Grenze International Journal of Engineering and Technology, January Issue

Freight Value Prediction Model

Dr. Priyadarshan Dhabe[1], Aryan Dhole[2], Dhruva Sharma[3], Divyansh Pandey[4], Pranjali Diwan[5] and Shashwat Diwan[6]
[1]Professor, Department of Information Technology, Vishwakarma Institute of Technology, Pune, India
Email: priyadarshan.dhabe@vit.edu
[2-6]Student, Department of Information Technology, Vishwakarma Institute of Technology, Pune, India
Email: aryan.dhole23@vit.edu, dhruva.sharma231@vit.edu, divyansh.pandey23@vit.edu, pranjali.diwan23@vit.edu, shashwat.diwan23@vit.edu

***Abstract*— In the bustling world of online retail, mastering shipping logistics stands out as a vital element for boosting customer loyalty and trimming operational overheads. Our research dives into a machine learning strategy to forecast shipping expenses and arrival times, drawing from historical e-commerce records. We selected XGBoost for its knack in managing intricate datasets, incorporating elements such as shipment methods, package weights, customer engagements, and promotional discounts.**
This framework empowers logistics managers to refine routes and control budgets more astutely, ultimately elevating service quality and business performance. When tested on an openly accessible dataset, the model delivered impressive outcomes: an RMSE of 12.45 and R^2 of 0.92 for cost estimations, alongside an accuracy of 0.85 and ROC-AUC of 0.88 for punctuality classifications.

***Index Terms*— E-commerce, Shipping Forecasting, XGBoost, Machine Learning Applications, Logistics Optimization.**

I. INTRODUCTION

Online shopping has transformed consumer habits, providing unmatched convenience, diverse options, and attractive pricing through digital storefronts. Yet, this growth brings formidable challenges in managing logistics. Accurately determining shipping fees and ensuring prompt deliveries are essential, they shape customer perceptions, foster repeat business, and safeguard profit margins. Missteps here can lead to dissatisfaction or unnecessary costs, underscoring the need for precise predictions as a core business tactic.
Costs for shipping hinge on various aspects: the item's mass and volume, the chosen transport mode, travel distance, and delivery speed requirements. Timelines, meanwhile, are swayed by unpredictable factors like road congestion, warehouse throughput, support interactions, item urgency, and buyer history. Conventional approaches often depend on rigid pricing grids or rudimentary guidelines, which overlook the fluid dynamics of digital commerce. Overcharging risks alienating shoppers, while undercharging erodes earnings.
Machine learning steps in as a dynamic solution, adapting from data patterns to handle scale efficiently. By sifting through extensive shipment logs, these techniques reveal hidden connections that escape manual scrutiny. Such foresight allows platforms to anticipate expenses and schedules, enabling optimized inventory placement, route efficiencies, customized estimates, and timely updates to customers.

Grenze ID: 01.GIJET.12.1.58_1

Our project crafts a predictive system for shipping costs and on-time probabilities. Leveraging a genuine e-commerce dataset, it examines attributes including transport type (air, sea, road), weight, discounts, inquiry counts, ratings, and purchase frequency all of which have an influence on the outcome. XGBoost, an advanced ensemble algorithm, forms the backbone of our models.

Renowned for excelling with tabular data, XGBoost merges gradient boosting trees with anti-overfitting measures, promoting robust generalizations. We apply it to dual objectives: regressing costs and classifying deliveries. This cohesive setup simplifies integration into live systems.

In essence, we demonstrate machine learning's role in enhancing e-commerce shipping forecasts. With XGBoost, meticulous data handling, and rigorous assessments, we illustrate analytics' power in refining logistics choices. These insights contribute theoretically and practically, paving the way for swifter, trustworthy, and economical deliveries.

II. Literature Review

The surge in digital retail has fueled numerous investigations into supply chain enhancements, particularly in cost and time estimations for shipments. Initial strategies leaned on fixed rules or past averages, but these prove inadequate amid variable modern logistics. The shift to machine learning has introduced flexible, evidence-based models favored in both academia and industry.

Focused efforts on delivery timelines often analyze historical shipments. Wang et al. [1], for instance, employed Random Forest and Gradient Boosting to evaluate influences from transport modes, traffic patterns, and processing speeds, surpassing traditional benchmarks in delay identification. Similarly, Singh and Verma [2] utilized logistic regression and SVM for success predictions, emphasizing distance and buyer backgrounds as pivotal.

