10-9-2023

Punctuality Predictions in Public Transportation: Quantifying the Effect of External Factors

Tim Meyer-Hollatz
Fraunhofer Institute for Applied Information Technology FIT, Branch Business & Information Systems Engineering, Augsburg, Germany, tim.meyer-hollatz@fit.fraunhofer.de

Nina Schwarz
Fraunhofer Institute for Applied Information Technology FIT, Branch Business & Information Systems Engineering, Augsburg, Germany; University of Applied Science Augsburg, Augsburg, Germany, nina.schwarz@fit.fraunhofer.de

Tim Werner
Fraunhofer Institute for Applied Information Technology FIT, Branch Business & Information Systems Engineering, Augsburg, Germany; University of Applied Science Augsburg, Augsburg, Germany, tim.werner@fit.fraunhofer.de

Punctuality Predictions in Public Transportation: Quantifying the Effect of External Factors

Research Paper

Tim Meyer-Hollatz[1], Nina Schwarz[1,2], Tim Werner[1,2]

[1] Fraunhofer Institute for Applied Information Technology FIT, Branch Business & Information Systems Engineering, Augsburg, Germany
{tim.meyer-hollatz,nina.schwarz,tim.werner}@fit.fraunhofer.de
[2] University of Applied Science Augsburg, Augsburg, Germany
{nina.schwarz,tim.werner}@hs-augsburg.de

Abstract. Increasing availability of large-scale datasets for automatic vehicle location (AVL) in public transportation (PT) encouraged researchers to investigate data-driven punctuality prediction models (PPMs). PPMs promise to accelerate the mobility transition through more accurate prediction delays, increased customer service levels, and more efficient and forward-looking planning by mobility providers. While several PPMs show promising results for buses and long-distance trains, a comprehensive study on external factors' effect on tram services is missing. Therefore, we implement four machine learning (ML) models to predict departure delays and elaborate on the performance increase by adding real-world weather and holiday data for three consecutive years. For our best model (XGBoost) the average MAE performance increased by 17.33 % compared to the average model performance when only trained on AVL data enriched by timetable characteristics. The results provide strong evidence that adding information-bearing features improve the forecast quality of PPMs.

Keywords: *Punctuality Prediction, Public transportation, Machine Learning, Automatic Vehicle Location*

1 Introduction

While research on human movement in time and space has already been around for several decades (Weiner 1997), disruptive technologies and new mobility concepts in smart cities usher the need to leverage large amounts of data for providing cleaner, safer, and more efficient urban transportation (Mehmood et al. 2017). Especially local public transportation can be a factor in reducing emissions, preventing accidents, and democratizing mobility (Torre-Bastida et al. 2018; Yang et al. 2015). Previous research highlights the impact of punctuality and arrival predictions as a significant factor for potential passengers to switch to public transportation (Olsson and Haugland 2004; Ibrahim and Borhan 2020), thereby leveraging artificial intelligence (AI) to solve societal challenges and provide more sustainable mobility. To this end, data-driven punctuality

18th International Conference on Wirtschaftsinformatik,
September 2023, Paderborn, Germany

prediction models (PPMs) can significantly improve prediction performance (Shi et al. 2021) and amplify the reliability of public transportation when forecasting potential delays. Furthermore, data-driven PPMs allow local transportation companies to refine driving schedules, predict high volumes of passengers, and plan supporting trips.

Data-driven PPMs use historic train movements (HTM) consisting of AVL and timetable data as well as external factors to predict future transportation punctuality by abstracting rules and dependencies from previous disturbances such as labor strikes or severe snowfall and their effect on the mobility services (Ge et al. 2022; Zakeri and Olsson 2018). Although previous research has exemplarily shown that extreme weather conditions adversely affect punctuality in PT (Oneto et al. 2016; Tao et al. 2018) it has not yet investigated the potential increase in prediction performance by combining features from versatile data sources (Wang and Work 2015; Pongnumkul et al. 2014) and quantify the resulting impact. Therefore, it remains to be seen to what extent additional factors improve the PPMs overall performance over a long time period when compared to models solemnly built on HTM. Against this backdrop, we formulate the following research question (RQ):

How does enriching historical datasets with external factors influence the performance quality of punctuality predictions in public tram transportation?

We address the RQ by implementing and training four state-of-the-art ML models on different combinations of features aggregated from HTM, weather, and holiday data to predict the expected departure delay with a time horizon of a few days in advance ahead as the limit for small-scale weather projections is between hours and days (Bauer et al. 2015; Jung et al. 2010). We compare their performance in an empirical case study to evaluate the effect of enriching HTM data with additional features. The historical punctuality dataset contains real-world AVL data of a medium-sized German city over three consecutive years resulting in 1.4 million Trips at 222 stations.

We contribute to existing research by quantifying the increase in prediction accuracy when adding exogenous features to a large HTM dataset. Compared to existing research, we increase the amount of exogenous data sources and refine their granularity with respect to time (hourly) and events (public holidays, vacations) for weather and holiday data, respectively. Then we utilize four different state-of-the-art ML models proposed by the literature as PPMs to test our RQ. We could not only confirm extant research on the impact of weather and holidays but could further extend the findings to a higher granularity and for a complete tram network over multiple years. Understanding the interplay between different data streams for additional prediction performance potential is essential for policymakers, mobility companies, and passengers alike to increase the service level in PT (Olsson and Haugland 2004; Ibrahim and Borhan 2020).

