Data Mining Project Outline HWS24

Gradient Descent Gururs - Predictive Modelling Train Delays

Raphael Ebner[2111615], Anh Tu Duong Nguyen[2115931], Adrian Scheibelhut[2110910], Georgios Terzidis[2114215] Harrison Walker[2118520]

October 13, 2024

Submitted to

Data and Web Science Group
Prof. Dr. Hertling
University of Mannheim

1 Introduction

Railway systems are crucial for public transportation, as maintaining punctual schedules is essential for operational efficiency and passenger satisfaction. In recent years, delays have become a significant concern for Deutsche Bahn, disrupting commuters and long-distance travellers. This project aims to develop a predictive model that categorises train delays based on various factors, providing detailed insights into their nature and severity. This will create a more structured framework for delay prediction.

The analysis will use historical data from Deutsche Bahn, including location, train type (e.g., RE, RB, S-Bahn), and external factors like weather and time of day. The project applies scientific methodologies and data mining techniques to address a real-world transportation problem, incorporating machine learning models aligned with state-of-the-art approaches.

From a business perspective, predicting delays could enable Deutsche Bahn to implement proactive strategies, optimise resource allocation and scheduling, and enhance customer satisfaction. Passengers would benefit from more accurate information on expected delays, improving their journey planning. Therefore, the project presents an academic challenge and the potential for meaningful business impact.

2 Methodology

This chapter outlines the methodology that is pursued for our group project. Section 2.1 explains the overall process on compiling the data for our project. Section 2.2 provides an overview on how we want to tackle the problems we have stated in Chapter 1.

2.1 Data Collection Methodology

We will either use a dataset from Kaggle[1] or collect our data by crawling. Although a dataset is available on Kaggle, it is rather small and only covers a week in July 2024 (08.07.2024 to 14.07.2024). To enhance the quality and performance of our analysis, we propose gathering additional data from the same source using the Timetables API and the StaDa API from the DB API marketplace. This approach allows us to collect more recent and extensive data while, if required, maintaining consistency with the source used for the Kaggle dataset. However, if crawling proves too challenging or time-consuming, we will use the existing Kaggle dataset. The dataset will contain the following data:

- Station
- Line
- Path
- Size of Station
- State
- City
- Postcode
- Planned Arrival
- Planned Departure
- Actual Arrival
- Actual Departure
- Announcements, i.e. construction on the track

[1] https://www.kaggle.com/datasets/nokkyu/deutsche-bahn-db-delays

2.2 Data Mining Methodology

Several pre-processing steps are required before beginning the main analysis.

Our pre-processing steps may include:

1. Aggregating data from multiple sources into a single dataset (when required).
2. Handling missing data, such as routes without end stations.
3. Transforming categorical variables into numerical values.
4. Binning the data based on factors such as delay duration, geographic regions, and station identifiers, to enable subsequent clustering.
5. Detecting and removing outliers, such as train rides from less frequented regions, as there may be insufficient data points for meaningful comparisons.

After pre-processing, we will explore the data to identify underlying trends and biases, such as geographic patterns in train journeys. We will use these insights to inform feature engineering for our models if we discover significant trends, such as increased delays during peak hours.

We will focus on regression algorithms for delay prediction when working with the kaggle dataset. Since most delays are less than six minutes, binning the data into multiple classes would result in overly narrow time intervals. Therefore, we plan to use interpolating regression trees to provide reasonable predictions based on the intervals within the training dataset.

If we work with the aforementioned custom dataset, we will use multiclass classification, as it performs better with imbalanced data. We plan to explore methods such as Decision Trees and Nearest Centroid classifiers for this purpose.

To ensure clarity and consistency, when comparing models to select the best one, we will only compare regression models against other regression models and classification models against other classification models. Cross-comparisons between regression and classification models will not be performed.

3 Evaluation of the Model

The delay prediction model will be evaluated by partitioning the dataset, with 20% of the data reserved for testing. This split is commonly regarded as a standard practice in data mining. To prevent overfitting, we will apply 10-fold cross-validation to the remaining 80% of the data. This approach allows us to make better use of the training data while also facilitating hyperparameter optimisation.

For multiclass classification, the performance will primarily be measured using the F1 score, as it balances precision and recall, which is particularly relevant when dealing with discrete classes. Additionally, a confusion matrix will be used to provide insights into how well the model distinguishes between delay categories, such as short and medium delays. We will also conduct a cost analysis to assess the impact of misclassifications, placing particular emphasis on high-impact errors. For example, from the user's perspective, incorrectly predicting a delay of 0–5 minutes when the actual delay is 30–35 minutes would be more detrimental than an error within smaller time intervals, such as predicting a 5–10 minute delay.

If regression is used, we will evaluate the model with the R^2 score and Root Mean Squared Error (RMSE). The R^2 score focuses on how much of the variance in the data is explained by the model, thus providing insight into its reliability. In contrast, RMSE measures the average difference between the predicted and actual values, offering a clear interpretation in the same units as the target variable.

4 Expected Results

The project aims to develop a model that accurately predicts train delays by both category and duration. While perfect accuracy is not expected, the goal is to provide reliable predictions that can support decision-making in train scheduling.

The expected outcomes include uncovering patterns that associate delays with specific factors, such as stations, times of day, and geographic regions. For example, certain stations may emerge as bottlenecks, experiencing more frequent or prolonged delays. Similarly, high-traffic regions may reveal delay trends. Additionally, external factors, such as holidays or local events, could be linked to extended delays in particular areas. Identifying these trends will contribute to a better understanding of the causes of train delays, ultimately improving scheduling and resource allocation.

Declaration of honor

I hereby declare that I have written the enclosed Bachelor's, Master's, seminar or project work without the help of third parties and without using any sources other than those listed and the aids listed in the table below, and that I have marked the passages taken from the sources used as such, either literally or in terms of content. This thesis has not yet been submitted to any examination authority in the same or a similar form. I am aware that a false declaration will have legal consequences.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT/GPT-4o	Rephrasing and Grammar correction	Throughout	++
Grammarly	Grammar correction	Throughout	++
ChatGPT/GPT-4o	Grammar correction	Throughout	+

Signature
Mannheim, den 13.October.2024