

Algoritme

Korte beschrijving van het algoritme.

Naam

De naam die gebruikt wordt om het algoritme aan te duiden.

Random Forest voorspelmodel

Organisatie

De volledige naam van de organisatie waar het algoritme ingezet wordt.

DataFryslân

Korte omschrijving

Een korte beschrijving van het algoritme.

DataFryslân heeft een model ontwikkeld om langdurige armoede in Fryslân te kunnen voorspellen. Langdurig is in deze context 3 jaar.

Hierbij wordt gebruikgemaakt van de verschillende sociaal minimagrenzen die gebruikt worden voor beleid: 101%, 110%, 120%, 130%, 140%, 150%.

Type algoritme

Is het algoritme zelflerend? In een niet-zelflerend algoritme specificeert de mens de regels die de computer moet volgen. Als het een zelflerend algoritme is, leert de machine over de patronen in de data.

Supervised leermodel

Methoden en modellen

Standaard methoden of modellen die het algoritme gebruikt.

Random Forest in R

Beleidssterrein

Trefwoorden over het beleidssterrein waarin het algoritme wordt ingezet.

- Sociaal
- Economisch

Status

Alleen gebruikt tijdens analyse van energiearmoede in de [Remote Access omgeving van het CBS](#)

De status van het algoritme: in ontwikkeling, in gebruik, of buiten gebruik.

Doel

Het doel waarvoor het algoritme ontwikkeld is en/of hoe de inzet ervan bijdraagt aan het behalen van die doelen.

Het doel van het model is om op de variabele importance van de verschillende variabelen inzichtelijk te hebben per inkomensgrens. De vraag is of dezelfde factoren van belang zijn in het correct voorspellen van armoede. In een Random Forest zit de feature importance ingebouwd en krijgen we inzichtelijk per afhankelijke variabele (3 jaar in de armoede per beleidsmatig sociaal minimumgrens van 101%, 110%, 120%, 130%, 140% of 150%) wat de belangrijkste factoren zijn.

Impact

*De impact van het algoritme op burgers en bedrijven.
Bijvoorbeeld: hoe werkt het algoritme en wat zijn de verwachte consequenties daarvan voor het individu of bedrijf?*

Alle data die vanuit de beschermde omgeving van het CBS worden geëxporteerd worden gecontroleerd door het CBS. Het proces staat [hier](#) beschreven. Door de waarborging van het CBS op de data output, is de impact van het algoritme op burgers daarom zeer klein, zeker gezien dat de gebruikte gegevens niet tot personen te herleiden zijn.

Proportionaliteit

Een afweging van de voor- en nadelen van de inzet van het algoritme en waarom dit redelijk gerechtvaardigd is.

Omdat de gepresenteerde waarden niet tot personen te herleiden vallen, zijn er geen nadelen op dit vlak. Daarnaast betreffen het schattingen die niet 1-op-1 passen op de realiteit. Daarnaast zijn er geen voorspelde waarden uit de Remote Access omgeving van het CBS gehaald. Het model is gebruikt om de samenhang en onderlinge rangorde van variabelen te bepalen. Met andere woorden: Welke variabelen zijn belangrijk in het voorspellen van langdurige armoede volgens verschillende armoedegrenzen? Daarnaast veronderstellen de uitkomsten geen causaliteit, maar samenhang.

Menselijke tussenkomst

Een omschrijving van hoe uitkomsten van het algoritme door een mens gecontroleerd en bijgesteld (kunnen) worden.

Op een schaal van 1 (volledig geautomatiseerd menselijk toezicht) t/m 5 (volledig menselijk toezicht).

De input voor het model zijn bepaald door de onderzoekers en de beschikbare data in het maatwerkarmoedebestand (gegevens over waar men woont, type huishouden, vermogen, rol in het gezin, ontvangen toeslagen, e.d.). Daarnaast zijn de precieze instellingen van het model ook aangepast door de onderzoekers om het model zodanig te tunen dat het beter presteert.

Dit hebben we gedaan voor de volgende reden:

- De accuratesse van het model is niet zozeer bepalend voor het beoordelen van de kwaliteit van het model, aangezien we een ongebalanceerde dataset hebben met relatief weinig huishoudens in langdurige armoede.
 - Belangrijk is dat het model correct langdurige armoede voorspelt (een '1') en niet wanneer huishoudens niet lang in de langdurige armoede zitten (een '0'). De kracht van het model is met name te zien bij de sensitiviteitswaarde. Een hoge sensitiviteitswaarde laat zien dat het model correct een 1 voorspeld.
-

Monitoring

Een overzicht van hoe de inzet van het algoritme wordt gemonitord.

Op een schaal van 1 (geen monitoring) t/m 5 (volledige monitoring).

De geautoriseerde onderzoekers van DataFryslân hebben zicht op de uitkomsten, het script en wat ermee gedaan wordt. Daarnaast heeft het CBS ook zicht op het model en de resultaten doordat alle data en scripts door de output controle moet.

Score 3 vanwege restricties tot monitoring in de Remote Access omgeving van het CBS.

Risico

Een overzicht van de voorziene risico's bij de inzet van het algoritme.

Op een schaal van 1 (zeer laag privacy risico) t/m 5 (zeer hoog privacy risico).

Het risiconiveau van de inzet van dit algoritme zijn beoordeeld als 'laag' omdat er een onafhankelijke organisatie tussen zit die expliciet de data en scripts controleert.

Score 1 vanwege de afwezigheid van gebruik in beleid, exportatie van persoonlijke gegevens, of gebruik van de voorspelde waarde.

Prestatienormen

Een omschrijving van de verwachte prestaties van het algoritme en hoe die worden gemeten.

De werking van het algoritme is gevalideerd op basis van het percentage juist herkende langdurige armoede. Daarnaast is er gewerkt met een train en valid set. De valideer dataset bestond uit 30% van de data. De train en valid dataset zijn at random gemaakt.

Databronnen

Een overzicht van de databronnen die op dit moment gebruikt worden door het algoritme en/of in het begin gebruikt zijn bij het maken van het algoritme.

CBS microdata en het maatwerkarmoedebestand.

Relatie met andere gebruikte algoritmen

Een kort overzicht van de relatie van het gebruikte algoritme met eerder werk in LABs, HUBs of andere projecten.

De gebruikte methodiek is eerder gebruikt in het project omtrent Energiearmoede en Stikstof Snuffelje.

Heeft u vragen? U kunt altijd mailen naar info@datafryslan.nl over de gebruikte AI.