EEN BIGDATA PROJECT MANAGEN

Module Samenvatting

Dit is de laatste module van de verkorte LOI HBO-opleiding big data specialist. Hierbij ga ik in de rol van projectmanager aan de slag met een eigen case.

Eerder opgedane kennis en vaardigheden aangaande big data, business intelligence, machine learning en ongestructureerde data spelen een centrale rol in deze case.

In vier stappen doorloop ik het project. De eerste stap is het opzetten van een big data projectplan, de tweede stap is het bepalen van het specifieke probleem dat onderzocht gaat worden, de derde stap betreft het analyseren en verzamelen van data en de vierde stap betreft de presentatie van het eindresultaat.

2 FEBRUARI 2020

JAN RUIJGROK

Inhoudsopgave

Revisieoverzicht

Datum	Versie	Status	Auteur
11-01-2020	*772Z1*	Draft ter review	Jan Ruijgrok
02-02-2020	*772Z2*	- Schrijffouten aangepakt - Voorblad/aanleiding toegevoegd - Inleiding toegevoegd - Revisieoverzicht - Inhoudsopgave toegevoegd - Opdracht 4 aangepast - Opdracht 5 aangepast - Opdracht 7 aangepast - Uitgebreid met probleemvaststelling	Jan Ruijgrok
19-03 -2020	*772Z3*	- Referentie inleiding aangepast	Jan Ruijgrok
18-04-2020	*772Z3*	- Opdrachten 9.2, 9.3, 9.5, 9.6, 9.8 en 9.9 aangepast - Lijst met afkortingen uitgebreid - Opdrachten 12 t/m 16 toegevoegd - De watervalplanning is aangepast - Alinea's 5, 6 en 9 van de inleiding aangepast - Opdracht 7 aangepast	Jan Ruijgrok
24-04-2020	*772Z4*	- Opdracht 16 aangepast - Ontdubbeling visualisaties prototype data-analyse	Jan Ruijgrok
26-04-2020	*772Z4*	- Uitwerking opdracht 17 - Uitwerking opdracht 18	Jan Ruijgrok

Lijst met afkortingen

Afkoring	Omschrijving
UAT	User Acceptance Test
WBS	Work Breakdown Structure
AI	Artificial Intelligence
PCA	Principal Component Analysis
SVM	Support Vector Machine
LR	Logistische Regressie

Inleiding

Dit document betreft de uitwerking van de laatste module van de verkorte LOI HBO-opleiding big data specialist. De laatste module staat in het teken van het managen van een bigdataproject.

Het managen van een bigdataproject bestaat uit de volgende stappen:
1. Het opzetten van een bigdataproject
2. Probleemvaststelling
3. Data verzamelen en data-analyse
4. Presentatie eindresultaat.

Iedere stap wordt in een apart hoofdstuk uitgewerkt aan de hand van opdrachten. De opdrachten zijn opgenomen in de inhoudsopgave van dit document.

In hoofdstuk 1 staat de opzet van het bigdataproject centraal. Het bigdataproject beschrijf ik met behulp van SMART-principes. Hierbij maak ik gebruik van de volgende referentie https://creately.com/blog/project-management/smart-in-project-planning/:
1. **S**pecifiek beschrijft het beoogde doel van het project.
2. **M**eetbaar benoemt de op te leveren producten.
3. **A**cceptabel benoemt de belanghebbenden.
4. **R**ealistisch beschrijft de onderkende risico's en risicobeperkende maatregelen.
5. **T**ijdgebonden beschrijft de doorlooptijd van het project.

In deze module neem ik aan dat iedere planning van een bigdataproject de volgende fasen kent: voorbereiding, increment 1 tot en met 3 en afsluiting.

In de voorbereiding wordt het projectplan opgesteld. In increment 1 staat de dataverzameling nodig voor de uitwerking van het project centraal. In increment 2 vindt de prototype van de data-analyse en dashboard plaats. In increment 3 staat het testscript centraal en levert het evaluatierapport UAT. Het project wordt afgesloten door het opsturen van de producten voortgebracht door het project, de management samenvatting en een presentatie van het resultaat aan de doelgroep en het management.

Mijn keuze is gevallen op het bigdataproject "Pilot intrusie detectiesysteem computernetwerken". Hierbij heb ik als risico onderkend dat het niet precies bekend is welke indicatoren van belang zijn om indringingen/aanvallen op computernetwerken te detecteren. De gekozen strategie is dan ook om gebruik te maken van de referentie http://faratarjome.ir/u/media/shopping_files/store-EN-1484204753-3159.pdf waarbij de relevante dataset vrij beschikbaar is gesteld door Kaggle.

In hoofdstuk 2 staat de probleemvaststelling uitgewerkt. De probleemvaststelling bestaat uit een plan van aanpak en een projectplan in Pivotal Tracker. Het bigdataproject wordt uitgevoerd met behulp van de SCRUM-methode.

Hoofdstuk 3 betreft de dataverzameling en data-analyse. De eerste stap is het omschrijven van de dataverzameling en data-analyse op de dataset. Dit geeft een idee

om welke data het precies gaat in het netwerkverkeer. De tweede stap is het ontwikkelen van een prototype data-analyse. Hierbij zet ik o.a. machine learning in en een paar slimme transformaties. De technische evaluatie betreft dan het vergelijken van het machine learning model met en zonder transformaties. De derde stap is het ontwikkelen van een dashboard met visualisaties die in samenspraak met de gebruikers tot stand is gekomen. Het dashboard heeft tot doel om mogelijke intrusie te visualiseren. De gestelde eisen van het dashboard worden getoetst aan de hand van een testscript. Het project levert een evaluatie op van de uitgevoerde gebruikersacceptatietest (UAT).

In hoofdstuk 4 presenteer ik het eindresultaat. Hierbij worden de producten opgeleverd, wordt een managementrapportage geschreven. Het project wordt afgerond met een PowerPointpresentatie aan de doelgroep en het management.

Hoofdstuk 1. Het opzetten van een bigdataproject

Opdracht 1: Reflectie op de artikelen big data

Artikel: Betekenis big data in de logistiek

Het artikel betreft een grote logistieke dienstverlener die gaten in het logistieke proces ontdekte door gebruik te maken van GPS-informatie en deze gegevens te vergelijken met de tijdstippen waarop chauffeurs de wachtknop van de boordcomputer indrukten. Hieronder volgt mijn reflectie.

Relatie tot big data. GPS-informatie vergelijken met (handmatige) input van chauffeurs.

Mogelijkheden/kansen. Starten met het analyseren van data in kleine stapjes waarna geleidelijk de introductie van big data kan plaatsvinden. Eerst door een aantal gebruiksmogelijkheden te bedenken, vervolgens de belangrijkste hiervan in de vorm van een proof of concept te realiseren en daar lering uit te trekken. Een paar gebruiksmogelijkheden zijn bijvoorbeeld:

Belangrijke informatie (zoals wachttijden, prestaties van het vervoermiddel) kunnen met behulp van sensoren en GPS informatie worden verkregen. Hierdoor kan de routeplanning worden geoptimaliseerd. Bovendien komt de informatie beschikbaar, die gebruikt kan worden om vroegtijdig storingen tijdens het vervoer te voorkomen.

