Vrije Universiteit

Masterthesis

Voorspellingsmodel voor treinvertragingen

Auteur:
L.P.A van der Breggen

Begeleider NS:
Kees Jong

Begeleiders VU:
Mark Hoogendoorn
Kristiaan Glorie

16 juni 2015

Vrije Universiteit

Masterthesis

Voorspellingsmodel voor treinvertragingen

L.P.A van der Breggen

VU Universiteit Amsterdam
Faculteit der Wetenschappen
De Boelelaan 1081a
1081 HV Amsterdam

Nederlandse Spoorwegen
Business Intelligence Competence Center
Laan van Puntenburg 100
3500 ER Utrecht

Begeleiders:
dr. Hoogendoorn
dr. Glorie
dr. Jong

16 juni 2015

Samenvatting

Betrouwbare reisinformatie is van belang voor de klanttevredenheid. Reizigers willen weten waar ze aan toe zijn en om deze reden bij een eventuele wijziging van hun reis op de hoogte te worden gebracht. In dit verslag is onderzocht of er een nieuw model kan worden opgesteld dat een betrouwbare voorspelling maakt van de aankomst- en vertrektijden bij een station.

Momenteel beschikt de Nederlandse Spoorwegen over een model dat deze tijden voorspelt. Dit huidige model kent zijn kracht in haar eenvoud, maar hanteert daarbij strakke aannames. Er is een vaste marge voor de inloop tijdens het rijden van een trein, en een vaste halteertijd die in een vertragingssituatie gehanteerd wordt. Een nadeel van het model is dat het uniform geldt voor alle (trein)situaties en dat het de onterechte aanname gebruikt dat een vertraging altijd zal worden ingelopen, terwijl het regelmatig voorkomt dat een vertraging toeneemt. Er is verbetering mogelijk wanneer deze restricties worden versoepeld en de specifieke context van de situatie wordt meegenomen in de voorspelling.

Er zijn twee nieuwe modellen gemaakt en geëvalueerd in dit verslag, beide kennen een sterke verbetering in nauwkeurigheid van de voorspelling ten opzichte van het huidige model. Het eerste model is het rijtijdmediaan model, deze schat de rijtijd op basis van historische data gegroepeerd op het traject, treinserie, initiële vertraging en geplande rijtijd. De tweede is een SVM model, gebaseerd op een zelflerend algoritme. Deze voorspelt de (negatieve) toename in vertraging gebaseerd op de treinserie, initiële vertraging en geplande rijtijd en is voor elk traject apart getraind en getest. Beide modellen bepalen de vertraging dus aan de hand van de specifieke situatie van de trein.

Er zijn diverse oorzaken bekeken waardoor een vertraging kan ontstaan, toenemen of verminderen, zodat deze - omgezet in attributen - kunnen worden toegevoegd aan het SVM model. Uiteindelijk zijn bovengenoemde attributen meegenomen in het model. Andere attributen welke de situatie van de trein specifieker zouden aanduiden, bleken met de gebruikte gegevens geen belangrijke indicator te zijn voor het verloop van de vertraging. Daarom is gekozen voor eenvoud.

Het resultaat is bij beide modellen aanzienlijk vaker tot 2 minuten of slechts 1 minuut nauwkeurig. Dit betekent onder andere dat overstappers betrouwbaardere informatie krijgen of ze een overstap wel of niet kunnen halen. Voor elk traject is de betrouwbaarheid bepaald en daarbij is over het algemeen te zeggen dat de voorspelling meer betrouwbaar is bij een korte geplande rijtijd en bij een kleinere initiële vertraging. Daarnaast heeft de frequentie van grote toename in vertraging op het traject een negatief effect op de betrouwbaarheid van de voorspelling.

Het SVM model valt qua betrouwbaarheid hoger uit dan de rijtijdmediaan en kent minder negatieve uitschieters in de betrouwbaarheid van trajecten dan de rijtijdmediaan. Het nadeel van het SVM model is dat de trainingstijd erg groot is en het niet duidelijk welke attributen het meest van belang zijn. Het model achter de rijtijdmediaan is daarentegen erg inzichtelijk. Beide modellen geven een sterke verbetering op de betrouwbaarheid van de aankomst- en vertrekvoorspellingen, wat mogelijk een positief effect kan hebben op de klanttevredenheid wanneer één van beide modellen wordt geïmplementeerd.

Voorwoord

Voor u ligt mijn masterscriptie welke geschreven is ter afsluiting van de master Business Analytics aan de Vrije Universiteit. In deze scriptie is mijn onderzoek vastgelegd dat ik heb uitgevoerd bij de Nederlandse Spoorwegen van januari tot en met juni 2015.

Allereerst wil ik Kees Jong, mijn begeleider bij NS, bedanken voor de hulp en advies die hij heeft gegeven. Hij maakte altijd even tijd vrij als ik een vraag had en aangezien hij naast me werkte kon ik ook mijn enthousiasme over een bepaald resultaat meteen delen. Daarnaast wil ik Karen Slijkhuis bedanken voor het bedenken van deze interessante onderzoeksvraag, zonder haar was dit onderzoek niet begonnen. Aad Smith, Peter Burghoorn en Merith Pelger wil ik bedanken voor de gesprekken en evaluaties tijdens mijn afstudeerproject, jullie enthousiasme hebben mij gestimuleerd in mijn onderzoek. Merith Pelger verdient nog een extra bedankje, omdat ze door mijn complete verslag kritisch door te lezen me erg heeft geholpen en me goede tips heeft gegeven om de laatste puntjes op de i te zetten.

Ook mijn begeleiders op de VU wil ik hier niet onbenoemd laten. Ik wil Mark Hoogendoorn bedanken voor de tijd die hij in mijn begeleiding heeft gestoken. Ik heb veel gehad aan onze (ongeveer) maandelijkse gesprekken en heb zijn kritische kijk erg gewaardeerd. Mijn tweede lezer, Kristiaan Glorie, wil ik bedanken voor het lezen en beoordelen van mijn scriptie en de aanwezigheid tijdens mijn afstudeerpresentatie.

Tot slot wil ik mijn familie en vrienden, in het bijzonder mijn vriend Geert, bedanken voor het aanhoren van al mijn verhalen over het wel en wee van mijn onderzoek. Ook mijn ouders, die zich door mijn scriptie heen hebben geworsteld om de laatste taalfouten en slordigheden te verbeteren, verdienen een bedankje op deze pagina.

Lisanne van der Breggen

Juni 2015

Inhoudsopgave

1 Inleiding

1.1 Aanleiding

Voor de Nederlandse Spoorwegen (hierna NS) is de mate van klanttevredenheid erg belangrijk. Een belangrijk aspect voor de reiziger is betrouwbare reisinformatie. Reizigers willen weten hoe laat de trein vertrekt die zij nodig hebben, hoe lang de rit duurt en of ze moeten overstappen. Op diverse manieren kunnen reizigers aan deze informatie komen: de reisplanner op de website, applicaties op smartphones en tablets en/of de schermen op de perrons.
Niet in alle gevallen rijden treinen volgens de geplande dienstregeling. In die gevallen is het wenselijk reizigers hier tijdig van op de hoogte te brengen. Door tijdig een wijziging van de tijden door te geven, weten reizigers beter waar ze aan toe zijn. Bij het laat doorgeven van een vertraging hebben bijvoorbeeld reizigers op het perron gewacht in de veronderstelling dat hun trein volgens dienstregeling zou rijden. Het komt nog voor dat een vertraging pas enkele minuten voor gepland vertrek wordt doorgegeven of dat de vertraging dan nog wordt aangepast. Dit heeft mogelijk een negatief effect op de klanttevredenheid.

1.2 Doelstelling

Momenteel wordt de vertraging van een trein bepaald met een eenvoudig model, dat mogelijk verbeterd kan worden. Het doel van dit onderzoek is om een model op te stellen dat nauwkeuriger voorspelt met welke vertraging treinen bij volgende stations zullen aankomen of vertrekken dan het huidige model.

1.3 Vraagstelling

De hoofdvraag luidt: Kan een model worden opgesteld met behulp van voorspellingstechnieken, zoals machine learning, dat een betrouwbare voorspelling maakt voor de treinvertraging bij aankomst en vertrek van een volgend station?

Er zijn hiervoor diverse deelvragen aan te wijzen die hiervoor van belang zijn:

1. *Hoe werkt en presteert het huidige model?*
   Om te zien in hoeverre een nieuw model een verbetering is op het huidige model moet onderzocht worden hoe het huidige model werkt. Daarnaast moet de prestatie van dit model worden geëvalueerd.

2. *Welke factoren spelen een rol bij het toenemen en verminderen van vertraging van een trein?*
   Om een goed beeld te krijgen van de vertragingen is het nodig om te onderzoeken waardoor vertragingen ontstaan. Vervolgens moet bekeken worden wat de ontwikkeling van die vertraging is. In welke situaties zal de vertraging verminderen en in welke toenemen?

3. *Wat is een geschikte methode om het model te evalueren?*
   Op welke manieren komt de reiziger aan reisinformatie en hoe worden vertragingen doorgegeven aan de reiziger? Wanneer is een gemaakte voorspelling van een trein betrouwbaar genoeg voor een reiziger?

4. *Wat zijn geschikte technieken om een beter model te maken?*
   Een nieuw model kan met diverse methoden worden gemaakt. Als bovenstaande vragen zijn

beantwoord kunnen modeltechnieken onderzocht worden. Zijn machine learning technieken geschikt om treinvertragingen te voorspellen?

1.4 Literatuur

Op dit onderwerp zijn eerder diverse onderzoeken naar dit onderwerp gedaan. In deze onderzoeken zijn verschillende methodes gebruikt, hieronder zal ik enkele hiervan toelichten. Ten eerste hebben Berger et all [8] een onderzoek uitgevoerd om vertragingen voor het Duitse spoornetwerk snel en betrouwbaar te voorspellen. Ze hebben een stochastisch model gemaakt waarin het spoornetwerk is weergegeven met behulp van een graaf. Er is aangenomen dat de aankomst- en vertrektijden een uniforme of unimodale verdeling hebben. Met name was het belangrijk bij dit onderzoek om de rekentijd te minimaliseren, zodat het model vaak ge-update kan worden. Kecman en Geverde [17] hebben zelf een process mining algoritme gemaakt, waarbij tevens gebruik is gemaakt van een 'timed event graph'. Dit model is erg specifiek en is enkel toegepast op het station in Den Haag. Over het algemeen is vaker gebruik gemaakt van grafen om treinvertragingen te voorspellen, zoals ook in de onderzoeken van Büker en Seybold [12], die een mesoscopisch graaf model hebben gemaakt en D'Ariano et al. [14] die een miscroscopisch graaf model hebben ontwikkeld. Aangezien dit onderzoek ingaat op machine learning technieken en niet zo zeer op stochastische en graaf modellen, zijn bovenstaande onderzoeken niet zeer relevant.

Er zijn eerder ook onderzoeken gedaan waarbij treinvertragingen zijn voorspeld met machine learning technieken, deze onderzoeken kunnen een basis vormen voor deze thesis. Zo heeft Yaghini [26] diverse machine learning technieken toegepast om op het Iraanse spoornetwerk vertragingen te voorspellen. Ze hebben gebruik gemaakt van een logistisch regressie model, een decision tree model en een neuraal netwerk met diverse verborgen lagen. Hierbij presteerde het neurale netwerk beter dan de andere twee modellen, waardoor een neuraal netwerk ook gebruikt zal worden. Het model van Yaghini heeft een beperkt aantal attributen toegevoegd, zoals het traject en de datum. Het verloop van treinvertragingen wordt regelmatig in de literatuur vergeleken met die van vliegtuig- of busvertragingen [26]. Bij het voorspellen van deze vertragingen zijn ook machine learning technieken gebruikt, zoals bij Yu et al. [27]. Zij maakten een voorspelling voor de aankomsten voor bussen, waarbij ze een neuraal netwerk en een SVM gebruikten. Het SVM model gaf een betrouwbaarder resultaat, waardoor in dit onderzoek ook de SVM wordt meegenomen. Als attributen zijn ook de tussentijd met de vorige bus en de gemiddelde rijtijd van de vorige bus meegenomen welke een positief effect hadden op de mean squared error. Voor het onderzoek naar treinvertragingen kan ook onderzocht worden in hoeverre deze attributen van belang zijn.

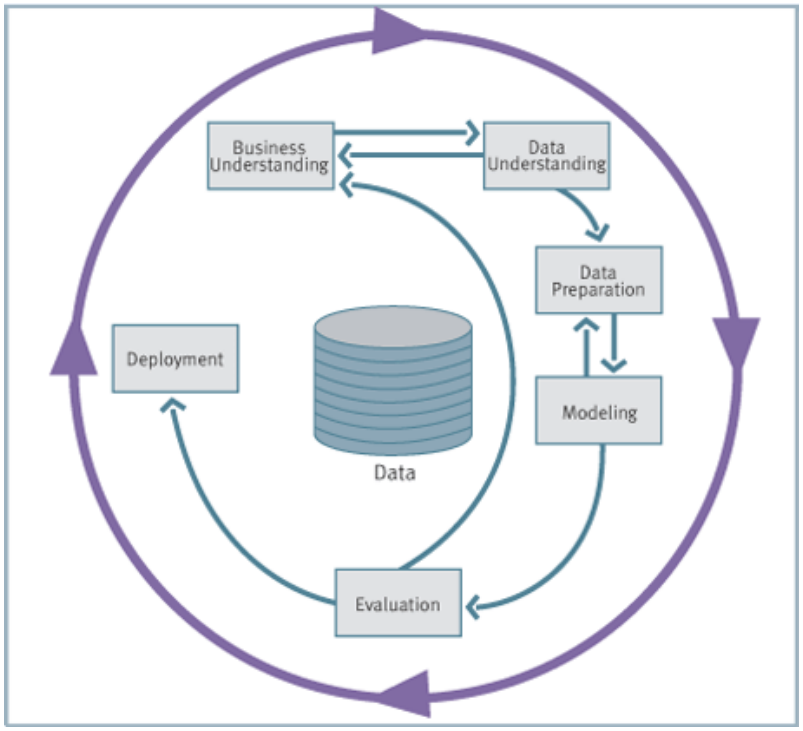
Er zijn ook onderzoeken gedaan waarbij het effect van een specifieke oorzaak van vertraging is onderzocht. Zo hebben Carey en Kwiecinski [13] een stochastische methode gebruikt om het verloop van vertraging te onderzoeken die is veroorzaakt door hinder van andere treinen. Deze oorzaak van vertraging wordt in dit onderzoek ook behandeld. Een ander onderzoek uitgevoerd door Huisman en Boucherie [16]onderzoeken hetzelfde, maar leggen hierbij de nadruk op de verschillen in snelheid tussen treinen die hinder van elkaar ondervinden. Blijkbaar is hinder van een andere trein voor eerder onderzoek een interessante factor geweest om de vertraging te voorspellen, ook in dit onderzoek zal deze factor verder worden onderzocht.

Tot slot is er een klein onderzoek bij NS uitgevoerd om een nieuw model te maken dat de vertragingen voorspelt [11]. Hierbij is onderzocht of uit de historische data af te leiden is wat de vertraging bij een volgend station is gegeven de bekende, huidige vertraging. Zo werd er per traject de mediaan bepaald van de volgende vertraging gegeven de huidige vertraging.

Bij de meeste onderzoeken is gebruik gemaakt van graafmodellen, voor dit onderzoek zijn deze minder relevant. Het onderzoek van Yaghini en Yu et al is interessant, aangezien ze machine learning technieken toepassen. Over het algemeen zijn niet veel onderzoeken op dit gebied toegepast, waardoor dit onderzoek mogelijk vernieuwende informatie kan geven over het voorspellen van de treinvertraging met machine learning technieken. De positieve algoritmen, neurale netwerken en SVM, zullen worden meegenomen evenals enkele attributen die ze hebben gebruikt in hun model.

