Scriptie

Machine Learning ten behoeve van het verrijken van klantprofielen voor marketingcampagnes

Auteur:	Tayfun Cakir
Functie:	Afstudeerder Data Scientist/Engineer
Avanade begeleider:	Jeroen van Steenbergen
HvA afstudeerbegeleider:	Ted van Gaalen
Datum:	
Periode:	februari 2017 – juni 2017
Versie:	

1 Versiebeheer

Versienummer	Datum	Wijziging
0.1	8-01-2017	Rapport opzet
0.2	15-01-2017	Aanpassingen hoofdvraag en deelvragen o.b.v. bedrijfsbegeleider.
0.3	6-03-2017	Intervisie 1: Hoofdvraag en deelvragen
0.4	11-04-2017	Intervisie 2: inhoudelijke aanpassingen
0.5	13-04-2017	Verwerking tussentijds feedback o.b.v. feedback van afstudeerbegeleider
0.6	13-4-2017	Toegevoegde waarde van ML-oplossingen beschreven
0.7	26-04-2017	Feedback tussentijdse scriptie o.b.v. afstudeerbegeleider
0.8	1-05-2017	Feedback verwerking o.b.v. bedrijfsbegeleider.
0.9	12-05-2017	Feedback verwerking o.b.v. afstudeerbegeleider

2 Voorwoord

Voor u ligt de scriptie 'Machine Learning ten behoeve van het verrijken van klantprofielen voor marketingcampagnes'. Het onderzoek voor deze scriptie naar machine learning is uitgevoerd bij en voor Avanade. Deze scriptie is geschreven in het kader van mijn afstuderen aan de opleiding Business IT & Management aan de Hogeschool van Amsterdam. Van februari 2017 tot en met juni 2017 ben ik bezig geweest met het onderzoek en het schrijven van deze scriptie.

Samen met mijn stagebegeleider, Jeroen van Steenbergen, heb ik de onderzoeksvraag voor deze scriptie geformuleerd. Het onderzoek dat ik heb uitgevoerd was complex. Na uitvoerig kwalitatief onderzoek en het toepassen van machine learning modellen heb ik de onderzoeksvraag kunnen beantwoorden. Tijdens dit onderzoek stonden mijn stagebegeleider, Jeroen van Steenbergen, en mijn afstudeerbegeleider vanuit mijn opleiding, Ted van Gaalen, altijd voor mij klaar. Door de ondersteuning van beide begeleiders heb ik steeds mijn vragen kunnen beantwoorden, waardoor ik verder kon met mijn onderzoek.

Ik bedank mijn begeleiders voor de fijne begeleiding en ondersteuning tijdens dit traject. Ook dank aan de respondenten Rick Meijvogel en Naween Badloe.

De andere collega's bij Avanade bedank ik voor de fijne samenwerking. Ik heb vaak kunnen sparren met jullie over mijn onderzoek. Tevens wil ik mijn vrouw en familie bedanken omdat zij altijd met me mee dachten.

Ik wens u veel leesplezier toe.

Tayfun Cakir

Utrecht, 9 juni 2017

Inhoudsopgave

3 Samenvatting

Om waarde op te leveren ten behoeve van het verrijken van klantprofielen als input voor marketingcampagnes is het belangrijk om inzicht te verkrijgen in de mogelijkheden met betrekking tot marketingmethoden en strategieën. Een gehanteerde marketingtechniek is het RFM. Dit zorgt ervoor dat op basis van verschillende attributen de recency, frequency en monetary te berekenen zijn. Dit resulteert in de customer value. De RFM-waardes kunnen in combinatie met de overige attributen gebruikt worden ter verrijking van klantprofielen middels machine learning. Echter, is het de bedoeling om op basis van de business situatie een machine learning techniek toe te passen. Wat de type machine learning betreft zal descriptive en predictive analyse gehanteerd worden. De vragen die hierbij beantwoord moeten worden is wat is er gebeurd en wat kan er gebeuren. Voordat machine learning daadwerkelijk toegepast kan worden is het van essentieel belang om een machine learning techniek te hanteren. De bekende machine learning technieken die mogelijk zijn binnen Azure waren supervised en unsupervised algoritmes. Omdat de resultaten gebaseerd zijn op descriptive en predictive analyses zijn de zojuist benoemde machine learning technieken gehanteerd. Hierdoor zal er clustering als classificatie plaatsvinden. Onder deze twee technieken vallen verscheidene algoritmes. Aan de hand van literatuur, die de voordelen en nadelen van de algoritmes beschreven zijn er verschillende algoritmes gehanteerd, namelijk: k-means voor het clusteren en logistische regressie en support vectormachines voor classificatie oftewel de predictive type van analyse. De overige modellen waren niet toepasbaar binnen de verkregenen datasets. Alhoewel de decision tree algoritme hoger scoorde dan support vectormachines, was het nadeel dat decision tree niet toepasbaar was op non-categorical data. Op basis van het toepassen en vergelijken van accuraatheden kwam naar voren dat logistisch de betere combinatie was met k-means voor het clusteren en voorspellen. Het onderzoek naar marktsegmentatie, waarbij verschillende algoritmes waren getoetst bleek te kloppen. Uiteraard is het verstandig om naast deskresearch ook field research uit te voeren. Hierbij zijn twee ervaring volle data engineers met ervaring van machine learning geïnterviewd volgens de ongestructureerde interviewtechniek, waarbij veel open vragen gesteld zijn en bevindingen naar voren kwamen. De combinatie van de k-means en logistische regressie algoritme bleek interessant te zijn. Echter, heeft het wel zijn nadelen. Doordat de situatie van de grote supermarktketen onduidelijk bleef, was het mogelijk dat de aantal clusters kon veranderen, attributen zouden kunnen wijzigen of toegevoegd/verwijderd zouden worden, maar ook hoeveel nieuwe klanten geanalyseerd zouden moeten worden met dit model. Verschillende factoren hebben ernaartoe geleidt dat de combinatie in de huidige toepassing geen meerwaarde had. Uiteindelijk is afgeweken van de oplossing, die werd beschreven aan de hand van literatuur. De resultaten uit het machine learning model, waarbij k-means is toegepast, leveren meerwaarde voor marketingstrategieën en campagnes. Doordat het RFM-marketingmodel is gebruikt voor het clusteren, kunnen zowel frequente als non frequente, veel opleverende klanten als niet opleverende klanten en actieve als non actieve klanten aan de hand van de customer lifecycle meerwaarde leveren voor differentiatie in aanbiedingen, proactieve retentie, kanaal strategie en verbeteringen in customer service. Deze resultaten leveren een hoger verkoopvolume op en een betere klantervaring. Wat de marketingcampagne betreft, zal e-mail segmentatie campagnes gehanteerd moeten worden.

1. Inleiding

Met de komst van internet en apparaten die met elkaar in verbinding zijn, wordt er steeds meer data gecreëerd. Data wordt gezien als een set aan waardes. Marktonderzoeker IDC (Reinsel, 2012) geeft aan dat het digitale universum (lees: data) ieder twee jaar verdubbelt en tussen 2013 en 2014 zal vertienvoudigen. Dit weergeeft de groeiende trend. Zo wordt in (Roth, 2015) vermeld dat 90% van data in de afgelopen twee jaar is geproduceerd. Tevens was in 2013 volgens IDC maar 22% van de digitale universum bruikbaar data en minder dan 5% van de bruikbare data is uiteindelijk geanalyseerd. Uiteraard is het belangrijk om de trend van toenemende hoeveelheid data op lange termijn te bestuderen. Met de opkomst van bijvoorbeeld Internet of Things (IoT), is de verwachting dat in het jaar 2020 meer dan 35% van de data bruikbaar zal zijn. (Reinsel, 2012). Dat is een opwaartse trend ten opzichte van de statistieken uit het jaar 2013.
Volgens een wereldwijd onderzoek uitgevoerd door Veritas Technologies (Data's Dark Side, 2017) blijkt zelfs dat 52% van de opgeslagen data van organisaties wordt beschouwd als ''dark" data. Dit betekent dat de waarde van de desbetreffende data niet bekend is. Tevens geeft het onderzoek aan dat bedrijven in het jaar 2020, 576 miljard pond zullen verspillen aan waarde, die verkregen kan worden uit data (Associates, 2016). Het creëren van waarde uit data kan op verschillende analytics mogelijkheden. Volgens onderzoek van Gartner (Gartner, Market share analysis: Business Intelligence and analytics software, 2013) is Machine Learning het snelst groeiende segment in de analytics markt. Een aantal redenen voor de groei van advanced analytics en daarmee Machine Learning zijn (Gartner, Machine learning drives digital business, 2015):

1. De hoeveelheid data die gegenereerd wordt door klant interacties, social media en voornamelijk sensoren van bijvoorbeeld verbonden apparaten en machines.
2. De realisatie dat traditionele computer engineering een bottleneck is geworden als het gaat om het leveren van kost effectieve oplossingen
3. De beschikbaarheid van minder dure geheugenopslag, snellere processors en cloudoplossingen.

Daarnaast geeft Microsoft aan dat de hedendaagse predictive analytics systemen vanwege de volgende trends een snelle groei meemaken (Barnes, 2015):

1. Exponentiele groei in data
2. Lage kosten voor digitale opslag
3. Computing power
4. De opkomst van big data analytics

Over het algemeen zijn de benoemde redenen omtrent de groei van machine learning overlappend. Zoals eerdere onderzoeken hebben aangetoond is er veel uit data te halen. Bill Gates, oprichter van Microsoft, geeft aan: "A breakthrough in machine learning would be worth ten Microsofts." De verwachtingen in de komende jaren omtrent bruikbaarheid van data geeft aan dat data gezien kan worden als "goldmines" voor business kansen en dat deze trends zullen groeien in de nabije toekomst. Met de huidige groei in data en de daarmee

mogelijke kansen is de vraag naar het toepassen van machine learning gegroeid. Verschillende bedrijven zien deze kansen en investeren in machine learning door de positieve return on investment (ROI) resultaten. (Briewald, 2016)

3.1 Over Avanade

Avanade is een joint venture van Accenture en Microsoft. De kracht van de drie maakt Avanade uniek en geeft het bedrijf een zeer groot voordeel op het gebied van concurrentiepositie. Dit komt o.a. door de werknemers, business development, brand en training. Avanade transformeert bedrijven voor het digitale tijdperk, waarbij het doel is de cliënt en daarmee de klanten van de cliënt te helpen. Dit wordt gerealiseerd door 23.000 professionals met verschillende achtergronden op 80 locaties in 23 landen. Avanade heeft sinds het jaar 2000 met meer dan 4000 klanten gewerkt, waarvan 43% klanten afkomstig zijn uit Global 500 en 34% uit Fortune 500. "Avanade levert innovatieve digitale en Clouddiensten, zakelijke en op gebruikerservaring gebaseerde oplossingen voor zijn klanten, aangedreven door de kracht van mensen en het Microsoft ecosysteem." (Avanade, 2017). Avanade realiseert resultaten in verschillende bedrijfstakken zoals o.a. bankwezen, retail, overheidsinstanties, ziekenhuizen, consumentengoederen en verzekeringen.

De oplossingen die Avanade o.a. biedt aan haar klanten zijn als volgt:

- **Enterprise Resource Planning(ERP)**: De aangeboden ERP-oplossingen geven de klanten van Avanade een flexibel en geïntegreerd infrastructuur, zodat de technologie van de klant afgestemd kan worden op de zakelijke behoeften. Het gaat hier voornamelijk om het efficiënter laten werken van bedrijven en het kunnen aanpakken van knelpunten.

- **Managed services**: Avanade is een specialist op het gebied van managed services binnen het ecosysteem van Microsoft technologie. Avanade helpt bedrijven te veranderen naar waardevolle managed services door kosten te verlagen, flexibiliteit te verhogen en groei te stimuleren.

- **Cloudtransformatie**: Door de realisatie van cloud transformatie, kan er sneller gereageerd worden op nieuwe kansen. Daarnaast wordt de time-to-market initiatieven verkort met bedrijfsapplicaties die passend zijn gemaakt voor de cloud en is het ook mogelijk om de beschikbaarheid van resources aan te passen.

- **Analytics**: Avanade biedt verscheidene oplossingen op het gebied van analytics. Bijvoorbeeld machine learning, powerBI en SQL.

- **CRM**: Avanade biedt oplossingen op het gebied van CRM door snel te reageren op de behoeften van de klant, nieuwe waarde te creëren en de betrokkenheid van de klant te vergroten d.m.v. digitale klantervaring.

De visie van Avanade is om het voortouw te nemen als digital innovator, waarbij resultaten worden gerealiseerd voor de klanten en hun klanten middels de expertise van de medewerkers en het ecosysteem van Microsoft. (Avanade, 2017).

Zoals eerder in paragraaf 1.1 is beschreven is Avanade werkzaam in verschillende sectoren. Avanade noemt de afdelingen ook wel taco's. Dit staat voor talent community. Mijn opdracht en functie behoren tot de analytics afdeling.

3.2 Relevante en actuele ontwikkelingen van Avanade

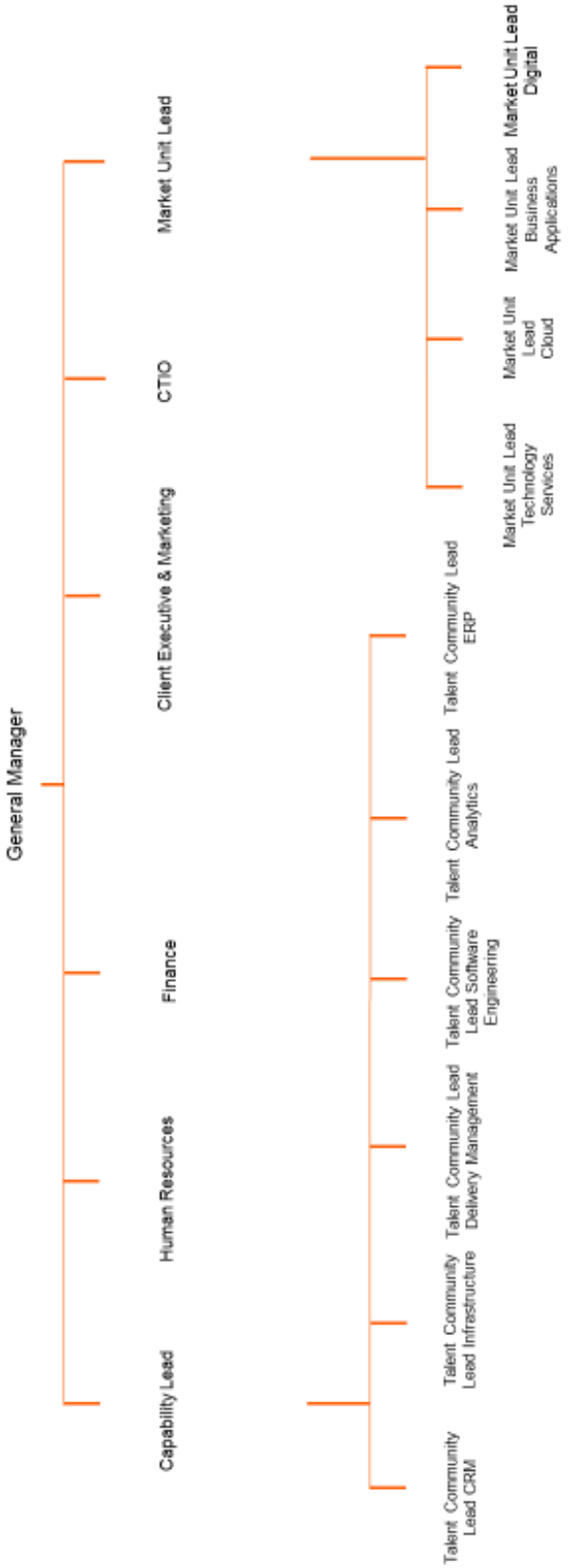
Digital en Cloud technologieën brengen wereldwijd gigantische veranderingen binnen alle branches. (Avanade, 2017) Zo is er al een aantal voorbeelden van gedigitaliseerde sectoren zoals ziekenhuizen en overheidsinstanties. In de toekomst zal er een aantal grote verstoringen plaatsvinden op het gebied van klant ervaring, connectiviteit, maar ook security en privacy. (Avanade, 2017) . Dit geeft de aandachtsgebieden voor in de toekomst weer.

De klanten van Avanade worden voornamelijk in de volgende 5 processen geholpen:

1. Digital customer journey
2. Transformatie naar de digitale werkplek
3. Moderne IT en Cloud transformatie
4. Gebruik van data en analytics
5. Opnieuw uitvinden van bedrijfsprocessen

De transformatie van de benoemde processen uit de activiteiten van Avanade resulteert in het bewust worden van de versmelting tussen fysieke en digitale belevingen. Dit betekent dat bedrijven inzien dat fysieke processen toepassingsmogelijkheden hebben voor het digitaliseren van de fysieke processen.

3.3 Organogram



3.4 Probleemstelling

In het algemeen zorgt de stijging van de hoeveelheid van data voor het overschrijden van de desbetreffende limiet van databronnen. Dit zorgt volgens (Goeyenbier, 2014) voor grote opslag en- verwerkingscapaciteit, waardoor de kosten oplopen. Tevens wordt er meer informatie opgeslagen en ontstaan er meer mogelijkheden voor het ontdekken van patronen en trends. Bij het analyseren van data kan er gebruik gemaakt worden van advanced analytics. Bij het toepassen van advanced analytics behoren data mining, patroon matching en machine learning tot de gebruikelijke technieken. Machine learning wordt gebruikt voor het uitvoeren van voorspellingen en ontdekken van patronen. Tijdens deze opdracht wordt er gekeken naar de mogelijkheden van het creëren van nieuwe inzichten en waardes op basis van advanced analytics m.b.v. machine learning.
Met betrekking tot deze opdracht kijkt een grote supermarktketen naar de mogelijkheden ter verbetering van het huidige verkoopvolume. Met de toename van data is het interessant geworden om te kijken naar welke waardes gegenereerd kunnen worden met gebruik van machine learning. De databronnen die gebruikt worden ten behoeve van het verbeteren van het verkoopvolume zijn:

1. Customer base date van Customer Relationship Management (CRM)
2. Sales data van point of sales (PoS)
3. Products data van datawarehouse (DWH)

Het is bovendien niet bekend welke inzichten en waarde er gecreëerd kan worden bij het toepassen van Machine Learning algoritmes. De vraag is dan ook hoe er waarde uit de verkregen databronnen gecreëerd kan worden met behulp van Machine learning, waarbij de gerealiseerde waarde dient als input voor marketingcampagnes en daarmee het verhogen van het verkoopvolume. Omdat de oplossing gericht is op marketing zal er gekeken worden naar het verrijken van klantprofielen, waarbij de mogelijkheden van het gebruik van marketingtechnieken onderzocht zullen worden. Het onderzoek naar marketingtechnieken zullen een bijdrage leveren voor het vinden van klant waarde. Het is overigens ook belangrijk om de evaluatie van de mogelijke data lake in acht te nemen. Dit is namelijk van essentieel belang, omdat de supermarktketen bewust is van o.a. de gelimiteerde DWH.

De Term data lake werd voor het eerst genoemd in een blog dat afkomstig is van de CTO van Business intelligence specialist Pentaho, James Dixon (Dixon, 2010):

"Een datamart/warehouse kan je vergelijken met een winkel voor flesjes bronwater. Het water is gezuiverd, gestructureerd verpakt en op deze manier geschikt voor eenvoudige consumptie. Een data lake is dan de waterbron in haar natuurlijke staat. De inhoud van de bron is ook water, maar ongezuiverd en nog niet verpakt. De waterbron kan bovendien ook voor andere doeleinden gebruikt worden."

De verschillen tussen een datawarehouse en data lake zijn volgens (Dull, 2015) als volgt:

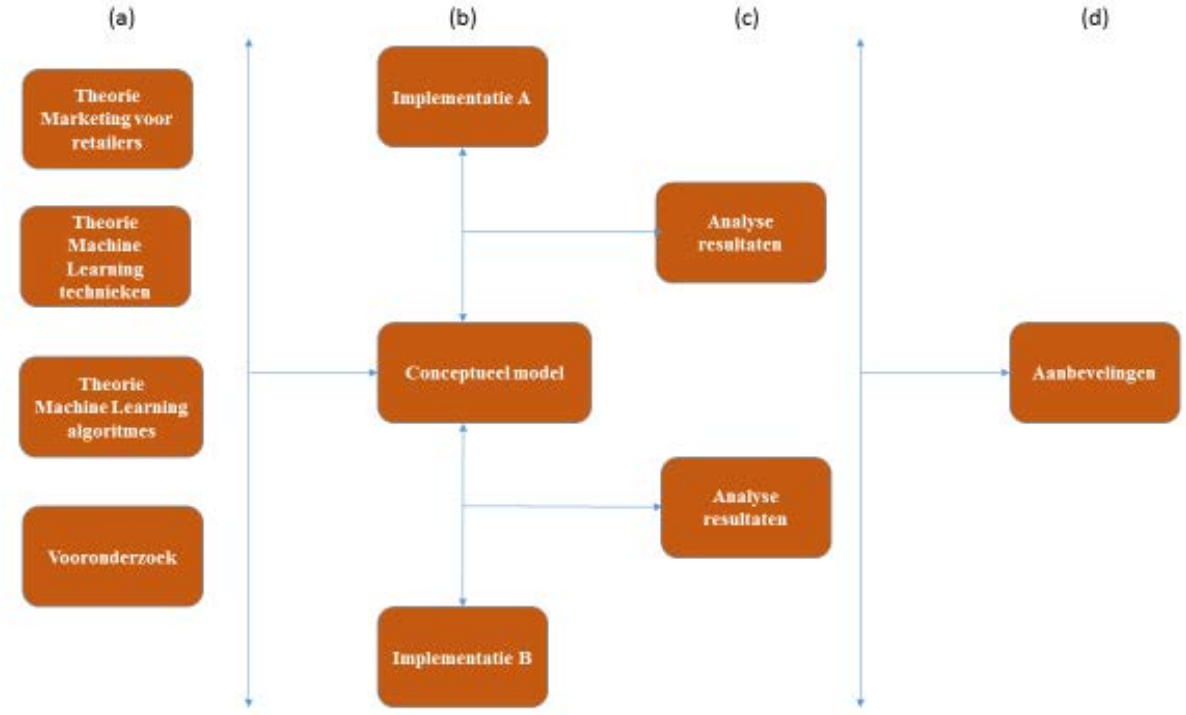
Data Warehouse	**VS**	Data lake
Gestructureerd, verwerkt	**DATA**	Gestructureerd/semi-gesturctureerd/ ongestructureerd/ raw
Schema-on-write	**VERWERKING**	Schema-on-read
Duur voor grote hoeveelheid aan data	**OPSLAG**	Ontworpen voor lage opslag kosten
Minder agile, fixed configuratie	**AGILITY**	Erg agile, configureer mogelijkheid indien nodig.
Mature	**BEVEILIGING**	Maturing
Business professionals	**GEBRUIKERS**	o.a data scientist

3.5 Doelstelling

Het doel van dit onderzoek is het doen van aanbevelingen aan Avanade over de mogelijkheden en toepassingen van Machine Learning. Deze aanbevelingen kunnen voor een hogere omzet/verkoop/ betere/specifiekere marketingcampagnes leiden voor de klanten van Avanade, waaronder de supermarktketen. Door het toepassen van Machine Learning kunnen naast verhogen van het verkoopvolume ook verbeteringen op de klantervaring tot stand komen. In dit onderzoek zijn de oplossingen gericht op het verhogen van het verkoopvolume. Dit wordt gedaan door inzicht te geven in de toepasbare classificatie, clustering of regressie modellen, die gehanteerd kunnen worden voor het voorspellen van gegevens (variabelen), waarbij wordt getracht waarde te leveren op basis van de beschikbare datasets. Naast het onderzoek zullen ook implementaties gerealiseerd worden met Azure Machine Learning, die op basis van de onderzochte literatuur tot stand komen. Het gebruik van Azure Machine Learning door Avanade is op basis van het gebruik van het Microsoft eco-systeem. De toepasbare technieken en modellen zullen hierop gebaseerd zijn.
Deze keuze zorgt ervoor dat huidige en toekomstige verbeteringen en oplossingen m.b.t. machine learning toegepast kunnen worden in projecten of opdrachten. Tevens worden de algoritmes en datasets verwerkt in de Cloud van Azure. Dit zal meerwaarde bieden voor de proof of concept uitwerking van het machine learning model voor zowel Avanade en de klant van Avanade. Het proof of concept zal deels ook geschreven worden in de programmeertaal R.

3.6 Onderzoeksmodel

De onderstaande onderzoeksmodel in figuur 1 geeft de stappen weer, die uitgevoerd moeten worden om uiteindelijk tot een aanbeveling te komen. Het onderzoeksmodel is tot stand gekomen op basis van het boek (Piet Verschuren, 2015).



Figuur 1. Onderzoeksmodel

3.6.1 Verwoording van het model

Een bestudering van toepassingen van predictive analytics voor Avanade, gebaseerd op bestaande literatuur omtrent marketing voor retailers, machine learning algoritmes en machine learning technieken, alsmede een vooronderzoek, (b) leveren een conceptueel model, waarmee de mogelijkheden van machine learning kan worden geëvalueerd op basis van implementatie op verschillende scenario's. (c) Een vergelijking van deze analyseresultaten resulteert in (d) aanbevelingen voor de mogelijkheden van het verrijken van klant profielen als input voor marketingcampagnes met machine learning.

4 Onderzoeksvraag

Hoe kan Machine Learning worden ingezet om waarde op te leveren ten behoeve van het verrijken van klantprofielen als input voor marketingcampagnes?

4.1 Deelvragen

1. Hoe kan Machine learning toegepast worden met betrekking tot marketing?
2. Welke learning technieken zijn toepasbaar op de verkregen datasets?
3. Welke Machine Learning algoritmes zijn geschikt om waardes uit de datasets nauwkeurig te voorspellen?
4. Wat is het verschil in performance tussen de verschillende toepasbare Machine Learning algoritmes?
5. Hoe kunnen de gecreëerde resultaten meerwaarde bieden voor marketingcampagnes?

4.2 Eindresultaat

Het eindresultaat zijn aanbevelingen van de mogelijkheden met Azure Machine Learning en een proof of concept middels de verkregen dataset, waarvan de gegevens uit de dataset getransformeerd en toepasbaar zijn gemaakt voor Machine Learning. Tevens levert het gemodelleerde Machine Learning model de beste waarde ten behoeve van het verrijken van klantprofielen als input voor marketingcampagnes. Dit wordt gerealiseerd op basis van onderzochte literatuur en theorieën omtrent marketingmogelijkheden met betrekking tot data, machine learning technieken en Machine Learning algoritmes oftewel modellen. Overigens zullen potentiële gevonden resultaten ook als eindresultaat dienen.

4.3 Onderzoeksstrategie

In het boek van (Piet Verschuren, 2015) wordt aangegeven dat het meest bepalende beslissing, die een onderzoeker maakt bij een technisch ontwerp, de keuze is van een onderzoekaanpak. In het boek wordt dit ook wel onderzoeksstrategie genoemd. Dit is een samenhang van beslissingen tijdens het onderzoek. In dit onderdeel gaat het voornamelijk over het vergaren van relevant materiaal tot het verkrijgen van antwoorden op de hoofd en deelvragen. De keuze bestaat uit vijf verschillende strategieën, die zich onderscheiden in de manier van onderzoek naar de onderzoeksvraag bestaande uit deelvragen.

