Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 03 ISSUE 12, 2025

Machine Learning Forecasts of Spot Truckload Prices Using Operational Carrier Data: A Comparative Study of XGBoost and Benchmark Models

Rabimov Nodir Rahmatilloevich[1], Akharov Akmal Rustamovich[1], Nazarov Fayzullo Makhmadiyarovich[1]
Samarkand State University named after Sharof Rashidov
[1] Department of Artificial Intelligence and Information Systems
Corresponding emails: **rabimovnodirrahmatilloevich@gmail.com, a-rustamovich@samdu.uz,** amaliy_fak@samdu.uz

Abstract:

This article creates and assesses a supervX`ised machine learning model in order to forecast spot truckload rates at the shipment level, this paper. We create a feature set with lane information, shipment date, distance, and cargo weight using operational data from carriers for 2022–2024. The target is defined as freight cost in USD per load and rate per mile. A global mean model, a lane mean model, multiple linear regression, random forest, and an XGBoost ensemble are the five methods we benchmark after formulating the problem as a regression task.

1. Introduction

Logistics is one of the fundamental principles of the modern economy. This movement provides goods and services, fueling global growth. Without an effective logistics system, no business can enter the international market. Currently, the share of freight transport officially registered using road transport varies from 13% to 90% depending on the country and geographic location. The United States plays a key role in the global regional sphere. According to preliminary data, in 2023, trucks

in the United States transported 11.18 billion tons of freight, demonstrating the share of road transport in the country's transportation system. Total industry revenue reached $987 billion, and the logistics sector employed 8.5 million people, including 3.55 million professional drivers. Furthermore, the United States is actively involved in cross-border trade: trucks accounted for 66.5% of overland transport between the United States and Canada and 84.5% of transport across the border with Mexico. These indicators confirm that the United States remains a leader in global trade and transportation. In this article, we use the US logistics industry and its data as a prototype for training our models, as Uzbekistan's logistics industry is not as developed as the US, and data is very limited. However, our models will be universal and applicable to all countries.

Forecasting freight shipping costs is a complex process with many variables. Shipping costs can be affected by many factors, including weather, cargo weight, cargo type, cargo value, final delivery address, the country's economic situation, and much more. The US logistics system is quite complex and consists of several players. The three main players in the process of shipping cargo from point A to point B are:

- Carrier
- Broker
- Costumer (Shipper)

A carrier is a company that provides cargo transportation services and has the vehicles to do so. A costumer is a company that has cargo and needs to transport it from one point to another. A broker is an intermediary that helps the costumer find a carrier that will deliver the cargo at the most cost-effective price. Carriers perform the bulk of the work here, as they physically transport the goods.

This industry is highly competitive. According to the U.S. Department of Transportation and the Federal Motor Carrier Safety Administration, there are more than 1.86 million carrier companies providing transportation services operating in the United States by 2024. Interestingly, 96% of companies have 10 or fewer trucks.

2. Data

For the study, data from several large and smaller carriers was collected to train our models. The data period selected runs from 2022 to 2024, as this period best represents the current state of the logistics market. Data was collected from four companies:

1. Halol Transport LLC,
2. PTI Cargo,
3. Zaki Transportation
4. Ansor Express LLC.

In total, our dataset contains over 15,000 rows of data. Each row represents a single completed shipment and contains the following elements:

Description	Example
Order_ID	5901
Load_ID	732400
Date	1.2.2022
Origin	GA 30043
Destination	NJ 08831

Origin coordinates	33.965958	-84.085840
Destination coordinates	40.352432	-74.477337
Distance	810	
Load type	53 Dry Van	
Load weight (lb)	4500	
Fuel price ($)	3.7	
Holiday indication	0	
Weather condition	Sunny	
Partial/Full load	Full	
Total Cost ($)	2800	

