Adviesraport

Gemaakt door: Cas Janssen & Daan Smienk

V2A

Groep 4



Inhoud

[1. ProRail 2](#_Toc180576062)

[1.1 Probleem 2](#_Toc180576063)

[1.2 Doelstelling 2](#_Toc180576064)

[1.3 Oplossing 2](#_Toc180576065)

[2. Data 3](#_Toc180576066)

[2.1 Beschrijving dataset 3](#_Toc180576067)

[2.2 Data Understanding 3](#_Toc180576068)

[2.3 Data Preparation 3](#_Toc180576069)

[2.4 Beschrijving opgeschoonde dataset 3](#_Toc180576070)

[3. Model 4](#_Toc180576071)

[3.1 Keuze voor bepaalde modellen 4](#_Toc180576072)

[3.2 Toelichting van hyperparameters 4](#_Toc180576073)

[3.3 Resultaten van experimenten 4](#_Toc180576074)

[3.4 Overzicht alle experimenten 4](#_Toc180576075)

[3.5 Evaluaties resultaten 4](#_Toc180576076)

[4. Dashboard 4](#_Toc180576077)

[4.1 Overzicht requirements 4](#_Toc180576078)

[4.2 Mock up 4](#_Toc180576079)

[4.3 Keuze ontwikkelomgeving 4](#_Toc180576080)

[4.4 Resultaten 4](#_Toc180576081)

[5. Conclusie en aanbeveling 5](#_Toc180576082)

[5.1 Conclusie 5](#_Toc180576083)

[5.2 Aanbeveling 5](#_Toc180576084)

# ProRail

## Probleem

ProRail heeft ons gevraagd om te helpen bij het inschatten van de tijd die nodig is om verschillende problemen op het spoor op te lossen. Dit kunnen uiteenlopende problemen zijn, zoals iemand die te dicht bij het spoor loopt of schade aan het spoor. Het probleem is dat het nu niet goed ingeschat kan worden, omdat het vooral door de aannemer wordt gedaan. Dit kan soms een goede inschatting zijn, maar soms klopt het ook niet. Daarom maken wij een model om dit beter te doen.

## Doelstelling

Voor dit project heeft ProRail een aantal eisen gesteld, zodat ze daadwerkelijk iets aan onze voorspelling hebben. Allereerst moeten we natuurlijk een zo nauwkeurig mogelijke schatting maken. Daarnaast moet het met een GUI duidelijk zichtbaar zijn hoe lang het duurt en hoe betrouwbaar die inschatting is.

## Oplossing

Voor dit project gaan we een model trainen dat de tijd zo nauwkeurig mogelijk kan inschatten, zodat ze direct weten hoelang het probleem zal duren en wanneer de treinen weer kunnen rijden.

*Figuur 1.1 Dit is het BPMN dus het proces van ProRail*

Afbeelding met tekst, diagram, schermopname

Automatisch gegenereerde beschrijving

Hier in kan je ook zien waar ons model in komt dus waar we onze schatting maken.

# Data

## Beschrijving dataset

We hebben een dataset gekregen met 139 kolommen. Veel van deze kolommen zijn gemarkeerd als n.v.t. of ?, wat betekent dat ze niet belangrijk zijn. Ook hebben we de regels met prioriteit 9 verwijderd. Dit komt omdat deze regels administratief zijn en geen functie herstel hebben. Ook dingen zoals kopieën, hoewel de dataset veel informatie bevat, is een groot deel hiervan niet bruikbaar voor onze modellen. Daarom zullen we veel kolommen verwijderen, wat we verder zullen uitleggen in de Data Preparation-fase.

## Data Understanding

Zoals al eerder gezegd, bevat deze dataset veel data. Door de data te onderzoeken, zijn we uitgekomen op een paar belangrijke kolommen. Deze kolommen kiezen we door overleg met PO. Daarnaast hebben we ook een correlatie tabel gemaakt, maar er waren geen correlaties sterke correlaties met de target variabele. Dit zijn de kolommen die de grootste invloed zullen hebben op het inschatten van de tijd. We hebben ook zelf een tabel toegevoegd aan de database; dit is de tijd die we gaan voorspellen. Dit betreft de tijd tussen het moment dat de aannemer zijn prognose maakt en het moment dat er weer functie herstel is. De belangrijke kolommen kunt u vinden in het notebook met de bijbehorende uitleg over wat de tabellen doen.

## Data Preparation

Er zijn veel keuzes gemaakt bij het opschonen van de database. We hebben kolommen uit de database verwijderd op basis van de volgende criteria: als het een beschrijvingen is, te veel lege plekken in de kolom, en geen toegevoegde waarde voor ons model. Een precieze lijst van de verwijderde kolommen kunt u vinden in het notebook. Ook hebben we nagedacht over het voorspellen van de tijd. Wij hebben uiteindelijk gekozen om 10 bins te maken die allemaal hetzelfde aantal datapunten heeft dit zorgt er voor dat je tijden die vaak voorkomen beter kan inschatten.

## Beschrijving opgeschoonde dataset

Na het opschonen van de database blijven er ongeveer 15 van de 139 kolommen over, en 48.000 van de 89.000 rijen. De belangrijkste kolommen voor onze modellen zijn: 'stm\_geo\_mld', 'stm\_prioriteit', 'stm\_oorz\_code', 'stm\_contractgeb\_gst', en 'stm\_techn\_mld'. Uitleg voor deze tabellen kunt u vinden in het notebook.

# Model

## Keuze voor bepaalde modellen

## Toelichting van hyperparameters

## Resultaten van experimenten

## Overzicht alle experimenten

## Evaluaties resultaten

# Dashboard

## Overzicht requirements

## Mock up

## Keuze ontwikkelomgeving

## Resultaten

# Conclusie en aanbeveling

## Conclusie

## Aanbeveling