Machine Learning Pipeline Automation

Amar Kisoensingh

25 februari 2021

Voorwoord

Samenvatting

Summary

Afkortingen

PaaS Platform as a Service 18, 19

Woordenlijst

cloud computing platform first 14, 19, 20

datamining 18

machine learning is het ontwikkelen van algoritmes en technieken waarmee computers kunnen leren. 14, 18–20

machine learning pipeline is een pijplijn waarin gedefiniëerd staat hoe data stroomt. first 14, 15, 18, 19

vendor lock-in is niet de mogelijkheid hebben om naar een andere dienst/provider te gaan die dezelfde service biedt. 18, 19

Lijst van figuren

Lijst van tabellen

Inhoudsopgave

1 Inleiding

1.1 Het bedrijf: NGTI

NGTI is een mobile app development bureau gevestigd in Rotterdam, Nederland en maakt hoogwaardige mobiele applicaties voor native, hybrid en webgebruik.

1.1.1 Klanten van NGTI

1.1.2 Tools die worden gebruikt

Om productief te zijn gebruikt NGTI een aantal tools en programma's om producten te maken en te communiceren met zowel collega's als klanten.

Slack

Interne communicatie gaat via Slack. Het programma faciliteren collega's om elkaar met een lage instap te benaderen en berichten die voor het hele bedrijf relevant zijn te versturen. Ook zijn er 'channels' beschikbaar over specifieke onderwerpen, zoals: *#dev*, *#ios* en *#test-automation*.

Google Workspace

Met Google Workspace kunnen bestanden en documenten gemaakt, opgeslagen en gedeeld worden. Omdat dit via een browser kan, hoeven werknemers geen software te installeren. NGTI gebruikt het ook om collaboratief en parallel te werken aan hetzelfde document.

Zoom

Voorheen werdt Zoom alleen gebruikt om te videobellen met collega's en interviewees. In de tijd van het pandemie is Zoom echter een belangrijke speler geworden om effectief samen te werken. Meetings zoals introducties van nieuwe collega's of demo's van producten worden online gehouden.

1.2 Opdracht

NGTI heeft voorzien dat ze haar applicaties 'slimmer' moet maken door machine learning in te zetten. Niet alleen zorgt dit voor een betere gebruikerservaring, maar geeft NGTI ook een voorsprong op haar concurrenten.

Om machine learning toe te passen is het raadzaam om een machine learning pipeline op te zetten. Een pipeline is een gestructureerde werkwijze om een model te trainen. Het opzetten van de pipeline en een competente model trainen is tijdrovend en vereist kennis in het domein. Vaak worden modellen getrained op een cloud computing platform zoals Azure, AWS of Google Cloud. Het probleem met platformen zoals deze is dat het ontzettend lastig is om te wisselen van platform en er is veel kennis vereist om een infrastructuur op te zetten.

De opdracht bestaat uit twee onderdelen:

1. onderzoek naar het automatiseren van het opzetten van een machine learning pipeline om het laagdrempelig en minder tijdrovend te maken.
2. onderzoek naar het maken van een platform agnostische oplossing

Hierbij zal een PoC gemaakt worden om aan te tonen of het haalbaar is. Een diepere duik in het probleem is te vinden in hoofdstuk 2.

1.3 Leeswijzer

2 Probleemanalyse

2.1 Probleemdefinitie

Zoals beschreven in paragraaf 1.2 is NGTI genoodzaakt om machine learning in te zetten om haar applicaties 'slimmer' te maken. Hier zijn een aantal redenen voor, namelijk:

1. gebruikerservaring verbeteren
2. voorsprong hebben op concurrenten

Het 'slimmer' maken van applicaties kan op verschillende manieren, maar met machine learning kan een platform gebouwd worden waar naar elke richting op gegaan kan worden. Om machine learning te implementeren in haar applicaties loopt NGTI tegen een aantal opstakelsop, namelijk: expertise vereist in het machine learning pipeline domein, tijd om een pipeline op te zetten en vendor lock-in.

2.1.1 Expertise ML

Machine learning is geen triviaal onderwerp. Om een model te trainen is kennis nodig van verschillende domeinen: datamining, software engineering en statistieken. In een multidisciplinair team is het voor één teamlid niet nodig om alle domeinen te beheersen.

Doordat er voorkennis nodig is om een model te trainen, is het vaak te hoog-drempelig om te beginnen voor developers. De expertise is daar bovenop niet in een korte tijd te vergaren.

2.1.2 Opzetting pipeline

Er bestaan verschillende manieren om een model te trainen. Een machine learning pipeline opzetten is daar een van. In een pipeline wordt voor het trainen van het model de data voorbereid. Het opzetten van een pipeline kost tijd. Op zichzelf niet zo zeer veel tijd, maar als er veel wordt geëxperimenteerd met het trainen van modellen kan de tijd opstapelen. Ook zijn de stappen in een pipeline over het algemeen hetzelfde, ongeacht wat voor model je traint. Hierdoor worden taken vaak herhaalt tussen het opzetten van verschillende pipelines.

2.1.3 Vendor lock-in

Er bestaan een aantal diensten, zogenoemde Platform as a Service (PaaS), waarbij je een pipeline kan opzetten. Een van de problemen met een PaaS is vendor lock-in. Dit betekent dat, als er eenmaal een pipeline is opgezet, de overdraagbaarheid van de pipeline naar een andere PaaS vrijwel onmogelijk is. Ook zijn de opties en mogelijkheden om uit te breiden in de toekomst gelimiteerd.

