Interpreteren Machine Learning modellen met LIME

Machine Learning modellen zijn aan een opmars bezig binnen het werkveld van verzekeraars. De meeste verzekeraars onderzoeken de mogelijkheden voor de toepassing van deze modellen voor onder andere pricing, fraudedetectie en marketing. Een praktische hindernis bij acceptatie van deze modellen is echter dat veel van de beter presterende modellen lastig te interpreteren zijn. Dit artikel gaat in op een techniek waarmee dit nadeel wordt gemitigeerd, en laat een voorbeeld zien van de toepassing ervan.

ACHTERGROND

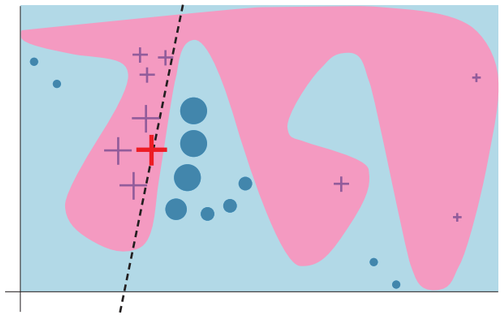
In Plat (2017)[1] zijn een aantal bekende Machine Learning modellen kort toegelicht. De relatief eenvoudige modellen, zoals (lineaire of logistische) regressie en een decision tree zijn goed interpreteerbaar. Dat wil zeggen, de uitkomsten van het model zijn goed te verklaren gegeven de input en de parameters. Echter, met name in het geval als er veel data beschikbaar is zullen de relatief complexere modellen vaak beter gaan scoren. Dat gaat vaak gepaard met minder inzichtelijkheid in het model en daarmee lastiger te verklaren uitkomsten, wat de acceptatie van het model in de organisatie lastiger maakt. Omdat dit een algemeen bekend probleem is bij Machine Learning modellen, is er de laatste tijd meer aandacht waar te nemen voor het verklaren van uitkomsten van de complexere modellen. In dit artikel wordt één zo'n techniek toegelicht.

LIME

In Ribeiro et al (2016)[2] wordt de 'Local Interpretable Model-agnostic Explanations' (LIME) techniek geïntroduceerd. Het doel van deze techniek is om een complex model 'lokaal' interpreteerbaar te maken door het benaderen van het complexe model met een eenvoudig, goed te interpreteren, model. Dit eenvoudige model kan bijvoorbeeld een lineair regressiemodel of een decision tree zijn. Het te benaderen complexe model kan ieder willekeurig complex Machine Learning model zijn. Lokaal betekent hier per datapunt, bijvoorbeeld een verzekeringspolis of een potentiële klant. De techniek wordt verder toegelicht in figuur 1.

Dr. Richard Plat AAG RBA is eigenaar van Richard Plat Consultancy en geeft in die hoedanigheid advies aan verzekeraars en pensioenfondsen op het gebied van waardering en risicomanagement.

De auteur dankt Pieter Marres voor zijn constructieve opmerkingen.



Figuur 1: toelichting LIME techniek (bron: Ribeiro et al (2016))

Figuur 1 representeert een classificatie probleem, waarbij de blauw / roze achtergrond de classificatie van punten is conform het (onbekende) complexe model. Dit model als geheel kan niet benaderd worden met een lineair model. Dit kan wel voor individuele datapunten. De volgende stappen worden genomen bij de LIME techniek:

a) De gebruiker kiest een datapunt waarvan met LIME de betrouwbaarheid vastgesteld moet worden, in dit geval het rode kruis in figuur 1;
b) Op basis van simulatie wordt *n* maal de input geschokt, en op basis hiervan worden door het complexe model *n* additionele voorspellingen gedaan (de blauwe stippen en paarse kruizen);
c) Het eenvoudige lokale model wordt gefit aan deze *n* 'nieuwe' datapunten, waarbij datapunten die veel op het oorspronkelijke datapunt lijken zwaarder meegewogen worden dan diegene die er minder op lijken (weergegeven door middel van de grootte van de stippen en kruizen);
d) De gestippelde lijn is het uiteindelijk gefitte lokale model, wat wél interpreteerbaar is.

De techniek is door Ribeiro en zijn medeauteurs geïmplementeerd in 'packages' in software programma's Python en R.

