into Jupyter Notebooks to have an analysis done in Python. Each of the three NCAA divisions and their respective years are ported in via Web Scraping and are saved as a DataFrame. The excel files for each of the divisions and their respective years are imported as a DataFrame too. Combining the two together, the dataset is prepared for analysis. The divisions are analyzed by year and then as a whole. The "describe" method offers insight into the minimum, maximum, standard deviation, mean and quartiles for the splits of each swimmer. This method gives a general overview of how fast a certain split is swam. To better visualize the quartiles, box plots are generated. The box plot shows the minimum and maximum as well as the 25^{th} percentile, 50^{th} percentile, 75^{th} percentile and any outliers that it sees fit. To better understand the distribution, a violin plot is used. Similar to a box plot, the violin plot shows the interquartile range of the 50 splits, as well as the kernel density. The kernel density represents the data distribution at different points in the interquartile range. While the "describe" method is insightful to see the range of values in a column, it does not keep rows together. For example, the fastest swimmer in prelims may have the minimum overall time as well as some of the fastest 50 splits, but some other swimmers may have swum a particular 50 faster. When the "describe" method is used, the fastest swimmer overall will not appear as the minimum split for every 50. To keep rows intact, a function is made to show the interquartile range based on overall time, displaying the 50 splits for all swimmers. This function displays how the fastest swimmer swims their race compared to the 25^{th}, 50^{th}, 75^{th} percentiles and the slowest swimmer. Additionally, the differences between 50s can be looked at by subtracting the 2^{nd} 50 from the 1^{st} and the 3^{rd} 50 by the 2^{nd} and so on. One can see how much time is added or subtracted between 50s. This is done by creating a differencing function.

Looking at the splits is insightful, but an easier method to quickly determine how a swimmer swims their race needed to be developed. Two new metrics are created. The first metric adds up the 50 splits for three different sections: the first third of the race (50s 2-4), the second third of the race (50s 5-7), and the last third of the race (50s 8-10). The first 50 is ignored because each swimmer starts outside the water before diving into the water with a height and momentum advantage not available to the other 50s. The first 50 will always be an outlier to other 50s as the dive can provide a two second or more advantage over the other 50s. The fastest third is recorded in a new column of the DataFrame. This new column–titled "Fastest"–signifies which third is the fastest. Values inside "Fastest" are represented by a string of either: "First Third", "Second Third" or "Last Third". Playing off this idea, another metric is created to show how a complete race is swum. If a swimmer swims the first third of the race the fastest followed by the last third and finally the second third, a new string will appear in a new column titled "swimming order". The new value inputted into "swimming order" will be "fls", an acronym for first third, last third and second third. With this idea in mind, a swimmer who swims the second third the fastest followed by the first third and finally the last third develops an acronym of "sfl". These metrics are applied to the different divisions and their respective years, the divisions as a whole, the top 16 in each division and year, the swimmers slower than the top 16 for each division and year, swimmers who improved on their season best, and finally swimmers

who did not improve on their season best. Recording the counts and percentages for each category in the "Fastest" and "swimming order" columns is done by tables. Box plots help to visualize these tables as well.

5.3 Regression

Analyzing the approach elite swimmers take in their races provides valuable insights. However, understanding the strategies that contribute to time improvements can offer a competitive advantage to any swimmer. Regression analysis serves as a powerful tool to assess the significance of these strategies and their impact on performance. Moreover, regression analysis gives a tool for predicting what improvements can be plausibly seen if one were to change the strategy of their race. The dataset from Swimcloud provides the time in full of each swimmer and the improvement from each swimmer's season best as a percentage. Using improvement as the dependent variable, dummy variables of the divisions and the fastest third act as the independent variables. If any of the independent variables appear to be statistically significant–meaning the occurrence is not random–then a prediction can be made about improvement. The 50 splits can also be independent variables, but they suffer from multicollinearity. As each 50 depends on the previous one, the impact of each independent variable diminishes. One cannot decipher how much an independent variable affects the dependent variable as all the independent variables are correlated with each other. There are solutions to minimizing this adverse effect. Lasso, Ridge, or Elastic-Net regressions identify and eliminate wasteful independent variables or limit the effect large erroneous coefficients may have. Lasso aims to prevent overfitting and promote feature selection. The feature selection encourages some coefficients to become exactly zero, effectively eliminating certain variables from the model. A simpler model can be formed from Lasso as wasteful independent variables are purged. Ridge regression addresses multicollinearity and overfitting by introducing a penalty term. Unlike Lasso, Ridge does not force coefficients to exactly zero but rather shrinks them, especially for highly correlated variables. Ridge is beneficial when dealing with datasets exhibiting multicollinearity, where some predictors are highly correlated. Elastic-Net combines the two regularization techniques to reap benefits from both parties. These machine learning techniques may prove to be most useful in purging unwanted variables. Heteroskedasticity produces another issue. The variability of the independent variables may not be constant. An accurate reading of the effect an independent variable has on the dependent variable may not be recorded if heteroskedasticity is present. Transformation of the dependent variable can mitigate the issue. Lastly, omitted variable bias is present in the regressions. Knowing a swimmer's height, weight, age, stroke count per lap, etc. would make the regression analysis much stronger. These data points are unavailable. One more tool to help identify the best independent variables is generating a Random Forest.

A Random Forest is a machine learning algorithm that helps to identify features that are important for prediction. In other words the Random Forest will provide another way of identifying the most influential independent variables like Lasso, Ridge and Elastic-Net regressions. By utilizing bootstrapping, the Random Forest will make several sample datasets by randomly selecting the existing data points with replacement. Decision trees will be made to help make a prediction on these samples, and by aggregating each decision that was made, a Random Forest is constructed. The independent variables will be ranked from most to least important for making a prediction. The Random Forest adds another way of looking at the most important variables.

