DATA SCIENCE INITIATIVE

Handleiding voor Data Science projecten

Victor Pereboom
CTO at Dutch Analytics

Sascha van Weerdenburg
Machine Learning Engineer at Dutch Analytics

Deze handleiding voor AI-projecten binnen de overheid is opgesteld in samenwerking met en aan de hand van het AI-project van Ariea Vermeulen en Arie van Kersen (Rijkswaterstaat). In 2019 startten zij een project om te onderzoeken of de combinatie van dronebeelden en Machine Learning zou leiden tot een verbeterde inspectie van bruggen en andere kunstwerken.

AUTEURS

Victor Pereboom
CTO at Dutch Analytics

Sascha van Weerdenburg
Machine Learning Engineer at
Dutch Analytics

OVER DUTCH ANALYTICS

Het software platform van Dutch Analytics, genaamd Xenia, helpt bedrijven om Artificial Intelligence (AI) algoritmen na ontwikkeling operationeel te maken, te beheren en te monitoren. Het idee achter de software is ontstaan uit de ervaring dat veel bedrijven moeite hebben om resultaten van Data Science projecten operationeel te maken. Veel algoritmen worden na een proof of concept nooit in gebruik genomen wat leidt tot verspilling van tijd, moeite en geld. Maar belangrijker nog: gemiste waarde. Door gebruik te maken van Xenia kunnen bedrijven eenvoudig de stap zetten van Data Science model tot schaalbaar eindproduct.

Meer informatie over het Dutch Analytics en Xenia is te vinden op **www.dutchanalytics.com**

REDACTIE

Victor Pereboom
CTO at Dutch Analytics

Sascha van Weerdenburg
Machine Learning Engineer at
Dutch Analytics

Ariea Vermeulen
Rijkswaterstaat

Arie van Kersen
Rijkswaterstaat

Francien Horrevorts
Rijkswaterstaat, Fran&Vrij Communicatie

Koen Hartog
DSI

Marloes Pomp
DSI

Data Science heeft zich in de afgelopen vijf jaar bewezen als een domein met grote maatschappelijke impact.

We zien steeds meer toepassingen verschijnen, zoals moderne smartphones met beeld- en spraakherkenningstechnologie. Zelfrijdende auto's zijn niet langer uitsluitend onderdeel van fictieve filmscenario's. Deze ontwikkelingen spelen niet alleen lokaal. Over de hele wereld investeren bedrijven in innovatieve technologieën om datagedreven oplossingen en nieuwe producten te ontwikkelen.

Echter, maar een klein percentage van deze initiatieven groeit uit tot een volwaardige eindproduct. Verrassend, als je ziet hoeveel men investeert om deze projecten tot een succes te maken. Dat werpt de volgende vragen op.

Wat zijn de factoren van een geslaagd Data Science project? In welke stappen kom je van een idee tot een goede oplossing?

Deze handleiding geeft meer inzicht in de levenscyclus van Data Science projecten en is gebaseerd op ervaringen met complexe data gedreven oplossingen.

Wat is Data Science 04

Soorten Machine Learning en deep learning 04

Lifecycle van een Data Science project 05

Focus op het beste model:
van business case naar proof of concept 06
Stap 1: een goede business case
Stap 2: de juiste data verkrijgen
Stap 3: data opschonen en verkennen
Stap 4: modellen ontwikkelen en evalueren

Focus op continuïteit:
van proof of concept naar volwaardig eindproduct 15
Stap 5: van werkend proof of concept naar implementatie
Stap 6: model beheren in operatie
Stap 7: van model beheren naar business case

Wat is Data Science?

Maar wat is nou eigenlijk Data Science? En hoe verhoudt het zich tot AI en bijvoorbeeld Machine Learning en Deep Learning? We zetten het voor je op een rij.

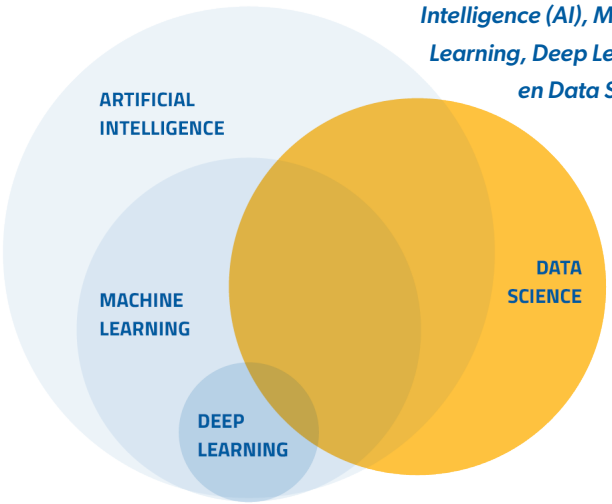
Data Science
Data Science is een vakgebied waarbij data onderzocht wordt op patronen en kenmerken. Het omvat een combinatie van methoden en technieken uit wiskunde, statistiek en computer science. Vaak worden visualisatie-technieken gebruikt om de data inzichtelijk te maken. De focus ligt op het begrijpen en gebruiken van de data om inzichten te verkrijgen die de organisatie verder helpen.

Artificial Intelligence
Bij Artificial Intelligence leren computers om menselijk gedrag en intelligentie na te bootsen.

