

Weekly meetings

- TA meetings: every Tuesday at 13:45 (online or on campus)
- Sprint planning meetings: every Tuesday after the TA meeting (on campus)

Research Questions

- Main research question:

“How can a data-driven model be used to predict the likelihood that a train path request by an operator will be granted by ProRail based on historical data?”

- Sub-research questions:

A: *“How does time tolerance (earliest and latest preferred departure) affect the probability?”*

B: *“How good is the match between the request of the operator and the response of ProRail?”*

C: *“Do route structure and stabling/service needs influence outcomes?”*

D: *“Are operator, planning phase, and request timing (hour/day/week) predictive?”*

Planning

Task	Definition of Done	8-sep	15-sep	22-sep	29-sep	6-okt	13-okt	20-okt	27-okt	3-nov
		Sprint 1	Sprint 2	Sprint 3	Sprint 4	Sprint 5	Sprint 6	Sprint 7	Sprint 8	Sprint 9
Group formation	Added to group chat									
Data exploration	Documented in backlog									
Meeting with TA	Minutes added to backlog									
Data cleaning	Defined rows and columns removed									
Milestone 1: Project proposal	Assignment handed in on BS									
Generate numeric dataset	Code runs without errors									
Try linear regression models	Code runs without errors									
Milestone 2: Midterm check	Midterm meeting									

Incorporate feedback	Session with professor									
Finish backlog	Documented in backlog									
Update Github	Documents in GitHub									
Milestone 3: Project portfolio submission	Assignment handed in on BS									
Milestone 4: Final presentation	Final presentation session									

Sprint

Ticket	Priority level (low-medium-high)	Category (preparation, organisation, input, code, output, reporting)	Assigned to	Review (1 per sprint)
Sprint 1: Forming (September 1)				
Group formation	low	organisation	Everyone	no comments
Sprint 2: Data collection & storming (September 8)				
Contact with company supervisor - Plan meeting to ask questions about certain columns	medium	input	Wendy, Quinten	improve communication between us and teaching team and company
Data exploration - First look at the dataset and getting familiar with it	high	preparation	Everyone	
Setup GitHub	low	organisation	Quinten	
Sprint 3: Problem definition (September 15)				
Data exploration (spillover) - First look at the dataset and getting familiar with it	high	preparation	Everyone	Create more structure
Meeting with company supervisor - Expectation management and	high	organisation, input	Quinten, Reno, Wendy	

discussing our end result				
Group meeting - Data cleaning, determine which columns are relevant	high	input	Everyone	
Schedule meeting with TA	low	organisation	Wendy	
Project proposal - Make a base for proposal	medium	reporting	Quinten	
Project proposal - Finish chapter 2: data preprocessing	medium	reporting	Reno	
Project proposal - Finish chapter 1: introduction, chapter 3: backlog and make additions to chapter 2: data preprocessing	medium	reporting	Wendy, Rosanne	

Sprint 4: Data cleaning & norming (September 22)

Clean sample data - Write code to read and clean dataset	high	code	Rosanne	Improve communication with teaching team and company, normalise data, see what works for us in terms of backlog
Meeting with TA - First meeting - Feedback on our work so far	medium	organisation	Everyone	
Assign project manager - To keep clear overview of our progress and clear communication	medium	organisation	Everyone: Wendy= PM	
Create user stories	low	preparation	Rosanne	
Create requirement list	low	preparation	Rosanne	
Submit project proposal	high	reporting	Quinten	
Create tickets in GitHub - More structured way of project management	low	organisation	Wendy	
Fix the sprint tasks - Make sprint tasks more	medium	organisation	Wendy	

specific				
Write a project backlog diary - Important for project management and reporting	high	organisation	Wendy	
Visualisations on cleaned data - Clear view on the trends of the data	medium	code	Quinten	
Make some tests on the cleaned data - Get familiar with updated dataset	high	code	Reno	
Sprint 5: Feature engineering & norming (September 29)				
Write a sprint review for each sprint - Nice to know what to keep into account for next sprint	low	reporting	Everyone	Missing data caused us to only be able to continue on Monday. We acted quickly with retrieving the missing data.
Upload documents to Github - Clear overview in one place	medium	organisation	Wendy	
Finish processing the data - Make data binary, categorical or numerical	high	code	Reno	
Try to run a regression model	high	code	Rosanne	Took longer than expected, managed to run simple logistic regression model per column. Next sprint: combine this into regression model