5. Results

The summary statistics reveal an unsurprising result for those who are knowledgeable about swimming: swimmers across all divisions each year ascend their splits and then make the final 50 extremely fast. Disregarding the first two laps–swimmers get to dive from outside the water on this 50 allowing for a much faster split than any of the others–the swimmers' 50 splits get slower by a few tenths or hundredths on average as the race progresses. The final 50 yielded a significant drop in time from the previous 50–on average 0.5 seconds to 1.0 seconds faster–revealing the awareness swimmers have that the end of the race is near. Swimmers should give everything they have left in the final laps of the race to place better or get a best time. From visual experiences at swim meets, this hypothesis is common among swimmers and spectators alike. The data corroborates this observation.

Unusual activity in the box plots is not present in the data. There are a few outliers, some swimmers had extremely slow times compared to most others, but the box plots confirm the information provided in the summary statistics. The violin plots, however, forthwith a peculiar phenomenon. Looking at the 50 splits across all divisions and years, a majority of the Division III 50 splits are normally distributed. This finding suggests Division III athletes are more evenly distributed around the mean i.e., Division III student-athletes have a relatively consistent level of performance across the board:

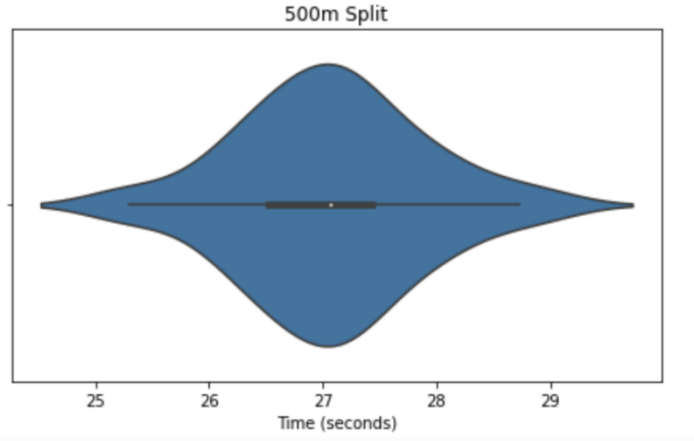

Fig. 1 the violin plot for the Division III (2022-2023) NCAA Championships. This shows the distribution for the 10th and final 50 of the 500-yard freestyle which is close to being normally distributed.

However, Division III swimmers are not entirely homogeneous in their spread. The violin plots can be left skewed, indicating there are some swimmers that outperform their competitors. Division I and Division II exhibit bimodal distributions. Many of the 50 splits for these two divisions show two peaks in the violin plots. Subpopulations among the divisions can be generated from this information. Division I athletes have a strong mix of left and right skewness as well–some athletes vastly outperformed or underperformed in comparison with others. Division II indicates a right skewness in the violin plots. Some swimmers in

Division II underperformed. The summary statistics and violin plots do not show any significant differences for a specific division across the years.

6.2 Fastest Thirds Analysis

The results moving forward encompass the divisions as a whole. The '18-'19, '21-'22 and '22-'23 results for the Division I NCAA Championship are combined into a single DataFrame known as "Division I". Other divisions follow suit so that the "Fastest" column can be inspected. Aforementioned in the Data and Analysis section, the "Fastest" column provides a new metric that adds up the 2nd-4th 50 splits, the 5th-7th 50 splits and 8th-10th 50 splits. The value outputted is stored in the "Fastest" column. These values give a better understanding of how hard a swimmer will push different sections of the race. The figure below shows the distributions for each division:

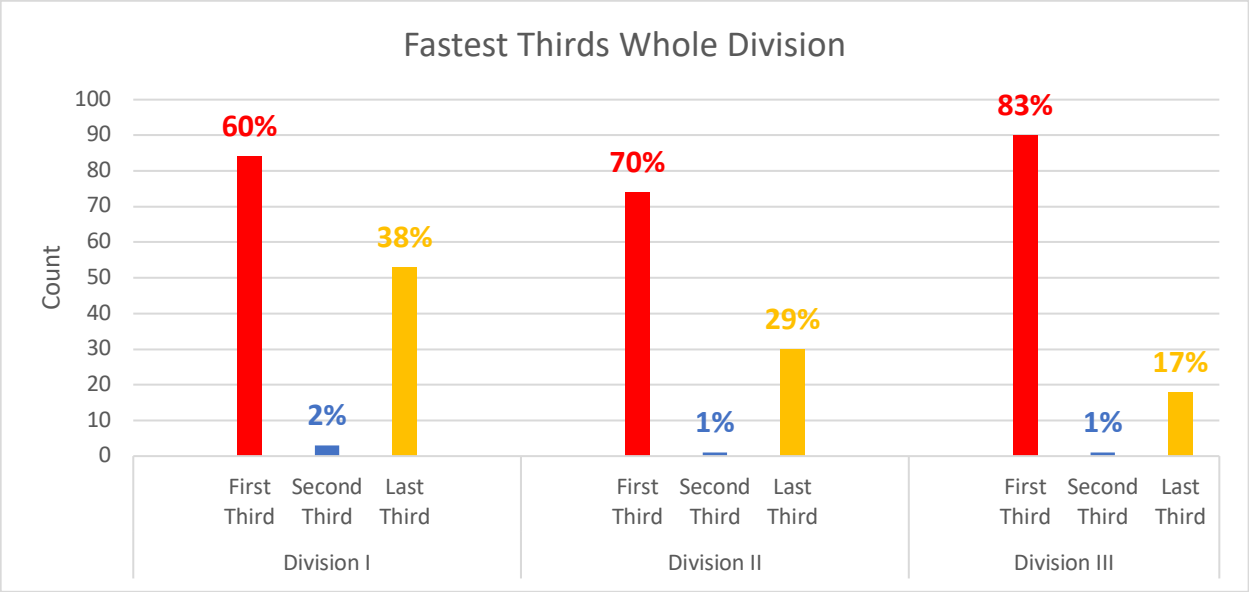

Fig. 2 shows the total counts for each third of the race that is fastest; furthermore, the percentages show the distribution of thirds for a given division.

