THE UNIVERSITY of ADELAIDE

Software Engineering & Project (COMP SCI 7015)

Snapshot Week 04 of Group RAIL PG-2

Rail Break Prediction ML

Tao Xu a1937511

Sheng Wang a1903948

Jinchao Yuan a1936476

Zi Lun Ma a1915860

Di Zhu a1919727

Xin Wei a1912958

Yifan Gu a1909803

Tianhua Zhang a1915934

Zihan Luo a1916700

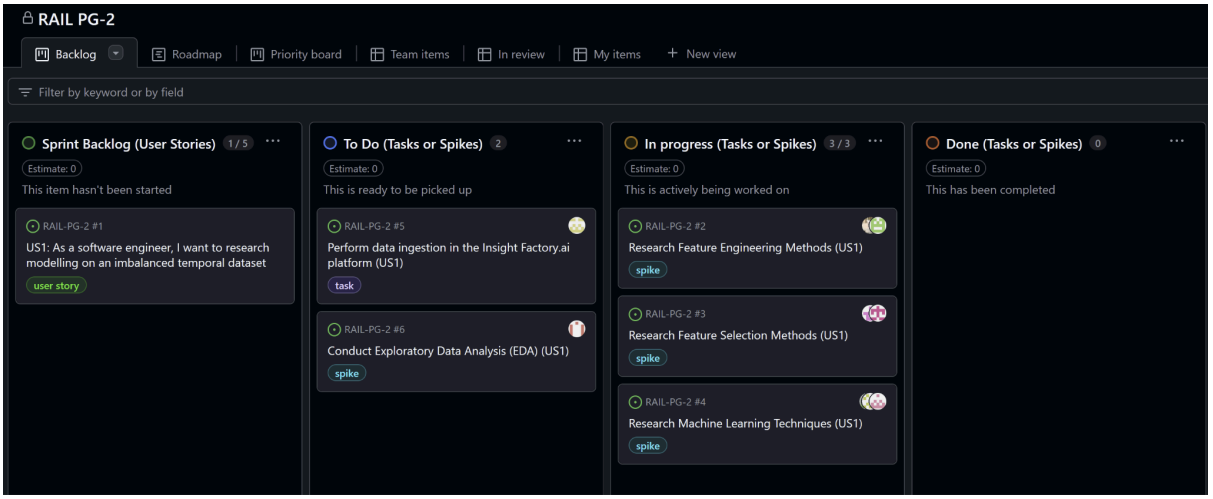
Supervisor : Murtaza Bootwala

1. Product Backlog and Task Board

1.1. The product backlog

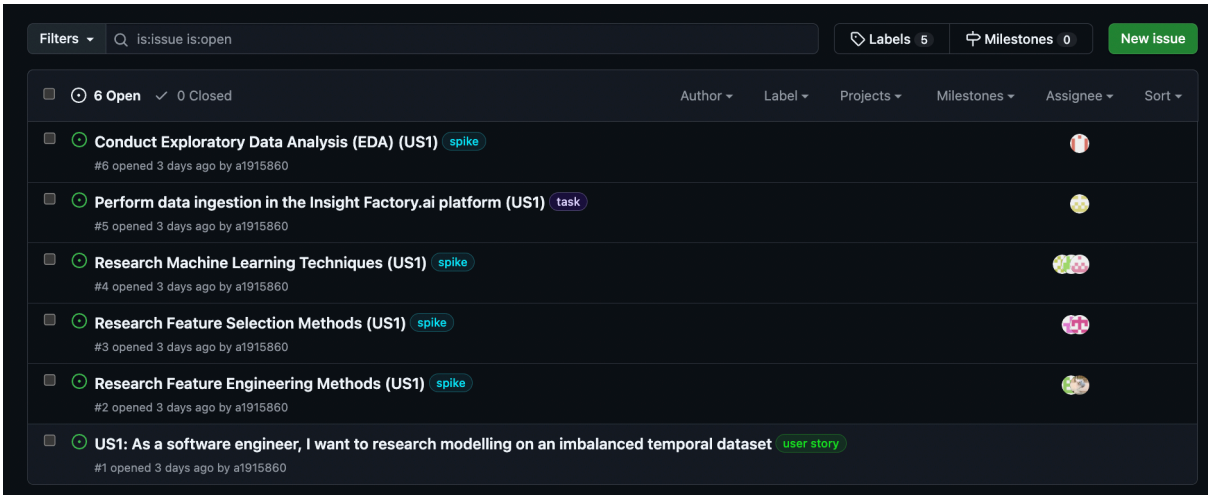
ID	Priority	User Story/Task/Spike	Description
PB1	1	Feature Engineering	Create new features based on domain knowledge and data patterns to improve model performance.
PB2	1	Feature Selection	Identify and retain the most relevant features to reduce noise and improve efficiency.
PB3	1	Model Research & Selection	Investigate suitable machine learning techniques for imbalance temporal datasets
PB4	2	Data Ingestion into InsightFactory.ai	Import the provided real-world production dataset into the InsightFactory platform.
PB5	2	Data Cleaning & Preprocessing	Handle missing values, outliers, and inconsistencies in the dataset.
PB6	2	Exploratory Data Analysis (EDA)	Analyze data distributions, trends, and anomalies to understand key characteristics.
PB7	3	Model Training	Train predictive models using the processed and engineered dataset.
PB8	3	Model Evaluation	Assess models using Accuracy, F1 Score, and AUCPR metrics.
PB9	3	Benchmark Comparison	Compare the model's performance against the InsightFactory bench mark model for potential bonus marks.
PB10	4	Model Optimization & FInalization	Fine-tune model parameters, optimize features, and prepare the final deliverable.

1.2. The task board



2. Sprint Backlog and User Stories

2.1. The Sprint backlog



2.2. User stories

User story 1: As a software engineer, I want to research techniques for modelling an imbalanced temporal dataset, including feature engineering, feature selection, and suitable machine learning methods. Ingest the data into the Insight Factory.ai platform. Conduct exploratory data analysis (EDA) to develop a plan to approach the project.

Related tasks:

1. Research Feature Engineering Methods
2. Research Feature Selection Methods
3. Research Machine Learning Techniques
4. Perform data ingestion in the Insight Factory.ai platform
5. Conduct Exploratory Data Analysis (EDA)

3. Definition of Done

A backlog item is considered “Done” when:

Spike:

- The research is complete, including findings, identified risks and challenges, and any recommendations.
- All relevant documentation is shared with the team.

Task*:

- Code (including database scripts) is implemented according to acceptance criteria.
- Code has been peer-reviewed and approved.
- All relevant tests (unit, integration) have been passed.
- Documentation (code comments, user guides) is updated.
- No major open defects remain.

* The current sprint is research-based, so the DOD for tasks might not be applicable. However, it's better to decide the expectation for tasks earlier.

4. Summary of Changes

This week, we were granted access to the InsightFactory platform and tested setting up production lines to ingest data into Databricks. We also had a first look at the tonnage and wagon datasets.

With this information, we started working on research works: techniques of feature engineering, feature selection, and machine learning.

We summarized the findings on the Github project board, and discussed which column could be useful for training the model. However, as the trainingContext table is not yet available, we don't have any information on rail breaks records. For now, we can't consider that we have finished the research, and will continue adding new findings.

Since we only have access to the Tonnagedata and Wagondata tables at the moment, the data ingestion and EDA can only be carried out on a limited dataset. These tasks will therefore be deferred to the next sprint.