Software Engineering & Project (COMP SCI 7015)

Snapshot Week 05 of Group RAIL PG-2

Rail Break Prediction ML

THE UNIVERSITY of ADELAIDE

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

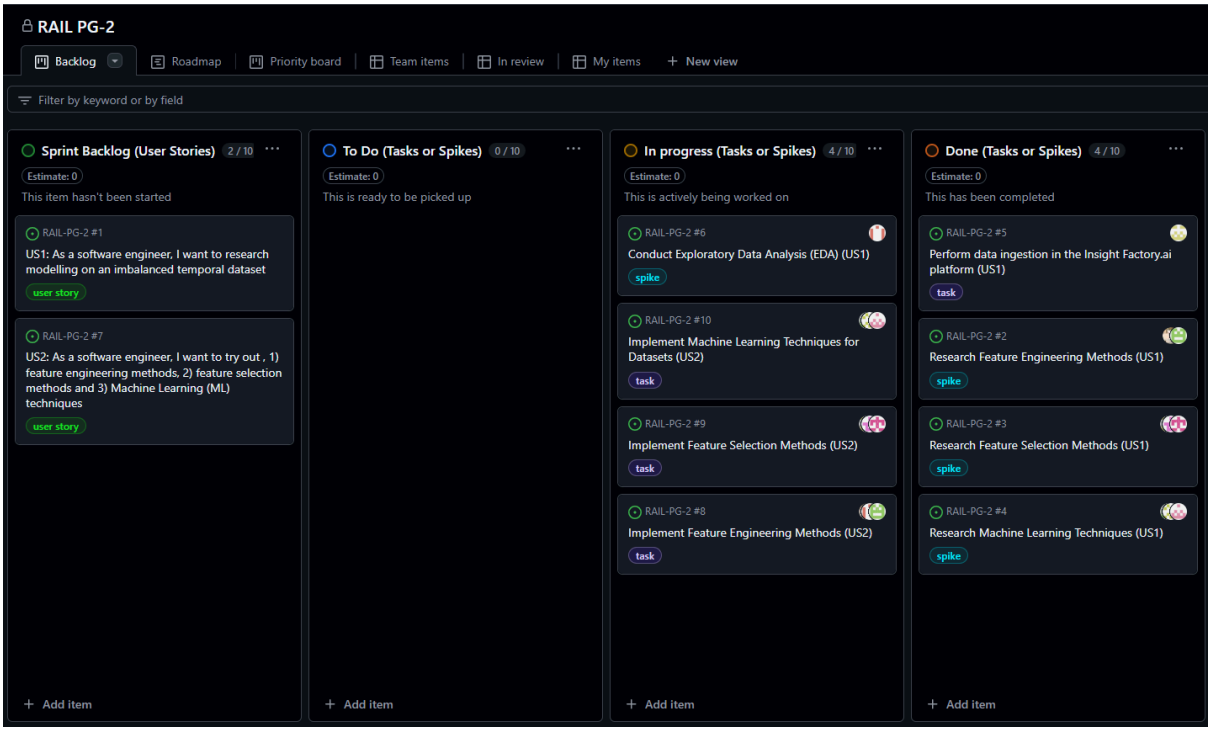
Supervisor : Murtaza Bootwala

1. Product Backlog and Task Board

1.1. The product backlog

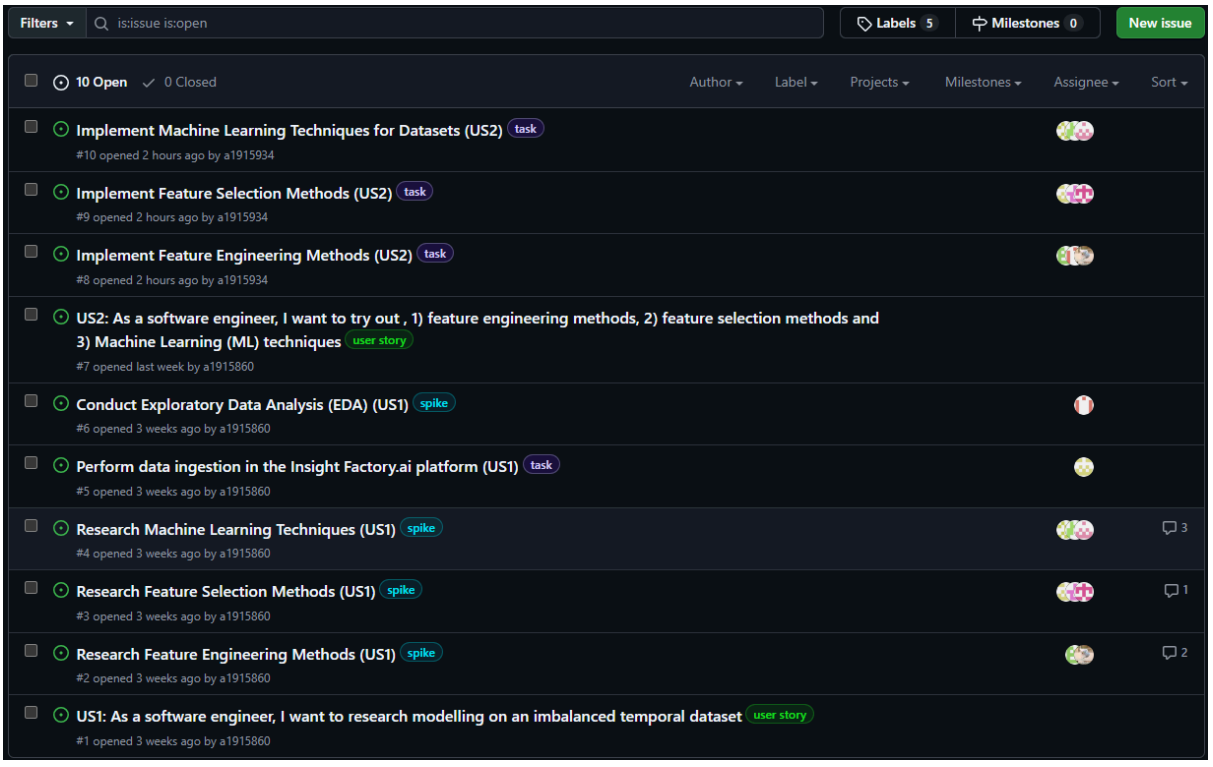
ID	Priority	User Story/Task/Spike	Description
PB1	1	Feature Engineering	Create new features based on domain knowledge and data patterns to improve model performance.
PB2	1	Feature Selection	Identify and retain the most relevant features to reduce noise and improve efficiency.
PB3	1	Model Research & Selection	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.
PB11	1	Implement Feature Engineering Methods	exploring and testing different feature transformation and construction approaches to enhance the predictive power of the dataset.
PB12	1	Implement Feature Selection Methods	applying statistical and algorithmic techniques to identify the most relevant features and reduce dimensionality for improved model efficiency.
PB13	2	Implement Machine Learning Techniques for Datasets	investigating specialized algorithms and resampling strategies to handle class imbalance effectively.

1.2. The task board



2. Sprint Backlog and User Stories

2.1. The Sprint backlog



2.2. User stories

Try different techniques including feature engineering methods, feature selection methods and Machine Learning (ML) techniques to approach a problem having an imbalanced dataset, to produce an initial model to InsightFactory leaderboard.

Related tasks:

1. Implement Feature Engineering Methods
2. Implement Feature Selection Methods
3. Implement Machine Learning Techniques for Datasets

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.

4. Summary of Changes:

Since the last sprint, our team has refined the new user story and expanded the scope of our work. Based on the new user story we add additional works in the Sprint backlog.

With the full dataset now available on the platform, we started addressing the Exploratory Data Analysis (EDA) tasks to better understand data distributions, correlations, and quality issues.

In the feature engineering stage, we created a training_dataset table and integrated it into the pre_main_model_training process.

Additionally, we created a new collaborative pipeline within the InsightFactory.ai platform to streamline group development and ensure reproducibility. This will allow team members to

work in parallel on different aspects of the workflow, including feature engineering, feature selection, machine learning techniques for imbalanced datasets, and model training.