The first third–the red bars in figure 2–is clearly dominant in all three divisions. Athletes have the most amount of energy to expend in the beginning of the race. Abundant stores of energy may be one reason why the first third appears dominant. The previously mentioned summary statistics also show that swimmers ascend their splits–they get slower as the race goes on. The final third of the race is the second most common. The final two laps of the race inflict a large time decrease compared to the other 50s and is usually one of the fastest 50s disregarding the first two laps of the race. The athletes are so close to the end they want to expel all their energy and might to finish the race. This phenomenon helps to explain why the final third is the second most common. Hardly any swimmer swims the second third the fastest. In the middle of the race, one can imagine it is hard to justify purging your body of energy when there is still so much swimming to be completed. Looking at the percentages, a stark difference between the divisions can

be witnessed. 82.57% of Division III swimmers take the first third of their race out the fastest compared to 70.48% of DII and 60.00% of DI. These results show that all swimmers across the divisions–especially Division III swimmers–find pushing the beginning of their races to be most valuable. Making the first third the fastest ensures all of one's energy is outputted into the race. Sometimes sprinting the last third of the race will result in improper use of energy. An athlete may not have used every inch of stamina and power they could muster, leaving even more time to be dropped in the future. Looking at the last third percentages, it could be that DI athletes are waiting until the end of the race to put forth their best effort. They could also, however, be more consistent in their splits. Rather than relying on one third of their race to be the best, a consistent swim may be the fastest. Swimming all the 50s at the same pace until the final 50–which experiences a large drop in time–would result in the final third of the race to be the fastest. A similar analysis of DII and DIII swimmers can be made. One way to see which hypothesis is correct regarding how the competitors swim their races is to look at the difference in median time from 50 to 50.

Differencing the 50 splits one after the other will show how consistent each division is at swimming the 500-yard freestyle. A more consistent swim will see a difference close to zero e.g., the third 50 is 26.71 and the fourth 50 is 26.73 which is a difference of 0.02 seconds. In this example the swimmer's fourth 50 is barely slower than their third making them a consistent swimmer. The table below outlines the median difference between 50s for each division:

Division	2 (100 Yards)	3 (150 Yards)	4 (200 Yards)	5 (250 Yards)	6 (300 Yards)	7 (350 Yards)	8 (400 Yards)	9 (450 Yards)	10 (500 Yards)
I	1.99	0.33	0.18	0.08	0.08	0.06	0.06	-0.01	-0.56
II	2.17	0.46	0.29	0.09	0.10	0.12	0.13	-0.05	-0.82
III	2.23	0.52	0.28	0.18	0.09	0.06	0.14	-0.06	-0.57

Table 1 displays the difference in median splits across the division for each 50 split in the 500-yard freestyle.

Table 1's columns present an interesting finding. Splits 2-4 have the largest difference in time which corroborates the data beforehand, the first third of the race is the fastest third for most swimmers. After 200 yards of competition, the splits continue to increase at a decreasing rate until the 450-yard mark. The increase in time suggests that swimmers are slowing down throughout the race until the end. The rate at which the splits increase, however, is much lower for Division I. DI athletes have less variation in their splits than the other two divisions. Going from 50 split to 50 split, DI swimmers are approaching their races with a different mindset: expending their effort evenly throughout the race. DIII athletes appear to slow down the most going into the middle section of their races due to the higher difference in splits. Division II

and Division III swimmers may find they can perform more closely to DI swimmers if they try to carry their speed with them throughout the whole race rather than extricating too much effort in the beginning, forcing them to drag their feet across the finish line. Table 1 shows how the divisions as a whole swim the 500. It is possible that the fastest in each division–the top 16 fastest times for each prelim swim–may have a different style.

The NCAA Championships, as with most swim meets, have preliminary swims in the morning and final swims at night. The top 16 swimmers in prelims get to compete at night in the finals. The top 16 are separated into the A final–finishers 1st through 8th in the morning–and the B final–finishers 9th through 16th in the morning. Making this elite group of 16 swimmers defines the fastest swimmers in the nation in their respective divisions. Applying the same technique to see which third is the fastest for the divisions as a whole yields the following results for the top 16:

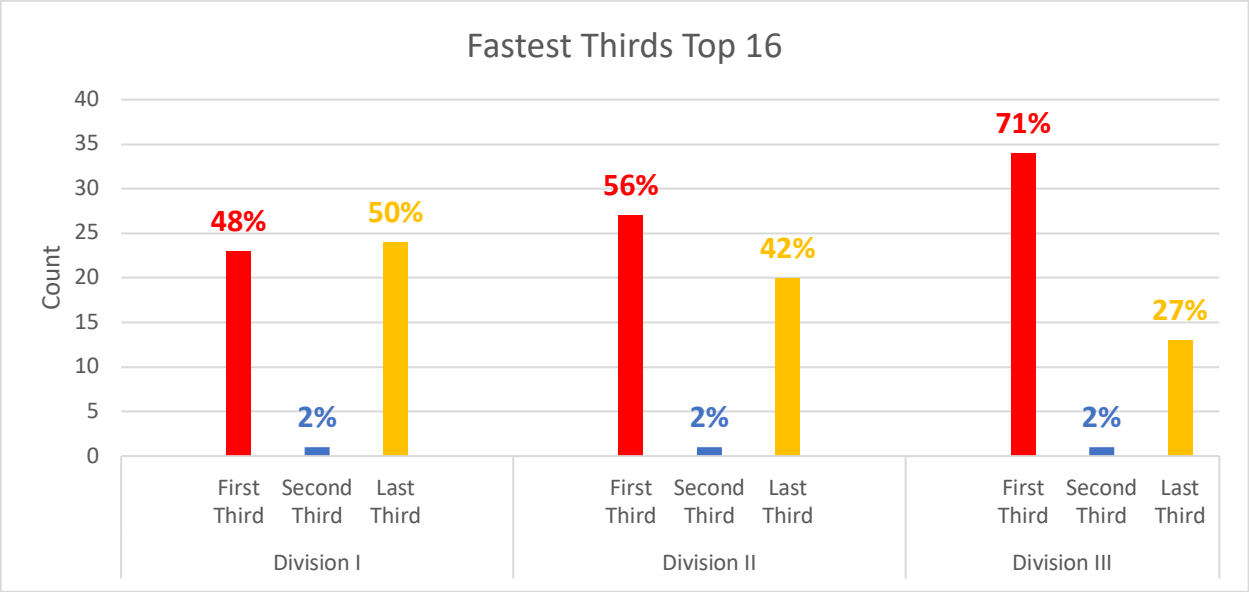

Fig. 3 pictures the distribution of the fastest thirds as counts and percentages across divisions for the top 16.

