2025-11

Predicting on-time deliveries in e-commerce: a machine learning approach to shipment performance analysis

Rahman, Md. Sohanur

School of Business and Entrepreneurship, Independent University, Bangladesh (IUB)

Predicting On-Time Deliveries in E-Commerce: A Machine Learning Approach to Shipment Performance Analysis

Md. Sohanur Rahman, Shumon Kumar Ray
Department of Industrial & Production Engineering, Rajshahi University of Engineering & Technology
Department of Industrial & Production Engineering,Shahjalal University of Science and Technology
[1]*rahman.ipe08@gmail.com;* [2]*shumon.ipe@gmail.com*

PAPER ID : ICEBTM-25-1294

Abstract - **Timely delivery is a critical driver of customer satisfaction and loyalty in the competitive e-commerce sector. This study applies machine learning techniques to predict shipment punctuality using a dataset of 10,999 delivery records with 12 features covering customer behavior, product characteristics, and logistics details. The target variable, Reached on Time, indicates whether shipments were delayed. Through exploratory data analysis, we identify key delivery factors, including the influence of shipment weight, product cost, and discounts. Four supervised models—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—are developed and evaluated using accuracy, F1-score, and AUC-ROC metrics. Results show that ensemble methods, particularly Random Forest, outperform traditional models in predicting late deliveries. The findings highlight the value of data-driven approaches in improving shipment reliability and provide actionable insights for optimizing logistics strategies and enhancing customer satisfaction.**

Keywords: **E-Commerce, On-Time Delivery, Shipment Prediction, Machine Learning, Logistics Optimization.**

I.INTRODUCTION

In the dynamic landscape of e-commerce, timely delivery has become a cornerstone of customer satisfaction and business success. As the volume and complexity of online orders grow, logistics systems are under increasing pressure to deliver efficiently and predictably. Late deliveries not only erode customer trust but also contribute to increased operational costs and negative brand perception. Given these stakes, there is a pressing need for data-driven strategies that can anticipate shipment delays and support proactive logistics management.

Recent research underscores the transformative potential of machine learning (ML) in understanding and optimizing delivery performance. [1] highlight how ML models can uncover hidden patterns in customer behavior and satisfaction, suggesting that predictive analytics can directly enhance service quality in e-commerce. Extending beyond prediction, [2]emphasize the shift from predictive to prescriptive analytics, arguing that data science must not only anticipate outcomes but also inform decision-making to optimize operations.

In the logistics domain, ML has shown promising applications in last-mile delivery optimization. [4] developed models to predict late deliveries using real-time data, showing how early alerts can mitigate service disruptions. Similarly, [5] review multiple use cases where ML improves forecasting accuracy and resource allocation in e-commerce supply chains. These works affirm the utility of supervised learning algorithms in classifying on-time versus delayed shipments.

Additional contributions further validate this perspective. [3] use simulation-based approaches to explore mismatched peer-to-peer delivery networks, a concept directly related to shipment timing inconsistencies. [9] take a more diagnostic view, analyzing real-world e-commerce delivery failures through historical data. Their findings highlight the value of predictive modeling in understanding why delays occur. Moreover, [7] introduced Random Forests—a model proven effective in logistics classification tasks—while [8] addressed class imbalance issues common in delivery datasets through the SMOTE technique.

After a handsome amount of literature review authors bridge the gap between theoretical ML applications and practical e-commerce logistics, offering a data-driven framework to improve delivery reliability—a critical need in today's competitive online retail landscape. The exposition of the paper is as follows: section 2 methodology; Section 3 problem formulation and Modeling the Problem; section 4 model evaluation; section 5 model limitation and finally section 6 conclusion.

II.METHODOLOGY

This study employs a structured machine learning workflow to predict shipment punctuality in e-commerce logistics. The methodological framework consists of six sequential stages: dataset description, data cleaning, preprocessing, exploratory data analysis (EDA), feature engineering, model development, and performance evaluation. Each stage is designed to ensure data quality, model robustness, and reproducibility.

1) *Dataset Description:* The dataset, sourced from an international e-commerce company, comprises 10,999 shipment records with 12 attributes spanning customer interactions, product characteristics, and logistics details. The target variable, Reached on Time (0 = on time, 1 = late), frames the study as a binary classification problem. Table 1 categorizes features into numerical, categorical, and target variables.