Table 1. Illustration of sample data

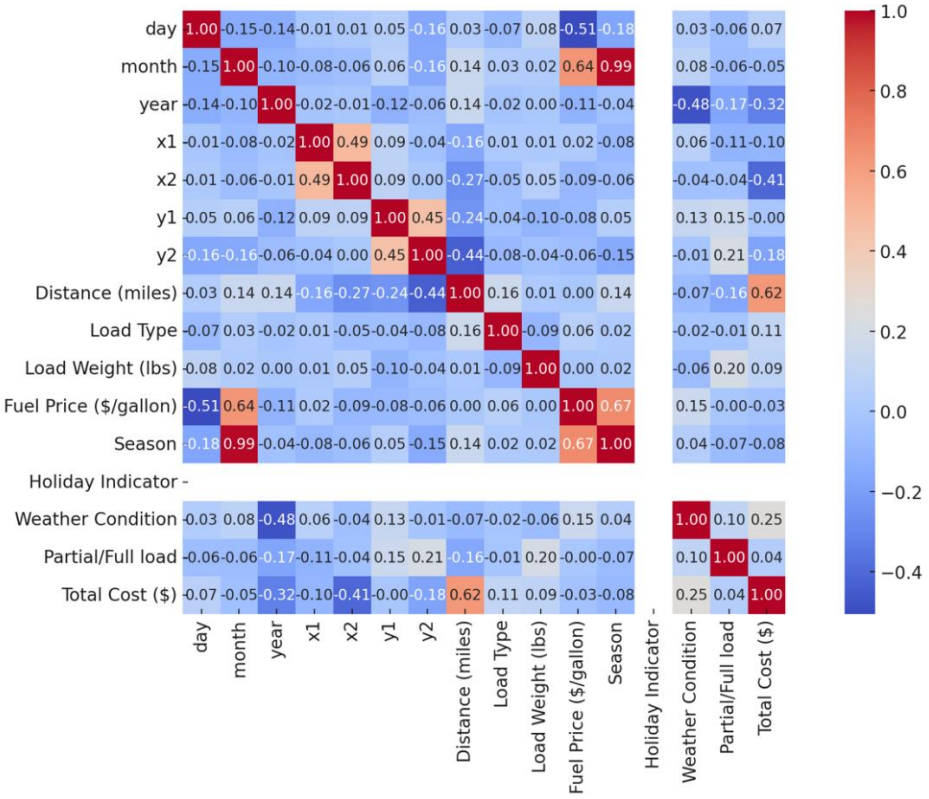
Additionally, we'll add an additional variable, rate per mile, to each row. It's calculated by dividing the final cost of transportation by the distance traveled.

$$rate\_per\_mile = \frac{Total\ Cost\ (\$)}{Distance\ (miles)}. \quad (1)$$

3. Date analysis

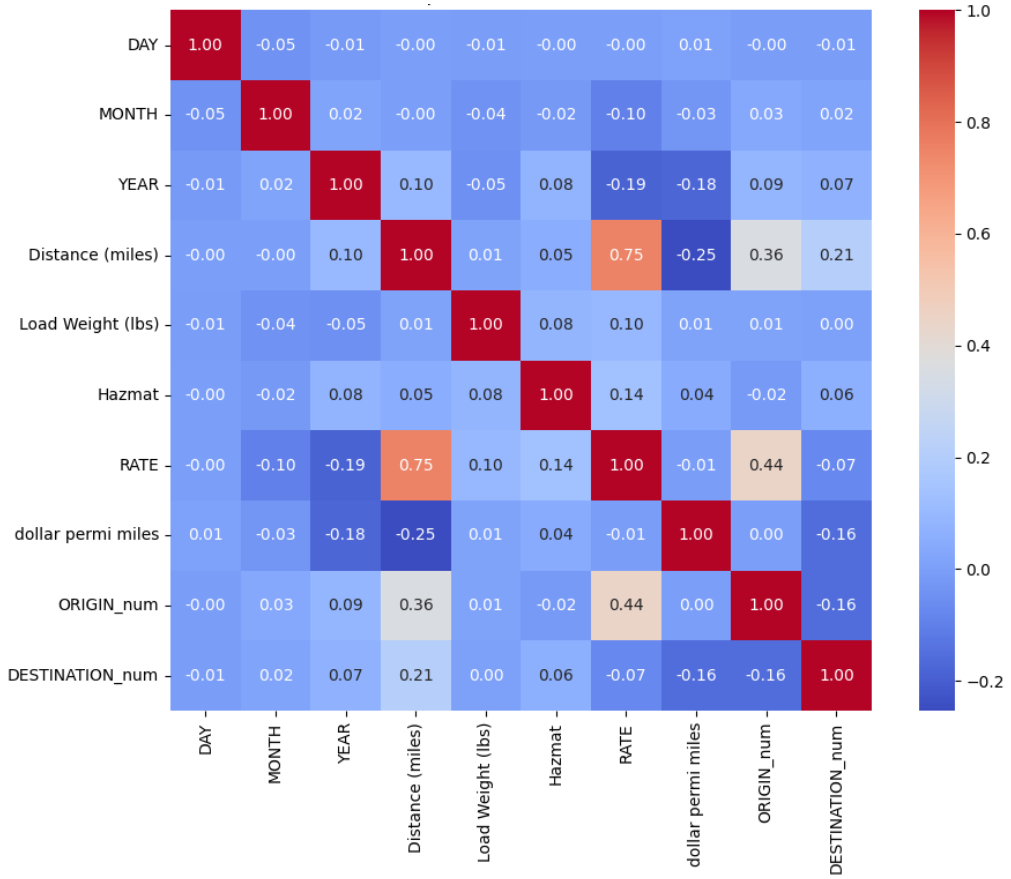
During the initial analysis, rows with outlier values that didn't represent the overall state of the logistics market were removed. Some shipments contained additional costs, such as loading and unloading expenses, which were not included in the unloading price. Additional compensation for late release of vehicles, etc., was also removed. Ultimately, we obtained 12,000 rows of useful data, which were ultimately used to train our models.