1.5 Aanpak

De aanpak van dit onderzoek volgt de methodiek van CRISP-DM. Dit is een methodology voor een onderzoeksproces voor een KDDM-proces: Knowledge Discovery and Data Mining. Het betreft hierbij een proces dat nieuwe kennis poogt te vergaren over een domein, waarvan data mining een van de vele stappen is[18]. Er zijn verschillende methodologiën, welke over het algemeen veel op elkaar lijken[18], [7] . Bekende methoden zijn SEMMA, CRISP-DM en de methode van Cios et all. Het verschil tussen deze laatste twee is dat Cios specifiek gericht is voor academisch onderzoek, waar CRISP-DM bedoelt is voor onderzoeken in het bedrijfsleven. Vandaar dat voor dit probleem is gekozen om de methodiek van CRISP te gebruiken. CRISP-DM staat kort voor: CRoss-Industry Standard Process for Data Mining[7].
In figuur 1 zijn de verschillende stappen weergegeven, welke hier vertaald zullen worden naar dit onderzoek. De eerste stap voor dit onderzoek betreft het creëren van een heldere context van de



Figuur 1: Methodiek: CRISP (cf. [7])

situatie. Dit betekent dat wordt ingegaan op de context rondom vertragingen: enkele definities dienen verklaard te worden en mogelijk oorzaken van het ontstaan van een vertraging te worden

onderzocht. Verder dient een beeld te worden gevormd over het systeem achter de reisinformatie die wordt meegegeven aan de reiziger. Dit alles valt onder de noemer 'Business Understanding' en wordt uitgewerkt in hoofdstuk 2 *Context*.
Wanneer de context is beschreven en belangrijke begrippen zijn verklaard kan een eerste data analyse worden uitgevoerd in hoofdstuk 3 *Data analyse*. Hierbij wordt onderzocht hoe vaak vertraging voorkomt en of er al eerste indicatoren opduiken in de data die een rol spelen bij het verloop van vertraging. In dit deel wordt dus meer ingezoomd op de beschikbare data: 'Data Understanding'.
Voordat er gestart wordt met een nieuw model, wordt de probleemstelling nauwkeuriger geformuleerd in hoofdstuk 4 *Probleemafbakening*. Hierbij wordt ook ingegaan op de 'Data preparation', waaronder het onderzoek naar de kwaliteit en het nut van de data valt: welke data wordt meegenomen in een model, welke bewerkingen moeten hierbij worden gemaakt. De 'Data Preperation' is te vinden in hoofdstuk 4.5 *Data voorbereiding*. In aanvulling hierop onderzoeken we hoe het huidige model in elkaar zit en welke factoren hierin zijn meegenomen, dit is te vinden in hoofdstuk 5 *Huidig model*.
Vervolgens wordt in meer detail de methodiek besproken voor het onderzoek. Aangezien één van de deelvragen is in hoeverre machine learning technieken gebruikt kunnen worden voor dit probleem, gaat dit hoofdstuk meer in op de achtergrond over deze methoden. Welke machine learning technieken zijn voor een dergelijk probleem te gebruiken en waar kennen deze technieken hun basis? Dit gedeelte is te vinden in hoofdstuk 6 *Methodiek*.

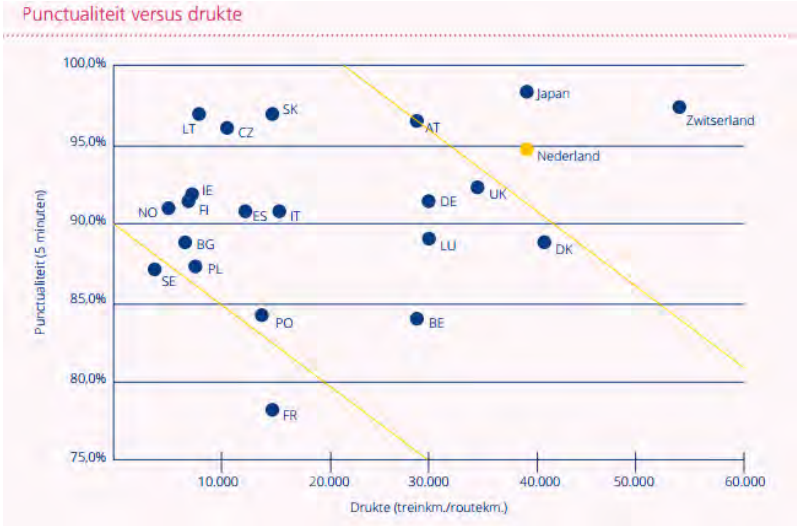
De fases 'Data Preperation', 'Modeling' en 'Evaluation' zullen - ietwat anders dan in de figuur weergegeven is - itererend worden herhaald. Na elk uitgevoerd model vindt de evaluatie plaats, waarbij geanalyseerd wordt in hoeverre het huidige model een verbetering is ten opzichte van de vorige stap. Hieruit kan een volgende hypothese voor een nieuw model volgen, waarna mogelijk opnieuw enkele aanpassingen aan de data moeten worden gedaan, waarna een nieuw model wordt uitgevoerd en geëvalueerd. De resultaten zijn terug te vinden in hoofdstukken 7 en 8.
Wanneer het model tot voldoende tevredenheid leidt, zal de stap 'Evaluation' tot in meer detail worden uitgevoerd in hoofdstuk 9 *Evaluatie* en kan de pijl naar de laatste stap gevolgd worden: 'Deployment', oftewel het in gebruik nemen van het model.

2 Context

Voordat we aan een nieuw model beginnen, is het nuttig de situatie en de achtergrond van het probleem beter in kaart te brengen. In dit hoofdstuk wordt ingegaan op de NS, enkele gebruikte begrippen en op de diverse factoren die een rol kunnen spelen bij het verloop van vertragingen. Vervolgens bespreken we het proces binnen NS waarin reisinformatie wordt ontwikkeld en gepresenteerd, het model dat eventuele vertragingen berekent valt hier uiteraard onder. Tot slot gaan we in op de reisinformatieproducten die hieruit ontstaan om te zien wat voor informatie de reiziger ontvangt. Afhankelijk van de informatiewensen van de reiziger volgt hoe het model geëvalueerd moet worden.

2.1 Nederlandse Spoorwegen

De Nederlandse Spoorwegen is de grootste vervoerder van reizigers over het treinnetwerk in Nederland. Bij de oprichting in 1917 was ook het spoornetwerk zelf van NS, maar dit is in de jaren negentig veranderd waardoor het spoornetwerk nu wordt beheerd door ProRail [2]. De NS beschikt momenteel over het hoofdrailnet, hieronder vallen alle trajecten waar binnenlandse intercity's rijden en enkele andere trajecten [6]. Het vormt het grootste deel van het Nederlandse spoornetwerk. Op andere trajecten mogen ook andere treinvervoerders treinen laten rijden. Dagelijks vervoert



Figuur 2: Punctualiteit versus drukte op spoor (cf. [6])

NS ruim 1,2 miljoen reizigers over ongeveer 4.800 treinritten [5]. Als men daarbij meeneemt dat het spoornetwerk 2100 kilometer lang is, blijkt dat het Nederlandse spoornetwerk tot een van de drukste ter wereld behoort en in Europa enkel Zwitserland een intensiever gebruikt spoor bezit [6]. Figuur 2 zet de punctualiteit van de treinen uit tegen de drukte op het spoor. Hieruit blijkt dat gezien het intensieve gebruik van het spoor het aantal verstoring minder is dan in andere delen van Europa. Japan echter heeft een even intensief gebruikt spoor maar minder verstoringen. Dit geeft aan dat er ruimte is voor verbetering.

Binnen NS vallen diverse onderdelen die elk andere verantwoordelijkheden hebben. De belang-

rijkste zijn NS Stations, NS Reizigers en NS Internationaal. De eerste houdt zich bezig met alles wat zich op de stations bevindt, waaronder alle winkels en andere voorzieningen. NS Internationaal houdt zich bezig met de internationale trajecten en onderhoudt samenwerking met internationale partners.
Dit onderzoek heeft plaatsgevonden bij de afdeling Business Intelligence Competence Center (BICC), welke valt onder ICT van NS Reizigers. Hierbinnen heeft de afdeling BICC de verantwoordelijkheid voor zowel het beheer, het ontsluiten als het rapporteren van gegevens van de NS. Vandaar dat veel data beschikbaar was op de afdeling en toegepast kon worden voor dit onderzoek.

2.2 Begrippenlijst

Er komt veel kijken bij het aanbieden van treinvervoer. Enkele begrippen zijn relevant voor het verdere model en worden hier toegelicht.

Dienstregeling

De dienstregeling geeft weer welke trein op welk tijdstip vanaf welk spoor vanuit welk station vertrekt. De dienstregeling kent in de basis een uurpatroon. Echter, afhankelijk van de dag van de week en het dagdeel kunnen er extra treinen worden toegevoegd (spits) of worden weggelaten (nachturen). De dienstregeling kent verkeersdagen die iets anders zijn dan normale dagen, aangezien alle treinen die vertrekken voor 02.00 uur nog bij de vorige dag horen.

Treintypes

Binnen de reisizgersstreinen kan onderscheid gemaakt worden tussen sprinters, intercity's en internationale treinen. Een intercity is een trein die ingezet wordt om middellange en lange trajecten en stopt enkel op (middel-)grote stations, een sprinter stopt daarnaast ook op de kleine tussengelegen stations [1]. In enkele gevallen stopt een intercity ook op een klein station. Internationale treinen, zoals de Thalys en IC Direct, stoppen op enkele grote stations en rijden ook naar het buitenland.
Een belangrijk verschil tussen de treintypes is dat bij het voorspellen van een aankomst die over 20 minuten plaats zou moeten vinden een sprinter waarschijnlijk meerdere stops in de tussentijd heeft gehad, terwijl de aankomst voor een intercity en een internationale trein met een grotere kans de eerste aankomst kan zijn. Hier wordt later verder op ingegaan.

Treinseries en treinnummers

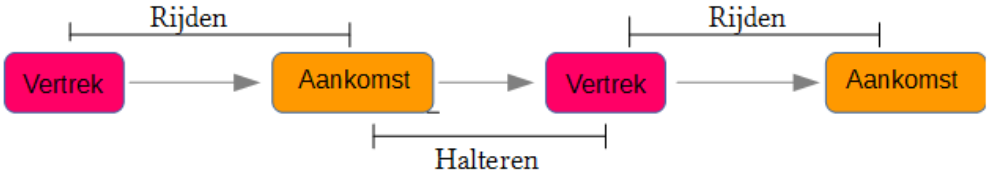
Om de volgens dienstregeling geplande treinen te classificeren en te benoemen wordt er gebruik gemaakt van treinnummers en treinseries. Eenzelfde treinserie wordt toegekend aan alle treinen die hetzelfde traject rijden en dezelfde haltes hebben, de richting mag wel verschillen. Zo behoren alle intercity's die van Groningen naar Rotterdam rijden en die van Rotterdam naar Groningen rijden tot serie 500. De richting wordt aangegeven met 'E' (even) of 'O' (oneven). De sprinter van Gronigen naar Utrecht rijdt wel op hetzelfde traject, maar stopt op andere haltes. Vandaar dat deze een ander serienummer krijgt.

Om een specifieke trein aan te duiden binnen een serie wordt een specifiek treinnummer gegeven aan elke trein. Het honderdtal van het treinnummer is (meestal) gelijk aan die van de treinserie, zodat de koppeling van trein naar treinserie aanwezig is in het treinnummer. De laatste twee cijfers worden bepaald door het tijdstip waarop de trein rijdt en loopt gedurende de dag op. Zo heeft de

intercity vanuit Rotterdam naar Groningen van 08.05 het treinnummer 527, en heeft de volgende intercity van Groningen naar Rotterdam van 08.16 het treinnumer 528. Het treinnummer is niet afhankelijk van de dag, en dus krijgt elke intercity die om 08.16 vanuit Groningen naar Utrecht gaat hetzelfde treinnummer.
Om in het model aan te duiden welke trein het betreft, kan het treinnummer worden gebruikt of er kan een ruimere classificatie worden gemaakt door enkel de treinserie mee te geven.

Rijden en halteren

Het is goed om kort uit te lichten dat een trein twee processen kent: rijden en halteren. Het halteerproces betreft het proces vanaf de aankomst bij een station tot het vertrek. Het rijproces is juist het proces vanaf het vertrek bij een station tot aan de volgende aankomst. Een treinrit ziet er als volgt uit, zie figuur 3. Om terug te komen op de treintypes: een sprinter heeft in verhouding meer halteerperiodes dan een intercity of een internationale trein.
Bij het voorspellen van een aankomst wordt het halteerproces op dat station niet meegenomen, terwijl dat bij het voorspellen van de vertrekvertraging wel wordt meegenomen. In een volgend hoofdstuk wordt ingegaan op mogelijke factoren voor vertragingen, waaruit zal blijken dat er verschillende factoren een rol spelen bij deze twee processen.



Figuur 3: Schematische weergave rijden vs halteren

Dienstregelpunten

Over het spoornetwerk heen liggen diverse dienstregelpunten. Deze dienstregelpunten liggen bij elk station, maar ook bij verschillende punten langs het spoor. Vaak bevinden die punten zich bij een bijzonderheid van het spoor, zoals een brug of een ander herkenbare plek. Bij deze punten liggen meetpunten waar wordt bijgehouden wanneer welke trein langsrijdt. Daarom zijn dienstregelpunten een belangrijke input voor het voorspellen van een treinvertraging.

Rijspeling

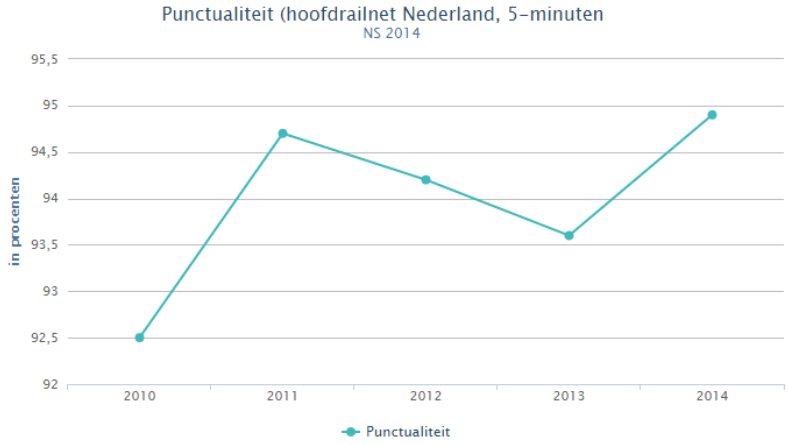
De dienstregeling is zodanig opgesteld dat er speling zit in de rijtijd tussen twee stations, waardoor een kleine verstoring kan worden opgevangen. Dit betekent dat een trein een traject sneller kan rijden dan volgens de dienstregeling wordt verwacht. Aangezien vanuit de dienstregeling bekend is hoe laat een trein langs elk dienstregelpunt behoort te rijden, worden deze gebruikt om te berekenen of een trein met vertraging rijdt. De rijtijdspeling komt grofweg neer op 8%, maar is niet tussen elke twee stations gelijk. Er is gekozen om de trajecten voor grote stations of knooppunten meer rijtijdspeling te geven. Dit heeft te maken met het feit dat de meeste reizigers bij grotere stations in- en uitstappen en de trein op die punten met name punctueel moet zijn.

Vertraging

Uiteraard mag een definitie voor treinvertraging niet ontbreken. Onder vertraging wordt verstaan: het negatieve tijdsverschil tussen het geplande moment van een actie ten opzichte van het werkelijke moment van een actie. Een actie kan een aankomst zijn, een vertrek of een passeermoment van een dienstregelingspunt zijn. Er wordt binnen NS een verschil gemaakt tussen een vertraging en een uitval van een trein: als een trein tijdens een rit uitvalt is er geen werkelijk moment van de actie en wordt er dus ook geen vertraging gemeld.

Punctualiteit

NS houdt bij hoeveel treinen op tijd aankomen en vertrekken, hiermee wordt de punctualiteit gemeten. Hierbij geldt de punticaliteitsnorm: een trein is punctueel wanneer de werkelijke vertrektijd minder dan 3 minuten afwijkt van de geplande vertrektijd, in sommige situaties wordt een afwijking tot 5 minuten ook als punctueel gezien. De treinpunctualiteit is in 2014 gestegen, deze ging van 93,6% in 2013 naar 94,9% in 2014, de waarde schommelt over het algemeen rond de 93% 4.