ISBN 978-984-35-5270-9 *https://icebtm.iub.edu.bd*

Table 1
Classifications of Attributes

Numerical Features	Categorical Features	Target Variable
Customer_care_calls Customer_rating Cost_of_the_Product Prior_purchases Discount_offered Weight_in_gms	Warehouse_block Mode_of_Shipment Product_importance Gender	Reachedno n.Time Y.N (0 = on time, 1 = late)

2) *Data Cleaning*: Rigorous data pre-processing was undertaken to ensure the integrity and reliability of the dataset. Redundant attributes, such as unique identifiers (ID Column) that do not contribute to predictive modeling, were systematically removed to reduce dimensionality. Comprehensive checks for missing values and duplicate records were conducted, confirming the completeness and uniqueness of the data. Outliers in key numerical variables—such as *Prior Purchases* and *Discount Offered*—were detected using the Interquartile Range (IQR) method. These anomalies were carefully addressed to mitigate their potential impact on model performance and to ensure that the learning process remained unbiased and representative of the underlying data distribution.

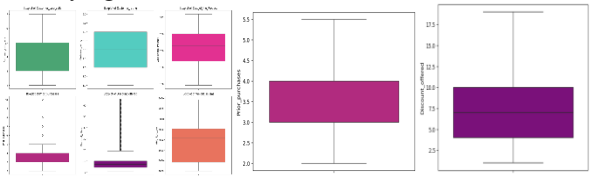

Fig.01. Outlier Detection and Handling

3) *Data Pre-Processing*: To ensure the dataset was suitable for machine learning algorithms, several preprocessing steps were applied. Categorical variables were encoded to facilitate numerical analysis. Specifically, nominal variables such as Warehouse Block, Mode of Shipment, and Gender were transformed using One-Hot Encoding, allowing the model to treat each category as an independent binary feature. For the ordinal variable Product Importance, Label Encoding was employed to preserve the inherent order among categories. Additionally, numerical features were standardized using StandardScaler, a technique that normalizes data by removing the mean and scaling to unit variance. This step is particularly critical for algorithms sensitive to feature scales, such as Logistic Regression and Support Vector Machines (SVM), as it ensures all features contribute proportionally to the model's performance.

4) *Exploratory Data Analysis (EDA):* Exploratory data analysis (EDA) was performed to uncover preliminary patterns and investigate relationships between features and the target variable. Univariate analysis, including distribution plots and frequency counts, was employed to assess the balance and variability of individual variables. Bivariate analysis was conducted to evaluate the influence of key features—such as shipment weight and product cost—on delivery outcomes, particularly punctuality. Additionally, correlation analysis identified moderate associations between certain features and delivery status, with discount rate emerging as a notable factor. The insights derived from EDA played a critical role in guiding feature selection and prioritization during subsequent model development phases.

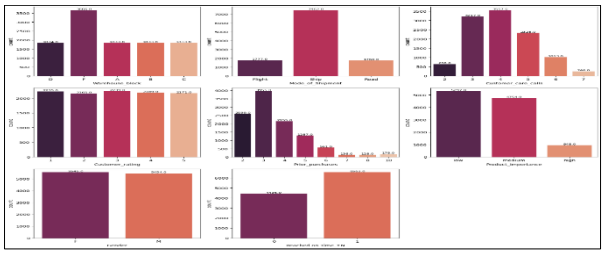

Fig.2. Analyze each data with respect to their Frequency

The analysis reveals several key insights into the current logistics and customer service operations. Warehouse F Block emerges as a critical hub, handling a significantly higher volume of shipments compared to other locations. Among the various transportation options, sea freight is identified as the most utilized mode, indicating a potential preference or cost-efficiency in maritime logistics. Customer engagement data shows that most shipments are associated with four customer care interactions, suggesting a moderate level of service requirement during the delivery process. Notably, customer satisfaction tends to cluster around a rating of 3, pointing to a generally average service perception that may warrant further improvement. The dataset also indicates that customers typically have a maximum of three prior purchases, reflecting either a relatively new customer base or low repeat purchase behavior. In terms of product characteristics, the inventory is predominantly composed of items classified as low or medium in importance, which may influence customer expectations and urgency. Demographic analysis highlights a well-balanced gender distribution among customers. However, a significant concern is that a higher percentage of products fail to reach customers on time, underscoring a pressing need for improvements in delivery efficiency and supply chain reliability.