- Survey
- Experiment
- Casestudy
- Gefundeerde theoriebenadering
- Bureauonderzoek

Voor dit onderzoek zal er gekozen worden voor bureauonderzoek, waarbij de informatie gebaseerd zal zijn op bestaande literatuur en/of door anderen bijeengebracht materiaal (bv nota's, archieven, databanken, verslagen van onderzoekers en literatuur etc.)
De wijze van onderzoek is voor bureauonderzoek in meeste gevallen kwalitatieve en kwantitatieve analyse. Dit kunnen machine learning model algoritmes met resultaten zijn, dat door onderzoekers samengebracht materiaal is. Tevens zal exploratief onderzoek worden

verricht op basis van de modellen die toepasbaar zijn op de verkregen dataset. De verschillende mogelijkheden zullen worden getoetst en tevens zullen de resultaten geanalyseerd worden om tot een concrete aanbevelingen te komen.

4.4 Voor- en nadelen

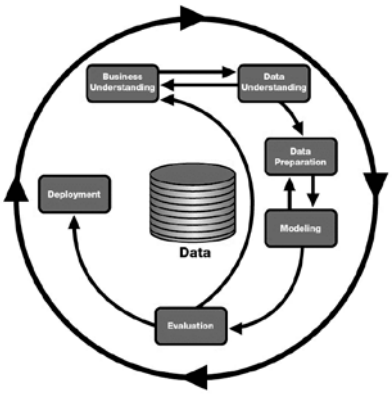
Dit stuk behandelt de voor- en nadelen voor de gekozen strategie, die gehanteerd zal worden voor het uitvoeren van het onderzoek. Het voordeel van bureauonderzoek is dat er veel beschikbare gegevens zijn, die gebruikt kunnen worden om het onderzoek uit te voeren. Het nadeel daarentegen is dat de onderzoeksvraag alleen beantwoord kan worden op basis van de beschikbare informatie uit literatuur en samengebrachte materiaal.
De onderzoeksvraag is dus afhankelijk van de beschikbare gevonden bronnen. Het risico is dat het doel en- onderzoeksvraag moeten worden aangepast. Zoals eerder aangegeven zal mijn onderzoek ook gedeeltelijk explorerend zijn bij het toepassen van de theoretische kennis omtrent machine learning en dus het vinden van relaties en waarden in de verkregen datasets. Waardoor de gewenste resultaten met betrekking tot de output van de waarde niet resulteert in een hogere omzet/verkoop/ betere/specifiekere marketingcampagnes leiden voor de klanten van Avanade.

4.5 Methodologie

Volgens Kdnugget (Kdnugget, 2014) is CRISP-DM de meest gebruikte methodologie voor data mining. CRISP-DM staat voor Cross Industrial Standard Process for Data Mining. Dit is een proces, waarbij machine learning algoritmes toegepast worden. De gehanteerde bron dateert uit 2014, waarbij in de tussentijd nieuwe modellen zijn ontstaan. De methodologie zou geüpdatet worden tussen 2006 en 2008, maar dat heeft niet plaatsgevonden. Tevens is de website van de originele CRISP-DM niet langer actief. (Have you seen ASUM-DM, 2015).
De website is in zijn laatste staat overgenomen en gehost door Smart Vision Europe. (Smart Vision Europe, 2000).

4.5.1 CRISP-DM

Onderstaande illustratie geeft de CRISP-DM-methodologie weer:



Figuur 2. Crisp-DM

De fases van CRISP-DM zijn als volgt (SPSS, 1999, pp. 10-11):

- **Business understanding**
  - Deze fase focust zich op het begrijpen van de business, waarbij er wordt gekeken naar de vereisten. Deze opgedane kennis wordt vervolgens getransformeerd naar een data mining probleemdefinitie, waarbij een plan nodig is om deze aan te pakken.

- **Data understanding**
  - In deze fase begint het verzamelen van data en dergelijke activiteiten om bekend te raken met de desbetreffende data. In deze fase wordt voornamelijk gekeken naar de datakwaliteit problemen en het ontdekken van inzichten of het bedenken van hypotheses op basis van de data.

- **Data preparation**
  - Deze fase bevat alle activiteiten die de uiteindelijke dataset moeten realiseren. Transformeren en schoonmaken van data zijn o.a. de activiteiten die in deze fase voorkomen.

- **Modeling**
  - Deze fase gaat om het modeleren van de data om de gewenste resultaat te krijgen.

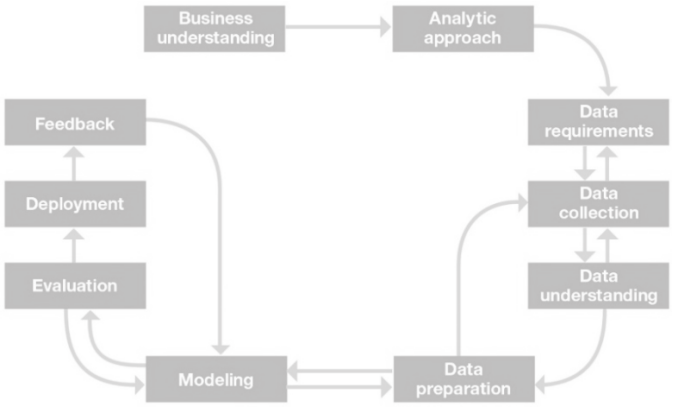
- **Evaluation**
  - De evaluatie vindt in deze fase plaats. In deze fase wordt er gekeken naar de verkregen resultaten en in hoeverre de resultaten, de business doelen bereikt hebben.

- **Deployment**
  - Het inzetten van het model dat in de fase modeling, gemodelleerd is.

4.5.2 ASUM-DM

In 2015 kwam IBM met een vernieuwde methodologie genaamd ASUM-DM. Dit staat voor Analytics Solutions Unified Method for Data Mining/Predictive Analytics (Have you seen ASUM-DM, 2015). Deze methodologie is gedetailleerder en is daarom een geschikter methodologie voor het toepassen van predictive analytics. Volgens (Have you seen ASUM-DM, 2015) dekt crisp-dm de infrastructuur/operationele gedeelte van het implementeren van datamining/predictive analytics projecten niet. Tevens bevat het weinig projectmanagement activiteiten en taken. En is het weinig gericht op de activiteiten en taken in de deployement fase. ASUM-DM wordt gezien als een extensie en verbeterd model van CRISP-DM. Een alternatief op de CRISP-DM en ASUM-DM is de methodologie SEMMA (SAS, 2016). SEMMA staat voor sample, explore, modify, model, en acces. Deze methodologie is ontworpen met de gedachtegang voor het toepassen van de ontwikkelde Enterprise miner tool van SAS (Hampton, 2011). Terwijl CRISP-DM en ASUM-DM open staat voor verschillende tools. Deze methodologie zal toegepast worden tijdens de uitwerking van de opdracht op basis van de informatie uit dit onderzoek. Tevens is het met dit model mogelijk om vaker een fase terug te gaan, indien blijkt dat in een van de fases geen correcte keuzes zijn gemaakt. Deze mogelijkheid is door de structuur van het model bij crisp-DM minder het geval. Volgens IBM (IBM Big Data & Analytics Hub, 2015) zorgt de ASUM-DM-methodologie ervoor dat het proces iteratief is. Modellen worden niet eenmalig gemaakt en ongewijzigd gelaten. Door feedback te vergaren vinden er aanpassingen plaats en kan het model verbeterd worden.

De fasen van ASUM-DM zijn (IBM Big Data & Analytics Hub, 2015):



Figuur 3. ASUS-DM

- **Business understanding**
  - Net zoals bij crisp-dm is het belangrijk om de business te begrijpen. Door het begrijpen van de business wordt de fundering gelegd voor een succesvolle oplossing van de business problemen. Het is belangrijk om het probleem, doelstellingen en requirements vanuit het business perspectief te definiëren.

- **Analytic approach**
  - Nadat de business probleem goed is geformuleerd kan de data scientist de analytische nadering toepassen voor het oplossen van het probleem. Hierbij

komen zowel statistische als machine learning technieken naar voren, waarbij de data scientist de technieken kan gebruiken die geschikt zijn.

- **Data requirements**
  - De keuze voor de analytische nadering bepaalt de data requirements. Hierbij zullen formats en o.a. data content een rol spelen bij de keuze van requirements. De requirements zijn geleid door domein kennis.

- **Data collection**
  - Het verzamelen en identificeren van verzamelde data. Deze kunnen zowel gestructureerd, ongestructureerd als semigestructureerd zijn. Deze resources zijn relevant voor het domein van het probleem. Bij het krijgen van gaps tijdens de verzameling, kan de data scientist ervoor kiezen om meer data te verzamelen en de requirements te herzien.

- **Data understanding**
  - Mogelijkheden van statistische en visualisatie helpen inzicht te geven in de data, waardoor de data scientist de data kan begrijpen. In deze fase is het ook belangrijk de kwaliteit en inzichten in de data te ontdekken en in beeld te brengen.

- **Data preparation**
  - Deze fase kent een aantal activiteiten voor het voorbereiden van de data. Hierbij kunnen we denken aan data cleaning, het combineren van data vanuit verschillende databronnen en het transformeren van data naar nuttige variabelen. Deze fase zal het meest tijd in beslag nemen. Volgens IBM (IBM Big Data & Analytics Hub, 2015) neemt deze fase het over het algemeen 70% van de tijd in beslag.

- **Modeling**
  - In eerste instantie wordt er gewerkt met de eerste versie van de voorbereide data set. Data scientists gebruiken hiervoor een training set. Historische data wordt gebruikt, waarbij de uitkomst meestal al bekend is. De bedoeling is om voorspellende modellen te ontwikkelen in combinatie met het beschrijven van de analytisch benadering. Belangrijk is dat de modeling fase iteratief is.

- **Evaluation**
  - Hierbij wordt er gekeken naar de kwaliteit van het model en wordt er gekeken of de resultaten de business probleem verhelpen en/of beantwoorden. Hiervoor

kunnen diagnostische metingen toegepast worden, maar ook tabellen en grafieken.

- **Deployment**
  - Zodra het model naar wens is gemodelleerd en de resultaten voldoen aan de verwachtingen, kan het model worden ingezet.

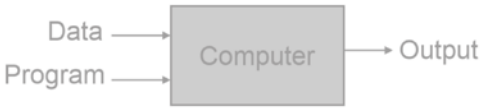
- **Feedback**
  - Door het verzamelen van resultaten van het geïmplementeerd model, krijgt de organisatie feedback op de performance, waardoor de data scientist de nuttigheid van het model kan verbeteren door o.a. de accuraatheid te verhogen.

5 Theoretisch kader

Voordat er in de komende hoofstukken getracht wordt de deelvragen ten behoeve van de hoofdvraag te beantwoorden, is het van belang om de context van deze scriptie duidelijk te krijgen. In paragraaf 3.1 zal de term machine learning toegelicht worden. Tevens zal in 3.2 het beeld van big data verduidelijkt worden en hoe machine learning een rol speelt in big data. In paragraaf 3.3 worden de termen overfitting en underfitting gedefinieerd, voor het verduidelijken van voorkomende problematieken bij het toepassen van machine learning. Vervolgens zal in 3.4 de ROC-curve toegelicht worden. Dit zal in de uitwerking van de opdracht vaak gebruikt worden voor het constateren van de accuraatheid van de machine learning modellen. Vervolgens zal in paragraaf 3.5 de evaluatie methodiek beschreven worden. De evaluatie methodiek van paragraaf 3.5 heeft veel samenhang met de ROC-curve in paragraaf 3.4. Tevens zal in 3.6 de analytics sector behandeld worden met onderzoeks- en adviesbureau in de informatietechnologie-sector genaamd Gartner. Hierbij zal de hype cycle besproken worden.

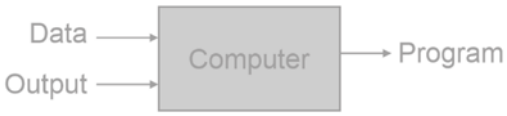
5.1 Wat is machine learning?

Volgens (Barnes, 2015, p. 13) kan Machine learning gezien worden als computing systemen die zich verbeteren aan de hand van ervaring. Data scientists hebben bepaalde methodes ontwikkeld die getraind en gebruikt worden met een hoog volume aan data, voor het voorspellen van patronen en bijvoorbeeld trends. Machine learning is een middel om de gewenste resultaten of voorspellingen te realiseren, door gebruik te maken van onder andere historische data. De beste manier om machine learning te omschrijven is door het te vergelijken met de hedendaagse moderne computer programming paradigma. Met traditionele programming modellen worden programma's en data voor gewenste resultaten verwerkt door de computer. Hierbij kan er gedacht worden aan het gebruiken van programma's voor het verwerken en produceren van bijvoorbeeld een report. (Barnes, 2015, p. 14).



Figuur 4. Traditionele programming systeem

Met machine learning is het verwerking paradigma drastisch veranderd. Data en de gewenste output zijn omgewisseld door de computer voor het produceren van nieuwe programma's. (Barnes, 2015, p. 14)



Figuur 5. Machine learning systeem

Voorbeelden van het toepassen van predictive analytics zijn: (Barnes, 2015, p. 19):

- Spam/junk email filters
  - Op basis van content, header, maar ook gedrag van de gebruiker zijn bruikbare informatie voor het filteren van bepaalde spam/junk mails.
- Patronen herkenning
  - Hierbij kunnen denken aan spraakherkenningen op smartphones, maar ook het herkennen van gezichten voor o.a. beveiligingscamera's.
- Huizenprijzen
  - Het voorspellen van huis prijzen op basis van de grootte van het huis.
- Creditcard fraude detectie
  - Het proces voor het herkennen van fraude is gebaseerd op bepaalde activiteiten, bij gebruik van een creditcard.
- Predictive onderhoud
  - Het monitoren van vliegtuigen, treinen, liften, en bijvoorbeeld auto's.

5.2 Big data

De hoeveelheid van data speelt een groot rol bij het gebruiken van Machine learning. Zoals eerder aangegeven wordt machine learning gebruikt voor het vinden van o.a. patronen en trends. Maar wanneer is data "big"?
Gartner analist Doug Laney (Gartner, 2013) omschreef in het jaar 2001 de 3V's. Deze 3V's omschrijven big data.

Volume: De omvang van de data speelt een rol, oftewel: hoeveel geheugen neemt het in beslag. Het kan gaan om terabytes waarvan de data afkomstig is van transacties, maar ook data uit social media of data uit sensoren.

Variety: Data kan afkomstig zijn uit verschillende bronnen. Deze databronnen kunnen gestructureerd zijn, maar ook ongestructureerd. Ongestructureerde data bestaat uit ruwe en

ongeorganiseerde data. Voorbeelden van ongestructureerde data zijn: video, social media en o.a. sensor data. (Sherpa Software)

Velocity: Data kan in batches worden verwerkt, maar door sensoren en het internet is streaming aan data gebruikelijk. Real-time aan data maakt data o.a. ook "big".

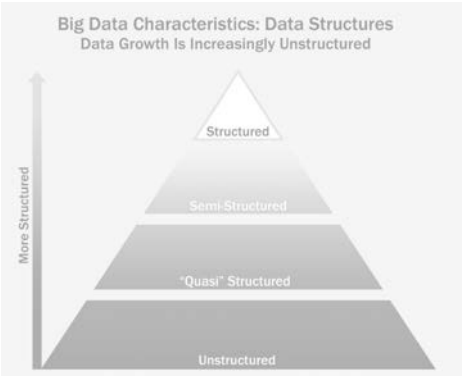
Op basis van de opgedane kennis uit de advanced analytics training, blijkt er een vierde v te bestaan, namelijk:

Veracity: Het detecteren en corrigeren van ruis en inconsistente data zijn belangrijk voor het uitvoeren van vertrouwelijke analyses.

De vierde V is afkomstig van IBM (The Four V's of Big Data, 2014). De volgende figuur illustreert de vier V's.

Figuur 6. 4V's Big Data

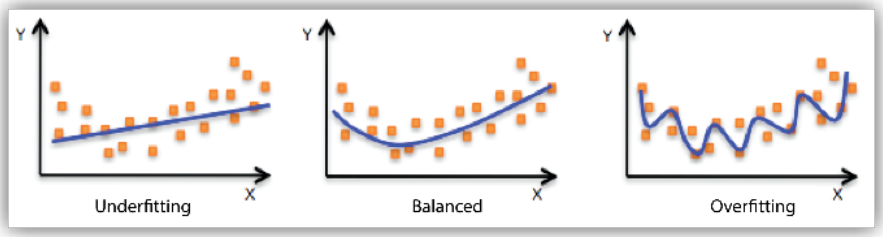
Volgens (EMC Education Services, 2015, p. 25) zal in de toekomst 80 tot 90% van data ongestructureerd zijn. De niveaus van gestructureerd tot ongestructureerd data zien er als volgt uit:



Figuur 7. Datastructuur verhouding

5.3 Overfitting en underfitting

Het kennen van de termen overfitting en underfitting zijn binnen machine learning van essentieel belang. Bij overfitting gaat het omtrent het model dat een te goede "fit" kent op de training set. Hierdoor presteert het model slecht op nieuwe samples uit de training set. (EMC Education Services, 2015, p. 204). Volgens (Model Fit: Underfitting vs. Overfitting, 2016) presteert het model goed op de training data, maar niet op de evaluatie data. Het model memoriseert de data, waardoor het niet toegepast kan worden op nieuwe data. Een voorbeeld van overfitting is dat het ervoor heeft gezorgd dat de accuraatheid van decision tree learning wordt verlaagd met 10-25% (Mitchell, 1997, p. 68). Volgens (Model Fit: Underfitting vs. Overfitting, 2016) kan op basis van de prediction error van de training data en evaluatie data beslist worden in hoeverre het model overfitting of underfitting is. Op het moment dat het model slecht presteert i.v.m. het niet kunnen realiseren van de relaties tussen de input en target waarden, is het model underfitting.



Figuur 8. Overfitting en underfitting visualisatie

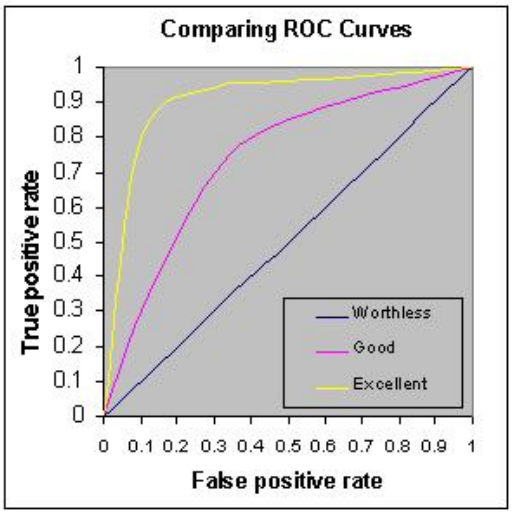
Om overfitting te voorkomen is het volgens (EMC Education Services, 2015, p. 177) gebruikelijk om het gehele dataset willekeurig te splitten in een training en testing set. Zodra het model is toegepast kan evaluatie plaatsvinden op basis van de testing set. Op het moment dat de dataset gering is voor het creëren van een training en testing set, kan de N-fold-cross validation techniek gebruikt worden. Deze techniek kan behulpzaam zijn bij het vergelijken van de "fitted" modellen. De N-Fold cross -validatie techniek werkt als volgt:

1. Het gehele dataset wordt willekeurig gesplitst in een aantal datasets van gelijkwaardige grootte.
2. Een model wordt getraind met het aantal gesplitste datasets en wordt vervolgens getest tegen de overige data in de dataset. In deze fase wordt de meting van het aantal errors gerealiseerd.
3. Het proces wordt herhaald op basis van het totale datasets met de verschillende combinaties van datasets.
4. Verzameling van het gemiddelde van het geobserveerde aantal foutmeldingen van de modellen.

De gemiddelde error van een model wordt vergeleken met het gemiddelde error van een ander model. De benoemde techniek kan overigens ook gebruikt worden voor het beslissen van extra toevoegingen omtrent variabelen dat ten koste van mogelijke overfitting kan plaatsvinden.

5.4 Receiver operating characteristics (ROC) curve

De term "receiver operating characteristic" dateert uit de tweede wereldoorlog, waarbij het diende voor het opvangen van object signalen en ruis. (Jerome Fan, 2006). Het is een vorm van objectief meten, dat gebruikt kan worden voor het vergelijken van de prestaties. (David J. Vining, 1992). Het doel is om een zo hoog mogelijke prestatie te realiseren door de juiste sensivity (true positives) en specificity (true negatives) te hebben. De prestatie wordt in dit geval verstaan als accuraatheid. De ROC-curve plot deze prestatie middels een lijngrafiek zoals te zien is in het onderstaande figuur.



Figuur 9. ROC Curve visualisatie

Zodra de waarden van sensivity en specificity hoog zijn, is de prestatie van het desbetreffende model erg hoog. De "area under the curve" oftewel AUC dient als onderscheidend vermogen van diagnostische tests (Jerome Fan, 2006). De AUC wordt berekend aan de hand van het gebied onder de ROC-curve (EMC Education Services, 2015, p. 227) De waardes liggen tussen de 0.5 en 1. Een AUC van 0.5 indiceert een willekeurig classificatie zonder waarde (Zhang Z. , 2016, p. 5). Dat betekent dat een hoge prestatie te realiseren is, waarvoor de AUC een waarde van 1.0 nodig heeft. (Jerome Fan, 2006).

5.5 Evaluatie methodieken

Op het moment dat de classificatie heeft plaatsgevonden, is het mogelijk om een evaluatie op de resultaten uit te voeren. Dit kan door o.a. te kijken naar de ROC-curves en AUC, maar ook de accuracy, precision, recall en confusion matrix (EMC Education Services, 2015, p. 230). De volgende termen worden o.a. gebruikt voor het berekenen van de ROC-curve:

1. **TPR**
   True positive ratio wordt berekend door het aantal true positives te delen door het aantal positives.

2. **FPR**
   False positive ratio wordt berekend door het aantal false positives te delen door het aantal negatives

Voor het berekenen van de accuracy, precision en recall is het van belang om de definities van TN, TP, FN en FP te kennen. Deze worden in de confusion matrix weergeven en is voor de evaluatie van het desbetreffende machine learning model.

TN
Dit staat voor true negative en dit vindt plaats als de casus negatief was en de voorspelling ook negatief bleek te zijn.

TP
Dit staat voor true positive en dit vindt plaats als de casus positief was en ook positief werd voorspeld.

FN
Dit staat voor false negative en dit vindt plaats als de case positief was, maar de voorspelling negatief bleek te zijn.

FP
Dit staat voor false postive en dit vindt plaats als de case negatief was, maar de voorspelling positief bleek te zijn

Voor het berekenen van de accuracy kan de volgende formule toegepast worden: TN + TP / TN + TP + FN + FP. Bij het berekenen van de recall oftewel de hoeveelheid van de true positives die gevonden zijn, kan de volgende formule worden toegepast: TP / FN + TP. Tevens is het mogelijk om de precision uit te rekenen en dat kan door het toepassen van de volgende formule: TP / FP + FN. (KDnuggets, 2017) (EMC Education Services, 2015, p. 280). Deze formules zullen in Azure Machine learning of andere machine learning tools normaliter geautomatiseerd zijn.

5.6 Machine learning trendanalyse

Zoals eerder aangegeven in paragraaf 3.1 kent analytics vier gebieden. Een van deze gebieden is predictive analytics. Machine learning is o.a. een onderdeel uit predictive analytics. Gartner is werelds leidend technologie en onderzoeksbureau (Gartner, 2017). Ieder jaar brengt Gartner de hype cyclus uit. De hype cycle illustreert de trends van technologieën. Deze informatie schetst de adoptie en verwachtingen van de markt. De verwachtingen van machine learning zijn hoog en is dus hot topic (Gartner, 2016). Tevens vindt de adoptie van machine learning zoals in figuur 10 wordt afgebeeld binnen twee tot vijf jaar plaats.

Figuur 10. Gartner hype cycle 2016

Dit komt in combinatie met toenemende hoeveelheid aan data en daarmee technologische mogelijkheden, zoals machine learning voor het ontdekken van o.a. patronen overeen met de benoemde punten in de probleemstelling. Databronnen worden gelimiteerd, uitbreidingsmogelijkheden zoals bijvoorbeeld Cloud behoren tot de oplossingen en door het verzamelen van meer attributen en informatie worden de mogelijkheden van machine learning toegankelijker gemaakt.

6 Machine learning en marketing

Machine Learning modellen kunnen op verscheidene manieren toegepast worden. In dit onderzoek gaat het om waarde, die gecreëerd kunnen worden, voor het verhogen van verkoopvolumes. De eerste deelvraag is.

> *"Hoe kan Machine learning toegepast worden met betrekking tot marketing?"*

6.1 Marketing

Zoals in hoofdstuk 1 beschreven staat is de hoeveelheid aan data exponentieel aan het groeien. Dit wordt bevestigd worden door het boek (The Complete Guide to B2B Marketing, 2015, p. 33). Het boek geeft namelijk aan dat gebruikers van een systeem of dienst data creëren omtrent persoonlijkheid en activiteiten. Deze gegevens worden in het geval van Avanade beschikbaar gemaakt ter uitvoering van analyses. Deze analyses worden uitgevoerd voor bijvoorbeeld het verbeteren van de klantervaring en klanttevredenheid. In dit onderzoek wordt er geconcentreerd op het creëren van waarde voor het verhogen van onder andere het verkoopvolume. Waardes kunnen ontdekt worden op verschillende manieren. Dit kan in sommige situaties op basis van persoonlijke informatie, waarbij er gecombineerd kan worden met andere variabelen. Onder persoonlijke informatie wordt het volgende verstaan: naam, functie, maar bijvoorbeeld ook leeftijd.

Het boek geeft tevens aan dat naast persoonlijke informatie ook demografische en firmograpische informatie kan worden geproduceerd. In het artikel van (Limborgh, 2016) wordt firmographic omschreven als traditionele segmentatie-variabelen, maar dan voor firma's(bedrijven) i.p.v. mensen. Tevens is het ook mogelijk om op basis van activiteiten van gebruikers gedragsdata te verzamelen. Dit gebeurt door middel van on-site activiteiten, offsite activiteiten en campagne activiteiten. Deze verscheidene voorbeelden van data kunnen gebruikt worden om de relatie, persoonlijke communicatie en digitale ervaring met de potentiele klant en het bedrijf te verbeteren. In het boek (King, 2015, p. 44) worden gerichte en gepersonaliseerde inspanningen vertaald naar een hoger consumptieratio en betere merkloyaliteit en het meest belangrijke is een hogere conversieratio.