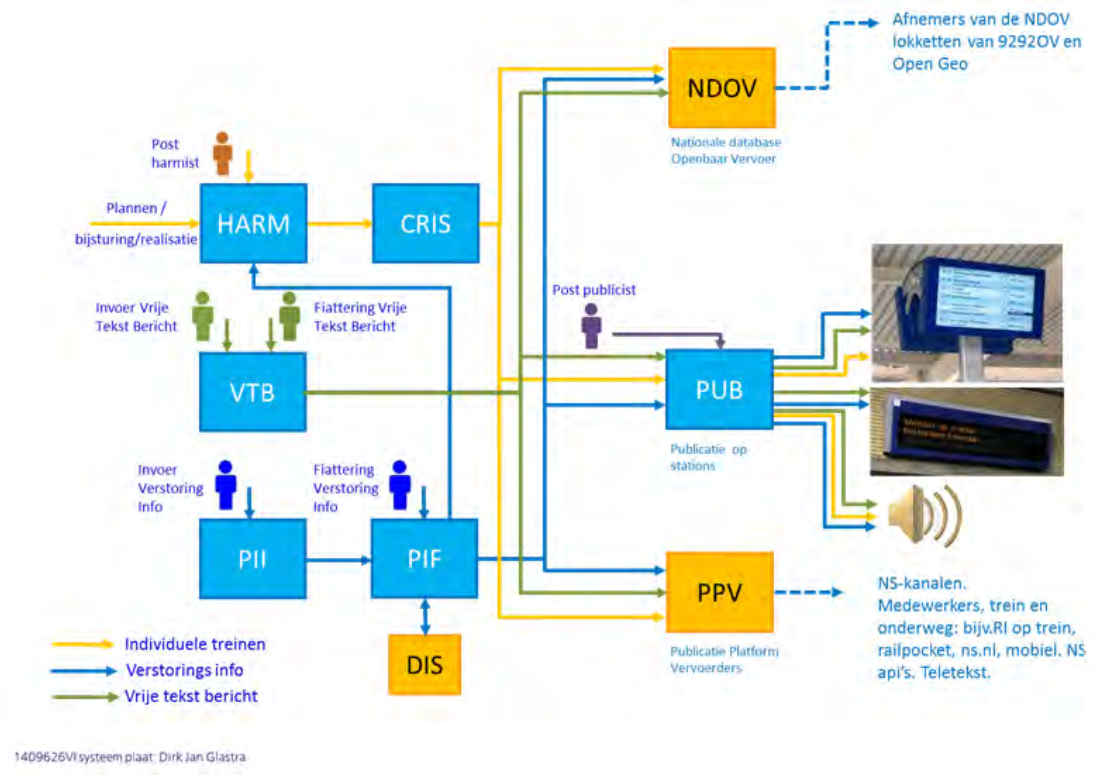
Figuur 4: Punctualiteit (cf. [6])

2.3 Systeemlandschap

NS kent twee reisinformatie systemen die de treinvertraging berekenen om ze aan verschillende reisinformatie-producten door te geven. In dit hoofdstuk werken we uit hoe deze systemen eruit zien, om welke producten het gaat en wat de toepassingen betekenen voor de eisen van het model.

2.3.1 Info-plus en AR-nu

De twee systemen die logistieke gegevens omzetten naar reisinformatie (waaronder het berekenen van de treinvertraging) zijn *Info-plus* en *AR-nu*. *Info-plus* is een nieuw systeem en zal samen met een nieuw publicatiesysteem PPV op den duur *AR-nu* overnemen. Beide systemen berekenen ongeveer op dezelfde wijze de vertragingen. In figuur 5 is het systeemlandschap van *Info-plus* te zien. De kern van het systeem bestaat uit drie deelsystemen: HARM, CRIS en PUB. Kortweg



Figuur 5: Systeemlandschap Info-plus (Bron: NS intern)

geldt dat de eerste de invoer van de data betreft, de tweede de verwerking en de derde de uitvoer. Het eerste deelsysteem HARM (Harmonisatie) stelt de invoer samen zodat het een goede input vormt voor CRIS (Centraal Reis Informatie Systeem). HARM krijgt onder andere de data binnen van de geplande dienstregeling. In deze data zijn vooraf geplande wijzigingen ook opgenomen. Hiernaast ontvangt HARM ook de werkelijke tijden waarop elke trein een meetpunt is gepasseerd. De combinatie van geplande en werkelijke tijden vormen de basis voor het rekenmodel dat CRIS uitvoert. Het precieze rekenmodel wordt later besproken. Op dit moment is het van belang te weten dat de uitvoer van CRIS de berekende tijden zijn waarop de trein het resterende deel van de rit zal uitvoeren. CRIS berekent dus welke treinen een vertraging zullen hebben en tevens ook de omvang van de vertraging. Het systeem vertaalt namelijk de bijgewerkte rit-informatie naar het gezichtspunt van een station in de producten 'Dynamische vertrekstaat' en 'Dynamische aankomststaat' voor elk station. Het laatste systeem, PUB (Publicatie), heeft als taak de dynamische vertrekgegevens te publiceren op de borden en omroepinstallaties van de stations.
Naast deze deelsystemen voor rit-informatie zijn er ook deelsystemen, VTB, PII en PIF, voor de verhalende reisinformatie (bij ernstige verstoringen en werkzaamheden aan het spoor). Waar PUB de aangepaste reisinformatie publiceert op stations, gaan in de toekomst PPV en NDOV de overige distributiekanalen thuis/onderweg en in de trein van actuele reisinformaite voorzien, zoals de website van NS of de externe website 9292ov. De producten die door deze deelsystemen mogelijk zijn gemaakt worden hieronder besproken.

2.3.2 Reisinformatie producten

Vertragingen worden op verschillende manieren doorgegeven aan de reizigers en kunnen gevat worden in 4 producten. De producten verschillen qua locatie en qua doelgroep. Elk product sluit op deze manier aan bij een andere groep reizigers.
Het eerste product is de 'Dynamische Vertrekstaat'. Deze publiceert het deelsysteem PUB van Info-plus op borden en via de omroep op de stations. Deze informatie is bestemd voor reizigers die zich reeds op het station bevinden. Deze reizigers zullen vaak de eerste reismogelijkheid nemen naar hun bestemming en zijn daarom met name geïnteresseerd in de vertrekvertraging van de eerstvolgende treinen. Via PPV zal dezelfde informatie worden verstrekt aan bijvoorbeeld mobiel internet en de Reisplanner applicatie.

De volgende twee producten betreffen de reisinformatie in de trein. Op de schermen in de trein wordt namelijk informatie getoond over de rit zelf: de actuele aankomst- en vertrektijden van de trein. Dit komt uit het product rit-informatie. Reizigers zullen hierbij vooral geïnteresseerd zijn in de aankomsttijden voor hun bestemming.
Daarnaast wordt het product actuele overstapinformatie getoond waarbij rekening is gehouden met actuele aankomst- en vertrektijden van de trein zelf als van de treinen waarop overgestapt kan worden. Door een verwachte vertraging van de trein zullen enkele overstappen wegvallen en als een andere trein vertraging heeft kan dat juist een extra overstapmogelijkheid opleveren.

Het laatste product is reisadvies welke aangeboden wordt door de reisplanner op de website, en op applicaties voor tablet en smartphone. Deze reisplanner verwerkt namelijk de actuele rit-informatie en past eventuele overstapmogelijkheden hierop aan. De reizigers die gebruik maken van de reisplanner zijn veelal nog niet vertrokken en zijn daarom meer geïnteresseerd in wijzigingen op langere termijn.

Het systeem *Info-plus* levert informatie aan stations, waar het systeem *AR-nu* nu nog informatie geeft via de (mobiel-) internet, de applicaties en een device voor het eigen personeel, de railpocket. Vertrektijden op de borden op stations en op de schermen in de trein zijn dus afkomstig van *Info-plus* en de reisplanner van de website maakt gebruik van de gegevens van *AR-nu*. Doordat beide systemen een klein verschil hebben in de manier waarop ze de vertraging berekenen en een bedienaar van HARM handmatig vertragingen kan aanpassen, kan het voorkomen dat een verschillende uitkomst wordt weergegeven op de applicatie en op het scherm in de trein. Tot slot is het relevant om op te merken dat een vertraging op het station rekenkundig wordt afgerond op 5 minuten, waar de reisplanner en de schermen in de trein dit op 1 minuut nauwkeurig vermelden.

2.3.3 Conclusie

De eisen voor een nieuw model hangen af van de strengste eisen die door de zojuist besproken toepassingen worden gesteld. Ten eerste blijkt dat zowel de aankomstvertraging als de vertrekvertraging van belang is om te voorspellen: de vertrekvertraging is van belang voor alle reizigers die hun treinreis moeten starten, de aankomstvertraging is van belang voor het inschatten van overstapmogelijkheden. Vanwege de toepassing via de reisplanner is het van belang dat vertragingen vrij lang van te voren worden doorgegeven. Voor de informatie op het station is dit minder van belang. De nauwkeurigheid is met name van belang voor reizigers die moeten overstappen, voor hen is een nauwkeurigheid van een minuut noodzakelijk.

Het model zal dus zowel aankomstvertragingen als vertrekvertragingen meenemen, zal dit op de minuut nauwkeurig voorspellen en zal pogen deze vertragingen zo vroeg mogelijk beschikbaar te

stellen.

2.4 Vertragingsfactoren

Vertraging kan op verschillende manieren worden veroorzaakt en kan aan de hand van de situatie ofwel worden teruggebracht ofwel verder oplopen. In dit onderdeel willen we verschillende factoren en situaties in kaart brengen. We maken hierbij onderscheid tussen het ontstaan, het toenemen en het verminderen van een vertraging. Het ontstaan van een vertraging is vanzelfsprekend: vanuit een situatie waarin de trein volgens dienstregeling reed is overgegaan naar een vertragingssituatie. Mocht de trein zich in een vertragingssituatie bevinden zijn er drie opties: de vertraging wordt verminderd, wordt toegenomen of blijft nagenoeg gelijk.
Het ontstaan van vertraging en het toenemen van vertraging betekent beide dat er meer vertraging is opgelopen. Het verschil zit erin dat het toenemen van vertraging te maken heeft met het verloop van een reeds aanwezige vertraging, waar het ontstaan van vertraging te maken heeft met een directe reden voor het ontstaan van een vertraging. Hieronder zijn enkele factoren uiteengezet.

2.4.1 Ontstaan van vertraging

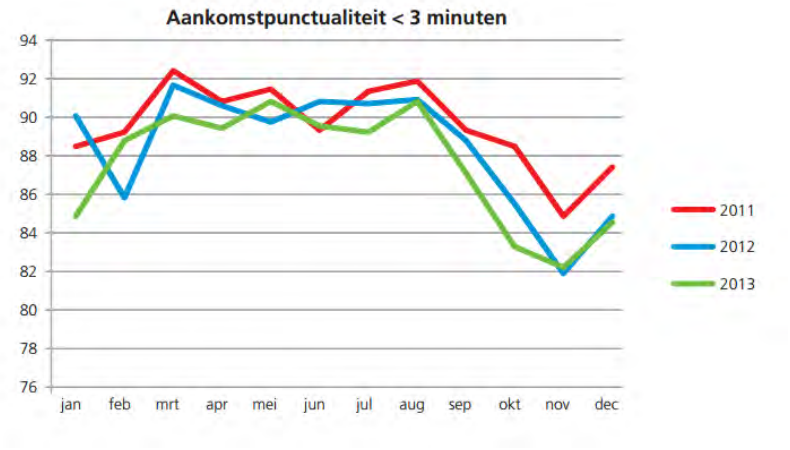
Hieronder vallen de oorzaken waardoor een vertraging ontstaat in een omgeving waarin treinen tot dan toe volgens dienstregeling reden.

Mankementen aan het spoor

Deze oorzaak bevat alle storingen die met het spoornetwerk te maken hebben of met mankementen aan de trein. Een bekend mankement is de wisselstoring, maar ook een kapotte bovenleiding en een seinstoring kunnen problemen veroorzaken.
Het is uiteraard niet te voorspellen wanneer een onderdeel stuk gaat, al kan het weer een indicator zijn. Dit komt doordat harde wind een oorzaak kan zijn van een kapotte bovenleiding en een seinstoring kan worden veroorzaakt door vrieskou. Wanneer de storing enige tijd actief is, kan deze worden gesignaleerd door het rijgedrag van voorgaande treinen te evalueren. Deze zullen waarschijnlijk extra vertraging hebben opgelopen door het mankement. Hier wordt later verder op ingegaan.

Ongeregeldheden bij het spoor

Hierbij horen alle externe oorzaken: oorzaken die niet zijn ontstaan door een verstoringen van het spoornetwerk of door een mankement aan de trein, maar er wel toe leiden dat een trein niet op de gewenste snelheid kan doorrijden. Als spoormedewerkers zich voor werkzaamheden dicht bij het spoor bevinden ontstaat bijvoorbeeld een gevaarlijke situatie bij het spoor waardoor de trein snelheid moet minderen. Een andere oorzaak die bij deze factor hoort is een blokkade op het spoor, zoals omgevallen bomen of een kapotte trein.
In figuur 6 zijn de punctualiteitscijfers van Prorail van de laatste paar jaar weergegeven, waardoor te zien is dat er inderdaad sprake is van een seizoenseffect. De punctualiteit komt niet overeen met die van NS die hierboven is besproken, aangezien hier de drie minuten norm is gehanteerd inplaats van 5 minuten. Bovenstaande factoren kunnen mogelijk deels herleid worden door een indicatie van het weertype mee te geven, zoals de windkracht of de temperatuur. Echter, in dit onderzoek wordt bekeken of er een beter voorspellingsmodel gemaakt kan worden enkel met behulp van data welke bij NS beschik baar is. Vandaar dat het weer niet wordt meegenomen in het model.



Figuur 6: Punctualiteit Prorail (cf. [3])

Storingen en werkzaamheden worden opgeslagen bij NS, echter deze data is niet in een bruikbaar format beschikbaar. Vandaar dat een aanwezige storing niet wordt meegenomen in het model.

2.4.2 Verminderen van vertraging

Als er een vertraging is ontstaan door een van bovenstaande oorzaken kan deze op verschillende manieren worden teruggebracht. Deze worden hieronder uitgezet.

Rijgedrag machinist

Tussen elke twee stations zit een rijtijdspeling ingebouwd om kleine vertragingen in te lopen. De mate waarin dit lukt, kan afhangen van de oplettendheid en rijgedrag van de machinist.

Gedrag conducteurs

Bij een vertraging is de halteertijd een moment waarop de vertraging ingelopen kan worden. In de dienstregeling staat vast hoe lang de halteertijd is op een station, maar bij een vertraging kan deze verminderd worden. Tijdens de haltering op een station spelen conducteurs een rol.

Het verminderen van vertraging heeft dus te maken met het personeel op de trein en is verschillend voor het rijproces en halteerproces. In het hoofdstuk data analyse wordt verder ingegaan op de mogelijkheden om tijdens deze processen vertraging in te lopen.
Het kan voorkomen dat een trein vertraging oploopt doordat personeel ziek is en gewacht moet worden op vervangen of zich heeft verslapen. Deze situaties zijn niet vooraf te voorspellen en worden deshalve niet meegenomen in het verslag.

2.4.3 Toenemen of gelijkblijven van vertraging

Er zijn situaties te bedenken waarin de vertraging niet kon afnemen: het neemt toe, blijft gelijk of wordt overgebracht op andere treinen.

Trein komt in conflict met andere trein

Door een vertraging kan een trein in conflict komen met een andere trein. In een dergelijke situatie beslist de bijsturing welke trein voorrang krijgt op een andere. Vaak wordt gepoogd een vertraging lokaal te houden, waardoor een trein die een grotere afstand aflegt voorrang krijgt. Bijsturing probeert om de vertraging te beperken, maar een conflict kan niet altijd vermeden worden. Als twee treinen achter elkaar aan rijden kan dit voor extra vertraging zorgen bij beide treinen als de voorste trein een extra stop moet maken om de achterste te laten passeren. Het gevolg is dat een trein die nog volgens dienstregeling reed hierdoor ook vertraging oploopt.
In het hoofdstuk Data analyse wordt gekeken naar de correlatie tussen het verschil in de aankomsttijden op eenzelfde locatie van twee treinen en de kans op een toename van vertraging.

