PARKRUN PERFORMANCE PREDICTOR

Owen George

INTRODUCTION

Many runners strive to improve their performance and achieve faster times, but setting realistic, personalized targets can be challenging without the right insights.

This project uses historic performance data and weather conditions to generate data-driven, tailored target parkrun times.

DATA SOURCING AND CLEANING

Scraped parkrun results

- Scraped the parkrun results web pages for a specific location and saves the results.
- This produced a DataFrame with columns: Date, Position, Runner name/id, Gender, Age Group, Time.
- The “Brighton” parkrun location was used, for events 1-826 (11/03/07 - 07/12/24), but this function can be customised for different dates/locations, or to add to the existing data.

Add weather data

- Weather data was added for each parkrun.
- A function looped through each date in the initial DataFrame.
- After inputting the location coordinates, the temperature, wind speed, and precipitation levels for each date were added, using the OpenMeteo API.

Processsing / Cleaning

- The columns were converted to appropriate data types.
- Personal bests and previous run time, and average times, were added for each runner, removing first appearances

DATA SUMMARY

826
Events

12 year 9 month
Period

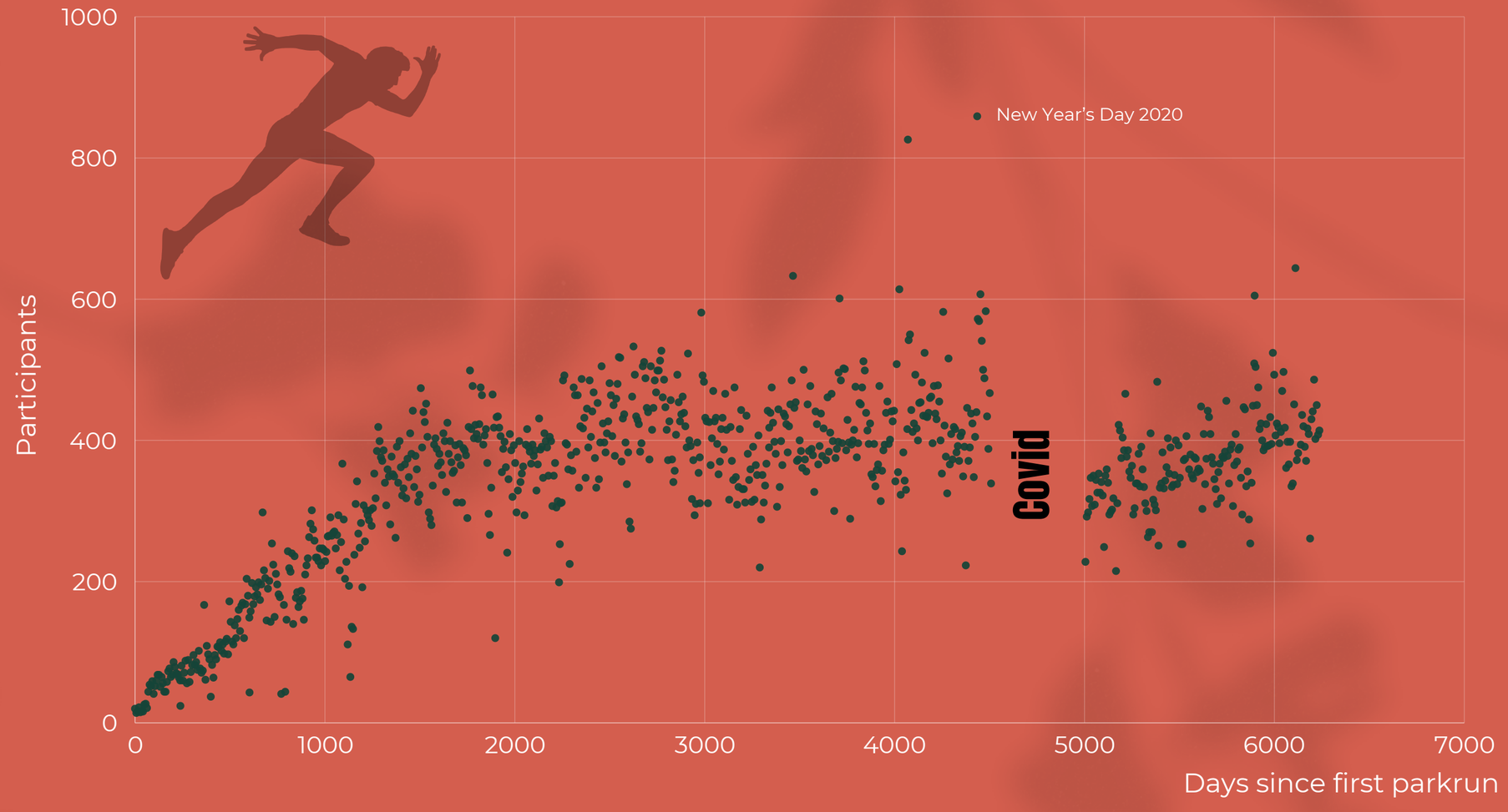
158K
Results

7,244
Unique runners

859
Peak participants

The popularity of parkruns increased dramatically since they began in 2007. Although there was a decline after it was paused for 16 months over the pandemic.

