Adviesrapport ProRail

In dit adviesrapport beschrijven we beknopt welke acties wij hebben ondernomen om tot de juiste oplossing te komen.

Context

We lichten kort de huidige en gewenste situatie toe.

Huidige situatie
Momenteel verloopt het proces als volgt:
Er wordt eerst een melding gedaan van een probleem op of bij het spoor, deze melding komt aan in de meldkamer. Vervolgens licht de meldkamer de regionale treindienstleiders in. Er wordt een aannemer gebeld en de aannemer begeeft zich naar de locatie van het probleem. Als de aannemer aangekomen is op locatie stelt hij een prognose van het probleem en maakt hierbij een schatting over de reparatieduur van het probleem. De schatting wordt gedaan aan de hand van een statische beslisboom en de intuïtie van de aannemer. De incidenten bestrijding wordt ingelicht zowel als de treinplanners van de NS. Hierna wordt de reizigersinformatie gepubliceerd.

Gewenste situatie
Door het gebruiken van ons model hoeven we niet meer te kijken naar de schatting die de monteur gedaan heeft, maar gaan we op basis van eerdere data voorspelling hoe lang het duurt voordat de dienstregeling opgepakt kan worden. Op basis van deze eerder verzamelde datum kunnen we het huidige ongeval vergelijken met een soortgelijk eerder ongeval en zo inschatten hoe lang de storing gaat duren.

Knel- en verbeterpunten
Er wordt nu nog vaak een verkeerde inschatting gemaakt door de monteurs; zij zijn vaak te pessimistisch waardoor de treinplanners van de NS nog bezig zijn met benodigde voorbereidingen terwijl de storing al opgelost is. Hierdoor ligt de dienstregeling vaak onnodig stil. Dit willen we oplossen met ons voorspellende model en onze applicatie. Ook willen we communicatie tussen de verschillende partijen verbeteren met onze applicatie: de applicatie gaat een dashboard tonen met alle relevante informatie, hierdoor hoeft er niet meer stap voor stap gecommuniceerd te worden tussen de stakeholders.

Voor meer informatie zie document: 1. Data understanding

Data understanding

Aan de hand van de data die wij van ProRail gekregen hebben willen wij onze voorspellingen gaan doen. In de stap data understanding gaan we een eerste blik werpen op de data, de data beschrijven en proberen de data kwaliteit te achterhalen.

De dataset oogt in eerste instantie wat rommelig. Het is een uitgebreide dataset die bestaat uit 140 kolommen. Om in een latere stap voorspellingen te kunnen doen moeten we de dataset zodoende ontleden dat we één target variabele (de variabele die we willen gaan voorspellen) en enkele feature variabelen (de variabelen waar uit we willen gaan voorspellen) overhouden.

De target variabele wordt de duur van een storing. De feature variabelen waar we naar gaan kijken zijn prioriteit, oorzaakgroep, oorzaakcode, equipment soort en equipment nummer. Prioriteit zegt iets over de ernst van de storing, oorzaakgroep en oorzaakcode zeggen iets over het soort ongeval en equipment soort en nummer geven aan waar in het materiaal het probleem zich bevindt.

Voor gedetailleerde stappen van de data understanding zie document: 2. Data understanding

Data preparation
In deze stap gaan we onbruikbare data uit de dataset verwijderen en de data die we wel willen gebruiken voorbereiden zodat het gebruikt kan worden in ons model.

Ten eerste zijn er een groot aantal kolommen die volledig, of voor een groot gedeelte uit lege waardes bestaan. Om een goede voorspelling te kunnen doen hebben we kolommen nodig die genoeg informatie bevatten, veel van de kolommen met weinig waardes verwijderen we daarom uit de dataset. Ook zijn er een aantal kolommen die dezelfde betekenis en waardes hebben, we verwijderen in dit geval de kolom met de meeste lege waardes om zo zo veel mogelijk data over te houden zonder dat we onnodig veel kolommen hebben. We hebben ook een aantal kolommen die ergens overlappen, zoals bijvoorbeeld meldtijd en melddatum. Bij overlappende kolommen zullen we deze samenvoegen zodat alle informatie in één kolom kan worden opgeslagen.

Voor gedetailleerde stappen van de data preparation zie document: 3. Data preparation

Modelling
In dit gedeelte van het adviesrapport tonen wij de verschillende modellen die we gebruikt hebben en geven wij een conclusie over welk model het beste werkt voor de probleemstelling.

Omdat wij een numerieke variabele gaan voorspellen hebben wij verschillende regressie modellen uitgeprobeerd.

1. Lineaire regressie
Score: -3.304
RMSE: 8209374840.113075

Toelichting: Als eerste model hebben we lineaire regressie uitgeprobeerd. De score van dit model is ook meteen het slechts van alle modellen. Ook is de RMSE erg hoog. Het is duidelijk dat onze data niet de trend van dit model volgt. Lineaire regressie is dus niet geschikt.

2. Ridge regressie
Score: 0.13
RMSE: 62.766

Toelichting: Waar er bij lineaire regressie geen rekening gehouden wordt met multicollineariteit, gebeurd dit bij ridge regressie wel. De score van ridge regressie is aanzienlijk beter dan die van het lineaire regressie model. Ook is de RMSE een stuk lager.

3. KNearest Neighbours regressie
Score: -0.1691
RMSE: 72.548

Toelichting: Bij KNearest Neighbours regressie wordt door middel van de gemiddeldes van dichtbij zijnde datapunten de voorspelling gedaan. De negatieve score zegt ons dat de data niet d trend van het model volgt en ook is de RMSE weer een stuk hoger dan bij de ridge regressie.

4. Decision tree regressie
Score: 0.0996
RMSE: 63.932

Toelichting: De score van de decision tree is niet slecht maar nog steeds niet zo goed als die van de ridge regressie.

5. Random forest
Score: 0.12
RMSE: 63.22

Toelichting: De score van het random forest model komt het dichtst in de buurt bij die van het ridge regressie model.

Omdat we voor het random forest model het beste de probability uit kunnen rekenen hebben we gekozen om verder te gaan met dit model en dit model in de applicatie te verwerken.

Evaluation
In deze stap blikken we terug naar de stappen die we hebben ondernomen en evalueren de resultaten.

De twee beste modellen zijn volgens ons dus de ridge regression en de random forest regressor. Afgezien van het feit dat dit de twee beste modellen zijn is de score nog niet hoog genoeg om een daadwerkelijk correcte voorspelling te kunnen doen. We zouden dus kunnen stellen dat op basis van de verkregen informatie en de gebruikte modellen we geen goede, betrouwbare voorspelling kunnen doen van hoe lang een storing gaat duren. De business objectives zijn dus niet behaald met onze oplossing.