Software Engineering & Project (COMP SCI 7015)

Snapshot Week 08 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

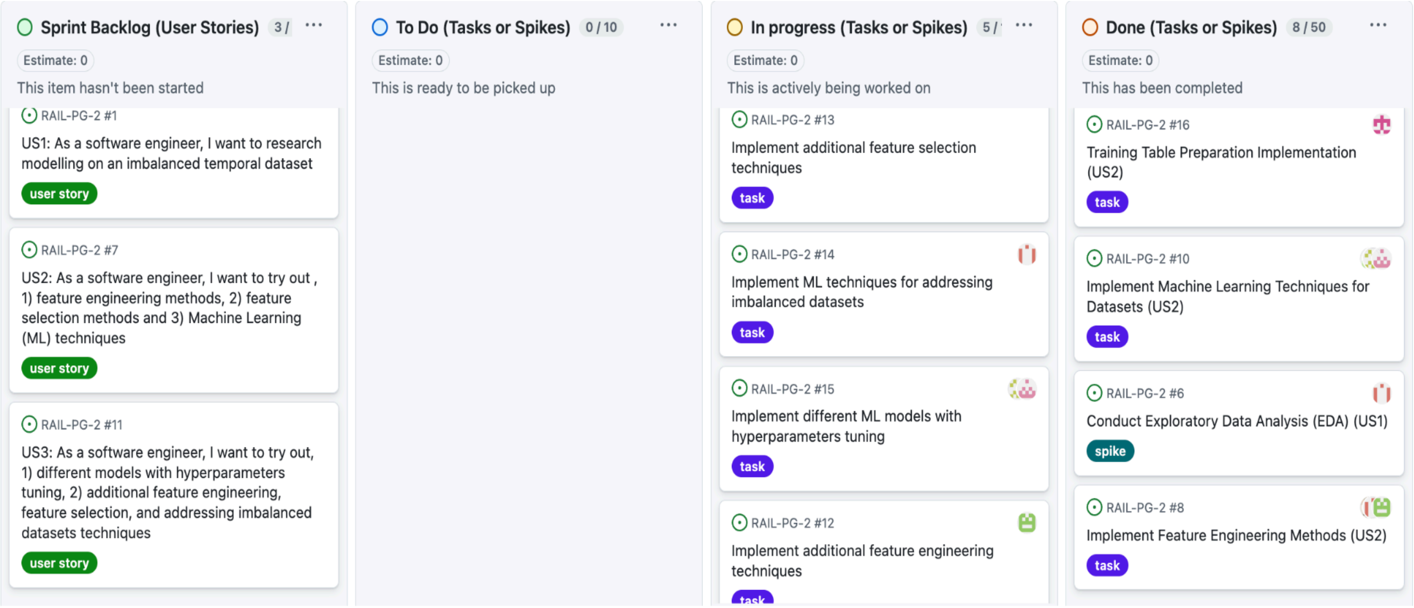
1. Product Backlog and Task Board

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.
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.
PB14	2	Training Table Preparation Implementation	Implement Training table preparation scripts, these scripts provide the fundamental data integration for the overall project pipeline.
PB15	1	Implement additional feature engineering techniques	Try out at least 2 more feature engineering techniques.

PB16	1	Implement additional feature selection techniques	Try out at least 2 more feature selection techniques.
PB17	1	Implement ML techniques for addressing imbalanced datasets	Try out at least 2 more techniques for handling imbalanced datasets.
PB18	2	Implement different ML models with hyperparameters tuning	Tuning hyperparameter of ● at least 3 different ML models ● feature selection, feature engineering, and imbalanced dataset handling techniques, if they have hyperparameters to tune

1.2. The task board



Sprint Backlog (User Stories) 3 /

Estimate: 0

This item hasn't been started

RAIL-PG-2 #1
US1: As a software engineer, I want to research modelling on an imbalanced temporal dataset
user story

RAIL-PG-2 #7
US2: As a software engineer, I want to try out , 1) feature engineering methods, 2) feature selection methods and 3) Machine Learning (ML) techniques
user story

RAIL-PG-2 #11
US3: As a software engineer, I want to try out, 1) different models with hyperparameters tuning, 2) additional feature engineering, feature selection, and addressing imbalanced datasets techniques
user story