Ook kan met big data een goede routeplanning worden gemaakt als bekend is welk product waar moet zijn.

Knelpunten/risico's/beren op de weg. Veel beelden in de hoofden van managers worden door data bevestigd of juist enorm ontkracht. Vooral dit laatste kan weerstand oproepen. Dat is jammer, immers een in het oog springend effect van big data zal zijn dat managers juiste beslissingen kunnen nemen die gebaseerd zijn op data en op algoritmen.

(Referentie: http://www.supplychainmagazine.nl/big-data-brengen-gaten-in-transportprocessen-aan-het-licht/)

Artikel: Big data en de noodzaak van snel datatransport

Het artikel gaat over een belangrijk gevolg van de komst van big data, namelijk de toegenomen behoefte aan technologie die data snel en betrouwbaar weet te delen. Bedrijven (vliegtuigmaatschappijen, verzekeringsmaatschappijen, Netflix) hebben een sterke behoefte om grote databestanden snel te kunnen versturen volgens het follow-the-sun principe, m.a.w. van het ene continent naar het andere continent.

Relatie tot big data. Toegenomen behoefte aan technologie als gevolg van de komst van big data.

Mogelijkheden/kansen. Software, zoals IBM Aspera, stelt bedrijven in staat om data te versturen met maximale snelheid, onafhankelijk van de bestandsgrootte, afstand of het netwerkinfrastructuur.

Knelpunten/risico's/beren op de weg. Het nadeel is dat bedrijven (bijv. Opta Sports) niet de volledige controle hebben over de bedrijfsdata. Bovendien is niet altijd duidelijk in welk land de server staat waarop de gegevens van het bedrijf worden opgeslagen.

(Referentie: https://www.smartbiz.be/blog/164741/data-delen-volgens-het-ritme-van-de-zon/)

Artikel: Toepasbaarheid van big data in de hotelsector

Het artikel begint met "Iedereen praat erover, maar niemand weet het echt": iedereen praat over big data, maar niemand weet eigenlijk welke kansen en risico's dit heeft binnen de longstay markt in de hotellerie. De longstay markt is een dynamische concept van tijdelijk wonen. Veel mensen verhuren bijvoorbeeld hun appartement als privé-accommodatie via alternatieve portals zoals Airbnb. Hieronder volgt mijn reflectie.

Relatie tot big data. Wat kunnen big data betekenen voor de longstay markt binnen de hotellerie?

Mogelijkheden/kansen. Operationele analyses voor rapportages over het functioneren van het dynamische concept van tijdelijk wonen, analyses waarmee de loyaliteit van medewerkers en online prestaties kunnen worden gemeten, voorspellende analyses om het gedrag van klanten inzake verblijf te voorspellen en correlatieanalyse om trends te ontdekken binnen de longstay markt.

Knelpunten/risico's/beren op de weg. Om hier big data in te kunnen zetten is het van belang de bestaande gegevens beter te onderhouden en minder aandacht te besteden aan de hoeveelheid gegevens dan aan de kwaliteit.

(Referentie: http://www.ahgz.de/marktdaten/big-data-ist-wie-teenage-sex,200012223194.html)

Artikel: Voorbeelden van toepassing van big data

Het artikel geeft vier mooie toepassingen van big data. De schrijver nodigt tot slot uit om na te denken over mogelijke big data applicaties. Hieronder volgt mijn reflectie.

Relatie tot big data. Voorbeelden van big data toepassingen.

Mogelijkheden/kansen. Fraudedetectiesysteem, analyse IT-logboeken, analyse van call centers en analyse van sociale media.

Knelpunten/risico's/beren op de weg. Los van de mooie toepassingen wordt er niet ingegaan hoe hierbij wordt omgegaan met persoonsgegevens.

(Referentie: http://www.ingrammicroadvisor.com/data-center/four-powerful-big-data-application-examples)

Artikel: Praktijkvoorbeelden uit de voetbalwereld laten zien wat er met data-analyse mogelijk is

Het artikel beschrijft praktijkvoorbeelden uit de voetbalwereld waarbij big data een belangrijke rol spelen. Hieronder volgt mijn reflectie.

Relatie tot big data. In de voetbalwereld is een databank opgezet met spelersprofielen met daarin uitgebreide statistieken over goals, uithoudingsvermogen, gemiddelde snelheid, gewonnen en verloren duels, gemaakte overtredingen, gele en rode kaarten (denk aan Opta Sports).

SAP heeft met Sport Solutions een totaaloplossing voor data-analyse ontwikkeld, die functioneert op basis van het In-Memory-Platform HANA.

Mogelijkheden/kansen. Big data analytics in de voetbalwereld geeft o.a. de volgende mogelijkheden en kansen: 1. een shortlist opstellen met de beste spelers in een bepaalde categorie. 2. het doorzoeken van de databank op schoten op doel, passes en kopballen. 3. het voorspellen en controleren van toekomstige ontwikkeling van spelers. 4. het voorspellen wanneer een speler weer geblesseerd dreigt te raken. 5. live-wedstrijdanalyse door het berekenen van alternatieve spelscenario's. 6. big data analytics inzetten ten behoeve van win/verlies kansen op basis van spelersprofielen.

Knelpunten/risico's/beren op de weg. Big data analytics kan ook ingezet ten behoeve van de transfermarkt. Men kan dan bijvoorbeeld voorspellende analyses doen met betrekking tot de blessuregevoeligheid van spelers.

(Referentie: http://cio.nl/big-data/89325-hoe-voetbalclubs-met-data-analytics-scoren)

Artikel: Analyse van cloud computing en de invloed van big data op bedrijven

Het artikel beschrijft de hypecyclus voor cloud computing. Hieronder volgt mijn reflectie.

Relatie tot big data. Big data bieden bedrijven binnen 2 tot 5 jaar transformatievoordelen als gebruik wordt gemaakt van cloud computing. Daarmee kunnen zij hun concurrenten met 20 % overtreffen in elke beschikbare financiële maatstaf.

Mogelijkheden/kansen. Op basis van een analyse van de Gartner Hype-cyclus voor cloud computing worden de beste resultaten bereikt door ondernemingen die zich richten op zoek zijn naar cloudgebaseerde technologieën om hun prestaties te versnellen. Gartner's nieuwste hype-cyclus voor cloud computing toont aan dat wanneer cloudgebaseerde platforms zijn afgestemd op goed gedefinieerde strategische initiatieven en bedrijfsdoelstellingen, ze waardevolle bijdragen leveren aan een onderneming. In ander woorden: bedrijven met een strategisch kader van doelen en doelstellingen verhoogt de kans op succes van een cloudgebaseerd platform volgens Gartner.

Knelpunten/risico's/beren op de weg. Bedrijven die alleen voor kostenreductie op cloudplatforms kijken, missen hun volledige potentieel aldus Gartner.

(Referentie: http://www.forbes.com/sites/louiscolumbus/2012/08/04/hype-cycle-for-cloud-computing-shows-enterprises-finding-value-in-big-data-virtualization/)

Artikel: Overschatting van big data

Wim van Dinten ontdekte dat er een verband is tussen de manier waarop mensen betekenis geven en hoe dit hun oordeel en gedrag bepaalt. Hieronder volgt mijn reflectie.

