Are You Still Booking Train Tickets Under the Assumption of No Delays?

Sumitrra Bala Subramaniam [*1] **Jiha Kim** [*2] **Frederik Panse** [*3] **Kim Isabella Zierahn** [*4]

Abstract

In 2022, Deutsche Bahn reported that one in three of its trains was late, underscoring the importance of finding the most reliable route, not just the fastest. We conduct route optimization to find the most reliable route going from Stuttgart to Frankfurt (Main), analyze factors that influence the train delays, and create a Random Forest model, using data provided by Deutsche Bahn. The optimal route has less than half of the mean delay of the fastest route that Deutsche Bahn offers. The train delay is most influenced by the route's proximity to the German border, the (relative) number of train rides and the betweenness centrality. This project highlights the fact that opting for the minimum time is not always the best choice.

1. Introduction

If a group of students from Tübingen wants to take a train trip to Frankfurt (Main), there are many potential routes they could take. Would they be able to decide beforehand on the most reliable route and the best time to go?

People determine their train schedules under the assumption of no train delays. However, when Deutsche Bahn announced that its punctuality quote fell to 65.2 percent in 2022, this assumption has become problematic. When traveling from one place to another via Deutsche Bahn, passengers often can't rely on arriving at their destination at the scheduled time. Although Deutsche Bahn provides real-time information on train delays, passengers are not informed beforehand of likely delays. Deutsche Bahn is able to identify the fastest route, without considering the mean delay of this route (e.g., travel information). Previous studies often focused on predicting train delay (Nair et al., 2019; Berger et al., 2011; Hauck & Kliewer, 2020). Our goal is to identify the most reliable train route going from Stuttgart to Frankfurt (Main).

To analyze the most reliable route, we are following two approaches: route optimization (Section 2.1) and analysing external influences (Section 2.2). The first approach is to find all possible routes going from Stuttgart to Frankfurt (Main) and then identifying the route with the least mean delay (i.e., the most reliable route). Our results show that the optimal route differs from the fastest route that Deutsche Bahn offers (Section 3.1). The second approach is to investigate whether e.g., the day of the week, holidays, the station's distance to the border or number of trains per station impact the delay. Our findings display several factors influencing the delay (Section 3.2).

We shed further light on one factor: weekdays versus weekends. Regular passengers often use the train either to commute to work on weekdays or to get home on weekends, so it is important to consider the day of travel when choosing a route. As this information is relevant to end-consumers, we analyze whether the overall most reliable route is also optimal when considering only weekdays or weekends.

We aim to fill a gap in train delay research by shifting the focus from predicting delays to providing passengers with pre-trip information on the most reliable route and the best time to travel. This allows passengers to make informed decisions about their travel, allowing them to choose the optimal route with confidence, rather than relying solely on the fastest route.

2. Data and Methods

We use train delay data provided by Deutsche Bahn (DB) which consists of stations and stops that measured the number of trains and minutes of delay of all trains passing by in 2016. We are also using the directory of operating sites, containing the names and coordinates of the stations. For geographical information on routes of the rail network, we

[*]Equal contribution [1]Matrikelnummer 6642271, sumitrra.bala-subramaniam@student.uni-tuebingen.de, MSc Quantitative Data Science Methods [2]Matrikelnummer 6640082, jiha.kim@student.uni-tuebingen.de, MSc Quantitative Data Science Methods [3]Matrikelnummer 5810899, frederik.panse@student.uni-tuebingen.de, MSc Quantitative Data Science Methods [4]Matrikelnummer 6635183, kim-isabella.zierahn@student.uni-tuebingen.de, MSc Quantitative Data Science Methods.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201). Style template based on the ICML style files 2023.

