TrHackathon 2025

Data visualisation and forecasting dashboard

Marley Young

Abstract

For TrHackathon 2025 I am presenting a prototype of an interactive dashboard which aims to intuitively visualise athlete performance data and its relationship with external variables, and moreover do predictive analysis using statistical models and neural networks.

Hopefully this serves as a proof of concept for how these notions could be leveraged to improve engagement and understanding for fans, commentators and athletes alike.

As someone who trains and competes in the shot put, I decided for convenience to restrict to only using the data from the heavy throws events, but the approach would work similarly for other events.

Description

Structure of the Dashboard

I built the dashboard using the *shinydashboard* R package. In the sidebar there are the following tabs:

- Athletes
  - Summary: overview of an athlete's career performance data.
  - Head to head: plots comparing athletes' career trajectories and effects of external variables on their performances.
- Forecasting: in the case of the men's discus throw, predictions for the 2024 European championships final (and events leading up to it) based only on prior data.
- Weather: plots showing the effect of external variables, such as temperature, on athlete performance.
- Events: plots showing differences in seasonal and overall performance trends between the different heavy throws events.

Innovation

Statistical Modeling and Forecasting

While none of the tools of data visualisation or statistical modeling used in this dashboard are themselves novel, many are new or under utilised in the context of athletics.

I used the R *fpp3* package to implement various typical time series models and compute forecasts. These models included:

- STL decomposition
- ARIMA models
- Neural network models

Exogenous Variables

I haven't seen many instances of athletics performances being linked to external data outside of a scientific context.

I used the free open-source weather API *open-meteo* to obtain data for the following variables (at hourly intervals, which were then averaged over competition timeframes):

- Temperature (at 2m above ground)
- Apparent Temperature
- Humidity
- Precipitation
- Wind speed (at 10m above ground)

I also used the R package *elevatr* to access elevation data. There are very strong associations between some of these variables (particularly temperature and humidity) and performance outcomes in the data.

Impact

Data Visualisation

Viewers of an athletics broadcast, prior to an event, are really only given a few pieces of information about each athlete: their personal best, season's best, and sometimes their accolades at major championships. The commentator may give extra context, but this is often limited, especially in the field events. One ought to raise a number of questions, for example:

- How consistent are the athlete's performances?
- Do they tend to underperform at major championships, or do they exceed expectations?
- What were the conditions like when they achieved their season's best, and how do they compare to the conditions of the present competition?
- Did they peak earlier in the season, or are they still improving?
- What has been their seasonal/overall trend in recent years?

Data Visualisation

All of these questions can weave a captivating narrative about the upcoming competition, and understanding them leads to a better appreciation of when something amazing or unexpected occurs. Hardcore fans who follow the athletes know this well, but a casual viewer misses out on much of the excitement.

Data visualisation is one of the quickest ways to digest at least some of this understanding, and I believe we need more of it in athletics.

Forecasting

It is interesting and entertaining to try and predict athletes' future performances based on past data. There is a large amount of variance and many unobserved variables that impact performance, so any method of forecasting should be taken with a grain of salt. However, we can certainly be more informative than a naïve approach such as predicting based on season's best or world ranking.

With more sophisticated models, one could make predictions that could enhance immersion for fans or offer actionable insight for coaches and athletes. These predictions could be further refined, for example, by collecting more nuanced data on individual athletes.

Exogenous Variables

It is clear that various external factors, such as the weather, have a significant effect on athletes' performance.

Quantifying these effects can help fans, commentators, and athletes put particular performances into context, and also improve our accuracy when modeling or predicting future results

Feasibility

Feasibility

It is certainly realistic that this dashboard, or its components, could be developed into a fully-functioning application, or integrated with broadcast graphics to improve viewer engagement.

There is of course much room for improvement in this prototype, particularly when it comes to the forecasting. I will discuss some of the current limitations, and how they could be overcome.

Limitations - Exogenous Variables

There are various limitations to my approach within the scope of the given TrHackathon dataset. For example:

- I don't think I 100% accurately geolocated every venue location.
- I was not able to retrieve the exact time of day of each event. This makes the estimated weather variables less accurate (especially in the case of precipitation, where there should only be a substantial effect if it is raining during the event).
- I did not try to take into account wind direction (as it should be considered relative to the orientation of the track); this is very important for the discus throw.

These will all substantially affect the accuracy of data visualisation and forecasting, however, they can in principle be overcome with more data collection.

Limitations - Forecasting

The time series forecasting approaches I used are not very effective for longer-term forecasts, as they have no way of anticipating long-term trends e.g. when the athlete's performance will peak and subsequently start declining. They also behave poorly with outliers and sparse data points.

This could be overcome by integrating the usual time series methods with a hierarchical/partially pooled model, which would base an athlete's forecasted career trend on the data present for older athletes. I think this would be particularly effective given a larger dataset, including lots of results for athletes later in their careers. However, I did not implement this at this time.

In general, as seems to always be the case with statistical modeling, I believe significant improvements could be made in terms of model and parameter selection given more time and thought, and the right scientific assumptions.

Thank You

Thank you for viewing this project, and thanks to those at AthTech who organised the TrHackathon event! I hope you enjoy playing around with the dashboard.

R packages used: tidyverse, shinydashboard, fpp3, imputeTS, openmeteo, elevatr, jsonlite, tidyjson, bruceR, DescTools, pkgcond, countrycode, comprehenr, slickR