Machine Learning
Machine Learning is een onderdeel van Artificial Intelligence dat zich focust op 'leren'. Algoritmen en statistische modellen kunnen taken 'leren' van data zonder instructies vooraf.

Deep Learning
Deep Learning is een onderdeel van Machine Learning waarbij artificiële neurale netwerken worden gebruikt. Dit type model is geïnspireerd op de structuur en werking van het menselijk brein.

Beeld van de verschillen tussen Artificial Intelligence (AI), Machine Learning, Deep Learning en Data Science



Soorten Machine Learning en Deep Learning

Er zijn verschillende type Machine Learning-benaderingen voor verschillende typen problemen. De bekendste drie zijn:

Supervised Learning
Supervised Learning is het meest gebruikte type Machine Learning. Een Supervised Learning model leert op basis van een set voorbeelden uit de data die gelabeld is met informatie. Bijvoorbeeld: bij het classificeren van schade aan beton worden foto's gebruikt waarvan bekend is of er wel of niet sprake is van schade. Deze foto's hebben een label meegekregen "wel schade" of "geen schade" om het verband tussen de foto en de classificatie te leren.

Unsupervised Learning
In Unsupervised Learning wordt er geen gebruik gemaakt van labels, maar probeert het model zelf verbanden te vinden in de data. Dit wordt vooral gebruikt voor het groeperen (clusteren) van de voorbeelden uit de data. Denk bijvoorbeeld aan het creëren van verschillende klantgroepen waarbij klanten met vergelijkbare kenmerken in dezelfde groep zitten. Van te voren is niet bekend welke groepen klanten er zijn en aan welke kenmerken ze voldoen, maar het algoritme kan groepen creëren die onderling zo veel mogelijk van elkaar verschillen.

Reinforcement Learning

Tot slot leert een Reinforcement Learning-model op basis van trial en error. Door goede keuzes te belonen en slechte te straffen, leert het model patronen te herkennen. Deze techniek wordt vooral gebruikt bij het doorgronden van spellen (zoals Go) en in de robotica (een robot die leert lopen door vallen en opstaan). Dit type Machine Learning valt gebruikelijk buiten Data Science, omdat hierbij het 'leren' van een taak het doel is en niet het begrijpen en gebruiken van de onderliggende data.

Lifecycle van een Data Science project

Nu we de verschillende termen hebben uitgelegd, focussen we op de Data Science projecten. Waar moet je op letten bij zo'n project, wat komt er bij kijken en welke best practices en learnings kunnen we je meegeven? We beginnen met iets meer achtergrond over de Lifecycle van deze projecten. De levensloop van een Data Science project bestaat uit twee fases, die in totaal 7 stappen bevatten.

FOCUS OP HET BESTE MODEL: VAN BUSINESS CASE NAAR PROOF OF CONCEPT

In deze fase ligt de focus op de ontwikkeling van het beste model voor de specifieke business case. Daarvoor is het definiëren van een goede business case essentieel. Vervolgens zal het Data Science team hiermee aan de slag gaan en werkt men toe naar een werkend prototype (proof of concept).

Deze eerste fase bestaat uit 4 stappen:

1. Een goede business case
2. De juiste data verkrijgen
3. Data opschonen en verkennen
4. Modellen ontwikkelen en evalueren

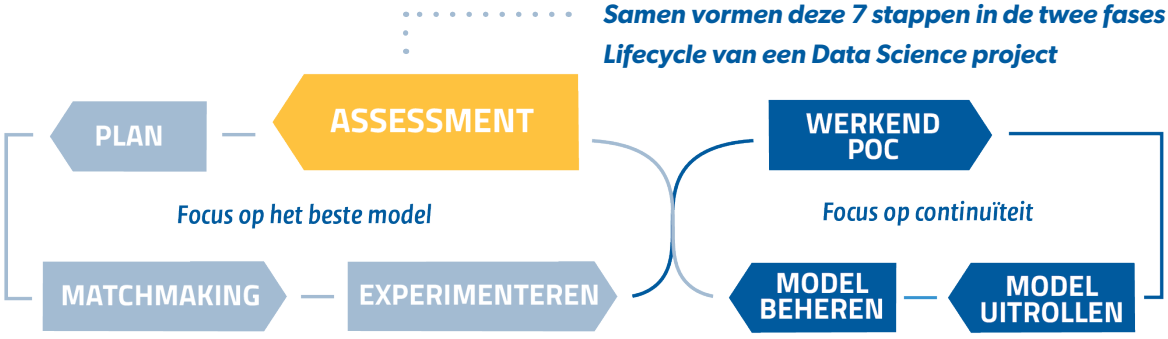
FOCUS OP CONTINUÏTEIT: VAN PROOF OF CONCEPT NAAR VOLWAARDIG EINDPRODUCT

In de tweede fase ligt de focus op continuïteit en wordt uit een werkend prototype een operationeel eind-product ontwikkeld.

Deze fase bestaat uit 3 stappen:

5. Van Proof of Concept naar implementatie
6. Model beheren in operatie
7. Van model beheren naar business case

Samen vormen deze 7 stappen in de twee fases de Lifecycle van een Data Science project



Maar hoe pak je dit proces als organisatie op een effectieve manier aan? In de volgende hoofdstukken bespreken we de 7 stappen van de Lifecycle en zetten we op een rijtje waar je op moet letten. Elk hoofdstuk sluiten we af met een korte samenvatting.