Since it returned it has been increasing again, and is now largely in line with pre-pandemic levels.



RUNNER DATA

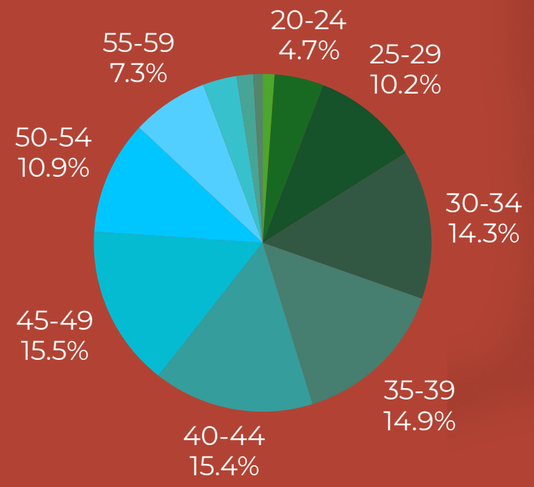
Improvement

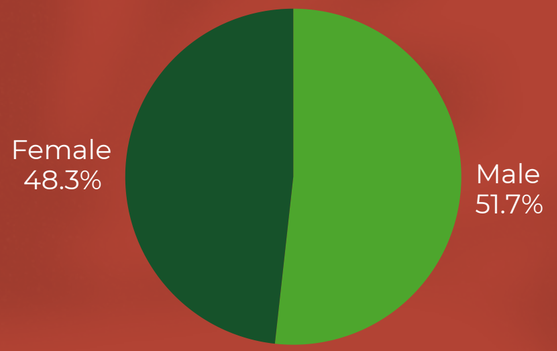
We can see that, especially for the first few parkruns, there is a clear improvement in run times