Having set up the first test data set and created a correlation table, we obtained the following table



Picture 1. Correlation matrix of the load features

Of the initial parameters, quite a few are unrelated to our target (Total Cost). The weakest linear relationships are for coordinates, fuel prices, weather conditions, and cargo type. After modifications, we retained the following parameters for our dataset and obtained the following table:



Picture 2.Modified correlation matrix of the load features

As we can see, the greatest correlation here is with distance and, as you can see, with the location from which we load the cargo. Indeed, the location of cargo acceptance has a strong influence, as cargo from the central states is much more expensive than cargo from the eastern states. Although our table doesn't show it, the location of cargo delivery also has a significant impact on the final shipping cost.

After analyzing the data, it was determined that the most common route is from the East (New Jersey) to the central states (Kentucky, Ohio, and Indiana). This statistic reflects the fact that New Jersey has many ports where imported goods are delivered and then distributed throughout the central states, where shipping by sea is not possible.

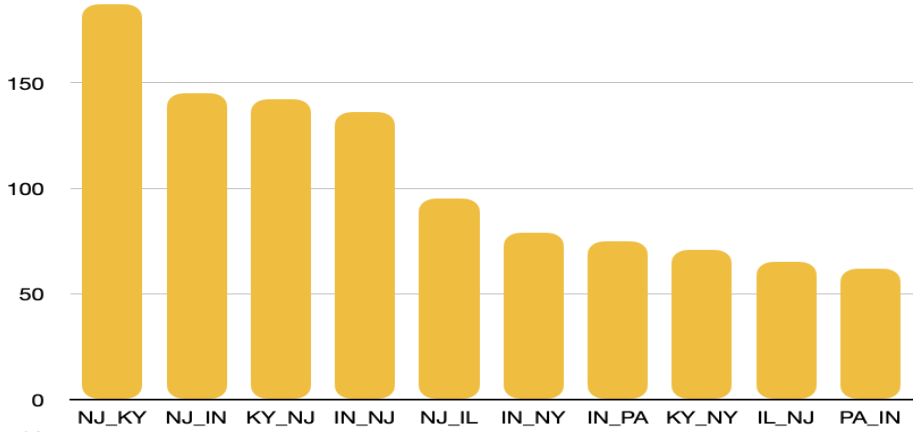
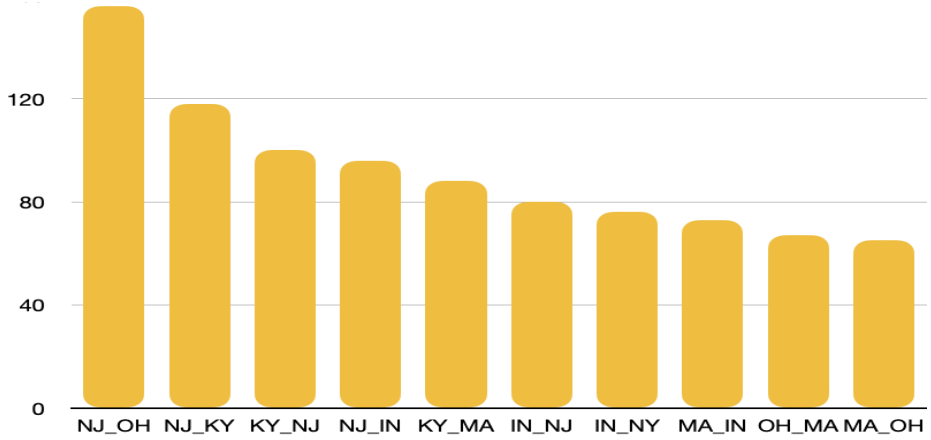
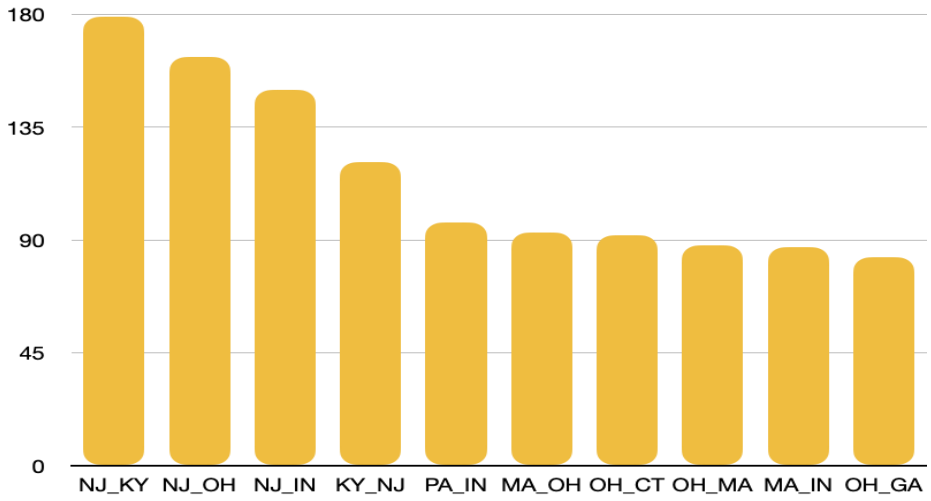
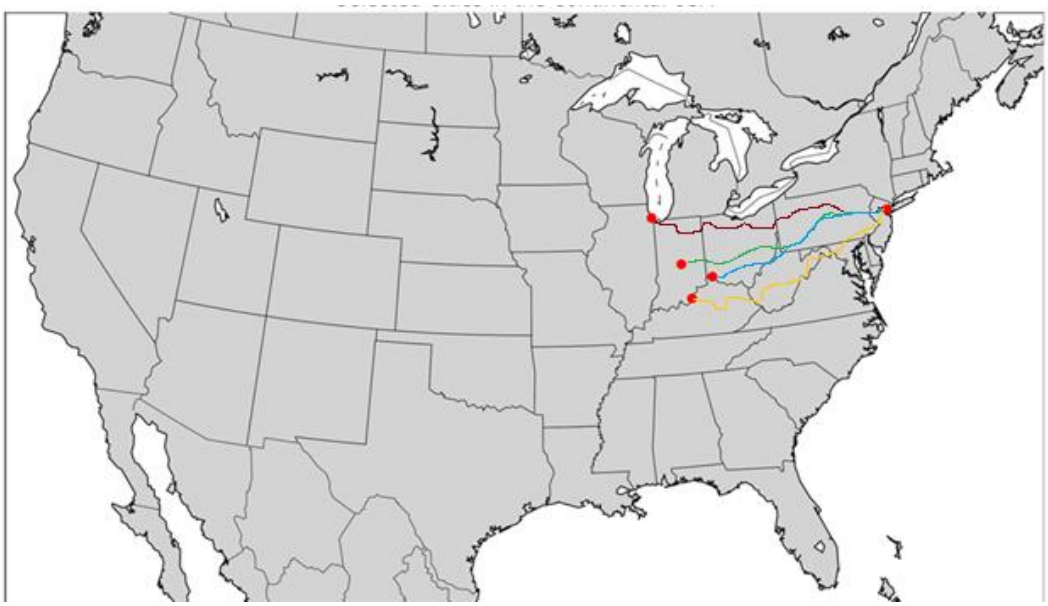
Year	Frequency
2022	
2023	
2024	

