**Deutsche Bahn Delay：**

**A Study of Official Statistical Bias**

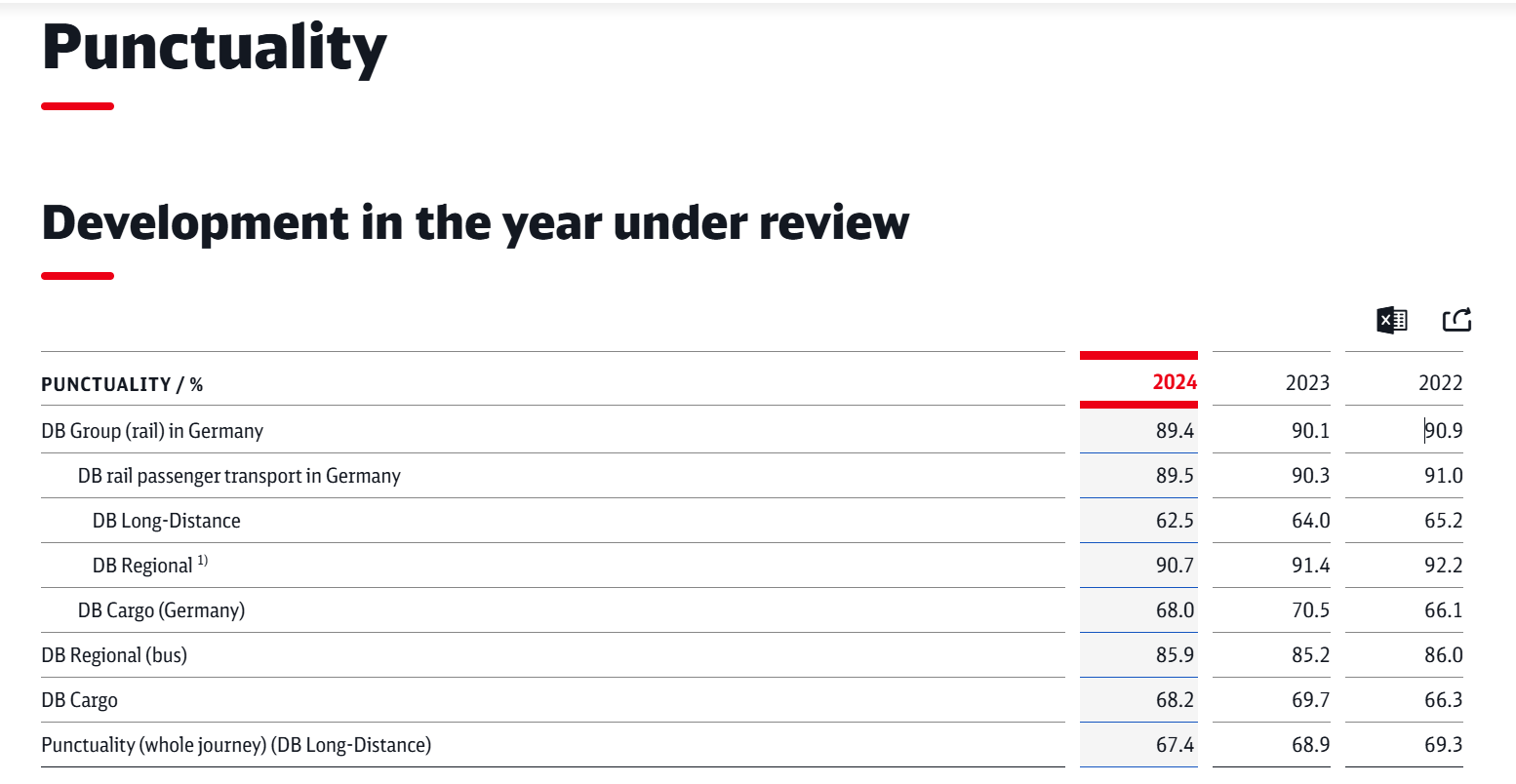
**and Empirical Modeling**

Xiao Shen, Yuanye Xu

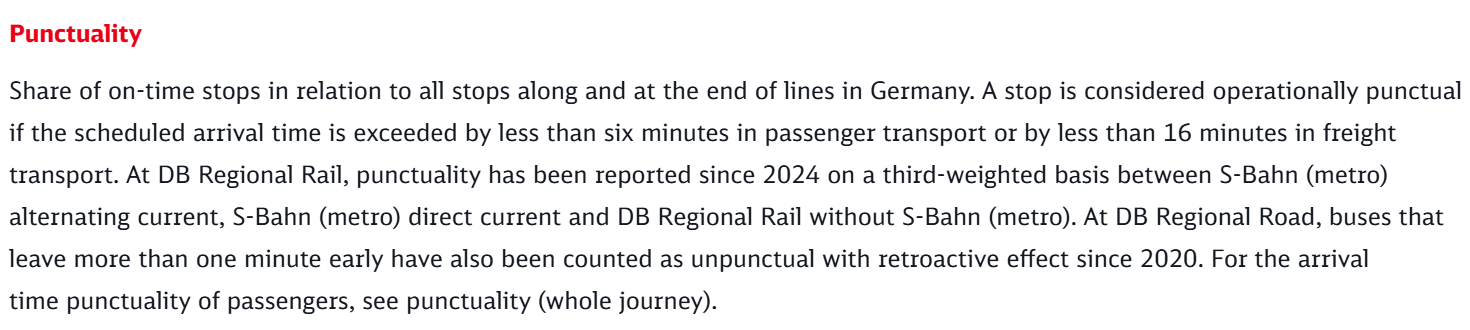
29.10.2025

**1.Introduction**

This study was inspired by a real travel experience. While taking trains operated by Deutsche Bahn (DB) in Germany, we were struck by the company’s highly unpredictable punctuality. Sometimes trains were only ten minutes late—acceptable by most standards—but often delays exceeded an hour, accompanied by last-minute platform changes or even complete cancellations. Out of curiosity, we examined DB’s official punctuality statistics as published on its website and annual report, only to find a surprising discrepancy between the reported data and passenger experience.

*Figure 1*  DB Integrated 2024

According to DB Geschäftsbericht 2024, the overall punctuality rate for passenger rail transport in Germany was reported as **89.5%.** However, a closer look reveals that long-distance trains (DB Long-Distance) achieved only **62.5%**, while regional trains (DB Regional) reached 90.7%. A simple weighted calculation shows that long-distance services account for just 4.3% of total passenger traffic, while regional services make up 95.7%. As a result, the high punctuality rate of regional trains statistically “dilutes” the poor performance of long-distance services, masking the severity of long-distance delays. The gap between these official statistics and actual passenger experience motivated us to re-examine how DB defines and measures punctuality. Our goal is to reassess DB’s reporting methodology and build a more realistic, data-driven delay prediction model that exposes structural biases in the official statistics.



*Figure 2* Grossary of Punctuaclity

1. **Methodological Biases in DB’s Punctuality Statistics**

**2.1. Overly Lenient Definition of “Punctuality”**

According to DB’s official documentation (Pünktlichkeitswerte – Zahlen & Fakten),“Züge gelten als pünktlich, wenn sie mit weniger als 6 Minuten Verspätung ankommen.”(Trains are considered punctual if they arrive with less than six minutes of delay.) In other words, any train arriving within 5 minutes and 59 seconds of schedule is classified as “on time.” This threshold is among the most lenient in Europe: France’s SNCF uses five minutes, and Japan’s JR uses one minute. By adopting the EU’s upper-limit tolerance of six minutes, DB effectively inflates its punctuality statistics.