2 Problem Context and Theoretical Background

Research on punctuality prediction in PT has grown significantly over the last few years (Ghofrani et al. 2018; Spanninger et al. 2020) and has been addressed in various studies

(Bešinović 2020; Ge et al. 2021). This development is strengthened by extended computing power and increased availability of larger datasets (Ghofrani et al. 2018). Following Ge et al. (2021), they can be comprised of endogenous data such as AVL, automatic fare collection (AFC), automatic passenger counting (APC), or rail network parameters (RNP) and exogenous data sources like weather, traffic, surveys, or social media. As a result, ML models have been developed to complement statistical approaches quantifying influencing factors on rail-based mobility (Hagenauer and Helbich 2017; Zúñiga et al. 2021).

The different approaches to predicting punctuality in public transportation can be classified by multiple aspects. One recent approach by Spanninger et al. (2020) distinguished research papers by categorizing them based on four characteristics. (I) First, which mathematical models are used for the analysis, e.g., time series analysis, artificial neural network. (II) Second, which input datasets are used to predict the delay. (III) Third, which type of output is considered, e.g., deterministic, single value prediction vs. stochastic, probability predictions. (IV) Fourth, based on the distinction between dynamic, based on real values at previous stations, and static forecasts, several hours or days in advance. As discussed in the introduction, this paper focuses on approaches that predict tram departure delays with a time horizon of a few days before the travel itinerary. Building on preliminary works by Ghofrani et al. (2018) and Spanninger et al. (2020), we focus on the theoretical background in characteristics (I) and (II) as they represent the focus of this research.

Spanninger et al. (2021) differentiate punctuality prediction models (PPM) as data-driven (based on historical punctuality records) or event-driven approaches (modeling the dependencies of sequential events). Due to the higher availability of data in the public transportation sector, data-driven approaches are increasingly the focus of ongoing research (Mesbah et al. 2015; Barabino et al. 2017; Shoman et al. 2020). Spanninger et al. (2021) further partition previous studies by their use of data sources, namely Historic Train Movements (HTM), Actual Delays (AD), Infrastructure Indicators (II), Timetable Properties (TP), and External Factors (EF) thereby differentiating between endogenous and exogenous factors as proposed by (Ge et al. 2021). Therefore, the key differences between data-driven PPMs in current research stem from different model selections or different datasets (Spanninger et al. 2020). Table 1 summarizes the most relevant publications with respect to this study, focusing on application area, data scope, selected features, implemented PPMs, and their evaluation metrics.

HTM datasets consisting of AVL and time table data are the basis for a wide range of data-driven approaches in the research field of punctuality predictions in public transportation (Wang and Work 2015; Shi et al. 2021; Spanninger et al. 2021; Ge et al. 2021). We can further differentiate these papers by comparing the different modes. Multiple research projects base their mobility predictions on HTM datasets for busses within urban regions (Pałys et al. 09.04.2022; Mandelzys and Hellinga 2010) or trains covering larger distances (Yaghini et al. 2013; Pongnumkul et al. 2014; Shi et al. 2021; Marković et al. 2015). However, less attention is paid to PPMs for highly intertwined rail-based transportation (i.e. trams) within urban environments that show higher sta-

tion frequencies and shorter distances (Rößler et al. 2021; Mesbah et al. 2015). Zychowski et al. (2018) predicted the travel time between two stations using a neural network.

Table 1. Overview of data-driven punctuality prediction models in current literature.

Literature	**Vehicle**	**Scope**	**Features**	**Models**	**Metric**
Wang and Work (2015)	Train	2 years, 100k trips	HTM, AD	Regression	RMSE
Mesbah et al. (2015)	Tram	5 years -	HTM, AD, EF	Regression	R^2
Laifa et al. (2021)	Train	1 year 12k trips	HTM, AD, EF	LigthGBM	RME, RMSE, R^2
Shi et al. (2021)	Train	6 months 30k trips	HTM, AD	XGB, Bayesian	RMSE
Arshad and Ahmed (2021)	Train	2 months 1k trips	HTM, AD, EF	RF	MAE, RMSE, R^2, Accuracy
Zychowski et al. (2018)	Tram	- 20k trips	HTM	ANN	Accuracy
Zhong et al. (2022)	Bus	7 days 10k trips	HTM, AD	RF, KNN, ANN	MAE, MAPE
Nimpanomprasert et al. (2022)	Bus	1 year -	HTM, AD	ANN, LSTM	RMSE
(Pałys et al. 2022)	Bus	30 days -	HTM, AD	KNN, Regression	MAE, STD
This study	Tram	3 years 1,4m trips	HTM, AD, EF	LightGBM XGBoost, RF, KNN	MAE, RMSE

The network was trained to conduct dynamic predictions using HTM features and information about the previous stop (Zychowski et al. 2018). In comparison, Mesbah et al. (2015) use weather and HTM features to predict travel time reliability. They leveraged an ordinary least square regression and limited the used HTM data to the morning peak timeframe (7 am - 9 am). The weather features were designed by calculating the average for the morning period as a stationary value (Mesbah et al. 2015). Many other studies extended their HTM data with weather events to test its effect on punctuality within PT on a time scale from several days to weeks (Ge et al. 2021). While the focus on weather impacts remains identical across all papers, the types of results vary.

