IRSTI 28.23.15 https://doi.org/10.26577/jpcsit20253205

Aruzhan Amanova * , **Nurbolat Amilbek** , **Diyar Mukhidenov** , **Zhanat Karashbayeva**

Astana IT University, Astana, Kazakhstan
*e-mail: aaruzhan232@gmail.com

SHORT-TERM BUS PASSENGER DEMAND FORECASTING USING MACHINE LEARNING: A CASE STUDY OF ROUTE №50 IN ASTANA

Abstract. This study introduced a machine learning-based approach for short-term forecasting of passenger traffic in the Astana city bus system, in particular, focusing on Route №50. Considering how rapidly cities are developing and increasing transport problems, accurate forecasting of bus demand is an important step towards optimizing resource allocation and improving the quality of service and passenger satisfaction. The research combines data from several sources – information about passenger traffic from a transport company, 15-minute traffic figures and weather conditions, and offers a predictive model that was developed using CatBoostRegressor. The data was collected over one week in December 2024 and covered 9,819 passenger traffic records with a total of 22,111 boardings. According to the results of the study, the model showed high performance with Root Mean Squared Error (RMSE) values of 2.920 and 2.516 for directions from A to B and from B to A, respectively, accurately reflecting the structure of demand at different times and in different places. Also, an analysis of the importance of features showed that factors such as the location of the stop, time of day and traffic congestion are the most significant factors affecting bus demand. The results serve as the foundation for dynamic bus allocation and timetable optimization in difficult urban conditions characterized by harsh winter conditions and traffic congestion. This research addresses critical gaps in the literature by developing a resource-efficient forecasting solution adaptable to evolving urban environments with limited historical data sets. This study offers resource-efficient solutions for forecasting bus demand, adaptable to evolving urban environments with limited historical data sets like Astana, and addresses gaps in the literature.

Key words: Public transport demand forecasting; Bus passenger flow; CatBoostRegressor; Machine learning; Traffic congestion; Weather impact; Feature engineering; Urban transit operations.

1. Introduction

Accurate prediction of bus demand is important for efficient allocation of resources, optimization of schedules and reducing overcrowding [1], especially in the context of dynamic population growth and expansion of urban areas [2]. Astana, the capital of Kazakhstan, is a prime example of a rapidly developing metropolis with specific transport needs. According to the Bureau of National Statistics of the Republic of Kazakhstan, the population of Astana at the beginning of 2025 is more than 1.3 million people [3]. Such rapid demographic growth creates significant pressure on the urban transport infrastructure, requiring new approaches to passenger transportation.

The Astana public transport system currently consists of 116 bus routes operated by a fleet of 1,670 units of rolling stock of various capacities [4]. Despite this extensive network, the system faces a number of significant problems: lack of buses, lack of bus drivers, insufficient integration of various modes of transport into a single system, and lack of flexible response to changes in passenger traffic. These disadvantages create service inequalities that result in high population dissatisfaction levels particularly during peak hours.

The evolution of public transport demand prediction models has evolved day by day and has advanced from traditional statistical approaches to advanced machine learning methods. Early demand forecasting relied on linear time series tools such as AutoRegressive Integrated Moving Average (ARIMA) and exponential smoothing; however, their limitations with nonlinearity and long-term dependencies have led to new solutions. Hybrid statistical-neural models decompose seasonality to learn residual dynamics – for example, Nagaraja et al.'s STL-LSTM (Seasonal-Trend decomposition using Loess combined with Long Short-Term Memory)

reduced Mean Absolute Error (MAE) at toll plazas in Karnataka by 15% compared to ARIMA [5].

Ensemble learning approaches also further enhance the accuracy of forecasting by combining the results of several weak learners. Random Forest [6] and XGBoost (Extreme Gradient Boosting) [7] improved forecasts for the Istanbul Metro by 12% with a grid search setting. Another example paper found that ensemble learners improve accuracy even further: Boateng's review shows up to 25% lower RMSE when aggregating weak models [8], and SHAP's analysis highlighted time of day and transfer activity as key factors [9].

Another type of model is techniques based on spatial-temporal graphs, which have been developed to encode the network architecture in transit systems. To directly encode network structure, spatiotemporal Graph Neural Networks (GNN) treat transit links as graphs: Diffusion Convolutional Recurrent Neural Network (DCRNN) reduced errors by 20% on traffic benchmarks [10], while Graph WaveNet's adaptive adjacency achieved similar gains [11].

Despite these advanced models, applications for specific regions, especially for the Astana bus network, are few in number. Taratynova et al. [12] investigated the forecast of the arrival time of shuttle buses in Astana using time series analysis, but an integrated ML-based passenger traffic forecasting system for the entire urban network has not yet been studied.

Additional optimization threads include behavior and outliers: queue-outlier models in Tsukuba reduced aggregate costs by 30% [13]; Model Predictive Control (MPC) on a Manhattan-like grid reduced vehicle and pedestrian delays by 18% and 25% [14]; Deep-Q GNN dispatching in Cairo reduced waiting times by 10%, but at high data and compute costs [15]. Thus, most state-of-the-art schemes assume dense datasets and powerful hardware, resources that remain scarce in fast-growing cities like Astana.

The goal of this study is to propose a machine-learning-based approach adapted to Astana's Bus №50, integrating multi-source data on ridership, 15 minute traffic congestion scores, and weather conditions to deliver reliable short-term forecasts and actionable insights for more efficient and resilient urban bus operations. This approach allows transit agencies to dynamically allocate buses, optimize driver schedules and adjust frequency in real time by accurately anticipating short-term surges and troughs in demand.

2. Materials and Methods

2.1. Data Collection

This paper presents a machine learning framework for Astana bus #50 that combines passenger flow records with 15-minute congestion estimates and local weather observations to create reliable short-term demand forecasts. The model identifies upcoming peaks and troughs in passenger flow, allowing the operator to reassign vehicles, refine driver lists, and change service frequencies in real time, ultimately improving system efficiency and resilience without compromising service quality.

City Transport Systems LLP (CTS) is responsible for the development and operational administration of the bus network. The company oversees fleet maintenance operations together with route planning activities and digital service implementation for enhanced customer convenience and timeliness. The implementation of dedicated bus lanes together with intelligent scheduling algorithms and real-time information displays at all 1,117 bus stations help minimize delays caused by heavy traffic and harsh winter weather conditions. However, the service capacity of the system does not match the daily demand because 622,000 people board buses each day while peak days reach 755,000. The highest volumes of passengers occur during the morning (07:00–09:00) and evening (17:00–19:00) rush hours. The operational planning becomes more complex because of severe winter temperatures reaching -40°C and frequent traffic jams which result in bus overcrowding and uneven route load distribution.

The study employed three datasets, covering key aspects of the Astana city transport environment: passenger traffic, traffic congestion and weather conditions.

The main dataset is bus passenger flow received from City Transportation Systems of the city of Astana. The dataset provides data from December 7, 2024 to December 13, 2024 inclusive and contains bus stop ID, bus ID, number of entered and exited passengers, timestamp of when the data was sent from the on-board computer, time when data is written to the system, and route number. Additionally, the data was manually verified and annotated with the direction of movement and the order of bus stops. There are two directions: from A to B – from Industrial railway station to Koktal Park, and from B to A – from Koktem residential complex to Industrial railway station. This data was manually extracted and structured to ensure the correct route building.

The second dataset is the traffic congestion in Astana. Information about traffic congestion was manually collected from the Yandex Traffic Jam web interface (Figure 1). For each 15-minute interval, day of the week, time, and congestion score from 0 (lowest) to 10 (highest) were recorded. The statistics are based on traffic information for the last two months.