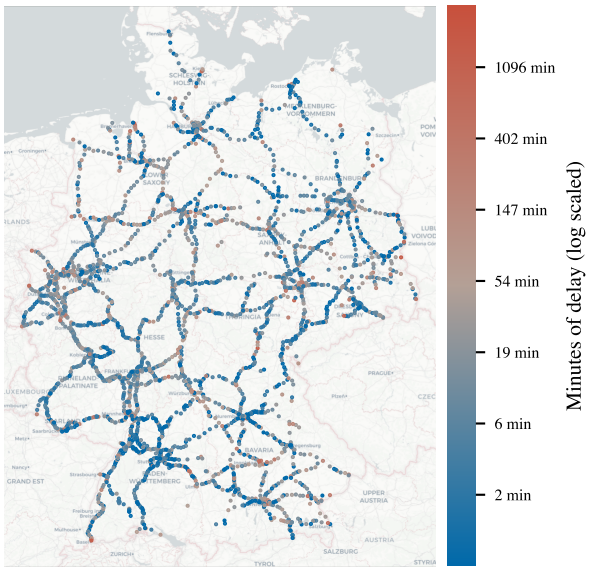

Figure 1. Mean delay of trains in 2016 for all available stations in Germany. Map data © OpenStreetMap contributors © CARTO.

use geographical route data. Our train delay data consists of 757 007 entries, measuring the delay all over Germany throughout the year 2016. Figure 1 shows an overview of all stops and stations and their mean delay in 2016. The mean delay of all stations in Germany in 2016 was 34.84 minutes. After merging all three datasets, our data consists of information about the stations and stops, their time stamp, their geographical data, their corresponding nearest routes and further information about the rails (e.g., railroad type, number of rails, speed limit or electrification). All analyses are conducted in Python (Van Rossum & Drake Jr, 1995).

2.1. Route Optimization

Our first approach is to find the optimal route from Stuttgart to Frankfurt (Main) (i.e., the route with the least mean delay). Using the geographical route data, we first get all continuous routes that exist in Germany. Then, all routes starting in Stuttgart are identified. From here, we identify all adjacent routes that connect to these starting routes. We repeatedly find adjacent routes until all possible routes are collected, starting in Stuttgart and terminating in Frankfurt (Main). Next, the route data is connected to the stations that measured the delay (i.e., the mean delay data). This is conducted by assigning the stations to the route to which it has a distance that is less or equal to 0.01 latitudes and longitudes. This translates to a maximum distance to the matching route of about 0.7 to 1.11 km, depending on the specific latitude or longitude. Finally, the mean delay is calculated for each possible route and the route with the lowest mean delay is determined.

To compare our optimal route with the fastest route offered by DB, we took the fastest route and its schedule from the DB website. We manually collect all the stations along the fastest route and calculate the mean delay for this route.

2.2. Factors Influencing Delays

The second approach is to identify factors that influence the delay. Several features are analyzed, including characteristics of the day (e.g., weekday, month, whether it is a holiday), characteristics of the station (e.g., the number of train rides at a station, the station's distance to the border) and centrality measures. We conduct a Random Forest model to examine which factors influence the train delay the most. It provides information about factors influencing the delay and their importance. According to Rogers & Gunn (2006), due to the random exploration of features, Random Forest lends itself to feature selection well and the measure of feature importance adopted here is the average information gain achieved during forest construction.

To determine if there is a more reliable route than the overall optimal one based on whether you are traveling on a weekday or weekend, we calculate the mean delay for each possible path on weekdays and weekends in addition to the overall mean. We then compare them visually.

2.2.1. Centrality Measures

Centrality measures demonstrate the level of interaction between stations, indicating how connected a station is to other stations. Betweenness centrality is a crucial indicator for a node (i.e., a train station) because it refers to how central it is for being between other stations (Freeman, 1977). It is crucial for the transportation of passengers (To, 2015) and measures the extent to which a station serves as a connection between other stations.

For a better understanding of the relationship between stations, we calculate the correlation between stations and only considered correlations that are 0.4 or higher. Then we calculate the centrality measures. Betweenness centrality for a station is calculated by counting how many times that station is part of the shortest paths between other stations. The more often it acts as a bridge or connector in the network, the higher its betweenness centrality (To, 2015).

The second centrality measure is closeness centrality. Closeness centrality for a station is a measure of how quickly it can reach all other stations in the network. If a station has high closeness centrality, it's relatively close to many other stations, indicating efficient communication or transportation within the network (Brandes, 2008).

3. Results

3.1. Route Optimization

There are 54 unique routes, all starting in Stuttgart and ending in Frankfurt (Main). We found the optimal route with the least mean delay, including 42 stations or stops. The mean delay for our optimal route is 14.19 minutes. The fastest route offered by DB contains 46 stations or stops with a mean delay of 35.65 minutes. Figure 2 displays a comparison of both routes.[1] Both routes start in a similar direction, via Vaihingen to Mannheim. Then they continue in two different directions. Our optimal route is via Worms, Mainz and Frankfurt Airport, while the fastest DB route is via Weinheim, Bensheim and Darmstadt to Frankfurt (Main).