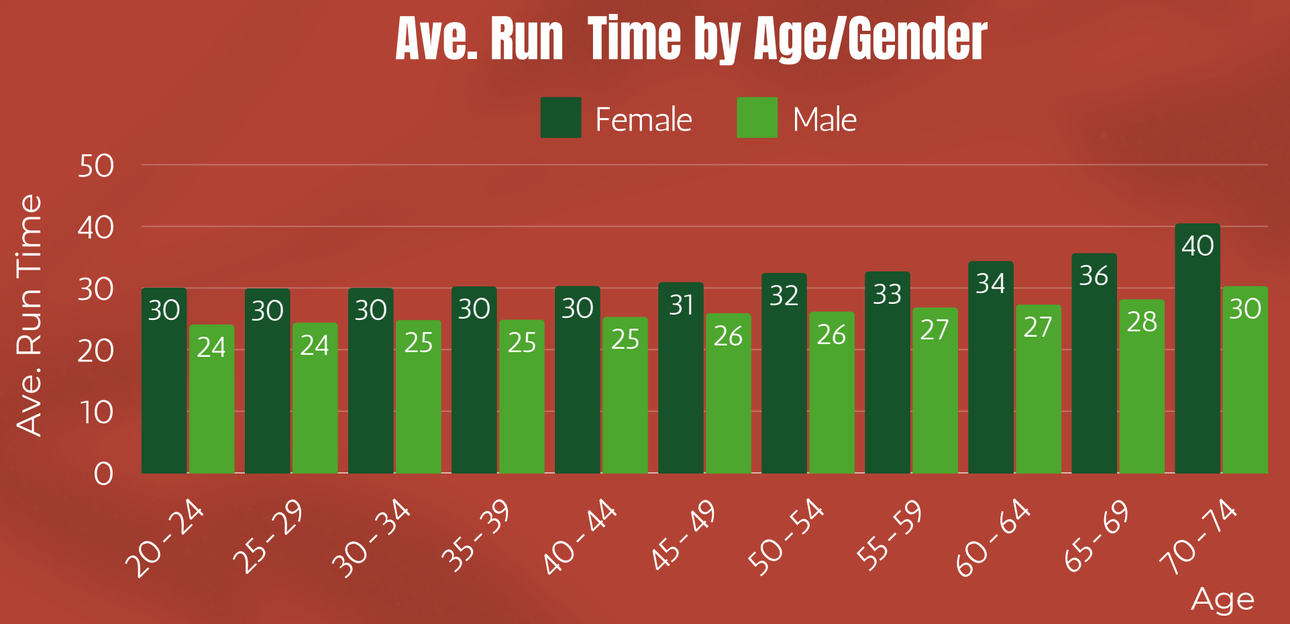
Age / Gender

Male run times tended to be quicker than female

Older runners get slower average times

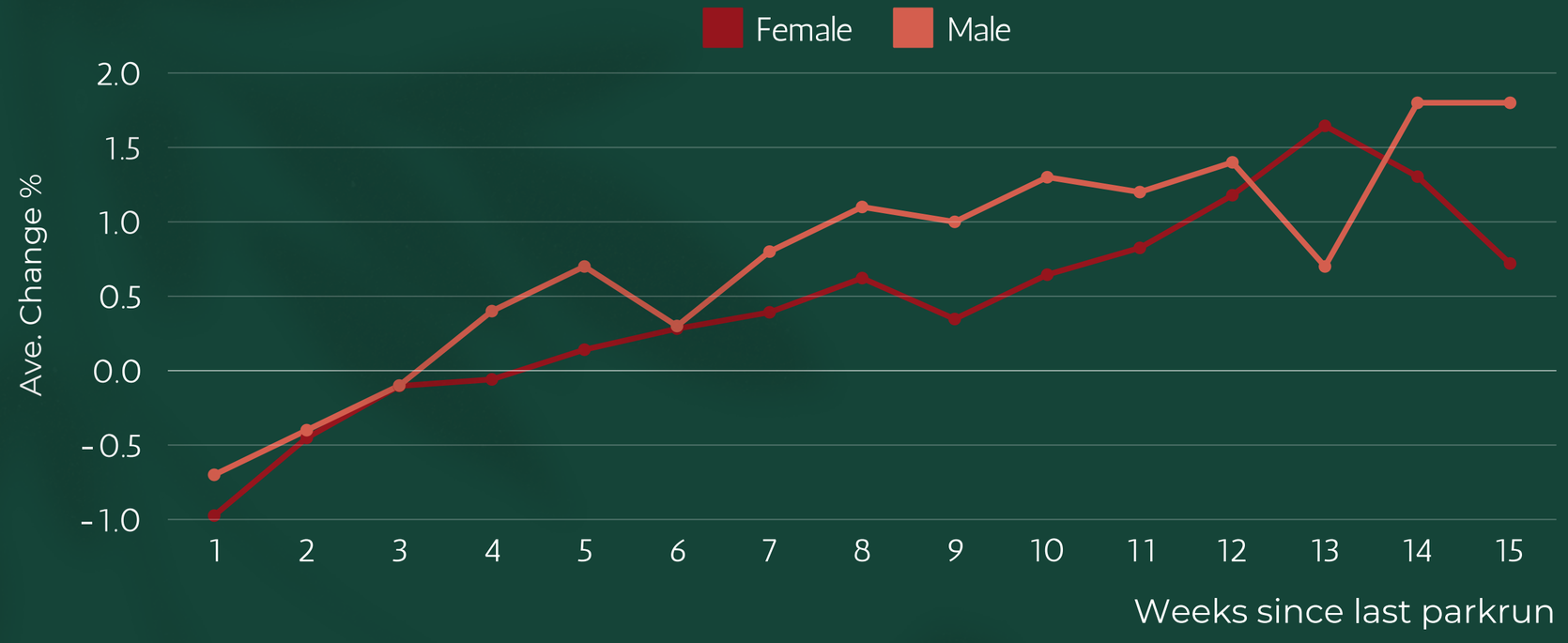


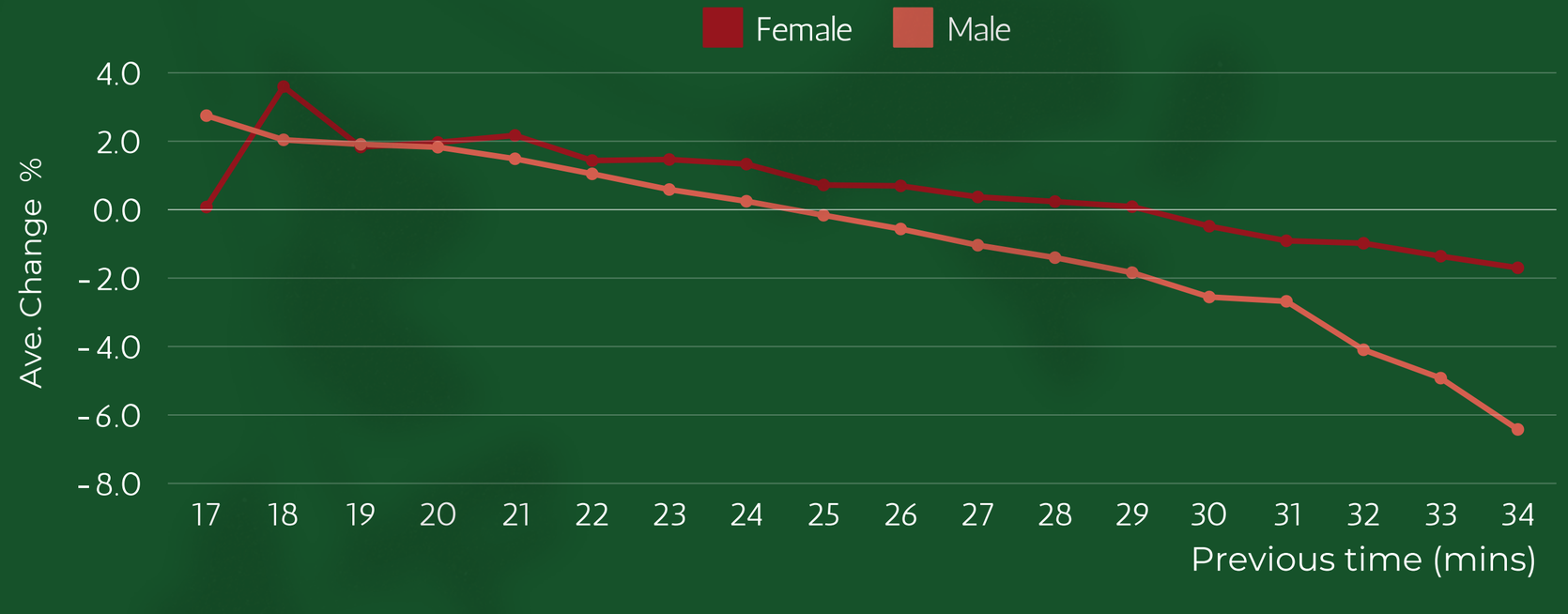




CHANGE IN RUN TIME

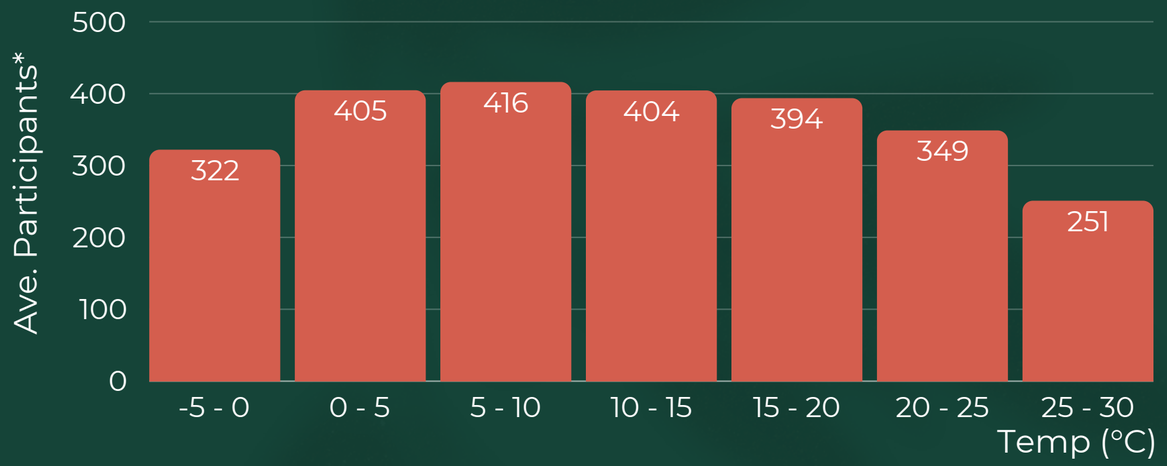
- For change in performance, the most important factors were time since last run and previous run timeds.
- Shorter gaps between parkruns resulted in greater improvements, and gaps over a month tended to result in slower next run time.
- Slower run times were more likely to be improved upon, and by a larger margin, than faster run times.

