Analysis of big data from New York taxi trip 2023: revenue prediction using ordinary least squares solution and limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithms

Sara Rhouas, Norelislam El Hami
Engineering Science Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Jun 14, 2024
Revised Sep 6, 2024
Accepted Oct 1, 2024

Keywords:

Big data
Data analysis
Linear regression
Machine learning
Spark

ABSTRACT

This study explores the prediction of taxi trip fares using two linear regression methods: normal equations (ordinary least squares solution (OLS)) and limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS). Utilizing a dataset of New York City yellow taxi trips from 2023, the analysis involves data cleaning, feature engineering, and model training. The data consists of over 12 million records, managed, and processed that involves configuring the Spark driver and executor memory to efficiently process the Parquet-format data stored on hadoop distributed file system (HDFS). Key features influencing fare amount, such as passenger count, trip distance, fare amount, and tip amount, were analyzed for correlation. Models were trained on an 80-20 train-test split, and their performance was evaluated using root-mean-square error (RMSE) and mean squared error (MSE). Results show that both methods provide comparable accuracy, with slight differences in coefficients and training time. Additionally, vendor performance metrics, including total trips, average trip distance, fare amount, and tip amount, were analyzed to reveal trends and inform strategic decisions for fleet management. This comprehensive analysis demonstrates the efficacy of linear regression techniques in predicting taxi fares and offers valuable insights for optimizing taxi operations.

Corresponding Author:

Rhouas Sara
Engineering Science Laboratory, National School of Applied Sciences, Ibn Tofail University
Av. de L'Université, Kénitra, Marocco
Email: rhouas.sara@gmail.com

1. INTRODUCTION

Predicting taxi trip fares accurately is a critical task for both fleet operators and passengers in urban transportation systems. With the advent of big data and advanced analytical tools, it is now possible to leverage extensive datasets to gain insights into fare determinants and improve fare prediction models [1]. This study focuses on utilizing linear regression techniques to predict taxi trip fares using data from New York City's yellow taxi fleet for the entire year of 2023 [2]. By comparing two prominent regression methods, normal equations ordinary least squares (OLS) solution and limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), we aim to identify the most effective approach for fare prediction. Accurate fare predictions can enhance operational efficiency, optimize pricing strategies, and improve customer satisfaction by providing transparent and predictable fare estimates [3], [4].

New York City’s yellow taxi dataset provides a rich source of information, encompassing millions of trip records with diverse attributes such as trip distance, passenger count, fare amount, tip amount, and temporal details [5]. The large volume of data allows for a detailed analysis of fare determinants and the development of robust predictive models. However, the presence of null values and outliers necessitates rigorous data cleaning and preprocessing. This study systematically addresses these challenges, ensuring the integrity and reliability of the dataset. Feature engineering techniques are employed to extract meaningful insights from the data, such as temporal patterns in trip frequencies and fare variations across different vendors [6], [7].

In addition to building predictive models, this study conducts a comprehensive correlation analysis to understand the relationships between various trip attributes and the fare amount. By examining these correlations, we identify the most significant features influencing fare predictions. The performance of the regression models is evaluated using metrics such as root-mean-square error (RMSE) and mean squared error (MSE), providing a quantitative measure of their accuracy. Furthermore, the study delves into vendor performance analysis, comparing key performance indicators like total trips, average trip distance, fare amount, and tip amount across different vendors. This holistic approach not only highlights the effectiveness of linear regression techniques in fare prediction but also offers valuable insights into vendor operations, contributing to the overall optimization of taxi services in New York City [8].

2. METHOD

Our approach to analyzing New York City taxi trip data in 2023 combines the Apache spark platform and linear regression models for fare prediction. Spark handles large datasets, enabling efficient data cleaning, transformation, and analysis. After loading the data from parquet files and filtering invalid records, we use linear regression to predict fares based on features like passenger count, trip distance, fare, and tips. We implement two methods for linear regression, normal equations for smaller data and L-BFGS for high-dimensional data [9]. The data is split into training and test sets to evaluate performance using RMSE and MSE. We also assess taxi vendors' performance by analyzing metrics such as trip counts, average distance, fare, and tips, visualized through bar charts to highlight performance differences. This integrated approach enhances taxi service efficiency and supports strategic decision-making in transportation [10].

