Proceedings of CONF-CIAP 2026 Symposium: International Conference on Atomic Magnetometer and Applications
DOI: 10.54254/2753-8818/2026.HZ27848

Random Forest Regression Analysis for Package Delivery Time Prediction

Allen He

BASIS International School Park Lane Harbour, Huizhou, China
allen.he16839-biph@basischina.com

Abstract. Predicting delivery times is a central challenge in logistics, especially during the last-mile phase, which is the final step in the supply chain where goods are delivered to the customer. This work studies previous approaches and proposes Random Forest Regression for Time Prediction (hereafter RFRTP), a Random Forest Regression-based model to improve delivery time predictions in logistics systems. This study examines the effectiveness of RFRTP compared to baseline statistical and machine learning models when accounting for the complex factors affecting delivery times, such as traffic conditions, order density, and route assignments. A dataset based on delivery data provided by JD Logistics, containing extensive information on last-minute deliveries in an urban environment, was used to train and test RFRTP and multiple baseline regression models, specifically Linear, Lasso, and Ridge, which were analyzed on their performances in delivery time prediction. The experimental results show that the proposed RFRTP model obtains the best results in last-minute delivery time prediction, which strongly contributes to more reliable time estimates, improved customer satisfaction, and greater efficiency. Through feature importance analysis on the RFRTP model, this study found the most critical dataset features for prediction, providing insight into the most important data. This study also calculated prediction intervals for the results of RFRTP, allowing for margins of error and transparency in the confidence of the model's estimates.

Keywords: Last-mile delivery, delivery time prediction, random forest regression.

1. Introduction

The recent rapid growth of e-commerce has placed enormous pressure on last-mile logistics, as even small inefficiencies can have a significant impact on customer satisfaction and operational costs. This study focuses on a critical challenge for logistics providers: the accurate prediction of delivery times. Precise estimates are crucial for companies to provide customers with reliable time windows and improve company revenue and credibility.

This study examines statistical modeling as a tool for addressing this challenge. By analyzing historical delivery data, patterns and relationships between variables such as delivery location, time of day, driver performance, and route characteristics are uncovered. Machine learning models, especially this study's Random Forest Regression for Time Prediction (hereafter RFRTP), which is

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.HZ27848

capable of capturing non-linear data, can leverage these relationships to produce accurate delivery time predictions.

This study used a real-world dataset containing over 430,000 delivery records from JD Logistics, one of the largest logistics companies in China. Each record has multiple features detailing the time, courier, and delivery. After applying statistical analysis to this dataset and training machine learning models on it, this study evaluates the models' predictive performance and identifies the most significant factors influencing delivery duration.

This study compares RFRTP with three baseline regression approaches: Linear Regression, Ridge Regression, and Lasso Regression. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2) were selected as performance metrics. A lower MAE and RMSE, and a higher R^2 indicate a strong ability to analyze the complex factors present in delivery data.

Beyond model accuracy, this study investigated feature importance to understand the key factors of delivery time. Notably, historical average delivery time (avg_time), time of day (hour), and geographic location (lng, lat) emerged as the most influential predictors. 95% prediction intervals for model results were also calculated, providing transparency in prediction uncertainty and offering insight for logistic companies' decision-making.

Although these studies show evident progress, several gaps remain. Many models focus on either route optimization or time estimation, but do not include both in a single framework. While some datasets, such as LaDe and the Amazon challenge dataset, provide valuable resources, there is still limited availability of public, fine-grained delivery operation data. This study also finds that current research lacks prediction intervals and uncertainty estimates in its delivery time forecasting, which is essential for leaving room for error and could improve decision-making.

This work has made the following highlights:

1. Developed a Logistics Random Forest Regression (RFRTP) model for package delivery time prediction, which produced strong results compared to linear and regularized baselines (model predictions achieving a mean absolute error of within 5 minutes, ~8 times more accurate than baselines)

2. Performed a comprehensive analysis on feature importance and found that average past delivery time and geographical location were the most influential features considered when predicting delivery time.