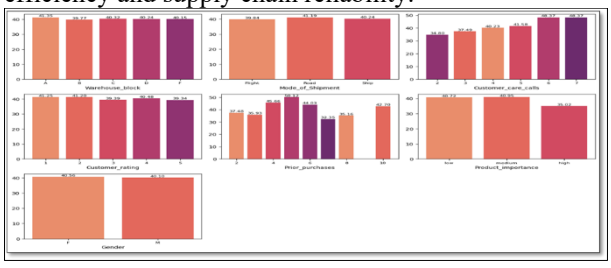

Fig.3. Category vs Impact on Delivery
[result show in percentage]

The analysis reveals that all warehouse blocks exhibit on-time delivery rates below 50%, indicating systemic inefficiencies

ISBN 978-984-35-5270-9

across the entire logistics network. Furthermore, all modes of transportation show relatively low delivery performance, suggesting that transportation infrastructure or operational processes may be contributing factors. A clear correlation is observed between the volume of customer care calls and on-time delivery rates, implying that delivery delays significantly impact on customer satisfaction and drive support inquiries. Interestingly, customer ratings do not appear to have a statistically significant relationship with on-time delivery, which may point to other factors influencing customer perceptions. The data also shows a notable improvement in on-time delivery percentages between categories 2 and 5, marking a performance peak that warrants further investigation. Additionally, products classified as low and medium in importance exhibit the highest on-time delivery rates, raising questions about prioritization strategies within the supply chain. Finally, the analysis finds no significant correlation between customer gender and on-time delivery performance, indicating that delivery outcomes are consistent across demographic lines.

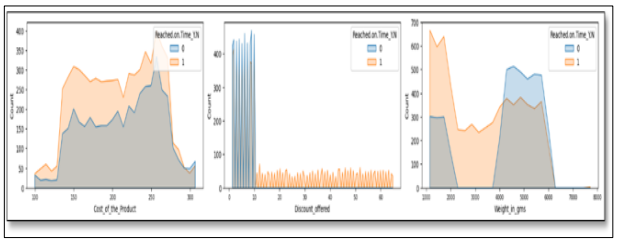

Fig.4. Finding any impact on 'Delivery on time' by considering Cost, Discount offer or Weight of products

The analysis reveals that most products are priced within the $201–$275 range, indicating a central tendency in product pricing. Interestingly, the highest-priced products demonstrate the best on-time delivery performance; however, their delivery reliability remains below 50%, suggesting potential operational inefficiencies even at premium price points. In terms of discounting, most products fall within the 1%–10% discount bracket, where on-time delivery performance maintains a relatively balanced and acceptable level. Regarding product weight, the highest concentrations are found in the 1,000–2,000 grams and 4,000–6,000 grams categories. Notably, products in the 4,000–6,000 grams range show significantly better on-time delivery performance, exceeding 56%, highlighting a possible correlation between product weight and logistics efficiency.

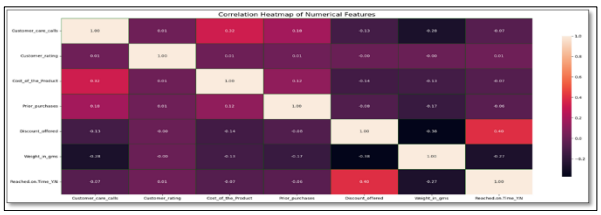

Fig.5. Detecting any Correlations between the features

The analysis revealed no strong correlations among the examined features. However, discount offers demonstrated a moderate relationship with timely delivery, accounting for approximately 40% of the variation. Additionally, customer care calls showed a moderate association with both the cost of the product (32%) and prior purchase behavior (18%). These findings suggest that while most variables operate independently, targeted incentives and customer engagement efforts may influence key outcomes such as delivery punctuality and purchasing patterns.

III.BUILDING MACHINE LEARNING MODELS

The primary objective of this phase is to develop and evaluate machine learning models capable of accurately predicting whether a shipment will be delivered on time. Given the binary nature of the target variable (Reached.on.Time_Y.N), the problem is formulated as a supervised binary classification task. To ensure a comprehensive comparison of predictive performance, four widely used classification algorithms were selected: Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM).

1) *Data Splitting:* To assess model generalization and avoid overfitting, the preprocessed dataset was partitioned into two subsets: 80% of the data was allocated to the training set, while the remaining 20% was reserved for testing. A fixed random seed was used to ensure reproducibility of the data split. This approach allows for a fair evaluation of each model's performance on previously unseen instances, ensuring reliable estimation of real-world effectiveness.

