Adviesrapport: Data-driven Business

Yujian Jiang, Max Jansen

July 2025

ProRail

Inhoudsopgave

1 Inleiding

1.1 Achtergrond van het project

Voor het vak DataDriven Business is ProRail naar de HU gekomen met de vraag of wij hulp kunnen bieden voor het DataLab. Het DataLab is een afdeling van ProRail, waar in grote hoeveelheden data verzameld wordt, wat daarna geanalyseerd en verwerkt wordt om verbeteringen toe te passen op het spoorwegnet. Het spoorwegnet is erg groot en er zullen altijd wel problemen opduiken van klein naar groot. Dit is voor iedereen erg vervelend. ProRail is daardoor veel bezig met het oplossen van de problemen en de kosten lopen daardoor ook op. En voor de reizigers is het ook vervelend. Die moeten omreizen en zullen zich ergeren als het herhalend optreed. Het is dus van groot belang dat problemen snel en efficiënt worden opgelost om zowel kosten te besparen als de tevredenheid van de reizigers te behouden.

1.2 Het belang van een snelle oplossing

Het is belangrijk dat er snel gehandeld moet worden. Daarvoor moet er wel duidelijk zijn wat de problemen zijn en hoe lang die gaan duren, zodat er vooruit gepland kan worden om de minste vertragingen op te lopen. Als er bijvoorbeeld een storing is, moeten de planners weten hoe lang het ongeveer gaat duren voordat het opgelost is, zodat ze alternatieve routes kunnen plannen of vervangend vervoer kunnen regelen. Nu is er dus gevraagd of er hulp geboden kan worden bij het verduidelijken van de problemen. Er is gevraagd of het mogelijk is om te voorspellen hoe lang een storing gaat duren zodat er op tijd en beter omheen gepland kan worden en de vertragingen zo beperkt mogelijk kunnen blijven.

2 Opdracht

2.1 Doel van de opdracht

De opdracht die wij gekregen hebben is om uit te zoeken of het mogelijk is om te voorspellen hoe lang een storing gaat duren. Dit willen ze gaan doen door gebruik te maken van een applicatie die dit kan voorspellen. Op dit moment is de applicatie er nog niet. Die gaan wij maken. Maar voor dat we dat kunnen doen moeten wij inzicht krijgen over het gehele proces. We weten namelijk niet hoe ze bij ProRail aan het werk gaan. Ook moet de data grondig doorgenomen worden. De data moet worden geanalyseerd, opgeschoond en voorbereid voor de modellen die de voorspellingen gaan doen. Dan moet de applicatie gemaakt worden. Hoe het er uit gaat zien, welk model er gekozen word, en hoe het gebruikt gaat worden. Als laatste moet er gedocumenteerd worden. Er moet een duidelijke uitleg zijn, en alles moet gerapporteerd worden. Omdat het proces behoorlijk groot is, gaan we het verdelen in verschillende delen, wat uit eindelijk een geheel moet vormen.

2.2 Deelopdrachten

Hieronder zullen verder in gaan op onze deelopdrachten. Alles valt beter te begrijpen als het individueel uitgelegd kan worden.

2.2.1 Business Understanding

Het eerste deel is de Business Understanding, zodat ons duidelijk wordt waar wij onze taken moeten vervullen en hoe het proces loopt. We doen namelijk een opdracht voor een extern bedrijf waar wij nergens bekend mee zijn. Ons moet duidelijk worden wat er van ons allemaal gevraagd word. We moeten weten wie de stakeholders zijn, waar onze data vandaan komt, en wat de knelpunten zijn. Maar ook hoe hun hele proces loopt, hoe het bedrijf aan het werk gaat, wat onze rol is. Pas als wij snappen hoe ProRail werkt. Kunnen wij een opdracht voor ze maken.

2.2.2 Data

Als het eenmaal duidelijk is wat we moeten gaan doen, is de volgende stap het begrijpen van de data. Zonder de data kunnen we namelijk geen voorspellingen maken. De data is erg ingewikkeld en ver van schoon. We hebben 58 kolommen met allemaal onduidelijke namen. Ook zijn er $\pm$ 800.000 rijen waar we mee werken, en er staan heel veel lege/verkeerde waarden in. Het is dus een zooitje. Nu is het belangrijk om te begrijpen wat alles is, wat er uit gefilterd mag worden en waar we mee doorgaan. Daarnaast moet er goed worden geanalyseerd, zodat er vervolgens drie modellen uitgewerkt kunnen worden.