Relatie tot big data. Wim van Dinten geeft zijn visie op big data.

Mogelijkheden/kansen De auteur geeft aan dat big data (digitalisering) van nut kunnen zijn bij profiling en hiermee kom je een eind in de marketing.

Knelpunten/risico's/beren op de weg. Met big data kunnen mensen niet geholpen worden in het vinden bij betekenisgeving aan het leven. Voor ieder mens zijn de genegeerde vormen van betekenisgeving immers onmisbaar, aldus Wim van Dinten.

(Referentie: http://sezen.nl/overschatting-van-big-data/)

Artikel: Belang van data-analyse in de logistiek

Het artikel handelt over een onderzoek dat is uitgevoerd onder 200 logistieke dienstverleners wereldwijd en het gebruik van big data. Hieronder volgt mijn reflectie.

Relatie tot big data. Onderzocht is het gebruik van big data in de logistieke sector.

Mogelijkheden/kansen. 88 procent van de respondenten ziet geavanceerde analytische mogelijkheden als een 'uitstekende' of 'goede' kans voor de organisatie. Bij deze organisaties worden de mogelijkheden voor analytics uitgebreid of is dit benoemd als topprioriteit in de komende periode. Hoewel geavanceerde analytics als belangrijk wordt gezien, geeft slechts 10 procent van de logistieke organisaties aan dat zij beschikken over voldoende flexibele systemen, empowerment, data-visualisatie en supply chain risk managementfunctionaliteiten.

Knelpunten/risico's/beren op de weg. 40 procent van de logistieke bedrijven benut de mogelijkheden van data-analyse nog niet voldoende. Data-analyse wordt voornamelijk ingezet om terug te blikken op het verleden, terwijl de waarde juist ligt in de geavanceerde en voorspellende analytics.

(Referentie: http://www.ict-en-logistiek.nl/newsitem/19087)

Artikel: Big data in de sociale media

Het artikel betreft het gebruik van big data in de sociale media. Sociale media lopen voorop in de big data technologie. Hieronder volgt mijn reflectie.

Relatie tot big data. Big data vormen een cruciale hoeksteen van de meeste sociale mediabedrijven (Pinterest, Facebook).

Mogelijkheden/kansen. Er worden o.a. de volgende mogelijkheden en kansen benoemd:
1. een zoekfunctie waarmee gebruikers slechts een deel van een afbeelding kunnen

selecteren en vervolgens naar andere vergelijkbare afbeeldingen op de site kunnen zoeken (Pinterest). 2. een nieuwe functie toevoegen aan de Messenger-app, die de camerarol van je telefoon bekijkt voor eventuele foto's die je hebt gemaakt van je Facebook-vrienden en je vervolgens vraagt om ze met die vrienden te delen. 3. foto's kunnen worden geanalyseerd door 'robotalgoritmen' om ze structuur te geven - wat zit erin, welke kleur is het, waar is het genomen, wie zit erin, trekken de mensen een blij of verdrietig gezicht, enz.

Knelpunten/risico's/beren op de weg. Gezichtsherkenning en privacy zijn zeker een knelpunt. Gebruikers van sociale media kunnen zich afmelden voor gezichtsherkenning of zich aanmelden voor specifieke toepassingen, die privacy gevoelig kunnen zijn.

(Referentie: https://www.linkedin.com/pulse/pinterest-facebook-take-big-data-another-level-bernard-marr)

Opdracht 2: Vind 3 artikelen en omschrijf de mogelijkheden

Hieronder heb ik 3 artikelen benoemd, waarbij ik de reden voor de keuze van het artikel aangeef en waarom big data van toepassing (kunnen) zijn.

Splunk brengt structuur aan in rommelige digitale wereld

Reden. Mijn organisatie heeft Splunk Enterprise als databron.

Big data is van toepassing/kan van toepassing zijn. Het is duidelijk dat big data van toepassing zijn en dit is het eerste artikel die ik tegenkom waarin rommeligheid wordt genoemd in relatie tot big data.

(Referentie: https://computerworld.nl/big-data/107621-splunk-brengt-structuur-aan-in-rommelige-digitale-wereld)

Big data: de 6 V's die je moet bekijken voor belangrijke inzichten

Reden. De 6 V's zijn belangrijke kenmerken van big data en 1 of meerdere kenmerken kan/kunnen tot belangrijke inzichten leiden.

Big data is van toepassing/kan van toepassing zijn. In het artikel worden voor de kenmerken interessante big data toepassingen benoemd.

(Referentie: https://www.frankwatching.com/archive/2017/05/09/big-data-de-6-vs-die-je-moet-bekijken-voor-belangrijke-inzichten-2/)

Artikel: Hoe big data netwerken beschermen

Reden. Het artikel sluit goed aan bij de keuze van het big data project dat ik als afsluiting van deze opleiding ga doen.

Big data is van toepassing/kan van toepassing zijn. Met big data kom je meer te weten over je netwerk en kun je op die manier eerder indringers opsporen.

(Referentie: https://computerworld.nl/security/75374-hoe-big-data-netwerken-beschermt)

Opdracht 3: Geef uw definitie van big data

Big data vormen een verzameling van gestructureerde, ongestructureerde en semigestructureerd gegevens uit traditionele en digitale bronnen binnen en buiten organisaties. Big data vormen hiermee een bron voor voortdurende ontdekking, analyse en toepassingen door gebruik te maken van wiskundige modellen en technieken.

Opdracht 4: Omschrijf twee mogelijke bigdataprojecten uit uw eigen omgeving

Bigdataproject 1: Pilot energiebesparing Hoftoren

Het project heeft tot doel om in april 2020 drie maatregelen te benoemen om energie te besparen in de Hoftoren te Den Haag.

Het project levert de volgende producten op: het projectplan, datasets, prototype data-analyse, dashboard, testscript, evaluatierapport UAT, managementsamenvatting en presentatie eindresultaat.

Het project wordt uitgevoerd in opdracht van vier partijen. Voorts zijn er 19 stakeholders die belang hebben bij de uitvoering van het project. De stakeholders worden maandelijks bijgepraat over de voortgang van het project.

De volgende risico's worden onderkend: 1. het projectteam heeft maar 8 uur per week beschikbaar om aan het project te werken. 2. het projectteam heeft geen ervaring in het werken met grote hoeveelheden data.

Als risicobeperkende maatregel wordt de scope van het project beperkt tot de reeds beschreven op te leveren producten.

Het project start op 1 december 2019 en eindigt op 30 april 2020. De kick off van het project is in december. Het project wordt uitgevoerd met behulp van de SCRUM-methode: in januari wordt het dataplatform ingericht, in februari vindt de dataverzameling en -analyse plaats, in maart wordt het dashboard met visualisaties ontwikkeld en tevens een PowerPointpresentatie van de resultaten gemaakt. Tot slot wordt in april de evaluatie met aanbevelingen gepresenteerd aan de opdrachtgevers en de stakeholders waaruit een go/no go volgt voor een vervolgtraject.

Bigdataproject 2: Pilot steekproefsysteem zelfscankassa's

Het project heeft tot doel om in april 2020 een prototype van een steekproefsysteem zelfscankassa's te hebben bij de AH in Delft. Er bleken namelijk te veel frustraties te zijn bij de klanten als er weer eens een medewerker vijf artikelen ging scannen om te controleren of de klant wel goed gescand heeft.