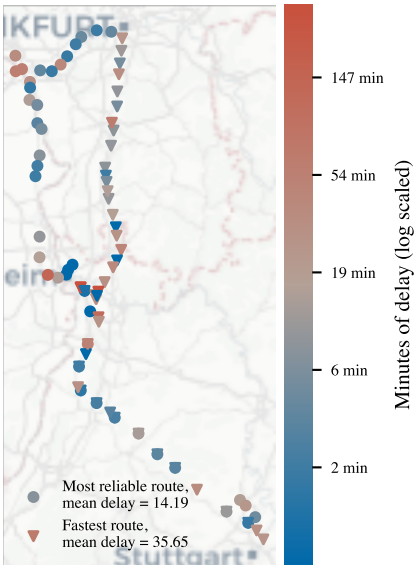

Figure 2. Most reliable route by our analyses and fastest route that Deutsche Bahn offers. Map data © OpenStreetMap contributors © CARTO.

3.2. Factors Influencing Delays

The dataset we use in this study is characterized by a significant imbalance, with a large proportion of non-delayed data. To tackle this, we apply downsampling, focusing on the count of delayed instances. The accuracy of the Random Forest model is 0.8229. Figure 3 gives an overview of all features and their importance for the delay.

[1]More visualizations and the code can be found on our github repository.

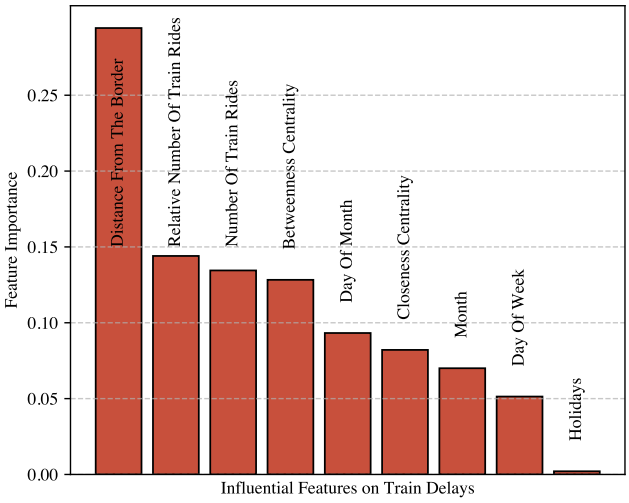

Figure 3. Feature Importance of Random Forest

3.2.1. Characteristics of the Day

Of all the factors we analyzed, the characteristics of the day, such as the weekday, month, or whether it is a holiday, have the least influence on the delay (Figure 3).

As shown in Figure 4, the overall most reliable route is also the most reliable route when only considering weekdays or weekends. The overall 7th most reliable route, however, has only about 3 seconds more mean delay on weekends than the the most reliable route. Figure 4 also reveals the systematic pattern that weekend travelling is more reliable than on weekdays. Trains have about 19 minutes less delay on weekends than on weekdays.

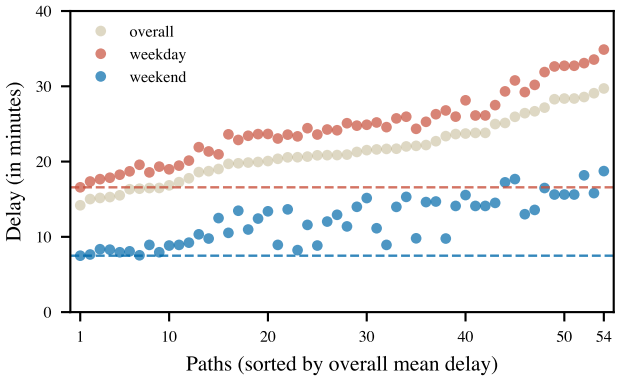

Figure 4. Comparing weekday and weekend mean delay for each route. The red and blue dashed lines show the mean delay for the most reliable route on weekdays and weekends, respectively.

3.2.2. Characteristics of the station

We consider three factors related to train stations, listed in order of variable importance: distance from the border, the ratio of the given date's train rides to the median of the usual train rides (relative number of train rides), and the number of train rides.

