e-ISSN: 3089-5081 | p-ISSN: 1979-8911
Vol. 14 No. 2 December 2025: 52-56
DOI: 10.15575/istek.v14i2.2338

ANALYSIS OF CRITICAL FACTORS INFLUENCING ONLINE MOTORCYCLE TAXI DRIVER'S INCOME PER TRANSACTION USING RANDOM FOREST REGRESSOR AND FEATURE IMPORTANCE

Agung Wahana[1], Cecep Nurul Alam[2]

[1,2] Informatics Department, UIN Sunan Gunung Djati Bandung, Indonesia

Article Info

Article history:

Received November 11, 2025
Revised December 20, 2025
Accepted December 28, 2025

Keywords:

Random Forest Regressor
Driver Income
Feature Importance
Platform Rates

ABSTRACT

This study aims to identify and measure the main factors that most significantly affect the Income of Online Motorcycle Taxi Drivers Per Transaction in the gig economy sector. The Machine Learning Random Forest Regressor algorithm was used on driver transaction data. This methodology was chosen for its ability to handle the data's non-linearity and to objectively measure Feature Importance. Traditional linear regression models have limitations in these areas. The main results show the Random Forest model is highly accurate (R^2 = 0.9634). It confirms the absolute dominance of distance, which accounts for 94.98% of the total predictive importance of revenue. The Total Transaction Value factor (3.82%) is a secondary predictor. Demographic variables (Age and Gender) and temporal variables (Days and Hours) together had a minimal (less than 1%) influence on fare per trip. This research concludes that the rate per driver transaction is determined almost exclusively by the platform's distance-based pricing policy. It is neutral to the characteristics of the driver. These findings recommend that platforms focus on increasing order volume and optimizing operational costs, rather than modifying base rates.

Corresponding Author:

Agung Wahana,
Informatics Department, Faculty of Science & Technology, UIN Sunan Gunung Djati Bandung
Jl. A. H. Nasution No. 105, Cibiru, Bandung, Indonesia. 40614
Email : wahana.agung@uinsgd.ac.id

1. INTRODUCTION

The online transportation sector has become the backbone of urban mobility in Indonesia, where driver income is a central issue that determines the sustainability of their livelihoods. Daily and per-transaction revenue fluctuations are frequently discussed, influenced by a combination of internal (driver demographics) and external factors (transaction characteristics, temporal patterns, and fare systems). Previous studies have highlighted the important role of operational costs and time efficiency in determining drivers' net profits in the gig economy [1][2]. The gig economy is a labor market characterized by the dominance of short-term or contract jobs instead of permanent or full-time positions [3]. The term "gig" originates from the music industry, where musicians are paid for a single performance or concert[4].

Although the gig economy model offers work flexibility, income uncertainty and profit erosion remain central challenges, especially in developing countries like Indonesia. Research from 2023–2025 shows that while variables such as working hours and number of orders positively influence gross revenue, operational cost factors like fuel prices and vehicle maintenance are crucial and often overlooked, significantly eroding drivers' net income [5]. This uncertainty is exacerbated by the lack of social and health protections typically enjoyed by formal workers, collectively increasing their financial vulnerability [6]. Therefore, revenue analysis must go beyond descriptive statistics and account for complex cost structures.

Previous studies have often relied on Multiple Linear Regression (MLR), which has limitations in modeling non-linear relationships between demographic factors (age, gender) and income a hallmark of complex gig economy data. To overcome these limitations and provide more accurate analysis, this research adopted the Random Forest Regressor algorithm [7][8]. This machine learning approach not only achieves higher prediction accuracy (R^2 is shown to be high) but also inherently provides a feature importance metric[9]. This metric allows researchers to objectively measure and rank variable contributions, providing empirical clarity on the dominance of platform tariff policies versus the contribution of driver factors themselves. Random Forest Regressor is an ensemble-based machine learning model well suited for analyzing complex, non-linear data [10]. Its main advantage is the ability to measure feature importance objectively, ranking which factors are most dominant in predicting the target variable in this case, driver revenue. This ability is especially valuable for regression scenarios where predictive variables (such as time and demographics) tend to show marginal contributions [11].