2.1. Work methodology

In this section, we will explore the various methodologies and tools employed to handle big data, focusing on techniques that enable efficient processing and analysis of large datasets. We will delve into the application of linear regression in machine learning, discussing how different approaches, such as ordinary least squares (OLS) and limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), can be utilized to optimize model parameters for predictive accuracy. Additionally, we will examine the metrics used for evaluating model performance, shedding light on how they measure the effectiveness of predictive models, identify areas for improvement, and ensure that the chosen algorithms align with the goals of data-driven decision-making. Through this exploration, we aim to provide a comprehensive understanding of how big data processing tools, such as Apache Spark, and linear regression techniques can be leveraged to build, optimize, and evaluate predictive models.

2.1.1. Tools for handling big data

Our approach to managing big data relies on Apache spark, a distributed computing system known for its efficiency and scalability. Apache spark excels in processing large datasets by distributing tasks across a cluster of computers, which enables parallel processing. This capability significantly reduces data processing time compared to traditional single-machine methods [11].

Apache spark uses resilient distributed datasets (RDDs) to ensure fault tolerance and enhance performance. RDDs are cached in memory, allowing iterative algorithms to reuse intermediate results across multiple computations. This feature greatly speeds up machine learning algorithms and other tasks that require multiple data passes [12].

Spark's unified analytics engine supports diverse data processing needs, including batch processing, real-time stream processing, and machine learning. It includes several specialized libraries, such as Spark SQL for SQL queries, spark streaming for real-time data, MLlib for machine learning, and GraphX for graph processing. These libraries extend Spark's functionality and make it versatile for various data tasks [13].

One of spark’s notable advantages is its in-memory computing capability, which allows for rapid data processing by storing data in memory rather than on disk [14]. This feature is particularly beneficial for iterative algorithms and interactive data exploration [15]. Additionally, spark’s user-friendly APIs in Java, Scala, Python, and R simplify the creation of complex workflows and data pipelines while providing advanced controls for experienced users. Spark’s efficient processing and versatile capabilities make it essential for modern data analytics and machine learning [16], [17].

2.1.2. Linear regression in machine learning

Linear regression is one of the most fundamental and widely used techniques in machine learning for predicting a continuous target variable based on one or more predictor variables [18]. At its core, linear regression aims to model the relationship between the dependent variable (the target) and the independent variables (the predictors) by fitting a linear equation to observed data. The primary objective of linear regression is to determine the optimal values for these coefficients such that the sum of the squared differences between the observed actual values and the values predicted by the linear model (known as the residual sum of squares) is minimized. In our study, we utilized two specific methods to perform linear regression: the OLS. and the L-BFGS algorithm [19].

The ordinary least squares (OLS) method is a fundamental approach in linear regression used to estimate the coefficients that minimize the residual sum of squares between the observed values and the values predicted by the model. The goal is to find the best-fit line that captures the relationship between the independent variables (predictors) and the dependent variable (target) [20]. The OLS solution is derived using the normal (1):

$$\beta = (X^T X)^{-1} X^T y \tag{1}$$

where β represents the vector of coefficients, X is the matrix of input features (including a column of ones for the intercept term), y is the vector of observed values, X^T is the transpose of the matrix. This method provides an exact solution by solving the above equation, making it straightforward and computationally efficient for smaller datasets. However, for very large datasets, the matrix inversion can become computationally expensive, which is a limitation of this approach [21].