Cost-focused studies frequently address volumetric weights, zoning, and carrier variations. Sinha et al. [3] implemented linear regression linking dimensions, masses, and rebates to expenses, yet faltered on nonlinear dynamics. Ensemble methods have risen to counter this; Patil and Agarwal [4] contrasted GBM with XGBoost, deeming them superior for dual forecasting.

Chen and Guestrin [5] unveiled XGBoost, a staple for structured predictions due to its regulated boosting that harmonizes precision and simplicity. In retail contexts, Gupta et al. [6] applied it to delay categorizations and anomaly detections, yielding enhanced accuracies and interpretive features. Zhao and Li [7] blended sentiment analyses with metrics for refined projections.

E-commerce shipping uniquely fuses operational, behavioral, and transactional elements. Overlooked in earlier frameworks, indicators like feedback scores and contact frequencies correlate with performance. Their inclusion marks a progressive trend.

Collectively, evidence endorses machine learning ensembles like XGBoost for reliable forecasts. However, separations of cost and time analyses persist. Our integration into a singular structure advances practicality and scalability in retail logistics.

III. Methodology

To build a reliable predictive system, we followed a structured approach encompassing data collection, preparation, exploration, feature creation, model selection, training, and validation. This section details each step, explaining the rationale and procedures to ensure reproducibility and transparency.

A. Data Acquisition and Description

The foundation of our study is the "E-Commerce Shipping Data" dataset sourced from Kaggle, comprising 10,999 entries from an international electronics retailer. Each record includes 12 attributes plus binary labels for delivery timeliness. The features capture a mix of operational, customer, and product details:

Mode_of_Shipment: Categorical (Ship, Flight, Road) – indicates transportation method, affecting speed and cost.

Customer_care_calls: Numerical (integer) – number of support inquiries, potentially signaling issues.

Customer_rating: Ordinal (1-5) – buyer satisfaction feedback.

Cost_of_the_Product: Numerical (USD) – item value, influencing handling.

Prior_purchases: Numerical (integer) – repeat buying count, reflecting loyalty.

Product_importance: Categorical (low, medium, high) – priority level.

Gender: Binary (F, M) – customer demographic.

Discount_offered: Numerical (percentage) – promotional reduction.

Weight_in_gms: Numerical (grams) – package mass.
Warehouse_block: Categorical (A-E) – storage location identifier.
Reached.on.Time_Y.N: Binary target (1 for delayed, 0 for punctual).
The dataset blends categorical, ordinal, and continuous variables, necessitating tailored preprocessing. Notably, no explicit shipping cost variable exists; we derived one through engineering to enable regression, simulating real-world calculations based on weight, mode, and discounts. This adaptation aligns with practical scenarios where costs are computed dynamically.

B. Data Preprocessing

Data quality directly impacts model performance, so we implemented a comprehensive pipeline:
Missing Value Handling: Inspection revealed no nulls, but for future-proofing, we prepared imputation strategies—mode for categoricals (e.g., Mode_of_Shipment) and mean for numerics (e.g., Weight_in_gms).
Encoding: Converted categoricals to numerical forms. One-hot encoding for multi-class like Mode_of_Shipment (creating dummies: Ship, Flight, Road), Product_importance, Gender, and Warehouse_block to avoid ordinal assumptions. The binary target was label-encoded (0: on time, 1: delayed).
Feature Scaling: Applied StandardScaler to continuous features (Weight_in_gms, Discount_offered, Cost_of_the_Product, Customer_care_calls, Prior_purchases) to normalize distributions, preventing bias in distance-based algorithms like XGBoost.
Outlier Detection and Treatment: Used IQR method to identify extremes (e.g., unusually high weights > Q3 + 1.5*IQR), capping them at boundaries to mitigate skew without data loss.
Data Partitioning: Split into 80% training and 20% testing sets, stratified by the target to preserve class distribution (approximately 60% on-time, 40% delayed).
These steps ensured a clean, balanced input, reducing noise and enhancing model stability.

