PAUL DE OCTOPUS 2.0

Hoe we met statistiek en machine-learning het WK voetbal probeerden te voorspellen

Hans Van Eetvelde & Christophe Ley

Afgelopen zomer vond de 21e editie van het wereldkampioenschap voetbal plaats in Rusland. Een editie zonder Nederland helaas, dat zich voor het eerst sinds 2002 niet wist te kwalificeren. Desalniettemin blijft het WK voetbal één van de grootste sportevenementen ter wereld en een vierjaarlijks hoogtepunt voor de voetbalfans. Het is een populair onderwerp in de media, maar ook in de kroeg of in de huiskamer met als hoofdvraag: wie wordt wereldkampioen? Vaak worden tussen fans onderling pronostiekjes georganiseerd en ook voor de gokindustrie zijn dit hoogtijdagen, aangezien vele voetbalfans een gokje durven wagen op hun toernooifavoriet. De beroemde Paul de Octopus werd in 2010 bekend door het voorspellen van de wedstrijden van Duitsland op het WK voetbal in Zuid-Afrika. Ook wij waagden onze poging om de winnaar te voorspellen, niet op basis van ons buikgevoel of een octopus, maar met een nodige portie statistiek en machine-learning-technieken.

We gingen als volgt te werk: we verzamelden een heleboel gegevens over de ploegen die deelnamen aan het WK. Deze gegevens vormen de input voor een model dat het verwacht aantal doelpunten voor beide teams voorspelt in een wedstrijd. Met dit model kunnen we dan de wedstrijden van het WK gaan simuleren en de winnaar voorspellen. Zie ook figuur 1.

De informatie over de ploegen

De data die we in ons model ingeven is zeer gevariëerd. We verzamelden geografische en economische factoren van elk land, zoals de populatie en het BNP per persoon. Het model houdt ook rekening met het thuisvoordeel, zowel op nationaal als continentaal vlak. De kwaliteit van de spelerskern werd geïncorporeerd aan de hand van een aantal variabelen zoals de gemiddelde leeftijd van de spelers, het aantal spelers dat in de halve finales van de Champions League en Europa League stonden in het jaar van het WK, enzovoort. Ook de leeftijd en het aantal jaren ervaring van de trainer werden mee in rekening gebracht. De belangrijkste factoren zijn echter degene die gebaseerd zijn op de voorbije resultaten van de nationale teams, zijnde de World Football Elo ratings, de kansen op eindwinst volgens de bookmakers en ten slotte een eigen rating, gecreëerd met behulp van de maximum-likelihood-methode (zie beneden). We verzamelden deze gegevens niet enkel voor het laatste wereldkampioenschap, maar ook voor deze van 2002, 2006, 2010 en 2014 om te kijken op welke manier deze variabelen invloed hebben op de uitslag, met andere woorden om ons model te trainen.

Onze eigen teamratings

Om een rating toe te kennen aan de nationale ploegen bekeken we de wedstrijden van alle nationale teams in de voorbije acht jaar. Veronderstel dat we een wedstrijd hebben tussen ploegen A en B, waarbij ploeg A thuis speelt. We veronderstellen dat het verwacht aantal goals g in een wedstrijd gemodelleerd wordt als volgt:

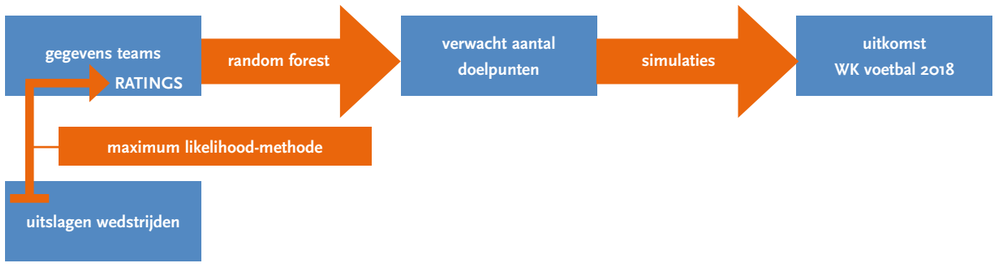
$$g_A = c \times t \times \frac{r_A}{r_B}$$