In contrast with figure 2, the top 16 in Division I swim the last third of the 500-yard freestyle the fastest a majority of the time. No other division shows this result. It is possible that the top 16 swimmers in DI may save up until the end to swim the fastest; or, they may swim at a steady pace throughout the race and bring it home with incredible speed. Differencing the splits will provide more insight. The percentages between the first third and the last third have gotten significantly closer together as well, especially in Division I and II. The first third in figure 2 for DI is 60.00% and DII is 70.48%. The last third in figure 2 for DI is 37.86% and DII is 28.57%. Figure 3 reveals an almost 50% split for both divisions between the first third and last third. Athletes in the top 16 are shifting to a different strategy. The last third gains ground in DIII as well thanks to an 11.74% drop in the first third from figures 2 to 3. With these shifts in mind, it is important to look at the differences in splits for the top 16:

Division	2 (100 Yards)	3 (150 Yards)	4 (200 Yards)	5 (250 Yards)	6 (300 Yards)	7 (350 Yards)	8 (400 Yards)	9 (450 Yards)	10 (500 Yards)
I	2.01	0.33	0.17	0.07	-0.02	0.01	0.04	-0.10	-0.51
II	2.16	0.44	0.20	0.09	0.05	0.01	0.09	-0.02	-0.85
III	2.23	0.49	0.25	0.17	0.09	0.01	0.07	-0.07	-0.58

Table 2 shows the differences in median splits across divisions for the top 16 in the 500-yard freestyle.

The splits across the divisions appear close to one another. The middle portion of the race–300 through the 450-yard mark–is the closest to a 0.00 second change from 50 split to 50 split. Overall, a large ramp up in time is seen in the beginning of the race followed by almost perfect pacing until the last 100 yards where the swimmers really took off time. Division II and III experience a shift in the swimmers' strategies when comparing the top 16 and the divisions as a whole. Five splits are under 0.10 seconds difference for Division II, and four splits are under 0.10 seconds difference for Division III in table 2. Only two splits are under 0.10 seconds difference for Division II and three are under 0.10 seconds difference for Division III in table 1. As a frame of reference, anything under a tenth is very close in swimming, and anything above a tenth is quite long. This transference from inconsistency to consistency demonstrates the top 16 in each division are better at consistently pacing the 500-yard freestyle than each division as a whole; but, does this hold true against the swimmers who were out of the top 16 at each meet?

Swimmers who do not make the top 16 are still excellent athletes and have trained fervently their whole lives to get to the NCAA Championships. It is no small feat to compete at these meets and being unable to qualify for the top 16 in no way discredits the dedication and time it took to get to where they are. The figure of the fastest thirds below visualizes a divergence from the other figures:

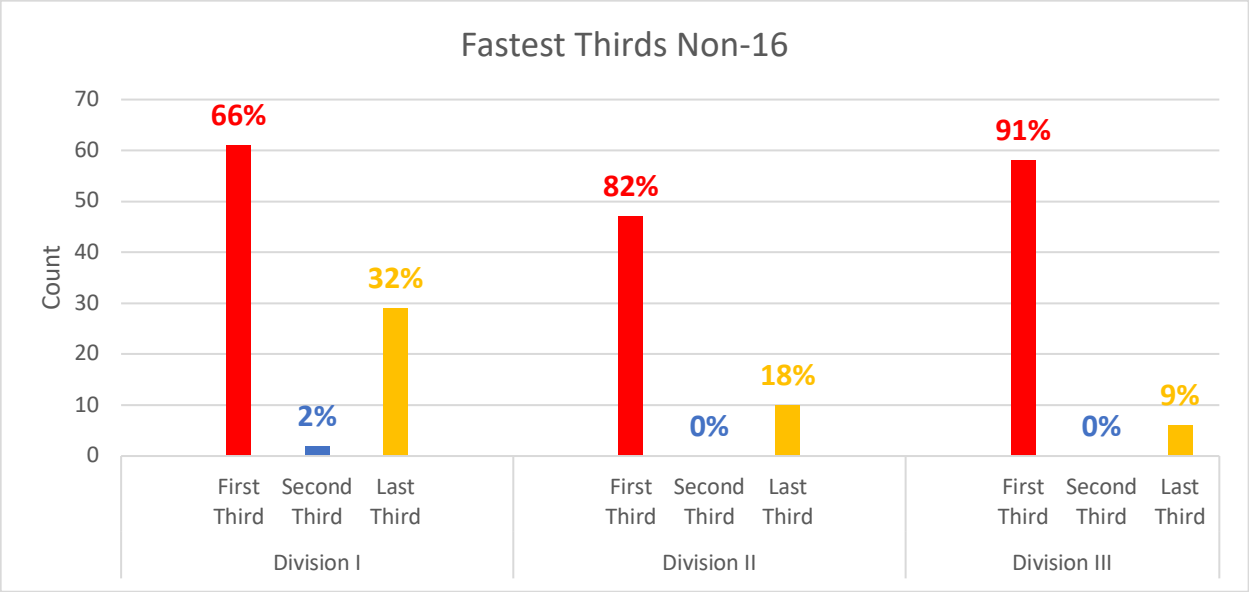

Fig. 4 shows the fastest thirds for each division broken down by count and percentages for swimmers not in the top 16.

