DDB Project 2022

- Adam Chebil

- 10-11-2022

1. Inleiding

Dit onderzoek wordt gedaan in opdracht van ProRail met als doel het in kaart brengen van de huidige situatie omtrent storingen en het opstarten van het treinverkeer na een storing. Er wordt hierbij uitgebreid analyse gedaan naar de data waarnaar er modellen worden opgezet die dit visueel ondersteunen. Dit model zal een voorspelling geven van de duur van de herstelperiode na een storing. We willen hiermee aantonen dat er een correlatie bestaan tussen de aspecten van een storing en zijn hersteltijd

Hiernaast zal er ook gekeken worden naar een hypothese, 'is de prognose van de aannemer te behoudend?' is de onderzoeksvraag van deze hypothese. De hypothese is 'De aannemer is te conservatief in zijn analyse over de hersteltijd.'

Dit onderzoek wordt uitgevoerd in opdracht van ProRail. Het is vaak voorkomend dat er storingen zijn op het spoor. Het probleem is dat het nog altijd niet helemaal precies is in te schatten hoe lang een storing zal duren. Er wordt hier gekeken of er verbeteringen kunnen worden gemaakt met de voorspelling.

Het proces begint eerst met de monteurs die een melding van het probleem ontvangen, met een schatting van hoe lang het gaat duren tot de storing is opgelost volgens de verwachtingen. Hierbij worden reizigers ingelicht over de vertraging. Er worden ook voorbereidingen gedaan over het voorzetten van de dienstregeling

De belangrijkste stakeholders voor dit project zijn op dit moment:

- De monteurs die de problemen oplossen
- De planners van ProRail
- Reizigers
- Meldkamer spoor
- Treindienstleiders
- Aannemer

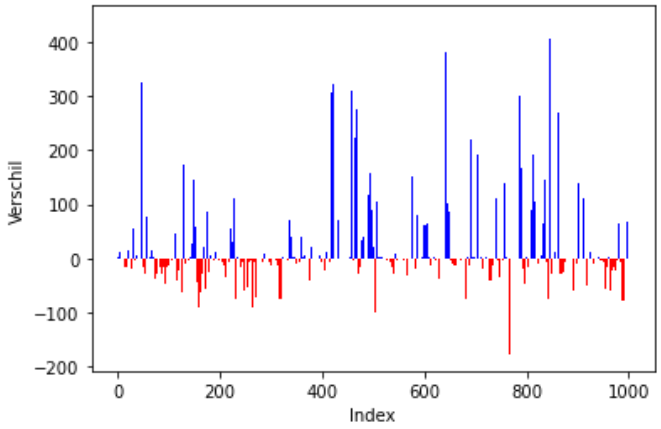
2. Data

Bij het analyseren van de data was al heel snel duidelijk dat er veel gegevens waren dat onbruikbaar was. Er was een aantal target variabelen die duidelijk niets betekende, en verwijderd konden worden. Hetzelfde geld voor de prognose tijd.

Er is gekeken naar welke kolommen nou daadwerkelijk bruikbaar zijn door eerst naar elke beschrijving te kijken. Vaak stond er in dat het niet van toepassing was of dat er niet verteld werd wat het betekende, deze konden dan ook verwijderd worden.

Daarbij is ook gekeken naar welke features te weinig data bevatten. Als er te veel NaN waardes stonden, maakte dat de feature onbruikbaar.

Daarbij was er wel al snel te merken dat er een hele sterke correlatie was tussen de prognose en daadwerkelijke herstel duur. Dat meegenomen kon worden in het uiteindelijke model. Voor de hypothese was ook gebleken dat het klopte dat de aannemer te behoudend was in zijn prognose. In alle toegestane maximale waardes van de target was het gemiddelde prognose duur lager dan de herstel tijd. Met bijna gemiddeld 2x zo een groot verschil



Figuur 1: Verschil prognose en actuele hersteltijd (hersteltijd – prognosetijd)

Daarbij is er ook gekeken naar de meldtijd van de storing, waarbij elk moment van de dag in ingedeeld in verschillende categorieën. Bijvoorbeeld dat een meldtijd van 0:00 – 6:00 betekende dat bij de kolom “nacht” een 1.0 kwam te staan. En bij elk andere kolom een 0.0

Als laatst zijn alle mogelijk bruikbare waardes omgezet in ordinale getallen, die te gebruiken zijn in een regressor model. Daarna zijn alle rijen genormaliseerd in plaats van de kolommen. Uit experimenten bleek deze oriëntatie betere uitkomsten te bieden.