Table 2. Top 10 states that has the most loads in 2022-2024

Three years of data show that New Jersey has the highest number of freight transportation options. These freights are delivered to major cities in the central states, typically the county seats.

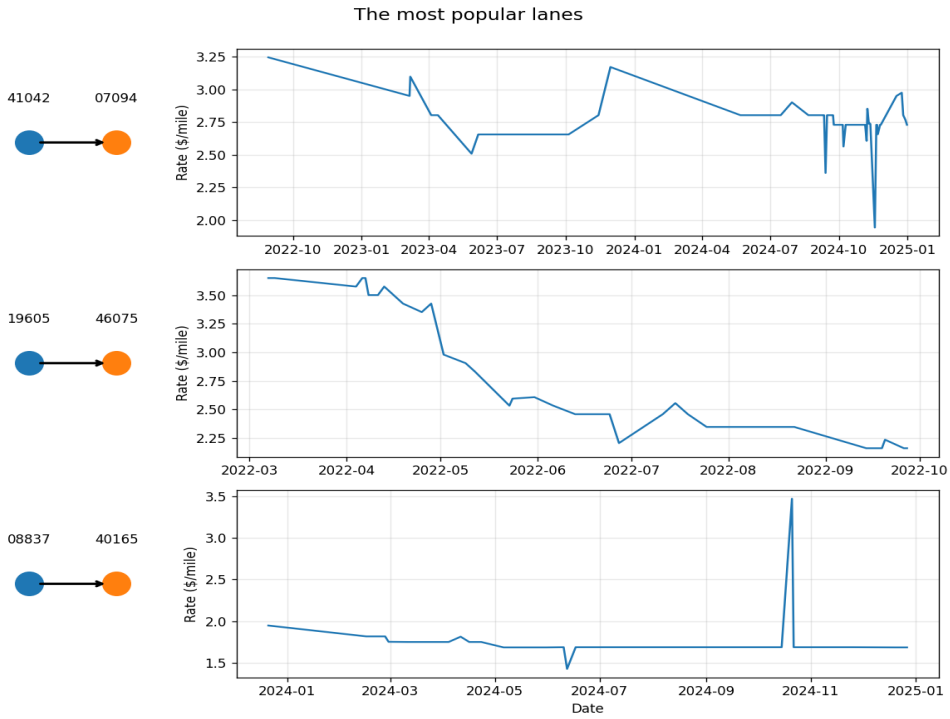


Picture 3. The most popular routes

As for prices, a gradual decline can be seen from 2022 to 2024.

- 2022 avg *~3.17 $/mile*
- 2023 avg *~2.39 $/mile*
- 2024 avg *~2.21 $/mile*

Average rates decreased between 2022 and 2024, consistent with the post-pandemic downturn in the freight market. In 2022, post-pandemic, prices were higher across all states. A sharp decline is seen in prices between 2022 and 2023. Prices will stabilize in 2023 and 2024, although on average they are lower than in 2023. The graph does not show the sharp declines seen between 2022 and 2023. Analysis of the most popular routes shows that freight from New Jersey is typically profitable or even unprofitable. Freight is typically sourced from there to deliver to states where profitable prices can be achieved. For example, average freight rates from New Jersey to Kyrenia are approximately $1.60-$1.90 per mile, while from Kyrenia to New Jersey they are $2.80-$3.00 per mile. So if you make a full circle from NJ to KY and back, you'll get an average of $2.35 per mile.



Picture 4. Rate per mile in the most popular lanes.

