Predictive Analytics II
Individual Project Report
Daniel Burkhalter
12/10/2023

Contents

Proposal

Competitive running is a sport which garners little attention from the public in the United States other than weird looks, yet often compliments. However, recreational running is a common practice in summer 5k's and marathons. Our goal is to promote competitive running in South Dakota for high school students by creating a power rankings system for cross country (otherwise known as XC).

Definition

What is a power ranking system? A power ranking system takes cross country results from all XC meets in SD and ranks individuals (and teams) based on head-to-head competition. For example, runners in Northwest SD (middle of nowhere, Bison SD) will occasionally run against Rapid City schools but will never see Sioux Falls schools until the end of the year. Rapid City schools will race Sioux Falls schools several times throughout the year. My goal is to compare the matchup between Bison-Rapid City and Rapid City-Sioux Falls, to determine Bison-Sioux Falls matchup.

Current System

The current system ranks individuals based on 5k times throughout the state. There are two major drawbacks:

1) The difficulty (and distance) of courses varies widely throughout the state. A flat course in Sioux Falls is not comparable to a course at altitude in Custer, SD. Hence times are not standardized.
2) Head-to-head competitions are not considered. If individual A beats individual B three times throughout the year in Custer, SD. Yet, individual B races in Sioux Falls (individual A doesn't race in Sioux Falls) and runs a "faster time" than individual A, then individual B will be ranked higher than individual A.

Data

There is a centralized database for all SD high school cross country meets housed by Athletic.net. The displayed results are proprietary information and permission was denied for using their results.

Given denied permission from Athletic.net to access their database, we contacted all cross-country coaches/administrators in SD that hosted a XC meet and request their results from 2023. There were 112 XC meets in SD for the Fall of 2023. After contacting the hosts, we obtained the majority of results from the meets. We had 46 unique data sources. One data source, Dakota Timing, is a major meet manager within SD. With 46 different data sources, the format of the results was in various formats including: .pdf, .xlsx, .html, and of course the USPS snail mail.

Ranking Methodology

The proposed ranking methodology is based on the ELO rating system (used for Chess rankings) where in a matchup each individual bets a certain percentage of their points for a win/loss. In a XC race, there are more than 2 individuals. However, everyone is running against the other runners one at a time.

For example, runner 1 is competing against runners 2,3,4,5,…; runner 2 is competing against runners 1,3,4,5,….; runner 3 is competing against runners 1,2,4,5,….; and so on. Given

$$X = Number\ of\ Runners\ in\ Race,$$

Then the number of head-to-head matchups can be calculated as:

$$n = \frac{(X-1)X}{2}.$$

In our points method, for each matchup the athlete places 5% of their existing points on the line for the win/loss. Although, the percentage of points seems marginal, with X=99 runners there are n=4851 head-to-head competitions going on in a single race. If an athlete hasn't competed, they are given 1000 points.

Ranking Example

Our initial goal for testing the ranking system based on the points system vs the current times system was to predict the results from the boy's Region 3A XC meet. We used performances from 13 meets (Table 1) from the schedule of schools in Region 3A (Table 2). The teams underlined in Table 2 have thin representation within the 13 XC meets assessed.

Table 1

Meet #	Date	XC Meet
1	8/25/2023	Beresford
2	8/29/2023	Dakota Valley
3	8/29/2023	McCook-Central
4	9/1/2023	Augie Twilight
5	9/5/2023	Canton
6	9/5/2023	Scotland
7	9/11/2023	Viborg
8	9/12/2023	SF Christian
9	9/14/2023	Chamberlain
10	9/18/2023	Alcester
11	9/28/2023	Lennox
12	10/5/2023	Big East Conference
13	10/5/2023	Dakota 12 Conference
Target	**10/11/2023**	**Region 3A**

For the first meet we used, Beresford, all athletes were assigned 1000 points. Following the race, the points were updated. In essence, the first-place individual took 5% of 2nd through 99th 1000 points resulting in 5900 points. Second place took 5% of the remaining 950 points from 3rd through 99th, third place took 5% of the remaining 902.5 points from 4th through 99th, and the 99th place lost points to everyone.