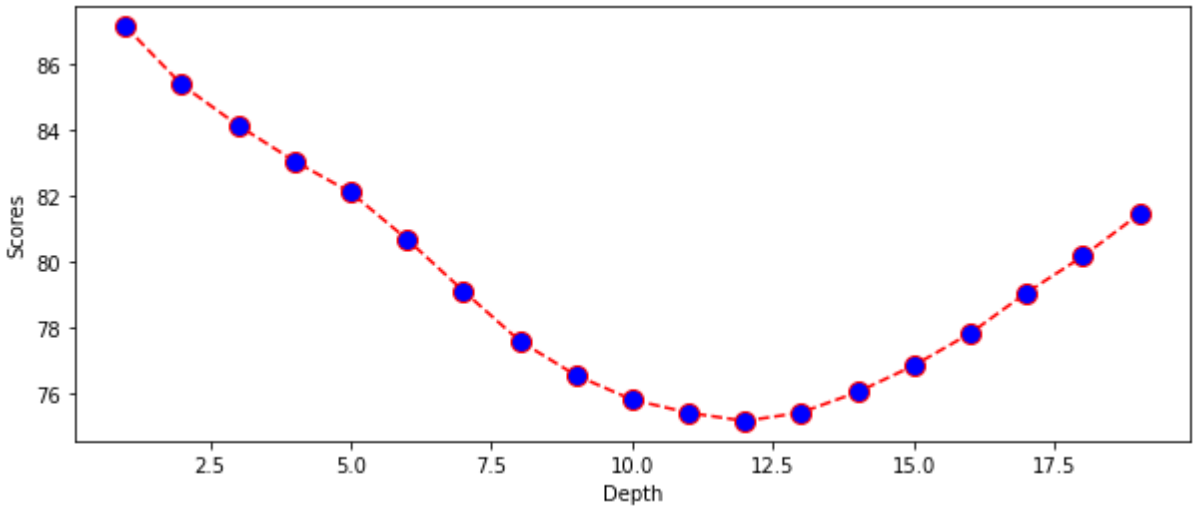
Naast deze aanpassingen waren natuurlijk ook de gebruikelijk opschoning van toepassing. Het verwijderen van NaN values, duplicaties, en outliers verwijderen.

3. Modeleren

Er is eerst begonnen met het gebruiken van een baseline. Hieruit was te vinden dat het aanhouden van de mediaan of de prognose een RMSE van 91.94 en 102.33. Hierna zijn we gaan kijken naar veschillende modellen voor een betere RMSE. We beginnen met het gebruiken van een KNN regression model met een K=9. Hierbij kwam een RMSE van 77.85 uit. Wat al een redelijke verbetering is van de baseline. Hierna gingen we kijken naar ook een Decision tree regressor, en een linear regressor. Waarbij de lineare regressie methode verslechterde resultaten opleverde, had de decision tree betere resultaten met een RMSE van uiteindelijk 75.11. Dit model heeft een max_depth van 12

Vanuit deze resultaten is er gekozen om de decision tree te gebruiken.

Bij het bekijken van de modellen is er ook wat gewerkt met verschillende hyperperameters. Bij de Decision tree model is er bijvoorbeeld naar verschillende depth values gekeken. Hieronder is een afbeelding van het experiment. Na het gebruiken van andere features bleek uit hetzelfde experiment dat een max_depth=12 de laagste



Daarna is er ook gekeken naar de correlatie van verschillende features door het maken van een correlatie matrix, hier zijn de top 6 features gebruikt voor het model. Bij het gebruik van 7 gingen de score achteruit.