The L-BFGS algorithm is an iterative optimization technique particularly well-suited for large-scale and high-dimensional datasets [22]. It is a variant of the BFGS algorithm that uses limited memory to approximate the inverse Hessian matrix, which is essential for determining the direction of the steepest descent in optimization problems. The iterative process follows these steps (2):

$$\beta_{k+1} = \beta_k - \alpha_k H_K^{-1} \nabla f(\beta_k) \tag{2}$$

where β_k is the coefficient vector at iteration k, α_k is the step size (learning rate), H_K^{-1} is the inverse hessian matrix approximation at iteration k, and $\nabla f(\beta_k)$ is the gradient of the cost function at β_k. Unlike the OLS method, L-BFGS does not require matrix inversion, making it more scalable and efficient for handling large datasets. It iteratively adjusts the coefficients by following the gradient of the cost function, gradually converging to the optimal solution. This makes L-BFGS particularly advantageous for scenarios where the dataset size or the number of features is large [23]. Despite the simplicity and interpretability of linear regression, it is essential to evaluate the underlying assumptions—such as linearity, independence, homoscedasticity (constant variance of errors), and normality of error terms—to ensure the validity and reliability of the model's predictions. By carefully selecting the appropriate method and validating the assumptions, linear regression remains a powerful tool for understanding and predicting the relationships within the data across various domains [24].

2.1.3. Scoring metrics

To fit the linear regression model using these methods, we first prepare the data by consolidating the selected features into a single vector using a VectorAssembler. The dataset is then divided into training and test sets, which allows us to evaluate the model's performance. Evaluation metrics such as RMSE and MSE are used to assess how well the model generalizes to unseen data. These metrics are essential for determining the accuracy of our predictions, offering insights into the model's effectiveness and its ability to handle new data [25].

MSE is a widely used metric for evaluating the accuracy of predictive models. It quantifies the mean of the squared differences between predicted and observed values. MSE essentially measures the average magnitude of the squared deviations across all data points, providing a detailed assessment of model performance. This metric is valuable for understanding the overall quality of the model's predictions, as it captures the extent of prediction errors in a continuous manner [26].

RMSE is another key metric that offers a straightforward measure of prediction error. By taking the square root of the MSE, RMSE presents an error metric that maintains the same units as the target variable, making it more intuitive. RMSE places a higher emphasis on larger errors due to the squaring of differences, which means it penalizes significant deviations more. This characteristic makes RMSE particularly useful for understanding the model's performance with respect to outlier predictions [27].

2.2. Application method

In this analysis, a combination of machine learning techniques, including linear regression and Spark's distributed computing capabilities, were employed to predict taxi trip revenues in New York City for the year 2023. Leveraging Spark's powerful data processing platform, the analysis aimed to provide accurate revenue predictions by incorporating key features such as passenger count, trip distance, fare amount, and tip amount. The utilization of linear regression, a well-established and interpretable modeling technique, ensured a comprehensive and effective approach to revenue prediction. Furthermore, Spark's distributed computing capabilities enabled the efficient handling of large-scale datasets, allowing for timely and accurate predictions even with massive amounts of data.

2.2.1. Data used

The dataset utilized in this analysis consisted of New York City taxi trip data for the year 2023, sourced from Parquet files. These files contain detailed information about taxi trips, including attributes such as pickup datetime, passenger count, trip distance, fare amount, tip amount, and total amount. The dataset was meticulously cleaned and preprocessed to ensure data quality and reliability for subsequent analysis. Invalid records and missing values were filtered out, and the pickup datetime column was cast to a date type for temporal analysis. This refined dataset served as the foundation for building and training the linear regression model for revenue prediction [28].

Through exploratory data analysis and feature engineering, insights were extracted from the dataset to enhance the predictive model's performance. Key features such as passenger count, trip distance, fare amount, and tip amount were identified based on their potential impact on trip revenues. These features were then used to train the linear regression model, which served as the predictive engine for estimating taxi trip revenues. By leveraging Spark's distributed computing capabilities, the model was able to efficiently process and analyze large-scale datasets, providing stakeholders with accurate and timely revenue predictions.

2.2.2. Process

This process outlines a data-driven approach for predicting taxi trip revenues in New York City for 2023. It begins with data loading and initial processing, where Spark is configured for efficient handling of large datasets. The data, stored in parquet format on Hadoop distributed file system (HDFS), is verified, loaded, and combined into a single data frame for the entire year. Following this, data cleaning and preprocessing ensure the dataset's integrity by removing rows with null or invalid values, reducing the data to 37,000,870 rows. Feature selection and engineering identify key insights, including peak operational periods and feature correlations, setting the stage for model training.