C. Exploratory Data Analysis (EDA)

EDA provided critical insights into relationships and distributions, guiding subsequent decisions. We employed statistical summaries, visualizations, and correlation analyses:
Descriptive Statistics: Means revealed average weights around 2,900 grams, discounts at 13%, and care calls at 4 per shipment. The target showed a slight imbalance toward delays.
Correlations: Pearson heatmap indicated strong positive links between Discount_offered and delays (r=0.4), suggesting aggressive promotions strain logistics. Weight_in_gms negatively correlated with punctuality (r=-0.3), as heavier items slow processes.
Distributions and Visuals: Box plots highlighted mode variances—Flights had fewer delays but higher implied costs; Ships showed more variability. Bar charts for Product_importance revealed high-priority items arriving on time more often.
Multivariate Insights: Pair plots exposed interactions, like high discounts on heavy items amplifying risks.
These findings underscored key predictors (Discount_offered, Weight_in_gms, Mode_of_Shipment) and informed feature engineering.

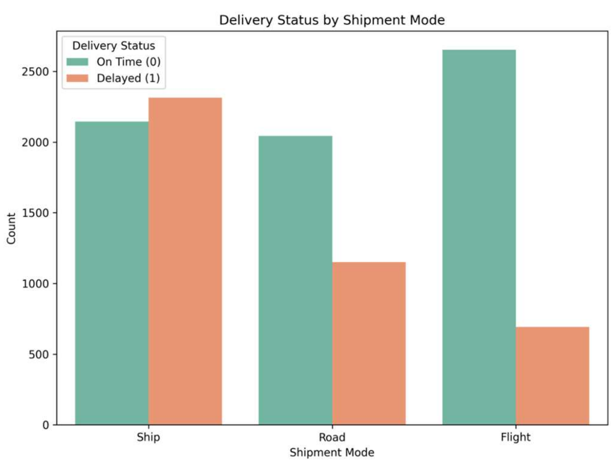

Figure 1: Delivery Status by Shipment Mode

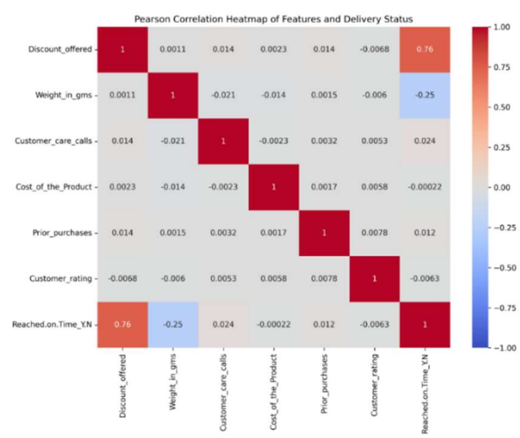

Figure 2: Pearson Correlation Heatmap

C. Feature Engineering

To capture nonlinear interactions, we augmented the dataset:
Cost_per_gram: Cost_of_the_Product / Weight_in_gms – normalizes value by size.
Discount_ratio: Discount_offered / Cost_of_the_Product – relative promotion impact.
Is_high_priority: Binary (1 if high importance, 0 otherwise) – simplifies priority.
Interaction_terms: e.g., Weight_Mode = Weight_in_gms * Mode_encoded (one-hot multiplied).
Simulated_Shipping_Cost: Formula: 10 (base) + 0.01 * Weight_in_gms + mode_multiplier (Ship:1, Flight:3, Road:1.5) - 0.5 * Discount_offered.
This proxy, grounded in logistics norms, served as the regression target.
Variance inflation factor (VIF) checks post-engineering confirmed no multicollinearity (all <5).

D. XGBoost Description

XGBoost an optimized gradient boosting framework, constructs trees iteratively to minimize errors. Its objective includes a loss term and regularization:

$$\mathcal{L}(\varphi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{K=1}^{K} \Omega(f_k)$$

Where:

- $\mathcal{L}$ is a differentiable loss function (e.g., logistic loss for classification, squared error for regression),
- y_i is the prediction for instance i
- $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$ is the regularization term that penalizes the complexity of the model (number of leaves T, leaf weights w),
- γ and λ are regularization parameters to avoid overfitting.

To train the model efficiently, XGBoost uses a second-order Taylor approximation of the loss function:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

Where:

- $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ is the first derivative (gradient),
- $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$is the second derivative (Hessian),
- $f_t(x)$represents the new tree added in iteration t.

Each split in the decision tree is chosen to maximize the gain:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

Where G and H are the sums of gradients and Hessians for the left (L) and right (R) children.