$$g_B = c \times \frac{r_B}{r_A}$$

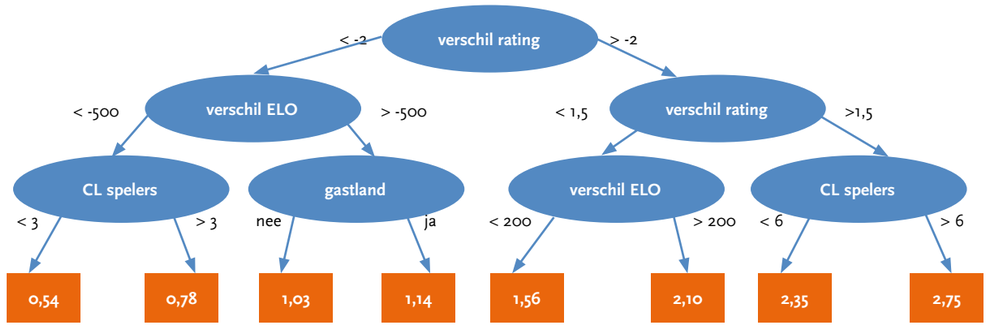
waarbij r_i de rating van team i weergeeft, en t een parameter is die het effect weerspiegelt van in het eigen stadion en voor het eigen publiek te spelen. Als de wedstrijd op neutraal terrein plaatsvindt valt deze parameter weg uit de formule. c kan geïnterpreteerd worden als het verwacht aantal doelpunten wanneer beide teams even sterk zijn en waarbij geen van hen een thuisvoordeel heeft. Al deze parameters moeten positief zijn, zodat ook het verwacht aantal doelpunten positief is. We veronderstellen nu dat het aantal doelpunten voor team A en team B in een wedstrijd een Poissonverdeling volgen. Daardoor kunnen we aan elke uitslag een bepaalde kans toekennen. Om deze ratings r_i, het thuiseffect t en de constante c te schatten, maken we gebruik van de maximum-likelihood-methode. Daarbij voegen we gewichten toe aan de wedstrijden, zodat meer recente wedstrijden meer doorwegen dan wedstrijden van een aantal jaar geleden. Ook geven we meer gewicht aan belangrijke wedstrijden, zoals het wereldkampioenschap en de continentale kampioenschappen, dan aan vriendschappelijke wedstrijden.

Random Forest

Als input voor het random forest gebruiken we de gegevens van een ploeg en zijn tegenstander of – voor sommige variabelen – het verschil tussen de waarde voor beide



Figuur 1. Schematische weergave van de aanpak om te komen tot het voorspellen van de uitslag van het WK voetbal



Figuur 2. Voorbeeld van een regressieboom

teams. Het random forest geeft ons dan een schatting voor het aantal doelpunten dat deze ploeg zal maken tegen deze tegenstander. Een random forest is een samenstelling van vele verschillende regressiebomen. De regressiebomen worden getraind op basis van de voorbije WK's. Een regressieboom kan bijvoorbeeld de vorm hebben, zoals weergegeven in figuur 2.

Het random forest genereert vele verschillende regressiebomen, door telkens een random steekproef uit de trainingsdata te nemen en bij elke vertakking een steekproef uit de variabelen. Een schatting wordt nu gemaakt door het gemiddelde te nemen van de uitkomsten van al deze regressiebomen (hier 5000). Zo zie je dat het model heel wat meer armen heeft dan Paul de Octopus.

Simulaties

Nu we voor elke (mogelijke) wedstrijd op het wereldkampioenschap het aantal goals voor beide ploegen kunnen schatten, zullen we het WK simuleren. Het aantal goals voor beide teams in elke wedstrijd wordt gesimuleerd door een trekking te nemen uit een Poissonverdeling met als enige parameter het gemiddelde dat we berekend hebben aan de hand van het random forest. Nadat we elke wedstrijd in de groepfase gesimuleerd hebben, maken we de stand in elke groep op en bekijken we welke ploegen tegen elkaar spelen in de volgende ronde. Deze wedstrijden worden op dezelfde manier gesimuleerd. Bij gelijkspel worden ook de verlengingen gesimuleerd en bij penalty's heeft elke ploeg 50% kans om door te stoten. Zo simuleren we ook de verdere rondes totdat we de winnaar van het toernooi kennen. Deze procedure herhaalden we 100.000 keer, waaruit we de kansen op winst voor elke ploeg haalden.

Resultaten

Het model voorspelde dat Frankrijk 10,8% kans had om het WK te winnen. Enkel Spanje (13,7%) en Duitsland (11,5%) kregen een hogere kans toegewezen. Brazilië (10,3%) en België (9,9%) vervolledigden onze top vijf. Als we kijken op het niveau van de wedstrijden zien we dat het model in staat was om meer wedstrijden correct te voorspellen dan sommige bookmakers en dat we ook een betere 'Rank Probability Score' haalden, wat als de beste maat wordt beschouwd voor het beoordelen van voorspellingen van voetbaluitslagen. Ten slotte namen we deel aan een online competitie http://fifaexperts.com waar we tweede eindigden op meer dan 500 teams (je vindt ons resultaat onder de naam Andreas Groll). Het blijkt dus dat het combineren van klassieke statistische methoden, zoals de maximum-likelihood-methode, en machine-learning-technieken als random forests uitstekende voorspellingen oplevert.