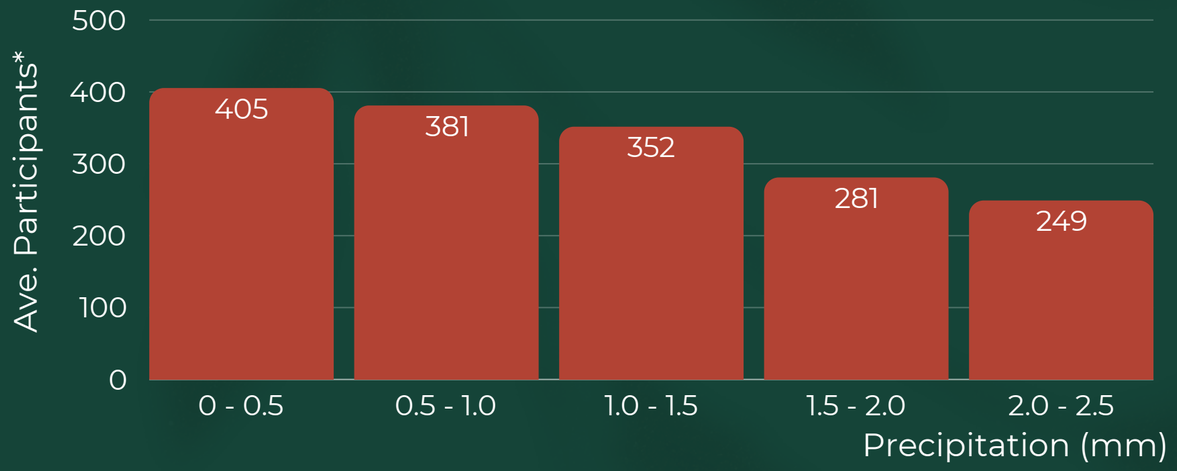


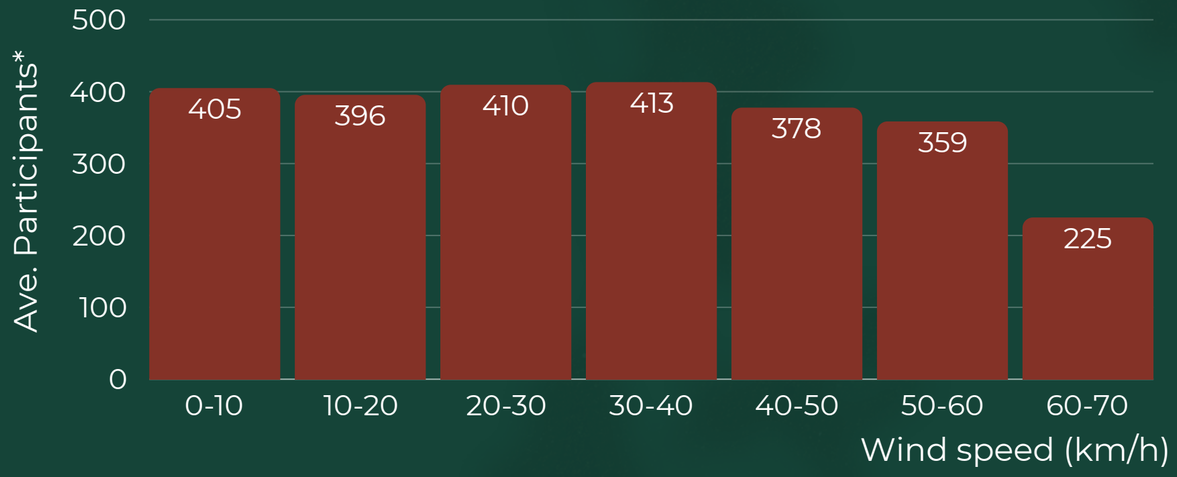


WEATHER CONDITIONS

*2014 onwards







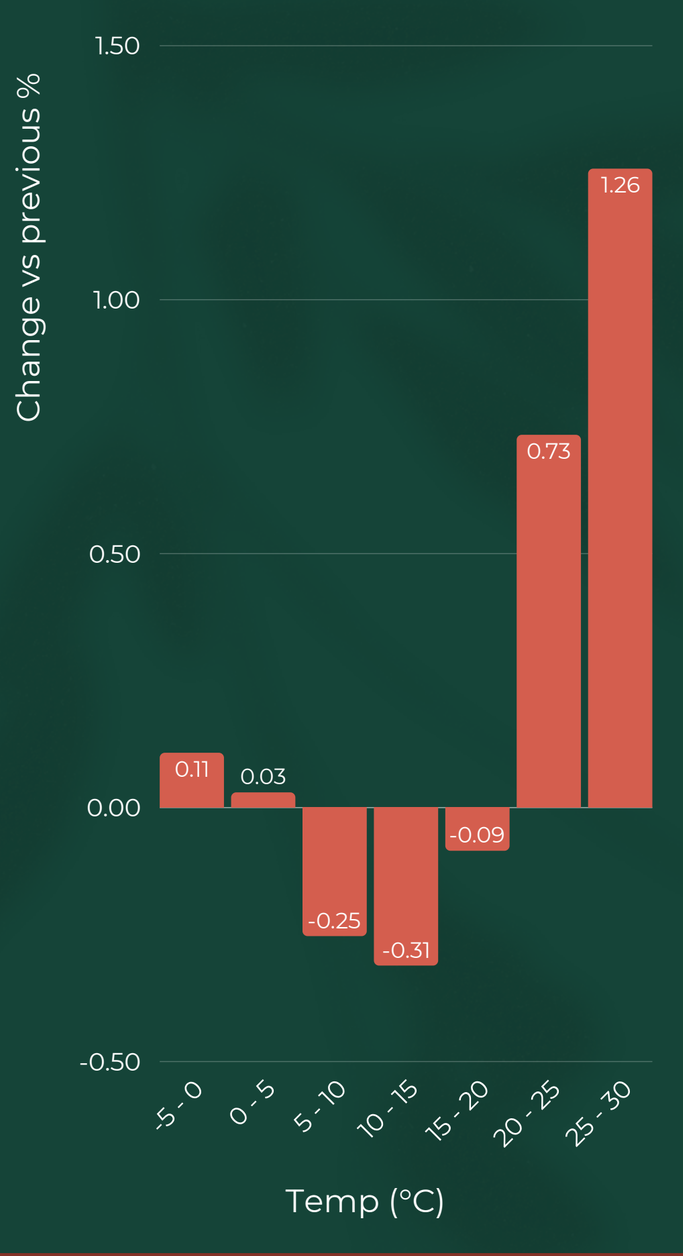
169/826
Rainy events
(20%)