3. Unlike previous works, the proposed approach calculates 95% prediction intervals to quantify model uncertainty, which enables logistics companies to anticipate potential delays and optimize scheduling decisions with room for error.

2. Literature review

2.1. Machine learning and optimization approaches

Statistical machine learning has recently emerged as a focus for last-mile delivery studies. Arıkan et al. combined operations research and machine learning to optimize delivery routes at Deutsche Post [1]; the study resulted in improvements in both efficiency and customer satisfaction. Merchán et al. presented the 2021 Amazon Last Mile Routing Research Challenge dataset [2], which has become a benchmark for testing routing and scheduling algorithms, a crucial problem for package delivery. Mo et al. applied a pair-wise attention-based pointer neural network to predict driver route trajectories, improving the adaptability of delivery planning [3].

DOI: 10.54254/2753-8818/2026.HZ27848

In a related field, Huang et al. proposed a multitask attention network for food delivery time prediction [4], which holds value for logistics as a related task and shows the ability of attention networks to do well in delivery time prediction. Wang et al. proposed the CourIRL method, which uses crossed-attention inverse reinforcement learning to model courier behavior [5]. The study found insights on delivery drivers' habits, which is very important when predicting delivery windows. Wu et al. introduced the LaDe dataset, the first industry-provided last-mile delivery dataset, which supports research on delivery time estimation and route optimization [6].

2.2. Predictive modeling for delivery time estimation

A variety of modeling techniques have been employed when modeling delivery time. Bas et al. explored median dendritic artificial neural networks for robust time series forecasting, creating a method with improved stability in noisy environments [7]. Chen et al. improved travel time prediction using deep learning models that process both recent chronological data, which encapsulates short-term traffic data leading up to the current moment, and historical traffic data reflecting larger trends [8]. Müller et al. reviewed predictive methods for delivery delays in supply chains and identified a need for hybrid models that combine easy interpretability with strong, accurate predictions [9]. Özarık et al. found that integrating domain-specific knowledge with machine learning can lead to better real-world performance [10].

3. Methodology

3.1. Dataset overview

The dataset used in this study, provided by JD Logistics, contains delivery records of including the following variables: timestamp of the order (create_time), unique order identifiers (waybill_code), delivery personnel identifiers (operator_user), delivery geographical coordinates (lng, lat), the road area that the package was delivered to (luqu), the historical average time each courier used to deliver packages (avg_time), and the actual time taken to complete each delivery (total_time). The central objective of this analysis is to model and predict the delivery duration based on available features, with a focus on identifying patterns and statistically significant factors that influence last-mile delivery efficiency. Figures 1, 2, and 3 depict the distributions of delivery time under different filters in the dataset, which are right-skewed with a few remarkably higher outlying points. This study ran code to assess the distribution of the data and concluded that the best-fitting distribution of total_time in the data was a logarithmic normal distribution. This study calculated the dataset to have a mean total_time value of 198.80 minutes, a median total_time of 187.35 minutes, and a total_time standard deviation of 63.50 minutes.

DOI: 10.54254/2753-8818/2026.HZ27848



Figure 1. Delivery time distribution by day of week (picture credit: original)

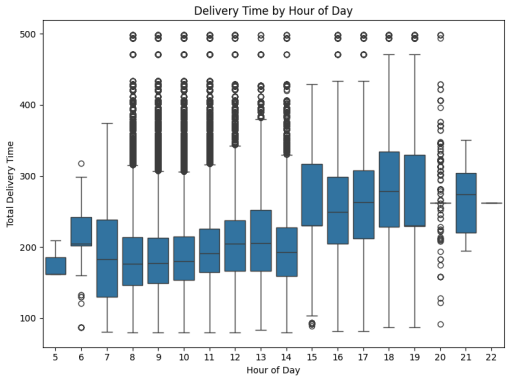


Figure 2. Delivery time distribution by hour of day (picture credit: original)