Among others, Zakeri and Olsson (2018) focus on the effects of serve weather conditions on the punctuality of trains in Norway. Oneto et al. (2016) followed a similar approach by dynamically predicting the delay of trains in Italy. Compared to Zakeri and Olsson (2018), Oneto et al. (2016) focused on the influence of regular weather conditions and compare multiple models. Besides weather, (Laifa et al. 2021; Mesbah et al. 2015) use holidays as an additional factor. However, evaluating the impact of holiday variables on the overall prediction quality is rare.

While the effects of external factors such as weather and holidays on public transportation have already been the focus of several studies (Oneto et al. 2016; Zakeri and Olsson 2018; Wei et al. 2018; Liu et al. 2017) it remains unclear to what extent these findings are applicable for tram-based mobility. Mesbah et al. (2015) and Zychowski et al. (2018) provide the first evidence but focus on either small sample data sets or limited time frames. Thus, an analysis of a complete tram network for multiple timeframes and years has – to the best of our knowledge – not been investigated before. Furthermore, we increase the granularity as compared to daily or weekly weather aggregations for weather features (Zakeri and Olsson 2018; Chen et al. 2004). In addition, this study contributes to the literature by answering the call for further research regarding the effect of holidays (Laifa et al. 2021). Lastly, the aforementioned results are verified by implementing four different PPMs, whichave seen promising results in previous works (Shi et al. 2021; Nimpanomprasert et al. 2022).

3 Methodology

This work builds on the Cross-Industry Standard Process for Data Mining (CRISP-DM), commonly applied in empirical studies, to quantify the effect of external factors on punctuality prediction in public transportation. The proposed method increases understanding and cross-project comparability by avoiding mistakes through a standardized and structured research approach (Wirth and Hipp 2000). We slightly modified the proposed method by omitting the deployment step, as it is not within the scope of this study, and extending the evaluation step with a comprehensive comparison. The modified CRISP-DM process consists of five different steps, namely "Business Understanding", "Data Understanding", "Data Preparation", "Modelling" and "Evaluation & Comparison" (Wirth and Hipp 2000).

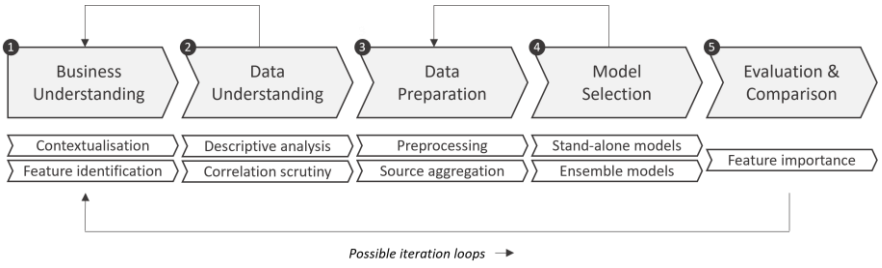

Figure 1. Derived five-step process for the quantification of external punctuality factors.

"Business Understanding": First, CRISP-DM starts by covering business understanding to identify specific challenges and strategic barriers. For our study, business understanding was primarily used to identify that weather and holiday events as features to be considered in improving the prediction of punctuality in public transportation (Laifa et al. 2021; Zakeri and Olsson 2018; Oneto et al. 2016) and to better understand the impact of delays in public transportation services for mobility providers and their passengers (Ge et al. 2021). Furthermore, this step includes general knowledge of the punctuality predictions in public transportation which has been displayed in Section 2.

"Data Understanding": Second, CRISP-DM ensures a profound understanding of the collected data, tightly linking it to the subsequent data preparation. We collected our combined dataset from three different sources. Subsequently, we implemented correlation scrutiny and descriptive analysis. Section 4 introduces both the dataset and the preprocessing steps.

"Data Preparation": Third, CRISP-DM provides a framework to fulfill the quality requirements of the underlying data. We cleansed the data from outliers and invalid, false data points during this process.

"Model Selection": Fourth, CRISP-DM provides a standardized guideline for the modeling and selection process detailed in Section 5. This study utilizes four PPMs to quantify the effect of external features on punctuality performance in public transportation.

"Evaluation & Comparison": Fifth, CRISP-DM assists in evaluating and comparing derived results. In this study, we assess two different error measures to substantiate our findings and underline the robustness of our results since single error measurements may not provide a holistic picture (Botchkarev 2019). For better comparison to the literature, we use the mean absolute error (MAE) as the primary evaluation metric when presenting the results and derived our own metric to ensure comparability with the existing research in Section 6. Afterward, Section 7 critically reviews specific results and derives implications, limitations, and prospects for further research.

4 Data Understanding and Preparation

As formulated in our research question, we aim to quantify the impact of additional features on the punctuality prediction of intra-urban public transportation. Therefore, we integrated data from different sources to cover the various aspects that influence the punctuality of tramways. In a second consecutive step, we processed the aggregated datasets. Finally, we derived the features for the machine learning algorithms.

Data Aggregation

In Section 2, we discussed which features were used in the existing literature to predict the delay of public transportation. Based on this research, we identified three available datasets for our region of interest. The primary dataset provided by Stadtwerke Augsburg GmbH includes parts of AVL data of the tramline network of Augsburg from May 2019 until August 2022. It comprises over 25 million records for five tram lines, each detailing a tram stop, including the arrival and departure times. The weather dataset provided by Meteostat contains weather information for each station and hour within

the historic train movements. Finally, the holiday dataset, provided by the German Ministerial Culture Conference, includes information about holidays, public holidays and long weekends during the same timeframe. The latter two are publicly accessible.