Model training and evaluation involves splitting the data into training and testing sets and applying two linear regression methods—normal equations and L-BFGS. The models are evaluated using RMSE and MSE metrics for accuracy. Performance evaluation and visualization examine feature impacts and vendor metrics, such as trip counts, average fares, and tips, while insights and decision-making leverage these results to optimize taxi operations and enhance customer satisfaction. This structured analysis offers actionable insights to improve service efficiency and profitability.

3. RESULTS AND DISCUSSION

The analysis focuses on evaluating the performance of two major taxi vendors in New York City. Using comprehensive trip data, key metrics such as the total number of trips, average trip distance, average fare amount, and average tip amount are analyzed to assess each vendor's operational efficiency and market positioning. The following sections provide a detailed examination of these metrics, highlighting the strengths and weaknesses of vendor 1 and vendor 2.

3.1. Performance analysis for each vendor

In this analysis, we examine the performance of two major taxi vendors in New York City using key metrics derived from comprehensive trip data. By evaluating the total number of trips, average trip distance, average fare amount, and average tip amount, we aim to understand the operational efficiency and market positioning of each vendor. The data spans a significant period and provides a robust foundation for comparing these vendors' effectiveness in meeting passenger demand and generating revenue. The following paragraphs delve into each metric, offering insights into the strengths and weaknesses of vendor 1 and vendor 2.

As shown in Figure 1, vendor 2 demonstrates a significantly higher volume of total trips compared to vendor 1. Specifically, vendor 2 recorded 27,471,887 trips, whereas vendor 1 recorded 9,528,983 trips. This disparity indicates that vendor 2 has a larger share of the market, which could be due to a variety of factors such as a more extensive fleet, more efficient dispatch and routing systems, or stronger brand recognition. The higher trip volume also suggests that vendor 2 is better at meeting passenger demand and potentially has wider operational coverage across New York City. This large volume of trips provides vendor 2 with a robust revenue base and enhances its ability to generate significant income from a high number of service transactions.

In Figure 2 we can see that the average trip distance for vendor 2 is slightly longer than that for vendor 1, with vendor 2 averaging 3.64 miles per trip and vendor 1 averaging 3.42 miles. While the difference may seem minimal, it has important implications for revenue. Longer trips typically result in higher fares, contributing more significantly to total revenue. Vendor 2's slightly longer average trip distance could indicate that they serve areas with greater distances between common pick-up and drop-off points or that they attract trips that tend to cover more distance. This could be a result of strategic operational decisions or a focus on areas with higher fare potential. The longer trip distances might also suggest that vendor 2 has a higher proportion of trips to and from major hubs like airports or business districts, which typically involve greater distances.

Vendor 2 also outperforms vendor 1 in terms of average fare amount, with an average fare of $19.67 compared to vendor 1's $18.71. This difference in fare amounts is likely linked to the longer average trip distances mentioned earlier. Higher average fares not only boost per-trip revenue but also suggest that vendor 2 may be operating more in premium segments of the market where passengers are willing to pay more for better service or convenience. Additionally, the higher fares could be a result of effective dynamic pricing strategies, where vendor 2 adjusts prices based on demand and supply conditions to maximize revenue. This ability to command higher fares strengthens vendor 2's overall financial performance and competitive advantage in the market.

The average tip amount is another area where vendor 2 leads, with an average tip of $3.65 compared to vendor 1's $3.26. Tips are often indicative of customer satisfaction and service quality. The higher average tips for vendor 2 suggest that passengers perceive the service quality to be better or feel more satisfied with their rides. This could be due to various factors such as cleaner vehicles, more courteous drivers, better ride experiences, or more reliable service. Higher tips contribute directly to the drivers' earnings and can also boost overall driver morale and retention. From a business perspective, higher tips indicate a positive customer experience, which is crucial for customer loyalty and repeat business.