The lower ranking swimmers very much favor the first third of their race as the fastest portion: 66.30% of DI athletes, 82.46% of DII and 90.63% of DIII. These percentages drastically diverge from the trend of the top 16 athletes. This intuitively makes sense. As a whole, the divisions favor the first third much more than the top 16 as seen with figure 2 compared to figure 3. Since there is a larger share of the population unable to qualify for the top 16, figure 4 should have a higher percentage of swimmers favoring the first third. Now, the first third being the fastest does not guarantee that the swims provided are not at an even pace. The table below will test that hypothesis:

Division	2 (100 Yards)	3 (150 Yards)	4 (200 Yards)	5 (250 Yards)	6 (300 Yards)	7 (350 Yards)	8 (400 Yards)	9 (450 Yards)	10 (500 Yards)
I	1.98	0.34	0.21	0.10	0.10	0.08	0.08	0.00	-0.60
II	2.19	0.49	0.36	0.09	0.15	0.21	0.17	-0.07	-0.82
III	2.21	0.53	0.34	0.17	0.13	0.16	0.15	-0.05	-0.63

Table 3 displays the difference in median splits for athletes who did not qualify for the top 16.

The swimmers who are not in the top 16 experience larger differences between median splits from 50 to 50. Division I in table 3 has only three 50s that are under 0.10 seconds difference whereas table 2 has four 50s under 0.10 seconds difference. Two 50s are under 0.10 seconds difference for Division II in table 3. The top 16 have five median split differences under a 0.10 seconds. With only one 50 below 0.10 seconds difference, Division III displays a large increase in time from 50 to 50. There is a surge at the last 50 across all divisions, consistent with all the other tables, but the bottom portion of swimmers find themselves going out too fast and dying off, unable to maintain their speed. These swimmers may also face a mental barrier in the pool, as well. They could be discouraged in the middle of the race as other competitors swim past them.

The top competitors in each division and Division I as a whole appear to swim their races with a more even distribution of energy. Knowing that the best swimmers take this approach in the 500-yard freestyle is valuable. While these elite athletes favor this strategy, it does not mean they are improving their times. Having established a method above by showing the fastest thirds for each division and then the differences in median splits, the same process can be used to look at those individuals who improved their season best times and those who did not. The fastest thirds for each division whose swimmers improved their season best time can be seen below:

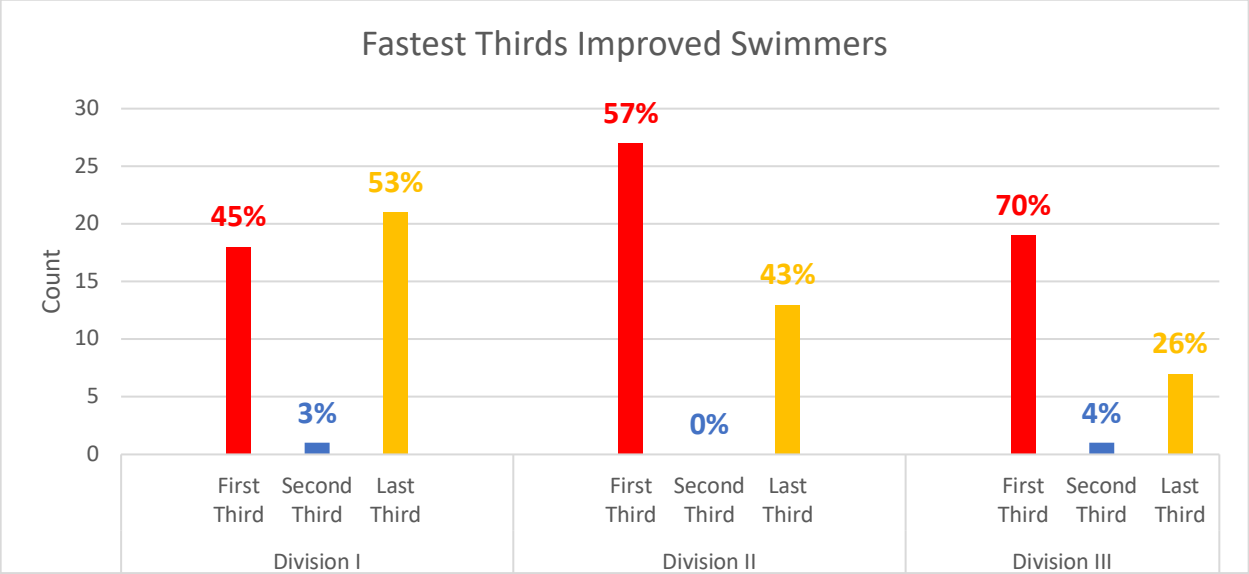

Fig. 5 shows the counts and percentages for the fastest thirds by division for the swimmers who improved from their season best.

Similar to figure 3, Division I athletes who improved their time favor the last third, and percentages for DII and DIII are nearly identical between figures 3 and 5. This commonality between the top 16 and the athletes who improved provides valuable insight into race strategy–consistent pacing is key. To corroborate this claim, the difference in median splits is below:

Division	2 (100 Yards)	3 (150 Yards)	4 (200 Yards)	5 (250 Yards)	6 (300 Yards)	7 (350 Yards)	8 (400 Yards)	9 (450 Yards)	10 (500 Yards)
I	2.01	0.31	0.15	0.13	0.02	0.02	-0.01	-0.05	-0.64
II	2.12	0.46	0.22	0.01	0.05	0.01	0.06	0.02	-0.81
III	2.20	0.46	0.21	0.04	0.01	0.05	0.00	-0.02	-0.63

Table 4 describes the difference in median splits across divisions for swimmers who improved their season best time.

The median splits for the swimmers who improved their best times are incredibly close to 0.00 seconds difference. While Division III swimmers tend to swim their first third the fastest, across all divisions the second half of the race is almost perfectly consistent. Division II and III swimmers have five 50s that are under 0.10 seconds difference with Division I having four. It appears from the data that an even pace throughout the 500-yard freestyle gives the swimmer the best chance at improving their overall time and being among the fastest swimmers at the meet.