Het project levert de volgende producten op: het projectplan, datasets, prototype data-analyse, dashboard, testscript, evaluatierapport UAT, managementsamenvatting en presentatie eindresultaat.

Het project wordt uitgevoerd in opdracht van de Raad van Bestuur van AH.

De volgende risico's worden onderkend: 1. de klant moet de boodschappen in het winkelwagentje laten zitten. 2. het projectteam heeft geen ervaring met AI-toepassingen.

Als risicobeperkende maatregel wordt een training AI verzorgd voor het projectteam en er wordt gebruik gemaakt van bestaande image recognition technieken.

Het project start op 1 februari 2019 en eindigt op 30 juni 2020. In februari is het projectplan gereed. Het project wordt uitgevoerd met behulp van de SCRUM-methode. In juni 2020 wordt besloten of er een go/no go is voor een vervolgtraject.

Opdracht 5: Omschrijf twee mogelijke bigdataprojecten in uw eigen organisatie

Bigdataproject 1: Pilot intrusiedetectiesysteem computernetwerken

Het project heeft tot doel om in april 2020 een prototype van een intrusiedetectiesysteem voor computernetwerken beschikbaar te hebben. Het systeem heeft als invoer netwerkverkeer en detecteert of er sprake is van indringing of een aanval.

Het project levert de volgende producten op: het projectplan, datasets, prototype data-analyse, dashboard, testscript, evaluatierapport UAT, managementsamenvatting en presentatie eindresultaat.

Het project wordt uitgevoerd in opdracht van het management van Netwerk Ontwikkeling.

Hierbij heb ik als risico onderkend dat het niet precies bekend is welke indicatoren van belang zijn om indringingen/aanvallen op computernetwerken te detecteren. De gekozen strategie is dan ook om gebruik te maken van de referentie https://pdfs.semanticscholar.org/1b34/8002c4ab0f632efa99e01a9b073903c5554.pdf waarbij de relevante dataset vrij beschikbaar is gesteld door Kaggle.

Als risicobeperkende maatregel wordt de scope van het project beperkt tot de reeds beschreven op te leveren producten.

Het project start op 1 februari 2020 en eindigt op 31 maart 2020. In februari is het projectplan gereed. Het project wordt uitgevoerd met behulp van de SCRUM-methode. In maart 2020 wordt besloten of er een go/no go is voor een vervolgtraject waarbij de data uit de eigen bronnen gebruikt mogen gaan worden.

Bigdataproject 2: Pilot classificatiesysteem incidenten

Het project heeft tot doel om in april 2020 een prototype van een classificatiesysteem over incidenten beschikbaar te hebben. Voor het systeem zijn 45.000 meldingen beschikbaar die te relateren zijn aan een applicatie (Outlook, Word, smartphone e.d.). Iedere melding is gerelateerd aan een applicatie en de kunst is om uit de bijbehorende teksten te kunnen achterhalen op welke applicatie deze van toepassing is. Er is een lijst van 10 meest voorkomende applicaties beschikbaar. Een melding die niet geclassificeerd kan worden heeft de klasse 'Overig'.

Het project levert de volgende producten op: het projectplan, datasets, prototype data-analyse, dashboard, testscript, evaluatierapport UAT, managementsamenvatting en presentatie eindresultaat.

Het project wordt uitgevoerd in opdracht van het management Innovatie.

De volgende risico's worden onderkend: 1. de dataset is beperkt tot 45.000 meldingen.

Als risicobeperkende maatregel wordt de scope van het project beperkt tot de reeds beschreven op te leveren producten.

Het project start op 1 februari 2020 en eindigt op 30 april 2020. In februari is het projectplan gereed. Het project wordt uitgevoerd met behulp van de SCRUM-methode. In april 2020 wordt besloten of er een go/no go is voor een vervolgtraject.

Opdracht 6: Schrijf een tekst (motivatiebrief) voor uw manager waarin u uw keuze toelicht en verkoopt

Op 3 november 2019 hebben wij in een bilateraal gesprek reeds gesproken over het toepassen van big data om bepaalde werksituaties te verbeteren, namelijk intrusiedetectiesysteem computernetwerken en classificatiesysteem incidenten .

Hiervoor heb ik twee opties aangereikt. Omdat wij verantwoordelijk zijn voor netwerkontwikkeling gaat mijn voorkeur uit naar het starten van een pilot, waarbij we een intrusiedetectiesysteem te ontwikkelen die mogelijke indringing of een aanval op ons netwerk kan identificeren. Hieronder volgt mijn motivatie.

De huidige situatie is dat wij een dataplatform hebben voor managementrapportages binnen de afdeling Netwerkontwikkeling. De rapportages gaan in het bijzonder over onze netwerkcomponenten en hoe deze met elkaar samenhangen in het netwerk. De dreiging van aanvallen op onze netwerken is altijd aanwezig. We willen de digitale veiligheid op orde hebben en dit ook inzichtelijk maken en kunnen aantonen. Dit laatste is mijns inziens nog niet het geval, omdat er nog geen rapportages zijn hierover. Derhalve is het nodig om een intrusiedetectiesysteem te hebben, dat we kunnen gebruiken voor rapportages aan managers over mogelijke indringers en aanvallen op ons netwerk.

Om dit te realiseren stel ik voor om een pilot uit te voeren, waarin een intrusiedetectiesysteem wordt ontwikkeld. Dit systeem is volledig gebaseerd op een eerdere studie ("A Study on NSL-KDD Dataset for Intrusion Detection System based on Classification Algorithms") uitgevoerd met een bestaande dataset. Na afloop van de pilot wordt besloten of er een go/no go is voor een vervolgtraject, waarbij de data uit de eigen bronnen gebruikt mogen gaan worden.

Opdracht 7: Maak een uitgebreide probleemstelling

Onze organisatie is nu niet in staat om aan te tonen dat de kwaliteit en de veiligheid van het computernetwerk is gewaarborgd. De organisatie kent veel databronnen, maar het is niet bekend welke data nu precies nodig zijn om intrusie inzichtelijk te maken. Dit

probleem ga ik oplossen door gebruik te maken van een bestaande studie. De studie heet "A Study on NSL-KDD Dataset for Intrusion Detection System based on Classification Algorithms" (https://pdfs.semanticscholar.org/1b34/8002c4ab0f632efa99e01a9b073903c5554.pdf). De studie is gebaseerd op een door Kaggle beschikbare gestelde dataset.

De eerste stap is het beschrijven van de dataverzameling en data-analyse. Dan is bekend om welke data het nu precies gaat in het netwerkverkeer. De tweede stap is het ontwikkelen van een prototype data-analyse. Hierbij zet ik o.a. machine learning in en enkele slimme transformaties. De technische evaluatie van het prototype betreft het vergelijken van het machine learning model met en zonder transformaties.

De tweede stap is het ontwikkelen van een dashboard met visualisaties die in samenspraak met de gebruikers tot stand is gekomen. Het dashboard heeft tot doel om mogelijke intrusies te visualiseren en overige relevante informatie.

