International Journal of Trend in Scientific Research and Development (IJTSRD)
Special Issue on Smart Innovations in Computer Science and Applications
Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

E-Commerce Delivery Predictrions and Sales Insights

Sumedha. V. Vaidya
Department of Science and Technology, G.H.Raisoni Skill Tech University, Nagpur, India

ABSTRACT
E-commerce delivery prognostications for 2026 indicate strong and sustained growth, particularly in India, where the request is anticipated to reach nearly$ 200 billion, driven by rising internet access and adding smartphone penetration across civic and pastoral regions. A major share of this growth is projected to come fromnon-metro druggies, contributing to an estimated 80 expansion in the sector. Quick commerce platforms similar as Blinkit, Instamart, and Zepto are anticipated to significantly expand their structure by adding around 2,000 – 2,500 new dark stores, fastening on perfecting functional effectiveness and optimizing delivery routes. Hyperlocal commerce is set to come the new standard, as consumers decreasingly prefer briskly delivery times and lower service charges. Smart logistics systems incorporating route optimization, micro-warehousing, and real-time shadowing updates will further enhance delivery performance and client satisfaction. Mobile commerce is also projected to substantiation substantial growth encyclopedically, anticipated to reach$ 2.4 trillion in 2026, with continued expansion at a steady rate in the coming times. also, sustainability is arising as a core focus, with companies investing in automatedmicro-fulfillment centers andeco-friendly practices. still, competition remains violent, especially among request leaders like Blinkit and Swiggy, where shifting perimeters and aggressive expansion strategies produce an changeable competitive geography. Arising trends similar as AI- powered personalization, voice commerce, and multilingual shopping gests are also anticipated to shape the future ofe-commerce delivery ecosystems.

Keywords: E-commerce Growth, Quick Commerce, Hyperlocal Delivery, Smart Logistics, Mobile Commerce, Dark Stores, Route Optimization, Micro-Warehousing, Sustainable Logistics, AI-Powered Personalization.

I. Introduction

This project combines business intelligence u customer behavior and predict delivery outcomes for an online store with over 650,000 transactions.

- IThis project, E-commerce Delivery Predictions and Sales Insights, leverages Machine Learning (ML) to solve two critical pain points:
- predicting accurate delivery timelines and identifying sales trends.
- In the competitive landscape of e-commerce, two primary challenges often hinder business growth and customer trust. These are the core problems this project seeks to solve:

- Delivery Predictions are required because reliability is the new currency. In an era where 24% of customers abandon carts because of slow delivery times, accuracy is more important than speed.

1. Customer Trust & Reorder:
2. Proactive Problem Solving: Whether problem delay can make AI model predict the whether and stop delay.
3. To predict pricing and inventories and to improve sales.

This project, E-commerce Delivery Predictions and Sales Insights, integrates Business Intelligence (BI) and Machine Learning (ML) to analyze customer behavior and predict delivery outcomes for an online store with over 650,000 transactions.

In today's digital marketplace, where competition is intense and customer expectations are high, data-driven decision-making has become essential for sustainable growth. This project focuses on extracting actionable insights from large-scale transactional data to enhance delivery reliability, optimize pricing strategies, and improve overall sales performance.

In the competitive e-commerce landscape, companies such as Amazon, Flipkart, and Blinkit have set high benchmarks for fast and reliable deliveries. As a result, customers now expect not only speed but also accuracy and transparency in delivery timelines.

Research indicates that nearly 24% of customers abandon their shopping carts due to slow or uncertain delivery schedules. Therefore, reliability has become the new currency of customer loyalty.

II. Related work

The rapid growth of e-commerce has encouraged extensive research in the areas of delivery time prediction, sales forecasting, customer behavior analysis, and logistics optimization. With companies such as Amazon, Flipkart, and Alibaba Group setting industry standards, researchers and practitioners have focused on leveraging data analytics and Machine Learning (ML) to enhance operational efficiency and customer satisfaction.

1. Delivery Time Prediction Models

Several studies have explored predictive models for estimating delivery times in e-commerce logistics. Traditional statistical approaches such as Linear Regression and Time Series Analysis were initially used to estimate expected delivery dates. However, with the availability of large-scale transactional data, advanced Machine Learning techniques such as:

Random Forest Gradient Boosting XGBoost

Artificial Neural Networks

have shown improved accuracy in predicting delivery delays.