The data used include transaction characteristics such as distance and total transaction value, driver demographic data including age and gender, and time variables such as days of the week and transaction hours. Through this approach, the study gains an in-depth understanding of the revenue structure, confirms the dominance of platform fare factors, and measures the true influence of demographic characteristics and time patterns on the income of online motorcycle taxi drivers[12].

2. METHOD

Figure 1 Flow of data science lifecycle. The Team Data Science Process (TDSP) is an Agile and iterative framework developed by Microsoft, similar to CRISP-DM but placing a greater emphasis on collaboration and the use of modern tools.

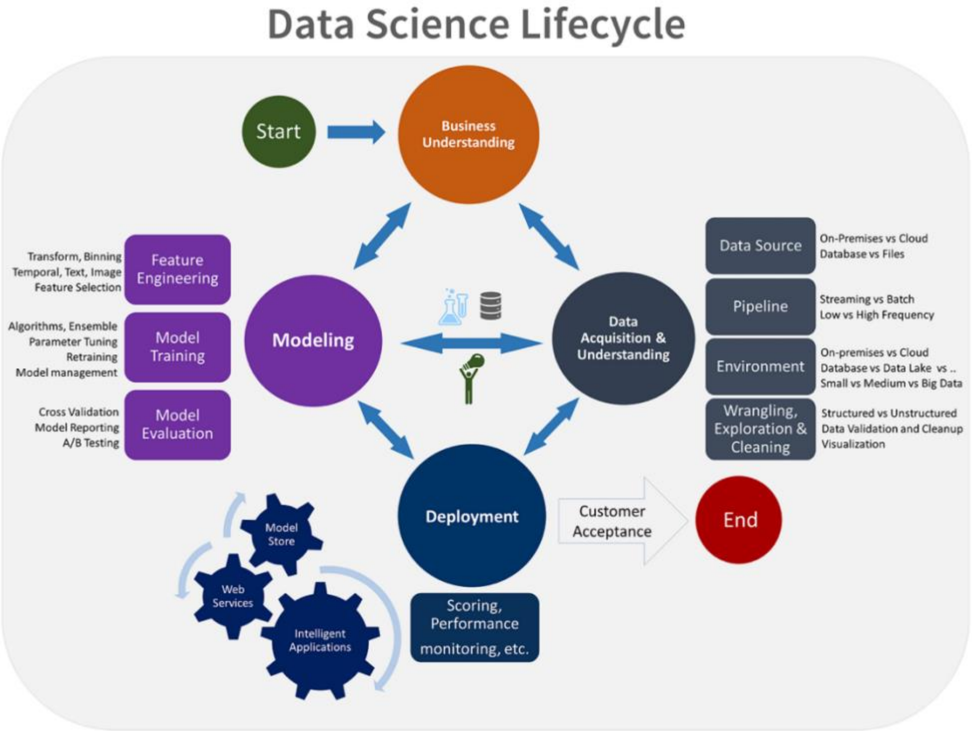

Figure 1. Research Methodology

This study focuses on identifying and measuring the main factors that most significantly affect Online Motorcycle Taxi Driver Income using the Random Forest Regressor algorithm.

3. RESULT AND DISCUSSION

Training and Testing used 4 data split scenarios namely 90:10, 80:20, 70:30, 60:40 with the following results:

Table 1. Training and Model Testing Results with 4 Data Split Scenarios

Scenarios	Split Data	R² Score
A	90:10	99,78%
B	80:20	98,03%
C	70:30	97,67%
D	60:40	93,73%

Table 1. It shows that scenario A gives the highest R^2 score of 99.78%, but scenario B is chosen with a data split of 80% for training and 20% for testing, even though scenario A gives the highest R^2 Score but has the risk of overfiting. Model Robustness with performance at R^2=98.03 indicates that the Random Forest model is quite robust and capable of generalizing very well, even when the training data is reduced. The Random Forest Regressor model was trained using 80% of transaction data and evaluated on 20% of the test data. Model performance is measured using the R^2 (R-squared) metric. The model is able to explain 96.34% of the variability of the total Driver Revenue. Very high values indicate the model's outstanding predictive ability.