Het voorgaande biedt kansen. Het is nu mogelijk gericht te zoeken naar data in de bronnen van onze organisatie, o.a. Splunk Enterprise en Netflow. Dan kan ik connectievectoren (zie artikel) ontsluiten, die invoer zijn voor het prototype en waarvan het gedrag wordt gevisualiseerd op het dashboard. Met het dashboard kan dan voor een willekeurig tijdraam mogelijke intrusies worden gevisualiseerd. Bij gebleken succes kan het prototype wellicht worden verbeterd en real-time worden ingezet.

Opdracht 8: Elevator pitch

https://youtu.be/YCv1HgWPMYU

Graag vertel ik jou over mijn project dat gaat over een intrusiedetectiesysteem voor onze netwerken. Het zal jou niet ontgaan zijn dat big data ook in onze organisatie van belang is.

We maken immers gebruik van de vele databronnen, die onze organisatie rijk is en maken daarmee ook mooie rapportages. De rapportages gaan in het bijzonder over onze netwerkcomponenten en hoe deze met elkaar samenhangen in het netwerk. Maar ze zeggen helemaal niets over mogelijke indringers en aanvallen op ons netwerk. De inzet van een intrusiedetectiesysteem in combinatie met onze big data tool Splunk kan ons hierover wat meer inzicht geven.

Het intrusiedetectiesysteem is een bestaand systeem dat is gebaseerd op een wetenschappelijke studie. Hierdoor is bekend welke data het systeem nodig heeft om indringers en aanvallen te herkennen. In combinatie met de data uit onze eigen tool Splunk Enterprise biedt het de perfecte kans om indringers en aanvallen tegen ons eigen netwerk te herkennen. Het systeem berekent namelijk de kans op een mogelijke indringer of aanval en zegt daarmee iets over de kwaliteit van het netwerk.

Met deze informatie verrijken we onze dagelijkse rapportages over de kwaliteit van ons netwerk en kunnen wij, indien nodig, ook snel maatregelen nemen om die kwaliteit te verbeteren, zodat het netwerk beter beschermd kan worden.

Vandaar dat ik mijn manager heb voorgesteld op korte termijn te starten met het maken van een prototype van een Intrusie Detectie Systeem voor onze organisatie.

Hoofdstuk 2. Probleemvaststelling

Opdracht 9: Plan van aanpak

1. Scope Opdracht

De scope van de opdracht bestaat uit:
1. Het ontwikkelen van een prototype intrusiedetectiesysteem dat in staat is om normaal netwerkverkeer te onderscheiden van abnormaal netwerkverkeer.
2. Een dashboard dat mogelijke intrusies visualiseert.

Het ontsluiten van data uit de eigen organisatie behoort niet tot de scope van het project.

2. Aannames

Het project kent aannames. Het project maakt gebruik van een bestaande representatieve dataset. Het prototype en dashboard gaan draaien op het dataplatform van de organisatie.

3. Beperkingen

Het project kent de volgende beperkingen. Het project dient uiterlijk 30 april 2020 te zijn opgeleverd en het beschikbare budget is 10000 euro. De ervaring, kennis en inzicht van het projectteam kan een belemmerende factor zijn.

Binnen de organisatie is men niet gewend Agile te werken. Derhalve is een Agile coach toegevoegd aan het team.

Het is onzeker tot welke inzichten de data-analyse gaat leiden. Bovendien is de domeinkennis beperkt.

4. Kritische voorwaarden

De kritische voorwaarden worden in hoge mate bepaald door de gebruikers. De gebruikers dienen vroegtijdig bij het project betrokken te worden waardoor het project gaat voldoen aan hun wensen en eisen ten aanzien van het opgeleverde product. Hierbij is het van belang om zorg te dragen voor een eerste prototype van een werkend product.

5. Belanghebbenden en communicatie

De belanghebbenden in de organisatie zijn het management Netwerkontwikkeling en de gebruikers van het dashboard. Daarnaast wordt er gecommuniceerd richting het projectteam, stakeholders, opdrachtgever en de projectleider.

6. Rollen en verantwoordelijkheden

Naam	Rol	Verantwoordelijkheid
Jan	Big data projectmanager	Projectleiding
Adrie	Data engineer	Data verzamelen
Mandy	Data analist	Data-analyse en Dashboard
Jan	Data scientist	Ontwikkelen prototype
Dennis	Tester	Uitvoeren UAT
Folkwin	Agile Coach	Begeleiding van het Agile proces
Klaas	Gebruiker	Testen
Piet	Gebruiker	Testen
Kees	Gebruiker	Testen

7. Mijlpalen en meten

Mijlpalen zijn producten die worden voortgebracht door het project. De op te leveren producten zijn:

1. Projectplan.
2. Datasets (train, validatie).
3. Prototype .
4. Dashboard.
5. Testscript .
6. Evaluatie rapport UAT.
7. Managementsamenvatting.
8. Presentatie eindresultaten.

8. Rapportage en afstemming

Wekelijks stelt de projectmanager zich op de hoogte van de status van de op dat moment actuele projectactiviteiten. Hij spreekt daarover met de individuele projectmedewerkers. Dit gebeurt middels een wekelijkse stand up.

In maart 2020 maakt de projectmanager de managementsamenvatting voor het management Netwerkontwikkeling. Het project wordt afgesloten met een PowerPoint-presentatie over de resultaten. Na afloop van de presentatie wordt besloten of er een go/no go is voor een vervolgtraject, waarbij de data uit de eigen bronnen gebruikt mogen gaan worden.

Het project kent diverse rapportages. Deze rapportages geven inzicht in 3 belangrijke aspecten van een project, namelijk geld, tijd en kwaliteit.

De financiële rapportage geeft inzicht in het beschikbare budget en de budgetuitputting. Een communicatieplan is handig waarin duidelijk beschreven staat hoe, wanneer en met wie wordt gecommuniceerd in termen van verantwoordelijkheden. De risicoanalyse betreft een rapportage die de onderkende risico's beschrijft en per risico de risico beperkende maatregel (en) benoemd. De voortgangsrapportage beschrijft de status van het project in relatie tot de gemaakte planning en de activiteiten. De rapportage over kwaliteit beschrijft de eisen die aan de kwaliteit van het product worden gesteld. Tot slot zijn acceptatiecriteria heel belangrijk en komen in samenspraak met de gebruikers tot stand.

Het onderstaande overzicht geeft weer voor wie de rapportages bedoeld zijn:

	Project-team	Stakeholders	Opdrachtgever	Projectleider	Gebruikers
Financiën			X		
Communicatie	X		X	X	
Risicoanalyse	X		X	X	
Voortgang	X	X	X	X	
Kwaliteit	X		X		
Acceptatiecriteria					X

9. Resultaten

Naast de op te leveren producten levert het project nog meer resultaten op. Er is ervaring opgedaan met data-analyse op een complexe dataset met behulp van wiskundige methoden en technieken. Bovendien is er kennis en inzicht opgedaan in de data in het computernetwerkverkeer die veel gebaseerd is op correlatie i.p.v. causaliteit.

Het project leidt tot aanbevelingen en conclusies waaraan kan worden gewerkt in een vervolg traject na een GO hiervoor.

Opdracht 10: Watervalplanning

Om tot schattingen te komen wordt gebruik gemaakt van de WBS methode. De schattingen zijn in uren. Hieronder volgt het overall plan van het project. In een watervalplanning wordt dit sequentieel uitgevoerd, terwijl dat in een Agile omgeving parallel mag.