3.2.3. Centrality Measures

As seen in Figure 3, the betweenness centrality can serve as a very useful tool to plan their schedule because the more 'central' stations experience much more stress, resulting in high delays (To, 2015).

4. Discussion & Conclusion

The aim of our study is to find the most reliable route from Stuttgart to Frankfurt (Main). To do so, we conduct route optimization and propose an optimized route that has less than half of the mean delay than the fastest route proposed by DB. In addition to that, we analyze external factors that might influence train delay and find that the route's distance to the border of Germany, the (relative) number of train rides and the betweenness centrality influence train delay the most. Thus, the relative number of train rides is more significant than the absolute number. Each station has a capacity to manage operations, and if this capacity is exceeded, it will have a more substantial impact on delays than a simple increase in the number of trains.

We show that it is necessary to consider factors beyond simple travel time, such as the operational schedules of other trains and the centrality of specific stations depending on date train system. Additionally, DB operators might want to consider these factors when deciding train routes and schedules. Centrality metrics and capacity of train stations can be applied to train systems.

We consider data from 2016, which might be outdated in 2024 and had some missing values. For example, more data was available for weekdays than for weekends, possibly influencing our results. The data from the database used for the Random Forest model development is incomplete, incorporating only about 35.31% of the available records. Specifically, we consider data from stations with records available for at least 288 days, accounting for over 90% of this period. We also limit the route optimization for the route from Stuttgart to Frankfurt (Main), for now only outlining what a universal tool for finding reliable routes might look like.

The delays are computed on a daily average basis, potentially leading to varying results based on different time intervals. Furthermore, our analyses exclusively focus on delay times. A more comprehensive understanding could be achieved by constructing a model that considers both travel time and the probability of delays.

Our results propose that there is a more reliable route than the fastest route provided by DB. This and the identification of factors influencing the train delay enables passengers to knowingly decide on the optimal route and time to go from Stuttgart to Frankfurt (Main). Our results indicate that for a trip to Frankfurt (Main), students from Tübingen should take our proposed route, preferably on the weekend, and avoid routes or stations near the border with a high number of trains passing or high betweenness centrality.

Contribution Statement

Sumitrra Bala Subramaniam worked on connecting the datasets and route optimization. Kim Zierahn worked on connecting the datasets and on map visualizations. Jiha Kim worked on the factors influencing the delay and the Random Forest. Frederik Panse worked on factors influencing the delay and analyzing the difference between weekdays and weekends. All authors jointly wrote the text of the report.

References

Berger, A., Gebhardt, A., Müller-Hannemann, M., and Ostrowski, M. Stochastic delay prediction in large train networks. In *11th Workshop on Algorithmic Approaches for Transportation Modelling, Optimization, and Systems*, 2011. doi: https://doi.org/10.4230/OASIcs.ATMOS.2011.100.

Brandes, U. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008. ISSN 0378-8733. doi: https://doi.org/10.1016/j.socnet.2007.11.001.

Freeman, L. C. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977. doi: https://doi.org/10.2307/3033543.

Hauck, F. and Kliewer, N. Data analytics in railway operations: Using machine learning to predict train delays. In Neufeld, J. S., Buscher, U., Lasch, R., Möst, D., and Schönberger, J. (eds.), *Operations Research Proceedings 2019*. Springer International Publishing, 2020. doi: https://doi.org/10.1007/978-3-030-48439-2_90.

Nair, R., Hoang, T. L., Laumanns, M., Chen, B., Cogill, R., Szabó, J., and Walter, T. An ensemble prediction model for train delays. *Transportation Research Part C: Emerging Technologies*, 2019. doi: https://doi.org/10.1016/j.trc.2019.04.026.

Rogers, J. and Gunn, S. Identifying feature relevance using a random forest. In Saunders, C., Grobelnik, M., Gunn, S.,

and Shawe-Taylor, J. (eds.), *Subspace, Latent Structure and Feature Selection. SLSFS 2005.*, volume 3940 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, 2006. Springer. doi: 10.1007/11752790_12.

To, W. Centrality of an urban rail system. *Urban Rail Transit*, pp. 249–256, 2015. doi: https://doi.org/10.1007/s40864-016-0031-3.

Van Rossum, G. and Drake Jr, F. L. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.