Vendor 2's higher trip volume, longer average trip distance, higher average fare amount, and greater average tip amount collectively paint a picture of a more dominant and financially successful operator. The higher trip volume indicates a larger operational scale and better market penetration, while the longer trip distances and higher fare amounts suggest a focus on higher-value segments of the market. The greater average tips reflect superior service quality, leading to higher customer satisfaction and loyalty. These factors combined position Vendor 2 as a more robust and competitive player in New York City's taxi industry, with a stronger ability to generate revenue and sustain long-term growth compared to vendor 1.

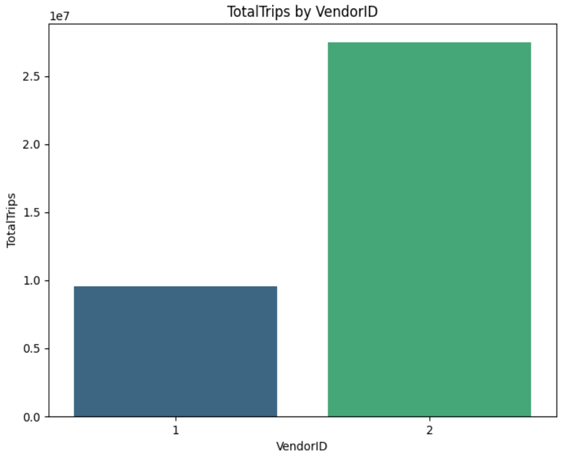

Figure 1. Total trips per vendor ID

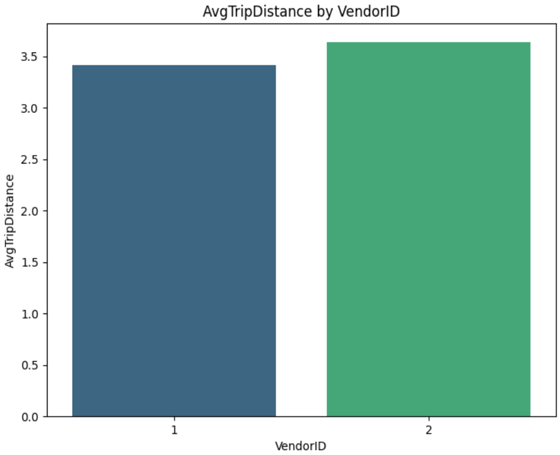

Figure 2. Average trip distance per vendor ID

3.2. Analysis of the regression performances

In analyzing yellow taxi trip fare prediction, linear regression models were employed to understand the impact of various factors on the total fare. Two methods, OLS and L-BFGS, were used to build these models. Both methods offer distinct advantages in terms of computational efficiency and scalability, making them suitable for different contexts depending on the size and complexity of the dataset. This section delves into the results obtained from both regression methods, providing a detailed comparison of their performance metrics, computational requirements, and the significance of the derived coefficients. By examining the coefficients and their implications, we gain insights into the primary drivers of taxi fares, enhancing our understanding of fare structures and customer behaviors.

As shown in Table 1, both the OLS and L-BFGS linear regression models yielded nearly identical coefficients, demonstrating the robustness of the findings. The coefficient for passenger count is approximately 0.0702, indicating that each additional passenger has a small but positive impact on the total fare. This suggests that while having more passengers slightly increases the fare, their influence is relatively minimal compared to other factors. The trip distance coefficient, around 0.0010, also shows a very small impact on the total fare, indicating that trip distance contributes marginally to fare calculations. This small impact might reflect a fare structure where fixed costs or time-based charges are more significant than distance, potentially due to minimum fare policies or the inclusion of initial service fees that overshadow the distance-based component.

Table 1. The results of each method

Metric	OLS	L-BFGS
Training time (seconds)	37.96	122.69
RMSE	4.691598018	4.6915980186
MSE	22.01109196	22.01109197
Passenger count coefficient	0.070184284	0.070184285
Trip distance coefficient	0.0009725934876	0.00097259340671
Fare amount coefficient	1.0036740054	1.0036740051
Tip amount coefficient	1.35752071718	1.357520719
Intercept	4.008011703	4.008011700

The fare amount, with a coefficient of about 1.0037, shows a near one-to-one relationship with the total fare, confirming that base fare calculations are the primary determinant of the total fare. In contrast, the tip amount, with a coefficient of approximately 1.3575, indicates that tips significantly boost the total fare. This higher coefficient suggests that tipping not only adds directly to the fare but also correlates with scenarios involving higher service quality or more expensive rides. The intercept, around 4.0080, represents the baseline total fare, ensuring a minimum charge regardless of other factors. This baseline underscores the importance of initial fees in the fare structure. Collectively, these coefficients reveal that while passenger count and trip distance play secondary roles, the fare amount and tips are crucial drivers of the total fare, reflecting a fare structure heavily influenced by base charges and customer tipping behavior.