VOORBEELD

De LIME techniek is toegepast op een voorbeeld in de context van autoverzekeringen. Achtereenvolgens worden de gebruikte data, het gehanteerde Machine Learning model en de resultaten van LIME besproken.

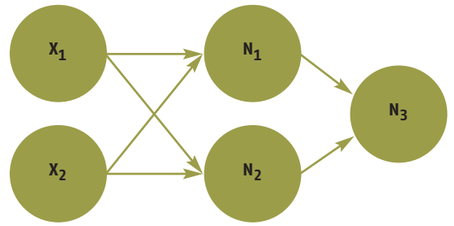
Data

Voor het voorbeeld in dit artikel gebruiken we data van de website Kaggle, waarop Data Science competities georganiseerd worden. De data in dit voorbeeld betreft een autoverzekering portefeuille van een Amerikaanse bank die ook verzekeringen verkoopt (zie https://www.kaggle.com/kondla/carinsurance). Deze bank-verzekeraar organiseert regulier telefonische marketing campagnes. De dataset bevat data van 4.000 potentiële klanten, met onder andere algemene gegevens van deze klanten (leeftijd, baan, etc.) en specifieke gegevens van de vorige campagnes (laatste contact, uitkomst daarvan, etc.). Het doel van het Machine Learning model is om te voorspellen of de potentiële klant een verzekering zal afsluiten ('Succes') of niet ('Failure'). Opgemerkt moet worden dat 4.000 datapunten veelal niet voldoende is om een complex Machine Learning aan te kalibreren, de exercitie in dit artikel is daarom ook alleen een illustratie van de uitkomsten van de LIME techniek.

Gebruikte Machine Learning model

Er zijn een veelheid van Machine Learning modellen beschikbaar, zoals verschillende regressie modellen, decision tree gebaseerde methodes en Neurale Netwerken. In dit voorbeeld is een Neuraal Netwerk gehanteerd.

Neurale netwerken zijn gebaseerd op de werking van de hersenen en zenuwen: het zijn zeer uitgebreide en onderling verbonden netwerken van kunstmatige neuronen. Zo'n netwerk bestaat uit verschillende lagen ('layers') van neuronen, waarbij de neuronen in de verschillende lagen met elkaar gelinkt zijn middels te kalibreren gewichten. Een voorbeeld van een eenvoudig Neuraal Netwerk is gegeven in figuur 2.



Figuur 2: eenvoudig Neuraal Netwerk

Dit Neuraal Netwerk bestaat uit 3 layers, een input layer (X_1 en X_2), een 'hidden layer' (N_1 en N_2) en een output layer (N_3, ofwel output y). De input van de neuron N_1 is een lineaire combinatie van de input variabelen X_1 en X_2, waarop een niet-lineaire functie is toegepast. Een mogelijke functie is bijvoorbeeld:

$$f(X_1,X_2) = max\,(0,\, w_1^1\, X_1 + w_2^1\, X_2)$$

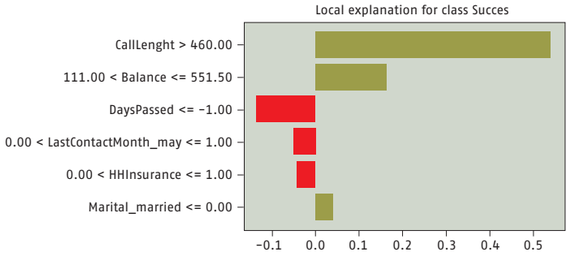
waarbij w_1^1 en w_2^1 de te kalibreren gewichten zijn voor neuron N_1. Door het toepassen van het maximum van de lineaire combinatie en 0 wordt een niet-lineair element toegevoegd, wat mede maakt dat een Neuraal Netwerk ook complexe niet-lineaire effecten kan adresseren. Voor neuron N_2 geldt eenzelfde functie (in dit voorbeeld). Neuron N_3 is de 'output layer' en volgt uit de output van N_1 en N_2 na eventueel toepassen van een niet-lineaire functie.

Neurale netwerken zijn al tientallen jaren geleden ontwikkeld, maar grote vooruitgang werd meer recent geboekt met onder andere het toevoegen van meerdere lagen in de modellen en het samenvoegen en bewerken van groepen neuronen (convolutionele neurale netwerken).