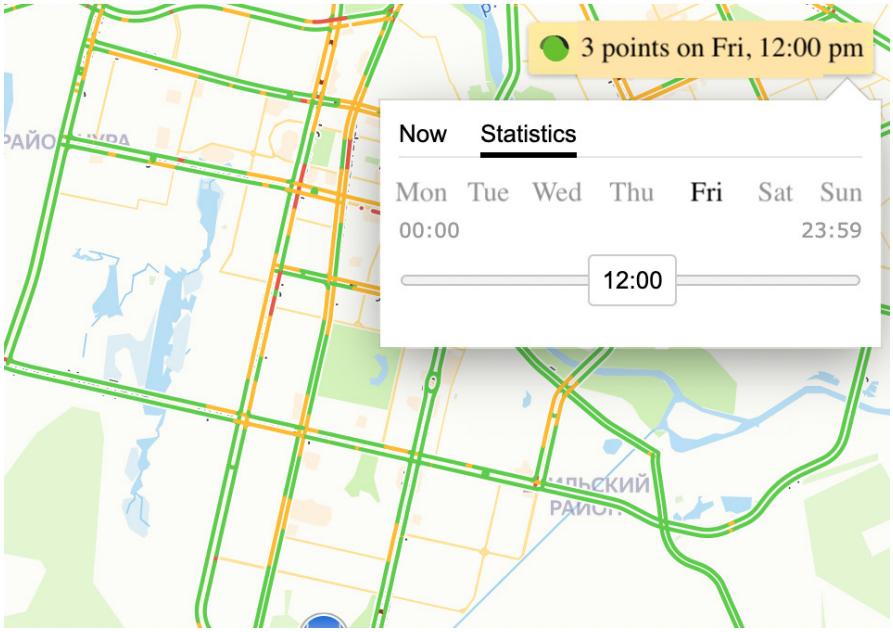


Figure 1 – Yandex Traffic Jam web interface.

The next dataset is weather conditions, received via the Open-Meteo Historical Weather API. The dataset contains weather parameters with an hourly interval from December 7, 2024 to December 13, 2024 inclusive. It includes date and time, temperature, relative humidity, dew point, perceived temperature, the amount of precipitation, precipitation characteristics (snow, rain), snow depth, atmospheric pressure, and conditional weather code (0 – Clear Sky, 99 – Thunderstorm with slight and heavy hail).

2.2. Data Preprocessing

To provide a comprehensive analysis of the city's transport system, data from three sources were synchronized and combined into a single dataframe. Given that the timestamps in each dataset are presented in different forms and with different accuracy, a step-by-step approach to combining was implemented. Firstly, the time values from onboard equipment have been converted to the datetime type. The day of the week and the time rounded up to the nearest 15 minutes were extracted from them, which allowed them to be synchronized with the traffic data. After preprocessing, the traffic data containing the congestion score was aggregated by day of the week and 15-minute intervals. The integration with bus data took place using the day and time keys using the left join in order to preserve all observations of passenger traffic. Following this, the meteorological parameters obtained with hourly discreteness contained timestamps, from which the date and hour were extracted. This made it possible to bring weather data to a format compatible with bus data, where the same parameters were similarly extracted. The final consolidation was carried out by date and hour, which made it possible to supplement each observation of passenger traffic with appropriate weather parameters. At the same time, left-hand pooling was also used to minimize data loss.

The dataset features used for model training are summarized in Table 1, including variable names, data types, descriptions, and units [16].

Table 1 – Data types.

#	Column name	Data type	Description	Unit
1	bus_stop_id	Integer	unique identifier of the bus	–
2	position	Integer	order of the stop on the route (e.g., 1 = first stop, 2 = second, etc.)	–
3	day	Integer	day of the month (0-31)	–
4	day_of_week	Integer	day of the week (0=Monday, 6=Sunday)	–
5	is_weekend	Boolean	indicator if the day is Saturday or Sunday (1 = weekend, 0 = weekday)	–
6	hour	Integer	hour of the day (0-23)	–
7	minute	Integer	minute of the hour (0-59(	–
8	is_peak	Boolean	whether the time is during peak hours (1 = peak, 0 = off-peak)	–
9	score	Integer	traffic congestion score from Yandex (0 = no traffic, 10 = heavy traffic)	–
10	temperature_2m	Float	air temperature measured at 2 meters above ground	°C
11	relative_humidity_2m	Float	relative humidity at 2 meters	%
12	dew_point_2m	Float	dew point temperature at 2 meters	°C
13	precipitation	Float	total precipitation (snow, rain)	mm
14	rain	Float	rainfall intensity	mm
15	snowfall	Float	snowfall amount	cm
16	snow_depth	Float	snow depth on the ground	meters
17	weather_code	Integer	weather condition as a numeric code (0=clear sky, 1=mainly clear sky, 51=light drizzling rain)	World Meteorological Organization (WMO) code
18	enter_sum	Float	number of passengers	–

Descriptive statistics were computed for the final set of features used in model training. Numerical features included weather and environmental indicators, as well as the target variable enter_sum representing the number of passengers boarding per stop-time instance. Table 2 presents summary statistics, including the mean, standard deviation, and range (min/max). Air temperature at 2 meters had a mean of −12.64°C (SD = 3.75°C), with a minimum of −21.18°C and maximum of −5.73°C. Relative humidity averaged 75.99% (SD = 6.25%), and precipitation, snowfall, and snow depth exhibited very low average values, indicating overall dry winter conditions during the observation period. The target variable enter_sum showed a mean of 2.26 passengers per stop event, with a maximum of 33.

Table 2 – Numerical columns summary.

#	Column name	mean	std	min	max
1	temperature_2m	-12.64	3.75	-21.18	-5.73
2	relative_humidity_2m	75.99	6.25	65.38	90.47
3	dew_point_2m	-16.06	3.07	-22.53	-9.53
4	precipitation	0.00	0.01	0.00	0.10
5	rain	0.00	0.00	0.00	0.00
6	snowfall	0.00	0.01	0.00	0.07
7	snow_depth	0.17	0.00	0.17	0.18
8	enter_sum	2.26	3.18	0.00	33.00

Categorical variables included the stop identifier (bus_stop_id), traffic congestion score (score), and weather condition code (weather_code). There were 95 unique bus stop IDs, with stop 1958 being the most frequently recorded (112 occurrences). The traffic score variable had 8 levels, with a mode of 3.0 (observed in 2,851 instances), while weather conditions were represented using 4 codes based on the WMO scheme, with 3.0 being the most frequent (8,846 instances). Sample values for each categorical feature are shown in Table 3.

Table 3 – Categorical columns summary.

#	Column name	Unique Categories Count	Most Frequent Category	Most Frequent Category Count	Sample Categories
1	bus_stop_id	95	1958	112	[1583, 2366, 1963, 6027, 2435]
2	score	8	3	2851	[4.0, 0.0, 3.0, 2.0, 5.0]
3	weather_code	4	3	8846	[3.0, 71.0, 2.0, 1.0]

2.3. Data Analysis

The dataset contains 9,819 rows of bus passenger traffic records and each of them corresponds to a unique bus arrival at a specific stop on Route 50 in Astana during the observation period (December 7–13, 2024). The target variable, enter_sum, represents the number of passengers who boarded the bus during that specific arrival event. The aggregated total of these values across all rows yields the full boarding count of 22,111 passengers, split between 10,621 in direction A to B and 11,590 in direction B to A. An analysis of the network structure showed the presence of 95 unique bus stops. In the direction A to B, the system detected movement through 46 stops, and in the direction B to A – 49 (according to bus_stop_id data, unique by direction). Figure 2 shows the structure of routes divided by directions: red dots indicate stops and a green line indicates the A to B direction, blue dots and a purple line belong to B_to_A. The trajectory of the routes is visualized using the geo_points column.

Figure 3 contains a visualization of the top 10 active bus stops by the number of passengers entering. Among them, the Nurly-zhol Railway Station stop in the A to B direction experienced the greatest load, with a total flow of 939 passengers.