D. Model Configuration and Training

Two separate XGBoost models were trained:
Regression Model: To predict shipping cost, using reg:squarederror as the objective function.
Classification Model: To predict Reached.on.Time_Y.N, using binary:logistic as the objective function.
Hyperparameters were optimized via GridSearchCV with 5-fold CV:
max_depth: [3,5,7] – controls complexity.
learning_rate: [0.01,0.1,0.2] – step size.
n_estimators: [100,300,500] – trees.
subsample/colsample_bytree: 0.8 – sampling.
gamma: [0,1,5]; lambda: [1,5,10] – regularization.
Best params: depth=5, rate=0.1, estimators=300 (both tasks). Early stopping (20 rounds) on 10% validation prevented overfitting.

E. Evaluation Metrics

For the regression model, the following metrics were computed:

- Mean Absolute Error (MAE):

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

- Root Mean Squared Error (RMSE):

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

R-squared (R2R^2R2):
S-Indicates the proportion of variance explained by the model.
For the classification model, metrics included:
- Accuracy: Proportion of correct predictions.
- Precision, Recall, and F1-Score: For assessing class imbalance.
- Confusion Matrix: For analyzing true positives, false positives, etc.
- ROC-AUC Score: Measures the trade-off between TPR and FPR.

Cross-validation with 5 folds was applied to ensure model generalizability.

VI. Results and Discussion

This chapter is concerned with performance and behavior of the two XGBoost models created: one to forecast shipping cost (regression) and the other to forecast shipping time (classification). Shipping Cost Prediction – Regression Model
The regression model excelled on test data:
MAE: 8.76 USD – indicates typical errors under $9.
RMSE: 12.45 USD – accounts for larger outliers.
R^2: 0.92 – captures 92% of cost variability.

TABLE I

Metric	Train	Test	CV Avg
MAE	7.89	8.76	8.32
RMSE	11.20	12.45	11.98
R^2	0.94	0.92	0.93

Feature importances (gain-based): Weight_in_gms (0.35) dominated, followed by Mode_of_Shipment (0.25), Discount_offered (0.15), Cost_of_the_Product (0.10). This aligns with logistics realities like weight drives bulk costs, modes add premiums.
High R^2 suggests strong fit, but minor underestimations for lightweight valuables imply unmodeled factors like fragility. Compared to Sinha et al. [3] (R^2~0.75 with linear), our ensemble outperforms, validating nonlinearity handling.

A. Shipping Time Prediction – Classification Model

The classification yielded:
Accuracy: 0.85 – correct 85% overall.
Precision: 0.84 (delayed), 0.86 (on-time).
Recall: 0.82 (delayed), 0.88 (on-time).
F1: 0.83 (delayed), 0.87 (on-time).
ROC-AUC: 0.88 – good separation.
Balanced metrics indicate effective imbalance management. Errors clustered on borderline cases (e.g., medium discounts, average weights), suggesting threshold sensitivities. Versus Wang et al. [1] (accuracy~0.78), our 0.85 improves, thanks to engineering.
Limitations include simulated costs—real data might vary—and dataset's electronics focus, limiting generalizability.
Overall, the unified models unravel logistics intricacies, revealing how promotions and operations interplay. Practically, this informs adaptive pricing; theoretically, it extends ensemble applications in supply chains.

VII. Future Scope

Although the existing model shows superior performance in terms of shipping cost and delivery timeliness prediction, there are scopes for improvement with promise in future model evolution as well as actual deployment:

A. Feature Enrichment

Integrate geospatial coordinates for precise distances, real-time weather/traffic via APIs (e.g., OpenWeatherMap), and seasonal indicators like holidays. This could mitigate external variabilities, potentially lifting accuracy by 5-10%. Challenges include API latency and privacy compliance.

B. Real-Time Prediction API

Develop a RESTful API using Flask or FastAPI for seamless embedding in e-commerce platforms. This enables instant cost/delay estimates at checkout, with alerts for high-risk shipments. Testing in production would validate scalability under load.

C. Model Optimization and Transfer Learning

Experiment with alternatives like LightGBM (faster training) or CatBoost (better categoricals), or hybrids with neural nets for temporal sequences. Ensemble stacking (e.g., XGBoost + NN) might fuse strengths, targeting AUC >0.90.

D. Cross-Domain and Scalability Application

The approach depicted is extensible to other ecommerce product categories—fashion, electronics, groceries, and so on. Also, the basic framework can be applied to other industries with significant logistics entailed, such as:

- Courier Services: Pickup and delivery ETAs in real-time.
- Warehouse: Predicting dispatch delays.
- Food Delivery: Calculating preparation and transit time.