Preprocessing of data

During the preprocessing steps, we excluded tram runs that had either transmission errors (e.g., wrong line number, no actual departure time) or corrupt data points (e.g., significant outliers). This procedure had several implications for our results. Foremost, our prediction focuses on tram delays and excludes tram breakdowns since they are indistinguishable from corrupted data points. Furthermore, our preprocessing methods reduced the number of records from 25 million to 14 million.

Figure 2 shows an excerpt from the final dataset. The plot on the left-hand side illustrates the distribution of the departure delay across the considered data points. Most data records exhibit low and positive departure delays, indicating that trams tend to leave stations belated rather than too early, which aligns with the policy of tram providers. The remaining plots highlight increased departure delays during usual working times compared to evening hours and the effect of different geolocations of tram stations concerning departure delay.

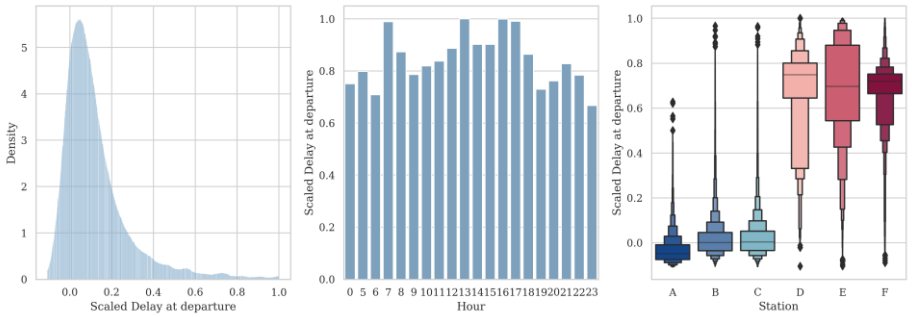

Figure 2. Summary plots for the historic train movements.

Data transformation & Feature selection

After the preprocessing steps, our final dataset contains 14 million punctuality records. Based on the literature review conducted in Section 2 and multiple explorative data analysis, we focused on 16 features divided into three feature sets. The first set contains features describing the historic tram movements and time data. The remaining two sets of features describe external influences on the tram movements, namely weather and holidays. Table 2 displays the variables and their respective specifications. As introduced in Section 1, this study forecasts the delay from multiple hours to days afront. The delay can be expressed in various formats reaching from different categories such as arrival delay (AD), dwell delay (DWD), driving time delay (DTD), and departure delay (DED) to different calculation methods, i.e., cumulative versus relative

Table 2. Description of all features used in this study.

Variables	Values
Historic Tram Movements (HTM)	
Daily number of departures at a station	Numeric
Line number	Categorical
Direction	Binary
Weekday	Categorical
Hour	Categorical
One week rolling mean departure delay by stop	Numeric (s)
One year rolling mean departure delay by stop	Numeric (s)
One week rolling mean departure delay by line number	Numeric (s)
One year rolling mean departure delay by line number	Numeric (s)
Weather (W)	
Temperature	Categorical
Snow height	Numeric (mm)
Windspeed	Categorical
Precipitation	Categorical
Holiday (H)	
School holidays	Binary
Public holidays	Binary
Long weekend (bank holiday + extra day)	Binary

*The categories for the weather features are based on Baumgarte et al. (2022).

This study focuses on the DED, which is calculated as follows:

$$DED_i = AD_i + DWD_i, \qquad i \in \{1, \dots, N\}, \tag{1}$$

where the arrival delay at station *i* is determined by

$$AD_i = AD_{i-1} + DWD_{i-1} + DTD_{i-1,i} \tag{2}$$

and $N \in \mathbb{N}$ is the number of stations within the transportation network. Altogether the DED combines potential arrival delays at station *i-1* (previous station), driving time delays between station *i* and *i-1*, and unscheduled waiting times at stations *i-1* and *i* and takes positive values in case of a departure delay and negative values in case of an early departure.

5 Model Selection, Fitting & Evaluation

After the data preprocessing, we conducted an intensive literature review to identify various ML models suited for DED prediction with reasonable optimization runtime and scalability for our datasets. We evaluated the selected models based on performance, accuracy, interpretability, and runtime.

Model selection

The ensemble models XGBoost, LightGBM, and the stand-alone KNN model improved the punctuality predictions or outperformed other comparable models' accuracy

or runtime (Shi et al. 2021; Laifa et al. 2021; Pongnumkul et al. 2014). These studies were all based on HTM data. In addition, we chose a Random Forest Regressor (RF) as a fourth model since its implementation by Arshad and Ahmed (2021) has already achieved good prediction results for the combination of HTM and weather data considering different error measures.

Model fitting

Recent studies underline that tuned hyperparameters outperform the default settings of machine learning libraries. Therefore, searching for the best hyperparameter configuration has become increasingly important (Truong et al. 2019). However, tuning a model to find the best combination is costly. Besides higher costs, manual hyperparameter tuning leads to reduced reproducibility. Since this problem is widely acknowledged, several publications and libraries provide possible solutions (Hutter et al. 2019). Automated Machine Learning (AutoML) summarizes these solutions aiming to improve the availability of ML functionalities, reduce training costs, and improve reproducibility while maintaining high accuracy (Truong et al. 2019). Therefore, we leverage a recent AutoML library called FLAML. Researchers at Microsoft developed FLAML to balance the tradeoff between trial costs (i.e., CPU costs for training) and trial error (i.e., accuracy of the model) (Wang et al. 2021b). The framework uses Blendsearch for hyperparameter tuning, combining global and local search (Wang et al. 2021a). Overall, the first benchmarking results of FLAML suggest that it outperforms existing AutoML frameworks for shorter (1h) and longer runs (4h) (Wang et al. 2021b; Gijsbers et al. 2019).