Table 2. Feature Importance

Rating	Factors	Feature Importance
1	Distance	94,98%
2	Total Transaction Value	3,82%
3	DayOfWeek_Sunday	0,34%
4	Hour	0,29%
5	DayOfWeek_Saturday	0,17%
6	Other Factors: Age, Gender, Other Days	< 0,20%

Table 2 explains that these results confirm that the most dominant factor that affects driver revenue per transaction is distance. These results confirm that the most dominant factor that affects driver revenue per transaction is distance. The finding that Distance dominates the model up to 94.98% is very consistent with the platform fare structure which is linear to the distance per kilometer. Another transaction factor, Total Transaction Value (3.82%), being the second important predictor, shows the marginal contribution of commission to the value of goods. The time factor (DayOfWeek, Hour, DayOfMonth) also has very little importance, suggesting that revenue per transaction is stable over different time periods. Demographic factors (Age and driver_gender) account for less than 1% of feature importance. This indicates that the system of determining rates per trip is neutral to demographics.

Figure 2 explains the Linear Pattern; The data points (transactions) are clearly arranged to form a straight line or narrow band, moving from the bottom left to the upper right. Positive correlation; When Mileage increases, Driver Revenue. This pattern is the most powerful visual evidence confirming that the driver's rate per transaction is determined by the linear function of Mileage. If the data forms a random pattern or a non-linear curve, then other factors will dominate. This graph directly explains why the distance variable has a Feature Importance of 95% in the Random Forest Regressor model. This very close correlation makes distance an almost perfect predictor of income. At each distance point (e.g., 5 km or 10 km), the grouped income points are very close. This shows that for the same mileage, the variation in the rates charged by the platform is very minimal and consistent. The most fundamental visualization in this graph validates that driver revenue per transaction is a function of distance, and strongly supports the high accuracy of the Machine Learning model ($R^2 > 0.98$).