2.2.3 Applicatie

Het belang van de opracht is een werkende applicatie. Dit gaat namelijk gebruikt worden door het personeel in de meldkamer, waar de storingen verwerkt worden. Hier komt het hele project samen. De applicatie gaat een van de gekozen modellen bevatten die de voorspelling gaat doen. Die gecreëerd is uit de analyse van de data. Maar ook omdat we inzicht hebben gecreëerd voor het bedrijf. Ook kan je alle oude data terug vinden om te kijken hoe de oude storingen waren, maar ook hoe lang het duurde voordat ze waren opgelost.

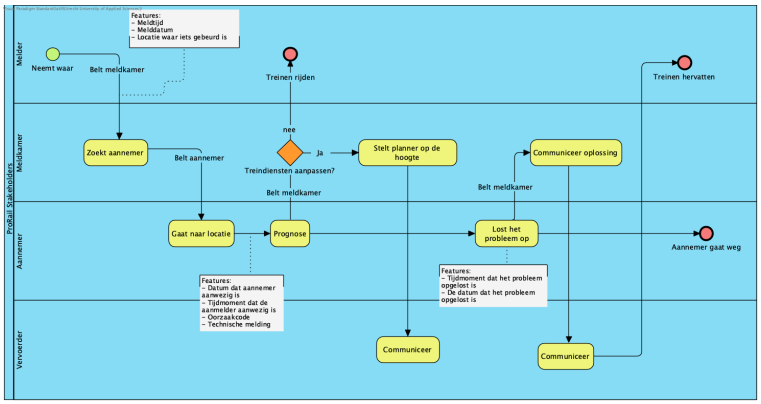
2.2.4 Documentatie

Om het project af te ronden is er goede documentatie nodig. Je kan wel een werkend product neer zetten maar als het niet duidelijk is hoe het gebruikt moet gaan worden kom je niet ver. Er wordt een duidelijke uitleg verwacht van hoe de applicatie gebruikt moet worden, maar ook hoe het model tot stand gekomen is. Er moet namelijk wel een goede reden zijn om geld en tijd te investeren in deze applicatie.

3 Business Understanding

3.1 Process

De eerste stap in het project, is het in kaart brengen van het probleem, en de oplossing waar we naartoe moeten werken. Om het probleem in kaar te brengen zijn wij druk bezig geweest met de mensen van ProRail om alles in kaart te brengen. Het proces begint wanneer iemand (de melder) een probleem opmerkt en contact opneemt met de meldkamer. De meldkamer zoekt vervolgens een aannemer die beschikbaar is en belt deze op. Zodra de aannemer onderweg is, maakt hij een prognose van het probleem. Op basis daarvan wordt bepaald of de treindienstregeling moet worden aangepast. Als dat zo is, wordt de planner op de hoogte gebracht. De aannemer komt aan op locatie en lost het probleem op. Tijdens dit hele proces worden belangrijke gegevens vastgelegd, zoals de oorzaak, locatie en duur van de storing. Zodra het probleem is opgelost, communiceert de meldkamer dit terug naar de betrokken partijen. De aannemer vertrekt en als alles in orde is, worden de treinen weer hervat. Dit hele proces is belangrijk om te snappen, omdat wij hierdoor weten op welk moment er welke data beschikbaar is, en waar onze applicatie het meest van waarde kan zijn.



Figuur 1: BPMN (Business Process Model and Notation)

3.2 Stakeholders

Tijdens het project hebben we gekeken naar wie er allemaal betrokken zijn bij het proces rondom een storing, en wie er dus belang hebben bij een betere manier van voorspellen. Deze belanghebbenden (oftewel stakeholders) zijn de mensen of partijen die direct of indirect geraakt worden door een storing, en dus ook baat hebben bij een goede oplossing. Omdat we te maken hebben met een groot proces met meerdere stappen en afdelingen, zijn er best veel belanghebbenden.

- **Meldkamer van ProRail:** Hier begint het hele proces. De meldkamer ontvangt de melding van de storing en moet direct actie ondernemen. Ze

schakelen een aannemer in en houden contact met alle betrokkenen. Voor hen is het handig om zo snel mogelijk te weten hoe lang iets ongeveer gaat duren, zodat ze kunnen schakelen.

- **Planners van ProRail:** Zodra een storing invloed heeft op het spoor, moeten planners beslissen of er een aangepaste dienstregeling nodig is. Hoe beter ze weten hoe lang een storing gaat duren, hoe beter ze dat kunnen inschatten. Zonder goede informatie is het vaak gokken, en dat zorgt voor extra overlast.