We fitted every model separately using the provided learners and their corresponding hyperparameter spaces of the FLAML library to answer our RQ, leading to the best possible configurations within the selected time budget. We trained each model to minimize the mean absolute error using a holdout validation set containing 25 % of the data, a 5-fold cross-validation on the test set containing the remaining 75 % of the entire dataset and no additional preprocessing. Considering the call for more comprehensible AI research, we provide an in-depth description of the hyperparameter tuning according to the specifications of Kühl et al., along with corresponding results upon request (Kühl et al. 2021).

Model evaluation

A common problem in punctuality prediction is that historic punctuality data underlies strict regulations in terms of publishing. The dependencies of the ML evaluation functions (e.g., MAE and RMSE) on the volatility of the test data set make them directly connected. This results in the fact that the MAEs per research project are not directly comparable, limiting the comparability between different studies. Based on prior works by Wang and Work (2015), we introduce an error metric that compares the mean of the differences between the planned departure times according to the timetable and the historic departure times ($DED_{timetable}$) and the mean of the predicted departure delay of the model ($DED_{predicted}$). The ratio between these two mean absolute errors, as defined in Equation 3, acts as a quality metric for the result that is independent of the granularity of the baseline values. By reporting the quality metric and the model characteristics, such as mean absolute error (MAE), and root mean squared error (RMSE), we guarantee the comparability of our results.

$$QM = 1 - \frac{MAE\left(DED_{predicted}\right)}{MAE\left(DED_{timetable}\right)} = 1 - \frac{MAE_{model}}{MAE_{baseline}} \quad (3)$$

6 Results

This section presents the prediction performance results of our large-scale data-driven PPMs, focusing on the differences between historical and enriched feature sets. We evaluate our models using a 5-fold cross-validation on the training dataset and summarize the resulting error metrics. To this end, our PPMs are presented on the x-axis. In contrast, the y-axis shows the previously introduced quality metric (3) (left-hand side) and the improvements compared to the baseline RMSE, which is calculated by using the RMSE instead of the MAE in Equation 3 (right-hand side). Figure 3 illustrates our results. The y-axis of the RMSE figure is scaled to ensure better readability of the results.

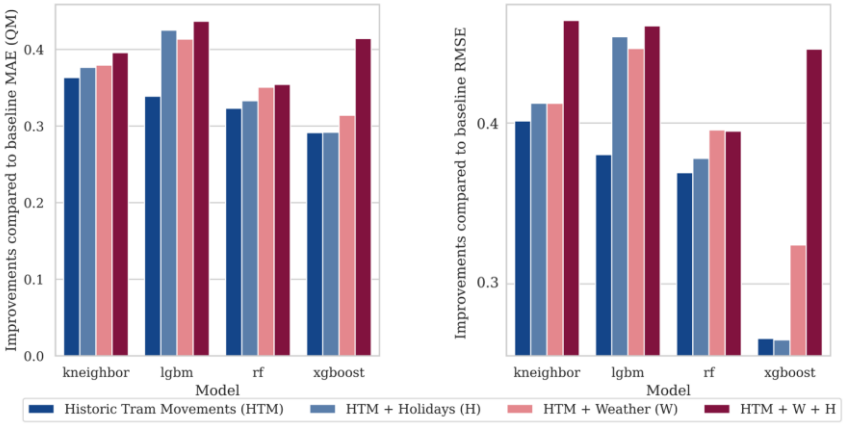

Figure 3. Performance evaluation for PPMs benchmarked against the deviations in the train schedule.

Considering both evaluation metrics, all models tend to improve performance when trained on enriched datasets. XGBoost exhibits the most significant relative improvements of 17.33 % comparing historical data and the complete feature set. While the relative percentage might vary between the models, the overall tendency prevails - the larger the set of uncorrelated features, the better the quality metric. Overall, the RF model and the KNN improved gradually over each feature extension. Table 3 additionally provides an overview of the two-error metrics and our quality metric for each of the four data-driven PPMs grouped by the underlying dataset, HTM, HTM + W, HTM + H, and HTM + W + H, to increase understandability and comparability. We notice that all PPMs trained on the enriched dataset outperform the models trained exclusively on HTM data.

Table 3. Error metrics and standard deviation for all PPMs.

Data-set	Model	MAE	RMSE	QM	Relative Improve-ment
HTM	KNN	57.75	81.70	36.33	-
	LightGBM	59.95	84.91	33.90	-
	RF	61.38	86.55	32.33	-
	XGBoost	64.26	99.37	29.15	-
HTM + H	KNN	56.54	79.94	37.66	2.11 %
	LightGBM	52.13	72.78	42.52	13.04 %
	RF	60.49	85.25	33.30	1.45 %
	XGBoost	64.22	99.47	29.19	0.06 %
HTM + W	KNN	56.27	79.94	37.96	2.56 %
	LightGBM	53.19	74.11	41.36	11.28 %
	RF	58.89	82.57	35.07	4.06 %
	XGBoost	62.21	92.61	31.41	3.20 %
HTM + W + H	KNN	54.79	70.91	39.59	5.13 %
	LightGBM	51.08	71.54	43.68	14.81 %
	RF	58.55	82.70	35.44	4.61 %
	XGBoost	53.12	74.18	41.43	17.33 %

Moreover, they exhibit greater robustness and generalization abilities than their counterparts trained on singular aggregated data.

