Cycling, cycling, and more cycling

2024/25

The assignment is an implementation and analysis of the data mining pipeline, as highlighted during the course.

Deliverable. The deliverable is i) a report of maximum 25 pages (including figures), and ii) the source code written to create the analysis. The code is to be delivered through Github (details to be defined). Both contribute to grading.
Teams. You should work on the deliverable in teams of 3 (three) students.
Tools. The project must use tools and techniques presented during the course

Dataset description

Domain

The dataset comprises of a set of professional cycling races, spanning several years in time.

Data

The dataset comprises 2 tables, `cyclists` and `races`, detailing single races, and the participating cyclists.

Features

Features include:

Feature	Description	Example
`_url`	Identifier of the race	`tour-de-france/1978/stage-6`
`name`	The name of the race	`Tour de France`
`points`	Points assigned to the race, higher points, higher prestige	`100`
`uci_points`	Alternative points assigned to the race	`100`
`length`	Length of the race	`162000`
`climb_total`	Total meters climbed during the race	`3512`
`profile`	Race profile of increasing difficulty: flat, hilly, mountainous, high mountains, high mountains with uphill finish	
`startlist_quality`	How strong are the participants at the race?	`1241`
`date`	Race date. Starting time is irrelevant and noisy	
`position`	Finish position of the given cyclist	`3`
`cyclist`	ID of the cyclist	`sean-kelly`
`is_X`	Is the race also run on a `X` surface, e.g., on gravel?	
`cyclist_team`	Team the cyclist belongs to	`visma-lease-a-bike-2024`
`delta`	How many seconds after the first-placed did the cyclist get to the finish?	
`weight`	Weight of the cyclist	`64.2`
`height`	Height of the cyclist	`178`

Report

General guidelines. This section provides a guide on the expected report's contents. Please keep in mind that the report has the goal of detailing and highlighting your work. As such, all analyses **must** contain

- Motivations: why did you perform this analysis, rather than another one? Why did you look to create one feature/representation, rather than another?
- Thorough analysis: each analysis should be performed in a reasonably large set of settings, e.g., considering several hyperparameters for the algorithms you run, and, when appropriate, choosing a set of hyperparameters. Please justify your choices.
- Observations: what insight and/or information did you gain from each analysis you performed?
- Limitations: how strong are the analytical results, and observations you have found?

The report comprises of three tasks, detailed below.

Task 1: Data understanding

10 points

Analyze the dataset, including:

- assessing data quality
- data distribution
- relationships between features

Task 2: Data transformation

20 points

The data transformation task includes three subtasks:

1. Feature engineering and/or novel feature definition
2. Outlier detection
3. A revamped data understanding task, now including the features of point 1, and eventual considerations of point 2.

As per subtask 1., improve the quality of your data by tackling eventually missing/incorrect values, either engineering or defining novel features of interest. Features may involve the single cyclist, the single race, the team, etc. For example,

- You may partition the season into different segments, studying each cyclist on said segments
- You may study the cyclists on different terrains and race profiles
- You may study the cyclists at different ages

As examples, the above *are not* compulsory, but feature engineering and definition is. Whenever you generate novel features, please report how you did so, motivating your choices, and indicating any parameter you have set, and why. You may leverage data representation algorithms to visualize the data, and gain insight on what type of features you may be interested in.

Task 3: Clustering

30 points. 2 bonus points for additional work

Leverage clustering algorithms to i) identify, and ii) describe the groups of instances you have found. Keep in mind the general guidelines as it pertains to thorough analysis, and observations. The dataset lends itself to create clusters considering cyclists, races, or both, **including eventual features engineered in the previous data transformation task**. In any case, **all features used in the clustering task must be previously defined in the data transformation task**.
The section should consider all clustering algorithms tackled in the course:

- k-means clustering
- Density-based clustering
- Hierarchical clustering

The task **must** also present final observations and comparisons on different clusterings. Additionally, the group can experiment with additional clustering algorithms available here. Additional algorithms can yield up to 2 *bonus* points in evaluation.

Task 4: Prediction

30 points.

Using learning algorithms seen in the course, fit a prediction model that will predict the final position in a race of a given cyclist.

Data. Of all the races in the dataset, please consider as test set only the races from 2022 (included) onward. You can use the rest of the dataset as you see fit.
Labels. Rather than using raw position, consider a top-20 placement, thus defining the learning task as a binary classification task: one class indicating top placement, the other vice versa.

The analysis and model selection should include a reasonable set of models, and proper model search, tuning, and validation. Make sure to include all validation measures you deem appropriate, and compare them across different models and model families. The end result may also be a small set of, rather than a single, models.

Task 5: Explanation

30 points.
Provide explanations for the model(s) of the previous step, focusing on:

- Feature importance
- Rule explanation
- Counterfactual instances

Analyze the explanations both in terms of their own properties, e.g., fidelity and complexity, and with respect to your findings in the previous tasks of data understanding and clustering, e.g.:

- Has the model learned the same patterns the data understanding has highlighted?
- Did it highlight some new ones?
- Are there some unexpected and/or nonsensical patterns?