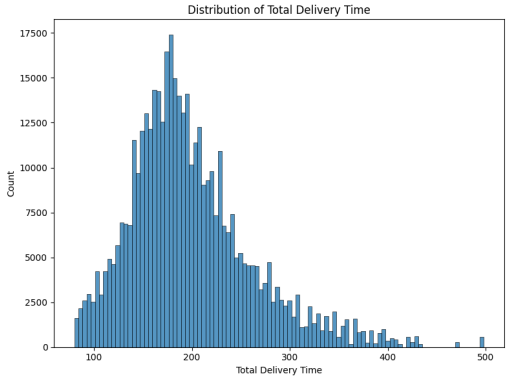


Figure 3. Distribution of total delivery time (picture credit: original)

3.2. Preprocessing and feature engineering

To prepare the data for modeling, this study performed multiple preprocessing and feature engineering steps, extracting important information, as follows:

1. Temporal Features: The time field was converted into a date-time format and decomposed into multiple variables, including hour of day, day of week, and a binary indicator for whether it was a weekend. These were included to capture trends in delivery performance, including traffic congestion, operational scheduling, and consumer behavior.

DOI: 10.54254/2753-8818/2026.HZ27848

2. Driver-Based Features: This study treated the driver ID field as a categorical variable. Since driver performance varies, the models were enabled to capture fixed effects related to specific delivery personnel. Group-level aggregates, such as mean delivery time per driver, were also considered to model experience or efficiency differences.

3. Geographic Features: While the address field was available as text, this study parsed and encoded address components, such as road number or district, where possible, to approximate urban density.

4. Target Variable: The delivery duration (time it took to deliver) was treated as a continuous numerical target variable. The modeling objective was to predict this value based on the other available fields.

3.3. Random Forest Regression for Time Prediction (RFRTP)

To capture non-linear interactions and complex feature relationships, this study implemented and trained the RFRTP model. This method is robust to outliers and does not assume linearity, suggesting that it is highly suitable for complex delivery datasets where patterns may vary depending on driver, traffic, and geography.

The model was implemented in Python. The model has four main parameters that this study explored, namely n_estimators (number of trees, default: 100), max_depth (maximum depth of a tree, default: ∞), min_samples_leaf (controls whether a node should be split or be a leaf; higher = simpler trees, default: 1), and max_features (the number of features that each tree considers at a split; lower means more randomness, default: auto). A fifth parameter, random_state, was set at 42 for all cases. This study tested multiple sets of parameters, which obtained different results as detailed in the table 1:

Table 1. RFRTP Performance with different parameters

Model Parameters	MAE	RMSE	R^2
Default	4.378837	13.713389	0.953004
n_estimators = 300	4.357127	13.650291	0.953436
max_depth = 10	33.523161	45.793978	0.475933
min_samples_leaf = 5	8.156300	17.803475	0.920790
max_features = sqrt	5.815369	15.109786	0.942946

This study found that changing parameters besides n_estimators did not lead to significant improvements beyond the default parameters. Although increasing the number of trees did lead to very slight performance increases, training time was also increased, so this study concludes that any larger n_estimators are suboptimal from a comprehensive perspective. During implementation, feature importance scores were extracted to gain insights into the variables that contributed most to the prediction.

3.4. Evaluation metrics

Models were evaluated on a separate test set. Evaluation was performed using performance metrics as detailed below:

Mean Absolute Error (MAE): Represents the average magnitude of prediction errors, useful for direct interpretation of accuracy (e.g., average delivery time predictions were off by y minutes).

DOI: 10.54254/2753-8818/2026.HZ27848

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

Where: $n =$ number of observations, $y_i =$ actual time in minutes, $\hat{y}_i =$ predicted time in minutes.

Root Mean Squared Error (RMSE): Penalizes larger errors more heavily and reflects variance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

R-Squared (R^2): Indicates the proportion of variance explained by the model's predictions; R^2 is helpful as a measure of how well the model fits the dataset.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{n=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{3}$$

Where: $\bar{y} =$ mean of actual time in minutes.