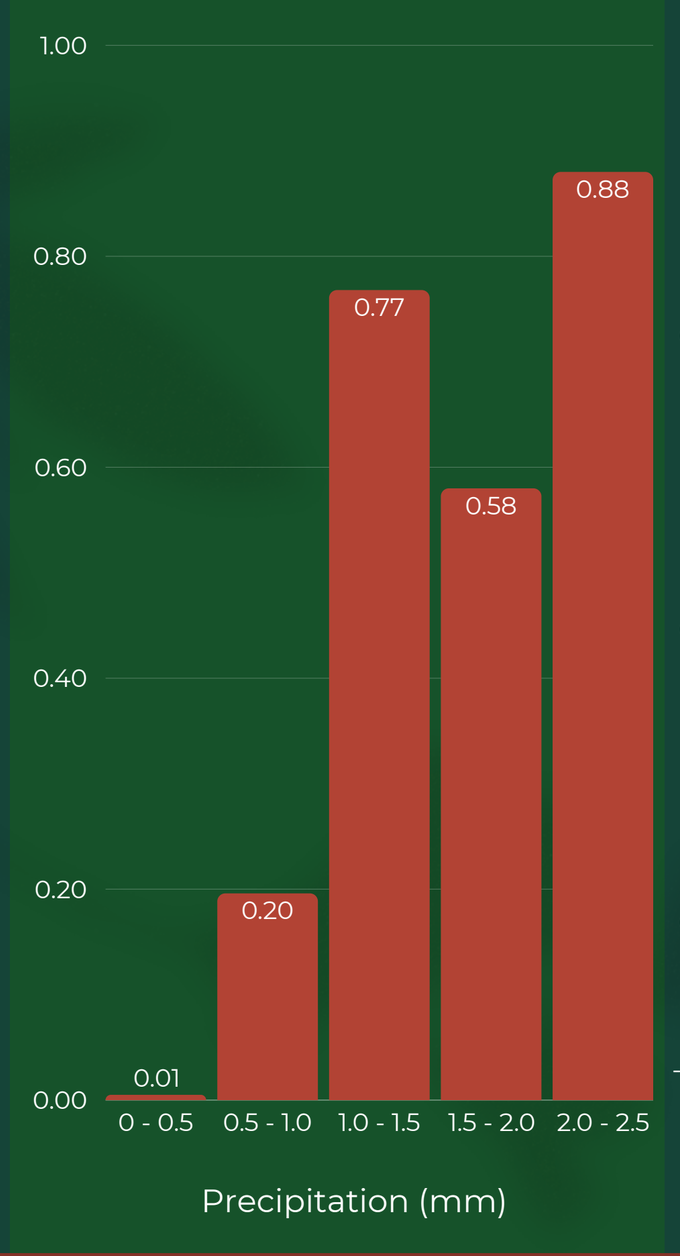
11.3 °C
Ave. Temp

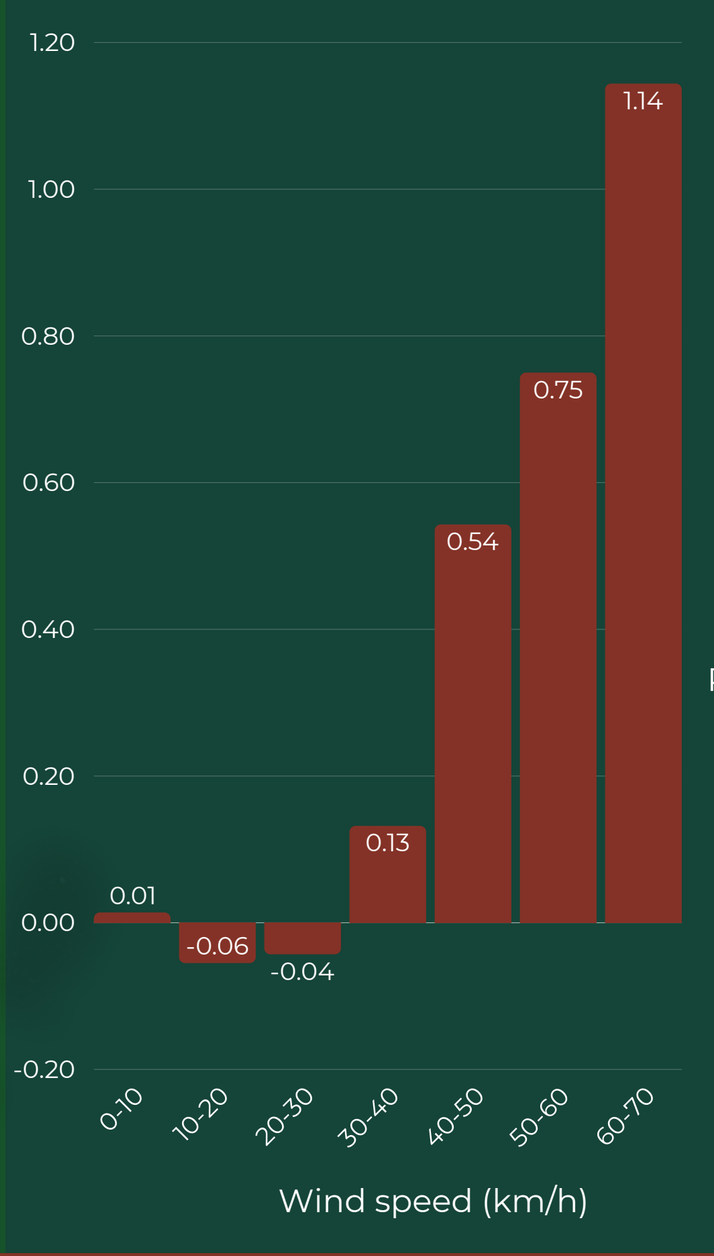
22.5 km/h
Ave. Wind speed

Weather conditions strongly impact partipation, with lower attendance at more extreme conditions

WEATHER CONDITIONS

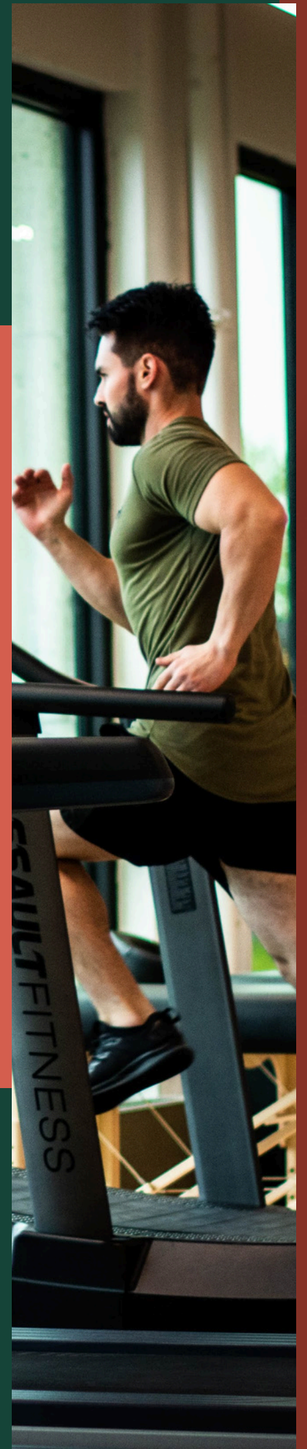






Impact of weather on change in run time was subtle, usually below 1% change, but more extreme conditions are more likely to result in worse times.

MODEL BUILDING



Data processing

- Outliers removed from “run time” and “previous run time”
- Age-group and gender converted to numeric values
- “Days since first parkrun” and “days since last parkrun” columns added
- Dropped unnecessary columns:
  - Date, Position, Name, ID, etc.