We argue that this long-standing rule is driven more by institutional convention than by empirical reasoning and fails to reflect actual passenger experience. Therefore, in this study we adopt a **five-minute threshold** for punctuality, aligning with common European standards and providing a more realistic reflection of traveler perception.

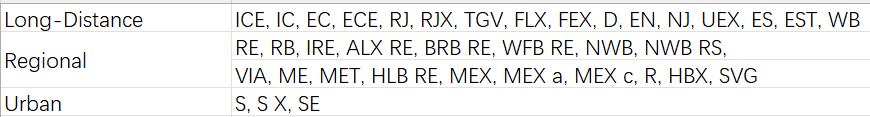
**2.2. Exclusion of Cancelled and Substitute Services**

DB’s official punctuality statistics explicitly exclude Ersatzverkehr (substitute bus services) and Ausfälle (cancelled trains). This means that services cancelled due to construction, weather, or operational failures are completely removed from the dataset, thereby underestimating true delay levels. From a passenger perspective, a cancellation represents the most severe form of delay. In our analysis, **cancelled trains** are therefore treated as cases of extreme delay and included in the dataset.

**2.3. Oversimplified Classification System**

DB divides passenger trains only into “long-distance” (ICE, IC) and “regional” (RE, RB, S-Bahn) categories. This simplification overlooks the operational and scheduling differences across service tiers. Although S-Bahn services account for the largest number of trains, they mainly operate within metropolitan areas and have limited relevance to intercity travel. In contrast, long-distance and regional trains (particularly RE and RB) determine the quality of medium- and long-haul passenger experience. Moreover, the high punctuality of S-Bahn services artificially inflates the average punctuality of the “regional” category.

To better reflect real operational structures, we introduced a three-tier classification: **long-distance, regional, and urban**. This refined taxonomy aligns with functional and scheduling characteristics and allows a more accurate understanding of how delays vary across the rail network.