Hans Van Eetvelde is een doctoraatstudent in de wiskunde aan de Universiteit Gent onder begeleiding van prof. Christophe Ley.
E-mail: hans.vaneetvelde@ugent.be

Christophe Ley is docent Wiskundige Statistiek aan de Universiteit Gent, en vice-president van de Luxembourg Statistical Society.
E-mail: christophe.ley@ugent.be

20

Van droom naar glorie

Toen ik in 1998 voorzitter werd van de Vereniging voor Statistiek en Operations Research stelde ik mezelf twee hoofddoelen. Beide hadden te maken met een zoektocht die al jaren binnen de Vereniging speelde. De eerste zoektocht betrof de publicaties van de Vereniging. De tweede zoektocht betrof de vraag of de SOR (Sectie Operations Research), die haar naam wijzigde naar NGB (Nederlands Genootschap voor Besliskunde), nu wel of niet binnen de Vereniging moest blijven.

De Vereniging kende in de jaren negentig drie soorten publicaties. Dat gaf een zware financiële last en de tevredenheid over de bladen was niet groot. Het *VVS-bulletin*, waarmee mededelingen van de Vereniging door de postbode bij de leden thuis werd bezorgd, was door de tijd ingehaald. De website en e-mail werkten toen al beter. *Statistica Neerlandica* werd door een deel van de Vereniging zeer gewaardeerd. De leden met een besliskunde achtergrond haalden het blad echter meestal niet uit de verpakking. Voor hen bestond het blad *Kwantitatieve Methoden*. Dat had echter een roestige uitstraling en artikelen werden niet breed gewaardeerd. Het besluit om het *VVS-bulletin* als mededelingen blad af te schaffen was niet moeilijk. Echter het *VVS-bulletin* was feitelijk ook het enige bindmiddel tussen de Secties van de Vereniging. Om dat op te lossen wilden we twee vliegen in één klap slaan. Een nieuw blad dat *Kwantitatieve Methoden* kon vervangen; dat zou helpen in de popularisering van ons vakgebied en dat voor zowel de leden met een besliskunde als een statistiek oriëntatie interessant zou zijn. *STAtOR* was natuurlijk de ideale naam. De stator is het stilstaande gedeelte in een elektromotor. Onze *STAtOR* zou het vaste, bindende element worden van onze Vereniging. Tegelijk was de naam een elegante samenvoeging van STAtistiek en OR. Om het project te laten slagen was een frisse en een meer journalistieke uitstraling nodig. Op dit vlak verrichtte Monique van Hootegem veel goed werk. De layout van het eerste (proef)nummer leeft nog steeds voort in de twintigste jaargang van *STAtOR*! Het bestuur zette zich er vol achter. De redactie van het eerste nummer bestond vooral uit bestuursleden en die hadden ook een grote rol in het schrijven van de artikelen voor de eerste nummers. Gelukkig was Dick den Hertog direct bereid om als hoofdredacteur de kar te trekken.

De komst van *STAtOR* maakte mijn tweede zelf gekozen hoofdtaak veel makkelijker. Met *STAtOR* viel de behoefte van de OR-NGB sectie om zich als een losse Vereniging te organiseren grotendeels weg. Daar ben ik tot op de dag van vandaag blij mee. Om impact te kunnen hebben is het veel beter de krachten te bundelen. De logica van die bundeling zie ik alleen maar groeien. De explosie aan beschikbare data en de toenemende (reken) snelheid van processoren creëert ongekende mogelijkheden voor onze vakgebieden. In de praktijk zie ik steeds meer toepassingen waarin technieken uit OR, Econometrie (in engere zin) en Statistiek gecombineerd worden toegepast. Big Data, Data Science, vele nieuwe termen komen op. Een gedegen achtergrond in Statistiek en OR is echter cruciaal voor gedegen, zinvolle en verantwoorde toepassingen.

Met tevredenheid kijk ik terug op het ontstaan van *STAtOR*. Het voorwoord van het eerste (proef)nummer eindigde met het uitspreken van de hoop dat *STAtOR* zou uitgroeien tot een blad waar de Vereniging trots op kan zijn. Dat is gelukt! Daarvoor dank aan velen die daaraan de afgelopen jaren hebben bijgedragen.

Gerrit Timmer, oud-voorzitter NGB en VVSOR
e-mail: gerrit.timmer@ortec.com