4. Methodology

4.1.Problem formulation

The data set we collected, consisting of cargo completed during 2022-2024, is indexed as $i = 1, \dots, N$. Where for each i is the actual price per mile $y_i \in \mathbb{R}$; and also is a vector $x_i \in \mathbb{R}^p$ which describes the operational characteristics of the transportation.

The goal is to teach the function:

$$f: \mathbb{R}^p \to \mathbb{R}, \qquad \hat{y}_i = f(x_i) \tag{2}$$

where $\hat{y}_i$ is a good approximation of the price per mile for future shipments. Our learning problem can be formulated as a supervised regression problem over a sample

$$\mathcal{D} = \{(x_i | y_i)\}_{i=1}^{N} \tag{3}.$$

This model is designed in such a way that the conditions for forecasting prices for future cargo depend on:

- The Lane, i.e., starting address and ending address (by ZIP codes)
- Date (day, month, and year)
- Weight of cargo
- As well as the distance from the starting point to the end point.

Thus, by entering these 4 characteristics, we should get an approximate forecast of the final cost of transportation.

4.2.Baseline regression model

We start with a multiple linear regression of the form

$$y_i = \alpha + x_i\beta + \varepsilon_i, \tag{4}$$

and in vector notation

$$y = \alpha 1_N + X\beta + \varepsilon, \tag{5}$$

where $X \in \mathbb{R}^{Nxp}$ is the planning matrix and ε is the error vector.

In the LAG-WMR model [1], inter-route and time correlations are modeled with a lag weight matrix W, resulting in a lag specification

$$y = \alpha 1_N + \rho W_y + X\beta + \varepsilon. : contentReference\ [oaicite:6]index\ = 6 \qquad (6)$$

Instead of parametric estimation (ρ,W,β), we use a nonlinear ensemble model (XGBoost) to approximate the unknown feature mapping.

4.3.XGBoost regression model

The XGBoost model is based on a decision tree and performs better than other methods such as random trees and gradient boosting. This model is well suited for a complex and large dataset like ours because it uses various optimization methods. It models the forecast as an ensemble of regression trees:

$$\widehat{y_i^{(K)}} = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in \mathrm{F} \qquad (7)$$

where for each f_k there is a regression tree

$$f_k(x) = \omega_{qk}(x) \qquad (8)$$

where *qk* routes the input x to the index of nodes $j \in \{1, \ldots, T_k\}$; ω_i is the node value; и T_k is the number of nodes in tree *k*.

In the objective function, at boosting iteration t the model adds a new tree f_t that minimizes the regularized objective

$$\mathcal{L}^t = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \qquad (9)$$

and has a loss

$$\ell(y, \hat{y}) = (y - \hat{y})^2, \qquad (10)$$

with the structure

$$\Omega(f) = \gamma T + \frac{\lambda}{2}\sum_{j=1}^{T} w_j^2. \qquad (11)$$

T is the number of nodes in γ complex trees, and λ is L_2 penalty on the weight of nodes.

Using the Taylor expansion $\hat{y}_i^{t-1}$,

$$\mathcal{L}^t \approx \sum_{i=1}^{N}[g_i, f_t(x_i) + \frac{1}{2}h_i, f_t(x_i)^2] + \Omega(f_t) \qquad (12)$$

where

$$g_i = \frac{\partial \ell(y, \hat{y})}{\partial \hat{y}}\Big|_{\hat{y}=\hat{y}_i^{t-1}}, \qquad h_i = \frac{\partial^2 \ell(y, \hat{y})}{\partial \hat{y}^2}\Big|_{\hat{y}=\hat{y}_i^{t-1}}$$

And for the quadratic error loss $g_i = -2(y_i - \hat{y}_i^{t-1})\ and\ h_i = 2$. Let $\mathcal{L}_j$ contain the set of samples assigned to node j, then the optimal weight is

$$\omega_j^* = \frac{\sum_{i \in x_i} g_i}{\sum_{i \in x_i} h_i + \lambda}, \qquad (13)$$

thereby

$$\mathcal{L}_j^* = -\frac{1}{2}\frac{(\sum_{i \in x_i} g_i)^2}{\sum_{i \in x_i} h_i + \lambda} + \gamma. \qquad (14)$$

This mechanism helps our model learn the relationship between time, operational characteristics such as cost per mile, and distance, which classical linear models are unable to learn well.