The fare amount, with a coefficient of 1.0037, shows a near one-to-one relationship with the total fare, confirming that base fare calculations are the primary factor. Meanwhile, the tip amount, with a coefficient of 1.3575, has a more significant influence, indicating that tips not only increase the fare directly but also correlate with scenarios involving higher service quality or more expensive rides. The intercept, around 4.0080, ensures a minimum fare, emphasizing the importance of base charges. Overall, fare amount and tips are the main drivers of the total fare, with passenger count and trip distance playing smaller roles.

The OLS and L-BFGS linear regression models were used to predict taxi fares, with both showing nearly identical performance metrics. The OLS model, using the normal equations method, had an RMSE of 4.6916 and an MSE of 22.0111, and completed in 37.96 seconds, making it efficient for datasets that fit within memory limits. This efficiency comes from the closed-form solution of the Normal Equations, which allows for quick calculations when data size is manageable.

The L-BFGS model, an iterative optimization method for larger datasets, achieved the same RMSE and MSE as the OLS model. However, its computational time was significantly longer, at 122.69 seconds, reflecting its iterative nature. Despite this, the L-BFGS method is more flexible and scalable, making it suitable for large datasets that exceed memory limits. Its performance and coefficient alignment with the OLS model confirm its effectiveness in capturing the dataset's linear relationships.

Comparing the two models, both showed similar predictive accuracy, but the OLS method was faster and more efficient for smaller datasets, while the L-BFGS method excelled in handling larger, more complex datasets. The choice between the two depends on the dataset size and computational needs, with OLS favored for speed and L-BFGS for scalability. Understanding these trade-offs ensures the appropriate model is used for efficient and accurate analysis.

4. CONCLUSION

The growing interest in big data and machine learning has revolutionized numerous industries, including urban transportation. Leveraging these advanced technologies allows for more informed decision-making, operational efficiency, and enhanced customer experiences. In this context, analyzing extensive datasets, such as those generated by New York City's yellow taxi services, provides valuable insights into the performance and market dynamics of competing vendors. This study harnesses the power of big data and machine learning to evaluate the operational metrics of two major taxi vendors, offering a detailed comparison of their effectiveness in meeting passenger demand and generating revenue.

In conclusion, the integration of big data and machine learning in analyzing New York City's yellow taxi industry reveals vendor 2 as the more dominant and financially successful operator. Higher trip volumes, longer average trip distances, higher fare amounts, and greater tips position vendor 2 as a stronger competitor with a better ability to meet passenger demands and generate revenue. These insights are instrumental for both vendors in optimizing their operations, improving service quality, and making data-driven decisions that enhance customer satisfaction and operational efficiency. This study exemplifies the transformative potential of big data and machine learning in urban transportation, paving the way for more effective and competitive service delivery.

REFERENCES

[1] B. Itri, Y. Mohamed, B. Omar, E. M. Latifa, M. Lahcen, and O. Adil, “Hybrid machine learning for stock price prediction in the Moroccan banking sector,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, pp. 3197–3207, Jun. 2024, doi: 10.11591/ijece.v14i3.pp3197-3207.

[2] Q. Hu, L. Zhu, C. Chang, and W. Zhang, “A truncated three-term conjugate gradient method with complexity guarantees with applications to nonconvex regression problem,” *Applied Numerical Mathematics*, vol. 194, pp. 82–96, Dec. 2023, doi: 10.1016/j.apnum.2023.08.006.

[3] A. L. Burton, “Ordinary least squares (linear) regression,” in *The Encyclopedia of Research Methods in Criminology and Criminal Justice*, Wiley, 2021, pp. 509–514.