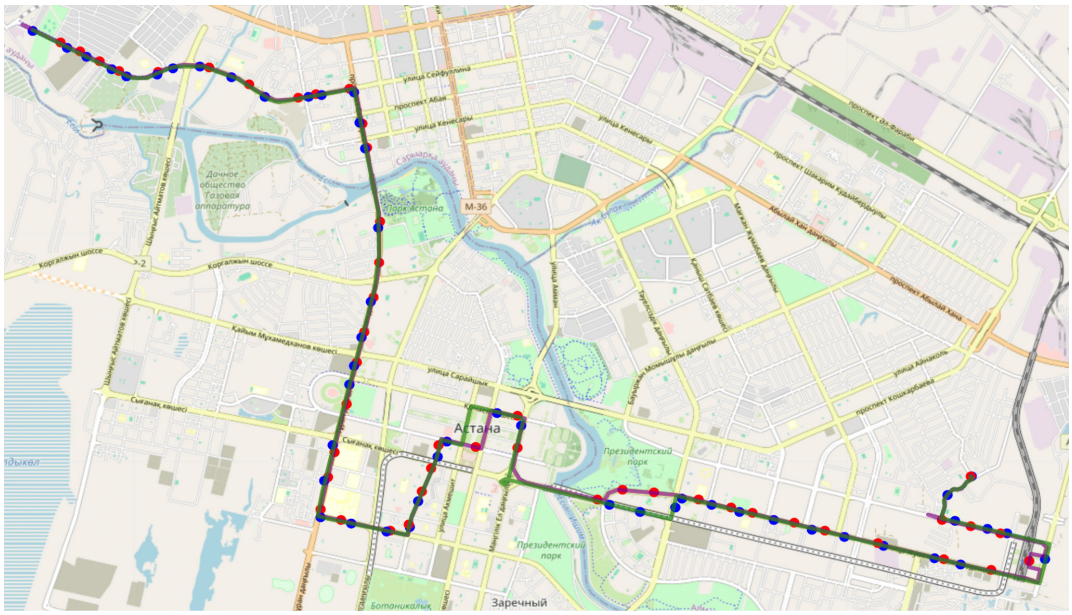
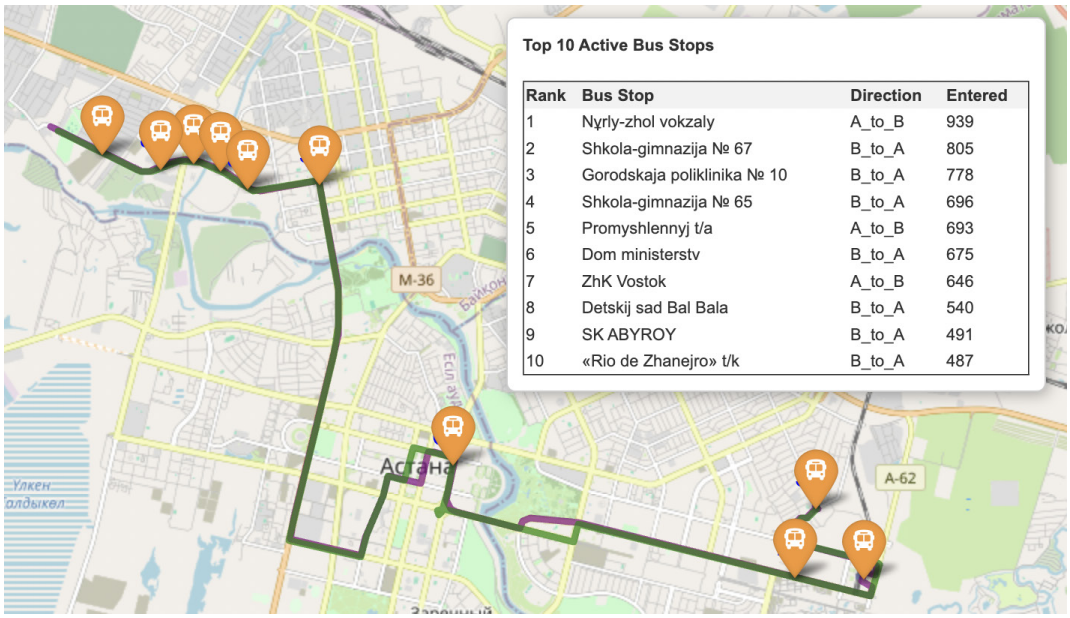


Figure 2 – Route 50 and bus stops.



Rank	Bus Stop	Direction	Entered
1	Nұrly-zhol vokzaly	A_to_B	939
2	Shkola-gimnazija № 67	B_to_A	805
3	Gorodskaja poliklinika № 10	B_to_A	778
4	Shkola-gimnazija № 65	B_to_A	696
5	Promyshlennyj t/a	A_to_B	693
6	Dom ministerstv	B_to_A	675
7	ZhK Vostok	A_to_B	646
8	Detskij sad Bal Bala	B_to_A	540
9	SK ABYROY	B_to_A	491
10	«Rio de Zhanejro» t/k	B_to_A	487

Figure 3 – Top 10 most active bus stops.

Figure 4 illustrates the daily activity of passengers in each direction. Peaks can be clearly traced in the morning and evening hours, especially in the 7-9 and 17-19 hours intervals, which is typical for working days and coincides with the hypothesis of the standard structure of urban mobility.

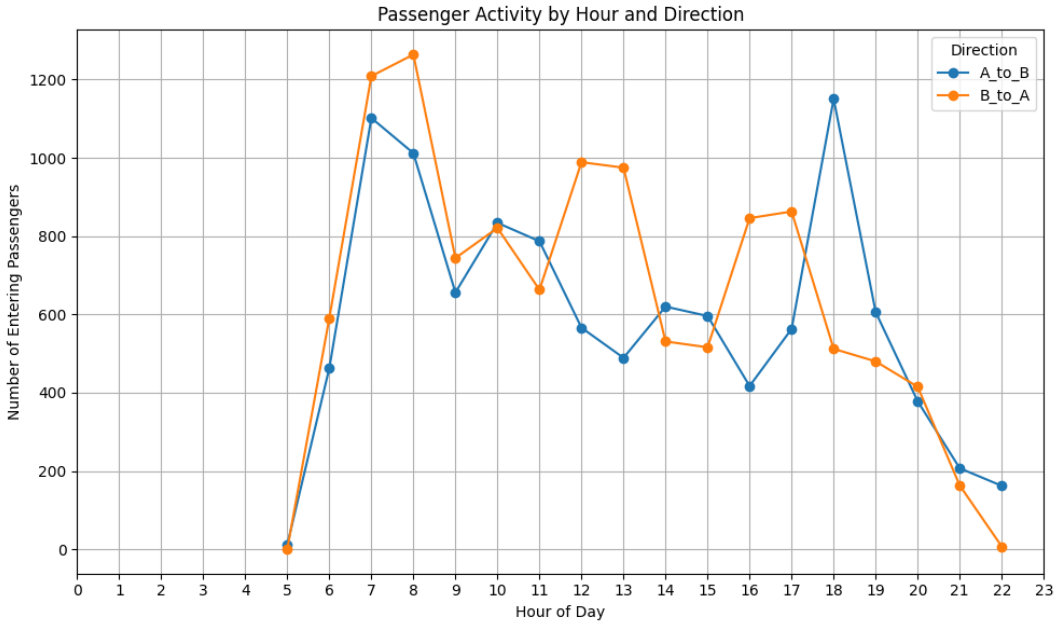


Figure 4 – Passenger Activity by Hour and Direction.

Figure 5 shows a weekly pattern – the highest activity is observed in the middle of the week (Wednesday and Thursday), and the lowest on Monday and Friday, which may indicate a shift in weekends or a change in user habits.

Figure 6 presentes a heat map of traffic jams by time of day and days of the week, where the intervals between 8:00 and 9:00, as well as 17:00 and 19:00 – morning and evening rush hours, typical for weekdays, stand out.