We benchmarked each feature against the results of the HTM data to link the improved model performance to the selected features. We achieved this by implementing four ML models for each feature set and comparing the tuple (model and feature set) to the HTM result of the same model while ruling out secondary effects since they are inherent in both training sets. Table 4 shows the average improvement for selected categorical features.

Table 4. XGBoost MAE comparison for selected exogenous features.

Exogenous event	XGBoost (HTM)	XGBoost (HTM+W/HTM+H)
Temperature below -10 °C	58.61	55.68 (-5.0%)
Precipitation over 50 $\frac{mm}{cm^2}$	59.18	56.39 (-4.7%)
Fresh snow (30 – 60 cm)	56.50	53.89 (-4.6%)
Evening hours of public holidays	58.53	58.37 (-0.3%)

7 Discussion and Implications

This study quantifies the effect of exogenous factors on tram punctuality predictions building on the work of Mesbah et al. (2015). We implement four different ML models derived from relevant literature and analyze prediction performance by adding seven exogenous features to the HTM baseline dataset. We train all models on permutations of the large-scale exogenous datasets to study performance for each category as well as individual features to rule out secondary effects. The best overall model, LightGBM, yields an MAE of 51.08s or 43.76 % QM when trained on HTM enriched with four weather and three holiday features.

Our results have various managerial and academic implications. First, we provided academia with a method to quantify punctuality prediction results based on a quality metric that allows researchers to compare their findings across different datasets. Second, we show that AutoML methods can be used in the field of punctuality prediction to train profound PPMs with reasonable computational time. Third, we verify previous findings for urban tram networks and quantify the effect of enriching HTM data with exogenous factors while ruling out secondary effects. Fourth, we confirm our results on an entire transportation network containing multiple stations and lines compared to previous research, mainly focusing on single lines or single stations. Fifth, we study the effects of more exogenous data (holidays and weather) for a longer time period with finer granularity compared to relevant literature.

Although, to our knowledge, this is the first study to investigate an entire tram network, our findings are limited to one transportation network within one city. Furthermore, the number of weather stations could be increased to more precisely depict hyperlocal weather development. While we see a performance increase when adding more exogenous factors, further research should investigate its impact on computational power as well as financial and ecological costs. Finally, we encourage the research community to study both the influence of endogenous and exogenous factors based on explainable ML models and broaden the scope by adding more transportation networks to foster well-informed managerial decision-making within the PT sector.

8 Conclusion

In this study, we address the RQ of whether and to what extent exogenous factors influence the performance of PPMs in PT after adequate training and tuning efforts. We derive four state-of-the-art ML models from relevant literature presented in Section 2 and implement them on a large-scale mobility dataset with over 14 million real-world punctuality records of a medium-sized German city. Our results suggest a fundamental link between PPM performance improvement and enriched HTM data for tram mobility. By answering our research question, we contribute to investigating the correlation between external features and data-driven punctuality prediction models in public rail-based transport. We confirm extant research that weather and holiday features positively influence the overall model performance. We further extended the existing discourse by proving fine granular weather and holiday features influence the prediction quality and, therefore, the delay of trams within a whole tram network.

References

Arshad, Mohd/Ahmed, Muqeem (2021). Train Delay Estimation in Indian Railways by Including Weather Factors Through Machine Learning Techniques. Recent Advances in Computer Science and Communications 14 (4), 1300–1307. https://doi.org/10.2174/2666255813666190912095739.

Barabino, Benedetto/Di Francesco, Massimo/Mozzoni, Sara (2017). An Offline Framework for the Diagnosis of Time Reliability by Automatic Vehicle Location Data. IEEE Transactions on Intelligent Transportation Systems 18 (3), 583–594. https://doi.org/10.1109/TITS.2016.2581024.

Bauer, Peter/Thorpe, Alan/Brunet, Gilbert (2015). The quiet revolution of numerical weather prediction. Nature 525 (7567), 47–55. https://doi.org/10.1038/nature14956.

Baumgarte, Felix/Keller, Robert/Röhrich, Felix/Valett, Lynne/Zinsbacher, Daniela (2022). Revealing influences on carsharing users' trip distance in small urban areas. Transportation Research Part D: Transport and Environment 105, 103252. https://doi.org/10.1016/j.trd.2022.103252.

Bešinović, Nikola (2020). Resilience in railway transport systems: a literature review and research agenda. Transport Reviews 40 (4), 457–478. https://doi.org/10.1080/01441647.2020.1728419.

Botchkarev, Alexei (2019). A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms. Interdisciplinary Journal of Information, Knowledge, and Management 14, 45–76. https://doi.org/10.28945/4184.

Chen, Mei/Liu, Xiaobo/Xia, Jingxin/Chien, Steven I. (2004). A Dynamic Bus-Arrival Time Prediction Model Based on APC Data. Computer-Aided Civil and Infrastructure Engineering 19 (5), 364–376. https://doi.org/10.1111/j.1467-8667.2004.00363.x.

