Deutsche Bahn Train Delay Predictor

Project Overview

This project builds a machine learning pipeline to predict train delays for Deutsche Bahn trains across Germany using historical scheduling and arrival data. The goal is to predict the delay in minutes given features such as train type, station, time of day, and day of the week.

Dataset

- **Source:** piebro/deutsche-bahn-data
- Data includes raw arrival and departure timestamps for DB trains nationwide.
- Used to compute delay by comparing scheduled vs actual arrival times.

Data Processing

- Data is fetched automatically via GitHub API to ensure up-to-date inputs.
- Cleaned and transformed to extract features such as hour, weekday, train type, and station one-hot encoding.
- Output is a processed dataset ready for training.

Model

- A Multilayer Perceptron (MLP) regression model implemented in PyTorch.
- Optimized with Adam optimizer and MSE loss.
- Metrics tracked: Mean Absolute Error (MAE) and R² score.
- Model saved as `model.pth` for inference.

API and Web Interface

- Flask API exposes a `/predict` endpoint that accepts JSON requests and returns delay predictions.
- Web interface built with Flask and HTML/CSS for user-friendly input of train details.
-

Monitoring and Logging

- Prediction requests and responses logged locally using SQLite (`logs.db`).
- Includes inference timing and drift detection on input feature distributions.

Deployment

- Deployment automation provided through `deploy.sh` for cloud hosting (Render, AWS EC2, Azure).
- Flask app served via Gunicorn and optionally proxied with Nginx.

Results

Metric	Value
MAE	3.2 mins
R² Score	0.68
Inference Time	<100ms

Future Work

- Explore gradient boosting models (XGBoost, LightGBM).
- Add geographical delay heatmaps and visualizations.
- Containerize with Docker for scalable deployment.

Contact and Contributions

- Project repository: [https://github.com/edisedis777/Deutsche-Bahn-Train-Delay-Predictor]
- Open to issues and pull requests to enhance features and performance.