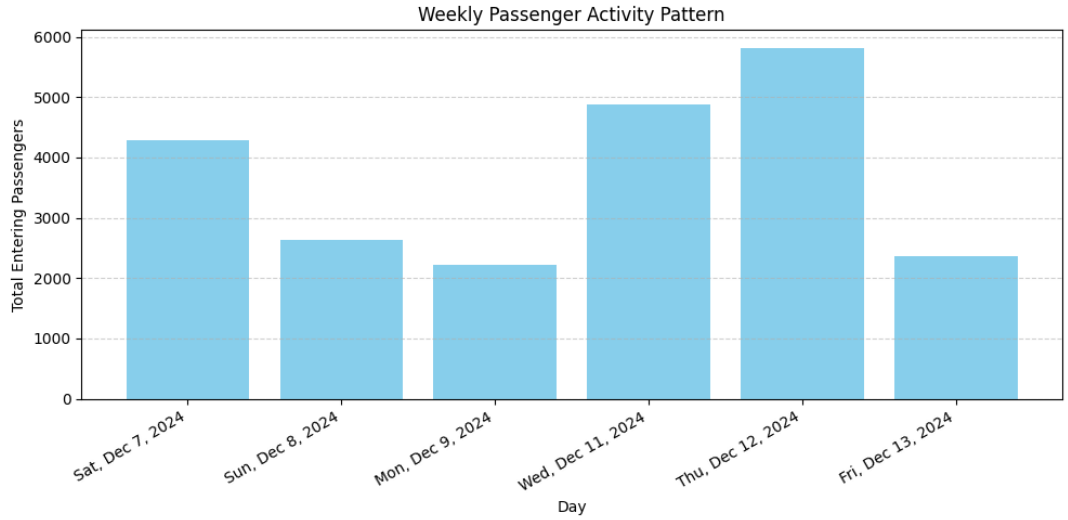


Figure 5 – Weekly Passenger Activity Pattern.

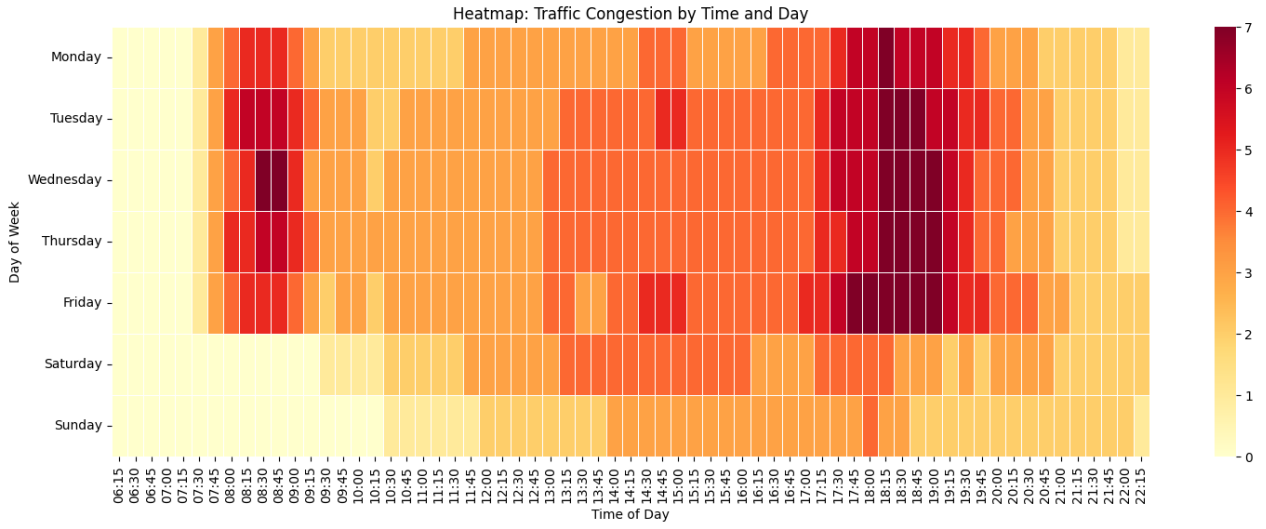


Figure 6 – Traffic Congestion by Time and Day.

Figure 7 shows the correlation matrix between the variables: entered passengers count (enter_sum), traffic (score), temperature (temperature_2m), and hour. The correlation between time of day and traffic was the most significant (r = 0.47), which confirms the load on the network during peak hours.

2.4. Feature Engineering

At the stage of data preparation, feature engineering was carried out to build a passenger traffic forecast model for extracting additional informative characteristics from timestamps and related traffic conditions. The initial data set was

limited to only records for one route direction (A to B or B to A), after which the timestamp (bus_board_computer_sent_time) was converted to datetime format and sorted in ascending time for the correct time sequence. The new features are extracted from the time information. For instance, hour and minute to reflect the intraday dynamics of passenger traffic, day of the month to account for the calendar day's impact on the number of passengers, day of the week for identifying differences between weekdays and weekends, the weekend attribute (is_weekend) is a binary characteristic that indicates whether an event occurs on Saturday or Sunday, and the peak hour attribute (is_peak) is a binary variable indicating that the event occurred at key intervals of heavy passenger traffic: morning peak (8:30-8:45) and evening peak (18:00-18:50) [17].

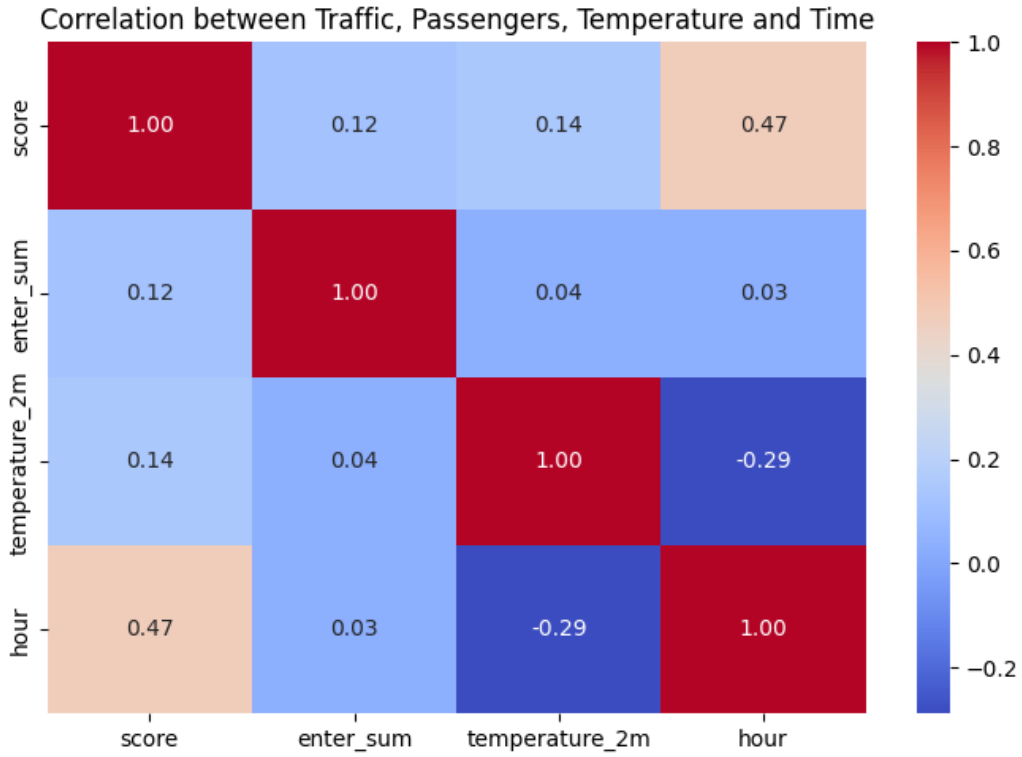


Figure 7 – Correlation between Traffic, Passengers, Temperature and Time.

In addition, categorical variables such as precipitation, rain, snowfall, weather conditions (weather_code), and traffic congestion (score) were converted to a string type and supplemented with the 'missing' value to process missing data without losing information.