4.4.Prediction for new loads

By specifying future features that are characterized by the feature x_{new} (this feature contains the same cargo characteristics that were used to train our model, such as date, zip codes, weight, price, distance, etc.), our trained ensemble produces a point forecast

$$\hat{y}_{new} = f^*(x_{new}) = \sum_{i=1}^{N} f^*_k(x_{new}) \quad (15)$$

determining the projected cost per mile for the expected load.

Since our vector contains such features as loading and unloading date and addresses, prices may change at the same address depending on the loading date.

In order to correctly evaluate the predicted results, we calculate several indicators on the test set.

Let T denote a set of test indices, and $|T| = N_{test}$. The predicted and realized prices $\hat{y}_i, y_i$ respectively.

1. *Mean Absolute Error (MAE)*

$$MAE = \frac{1}{N_{test}} \sum_{i \in T} |y_i - \hat{y}_i| .$$

2. *Root Mean Squared Error (RMSE)*

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{i \in T} (y_i - \hat{y}_i)^2}.$$

3. *Coefficient of determination (R^2)*

$$R^2 = \frac{\sum_{i \in T} (y_i - \hat{y}_i)^2}{\sum_{i \in T} (y_i - \bar{y}_i)^2}, \qquad \bar{y} = \frac{1}{N_{test}} \sum_{i \in T} y_i$$

These indicators together help to qualitatively evaluate the average deviation and relative errors when obtaining results.

5. Results

The XGBoost model was trained on a dataset from 2022, 2023, and 2024. The target was freight costs. We tested our dataset on five models:

1. Global mean
2. Lane mean
3. Linear regression
4. Random forest
5. XGBoost model

Model	MAE (USD)	RMSE (USD)	(R^2)
Global mean baseline	772.277	992.679	-0.000
Lane mean baseline	392.506	568.136	0.672
Linear regression	454.184	582.503	0.656
Random forest	287.040	446.527	0.799
XGBoost (proposed)	**242.640**	**376.390**	**0.863**

Table 3. Performance of baseline models and XGBoost

The results show that XGBoost outperforms all baseline models and shows the best result across all criteria.

Considering that the average cost of freight transportation for the period 2022-2024 is $1991, we find that for MAE we have an error of approximately 12%, and for RMSE, about 19%. The R^2 indicator = 0.863 indicates that our model can explain more than 86% of freight transportation prices.
We also ran tests where we randomly selected cargo data for 2025 and tested it on our model. The test consisted of 90 lines of cargo information.

DATE	Orgin code	Destin coded	origin	destination	Distance	Weight	RATE	Predicted Rate	Difference	load number
2-Jan	6	16	CT 06422	KY 41076	744	12809	1400	1423	23	4297526
11-Feb	22	10	IN 47807	PA 18902	761	35000	2200	2147	-53	101596
7-Mar	18	15	GA 30549	OH 43125	540	30000	950	938	-12	1480814
8-Apr	18	15	GA 30549	OH 43125	602	30000	1100	1142	42	1498401
7-May	4	22	MA	IN	933	10000	1200	1173	-27	134037
9-Jun	8	15	NJ 07105	OH 45240	660	40000	900	977	77	150152
9-Jul	6	15	CT 06907	OH 43228	683	10000	1000	984	-16	2002140277
12-Aug	8	15	NJ, 08810	OH, 45502	591	38000	1000	973	-27	6343
9-Sep	24	20	IL 60632	FL 33716	1225	30000	3375	3370	-5	609305
7-Oct	21	6	MI 48393	CT 06484	732	12600	2100	2106	6	2E+09
11-Nov	16	37	KY 40215	TX 78041	1265	35872	2600	2580	-20	9254286

Table 4. XGBoost model test result on dataset of 2025 loads

The results show that the model is quite successful in describing and capturing spot prices, but there is also an error due to market volatility, which is difficult to predict.
As we can see, the XGBoost model is more efficient in terms of MAE estimation:

- 69% more efficient than the Global Mean Baseline
- $38 more efficient than the Lane Mean model
- 47% more efficient than Linear Regression
- 15% more efficient than Random Forest

According to the RMSE score, our model reduces errors by 62% more than the Global mean Baseline, 34% more than the lane mean model, 35% better than linear regression, and 16% better than Random Forest.
Regarding variance, our model also performs better and increases R^2 by:

- +0.191 relative to the lane mean model
- +0.207 relative to the linear regression
- and +0.064 relative to the random forest.

