Predicting Collegiate Track and Field Results

Ellie Talius and Erika Malaspina
Department of Computer Science
Stanford University
etalius@stanford.edu, emalasp@stanford.edu

Abstract

Machine learning has long been used to predict the outcome of sporting events or the value of certain players at the professional level. However, it has never been applied to the sport of track and field. As such, we propose the application of linear regression to the problem of predicting collegiate track and field results for athletes based on their performances in previous years, the school they attend, the conference they are in, and the year they graduate college. Concentrating on men and women in the 100m, 5000m, pole vault and triple jump, we find that by using a linear regression model, we can accurately predict an athlete's senior year performance from these features. We achieve a top R^2 value of 0.98 on Men's Pole Vault, suggesting that our model is highly accurate.

1 Introduction

We decided to tackle the problem of predicting the performance of collegiate track and field athletes in their senior years. We are interested in this project because we are both track and field athletes and have seen many of our teammates improve over the years and sign professional contracts. Some of these teammates were top recruits in high school, while others made drastic improvements in college, which makes the problem of predicting who will improve enough to turn professional a complex and exciting problem.

We apply various machine learning algorithms to the task of predicting collegiate track and field results in the 100m, 5000m, pole vault and triple jump for both men and women. Specifically, we focus on the task of predicting senior year season bests based on the athletes' season bests from their first three years, the school they attend, the year they complete their senior year of eligibility, and the conference they participate in. We accomplish this by utilizing different types of linear regression models.

Senior year personal bests are particularly relevant, as the top collegiate athletes typically pursue professional running after they graduate, but commonly only hit the standards needed to pursue professional running in this final year. Thus, there is value for agents in being able to predict which athletes will do well while they are still underclassmen, and our prediction can be used as a proxy for determining which collegiate track and field athletes will perform well enough to turn professional. However, the progression of athletes in collegiate track and field does not follow the ideal linear trajectory, as injuries and other circumstances can lead to athletes failing to improve one year, and breakthrough seasons are quite common. Athletes improve at different rates, so a basic modeling technique of using the average rate of improvement of an athlete over their first three years of college to predict their final year's results is not sufficient. As a result, more sophisticated models are needed.

2 Related Work

Machine learning has been applied to a variety of tasks relating to predicting the outcomes of sporting events and predicting various sporting metrics. Previous research includes predicting the outcomes of NFL games, in which Purucker achieved 61% accuracy by using a neural network and Kahn improved upon to reach 75% accuracy, surpassing human expert performance of 63%. Additionally, regression analysis has been performed to predict the outcomes of horse races, elite female swimming at the 2000 Sydney Olympics, and PGA winning golf scores (Bunker). However, no such application of machine learning has been seen in the sport of track and field, and previous approaches have focused on professional sports, largely ignoring the collegiate level which has far more athletes participating and thus more data.

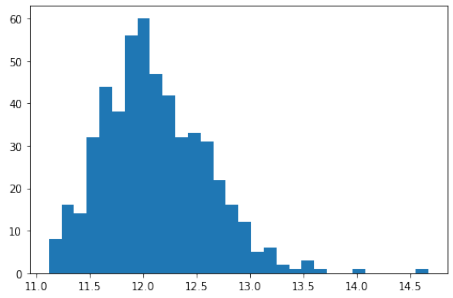
3 Dataset and Features

To obtain our data, we performed a scrape of the website athletic.net, which tracks the top collegiate athletes in each event, top athletes in different conferences and schools, as well as the meet results and season PRs for each individual athlete. Since

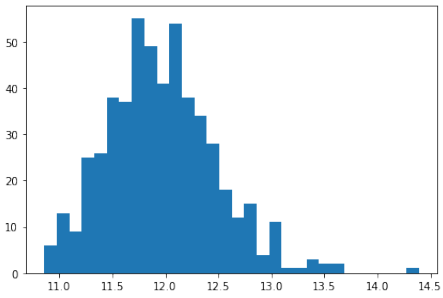
athletic.net contains a plethora of data, we focused on pole vault, triple jump, the 100m, and the 5,000m for both men and women in order to narrow the scope of our analysis while still surveying a wide range of event types. To scrape the data, we first compiled a list of the top 25 athletes in each Division 1 conference for the years 2010 to 2021 for each of the 4 events we are considering. From this list, we then pulled data for each athlete about their outdoor PR's over their 4 years of eligibility, removing athletes that did not have 4 years worth of results in the event. Our average dataset size for each gender for each event was about 300 to 500 examples. We then split this data into a 85% train and 15% test split for our preliminary data analysis.