FOCUS OP HET BESTE MODEL: VAN BUSINESS CASE NAAR PROOF OF CONCEPT

Stap 1: Een goede business case

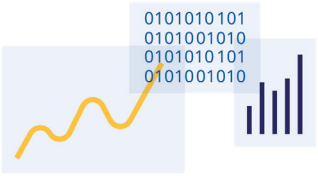
De uiteindelijke waarde van een Data Science project hangt af van een duidelijke business case. Wat wil je bereiken met het project? Wat is de toegevoegde waarde voor de organisatie en hoe gaat de informatie uit een algoritme uiteindelijk gebruikt worden? Hieronder een aantal richtlijnen voor het definiëren en toetsen van de business case.

Betrek eindgebruikers

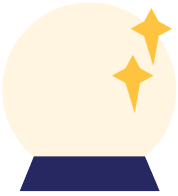
Een business case is sterk wanneer deze van de mensen dicht op de praktijk komt, omdat zij de mensen zijn die de informatie uit de modellen moeten gebruiken en erop moeten vertrouwen. Het is daarom belangrijk dat zij het nut van de Data Science oplossing inzien. Waar ligt de behoefte van de gebruiker en op welke wijze kan het eindproduct waarde creëren voor de gebruiker?

Biedt AI de beste oplossing voor het probleem?

Na het definiëren van de business case is het verstandig te beoordelen of AI de beste oplossing is voor het probleem. AI is erg geschikt voor het vinden van patronen in data, die voor mensen te groot en te complex zijn. AI heeft zich hierbij bewezen in bijvoorbeeld beeld- en spraakherkenning. AI helpt bij het automatiseren van deze taken, maar in sommige gevallen is AI minder bruikbaar:

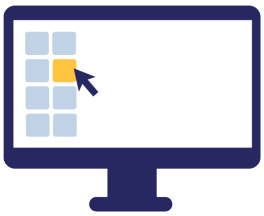


AI modellen leren van (grote hoeveelheden) data. Als er weinig relevante data beschikbaar is of de contextuele informatie niet in de data aanwezig is, kan er beter niet voor een AI oplossing worden gekozen.



AI is niet goed in onvoorspelbare situaties waarbij creativiteit en intuïtie nodig zijn om ze op te lossen.

Als transparantie van het algoritme van groot belang is. Van (met name Deep Learning) modellen is vaak moeilijk in te schatten waarom deze bepaalde uitkomsten geven.



Er kunnen simpelweg eenvoudigere, effectievere of goedkopere oplossingen bestaan dan Data Science oplossingen, zoals traditionele software.

BENOEM STAKEHOLDERS

Om het Data Science project te laten slagen, zijn drie stakeholders van belang:

DE OPDRACHTGEVER

(iemand vanuit beleid/ strategie)

De opdrachtgever is verantwoordelijk voor het behartigen van de toegevoegde waarde van de business case.

DE EINDGEBRUIKER / UITVOEREND EXPERT

(iemand vanuit de operatie)

Deze persoon is verantwoordelijk voor acceptatie bij de gebruikers. Zonder een duidelijk aanspreekpunt en voorstander bij de gebruikers is het mogelijk dat er een Data Science model wordt gemaakt dat in de praktijk niet gebruikt gaat worden.

IEMAND UIT HET DATA SCIENCE TEAM

De data scientist is verantwoordelijk voor het beoordelen van de kans van slagen van het project op basis van de oplossing en beschikbare data.

Deze drie stakeholders moeten vooraf op één lijn zitten over wat er van het eindproduct verwacht wordt. Als er geen Data Science team binnen de organisatie beschikbaar is, kan je er voor kiezen om het project uit te besteden. In dat geval is de opdrachtgever het aanspreekpunt voor het Data Science team.

STAP 1 SAMENGEVAT

- Bedenk de business case samen met de eindgebruikers.
  - *Waar hebben zij behoefte aan?*
  - *Aan welke eisen moet de oplossing voldoen om waarde voor hen te creëren?*
- Biedt AI de beste oplossing voor het probleem?
- Benoem stakeholders vanuit beleid en strategie (de opdrachtgever), de uitvoering (de eindgebruiker) en het Data Science team.

Stap 2: De juiste data verkrijgen

Het ontwikkelen van een AI algoritme vraagt vaak om grote hoeveelheden kwalitatief goede data, omdat de intelligentie gebaseerd wordt op de informatie die in de data aanwezig is. Hoe kom je aan deze goede data en informatie?

Zorg dat data beschikbaar is

Voordat een Data Science team kan beginnen met het ontwikkelen van een model, zijn de juiste data nodig. Data Science teams zijn afhankelijk van data engineers en database beheerders, omdat zij de juiste rechten hebben om deze data te ontsluiten. Daarnaast zijn Data Science teams afhankelijk van de uitvoerende experts, omdat zij weten in welke context de data geplaatst moet worden. Daarom heb je de kennis van zowel het Data Science team, data engineers of database beheerders en de uitvoerend experts nodig om de benodigde data te kunnen vinden. Het kan ook zijn dat de benodigde data binnen de organisatie niet beschikbaar is. Je kunt er dan voor kiezen om data te verzamelen, bijvoorbeeld bij externe bronnen.