Ge, Liping/Sarhani, Malek/Voß, Stefan/Xie, Lin (2021). Review of Transit Data Sources: Potentials, Challenges and Complementarity. Sustainability 13 (20), 11450. https://doi.org/10.3390/su132011450.

Ge, Liping/Voß, Stefan/Xie, Lin (2022). Robustness and disturbances in public transport. Public Transport 14 (1), 191–261. https://doi.org/10.1007/s12469-022-00301-8.

Ghofrani, Faeze/He, Qing/Goverde, Rob M.P./Liu, Xiang (2018). Recent applications of big data analytics in railway transportation systems: A survey. Transportation Research Part C: Emerging Technologies 90, 226–246. https://doi.org/10.1016/j.trc.2018.03.010.

Gijsbers, Pieter/LeDell, Erin/Thomas, Janek/Poirier, Sébastien/Bischl, Bernd/Vanschoren, Joaquin (2019). An Open Source AutoML Benchmark. AutoML Workshop at ICML 2019.

Hagenauer, Julian/Helbich, Marco (2017). A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications 78, 273–282. https://doi.org/10.1016/j.eswa.2017.01.057.

Hutter, Frank/Kotthoff, Lars/Vanschoren, Joaquin (2019). Automated Machine Learning. Cham, Springer International Publishing.

Ibrahim, Ahmad Nazrul Hakimi/Borhan, Muhamad Nazri (2020). The Interrelationship Between Perceived Quality, Perceived Value and User Satisfaction Towards Behavioral Intention in Public Transportation: A Review of the Evidence. International Journal on Advanced Science, Engineering and Information Technology 10 (5), 2048. https://doi.org/10.18517/ijaseit.10.5.12818.

Jung, T./Miller, M. J./Palmer, T. N. (2010). Diagnosing the Origin of Extended-Range Forecast Errors. Monthly Weather Review 138 (6), 2434–2446. https://doi.org/10.1175/2010MWR3255.1.

Kühl, Niklas/Hirt, Robin/Baier, Lucas/Schmitz, Björn/Satzger, Gerhard (2021). How to Conduct Rigorous Supervised Machine Learning in Information Systems Research: The Supervised Machine Learning Report Card. Communications of the Association for Information Systems 48 (1), 589–615. https://doi.org/10.17705/1CAIS.04845.

Laifa, Hassiba/khcherif, Raoudha/Ben Ghezalaa, Henda Hajjami (2021). Train delay prediction in Tunisian railway through LightGBM model. Procedia Computer Science 192, 981–990. https://doi.org/10.1016/j.procs.2021.08.101.

Liu, Chengxi/Susilo, Yusak O./Karlström, Anders (2017). Weather variability and travel behaviour – what we know and what we do not know. Transport Reviews 37 (6), 715–741. https://doi.org/10.1080/01441647.2017.1293188.

Mandelzys, Michael/Hellinga, Bruce (2010). Identifying Causes of Performance Issues in Bus Schedule Adherence with Automatic Vehicle Location and Passenger Count Data. Transportation Research Record Journal of the Transportation Research Board 2143 (1), 9–15. https://doi.org/10.3141/2143-02.

Marković, Nikola/Milinković, Sanjin/Tikhonov, Konstantin S./Schonfeld, Paul (2015). Analyzing passenger train arrival delays with support vector regression. Transportation Research Part C: Emerging Technologies 56, 251–262. https://doi.org/10.1016/j.trc.2015.04.004.

Mehmood, Rashid/Meriton, Royston/Graham, Gary/Hennelly, Patrick/Kumar, Mukesh (2017). Exploring the influence of big data on city transport operations: a Markovian approach. International Journal of Operations & Production Management 37 (1), 75–104. https://doi.org/10.1108/IJOPM-03-2015-0179.

Mesbah, Mahmoud/Lin, Johnny/Currie, Graham (2015). “Weather” transit is reliable? Using AVL data to explore tram performance in Melbourne, Australia. Journal of Traffic and Transportation Engineering (English Edition) 2 (3), 125–135. https://doi.org/10.1016/j.jtte.2015.03.001.

Nimpanomprasert, Thummaporn/Xie, Lin/Kliewer, Natalia (2022). Comparing two hybrid neural network models to predict real-world bus travel time. Transportation Research Procedia 62, 393–400. https://doi.org/10.1016/j.trpro.2022.02.049.

Olsson, Nils O.E./Haugland, Hans (2004). Influencing factors on train punctuality—results from some Norwegian studies. Transport Policy 11 (4), 387–397. https://doi.org/10.1016/j.tranpol.2004.07.001.

Oneto, Luca/Fumeo, Emanuele/Clerico, Giorgio/Canepa, Renzo/Papa, Federico/Dambra, Carlo/Mazzino, Nadia/Anguita, Davide (2016). Advanced Analytics for Train Delay Prediction Systems by Including Exogenous Weather Data. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17.10.2016 - 19.10.2016. IEEE, 458–467.

Pałys, Łukasz/Ganzha, Maria/Paprzycki, Marcin (2022). Applying machine learning to predict behavior of bus transport in Warsaw, Poland.

Pałys, Łukasz/Ganzha, Maria/Paprzycki, Marcin (2022). Machine Learning for Bus Travel Prediction. In: International Conference on Computational Science. Springer, Cham, 703–710.