The model was trained on features including the stop ID, calendar and time characteristics, weather and traffic situation indicators. The target variable used in this study is enter_sum, which represents the number of passengers boarding at a specific bus stop during a particular time interval. For training and testing, the data was divided into training and test samples with a 20% test share without random shuffling, which preserves the time structure.

2.5. Model Development

Briefly about the tools used in the work: the modeling process was implemented using Python, an open-source programming language widely used in data science and machine learning; Pandas and NumPy libraries were used for data processing and preprocessing, which provide flexible tools for working with structured and numerical data; meanwhile visualizations and diagnostic plots were created using Matplotlib and Seaborn to better understand temporal patterns and model behavior. Additionally, Scikit-learn was used for tasks such as model evaluation, data splitting, and clustering, providing a reliable framework for implementing standard machine learning procedures.

When talking about the choice of concrete model, it is worth referring to the previously made literature review. While the literature review highlights a wide range of forecasting models, from classic ARIMA and exponential smoothing to LSTM, XGBoost, and even spatiotemporal GNNs, the choice of CatBoostRegressor was shaped by the specific characteristics of our dataset and the limitations of the study. Many of the more complex models considered require either large amounts of long-term historical data or dense sensor networks, which were not available in this case. For example, LSTM-based architectures are known to perform well on extended time series, but tend to overfit or fail when trained on limited time windows, such as the one-week dataset used here. Similarly, ARIMA assumes stationarity and cannot model nonlinear demand spikes common during peak hours or with changing weather and traffic conditions. Even ensemble methods like XGBoost, while powerful, require additional tuning and preprocessing for categorical variables, which adds overhead without a clear performance benefit in this context.

In contrast, CatBoostRegressor provided a practical and effective middle ground. It handles high-power categorical features natively – such as bus stop identifiers and weather codes – without extensive feature engineering. It also includes built-in mechanisms for dealing with missing data and preventing leakage of target data. Given these advantages and the need for a model that is both interpretable and robust in the face of limited data, CatBoost emerged as a strong candidate. Besides, the model, which is developed by Yandex, is claimed as gradient boosting and specifically designed to handle categorical features efficiently without large preprocessing [18].

The CatBoostRegressor algorithm was fine-tuned using a set of carefully selected hyperparameters to address the task of forecasting passenger boarding counts at bus stops, conditioned on temporal and meteorological factors. Firstly, the number of boosting iterations was set to 1,000, providing a sufficient number of ensemble rounds to ensure model convergence while controlling for overfitting and reducing prediction variance. Secondly, the learning rate was configured at 0.01, enabling gradual learning and stabilizing the training process by moderating the contribution of each successive tree. The maximum tree depth was specified as 8, offering an effective trade-off between model complexity and generalization capability. This depth facilitates the modeling of non-linear interactions such as rush-hour effects, weekday/weekend variability, and bus-stop-specific demand patterns. Lastly, the model's random seed was fixed at 42 to guarantee reproducibility of experimental results and ensure consistency across multiple training runs.

CatBoost natively supports categorical features and applies an advanced encoding strategy called Ordered Target Statistics, which prevents target leakage by using permutations of the dataset. This encoding is performed internally and dynamically, making it especially advantageous for high-cardinality features such as bus stop ID (bus_stop_id), traffic congestion score, weather code and binary weekend feature (is_weekend).

To perform a structural analysis of the trained CatBoostRegressor model, the model was exported to JSON format, after which a programmatic traversal of all the trees of the ensemble was implemented to extract and register the involved features at the tree level. In CatBoost, each element of the ensemble is an oblivious decision tree, where the same split condition applies at all nodes of the same level. This results in a compact and predictable structure, but it also means that a strictly limited subset of features is involved in a single tree.

Individual trees use only one feature, especially in the early stages of boosting, for example, Tree 0 uses only one feature (day), which indicates its high importance in initializing the model. Most trees use 2-5 features, which indicates a more complex configuration of gradient responses as the model deepens. Some trees use up to 7-8 features, for example, Tree 720 uses features such as is_peak, temperature_2m, day, day_of_week, score, is_weekend, hour (Figure 8); this indicates complex multidimensional interactions between variables. A small number of trees (less than 2%) did not use features at all, which may be a consequence of tree growth stalling due to the lack of informative splits in the current iteration.

The visual distribution of the number of features across the trees showed that the features with indexes temperature_2m, day, day_of_week, score, and is_weekend were used most often, which allows us to conclude that they play a key role in predicting the target variable.

3. Results

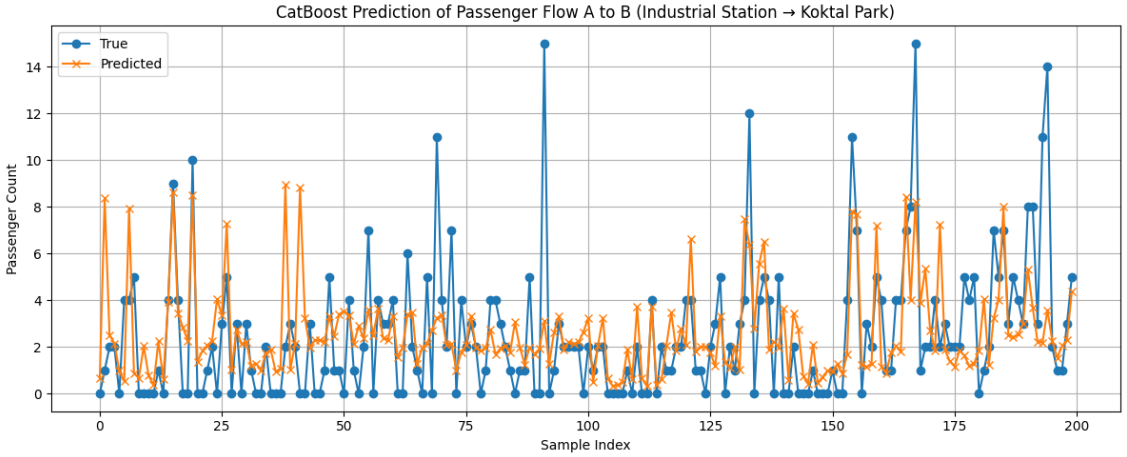
As a result of training the CatBoostRegressor model, metrics for predicting passenger traffic at stops in two directions were obtained: A to B and B to A. The model has demonstrated consistent results

in Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Coefficient of Determination (R^2). As can be seen from Table 4, the model predicts passenger traffic in the direction of B to A with a slightly lower error compared to A to B, which may be due to differences in passenger behavioral patterns and schedule stability in different parts of the route.

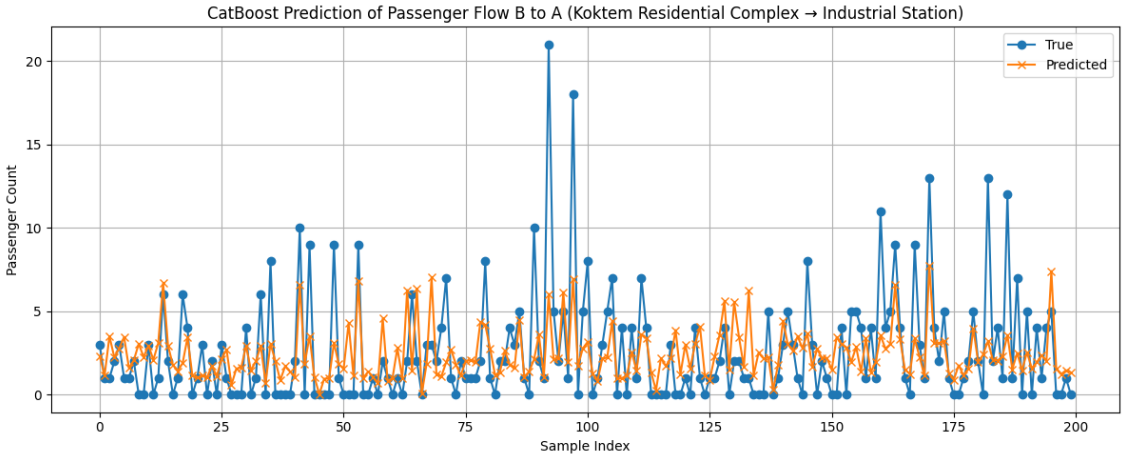
Table 4 – Performance Metrics.