Perron bezet

Een vertraagde trein wil op een afwijkende tijd een station binnenrijden, waardoor het mogelijk is dat op dat moment het perron bezet is voor een andere trein. Hierdoor loopt de trein extra vertraging op.
Wanneer een trein met vertraging een station wel binnen kan rijden, neemt het op een afwijkende tijd plaats in bij een perron. Dit kan als gevolg hebben dat een trein die volgens dienstregeling rijdt niet het station binnen kan rijden en moet wachten tot de vertraagde trein het station heeft verlaten. Nu ontstaat een vertraging bij een trein die nog volgens dienstregeling reed.

Drukte in de trein of op het perron

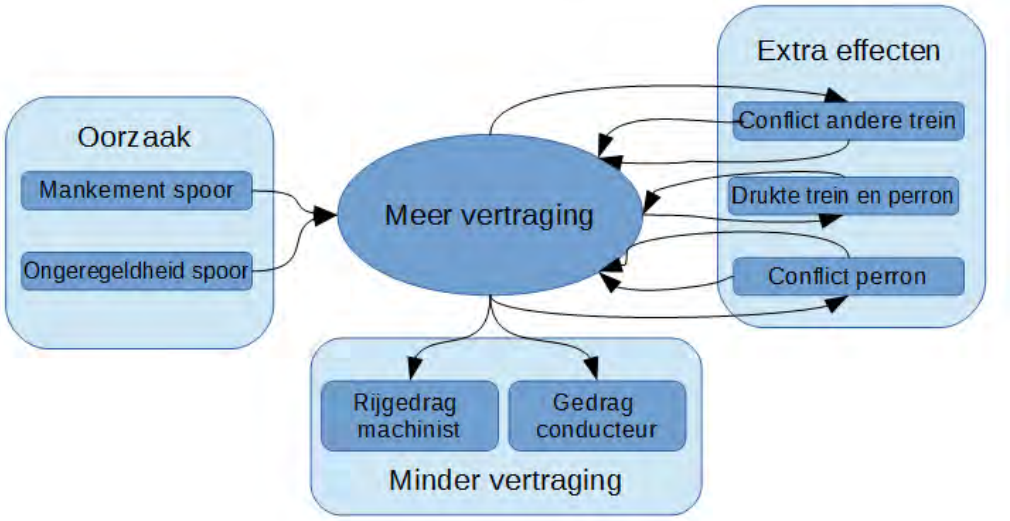
Hierboven is beschreven dat de halteertijd verkort kan worden om de vertraging te verminderen. Wanneer een trein met vertraging een station binnenkomt bepaalt de drukte hoe lang de trein moet halteren. In een rustige situatie kan de trein soms tot enkele minuten korter halteren, maar bij drukte is dit niet altijd mogelijk.
In het hoofdstuk Data analyse wordt het verkorten van de halteertijd verder onderzocht.

Gemiste aansluiting

Het kan voorkomen dat het materieel van een trein vrij snel na aankomst bij een (eind-)station wordt ingezet voor een andere treinrit. In dat geval zal een vertraging van het materieel in de ene treinrit voor een vertraging zorgen voor de volgende treinrit.
In het hoofdstuk Probleemstelling zal worden toegelicht dat deze situatie niet wordt meegenomen in het model.

2.4.4 Schema verloop vertraging

Hierboven zijn enkele facetten genoemd die een rol spelen bij het oplopen, toenemen of verminderen van vertraging. In onderstaand schema, zie figuur 7, is weergegeven hoe het verloop van een vertraging kan zijn en welke aspecten hierbij een rol spelen. Vanuit 'conflict trein' en 'conflict perron' lopen twee pijlen terug naar 'meer vertraging', omdat een andere trein hierdoor ook (extra) vertraging kan oplopen. Uit dit schema blijkt dat een vertraging ofwel door rijtijdspeling kan worden opgelost, maar dat het ook mogelijk is dat het probleem escaleert doordat er verschillende conflictspunten zijn met andere treinen. De hierboven genoemde oorzaken zullen als leidraad worden meegenomen bij het bepalen van mogelijke factoren voor een nieuw model.



Figuur 7: Schematisch verloop vertraging

2.4.5 Voorbeeld: Oorzaak vertraging voor specifiek geval

In het volgende voorbeeld wordt een situatie uitgewerkt, waarbij uit de data wordt gepoogd af te leiden wat de oorzaak was voor een bepaalde vertraging. Dit betreft de grote vertraging van een trein op het traject van het meetpunt bij Meppel tot aan de aankomst in Zwolle op 25 september bij treinnummer 766. In tabel 1 is weergegeven hoe de vertraging is opgelopen van de 2 minuten bij het meetpunt bij Meppel tot aan de bijna 9 minuten vertraging bij het vertrek in Zwolle. Er is te zien dat de vertraging met 5 minuten is vergroot tot 7 minuten in het traject tussen de meetpunten bij Meppel (Mp) en Dedemsvaart (Ddv). Deze vertraging kan in de resterende 10 minuten tot aan de aankomst in Zwolle slechts met een kleine minuut worden ingelopen.

Dienstregelpunt	Geplande tijd	Gerealiseerde tijd	Vertraging
Mpa	18:27:00	18:29:00	120s
Mp	18:28:00	18:31:18	178s
Ddv	18:34:00	18:41:46	466s
Hea	18:39:00	18:46:54	474s
Zlgea	18:41:00	18:48:23	443s
Zl (A)	18:43:00	18:50:25	445s
ZL (V)	18:45:00	18:53:36	516s

Tabel 1: Deel rit treinnummer 766

Opvallend is dat andere treinen die iets eerder dit traject aflegde ongeveer hetzelfde vertragingspatroon hadden: voor en bij het meetpunt van Meppel was er geen of een kleine vertraging welke fors toenam in het traject tussen Meppel en Dedemsvaart. De vertraging hield vervolgens aan tot de aankomst in Zwolle. Als voorbeeld zijn de rijtijden van een trein van treinserie 12500 weergegeven, zie tabel 2.
Verder zijn in tabel 3 de treinen weergegeven die rond deze periode langs het meetpunt 'Ddv'

Dienstregelpunt	Geplande tijd	Gerealiseerde tijd	Vertraging
Mpa	18:21:00	18:23:00	120s
Mp	18:22:00	18:24:04	124s
Ddv	18:28:00	18:35:44	484s
Hea	18:33:00	18:40:34	454s
Zlgea	18:35:00	18:42:11	431s
Zl (A)	18:38:00	18:44:08	368s
ZL (V)	18:48:00	18:48:23	23s

Tabel 2: Deel rit treinnummer 12566

zijn gereden met hun bijbehorende richting. Treinen die de andere richting op gaan lopen geen vertraging op. Dit feit kan erop duiden dat er twee sporen liggen op dit traject en dat meerdere treinen op hetzelfde spoor vertraging oplopen doordat ze hinder van elkaar ondervinden. We zien inderdaad dat een sprinter een ruime vertraging heeft en dat de twee intercity's het meetpunt al gepasseerd moesten hebben op het moment dat de sprinter het meetpunt heeft bereikt. Een mogelijke oorzaak voor deze vertraging is dat de intercity's hinder ondervinden van de vertraagde sprinter. Doordat ze tot aan Zwolle achter de sprinter aan moeten rijden is ook verklaard waarom de intercity's vrijwel geen vertraging inlopen.

Richting	Treintype	Treinnummer	Geplande tijd	Gerealiseerde tijd	Vertraging
Zwolle	SPR	9166	18:23:00	18:31:03	483s
Meppel	SPR	9161	18:35:00	18:35:10	10s
Zwolle	IC	12566	18:28:00	18:35:44	484s
Zwolle	IC	766	18:34:00	18:41:46	486s

Tabel 3: Diverse treinen langs meetpunt Dedemsvaart

3 Data analyse

NS beheert veel data, waaronder een dataset over de geplande aankomst-/vertrek- en doorkomsttijden van treinen bij meetpunten. Deze dataset, genaamd 'Treinactiviteiten', bevat veel relevante gegevens en in principe zal de data voor dit onderzoek uit deze dataset worden verkregen. In dit hoofdstuk wordt kort uitgelegd waar de dataset informatie over geeft, waarna enkele eerste analyses worden uitewerkt.

Dataset

De dataset 'Treinactiviteiten' is zodanig opgebouwd dat het in elke regel informatie geeft over een doorgekregen meting. Een meting indiceert de tijd waarop een trein langs een bepaald meetpunt is gereden. Een meetpunt wordt gekarakteriseerd door een afkorting van de locatie en het soort meetpunt. Bij een station komt namelijk zowel een meting door van een aankomst als van een vertrek. Deze hebben dezelfde naam als meetpunt, namelijk die van het station. Door het type meetpunt toe te voegen worden 'aankomst' en 'vertrek' van elkaar onderscheiden. Een derde type is 'doorkomst', bedoelt voor meetpunten die enkel gepasseerd worden.
In de dataset is de meting reeds gekoppeld aan de treinserie en treinnummer van de trein die de meting heeft veroorzaakt. Hierdoor is naast het tijdstip en de locatie van de meting ook informatie weergegeven over de trein zelf: het type trein, de vervoerder en de treinrichting. Vervolgens wordt de geplande tijd waarop de trein het meetpunt had moeten passeren weergegeven en is het verschil tussen de geplande tijd en de werkelijke tijd berekend.

Cijfers

NS beschikt over de vertrek- en aankomsttijden voor alle treinen van de afgelopen jaren. In de data analyse is de data gebruikt welke in een later stadium wordt gebruikt om een nieuwe model op te trainen en te testen. Dit is de date over 28 dagen, informatie over deze dagen is weer gegeven in appendix B.1. Elke dag van de week is tweemaal gerepresenteerd en de data liggen door het jaar heen. Er is ook bekeken of de dag qua vertragingen en uitval een gemiddelde dag was voor die periode om op deze manier een zo representatief mogelijke data set te vormen.
In tabel 4 zijn enkele aantallen te zien om inzicht te krijgen in de hoeveelheid data. De percentages intercity en sprinter ritten tellen niet op tot 100%, dit komt doordat er treinen zijn die halverwege hun rit van treintype overgaan. Die ritten worden dubbel meegeteld.

Tabel 4: basisgegevens

Gegevens	Aantal
Aantal treinseries	91
Aantal treinnummers	4789
Aantal meetpunten	500
Aantal stations	310
Intercity-ritten per dag	35.4%
Sprinter-ritten per dag	66.1%
Aantal te voorspellen gevallen	2.438.617
300s $>$ Geplande rijtijd	61.1%
300s $\leq$ Geplande rijtijd $<$ 600s	23.3%
600s $\leq$ Geplande rijtijd $<$ 900s	8.8%
900s $\leq$ Geplande rijtijd $<$ 1200s	4.4%
1200s $\leq$ Geplande rijtijd $<$ 1800s	2.0%
1800s $\leq$ Geplande rijtijd $<$ 3600s	0.4%

Vertraging

Hieronder (figuren 8a en 8b) is te zien hoeveel treinen met een zekere afwijking zijn vertrokken of aangekomen bij een station.

(a) Vertraagde treinen bij aankomst

(b) Vertraagde treinen bij vertrek

Figuur 8: Verdeling vertraagde treinen

Uit de figuren blijkt dat een groot deel van de treinen vertrekt danwel aankomt met minder dan een minuut vertraging. Slechts 5% van de treinen heeft een vertraging van meer van 2.5 minuten. Aangezien we de vertraging per minuut willen voorspellen, vallen alle treinen met een vertrek-/aankomstvertraging van minstens 30 seconden onder de noemer 'vertraagd'. Met deze definitie hebben ruim 40% van de vertrekken te laat plaatsgevonden en 50% van de aankomsten. Dit verschil tussen de punctualiteit van vertrek en aankomst is opmerkelijk. Uit dit verschil blijkt dat treinen blijkbaar tijdens het halteren vertraging kunnen inlopen, zodat een deel van de treinen die met vertraging bij een station aankomen niet met vertraging vertrekken.

Opkomen vertraging

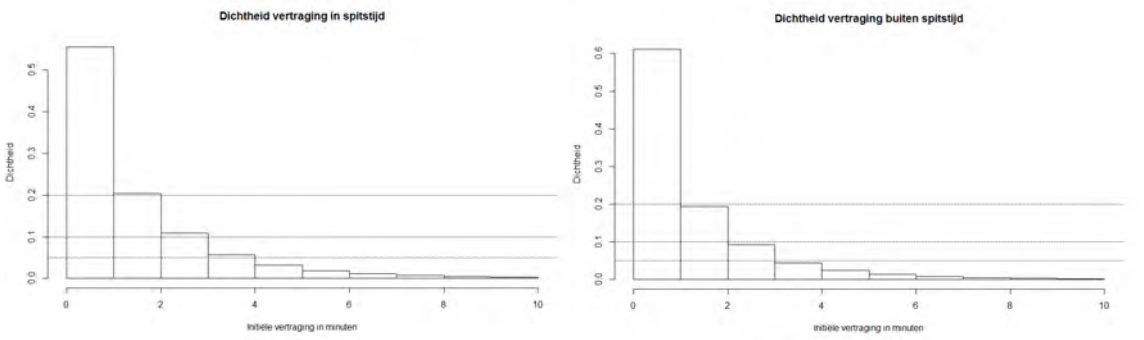
Onder het opkomen van vertraging verstaan we de situatie waarin de initiële vertraging, de vertraging die aanwezig is op het moment dat de trein het meetpunt passeert, minder is dan 30 seconden en overgaat naar een vertraging van ten minste 30 seconden bij het volgende station. In onderstaande tabel 5 is weergegeven welk deel van de data vertraging oploopt. Er is te zien dat 73.5 % van de treinen zonder vertraging bij het huidige meetpunt geen vertraging oploopt. Slechts 13.7% krijgt een vertraging van minstens 1 minuut, 3.3% een vertraging van ten minste 2 minuten. Om begripsverwarring te voorkomen wil ik benoemen dat de data waarbij de vertraging groter is dan 60 seconden ook valt onder de data waarbij de vertraging groter is dan 30 seconden.

Tabel 5: Opkomen vertraging

Vertraging bij meetpunt	Deel dataset	Vertraging volgend station		
		> 30 s	> 60 s	> 120 s
minder dan 30 s	44.1%	26.5%	13.7%	3.3%
minder dan 60 s	57.6%	33.3%	15.8%	3.8%

Spits

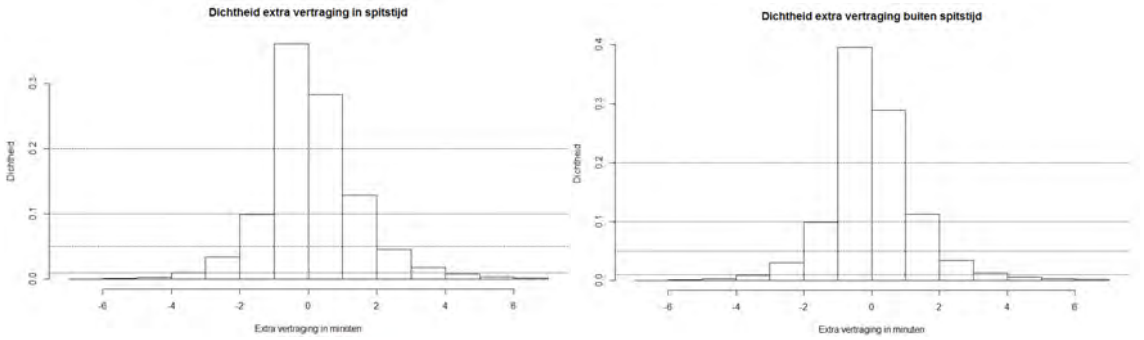
In figuur 9 is verdeling van de initiële vertraging weergegeven voor treinen binnen en buiten de spits. Er is te zien dat buiten de spits meer treinen rijden zonder een vertraging. Over het algemeen komen vertragingen iets vaker voor binnen spitsuren dan buiten de spits.
Figuur 10 toont vervolgens dat ook de toename bij vertraging groter is dan buiten de spits. Een



(a) Binnen de spits **(b)** Buiten de spits

Figuur 9: Verdeling vertraging binnen en buiten de spits

toename van vertraging komt binnen de spits voor bij 49.1% van de data en bij 46.0% van de data buiten de spits. Met behulp van een Wilcoxontest, zie hoofdstuk 4.6, is bekeken of de verdelingen gelijk zijn. De p-waarde was kleiner dan $2.2 \cdot e^{-16}$, waardoor deze hyptohese wordt verworpen. De verdelingen zijn dus niet identiek.