Nr.	Voorbereiding: activiteit	Uren
1.1	Opzetten bigdataproject	16
1.2	Probleemvaststelling	40
1.3	Opstellen projectplan -> mijlpaal	16

Nr.	Increment 1: activiteit	Uren
2.1	Downloaden van de dataset	4
3.1	Omschrijven van de dataverzameling	10
3.2	Omschrijven van de keuzes voor de data-analyse	3
4.1	Bijstellen plan van aanpak	8

Nr.	Increment 2: activiteit	Uren
5.1	Prototype van de data-analyse	4
5.2	Technische evaluatie van de data-analyse	10
6.1	Machine Learning model -> mijlpaal	160
7.1	Prototype Dashboard Shiny	16

Nr.	Increment 3: activiteit	Uren
8.1	Testscripts opstellen in samenspraak met de gebruikers.	4
9.1	Uitvoeren UAT	8
9.2	Evaluatie UAT -> mijlpaal	8
10.1	Opleveren eindproduct -> mijlpaal	

Nr.	Afsluiting: activiteit	Uren
10.1	Managementsamenvatting > mijlpaal	8
11.1	Presentatie management	0.5
12.1	Opleveren producten door het project voortgebracht -> mijlpaal	8

Opdracht 11: Creëer een projectplan in Pivotal Tracker

Oplossing

https://www.pivotaltracker.com/projects/2426171

Hoofdstuk 3. Data verzamelen en data-analyse

Opdracht 12: Data

Omschrijving van de dataverzameling

Het project maakt gebruik van de dataset die gebruikt is in het volgende onderzoek:

- http://faratarjome.ir/u/media/shopping_files/store-EN-1484204753-3159.pdf

De relevante dataset is vrij beschikbaar gesteld door Kaggle (https://www.kaggle.com/galaxyh/kdd-cup-1999-data#kddcup.names)

Voor de gebruikte dataset geldt het volgende:

Nr.	**Eigenschap**	**Waarde**
1	Bron	Kaggle
2	Auteursrechtelijke problemen	Geen
3	Uitdagingen	Geen
4	Kwaliteit	Representatief
5	Kosten	Geen
6	Is de data compleet?	Ja
7	Is de data correct?	Ja
8	Is de data representatief?	Ja
9	Is de data courant?	Ja

De dataverzameling bestaat uit netwerkconnectie vectoren. Iedere vector bestaat uit 41 features die functioneel als volgt zijn opgebouwd:

1. Algemene features.
2. Content gerelateerde features.
3. Tijd gerelateerde 'traffic' features.
4. Host gebaseerde 'traffic' features.

Hieronder volgt een korte beschrijving van iedere feature .

De algemene features zijn:

Nr.	Feature	Omschrijving	Type
1	duration	Duur van de connectie (sec).	Continu
2	protocol_type	Type netwerk protocol (tcp, udp e.d.).	Discreet
3	service	Netwerk service bestemming (http, telnet e.d.).	Discreet
4	flag	Normale of fout status van de connectie.	Discreet
5	src_bytes	Aantal data bytes van bron naar bestemming in enkelvoudige connectie.	Continu
6	dst_bytes	Aantal data bytes van bestemming naar bron enkelvoudige connectie.	Continu
7	land	1 als bron en bestemming IP en port nummers gelijk zijn, anders 0.	Discreet
8	wrong_fragment	Aantal foute fragmenten in de connectie.	Continu
9	urgent	Aantal dringende pakketten in de connectie. Dringende pakketten zijn pakketten met de 'urgent' bit geactiveerd.	Continu

De content gerelateerde features zijn:

Nr.	Feature	Omschrijving	Type
10	hot	Aantal 'hot' indicatoren in de content zoals binnenkomen systeem directory, maken en uitvoeren van programma's.	Continu
11	num_failed_logins	Aantal foute login pogingen.	Continu
12	logged_in	1 als login is gelukt; 0 anders.	Discreet

13	num_compromised	Aantal gecompromitteerde voorwaarden.	Continu
14	root_shell	1 als root shell wordt verkregen; 0 anders.	Discreet
15	su_attempted	1 als 'su root' command wordt verkregen of gebruikt; 0 anders.	Discreet
16	num_root	Aantal 'root' toegangen of aantal operaties uitgevoerd als een 'root' in de connectie.	Continu
17	num_file_creations	Aantal bestand creatie operaties in de connectie.	Continu
18	num_shells	Aantal shell prompts.	Continu
19	num_access_files	Aantal operaties op access control files.	Continu
20	num_outbound_cmds	Aantal outbound SMDs in een ftp sessie.	Continu
21	is_hot_login	1 als de login behoort tot de hot lijst (root of admin); 0 anders.	Discreet
22	is_quest_login	1 als de login behoort tot de quest lijst; 0 anders.	Discreet

Tijd gerelateerde 'traffic' features zijn:

Nr.	**Feature**	**Omschrijving**	**Type**
23	Count	Aantal connecties naar dezelfde bestemmingshost als de huidige connectie in de laatste 2 seconden.	Continu
24	srv_count	Aantal connecties naar dezelfde service (port nummer) als de huidige connectie in de laatste 2 seconden.	Continu
25	serror_rate	Percentage connecties die de feature flag s0, s1, s2 of s3 hebben geactiveerd tussen de connecties geaggregeerd in count.	Continu
26	srv_serror_rate	Percentage connecties die de feature flag s0,	Continu

		s1, s2 of s3 hebben geactiveerd tussen de connecties geaggregeerd in srv_count.	
27	rerror_rate	Percentage connecties die de feature flag REJ hebben geactiveerd tussen de connecties geaggregeerd in count.	Continu
28	srv_rerror_rate	Percentage connecties die de feature flag REJ hebben geactiveerd tussen de connecties geaggregeerd in srv_count.	Continu
29	same_srv_rate	Percentage connecties naar dezelfde service tussen de connecties geaggregeerd in count.	Continu
30	diff_srv_rate	Percentage connecties naar verschillende service tussen de connecties geaggregeerd in count.	Continu
31	srv_diff_host_rate	Percentage connecties naar verschillende bestemmingsmachines tussen de connecties geaggregeerd in srv_count.	Continu

Host gebaseerde 'traffic' features zijn:

Nr.	**Feature**	**Omschrijving**	**Type**
32	dst_host_count	Aantal connecties die dezelfde destination host IP address hebben.	Continu
33	dst_host_srv_count	Aantal connecties die dezelfde port nummer hebben.	Continu
34	dst_host_same_srv_rate	Percentage connecties naar dezelfde service tussen de connecties geaggregeerd in dst_host_count.	Continu
35	dst_host_diff_srv_rate	Percentage connecties naar verschillende service tussen de connecties geaggregeerd in dst_host_count.	Continu
36	dst_host_same_src_port_rate	Percentage connecties naar dezelfde bron poort tussen de connecties	Continu

		geaggregeerd in dst_hst_srv_count.	
37	dst_host_srv_diff_host_rate	Percentage connecties naar verschillende bestemmingsmachines tussen de connecties geaggregeerd in dst_host_srv_count.	Continu
38	dst_host_serror_rate	Percentage connecties die de feature flag s0, s1, s2 of s3 hebben geactiveerd tussen de connecties geaggregeerd in dst_host_count.	Continu
39	dst_host_srv_serror_rate	Percentage connecties die de feature flag s0, s1, s2 of s3 hebben geactiveerd tussen de connecties geaggregeerd in dst_host_srv_count.	Continu
40	dst_host_rerror_rate	Percentage connecties die de feature flag REJ hebben geactiveerd tussen de connecties geaggregeerd in dst_host_count.	Continu
41	dst_host_srv_rerror_rate	Percentage connecties die de feature flag REJ hebben geactiveerd tussen de connecties geaggregeerd in dst_host_srv_count.	Continu

Voorts worden in de dataset een heleboel aanvalstypen benoemd die grofweg in 4 klassen kunnen worden ingedeeld (zie onderstaande tabel).