Event	Training Data Size	Test Data Size
Women's 100m	445	78
Men's 100m	349	61
Women's 5000m	312	55
Men's 5000m	320	57
Women's Pole Vault	310	55
Men's Pole Vault	251	44
Women's Triple Jump	275	48
Men's Triple Jump	185	32

After obtaining our data, we then filtered it for any outliers and removed them by using histograms to evaluate the shape and spread of the data. We plotted a histogram for the distribution of PRs for each of the 4 years and then also compared the distributions from year to year. We found that the variances typically increase from freshman to senior year. We hypothesize this is because the experience of each athlete in college is so unique that athletes can make drastic leaps in performance. The figure below shows one such histogram that was used for the women's 100m. We note that in this histogram, all values are in a normal range.



(a) Freshman Year PRs



(b) Senior Year PRs

Figure 1: Distribution of PRs in the Women's 100m

We then also looked at the 4 year trends for 7 random samples of the data set for each gender and event to understand the what the 4 year trends looked like across different events. For example, in the Women's 100, we saw both improvements, as well as decreased performance over the 4 years.

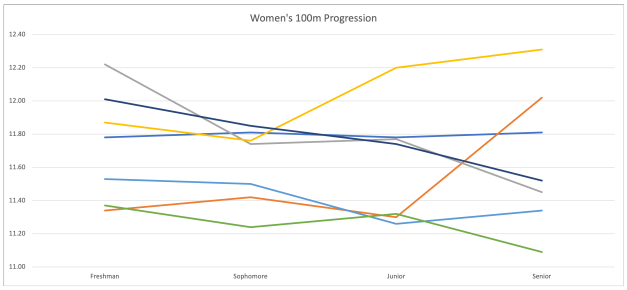


Figure 2: Random samples of progressions in the Women's 100m

The features that we chose to consider were the athlete's PRs from their Freshman to Junior year, the school they attended, the conference that they participated in, and the year that they graduated. We chose these because we hypothesized that the school and conference that an athlete attended would impact their progression over the 4 years, as well-funded schools typically put more into their athletes' development in order to be competitive within their conference. We also have seen a general trend of performances getting stronger over the past few years in the NCAA, and since our data comes from such a large time span, we felt that the graduation year of the athlete may be informative.

4 Methods

Our primary method used was linear regression. Our prediction took the form

$$h(x) = \theta^T x$$

where θ were the weights learned by our model, x was a vector representing one sample, and the output $h(x)$ was a value representing the predicted time or distance for the sample. We define the cost function used for linear regression as

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$$

where $y^{(i)}$ is the actual senior year PR for athlete i and $h_\theta(x^{(i)})$ is our predicted senior year PR for athlete i. Intuitively, this cost function measures the sum of the squared difference between our prediction of the senior year PR and the actual senior year PR.

We then train a linear regression model with the goal of minimizing this cost function. Taking the derivative of the cost function with respect to theta gives

$$(y^{(i)} - h_\theta(x^{(i)}))x^{(i)}$$
$$= (y^{(i)} - \theta^T x^{(i)})x^{(i)}$$

We can then use this derivative to define an update rule for the values of theta after seeing a single training example as below:

$$\theta := \theta + \alpha(y^{(i)} - \theta^T x^{(i)})x^{(i)}$$

We then use this update rule in the stochastic gradient ascent algorithm. We began by running this simple linear regression on each of our 8 data sets, using only the numerical features of the athlete's freshman through senior year PRs. After we had this numerical baseline, we then decided to consider adding higher order terms to our features to add some non-linearity to the model to better capture the data.

To add these higher order terms, we added 3 new features, each of which corresponded to the square of the athlete's freshman, sophomore and junior year PRs. Mathematically, we defined our feature map $\phi : R^4 \to R^7$ as

$$\phi(x) = \begin{bmatrix} 1 & x_1 & x_2 & x_3 & x_1^2 & x_2^2 & x_3^2 \end{bmatrix}^T$$

We then used the same gradient ascent algorithm as derived above, but replaced x with $\phi(x)$ to get

$$\theta := \theta + \alpha(y^{(i)} - \theta^T \phi(x^{(i)}))\phi(x^{(i)})$$

After using this method, we then decided to consider the categorical variables present in our data, which were the school the athlete attended, the conference they participated in, and the year that they graduated. In order to use these categorical terms in our linear regression model, we needed to use one-hot encoding for them. In this form of encoding, for each categorical feature, we take all the categorical values of that features and turn them into individual features. Then, we use a 0 to represent the absence of that feature and a 1 to represent the presence of that feature. Thus, each example will only have a single 1, corresponding to its categorical value, and the rest of the features will have the value 0, hence the name one-hot encoding.