Datadump voor eerste fase

In de eerste fase van de Data Science Lifecycle voldoet meestal een eenmalige datadump

Deze zal door het Data Science team voor een deel gebruikt worden om het model op te “trainen” en voor een deel om op te “testen”.

Als meerdere Data Science teams concurreren om dezelfde business case, moeten alle teams dezelfde dataset krijgen. Zo hebben alle teams een gelijke kans om het beste model te vinden. Bovendien kun je de modellen daardoor eerlijk onderling vergelijken.

Het is verstandig om een extra “test dataset” te delen waarvan de te voorspellen waardes achter de hand worden gehouden. Op basis van hoe goed de voorspellingen van de Data Science teams overeenkomen met de te voorspellen waardes in deze ‘test dataset’, kan objectief worden beoordeeld welk model het beste presteert. Dit is de manier hoe bijvoorbeeld het platform ***Kaggle*** opereert dat openbare competities namens bedrijven uit schrijft voor data scientists over de hele wereld.

Automatisch uitlezen data in tweede fase

In de tweede fase van de Data Science Lifecycle, voldoet een eenmalige datadump niet meer

De data moet dan automatisch het model bereiken. In de praktijk is dit vaak ingewikkeld, omdat je te maken krijgt met data silo's. Dit zijn afgesloten databases die lastig te integreren zijn in een applicatie. Dit komt doordat veel systemen niet zijn ontwikkeld om met elkaar te communiceren. Of deze communicatie is lastig door interne IT-beveiligingsmaatregelen van het bedrijf. Het is aan te raden om al in de eerste fase over de tweede fase na te denken.

Begin tijdig met inrichting infrastructuur

Investeren in een goede infrastructuur voor opslag en uitwisseling van data is essentieel voor het slagen van het Data Science project. Een data engineer kan in dit proces helpen om een robuuste data infrastructuur op te zetten. Begin hier vroeg mee en houd rekening met beveiliging, toegangsrechten en bescherming van persoonsgegevens.

STAP 2 SAMENGEVAT

- Zorg ervoor dat alle data beschikbaar is voor Data Science teams i.s.m. data engineers.
  - *Voor de eerste fase voldoet een eenmalige datadump.*
  - *Voor de tweede fase is het belangrijk dat de data automatisch uitgelezen kan worden en aan het model gevoed kan worden.*
- Begin tijdig met de inrichting van een goede data infrastructuur en toegangsrechten.

Stap 3: Data opschonen en verkennen

Als de dataset beschikbaar is, kan het Data Science team beginnen met de ontwikkeling van de oplossing. Een belangrijke stap is om de verkregen data eerst op te schonen en te verkennen. De data moet aan een aantal voorwaarden voldoen voor het ontwikkelen van AI modellen. De data moet representatief zijn en van goede kwaliteit.

REPRESENTATIEVE DATA

Het is belangrijk dat de data die gebruikt wordt voor het ontwikkelen van een model een zo goed mogelijke weerspiegeling van de werkelijkheid is. Een zelflerend algoritme wordt immers slim gemaakt door te leren van voorbeelden in de gegeven data. Vereisten aan de data zijn hierbij de verscheidenheid en volledigheid van de datapunten. Verder moet de data voor veel business cases actueel zijn, omdat de data van lang geleden niet meer representatief is voor de huidige situatie. Wees er ten slotte alert op dat er geen onbedoelde bias in de data aanwezig is.

VOORBEELD

Als je een model ontwikkelt voor het classificeren van afbeeldingen moet rekening worden gehouden met de verscheidenheid in de afbeeldingen. Als je op basis van foto's schade in beton wilt herkennen en alle foto's met schade zijn gemaakt op een bewolkte dag en alle foto's zonder schade op een zonnige dag, kan het zijn dat het Data Science model zijn keuze gaat baseren op de achtergrondkleuren. Een nieuwe afbeelding van beton op een zonnige dag kan daardoor foutief als niet 'beschadigd' worden geclassificeerd.

Kwaliteit van de data
Ook de kwaliteit van de data is van groot belang. Een goede datastructuur helpt hierbij, zodat de data zo compleet, consistent en duidelijk mogelijk is. Daarnaast is het essentieel om menselijke (invoer)fouten zoveel mogelijk te voorkomen. Hiervoor kunnen verplichte velden, checks en categorieën in plaats van open vakken tekst bij de data invoer van pas komen.

Vertrouwen
Representatieve data en de kwaliteit van de data hebben een positief effect op het vertrouwen in het project en de oplossing. En dit vertrouwen draagt bij aan de acceptatie en ingebruikname door de eindgebruikers.

Korte feedback cyclus
Tijdens het verkennen van de data is het van cruciaal belang dat de data juist geïnterpreteerd wordt. Dit gebeurt door feedback te vragen van de uitvoerende experts. Hiervoor is een korte feedback cyclus tussen het Data Science team en deze experts nodig. Dit kan bijvoorbeeld door elke paar weken een presentatie van de bevindingen te geven aan de experts en de opdrachtgever.