- **Aannemers:** De mensen die daadwerkelijk naar de storing toegaan en het probleem oplossen. Zij geven belangrijke informatie door, zoals de oorzaak en wanneer ze ter plekke zijn. Deze info gebruiken wij in onze modellen om te voorspellen hoe lang het nog gaat duren.

- **Vervoerders:** Als een storing impact heeft op het treinverkeer, zijn de vervoerders daar de dupe van. Zij moeten hun treinen en personeel opnieuw inplannen en zorgen dat reizigers weten waar ze aan toe zijn. Hoe sneller zij weten wat de impact is, hoe sneller ze kunnen schakelen.

- **Reizigers:** De mensen die elke dag met de trein reizen. Zij hebben natuurlijk geen invloed op het proces, maar ze merken de gevolgen wel direct. Door betere voorspellingen kunnen we de vertraging voor hen zoveel mogelijk beperken en beter communiceren wat er aan de hand is.

- **DataLab van ProRail:** Het DataLab is de afdeling binnen ProRail die zich bezighoudt met innovatie en data. Zij hebben ons gevraagd om dit project op te pakken. Ze willen weten of machine learning hierbij kan helpen, en of onze aanpak in de praktijk toegevoegde waarde biedt.

- **Ons projectteam (HU-studenten):** Tot slot zijn wij zelf ook belanghebbenden. Voor ons is het belangrijk dat het project duidelijk afgebakend is, dat de data bruikbaar is, en dat we goede feedback krijgen. Alleen dan kunnen wij iets maken wat echt gebruikt gaat worden.

Elke partij heeft een eigen rol en een eigen doel, maar uiteindelijk willen we allemaal hetzelfde: zo snel mogelijk een storing oplossen, met zo min mogelijk gedoe.

3.3 Knelpunten

Tijdens het project kwamen we verschillende knelpunten tegen, zowel in het proces als in de data. Deze punten zorgen ervoor dat het moeilijk is om direct een goed werkend model of systeem te maken. Hieronder de belangrijkste knelpunten die we zijn tegengekomen:

- **Inzicht in het proces:** In het begin was het onduidelijk hoe het proces precies verliep binnen ProRail. Omdat we extern zijn, wisten we niet

wat er allemaal gebeurt vanaf het moment dat een storing gemeld wordt tot aan de oplossing. Daardoor was het lastig om te bepalen waar onze applicatie echt waarde kon toevoegen.

- **Datakwaliteit:** De data waar we mee werken is allesbehalve perfect. Er zitten veel lege velden in, foutieve of inconsistente waarden, en soms ontbreekt essentiële informatie. Dit maakt het moeilijk om betrouwbare voorspellingen te doen. Ook is de data erg technisch en soms lastig te interpreteren.
- **Tijdstip van datatoegang:** Niet alle informatie is meteen beschikbaar wanneer een storing net gemeld is. Sommige data komt pas later in het proces binnen, bijvoorbeeld als de aannemer ter plekke is. Daardoor kunnen we niet altijd meteen een goede voorspelling doen.
- **Complexiteit van het probleem:** Het voorspellen van de duur van een storing is lastig, omdat elke storing anders is. De ene keer is het een kapotte bovenleiding, de andere keer een wisselstoring. Er zijn veel factoren die invloed hebben, en die zijn lang niet altijd duidelijk of meetbaar.
- **Gebrek aan standaardisatie:** Veel waarden in de data zijn handmatig ingevoerd en worden niet altijd op dezelfde manier geregistreerd. Dit zorgt voor ruis in de data en bemoeilijkt het trainen van een goed model.

Deze knelpunten hebben we in kaart gebracht zodat we er tijdens het project rekening mee konden houden. Het helpt ook om realistische verwachtingen te scheppen over wat wel en niet mogelijk is binnen de beschikbare tijd.

4 Data

4.1 Beschikbaarheid van de data

Het is dus ook belangrijk om te weten hoe de data in elkaar zit en wat we gaan gebruiken. Door het proces goed te begrijpen, kunnen we bepalen op welk moment in de workflow onze voorspellingen het meest nuttig zijn. Bijvoorbeeld, zodra een storing gemeld wordt, is er nog weinig informatie beschikbaar. Maar wanneer de aannemer ter plekke is, hebben we meer data om een nauwkeurige voorspelling te maken.

4.2 Eerste onderzoek

Daarna zijn we ons gaan focussen op de data. Dit is een van de belangrijkste onderdelen aangezien we moeten weten waar we mee werken. Er moet zekerheid zijn over de data of alles wel juist is. Er staan vaak namelijk verkeerde en/of zelfs lege plekken in de data. Als eerste zijn we gaan onderzoeken wat er allemaal in de data zit.