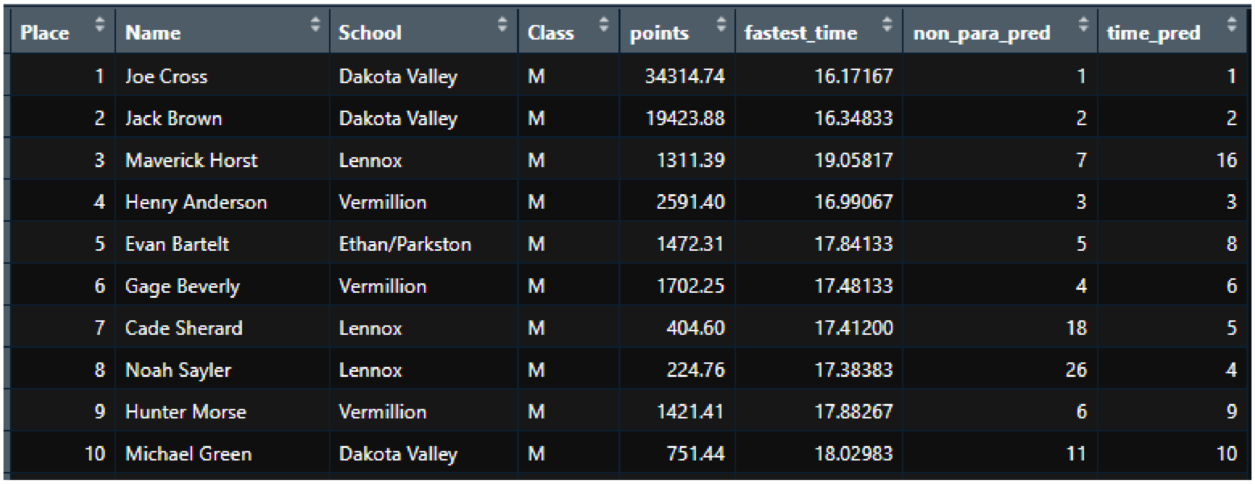
For the second meet, Dakota Valley, individuals that hadn't competed yet were assigned 1000 points. The compiled rankings after the 2nd meet are shown in Table 3. After 13 XC meets, the points system and the fastest season time were used to predict placing for the Region 3A XC meet. The top 10 placers are shown in Figure 1.

Table 2

Region 3A Teams
Beresford
Bon Homme
Canton
Dakota Valley
Elk Point-Jefferson
Ethan/Parkston
Hanson
Lennox
Mount Vernon/Plankinton
Parker
Sanborn Central/Woonsocket
Vermillion
Wagner

Table 3

Rank	**Name**	**ID**	**School ID**	**Points**
1	Jack Brown	3	9	7142
2	Joe Cross	2	9	6828
3	Owen Janiszeski	1	17	5900
4	Levi Vander Leest	4	22	4930
5	Alex Oberloh	5	22	4643
6	Ryan Fick	6	17	4372
7	Benjamin Strunk	7	26	4117
8	Tavin Schroeder	8	14	3876
9	Jonathon Roth	9	26	3649
10	Reid Hammerquist	16	23	3516



Place	Name	School	Class	points	fastest_time	non_para_pred	time_pred
1	Joe Cross	Dakota Valley	M	34314.74	16.17167	1	1
2	Jack Brown	Dakota Valley	M	19423.88	16.34833	2	2
3	Maverick Horst	Lennox	M	1311.39	19.05817	7	16
4	Henry Anderson	Vermillion	M	2591.40	16.99067	3	3
5	Evan Bartelt	Ethan/Parkston	M	1472.31	17.84133	5	8
6	Gage Beverly	Vermillion	M	1702.25	17.48133	4	6
7	Cade Sherard	Lennox	M	404.60	17.41200	18	5
8	Noah Sayler	Lennox	M	224.76	17.38383	26	4
9	Hunter Morse	Vermillion	M	1421.41	17.88267	6	9
10	Michael Green	Dakota Valley	M	751.44	18.02983	11	10

Figure 1

Based on Figure 1 we see the Top 10 places for boys Region 3A XC meet, their school, points accumulated throughout the season, their fastest time (in minutes), their points (non-parametric) rank, and their fastest time rank. For example, Maverick Horst from Lennox placed 3rd, was ranked 7th using points and 16th using time.

There were 65 runners in the race (where Region wasn't their first competition). We used the Mean Absolute Error (MAE) to predict the accuracy of the model where:

$$MAE_i = \frac{1}{n}\sum_{i=1}^{n} |Actual_i - Predicted_i|$$

Where $n = 65$ runners. The results of the two ranking methods are shown in Table 4. The points method predicted on average within 10 places of finishing whereas the time method predicted on average within 5 places of finishing. Hence, times throughout the season appear to be a better ranking method than the points method in this particular case.

Table 4

	Points Method	Fastest Time
MAE	10.03	5.72

Discussion

Based on our initial findings, the current method of time is still superior in predicting race finish. Some drawbacks of the points system are

- More racing causes an above average individual to garner more points than an individual with fewer races
- Schools that raced fewer times (within the 13 races we used) may be underrepresented in magnitude of points
- The ratio of points sacrificed impacts the sensitivity of wins/losses and how easily an individual rises or falls within the rankings

A complete ranking system using all the SD meets to predict finish at the State XC meet could yield different results due to equal representation from schools.