(a) Binnen de spits **(b)** Buiten de spits

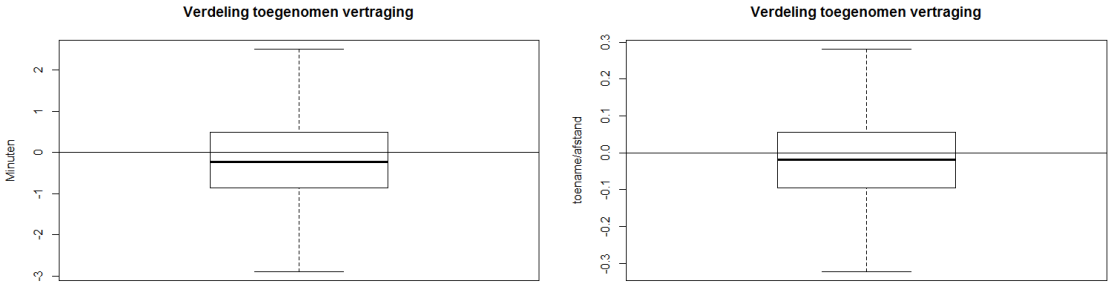
Figuur 10: Verdeling extra vertraging binnen en buiten de spits

Verloop vertraging

In figuur 11a is de verdeling te zien van de extra vertraging naar het eerstvolgende meetpunt voor de data met een initiële vertraging van minstens 30 seconden om te zien wat het verloop is van een bestaande vertraging. Een negatieve waarde geeft aan dat de trein vertraging heeft ingelopen, een positieve waarde betekent dat de trein extra vertraging heeft opgelopen.

Aangezien de afstand tussen een meetpunt en een actie bij een station verschilt, en dit mogelijk een relevante factor is voor de mate waarop een vertraging oploopt of inloopt, is in figuur 11b de toegenomen vertraging weergegeven ten opzichte van de geplande rijtijd, ofwel afstand. Ruim 50 % van de vertragingen wordt deels ingelopen.

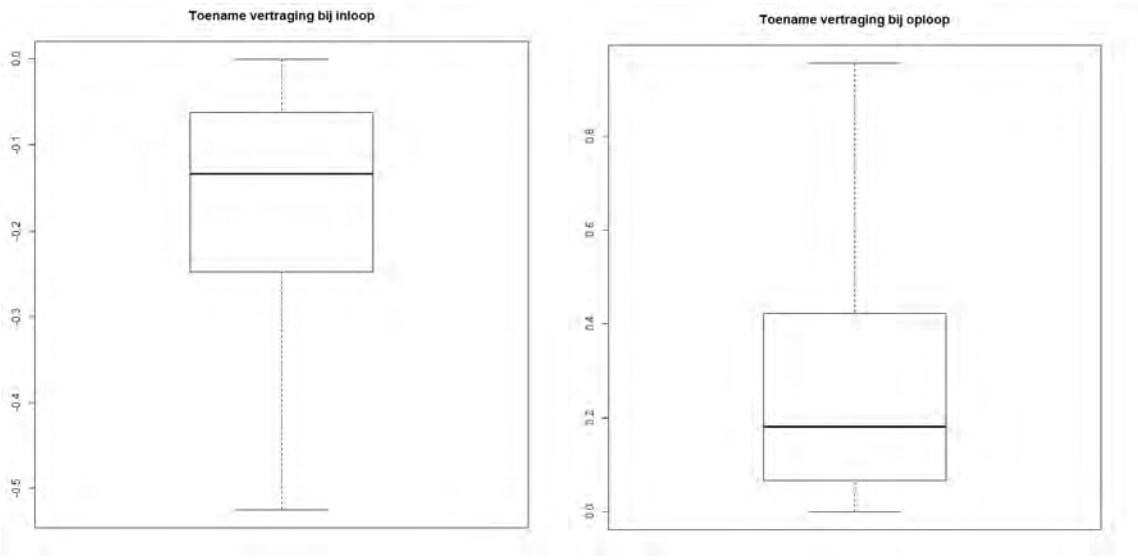
We kunnen de data splitsen op basis van de inloop of oploop op het volgende traject en krijgen



(a) Verdeling absolute vertraging **(b)** Verdeling relatieve vertraging

Figuur 11: Verdeling toename vertraging

onderstaande verdelingen in figuur 12a en 12b. We merken op dat bij oploop gemiddeld 19% van de rijtijd extra aan vertraging wordt opgelopen en dat bij inloop 12% van de rijtijd aan vertraging wordt ingelopen.



(a) Verdeling relatieve vertraging bij inloop **(b)** Verdeling relatieve vertraging bij oploop

Figuur 12: Verdeling toename vertraging bij inloop/oploop

4 Probleemafbakening

In de inleiding is het probleem geïntroduceerd en is de doelstelling van dit onderzoek gedefinieerd. Nu de context qua reisinformatie in kaart is gebracht en een eerste data analyse is uitgevoerd kan dieper op het probleem worden ingezoomd. Dit hoofdstuk gaat verder in op welke situaties het model een voorspelling dient te maken, welke attributen daarbij gebruikt worden en hoe het model geëvalueerd wordt.

4.1 Te voorspellen gevallen

Een eerste vraag hierbij is wanneer een voorspelling gemaakt dient te worden. Als input voor het voorspellingsmodel nemen we de doorkomsttijden van treinen die gemeten worden bij de meetpunten bij dienstregelpunten. Deze doorkomsttijden geven aan of de trein volgens dienstregeling rijdt of dat er op dat moment een vertraging aanwezig is. Aan de hand van het resultaat bij dit meetpunt wordt onderzocht hoe de trein haar traject zal vervolgen: hoe laat zal de trein op de eerstvolgende stations aankomen en vertrekken?
Er kan dus vanaf elk meetpunt een voorspelling worden gemaakt voor elk vertrek en aankomst bij alle volgende stations die de trein zal aandoen. Echter, gezien de rijtijd van bepaalde treinseries en de mogelijke onvoorziene hinder op het vervolgtraject is het voor de betrouwbaarheid goed om een beperking op deze gevallen te leggen. Onvoorziene hinder kan zowel plaatsvinden tijdens het rijtraject als tijdens de haltering, vandaar dat op beide punten een beperking wordt gesteld. In de dataset zijn maximaal 3 volgende stations - aankomsten en vertrekken - vanaf het meetpunt meegenomen, waarvan de geplande rijtijd hooguit één uur betreft.
In de eindevaluatie zal gekeken worden of de betrouwbaarheid van lange trajecten en trajecten met meerdere tussenliggende stations significant lager is. Aan de hand van dit resultaat kan in vervolgonderzoek de set met te voorspellen gevallen worden vergroot danwel verkleind.

4.2 Verloop voorspelling

Als voorbeeld bespreken we een trein die rijdt van Utrecht naar Zwolle, zie tabel 6. Bij het vertrek in Utrecht komt een meting binnen met de doorkomsttijd. Aan de hand van deze gegevens wordt

Tabel 6: Voorbeeld deel treinrit 519

Dienstregelpunt	Type	Tijd	Voorspeld voor
Utrecht	vertrek	6:50	A'foort aankomst en vertrek, Zwolle aankomst
Utrecht Overvecht	doorkomst	6:53	A'foort aankomst en vertrek, Zwolle aankomst
Bilthoven	doorkomst	6:56	A'foort aankomst en vertrek, Zwolle aankomst
Soestduinen	doorkomst	7:00	A'foort aankomst en vertrek, Zwolle aankomst
Amersfoort	aankomst	7:04	A'foort vertrek, Zwolle aankomst
Amersfoort	vertrek	7:07	A'foort vertrek, Zwolle aankomst
A'foort Vathorst	doorkomst	7:11	Zwolle aankomst
Putten	doorkomst	7:17	Zwolle aankomst
Nunspeet	doorkomst	7:27	Zwolle aankomst
Wezep	doorkomst	7:35	Zwolle aankomst
Zwolle	aankomst	7:42	

voorspeld hoe laat de trein aankomt in Amersfoort, vertrekt in Amersfoort en aankomt in Zwolle.

Tussen Utrecht en Amersfoort liggen meerdere dienstregelpunten, in tabel 6 is slechts een deel van de dienstregelpunten weergegeven, waar opnieuw de passeertijd wordt gemeten en opnieuw de bovengenoemde gebeurtenissen worden voorspeld. Het model blijft zich updaten.
Het tijdsinterval tussen het passeren van een meetpunt en de te voorspellen gebeurtenis kan zeer verschillend zijn. In Utrecht Overvecht wordt de aankomst in Amersfoort voorspeld, waarvan de geplande rijtijd 11 minuten bedraagt, en tegelijk ook de aankomst in Zwolle, wat volgens dienstregeling pas na 49 minuten zal gebeuren. Er is een maximum gesteld op de geplande rijtijd, zodat deze niet langer dan 60 minuten zijn.

4.3 Overige aannames en beperkingen in het model

Er zijn nog enkele aannames gemaakt welke hieronder zijn toegelicht.

Een trein zal niet voor de geplande tijd aankomen of vertrekken bij het station. Er is dus enkel sprake van positieve vertraging. Deze aanname wordt gedaan zodat er geen negatieve vertragingen zullen worden voorspeld. Het is niet de bedoeling dat treinen te vroeg aankomen of vertrekken en deze informatie zal ook niet worden doorgegeven aan de reiziger.

Aangezien een voorspelling wordt gedaan aan de hand van passeertijden bij meetpunten, wordt er geen vertraging voorspeld voor een trein die nog niet is vertrokken vanaf haar beginpunt. Het betreft hierbij onvoorziene wijzigingen, welke vanaf het begin van de treinrit zijn opgelopen. In een definitief model is dat wel de bedoeling, aangezien de trein al wel vertraging kan hebben opgelopen. Ook voor uitgevallen treinen worden geen vertragingen voorspeld. Immers, er wordt geen doorrijtijd gemeten als de trein niet meer rijdt, waardoor een voorspelling ook niet gemaakt kan worden. Hierbij moet worden opgemerkt dat een dergelijke voorspelling ook niet zinvol is, dus dit heeft geen negatieve consequentie.

4.4 Attributen

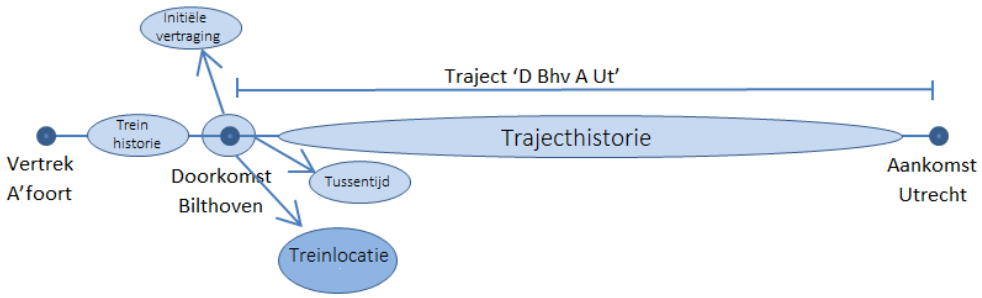
Het huidige model bleek een zeer eenvoudig model te zijn, dat met slechts een beperkt aantal factoren rekening hield. In hoofdstuk 2 zijn diverse oorzaken besproken waardoor vertraging kan ontstaan, toenemen of afnemen. In dit hoofdstuk zullen we proberen om enkele van deze oorzaken te vangen in attributen die meegenomen kunnen worden in een nieuw model. Verschillende attributen worden hieronder besproken en in figuur 13 is schematisch weergegeven op welk aspect van de rit de attribuut betrekking heeft. In appendix A wordt besproken hoe deze attributen uit de data zijn gemaakt.

Treinserie

Een treinserie omvat slechts een beperkt aantal te voorspellen gevallen, welke allen voor eenzelfde type trein van toepassing zijn. Hierdoor worden intercity's en sprinters, waarvan het rijgedrag onderling verschilt, van elkaar onderscheiden.

Traject

Waar een treinserie meerdere trajecten omvat, worden deze bij dit attribuut van elkaar onderscheiden. Een traject wordt gedefinieerd als een combinatie van meetpunt tot een te voorspellen



Figuur 13: Schematische weergave van enkele attributen

aankomst of vertrek. In figuur 13 is het traject vanaf de doorkomst in Bilthoven tot de aankomst in Utrecht weergegeven, in hoofdstuk A wordt verklaard hoe de afkorting voor het traject wordt gevormd. Wanneer het vertrek in Utrecht voorspelt wordt, is dus sprake van een ander traject. Hierbij kunnen echter wel verschillende treinseries worden omvat die dit traject afleggen.

Geplande rijtijd

Zoals reeds genoemd kan de geplande rijtijd een rol spelen voor de mate van betrouwbaarheid van een voorspelling, aangezien bij een grotere geplande rijtijd er meer onzekerheden zijn. Ook de mate van inloop hangt af van de geplande rijtijd, vandaar dat dit als attribuut wordt meegenomen.

Initiële vertraging

Het meetpunt geeft de initiële vertraging mee wat aan de basis staat van de voorspelling. De vertraging bij een volgend station is uiteraard afhankelijk van de initiële vertraging.

Spits

In de spits zijn er over het algemeen meer treinreizigers dan buiten de spits. Spits is gedefiniëerd als de tijd tussen $7.00-9.00$ uur en $16.00-19.00$ uur op doordeweekse dagen. In vertragingsfactoren is besproken dat de halteertijd mogelijk langer duurt of minder verkort kan worden wanneer er veel reizigers in de trein of op het perron rijden. Daarnaast rijden er op enkele trajecten meer treinen in de spits dan buiten de spits. Dit attribuut gaat dus in op de mogelijke oorzaak of toename van vertraging genaamd 'Drukte in de trein of op het perron', welke is besproken in hoofdstuk 2.4.3.

Treinhistorie

Mocht een trein met vertraging langs een meetpunt rijden, dan is het nog onduidelijk wat de situatie is. Het kan zijn dat de vertraging eerder op het traject groter was en dat de trein bezig is met het inlopen van vertraging, het kan ook zijn dat een trein sinds het vorige meetpunt juist vertraging heeft opgelopen, zie figuur 13. Met dit attribuut wordt onderzocht of het rijgedrag op het gedeelte tussen het vorige meetpunt en het huidige meetpunt een indicatie is voor het rijgedrag in het volgende deel. Door dit attribuut wordt onder andere ingezoomd op het (rij)gedrag van de machinist en conducteurs welke besproken is als mogelijke oorzaak om vertraging te verminderen

in hoofdstuk 2.4.2 en deze attribuut kan een indicatie geven dat de trein hinder ondervindt van een andere trein wanneer deze langzaam rijdt. Dit laatste attribuut is besproken in hoofdstuk 2.4.3.

Trajecthistorie

De attribuut trajecthistorie heeft te maken met problemen aan het spoor of een ander trajectspecifieke hinder. Hinder bij het traject betekent namelijk dat voorgaande treinen die over het traject rijden hiervan ook al last hebben ondervonden. Is het mogelijk om deze verstoringen op te sporen en valt daar van af te leiden of de betreffende trein ook vertraging zal oplopen op het komende traject? Dit attribuut heeft te maken met het traject dat de trein nog af zal gaan leggen, zie figuur 13. Hierbij wordt specifiek bekeken of situaties gevonden kunnen worden waar een trein vertraging zal oplopen in plaats van inlopen. Hierbij wordt ingezoomd op de factoren 'ongeregeldheden/mankementen bij het spoor' die zijn behandeld in hoofdstuk 2.4.1.