2) *Feature Scaling:* Feature scaling was applied to numerical variables using the StandardScaler method, which standardizes features by removing the mean and scaling to unit variance (i.e., a mean of 0 and a standard deviation of 1). This step is particularly critical for algorithms like Logistic Regression and SVM, which are sensitive to the scale of input features and may otherwise yield biased or suboptimal results.

3) *Underlying Algorithms:* Each algorithm was initially trained using default hyperparameters, followed by basic hyperparameter tuning to improve performance. The models, Logistic Regression, Decision Tree Classifier, Random Forest Classifier and Support Vector Machine (SVM), were implemented using Scikit-learn, a robust and widely adopted Python-based machine learning library. Performance was evaluated on both training and test sets using the following metrics: Accuracy (overall correctness), Precision (proportion of predicted late deliveries that were actually late), Recall (the model's ability to identify actual late deliveries), and ROC AUC Score (the model's discriminative power across classification thresholds).

Given the business context—where undetected late deliveries can negatively impact customer satisfaction—Recall and ROC AUC Score were prioritized as key performance indicators. Among the

ISBN 978-984-35-5270-9 https://icebtm.iub.edu.bd

evaluated models, the Random Forest Classifier emerged as the most effective, achieving the highest overall performance across all metrics and demonstrating superior generalization capabilities. This makes it a strong candidate for deployment in real-world shipment delivery prediction systems. By leveraging ensemble learning, Random Forest effectively balances bias and variance, reducing the risk of overfitting. Moreover, its interpretability through feature importance analysis offers actionable insights into the key drivers of late deliveries. Deploying such a model not only enhances predictive accuracy but also empowers management with data-driven strategies to optimize logistics operations and improve service reliability.

IV.MODEL EVALUATION

To rigorously evaluate the predictive models developed in this study, a comprehensive and methodologically robust assessment framework was implemented, designed to capture both overall performance and the nuanced implications of classification errors in a real-world business context. Given the significant asymmetric cost associated with false negatives—specifically, the failure to accurately identify late shipments—evaluation efforts prioritized not only aggregate predictive accuracy but also detailed class-specific performance metrics. The analysis employed a set of canonical classification measures to assess model efficacy on an independent test set. These included Accuracy, quantifying the overall proportion of correct classifications; Precision, reflecting the proportion of predicted late shipments that were truly late, thus measuring the reliability of positive predictions; Recall (Sensitivity), which denotes the ability of the model to identify the full complement of actual late shipments; and the F1 Score, the harmonic mean of precision and recall, serving as a balanced metric particularly valuable in contexts characterized by class imbalance. In addition, the Receiver Operating Characteristic Area Under the Curve (ROC AUC) was utilized as a threshold-agnostic performance indicator, offering a robust measure of the model's discriminatory power across varying classification thresholds. Collectively, this rigorous evaluation paradigm ensures that the predictive models deliver not only statistically sound performance, but also practical utility aligned with the operational priorities and risk management imperatives inherent to late shipment detection.

V. RESULTS ASSESSMENT

This study evaluates four distinct machine learning classifiers—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)—using multiple performance metrics on both training and testing datasets. The metrics considered include Accuracy, Precision, Recall, F1 Score, and ROC AUC Score, providing a comprehensive overview of model effectiveness in classification tasks.

Table 2

ML Models Result Comparison

ML_Models	Dataset	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Logistic Regression	Training	63.65%	69.38%	70.12%	69.75%	62.09%
	Testing	63.14%	68.03%	71.42%	69.68%	61.24%
Decision Tree	Training	100%	100%	100%	100%	100%
	Testing	64.14%	69.43%	70.65%	70.03%	62.64%
Random Forest	Training	100%	100%	100%	100%	100%
	Testing	65.32%	73.81%	64.37%	68.77%	65.54%
SVM (Support Vector Machine)	Training	70.55%	86.95%	59.68%	70.78%	73.19%
	Testing	65.77%	81.08%	55.17%	65.66%	68.20%

1) *Logistic Regression:* Logistic Regression exhibits moderate and consistent performance across both training and testing datasets. The training accuracy stands at 63.65%, with testing accuracy closely following at 63.14%, indicating the model generalizes reasonably well without significant overfitting. Precision and recall metrics are balanced, with precision slightly higher in training (69.38%) compared to testing (68.03%), and recall marginally improving from 70.12% in training to 71.42% in testing. The F1 score remains steady (~69.7%), while the ROC AUC score decreases slightly from 62.09% to 61.24% in testing. These results suggest Logistic Regression provides stable but moderate predictive capability, making it a reliable baseline model.