[4] S. Rhouas, A. El Attaoui, and N. El Hami, “Optimization of the prediction performance in the future exchange rate,” in *2023 9th International Conference on Optimization and Applications (ICOA)*, Oct. 2023, pp. 1–6, doi: 10.1109/ICOA58279.2023.10308858.

[5] A. El Attaoui, S. Rhouas, and N. El Hami, “ETL applied to klarna e-commerce dataset,” in *2023 9th International Conference on Optimization and Applications (ICOA)*, Oct. 2023, pp. 1–4, doi: 10.1109/ICOA58279.2023.10308808.

[6] M. B. Ulak, A. Yazici, and M. Aljarrah, “Value of convenience for taxi trips in New York City,” *Transportation Research Part A: Policy and Practice*, vol. 142, pp. 85–100, Dec. 2020, doi: 10.1016/j.tra.2020.10.016.

[7] M. S. Ansar, Y. Ma, S. Chen, K. Tang, and Z. Zhang, “Investigating the trip configured causal effect of distracted driving on aggressive driving behavior for e-hailing taxi drivers,” *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 8, no. 5, pp. 725–734, Oct. 2021, doi: 10.1016/j.jtte.2020.12.001.

[8] X. Dong, E. Guerra, and M. S. Ryerson, “Investigating the recovery of for-hire-vehicle, taxi, and airtrain at two New York City airports during the COVID-19 pandemic,” *Travel Behaviour and Society*, vol. 33, Oct. 2023, doi: 10.1016/j.tbs.2023.100646.

[9] D. Katić, H. Krstić, I. Ištoka Otković, and H. Begić Juričić, “Comparing multiple linear regression and neural network models for predicting heating energy consumption in school buildings in the Federation of Bosnia and Herzegovina,” *Journal of Building Engineering*, vol. 97, Nov. 2024, doi: 10.1016/j.jobe.2024.110728.

[10] B. V. Surya Vardhan, M. Khedkar, I. Srivastava, and S. K. Patro, “Impact of integrated classifier — regression mapped short term load forecasting on power system management in a grid connected multi energy systems,” *Electric Power Systems Research*, vol. 230, May 2024, doi: 10.1016/j.epsr.2024.110222.

[11] M. Armanur Rahman, A. Hossen, J. Hossen, V. C, T. Bhuvaneswari, and A. Sultana, “Towards machine learning-based self-tuning of hadoop-spark system,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 2, pp. 1076–1085, Aug. 2019, doi: 10.11591/ijeecs.v15.i2.pp1076-1085.

[12] A. Manconi, M. Gnocchi, L. Milanesi, O. Marullo, and G. Armano, “Framing Apache spark in life sciences,” *Heliyon*, vol. 9, no. 2, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13368.

[13] P. Jha, A. Tiwari, N. Bharill, M. Ratnaparkhe, M. Mounika, and N. Nagendra, “Apache spark based kernelized fuzzy clustering framework for single nucleotide polymorphism sequence analysis,” *Computational Biology and Chemistry*, vol. 92, Jun. 2021, doi: 10.1016/j.compbiolchem.2021.107454.

[14] M. B. Al-Masadeh, M. S. Azmi, and S. S. Syed Ahmad, “Tiny datablock in saving hadoop distributed file system wasted memory,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 2, pp. 1757–1772, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1757-1772.

[15] F. Ashkouti, K. Khamforoosh, and A. Sheikhahmadi, “DI-Mondrian: distributed improved Mondrian for satisfaction of the L-diversity privacy model using Apache spark,” *Information Sciences*, vol. 546, pp. 1–24, Feb. 2021, doi: 10.1016/j.ins.2020.07.066.

[16] Y. Liu and S. Cao, “The analysis of aerobics intelligent fitness system for neurorobotics based on big data and machine learning,” *Heliyon*, vol. 10, no. 12, Jun. 2024, doi: 10.1016/j.heliyon.2024.e33191.

[17] S. G Purohit and V. Swamy, “Enhancing data publishing privacy: split-and-mould, an algorithm for equivalent specification,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1273-1282.

