Adviesrapport

DataDriven Business:
ProRail

Latif Azaïm - 1851448
Mathias Hendriks – 1740564
Max Jansen – 1848848
Yujian Jiang – 1835919

07 / 11 / 2024

Inhoudsopgave

07 / 11 / 2024

Inleiding

1.1 Achtergrond

Voor het vak DataDriven Business is ProRail naar de HU gekomen met de vraag of wij hulp kunnen bieden voor het DataLab. Het DataLab van ProRail is een afdeling waar 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 efficient worden opgelost om zowel kosten te besparen als de tevredenheid van de reizigers te verhogen.

1.2 Belang

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 omheen gepland kan worden en de vertragingen zo beperkt mogelijk kunnen blijven.

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. Omdat het best een groot proces is dat we door lopen. Is het geheel eigenlijk opgedeeld in drie kleinere opdrachten die uiteindelijk een geheel moeten vormen.

2.2 Deelopdrachten

Het eerste deel is de Business Understanding, zodat ons duidelijk wordt waar wij onze taken moeten vervullen en hoe het proces loopt. Daarna moeten wij door gebruik te maken van slimmen systemen, die kunnen leren van oude data, voorspellende modellen kunnen maken. Hiermee onderzoeken we of het mogelijk is om de hersteltijden te voorspellen van de storingen. Als laatste moet er een gebruikersapplicatie gemaakt worden die gebruikt maakt van een voorspellend model, waar de gebruiker een storing kan invoeren en een verwachte hersteltijd terug krijgt. Het DataLab heeft ons daarom toegang gegeven tot een verzameling data waarin alles van oude storingen in staan. Ook hebben wij inzicht gekregen over hoe zij te werk gaan en wat de verwachtingen zijn van ons.

Business Understanding

3.1 Procesanalyse

De eerste stap is om het proces te verduidelijken om het door te nemen. Dit is belangrijk omdat wij moeten weten waar wij onze applicatie en voorspellingen toe kunnen passen. Daarvoor hebben wij een BPMN1 (Business Process Model and Notation) gemaakt om een duidelijk inzicht te krijgen over hoe het proces loopt als er een storing optreed. BPMN is een standaard methode om bedrijfsprocessen visueel weer te geven, zodat iedereen begrijpt hoe de stappen verlopen. Daaruit kunnen wij ook uitzoeken welke data op de momenten beschikbaar is. Het kan namelijk zo zijn dat als je een voorspelling wilt maken, dat je een deel van de benodigde data dan nog niet hebt.

3.2 Beschikbaarheid van 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.

Data

4.1 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.

4.2 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. Sommige storingen moeten sneller opgelost worden dan andere.

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.

4.3 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.

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

Als eerste hebben we lineaire regressie gebruikt. Lineaire regressie is een - eenvoudig model dat probeert een rechte lijn te vinden die het beste past bij de data. Het voorspelt de hersteltijd op basis van een lineaire combinatie van de inputvariabelen. Dit model gaf een redelijk resultaat, maar was niet nauwkeurig genoeg voor onze doeleinden. Het kon niet goed omgaan met de complexe relaties tussen de variabelen.

5.1.2 Random forest

Daarna hebben we een random forest model geprobeerd. Random forest is een ensemble van decision trees. Het bouwt meerdere decision trees en neemt het gemiddelde van de voorspellingen. Dit model presteerde beter dan de lineaire regressie, omdat het beter om kan gaan met niet-lineaire relaties en interacties tussen variabelen. Maar het was nog steeds niet optimaal.

5.1.3 Decision tree

Uiteindelijk hebben we besloten om een decision tree model te gebruiken. Een decision tree is een boomstructuur waarbij elke knoop een beslissing voorstelt op basis van een variabele. Het verdeelt de data in kleinere groepen op basis van de meest significante variabelen. Dit model gaf de beste resultaten en was het meest nauwkeurig in het voorspellen van de hersteltijden. Met de decision tree konden we de complexe relaties in de data beter in kaart brengen en nauwkeurigere voorspellingen doen.

5.2 Resultaten van de modellen

We hebben de prestaties van elk model geëvalueerd door middel van de Mean Squared Error (MSE), wat een maat is voor de gemiddelde fout tussen de voorspelde en de werkelijke waarden. Het decision tree model had de laagste MSE, wat betekent dat het de meest nauwkeurige voorspellingen gaf.

Daarnaast is het decision tree model ook interpreteerbaar. We kunnen zien welke variabelen het meest bijdragen aan de voorspelling. Dit is nuttig, omdat we zo kunnen begrijpen welke factoren de hersteltijd het meest beïnvloeden.

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.

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.

07 / 11 / 2024