Het effectief uitvoeren van personalisatie en targeting kan op lange termijn klant loyaliteit en een hogere customer lifetime waarde opleveren. In het boek (King, 2015, p. 60) worden tools omschreven die i.v.m. de opkomende technologie en data, bruikbaar zijn en in de toekomst meer waarde kunnen creëren. De tools zijn: analytics, experiment en optimalisatie, marketing automatisering, targeting en personalisatie. Data kan gebruikt worden voor het verbeteren van marketing prestaties, verhogen van klant behoud, verbeteren van conversie ratio's en het aanbieden van een betere ervaring voor de desbetreffende klanten. Analytics is volgens het boek big business en de verwachtingen zijn, dat de markt zich in de komende vijf jaar zal gaan verdubbelen. In 2014 bedraagt de markt één miljard dollar en de verwachtingen zijn dat het meer dan drie miljard zal bedragen. Analytics is onderverdeeld in vier gebieden, namelijk: web analytics, marketing analytic, customer analytics en predictive analytics (King, 2015, p. 64). Deze gebieden van analytics richten zich op o.a. de decision forming. Hierbij worden keuzes gevormd a.d.v. verkregen inzichten en kennis. Dit zorgt bijvoorbeeld voor betere beslissingen, dat o.a. een bijdrage levert voor het voorspellen van variabelen. In het stuk van (Singh, 2015) wordt aangegeven dat Analytics binnen drie jaar een nieuwe generatie van

oplossingen zal gaan realiseren. Deze nieuwe generatie wordt omschreven als predictive analytics.

6.2 Marketing methodieken en strategieën

In dit onderzoek is het van belang om te kijken naar mogelijkheden voor het verrijken van klant profielen voor marketingcampagnes. Een mogelijke toepassing hierbij is personalisatie. De mogelijkheden tot personalisatie zijn afhankelijk van de beschikbare data. Daarnaast is het ook belangrijk om veel soorten informatie van de desbetreffende persoon te verzamelen. De hoeveelheid heeft ook effect op de prestatie en accuraatheid. Uiteraard hangt dat van het machine learning algoritme af, maar in het algemeen zal meer informatie leiden naar een betere prestatie en daarbij accuraatheid. Voordat personalisatie toegepast kan worden geeft het boek (King, 2015, p. 180) aan dat in eerste instantie segmentatie gerealiseerd moet zijn, voordat targeting en personalisatie tot stand kunnen komen.

Figuur 11. Stappen voor personalisatie

Informatie die verzameld kan worden ten behoeve van het segmenteren van klanten zijn: Naam, functie, postcode, provincie en geografische informatie. De essentie is dat de groepen vergelijkbare karakteristieke bevatten. In het boek (King, 2015, p. 183) wordt er een aantal strategieën voor segmentatie uitgelicht. De volgende opsommingen van strategieën zijn een selectie van strategieën, die toepasbaar zijn in de verkregen datasets:

1. ***Job titel***
   Het groeperen van personen op o.a. gebruikers, kopers etc., zodat diegene alleen informatie ontvangt die nuttig is. Uiteraard zal dit in de verkregen datasets gericht zijn op kopers alleen. De dataset bevat namelijk alleen afnames van producten door klanten. Dit betekent dat het gaat om kopers.

2. ***Functioneel gebied***
   Het segmenteren van de groepen van personen op basis van hen niveau zoals bijvoorbeeld educatie.

3. ***Product interesse***
   Het segmenteren van personen op basis van product aankopen.

4. ***Geografische locatie***
   Het segmenteren van personen op basis van verkoop gebied, postcode etc.

In het onderzoek van (Raquel Florez-Lopez, 2008, p. 97) wordt aangegeven dat voor customer segmentatie een combinatie van cluster- en discriminantanalyse oftewel logistische regressie

wordt toegepast. Een interessant alternatief voor klant segmentatie is het decision tree algoritme. De benoemde algoritmes worden in hoofdstuk 6 toegelicht ter verduidelijking. Daarnaast wordt in het onderzoek van (Raquel Florez-Lopez, 2008, p. 100) ook aangegeven, dat als segmentatie zonder enige relatie met het bedrijf toegepast wordt op marketingpolitiek voor huishoudens, dat de keuzes dan gebaseerd zijn op basis van de relatie tussen de onafhankelijke attributen en de reacties van de mailing test. Dit kan geanalyseerd worden met een extreme vorm van apriori predictive segmentatie. Dit algoritme wordt in paragraaf 6.1.4 toegelicht. In deze situatie zijn er twee types (koper en geen koper) en het aantal segmenten, waarbij de twee worden gedefinieerd als apriori en een set van onafhankelijke variabelen voor het voorspellen van cluster lidmaatschappen. Dit kan uiteraard vertaald worden naar de informatie, die beschikbaar is binnen de aangeleverde databronnen van Avanade. De concrete marketing toepassingen worden op basis van literatuuronderzoek en het analyseren van de huidige structuur in paragraaf 3.5 en in paragraaf 3.8 beschreven.

6.3 Klant segmentatie

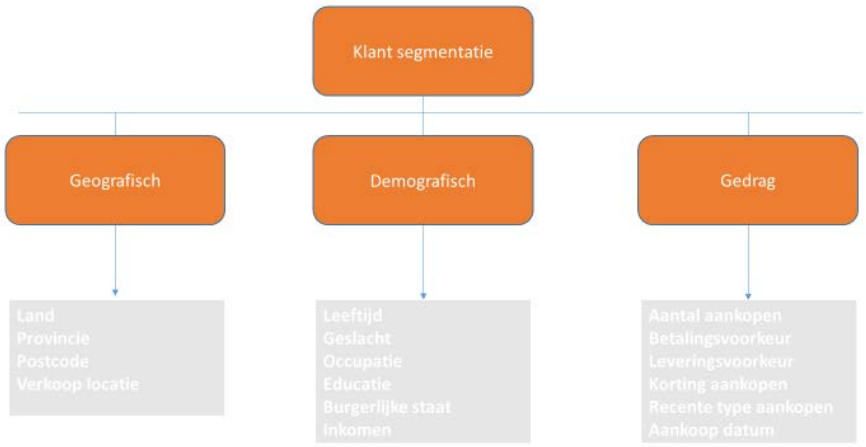
In de white paper van (Synchrony Financial, 2016, p. 2) wordt verteld dat bedrijven via segmentatie inzichten kunnen verzamelen ten behoeve van het vaststellen van marketingstrategieën en het verhogen van klant loyaliteit. Klant segmentatie is volgens (Synchrony Financial, 2016) een tool die mogelijkheden biedt voor marketeers om aanpassingen in de inspanningen uit te voeren op basis van het gedrag van klanten. Daarnaast is klantsegmentatie nuttig voor het gericht versturen van aanbiedingen of services. De voordelen van klant segmentatie is het kunnen realiseren van 20% omzet groei. (Wieland, 2014). In het onderzoek van (Raquel Florez-Lopez, 2008) wordt uitgelegd dat segmentatie in twee categorieën kan worden onderverdeeld. De eerste is segmentatie als een strategie, dat gerelateerd is aan targeting van producten voor een selectie van bepaalde klanten. Tweede categorie is segmentatie als een methodologie, die gerelateerd is aan een bepaalde techniek en methode.

Op basis van de zojuist twee benoemde bronnen kan klant segmentatie omschreven worden als het categoriseren van specifieke personen/klanten, die kenmerken van elkaar vertonen. Dit zorgt ervoor dat segmentatie strategieën opgezet worden om vervolgens aanpassing in de inspanning omtrent marketing te verwezenlijken. Klant segmentatie kan gebruikt worden voor verschillende doeleinden. Het is mogelijk om klanten onder te verdelen in het aantal verkopen, seizoen aankopen en korting en volle prijs aankopen van klanten. Door het gebruik van machine learning kan het toekomstige gedrag van klanten voorspeld worden. Daarom is het volgens (Synchrony Financial, 2016) belangrijk om te kijken naar informatie dat het gedrag van klanten weergeven. Hierdoor kunnen klanten die veel opleveren, maar de laatste periode weinig producten hebben afgenomen, resulteren in een klant dat van leverancier veranderd. Met deze informatie kan een proactieve strategie worden toegepast om de klanten te behouden. Het gaat niet om het voorspellen van de verkopen, maar het begrijpen wat klanten doen en waar het gedrag van klanten op gebaseerd is.
Het modelleren en segmenteren zorgen ervoor dat marketeers het budget kunnen optimaliseren en een hogere return on investment (ROI) kunnen realiseren (Synchrony

Financial, 2016, p. 5). Zodra segmentatie is toegepast kunnen marketingstrategieën gebruikt worden. Voorbeelden hiervan zijn:

1. Differentiatie in aanbiedingen
2. Proactieve retentie
3. Kanaal strategie
4. Customer service

Om de genoemde marketingstrategieën toe te passen is in paragraaf 4.5 de huidige structuur van de desbetreffende datasets omschreven, waarbij in paragraaf 4.6 de wenselijke structuur van de datasets omschreven staan. Klant segmentatie zal op de volgende onderdelen gecategoriseerd worden: Geografisch, demografisch en op basis van gedrag. De attributen die hierbij horen zijn afgebeeld in figuur 12.



Figuur 12. Klant segmentatie gebieden

6.4 RFM

In het boek (King, 2015, p. 186) wordt ook aangegeven dat de informatie voor segmentatie, aangevuld kan worden met het RFM-marketingmodel. RFM staat voor recency, frequency en monetary. Met RFM is het mogelijk om de waarde van de klanten te berekenen (Andale, 2015). Correcte segmentatie kan door RFM, oftewel recency, frequency en monetary, zoals eerder vermeld, ondersteund worden (King, 2015, p. 186).

Recency

In dit onderdeel wordt er gekeken naar de recentelijke aankopen van de klant. Volgens (Andale, 2015) zijn kopers, die recentelijk aankopen doen eerder geneigd om weer een aankoop te doen, dan klanten die al voor een lange periode nog geen aankopen hebben gedaan.

Frequency
In dit onderdeel wordt er gekeken naar hoe vaak een klant aankopen heeft gedaan. Klanten die bijvoorbeeld wekelijks aankopen doen, zijn eerder geneigd een nieuwe aankoop te doen, dan klanten die jaarlijks aankopen doen en dus minder aankopen doen dan frequente klanten.

Monetary
In dit onderdeel wordt er gekeken naar de hoeveelheid geld die een klant uitgeeft. Klanten die hoge uitgaves doen, zijn eerder geneigd om weer een aankoop te verrichten. Tevens verrichten de klanten duurdere aankopen.

Om deze onderdelen te berekenen is een aantal data nodig. Daarnaast is het belang van de benoemde punten gesorteerd op prioriteit (Mutyala, 2011):

1. Recentelijke aankoopdatum
2. Aantal aankopen binnen een bepaalde periode
3. Totale aankopen per klant. (Berekening van het gemiddelde is ook mogelijk)

De benoemde informatie kan verkregen worden op basis van de verkregen datasets. In dit geval gaat het om de transactie dataset, dat in paragraaf 4.5 wordt verduidelijkt. Overigens is het belangrijk om klanten, die op basis van het RFM-model minder waarde hebben, niet te verwaarlozen. Om dit model toe te kunnen passen in de verkregen datasets is het belangrijk om transformaties te realiseren. Dit zal in de volgende paragraaf behandeld worden in de vorm van een GAP-Analyse.

6.5 Huidige structuur van de datasets

De verkregen databronnen bestaan zoals eerder beschreven uit: Customer base date van Customer Relationship Management (CRM), Sales data van point of sales (PoS) en producten data van datawarehouse (DWH).

De onderstaande tabellen (1,2,3) illustreren de kolommen met de attributen, die de kolommen bevatten. De weergegeven kolommen zijn een selectie uit bruikbare informatie uit de desbetreffende databronnen. Tevens zijn de duplicatie kolommen niet meegenomen

iD	Date of Birth	Gender	State-province	Country	Postalcode
NUM	NUM	CHAR	CHAR	CHAR	NUM

Tabel 1. Customer base data

ProductiD	Price	ProductStatus	StoreiD
NUM	NUM	CHAR	NUM

Tabel 2. Sales data

CustomeriD	TransactioniD	Quantity	ProductiD	Discount	Total Price	Payment Method	DateTime	Sale location	StoreiD
NUM	NUM	NUM	NUM	NUM	NUM	CHAR	NUM	NUM	NUM

Tabel 3. Transactie data

6.6 Wenselijke structuur van de datasets

Op basis van de segment strategieën benoemd in paragraaf 4.2, zijn er een aantal extra attributen nodig om een beter klant segment te realiseren. De volgende tabellen illustreren de extra attributen, waarbij sommige kolommen zich al in de databronnen bevinden, maar geen informatie bevatten in verband met het anonimiseren van gevoelige informatie.

Occupation	Education	Income	Marital status	Age
CHAR	CHAR	NUM	CHAR	NUM

Tabel 4. Customer base data transformatie

Zoals te zien is, is de kolom leeftijd qua informatie niet afwijkend van de bestaande kolom date of birth. De bestaande kolom met de daarbij horende gegevens moeten vertaald worden naar leeftijd, met numerieke waarden tussen de 1 en 100. De verandering zal de bestaande date of birth kolom niet vervangen. De transformatie, die plaats zal vinden voor de verandering van de kolom zal in het uitwerkingsdocument van machine learning toegelicht worden.

Order Total	Order date	Product Categorie
NUM	CHAR	CHAR

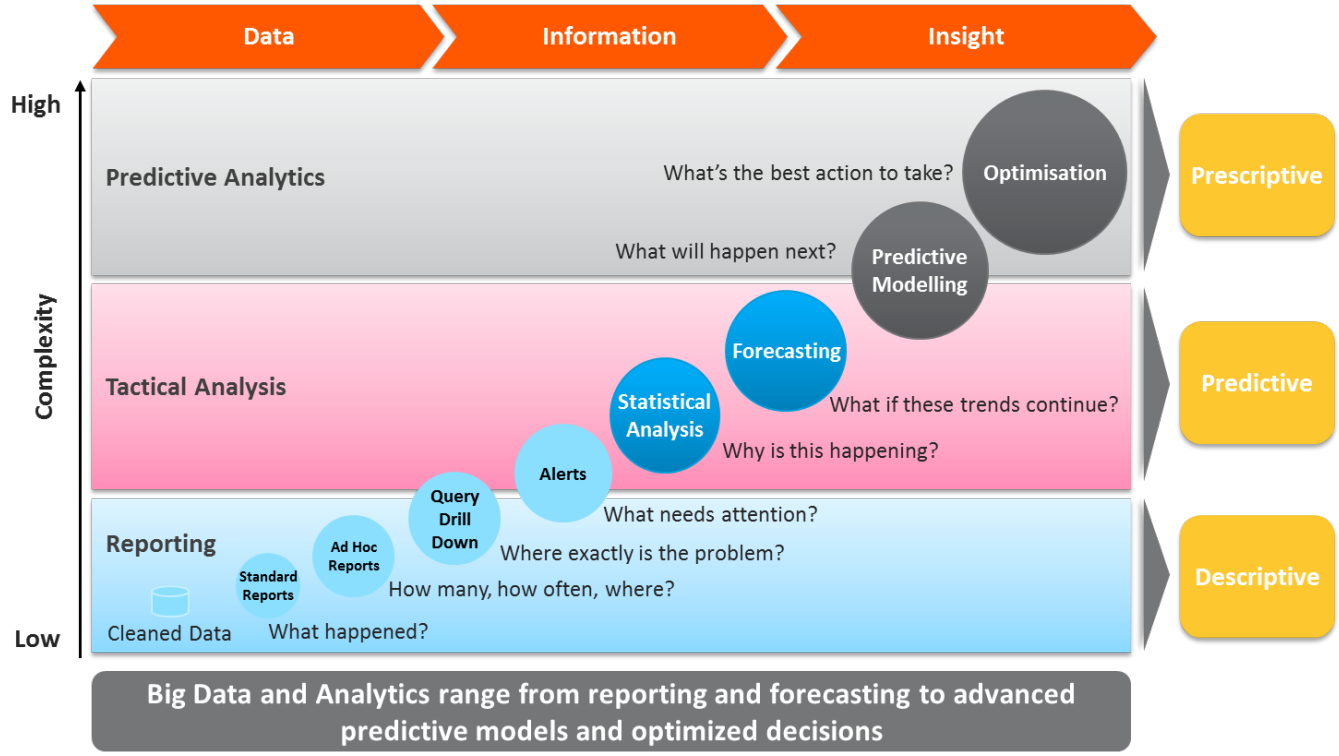
Tabel 5. Product transactie data transformatie

Door het toevoegen van een nieuwe kolom met de naam order total kan er gekeken worden naar de totale aankopen van een klant. Dit is een attribuut, dat belangrijk is bij het segmenteren van klanten, die veel of weinig aankopen doen en daardoor meer opleveren voor de supermarktketen in kwestie. Tevens is order date een bestaande kolom, dat aangepast moet worden, zodat de date op de gregoriaanse kalender wijze toegepast is. Dat is namelijk in de huidige situatie niet het geval. In de kolom van product id zijn er naast numerieke waardes ook karakter waardes. Hierbij zal de vraag zijn welke machine learning modellen met verschillen in datatype om kunnen gaan. Dit zal verder in hoofdstuk 6 behandeld worden.

6.7 Type Machine learning analyse

Het analyseren van machine learning resultaten en het verwerken van de gegevens vindt plaats door verschillende vragen. In onderstaand figuur 13, dat afkomstig is uit een deelgenomen machine learning training, is te zien, dat het niveau van complexiteit afhankelijk is van de type analyse. De onderverdeling met de voor opgestelde vragen zijn als volgt:

- Wat is er gebeurd? – Descriptive analyse
- Wat zou er kunnen gebeuren? – Predictive analyse
- Wat zou er moeten gebeuren? – Prescriptive analyse



Figuur 13. Machine learning analyse types

Voor de geselecteerde RFM-marketingtechniek is het segmenteren van klanten, de juiste aanpak. De analyse type is in dit geval descriptive. Er zal voornamelijk gekeken worden wat er gebeurd is in plaats van te voorspellen wat er gedaan kan worden. Volgens (Etaati, 2016) zorgt descriptive analyse ervoor om analyses uit te voeren voor het vinden van trends, gedrag, en structuur van de data. Een van de belangrijkste benadering voor descriptive analytics is het classificeren van data in verschillende clusters met gebruik van k-means. Cluster algoritmes kunnen gebruikt worden voor vragen zoals bijvoorbeeld: Welke klanten hebben dezelfde voorkeuren.

Tevens nemen deze types een informerende en waarschuwende rol (Avanade Analytics Training, 2017). In paragraaf 4.1 is aangegeven dat de combinatie van cluster en logistische regressie een oplossing is voor klant segmentatie. Hierbij wordt er deels predictive analyse toegepast. De vraag is uiteraard in hoeverre dit de oplossing is en wat daarvan de toegevoegde waarde van is. Dit zal in het realisatie document toegelicht worden en in een later hoofdstuk in dit onderzoek worden.

6.8 Conclusie

In deze paragraaf wordt de conclusie evenals marketing toepassingen en strategieën beschreven in combinatie met voorbeelden ter verduidelijking van de eerste deelvraag, namelijk: Hoe kan marketing toegepast worden met betrekking tot machine learning?

Machine Learning gebruikt data als input voor verscheidene doelen. Gegevens die zich in de datasets bevinden richten zich op de demografische, geografische en gedragsgebieden. Deze gegevens kunnen gebruikt worden ten behoeve van het segmenteren van klanten. Doordat klanten gecategoriseerd worden is het verhogen van het verkoopvolume realiseerbaar d.m.v. marketingcampagnes. Dit kunnen direct mailing en gepersonaliseerde aanbiedingen zijn. Het doel is om op korte of lange termijn de consumptie, conversie ratio en loyaliteit te verhogen. Om personalisatie toe te passen is het belangrijk om eerst segmentatie en vervolgens targeting uit te voeren in de roadmap naar marketing personalisatie.
Klant segmentatie is met betrekking tot marketing een toepasbare strategie met de verkregen data. Door klant segmentatie kunnen marketeers en de grote supermarktketen, aanpassingen in de activiteiten van campagnes verrichten. Segmentatie van klanten kan op verschillende strategieën gebaseerd zijn. Segmentatie strategieën die in combinatie gebruikt kunnen worden zijn: job titel, functioneel gebied, product interesse en geografische locatie. Om een volledige klant segmentatie te realiseren, waarbij klanten op basis van verschillende attributen worden gecategoriseerd, moeten in eerste instantie de lege kolommen: Occupatie, inkomen, burgerlijke staat en betalingsvoorkeur door middel van gerandomiseerde waardes ingevuld worden. Kolommen date of birth van customer data en date time van transactie data moeten getransformeerd worden naar geschikte waardes. Tevens is het aanbevolen om nieuwe kolommen toe te voegen, namelijk: totale order per klant, totale opbrengst per klant, leeftijd en productcategorie.
De klant segmentatie richt zich in dit geval zoals aangegeven op de geografische, demografische en gedragsgebieden in de verkregen datasets. Door het toevoegen van deze extra kolommen zal de segmentatie op meerdere karakteristieken en kenmerken gebaseerd zijn en daardoor kunnen er gerichter campagnes uitgevoerd worden. Voor klant segmentatie zal er tevens ook gebruik gemaakt worden van het RFM Marketingmodel. De onderdelen recency, frequency en monetary zorgen er namelijk voor dat de waarde van de klant berekend kan worden. Deze waardes zullen uiteindelijk meegenomen worden bij het segmenteren van klanten. Het is voornamelijk de bedoeling om uit de bestaande datasets waardes te creëren en vervolgens daaruit verdere ontwikkelingen te realiseren ten behoeve van het opleveren van extra waardevolle informaties.
Op het moment dat de transformatie, toevoegingen en aanpassingen in de databronnen gerealiseerd zijn, kan klant segmentatie toegepast worden. Op het moment dat klant segmentatie is toegepast, kan de supermarktketen verschillende marketingstrategieën toepassen, namelijk: differentiatie in aanbiedingen, proactieve retentie, kanaal strategie en customer service.

De marketing toepassingen zijn tot stand gekomen op basis van informatie uit literatuur, die in de vorige paragrafen beschreven zijn, maar ook door de huidige structuur te analyseren en op basis daarvan de gewenste structuur van de datasets te benoemen. De volgende punten zijn activiteiten die naast het transformeren, aanpassen en toevoegen van data gerealiseerd zullen worden omwille van het creëren van waarde die nuttig zijn voor marketingcampagnes:

- Voorspelling betalingsvoorkeur
- Voorspelling leveringsvoorkeur
- Klant segmentatie

Het is van essentieel belang om de toegevoegde waarde van deze waarde creatie toe te lichten. Deze dienen als verrijking van klantprofielen als input voor marketingcampagnes. Doordat de achterliggende werkprocessen van de grote supermarkt niet bekend zijn, is de waarde creatie gebaseerd op creatieve mogelijkheden.
De toegevoegde waarde van het voorspellen van de betalingsvoorkeur is, dat er een beter klantervaring gerealiseerd kan worden bij het afrekenen in de supermarkt of webshop. Zo kan de betaling pop-up op basis van de voorspelling weergeven worden. Hierdoor duurt een transactie korter en kunnen de klanten sneller geholpen worden. Omdat dit onderzoek zich richt op marketingcampagnes kan er gekozen worden om klanten met bijvoorbeeld creditcardbetalingen korting te geven.
Het is belangrijk om erbij te vermelden dat het voorspellen van de leveringsvoorkeur als waarde ook het verbeteren van klantervaring kent. Klanten kunnen naast betere klantervaring ook gerichte aanbiedingen ontvangen. Hierbij kan gedacht worden aan het aanbieden van kortingen bij het ophalen of laten versturen van producten. Als de voorspelling van leveringsvoorkeur van een product ophalen is en een bepaald product bijna is uitverkocht, dat de klant dan een bericht krijgt met een kortingscode voor bij het ophalen van het product. Uiteraard kan dit ook toegepast worden op producten die uit gefaseerd worden, waardoor bijvoorbeeld de kosten voor voorraadbeheer omlaag kan.
Voor klant segmentatie kunnen klanten gesegmenteerd worden op overeenkomende kenmerken, waardoor marketingcampagnes gericht kunnen worden uitgevoerd. Niet iedere klant is namelijk winstgevend. De bedoeling van het segmenteren van klanten is het realiseren van een hogere omzet, door effectieve marketingcampagnes uit te voeren voor de desbetreffende segmenten. Zoals eerder aangegeven zorgt het correct segmenteren van klanten voor een omzetgroei van 20%.
Tevens zal descriptive als prescriptive machine learning toegepast worden. Dit betekend dat er gekeken zal worden wat er gebeurd is en wat er voorspelt kan worden. Dit is op basis van het onderzoek dat verricht is door (Raquel Florez-Lopez, 2008), waarbij wordt aangegeven dat een combinatie tussen cluster en logistische regressie een oplossing is voor klant segmentatie.

7 Machine Learning technieken

Machine learning modellen werken op verschillende manieren. De methode is gebaseerd op de learning technieken. In dit hoofdstuk wordt er getracht duidelijkheid te vormen omtrent learning technieken, die toepasbaar zijn binnen de verkregen databronnen. Uiteindelijk zal dit uitmonden in een conclusie, die de vraag beantwoord omtrent technieken, die van toepassing zullen. De tweede deelvraag is:

"Welke learning technieken zijn op basis van de beschikbare databronnen toepasbaar"

7.1 Supervised learning

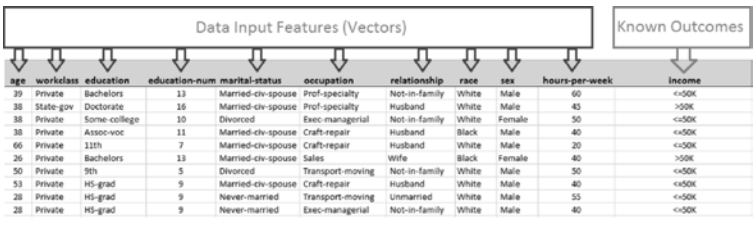
Supervised learning wordt gezien als een onderliggende techniek m.b.t. machine learning, die "bekende" datasets gebruikt om vervolgens daarmee een datamodel te creëren ten behoeve van het maken van voorspellingen. Op basis van de desbetreffende trainingssets worden er pogingen vanuit het algoritme gemaakt om een nieuw model te bouwen voor het maken van voorspellingen, dat gebaseerd is op de nieuwe input waarden gecombineerd met de bekende verwachtingsresultaten (Barnes, 2015, p. 28). Tevens wordt de supervised learning techniek vaker gebruikt in tegenstelling tot de andere machine learning technieken, namelijk 70% van de gebruikte machine learning type algoritme (Toolbox, 2016). Supervised learning kan zich onderscheiden in twee categorieën van algoritmes:

1. Classificatie
2. Regressie

Figuur 14. Formule supervised learning

De essentie van supervised learning is, dat de methode gebaseerd is op het labelen van de input data en verwachtingsresultaten. De eisen voor het toepassen van supervised learning is dat er een training dataset is, waarvan de input kolommen minimaal één van de volgende twee mogen bevatten (Barnes, 2015, p. 29):

1. **Features/Vectoren** –Data, waarvan de kolommen worden gebruikt voor het realiseren van voorspellingen.
2. **Labels/Supervisory signal** – Dit vertegenwoordigt het verwachtingsresultaat oftewel dat wat er voorspeld moet worden.



age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	hours-per-week	income
39	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Not-in-family	White	Male	60	<=50K
38	State-gov	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	45	>50K
38	Private	Some-college	10	Divorced	Exec-managerial	Not-in-family	White	Female	50	<=50K
38	Private	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Black	Male	40	<=50K
66	Private	11th	7	Married-civ-spouse	Craft-repair	Husband	White	Male	20	<=50K
26	Private	Bachelors	13	Married-civ-spouse	Sales	Wife	Black	Female	40	>50K
50	Private	9th	5	Divorced	Transport-moving	Not-in-family	White	Male	50	<=50K
53	Private	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	40	<=50K
28	Private	HS-grad	9	Never-married	Transport-moving	Unmarried	White	Male	55	<=50K
28	Private	HS-grad	9	Never-married	Exec-managerial	Not-in-family	White	Male	40	<=50K

Figuur 15. Voorbeeld dataset input en output

7.2 Unsupervised learning

In het geval van unsupervised learning is het ingewikkelder om bepaalde gegevens te voorspellen. Het algoritme krijgt geen bekende input data of verwachtingsresultaten voor het bouwen van een predictive model. Bij unsupervised learning hangt het af van de bekwaamheid van het algoritme in het ontdekken van patronen, structuren en relaties in de dataset. De essentie is dat er wordt gekeken naar vergelijkbare objecten binnen de data, waarmee het zich kan associëren. (Barnes, 2015, p. 33)

Volgens het boek (EMC Education Services, 2015, p. 118) wordt unsupervised learning gerefereerd naar het probleem van het vinden van verborgen structuren in non-labeled data. Clustering technieken behoren tot de unsupervised learning algoritme, waarbij de data scientist bij voorbaat niet de labels bepaald voor het toepassen van de clusters. Bij het gebruiken van unsupervised learning, kan er gekeken worden naar twee verschillende benaderingen (Barnes, 2015, p. 34).

