Memo: over de doorontwikkeling van het voorspelmodel uitval

Aanleiding

Het doel van dit memo is om toe te lichten welke stappen zijn ondernomen om het voorspelmodel voor uitval in het eerste studiejaar van nieuw instromende MBO-studenten op niveau 2 te verbeteren. In het kader van de pilot 'datagedreven werken in het MBO' die is uitgevoerd in het voorjaar van 2020 is door de werkgroep naar tevredenheid een eerste versie van een dergelijk voorspelmodel ontwikkeld. In het voorjaar van 2021 werd een vervolg aan deze pilot gegeven. Een van de speerpunten van dit vervolg is de doorontwikkeling van het onderliggende voorspelmodel.

Wat betekent doorontwikkeling?

De focus bij de doorontwikkeling ligt op het realiseren van een verbetering in de voorspellende kracht van het model: het doel is om een model te realiseren dat, wanneer we het model wordt gebruikt om de uitkomst voor studenten te voorspellen van wie de uitkomst al bekend is, minder voorspelfouten maakt.

De kwaliteit van het voorspelmodel meten we aan hand van de volgende kengetallen:

- De nauwkeurigheid van het model: welk percentage van de gevallen wordt correct geïdentificeerd?
- De sensitiviteit: het percentage van de daadwerkelijk uitvallende studenten dat correct wordt geïdentificeerd door het model. Een student met een voorspelde kans van meer dan 50 procent op uitval wordt als uitvaller geïdentificeerd;
- De specificiteit: het percentage van de studenten dat door het voorspelmodel als uitvaller wordt geïdentificeerd dat daadwerkelijk uitvalt;
- De AUC-score: een algehele kwaliteitsmaat voor voorspelmodellen, op een schaal van 0.5 voor een model zonder enige voorspellende waarde tot 1.0 voor een model dat alles perfect voorspelt.

Het doorlopen stappenplan

Stap 1: extra data toegevoegd

Om een beter voorspelresultaat te realiseren, hebben we een set nieuwe data toegevoegd. Concreet hebben we van de studenten waarvan dit bekend is gegevens toegevoegd over de prestaties in primair onderwijs. Allereerst: heeft de student in kwestie al dan niet op het speciaal basisonderwijs gezeten. En daarnaast een inschaling van hoe de student in kwestie heeft gescoord op de eindtoets aan het einde van het basisonderwijs. Deze inschaling is vanwege de diversiteit van het aanbod aan verschillende eindtoetsen geformuleerd in termen van het aantal standaarddeviaties dat de behaalde toetsscore van het gemiddelde voor alle deelnemers aan dezelfde toets afwijkt.

Stap 2: data met extra detail verwerkt

Een volgende stap ter verbetering is de toevoeging van meer detail aan de data die het verleden van de student in het VO beschrijven. Waar we in de eerdere iteraties van het voorspelmodel werkten met samengestelde indicatoren op basis van de onderliggende data over het VO-verleden van de student, hebben we nu voorgenomen om te werken met voorspeltechnieken die beter in staat zijn om de veel gedetailleerdere onderliggende data te verwerken. Bijvoorbeeld: waar we in eerdere iteraties studenten de indicator 'VO-afstromer' gaven op het moment dat de student in kwestie ergens in de loop van zij/haar

VO-carrière naar een lager niveau is doorgestroomd, nemen we nu direct de data mee die per jaar van de VO-carrière van de student aangeeft aan welk onderwijstype de student heeft deelgenomen.

We hebben met verschillende meer geavanceerde voorspeltechnieken gewerkt om met de fijnmazigere inputdata te kunnen werken die we door deze wijzigingen zijn gaan gebruiken. We hebben geëxperimenteerd met verschillende 'Deep Learning'- en 'Gradient Boosting'-technieken om deze data efficiënt te kunnen gebruiken.

Stap 3: definitie te voorspellen uitkomst aangescherpt

De derde stap ter verbetering die we genomen hebben is het hanteren van een eenduidiger definitie van de uitkomst die we proberen te voorspellen. In de voorgaande iteratie hebben we ervoor gekozen om drie verschillende uitkomsten te definiëren:

- 'Succes': de student behaalt zijn/haar diploma of stroomt uit naar een hoger MBO-niveau;
- 'Reguliere doorstroom': de student in kwestie gaat door met een opleiding op niveau 2;
- 'Uitval/afstroom' voor de studenten die ofwel afstromen naar MBO-niveau 1 ofwel uitvallen uit het MBO.

Om een beter presterend model en eenvoudiger te interpreteren modeluitkomsten te krijgen, hebben we ervoor gekozen om de te modelleren uitkomsten te reduceren tot twee uitkomsten: 'succes' of 'uitval'. 'Succes' betekent in dezen dat de student in kwestie ofwel een diploma behaalt, ofwel verder onderwijs blijft volgen in het MBO, al dan niet aan dezelfde instelling. De uitkomst 'uitval' hebben we beperkt tot die studenten die in het opvolgende jaar niet meer ingeschreven staan in een MBO-opleiding, zonder een diploma te hebben behaald in het voorgaande jaar.

Waar heeft dit toe geleid?

Het voorspelmodel zoals ontwikkeld in 2020 functioneerde redelijk tot goed. De focus van dat specifieke model lag op het voorspellen van meerdere uitkomsten. Deze keuzes leidden tot een model dat was geoptimaliseerd voor sensitiviteit. De sensitiviteit van het model in kwestie met betrekking tot de uitkomst 'Uitval of afstroom' was derhalve hoog: 89 procent: de specificiteit was lager met 29 procent. Dit correspondeerde et een AUC-score voor deze uitkomst van ongeveer 0.78 Wat betreft het model dat we voor deze iteratie hebben opgeleverd lagen de prioriteiten anders: specificiteit wordt belangrijker sensitiviteit. De belangrijkste reden hiervoor is dat een model dat geoptimaliseerd voor specificiteit beter aansluit bij de beoogde functie van het model in de applicatie: het optimaal inschatten van de kans op uitval, gegeven een set variabelen dat aangeeft wat we weten over de student bij instroom op de opleiding.

Dit heeft geleid tot een model met een specificiteit van 0.73, tegen een sensitiviteit van 0.29. Dat betekent dat we niet alle uitvallers als zodanig voorspellen, maar dat de studenten die we voorspellen wel in het overgrote deel van de gevallen daadwerkelijk uitvalt. Dat het model in geheel beter voorspelt kunnen we echter het meest duidelijk zien aan de hand van de AUC-score. Voor dit model bedraagt deze 0.81.