2) *Decision Tree:* The Decision Tree model achieves perfect scores (100%) on all metrics during training, a classic indication of overfitting. Such overfitting is further confirmed by the significant drop in testing performance, with accuracy plummeting to 64.14%. Although testing precision (69.43%), recall (70.65%), and F1 score (70.03%) are somewhat improved compared to Logistic Regression, the gap between training and testing metrics signals poor generalization. The ROC AUC on testing remains low at 62.64%, reinforcing the model's tendency to memorize training data rather than learning robust patterns.

3) *Random Forest:* Random Forest also exhibits perfect training metrics, indicating overfitting like the Decision Tree. However, its testing performance is notably better than that of the Decision Tree and Logistic Regression, achieving the highest testing accuracy (65.32%) and ROC AUC score (65.54%) among all models tested. Precision on testing is significantly higher (73.81%), though recall decreases to 64.37%, suggesting the model is better at correctly identifying positive predictions but misses some true positives. The F1 score (68.77%) reflects this balance. These results highlight Random Forest's enhanced robustness and improved generalization compared to single decision trees, despite some overfitting during training.

4) *Support Vector Machine (SVM):* SVM shows a different pattern with moderate training accuracy (70.55%) but a notable drop in recall (59.68%) despite high precision (86.95%). This suggests the model is highly precise but misses a substantial portion of true positive cases. On testing, the accuracy (65.77%) is comparable to Random Forest, with precision slightly reduced to 81.08% and recall further decreased to 55.17%. The F1 score drops accordingly to 65.66%, and the ROC AUC score decreases from 73.19% in training to 68.20% in testing, still the highest

ISBN 978-984-35-5270-9

ROC AUC among the tested models. This pattern implies that while SVM is conservative in positive classification (high precision), it may be less sensitive in detecting all relevant instances (low recall), which is critical depending on application context.

● MODEL SELECTION AND RECOMMENDATIONS

Based on the comparative performance analysis of the four machine learning models—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM)—this section presents the rationale for model selection and the final recommendation.

The selection process considered five key performance metrics across both training and testing datasets: Accuracy, Precision, Recall, F1 Score, and ROC AUC Score. These metrics were evaluated to balance the trade-offs between prediction correctness, class imbalance handling, and generalization capability.

Random Forest emerged as the most promising model overall. Despite achieving perfect scores on the training dataset—an indicator of potential overfitting—it delivered the highest testing accuracy (65.32%), precision (73.81%), and a strong ROC AUC score (65.54%). Its F1 score (68.77%) also reflects a well-balanced trade-off between precision and recall. Compared to the single Decision Tree, Random Forest's ensemble approach significantly improved generalization and robustness, mitigating some of the overfitting effects.

Support Vector Machine (SVM) demonstrated the highest ROC AUC score (68.20%) on the test data, suggesting superior class separation capability. Additionally, it maintained high precision (81.08%), making it suitable for scenarios where false positives carry a higher cost. However, the notably lower recall (55.17%) indicates that the model may miss a substantial number of positive cases, which could be detrimental in recall-sensitive applications.

Logistic Regression, while not outperforming the other models in any individual metric, showed consistent and balanced performance between training and testing sets. Its minimal overfitting, decent F1 score (~69.7%), and stability make it a reliable baseline, particularly for simpler applications where interpretability and consistency are prioritized over marginal performance gains.

Decision Tree, in contrast, is not recommended for deployment due to severe overfitting. While it attained perfect scores on the training set, its testing performance did not translate proportionally, with accuracy (64.14%) and ROC AUC (62.64%) lagging behind the ensemble and kernel-based methods.

Final Recommendation: considering the balance between predictive performance, generalization ability, and applicability to real-world data, the Random Forest classifier is recommended as the most suitable model for the given classification task. It offers the best combination of accuracy, precision, and F1 score on unseen data, indicating its capability to handle complex feature interactions and noise in the dataset.