In dit voorbeeld is het 'grid search' proces gevolgd zoals beschreven in Plat (2017), om de meest gunstige structuur van het Neurale Netwerk te vinden. Dit heeft geleid tot een Neuraal Netwerk met één hidden layer met 50 neuronen. Met dit model worden voor deze dataset 70% van de potentiële klanten juist geclassificeerd (in-sample). Zoals verwacht is dat niet heel indrukwekkend, mede vanwege het geringe aantal datapunten.

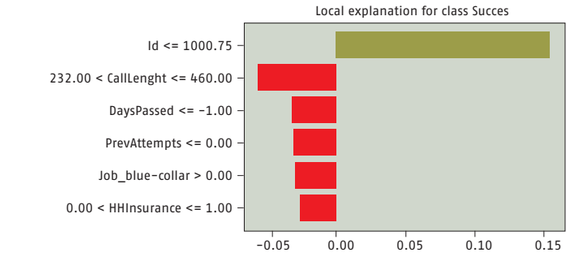
TOEPASSING LIME

Op bovenstaand model is de LIME methodiek toegepast, gebruik makend van de package in Python. In figuur 3 en figuur 4 zijn resultaten van LIME voor respectievelijk potentiële klanten A en B weergegeven. Voor beiden klanten is de kans op succes door het Neurale Netwerk model ingeschat op 100%. De grafieken laten de 6 factoren zien die de meeste impact hebben op de door het model ingeschatte kans op succes. De groene staven geven de factoren aan die bijdragen aan de inschatting van succes, de rode staven geven de factoren aan die bijdragen aan het tegenovergestelde (failure). De x-as geeft een benadering van de impact op de kans op succes aan.



Figuur 3: LIME uitkomsten potentiële klant A

Figuur 3 laat zien dat de lengte van het telefoongesprek van meer dan 460 seconden bij de vorige marketing campagne ('CallLength') een belangrijke bijdrage levert aan de kans op succes voor potentiële klant A. Als deze factor zou wegvallen, zou de kans op succes ongeveer 50% lager zijn. De andere factoren zijn respectievelijk: uitstaand bedrag bankrekening, dagen sinds vorige campagne waar -1 betekent dat de potentiële klant nog nooit benaderd is, de maand van het laatste contact, het reeds in bezit hebben van verzekeringen bij deze bank en huwelijkse staat. Dit zijn ook factoren waarvan we ons kunnen voorstellen dat ze een rol spelen in de inschattingen van het model. Op basis hiervan zouden we ons voor kunnen stellen dat het model zijn werk goed doet en logische uitkomsten genereert voor dit specifieke datapunt (potentiële klant A).



Figuur 4: LIME uitkomsten potentiële klant B

Figuur 4 laat echter een ander beeld zien. Ook hier was de kans op succes ingeschat op 100%, maar de belangrijkste bijdrage hieraan wordt geleverd door het ID nummer van de potentiële klant. Het ID nummer is geen logische factor voor de inschatting van de kans op succes. De volgende stap in dit proces zou daarom zijn om het ID nummer te verwijderen uit de dataset met factoren en indien dit geen soelaas biedt, het model als ongeschikt classificeren.

Bovenstaande analyse is een voorbeeld hoe LIME gebruikt kan worden voor de beoordeling van een Machine Learning model. Naast classificatie vraagstukken zoals in dit voorbeeld ondersteunt de package ook onder andere vraagstukken zoals tekst classificatie en beeldherkenning.

CONCLUSIE

In dit artikel is een voorbeeld gegeven van de werking van LIME, een techniek voor het verklaren van uitkomsten van complexe Machine Learning modellen. Omdat het 'black-box' gehalte een algemeen bekend probleem is bij complexe Machine Learning modellen, is de verwachting dat de komende jaren meerdere technieken zoals LIME worden ontwikkeld om de werking en uitkomsten van complexe Machine Learning modellen te verklaren. Hiermee wordt één van de belangrijkste bezwaren tegen dit soort modellen deels weggenomen. ■

1 – Zie Plat – 'Data Science en Machine Learning: concreet voorbeeld verzekeringsportefeuille', De Actuaris (september 2017).
2 – Zie Ribeiro, Singh en Guestrin (2016) – 'Why should I trust you? Explaining the predictions of any classifier'