- Cluster analyses: Dit wordt gebruikt voor het vinden van verborgen patronen of groepen in de datasets. Voorbeelden van cluster analyses zijn:

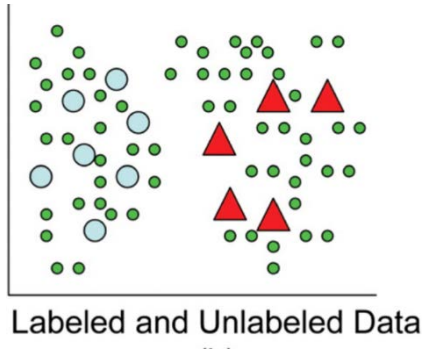
    - Social network graphs: groepen mensen die gerelateerd zijn aan jou op basis van familie, vrienden, werk of school.
    - Aankooppatronen: Hierbij speelt prijsklasse, intensiviteit van gebruik, keuze voor retail company, koper of geen koper en intensiviteit van aankopen een rol.

Figuur 17. Clustering van data objecten voorbeeld

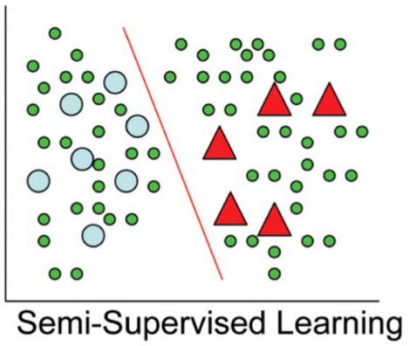
Figuur 16. Clustering proces voorbeeld

7.3 Semi-Supervised learning

Semi supervised learning is een combinatie van supervised en unsupervised learning. Volgens een artikel geschreven door Xiaojin Zhu (Zhu, 2007, p. 4) is het verkrijgen van labeled data voornamelijk moeilijk, prijzig en tijdrovend. Dit komt door bijvoorbeeld inspanning, maar ook door de nodige ervaring. Unlabeled data is daarentegen volgens het artikel sneller te verzamelen, maar zijn er nauwelijks manieren om er gebruik van te maken. Semi-supervised learning lost dit probleem op door een grote hoeveelheid van unlabeled als labeled data te gebruiken ten behoeve van het verbeteren en bouwen van classificaties. Semi-supervised learning vereist minder inspanning en geeft een hogere accuraatheid (Zhu, 2007, p. 4). In de huidige situatie is de dataset verkregen en vallen de nadelige opsommingen af. Volgens het boek Semi-Supervised learning (MITPress, 2006, p. 4) is het in sommige scenario's verstandiger om semi-supervised learning te gebruiken. Bij het gebruik van labeled data en toevoeging van unlabeled data voor een hogere accurariteit, moet het classificatie probleem van beide data gerelateerd zijn aan elkaar. De informatiewaarde van unlabaled data moet een verlengstuk kunnen zijn voor de waarde uit de labaled data. Als dit niet het geval is, dan is de toegevoegde waarde van semi-supervised learning nihil en kan er beter gebruik gemaakt worden van het supervised learning algoritme



Figuur 19. Labeled en unlabeled voorbeeld



Figuur 18. Semi-supervised voorbeeld

7.4 Reinforcement learning

Volgens onderzoek (Ghahramani, 2004, p. 3) is reinforcement learning de interactie van de machine met de omgeving door het produceren van acties. De machine ontvangt de input en genereert vervolgens acties. De acties hebben effect op de staat van de omgeving, waarvoor het resultaat voor de machine een beloning of afstraffing geldt. De essentie van het reinforcement learning algoritme is dat het leert om op de juiste manier te handelen op basis van het maximaliseren van de beloningen of door het minimaliseren van de afstraffingen. In het boek (Richard S. Sutton, 2012, p. 4) wordt reinforcement learning omschreven als een algoritme, dat leert om bepaalde handelingen uit te voeren in bepaalde situaties. De machine wordt dus niet aangeleerd wat er uitgevoerd moet worden, maar leert op basis van acties, die beloningen opleveren. De twee kenmerken van reinforcement learning zijn:

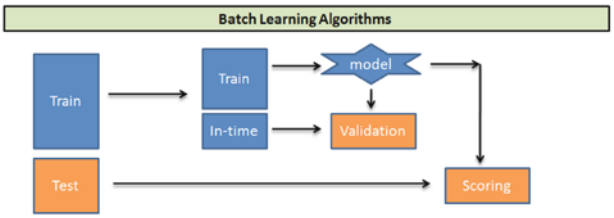
- Trial and error
- Delayed reward

Voorbeeld met betrekking tot reinforcement learning (Richard S. Sutton, 2012, p. 6):

- Een robot moet een keuze maken met betrekking tot het binnentreden van een kamer voor het zoeken en verzamelen van afval of terug te keren naar zijn oplaadpunt. De beslissing wordt genomen op basis van hoe snel en gemakkelijk de oplaadpunt in het verleden gevonden is.

7.5 On-line en offline learning

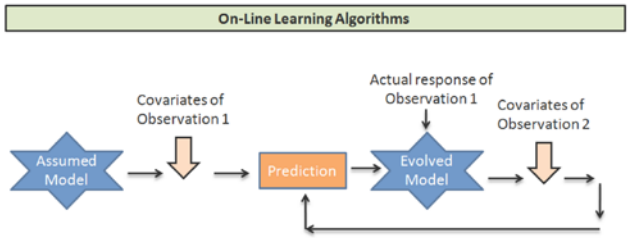
Online en offline learning technieken staan los van de voorgaande beschreven technieken en kunnen in samenhang gebruikt worden. De termen online en offline learning technieken hebben geen overeenkomsten met betrekking tot de netwerkverbinding definities. Batch learning oftewel offline learning maakt het mogelijk om de training data voor het trainen in losse stukken te gebruiken. In dit proces wordt vervolgens op basis van de gevonden relaties in de training data (Srivastava, 2015) de test sample voorspeld. Dit kan bevestigd worden door (University of Regina, 2012). In de aangegeven bron is het proces van batch learning opgedeeld in fasen. De training fase wordt namelijk als eerst uitgevoerd en vervolgens daarna de testing fase. Figuur 19 illustreert de werking van batch learning.



Figuur 20. Batch learning techniek

In het geval van online learning wordt de verwerking niet in fasen opgedeeld en wordt het leren tijdens de verwerking van data uitgevoerd. In (University of Regina, 2012) wordt aangegeven dat er een keuze gemaakt moet worden omtrent acties. De opties zijn

1. Het model direct goed laten presteren
2. Het model op lange termijn kennis laten ontwikkelen om het vervolgens op een later stadium beter te laten presteren.



Figuur 21. On-line learning techniek

Bij het kiezen tussen de twee learning technieken zijn er een aantal afwegingen, die in acht genomen moeten worden.

1. *Snellere berekeningen en efficiëntere opslag van on-line*
   Doordat on-line learning geen verschillende fases in het proces bevat, worden de gegevens sneller verwerkt. De data wordt niet opgeslagen en gebruikt voor weergave tijdens het leer proces, waardoor de opslag ook efficiënter is.

2. *Normaliter makkelijker te implementeren*
   De data wordt bij on-line learning niet in fasen opgedeeld en iedere keer wordt er een sample verwerkt. De data is afkomstig vanuit een stream, waardoor het algoritme simpel toegepast kan worden.

3. *Lastiger te onderhouden tijdens uitvoering*
   On-line learning verwacht, indien die aanwezig zijn, constant zelfde data punten en vectoren. Zodra een bepaald waarde uit de databron wijzigt, veranderd de output daarmee, waardoor de mogelijk is output geen waardevolle informatie meer bevat.

4. *Lastiger om online te evalueren*
   Bij het toepassen van on-line learning kan de test set niet tijdelijk vast gehouden worden voor evaluatie.

5. *Normaliter lastiger om het correct toe te passen*
   Doordat het leren geautomatiseerd wordt tijdens het verwerken van data bij on-line learning, is het lastig om het algoritme consistent te houden met betrekking tot performance en accuraatheid. Het is daardoor dus moeilijk om een diagnose uit te voeren.

7.6 Conclusie

In hoofdstuk vier zijn de mogelijkheden omtrent het creëren van waardes op basis van de huidige dataset en de waardes die tot stand kunnen komen door het transformeren en toevoegen van data en kolommen besproken. De mogelijkheden zijn variërend. Omdat de uitkomst (output) in sommige gevallen al bekend is, is er gekozen voor de supervised learning techniek. Voor het segmenteren van klanten is er gekozen voor de unsupervised learning techniek. Tevens is het de bedoeling om bij segmentatie van klanten een cluster en classificatie model toe te passen. Hierdoor zal er gekeken worden naar unsupervised als supervised algoritmes binnen Azure. De zojuist benoemde combinatie verschilt van de beschreven Semi-supervised learning. Bij semi-supervised learning is het belangrijk dat zowel de labeled als unlabeled data een verlengstuk van elkaar zijn. Bij het segmenteren van klanten worden verschillende kenmerken met elkaar geclusterd om klanten met dezelfde karakteristieken en kenmerken te categoriseren. De attributen zijn verschillend van elkaar en daarom zal semi-supervised learning geen oplossing bieden. Tevens is reinforcement learning niet toepasbaar voor het realiseren van de benoemde waardes. Reinforcement learning valt voornamelijk in de categorie van robotica, waarbij het systeem vaak te maken heeft met een omgeving waarvoor beloningen en afstraffingen gelden.

Omdat de databronnen van de grote supermarktketen in kwestie geëvalueerd moeten worden voor het verwezenlijken van een data lake, is het dus van essentieel belang om een stream aan data te hebben. De definitie van big data en een data lake is in paragraaf 3.2 beschreven. De verkregen databronnen zijn type gegevens, die voldoen aan de criteria van voor grote hoeveelheid data. Dit is ook terug te vinden in de omschrijving, waarbij geconstateerd wordt dat het datawarehouse gelimiteerd wordt. Om deze reden is het toepassen van on-line learning de juiste keuze. Uiteraard zal de definitie van big data eerder gelden voor data, die afkomstig zijn van bijvoorbeeld auto's, waarbij data gecreëerd wordt door sensoren. Hierdoor zullen de verkregen datasets in zijn geheel niet onder big data vallen, maar met de mogelijkheid van een data lake, is de keuze voor on-line learning gunstig. Echter, is het met Azure ML niet mogelijk om on-line learning toe te passen. Daarnaast is het niet mogelijk om meer informatie te verschaffen omtrent de situatie van de grote supermarktketen. Indien verschillende attributen veranderen of nieuwe toegevoegd worden, zal on-line learning niet goed toe te passen zijn. On-line learning zou wat een data lake betreft een goede keuze zijn, zolang de attributen op lange termijn niet wijzigen.

8 Machine Learning modellen

In het vorig hoofdstuk was de conclusie om gebruik te maken van de supervised en unsupervised learning technieken. Om deze reden zijn de meest besproken en bekende algoritmes ter verduidelijking gekozen in dit hoofdstuk. In de vorige subhoofdstukken zijn de learning technieken beschreven, die behoren tot specifieke machine learning modellen. Ieder machine learning model is geschikt voor een bepaald scenario. In dit hoofdstuk wordt er getracht om bekende modellen te beschrijven, zodat de toepasbare modellen inzichtelijk worden gemaakt ter evaluatie van de toepasbare modellen ten behoeve van het verrijken van klant profielen. De derde deelvraag is:

"*Welke Machine Learning algoritmes zijn geschikt om waardes uit de datasets nauwkeurig te voorspellen*"

8.1 Unsupervised learning

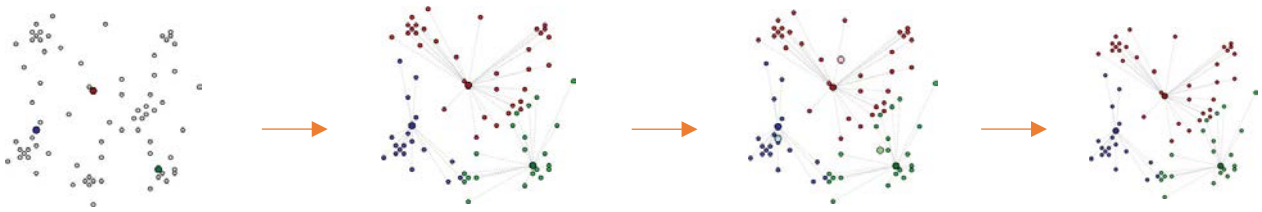
Volgens (EMC Education Services, 2015, p. 118) wordt over het algemeen voor het clusteren van vergelijkbare objecten de unsupervised learning techniek gebruikt. In machine learning wordt de unsupervised learning techniek gerefereerd naar het probleem voor het vinden van verborgen structuren in een unlabeled data omgeving. De gebruiker oftewel de data scientist kan bij het kiezen van clustering technieken geen labels toepassen op de desbetreffende cluster en dataset. Dit is tevens ook de grootste kenmerk voor de unsupervised learning techniek. In de voorgaande hoofdstuk omtrent doelstelling, is het gebruik van Azure Machine Learning vermeld. Voor unsupervised learning is er één mogelijk toepassing van clusteren en dat is K-means. Om deze reden zal voor unsupervised learning alleen k-means behandeld worden.

8.1.1 K-Means

De k-means algoritme kijkt naar de verzameling van data, waarbij er aangegeven wordt hoeveel clusters er gevonden mogen worden. Dit wordt aangegeven door middel van de *k.* De clusters worden gevonden op basis van de aritmetische oftewel wiskundige gemiddelde. In onderstaand figuur is het aantal k=3.
De kleuren geven de verschillende clusters aan en in dit geval zijn blauw, groen en rood de clusters.
In eerste instantie wordt in de verzameling van objecten een centroid geplaatst. Centroid staat voor het gemiddelde van de geclusterde objecten. De objecten behoren bij iedere fase tot de dichtstbijzijnde centroid. In de onderstaande figuren is het aantal clusters aangegeven met k=3. Bij de eerste stap van het clusteren van de objecten wordt de centroid geschat om vervolgens in de navolgende stappen de Euclides afstand op de clusters toe te passen. Het berekenen van de centroid na de eerste stap wordt gedaan door middel van een formule, namelijk: $(x_c, y_c) = \left(\frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m} \right)$.

In dit voorbeeld bestaat de verzameling van objecten uit twee dimensies (respectievelijk x en y). De formule van de Euclides afstand is in dit geval: d= √ (x1-x2) ^2+(y1-y1) ^2. Uiteraard kan er ook gekozen worden voor de cosinus similarity afstand en Manhattan afstand. Cosinus similarity wordt voornamelijk gebruikt voor document frequentie en bij het berekenen van

rechthoekige afstanden kan de manhattan distance toegepast worden. Deze centroid wordt telkens bij ieder fase opnieuw vastgelegd tot dat er geen verandering meer plaatsvindt.



Zodra de dataobjecten van een cluster niet meer wijzigen, kan er gekeken worden naar de elbow-point methode. Hierbij wordt er gekeken naar het percentage van de variantie. De lijn in de grafiek daalt naarmate het aantal clusters wordt gespecificeerd en zodra er geen wijzigingen in de lijn worden geconstateerd, kan het aantal clusters vastgesteld worden. Het punt waarop de lijn geen daling meer laat zien wordt ook wel de elbow-point genoemd.

Volgens (Chinedu Pascal Ezenkwu, 2015) is de commerciële wereld over de jaren heen competitiever geworden. De organisaties moeten voldoen aan de behoeftes van de klanten, maar ook moeten nieuwe klanten aangetrokken worden. De taak van het identificeren en voldoen van de behoeftes van iedere klant is in de business een zeer moeilijke taak. Dit komt doordat de klanten verschillen in behoeftes, geografie, maar ook in smaak en voorkeur. Het is daarom verkeerd om iedere klant gelijk te behandelen. Dit heeft de adoptie van klantsegmentatie gemotiveerd, waarbij klanten worden gesegmenteerd in kleinere groepen. Volgens (Goyat, 2011, p. 45) is klantsegmentatie een strategie om de markt onder te verdelen in homogene groepen. Jerry W. Thomas geeft in "Market Segmentation" het volgende aan: "The purpose of segmentation is the concentration of marketing energy and force on the subdivision (or the market segment) to gain a competitive advantage within the segment" (W.Thomas, 2016, p. 1). K-means kan gebruikt worden voor het toepassen van klant en marktsegmentatie. Volgens (Goyat, 2011) zijn de volgende strategieën bruikbaar voor marktsegmentatie:

1. Geografisch
2. Demografisch
3. Psychologisch
4. Gedrag

Deze strategieën zijn expliciet bedoeld voor marktsegmentatie, maar zijn bruikbare voorbeelden voor toepassingen voor klantsegmentatie. Overigens komen deze strategieën ook overeen met de strategieën benoemd in paragraaf 4.3. Dit geeft aan dat klant segmentatie een mogelijkheid is voor het toepassen met machine learning. Het verwezenlijken van de klant segmentatie zal klant profielen verrijken als input voor marketingcampagnes.
Bij geografische segmentatie van klanten kan er gekeken worden naar regio, klimaat en populatiedichtheid. Als er gekeken wordt naar demografische gegevens dan kan leeftijd, geslacht, inkomen, educatie, nationaliteit, sociale klasse en gezinsgrootte een belangrijk factor spelen bij het segmenteren van de klant. In dit document zullen in een later hoofdstuk de mogelijkheden behandeld worden op basis van de beschikbare dataset van de supermarktketen.

Psychologische gegevens hebben volgens (Goyat, 2011) ook een grote waarde omtrent het segmenteren van klanten zoals bijvoorbeeld: interesses, activiteiten en meningen. Het verzamelen van deze gegevens zullen meer inspanning vereisen, omdat de gegevens van externe partijen afkomstig zullen zijn. Als laatste wordt er vermeld om gedrag informatie te verzamelen voor het segmenteren van de markt. Dit zijn o.a. gegevens over loyaliteit van klanten aan een merk, maar ook de bereidheid van het doen van aankopen. Voor deze gegevens geldt ook het inspanningsniveau zoals bij het verzamelen van psychologische gegevens. Het verzamelen van deze gegevens zal problematisch zijn.
In het artikel Predicting customer loyalty using the internal transactional database (Wouter Buckinx, 2007) wordt aangegeven dat CRM gegroeid is en een grote trend is binnen marketing. De oplossing voor het voorspellen van klant loyaliteit is in eerste instantie gebaseerd op response modeling. De vraag die in dat domein wordt beantwoord is of een klant reageert op speciale aanbiedingen en brochure. Dit is tevens de centrale applicatie binnen dat domein. Cross selling analyses zijn ook nodig, zodat het juiste product aangeraden kan worden aan de klant. Daarnaast is upselling analyse ook van toepassing doordat het gericht is op het verkopen van meer producten. Beide technieken zorgen ervoor dat de relatie van de klant wordt geïntensiveerd, door het verhogen van het aandeel van de producten, dat gekocht is bij de lokale bv. Dit zorgt ervoor dat de producten niet bij de concurrentie worden aangeschaft. Het verlies van klanten met betrekking tot aankopen, die gedaan worden bij de concurrentie resulteert in een churn analyse.
Volgens (Frederick F. Reichheld, 1990) heeft “customer deflaction” een grote impact op de winst van het bedrijf. In het artikel customer segmentation and strategy development based on customer life time value (Su-Yeon Kim, 2006) zijn live time value (LTV) analyses een veelvuldig gebruikte techniek voor het voorspellen van mogelijke klanten. In het onderzoek zijn multiple lineaire regressie, decision tree en random forest toegepast voor het voorspellen van klant loyaliteit.

De voordelen en nadelen van de k-means algoritme worden door (Manju Kaushik, 2014, p. 2) als volgt geformuleerd:

Voordelen:

1. Simpel algoritme, dat makkelijk te begrijpen en implementeren is.
2. Efficiënt algoritme, waarbij n staat voor het aantal data punten, de k voor het aantal clusters en T voor het aantal iteraties.
3. K en T zijn over het algemeen van kleine aantallen en dat maakt het algoritme lineair.

Nadelen:

1. De gebruiker moet het aantal clusters specificeren.
2. Het algoritme is gevoelig voor uitschieters en ruis. (Dit zijn data punten die ver van elkaar liggen)

Daarnaast wordt in applying unsupervised learning (MathWorks, 2016) aangegeven dat K-Means gebruikt wordt als het aantal clusters bekend is en als snelheid van clusteren een requirement is in het geval van grote datasets.

8.1.1.1 Use cases k-means

Image processing

Zoals in subhoofdstuk 3.2 aangegeven is, is bijvoorbeeld video ongestructureerd data. Voor iedere frame uit de video kan k-means clustering worden toegepast. Het doel bij het toepassen van k-means in deze situatie is dat er gekeken wordt naar de pixels, die vergelijkbaar met elkaar zijn. De attributen die behoren tot de pixel zijn o.a. contract, kleur en locatie. Deze worden opgeslagen in een x en y positie uit de video frame. De veranderingen in bijvoorbeeld veiligheidsopname kan aangeven dat er ongeautoriseerde toegang heeft plaatsvonden.

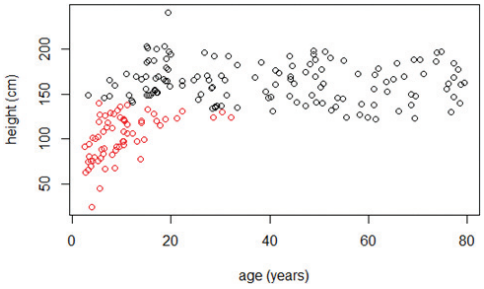
Medicatie

Attributen van patiënten zoals: leeftijd, lengte, gewicht, bloeddruk, cholesterolniveau etc. kunnen voorkomende clusters identificeren. Deze clusters kunnen vervolgens gebruikt worden voor specifieke individu's, waarbij preventieve metingen worden uitgevoerd. Clusteren is o.a. nuttig in biologische vraagstukken met de daarbij horende classificatie van planten, dieren en menselijke genetica.

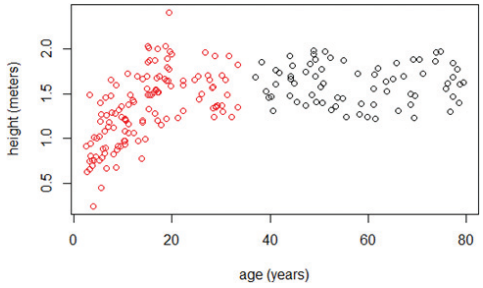
Klant segmentatie

Voor het identificeren van klanten, die vergelijkbare uitgavenspatronen en gedrag hebben kan er k-means gebruikt worden. Volgens (EMC Education Services, 2015) wordt dit model gebruikt voor het segmenteren van klanten. De attributen die behoren tot de klanten in het geval van een telefoniebedrijf zijn o.a. maandelijkse facturen, aantal sms berichten, data gebruik, aantal gebruikte belminuten en aantal jaar als klant zijnde. Het bedrijf kan dan kijken naar de gerealiseerde clusters om vervolgens een strategie te bedenken voor het verhogen van sales of het verlagen van klant churn rate oftewel het aantal abonnees/klanten, die in een bepaalde periode het bedrijf verlaten.

Alhoewel k-means een unsupervised learning techniek is, is het belangrijk om bepaalde keuzes te maken met betrekking tot het meenemen van bepaalde attributen, wat voor eenheid van meten er gebruikt wordt voor ieder attribuut, maar ook of de attributen opnieuw geschaald(rescaled) moeten worden, zodat een bepaald attribuut geen disproportioneel effect heeft op de resultaten. De essentie is dat de aantallen van de attributen vergelijkbaar kunnen zijn. Zoals aangegeven is het belangrijk om de attributen specifiek een eenheid van meting mee te geven. Bij het clusteren van patiënten, kan de leeftijd in jaren worden meegegeven en de lengte in centimeters. (Waarbij k=2), dan zouden de clusters als onderstaande figuur 23 uit zien, maar zodra de centimeters naar meters geschaald zijn, zouden de clusters als onderstaand figuur 22 zijn.



Figuur 23. Clustering voorbeeld patiënten



Figuur 22.. Clustering voorbeeld patiënten

8.1.2 Association rules

Bij Association rules wordt er gekeken naar hoe vaak een product wordt verkocht als een ander product gekocht wordt, maar ook of een klant lid wordt van een abonnement als het gelijkenissen toont met een ander klant. De Association rule kijkt naar de relatie en verbanden tussen de variabelen. Het doel van de rule is om interessante relaties tussen variabele/items te vinden. In het boek van (EMC Education Services, 2015, p. 139) wordt de apriori algoritme gebruikt om de Association rules toe te passen. Apriori is een van de eerste en fundamenteelst algoritme voor het genereren van Association rules. Het algoritme van apriori werkt als volgt:

1. Eerst wordt er een minimumsupport meegegeven. Indien deze op 0,5 staat, dan kan er geconcludeerd worden dat van alle transacties de helft (50%) meegenomen moet worden voor de berekening.
2. In de tweede stap worden de frequente paren samengesteld. Bij deze stap wordt de eerste stap herhaald.
3. Dit proces wordt herhaald tot de meest frequente paar van variabele/items overblijven.