To Do (Tasks or Spikes) 0 / 10

Estimate: 0

This is ready to be picked up

In progress (Tasks or Spikes) 5 /

Estimate: 0

This is actively being worked on

RAIL-PG-2 #13
Implement additional feature selection techniques
task

RAIL-PG-2 #14
Implement ML techniques for addressing imbalanced datasets
task

RAIL-PG-2 #15
Implement different ML models with hyperparameters tuning
task

RAIL-PG-2 #12
Implement additional feature engineering techniques
task

Done (Tasks or Spikes) 8 / 50

Estimate: 0

This has been completed

RAIL-PG-2 #16
Training Table Preparation Implementation (US2)
task

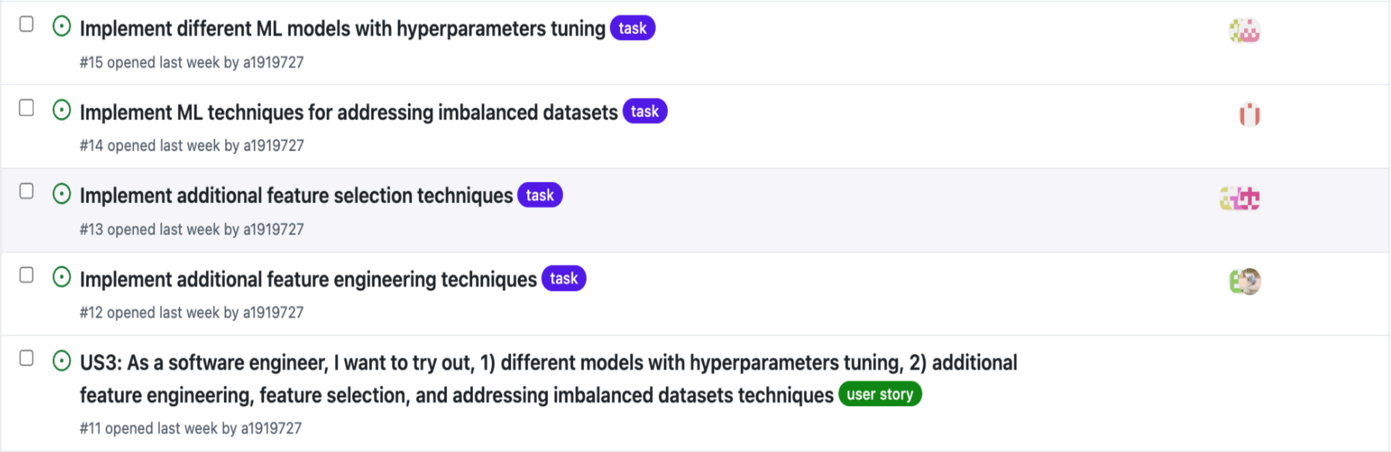
RAIL-PG-2 #10
Implement Machine Learning Techniques for Datasets (US2)
task

RAIL-PG-2 #6
Conduct Exploratory Data Analysis (EDA) (US1)
spike

RAIL-PG-2 #8
Implement Feature Engineering Methods (US2)
task

2. Sprint Backlog and User Stories

2.1. The Sprint backlog



☐ Implement different ML models with hyperparameters tuning task
#15 opened last week by a1919727

☐ Implement ML techniques for addressing imbalanced datasets task
#14 opened last week by a1919727

☐ Implement additional feature selection techniques task
#13 opened last week by a1919727

☐ Implement additional feature engineering techniques task
#12 opened last week by a1919727

☐ US3: As a software engineer, I want to try out, 1) different models with hyperparameters tuning, 2) additional feature engineering, feature selection, and addressing imbalanced datasets techniques user story
#11 opened last week by a1919727

2.2. User stories

User story 3:

As a software engineer, I want to 1) try out different models while fine-tuning their hyperparameters, and 2) tryout additional techniques for feature engineering, feature selection, and methods for addressing imbalanced datasets, so that I can improve the test set predictions to achieve an F1 score exceeding 55%.

Related tasks:

1. Implement additional feature engineering Methods
2. Implement additional feature selection Methods
3. Implement machine learning techniques for addressing imbalanced datasets
4. Implement different machine learning models with hyperparameters tuning

3. Definition of Done

A backlog item is considered “Done” when:

Task:

- Create different branches for different tasks
- Code (including database scripts) is implemented according to acceptance criteria.
- Code has been peer-reviewed and approved if a new table is created in the final schema.
- All relevant tests (unit, integration) have been passed.
- Documentation (code comments, user guides) is updated.
- No major open defects remain.

4. Summary of Changes

This week, the feature engineering team applied resampling techniques: stratified and random oversampling/undersampling across three training tables. In total, 12 training datasets (2 [stratified & random] × 2 [oversampling & undersampling] × 3 [total_training_table, preprocess_table, fe_training]) were generated for model evaluation. On top of that, the feature engineering script was refactored to make it more adaptable for adding new features. We experimentally added lag features for all numeric columns to observe their impact on model performance, and introduced interaction features (e.g., curvature × speed, speed × twist) to assess potential performance improvements.

The feature selection team explored feature selection strategies with varying time windows, Group Lasso settings, and parameter tuning in Focal Loss for imbalance handling. Also applied a new method, gating, for feature selection, finally achieving an F1 score of 53.64%.

The model training team trained models with datasets that were handled by different feature selection and feature engineering techniques. This process is to find which feature selection and feature engineering techniques are suitable and facilitate us to achieve a higher F1 score. Also, they tuned hyperparameters for three models, including SVM, DNN, and Transformer. After submitting inferences, the highest F1 score for the SVM is 52.51%, for the DNN is 30.00%, and for the Transformer is 53.11%.