TABLE VII : MAPPING OF ATTACK CLASS WITH ATTACK TYPE

Attack Class	Attack Type
DoS	Back, Land, Neptune, Pod, Smurf,Teardrop,Apache2, Udpstorm, Processtable, Worm (10)
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (6)
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

Dit is een heel technisch verhaal. Derhalve beperk ik mij door het computernetwerkverkeer te classificeren als normaal en abnormaal.

Omschrijving van keuzes voor de data-analyse

Het uitgangspunt van de analyse is om een baseline model te ontwikkelen met behulp van machine learning om zodoende inzicht in de data te verkrijgen. Hierbij wordt gebruik gemaakt van het logistische regressiemodel. Het model rekent snel de bijdragen van de variabelen en dus hun belangrijkheid. Het model wordt gevalideerd door gebruik te maken van een validatiedataset.

De keuzes voor de data-analyse is hieronder beschreven. De concrete uitwerking ervan is terug te vinden in opdracht 14: het prototype.

De eerste stap in de analyse is het numeriseren van de dataset. Dit resulteert in 2 typen variabelen: continu en binair.

De tweede stap betreft de analyse van de continue variabelen. Een variabele met variantie 0 levert geen bijdrage aan informatie en wordt daarom ook uit de dataset verwijderd. Daarnaast wordt gekeken naar het domein van de variabelen. De meeste variabelen hebben waarden die behoren tot het domein [0.1]; er zijn 11 variabelen waarvan de waarden behoren tot het domein [0, x] met x > 1. Dus er zijn 2 groepen continue variabelen resp. groep 1 en groep 2.

De derde stap betreft het onderzoeken van de correlaties tussen de variabelen onderling. De variabelen met een correlatie coëfficiënt groter dan 0.75 worden berekend. Indien er sprake is van een sterke multicollineariteit (dat hier het geval is), kan dit een negatieve invloed hebben op het baseline model.

De vierde stap is het visualiseren van het computernetwerkverkeer in het XY- vlak met behulp van principal component analysis. De volgende inzichten kunnen hieruit worden afgeleid met betrekking tot o.a. uitschieters en clusters.

In de vijfde stap worden de groep 2 continue variabelen behandeld; variabelen zoals duration, src_bytes en dst_bytes kunnen exceptioneel groot zijn. Op de groep 2 variabelen worden daarom een 3-tal bewerkingen uitgevoerd:

- er wordt een zeer kleine waarde bij de variabele opgeteld indien deze 0 is (om logaritme te kunnen berekenen).
- de logaritme wordt berekend over de variabele, waardoor de waarden hiervan aanzienlijk kleiner worden.
- een translatie wordt uitgevoerd op de variabele zodanig dat iedere waarde weer groter of gelijk aan 0 zijn.

Voorts zijn er 2 variabelen die betrekking hebben op een enkelvoudige connectie, namelijk de src_bytes (bytes van bron naar bestemming in enkelvoudige connectie) en dst_bytes (bytes van bestemming naar bron in enkelvoudige connectie). Als de src_bytes als functie van dst_bytes worden weergegeven in het XY-vlak dan kan aan ieder punt een waarde worden toegekend, die bijvoorbeeld de afstand is van de oorsprong naar dit punt. Dus deze variabelen mogen vervangen worden door de variabele bytes, waarbij aangenomen is dat er dan geen informatie verloren gaat.

De zesde stap betreft het vinden van mogelijke uitschieters. Hiervoor is gekozen voor de visuele aanpak. Er zijn dan 13 datapunten rondom de oorsprong in 2-dimensionale visualisaties te vinden die als uitschieters zijn aan te merken. Deze datapunten worden derhalve ook niet gebruikt voor het trainen van het logistisch regressie model.

De laatste stap heeft betrekking op het trainen en valideren van het logistisch regressie model. De numerieke dataset wordt gesplitst in 70% traindata en 30% testdata. Daarnaast is een validatieset beschikbaar om het model te valideren. Op deze datasets worden de bewerkingen uit stap 5 eerst uitgevoerd. Het valideren van het getrainde model gebeurt door gebruik te maken van classificatierapporten voor de 3 datasets. De resultaten hiervan worden vergeleken zonder de bewerkingen uit stap 5.

Opdracht 13: Evalueer uw pva en pas deze zo nodig aan

Het plan van aanpak kan op een aantal punten worden verbeterd.

Normaliter wordt een watervalplanning in een MS-project gezet of Excel, zodat er een totaal overzicht komt over de doorlooptijd met betrekking tot uren, acties en actoren; binnen de organisatie zijn deze mogelijkheden helaas niet direct voorhanden.

Voorts zijn naar aanleiding van review de volgende onderdelen van opdracht 9 aangepast: 9.2, 9.3, 9.5, 9.6, 9.8, 9.9.

Opdracht 14: Prototype

De prototype van de analyse is in een r-bestand bijgevoegd.

De uitvoering van de analyse en de technische evaluatie hiervan zijn terug te vinden in het bijgevoegde HTML. De analyse is sterk gericht geweest op het komen tot een voorspellend model dat abnormaal computernetwerkverkeer kan onderscheiden van normaal computernetwerkverkeer. Het voorspellend model wordt geëvalueerd door het gedrag van het model op de 3 datasets (train- , test-, en de validatiedataset) met elkaar te vergelijken. Hierbij is ernaar gestreefd om de nauwkeurigheid en de kappa-statistiek voor de validatiedataset resp. groter dan 0.8 en 0.6 te krijgen. Dat is gelukt dankzij de uitgevoerde transformaties. Het resultaat hiervan is vergeleken met een voorspellend model, waarbij de transformaties niet zijn toegepast.

Opdracht 15: UserAcceptance

Het product bestaat uit het voorspellend model, een dashboard en de validatiedataset. Hiervoor wordt een testscript geschreven. Het testscript geeft ook meteen een goed overzicht van wat er getest moet worden en welke functionaliteiten beschikbaar dienen te zijn. Het testscript is gemaakt in Excel en als bijlage van dit document bijgevoegd.