6. Conclusion

In this paper, we developed and evaluated a data-driven spot truck loading rate forecasting model using a modified gradient-boosted decision tree model (XGBoost) and the LAG-WMR model. Using real-world operational data for 2022–2024, we constructed a rich set of explanatory features that integrate calendar effects, lane information (shipper-sender pairs), distance, weight, and other operational attributes. The problem was formulated as a supervised regression problem with the total cost of a load (RATE, USD per load) as the target variable.

To evaluate the advantages of the proposed approach, we compared XGBoost with several intuitive and widely used benchmark models: the global mean model, the lane mean model, multiple linear regression, and a random forest ensemble. On the withheld test set, the XGBoost model achieved a mean absolute error of $242.64, a root-mean-squared error of $376.39, and an R^2 of 0.863. These values represent significant improvements over all baselines: the MAE is reduced by approximately 69% compared to the global mean, 38% compared to the band mean, 47% compared to linear regression, and 15% compared to random forest, while R^2 increases to the highest level among all tested models.

References

1. J.W. Miller, "ARIMA Time Series Models for Full Truckload Transportation Prices," *Forecasting*, vol. 1, no. 1, pp. 121–134, 2019.
2. A. Budak, A. Ustundag, B. Guloglu, "A Forecasting Approach for Truckload Spot Market Pricing," *Transportation Research Part A: Policy and Practice*, vol. 97, pp. 55–68, 2017.
3. A. Budzyński, M. Cieśla, "Application of a Machine Learning Model for Forecasting Freight Rate in Road Transport," *Scientific Journal of Silesian University of Technology, Series Transport*, vol. 126, pp. 23–48, 2025.
4. A. Budzyński, "Enhancing Road Freight Price Forecasting Using Gradient Boosting Machines," *Mathematics*, vol. 13, no. 18, article 2964, 2025.
5. T. Green, A. Rokoss, K. Kramer, M. Schmidt, "Application of Machine Learning on Transport Spot Rate Prediction in the Recycling Industry," in *Proceedings of the Conference on Production Systems and Logistics (CPSL 2022)*, pp. 554–563, 2022.
6. B. Mrówczyńska, M. Cieśla, A. Król, A. Sładkowski, "Application of Artificial Intelligence in Prediction of Road Freight Transportation," *Promet – Traffic & Transportation*, vol. 29, no. 4, pp. 363–370, 2017.
7. J.W. Miller, A. Scott, B.D. Williams, "Pricing Dynamics in the Truckload Sector: The Moderating Role of the Electronic Logging Device Mandate," *Journal of Business Logistics*, vol. 42, no. 4, pp. 388–405, 2021.
8. H. Bousqaoui, S. Achchab, K. Tikito, "Machine Learning Applications in Supply Chains: An Emphasis on Neural Network Applications," in *Proc. 2017 3rd Int. Conf. on Cloud Computing Technologies and Applications (CloudTech)*, pp. 1–7, 2017.
9. T. Chen, C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 785–794, 2016.
10. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

11. Y. LeCun, Y. Bengio, G. Hinton, “Deep Learning,” *Nature*, vol. 521, pp. 436–444, 2015.
12. W.B. Powell, “Maximizing Profits for North American Van Lines’ Truckload Division,” *Interfaces*, vol. 18, no. 1, pp. 21–41, 1988.
13. V. Ram, “Use Machine Learning to Forecast Trucking Rates,” Penske Logistics, online article, approx. 2018–2019 (accessed 2025).
14. Truckstop, “Understanding Spot Freight Rates,” Truckstop.com, online article, Apr. 5, 2022 (accessed 2025).
15. Flock Freight, “Using Trucking Spot Rates to Find Higher-Paying Loads,” FlockFreight.com, online article, May 9, 2023 (accessed 2025)