STAP 3 SAMENGEVAT

- Zorg ervoor dat de data representatief is. De data moet volledig, compleet in verscheidenheid, actueel en vrij van onbedoelde bias zijn. Belangrijke concepten van de werkelijkheid die de voorspelling kunnen beïnvloeden moeten aanwezig zijn in de data. Gebruik hiervoor de kennis van de uitvoerend experts.
- Zorg voor hoge kwaliteit van de data. Voorkom fouten in de data door deze zo goed mogelijk af te vangen bij de invoer ervan. Evalueer of alle data compleet en consistent is.
- Zorg voor vertrouwen bij alle betrokkenen.
- Zorg voor een korte feedback cyclus voor juiste interpretatie van de data.

Stap 4: Modellen ontwikkelen en evalueren

In deze fase heeft de data scientist veel vrijheid. Het doel is om zo snel mogelijk veel business waarde te creëren. Daardoor is geen enkele Data Science oplossing precies hetzelfde en heeft Data Science een sterk experimenteel karakter. Maar waar moet je aan denken en waarmee wordt geëxperimenteerd? Dit zijn zowel het type algoritme en zijn parameters, als de variabelen uit de dataset (de features). Er zijn verschillende categorieën AI modellen die je kunt gebruiken, afhankelijk van het probleem en de beschikbare data. Bij het ontwikkelen en evalueren van de modellen, moet je op een aantal punten letten.

Labelen van de dataset

De meest gebruikte modellen vallen in de categorie Supervised Learning, waarbij een model leert van een set voorbeelden uit de data met labels. Denk aan het eerder genoemde voorbeeld over herkenning van betonschade. Het model wordt getraind op een set van foto's van betonstructuren waarvan het bekend is of ze beschadigd zijn. Na training kan het model nieuwe foto's classificeren.

Helaas is er niet altijd een dataset beschikbaar waarbij dit soort labels bekend zijn. Dan is het nodig om de data te "labelen" of "annoteren". Dit is vaak grotendeels handmatig werk. Voor het voorbeeld van de betonschade betekent dit dat iemand handmatig door zo'n 2000 foto's heen gaat en aangeeft of er schade zichtbaar is. Er bestaan ook snellere methodes waarbij de expert enkel foto's krijgt waarvan het Machine Learning model het meest onzeker is. Veel modellen geven namelijk zelf aan met welke zekerheid de voorspelling is gedaan.

Beoordelen van het model

Labels zijn ook van belang wanneer het model moet worden beoordeeld. Het kan zijn dat de werkelijke waarde automatisch in de data verschijnt als deze bekend wordt. In dat geval kun je de werkelijke waarde direct vergelijken met wat het model voorspeld had. Als de werkelijke waarde niet automatisch in de data verschijnt, zoals bij het voorbeeld over betonschade, is het slim om een feedback-loop in te bouwen in de eindoplossing. Dan krijgt de gebruiker de vraag om feedback te geven over de juistheid van de classificatie. Deze informatie is van belang voor het monitoren van de kwaliteit van het Data Science model.

Optimalisatie van het algoritme

Een algoritme moet geoptimaliseerd worden voor het betreffende probleem en de data. Door het aanpassen van hyperparameters , "de spelregels van het model", wordt het algoritme passend gemaakt voor de toepassing. Deze optimalisatie stap volgt vaak uit een grid search, waarbij een grote set aan waardes wordt geprobeerd en de beste worden gekozen.

Kiezen beoordelingsmethode

Om te kunnen bepalen wat het beste model is, moet er een beoordelingsmethode worden gedefinieerd, die past bij het doel van de oplossing. Zo is het bijvoorbeeld in het medisch domein van belang dat extreme fouten van het Data Science model niet voorkomen, terwijl in andere domeinen extreme fouten misschien zijn ontstaan door extreme meetpunten in de data waarvan is besloten er minder waarde aan te hechten. Bij classificatiemodellen kun je denken aan de balans tussen inclusiviteit (alle betonschade vinden) en precisie (alleen betonschade vinden). De beoordelingsmethode moet aansluiten bij de business case en hoe het algoritme in de praktijk gebruikt gaat worden.

Monitoren van de modellen

Het monitoren van de kwaliteit van Data Science modellen is erg belangrijk. Dit komt door de afhankelijkheid van de data.

Data kan over de tijd veranderen waardoor een Data Science model minder goed kan gaan presteren. Zodra de buitenwereld verandert, verandert de data immers mee. Zo kunnen de seizoenen en jaargetijden invloed hebben op foto's en dientengevolge op de resultaten van het model. Het kan zinvol zijn om een onderbouwing toe te voegen aan de voorspelling van een Data Science model, zodat de gebruiker begrijpt waar de voorspelling op gebaseerd is. Dit geeft meer inzicht en transparantie in hoe het model redeneert.

Uitlegbaarheid van de voorspellingen

Het maken van een Data Science model bestaat vooral uit uitproberen en evalueren. Ook hierbij is de feedback van gebruikers belangrijk. Zijn de voorspellingen van het model logisch? Missen essentiële variabelen die invloed zouden kunnen hebben? Hebben de gevonden relaties ook een causaal verband? Hierbij speelt de transparantie van het algoritme een rol. Sommige typen algoritmen zijn erg ondoorzichtig waardoor gegeven uitkomsten niet terug te leiden zijn naar de input data. Dit is onder andere het geval bij neurale netwerken. Meer lineaire methoden of traditionele statistische benaderingen zijn vaak makkelijker te interpreteren. De vraag naar transparante algoritmen is terug te zien in de recente ontwikkelingen omtrent Explainable AI. Neem in de keuze voor het model de eisen vanuit de praktijk mee voor de uitlegbaarheid van de voorspellingen.