1) *Support Vector Machine (SVM):* The selected predictive model introduces transformative capabilities within logistics and supply chain operations by enhancing decision-making through data-driven insights. Firstly, it enables proactive intervention by allowing operations teams to identify and address high-risk shipments before disruptions occur, thereby minimizing delays and their downstream impacts. Secondly, it drives resource optimization by informing the dynamic reallocation of logistical assets—such as personnel, transportation, and warehousing—to prioritize vulnerable or time-sensitive orders, ultimately improving operational efficiency. Most critically, the model significantly enhances customer satisfaction by ensuring greater consistency and reliability in delivery performance, which contributes to improved client retention, elevated service quality, and strengthened brand credibility. Taken together, these capabilities position the model as a strategic asset for achieving resilient, adaptive, and customer-centric supply chain systems, aligning with contemporary demands for agility and service excellence in global logistics networks.

1) *Limitations:* Despite encouraging results, this study has several important limitations. The dataset omitted key real-world factors such as route distance, carrier characteristics, seasonal demand, and external disruptions like weather, which limits the model's predictive accuracy. Data quality concerns and inconsistent labeling of "on-time delivery" further challenge the reliability of the target variable. Methodologically, the use of random train–test splits fail to capture the temporal dynamics of shipment data, potentially leading to overly optimistic performance estimates. Additionally, limited hyperparameter tuning and the exclusion of advanced machine learning algorithms like XGBoost constrained the models' full potential. The evaluation framework did not address class imbalance or incorporate cost-sensitive metrics, reducing practical business relevance. Moreover, interpretability was limited to basic feature importance without robust explainability tools, and deployment considerations were not discussed. Addressing these gaps in future work is essential to enhance model robustness, fairness, and real-world applicability in predicting shipment delays.

VII. CONCLUSION

In this study, we developed and rigorously evaluated machine learning models to predict on-time delivery performance within an e-commerce context. Among the tested models, the Random Forest Classifier emerged as the most effective, demonstrating superior predictive capabilities characterized by high recall and ROC AUC scores. These metrics are particularly critical for minimizing the occurrence of undetected late deliveries, thereby safeguarding customer satisfaction and trust. Our findings underscore the potential of data-driven approaches to

ISBN 978-984-35-5270-9 https://icebtm.iub.edu.bd

revolutionize logistics decision-making by accurately identifying shipments at risk of delay, enabling timely and targeted interventions.

Moreover, the study highlights the essential role of comprehensive feature engineering, stringent validation procedures, and cost-sensitive evaluation frameworks in operational environments. By incorporating additional contextual variables, leveraging more advanced algorithms, and integrating interpretability tools, future models can achieve enhanced predictive accuracy and provide more actionable insights. Overall, this research confirms the feasibility and tangible benefits of employing machine learning for shipment performance prediction. It lays a robust foundation for subsequent efforts aimed at optimizing supply chain efficiency, elevating customer experience, and driving improved business outcomes through data-informed logistics management.

REFERENCES

[1] Ghosh, D., & Kankanhalli, A. (2018). Understanding customer satisfaction in e-commerce using machine learning. Information Systems Research, 29(2), 383–404. https://doi.org/10.1287/isre .2018. 0777

[2] Bertsimas, D., Kallus, N., & Weinstein, A. M. (2016). From predictive to prescriptive analytics. Management Science, 62(3), 843–865. https://doi.org/10.1287/mnsc.2015.2176

[3] Huber, N., & Spinler, S. (2018). Mismatched peer-to-peer networks in last-mile delivery: A data-driven simulation approach. International Journal of Physical Distribution & Logistics Management, 48(3), 308–329.

[4] Nguyen, T. T., Nguyen, H. T., & Vo, H. M. (2020). Predicting late delivery in last-mile logistics using machine learning. Procedia Computer Science, 176, 2452–2461. https://doi.org/10.1016/j.procs. 2020.09.297

[5] Xu, Y., & Zhang, X. (2021). Logistics and supply chain predictive analytics: Machine learning applications in e-commerce. Journal of Business Logistics, 42(4), 435–452. https://doi.org/10.1111/jbl. 12270

[6] Kaur, H., & Singh, S. P. (2017). Heuristic modeling for e-commerce logistics optimization. Computers & Industrial Engineering, 112, 762–774.

[7] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A: 10109334 04324

[8] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

[9] Singh, A., & Yassine, A. (2020). Data-driven analysis of e-commerce delivery failures. Logistics Research, 13(1), 1–12. https://doi.org/10.23773/2020.13.1

[10] Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. Journal of Business Logistics, 34(2), 77–84.

ISBN 978-984-35-5270-9 *https://icebtm.iub.edu.bd*