Rail Break Prediction AI

Software Engineering Project 2024

Introduction

This project provides a rare opportunity to work with real-world data to solve a practical problem faced in industry. You are required to build a data pipeline to extract & ingest data from various endpoints, enrich it, and then use machine learning models to extract meaningful insights and predictions. This process will be streamlined with the easy-to-use Insight Factory platform, a tool built by Data & AI Practitioners at insightfactory.ai to deliver data initiatives to clients around the world. This project involves a competitive element – the performance of your model in determining rail breaks which will be compared with rival groups on a leaderboard. **The most accurate models will win prizes at the end of the semester** – more on this later. You can expect to use SQL and Python for this task, utilising whichever libraries give you the best result. We will provide each team with an Insight Factory platform and encourage you to be as creative as possible to edge out your peers.

Problem context

Over 33,000 kilometres of railway stretch across the Australian continent, providing a transportation network crucial for industrial freight operations. Every day, hundreds of trains traverse this network, often pulling tens of multi-tonne carriages behind them. Understandably, this leads to wear and tear, and in some cases, breakages in rail sections. Not only can this be dangerous, but the delays also caused by such events can be costly for freight companies and the industries that rely on them. To combat this, The Australian Rail Track Corporation (ARTC) has provided us with extensive sensor data from trains using the rail network.

Dataset

Thanks to ARTC for sharing the dataset for the project.

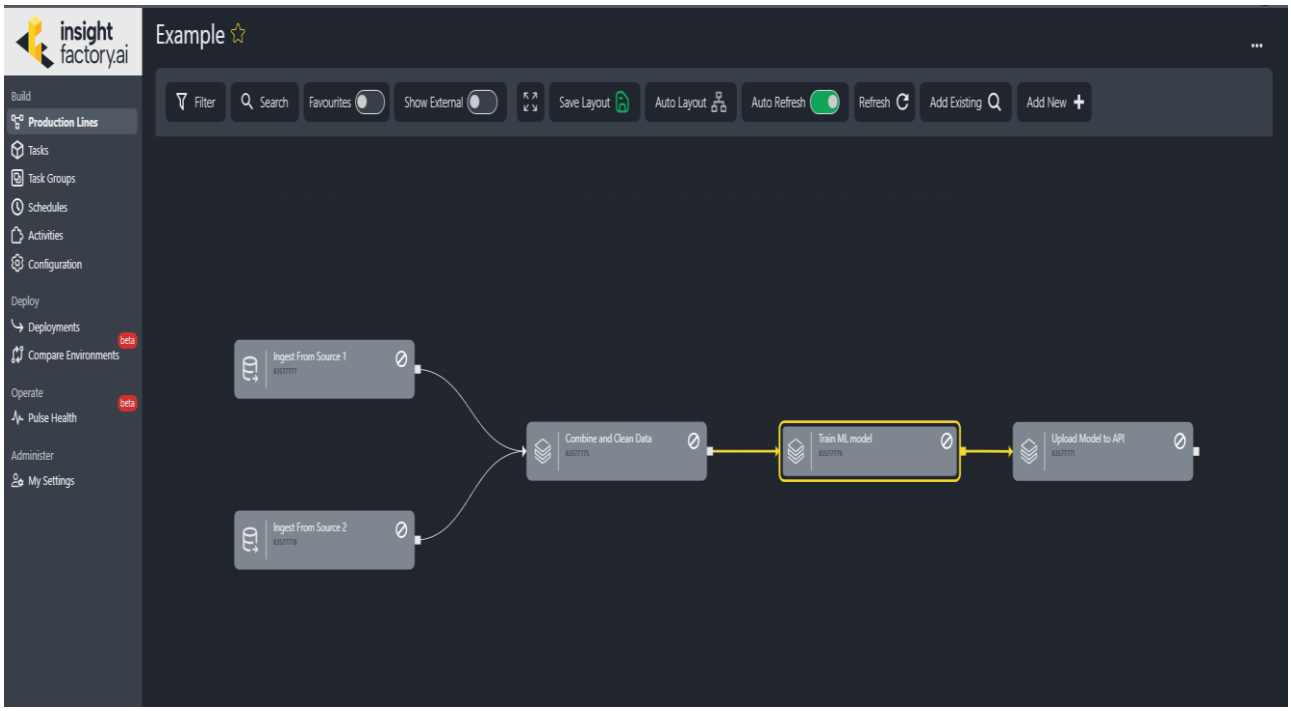
The dataset consists of two key components, the first being a history of observed rail breaks (where and when rail breaks have occurred) and the second component is the feature data, this includes railway sensor readings accumulated through special instruments that measures characteristics of the tracks current condition as trains traverse each segment. With this data, we can build predictive machine learning models to detect specific sections of track that have an elevated probability of breakage in a forward time window. In industry this is called a Predictive Maintenance model and allows engineers to take pre-emptive maintenance to prevent rail breaks from occurring.

Machine Learning Model – Predictive Maintenance

In this competition you will construct a machine learning classification model for Predictive Maintenance. The target variable (last 60 days) is constructed from the history of observed rail breaks, you will then use the feature data set to predict if a rail segment is in its last 60 days before a break occurs (True or False). We encourage you to create new features and test many different approaches to building and optimising your machine learning models.

The Insight Factory Platform

Should you choose to take on this project, you will be given access to the Insight Factory platform. The Insight Factory provides a logical way to organise and structure a series of interdependent tasks in what is called a 'Production Line'. Importantly, it leverages Azure Databricks to manage data and execute your code using cloud resources. An example production line for this project may look like this:



The first tasks on the left are responsible for connecting to an SQL Server Database, and ingesting the data into a Data Lake. The next task then calls a Databricks notebook which will access the data, aggregate the separate Database tables into one table and then performs some basic data cleaning. In the subsequent task, a machine learning model is trained on the data in the new table. The final task then uploads the model to our testing API, which calculates accuracy metrics and displays it on our live leaderboard.

If it is our project you decide to enrol in, we will run through how to use the platform in greater detail in a session later in the week.

Competition Details

Once every 2 weeks, your team will be required to submit your best performing model to our leaderboard API. The results of this will then be published, and your ranking position will be tied to your team for that round. We will also submit an insightfactory.ai benchmark model, which if you manage to beat, will award bonus marks. At regular intervals throughout the competition, we will provide guidance and advice on how you can improve your models.

If this all seems a little confusing at first, do not worry! There is plenty of time to familiarise yourself with the platform and the world of model building, and we will run support sessions to assist through the project. We believe that this is a really exciting opportunity to work with industry tools on a real project with real data. We look forward to working with you soon!