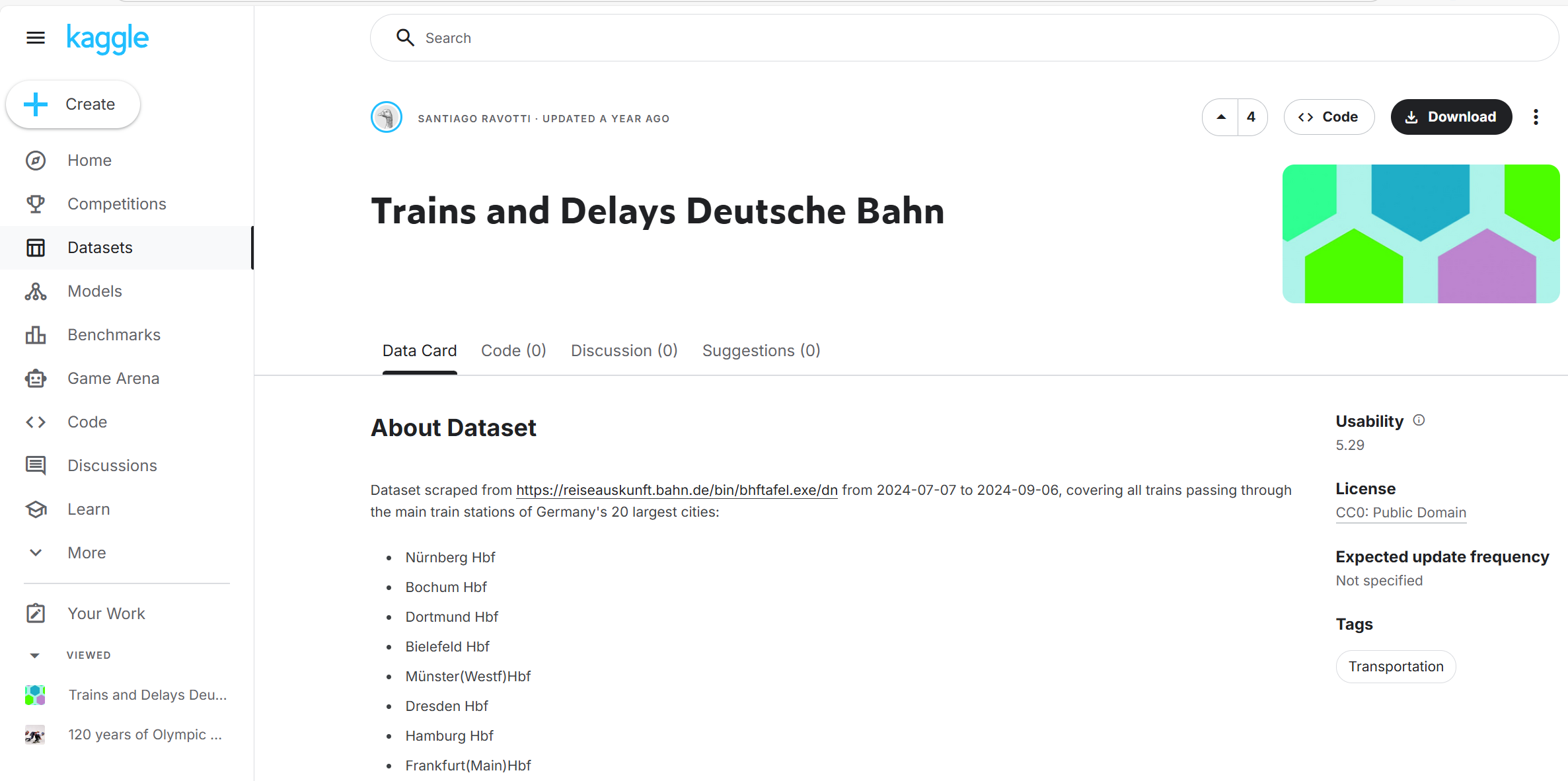


*Figure 3* Three-tier Classification

**3. Data Collection and Processing**

Because DB does not publicly release detailed delay data, obtaining reliable samples was the first major challenge. The only source available is DB’s official real-time departure and arrival board, which refreshes every few seconds and displays only short-term operational data. This means that even if we conducted our own web scraping, the accessible data would be limited to a few days of recent records (e.g., several days in October), making it impossible to construct a long-term dataset.

To overcome this limitation, we therefore used a third-party dataset from Kaggle, compiled by independent developers through automated web scraping of DB’s real-time arrival and departure boards. The dataset includes train records from July 20–25, 2024, and September 1–2, 2024, covering 20 major German cities. Although limited in time span, these samples are representative of typical operational conditions and are suitable for analyzing daily punctuality patterns.

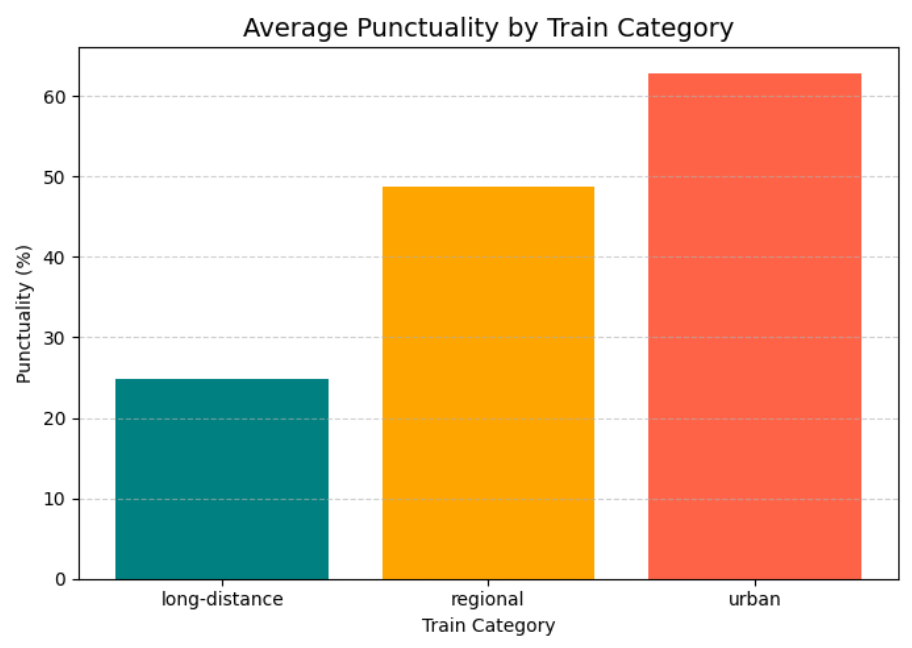
*Figure 4*  Datasets from Santiago Ravotti