Tussen stap twee en drie wordt ook de confidence rule gebruikt ter evaluatie van de samengestelde paar van variabele/items. De formule ziet er als volgt uit: Confidence (X->Y)=Support(X^Y) / Support(X). Volgens (EMC Education Services, 2015, p. 141) is de confidence gedefinieerd als de measure voor zekerheid en betrouwbaarheid, die geassocieerd zijn met ieder gevonden rule. Een voorbeeld die gebruikt kan worden ter verduidelijking is dat de confidence het percentage van x en y uit alle transacties is, waarvan de transactie bestaat uit x.

De support en de confidence zeggen veel over de zekerheid en betrouwbaarheid, maar neemt y niet in acht bij het kijken in hoeverre x en y gerelateerd zijn aan elkaar, in plaats van het toeval dat ze samen voorkomen. De formule voor de lift ziet er als volgt uit: Lift(X->Y)=Support(X^Y) / Support(X)*Support(Y). De lift bekijkt de ratio tussen de twee. Leverage lijkt op de wijze hoe lift berekend wordt, maar is meer toepasselijk bij het uitrekenen van de verschillen in mogelijkheden van x en y, die samen voorkomen, vergeleken met wat er individueel van elkaar verwacht werd van x en y.

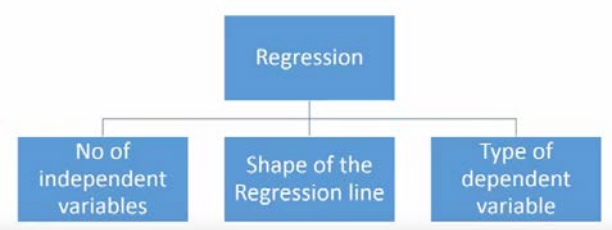
De leverage formule ziet er als volgt uit: Leverage (X -> Y)=support(X^Y)-Support(X)*Support(Y).

Voorbeelden waar Association rules gebruikt wordt zijn recommender systemen, maar ook clickstream analysis. Amazon en Netflix gebruiken o.a. de recommender system, waarbij Association rules wordt toegepast. Hierbij worden interessante ontdekkingen gemaakt met betrekking over films, die op elkaar lijken of gelijkenissen tussen klanten (lees: klanten die product A hebben aangeschaft, hebben ook product B aangeschaft).

8.2 Supervised learning

8.2.1 Regressie

Over het algemeen worden regressieanalyses gebruikt om te kijken in hoeverre een set van variabele invloed heeft op de uitkomst. In meeste gevallen worden de uitkomsten dependent variables genoemd. Volgens (Ray, 7 Types Of Regression Techniques., 2015) wordt de dependent als target omschreven. Omdat de uitkomst afhankelijk is van de set van variabelen wordt dit dependent genoemd. In sommige gevallen worden variabelen ook input en independent(predictor) variables genoemd. Regressieanalyses kunnen opgedeeld worden in twee soorten. Lineaire regressie en logistische regressie. Deze algoritmes behoren tot de supervised learning techniek.



Figuur 24. Regressie type

In het volgende stuk is het de bedoeling om duidelijkheid te verschaffen omtrent gebruik van regressieanalyse modellen. (Lees: lineair en logistisch). De type van regressie algoritmes zijn gebaseerd op verschillende factoren.

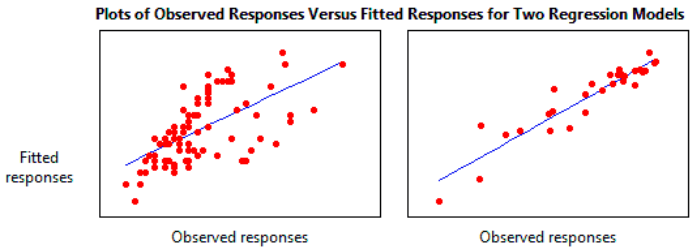
8.2.1.1 Lineaire regressie

Lineaire regressie is een van de meest bekende modeleer technieken. (Ray, 7 Types Of Regression Techniques., 2015). Volgens (EMC Education Services, 2015, p. 162) is het doel van lineaire regressie modellen om relaties tussen verschillende variabelen en uitkomsten te modeleren. In dit model gaat men ervan uit dat de relaties van de input variabelen en uitkomst lineair zijn. De relaties worden gerealiseerd aan de hand van de dependent variabele (Y) en een of meer independent variabelen(X), door gebruik te maken van de ”best fit straight line” ook wel bekend als de regressie lijn (Ray, 7 Types Of Regression Techniques., 2015). De formule voor de regressie lijn is Y=a + b* x + e, waarvan de a de intercept is, b de slope, oftewel de helling van de lijn en e de error term. Op basis van de bekende input waardes realiseert het model een verwachte uitkomst.
Met betrekking tot de lineaire regressie bestaat deze uit de *simple* lineaire regressie en *multiple* lineaire regressie. Het verschil is dat multiple lineaire regressie meer dan één independent variabele bevat, terwijl simple lineair regressie maar één independent variabele heeft. In het boek Azure Machine Learning (Barnes, 2015, p. 144) wordt aangegeven dat er een derde type bestaat, namelijk: multivariate. In deze variatie zijn meerdere gecorreleerde dependent variabelen voorspeld in tegenstelling tot simple en multiple lineair regression, waarbij er één dependent variable als voorspelling geldt. De waardes die gegenereerd worden zitten om de lineair lijn heen, echter is de vraag hoe “the best fit” te behouden is. Dit kan opgelost worden

door de "least square method". (Ray, 7 Types Of Regression Techniques., 2015) Dit is de meest bekende methode voor het verkrijgen van een passende regressie lijn. Deze methode berekent de best passende lijn van de geobserveerde data, door de sum of squares van de verticaal deviaties van ieder data punt van de lijn te minimaliseren.
De evaluatie van de regressie lijn vindt plaats door te kijken naar de metriek R-Squared. Volgens (Frost, 2013) is R-Squared een statistische measure omtrent de afstand van de data observatie punten ten opzichte van de regressie lijn. Dit wordt soms ook wel de coëfficiënt van determinatie genoemd. De definitie is het percentage van de response variabel variatie. De uitkomst van de R-Squared is tussen de 0 en 100%

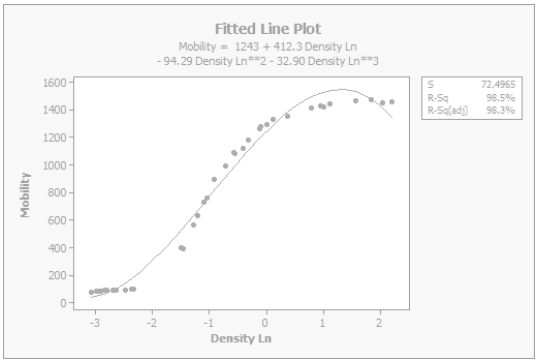
- 0% indiceert dat het model geen variabiliteit van de response data om het gemiddelde heeft oftewel de regressie lijn.
- 100% indiceert dat het model alle variabiliteit van de response data om het gemiddelde heeft (regressie lijn)

Dit betekent dat een hoger r-squared percentage aangeeft dat het model passend is op de data.



Figuur 25. R-Squared voorbeeld

In figuur 24 zijn twee resultaten van de regressie model afgebeeld. De linker illustratie geeft een r-squared percentage weer van 38 en het rechter model geeft een r-squared percentage weer van 87. Een hoger r-squared betekent niet in alle gevallen, dat het een betere uitkomst is. Soms is de verwachting dat de R-squared aan de lage kant zal zijn. Het voorspellen van menselijk gedrag zoals psychologie heeft typerend een lage r-squared waarde. In meeste gevallen is dit lager dan 50%. Als de r-squared laag is, maar er statistisch significante predictors zijn, dan is het mogelijk om te concluderen dat de verschillen in de predictor waarde geassocieerd zijn met de veranderingen in de response waarde. Een hogere r-squared waarde is in veel gevallen positief, maar het kan voorkomen dat de regressie lijn boven en onder de data valt(bias).



Figuur 26. R-Squared voorbeeld

Zoals in de bovenstaand figuur 25 is afgebeeld, kan bias voorkomen op het moment dat de lineaire regressie mode o.a. belangrijke missende predictor, polynomiaal termen en interactie termen bevatten. Statistici noemen dit specificatie bias. Voor het oplossen van dit probleem kunnen er correcte termen toegevoegd worden (Frost, 2013). Lineaire regressie kan volgens (MathWorks, 2016) het best worden toegepast als het algoritme makkelijk te geïnterpreteerd moet worden. Tevens kan het model gebruikt worden als vergelijkingsmateriaal voor het evalueren van complexe regressie modellen.

8.2.1.1.1 Use cases lineaire regressie

Huizenprijzen

Met lineaire regressie is het mogelijk om de huis prijzen te voorspellen op basis van het gebied waar de woning zich bevindt. Dit model helpt de prijzen in de markt te evalueren. Dit model is verder ook uit te breiden in dezelfde situatie, door gebruik te maken van de input variabele, zoals het aantal badkamers, slaapkamers en criminaliteit statistieken.

Demand forecasting

Business en regeringen kunnen de lineaire regressie model gebruiken voor het voorspellen van de vraag van producten en services. Een restaurantketen kan de type en kwaliteit van voedsel, die klanten consumeren voorspellen op basis van het weer, dag van de week, product aangeboden wordt als actieprijs, tijd van de dag en reservatie hoeveelheid. Dit model kan ook nuttig zijn voor o.a. retail verkopen

Medisch

Een lineair regressiemodel kan gebruikt worden voor het analyseren van het effect van radiatie behandeling voor het verminderen van tumor. Input variabelen kunnen o.a. duur van radiatie behandeling zijn, de frequentie van radiatie behandeling en patiënten attributen zoals leeftijd en gewicht.

8.2.1.2 *Logistische regressie*

Logistische regressie werkt anders dan lineair regressie. Als bijvoorbeeld het inkomen van een persoon ter beoordeling van de voorspelling niet nodig is, maar eerder of iemand rijk of arm is en de uitkomst variabel categorisch van nature is, dan is de logistische regressie te gebruiken ten behoeve van het doen van voorspellingen. Logistische regressie kan gebruikt worden voor de aannemelijkheid van de uitkomst, dat gebaseerd is op de input variabelen. De uitkomst van een logistische regressie is binair oftewel een nul en een één. Dit kan vertaald worden naar ja en nee en in bepaalde situaties naar een true/false of een pass/fail resultaat.

De logistische regressie model is gebaseerd op de logistische functie: $f(y) = \frac{e^y}{1+e^y} \quad \text{for} -\infty < y < \infty$

Door de range van f(y) is (0,1) is de logistische functie een geschikt model voor de zekerheid van een specifiek uitkomst. Als de waarde van y stijgt, stijgt de zekerheid van de uitkomst. Volgens (Applying Supervised Learning, 2016) kan logistische regressie gebruikt worden, zodra de data door een lineair boundary onderscheiden kan worden. Tevens kan het model gebruikt worden ter vergelijking en evaluatie van complexe classificatie modellen.

8.2.1.2.1 Use cases logistische regressie

Medisch

Model dat gebouwd kan worden op basis van de logistische regressie is het voorspellen of bepaalde behandelingen van een patiënt succesvol zullen zijn. Input variabelen kunnen o.a. leeftijd, gewicht, bloeddruk en cholesterolniveaus zijn.

Financieel

Met het gebruik van geschiedenis uitbetalingsgegevens en details van het loon, kan de waarschijnlijkheid dat een aanvrager op de lening in gebreke zal blijven voorspeld worden.

Marketing

In de marketingwereld is het belangrijk om klanten aan te trekken, maar ook customer churn is belangrijk. Customer churn wordt ook wel customer attrition genoemd gaat omtrent klanten die de leverancier verlaten. Denk bijvoorbeeld aan klanten met een tv-abonnement. In het geval van telefoniebedrijf zal de logistische regressie nuttig zijn voor het weerhouden van “churning”. Dit model kan gevoed worden op basis van leeftijd, aantal familie leden die dezelfde abonnement gebruiken, resterende maanden op het abonnement en sociale netwerk contacten. Met de informatie kan de marketingafdeling aanbiedingen aanbieden.

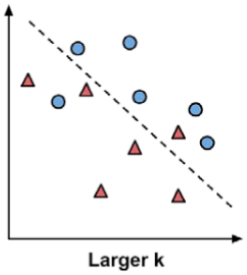
Engineering

Logistische regressie modellen kunnen gebruikt worden ten behoeve van het voorspellen van de zekerheid dat een elektrisch apparaat last krijgt van storingen. Met deze inzichten kan onderhoud gepland worden.

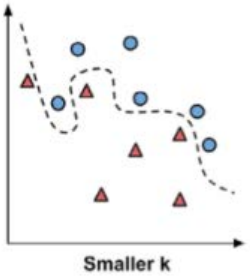
8.2.2 K- Nearest Neighbor

Volgens (Zhang Z. , 2016, p. 2) is kNN een unsupervised learning techniek. Terwijl (Peterson, 2009) aangeeft dat kNN een supervised learning techniek is. In het onderzoek van (Xindong Wu, 2007, p. 22) wordt aangegeven dat kNN bestaat uit drie sleutel elementen: set van labeled objecten, similarity metriek en het aantal k's. In het onderzoek van (Zhang Z. , 2016) wordt de werking van kNN als volgt omschreven: de unlabaled observaties worden toegewezen aan de observaties, die de meeste gelijkenissen vertonen. De karakteristieken van de observaties worden verzameld voor de training en test dataset. Fruit en groente kunnen bijvoorbeeld onderscheiden worden door de structuur en zoetigheid. Zodra granen als observatie in de dataset voorkomen, wordt er gekeken naar welke toegewezen observaties het meest overeenkomen.

Op het moment dat er een meerderheid is van de gelijkenissen met de al toegewezen observaties, behoort de nieuwe observatie tot dat gebied. Voor het berekenen van deze gelijkenissen wordt normaal gesproken de euclidische afstand gebruikt. Tevens kan het aantal k aangegeven worden. De correcte keuze van de hoeveelheid van k kan een significante impact hebben op de performance van de kNN algoritme. Zodra een hoog aantal wordt aangegeven, wordt de impact van de variantie met betrekking tot random foutmeldingen, geminimaliseerd. Dit zorgt overigens dat kleine, belangrijke patronen worden genegeerd. Tevens vindt er op basis van de hoeveelheid k een wijziging plaats op de decision boundry (lijn tussen de observaties), zoals in de onderstaande illustraties afgebeeld is.



Figuur 28. Decision boundary 1



Figuur 27. Decision boundary 2

Hawkins geeft in zijn artikel (Hawkins, 2004, p. 72) aan dat de sleutel voor underfitting en overfitting, het vinden van de balans door middel van de juiste hoeveelheid k is. Lantz, Brett (Lantz, 2015, p. 71) Concludeert in zijn boek dat het aantal k gelijk moet zijn aan de square root van het aantal observaties in de training dataset voor toepassen van het juiste aantal k. Volgens (Saxena, 2016) zijn de voordelen van kNN als volgt:

1. Simpele implementatie
2. Snel resultaat met kleine dataset
3. Performance gelijk aan bayesian classifier
4. Heeft geen voorkennis nodig omtrent de structuur van de training set.
5. Hertrainen is niet nodig, indien nieuwe trainingspatroon is toegevoegd aan het bestaande training set.

In het rapport van (Saxena, 2016) wordt ook aangegeven dat de limitatie 's van k-Nearest Neighbor (K-NN) o.a. op het moment dat de training set groot is, het geheugengebruik is en voor ieder test data moeten de afstanden tussen de test data en training set berekend worden. Dit neemt veel tijd in beslag. Het is dus belangrijk om te kijken naar het geheugengebruik. K-nn is een toepasbaar algoritme als geheugen gebruik geen invloed heeft op de keuze voor het kiezen van een algoritme. Tevens geeft (Vishwanathan, 2010, p. 25) aan dat de voorspellingen "noisy" kunnen zijn. In (MathWorks, 2016) aangegeven dat k-nn gebruikt kan worden voor het tot stand brengen van benchmarking learning regels met een simpel algoritme. Verder is het gebruik van k-nn ook interessant als voorspellingssnelheid geen hoge prioriteit heeft.

8.2.3 Naive Bayes

De naive bayes classificatie methode is gebaseerd op de Bayes theorie en is volgens (Xindong Wu, 2007, p. 24) een supervised learning techniek. De Bayes theorie geeft de relatie tussen twee waarschijnlijkheden van twee gebeurtenissen. De classificatie gaat ervan uit dat de aanwezigheid of afwezigheid van een bepaald kenmerk van een class niet gerelateerd is aan de ander kenmerk. In het boek (EMC Education Services, 2015, p. 211) wordt aangegeven dat een object geclassificeerd kan worden op basis van vorm, kleur en gewicht. Als het object rond, geel en minder dan 60 gram weegt, is de kans groot dat dit een tennisbal is. Naive Bayes gaat ervan uit dat ieder kenmerk individueel een rol speelt bij het classificeren van het object. In het algemeen zullen de input variabelen categorisch zijn. Daarnaast kan het algoritme ook omgaan met doorlopende variabelen. Volgens (EMC Education Services, 2015, p. 211) zijn er verschillende manieren om doorlopende variabelen te converteren naar categorische variabelen. Dit proces wordt ook wel "discretization of continuous variables" genoemd. Het voorbeeld omtrent inkomen zou als volgt kunnen worden onderverdeeld:

- Lage inkomen: inkomen < 10,000
- Arbeiders: 10,000 ≤ inkomen < 50,000
- Middenklasse: 50,000 ≤ inkomen < 1,000,000
- Hoge klasse: inkomen ≥ 1,000,000

Net zoals bij kNN is naive Bayes een classificatie methode die gemakkelijk te implementeren is en efficiënt is in de uitvoering. Hiervoor heeft het model geen voorkennis nodig. Net zoals ieder ander machine learning algoritme heeft naive bayes een toepasbaare formule. De voorwaardelijke kans dat gebeurtenissen C voorkomen, waarvan de gebeurtenis A al is voorgekomen wordt aangeduid als $P(C|A)$. De formule hiervoor is $P(C|A) = P(A \cap C) / P(A)$ een ander Bayes theorie is de aangepaste conditionele kans formule dat gebaseerd is op de vorige formule. De formule is $P(C|A) = P(A|C) * P(C) / P(A)$. De C staat in dit geval voor class label en de A voor geobserveerde attributen. De aangepaste conditionele kans formule met betrekking tot de Bayes theorie is tevens de meest bekende Bayes theorie.
Wiskundig geeft het Bayes theorie de relatie tussen de zekerheden van C en A, P(C) en P(A) en de aangepaste conditionele kansen van C met A en A met C, namelijk: $P(C|A)$ en $P(A|C)$. De Bayes theorie is significant, omdat in veel gevallen vanuit de training data $P(C|A)$ ingewikkelder is om te berekenen dan $P(A|C)$ en P(C). Een voorbeeld dat in het boek (EMC Education Services, 2015, p. 212) wordt aangegeven is John, die frequent vliegt en graag zijn boeking wil upgraden. John is van mening dat als hij twee uur voor vertrek wil inchecken, de kans naar een upgrade 0.75 oftewel 75% is. Door zijn drukke schema is hij 40% van zijn vluchten, twee uur voor vertrek aanwezig om in te checken voor zijn vlucht. Om de kans uit rekenen dat John niet twee uur voor zijn vlucht aanwezig is, kan het volgende toegepast worden: C. = {John arriveert twee uur voor vlucht}, A = {John heeft een upgrade}, ¬C = {John is niet twee uur voor vertrek gearriveerd} en ¬A {John heeft geen upgrade}. Omdat John 40% van alle vluchten twee uur voor zijn vlucht heeft ingecheckt is P(C) = 0.4, dus P(¬C) = 1-P(C) = 0.6. De kans dat John een upgrade heeft ontvangen, op het moment dat John twee uur voor vertrek aanwezig was, is 0.75, dus $P(A|C) = 0.75$.

De kans dat John een upgrade heeft ontvangen op het moment dat John twee uur voor vertrek niet aanwezig was, is 0.35, dus P (A | ¬C) = 0.35 en vervolgens weer P (¬A | ¬C) = 0.65. De kans dat John een upgrade heeft ontvangen P(A) kan als volgt worden berekend:

P(A) = P(A ∩ C) P+(A ∩¬C)
= P(C)*P(A|C)+P(¬C)*P(A | ¬C)
= 0.4 *0.75+0.6*0.35
= 0.51

Het resultaat van de theorie van Bayes geeft aan, dat John 49% kans heeft op het niet ontvangen van een upgrade.

De naive bayes algoritme heeft volgens (Hackerearth, 2017) drie verschillende types:

1. Gaussian – deze algoritme gaat ervan uit dat de distributie van kenmerken normaal verdeeld zijn. Gaussian is vernoemd naar de Duitser Carl Friedrich Gauss).
2. Multinomiaal – Dit wordt gebruikt op het moment dat de data multinomiaal is gedistribueerd en de gebeurtenissen (incidenten) belangrijk zijn.
3. Bernoulli – dit wordt gebruikt op het moment dat de kenmerken in de dataset binair waardes bevatten.

Het kan voorkomen dat de dataset missende waardes bevatten. Naive bayes kan goed omgaan met dit probleem door Laplace smoothing toe te passen. In het boek van (EMC Education Services, 2015, p. 217) wordt er namelijk aangegeven dat smoothing de nul waarde omwisselt naar een minimale waarde. Met deze oplossing kan naive bayes goed omgaan met missende waardes. Tevens is het model ook robuust tegen irrelevante variabelen. Dit zijn variabelen, die onder alle classes zijn gedistribueerd, waarvan het effect niet is uitgesproken. Daarnaast is het voordeel van naive bayes, dat het een efficiënt berekening heeft en niet veel rekenkracht nodig heeft.
Dit model kan overigens goed en efficiënt omgaan met data doe een hoge dimensionaliteit kent. Volgens onderzoek (Zhang H. , 2004) blijkt dat de naive bayes classificatie concurrerend is met andere machine learning algoritmes. Verder geeft (Hackerearth, 2017) aan dat dit model sneller is met voorspellingen en dat het op basis van kleine datasets ook gemakkelijk getraind kan worden. Nadeel naast de eerder benoemde “zero conditional probability problem”, waarvoor Laplace smoothing de oplossing is, is de sterke onafhankelijkheid veronderstelling.

8.2.3.1.1 Use cases naive Bayes

Filteren van spam

Dit is een voorbeeld van tekst classificatie en is een populair methode om e-mails te onderscheiden van spam.

Fraude detectie

Volgens het onderzoek van (Bhowmik, 2008, p. 48) kan naive bayes gebruikt worden voor fraude detectie en bijvoorbeeld cijfer herkenning. Het onderzoek geeft aan dat het model betrouwbaar is met betrekking tot het gebruik voor fraude detectie.

Voorraad beheer management

Dit algoritme kan toegepast worden bij het voorspellen van voorraadbeheer. In het onderzoek van (Manas Gaur, 2015) wordt geconcludeerd dat dit model betere resultaten leverde ten opzichte van kNN omtrent voorraadbeheer voorspellingen.

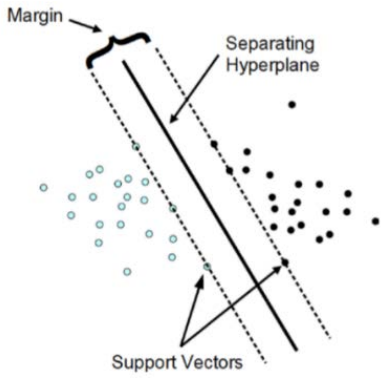
Financiële voorspellingen

Uit een recentelijk onderzoek (Mahajan Shubhrata D, 2016, p. 123) komt naar voren dat naive bayes de potentie heeft om gebruikt te worden voor het voorspellen van aandeelprijzen.

8.2.4 Support Vector Machine (SVM)

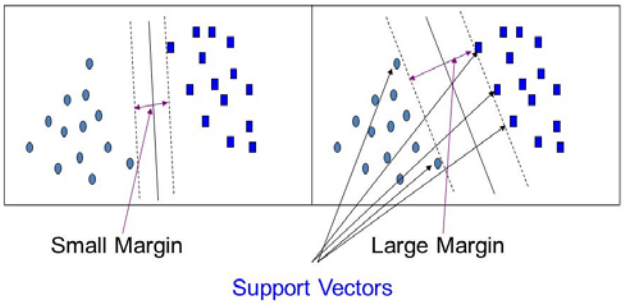
Support Vector Machine (SVM) is een training algoritme voor het leren van classificatie en regressie regels. Volgens een onderzoek van (Kotsiantis, 2007, p. 260) is SVM de nieuwste supervised learning techniek. In het onderzoek van (Robert Burbidge, 2011, p. 5) wordt aangegeven dat SVM supervised en unsupervised learning technieken kan toepassen in verschillende situaties. De input data kan namelijk gelabeld, maar ook unlabeled zijn. SVM kan omgaan met zowel lineaire als non-lineair data. SVM-modellen lijken zeer veel op neurale netwerkmodellen (Ayodele, 2016, p. 7). De twee belangrijke elementen van dit algoritme, zijn de technieken van wiskundige codering en kernel functies (Robert Burbidge, 2011, p. 5). De kernel functies zijn een alternatieve trainingsmethode voor polynomiaal, radiaal basisfunctie en multilagen perceptron classificaties, waarvan het gewicht wordt gevonden door het berekenen van kwadratische programmeer problemen met lineaire lijnen.
Het algoritme is een van de meest robuuste en accurate methoden (Xindong Wu, 2007, p. 10). Tevens kan dit algoritme goed omgaan met hoge dimensie data. Op het moment dat SVM wordt toegepast in twee klassen learning, is het doel van SVM om de beste classificatie functie te vinden voor het onderscheiden van de objecten training set. De metriek voor de beste classificatie functie kan geometrisch gerealiseerd worden. In het geval van een lineair onderscheiding tussen de data objecten, kan er gebruik gemaakt worden van onderscheidende hyperplane, dat diagonaal langs de objecten plaats vindt. Hyperplane wordt overigens gezien als een feature ook wel kenmerk genoemd. Het kan voorkomen dat de SVM geen onderscheidende hyperplane kan vinden. Dit probleem kan opgelost worden door gebruik te maken van de soft margin, die misclassificaties van de trainingstest accepteert. Naast soft margin als oplossing is het ook mogelijk om de kernel aan te passen ten behoeve van het mogelijk maken van classificatie van non lineaire data. Voor deze wiskunde oplossing worden de simpliciteit van onderscheidende hyperplanes behouden. Deze aanpak beroept zich op de wiskundige methode van hyperplanes. Het is ook mogelijk om een nieuwe feature toe te voegen door de volgende formule toe te passen: z=x^2+y^2.
Volgens (Deshpande, When do support vector machines trump other classification methods, 2013) wordt in meeste gevallen SVM kernel functies automatisch toegepast, zonder dat de gebruiker handmatig de transformatie van data moet uitvoeren. Het kan voorkomen dat de dataset uit verschillende attributen bestaat en het daardoor moeilijk wordt om een specifieke kernel te gebruiken. De meest gebruikte kernels zijn de eerdere benoemde alternatieve traingingsmethode. Aan de hand van onderzoek (Hsin-Yuan Huang) wordt aangegeven dat non lineair classificatie de voorkeur krijgt indien men accuraatheid belangrijk vindt. Voor het toepassen van non lineair classificatie wordt de non lineair kernel toegepast. Daarnaast wordt in het onderzoek vermeld dat lineair classificatie de voorkeur krijgt als het aantal features (kenmerken) zeer hoog zijn (bv: bij een document classificatie). De voorkeur is gebaseerd op snelheid.
De werking van SVM van een lineaire oplossing is opgedeeld in verschillende fasen. In eerste instantie worden de data objecten op de X en Y-as geplot. Vervolgens wordt door het toepassen van 1-dimensionale hyperplane (lijn) getracht de objecten van elkaar te onderscheiden.

