Analiza danych maratonu nowojorskiego 2022

Michał Klatkowski

June 2023

Contents

1 Temat projektu

Projekt analizy danych będzie związany z danymi biegaczy maratonu nowojorskiego z roku 2021. Do analizy używane będą dane związane z: płcią, narodowością i wiekiem - jako parametrami oddziałującymi na uzyskany przez konkretnego biegacza wynik - czas, który potrzebny jest temu biegaczowi na przebiegnięcie 42,195 km. Celem projektu będzie estymacja wyników na podstawie wcześniej wspomnianych parametrów.

2 Flow projektu

Projekt został podzielony na poszczególne części:

1. Zbadanie wycinka rzeczywistości, w celu znalezienia biegu, który charakteryzuje się znacznym zróżnicowaniem pod względem określonych wyżej parametrów (szczególnie zwrócono uwagę na międzynarodoą skalę biegu)
2. Uzyskanie danych biegaczy, w celu opracowania na nich projektu

3. Analiza uzyskanych danych oraz oczyszczenie danych z granicznych przypadków bądź rekordów niedających się uwzględnić w analizie

4. Wybór modelu oraz wnioski

3 Zbadanie wycinka rzeczywistości

Początkowo, należało wybrać taki bieg, który charakteryzował się poszczególnymi cechami:

- Brało w nim udział wielu biegaczy (rzędu paru tysięcy)
- Był zróżnicowany narodowościowo (bieg o międzynarodowym charakterze)
- Możliwość uzyskania szczegółowych danych o biegaczach

Wszystkie te trzy wymagania spełniał Maraton Nowojorski. Jest on bowiem biegiem znacznej rangi, w którym rokrocznie bierze udział kilkadziesiąt tysięcy biegaczy z całego świata. Dodatkowo, strona z wynikami udostępnia dane biegaczy takie jak płeć, narodowość oraz konkretny wiek (a nie kategorię wiekową) biegacza. Te czynniki spowodowały, że wybór padł właśnie na ten bieg. Wszystkie dane, które w dalszej części będą wykorzystywane, pochodzą z biegu, który odbył się w roku 2022. Wiele innych znanych maratonów (np. londyński, paryski) udostępniały dane tylko i wyłącznie o kategorii wiekowej uczestnika, a nie o konkretnym wieku, co zawężałoby zakres możliwości projektu.

4 Uzyskanie danych biegaczy

Strona z wynikami tego biegu nie udostępnia publicznego API, zatem konieczne było pozyskanie danych za pomocą metody zwanej web scrappingiem. Dane związane z wynikami nie były zakodowane w samym kodzie HTML, a były dynamicznie generowane, zatem do scrappingu zastosowana została biblioteka Selenium, która imitowała aktywność użytkownika przy wchodzeniu na stronę i stamtąd pobierała dane z konkretnych elementów strony.

Poniżej przedstawiono przykład tego, jak wyglądają dane, które zostały zebrane:

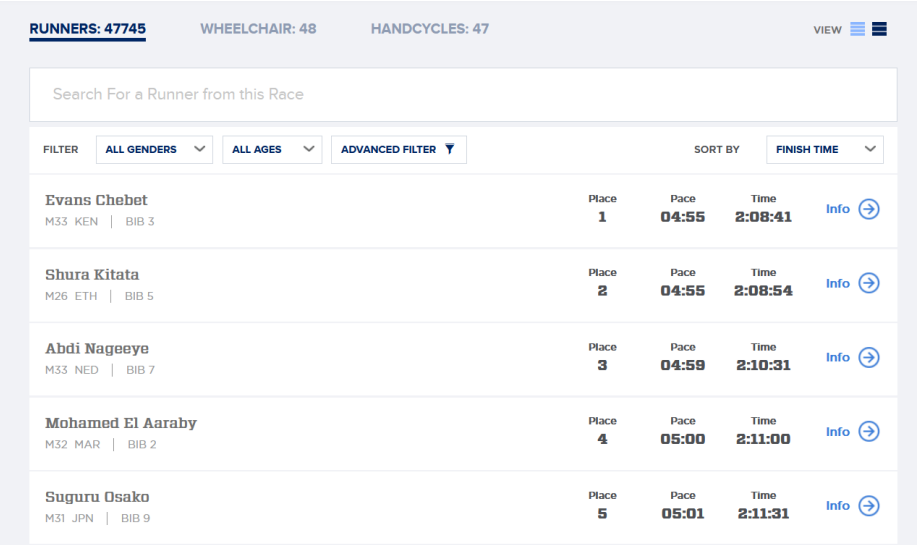

Figure 1: Przykład danych na stronie maratonu nowojorskiego

W trakcie scrapowania danych, zebranych zostało 47745 rekordów, które zostały zapisane w pliku csv o nazwie "runners-data.csv" w kolumnach o postaci:

- imię (Name)
- nazwisko (LastName)
- płeć (Sex)
- wiek (Age)
- narodowość (Country)
- numer startowy (BibNumber)
- wynik w postaci HH:mm:ss (Result)
- zajęte miejsce (Place)

Plik "runners-data.csv" z danymi oraz skrypt scrapujący "ny-marathon-scrapper.py" został załączony do dokumentacji projektu.