The raw dataset originally contained records for more than 30 train categories, including both rail and substitute bus services. To ensure consistency with official railway operations, we removed all bus (Ersatzverkehr) records and reclassified the trains into three operational categories according to their service characteristics:long-distance, regional, and urban.

After extensive cleaning, each record contained the following attributes: **date, departure station, arrival station, train category, departure time bucket, and punctuality status** (0 = delayed, 1 = on-time). We defined a train as delayed if its arrival delay exceeded five minutes, aligning with the revised punctuality standard adopted in this study. In addition, trains that were cancelled for any reason (e.g., technical failure, construction, or substitution) were also treated as delayed and assigned a value of 0 in the binary variable has\_delay (where 1 indicates on-time and 0 indicates delay).

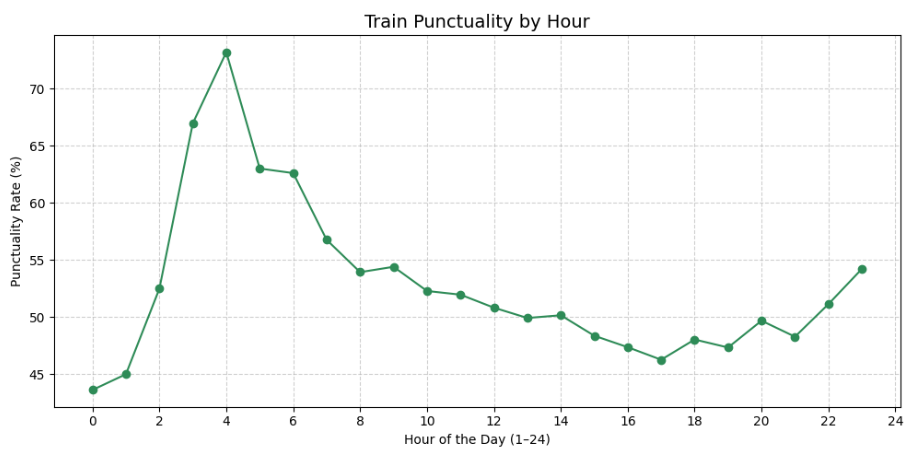
1. **Descriptive Analysis**

The analysis revealed clear structural patterns. Long-distance trains achieved an on-time rate of only 24.84%, meaning roughly one in four arrived as scheduled. Regional trains reached 48.74%, showing moderate stability. Urban trains (S-Bahn) performed best at 62.90%, reflecting the resilience of short-distance commuter operations.



*Figure 5* Punctuality by train category

Across the day, punctuality followed an inverted-U pattern: highest between 03:00–04:00 (67–74%), lowest between 10:00–18:00 (46–50%), and slightly higher again in the late evening. This pattern highlights how traffic density and infrastructure load strongly affect on-time performance, with daytime peak hours suffering the heaviest congestion and operational interference.



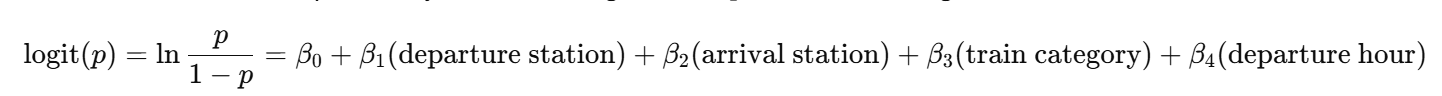
*Figure 6* Punctuality by hour

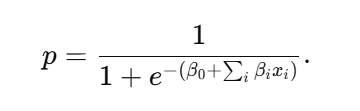
**5. Modeling and Results**

To quantify the influence of various factors on train punctuality, we applied machine learning methods. Because the available data covered only two time windows, we merged July and September samples and performed a random split (80% training, 20% testing) to ensure balanced distributions and robust generalization. Four models of increasing complexity were trained and compared.