Het verschil in afstand tussen de lijnen en objecten wordt de margin genoemd, zoals in het onderstaande illustratie is afgebeeld.



Figuur 29. SVM

De objecten die zich rondom de lijn bevinden, worden support vectoren genoemd. Een vector is een rij van data, dat waardes voor verscheidene attributen bevat (Deshpande, When do support vector machines trump other classification methods, 2013). Volgens (Kotsiantis, 2007) worden andere data punten genegeerd en wordt het eindresultaat gerealiseerd op basis van de support vectoren.
Omdat er verschillende hyperplanes manieren zijn, is de vraag wat de beste hyperplane is en hoe deze optimale lijn bepaald wordt. De SVM-analyse, zoals in het onderzoek van (Gonzalez-Abril, 2005) wordt aangegeven, kan de beste en optimale lijn gevonden worden door de margin tussen het aantal support vector te maximaliseren, zoals in het onderstaande illustratie is afgebeeld.



Figuur 30. Support vectors voorbeeld

De voordelen van het gebruik van SVM als machine learning algoritme zijn (Deshpande, When do support vector machines trump other classification methods, 2013):

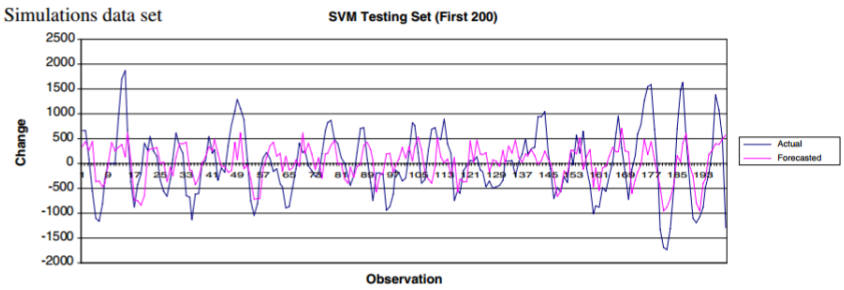
1. Zodra een boundary(van de hyperplane) is vastgelegd, wordt de meerderheid van de desbetreffende training data overtollig. Hierdoor zullen kleine aanpassingen geen effect hebben op de hyperplane. Dit betekent overigens dat SVM de neiging heeft om goed te generaliseren.
2. Resistent tegen overfitting
3. Zeer hoge accuraatheid wat ten koste gaat van de snelheid.
4. Effectief in data met een hoge dimensionaliteit

Nadelen van het gebruik van SVM als machine learning algoritme zijn (Ray, Understanding Support Vector Machine algorithm from examples, 2015):

1. Performance is slecht in het geval van een grote dataset.
2. Performance is slecht als data ruis bevat.
3. Lage snelheid m.b.t. training van dataset. (Kotsiantis, 2007, p. 261)

8.2.4.1.1 Use cases SVM

Op basis van onderzoek (Real Carbonneau, 2007) kan geconcludeerd worden dat SVM toepasbaar is op het voorspellen van voorraadbeheer. In het onderzoek komt naar voren dat SVM vergeleken met andere machine learning modellen een hoger accuraatheid realiseert.



Figuur 31. SVM-toepassing op voorraadbeheer voorspelling

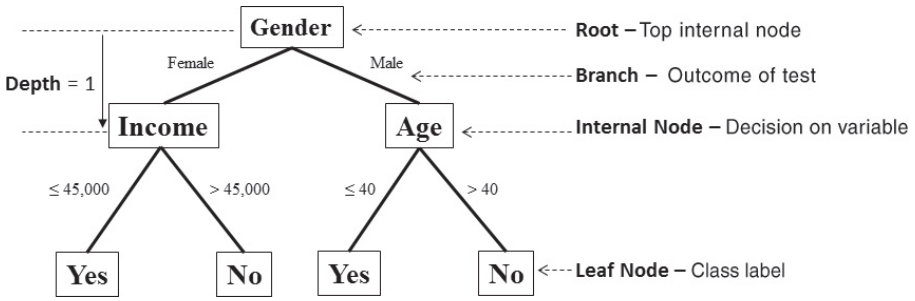
8.2.5 Decision Tree

Volgens een gepubliceerd artikel (Center For Computational Science - University of Miami, 2009) is decision tree een krachtige methode voor classificatie en het maken van voorspellingen. Het model gebruikt een boomstructuur voor het specificeren van volgordes omtrent keuzes en consequenties (EMC Education Services, 2015, p. 192). Bij het toepassen van decision tree zijn de input waarden categorisch en continu. Door de flexibiliteit en visualisatie mogelijkheden is dit model veelal gebruikelijk bij data mining applicaties met betrekking tot het classificeren. De decision tree heeft twee variaties (EMC Education Services, 2015, p. 193):

1. Classification trees
2. Regression trees

Classificatie tree zijn voornamelijk van toepassing, waarbij de output variabelen categorisch en vaak binair zijn. Regression trees zijn van toepassing, waarbij de output variabelen numeriek en continu zijn, zoals het voorspellen van de prijs van een product of de kans dat een abonnement aangeschaft zal worden. Het resultaat is gebaseerd op de if-then statements. Een onderliggende onderstelling van een lineair of non-lineair relatie tussen de input en response variabelen is bij decision tree niet het geval.

In het onderstaande illustratie refereert de branche de uitkomst van de beslissing en wordt als lijn tussen de twee nodes (root en internal) gevisualiseerd. De internal nodes zijn de beslissingen, waarvan aftakkingen de consequenties of resultaten zijn. In dit voorbeeld zijn de leaf nodes oftewel een node zonder verdere branches, geplot in binaire resultaten.



Figuur 32. Decision tree structuur

Het limiteren van splits oftewel aftakkingen, worden deze korte trees gecreëerd. Deze korte trees worden vaak gebruikt als componenten (weak learners/base learners) in ensemble methoden (EMC Education Services, 2015, p. 194). Ensemble methoden zijn zoals in het onderzoek (III, 2012) o.a.: Bagging, AdaBoost en Random Forrest. Deze ensemble methoden gebruiken meerdere voorspellende modellen gebaseerd op de decision tree, waarvan het resultaat een combinatie van de meerderheid uit de voorspellingen zijn. Naast het limiteren van splits, kan er ook gekozen worden om een decision stump te creëren. Hierbij is de root direct verbonden met de leaf node.

In het algemeen is de werking van decision trees het creëren van een boomstructuur op basis van de attributen in de training set. Als alle data uit de trainingstest behoren tot bepaalde

klassen (abonnement aanvraag = Ja), dan wordt de desbetreffende node een leaf node. Op het moment dat niet alle data uit de training set behoren tot een klassen, selecteert het algoritme de meest informatieve attribuut en verdeelt de training set corresponderend op de informatieve attribuut.

Het algoritme creëert splits (aftakkingen) voor de subsets uit de training set recursief tot dat een van de criteria zijn voldaan:

1. Alle leaf nodes voldoen aan de minimum purity treshhold
2. De decision tree kan niet verder gesplit worden met de aangegeven minimum purity treshhold
3. Overige stopping criteria is voldaan (o.a. maximum depth van de decision tree)

De eerste stap in het creëren van een decision tree is de keuze van de meest informatieve attribuut. In het boek wordt aangegeven dat het identificeren van de informatieve attribuut uitgevoerd kan worden door te kiezen voor de entropie gebaseerde methode (EMC Education Services, 2015, p. 197). Dit zijn learning algoritmes zoals ID3 en o.a. C4.5. De meest informatieve attribuut wordt geselecteerd door twee metingen:

1. Entropie – dit meet de impurity van de attribuut
2. Information gain – dit meet de purity van een attribuut

De information gain wordt berekend op basis van het vergelijken van de purity van de parent node voordat de split met de purity van de child node wordt uitgevoerd. Het hoogste aantal geeft aan in hoeverre het attribuut “pure” is. Zoals aangegeven zijn er twee manieren om de informatieve attribuut de identificeren. ID3 staat voor iteratief dichotomiser 3 (Wikipedia, 2017). Dit algoritme is een van de eerste decision tree algoritmes en is ontwikkeld door John Ross Quinlan. Volgens het boek (EMC Education Services, 2015, p. 203) is C4.5 een van de opvolgers van ID3 en kent verscheidene verbeteringen t.o.v. ID3. De C4.5 algoritme kan namelijk goed omgaan met missende data. Tevens kan de ID3 een diepe en complexe tree creëren, waardoor overfitting kan plaatsvinden. De C4.5 lost dit probleem op door een bottom-up techniek toe te passen dat pruning heet.

Voor decision tree learning kan overfitting door het gebrek aan training data of biased training data zijn. In het boek (Mitchell, 1997, p. 68) worden twee oplossing voorgesteld.

1. De tree op een vroeg stadium laten stoppen met groeien, voordat het punt bereikt wordt waar het gehele training data perfect is geclassificeerd.
2. De tree laten overfitten, waarna vervolgens prune toegepast kan worden.

8.2.5.1.1 Use cases decision trees

Classificeren

Met decision trees is het mogelijk om dieren te classificeren. Dieren kunnen namelijk onderscheiden worden in koudbloedig, warmbloedig en zoogdier of geen zoogdier.

Videogame

Artificiële intelligentie engine uit een videogame gebruikt decision tree voor het controleren van autonome acties van karakters in specifieke situaties, die tijdens de game plaatsvindt.

Retailers

Bedrijven kunnen decision trees gebruiken ten behoeve van het segmenteren van klanten of voor het voorspellen van reacties op marketing en promoties.

Financieel

Decision trees kunnen banken en financiële instanties helpen bij het beslissen en goedkeuren van leningen. In dit geval kan de if-then statements toegepast worden voor het voorspellen of een klant van de bank zal gaan falen op het loon.

Het achterliggende proces van decision tree is over het algemeen goedkoop en het is ook makkelijk om data te classificeren. Zoals aangegeven kunnen verscheidene decision tree algoritmes de belangrijkheid van de input variabelen detecteren. Dit model kan omgaan met numerieke en categorische attributen en is robuust als het gaat om redundant en gecorreleerde variabelen. Als zich een groot aantal gecorreleerde variabelen in de dataset bevinden, kan het model daar minder goed mee omgaan. Daardoor zal instabiliteit en overfitting optreden. Om dit probleem op te lossen kan er gekozen worden voor het combineren van beslissingen oftewel het toepassen van random forest. Volgens het boek hebben ensemble methodes zich bewezen omtrent het verbeteren van voorspellende kracht t.o.v. een single decision tree. Tevens kan het model ook omgaan met non-lineaire effecten op het resultaat. Volgens het boek (EMC Education Services, 2015, p. 205) werkt dit model voor hoge non lineaire problemen daarom beter dan lineaire modellen. Het model is zeer sensitief voor kleine veranderingen in de dataset. Op het moment dat er twee verschillende subsets zijn gecreëerd, is de kans groot dat de decision trees van elkaar zullen verschillen. Het gebruik van decision trees wordt afgeraden indien er irrelevante variabelen in de dataset bevinden. Om het probleem van irrelevante variabelen op te lossen, wordt er aangeraden om feature selection toe te passen tijdens de data pre processing fase.

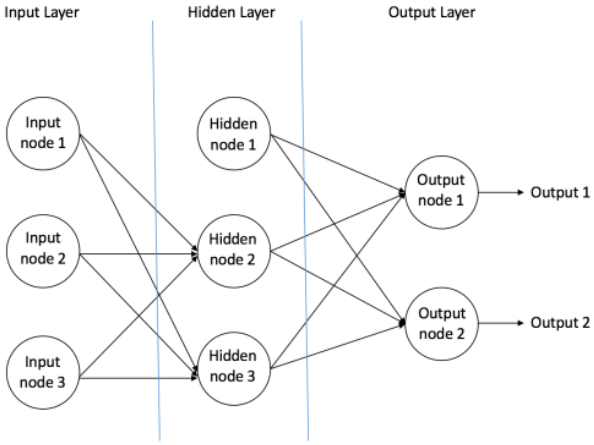
In (MathWorks, 2016) wordt aangegeven dat decision tree gebruikt kan worden als er vraag naar een gemakkelijk interpreteerbaar algoritme is. Daarnaast minimaliseert het geheugen gebruik. Indien een hoog accuraatheid percentage geen requirement is, is decision tree een geschikt model om te gebruiken. Uiteraard hangt dit af van verschillende factoren.

8.2.6 Neural network

Neurale netwerk is een machine learning model, dat bedacht en gecreëerd is op basis van de biologische werking van de neurale netwerk in het menselijk brein bij het verwerken van informatie (ujjwalkarn, 2016). Neurale netwerk is een alternatief op logistische regressies, waarvan de statistische techniek het meest overeen komt (Tu, 1996, p. 1) Het model werkt o.a. met perceptrons. Perceptrons zijn wiskundige modellen van biologische neuronen in het menselijke brein. Neurale netwerk heeft de bekwaamheid om wiskundig de relaties tussen de input en output te leren (Stanford University, 2017). Volgens onderzoek (Kotsiantis, 2007, p. 255) kunnen perceptronen alleen lineaire sets van data classificeren. Multilayered perceptronen zijn gecreëerd ten behoeve van het oplossen van het probleem, waarbij sets van data niet allemaal correct worden geclassificeerd i.v.m. de restricties tot het lineair classificeren. De multilayered neural network bestaat uit een groot aantal neuronen, die met elkaar in verbinding staan. De multilayered network bestaat uit drie lagen:

1. Input layer
2. Hidden layer
3. Output layer

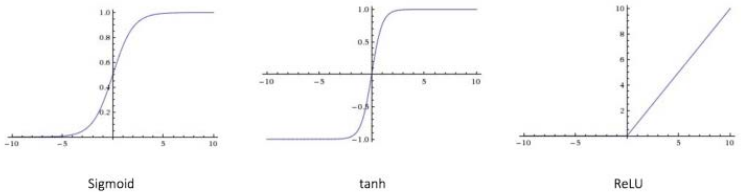
In het onderzoek van (Kotsiantis, 2007) worden de layers als units beschreven. Feed forward neural network oftewel FFNN maakt het mogelijk om signalen eenmalig van de input naar de output te laten verwerken.



Figuur 33. Neurale Netwerk (FFNN)

Het netwerk wordt als eerst getraind voor het beslissen van de mapping tussen de input-output. Vervolgens worden de gewichten van ieder connectie gerealiseerd, waardoor het netwerk de classificatie kan uitvoeren. De input nodes bevatten informatie van de buitenwereld. In de input layer worden er geen berekeningen uitgevoerd. De input nodes zijn puur voor het doorgeven van informatie naar de hidden node. In de hidden nodes zijn er geen directe connecties met de buitenwereld. Berekeningen worden in de hidden node uitgevoerd. En verstuurd de informatie van de input node naar de output node. De berekeningen kunnen afhankelijk van de type activatie functie verschillen. Een verzameling van de hidden nodes creëren de hidden layer. Tevens zijn de output nodes verantwoordelijk voor het berekenen en versturen van informatie naar de buitenwereld.

Op basis van de dataset kunnen er verschillende activatie functies worden toegepast. Deze bestaan uit: sigmoid, tanh en relu. In het geval van sigmoid zijn de input waarden real-valued en transformeert het de real-valued data type naar een numeriek getal tussen de nul en één. De formule die hierbij geldt is: ($\sigma(x) = 1 / (1 + \exp(-x))$). Terwijl de tanh functie, de waardes tussen de min één en één transformeert. De fomule die hierbij geld is: ($\tanh(x) = 2\sigma(2x) - 1$). Een derde activatie functie is Relu. Relu staat voor rectified linear unit. Deze functie neemt real-valued en legt een drempel op de nul. Negatieve waarden worden vervangen door het cijfer nul. Voor deze functie geldt de volgende formule: ($f(x) = \max(0, x)$)



Figuur 34. FFNN Activatie types

Het doel van de activatie functies is het introduceren van non-lineairity naar de output vanuit de neuronen. Dit is volgens (ujjwalkarn, 2016) belangrijk, omdat de huidige data die tegenwoordig gegenereerd wordt non lineair is. Het learning algoritme van neurale netwerken is de backpropagation algoritme (Richard D. De Veaux, 2017, p. 3). Dit algoritme is een supervised learning methode en leert van fouten die uitgevoerd worden. De werking van het algoritme is door middel van het willekeurig toevoegen van gewichten aan de edges oftewel connecties van de input nodes. De output wordt vergeleken met de gewenste output, waarvan de error wordt aangetoond aan de voorgaande layer. De error wordt onthouden en de gewichten worden vervolgens ten behoeve van het verminderen van het aantal foutmeldingen aangepast tot dat het minimaal aantal bereikt wordt. Machine learning modellen zoals neurale netwerken worden in praktijk gebruikt voor het vinden van patronen en trends.

In het onderzoek (Tu, 1996) worden een aantal voordelen en nadelen van het neurale netwerk omschreven. Zo is het voordeel o.a. dat het model de bekwaamheid heeft om alle mogelijke relaties van de predictor variabelen te detecteren. De nadelen zijn o.a. dat het model gevoelig is voor overfitting en dat het model veel rekenkracht nodig heeft. Tevens wordt in (MathWorks, 2016) aangegeven dat neurale netwerk in acht genomen kan worden indien data incrementeel aangeboden wordt en daarmee telkens het model mee wilt updaten. Daarnaast kunnen neurale netwerken ook goed om gaan met onverwachte wijzigingen in data. Als er geen eisen zijn voor het interpreteren van de output, dan is neurale netwerk een model dat toegepast kan worden. Uiteraard hangt dit af van andere factoren en zijn deze redenen onafhankelijk van de scenario's.

8.2.6.1.1 Use case neural network

Volgens (Stergiou) zijn er een aantal voorbeelden waarbij het model is toegepast op verscheidene situaties: verkoop voorspellingen, klant onderzoek, data validatie, risicomanagement, target marketing, gezichtsherkenning en onderhoud voorspelling.

8.2.7 Vergelijking supervised modellen

In het onderzoek van (Kotsiantis, 2007, p. 263) worden de supervised classificatie modellen op basis van verschillende punten met elkaar vergeleken. De benoemde punten in het vergelijkingstabel zijn overlappend op de punten die behandeld zijn in hoofdstuk vijf. In het onderzoek van Kotsiantis is k-means niet behandeld, omdat het een unsupervised learning techniek betreft. Om deze reden is het niet mogelijk om k-means te gebruiken ter vergelijking van de prestaties met andere machine learning algoritmes.

	Decision tree	Neural network	Naïve bayes	kNN	SVM
Accuraatheid	**	***	*	**	****
Snelheid learning	***	*	****	****	*
Snelheid classificatie	****	****	****	*	****
Tolerantie voor missende waarden	***	*	****	*	**
Tolerantie voor irrelevante attributen	***	*	**	**	****
Tolerantie voor overtollige attributen	**	**	*	**	***
Omgang met onopvallende/binair/voortdurende attributen	****	***	***	***	**
Tolerantie voor ruis	**	**	***	*	**
Omgang met overfitting	**	*	***	***	**
Totaal	25	18	25	19	24

*Tabel 6. Vergelijken machine learning algoritmes (**** representeert een goed en * een slechte performance)*

Zoals in het bovenstaande tabel te zien is, hebben de volgende modellen een hoge score: decision tree, naive bayes en SVM. De regressie modellen zijn in het onderzoek niet behandeld, maar doordat kmeans in dit onderzoek de enige behandelde unsupervised learning methode is en in een eerder vermeld onderzoek in combinatie met logistische regressie aanbevolen wordt, zal dit niet gebruikt worden ter vergelijking. Op basis van de plus -en minpunten zal SVM gebruikt worden ter vergelijking met de logistische regressie. De vraag is uiteraard of de oplossing voor klant segmentatie zoals in paragraaf 4.1 beschreven staat, de juiste aanpak is.

8.3 Conclusie

In hoofdstuk twee uit dit onderzoek kwam naar voren dat klant segmentatie, betalingsvoorkeur en afleveringsvoorkeur mogelijk zijn. Op basis van benoemde mogelijkheden voor het verrijken van klant profielen als input voor marketingcampagnes, zijn deze Machine Learning modellen toepasbaar:

1. K-means
2. Logistische regressie
3. SVM

De reden waarom k-means is gekozen is omdat het algoritme simpel te begrijpen en makkelijk te implementeren is. Daarnaast is de snelheid van clusteren belangrijk, omdat het te maken heeft met een grote set aan data met in acht neming van de mogelijke data lake, dat ingericht kan worden. Het is overigens ook duidelijk hoeveel clusters er toegepast zullen worden bij het segmenteren van klanten. Decision tree zal gebruikt worden doordat in eerste instantie waardes uit de bestaande dataset gevonden zullen worden, waarbij a.d.v. hypothesis de output interpreteerbaar moet zijn. Dit is tevens een sterke punt van decision trees. Daarnaast kan de decision tree goed om gaan met geheugen gebruik en dit zou in het geval van de supermarktketen het geval kunnen zijn. Uiteraard kan het bedrijf investeren in een groot Cloud opslag omgeving, maar het is verstandig om ervan uit te gaan dat de klant in kwestie het geheugen gebruik minimaal wil houden. Bij het beantwoorden van de hypotheses is een zeer hoge accuraatheid niet van belang omdat er eerder een trend ontdekt moet worden. De logistische regressie zal in combinatie met k-means gebruikt worden bij klant segmentatie voor het voorspellen van klant gedrag. Alhoewel de accuraatheid van SVM zeer hoog is, valt de snelheid van het leren nadelig. In paragraaf 4.6 is namelijk aangegeven dat online learning toegepast moet worden in het geval van een data lake. Hierbij is het belangrijk dat de training data niet een lange tijd in beslag neemt. Omdat Avanade de oplossing met het eco systeem van Microsoft aanbiedt, zal snelheid in mindere mate effect hebben. De verwerking van data gebeurt namelijk via de Cloud. Machine learning algoritmes die niet gebruikt zullen worden zijn:

1. Naive bayes
2. Neurale Netwerken
3. K-NN
4. Decision tree

Naive bayes kan goed omgaan met missende waardes en is vergeleken met andere classificatie modellen zeer snel. Nadeel van dit model is dat het de laagste accuraatheid heeft. Om deze reden is het geen goed model in het creëren van waardes in de des betreffende datasets. Tevens wordt naive bayes voornamelijk gebruikt bij kansberekeningen en tekst mining. In het vergelijkingstabel (paragraaf 6.2.7) heeft neurale netwerken in totaal achttien punten en is in dat lijst een model dat op basis van de gestelde kenmerken het laagst scoort. Tevens is de output van neurale netwerk moeilijk te interpreteren, waardoor er verkeerde conclusies getrokken kunnen worden indien de ervaring ontbreekt. K-NN volgt na neurale netwerk in de vergelijkingstabel en is op basis van de kenmerken die voor de desbetreffende datasets van

belang zijn niet geschikt. Op basis van de huidige dataset is het ook mogelijk om de vraag voor producten te voorspellen. Daarvoor zou de lineaire regressie model toegepast kunnen worden, maar omdat dit onderzoek zich richt op het verrijken van klant profielen als input voor marketingcampagnes, zal dat model niet gebruikt worden. Decision tree daarentegen kwam als beste uit de vergelijkingstabel 6. Decision tree kan goed om gaan met categorische waardes, waardoor een tree (boom) met aftakkingen goed gevisualiseerd wordt voor makkelijke interpretatie. De type gegevens uit de verkregen datasets zullen zich voornamelijk richten op de RFM-informatie en demografische en geografische gebieden van klant segmentatie. Deze gegevens zijn in vele gevallen niet categorisch en om deze reden zal decision tree niet toepasbaar zijn.

9 Performance Machine Learning modellen

In dit hoofdstuk zullen de performances oftewel de accuraatheden van de verschillende geselecteerde modellen in kaart gebracht worden. Hierbij is het doel om duidelijkheid te geven in de verschillen en de redenen van de accuraatheden van de algoritmes.

De vierde deelvraag is als volgt:

"Wat is het verschil in performance tussen de verschillende toepasbare Machine Learning algoritmes?"

In mijn realisatie document zijn de toepassingen en transformaties omschreven. Hierbij zijn ook de keuzes onderbouwd en dienen de voorgaande hoofdstukken als fundering. Op basis van de toepassingen zijn er verschillen tussen de accuraatheden. Voor het segmenteren van klanten aan de hand van karakteristieken uit de transactie en klant dataset zijn de combinatie k-means en logistische regressie gebruikt.

ML-model	Clusters	Tune hyperparameters	Accuraatheid
Two class Logistic Regression	2	NO	93
Two class Logistic Regression	4	NO	91
Two class Logistic Regression	2	YES	93
Two class Logistic Regression	4	YES	95
Multi class Logistic Regression	2	NO	93
Multi class Logistic Regression	4	NO	80
Multi class Logistic Regression	2	YES	95
Multi class Logistic Regression	4	YES	85
Support Vector Machines (SVM)	2	NO	83
Support Vector Machine (SVM)	4	NO	68

Tabel 7 vergelijking gerealiseerde modellen

Opvallend is de accuraatheid van SVM. Bij 4 clusters kwam de accuraatheid als laagst uit de vergelijking. Dit komt doordat de SVM-algoritme in Azure binaire resultaten geeft. Zodra het aantal clusters vastgelegd worden op vier en de output via SVM maximaal uit twee cijfers kan bestaan, resulteert dat in een lagere accuraatheid. Volgens (Feature Selection for SVMs, 2001) zijn andere mogelijke redenen voor een lage accuraatheid onder andere irrelevante attributen. Bij het toepassen van de permutation feature selection operator bleken de attributen geen irrelevantie te vertonen.

9.1 Bevindingen Machine Learning feature generation

Op basis van het onderzoek van (Raquel Florez-Lopez, 2008) kwam naar voren dat een oplossing voor marktsegmentatie een combinatie van cluster en logistische regressie bleek te zijn. Bij het toepassen van de onderzochte theorieën en de combinatie van de zojuist benoemde oplossing was de vraag of dit daadwerkelijk de juiste oplossing was voor klant segmentatie. De accuraatheid van de toegepaste algoritmes waren zeer hoog. Tevens is het belangrijk om de oplossing van combinatie in algoritmes, die gevonden is op basis van desk research te toetsen middels field research binnen Avanade, hierdoor wordt er namelijk op basis van field en deskresearch een aanbeveling en toepassing gehanteerd. De interviews zijn afgenomen op kantoor van Avanade bij medewerkers van Avanade. In de bijlage zijn de interview vragen met de data engineers opgenomen.

De gehanteerde interviewtechniek bestaat uit open en gesloten vragen, waarbij het merendeel van de vragen bestaat uit open vragen. Hierbij is het van belang om door te vragen, zodat de situatie en antwoord helder wordt. Uit de verschillende soorten interviews is de ongestructureerd interview gehanteerd (Scribbr, 2017). Het voordeel hiervan is dat de onderzoeker kan doorvragen.