Prototype af?

Is het prototype af? Is er bijvoorbeeld een interface beschikbaar waarop het resultaat te zien is? Zodra het prototype waarde genereert, zijn alle stappen van de eerste fase doorlopen en begint fase twee, de operationalisatie.

STAP 4 SAMENGEVAT

- Zorg ervoor dat de dataset gelabeld is.
- Zorg ervoor dat de beoordeling van het model onderdeel is van de eindoplossing.
    - *Kan de werkelijke waarde automatisch uit de data verkregen worden?*
    - *Kan de gebruiker feedback geven?*
- Wordt het algoritme met de juiste beoordelingsmethode beoordeeld?
    - *Sluit de gebruikte beoordelingsmethode aan bij de probleemstelling en de gebruikte data?*
- Zorg voor monitoring van de kwaliteit van het model en de veranderingen in de data. Voeg bijvoorbeeld een onderbouwing toe aan de voorspelling van het model, zodat de gebruiker de voorspelling begrijpt.
- Zorg ervoor dat de Data Science oplossing transparant is. Bespreek de prestatie van het model met de gebruiker.
    - *Zijn de voorspellingen logisch?*
    - *Voldoet het model aan de verwachtingen?*
    - *Hoe kunnen de voorspellingen gebruikt worden door de gebruiker?*
- Als het prototype waarde genereert, kan men verder naar de volgende stap: operationaliseren.

FOCUS OP CONTINUÏTEIT: VAN PROOF OF CONCEPT NAAR VOLWAARDIG EINDPRODUCT

We zijn in de tweede fase beland van de Lifecycle van een Data Science project, de operationalisatie. In deze fase wordt uit de proof of concept een volwaardig eindproduct ontwikkeld. Dit bestaat uit drie stappen.

Stap 5: Van werkend proof of concept naar implementatie

In stap 5 wordt het model en de organisatie klaargestoomd om het model in gebruik te nemen. Belangrijke elementen van deze stap zetten we voor je op een rij.

Kwaliteit van de code

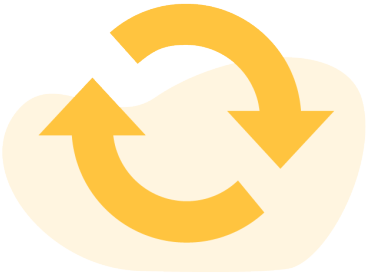
Data Science en software ontwikkeling zijn twee verrassend verschillende werelden. Data scientists richten zich op het bedenken en uitvoeren van experimenten en software ontwikkelaars zijn gericht op het bouwen van stabiele, robuuste en schaalbare oplossingen. Dit gaat soms moeilijk samen.

In de eerste fase van het project ligt de focus op snel een zo goed mogelijk prototype realiseren. Het draait om de prestaties van het model, en het zo snel mogelijk creëren van een grote business waarde. Daardoor is er vaak weinig tijd en aandacht voor de kwaliteit van de programmeercode. In de tweede fase wordt de kwaliteit van de programmeercode juist erg belangrijk. Die kwaliteit bepaalt de afhandeling van mogelijke toekomstige fouten, duidelijke documentatie en de snelheid van het uitvoeren van de code.

Daarom wordt in de praktijk de code van een oplossing in de tweede fase vaak opnieuw gestructureerd. Hierbij wordt het uiteindelijke model losgemaakt van alle experimenten die zijn gedaan tijdens de vorige fase.

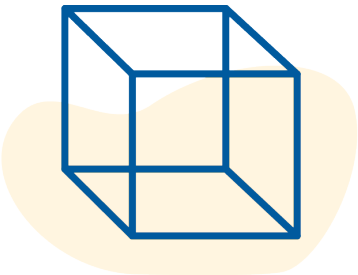
Integratie in bestaande processen

Het prototype moet opgenomen worden in de bestaande processen in de organisatie. Dit bestaat uit twee stappen:



Het automatiseren van datastromen

Het ophalen en wegschrijven van data in externe systemen en databases moet automatisch gebeuren met queries. Dit kan een complex proces zijn dat een data engineer kan uitvoeren.



Het opzetten van een infrastructuur voor het hosten en beheren van het model

Het model moet beschikbaar zijn voor alle gebruikers en moet kunnen meeschalen met het gebruik ervan.

Eisen aan de productieomgeving

Er zijn verschillende eisen die je kunt stellen aan het hosten van het model in een productieomgeving.

