

Sport Analytics with Statistical Learning

Barabás Eszter

ELTE Eötvös Loránd Tudományegyetem

Supervisor: Csáji Balázs Csanád

Budapest, 2025.

Introduction and Data

- athletic performance and external environmental factors (e.g., temperature, wind, humidity)
- 1,258 races (rows) held between 1936 and 2019 across 42 countries, encompassing data from 7,867 athletes.
- information about the races (columns) includes type of the race, date, location data, weather parameters, previous records and top finish times

Statistical methods

Linear regression: Given a data set $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ of n independent sampling units and p observed features. Each $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ is a vector of feature measurements for the i th case. The relationship between the dependent variable y_i and the input vector x_i takes the form

$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j.$$

Given in matrix notation $\mathbf{y} = \mathbf{X}\beta$, where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & & & \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Shrinkage methods improve model fit by "shrinking" the regression coefficients to reduce variance with introducing bias: $\lambda \geq 0$ is a parameter that controls the amount of shrinkage

Ridge regression:

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}.$$

Lasso regression:

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}.$$

Decision Tree: Take our data (x_i, y_i) for $i = 1, 2, \dots, N$, with $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and suppose we have a partition into M regions R_1, R_2, \dots, R_M , and we model the response as a constant c_m in each region:

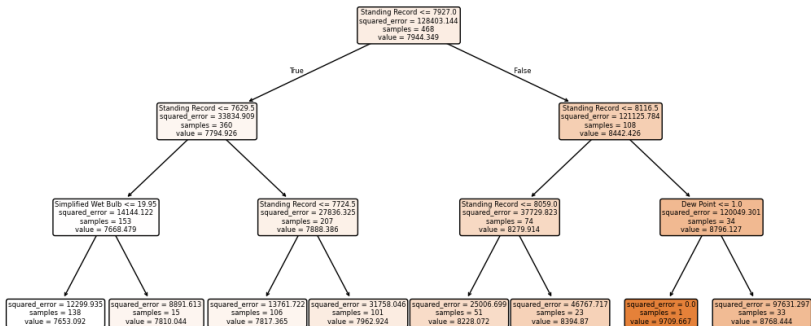
$$f(x) = \sum_{m=1}^M c_m I(x \in R_m)$$

Binary partition is constructed via greedy algorithm. Define

$$R_1(j, s) = \{X | X_j \leq s\}, \quad R_2(j, s) = \{X | X_j > s\}$$

The next step is solving the following for j and s

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$



Model Performance






$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Model	NRMSE		R^2 score	
	Train data	Test data	Train data	Test data
Ridge	0.487	0.506	0.763	0.744
Lasso	0.487	0.502	0.762	0.747
Decision Tree	0.342	0.541	0.883	0.706

Table 1. Performance of models based on NRMSE and R^2 score

References

-  Mantzios, Konstantinos, et al. Effects of weather parameters on endurance running performance: discipline-specific analysis of 1258 races. *Medicine and science in sports and exercise* 54.1 (2021): 153.
-  Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. (2009).
-  Abdulhafedh, A. (2022) Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest.
-  R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, IJCAI'95*, Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143
-  Brimicombe, Chloe, et al. Wet bulb globe temperature: Indicating extreme heat risk on a global grid. *GeoHealth* 7.2 (2023): e2022GH000701.