Prediction Intervals: For the RFRTP model, this study computed 95% prediction intervals to quantify uncertainty in predicted delivery times.

$$PI_{95\%} = \bar{y} \pm z_{0.975} \bullet s \bullet \sqrt{1 + \frac{1}{n}} \tag{4}$$

Where: $\bar{y} =$ sample mean in minutes, $s =$ sample standard deviation.

3.5. Implementation details

The model and analysis were implemented in Python. Libraries including pandas and np were used for data manipulation, and scikit-learn was used for model training and evaluation. Cross-validation was used during implementation to avoid overfitting, and all preprocessing steps were ensured to be reproducible.

4. Results

4.1. Baseline models

After obtaining structured data from preprocessing, this study trained and implemented multiple linear baseline models alongside RFRTP on the provided dataset, detailed below:

1. Linear Regression: As a baseline model, multiple linear regression was employed to identify linear relationships between features and delivery time. This linear approach allowed for a transparent view into the significance and direction of influence for each predictor in the dataset. During implementation, features were standardized, and variance inflation factors (VIFs) were used to assess multicollinearity and ensure model stability.

2. Lasso and Ridge Regression: Recognizing that potential problems such as overfitting and high dimensionality (particularly due to one-hot encoding of categorical variables) might occur in basic linear models, this study applied Lasso and Ridge regression models as baselines. These two models introduce regularization terms that penalize significant coefficients, promoting model sparsity (Lasso) or shrinkage (Ridge).

DOI: 10.54254/2753-8818/2026.HZ27848

4.2. Model performance

After dataset preprocessing and model training were completed, this study evaluated the predictive performance of four methods as previously detailed: Linear Regression, Ridge Regression, Lasso Regression, and RFRTP. The models were assessed using three performance metrics: MAE, RMSE, and R^2. Through these metrics, this study gained insight into both the accuracy and the goodness of fit of the models, through both the highest and the overall performances (mean and standard deviation), as displayed in Tables 2 and Table 3.

Table 2. Highest performance per model

Model	Minimum MAE	Minimum RMSE	Maximum R^2
Linear Regression	44.401	58.386	0.149
Ridge Regression	44.400	58.386	0.149
Lasso Regression	44.568	58.762	0.137
RFRTP	4.339	13.417	0.955

Table 3. Model performance

Model	Mean MAE	MAE SD	Mean RMSE	RMSE SD	Mean R^2
Linear Regression	44.478	0.054	58.559	18.871	0.145
Ridge Regression	44.478	0.054	58.560	18.835	0.145
Lasso Regression	44.666	0.068	58.940	19.256	0.134
RFRTP	4.389	0.035	13.624	4.719	0.954

As reflected in the table, RFRTP performed the best, with the highest R^2 and the lowest MAE. Random Forest achieved an MAE of 4.38, an RMSE of 13.71, and an R^2 score of 0.953. These results show that RFRTP provides the most accurate delivery time predictions among the tested models. Lasso Regression had the worst performance, with an MAE of 44.68. This reflects Lasso's inability to capture the non-linearity in the dataset.

4.3. Feature importance

To further evaluate the performance of the RFRTP model, this study calculated and examined feature importance scores, which measure the impact of each feature in the dataset on the model's predictions and sum to 1. The most contributing features were as follows:

1. avg_time: 0.378121
2. hour: 0.150111
3. lat: 0.120745

The avg_time feature, scoring 0.378121, was the most influential. This shows that the historical average time taken by a driver to complete a delivery plays a significant role in predicting future delivery times. A higher avg_time implies slower deliveries, which is a crucial factor to consider when predicting delivery times.