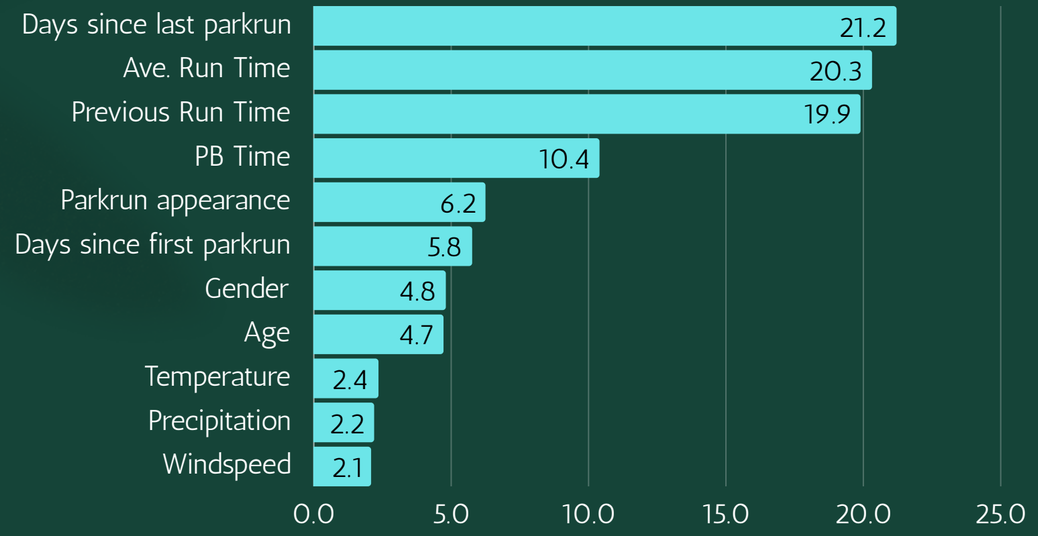
Training models

Various models were built, using a range of methods. Initial models overestimated times for faster runners and underestimated for slower runners. Therefore these were adjusted to predict the time-change-index (new time / previous time), and then calculate the time.

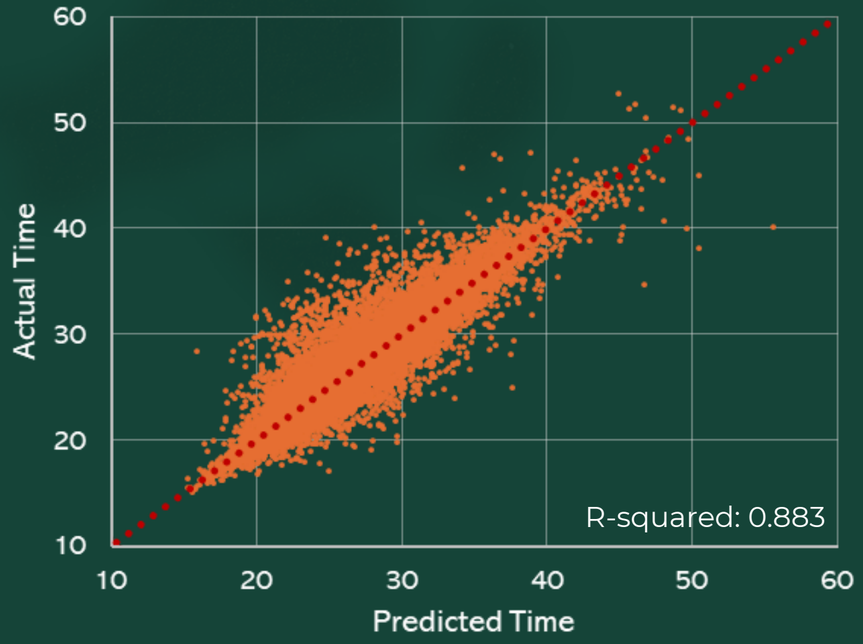
Optimisation

XGBoost was the method of the initial models that produced the best predictions, so this was further optimised.

Optimised Model - Feature Importance



Optimised Model Test - Prediction vs Actual



USING THE PREDICTOR

The predictor is designed so that you can manually input your stats and it will produce a target time using the model

OR

If you have your parkrun ID, there is a function that will scrape your stats from the parkrun website and use these to produce the prediction.

Link

POTENTIAL IMPROVEMENTS

- Current reliance on scraping from the parkrun website makes the functionality susceptible to format changes. Supplementary data sources or full integration with park run data could future-proof the model.
- Further tuning and optimisation of the models could be completed, or incorporation of additional features such as terrain type, gradient, etc. could enhance predictions.
- Extend the model to support other running distances beyond 5km park runs.
- Incorporate weather forecasting API to automatically factor into targets.
- Create a more user-friendly web-interface or app with instant predictions.

THANK YOU.

GitHub Link