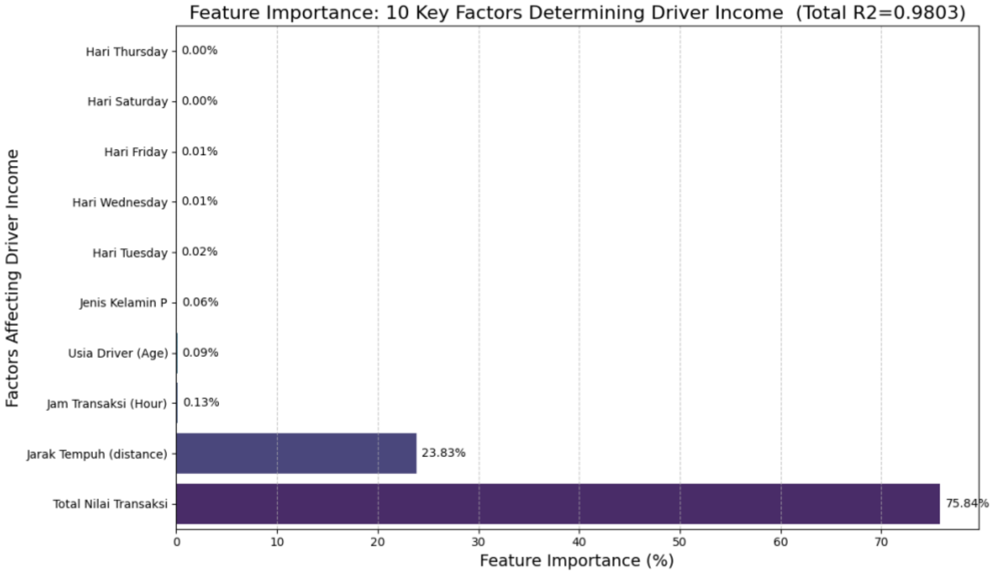

Figure 2: Feature Importance

4. CONCLUSION

The main results of the Random Forest Regressor model unequivocally show that Distance is the most dominant determinant of revenue per transaction, accounting for nearly 95% of the total feature importance. The high stability and accuracy of the model ($R^2=0.9634$) corroborate the fact that platform pricing policy is the macroeconomic variable that most strongly influences drivers. Demographic factors (Age and Gender) and time pattern factors (Transaction Day/Hours) have minimal contributions (total collective contribution less than 1%) to the amount of revenue per transaction.

The Gender and Day analysis confirmed that no significant bias was detected in rates per trip based on age or gender. The median income between Male and Female drivers is almost identical. Revenue fluctuations (payday effect or weekend surge) tend not to affect the amount of the rate per transaction, but may affect the Transaction Volume (the number of orders obtained by the driver).

REFERENCES

[1] S. J. Hong, J. M. Bauer, K. Lee, dan N. F. Granados, “Drivers of Supplier Participation in Ride-Hailing Platforms,” *Journal of Management Information Systems*, vol. 37, no. 3, hlm. 602–630, Jul 2020, doi: 10.1080/07421222.2020.1790177.

[2] S. Darmastuti, A. Rahmawati, R. L. Putri, dan A. Sundusiyah, “The Rise of the Gig Economy: Job Creation and the Paradox of Working Relationship on the Digital Transportation Platform.”

[3] M. Li, X. Hu, K. Jin, dan J. Han, “Exploring factors influencing entry into the gig economy: A study of Chinese workers,” *Acta Psychol (Amst)*, vol. 259, Sep 2025, doi: 10.1016/j.actpsy.2025.105301.

[4] J. Woodcock dan M. Graham, “A Critical Introduction.”

[5] E. Mariano, V. Amkeun[2], N. Bau[3], dan Y. P. Lian, “Analisis Pendapatan Pengemudi Ojek Online Maxim-Bike,” vol. 2, 2023, [Daring]. Tersedia pada: http://jurnal.jomparnd.com/index.php/jk

[6] D. S. Pratomo, P. M. A. Saputra, D. A. N. Asrofi, C. Natalia, dan S. L. Zenritami, “Gig Workers In The Digital Era In Indonesia: Development, Vulnerability And Welfare,” 2024, hlm. 47–60. doi: 10.2991/978-94-6463-525-6_6.

[7] L. Breiman, “Random Forests,” 2001.

[8] Y. Yang dan H. Wang, “Random Forest-Based Machine Failure Prediction: A Performance Comparison,” *Applied Sciences (Switzerland)*, vol. 15, no. 16, Agu 2025, doi: 10.3390/app15168841.

[9] T. Tulabandhula dan C. Rudin, “Machine Learning with Operational Costs,” 2013.

[10] K. R. Putra, “Comparison of Prediction Models: Decision Tree, Random Forest, and Support Vector Regression,” vol. 6, no. 1, hlm. 39–49, 2025, doi: 10.33365/jatika.v6i1.18.

[11] E. Bhaduri dan S. Pal, “Bhaduri, Pal, and Goswami Analysing Factors Affecting the Adoption of Ride-Hailing Services (RHS) in India? A Step-Wise LCCA-MCDM Modeling Approach.”

[12] X. Chen, S. Bai, Y. Wei, dan H. Jiang, “How income satisfaction impacts driver engagement dynamics in ride-hailing services,” *Transp Res Part C Emerg Technol*, vol. 157, hlm. 104418, Des 2023, doi: 10.1016/j.trc.2023.104418.