DOI: 10.54254/2753-8818/2026.HZ27848

4.4. Prediction intervals and model uncertainty

To quantify uncertainty in delivery time predictions, this study employed a bootstrapping approach that trained 1000 Random Forest models on resampled subsets of the training data. For each test instance, the 2.5th and 97.5th percentiles of the resulting predictions were used to construct a 95% prediction interval. Upper and lower bounds for the interval were calculated, with a lower bound of 103.94 and an upper bound of 354.82. Therefore, for 95% of predictions, the actual delivery time is within 103.94 minutes and 354.82 minutes. This uncertainty can help logistics companies while planning for possible delays and optimizing delivery schedules.

4.5. Model diagnostics and results

Residual diagnostics were performed for the RFRTP model to assess model fit. The residual plot shows a scatter of points around zero with no clear pattern, suggesting that the model performs well on the majority of the data but may struggle with outliers. Since the mean of residuals was close to 0, the model is overall unbiased.

4.6. Overview of results

The RFRTP model outperforms linear models in terms of accuracy and predictive power. This is likely due to the non-linearity of the delivery time data, which linear models cannot fully capture. It achieved an MAE of under 5, meaning that on average, it was able to predict delivery time to within 5 minutes of the actual time.

Through feature importance analysis, this study found that avg_time plays a crucial role in predicting delivery times, suggesting that a driver's historical average delivery duration strongly impacts predictions of future performance. This is likely due to avg_time reflecting both personal efficiency and consistent route conditions.

5. Discussion

5.1. Model diagnostics and residual analysis

Residuals from the RFRTP model showed no clear pattern, so the model did not suffer from bias or overfitting. However, a few outliers were observed, where deliveries took significantly longer than predicted; these outliers represent rare delays that are difficult to predict. From a comprehensive view, these outliers hold little meaning because there is always inherent uncertainty in delivery times that no model can fully capture.

The RFRTP model outperformed the linear baseline models due to its ability to capture non-linearity and handle feature interactions. These advantages make RFRTP especially suitable for complicated traffic situations, while linear models struggle to grasp the connections between features.

5.2. Feature insights and real-world implications

The feature important analysis revealed that historical delivery time (via avg_time) and time of day (via hour) are the most significant predictors of delivery performance. When predicting delivery times, logistics providers should track driver efficiency and consider time-of-day traffic patterns, as both have a substantial amount of variance. Geographic location (latitude) also proved crucial, as

DOI: 10.54254/2753-8818/2026.HZ27848

traffic conditions vary between locations. This highlights the importance of systems that consider area- and time-specific traffic conditions, such as the work of Lin et al. in 2023, which utilized transformer networks to estimate the time of arrival while taking local traffic congestion data into account [11].

5.3. Limitations and future work

Given the results of this study, there is still room for improvement. Future work should focus on, firstly, improving models' performance when faced with outliers, and secondly, providing more accurate prediction windows. This study proposes that future work incorporate real-time data on traffic congestion and weather conditions into models and further research deep learning methods and multi-objective optimization to improve their performance when facing outliers. The use of real-time data on traffic conditions and weather may be incorporated through the connection of this model to company server systems, providing crucial and timely data on the variables that affect delivery time.

6. Conclusion

This study explored the application of statistical and machine learning methods to predict delivery times in last-mile logistics. Using a dataset containing delivery records that featured order details, courier identifiers, geographical coordinates, and average processing times, this study evaluated multiple statistical and machine learning models: Linear Regression, Ridge Regression, Lasso Regression, and RFRTP.

In implementation, the RFRTP model showed a clear lead in performance, achieving the lowest MAE and RMSE as well as the highest R^2 score. Therefore, it is concluded that the Random Forest model is best able to capture the complex, nonlinear relationships between features present in the data. Feature importance analysis revealed that average delivery time, hour of day, and geographical position were the most influential predictors, since both operational efficiency and temporal factors significantly affect delivery times.

This study also calculated 95% prediction intervals to improve the reliability of the Random Forest model's results. The quantified uncertainty that it provides is greatly valuable for logistics planners' scheduling and planning. Contrarily, the model cannot accurately predict certain outlying cases, as the variability in the prediction intervals still shows. These outliers likely involve unpredictable human behavior or traffic conditions, which are difficult for this study's model to anticipate.