After adding these categorical terms, we noted that the model was overfitting, so we chose to explore using ridge and LASSO regression. For ridge regression, a penalty term is added to the cost function. The penalty term is the L2 norm of the weights squared, and thus the model to chooses smaller weights and decreases overfitting (Marquardt). The cost function becomes:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda||\theta||_2^2$$

And thus the gradient ascent update rule is

$$\theta := \theta + \alpha((y^{(i)} - \theta^T x^{(i)})x^{(i)} + 2\lambda||\theta||_2)$$

Next, we considered LASSO, which uses the L1 norm as a regularizing term, instead of the L2 norm in ridge regression. Instead of shrinking all the weights as in ridge regression, LASSO shrinks the weights of some features to 0 to avoid overfitting, which is important in our case as we introduced more features (Ranstam). The cost function is

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda||\theta||_1$$

Since the L1 norm is not differentiable, we must use coordinate descent to get the update rule (Ranstam).

To evaluate our data, we used 2 metrics, the coefficient of determination (R^2) and the root mean squared error (RMSE) and applied them to both our train and test data sets.

The coefficient of determination measures the proportion of variance in the dependent variable, the senior year PR of the athlete, that is predictable from the independent variables (the features we considered). We define R^2 as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where $SS_{res} = \sum_{i=1}^{n}(y^{(i)} - h_\theta(x^{(i)}))^2$ and $SS_{tot} = \sum_{i=1}^{n}(y^{(i)} - \bar{y})$ where $\bar{y}$ is the mean of y values across all samples (Ozer). In this case, an R^2 value of 0 depicts a model that always predicts the mean, and a negative value means the model does worse than this baseline, while a positive value means it does better (Ozer).
The RMSE is defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} h_\theta(x^{(i)}) - y^{(i)}}{n}}$$

This error measures the average distance between the predictions the model made and the actual values of the dependent variables (Nevitt). Thus, the RMSE also depends on the units and sizes of values being predicted.

5 Experiments, Results, and Discussion

After obtaining our data, we first ran a simple linear regression to predict senior year PRs from freshman, sophomore, and junior year PRs only. We did this for both men and women in the pole vault, triple jump, 100m, and 5,000m. These results can be seen in the table below.

Event	Train R^2	Train RMSE	Test R^2	Test RMSE
Women's 100m	0.79	0.23	0.74	0.25
Men's 100m	0.74	0.21	0.61	0.29
Women's 5k	0.69	25.5	0.57	29
Men's 5k	0.65	25	0.69	32.2
Women's Pole Vault	0.97	0.24	0.86	0.19
Men's Pole Vault	0.97	0.24	0.98	0.19
Women's Triple Jump	0.96	0.46	0.65	0.47
Men's Triple Jump	0.98	0.52	0.57	0.65

We interpreted the results for the women's and men's 100m and pole vault as acceptable baselines to use when exploring the rest of our approaches, as they perform well.

However, our predictions for the 5000m and triple jump were worse, as evidenced by the poor test R^2 and test RMSE data. We hypothesize that for the 5000m, this comes from a larger variation in the results as runners adapt to higher mileage and are able to make bigger improvements than in the 100m. Additionally, as the performance values are generally much larger for the 5000m due to the nature of the performance marks in this event, the RMSE values will also tend to be much larger than those for other events. We also hypothesize that for the triple jump, this poor performance is due to a much smaller sample size for this event.

After seeing how well these simple linear regressions performed, we then attempted to predict senior year PRs from subsets of their previous years' PRs: only freshman year; only freshman and sophomore year; only sophomore and junior year; and only junior year. A model that performs well with only a subset of PRs would be useful, as it could help agents identify top athletes earlier or with less available data. Since this is a lot of models to examine, we will only be reporting the results of the women's 100m.

Class Year Results Used	Train R^2	Train RMSE	Test R^2	Test RMSE
Freshman	0.6	0.33	0.5	0.3
Freshman and Sophomore	0.79	0.24	0.64	0.27
Sophomore and Junior	0.79	0.24	0.74	0.25
Junior	0.75	0.25	0.78	0.26

From these results, we can see that the models involving the junior year PRs improved the R^2 values significantly over the models that did not have access to this data.

We also edited our model by adding squared terms for the freshman, sophomore, and junior PRs. For the women's 100m, the results of this were: Train R^2 of 0.78, Train RMSE of 0.24, Test R^2 of 0.83, and Test RMSE of 0.22. Thus, we see that adding non-linearity in the model led to better results.