> *“Wanneer een model intensief gebruikt gaat worden, wordt schaalbaarheid relevant.”*

- Wanneer een model intensief gebruikt gaat worden, wordt schaalbaarheid relevant. Dit betekent dat een model meerdere keren parallel wordt gestart, zodat meerdere verzoeken tegelijk kunnen worden verwerkt. Als het aantal verzoeken erg varieert, is het handig als de productieomgeving automatisch schaalt. Dit zorgt ervoor dat servers zo efficiënt mogelijk kunnen worden gebruikt. Dit is te vergelijken met een supermarkt waar meerdere kassa’s open gaan als het drukker wordt en weer sluiten als het rustiger wordt. Automatisch schalen zorgt er ook voor dat het model klaar is voor de toekomst. Het kan intensiever gebruik aan zonder extra investering in de infrastructuur.
- Een belangrijke voorwaarde aan de productieomgeving is de beschikbaarheid. De gebruikers moeten het model altijd kunnen gebruiken. Het model moet dus bijvoorbeeld automatisch herstarten als systemen uitvallen of automatisch back-ups maken.
- Een derde eis is authenticatie en beveiliging. Alleen mensen en systemen die de productieomgeving, het model en de data mogen bekijken en gebruiken, moeten toegang krijgen. Hiervoor is een veilig en betrouwbaar rechtensysteem nodig.
- Transparantie en auditing in de productieomgeving is ook een aandachtspunt. Je wilt weten wie wat wanneer veranderd heeft. Wijzigingen moeten traceerbaar zijn, zodat te herleiden is waar fouten vandaan komen. Dankzij een goede logging kan je snel terugvinden wat er precies gebeurd is en kun je fouten makkelijker oplossen. Koppel auditing aan het monitoren van het model, zodat ook duidelijk wordt of er een relatie is tussen de prestatie en een update van het model.

Beheer van de productieomgeving

Als het model foutmeldingen geeft, zal iemand deze moeten oplossen. Het is goed om van te voren duidelijk uit te spreken wie hier verantwoordelijk voor is. Dit kan de IT afdeling zijn, of het Data Science team zelf. In ieder geval heb je er baat bij als de code gestructureerd is en foutmeldingen in een duidelijke log worden beschreven.

Omgaan met meerdere stakeholders: ownership van data en code

Het komt regelmatig voor dat een organisatie samenwerkt met een externe partij voor de ontwikkeling van Data Science modellen of het aanleveren van data. Data heeft vaak een gevoelig karakter en veel organisaties willen hun data beschermen en controle hebben over wie hier toegang toe heeft. Ook kan dit vereist zijn door bijvoorbeeld de AVG wetgeving. Voor de modelcode van een data scientist geldt hetzelfde.

Er kunnen discussies ontstaan tussen partijen over wie eigenaar is van de data of het algoritme. In een samenwerking met meerdere partijen, kan een neutraal platform uitkomst bieden. Dit platform is te bereiken door de data scientist met zijn modelcode en door de data leverancier met zijn data. Via dit platform kunnen de eigenaren de juiste rechten krijgen. Op zo'n platform kan je het model van de data scientist testen zonder de broncode ervan te bekijken. Zo kun je een test set maken en daarmee meerdere modelvarianten vergelijken.

De juridische aspecten van een AI-project

Er zijn diverse juridische aandachtspunten als je een AI-project start, zoek dus vroegtijdig contact met een juridisch medewerker of een externe juridische expert.

Enkele voorbeelden van waar je rekening mee moet houden:

- Zorg dat je grip krijgt op de gemarkeerde data, zodat je niet afhankelijk bent van één ontwikkelaar.
- Leg specifieke kwaiteitsnormen vast voor te ontwikkelen AI-modellen (supervised learning).
- Intellectueel eigendom: de toegevoegde waarde van een model vloeit voort uit de het afstellen van de parameters. Om de IE-rechten daarop te hebben, dienen expliciete afspraken te worden gemaakt.

Advocatenkantoor Pels Rijcken ontwikkelde in samenwerking met de Gemeente Amsterdam ***algemene voorwaarden voor de inkoop van AI-toepassingen*** die specifiek worden gebruikt voor het nemen van besluiten over burgers. Hoewel inkoopvoorwaarden afhankelijk zijn van de precieze casuïstiek, kunnen onderdelen uit die voorwaarden wellicht worden hergebruikt.

Acceptatie

Enthousiasme van gebruikers voor de nieuwe oplossing komt niet altijd vanzelf. Vaak heeft de gebruiker het gevoel dat de techniek een bedreiging vormt voor de rol van de werknemer in een organisatie. Toch is er een scherp contrast tussen de kracht van mensen en die van een algoritme. Waar computers en algoritmen goed zijn in het monitoren van de continue inkomende stroom aan data, zijn mensen een stuk beter in het begrijpen van de context waarin de uitkomst van een algoritme geplaatst moet worden. Dit pleit voor oplossingen die gericht zijn op samenwerking tussen mens en algoritme. Het algoritme kan de datastroom filteren en verrijken met verbanden of eigenschappen, en de medewerker kan tijd besteden aan de interpretatie van de resultaten en het nemen van beslissingen.