**5.1. Logistic Regression (Baseline Model)**

As a linear model, logistic regression offers interpretability and efficiency. Using OneHotEncoder for all categorical variables (departure station, arrival station, train type, and departure time bucket) eliminated false ordinal relationships and allowed each category to be represented independently. The model estimates the probability of a train being on time (*p*) based on the logistic function:



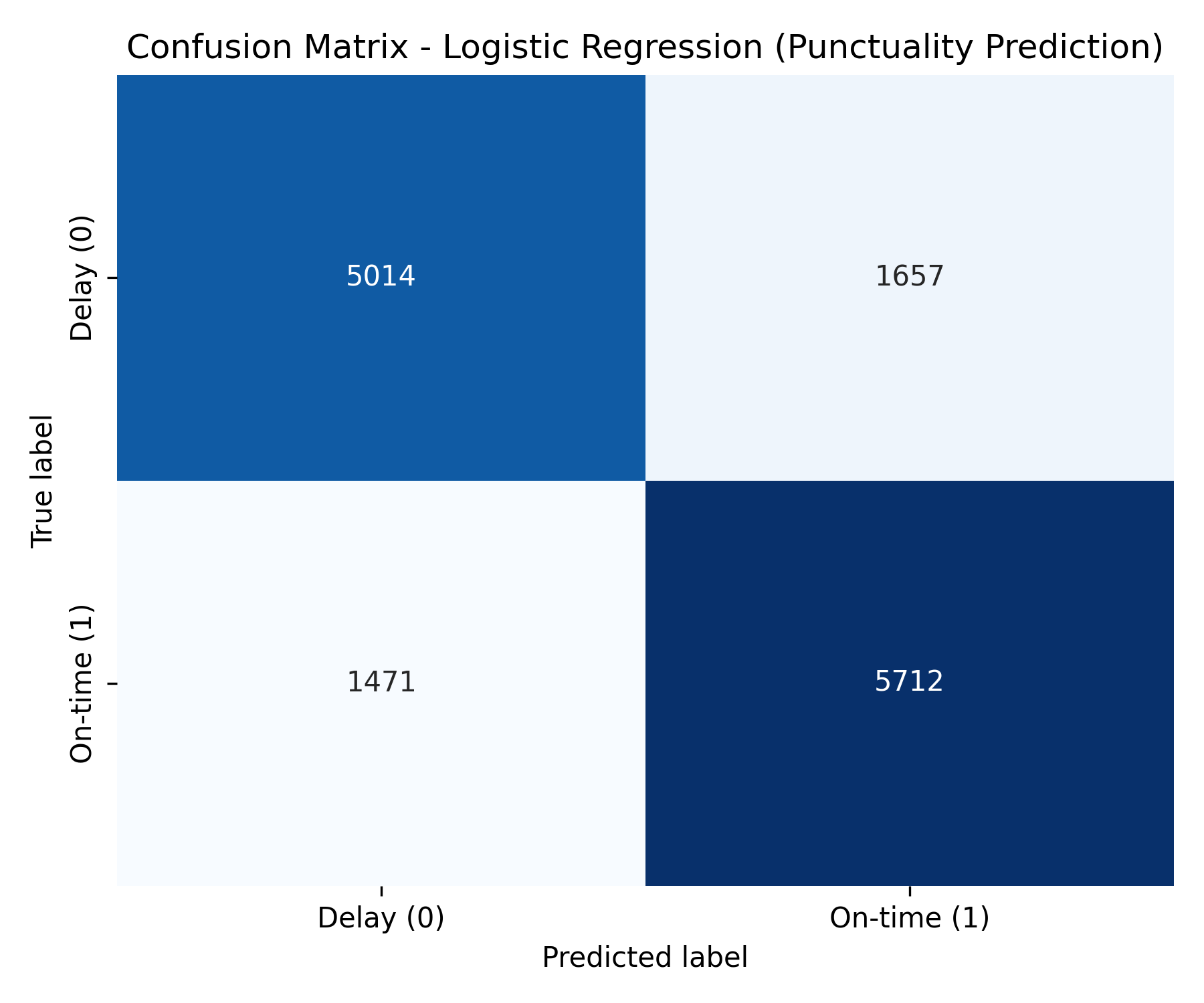
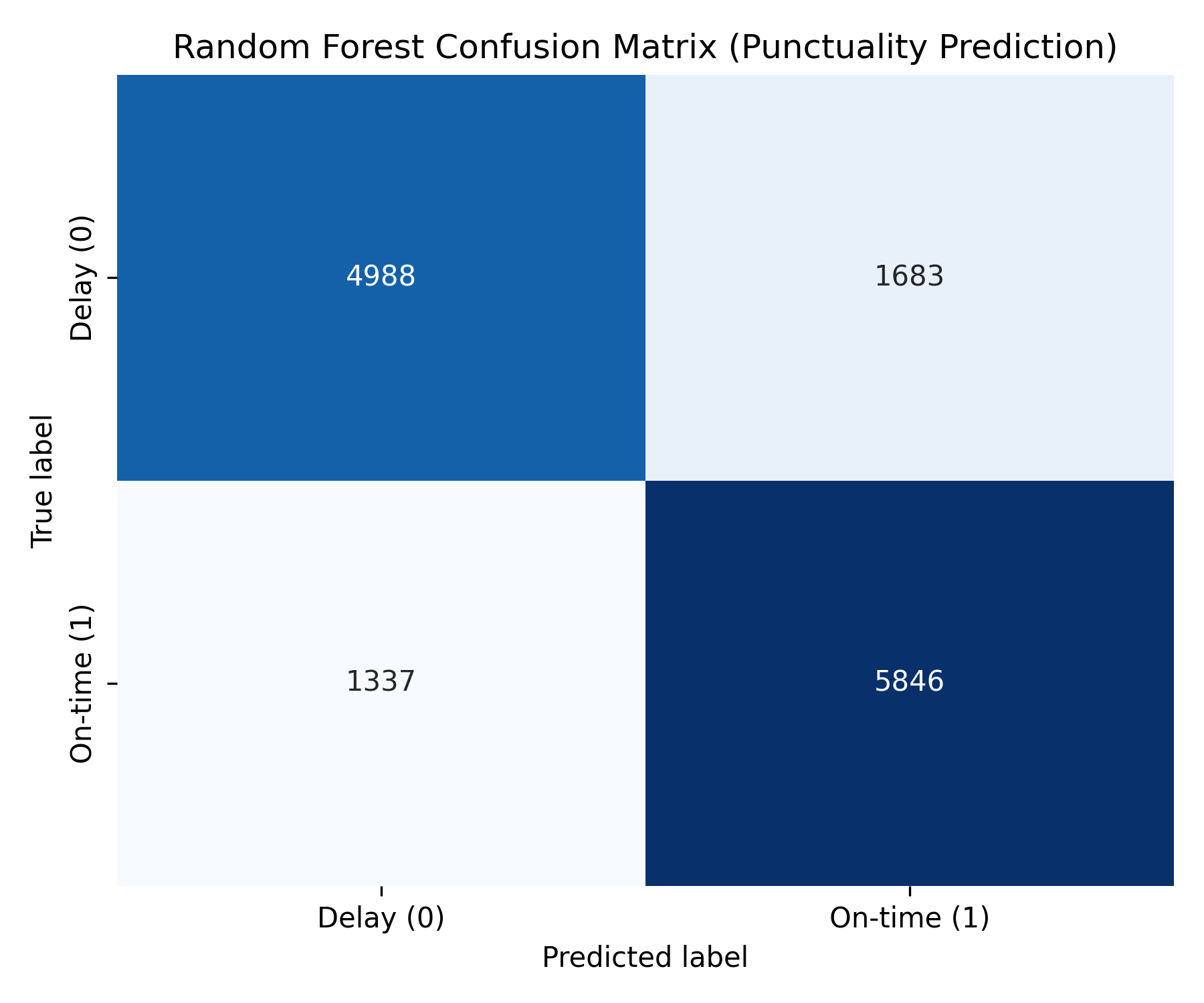
where β0 is the intercept and βi are the coefficients associated with each feature. The probability of punctuality is then given by:

In this formulation, a positive coefficient increases the log-odds of a train being on time, whereas a negative coefficient implies a higher likelihood of delay. This linear model provides a transparent baseline for comparison with more complex nonlinear models such as Random Forest and Gradient Boosting, balancing simplicity, interpretability, and predictive capability.

The model achieved 77.4% accuracy and an F1-score of 0.77, with recall values of 0.80 for on-time and 0.75 for delayed samples. Errors concentrated near borderline cases, such as minor delays close to the threshold. Overall, logistic regression provided a stable, interpretable baseline.

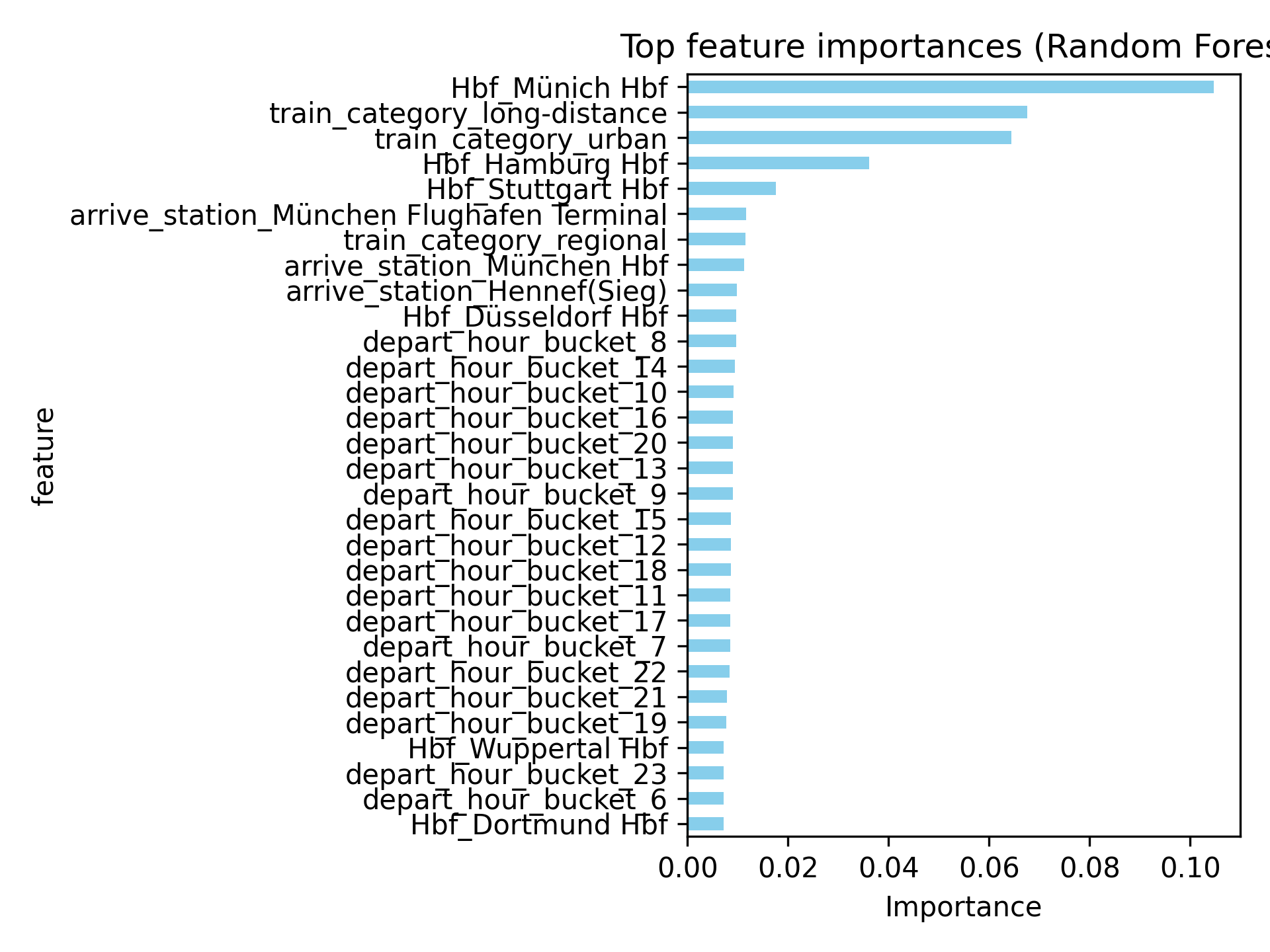
**5.2. Random Forest**

Building on the logistic regression baseline, a Random Forest model was developed to capture the nonlinear relationships that the linear model could not. The model achieved 78.2% accuracy and an F1-score of 0.78, showing a modest improvement over logistic regression while maintaining balanced precision and recall across both delayed and on-time classes.

****

*Figure 7* Comparison of Two Confusion Matrices

Comparing the confusion matrices of Logistic Regression and Random Forest reveals that the latter provides a modest but meaningful improvement in classification performance. The Random Forest correctly identifies more on-time trains (5846 vs. 5712) and reduces false negatives (1337 vs. 1471), resulting in higher recall for the on-time class. This improvement stems from its ability to capture nonlinear interactions between spatial and temporal variables, such as station combinations and departure times. While Logistic Regression remains more conservative and balanced, Random Forest offers stronger generalization and better recognition of borderline cases, reflecting the clustered and systemic nature of DB’s delays.