Tussentijd

Een attribuut is het verschil in de geplande en werkelijke tussentijd tussen de trein en de voorgaande trein. Mocht dit verschil groot zijn en de treinen dichter achter elkaar rijden, dan is het mogelijk dat ze hinder van elkaar gaan ondervinden. Deze attribuut heeft betrekking op de tussentijd bij het meetpunt zelf, zie figuur 13. Hierbij wordt ingegaan op de factor 'Trein komt in conflict met andere trein' welke is besproken in hoofdstuk 2.4.3 en waardoor een vertraging kan ontstaan of toenemen.

4.5 Data voorbereiding

De dataset 'Treinactiviteiten' is gebruikt als basisset voor het model. In SAS Enterpise Guide is deze tabel opgehaald en aangepast zodat alle besproken te voorspellen gevallen per regel worden weergegeven. In dit hoofdstuk wordt beschreven hoe dit gedaan is en hoe de dataset eruit ziet, welke data is verwijderd en hoe de extra attributen zijn gevormd.

In hoofdstuk 3 *Data Analyse* is vermeld hoe de dataset is opgebouwd: in elke regel staat informatie over een doorgekomen meting. Ten eerste is per doorgekomen meting bepaald welke gebeurtenissen van hieruit voorspeld dienen te worden. Hierbij is rekening gehouden met de twee beperkingen qua geplande rijtijd en aantal halteringen. Enkele observaties moeten echter uit de data worden gehaald of enigszins worden aangepast. Hieronder worden deze aanpassingen besproken.

Enkel treinen in doelgroep

De dataset 'Treinactiviteiten' bevat informatie over alle treinen die over het Nederlandse spoornetwerk rijden, zowel reizigerstreinen als leeg materieel. Een vertragingsvoorspelling dient enkel gemaakt te worden voor reizigerstreinen, zodat de observaties van leeg materieel uit de data is gehaald. Verder zijn internationale treinen uit de dataset gehaald, aangezien we enkel voor binnenlandse treinen vertragingen zullen voorspellen. In de dataset zijn ook gegevens aanwezig van treinen die onder een andere vervoerder hebben gereden. Deze treinen zijn ook uit de data gehaald, omdat we alleen NS treinen zullen voorspellen.
Tot slot hebben we alleen data gebruikt van treinen die allemaal daadwerkelijk hun rit hebben gereden en niet zijn uitgevallen, en waarbij er geen aanpassing op de dienstregeling van toepassing was op de treinrit.

Korte stop

In de meeste gevallen geldt dat er zowel de aankomst als de vertrektijd van een trein bij een station bekend is, maar dit is niet voor alle stations het geval. Het komt namelijk voor dat een trein een korte stop maakt op een station, waardoor de aankomsttijd en vertrektijd in dezelfde minuut vallen.
Aangezien de dienstregeling alle tijden in gehele minuten weergeeft zou de aankomsttijd en vertrektijd gelijk zijn. Voor een korte stop wordt daarom maar één tijd genoteerd. Wij zetten deze korte stoppen om in een aparte aankomsttijd en vertrektijd, zodat zowel de aankomst als het vertrek voorspeld kan worden. De aankomsttijd is 30 seconde voor de vertrektijd gezet om een benadering te geven. Het gevolg van de afwijking zal vermoedelijk gering zijn, aangezien het verschil in de orde van secondes ligt.

Te vroege vertrek- en aankomsttijden

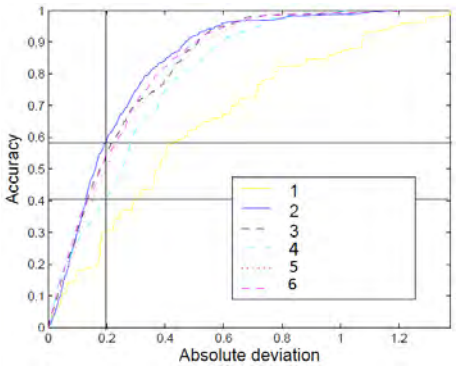
Hierboven is al genoemd dat de dienstregeling in gehele minuten wordt opgesteld. De werkelijke passeertijden worden echter in seconden doorgegeven. Hierdoor kan ook worden opgemerkt dat een trein enkele seconden te vroeg vertrekt, of in een uitzonderlijke situatie enkele minuten te vroeg aankomt. Aangezien een trein in principe niet vertrekt voor de geplande tijd, zal een veel te vroeg vertrek vrijwel nooit voorkomen. Het is wel mogelijk dat een te snel gereden trein enkele minuten vroeger dan gepland binnenkomt.
Aangezien dit uitzonderingen zijn en we niet het doel hebben negatieve vertragingen te voorspellen, zijn in de datasets deze 'vertragingen' op 0 seconden gezet. Een ander gevolg hiervan is dat als de initiële vertraging bijvoorbeeld (-90) seconden was en de vertraging op het volgende station (+10) seconden is er niet meer wordt aangegeven dat er 100 seconden aan vertraging is opgelopen. Immers, de eerste 90 seconden oploop van vertraging hebben geen negatieve waarde en moet daar ook niet mee verward worden in de data.

4.6 Evaluatie

Voor de beoordeling van het model hebben we in hoofdstuk 2.3 al enkele eisen gevonden. Het uiteindelijke model zal een vertraging in minuten weergeven, maar aangezien de data in seconden nauwkeurig is gegeven, zullen tussentijdse resultaten ook in seconden worden weergegeven. Voor de beoordeling van het model hebben we in hoofdstuk 2.3 al enkele eisen gevonden. De resultaten zullen op de diverse manieren worden weergegeven, deze methoden worden hieronder uitgewerkt.Het uiteindelijke model zal een vertraging in minuten weergeven, maar aangezien de data in seconden nauwkeurig is gegeven, zullen tussentijdse resultaten ook in seconden worden weergegeven. De resultaten kunnen op de volgende manieren worden weergegeven.

Resultaat in REC-plot

De gevonden afwijking kan worden weergegeven in een REC-plot (Regression Error Characteristic) [9]. Dit type plot is een afgeleide van de ROC (Receiver Operating Characteristic), welke gebruikt wordt voor de evaluatie bij classificatie problemen.
Een REC curve toont een error tolerantie op de horizontale as en de nauwkeurigheid van de regressie functie op de verticale as [9]. In het getoonde voorbeeld in figuur 14



Figuur 14: Voorbeeld REC [9]

zien we dat 68% van de resultaten bij 'lijn 2' binnen een marge van 0.2 ligt, terwijl dit slechts 40% is voor 'lijn 4'. Bij het voorspellen van treinvertragingen laat de plot zien welk deel van de voorspellingen valt binnen een afwijking ten opzichte van de werkelijke vertraging die op de horizontale as getoond wordt. Uiteraard is het wenselijk om het deel bij een kleine afwijking te maximaliseren. Er wordt hierbij geen rekening gehouden tussen een negatief of positief verschil in voorspelde vertraging en werkelijke vertraging. Het voordeel van een REC plot is dat het inzicht geeft in de prestatie van het model met betrekking tot een toegestane afwijking. Voor overstappers is een nauwkeurigheid op 1 minuut van belang, maar voor andere reizigers kan dat anders liggen. Deze plot geeft hierover een inzichtelijk beeld.

Werkelijke vertraging vs voorspelde vertraging

Een andere interessante plot zet de werkelijke vertraging of rijtijd uit tegen de voorspelde waarden. Perfecte voorspellingen zullen op de diagonaal terechtkomen, afwijkingen zullen ernaast worden weegegeven. Het interessante van deze plot is dat het wel onderscheid maakt tussen negatieve en positieve verschillen en dat uitschieters worden weergegeven. Het nadeel is dat een hoeveelheid dicht bij elkaar liggende punten niet goed van elkaar te onderscheiden zijn, waardoor geen goede inschatting van de verdeling van de afwijking gemaakt kan worden. Voor dit doel is een REC plot juist zeer geschikt.

Verbetering in minuten

Een histogram kan worden gebruikt om de frequentie van een zekere verbetering in de voorspelling tussen twee methoden weer te geven. De verbetering is als volgt weergegeven:

$$\begin{aligned}\text{Verbetering} = &|\text{vertraging huidige model} - \text{werkelijke vertraging}| \\ &-|\text{vertraging nieuw model} - \text{werkelijke vertraging}|\end{aligned}$$

en valt te interpreteren als het absolute verschil tussen de werkelijke vertraging met de voorspelling met het huidig model en met de voorspelling van het nieuwe model. Waar de vorige plotjes een totaalbeeld geven, wordt bij deze methode per voorspelling gekeken wat het verschil is. Een positieve verbetering van 1 minuut betekent dat de afwijking van de voorspelling van het nieuwe model 1 minuut is afgenomen ten opzichte van de voorspelling van het huidig model. Het kan zijn dat het huidige model een vertraging voorspelt die 3 minuten te laag uitvalt en het nieuwe model juist 2 minuten te hoog voorspelt. Het absolute verschil hiertussen is 5 minuten, maar de verbetering ten opzichte van de werkelijke vertraging is 1 minuut.

Afstandsintervallen

Aangezien de geplande rijtijd, ook wel afstand, tussen het meetpunt en de te voorspellen aankomst- of vertrektijd erg verschilt, zullen de uitkomsten ook weergegeven worden ten opzichte van deze geplande rijtijd. Zo zullen de uitkomsten gesplitst worden in de volgende afstandsintervallen: minder dan 5 minuten, 5 - 10 , 10 - 15 , 15- 20, 20 - 30, 30 - 60 minuten. Op deze manier wordt bekeken wat het resultaat is voor treinen die binnen 5 minuten de volgende actie hebben ten opzichte van treinen die een volgende actie pas tussen 15 en 20 minuten hebben. De reden voor de splitsing is dat we vermoeden dat een voorspelling minder betrouwbaar wordt als afstand groter wordt, aangezien de kans groter is dat de trein nog onvoorziene hinder ondervindt.

Het is goed te realiseren dat het mogelijk is dat er meerdere te voorspellen gevallen van één trein binnen dezelfde categorie vallen. Zo kan een sprinter 2 stations aandoen in minder dan 5 minuten, waardoor er twee aankomsten en twee vertrekken worden voorspeld vanaf een zeker meetpunt die allen bijvoorbeeld in de afstand 5-10 minuten vallen. Aangezien de geplande rijtijd tussen twee stations bij een sprinter over het algemeen korter is dan bij een intercity zal dit bij een intercity minder vaak voorkomen.

Verdelingswaarden

Tot slot kan de prestatie van het model worden gemeten door de mediaan, het gemiddelde en de standaard deviatie te bekijken van de absolute afwijking van de voorspelling. Het voordeel van de mediaan ten opzichte van het gemiddelde, is dat deze maat minder gevoelig is voor uitschieters. De standaard deviatie geeft de spreiding aan van de afwijking. Zo kan een resultaat met een hogere mediaan, maar met een kleinere standaard deviatie worden verkozen boven een andere model dat een lagere mediaan gaf.

Significantie test

Om te testen of een verandering aan een model tot een significante verbetering lijdt, wordt een significantie test uitgevoerd. Er is gekozen voor de Mann-Whitney-Wilcoxon test ofwel de Wilcoxon rank-sum test, aangezien deze niet veronderstelt dat de data normaal verdeeld is. Deze test, hierna Wilcoxon test, onderzoekt of beide verdelingen aan elkaar gelijk zijn. Anders gezegd, de kans dat een observatie met een zekere afwijking tot het oude of de nieuwe model hoort is beide 0.5. Deze nul hypothese wordt verworpen wanneer de p-waarde groter is dan 0.05.

Evaluatie te voorspellen gevallen

In het definitieve model wordt zoals eerder besproken onderzocht welk deel van de te voorspellen gevallen daardwerkelijk een betrouwbaare voorspelling geeft. In een laatste evaluatie zal de betrouwbaarheid van elk te voorspellen geval worden onderzocht. Er wordt onder andere bekeken of een geplande rijtijd van 60 minuten alsnog een betrouwbaar resultaat geeft, of dat dit nog van andere factoren afhankelijk is.

6 Methodiek

Via twee verschillende technieken zal gepoogd worden een nieuw model te ontwikkelen. De eerste methode maakt direct gebruik van historische data om de rijtijd te schatten. Via deze rijtijd kan de vertraging op een volgend station worden bepaald. De tweede methode probeert met behulp van een machine learning techniek de vertraging te voorspellen. Hieronder worden beide methodes besproken.

6.1 Schatting werkelijke rijtijd

Het huidige model neemt aan dat vertraging in principe niet wordt opgelopen, maar enkel inloopt. Deze inloop hangt enkel af van de geplande rij- en halteertijd en zijn hierdoor uniform voor alle treinseries en trajecten. Echter, de mate waarop een vertraging ingelopen kan worden kan van het treintype of traject afhangen, bijvoorbeeld doordat binnen bepaalde stedelijke gebieden niet sneller gereden kan worden. Door deze inloop factor specifieker te maken voor de betreffende situatie van de trein, kan de factor betrouwbaardere resultaten geven.
Een andere reden om dit te onderzoeken, is het resultaat van het onderzoek dat eerder door NS is uitgevoerd [11]. Hierbij werd voor elke twee dienstregelpunten een kruistabel opgezet waarbij uiteen werd gezet hoe vaak een combinatie van initiële vertraging en volgende vertraging voor kwam. De voorspelde vertraging betrof dan ofwel het gemiddelde, ofwel de mediaan ofwel de trimean. Deze gaf in veel gevallen een significante verbetering.

Methode

Voor de data uit de trainingset wordt bekeken wat de werkelijke rijtijden zijn geweest tussen een meetpunt en een actie op een volgend station. Een interessante groep vormen de treinen mét initiële vertraging. De initiële vertraging is de vertraging die aanwezig is op het moment dat de trein het meetpunt passeert. Van deze treinen wordt verwacht dat ze sneller rijden dan treinen zonder vertraging om vertraging in te lopen.
In principe volgt uit de kortste rijtijd de werkelijke rijtijdspeling: dit is de geplande rijtijd minus deze kortste rijtijd. Immers, als het mogelijk is geweest om een traject in die kortste tijd af te leggen, is dat in theorie ook voor een volgende trein op dit traject mogelijk. We zijn echter niet geïnteresseerd in rijtijden die in theorie mogelijk zijn, maar in die rijtijden waarvan verwacht mag worden dat ze gereden worden indien het nodig is om een vertraging in te lopen. Vandaar dat we voor de werkelijke rijtijden verschillende percentielen berekenen met data uit de trainingset om hieruit af te leiden welk percentiel de meest betrouwbare indicatie geeft. De berekende percentielen zijn: 2%, 5%, 10%, 25%, 50%, 75% en 90%, zodat de spreiding van de rijtijden bekend is. Hierbij is 2% een relatief korte rijtijd en 90% een relatief lange rijtijd.
Vervolgens wordt voor data in de testset het verschil berekend tussen de geplande rijtijd en een bepaald rijtijdpercentiel van de werkelijke rijtijd. Een positief verschil vormt de tijd die gebruikt kan worden om een vertraging in te lopen. Bijvoorbeeld; stel een geplande rijtijd tussen het vertrek in Amersfoort en het vertrek in Utrecht is 21 minuten, maar 10 procent van de treinen legt het traject af in minder dan 19 minuten. Er kan op dit traject dan 2 minuten van een vertraging worden ingelopen, als het 10^e percentiel wordt toegepast.

Tot slot volgt dan op de volgende wijze de vertraging op het volgende station:

$$\text{vertraging} = \max\left(\text{initiële vertraging} - \text{mogelijke inloop}, 0\right) \tag{1}$$

met

$$\text{mogelijke inloop} = \max\left(\text{geplande rijtijd} - \text{rijtijd percentiel}, 0\right). \quad (2)$$

Er is aangenomen dat een trein enkel zijn snelheid zal verhogen wanneer dat nodig is om een vertraging in te lopen, vandaar dat geen negatieve vertragingen voorspeld worden. Een voorspelling vanaf een meetpunt naar een aankomst of vertrek met enkele tussenliggende stations wordt bij deze methode in één keer uitgevoerd. Er wordt niet meegenomen hoeveel halteringen er zijn, hoe lang deze gepland zijn te duren en wat de afstand is tussen beide punten. Er wordt enkel in historische data onderzocht wat de werkelijke tijd is waarin het traject is afgelegd. Mocht een bepaalde groepering uit de testset niet voorkomen in de de trainingset, dan is hierover geen historische data beschikbaar en zal deze set ook niet worden meegenomen in de resultaten.