#	Direction	RMSE	MAE	R^2
1	A to B (Industrial Station → Koktal Park)	2.920	1.883	0.227
2	B to A (Koktem Residential Complex → Industrial Station)	2.516	1.819	0.327

Figure 8 shows comparative graphs of the real and predicted number of passengers in the test sample (200 points) for directions A to B and B to A, respectively. The model as a whole accurately reproduces the general trend and peaks of passenger traffic, despite the presence of single emissions.



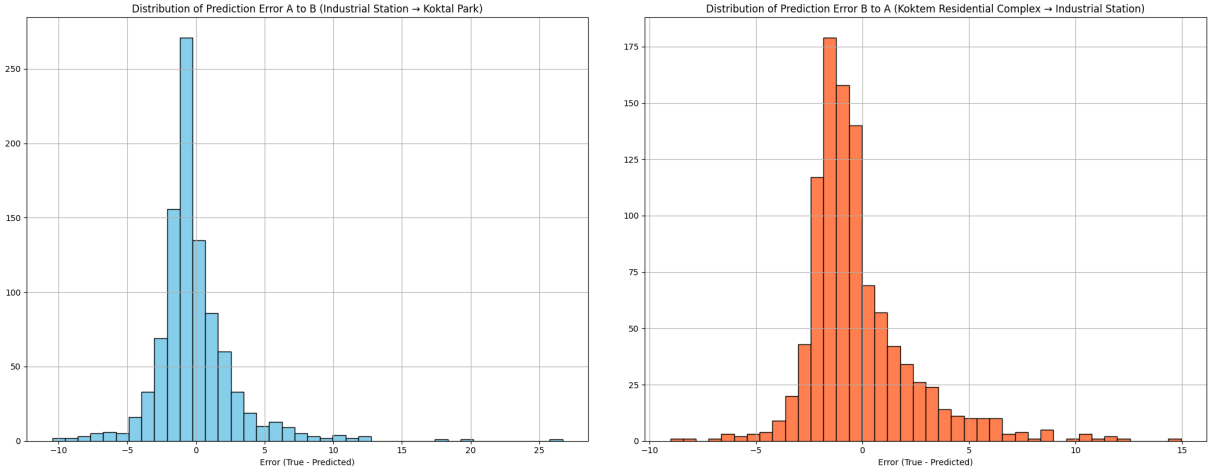
(a)



(b)

Figure 8 – CatBoost Prediction of Passenger Flow: (a) A to B (Industrial Station → Koktal Park); (b) B to A (Koktem Residential Complex → Industrial Station).

Figure 9 presents the distributions of prediction errors. Both histograms have a shape close to normal, with an offset to zero, which confirms the absence of a systematic bias of the model. Errors mostly lie in the range from -5 to +5 passengers.



(a) (b)

Figure 9 – Distribution of Prediction Error: (a) A to B; (b) B to A.

Figure 10 shows a diagram of the importance of the features obtained based on the simulation results, which shows the relative contribution of each feature in predicting the target variable enter_sum (the number of passengers entering the bus). The factors, namely unique stop identifier (bus_stop_id), the position of the stop on the route, hour and minute, and traffic score demonstrated the greatest importance for prediction. This confirms the assumption of a high dependence of passenger traffic on spatial and temporal characteristics, as well as on road conditions.

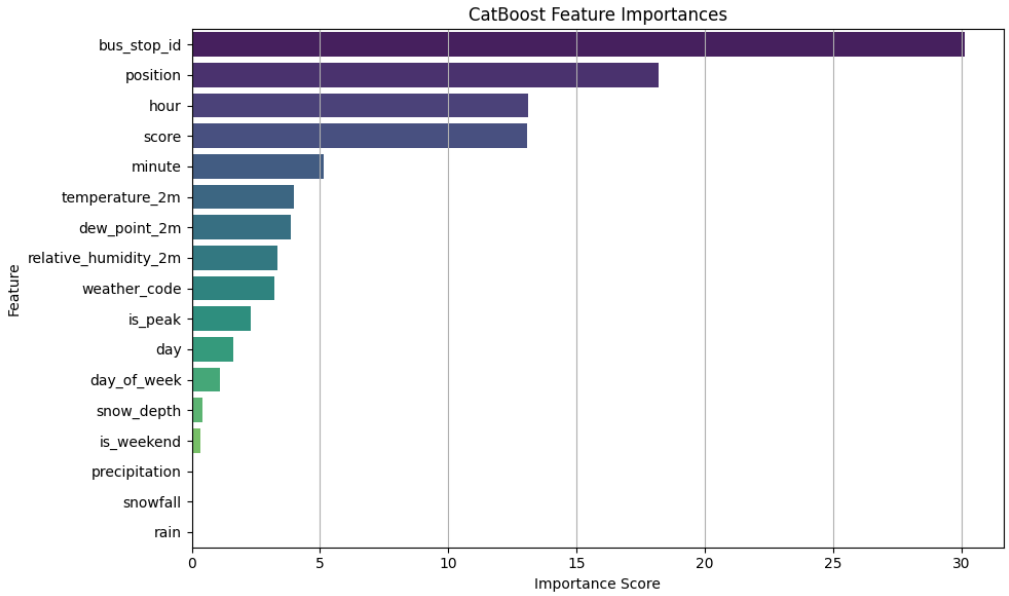


Figure 10 – CatBoost Feature Importances.

4. Discussion

In the conducted study, a passenger traffic forecasting model based on the CatBoost gradient boosting algorithm was developed and tested. The model was trained in the directions of route number 50 in Astana: A to B and B to A, using time, geographical and weather features.

The results showed that CatBoostRegressor is able to effectively capture complex relationships between time of day, stop location, weather conditions, and road congestion. Based on error analysis, it can be noted that the model is particularly good at predicting the background – low and moderate values of passenger traffic, but to a lesser extent – with sharp peaks (for example, during morning or evening rush hours). This is due both to the high dispersion of demand in these intervals and to potential sources of noise in the data (including possible inaccuracies in recording passenger traffic by on-board devices).

After a thorough analysis of two direction of the route №50, it can be seen that the model performs slightly better in the direction from B to A (RMSE = 2.516, MAE = 1.819) than in direction from A to B (RMSE = 2.920, MAE = 1.883). Even though this difference is small, it shows a more regular nature of direction from B to A, which is associated with evening return from work and school/study. In addition, the standard deviation of landings is higher in the direction from A to B that illustrates greater variability and leads to a greater margin of error. According to these statistics, it can be concluded that direction from B to A provides more stable demand data for model training, while the A-to-B direction may require adding additional information such as school calendar, working hours, to improve the accuracy of forecasting.

During peak hours (for instance, 08:00 – 09:00 and 18:00 – 19:00) the average number of boardings per stop on route №50 is approximately 50-80 passengers. By comparing these values with the RMSE values of 2.920 and 2.516, it can be concluded that the average deviation of the model is 3 passengers per forecast, which corresponds to the relative error of 4-5% during high loads. These results demonstrate the practical relevance of the model for short-term planning, even though it works with a limited set of data of 1 week.

For comparison, a simple statistical method – the rolling average was applied to the same dataset. This method calculates the average number of passengers over a fixed time interval and is usually used as a simplified forecasting approach in time series analysis. Although a rolling average may smooth out fluctuations and show general trends, it does not allow to consider contextual variables such time of day, weather or traffic congestion.