Figure 5 and table 4 produce informative conclusions for how a swimmer can improve their time. To contrast this, one must look at the swimmers who did not improve their performance. Those athletes who did not better their season best may be limiting themselves through inefficient race strategy. The fastest thirds for the unimproved athletes are below:

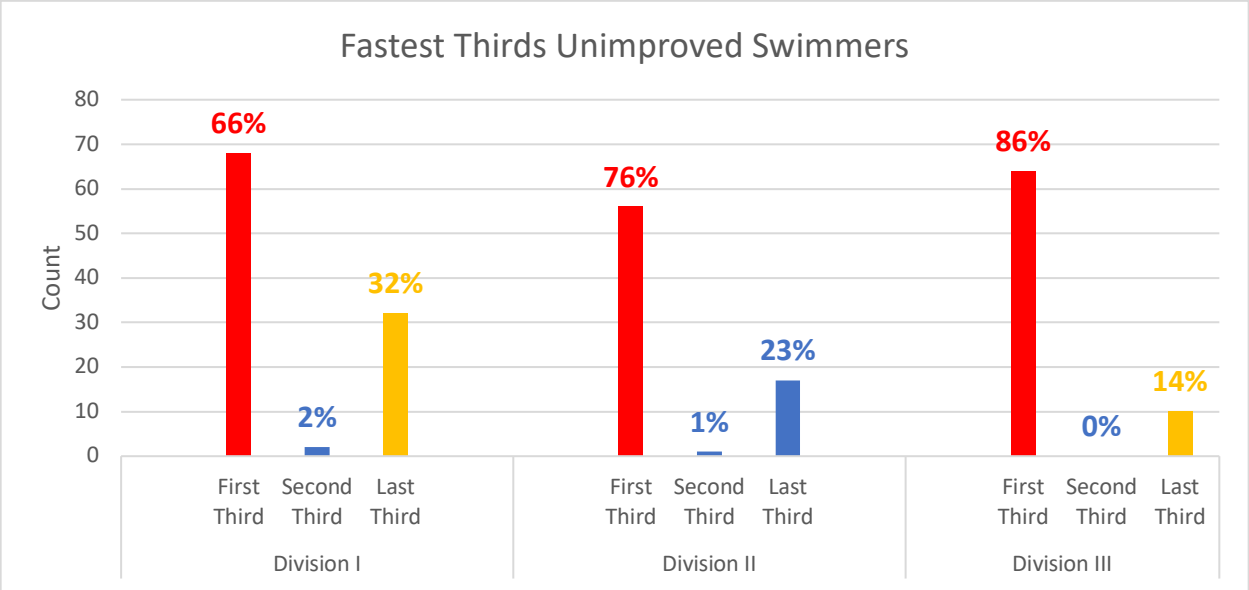

Fig. 6 shows the distribution of the fastest thirds for the swimmers who did not improve their season best.

Figure 6 shows a vast majority of the swimmers who did not better their season best swam the first third the fastest–a divergence from those that improved. The figure looks eerily similar to figure 4 whose percentages for the first third are as follows: 66.30% DI, 82.46% DII, 90.63% DIII. Athletes who did not make the top 16 heavily favor the first third of the race as the fastest and it appears that the athletes who show little improvement in their time do, as well. The splits between the unimproved athletes and the lower ranked swimmers may be similar too. To confirm this hypothesis, the difference in median splits can be seen below in table 5:

Division	2 (100 Yards)	3 (150 Yards)	4 (200 Yards)	5 (250 Yards)	6 (300 Yards)	7 (350 Yards)	8 (400 Yards)	9 (450 Yards)	10 (500 Yards)
I	1.99	0.34	0.21	0.07	0.09	0.08	0.13	0.02	-0.55
II	2.19	0.46	0.34	0.11	0.11	0.16	0.17	-0.06	-0.82
III	2.28	0.54	0.32	0.20	0.11	0.07	0.16	-0.06	-0.56

Table 5 visualizes the difference in median split times across divisions for the athletes who did not improve their season best.

A quick glance at Division II and III displays a large variance between each 50. Only two 50s are under 0.10 seconds difference for DIII and a single 50 in DII is under 0.10 seconds difference. Athletes are continually adding incredible amounts of time each 50, similar to the lower ranked swimmers in each division. Division I has a better distribution of effort throughout their 500-yard freestyles with a more consistent swim, but the difference in median times is still very close to or above 0.10 seconds. The splits in table 5 echo the same conclusions as table 3: athletes could be over swimming the first third and become too tired for the remainder of the race, or swimmers could also be giving up in the middle of the race. There are many reasons and possibilities for the ramp up in time for the unimproved swimmers; but, one thing is clear. Improving one's time is likely done by consistent pacing.

6.3 Regression Analysis

With organizing and reorganizing the swimmers into groups and subgroups, a plethora of information can be siphoned. To make the findings more robust, regressions can be introduced. Three regressions will be illustrated below in table 6. The columns of the table identify which independent variables are used to predict the dependent variable, improvement. The first column is titled "Division and Thirds". This regression makes predictions using the specific division a swimmer is in as well as which third is the fastest. Dummy variables, a 0 or 1, help to identify which swimmer is in what division and which third is the fastest. Division I and the first third are excluded from the regression to avoid falling into the dummy

variable trap–avoiding multicollinearity as much as possible. The second column titled "Splits Only" shows the effects each split in the 500-yard freestyle has on overall time improvement. The third column titled "Division, Thirds and Splits" combines the first two regressions to try to eliminate as much omitted variable bias as possible. While this attempt to mitigate omitted variable bias is made, there are still several variables omitted such as height, weight, etc. Therefore, the results can help draw some conclusions, but nothing should be taken as causal:

Table 11–MAGNITUDE OF THE EFFECTS OF THE DIVISION, FASTESET THIRDS AND SPLITS ON IMPROVEMENT

Dependent variable: Improvement from Season Best

	Division and Thirds	Splits Only	Division, Thirds and Splits
Division II	-0.2528	-	1.7279***
	(0.142)	-	(0.191)
Division III	0.1089	-	2.1898***
	(0.143)	-	(0.208)
Second Third	0.6738	-	-0.0018
	(0.493)	-	(0.382)
Final Third	0.6807***	-	-0.2359
	(0.131)	-	(0.139)
50	-	0.0467	-0.0300
	-	(0.161)	(0.191)
100	-	0.2012	0.1565
	-	(0.255)	(0.222)
150	-	0.4612	0.2342
	-	(0.299)	(0.260)
200	-	-0.2331	-0.1755
	-	(0.309)	(0.271)
250	-	0.3676	-0.0938
	-	(0.300)	(0.264)
300	-	-0.1063	-0.2968
	-	(0.283)	(0.247)
350	-	0.1196	0.0669
	-	(0.274)	(0.236)
400	-	-0.7203**	-0.6164**
	-	(0.249)	(0.215)
450	-	-0.1697	-0.2134
	-	(0.180)	(0.160)
500	-	-0.1842	-0.3466***
	-	(0.110)	(0.098)
Observations	351	351	351
R-Squared	0.088	0.328	0.510
Adjusted R-Squared	0.078	0.309	0.489
F-Statistic	8.366	16.63	24.96
P-Value (F-Stat)	0.000	0.000	0.000

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

Table 6 depicts the three regressions conducted to offer more robust findings.

The final third of the race is significant in the "Divisions and Thirds" regression; however, the Adjusted R-Squared of 0.078 is extremely low. Only 7.8% of the variability in overall time improvement can be explained by the limited number of independent variables. So, the significance of the final third does not hold much gravity. The second regression finds more success in an Adjusted R-Square of 0.309. The 400-yard split is statistically significant at the 5% level. An interpretation of this result means that a 0.01 second increase in time at the 400-yard mark leads to a 0.7203 percentage point decrease in improvement on average. If one wants to avoid adding time to their season best, the data suggests that the 8th 50 better be fast. Regression number three corroborates the importance of the 400-yard split. It too is statistically significant at the 5% level, averaging a 0.6164 percentage point decrease for every 0.01 seconds added to the split. The effect the 8th 50 has on improvement is less for the "Division, Thirds and Splits" regression than for the "Splits Only" regression. Unlike the second regression, the split at the 500-yard mark in the third regression is highly statistically significant at the 1% level. Adding an extra 0.01 seconds to the 10th 50 leads to a 0.3466 percentage point decrease in overall race improvement. Though, the 500-yard mark seems to be less important in improving one's time compared to the 400-yard mark. Sticking with the third regression, Division II and Division III are both statistically significant at the 1% level. Holding all other variables constant, Division II swimmers, on average, have an improvement that is 1.7279 percentage points higher than Division I swimmers; Division III swimmers have an improvement of 2.1898 percentage points higher than Division I swimmers holding all other variables constant on average. These coefficients dominate all other coefficients in the regression with their magnitude. Seeing such high coefficients leads to a potential explanation. Division I athletes often undergo more rigorous and stringent training, which allows them the opportunity to achieve their highest potential. The DI schools will often receive more funding over the DII and DIII schools as well, aiding in the progression of their swimmers. That is not to say that Division II and Division III swimmers cannot train like Division I swimmers, but it is less likely. Therefore, the DII and DIII competitors may not achieve their full potential. Someone who does not train and race to their fullest potential will have a much higher ceiling to improve than someone who is much closer to their ceiling such as Division I swimmers. These results offer explanations in the data, but omitted variable bias, heteroskedasticity and multicollinearity appear in these regressions. The effects of the independent variables cannot be causal and truly relied on. Some methods to mitigate the issues present in the regressions are Lasso, Ridge and Elastic-Net regressions.

6.4 Machine Learning Algorithms

Better modeling comes from eliminating multicollinearity and unnecessary variables that have little to no effect on improvement. Lasso, Ridge, and Elastic-Net provide some help in this process. A training set is forged for the models to help make predictions and understand the important variables in the regression. A test set is used to see the accuracy and precision of the models. Unfortunately, 97 swimmers positively improved while 247 swimmers unimproved, leading to an imbalance. The positive improvement rows are oversampled to get a better split between the two for testing. Oversampling takes random draws from the existing pool of positive improvements to add more to the training set. Measuring how well these

regressions did, the mean squared error, R-Squared, K-Fold Cross-Validation and the weight of coefficients are presented. The mean squared error measures the average difference between the prediction and actual values of improvement. A low MSE is best. The R-Squared previously mentioned explains the variation in the dependent variable. The K-Fold Cross Validation splits the data into subsets and evaluates the performance at each fold. The weight of coefficients shows the weight given to each independent variable. These tools of measurement for the regressions can be seen below in table 7:

Method	MSE	R-Squared	K-Fold 1	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5
Lasso	1.50	-0.25	-1.03	-0.94	-1.66	-1.72	-1.12
Ridge	0.78	0.35	-0.53	-0.68	-0.82	-0.71	-0.69
Elastic-Net	1.50	-0.25	-0.96	-0.89	-1.63	-1.69	-1.06

Table 7 shows the MSE, R-Squared and K-Folds for Lasso, Ridge and Elastic-Net.