The findings of this study showcase the potential of statistical and machine learning modeling, especially RFRTP, in the field of logistics; models trained on real-world datasets can improve delivery time predictions and raise efficiency. Accurate predictions are not only crucial for customer satisfaction, but they also allow companies to allocate resources effectively and reduce operational costs. As work in this direction carries on, the integration of real-time data and the development of stronger and well-trained models can further improve the accuracy and reliability of predictions.

References

[1] Arıkan, U., Kranz, T., Sal, B. C., Schmitt, S., & Witt, J. (2023). Human-Centric Parcel Delivery at Deutsche Post with Operations Research and Machine Learning. INFORMS Journal on Applied Analytics, 53(5), 359–371. https://doi.org/10.1287/inte.2023.0031.

DOI: 10.54254/2753-8818/2026.HZ27848

[2] D. Merchán, J. Arora, J. Pachon, K. Konduri, M. Winkenbach, S. Parks, and J. Noszek, “2021 Amazon Last Mile Routing Research Challenge: Data Set, ” Transportation Science, vol. 58, no. 1, pp. 65–77, 2022, doi: 10.1287/trsc.2022.1173.

[3] Mo, B., Wang, Q. Y., Guo, X., Winkenbach, M., & Zhao, J. (2023). Predicting Drivers’ Route Trajectories in Last-Mile Delivery Using A Pair-wise Attention-based Pointer Neural Network. ArXiv.org. https: //arxiv.org/abs/2301.03802.

[4] F. Huang, W. Jiang, and S. Chen, “A Multitask Attention Network for Food Delivery Time Prediction, ” Journal of Circuits, Systems, and Computers, vol. 33, no. 24, pp. 2450025-1–2450025-18, 2023, doi: 10.1142/S0218126624500257.

[5] Wang, S., Kong, T., Guo, B., Lin, L., & Wang, H. (2024). CourIRL: Predicting Couriers’ Behavior in Last-Mile Delivery Using Crossed-Attention Inverse Reinforcement Learning. CIKM ’24: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, 4957–4965. https: //doi.org/10.1145/3627673.368004.

[6] L. Wu, H. Wen, H. Hu, X. Mao, Y. Xia, E. Shan, J. Zhen, J. Lou, Y. Liang, L. Yang, R. Zimmermann, Y. Lin, and H. Wan, “LaDe: The First Comprehensive Last-Mile Delivery Dataset from Industry, ” arXiv preprint, arXiv: 2306.10675, 2023, doi: 10.48550/arXiv.2306.10675.

[7] Bas, E., Egrioglu, E., & Cansu, T. (2023). Robust training of median dendritic artificial neural networks for time series forecasting. Expert Systems with Applications, 238(Part C), 122080. https: //doi.org/10.1016/j.eswa.2023.122080.

[8] C. Chen, E. W. Sun, M.-F. Chang, and Y. Lin, “Enhancing Travel Time Prediction with Deep Learning on Chronological and Retrospective Time Order Information of Big Traffic Data, ” Annals of Operations Research, vol. 343, pp. 115–137, 2023, doi: 10.1007/s10479-023-05223-7.

[9] Müller, N., Burggräf, P., Steinberg, F., Sauer, C. R., & Schütz, M. (2025). An analytical review of predictive methods for delivery delays in supply chains. Supply Chain Analytics, 11, 100130. https: //doi.org/10.1016/j.sca.2025.100130.

[10] S. S. Özarık, P. da Costa, and A. M. Florio, “Machine Learning for Data-Driven Last-Mile Delivery Optimization, ” Transportation Science, vol. 58, no. 1, pp. 78–92, 2023, doi: 10.1287/trsc.2022.0029.

[11] Lin, S., Xu, Y., Zhao, S., Wang, Y., & Xu, J. (2023). TransETA: transformer networks for estimated time of arrival with local congestion representation. Applied Intelligence, 53(24), 30384–30399. https: //doi.org/10.1007/s10489-023-05139-6.