Voordat ik met de gebruikers ga zitten, is het handig om een voorzet te geven van het dashboard. Hiervoor heb ik een eenvoudig dashboard gemaakt, dat slechts een visualisatie betreft van netwerkverkeer en inzicht geeft in de prestatie van het voorspellend model. Hierbij is de validatiedataset gebruikt. Voor de gebruikers is het relevant dat het dashboard o.a. inzicht geeft in de eigenschappen van afwijkend computernetwerkverkeer, zoals het type protocol of de betreffende service. Hierbij is het maken van een testscript erg handig, omdat dan de functionaliteit direct kan worden vastgelegd. Er is dan ook meteen een overzicht van functies die nog aandacht behoeven, die eventueel naar een vervolgtraject kunnen worden meegenomen.

Opdracht 16: Reflecteer op het analyseproces

Het hart van het product is het voorspellend model om computernetwerkverkeer te kunnen classificeren als abnormaal of normaal. De invoer voor het model zijn data uit de netwerken. Het model classificeert voorts het netwerkverkeer en aan de gebruikers wordt een dashboard getoond met informatie.

De analyse van de data heeft tot een doorbraak geleid door gebruik te maken van een paar slimme transformaties op een specifieke groep van continue variabelen. Zonder deze transformaties was het haast onmogelijk patronen in de data te verkrijgen. Deze patronen zijn gevisualiseerd voor deze specifieke groep van variabelen (zie prototype opdracht 14). De patronen zijn verder niet diepgaand onderzocht m.u.v. het bepalen van mogelijke uitschieters in de dataset.

Het trainen van het model is volgens het boekje uitgevoerd, namelijk door de dataset op te splitsen in 70% traindata en 30% testdata. De traindata bleek hier voldoende voor een mooi resultaat. Onderzoek naar uitbreiding van de traindata met datapunten uit de

testdata heeft niet plaatsgevonden. Dus ik kan nu geen uitspraak doen over het resultaat als de volledige dataset wordt beschouwd als traindata.

Een groot voordeel van dit project is nu dat er nu een goed beeld is verkregen van de data die bepalend zijn voor afwijkingen in computernetwerkverkeer. Alleen is in het analyseproces weinig aandacht besteed aan domeinspecifieke variabelen met uitzondering van de hoeveelheid bytes in een enkelvoudige connectie.

De uitdaging in een vervolgtraject is nu om de prestaties van een voorspellend model te verbeteren, dat bijvoorbeeld is staat is om computernetwerkverkeer met een nauwkeurigheid van 90% of meer goed te kunnen classificeren. Denk hierbij o.a.:

1. Cross validaties
2. Andere modellen zoals SVM, AdaBoost e.d.
3. Het oplossen van een 4-klassen classificatieprobleem.
4. Het bestuderen van domeinspecifieke variabelen.
5. Cluster analyse.
6. Het inzetten van PCA voor de continue variabelen.
7. Tunen van hyperparameters.
8. Het ontsluiten van de data uit het netwerk met behulp van Netflow en het inzetten van Hadoop om voorts deze data loslaten op het model.
9. Onderzoek naar uitbreiding van de traindata met datapunten uit de testdata.

Hoofdstuk 4. Presentatie eindresultaat

Opdracht 17: Eindopdracht

Eindproduct

Zie het bijgevoegde zip bestand LOI-772Z4 Opdracht 17 Eindproduct.zip. Hierin bevindt zich:

1. De prototype van de data-analyse
2. Het dashboard voor de gebruikers
3. De gebruikte validatiedataset
4. Het voorspellend model
5. De r source code

Management Summary

Het pilotproject intrusiedetectiesysteem computernetwerken heeft meerdere producten opgeleverd, namelijk datasets, een prototype van de data-analyse, een dashboard en gebruikersacceptatietest. Het hart van het product is een voorspellend model, waarmee afwijkend computernetwerkverkeer kan worden herkend. Hierdoor kan tijdig worden ingegrepen. Hier ligt nu een mooie kans om de kwaliteit van het eigen netwerk goed te bewaken en om adequate maatregelen te nemen de kwaliteit te verbeteren.

Data-analyse en dashboard

Het belangrijkste onderdeel van het project is een prototype van de data-analyse. Hierbij hoort een dashboard met rapportages voor de gebruikers. Het prototype van de data-analyse heeft geleid tot een bruikbaar, voorspellend model om computernetwerkverkeer te kunnen classificeren als abnormaal of normaal. De bruikbaarheid van het voorspellend model is aangetoond met een technische evaluatie, waarin resultaten met elkaar worden vergeleken.

Benchmark

Het project heeft gebruik gemaakt van een traindataset (benchmark van gesimuleerde intrusies in een militaire netwerkomgeving), die veel gebruikt wordt in onderzoek naar intrusies. Het gebruik van deze dataset heeft geleid tot kennis van het type data waarmee intrusies kunnen worden vastgesteld. Met behulp van deze traindataset is een voorspellend model ontwikkeld, dat in staat is om afwijkend netwerkverkeer te herkennen. Het model is getest met de testdataset uit een andere omgeving dan die van de traindataset. Het voordeel hiervan is deze set intrusies bevat die niet voorkomen in de traindataset.

Verkregen inzichten

Het prototype van de data-analyse heeft gebruik gemaakt van transformaties op een specifieke groep van continue variabelen. Hierdoor werden patronen zichtbaar. In het

prototype zijn deze ook gevisualiseerd. De patronen zijn verder niet diepgaand onderzocht, omwille van de tijd. Deze worden in een vervolgonderzoek meegenomen.

Voorspellend model

Het model is getraind met logistische regressie, waarbij gebruik is gemaakt van 70% van de traindata en is getest met de testdataset. De technische evaluatie hiervan is gebeurd door het onderzoeken van de nauwkeurigheid van het aantal juiste voorspellingen en de kappa-statistiek. De nauwkeurigheid is een indicatie voor het aantal correcte voorspellingen, terwijl de kappa-statistiek een indicatie is voor de voorspelkracht van het model; indien kappa > 0.6 dan duidt dit op een goede voorspelkracht. Hieronder is het belangrijkste resultaat getoond, waarbij het model getraind is respectievelijk zonder en met transformaties.

	Zonder transformaties	**Met transformaties**
Nauwkeurigheid	0.7618	0.8303
Kappa	0.5377	0.6598

Dashboard

Een ander onderdeel van het project betrof de totstandkoming van een dashboard. De invulling hiervan is afgestemd met gebruikers. Een belangrijk onderdeel van het dashboard is het visualiseren van het netwerkverkeer in een XY-vlak. Daarnaast wilden gebruikers graag bij het afwijkende netwerkverkeer een overzicht zien van de protocollen en de services waarbij die afwijkingen optreden.

Conclusie en aanbevelingen

Dit project laat zien welke data nodig zijn om intrusies te herkennen. Met het prototype van de data-analyse zijn inzichten in de structuren van de data verkregen. De uitdaging in een vervolgtraject is om de prestaties van het voorspellend model verder te verbeteren, zodat dit model in staat is om computernetwerkverkeer met een nauwkeurigheid van 90% of meer goed te kunnen classificeren. Hierbij gaat het om o.a. het inzetten van andere modellen, het bestuderen van domeinspecifieke variabelen en het toetsen van het voorspellend model met data uit het eigen netwerk.

Opdracht 18: Eindpresentatie

Slidedeck van de presentatie

Zie het bijgevoegde zip bestand LOI 772Z4 Opdracht 18 Slides Eindpresentatie .pptx.

Video van de presentatie

https://youtu.be/yyrLTOmPXBM