Verfijning methode

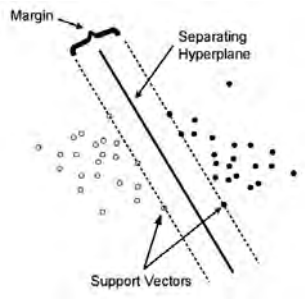
De voorspelling van de rijtijd kan specifieker gemaakt worden door meerdere attributen mee te nemen die besproken zijn in hoofdstuk 4.4. Wanneer enkel de geplande rijtijd wordt meegenomen, worden de rijtijdpercentielen gemaakt per geplande rijtijd die in de trainingset voorkomt. Dit resultaat wordt verfijnd als bijvoorbeeld de treinserie wordt meegegeven: dan worden de percentielen voor elke combinatie geplande rijtijd en treinserie berekend.

6.2 Machine Learning

De naam Machine Learning is gegeven aan een tak binnen computerwetenschappen waarbij een algoritme zelf uit de data leert, in plaats van dat er specifieke leerregels worden meegegeven [20], [22]. Algoritmes passen patroonherkenning toe om bepaalde voorspellingen te doen of beslissingen te nemen.
Machine learning technieken kunnen worden toegepast op verschillende typen problemen, waaronder classificatie en regressie problemen. Onder classificatie wordt verstaan dat de discrete klasse voorspeld moet worden, zoals het geslacht van een persoon of een type boom. Bij regressie problemen is er sprake van een numerieke uitkomst die voorspeld dient te worden, zoals de leeftijd van een persoon. In dit onderzoek wordt de (extra) vertraging voorspeld, wat een regressie probleem is.
Enkele belangrijke algoritmes binnen machine learning zijn: Multi-layer Perceptrons (MLP), Desicion Tree, Support Vector Machines (SVM) en Random Forests. Deze worden in een eerste iteratie vergeleken, waarna met het algoritme dat het beste resultaat geeft verdere onderzoeken worden gedaan. Er blijkt dat SVM het meeste nauwkeurige model gaf, deze zal om die reden voornamelijk worden gebruikt in dit onderzoeken. De werking van de SVM wordt hieronder uitgebreid besproken. In Appedix C worden de overige genoemde algoritmen ook kort besproken.

6.2.1 Support Vector Machines

Support Vector Machines (SVM) is een relatief nieuw algoritme dat gebruik maakt van machine learning technieken. Het is een aantrekkelijk algoritme, aangezien het zowel geschikt is voor regressie als voor classificatieproblemen en kan gezien worden als een nieuw trainings algoritme voor onder andere polynomische, RBF en multilayer perceptron netwerken [19].

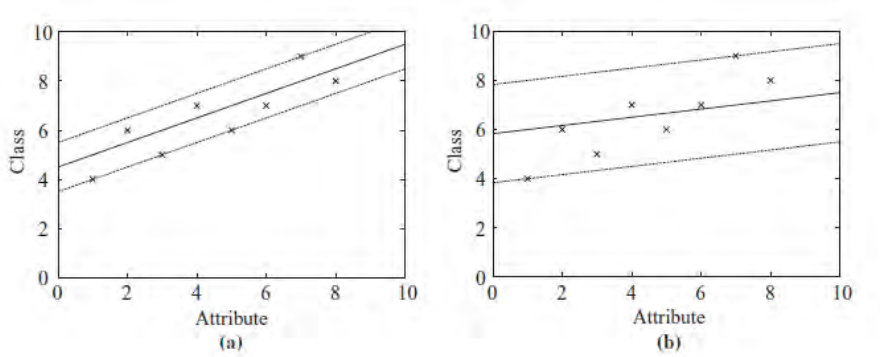


Figuur 20: Support Vector machine

Classificatieprobleem

SVM wordt van origine gebruikt voor classificatieprobleen. Het creëert een hypervlak om een scheiding aan te brengen tussen de verschillende klassen zodat de marge tussen de klassen is gemaximaliseerd [15]. Hiervoor selecteert de SVM een klein aantal grensgevallen van elke klasse die de support vectors worden genoemd[25]. In figuur 20 is het hypervlak en de bijbehorende support vectors te zien.
Het probleem en haar oplossing is vrijwel identiek aan die van de Support Vector Regression Machine en zal daarom enkel daar worden toegelicht.

Regressie probleem



Figuur 21: Support Vector machine bij regressie probleem [25]

Bij Support Vector Regression Machines (SVRM) wordt een tube om de regressiefunctie gebouwd. De afwijkingen op de regressiefunctie die binnen de tube liggen worden genegeerd [25]. In figuur 21 is te zien dat de regressiefunctie afhangt van de breedte van de tube, ϵ. In figuur 21a geldt $\epsilon = 1$, bij 21b geldt $\epsilon = 2$. Om het risico op overfitting te minimaliseren wordt de vlakheid van de functie gemaximaliseerd.
De support vector (regressie) functie is:

$$g(x) = b_0 + b^T x, \tag{3}$$

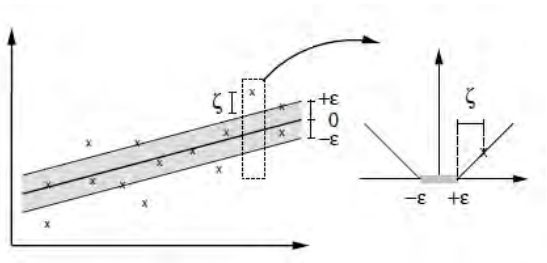
hierbij wordt x gevormd door de attributen en is b een vector die aan elk attribuut een bepaald gewicht hangt met ||b||=1. Tot slot is b_0 de bias, die de onzuiverheid van de schatter corrigeert. Deze functie wordt gevonden door het volgende probleem op te lossen [24]:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\zeta_i + \zeta_i^*) \\ \text{volgens} \quad & y_i - \langle w, b\rangle - \beta_0 \leq \epsilon + \zeta_i \\ & \langle w, x_i\rangle + \beta_0 - y_i \leq \eta + \zeta_i^* \\ & \zeta_i, \zeta_i^* \geq 0. \end{aligned}$$

Hierbij zijn de ζ gebruikt om de errors aan te geven, indien het niet mogelijk is om alle uitkomsten binnen de tube te laten vallen, zie figuur 22. De ζ is gedefinieerd als

$$|\zeta|_\epsilon = \begin{cases} 0 & \text{als } |\zeta| \leq \eta \\ |\zeta| - \eta & \text{anders.} \end{cases}$$

Dit optimalisatieprobleem kan worden opgelost met Langrange multipliers. Gezien de complexiteit om het hypervlak te vinden, is dit een kwadratisch probleem (QP) en is daarom een NP-compleet probleem. De rekentijd voor het trainen van het model is ωm^2 wat betreft de grootte van de trainingset[23].



Figuur 22: Afwijkingen bij regressie probleem [24].

Voor de iteraties is gebruik gemaakt van de package $e1071$ in R om de SVRM te trainen. Er is gevariëerd met de waarde van ϵ, de resultaten zijn verkregen met $\epsilon = 0.1$.

6.3 5-Fold cross-validation

Gezien de rekentijd die exponentieel toeneemt wanneer het aantal observaties stijgt, is het niet wenselijk de complete dataset te gebruiken voor het trainen en testen van het model. Vandaar dat we een willekeurige subset nemen uit de dataset. We hebben het model getraind op 10.000 observaties en getest op 5.000 observaties. Om het model minder gevoelig te maken voor overfitting is gebruik gemaakt van de 5-Fold cross-validation. Bij deze methode wordt de data in 5 subsets verdeeld, waarvan 1 groep als testset wordt gebruikt en de overige 4 groepen als trainingset. Door dit 5 maal te herhalen is elke subset gebruikt als testset. Uit de training- en testset is een willekeurige deelset genomen, zodat de trainingset uit 10.000 observaties bestaat en de testset uit 5.000.

10 Conclusie

In deze thesis is een model gemaakt dat de treinvertraging voorspelt aan de hand van historische gegevens met behulp van onder andere Support Vector Machines. In dit hoofdstuk zullen de voornaamste conclusies worden besproken.

Twee modellen

Het huidige model is door middel van twee methodes verbeterd. Bij de eerste methode is de werkelijke rijtijd geschat welke is gebruikt om te zien hoeveel vertraging ingelopen zal worden. Bij de tweede methode is de SVRM, ofwel een SVM voor regressie problemen, toegepast. Het trajectspecifieke SVM model uit hoofdstuk 8.3 gaf de meest betrouwbare voorspelling, dit geeft aan dat het traject een zeer belangrijk attribuut is. Dit model is per traject apart getraind en getest, wat betekent dat het bestaat uit 15.000 kleine modellen.

Attributen

Bij beide modellen zijn verschillende attributen toegevoegd. De attributen die een waardevolle toevoeging hadden voor het resultaat zijn in beide modellen gelijk, namelijk traject, treinserie, initiële vertraging en geplande rijtijd.
Er zijn diverse attributen getest welke geen significante verbetering gaven. Zo bleek het wel of niet rijden in de spits geen verschil te maken op het verloop van vertraging. Andere attributen waren toegevoegd om meer informatie te verkrijgen over het rijgedrag van de trein en de situatie in de omgeving. De reden dat deze attributen geen toevoeging zijn voor het model kan te maken hebben met de complexiteit en variabiliteit van de situatie, waardoor vooraf geen goed beeld te vormen valt. Een toelichting hierop is te vinden in hoofdstuk 9.3.

Verbetering

Om het resultaat en de verbetering ten opzichte van het huidige model te evalueren is gebruik gemaakt van een REC-plot, welke de nauwkeurigheid uitzet tegen een zekere toegestane afwijking in de voorspelling. Er is te zien dat met een toegestane afwijking van 1 minuut ongeveer 67% van het huidige model een juiste voorspelling geeft, tegenover 88% bij het trajectspecifieke SVM model. De mediaan van de rijtijdpercentielen geeft in 85% van de gevallen een juiste voorspelling.

Met behulp van de Wilcoxon-test is gebleken dat de resultaten van de drie modellen allen significant van elkaar verschillen, waardoor geconcludeerd kan worden dat het trajectspecifieke SVM model het meest betrouwbare resultaat geeft.

Vergelijking tussen de modellen

Het resultaat van de rijtijdmediaan gaf na het SVM model het meest betrouwbare resultaat. Er zijn daarom diverse testen uitgevoerd om de resultaten van deze twee modellen te vergelijken. Hiervoor is de betrouwbaarheid van de modellen per traject onderzocht. Hieruit kan worden besloten om geen voorspelling te maken voor een traject wanneer de betrouwbaarheid op dat traject te laag is en te wachten op een volgend meetpunt waarbij de betrouwbaarheid van de voorspelling is gestegen.
Op alle geteste aspecten - korte en lange geplande rijtijd, korte en lange initiële rijtijd - gaf de SVM een beter resultaat. De mediaan van de betrouwbaarheid is op deze vlakken hoger en er

zijn minder negatieve uitschieters te zien. Over het algemeen wordt de betrouwbaarheid bij beide modellen groter voor een traject met een kortere geplande rijtijd en waarbij minder vaak een grote vertraging plaatsvindt.

Beperkingen van het model

Een nadeel van het trajectspecifieke SVM model is dat het minder inzichtelijk is dan de rijtijdmediaan, aangezien niet duidelijk is wat het model heeft geleerd. Daarnaast kost het trainen van het model veel tijd ten opzichte van de rijtijdmediaan. Aangezien de dienstregeling jaarlijks wijzigt zal minimaal jaarlijks een nieuw model moeten worden getraind. Aangezien in de eerste weken van het nieuwe jaar geen data bekend is waarop het model kan trainen, is bekeken wat de nauwkeurigheid is van het huidige SVM model in het volgende dienstregeljaar. Het onderzoek is getraind op 23 dagen in het dienstregeljaar 2014 en is getest op dagen van het dienstregeljaar 2015. Uit dit resultaat bleek dat de voorspelling weliswaar minder nauwkeurig wordt, maar alsnog een betrouwbare basis kan vormen tot een nieuw model getraind is. Dit is niet getest voor de rijtijdmediaan, maar gezien de eerdere resultaten wordt verwacht dat ook dit model in dienstregeljaar 2015 beter zal presteren dan het huidige model.

De modellen zijn gebaseerd op treinen die volgens dienstregeling reden. Treinen waarvan de dienstregeling vooraf zijn gewijzigd zijn niet meegenomen. De vertraging voor deze treinen kan wel worden getest. Het is mogelijke dat de voorspelling voor deze treinen een sterkere afwijking heeft, hier zal verder onderzoek voor nodig zijn.

Verder onderzoek

In verder onderzoek kan onderzocht worden of het gebruik van een ander machine learning algoritme tot betere resultaten leidt. In een eerste test zijn enkele algoritmen met elkaar vergeleken, maar het is mogelijk dat een ander algoritme nieuwe inzichten geeft.
Veel van de toegevoegde attributen bleken geen significante verbetering te zijn in het model. Het is mogelijk dat een verfijning hiervan wel tot verbetering leidt.
In dit onderzoek zijn de rijtijden bij de dienstregelpunten als input gebruikt. Een nadeel hiervan is dat de afstand tussen deze punten niet gelijk is en het rijgedrag onbekend is tot de trein een volgend punt is gepasseerd. Er zijn GPS signalen beschikbaar, waardoor het rijgedrag vaker te controleren is en in een eerder stadium een voorspelling gemaakt kan worden. Het nadeel van GPS signalen is dat deze niet in het hele land beschikbaar zijn, en alleen op de dienstregelpunten de geplande passeertijd bekend is. Verder onderzoek kan dit uitbreiden en toepassen om te zien of dit de voorspelling verbetert.

Aanbeveling

Er zijn twee modellen aangedragen die een significatie verbetering geven op het huidige model, wat aanleiding geeft om het huidige model te vervangen. Het trajectspecifieke SVM model gaf het beste resultaat, echter is mijn aanbeveling om het rijtijdmediaan model te implementeren. De kracht van dit model zit in de eenvoud qua gebruik en uitlegbaarheid, wat de overgang kan bemoedigen. Er zitten meer negatieve uitschieters in dit model vergeleken met het SVM model, maar over het algemeen geeft dit model een sterke verbetering in de nauwkeurigheid wat de klanttevredenheid kan helpen te vergroten.

Appendices

A Aanmaken attributen

Bij de probleemstelling zijn enkele attributen bespoken. Hieronder wordt beschreven hoe deze zijn opgesteld.

Initiële vertraging

De initiële vertraging bestaat uit het verschil tussen de geplande doorkomst en de werkelijke doorkomst. Wanneer de trein te vroeg langs het meetpunt is gekomen, en er een negatieve vertraging is, wordt deze op 0 gesteld. We behandelen hierdoor een negatieve vertraging gelijk aan een situatie waarbij geen vertraging was.

$$\text{initiële vertraging} = \max\left(\text{tijd}^{i}_{werkelijk} - \text{tijd}^{i}_{gepland}, 0\right)$$

Uiteindelijke vertraging

Deze vertraging bestaat ook uit het verschil tussen de geplande en werkelijke tijd waarop de trein is aangekomen of vertrokken bij het station dat voorspeld moest worden. Aangezien we geen negatieve vertragingen voorspellen worden ook hierbij deze op 0 seconden vertraging gesteld.

$$\text{uiteindelijke vertraging} = \max\left(\text{tijd}^{u}_{\text{werkelijk}} - \text{tijd}^{u}_{\text{gepland}}, 0\right)$$

Geplande en werkelijke rijtijd

Door het verschil in geplande doorkomsttijd en werkelijke doorkomsttijd bij het meetpunt en de te voorspellen actie te nemen wordt de geplande rijtijd en werkelijke rijtijd bepaald. Hierop is een aanpassing noodzakelijk wanneer de trein te vroeg het meetpunt of bij de te voorspellen actie passeerde. De rijtijd is langer dan gepland wanneer de trein te vroeg langs het meetpunt kwam, terwijl dit geen negatieve consequentie had. Eenzelfde situatie geldt wanneer de trein te vroeg aankwam of vertrok. Deze verschillen worden aangepast in de rijtijd, zoals eerder is gedaan bij de vertraging.
Als voorbeeld wordt hieronder weergegeven hoe de geplande rijtijd is opgebouwd:

$$\text{geplande rijtijd} = \text{tijd}^{u}_{\text{gepland}} - \text{tijd}^{i}_{\text{gepland}}.$$

$$\begin{aligned}\text{werkelijke rijtijd} = &\max\left(\text{tijd}^{u}_{\text{werkelijk}}, \text{tijd}^{u}_{\text{gepland}}\right)\\ &-\max\left(\text{tijd}^{i}_{\text{werkelijk}}, \text{tijd}^{i}_{\text{gepland}}\right).\end{aligned}$$

Treinhistorie

De treinhistorie betreft de toename in vertraging in een eerder traject, namelijk het traject vanaf de laatst gepasseerde aankomst of vertrek tot aan het meetpunt. Door niet een laatstgepasseerd doorrijpunt te nemen, wordt gezorgd dat de afstand over het traject over het algemeen niet klein is. Over een grotere afstand wordt verwacht dat een afwijking in de rijtijd meer zegt dan over een klein stuk van soms nog geen 2 minuten.

