University of Pisa

Master's Degree Program in Computer Science

Data Mining in Cycling: From Raw Data to Predictions

By

Riccardo Marcaccio

Margherita Pensa

Oleksiy Nedobiychuk

0.1 Introduction

Our work is about to analyze and extract a perfect clean data for usage of machine learning training. This work is applied to 2 different datasets: cyclists dataset and race dataset. Sequence of action made for elaborate datasets are:

1) Analyze distribution and correlation between features. 2) Feature engineering for managing missing values

0.2 Data understanding

0.2.1 Missing values

Scraping

During the data quality assessment task, we identified three fundamental issues with the races' dataset:

- (UCI) Points duplication:
  Each cyclist in a particular race received the same number of points, which were identical to those awarded to the first-place cyclist.

- Incorrect final position:
  We discovered that some cyclists were disqualified from certain races, yet the dataset retained their original positions unchanged. The substitute and subsequent riders were incorrectly shifted down by one position.
  Example: If 15th placed rider was disqualified, the replacement one was incorrectly listed in 16th position.

- Negative deltas:
  We found out the cause of negative deltas was a different representation of the same value. The first positioned rider has a finishing time, for example, 0:40:00, which stands for the hours, minutes, and seconds needed to arrive as number one. The subsequent ones were given either a positive value, which means the arrived time after the first one, or a negative value, which express the hours, minutes, and seconds required to cross the finishing line, for example: -0:42:00. Therefore, the correct delta was assumed to be 0:42:00–0:40:00 = 2:00.

Lastly, one issue with cyclists dataset:

- Missing cyclists:
  In addition, we realized that some cyclists were only present in the race dataset but not in the cyclist dataset.

After careful consideration, we decided to obtain a better version of the two dataset by scraping this website: *https://www.procyclingstats.com*

0.2.2 Correlation between features

We begin the feature correlation analysis with the correlation matrix between all the numerical features of the two datasets combined. We observed a strong expected

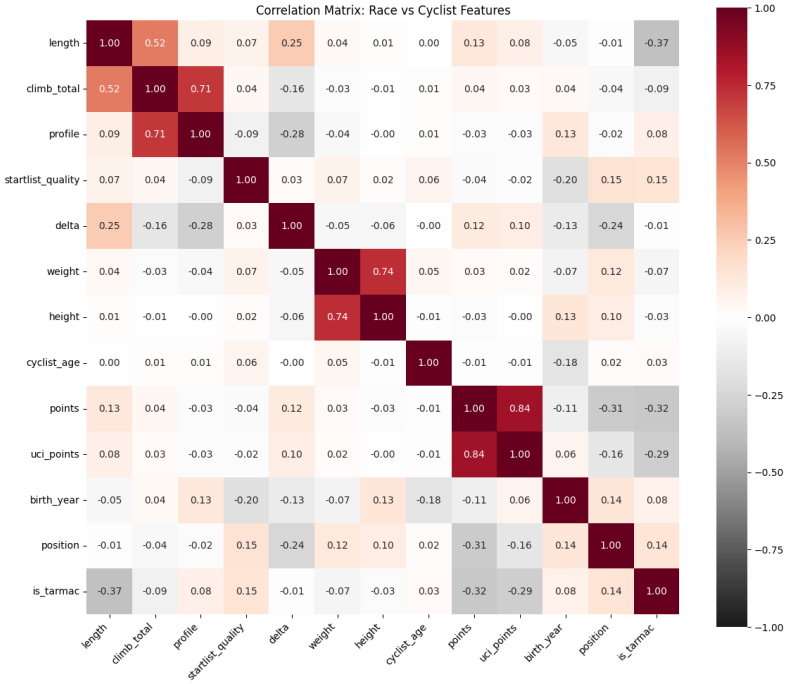

Figura 1: Correlation matrix between all relevant features.

positive correlation between *height* and *weight* (0.74), *climb_total* and *profile* (0.71), *points* and *uci_points* (0.84). In addition, there is a significant positive correlation between *climb_total* and *length* (0.52). Less evident positive correlation is *delta* and *length* (0.25).

Regarding the negative correlation, we noted that there is a weak negative correlation between *length* and *is_tarmac* (0.36), which might indicate that the longer the race, the less likely it is to be primary on tarmac. As the second most significant negative correlation, we have the one between *points* and *is_tarmac* (0.32), which again might indicate that a race has more points, on average, the more likely it is to have different types of terrain. Furthermore, there are negative correlations such as *points* and *position* (0.31), *profile* and *delta* (0.28). The last one is counterintuitive, even though it is a feeble correlation, which might indicate that as the race profile becomes more challenging (moving from flat towards high mountains with uphill finish), the time gaps (delta) between riders tend to decrease because of riders grouping?

0.2.3 Correct values check

0.3 Data transformation

0.3.1 Fill missing data

0.3.2 Feature engineering

0.3.3 Outlier detection

0.4 Data clustering

0.4.1 K-means

0.4.2 DBSCAN clustering

0.4.3 Hierarchical clustering

0.5 Predictive models

0.5.1 Linear classification

0.5.2 Neural network

0.5.3 Support vector machine

0.5.4 Autoencoder

0.5.5 Decision tree

0.5.6 Random forest

0.6 Conclusion

Test cite[1]

Bibliografia

[1] Book Author. *Book title*. Book publisher, 2000.