9.1.1 Bevinden interview 1:

De geïnterviewde persoon is werkzaam bij Avanade en heeft vijf jaar ervaring op het gebied van Analytics. De geïnterviewde heeft verschillende opdrachten gedaan, deze variëren van BI-opdrachten tot het toepassen van Machine Learning. Verschillende vraagstukken, die de geïnterviewde heeft mogen uitvoeren zijn het opzetten van de ETL-proces tot het toepassen van machine learning algoritmes ten behoeve van het analyseren van churn en risico's van klanten met betrekking tot banken. De geïnterviewde heeft verder ook ervaring met segmenteren en gebruikt daarvoor k-means. Volgens de geïnterviewde is k-means de populairste cluster algoritme, omdat dit makkelijk te implementeren is en duidelijk werkt. Dit is omdat het algoritme geen moeilijke criteria bevat, behalve dat het niet goed werkt met missende waardes.
Voor klant segmentatie raad de geïnterviewde de k-means algoritme aan. Uiteraard kan er gekeken worden naar andere cluster algoritmes, maar dit is alleen als de vraagstelling niet beantwoord kan worden met k-means. Tevens is volgens de geïnterviewde de combinatie van k-means en logistische regressie erg interessant. Je hoeft bij nieuwe klanten namelijk niet opnieuw te segmenteren, maar dit is uiteraard afhankelijk van de aantal klanten. Als het toepassen van k-means binnen kortere sessies uitgevoerd kan worden, is het toepassen van logistische regressie niet nodig. Tevens is voor het segmenteren van nieuwe klanten ook verwerkingstijd nodig, waarbij afhankelijk van performance en resultaat liever k-means gebruikt kan worden.
Daarnaast is het nadeel volgens de geïnterviewde van het combineren van data, ook wel feature generation wordt genoemd, dat het aantal clusters in de toekomst gewijzigd kan worden door bijvoorbeeld marketingcampagne doeleindes. Om de clusters te evalueren moet er buiten de componenten van Azure gekeken worden. De mogelijke elbow-point methode is namelijk niet in Azure beschikbaar, maar kan in R-studio geanalyseerd worden. Wat accuracy betreft is volgende de geïnterviewde waarden tussen de 70-90% erg goed. Tevens kan de

performance van het SVM-algoritme door een lage data sample verlaagd worden. Als laatst gaf de geïnterviewde aan dat de sterke punt van Azure, de koppeling is met andere Microsoftproducten zoals bijvoorbeeld CRM en PowerBi. Voor het aanbieden van oplossingen, die geautomatiseerd kunnen worden, zal dit altijd een groot voordeel zijn.

9.1.2 Bevindingen interview 2:

De geïnterviewde is drie jaar werkzaam bij Avanade en heeft voorheen bij Microsoft gewerkt bij afdeling service engineering. De projecten die de geïnterviewde heeft uitgevoerd zijn bij de Nationale Nederlanden. De use cases van de geïnterviewde zijn: het reduceren van churn en verhogen van cross en upselling, klant segmentatie, voorspelling op marketingcampagnes en verzekeringen. De tools die de geïnterviewde gebruikt zijn Azure SQL, PowerBi en Azure ML. Het voordeel van Azure ML is volgens de geïnterviewde het kostenmodel. Tevens is de koppeling met andere Microsoftproducten erg goed geregeld.
De favoriete use case van de geïnterviewde was het opleveren van waarde voor hypotheken. Dit is zijn favoriete use case, doordat het de meeste business waarde heeft opgeleverd. Een favoriete algoritme heeft de geïnterviewde niet. Dit komt doordat iedere algoritme geschikt is voor een specifieke situatie. Wat klantsegmentatie betreft heeft de geïnterviewde k-means toegepast en heeft het marketingmodel RFM gebruikt voor het creëren van klant waarde. Voor het vinden van de juiste clusters, gaf de geïnterviewde aan om gebruik te maken van R. Verder gaf de geïnterviewde aan dat de combinatie tussen twee algoritmes interessant zijn, maar dat, dat afhangt van de resultaten en de business oftewel de eisen vanuit de opdrachtgever.
Indien de resultaten voorspellingen moet zijn op verscheidene attributen, dan is het belangrijk om een classificatie algoritme toe te voegen. Tijdens het interview kwam naar voren dat het aantal clusters zou kunnen wijzigen, waardoor de combinatie van cluster en classificatie nadelig is. Wat classificatiemodellen betreft zijn accuracy waarden tussen de 70 en 90 procent goed, tenzij er tuning heeft plaatsgevonden. Als dat het geval is, zijn accuracy waardes tussen 90 en 95 procent correct.

9.2 Conclusie

Volgens de gevolgde analytics bootcamp training zal in praktijk de beschreven performances en daarmee accuraatheden van theoretische algoritmes vaak anders zijn. Dit komt door de datasets, die geoptimaliseerd moeten worden. Tevens is field research uitgevoerd om een zo realistische beeld te krijgen omtrent de verschillen in performances van algoritmes bij het toepassen in praktijk. De combinatie tussen desk en field resulteert in een zo correct mogelijke conclusie.

Aan de hand van de afgenomen interviews bij zeer ervaren medewerkers van Avanade en aan een deelgenomen training, komt naar voren dat voor klant segmentatie k-means wordt gebruikt. K-means is de populairste algoritme voor clusteren, doordat het geen moeilijke criteria bevat en makkelijk te implementeren is. Tevens is het RFM-marketingmodel al eerder voor klant segmentatie toegepast door een van de geïnterviewden. Dit bevestigt de aanpak met het marketingmodel RFM binnen dit onderzoek met k-means ten behoeve van het creëren van klant waarde. Dit antwoord is gebaseerd op de geïnterviewden, die circa 2-3 jaar ervaring hebben met betrekking tot het toepassen van machine learning oplossingen. Beide geïnterviewde geven overigens aan dat de combinatie van twee algoritmes ten behoeve van het resultaat van toegevoegde waarde kan zijn. De geïnterviewden geven aan dat dit uiteindelijk afhangt van de resultaten en de business eisen. Voor klantsegmentatie blijkt het clusteren met k-means de correcte keuze. Indien er gekozen zou worden voor de combinatie van logistische regressie, is het belangrijk om bepaalde vraagstukken goed te evalueren.

1. Wat is de hoeveelheid van nieuwe klanten, dat middels een tweede algoritme en in dit geval logistische regressie uitgevoerd kan worden
2. Hoe groot is de kans dat de segmenten toegevoegd, verwijderd of aangepast moeten worden voor bijvoorbeeld nieuwe marketingcampagnes.
3. Wat is het verschil tussen het opnieuw uitvoeren van logistische regressie op nieuwe klanten of k-means voor het clusteren van nieuwe klanten.
4. In hoeverre is voorspellen belangrijk i.v.m. de gewenste output (resultaten).

Door het beantwoorden van deze vragen komen we tot de conclusie, dat het toepassen van k-means de correcte keuze is binnen dit onderzoek en het realiseren van waarde ten behoeve van het verrijken van klantprofielen als input voor marketingcampagnes. De hoeveelheid aan nieuwe klanten is niet bekend en er is kans dat segmenten gewijzigd, verwijderd of toegevoegd moeten worden. Tevens is er geen verschil bij het toepassen van beide situaties (k-means + logistisch en k-means alleen of logistische regressie alleen). Hierdoor is het belangrijkste welke vorm de beste resultaten oplevert en in dit geval is dat door het toepassen van k-means zonder enkele combinatie. Overigens is het binnen de scope van dit onderzoek, de bedoeling om waarde te creëren voor het verrijken van klant profielen als input voor marketingcampagnes en daarbij is een voorspellende factor niet belangrijk. Om deze redenen is er afgeweken van de onderzochte oplossing volgens het onderzoek van (Raquel Florez-Lopez, 2008). Wat de accuraatheden betreft komen deze overeen met de bevindingen uit de interviews. De accuracy zitten namelijk tussen de 70-90 procent, waarbij de geoptimaliseerde algoritmes tussen de 90 en 95 procent liggen.

10 Verrijking van klantprofielen

In dit hoofdstuk zal er getracht worden de meerwaarde van de gecreëerde gegevens voor klantprofielen duidelijk en concreet te maken, zodat op basis van deze informatie marketingcampagnes bedacht kunnen worden voor het verhogen van het verkoopvolume, maar ook de klantervaring.

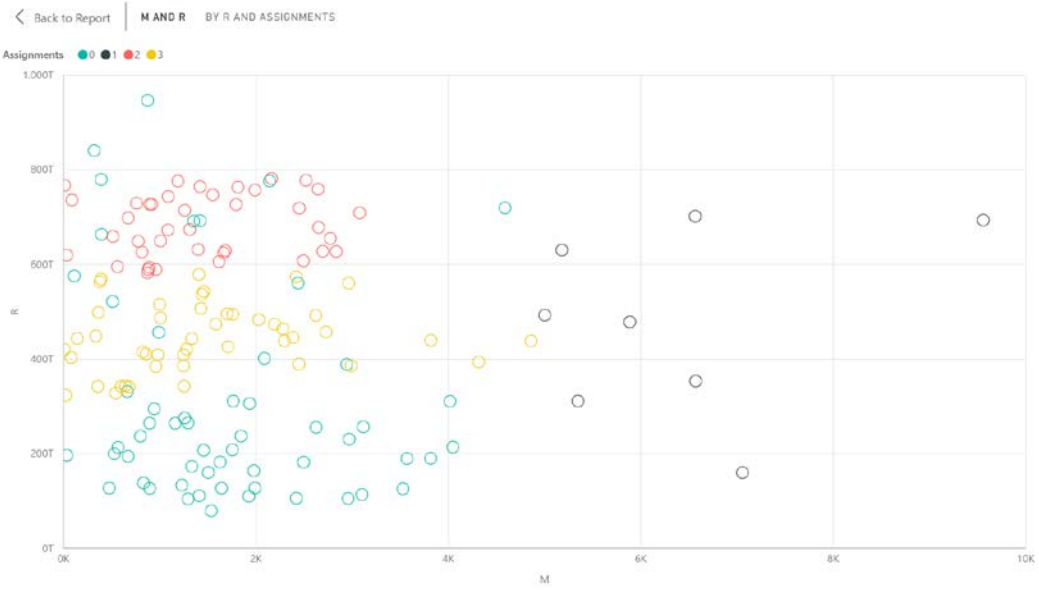
De deelvraag luidt als volgt:

"Hoe kunnen de gecreëerde resultaten meerwaarde bieden voor marketingcampagnes?"

Om klantprofielen te verrijken is op basis van machine learning waarde gecreëerd. Voor het creëren van waarde is in het model het RFM marketingmodel toegepast. Zoals in hoofdstuk 4 staat omschreven is dat een marketingmodel om klant waarde te berekenen. Dit is tot stand gekomen door RFM samen met andere attributen waaronder: inkomen, educatie, geslacht, leeftijd, burgerlijke staat, postcode, provincie, voorkeur verzendmethode, en betalingstype te analyseren en clusteren. Deze attributen vallen onder de drie onderdelen van klant segmentatie, namelijk: geografisch, demografisch en gedrag. Het proces voor het creëren van de resultaten op basis van de benoemde attributen kan bestudeerd worden in het realisatie document.

10.1 RFM Cluster resultaten

De onderstaand figuur illustreert de vier clusters, die op basis van de benoemde attributen tot stand zijn gekomen. Aan de hand van de informatie uit de clusters zal er verschillende marketingcampagnes toegepast moeten worden voor het verhogen van het verkoopvolume en verbeteren van klantervaring. Dit zal onder verschillende sub paragrafen onderverdeeld worden. Maar eerst worden de details en klanten uit de clusters verduidelijkt, zodat de marketingcampagnes begrijpbaar zijn.



Figuur 35. Klant segmentatie - Clusters

10.1.1 Inkomen per cluster

Door te kijken naar de gemiddelde inkomen per cluster, kan er geanalyseerd worden wat de redenen zijn van de RFM-cluster resultaten.

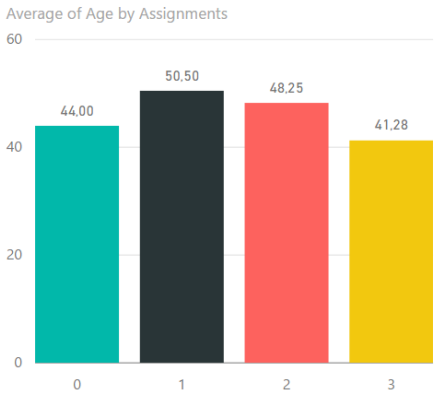
Figuur 36. Inkomen per cluster

Assignments	Average of Income
0	455.765.542.240.757,25
1	640.195.586.760.281,00
2	527.736.208.625.878,00
3	569.439.841.262.919,13

Aan de hand van de visualisatie en de daarbij horende inkomen per cluster, is te zien dat cluster één een hoger inkomen heeft dan de andere clusters. Dit resulteert volgens analyse, dat klanten met een hoger inkomen, meer opleveren voor de grote supermarktketen. Tevens is te zien, dat klanten die recentelijk een aankoop verricht hebben, vaker een aankoop doen. Dit is een cluster, waarbij het inkomen het laagst zijn.

10.1.2 Leeftijd per cluster

Zoals in figuur 37 wordt afgebeeld, zijn de leeftijden per cluster verschillend van elkaar. De leeftijden bieden meerwaarde voor doelgerichte campagnes, waarvan producten of aanbiedingen leeftijdgebonden zijn.

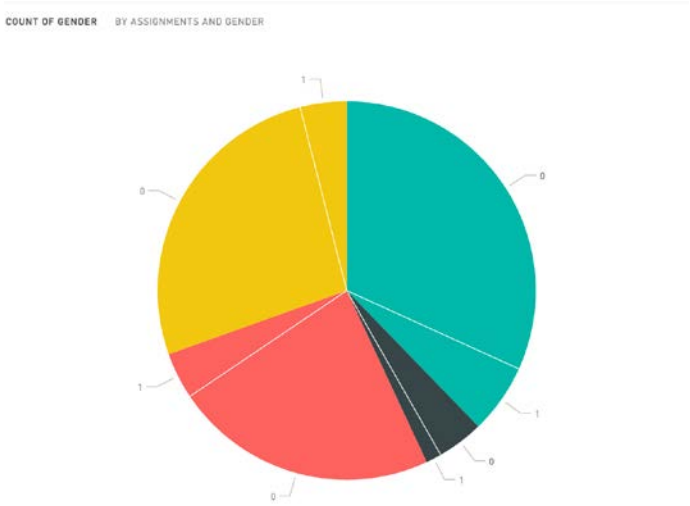


Figuur 37. Leeftijd per cluster

10.1.3 Geslacht per cluster

Door het geslacht per cluster te analyseren, kan er verschillende productcampagnes gehanteerd worden voor mannen of vrouwen. De onderstaand figuur (38) illustreert de geslacht aantallen per cluster. De klanten van de grote supermarktketen, waarvan de data door mij verkregen is, bestaat voornamelijk uit vrouwen. Dit was een attribuut, waarvan een aantal regels missend waren. Hiervoor is een functie toegepast, dat het gemiddelde neemt en die invoert voor de missende regels. Tevens zijn de alfabetische benamingen getransformeerd naar getallen. Dit was een nodige transformatie voor de k-means algoritme. Het proces is verder te bestuderen in het realisatie document.

Vrouw = 0
Man = 1



Figuur 38. Geslacht per cluster

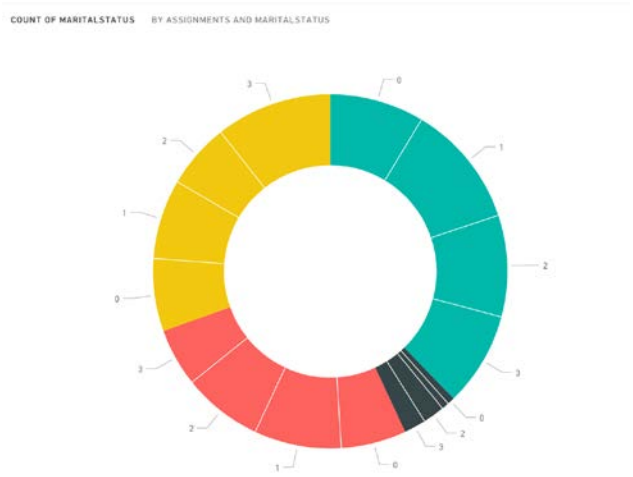
Assignments	0	1
0	48	9
1	6	2
2	34	6
3	40	6

10.2 Burgerlijke staat per cluster

De burgerlijke staat was een onderdeel uit de verzameling van attributen voor het clusteren van klanten. Aan de hand van deze informatie kan er in combinatie met een ander attribuut gekozen worden voor een marketingstrategie. Klanten die getrouwd zijn, zullen vaker voor bepaalde producten gaan i.p.v. alleenstaande klanten. Onderzoek naar welke type producten hiervoor van toepassing kunnen zijn staan buiten de scope van dit onderzoek. Desalniettemin is de informatie omtrent burgerlijke staat per cluster een belangrijke asset voor marketingcampagnes.

De alfabetische benaming voor burgerlijke staat heeft transformatie nodig gehad voor het verwerken met de k-means algoritme.

Undisclosed = 0
Single = 1
Married = 2
Divorced = 3



Figuur 39. Burgerlijke staat per cluster

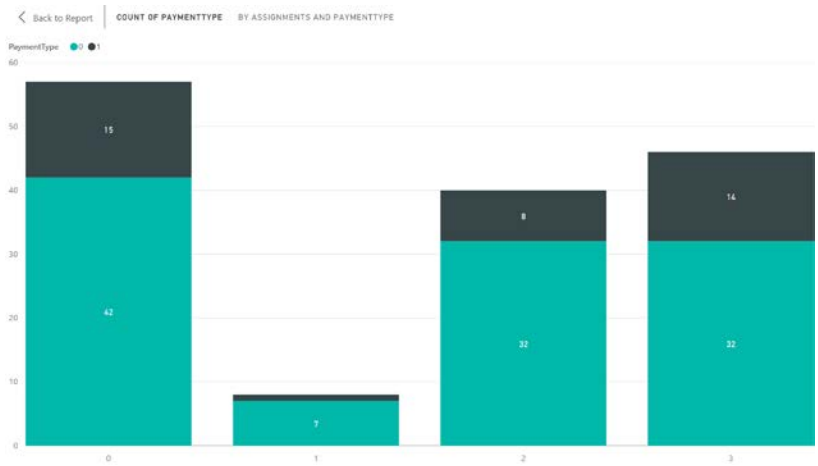
Assignments	0	1	2	3
0	13	17	14	13
1	1	1	3	3
2	9	12	11	8
3	10	11	9	16

10.2.1 Betalingsvoorkeur per cluster

Door het verzamelen van de betalingsvoorkeur is het mogelijk om deze per cluster te onderverdelen. Hierbij wordt er een verschil gemaakt in cash of creditcard. Dit waren namelijk de twee betalingsmogelijkheden, die zich in de dataset bevonden. De toegevoegde waarde hiervan is het kunnen aanbieden van aanbiedingen van internetproducten, die niet op voorraad zijn in de lokale winkels, waardoor alleen clusters oftewel segmenten benaderd kunnen worden, die voornamelijk met creditcard hebben betaald. Dit maakt de mogelijkheid voor een aankoop groter.

In het geval van betalingsvoorkeur is er transformatie plaatsgevonden om zodoende deze informatie mee te nemen voor het clusteren van klanten. De transformatie en daarmee de veranderingen in de benamingen naar getallen zijn als volgt:

Creditcard = 0
Cash = 1



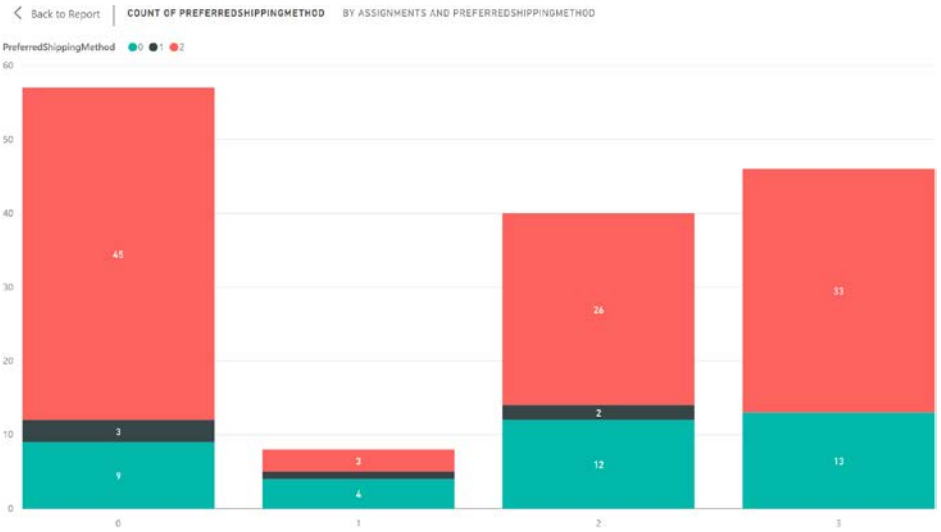
Figuur 40. Betalingsvoorkeur per cluster

10.2.3 Verzendingsvoorkeur per cluster

Het doel voor het overzetten van dit attribuut uit de klanten dataset naar de transactie dataset, is om de clusters zoveel mogelijk karakteristieken met elkaar te vormen. Tevens is dit een waardevolle informatie per clusters, doordat er per cluster gekeken kan worden naar de voorkeuren van de verzendingsvoorkeur.

Voor dit attribuut is transformatie toegepast, waarbij alfabetische waarden veranderd zijn naar numerieke getallen. Dit is gedaan voor het verwerken van de informatie met de machine learning algoritme k-means. De transformaties zijn als volgt:

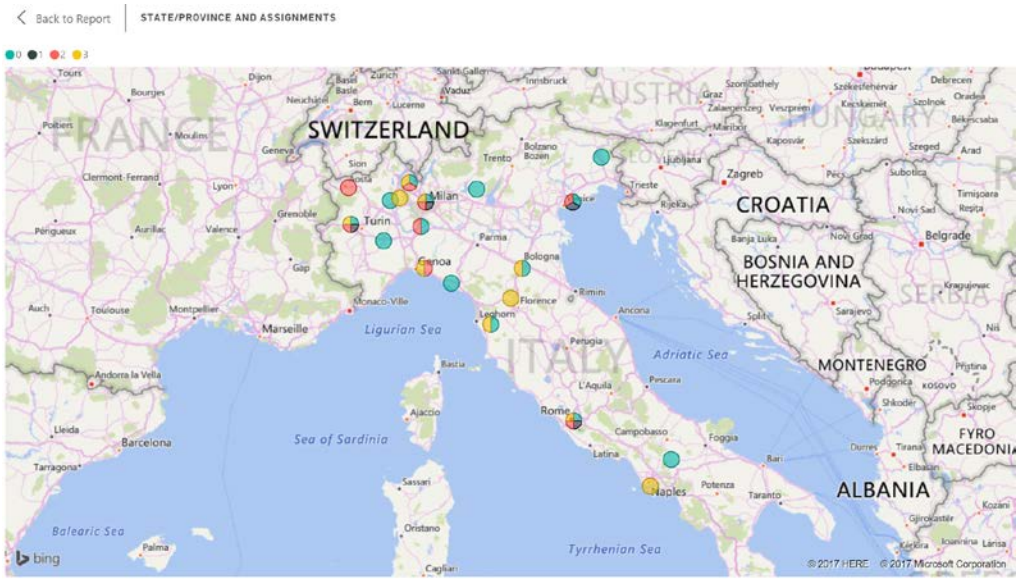
Delivery = 0
PickUp = 1
Shipping = 2



Figuur 41. Verzendingsvoorkeur per cluster

10.2.5 Locatie per cluster

De informatie en attributen zijn afkomstig van een grote supermarktketen, dat zich bevindt in Italië. Iedere transactie en daarmee klant hebben een postcode en provincie gegevens. Figuur 42 illustreert de clusters, die zich in Italië bevinden. Met deze informatie kunnen er verscheidene marketingcampagnes op basis van geografische gecombineerd met demografisch en gedrag uitgevoerd worden.



Figuur 42. Locatie van clusters.

Het is o.a. mogelijk om op basis van de RFM-clusters de hoge monetary klanten te benadrukken. Hierdoor heeft de grote supermarktketen inzicht in de locatie van de segmenten en kan voor de specifieke segmenten gelokaliseerde aanbieden gehanteerd worden. Figuur 43 illustreert

Figuur 43. Highlight segment

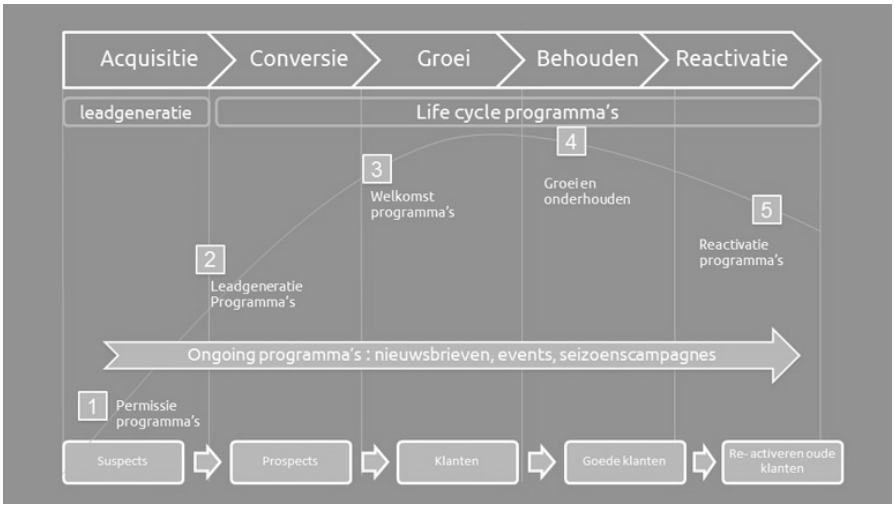
10.3 Marketingcampagnes

Voor het toepassen van marketingcampagnes wordt voor de gerealiseerde resultaten o.a. e-mail marketing segmentatie gehanteerd. E-mail marketing zorgt voor open en clickthrough rates. Volgens de data uit 2016 afkomstig van MailChimp, blijkt dat gesegmenteerde campagnes 14.64% meer open en 59.9% meer clicks genereren dan non-gesegmenteerde campagnes. (Raso, 2016). De marketingcampagnes die toegelicht zullen worden, zijn gebaseerd op twee verschillende gegevens. Deze verschillende campagnes zullen leiden naar een hoger verkoopvolume en verbeteringen in de klantervaring. De campagnes zullen onderverdeeld worden onder de volgende twee inzichten:

1. Bevindingen tijdens exploratieve analyse
2. Resultaten aan de hand van Machine learning.

10.3.1 Marketingcampagne klant registratie

Aan de hand van exploratieve analyse waren de bevindingen, dat het aantal transactie niet overeenkomt met de aantal klanten in de klanten dataset afkomstig uit CRM. Exploratieve analyse betekent het maximaliseren van inzichten, detecteren van outliers, toevoegen van extra attributen (Engineering Statistics Handbook, 2013). Uit de 3380 klanten in de transactie dataset, waren in de klanten dataset 966 klanten. Daaruit kwamen totaal 303 klanten overeen met elkaar op basis van de klant id. Voor het inzetten van marketingcampagnes kan de customer lifecycle toegepast worden. (Beelen, 2015)



Figuur 44. Customer lifecycle

De eerste stap betreft de acquisitie, waarbij de bedoeling is om suspects over te halen voor een aankoop. In dit geval is het overhalen van klanten voor het registeren. Op basis van de bevindingen doen een aantal klanten namelijk een aankoop, maar registreren zij zich niet. Door klanten te laten registeren, kan meer informatie verzameld worden, waardoor segmentatie en in een latere stadium personalisatie en daarmee target toegepast kan worden. Dit zal uiteindelijk resulteren in meer verkopen en kunnen klanten voordeel uit verschillende processen ontvangen. De voordelen voor de klant en het bedrijf in kwestie wordt in de volgende sub paragraaf toegelicht.

