2025-09-08 Meeting Notes

📅 Date

08 Sep 2025

👥 Participants

- RAIL PG-2 project team

Tao Xu	a1937511
Sheng Wang	a1903948
Jinchao Yuan	a1936476
Zilun Ma	a1915860
Di Zhu	a1919727
Xin Wei	a1912958
Yifan Gu	a1909803
Tianhua Zhang	a1915934
Zihan Luo	a1916700

- Murtaza (Proxy Client)

🗣️ Discussion topics

1. **Github backlog**
   - Introduced completed tasks in sprint 2, including the implementation of feature selection, feature engineering, training table creation, and model training.
   - Divided tasks for sprint 3 based on the user story 3, including the implementation of additional feature selection, feature engineering, ML techniques for handling imbalance datasets, and hyperparameters tuning for ML models.
2. **Project progress**
   - **EDA team:**
     - Applied visualization techniques (e.g. heatmap, histogram, and time series plot) to analyze data.
   - **Feature selection team:**
     - Joined and integrated trainingContext, wagondata, and tonnagedata table into a total training table
     - Preprocessed the total training table

    - Implemented a basic version of transformer model with Group Lasso and REF with LightGBM
    - Tested and evaluated performance of combining feature selection methods with models
- **Feature engineering team:**
    - Goal: Enhance dataset prediction ability
    - Explored the available sensor features and set risk thresholds
    - Applied Baseline features, Fourier transform, and trend features methods
- **Model training team:**
    - Trained SVM, DNN, and transformer models
    - Evaluated model performance based on accuracy score
    - Created inferences to submit
- **Production line creation:**
    - Built a production line in the IF platform to connect feature selection, feature engineering, model training, prediction, and submission

3. **Sprint 2 results**

Leadboard accuracy results:

- Transformer: 66
- DNN: 65
- SVM: 43

Based on accuracy score, the best model in Sprint 2 is the transformer

4. **Blockers**
    - Our team did not get datasets at the beginning of Sprint 2, so we started to build this project a little late.
    - Building and training model time was limited, as team members first need to complete feature selection, data preprocessing, and training table creation.
    - The runtime duration on the IF platform was unstable. This increases uncertainty for training models and creating inferences.
5. **Next plan**

Goal: Improve models overall performance and achieve an F1 score over 55%.

Team tasks

- **Feature selection team:** Implement at least 2 additional methods and continuously iterate embedded methods and integrate with models to test
- **Feature engineering team:** Implement at least 2 additional methods and handling imbalanced dataset techniques.
- **Model training team:** Tune hyperparameters for every model to improve model performance and record tuned hyperparameters and related results.

6. **QA**

Q: For addressing imbalanced datasets, do we need to create a balanced table by preprocessing techniques, or can we directly apply handling imbalanced datasets techniques when training a model?

A: It depends on your choice. You can check scores in the leader board. The suitable techniques can help you to achieve a better score.

7. **Suggestions and feedback**
   - **Initial report**
     - Copy and paste the whole content of the snapshot in the appendix in the future report and retrospective. The cover page should be included.
     - Users of the solution are the real roles (e.g. maintenance engineer)
     - Writing reports with formal language
     - Build a full architecture in the final report rather than the initial one
   - **Snapshot**
     - Definition of done is required to follow the software development process
   - **Retrospective**
     - Do not copy and paste the same thing in each retrospective
     - Write detailed technical explanations and follow the software development methodology
     - Add details and description rather than generalization