Adviesraport

Gemaakt door: Cas Janssen & Daan Smienk



V2A

Groep 4

Inhoud

[1. ProRail 2](#_Toc181617094)

[1.1 Probleem 2](#_Toc181617095)

[1.2 Doelstelling 2](#_Toc181617096)

[1.3 Oplossing 2](#_Toc181617097)

[2. Data 3](#_Toc181617098)

[2.1 Beschrijving dataset 3](#_Toc181617099)

[2.2 Data Understanding 3](#_Toc181617100)

[2.3 Data Preparation 4](#_Toc181617101)

[2.4 Beschrijving opgeschoonde dataset 4](#_Toc181617102)

[3. Model 4](#_Toc181617103)

[3.1 Keuze voor bepaalde modellen 4](#_Toc181617104)

[3.2 Toelichting van hyperparameters 4](#_Toc181617105)

[3.3 Resultaten van experimenten 4](#_Toc181617106)

[3.4 Overzicht alle experimenten 4](#_Toc181617107)

[3.5 Evaluaties resultaten 4](#_Toc181617108)

[4. Dashboard 5](#_Toc181617109)

[4.1 Overzicht requirements 5](#_Toc181617110)

[4.2 Mock up 5](#_Toc181617111)

[4.3 Keuze ontwikkelomgeving 5](#_Toc181617112)

[4.4 Resultaten 5](#_Toc181617113)

[5. Conclusie en aanbeveling 5](#_Toc181617114)

[5.1 Conclusie 5](#_Toc181617115)

[5.2 Aanbeveling 5](#_Toc181617116)

# ProRail

## Probleem

ProRail is een bedrijf dat verantwoordelijkheid heeft voor alle sporen in Nederland voor de aanleg, het onderhoud, beheer en de veiligheid van het spoor. Dus moet ProRail problemen oplossen die te maken hebben met het spoor. Dit kunnen veel verschillende problemen zijn: dit kan iemand zijn die te dicht bij het spoor loopt of dat het spoor ergens kapot is. Het kost natuurlijk tijd om deze problemen te analyseren en op te lossen.

Nu wordt deze tijd ingeschat door de aannemers die het probleem gaan oplossen met vakmanschap. Deze prognose kan natuurlijk goed zijn, maar er wordt ook wel vaker een foute prognose gesteld. Volgens de gegevens in de database is de prognose van de aannemer 19% van de tijd goed.

## Doelstelling

Het doel van dit project is om de planners te ondersteunen bij het organiseren van de hervatting van het treinverkeer. Dit doen we door een nauwkeurige voorspelling te maken van hoe lang het duurt voordat de treinen weer kunnen rijden. Momenteel komt het vaak voor dat het probleem al is opgelost, terwijl de inschatting was dat het nog een uur zou duren, waardoor er geen personeel en trein klaarstaat. Dit gebeurt omdat aannemers vaak te pessimistisch zijn in hun prognoses.

Ons doel is om een meer precieze tijdsindicatie te geven, zodat de organisatie op de juiste momenten personeel en treinen gereed heeft. We gaan dit bereiken door een model te trainen en een GUI te ontwikkelen die de informatie weergeeft. Er zijn daarbij enkele eisen gesteld door ProRail, zodat ze de tool goed kunnen gebruiken. We moeten een GUI maken waarop de betrouwbaarheid en nauwkeurigheid van onze inschatting zichtbaar zijn, evenals de geschatte tijd.

## Oplossing

Om het probleem op te lossen, gaan we een model maken dat betrouwbaarder en nauwkeuriger is dan de prognoses van de aannemer. Dit model zal laten zien hoe nauwkeurig het is en hoe betrouwbaar het kan zijn. We doen dit door te kijken naar de percentages en te analyseren hoe vaak het model naast de juiste inschatting zit wanneer het fout is.

*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 bij het probleem aan komt 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 verschillende keuzes gemaakt bij het opschonen van de database. Kolommen zijn uit de database verwijderd op basis van de volgende criteria: als ze beschrijvende informatie bevatten, te veel lege waarden hadden, of geen toegevoegde waarde boden voor ons model. Een precieze lijst van de verwijderde kolommen is te vinden in het notebook.

Daarnaast hebben we overwogen hoe we de tijd het beste konden voorspellen. Uiteindelijk hebben we gekozen om 12 bins te maken met een gelijk aantal datapunten. Dit zorgt ervoor dat we veelvoorkomende tijden beter kunnen inschatten. Deze aanpak bracht echter het probleem met zich mee dat sommige bins te klein of juist te groot werden. Daarom hebben we 12 bins gemaakt waarbij geen enkele bin een tijdsduur van meer dan 2 uur heeft. Kleinere bins kunnen we indien nodig samenvoegen.

## 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

Voor de modellen die we naar hebben gekeken moest het model een probabilistische output te geven, aangezien dit gebruikt zal worden om een betrouwbaarheidspercentage te geven. Daarna hebben we gewoon een lijst van modellen getraind en zijn we verder gegaan met de modellen die de laagste loss heeft en de hoogsteaccuracy.

