

Algoritme

Korte beschrijving van het algoritme.

Naam

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

Voorspelmodel voor armoede op basis van gradient boosting.

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 armoede in Fryslân te kunnen voorspellen.

Hierbij is gekeken naar huishoudens met een huishoudinkomen onder 130% van het sociaal minimum.

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.

eXtreme Gradient Boosting (xgboost) in R
SHapley Additive exPlanations (shapviz) in R

Beleidssterrein

Trefwoorden over het beleidssterrein waarin het algoritme wordt ingezet.

- Sociaal
- Economisch

Status

Alleen gebruikt tijdens deze analyse 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 dit model was tweeledig.

1. Het bepalen van de relevante factoren voor het correct voorspellen van huishoudens wel of niet in armoede, waarbij geen direct met inkomen gerelateerde kenmerken zijn meegenomen. Bij de XGboost methode kan de relevantie van factoren worden bepaald aan de hand van de SHAP-waardes.
2. Een vergelijking maken met werking en uitkomsten van het random forest model.

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

1. Het algoritme is geschikt om interacties tussen variabelen mee te nemen. Hierdoor is een ruime set aan variabelen meegenomen. Vanwege het

doel van deze analyse zijn variabelen die direct zijn gecorreleerd aan inkomen niet meegenomen.

2. Er is een aparte analyse uitgevoerd om de parameters van het model te bepalen.
3. De tekortkomingen van het model door een ongebalanceerde dataset zijn onderdeel van de analyse.

Score 3, want inputvariabelen en parameters zijn bepaald door de onderzoekers.

Monitoring

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

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 een beschrijving van de methode door de output controle moet.

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

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.

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

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

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 voorspelde waardes. Er is gewerkt met een train- en testdataset. De testdataset bestond uit 20% van de data. De train- en testdataset zijn willekeurig gekozen.

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.

De methode is nog niet eerder gebruikt, maar is vergelijkbaar met de random forest methode die eerder

Relatie met andere gebruikte algoritmen

is gebruikt in de projecten Energiearmoede en Stikstof Snuffelje.

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

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