4.3 Gekozen kolommen

Op het eerste oog zijn wij breedschalig gaan kijken welke kolommen wij willen gebruiken. Uiteindelijk hebben we besloten om de volgende kolommen te gebruiken:

- stm_oorz_code: Dit is de oorzaakcode die geclassificeerd wordt door de aannemer. Het geeft aan wat de oorzaak is van de storing.
- stm_geo_mld: Dit is de geocode van waar het incident zich plaatsvindt. Hiermee weten we op welke locatie de storing is.
- stm_sap_melddatum: Dit is de datum van wanneer de melding gemaakt is. Dit kan invloed hebben, bijvoorbeeld bij seizoensgebonden problemen.
- stm_aanntpl_tijd: Dit is de tijd vanaf wanneer de aannemer aanwezig is. Vanaf dit moment kan de hersteltijd beginnen.
- stm_techn_mld: Dit is de categorie waar de incident bij hoort. Bijvoorbeeld of het een mechanisch probleem is of een elektrisch probleem.
- stm_prioriteit: Dit is de prioriteitcode waarmee aangegeven wordt hoe hoog de prioriteit ligt.

Deze kolommen hebben we gekozen omdat ze het meest relevant zijn voor het voorspellen van de hersteltijd. Ons target, oftewel wat we willen voorspellen, is targetherstel. Dit is de tijd die het duurt vanaf het moment dat de aannemer aanwezig is tot het moment dat het incident verholpen is.

4.4 Opschonen van de data

De volgende stap is om te zorgen dat er meer duidelijkheid komt over wat de data inhoudt. Het is opgevallen dat er inconsistenties zijn in de data, zoals ontbrekende waarden en foutieve invoer. We moeten deze issues oplossen om accurate voorspellingen te kunnen maken. We hebben daarom de data opgeschoond door ontbrekende waarden te vullen of te verwijderen waar nodig, en fouten te corrigeren. Bijvoorbeeld, als er een tijd ontbreekt, kunnen we die niet gebruiken in ons model. Ook hebben we ervoor gezorgd dat alle data in het juiste formaat staat, zoals het omzetten van datums naar numerieke waarden die het model kan begrijpen.

5 Modellen

5.1 Gebruikte algoritmes

Na het opschonen van de data zijn we begonnen met het bouwen van modellen. We hebben verschillende algoritmes getest om te kijken welke het beste resultaat zou geven.

5.1.1 Lineaire Regressie

Lorem Ipsum

5.1.2 Decision Tree

Lorem Ipsum

5.1.3 Random Forest

Lorem Ipsum

5.2 Resultaten van de modellen

Lorem Ispum

6 Applicatie

6.1 Ontwerp van de applicatie

Het model waarin de voorspelling wordt gemaakt is erg ingewikkeld en vereist technische kennis. Daarom is er een applicatie gemaakt waarin je op een gebruiksvriendelijke manier hetzelfde resultaat kunt behalen. Omdat een goede applicatie maken veel werk kost, zijn wij als eerste visuele en interactieve ontwerpen gaan maken om te laten zien hoe de applicatie eruit gaat zien. We hebben wireframes en prototypes gemaakt om het ontwerp te testen. De applicatie moet intuïtief zijn, zodat gebruikers zonder technische achtergrond er mee kunnen werken.

6.2 Implementatie en veiligheid

Na dit getest te laten hebben zijn wij begonnen met het uitwerken van de applicatie, die vervolgens op het web geplaatst wordt zodat de betrokkenen er altijd bij kunnen. Omdat de applicatie online staat hebben wij ons ook gefocust op de veiligheid, en is het dus niet toegankelijk voor iedereen. Er is een systeem gebouwd dat ervoor zorgt dat alleen de mensen met een account er toegang tot hebben. We hebben ook gebruik gemaakt van beveiligde verbindingen (SSL) om de data te beschermen. De applicatie bevat ook een helpsectie, zodat gebruikers snel antwoord kunnen vinden op hun vragen.

7 Conclusie

Door het combineren van data-analyse en machine learning is het mogelijk om nauwkeurige voorspellingen te doen over de hersteltijden van storingen. Dit stelt ProRail in staat om beter te plannen en de impact op reizigers te minimaliseren. De ontwikkelde applicatie maakt het eenvoudig voor gebruikers om deze voorspellingen te raadplegen en draagt bij aan een efficiënter spoorwegsysteem. Het decision tree model heeft bewezen het meest effectief te zijn en kan in de toekomst verder verbeterd worden met meer data.