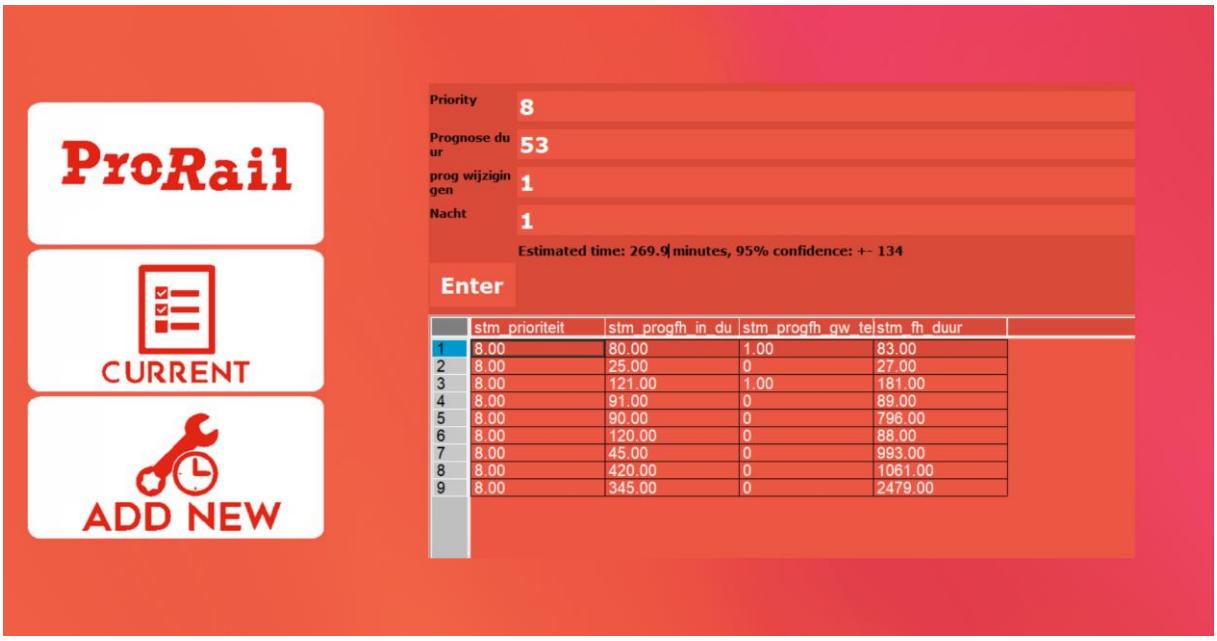
```
stm_afspr_aanvangtijd       0.033157
stm_progfh_gw_teller        0.044423
stm_progfh_gw_duur          0.060061
spits_ochtend               0.074600
nacht                       0.077281
stm_afspr_aanvangdd         0.119666
stm_afspr_func_hersteldd    0.146690
stm_prioriteit              0.234409
stm_progfh_in_duur          0.380649
stm_fh_duur                 1.000000
Name: stm_fh_duur, dtype: float64
```

Met dit model kan er worden gezegd dat er een verbetering is gemaakt rondom het probleem van ProRail. Het model is een verbetering bij het voorspellen van een hersteltijd vergeleken met een prognose van een aannemer.

4. Gebruikersapplicatie

Op onderstaande afbeelding is het uiteindelijk gemaakte applicatie te zien. Aan de linkerzijde bevinden zich buttons om van pagina te wisselen. Onderstaande afbeelding bevindt zich op de “add new” pagina, waarbij aan de rechterzijde een input veld te zien is. Een medewerker kan hier nieuwe data invoeren. Aan de hand daarvan krijgt de medewerker een schatting te zien van hoe lang de storing zal duren.

Voor het ontwerp is gekozen voor een simpele visualisatie die alles in een oogopslag duidelijk moet maken voor de betreffende medewerker. Het logo van ProRail en de kleuren komen terug in het ontwerp. Verder is er gekozen voor een simpel input veld en duidelijke knoppen.



5. Conclusies en aanbevelingen

Vanuit de analyse over de hypothese is duidelijk te concluderen dat de prognose te conservatief is in zijn uitslag. Daarbij is ook aan te tonen dat een model betere resultaten kan opleveren dan het aanhouden van een mediaan. Er is ook duidelijk een correlatie te vinden in de hersteltijd van een probleem, en de kwaliteiten van een probleem.

Het is zeer waarschijnlijk dat de scope van de gebruikte modellen niet goed genoeg kunnen zijn voor de data die is gegeven. Om een bruikbaar model te maken zou er waarschijnlijk gekeken moeten worden naar AI technieken die wat geavanceerder zijn dan voor dit project. Het is hiermee wel al aan te tonen dat het erg mogelijk is om een model te maken.