## Toelichting van hyperparameters

## Resultaten van experimenten

## Overzicht alle experimenten

## Evaluaties resultaten

# Dashboard

## Overzicht requirements

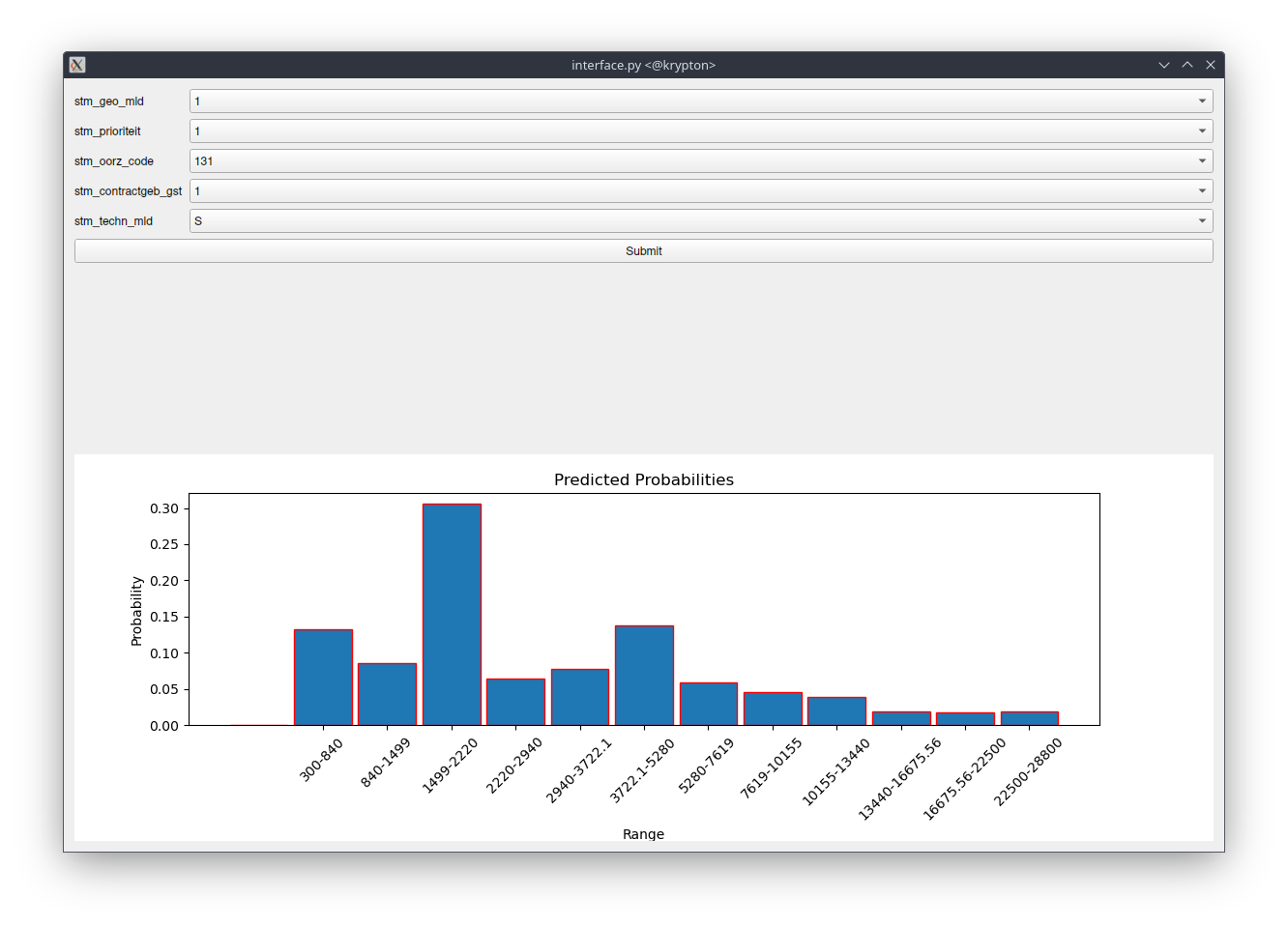
Het dashboard moet voldoen aan een aantal eisen. Ten eerste moet de gebruiker de feature variabelen kunnen invullen. Dit is van belang omdat de gebruiker anders niet een prediction kan krijgen.  
Ten tweede moet de dashboard een grafiek kunnen weergeven met informatie over de resultaten van het model. Deze informatie bestaat uit welke klassen het model voorspelt en hoe zeker hij is over die klassen. Dit is nodig zodat de gebruiker een geinformeerde beslissing kan maken over zijn keuze.

## Keuze ontwikkelomgeving

Er is gekozen om de interface in PyQT te maken. De reden hiervoor is omdat PyQT een breed scala aan customisability geeft in het maken van een GUI. Hiernaast is er ook ervaring hiermee in de ontwikkelteam, dit zorgt ervoor dat er gemakkelijk iets gemaakt kan worden. Ook had de PO aangegeven dat het niet uitmaakt wat voor GUI library gebruikt wordt.

## Resultaten

Met de requirements in gedachte zijn er 2 GUI’s gemaakt. De eerste is de GUI waar de gebruiken de feuture variabelen kan invullen en daar een voorspelling terug kan krijgen.



Hiernaast is er een tweede GUI gemaakt. Hierop is te zien waar de storing zich bevindt, en ook informatie over de storing zelf. Op de kaart zijn alle spoorwegen inclusief stations te zien. De stations zijn aangegeven met een blauwe markering en de storingen met een roode markering.  
De GUI is te zien op website die [hier te vinden is.](https://mightypiggie.github.io/data-driven-bussines-ai-2024/)

# Conclusie en aanbeveling

## Conclusie

## Aanbeveling