Het rijgedrag in een vorig traject is enkel interessant wanneer de initiële vertraging meer dan 30 seconden bevat. Een vertraging in het verleden heeft geen invloed meer op de toekomst wanneer deze is opgelost. Immers, de trein rijdt volgens dienstregeling en een situatie uit het verleden heeft daarop geen invloed meer. Deze attribuut wordt daarom enkel meegenomen wanneer de initiële vertraging meer dan 30 seconden bevat, dit betreft 56.7% van de data. Voor 30 seconden is gekozen, aangezien hierbij sprake is van een vertraging.

$$\text{treinhistorie} = \begin{cases} \text{initiële vertraging} - \text{vorige vertraging} & \text{als meetpunt}^i \neq \text{beginpunt treinserie} \\ \text{leeg} & \text{anders.} \end{cases}$$

Dit attribuut kan zowel de extra vertraging aangeven op het vorige traject of als een boolean aangeven of de trein in het vorige traject vertraging heeft ingelopen of opgelopen.

Trajecthistorie

Voor deze attribuut is onderzocht wat de extra vertraging op het traject is geweest voor eerdere treinen. Het traject vanaf het meetpunt tot de eerstvolgende actie wordt onderzocht en hierbij worden enkel treinen meegenomen die maximaal een uur voordat de betreffende trein het meetpunt heeft gepasseerd bij het te voorspellen station zijn aangekomen of vertrokken. Wanneer deze tijdsspanne wordt verkleind zal het voorkomen dat op verschillende trajecten te weinig treinen hebben gereden om een zinvolle conclusie uit te halen. Wanneer de tijdsspanne wordt vergroot is de kans groter dat een aanwezige blokkade immidels is verholpen. Van alle treinen die dit traject binnen de gegeven tijdsperiode hebben afgelegd wordt het aantal treinen meegegeven en de minimale extra vertraging die is opgelopen. Dit attribuut wordt slechts meegenomen wanneer ten minste twee treinen over het traject hebben gereden binnen de tijdsspanne.

$$\text{trajecthistorie} = \begin{cases} \min\limits_t \left(\text{tijd(t)}^u_{\text{werkelijk}} - \text{tijd(t)}^i_{\text{werkelijk}}\right) & \text{als conditie } A \text{ geldt} \\ \text{leeg} & \text{anders.} \end{cases}$$

$$\text{conditie } A \quad = (\text{aantal treinen dat traject in afgelopen uur hebben afgelegd } t \geq 2)$$

Dit attribuut kan worden weergegeven als de minimale extra vertraging die op het traject is opgelopen, of als een boolean waarbij wordt aangegeven dat er een blokkade is. Er wordt daarom eerst onderzocht vanaf welke mate van toename van vertraging er een vergrootte kans is van een oploop.

Tussentijd

Het attribuut tussentijd geeft het verschil aan tussen de geplande en werkelijke tussentijd van de trein met de voorgaande trein. Dit attribuut is niet veelzeggend wanneer het huidige meetpunt bij een station is, aangezien hier meerdere treinen tegelijk aanwezig zijn die geen hinder van elkaar hoeven te ondervinden. Vandaar dat enkel de data van meetpunten bij overige dienstregelpunten - niet bij een station - worden meegenomen. Daarnaast dient de geplande tussentijd kleiner te zijn dan 10 minuten, aangezien de kans dat het verschil tot hinder leidt te klein is.
Dit attribuut kan zowel het absolute verschil geven, als het relatieve verschil waarbij het verschil is gedeeld door de geplande tussentijd.

$$\text{als conditie } B \text{ geldt} = \begin{cases} \text{geplande tussentijd} & = \text{tijd(t)}^i - \text{tijd(trein voor t)}^i) \\ \text{werkelijke tussentijd} & = \text{tijd(t)}^u - \text{tijd(trein voor t)}^u) \\ \text{verschil tussentijd} & = \text{geplande tussentijd} - \text{werkelijke tussentijd}) \end{cases}$$

$$\text{conditie } B \quad = \text{dienstregelpunt}^i = \text{'D' \& tijd}^i \geq 30 \text{ seconden}$$

Overige attributen

De overige attributen die in het hoofdstuk Probleemstelling zijn genoemd betreffen een toewijzing welke aan alle data kan worden toegevoegd. Wanneer het tijdstip van geplande doorkomsttijd binnen de spitsuren valt, is de boolean 'Spits' gelijk aan 1.
Het traject wordt gekenmerkt door de combinatie van het meetpunt, het type meetpunt, het te voorspellen station en het type daarvan. Zo staat : DAmfVUt voor doorkomst (D) bij Amersfoort (Amf) tot het vertrek (V) bij Utrecht (UT).

B Dataset

B.1 Dataset dienstregeljaar 2014

Weekdag	Datum	Punctualiteit in %	Uitval in %	# grote verstoringen
Maandag	6 januari	87.9	1.1	4
Woensdag	22 januari	88.0	2	11
Dinsdag	11 februari	90.9	1.1	6
Donderdag	27 februari	89.1	3.2	6
Vrijdag	7 maart	87.9	1.1	9
Zaterdag	15 maart	95.5	0.2	3
Woensdag	26 maart	92.2	2	6
Donderdag	17 april	89.9	1.3	3
Zondag	27 april	92.9	0.6	5
Vrijdag	9 mei	89.6	1.9	11
Maandag	19 mei	85.9	2.7	15
Zaterdag	31 mei	94.2	0.7	4
Zondag	15 juni	92.0	1.5	3
Dinsdag	24 juni	92.2	1.4	6
Maandag	7 juli	92.7	1.1	3
Woensdag	23 juli	85.9	4	13
Donderdag	7 augustus	94.6	0.6	5
Dinsdag	19 augustus	89.9	2.7	4
Woensdag	10 september	92.0	1.8	12
Vrijdag	19 september	90.7	0.7	3
*Donderdag	25 september	90.7	0.6	5
*Zaterdag	4 oktober	88.6	1.1	6
Vrijdag	17 oktober	85.7	2.4	10
Zaterdag	8 november	90.9	0.6	3
Zondag	16 november	85.8	1.6	9
*Maandag	24 november	85.6	1.6	9
*Zondag	7 december	87.6	0.8	2
Gemiddeld		**Punctualiteit in %**	**Uitval in %**	**# grote verstoringen**
Over 2014		89.7%	2.0%	6.8
Over data set		90.0	1.4	6.4

Tabel B30: Gegevens data set 2014

B.2 Data dienstregeljaar 2015

Weekdag	Datum	Punctualiteit in %	Uitval in %	# grote verstoringen
Zaterdag	20 december 2014	89.5	1.3	2
Donderdag	22 januari 2015	85.6	7.2	14
Woensdag	11 februari 2015	92.2	1.0	6
Zaterdag	7 maart 2015	89.3	7.6	6
Maandag	27 april 2015	85.2	0.9	7

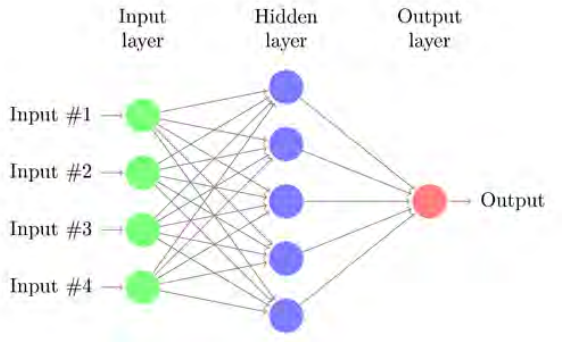
Tabel B31: Gegevens data set 2015

C Machine learning technieken

Zoals genoemd in hoofdstuk 6.2 zijn er verschillende technieken onderzocht voordat besloten is om verder te gaan met Support Vector Machines. In deze bijlage zullen deze methodes worden toegelicht.

Multilayer perceptron

Een multilayer perceptron is een neuraal netwerk en is een uitbreiding op Rosenblatt's perceptron en heeft als toevoeging één of meerdere verborgen lagen [15], zie figuur C56. Het aantal verborgen lagen maakt de classificatie danwel regressie complexer en specifieker.
De MLP is toegepast met behulp van de package RSNNS in R. Hierbij wordt het model gemaakt met behulp van back-propagation. De learning rate is gesteld op 0.1, net als de afwijkingsnorm. Deze norm stelt dat een afwijking kleiner dan die waarde niet erkend wordt.



Figuur C56: Voorbeeld multilayer perceptron

Decision Tree

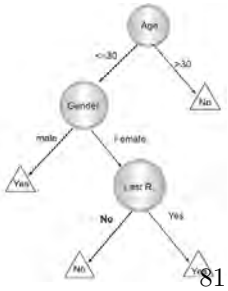
Decision trees, ofwel beslissingsbomen, worden onder andere gebruikt vanwege het grote voordeel dat ze erg inzichtelijk zijn [21]. Een voorbeeld van een beslissingsboom staat in figuur C57, waarin onder anders aan de hand van leeftijd en geslacht wordt bepaald of een e-mail wel of niet direct wordt beantwoord. De leafs of bladeren van de boom zijn 'Yes' en 'NO' en bovenstaande splitsingen zijn de knopen. Dit voorbeeld is een classificatie boom, in dit onderzoek is gebruik gemaakt van een regressie boom. In dit geval voorspeld een blad een getal in plaats van een klasse. In R is gebruik gemaakt van de package RPART, welke het CART-algoritme (Classification and Regression Trees) gebruikt. Het criteria waarop wordt gesplitst wordt bepaald aan de hand van de least squares error (anova methode) oftwel:

$$SS_T - (SS_L + SS_R), met SS_T = \sum (y_i - \bar{y})^2. \tag{5}$$

SS_T is de sum of squares van de knoop en SS_L en SS_R die van de linker en rechter zoon. Het blad krijgt de gemiddelde waarde en de afwijking van de knoop is de variantie van anova. Voor een specifieke observatie is de afwijking ($y_{voorspeld}$ - $\bar{y}$).

Random Forest

Random Forests is een verzameling van B individuele decision trees T_b [21], [**?**]. Voor elke boom wordt een bootstrap sample van grootte n genomen. Vervolgens wordt bij elk blad in de boom een willekeurige selectie gemaakt van de attributen waarop de beslissing wordt gebasseerd. Door de bootstrap sample en de willekeurige selectie van attributen wordt de correlatie tussen de bomen geminimaliseerd[10]. Vervolgens wordt de voorspelling gebasseerd



Figuur C57: Voorbeeld classificatie boom ([21])

op het gemiddelde van de uitkomsten:

$$\hat{f}_{rf} = \frac{1}{B} \sum_{b=1}^{B} T_b. \tag{6}$$

Het voordeel van Random Forest ten opzichte van een enkele decision tree, is dat de variantie wordt verkleind en daarmee de nauwkeurigheid wordt vergroot [21].
Random Forest zijn in R toegepast met behulp van de package 'randomForest'.

Referenties

[1] Defenitie treintypes. `http://www.ns.nl/over-ns/wat-doen-wij/ontdek-ns/treinen`. Accessed: 2015-02-17.

[2] Geschiedenis ns. `http://www.ns.nl/over-ns/wie-zijn-wij/profiel/geschiedenis`. Accessed: 2015-02-19.

[3] Jaarverslag prorail 2013. `http://www.jaarverslagprorail.nl/`. Accessed: 2015-02-17.

[4] Knmi klimatologie oktober 2014. `http://www.knmi.nl/klimatologie/maand_en_seizoensoverzichten/maand/okt14.html`. Accessed: 2015-05-18.

[5] Ns in cijfers. `http://www.ns.nl/over-ns/wat-doen-wij/spoorsector/verantwoordelijkheden-op-het-spoor`. Accessed: 2015-02-19.

[6] Jaarverslag ns 2014. 2014.

[7] Ana Isabel Rojão Lourenço Azevedo. Kdd, semma and crisp-dm: a parallel overview. 2008.

[8] Annabell Berger, Andreas Gebhardt, Matthias Müller-Hannemann, and Martin Ostrowski. Stochastic delay prediction in large train networks. *ATMOS*, 20:100–111, 2011.

[9] Jinbo Bi BIJ, RPI EDU, and Kristin P Bennett BENNEK. Regression error characteristic curves. In *Twentieth International Conference on Machine Learning (ICML-2003). Washington, DC*, 2003.

[10] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[11] John Brouwer. Voorspelmodel vertragingen – historische gegevens gebruiken om vertragingen op het spooor te voorspellen. *Stageverslag NS*, 2012.

[12] Thorsten Büker and Bernhard Seybold. Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning & Management*, 2(1):34–50, 2012.

[13] Malachy Carey and Andrzej Kwieciński. Stochastic approximation to the effects of headways on knock-on delays of trains. *Transportation Research Part B: Methodological*, 28(4):251–267, 1994.

[14] Francesco Corman, Rob MP Goverde, and Andrea D'Ariano. Rescheduling dense train traffic over complex station interlocking areas. In *Robust and Online Large-Scale Optimization*, pages 369–386. Springer, 2009.

[15] Simon S Haykin, Simon S Haykin, Simon S Haykin, and Simon S Haykin. *Neural networks and learning machines*, volume 3. Pearson Education Upper Saddle River, 2009.

[16] Tijs Huisman and Richard J Boucherie. Running times on railway sections with heterogeneous train traffic. *Transportation Research Part B: Methodological*, 35(3):271–292, 2001.

[17] Pavle Kecman and Rob MP Goverde. An online railway traffic prediction model. In *RailCopenhagen2013: 5th International Conference on Railway Operations Modelling and Analysis, Copenhagen, Denmark, 13-15 May 2013*. International Association of Railway Operations Research (IAROR), 2013.

[18] Lukasz A Kurgan and Petr Musilek. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01):1–24, 2006.

[19] Joseph O Ogutu, Hans-Peter Piepho, and Torben Schulz-Streeck. A comparison of random forests, boosting and support vector machines for genomic selection. In *BMC proceedings*, volume 5, page S11. BioMed Central Ltd, 2011.

[20] Edgar Osuna and Federico Girosi. Reducing the run-time complexity of support vector machines. In *International Conference on Pattern Recognition (submitted)*. Citeseer, 1998.

[21] Bruce Ratner. *Statistical modeling and analysis for database marketing: effective techniques for mining big data*. CRC Press, 2004.

[22] Lior Rokach. *Data mining with decision trees: theory and applications*. World scientific, 2007.

[23] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226, 2000.

[24] Greg Schohn and David Cohn. Less is more: Active learning with support vector machines. In *ICML*, pages 839–846. Citeseer, 2000.

[25] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[26] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[27] Masoud Yaghini. Railway passenger train delay prediction via neural network model. *Journal of advanced transportation*, 47-3:355–368, 2013.

[28] Bin Yu, Zhong-Zhen Yang, Kang Chen, and Bo Yu. Hybrid model for prediction of bus arrival times at next station. *Journal of Advanced Transportation*, 44(3):193–204, 2010.