[18] Y. Huang, W. Xu, P. Sukjairungwattana, and Z. Yu, “Learners’ continuance intention in multimodal language learning education: an innovative multiple linear regression model,” *Heliyon*, vol. 10, no. 6, Mar. 2024, doi: 10.1016/j.heliyon.2024.e28104.

[19] C. Kleiber, “Finite sample efficiency of OLS in linear regression models with long-memory disturbances,” *Economics Letters*, vol. 72, no. 2, pp. 131–136, Aug. 2001, doi: 10.1016/S0165-1765(01)00423-2.

[20] I. Ahmad *et al.*, “Spatial configuration of groundwater potential zones using OLS regression method,” *Journal of African Earth Sciences*, vol. 177, May 2021, doi: 10.1016/j.jafrearsci.2021.104147.

[21] E. Ghysels and H. Qian, “Estimating MIDAS regressions via OLS with polynomial parameter profiling,” *Econometrics and Statistics*, vol. 9, pp. 1–16, Jan. 2019, doi: 10.1016/j.ecosta.2018.02.001.

[22] A. Bemporad, “An L-BFGS-B approach for linear and nonlinear system identification under l_1 and group-Lasso regularization,” *arXiv:2403.03827*, Mar. 2024.

[23] F. Alpak *et al.*, “A machine-learning-accelerated distributed LBFGS method for field development optimization: algorithm, validation, and applications,” *Computational Geosciences*, vol. 27, no. 3, pp. 425–450, Jun. 2023, doi: 10.1007/s10596-023-10197-3.

[24] D. Chang, S. Sun, and C. Zhang, “An accelerated linearly convergent stochastic L-BFGS algorithm,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3338–3346, Nov. 2019, doi: 10.1109/TNNLS.2019.2891088.

[25] M. W. Liemohn, A. D. Shane, A. R. Azari, A. K. Petersen, B. M. Swiger, and A. Mukhopadhyay, “RMSE is not enough: guidelines to robust data-model comparisons for magnetospheric physics,” *Journal of Atmospheric and Solar-Terrestrial Physics*, vol. 218, Jul. 2021, doi: 10.1016/j.jastp.2021.105624.

[26] S. Rhouas, A. El Attaoui, and N. El Hami, “Enhancing currency prediction in international e-commerce: Bayesian-optimized random forest approach using the Klarna dataset,” *International Journal of Electrical and Computer Engineering*, vol. 14, no. 3, Jun. 2024, doi: 10.11591/ijece.v14i3.pp3177-3186.

[27] S. Hadiyoso, H. Nugroho, T. L. Erawati Rajab, and K. Surendro, “Data prediction for cases of incorrect data in multi-node electrocardiogram monitoring,” *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, pp. 1540–1547, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1540-1547.

[28] “TLC trip record data,” *Taxi and Limousine Commission*. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page (accessed Jun. 13, 2024).

BIOGRAPHIES OF AUTHORS

Sara Rhouas is currently pursuing her Ph.D. in computer science at the National School of Applied Sciences, Ibn Tofail University in Morocco. She earned her engineering degree in Industrial Engineering in 2019 from the same institution. Her academic experience includes a strong focus on automobile technologies, with a particular interest in braking systems. She has carried out research in the field of optimization algorithms and her research interests extend to areas such as big data, interoperability, artificial intelligence, machine learning, and deep learning. She has authored and co-authored several publications in both conferences and scientific journals. She can be contacted at email: rhouas.sara@gmail.com.

Norelislam El Hami is a professor of computer science at the National School of Applied Sciences, Ibn Tofail University, in Kenitra, Morocco. He earned a diploma of state engineer in 2000, specializing in computer and telecommunications from the Polytechnic Faculty of Mons (FPMS) in Belgium. He holds a Ph.D. in computer science from the National Institute of Applied Sciences (INSA) of Rouen, France, as well as a Ph.D. in applied mathematics and computer science from Mohammed V University in Rabat, Morocco. His work includes numerous scholarly publications in conferences and journals. He can be contacted at email: norelislam@outlook.com.