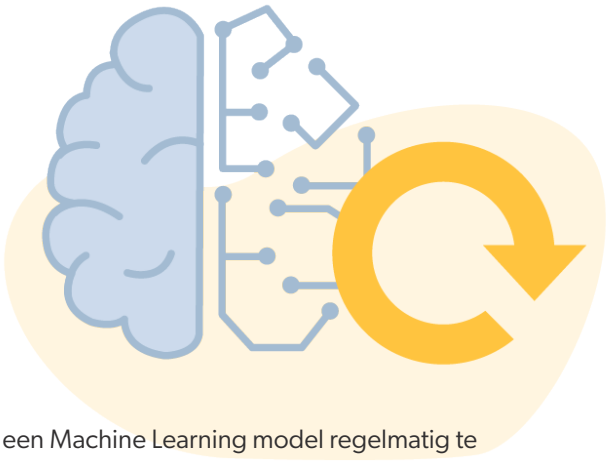
Aansprakelijkheid
Hoe het model gebruikt wordt, bepaalt deels wie verantwoordelijk is voor de beslissingen die gemaakt worden. Wanneer het Data Science model gebruikt wordt als filter, moet je erover nadenken wat er gebeurt als het model een fout maakt. Wat zijn de effecten en wie is hiervoor verantwoordelijk? Hoe zit dit bijvoorbeeld bij een zelfrijdende auto, wie kan men dan aanspreken als de auto een ongeluk veroorzaakt? Dit zijn vaak lastige ethische vragen, maar het is belangrijk om hierover na te denken wanneer een model geïntegreerd wordt in bedrijfsprocessen.

STAP 5 SAMENGEVAT

- Zorg voor een goede kwaliteit van de programmeercode.
- Integreer het model in de bestaande bedrijfsprocessen.
  - *Automatiseer datastromen.*
  - *Zorg voor infrastructuur voor hosten en beheer van het model.*
- Let op de eisen aan de productieomgeving: schaalbaarheid, beschikbaarheid, beveiliging en transparantie & auditing.
- Specificeer verantwoordelijken voor het beheer van Data Science modellen.
- Een neutraal platform kan helpen bij het beschermen van ownership over data en model code.
- Gebruikers moeten meegenomen worden in het project.
  - *Voorkom het gevoel van bedreiging door aan te sturen op samenwerking tussen mens en algoritme.*
  - *Help gebruikers bij het gebruik van de oplossing.*
- Specificeer wie aansprakelijk is bij foutieve voorspellingen van het Data Science model.
- Bespreek vroegtijdig de juridische aandachtspunten van het project met een juridisch expert.

Stap 6: Model beheren in operatie

Een model dat in een productieomgeving draait en gebruikt wordt, moet nog altijd gecontroleerd worden. Er moet beoordeeld worden of het model goed blijft werken en bruikbaar blijft. Spreek af wie dit doet en wie hiervoor verantwoordelijk is. Dit kan bijvoorbeeld het Data Science team zijn, of de eindgebruiker.



Het kan nodig zijn om een Machine Learning model regelmatig te 'hertrainen' op nieuwe data. Dit kan een handmatige taak zijn, maar kan ook ingebouwd zijn in de eindoplossing. In dat laatste geval is monitoring van de prestatie van het model erg belangrijk. Hierbij moet op een structurele manier opgeslagen worden welke prestatie en code hoort bij welke trainingsdata, zodat veranderingen in de prestatie van een model terug te herleiden zijn naar deze data. Dit wordt wel Data lineage genoemd. Hiervoor komt steeds meer tooling op de markt, bijvoorbeeld Data Version Control (DvC).

STAP 6 SAMENGEVAT

- Spreek duidelijk af wie het model beheert als het in gebruik genomen is.
- Wanneer het model frequent 'hertraind' wordt op nieuwe data is het belangrijk om consistent op te slaan wanneer welke versie wordt gebruikt, zodat men kan traceren welke prestatie bij welke training dataset hoort.

Stap 7: Van model beheren naar business case

Door de prestatie van een Data Science model vaak te controleren, kun je erachter komen dat een model niet (meer) goed werkt.

Dit komt door de sterke afhankelijkheid tussen de code van het algoritme, de data die gebruikt is om het algoritme te trainen en de continue stroom aan nieuwe data die door externe factoren beïnvloed wordt. Zo kan het voorkomen dat omgevingsfactoren veranderen waardoor bepaalde aannames niet meer blijken te kloppen of dat er nieuwe variabelen gemeten worden die voorheen niet beschikbaar waren.

De ontwikkeling van het algoritme gaat daarom op de achtergrond vaak door. Hierdoor ontstaan nieuwere modelversies voor dezelfde business case. Het gevolg is dat software updates nodig zijn. Vaak wil men eerst een periode meerdere modelversies tegelijk draaien, zodat de verschillen inzichtelijk worden. Hierbij heeft ieder model zijn eigen afhankelijkheden om rekening mee te houden.

Als meerdere modellen en projecten naast elkaar in de productieomgeving bestaan, moeten alle projecten overzichtelijk en controleerbaar blijven. Hiervoor zijn standaarden nodig om versplintering tegen te gaan. Denk aan een vaste locatie (allemaal in dezelfde omgeving) en dezelfde (code- en data-) structuur.

Uiteindelijk vormen Data Science projecten een gesloten cyclus en kan je altijd blijven doorontwikkelen. De steeds veranderende code, variabelen en data, maken Data Science producten zowel technisch complex, als zeer flexibel en bijzonder krachtig wanneer goed toegepast.

STAP 7 SAMENGEVAT

- Als het mogelijk is, is het goed om een nieuwe modelversie een tijdje tegelijk te laten draaien met de vorige modelversie, om ze goed te vergelijken.
- Maak standaarden en maak deze vereist voor nieuwe Data Science projecten, zoals:
  - *Vaste locatie (allemaal in dezelfde omgeving).*
  - *Dezelfde code-structuur (indien deze intern ontwikkeld wordt).*
  - *Dezelfde data-structuur.*