Next, we decided to return back to our simple linear regression model, but added in the school they attended, the conference they were in, and the year they graduated college. The results for this can be seen in the table below.

Event	Train R^2	Train RMSE	Test R^2	Test RMSE
Women's 100m	0.99	8.02E-08	0.52429	0.320869
Men's 100m	0.99	4.38E-08	0.256662	0.334662
Women's 5k	0.99	1.60E-05	0.162138	198.040768
Men's 5k	0.99	4.23E-06	0.464924	36.971454
Women's Pole Vault	0.99	1.81E-07	-0.475936	0.395972
Men's Pole Vault	0.99	1.56E-07	0.070962	1.950868
Women's Triple Jump	0.99	1.29E-07	0.907186	1.211311
Men's Triple Jump	0.99	2.23E-07	-2.201308	1.837482

Generally, these models performed much better on the training data but much worse on the test data as compared to all of our previous models. This suggests that these models were overfitting to the extra variables. We decided to run both ridge regression and LASSO in an attempt to help with the overfitting. We also ran ridge regression on this same model but removed the school variable, as we deemed that variable to be the most likely one responsible for overfitting. Again, due to the large amount of models to examine, we have only reported results here for the women's 100m.

Regularization Type	Train R^2	Train RMSE	Test R^2	Test RMSE
Ridge Regression	0.983313	0.067228	0.552597	0.311176
LASSO	0.818926	0.221455	0.328783	0.381143
Ridge Regression (no school)	0.903437	0.161719	0.459045	0.342166

These results show that the regularization methods performed generally did not have a large effect on the test R^2 and test RMSE values, and in some cases made the models perform more poorly.

6 Conclusion and Future Work

Our overall best-performing prediction model ended up being the model using only the 3 previous years' season bests with squared terms. Additionally, we found that models containing junior year marks performed notably better than models that did not contain this data.

These results have interesting implications in predicting which collegiate track and field athletes will perform well enough their senior year to turn professional. Our results suggest that that the school an athlete attends, the conference that athlete competes in, and the year that athlete completes their senior season are not all that useful in determining an athlete's ability to turn professional. This is somewhat surprising, as certain conferences and schools have reputations in the track community for producing more professional athletes. For instance, it is commonly believed that the SEC generally produces a lot of athletes who later become pro. Additionally, USC has a reputation for producing lots of good sprinters, and Stanford and Oregon have reputations for producing a lot of good distance runners. However, our findings suggests that the reputations of these conferences and schools are less important than people might think.

Our results also suggest that how an athlete has performed in the past is not as important as how they have performed in the last year, and that it is harder to predict who will turn professional based on an athlete's performance as an underclassman. This makes sense based on our prior knowledge of the potential for breakthrough seasons or performances.

In the future, it would be useful to explore adding more higher order terms to our most successful models to see how that affects the performance of those models. Additionally, there is the potential for our methods to be generalized to all of the track and field events. One could also attempt to make more complicated prediction models using this data in the hopes of finding something more accurate, such as creating a neural network with data augmented from bootstrapping. In short, there is a lot of potential for future work in the topic of predicting collegiate track and field athletic performances.

Contributions

Ellie worked on scraping the data, running the experiments and writing the Introduction, Related Works, Dataset, and Methods Sections of the report. Erika designed the experiments, collected data, and wrote the experiments, discussion and conclusion of the report.

References

[1] Rory P. Bunker, Fadi Thabtah, A machine learning framework for sport result prediction, Applied Computing and Informatics, Volume 15, Issue 1, 2019, Pages 27-33, ISSN 2210-8327, https://doi.org/10.1016/j.aci.2017.09.005. (https://www.sciencedirect.com/science/article/pii/S2210832717301485)

[2] Ozer, D. J. (1985). Correlation and the coefficient of determination. Psychological Bulletin, 97(2), 307-315. doi:10.1037/0033-2909.97.2.307

[3] Jonathan Nevitt Gregory R. Hancock (2000) Improving the Root Mean Square Error of Approximation for Nonnormal Conditions in Structural Equation Modeling, The Journal of Experimental Education, 68:3, 251-268, DOI: 10.1080/00220970009600095

[4] Donald W. Marquardt Ronald D. Snee (1975) Ridge Regression in Practice, The American Statistician, 29:1, 3-20, DOI: 10.1080/00031305.1975.10479105

[5] J Ranstam, J A Cook, LASSO regression, British Journal of Surgery, Volume 105, Issue 10, September 2018, Page 1348, https://doi.org/10.1002/bjs.10895