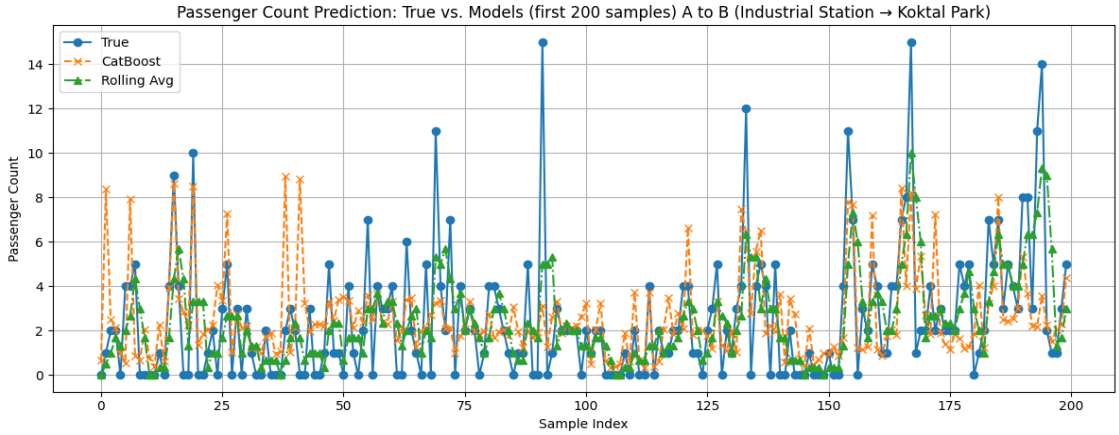
As can been seen on Figure 11, CatBoostRegressor constantly provided forecasts that more accurately corresponded to the actual number of passengers, especially during dynamic periods with sharp peak or sudden changes. The rolling average was usually underestimated during periods of high demand and overestimated during quiet periods.

Figure 12 shows comparison of RMSE values between models. In the direction A to B, CatBoost got an RMSE score of 2.92, while the moving average model showed a score of 2.60. In the direction of B to A, CatBoost showed an RMSE of 2.52, compared to 2.45 for the baseline. These results show that even though the rolling average is competitive in terms of average error, it lacks accuracy in dynamic environments where CatBoost's ability to use contextual features proves its benefits.

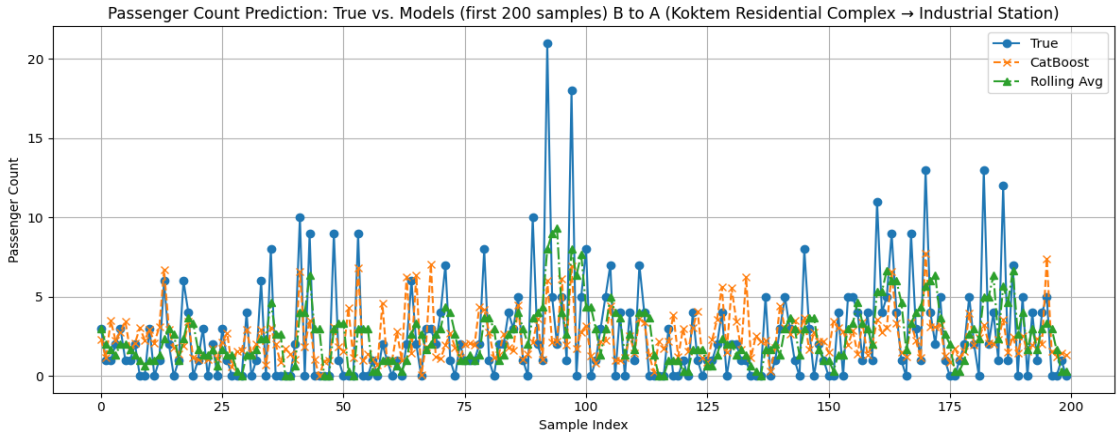
n order to further improve the reliability of the model and to eliminate the limitations related with dividing one route into tests, this study implemented cross-validation of time-series using the TimeSeriesSplit method from the scikit-learn library. The dataset was divided into 5 chronological groups without shuffling, which allowed to keep the temporal order of observations. This method allows for a more realistic assessment of the effectiveness of the model. The model achieved values of 2.476, 2.757, 3.240, 2..835 and 2.438 in five variants, resulting in an average RMSE of 2.749. These results confirm the model's ability to generalize beyond the training set, while maintaining consistency across multiple time segments.

The object importance chart showed the dominant influence of variables such as the bus stop ID, location, hours, minutes, and number of points. This highlights the high dependence of passenger traffic on the spatial location of the stop on the route, the time of day and the traffic situation. The significance of the weather signs turned out to be noticeable.

The feature importance chart showed the dominant influence of variables such as bus stop ID, position, hour, minute, and score. This highlights the high dependence of passenger traffic on the spatial position of the stop on the route, the time of day and the traffic situation. The significance of weather signs turned out to be noticeable.

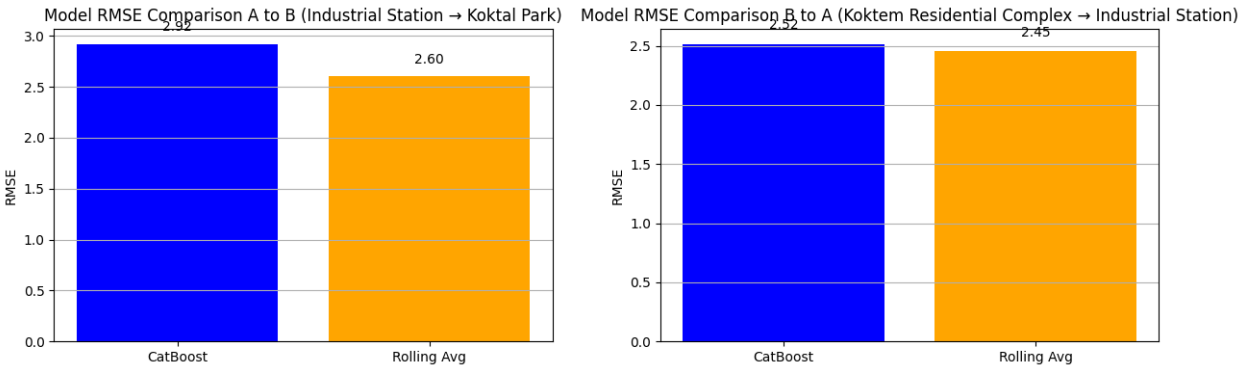


(a)



(b)

Figure 11 – Passenger count prediction: true vs. CatBoost and rolling average for the first 200 samples in both directions: (a) A to B; (b) B to A.



(a) (b)

Figure 12 – Root Mean Squared Error (RMSE) comparison between CatBoost and rolling average models: (a) A to B; (b) B to A.

While the results of this study are encouraging, there are several important limitations to consider. One major limitation is the short time frame of the data – just one week in December – which naturally limits the model's ability to capture longer-term patterns. Factors such as seasonal shifts, school holidays, or special events can significantly impact ridership, but are not captured here. Collecting data over a longer period would provide a more complete picture of demand behavior. Another limitation is that while CatBoostRegressor was chosen for its strengths in handling categorical variables and limited data, the research did not conduct a practical comparison based on data with other models discussed in the literature, such as XGBoost or LSTM. Additionally, while the model does a good job of estimating overall trends, it struggles a bit with sudden spikes in demand during peak hours, which tend to be more unpredictable due to a combination of social and behavioral factors. Capturing these peaks more accurately may require more detailed data or even external signals, such as event schedules or real-time updates. Finally, the model does not yet account for unexpected disruptions – such as road closures or technical failures – that can quickly throw off demand models. Addressing such outliers through anomaly detection or real-time corrections could be a valuable direction for future work.

5. Conclusions

This study presented a machine learning model for predicting bus demand on Route №50 in Astana, Kazakhstan based on real data. Using the CatBoostRegressor algorithm and combining several types of data – spatial, temporal and environmental data, the model was able to achieve a fairly good prediction accuracy of RMSE 2.920 and 2.516 for both directions of the route. In its forecasting, the model took into account not only regular changes in passenger traffic, but also its various variations, thereby showing efficiency in forecasting basic traffic and some limitations in forecasting demand during rush hour.