2.2 Doelstelling

De doelstelling is om een systeem te ontwikkelen waarbij developers met weinig tot geen kennis een model kunnen trainen, onderdelen van de pipeline geautomatiseerd zijn en platform agnostisch is.

Het trainen van een model is een iteratief proces omdat de data waarmee het model getrained is verouderd waardoor het model niet meer optimaal presteerd. Om een nieuw model te trainen en consistent te blijven met hoe een model getrained wordt kan een pipeline opgezet worden. Een pipeline is dus herbruikbaar.

2.3 Bestaand oplossingen

Er bestaan een aantal oplossingen voor vrijwel alle problemen dat NGTI ondervindt. Elk oplossing is een PaaS van een derde partij waarbij vendor lock-in inherent is. Dit maakt ze ongeschikt maar betekent echter niet dat ze nutteloos zijn. Er kan namelijk gekeken worden hoe een pipeline wordt opgezet en daar vervolgens (gedeeltelijk) het systeem op baseren.

De oplossingen kunnen gecategoriseerd worden in twee groepen:

1. Machine learning pipeline specifieke services
2. Cloud computing platformen

2.3.1 Machine learning pipeline specifieke service

Bedrijven zoals Algorithmia [1] en Valohai [2] bieden alleen diensten om pipelines op te zetten. Ze zorgen voor het databehoud dat door de gebruiker wordt geuploaded en het trainen van het model. Verder kan er toezicht gehouden worden op de kosten, beschikbaarheid en prestatie van het model.

Valohai heeft documentatie een aantal blog posts die bij het ontwerpen van het systeem relevant zouden kunnen zijn.

2.3.2 Cloud computing platformen

De drie grote cloud computing platformen Amazon, Azure en Google hebben meer te bieden dan alleen een pipeline opzetten, zoals het hosten van een website, database of virtuele server. De cloud computing platformen hebben hetzelfde probleem als de machine learning pipeline specifieke services; vendor lock-in is onvermijdelijk. Wat wel een mogelijkheid zou kunnen zijn is dat de andere services van de cloud computing platformen gebruikt kunnen worden als onderdeel van het systeem.

2.4 Hoofd- en deelvragen

Uitgaand van de drie focuspunten in paragraaf 2.1 kan de hoofdvraag als volgt worden geformuleerd:

Hoe kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform of lokale framework?

De hoofdvraag kan worden onderbouwd met vier deelvragen. Om te beginnen is het verstanding om te onderzoeken hoe een machine learning model wordt getrained:

Welke stappen moeten worden ondernomen om een machine learning-model te trainen?

Vervolgens kan er op de deelvraag voortborduurt worden om het opzetten van een pipeline in kaart te brengen:

Hoe wordt een machine learning-pipeline opgezet?

Verder zijn er verschillende cloud computing platformen en lokale frameworks waarmee machine learning modellen getrained kunnen worden. Doordat het systeem platform agnostisch moet zijn is het van belang om verschillende platformen en frameworks te onderzoeken:

Wat zijn de verschillende en overeenkomsten tussen cloud computing platforms en lokale frameworks waarmee machine learning-modellen kunnen worden getrained?

Ten slotte wordt een PoC gemaakt om te laten zien of het probleem oplosbaar is. Hiervoor is een doordachte voorbereiden onmisbaar:

Hoe ziet de architecturale blauwdruk van een applicatie, waarin een machine learning pipeline kan worden opgezet en die platformonafhankelijk is, eruit?

3 Onderzoeksmethodes

3.1 Kwalitatieve onderzoeksmethode

3.2 Kwantitatieve onderzoeksmethode

3.3 Onderzoeksmethodes bij de hoofd- en deelvragen

D1	**Welke stappen moeten worden ondernomen om een machine learning-model te trainen?**	
	Methode	Kwalitatief en kwantitatief

Tabel 3.1: Deelvraag 1

D2	**Hoe wordt een machine learning pipeline opgezet?**	
	Methode	

Tabel 3.2: Deelvraag 2

D3	**Wat zijn de verschillen en overeenkomsten tussen cloud computing platforms en lokale frameworks waarmee machine learning-modellen kunnen worden getraind?**

Tabel 3.3: Deelvraag 3

D4	**Hoe ziet de architecturale blauwdruk van een applicatie, waarin een machine learning pipeline kan worden opgezet en die platformonafhankelijk is, eruit?**

Tabel 3.4: Deelvraag 4

H	**Hoe kan een machine learning pipeline worden geautomatiseerd onafhankelijk van de onderliggende cloud computing platform of lokale framework?**

Tabel 3.5: Hoofdvraag

4 Requirementsanalyse

4.1 Stakeholders

4.2 Requirements

5 Hoe werkt Machine Learning?

5.1 Soorten Machine Learning

5.2 Wachttijden voorspellen

5.3 Conclusie

6 Hoe werkt een Machine Learning pipeline

6.1 Literatuur

6.2 POC

6.3 Conclusie

7 Cloud computing platformen en lokale frameworks

7.1 Literatuur

7.2 Conclusie

8 Architecturale ontwerp van de oplossing

8.1 Literatuur

8.2 Design

8.3 Conclusie

9 Oplossing

9.1 Scope definiëren

9.2 Research techstack

9.3 Wireframe

9.4 Mockup

9.5 POC

9.6 Conclusie

10 Conclusie

11 Aanbeveling

12 Discussie

13 Reflectie

Bibliografie

[1] Algorithmia Inc. *Algorithmia*. URL: `https://algorithmia.com`.

[2] Valohai. *Valohai*. URL: `https://valohai.com`.

Bijlagen