10.3.1.1 Voordelen marketingcampagne klant registratie

In dit sub paragraaf zullen de voordelen voor de suspects en de grote supermarktketen toegelicht worden. Door de voordelen te benoemen, wordt de keuze onderbouwd en ondersteunt.

Voordelen voor de klant:

1. *Verbeterde klantenservice* – De gegevens die bekend zijn, kan sneller gelezen worden en kan er eerder oplossingen aangeboden worden voor de klant. Het proces wordt hierdoor versneld, waardoor dit een voordeel voor de klant is.

2. *Snellere bestelproces* – Het bestellen is sneller, omdat de gegevens al bekend zijn, waardoor de klant bepaalde formulieren digitaal niet hoeft in te vullen.

3. *Aankoop geschiedenis* - De klant kan zijn bestelgeschiedenis bekijken en de producten sneller opnieuw bestellen zonder ernaar te zoeken, waardoor de klant tijd bespaart.

Voordelen voor het bedrijf:

1. *Gecentraliseerde datamanagement* - Door het verzamelen van klantgegevens kan er analyses uitgevoerd worden op klanten en kan doormiddel van email informatie targeting plaatsvinden.

2. *Verbetering marketinginspanningen* – Doordat klanten registeren met een e-mail kan het bedrijf klanten notificeren omtrent aanbiedingen voor producten. Tevens kan het bedrijf de email campagne open rate en response analyseren (Active Network, 2014).

10.3.1.2 Marketingcampagne

De marketingcampagne betreft hier een bonus systeem, waarbij klanten punten kunnen sparen voor kortingen. De klant wordt hierdoor aangetrokken doordat het een voordeel bevat, namelijk, producten kopen met korting op basis van het sparen van punten. De punten worden verdiend aan de hand van producten aankopen. De volgende figuur illustreert de campagnes die uitgevoerd kunnen worden voor klanten die zich registeren. Dit is tot stand gekomen op basis van het customer lifecycle fasen.

Acquisitie	Conversie	Groei	Behouden	Reactivatie
Klanten registratie	**Welkom mail voor nieuwe klanten**	**Activatie voor nieuwe klanten**	**Retentie en Loyaliteit**	**Klant terugwinnen**
Registratie bonus systeem	Dank mail	Herrinering notificatie	Verjaardagsmail	Openstaande winkelmand mail
	Speciale aanbieding tweede aankoop	Upsell mail	Promotie email	Aanbieding herrinering
		Crossell mail		

Figuur 45. Customer Lifecycle klant registratie

10.3.2 Marketingcampagne cluster resultaten

Zoals aangegeven is de cluster analyse gebaseerd op de RFM-marketingmodel en attributen op de drie segment gebieden. De F uit het RFM-model oftewel de frequency kan in combinatie met recency gebruikt worden voor marketingstrategieën om het verkoopvolume te verhogen. De klanten kunnen onderverdeeld worden in twee groepen. Het is belangrijk dat in beide gevallen de klanten geregistreerd zijn bij de grote supermarktketen.

Frequente kopers:
Deze groep koopt vaak producten bij de grote supermarktketen en doet dit minimaal maandelijks. Deze groep is geïnteresseerd in de producten die aangeboden zijn en heeft een hoge waarde voor de supermarktketen. Volgens (Raso, 2016) zijn er twee marketingstrategieën voor deze groep toepasbaar, namelijk:

1. Upselling van producten
2. Aanbiedingen van producten aanbieden
3. Promoten van nieuwe producten

Eenmalige klanten:
Deze groep koopt niet vaak producten bij de grote supermarktketen. De klanten binnen deze groep hebben minimaal zes maanden geen aankoop verrichten. De klanten zouden ook frequente afnemers kunnen zijn, maar een tijd niet meer actief is geweest bij de grote supermarktketen. Volgens (Raso, 2016) is een e-mailcampagne nodig om de klanten actief te krijgen. Het is belangrijk dat de mails persoonlijk zijn en dus met de desbetreffende klant naam worden aangesproken. De mails moeten het volgende bevatten:

1. Aanbiedingen op producten
2. Het benadrukken van de voordelen van de grote supermarktketen

Een bedrijf heeft een vergelijkbare email campagne uitgevoerd op basis van klant segmentatie en heeft volgens (Rueter, 2013):

1. Stijging van 278% in omzet.
2. Stijging van 183% CTR (click through ratio).
3. Stijging van 41% in open rates.

De M van monetary uit het RFM-marketingmodel weergeeft de klanten die het meest hebben opgeleverd. Een marketingstrategie die toegepast kan worden is het onderverdelen van klanten in drie niveaus. Hierbij kan de monetary, inkomen en frequency gebruikt worden. Een kledingwinkel voor vrouwen heeft een vergelijkbare strategie toegepast door de klanten te segmenteren in drie niveaus, namelijk:

1. VIP's – Klanten met een hoog inkomen en een hoge monetary.
2. Sales klanten – Klanten die gemotiveerd worden door aanbiedingen.
3. Merk klanten – Klanten die trouw zijn aan het bedrijf met een hoge frequency en lage monetary.

De VIP klanten kregen non-monetary aanbiedingen zoals exclusieve uitnodigingen, terwijl de andere twee segmenten kortingen hadden ontvangen tussen de 10 en 30 procent. Hierdoor heeft de kledingwinkel volgens (Moth, 2014) een stijging van 15 procent jaar omzet gerealiseerd.

De gecreëerde RFM marketingmodel waarde kan in combinatie met verschillende gegevens gecombineerd worden. Naast de zojuist benoemde marketingcampagne, kan ook de geografische gegevens omtrent klant bruikbaar zijn. Een bedrijf gevestigd in San Francisco, London en Boston gebruikte de geografische gegevens voor het versturen van email campagnes, waarbij de locatie voor aankopen van klanten veel invloed. Volgens (Send Targeted Messages with Geolocation, Device, and Engagement Data, 2014) heeft de campagne gezorgd voor 68% open rate, vergeleken met de 22% open rate bij algemene email campagnes.

De gegevens uit de segmenten die gebruikt kunnen worden in combinatie met de onderdelen uit het RFM marketingmodel zijn:

1. Verzendingsvoorkeur
2. Geslacht
3. Betalingsvoorkeur
4. Burgerlijke staat

Hierdoor kan het bedrijf emailcampagnes voeren, waarbij klanten met een lage recency, en frequency, of monetary waarvan het geslacht man of vrouw is en de voorkeur uit gaat naar opsturen, ophalen of levering. Het verschil tussen opsturen en levering is dat bij het opsturen het gaat om kleine producten, die op de post kunnen. Bij levering gaat het om grote producten, die niet per post verstuurd kunnen worden en door het bedrijf zelf geleverd moet worden (Difference Between Shipping and Delivery, 2013). De meerwaarde uit deze resultaten en toepassingen met de zojuist benoemde voorbeeld, is dat er campagnes op geografisch gebied plaats kan vinden. Het is belangrijk dat dit gebeurt in combinatie met de clusters uit het RFM-marketingmodel. Tevens kunnen de inactieve klanten email campagnes ontvangen, waarbij figuur 37 uit sub paragraaf 8.3.1.2 als fundering geldt.

10.4 Conclusie

Op basis van de resultaten en daarmee de klant verrijking, kan de meerwaarde voor marketingcampagnes aanbevolen worden. De clusters, die gevormd zijn, geven de recency, frequency en monetary waardes weer. De klanten doen in de verkregen dataset eenmalig een aankoop, waardoor de frequency van de klanten altijd dezelfde waarde hebben. Om deze reden wordt de frequency niet meegenomen in de conclusie. Dit is wel bruikbaar voor andere datasets, waarmee de correcte aantallen verkregen kunnen worden. Op basis van de bevindingen, zijn er een aantal campagnes onderzocht, die meerwaarde leveren voor de aanbevolen marketingcampagnes. De marketingcampagne die aanbevolen wordt is email marketing segmentatie strategieën. Het is belangrijk dat de klanten zich dus registeren, waardoor per mail specifieke campagnes gehanteerd kunnen worden. Aan de hand van de exploratieve analyse blijkt dat een deel van de klanten zich niet geregistreerd hebben. Het registeren moet aantrekkelijk zijn voor de klant. Dit moet gedaan worden aan de hand van een spaarsysteem. Door het hanteren van een spaarsysteem, kan er meer gegevens verzameld worden. Tevens is dit een belangrijke stap, waardoor de marketingcampagnes middels email marketing segmentatie gerealiseerd worden. Voor het verbeteren van de klanten ervaring is de customer lifecycle gehanteerd, waarbij per onderdeel verschillende campagnes worden toegepast ten behoeve van acquisitie, conversie, groei, behouden en reactivatie van klanten. Omdat verbetering in klantervaring niet altijd in cijfers zijn uit te drukken is d.m.v. het opsommen van voordelen voor de klant en de grote supermarktketen de keuze onderbouwd. De voordelen voor de klant omtrent het toepassen van de customer lifecycle en daarmee de campagnes zijn:

1. Verbeterde klantenservice
2. Snellere bestelproces
3. Inzicht in aankoopgeschiedenis.

De voordelen voor de grote supermarktketen zijn:

1. Gecentraliseerde datamanagement
2. Verbetering marketinginspanning

De clusters op basis van de RFM-marketingmodel vormen een leidraad voor de gehanteerde marketingcampagne. Een aanbevolen mail campagne is door de klanten onder te verdelen in frequente en non-frequente klanten. Hiervoor geldt te recency attribuut ter verrijking van klantprofiel. De marketingcampagnes die gehanteerd kunnen worden voor de frequente klanten zijn upselling van producten, aanbieden van producten en promoten van nieuwe producten. Dit zorgt voor stijgingen in de omzet, CTR en open rate. Voor non-frequente klanten is het belangrijk om de klanten weer actief te krijgen. Dit kan gedaan worden door producten aan te bieden en de sterke punten van het bedrijf te benadrukken. Dit kan gecombineerd worden met de demografische en geografische attributen uit de clusters ten behoeve van het verrichten van email segmentatie campagnes op microniveau. De RFM-waardes kunnen in combinatie met leveringsvoorkeur, geslacht, leeftijd, betalingsvoorkeur, burgerlijke staat en locatie van verkopen voor verhoging van verkoopvolume leiden. Deze campagnes zullen dus zorgen voor een hogere omzet. De fasen conversie, groei, behouden en reactivatie van klanten geldt ook voor de email marketing segmentatie.

11 Literatuurlijst

Active Network. (2014, 1 1). *The Top 10 Benefits of Online Registration for You and Your Participants*. Opgehaald van ActiveNetwork: http://www.activenetwork.sg/event-management-resources/articles/top-10-benefits-of-online-registration.htm

Amazon. (2016). *Model Fit: Underfitting vs. Overfitting*. Opgehaald van Amazon AWS: https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Andale. (2015, December 14). *RFM and Customer Value*. Opgehaald van Statisticshowto: http://www.statisticshowto.com/rfm-customer-value/

Associates, Q. (2016). *The databerg report: See what others don't - Identify the value, risk and cost of your data.* Q associates.

Avanade. (2017). *Over Avanade*. Opgehaald van Avanade: https://www.avanade.com/nl-nl/about-avanade

Avanade. (2017). *Over Avanade.* Opgehaald van Avanade: https://avanade.sharepoint.com/teams/merketingnl/_layouts/15/WopiFrame.aspx?sourcedoc=%7B6EEEFEDF-C1A5-48C7-8B17-9EB0DC2A349A%7D&file=About_Avanade_FY17_NL.pptx&action=default

Avanade. (2017). *Solutions*. Opgehaald van Avanade: https://www.avanade.com/nl-nl/solutions

Avanade Analytics Training. (2017, 4 27). *Avanade Analytics Training.* Opgehaald van Avanade Analytics Machine Learning Training.

Ayodele, T. (2016). *Types of Machine Learning Algorithms.* UK: University of Portsmouth.

Barnes, J. (2015). *Azure Machine Learning.* Microsoft.

Beelen, W. (2015, Maart 10). *Customerlifecyclecampagnes in e-mailmarketing - hoe pak je dit aan?* Opgehaald van MarketingFacts: http://www.marketingfacts.nl/berichten/customer-lifecycle-campagnes-in-e-mailmarketing-hoe-pak-je-dit-nu-aan

Bhowmik, R. (2008). *Data Mining Techniques in Fraud Detection.* Dallas: University of Texas at Dallas .

Briewald, L. (2016). How real businesses are using machine learning. *Tech Crunch*.

Center For Computational Science - University of Miami. (2009). *Decision Tree: Introduction*. Opgehaald van ccs.miami.edu: http://www.ccs.miami.edu/~hishwaran/papers/decisionTree_intro_IR2009_EMDM.pdf

Chinedu Pascal Ezenkwu, S. O. (2015). *Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services.* Akwa Ibom State: International Journal of Advanced Research in Artificial Intelligence,.

Coursera. (2017). *Machine learning*. Opgehaald van Coursera: https://www.coursera.org/learn/machine-learning

David J. Vining, G. W. (1992). *Receiver Operating Characteristic Curves: A Basic Understanding.* Louisiana: Department of Radiology, Louisiana State University Medical Center.

Deshpande, B. (2013, February 11). *How support vector machines use kernel functions to classify data*. Opgehaald van Simafore: http://www.simafore.com/blog/bid/113227/How-support-vector-machines-use-kernel-functions-to-classify-data

Deshpande, B. (2013, 1 28). *When do support vector machines trump other classification methods*. Opgehaald van Simafore: http://www.simafore.com/blog/bid/112816/when-do-support-vector-machines-trump-other-classification-methods

DifferenceBetween. (2013, July 1). *Difference Between Shipping and Delivery*. Opgehaald van DifferenceBetween: http://www.differencebetween.com/difference-between-shipping-and-vs-delivery/

Dixon, J. (2010, 10 14). *Pentaho, Hadoop, and Data Lakes*. Opgehaald van jamesdixon.wordpress: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Dull, T. (2015, September 1). *Data Lake vs Data Warehouse: Key Differences*. Opgehaald van kdnuggets: http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

EMC Education Services. (2015). *Data Science & Big Data Analytics.* Indianapolls: John Wiley & Sons, Inc.

Engineering Statistics Handbook. (2013, October 30). *What is EDA*. Opgehaald van ITL.NIST: http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm

Ensie. (2013, October 22). *Churning*. Opgehaald van Ensie: https://www.ensie.nl/aegon/churning

Etaati, L. (2016, 08 18). *What is Azure Machine Learning, why should we use it and how is it helpful?* Opgehaald van Theta: https://www.theta.co.nz/news-blogs/tech-blog/what-is-azure-ml-why-should-we-use-it-and-how-is-it-helpful

Frederick F. Reichheld, W. E. (1990, September-October). *Zero Defections: Quality Comes to Services*. Opgehaald van Harvard Business Review: https://hbr.org/1990/09/zero-defections-quality-comes-to-services

Frost, J. (2013, May 30). *Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?* Opgehaald van Minitab: http://blog.minitab.com/blog/adventures-in-

statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

Gartner. (2013, March 27). *Gartner*. Opgehaald van Forbes: http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#3e3a237f3bf6

Gartner. (2013). *Market share analysis: Business Intelligence and analytics software.* 2013: Gartner.

Gartner. (2015). *Machine learning drives digital business.* Gartner.

Gartner. (2016, August 16). *Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage*. Opgehaald van Gartner: https://www.gartner.com/newsroom/id/3412017

Gartner. (2017, 1 1). *About Gartner*. Opgehaald van gartner: http://www.gartner.com/technology/about.jsp

Ghahramani, Z. (2004). *Unsupervised Learning.* London: Gatsby Computational Neuroscience Unit.

Goeyenbier, P. (2014). *Big Data analytics: Kansen en Risico's.* Den Haag: Audit Magazine.

Gonzalez-Abril, L. (2005). *Unified dual for bi-class SVM approaches.* Sevilla: Elsevier Science.

Goyat, S. (2011). *The basis of market segmentation: a critical review of literature.* European Journal of Business and Management.

Hackerearth. (2017, february 2). *Introduction to Naive Bayes Classification Algorithm in Python and R*. Opgehaald van Hackerearth Blog: http://blog.hackerearth.com/introduction-naive-bayes-algorithm-codes-python-r

Hampton, J. (2011, February 16). *Jesshampton*. Opgehaald van SEMMA AND CRISP-DM: DATA MINING METHODOLOGIES: http://jesshampton.com/2011/02/16/semma-and-crisp-dm-data-mining-methodologies/

Hawkins, D. M. (2004). *The problem of overfitting.* Minnesota: School of Statistics, University of Minnesota.

Hsin-Yuan Huang, C.-J. L. (sd). *Linear and Kernel Classification: When to Use Which?* Taiwan: Department of Computer Science, National Taiwan University.

IBM. (2014, 1 1). *The Four V's of Big Data*. Opgehaald van IBM Big Data & Analytics Hun: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

IBM. (2015, October). *Have you seen ASUM-DM*. Opgehaald van developer.IBM: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/

IBM. (2015, August 24). *IBM Big Data & Analytics Hub*. Opgehaald van IBM: http://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

III, H. D. (2012). *A Course in Machine Learning.* CIML.

J.Weston, S. O. (2001). *Feature Selection for SVMs.* Massachusetts,: AT&T Research Laboratories.

Jerome Fan, S. U. (2006). *Understanding receiver operating characteristic.* Hamilton: Division of Emergency Medicine, McMaster University.

Kdnugget. (2014, October). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. Opgehaald van Kdnuggets: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

KDnuggets. (2017, 1 1). *How Are Precision and Recall Calculated?* Opgehaald van KDnuggets: http://www.kdnuggets.com/faq/precision-recall.html

King, K. A. (2015). *The Complete Guide to B2B Marketing.* Pearson Education .

Kotsiantis, S. B. (2007). *Supervised Machine Learning: A Review of Classification.* Tripolis: University of Peloponnese.

Lantz, B. (2015). *Machine learning with R - second edition.* Packt.

Limborgh, A. V. (2016, May 10). *Start met niet-traditionele vormen van B2B-klantsegmentatie*. Opgehaald van Marketingfacts: http://www.marketingfacts.nl/berichten/start-met-niet-traditionele-vormen-van-b2b-klantsegmentatie

Mahajan Shubhrata D, D. K. (2016). *Stock Market Prediction and Analysis Using Naïve Bayes.* Pune, India: IJRITCC.

Manas Gaur, S. G. (2015). *Comparison between Nearest Neighbours and.* New Delhi: INDIACom.

Manju Kaushik, B. M. (2014). *Comparative Study of K-Means and Hierarchical Clustering Techniques.* International Journal of Software & Hardware Research in Engineering.

MathWorks. (2016). *Applying Supervised Learning.* MathWorks.

MathWorks. (2016). *Applying Unsupervised Learning.* MathWorks.

Mitchell, T. M. (1997). *Machine Learning.* Portland: McGraw-Hill Science/Engineering/Math.

MITPress. (2006). *Semi-Supervised Learning.* London: The MIT Press.

Moth, D. (2014, Maart 20). *10 case studies that show the power of email segmentation*. Opgehaald van econsultancy: https://econsultancy.com/blog/64551-10-case-studies-that-show-the-power-of-email-segmentation/

Mutyala, S. (2011, January 3). *Using RFM to Identify Your Best Customers*. Opgehaald van Eightleaves: http://www.eightleaves.com/2011/01/using-rfm-to-identify-your-best-customers

Peterson, L. E. (2009). *Scholarpedia*. Opgehaald van K-nearest neighbor: http://www.scholarpedia.org/article/K-nearest_neighbor

Piet Verschuren, H. D. (2015). *Het ontwerpen van een onderzoek.* Amsterdam: Boom uitgevers.

Raquel Florez-Lopez, J. M.-J. (2008). *Marketing Segmentation Through: An Approach Based on Customer Relationship Management and Customer Profitability Accounting.* Spain: Sage Journals.

Raso, A. (2016, 1 1). *KissMetric Blog*. Opgehaald van 10 Quick and Easy Email Marketing Segmentation Strategies to Try Today: https://blog.kissmetrics.com/email-marketing-segmentation-strategies/

Ray, S. (2015, August 14). *7 Types Of Regression Techniques.* Opgehaald van Analyticsvidhya: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

Ray, S. (2015, 10 6). *Understanding Support Vector Machine algorithm from examples*. Opgehaald van Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/10/understaing-support-vector-machine-example-code/

Real Carbonneau, K. L. (2007). *Application of machine learning techniques.* Montreal,: Elsevier.

Reinsel, J. G. (2012). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East . 16.

Richard D. De Veaux, L. H. (2017). *A Brief Introduction to Neural Networks.* Pennsylvania: University of Pennsylvania.

Richard S. Sutton, A. G. (2012). *Reinforcement Learning:.* London: The MIT Press.

Robert Burbidge, B. B. (2011). *An Introduction to Support Vector Machines for Data.* UK: Computer Science Dept.

Roth, E. (2015, July 29). *How much data will you have in 3 years?* Opgehaald van Sisense: https://www.sisense.com/blog/much-data-will-3-years/

Rueter, T. (2013, September 26). *Apparel retailer Onward Reserve refines its e-mail targeting and gains revenue*. Opgehaald van Digitalcommerce360: https://www.digitalcommerce360.com/2013/09/26/apparel-retailer-onward-reserve-refines-its-e-mail-targeting/

SAS. (2016). *Enterprise Miner: Data exploration And Visualisation.* UK: SAS.

Saxena, R. (2016, December 23). *Dataaspirant*. Opgehaald van KNN CLASSIFIER, INTRODUCTION TO K-NEAREST NEIGHBOR ALGORITHM: http://dataaspirant.com/2016/12/23/k-nearest-neighbor-classifier-intro/

Scribbr. (2017, 1 1). *Soorten interviews*. Opgehaald van Scribbr: https://www.scribbr.nl/onderzoeksmethoden/soorten-interviews/

Sherpa Software. (sd). *Structured and Unstructured Data: What is It?* Opgehaald van Sherpa Software: http://sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/

Singh, V. (2015, April 23). *The Future Of Marketing Automation*. Opgehaald van Techcrunch: https://techcrunch.com/2015/04/23/the-future-of-marketing-automation/

Smart Vision Europe. (2000, 1 1). *About CRISP-DM*. Opgehaald van crisp-dm.eu: http://crisp-dm.eu/home/about-crisp-dm/

Smith, L. (2014, September 29). *Send Targeted Messages with Geolocation, Device, and Engagement Data*. Opgehaald van litmus: https://litmus.com/blog/send-targeted-messages-with-geolocation-device-and-engagement-data

SPSS. (1999). CRISP-DM. *Step-by-step data mining guide*, 10-11.

Srivastava, T. (2015, 1 27). *Introduction to Online Machine Learning : Simplified*. Opgehaald van analyticsvidhya: https://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/

Stanford University. (2017, 1 1). *The perceptron*. Opgehaald van Cs.Stanford.edu: https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/Neuron/index.html

Stergiou, C. (sd). *What is a Neural Network?* Opgehaald van Imperial College London: https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/cs11/article1.html

Su-Yeon Kim, T.-S. J.-H.-S. (2006). *Customer segmentation and strategy development based on customer.* South Korea: Elsevier.

Synchrony Financial. (2016). *Segmentation strategies for retailers: Leverage customer insights to drive profitable growth.* Synchrony Financial.

Toolbox. (2016, August 5). *Big Data vs. Machine Learning - To Whom Does the Future Belong?* Opgehaald van Toolbox: http://it.toolbox.com/blogs/inside-erp/big-data-vs-machine-learning-to-whom-does-the-future-belong-74085

Tu, J. V. (1996). *Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes.* Elsevier.

ujjwalkarn. (2016, August 9). *The data science blog*. Opgehaald van ujjwalkarn.me: https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/

University of Regina. (2012, 6 8). *Machine Learning*. Opgehaald van University of Regina: http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/1_ml.html

Veritas Technologies. (2017, 1 1). *Data's Dark Side*. Opgehaald van Veritas: https://www.veritas.com/dark-data.html

Vishwanathan, A. S. (2010). *Introduction to Machine Learning.* Cambridge: Cambridge University Press.

W.Thomas, J. (2016). *Market segmentation.* Decision analyst.

Wieland, E. (2014, 1 1). *Segmenteren van klanten*. Opgehaald van Salesgids: http://www.salesgids.com/marketing/segmenteren-klanten/

Wikipedia. (2017). *ID3 algorithm*. Opgehaald van Wikipedia: https://en.wikipedia.org/wiki/ID3_algorithm

Wouter Buckinx, G. V. (2007). *Predicting customer loyalty using the internal transactional database.* Ghent: Elsevier.

Xindong Wu, V. K. (2007). *Top 10 algorithms in data mining.* London: Springer-Verlag.

Zhang, H. (2004). *The Optimality of Naive Bayes.* New Brunswick: Faculty of Computer Science University of New Brunswick.

Zhang, Z. (2016). *Introduction to machine learning: k-nearest neighbours.* Jinhua, China: Department of Critical Care Medicine, Jinhua Municipal Central Hospital, Jinhua Hospital of Zhejiang University.

Zhu, X. (2007). *Semi-supervised learning survey.* Madison: University of Wisconsin.

12 Bijlage

Interview vragen

INTRO

1. Wie ben je en wat is je functie?

2. Wanneer ben je begonnen met het toepassen van machine learning?

3. Wat voor projecten heb je mogen uitvoeren?

4. Wat is je favoriete use case, die je hebt mogen uitvoeren?

ALGORITMES

5. Wat is je favoriete algoritme en waarom?

6. Wat is het meest populair cluster algoritme?

a. Waarom is deze populair?

7. Welke algoritmes zijn toepasbaar voor klant segmentatie?

TOEPASSINGEN

8. Is een combinatie van verschillende algoritmes zoals k-means en logistische regressie een betere optie, dan alleen klant segmentatie bij het voorspellen van clusters.

a. Is het verstandig om binaire gegevens als label te nemen voor voorspellingen?

EVALUATIE

9. Hoe evalueer je de kwaliteit van de clusters, die zijn gegenereerd door K-means?

10. Wat zijn goede accuracy waardes voor logistische regressie en SVM?

11. Werkt SVM goed met kleine sample data?

12. Wat is volgens jou de sterke punt van Azure?