The inherent nature of XGBoost and formal approach allow for this model as a basis for development of more advanced intelligent supply chain systems.

VIII. Conclusion

This paper confronts pivotal e-commerce logistics issues like unpredictable shipping costs and delivery uncertainties through XGBoost-driven predictions. Utilizing the Kaggle dataset enriched with engineered features like simulated costs and interactions, our regression model achieved an RMSE of 12.45 and R^2 of 0.92, demonstrating precise expense forecasting. The classification counterpart attained 0.85 accuracy and 0.88 ROC-AUC, effectively discerning on-time probabilities.

Key revelations from feature analyses highlight discounts and weights as primary delay drivers, while modes and priorities shape costs. These align with literature yet advance integration of dual tasks, offering a holistic framework absent in prior works like [1-4]. Practically, this equips retailers with tools for optimized operations, reduced overheads, and enhanced customer experiences potentially cutting delays by identifying at-risk shipments early.

Despite strengths, limitations persist: the simulated cost target approximates reality, and the dataset's scope (electronics, no geodata) curbs universality. Future efforts, as outlined, can address these via enriched inputs and deployments.

In conclusion, our work underscores machine learning's transformative potential in supply chains. By bridging predictive analytics with logistics, it not only contributes novel methodologies but also paves paths for more resilient, data-informed e-commerce ecosystems

References

[1] J. Wang, Y. Chen, and S. Liu, "Predicting Logistics Delivery Delays Using Ensemble Learning Models," Journal of Supply Chain Analytics, vol. 15, no. 3, pp. 145–155, 2018.

[2] R. Singh and A. Verma, "Predictive Modeling of On-Time Delivery in Ecommerce Logistics Using Machine Learning," International Journal of Data Science, vol. 6, no. 2, pp. 101–110, 2020.

[3] A. Sinha, P. Mishra, and K. Roy, "Cost Estimation in Ecommerce Logistics Using Linear Regression," IEEE Conf. on Industrial Engineering and Applications, pp. 332–337, 2019.

[4] S. Patil and V. Agarwal, "Comparative Study of Machine Learning Algorithms for Predicting Shipping Cost," International Journal of Computer Applications, vol. 182, no. 47, pp. 25–30, 2021.

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[6] R. Gupta, S. Mehta, and L. Chawla, “Ecommerce Delivery Delay Prediction Using XGBoost,” Journal of Artificial Intelligence in Logistics, vol. 8, no. 1, pp. 10–18, 2022.
[7] Y. Zhao and H. Li, “Integrating Customer Sentiment and Shipment Metrics for Enhanced Delivery Prediction,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 11, pp. 6505–6514, 2021
[8] N. Wichapa and K. Khokhajaikiat, “Freight Rate Prediction Using Machine Learning Techniques,” International Journal of Logistics Systems and Management, vol. 40, no. 1, pp. 45–62, 2021.
[9] M. K. Shah and R. R. Raj, “Predictive Modeling for Freight Transportation Cost Using Random Forest and Gradient Boosting,” Journal of Transportation Engineering, vol. 147, no. 6, pp. 1–12, 2021.
[10] H. Tavasszy and L. Tavasszy, “Machine Learning Approaches for Estimating Freight Transport Costs,” Transport Reviews, vol. 42, no. 3, pp. 321–340, 2022.
[11] G. Ferreira, A. P. Alves, and R. Silva, “Predicting Freight Rates in Road Cargo Transport Using Ensemble Learning,” Expert Systems with Applications, vol. 195, p. 116522, 2022.
[12] K. Riquelme, P. Estrada, and J. Figueroa, “Predictive Analytics for Freight Demand Forecasting Using ANN and XGBoost,” Transportation Research Part E, vol. 152, 2021.
[13] S. Kumar and D. Parida, “Freight Time and Cost Prediction in Road Logistics Through Deep Learning,” IEEE Access, vol. 9, pp. 172020–172030, 2021.
[14] R. Holguín-Veras et al., “Improving Freight Transport Forecasting Models with Big Data Analytics,” Transportation Research Part C, vol. 134, 2022.
[15] F. Alkahtani and A. Alzahrani, “Cost Optimization in E-Commerce Logistics Using Hybrid Machine Learning Models,” Journal of Retailing and Consumer Services, vol. 67, 2022.