Do you have what it takes to win

Le Tour De France?

Looking at historic race data and rider characteristics to try and predict future winners.

By: Raminderpreet Singh Khaira

Problem Statement/Background

Like most sports cycling has its fare share of conjecture. Conjecture that I believe can be challenged using Data Science. For my project I decided to take on the Tour De France, cycling's biggest event of the year, and using Data Science come up with a few characteristics that can help determine:

1. Is there an optimal rider type?
2. What factors influence a rider's success? (such as team they ride for)

Using this information, team recruiters can focus on riders with certain body types, and background. Additionally, riders can focus on getting on certain teams, or developing themselves in a way that would lead to success at cycling's most important race.

Data

The majority of data was web-scraped from procyclingstats.com. It is a website that very carefully tracks race results and metrics on the riders. Some missing information was manually inputted from Wikipedia.

The collected data included the overall race results over the past 10 years (2011-2021). This was done to ensure technology was not influencing the modeling and results as the 11-speed drivetrain was introduced in 2011 (it has been the standard since). Additionally, the winner of TDF 2010 and certain previous ones have been caught for doping, so to avoid PEDs having an influence on the results, it was decided to only collect data to the year 2011.

A list was then created for the individual riders who competed over the 10-year period, and their metrics such as height, weight, and year of birth (for age) were collected separately.

Data Cleaning, EDA, and Feature Engineering

The Data went through a series of transformations to be usable for modeling. Some of the steps included:

1. Converted object data types to integer or float for modeling.
2. Reduced int64 and float64 data types to smaller data types (such as int8 or float16 depending on the range of the data) to reduce memory usage and model run time.
3. One hot encoded categorical data such as the Team name or Rider name
4. Calculated age during competition by subtracting year of competition from year of birth
5. Created new features to binarize winners, and binarize podium finishers for the target variable
6. Missing values were imputed by cross referencing other riders with matching stats (ex: calculate median weight for a given height and impute to the rider of that height with the missing weight)

Data was also transformed using PCA to see if it improved modeling results, but this was not done as part of EDA. With initial EDA we learn that riders who have won the TDF over the last 10 years tend to weigh around 66kg, be slightly taller 1.8m and are around 30 years of age.

Modeling

In order to get results that are actionable, a variety of supervised learning methods were applied to the data to see is good testing scores can be achieved. The following classification methods were applied:

1. Logistic Regression
2. Support Vector Machine
3. K Nearest Neighbor
4. Decision Tree

Linear Regression model was also applied.

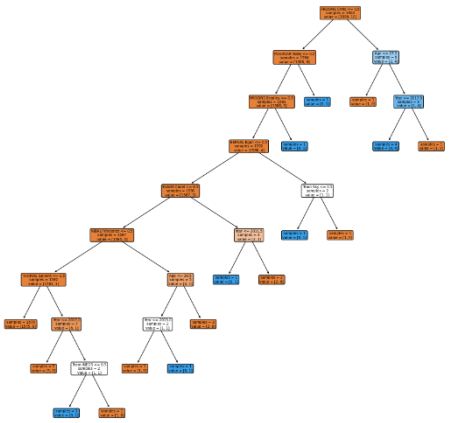
Different models, and different methods to classify the data presented their own unique challenges. While determining the 1st place is probably the most critical, it leads to a huge class imbalance, and while the models could achieve good scores, they were achieving this by labeling everyone as non-winner, defeating the purpose. A broader class system with 35 unique classes did not improve the outcome by much. The Linear Regression model trained well, but unfortunately its test scores were not that impressive.

In the end the model with the most reliable output was the Decision Tree. Being mindful the decision tree can very easily overfit to the training set, when its hyperparameters (depth) is optimized, it can very reliably predict if a rider has a chance of winning the next Tour De France. A slightly over fit Decision Tree model is not too problematic in this case, as you will not have a total random selection of cyclists enter the Tour De France, most riders are repeats, with 3 having ridden for all 10 years of data collected.

	Models	Train_Score	Test_Score
1	Logistic Regression Winner only	0.993750	0.992908
2	SVM Winner only	0.995000	0.992908
3	KNN Winner only	0.995000	0.985816
4	Decision Tree Winner only	1.000000	1.000000
5	Logistic Regression with H.O. & Points	0.785201	0.785100
6	KNN with H.O. & Points	0.788075	0.785100
7	Decission Tree with H.O. & Points	0.785201	0.785100
8	SVM with H.O. & Points	0.795259	0.785100
9	Logistic Regression with PCA & Points	0.785201	0.785100
10	KNN with PCA & Points	0.788793	0.785100
11	DT with PCA & Points	0.785201	0.785100
12	Linear Regression 1	0.872664	0.435675

Figure 1: Supervised Learning Models with Training and Test scores

Figure 2: Size of one of the Decision Trees fitted to our data



Conclusions

The modeling seems to be as not as effective as initially hoped. The data on which the modeling has been applied, along with the type of modeling leave something to be desired. A more detailed data set that collects various other metrics on the riders, such as rider type, history of racing, races won, along with more complex modeling processes such as ensemble and unsupervised learning could yield to a model that can more reliably predict who will win the next Tour De France.

The business case for this remains open, this project shows why there is so much conjecture in the sport of cycling, as clear signifiers for success are not that easy to predict in sports.