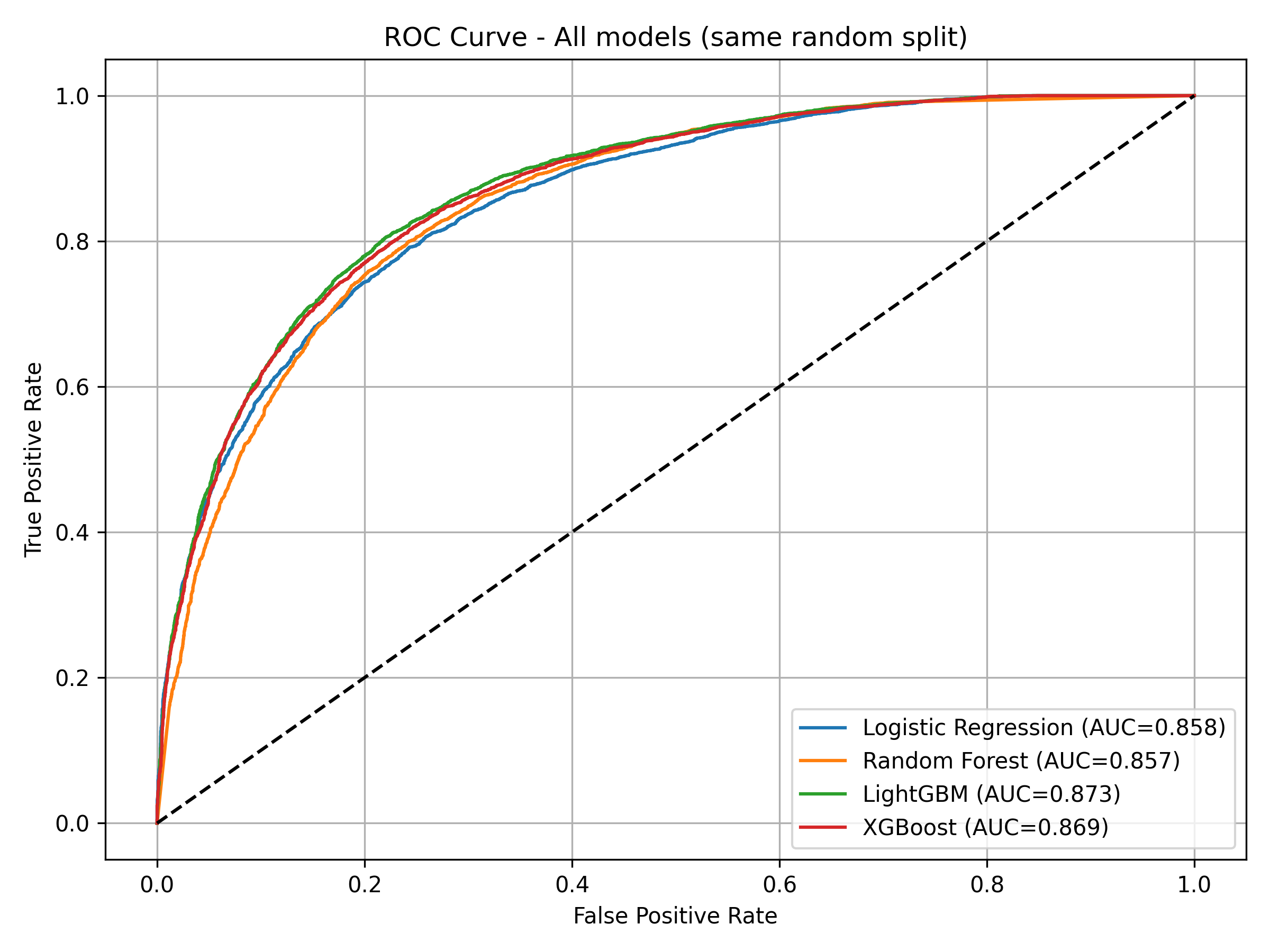
*Figure 8* Top feature importances

Feature importance analysis further revealed that departure and arrival stations were the most influential predictors of punctuality, accounting for about 63% of total importance, followed by departure time bucket (25%) and train category (12%). Major hubs such as Munich, Hamburg, and Stuttgart Hbf dominated the top ranks, confirming the spatial concentration of delays across Germany’s rail network. Temporal factors also played a secondary role, with peak-hour operations showing higher delay risks. These findings suggest that train delays are not random but are strongly associated with spatial and temporal factors, particularly along heavily congested routes such as the Frankfurt–Köln–München corridor, highlighting the clustered and systemic nature of DB’s delay patterns.

**5.3. Gradient Boosting Models: LightGBM and XGBoost**

For further performance gains, we implemented two gradient boosting decision-tree frameworks: LightGBM and XGBoost. Both are well suited to high-dimensional, sparse data and can efficiently capture feature interactions. LightGBM achieved the best results: accuracy 0.7925, F1 0.8019, AUC 0.8727. XGBoost performed similarly: accuracy 0.7864, F1 0.7969, AUC 0.8689.

LightGBM’s advantage stems from its histogram-based split and leaf-wise growth strategy, which enhances efficiency in sparse environments. XGBoost’s level-wise growth produces slightly lower accuracy but greater stability on smaller feature spaces.

****

*Figure 9* 4 Models ROC Curve

**6.Key Findings**

The modeling results demonstrate that DB’s punctuality prediction problem is not linearly separable but influenced by multiple nonlinear interactions. Train type, departure time, and departure station interact strongly—verified by the feature importance of LightGBM. The combination of departure and arrival stations remains the most significant factor (over 60% of total importance), underscoring the tight link between train reliability, geography, and network load.

From a traveler’s perspective, this finding is intuitive. Journeys between major hubs such as Frankfurt, Munich, and Cologne carry a significantly higher risk of delay. These corridors handle the densest traffic and most complex scheduling in the German rail network. In practical terms, travelers should plan for possible delays when moving between major cities—by allowing buffer time for transfers or avoiding peak hours and holidays. For DB itself, these high-risk routes represent critical areas for capacity management and scheduling optimization to improve overall punctuality and passenger satisfaction.

**7. Limitations and Future Work**

Although the four models differ in accuracy, their overall performance is similar (77–79%), indicating that the available features nearly exhaust the dataset’s explanatory power. Since the input variables are mostly categorical (stations, train types, time buckets), their relationships with punctuality are largely linear. Nonlinear models such as LightGBM and XGBoost offer stronger flexibility but limited improvement without additional dynamic features like weather, congestion, or maintenance data.

Nevertheless, complex models remain valuable. LightGBM and XGBoost exhibit better robustness on borderline samples and provide richer feature-importance insights, revealing hidden interactions that can guide future feature engineering. Future research should therefore expand the data scope—integrating weather, construction schedules, holidays, and historical delay patterns—to enable models to capture broader temporal dynamics and achieve higher predictive precision.

**References**

**1.<https://www.kaggle.com/datasets/santiagoravotti/trains-and-delays-deutsche-bahn?resource=download>**

**2.<https://ibir.deutschebahn.com/2024/en/combined-management-report/product-quality-and-digitalization/the-customer-is-at-the-center-of-our-actions/punctuality/?utm_source=chatgpt.com>**

**3.https://ibir.deutschebahn.com/2024/en/glossary/**

**4.**