The feature importance that was done through the analysis demonstrated that location of bus stops, temporal variables such as hour and minute, and traffic congestion are the most influential factors that affect passenger flow, confirming our hypotheses at the beginning of our research. Also, the analysis depending on the specific destination revealed interesting differences in predicting between the A to B and B to A routes, illustrating differences in commuting patterns that deserve further study.

This research addresses significant gaps in literature, providing a practical and computationally efficient solution for predicting passenger demand in a developing city with limited historical data. The model that was presented in this paper can be easily implemented and used by City Transportation Systems LLP for dynamic resource allocation, schedule optimization and improving service quality in the Astana bus network.

While this targeted approach allows one to build and evaluate the methodology in a controlled setting, it is noticeable that relying on a single route, in this case №50, limits the generalizability of the results. For this reason, the work should be considered as a pilot study. Future research areas include expanding the model for adding not just one, but more routes, integrating data transfer between routes, developing real-time forecasting opportunities and extending the time range of analysis to account for seasonal fluctuations. Additionally, in future work, computer vision methods [19] could be used to analyze passenger boarding and disembarkation patterns based on bus CCTV camera recordings as an additional source of data. Using and implementing this model of prediction potentially may significantly enhance the efficiency and resilience of Astana's urban transport system, which will also improve the quality of passenger service and optimize operational resources.

Funding

This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No.BR24992852 "Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens" quality of life improvement").

Author Contributions

Conceptualization, A.A. and Nurbolat Amilbek N.A.; Methodology, A.A.; Software, A.A. and D.M.; Validation, A.A., N.A. and D.M.; Formal Analysis, A.A.; Investigation, A.A. and N.A.; Resources, N.A.; Data Curation, A.A. and D.M.; Writing – Original Draft Preparation, A.A.; Writing – Review & Editing, D.M. and N.A.; Review, Zh.K.; Visualization, D.M.; Supervision, N.A.; Project Administration, A.A.; Funding Acquisition, N.A.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Goetz, A. R. (2019). Transport challenges in rapidly growing cities: is there a magic bullet? *Transport Reviews,* 39(6), 701–705. https://doi.org/10.1080/01441647.2019.1654201.

2. Gheorghe, C., & Soica, A. (2025). Revolutionizing Urban Mobility: A systematic review of AI, IoT, and predictive analytics in adaptive traffic control systems for road networks. *Electronics*, 14(4), 719. https://doi.org/10.3390/electronics14040719.

3. Astana city – Statistics of the regions of the Republic of Kazakhstan – Agency for Strategic planning and reforms of the Republic of Kazakhstan Bureau of National statistics. (n.d.). https://stat.gov.kz/en/region/astana/.

4. Официальный сайт ТОО «City Transportation Systems». (n.d.). Официальный Сайт ТОО «City Transportation Systems». https://cts.gov.kz/ru/company/activities/.

5. Nagaraj, N., Gururaj, H. L., Swathi, B. H., & Hu, Y. (2022). Passenger flow prediction in bus transportation system using deep learning. *Multimedia Tools and Applications,* 81(9), 12519–12542. https://doi.org/10.1007/s11042-022-12306-3.

6. Breiman, L. Random Forests. *Machine Learning 45,* 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

7. Chen, T., & Guestrin, C. (2016). XGBoost. *The 22nd ACM SIGKDD International Conference,* 785–794. https://doi.org/10.1145/2939672.293978.

8. Boateng, A., Adams, C. A., & Akowuah, E. K. (2023). Estimating passenger demand Using Machine Learning Models: A Systematic review. *E3S Web of Conferences,* 418, 03002. https://doi.org/10.1051/e3sconf/202341803002.

9. Aydogmus, H. Y., & Turkan, Y. S. (2022). Application of machine learning methods for passenger demand prediction in transfer stations of Istanbul's public transportation system. *In IGI Global eBooks* (pp. 1086–1106). https://doi.org/10.4018/978-1-6684-6291-1.ch057.

10. Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)* (pp. 3634–3640). https://doi.org/10.48550/arXiv.1709.04875.

11. Jiang, W., & Luo, J. (2022b). Graph neural network for traffic forecasting: A survey. *Expert Systems With Applications,* 207, 117921. https://doi.org/10.1016/j.eswa.2022.117921.

12. Taratynova, D., Kassenova, A., Dauletbayev, B., Al-Shedivat, M., & Pinsky, E. (2025). A Predictive Model of Arrival Times for Smart Shuttle Buses in Astana, Kazakhstan. In Computer Science and Education in Computer Science (pp. 29–49). https://doi.org/10.1007/978-3-031-84312-9_2.

13. Nakamura, A., Ferracina, F., Sakata, N., Noguchi, T., & Ando, H. (2025). Reducing Total Trip Time and Vehicle Emission through Park-and-Ride – methods and case-study. Journal of Cleaner Production, 144860. https://doi.org/10.1016/j.jclepro.2025.144860.

14. Tettamanti, T., Varga, I., & Peni, T. (2010). MPC in Urban Traffic Management. *In Sciyo eBooks.* https://doi.org/10.5772/9922.

15. Darwish, A., Khalil, M., & Badawi, K. (2020). optimising Public Bus Transit Networks Using Deep Reinforcement Learning. *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC),* 1–7. https://doi.org/10.1109/itsc45102.2020.9294710.

16. Weather Forecast Api. Open. (n.d.-a). https://open-meteo.com/en/docs

17. Лайфхаки. Яндекс Go – заказ поездок онлайн. (2018, April 5). https://taxi.yandex.ru/blog/kak-perekhitrit-chas-pik/.

18. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

19. Amirgaliyev Beibut, Miras Mussabek, Tomiris Rakhimzhanova, and Ainur Zhumadillayeva. 2025. "A Review of Machine Learning and Deep Learning Methods for Person Detection, Tracking and Identification, and Face Recognition with Applications" *Sensors 25*, no. 5: 1410. https://doi.org/10.3390/s25051410.

Information about authors

Aruzhan Amanova is a junior scientist and second-year Master's student in Computer Science and Engineering at Astana IT University (Astana, Kazakhstan, 231942@astanait.edu.kz), specializing in data science, machine learning, predictive modeling and data-driven decision-making. Her research focuses on developing and evaluating advanced forecasting models for public transportation demand prediction in urban settings. Aruzhan's work integrates diverse datasets, including passenger flow records, traffic congestion metrics, and meteorological information, to create data-driven solutions for optimizing bus operations in Astana. Driven by a commitment to smart city innovation, she aims to advance intelligent transit systems and contribute to sustainable urban mobility. ORCID iD: 0009-0003-9325-2863.

Nurbolat Amilbek is a junior scientist and graduate student at Astana IT University (Astana, Kazakhstan, 231986@astanait.edu.kz), specializing in data science, machine learning, and artificial intelligence. As a dedicated researcher, Nurbolat is actively involved in various projects that apply advanced technologies to solve real-world challenges. He is currently pursuing his graduate studies, where his academic focus includes intelligent systems, predictive modeling, and data-driven decision-making. Despite being early in his academic career, Nurbolat has already demonstrated a strong commitment to advancing research in smart technologies and their application to urban development. Through his work at Astana IT University, he aspires to contribute to the future of smart cities and innovative solutions for urban infrastructure. ORCID iD: 0009-0004-3973-2820.

Diyar Mukhidenov is a junior scientist and Master's student at Astana IT University (Astana, Kazakhstan, 232011@astanait.edu.kz), specializing in software engineering and computer science. A special focus is the active use of game engines paired with machine learning to expand the capabilities of conventional technologies, as well as conducting research on synthetically developed datasets. ORCID iD: 0009-0002-0275-719X.

Zhanat Karashbayeva is a postdoctoral researcher at Astana IT University (Astana, Kazakhstan, zhanat.karashbaeva@astanait.edu.kz), specializing in mathematical and computer modeling. ORCID iD: 0000-0001-7329-3121

Submission received: 30 April, 2025.
Revised: 26 May, 2025.
Accepted: 26 May, 2025.