Pongnumkul, Suporn/Pechprasarn, Thanakij/Kunaseth, Narin/Chaipah, Kornchawal (2014). Improving arrival time prediction of Thailand's passenger trains using historical travel times. 11th Int. Joint Conf. on Computer Science and Software Engineering (JCSSE), 307–312. https://doi.org/10.1109/JCSSE.2014.6841886.

Rößler, David/Reisch, Julian/Hauck, Florian/Kliewer, Natalia (2021). Discerning Primary and Secondary Delays in Railway Networks using Explainable AI. Transportation Research Procedia 52, 171–178. https://doi.org/10.1016/j.trpro.2021.01.018.

Shi, Rui/Xu, Xinyue/Li, Jianmin/Li, Yanqiu (2021). Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimization. Applied Soft Computing 109, 107538. https://doi.org/10.1016/j.asoc.2021.107538.

Shoman, Maged/Aboah, Armstrong/Adu-Gyamfi, Yaw (2020). Deep Learning Framework for Predicting Bus Delays on Multiple Routes Using Heterogenous Datasets. Journal of Big Data Analytics in Transportation 2 (3), 275–290. https://doi.org/10.1007/s42421-020-00031-y.

Spanninger, Thomas/Trivella, Alessio/Büchel, Beda/Corman, Francesco (2021). A Review of Train Delay Prediction Approaches. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3964737.

Spanninger, Thomas/Trivella, Alessio/Corman, Francesco (2020). Approaches for real-time train delay prediction. https://doi.org/10.3929/ethz-b-000420036.

Tao, Sui/Corcoran, Jonathan/Rowe, Francisco/Hickman, Mark (2018). To travel or not to travel: 'Weather' is the question. Modelling the effect of local weather

conditions on bus ridership. Transportation Research Part C: Emerging Technologies 86, 147–167. https://doi.org/10.1016/j.trc.2017.11.005.

Torre-Bastida, Ana Isabel/Del Ser, Javier/Laña, Ibai/Ilardia, Maitena/Bilbao, Miren Nekane/Campos-Cordobés, Sergio (2018). Big Data for transportation and mobility: recent advances, trends and challenges. IET Intelligent Transport Systems 12 (8), 742–755. https://doi.org/10.1049/iet-its.2018.5188.

Truong, Anh/Walters, Austin/Goodsitt, Jeremy/Hines, Keegan/Bruss, C. Bayan/Farivar, Reza (2019). Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), Portland, OR, USA, 04.11.2019 - 06.11.2019. IEEE, 1471–1479.

Wang, Chi/Wu, Qingyun/Huang, Silu/Saied, Amin (2021a). Economical Hyperparameter Optimization with Blended Search Strategy. In: The Ninth International Conference on Learning Representations (ICLR 2021).

Wang, Chi/Wu, Qingyun/Weimer, Markus/Zhu, Erkang (2021b). FLAML: A Fast and Lightweight AutoML Library. In: Fourth Conference on Machine Learning and Systems (MLSys 2021).

Wang, Ren/Work, Daniel B. (2015). Data Driven Approaches for Passenger Train Delay Estimation. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, 2015 IEEE 18th International Conference on Intelligent Transportation Systems - (ITSC 2015), Gran Canaria, Spain, 15.09.2015 - 18.09.2015. IEEE, 535–540.

Wei, Ming/Corcoran, Jonathan/Sigler, Thomas/Liu, Yan (2018). Modeling the Influence of Weather on Transit Ridership: A Case Study from Brisbane, Australia. Transportation Research Record Journal of the Transportation Research Board 2672 (8), 505–510. https://doi.org/10.1177/0361198118777078.

Weiner, Edward (1997). Urban Transportation Planning in the United States. An Historical Overview. U.S. Department of Transportation.

Wirth, Rüdiger/Hipp, Jochen (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Yaghini, Masoud/Khoshraftar, Mohammad M./Seyedabadi, Masoud (2013). Railway passenger train delay prediction via neural network model. Journal of Advanced Transportation 47 (3), 355–368. https://doi.org/10.1002/atr.193.

Yang, Wenyue/Li, Tao/Cao, Xiaoshu (2015). Examining the impacts of socio-economic factors, urban form and transportation development on CO2 emissions from transportation in China: A panel data analysis of China's provinces. Habitat International 49, 212–220. https://doi.org/10.1016/j.habitatint.2015.05.030.

Zakeri, Ghazal/Olsson, Nils O. E. (2018). Investigating the effect of weather on punctuality of Norwegian railways: a case study of the Nordland Line. Journal of

Modern Transportation 26 (4), 255–267. https://doi.org/10.1007/s40534-018-0169-7.

Zhong, Gang/Yin, Tingting/Li, Linchao/Zhang, Jian/Zhang, Honghai/Ran, Bin (2022). Bus Travel Time Prediction Based on Ensemble Learning Methods. IEEE Intelligent Transportation Systems Magazine 14 (2), 174–189. https://doi.org/10.1109/MITS.2020.2990175.

Zúñiga, Felipe/Muñoz, Juan Carlos/Giesen, Ricardo (2021). Estimation and prediction of dynamic matrix travel on a public transport corridor using historical data and real-time information. Public Transport 13 (1), 59–80. https://doi.org/10.1007/s12469-020-00255-9.

Zychowski, Adam/Junosza-Szaniawski, Konstanty/Kosicki, Aleksander (2018). Travel Time Prediction for Trams in Warsaw. In: International Conference on Computer Recognition Systems. Springer, Cham, 53–62.