5 Analiza danych

Analiza danych polegała na weryfikacji, czy dane występują w odpowiednim formacie oraz czy są na tyle zbalansowane, aby mogły brać udział w dalszej analizie problemu. Pierwszym napotkanym problemem była niepoprawna wartość

pola Place, które w wartościach od 1000 wzwyż zawierało przecinek oddzielający części tysięcy. Aby można było dalej działać na danych należało się tego pozbyć, ujednalicając format tak, żeby dało się go dalej przetwarzać.

Kolejnym problemem były rekordy uczestników o nazwie "Anonymous", które nie zawierały żadnych danych. Te rekordy należało usunąć, ponieważ przeszkadzały w analizie danych.

Analiza jakościowa danych polegała na stworzeniu wykresów zależności liczby uczestników od poszczególnych parametrów (osobno: płci, narodowości oraz wieku). Początkowe dane prezentowały się następująco:

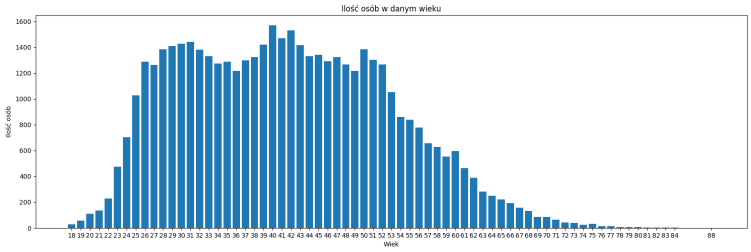

Figure 2: Początkowa zależność liczby uczestników od wieku

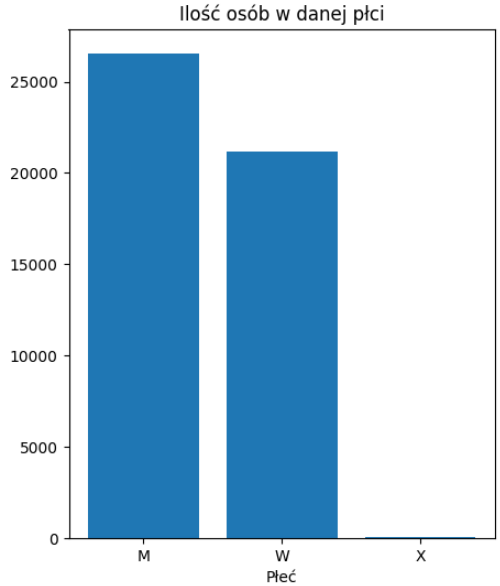

Figure 3: Początkowa zależność liczby uczestników od płci

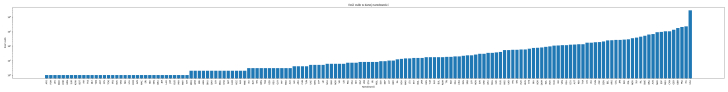

Figure 4: Początkowa zależność liczby uczestników od narodowości (skala logarytmiczna)

Z analizy powyższych wykresów można otrzymać wnioski dotyczące jakości danych:

- Niski odsetek biegaczy w wieku 75+, który mógłby spowodować zaburzenia w analizie danych
- Bardzo małoznacząca grupa ludzi o płci niedookreślonej (X)
- Duża liczba narodowości, z których pochodziła niewielka część biegaczy

Po przeanalizowaniu powyższych wniosków zastosowane zostały następujące zmiany:

- Usunięcie danych biegaczy, których wiek przekracza 75 lat
- Usunięcie danych biegaczy o niedookreślnej płci
- Usunięcie danych biegaczy z krajów, które były reprezentowane przez mniej niż 50 zawodników

Po powyższych zmianach zależności liczby biegaczy od poszczególnych parametrów wyglądają następująco:

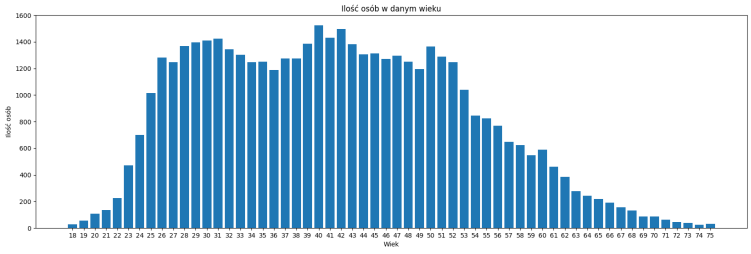

Figure 5: Końcowa zależność liczby uczestników od wieku

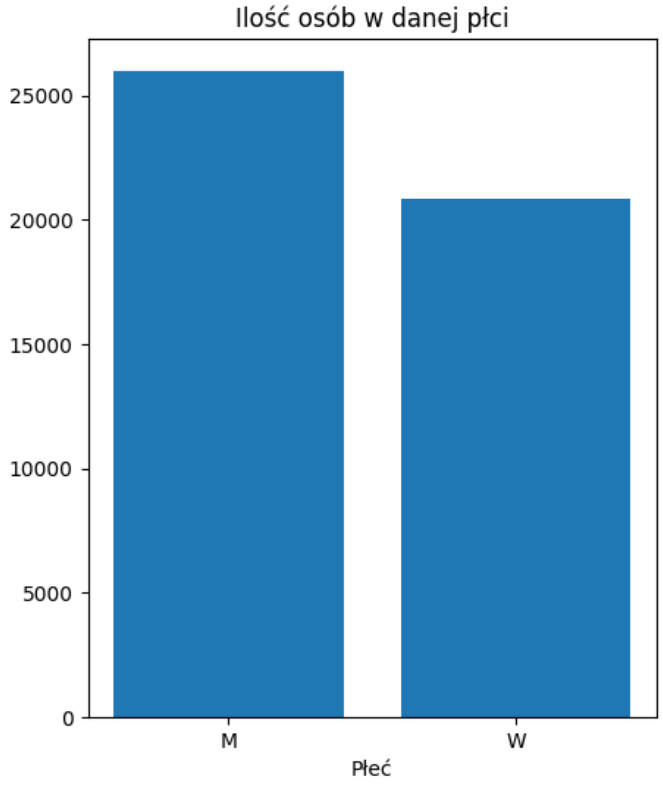

Figure 6: Końcowa zależność liczby uczestników od płci

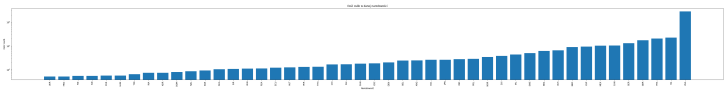

Figure 7: Końcowa zależność liczby uczestników od narodowości (skala logarytmiczna)

Po wszystkich zmianach, liczba danych zmniejszyła się do 46811 rekordów biegaczy. Plik "cleared-runners-data.csv" z danymi oczyszczonymi, skypt czyszczący "ny-marathon-clearing-data.py" oraz skrypt analizujący "ny-marathon-analysis.py" został załączony do dokumentacji projektu.

6 Wybór oraz trenowanie modelu

Celem projektu jest estymowanie czasu biegacza o zadanych parametrach. Ten etep został podzielony na parę mniejszych zadań:

- Dostosowanie danych do tego, aby można było zastosować mechanizmy udostępniane przez Pythona
- Podział danych na dane testowe oraz treningowe
- Dobór odpowiedniego modelu

Pierwszym etapem była transformacja danych nienumerycznych, tj. płci oraz narodowości do danych, które możnaby było dalej analizować. Zarówno do narodowości, jak i płci posłużył LabelEncoder, który zakodował poszczególne wartości z danych na liczby naturalne. Oprócz tego, czas biegu każdego biegacza został zamieniony z formatu "hh:mm:ss" na liczbę sekund.

Kolejnym krokiem był podział na dane treningowe oraz testowe. Zaproponowany został standardowy podział dostępnych danych w proporcjach 80 do 20.

Ostatnim etapem był dobór najlepszego modelu. Spośród modeli:

- regresji liniowej
- regresji wielomianowej
- drzewa decyzyjnego
- lasu losowego
- SVR

najlepszy rezultat dała regresja liniowa. Pomiar jakości modeli odbył się poprzez wyszukanie najmniejszego średniego błędu bezwzględnego (MAE). W przypadku regresji liniowej wyniósł on około 45 minut.

W przypadku wyboru danych biegaczy, którzy ukończyli bieg na pierwszych 1000 pozycjach, błąd ten wyniósł już tylko około 7 minut.

Skrypt "ny-marathon-model.py" z opisanymi wyżej działaniami został załączony do dokumentacji projektu.

7 Wnioski

Zbiór danych który został użyty do projektu był zbiorem bardzo zróżnicowanym. Biegacze występujący na maratonie byli zarówno amatorami oraz zawodowcami. Zarówno to, jak i znaczna liczba biegaczy, która sukcesywnie wbiegała na metę spowodowało, że dokładna predykcja wyniku danego biegacza na zadanym zbiorze jest bardzo trudna. Z modelu można uzyskać ramy czasowe, które ze znacznym prawdopodobieństwem będą mogły określić w jakim przedziale czasowym będzie znajdował się wynik biegacza o poszczególnych parametrach.

Ciekawym przypadkiem okazało się ograniczenie zbioru danych do pierwszych 1000 biegaczy. W takim przypadku model lepiej dopasowywał się do danych, oraz predykcja wyniku była łatwiejsza do określenia. Spowodowane to było profilem zawodników z początku stawki, których można określić jako zawodowców. W tej grupie wyniki były bardziej ustatkowane, a co za tym idzie łatwiejsze do estymacji.

Hipotezy które można było wysnuć jeszcze przed rozpoczęciem projektu okazały się prawdziwe: często z racji na warunki fizyczne szybsi byli mężczyźni, oraz osoby młodsze.