Make visualisations that support the research question - Data evaluation of cleaned data and visuals of statistics	medium	code	Quinten	New visualisations on the new dataset. Data from december 2022 seems to be missing still, but discussed with TA that this shouldn't be a problem.
Write an explanation on the visualisation	low	reporting	Wendy	
Sprint 6: Model selection & performing (October 6)				
Make visualisations - Relevant for midterm presentation	medium	code	Quinten	
Try a linear regression model	high	code	Reno	
Think of possible solutions - Check regression models and other models	medium	code	Rosanne	
Finish backlog - Process comments Mahnam and put in a notebook	high	reporting	Wendy	
Make PowerPoint	high	reporting	Everyone	
October 10 Midterm				
Sprint 7 (October 13)				
Sprint 8 (October 20)				
Sprint 9 (October 27)				
Sprint 10 (November 3)				
4 November Project deadline				

Project Backlog

Changes/updates in the backlog:

Date October 9 2025

- Add sprint planning to backlog
-

Date: September 23 2025

- Update sprint documentation
- Process information documented in the project proposal (e.g. cleaning of dataset)

Date: September 16 2025

- Create backlog

Minutes of meetings

- **October 10** (on campus):
 1. Midterm presentation with professors, ProRail and TA.
- **October 7** (on campus):
 1. [TA meeting 3](#)
 2. Preparing for midterm presentation
- **September 30** (on campus):
 1. [TA meeting 2](#)
 2. Group meeting: sprint planning
- **September 24** (on campus):
 1. Discussion on feedback
- **September 23** (on campus):
 1. [TA meeting 1](#)
- **September 22** (online and on campus)
 1. Finishing up milestone 1.
- **September 17** (on campus):
 1. Discussing how to clean the data.
- **September 16** (online and on campus):
 1. Meeting the group on campus.
 2. Online meeting with ProRail.

Current planned tech stack

1. Operating systems: MacOS and Windows
2. Server-Side Programming: Python
3. Frameworks and libraries
 - a. Machine learning techniques to predict probability
 - b. Data cleaning: pandas, numpy
 - c. Visualisation: matplotlib, seaborn
 - d. Machine learning models: scikit-learn (logistic regression, gradient boosting, random forests)
 - e. Model evaluation: built-in scikit metrics (AUC, precision recall, Brier score)
4. Techniques, iterative approach

- a. Baseline model: logistic regression with simple engineered features (time-of-day, day-of-week, stops). This provides interpretability and a benchmark.
 - b. Tree-based models: Random Forest and Gradient Boosting to capture nonlinear relationships.
 - c. Advanced feature engineering: incorporating categorical encodings to group orders.
 - d. Lastly for evaluation: apply probability calibration (e.g. Platt scaling, isotonic regression) to ensure probability outputs.
5. Version control: GitHub
 6. Development environment: VS Code and Jupyter Notebook
 7. Documentation: backlog document (it is not possible to create a wiki in a private repository)

Data cleaning

The code for cleaning the raw data set is saved in: cleaned data 23_09.ipynb.

What it does:

1. It deletes complete empty rows
2. It keeps rows where processtype is not ANNULERING (so deletes Annulering)
3. It keeps reaction type: aangeboden and geen mogelijk
4. It deletes rows without reaction
5. It keeps rows where processtype is not handmatig
6. It keeps rows where processtype is not NSR
7. It keeps rows where processtype is not JAARDIENST
8. It keeps only rows with most recent ordernummer

The cleaned dataset is then saved into a new csv file called "cleaned_dataset.csv".

Code

Numeric dataset

The code for generating a numeric dataset for machine learning is: test_2.ipynb.

What it does:

1. It filters the data by process type, leave only the NIEUW requests
2. It extracts the hours from the earliest departure time column, and remove the rows where this information is missing
3. The hours of a day of the earliest departure time column is then encoded in a cyclical way, so that 0:00 is close to 23:00
4. The reaction by ProRail is mapped into binary numbers. AANGEBODEN: 1, GEEN_AANBIEDING_MOGELIJK: 0
5. The stabling service request information (binary) and time tolerance are copied from the cleaned dataset
6. The days of week of the requests are also encoded in a cyclical way, so that Monday is close to Sunday
7. The origin and destinations are encoded with OneHotEncode

The numeric dataset is saved into a new csv file called "numeric_dataset.csv".