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 play a crucial role in public transportation, and maintaining on-time train schedules is essential for operational efficiency and passenger satisfaction. In recent years, delays have become a growing concern for Deutsche Bahn, leading to disruptions that affect commuters and long-distance travelers alike. To address this, we aim to create a predictive model that categorizes train delays based on various factors, offering more detailed insights into the nature and severity of delays. By focusing on delay categories, we can provide a more structured and practical prediction framework. Our project aims to address this issue by developing a predictive model that categorizes future train delays, providing insights into the extent of the delay. Using historical data from Deutsche Bahn, we will analyze variables such as location, train type (e.g., RE, RB, S-Bahn), and additional factors like weather or time of day. From an academic perspective, this project will contribute to the field of predictive analytics, particularly in transportation systems. The findings will not only advance data mining techniques but also offer a practical application of machine learning models to real-world problems. From a business perspective, the ability to predict delays could enable Deutsche Bahn to implement proactive measures to minimize disruptions. By understanding patterns in delays, Deutsche Bahn can allocate resources more effectively, optimize scheduling, and ultimately improve customer satisfaction. Furthermore, passengers may benefit from more accurate information about expected delays, helping them plan their journeys better. In this sense, our project offers both an academic challenge and the potential for a 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 the crawled dataset or the existing Kaggle[1] dataset. Although a dataset is already available on Kaggle, we propose crawling additional data using the same APIs: the Timetables and StaDa APIs from the DB API marketplace. This is because the existing dataset is limited to a single week in July (08.07.2024 to 14.07.2024), and we aim to enhance it with more recent data to improve performance, as larger datasets typically enable more accurate analysis. However, if the crawling process proves to be more challenging than anticipated, we will proceed with the project using the Kaggle dataset. The dataset will include the following data:

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

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

2.2 Data Mining Methodology

Due to the fact that we intend to crawl our data using the DB APIs, it becomes evident that there are several pre-processing steps required before the main work can start. At first we need to aggregate all data in a single dataset such that it mathces the existing dataset. Some additional potential steps are:

1. Tranformation, i.e. transforming discrete classes into numerical values
2. Binning according to delay, regions, stations etc. for following clustering
3. Potentially detecting and removing outliers such as train rides from unfrequent regions because there are hardly any data points to compare with

Subsequentely, we will explore the data to determine underlying trends and biases in the data, e.g., by analysing the geographical distribution of the train rides. Moreover, if certain trends are discovered, e.g., trains in the rush hours tend to have a much larger delay time, those findings can be used for feature engineering. In order to make a prediction, we focus on Regression algorithms if using the already existing dataset because the majority of the delays are less then six minutes which would make the time periods for binning the data in to multipile classes to small. To predict the delay, we intend to use the interpolating Regression Trees to give a reasonable prediction taken from the training dataset interval. Otherwise, if using a custom dataset, we want to utilise multiclass prediction because it tends to perform better with imbalanced data. Here, we aim to make use of other convential methods for classification like Decision Trees and Nearest Centroids. For reasons of comparability and simplicity between models, we compare only regression and classification among themselves and not perform cross checking.

3 Evaluation

The delay prediction model will be evaluated by partitioning the dataset, with 20% reserved for testing because this split is seen as the gold standard in data mining.. To avoid overfitting, 10-fold cross-validation will be applied to 80% of the data. Furthermore, this allows us to utilise more training data and perform hyperparamteter optimization. If doing multiclass classification, the performance will be primarily measured using the F1 score, which balances precision and recall, because we deal with discrete classes. Moreover, a confusion matrix will offer insight into misclassifications between delay categories, such as between medium and short delays. Additionally, a cost analysis will be conducted to assess the impact of these misclassifications, with a focus on high-impact errors because from a users perspective a predicted delay of 0-5 minutes vs. the actual delay of 30-35 minutes has a greater negative impact than predicting 5-10 minutes.
If using regression, we use R^2 score and Root Mean Squared Error (RMSE) because the first metric lies it's focus on the variance and therefore the reliabilty of the model, while the latter give insights in how far the predicted values are from the actual values, in the same units as the target variable.

4 Expected Results

The project aims to develop a model that accurately predicts train delays by category and duration, focusing on regional and express trains. While perfect accuracy is not expected, the model is intended to provide reliable predictions that inform decision-making in train scheduling. The expected results include identifying patterns that link delays to specific factors, such as stations, times of day, or regions. Certain stations may be identified as bottlenecks with more frequent or prolonged delays, while rush hours or high-traffic regions may show time-based delay trends. External factors, such as holidays or local events, may also contribute to longer delays in certain areas.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

Unterschrift
Mannheim, den 13.Oktober.2024