Lasso and Elastic-Net appear to be nearly identical in everything except for their K-folds which barely diverge from each other. The pair share the same MSE of 1.50 which is higher than the 0.78 MSE of the Ridge regression. This difference in MSE indicates that the Ridge regression did a better job at predicting the improvement of a swimmer based on the given independent variables than the other two methods. A clearer reason for this is the R-squared. The negative R-squared of Lasso and Elastic-Net indicate that they are performing so poorly that a naive, simple average may perform better. As for the Ridge regression, 35% of the variance in the dependent variable, improvement, can be explained by the independent variables. This positive R-squared performs quite well, but not as well as the "Division, Thirds and Splits" regression's R-squared of 48.9% from table 6. Consistency and positive numbers in the K-folds is important. There is little variation across the data split, but negative numbers in the K-folds present an issue. Negative numbers typically stipulate that the model is performing poorly, and the predictions are worse than a simplistic model that predicts the mean of the target variable for all observations. Negative scores suggest that the model is not capturing the underlying patterns in the data and might even be making systematic errors in its predictions. To better understand why the models may not be performing as well, the weight of the coefficients can be witnessed below:

Model	50	100	150	200	250	300	350	400	450	500	DII	DIII	2/3	3/3
Lasso	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ridge	-0.20	0.37	0.28	-0.15	-0.33	-0.37	0.16	-0.64	-0.13	-0.27	1.47	1.94	-0.05	-0.29
Elastic	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 8 displays the coefficient weights for Lasso, Ridge and Elastic-Net methods.

The weights for the coefficients for the Lasso and Elastic-Net are all 0, meaning that the pair of models do not think any of the predictors are useful. The Lasso regression is forcing itself to overpower the Ridge regression in the Elastic-Net. It is peculiar that Lasso shrinks all coefficients to zero. Occasionally, the Lasso estimator may favor independent variables with high collinearity if they have a better signal-to-noise ratio,[4] and these regressions do have high multicollinearity. But, it appears that Lasso is disregarding this tendency. Instead, it finds all the coefficients to be suboptimal. The Ridge regression–unable to send coefficients to zero–clearly favors Division II and Division III as its most helpful predictors. As discussed before, DII and DIII competitors have a higher ceiling for potential than Division I swimmers, suggesting larger weights in the Ridge regression. To better address the coefficient problem, another method for eliciting variable importance is utilized–the Random Forest.

Random Forests are unable to send coefficients to zero, similar to Ridge, but research has shown that the Lasso and Random Forest have very little difference in accuracy and specificity in certain cases.[5] This fact means that the Random Forest may put very little weight on all features similar to Lasso. To test this, the feature importance for the Random Forest can be seen below:

Model	50	100	150	200	250	300	350	400	450	500	DII	DIII	2/3	3/3
Random Forest	0.12	0.15	0.05	0.05	0.03	0.04	0.04	0.15	0.29	0.05	0.002	0.01	0.001	0.0001

Table 9 shows the feature importance for the Random Forest.

The 450-yard mark proves to be the most important feature for making a prediction according to the Random Forest. This result is a strong departure from the Ridge regression which found the 450-yard mark to hurt predictions with a -0.13. Furthermore, the Random Forest does not find Division II or Division III

[4] Jorge A. Chan-Lau "Lasso Regressions and Forecasting Models in Applied Stress Testing"

[5] Parzinger, Michael, et al. "Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection – using machine learning methods Random Forest and lasso."

to be helpful in identifying which swimmers may improve. The Ridge regression believes that Division II and Division III are the most important features for prediction. Each method generates their results in different ways, and it is evident there is a divergence in their methods in table 8 and 9. The MSE for the Random Forest is 0.69 and the R-Squared 0.43, meaning the Random Forest does the best job at making predictions out of the machine learning algorithms. Still, the results cannot be thought of as causal. Instead, these methods help to understand what may be important in improving one's time, but they also prove that there are many missing variables. The robustness and clear evidence of concept are lacking with the current dataset, but that does not in any way diminish the findings from the figures and tables of section 6.2 Fastest Thirds Analysis. Those findings still hold plenty of gravity. The top 16 and most improved swimmers swim a faster race compared to others by holding a consistent pace throughout their whole 500-yard freestyle.

7. Conclusion

Developing a concrete strategy for the 500-yard freestyle can be an arduous task. Many variables go into a race before it even happens: the nutrition of the athlete, training plan, pre-race routine, the mental preparedness and many more factors. This paper focuses on one variable: race strategy. The results have shown that the very best swimmers–both by collegiate division and within their divisions–swim a more consistent even paced race than less successful swimmers. Moreover, the swimmers who improved upon their best time in a given season swim their race akin to the top swimmers in the country. The results of this analysis lead to the conclusion that the best way to understand and implement this race strategy is to train a consistent pace in practice. Holding an ideal pace repeatedly in practice will train the body and the mind to resort to that steady pace during a race. This training will not guarantee a best time for anyone, but it will increase the likelihood of success. A coach and swimmer cannot control everything that goes into a race, but training a consistent pace is something that can be controlled.

In the future, gathering more variables for hypothesis testing will prove to be most useful. Height, weight, venue, stroke rate and many other factors go into a race. Having all of these factors will lead to better predictions from the models presented in this paper. Additionally, this paper hopefully leads to more analysis in swimming. Breaking down one's training regimen into thought out carefully analyzed practices will only lead to more success at meets.

References

Chan-Lau, J. (2017). Lasso regressions and forecasting models in applied stress testing. *IMF Working Papers*, *17*(108), 1. https://doi.org/10.5089/9781475599022.001

Hausman, S. (2022, March 12). *This professor studies each swimmer as a math problem. it's helped them to be faster*. NPR. https://www.npr.org/2022/03/12/1085542427/uva-professor-swimmer-math-faster

Honicky, B. (2016, April 20). *A guide to taper and trusting your training*. Swimming World News. https://www.swimmingworldmagazine.com/news/a-guide-to-taper-and-trusting-your-training/

Keith, B. (2022, October 28). *What happens when you put an 18.2 sprinter in the 500 free?*. SwimSwam. https://swimswam.com/what-happens-when-you-put-an-18-2-sprinter-in-the-500-free/

Parzinger, M., Hanfstaengl, L., Sigg, F., Spindler, U., Wellisch, U., & Wirnsberger, M. (2022). Comparison of different training data sets from simulation and experimental measurement with artificial users for occupancy detection — using machine learning methods Random Forest and lasso. *Building and Environment*, *223*, 109313. https://doi.org/10.1016/j.buildenv.2022.109313