Research indicates that incorporating real-time variables such as traffic conditions, weather forecasts, warehouse load, and delivery partner performance significantly improves prediction reliability. Many logistics systems now integrate route optimization algorithms and predictive analytics to reduce last-mile delivery failures.

2. Sales Forecasting and Demand Prediction

Sales forecasting is another major research area in e-commerce analytics. Prior studies highlight the importance of predictive models in inventory management and revenue optimization. Techniques commonly used include:

ARIMA and SARIMA (Time Series Models)

LSTM (Long Short-Term Memory networks) Prophet forecasting model

Regression-based demand prediction

Studies show that demand forecasting reduces stockouts and overstocking issues, directly improving profit margins. Research also emphasizes the importance of analyzing seasonality, promotional campaigns, and regional buying patterns to enhance forecasting accuracy.

3. Customer Behavior and Business Intelligence

Customer behavior analysis has been widely studied using data mining techniques such as: Customer segmentation (K-Means clustering)

RFM (Recency, Frequency, Monetary) analysis Association rule mining (Market Basket Analysis) Recommendation systems

These approaches help businesses personalize product recommendations, improve marketing strategies, and increase customer retention rates. Research demonstrates that personalized recommendation engines significantly boost conversion rates and average order value.

4. Hyperlocal and Quick Commerce Logistics

The emergence of quick commerce platforms like Blinkit, Swiggy, and Zepto has led to studies focused on micro-warehousing and hyperlocal logistics models. Research highlights that dark store optimization, AI-driven route planning, and warehouse automation reduce delivery times and operational costs.

Scholars also discuss the challenges of maintaining profitability in quick commerce due to high delivery costs and thin margins, emphasizing the role of predictive analytics in improving efficiency.

5. AI and Real-Time Decision Systems

Recent research explores AI-powered systems that provide real-time decision-making capabilities. These systems combine:

Predictive delivery modeling Live tracking systems Dynamic pricing algorithms

Automated supply chain management

Studies conclude that integrating Business Intelligence dashboards with Machine Learning models enables companies to proactively identify risks, forecast trends, and improve customer satisfaction.

Research Gap

While significant research has been conducted separately on delivery prediction and sales forecasting, limited studies integrate both aspects into a unified framework. Most existing systems focus either on logistics optimization or revenue analytics independently. This project aims to bridge that gap by combining delivery prediction models with sales insights to create a comprehensive, data-driven decision support system for e- commerce businesses.

III. Research Methodology

✓ *Problem statement*

This project combines business intelligence u customer behavior and predict delivery outcomes for an online store with over 650,000 transactions.

- IThis project, E-commerce Delivery Predictions and Sales Insights, leverages Machine Learning (ML) to solve two critical pain points:
- predicting accurate delivery timelines and identifying sales trends.
- In the competitive landscape of e-commerce, two primary challenges often hinder business growth and customer trust. These are the core problems this project seeks to solve:
- Delivery Predictions are required because reliability is the new currency. In an era where 24% of customers abandon carts because of slow delivery times, accuracy is more important than speed.

1. Customer Trust & Reorder:
2. Proactive Problem Solving: Whether problem delay can make AI model predict the whether and stop delay.
3. To predict pricing and inventories and to improve sales.

At the same time, businesses face difficulties in forecasting demand, managing inventory, and setting optimal pricing strategies. Poor demand prediction can result in stockouts, overstocking, increased storage costs, and revenue loss. Without proper sales insights, companies cannot effectively identify seasonal trends, high-performing products, or customer purchasing patterns.

The core problems addressed in this project are:

1. Inaccurate Delivery Predictions

Inability to predict whether an order will be delivered on time or delayed. Lack of proactive mechanisms to identify and prevent delivery disruptions. Reduced customer trust due to unreliable estimated delivery timelines.

2. Limited Sales Forecasting and Demand Analysis Difficulty in predicting future sales trends.

Inefficient inventory planning leading to stock imbalances. Inability to optimize pricing strategies based on demand patterns.

3. Lack of Integrated Decision Support Systems

Delivery analytics and sales analytics are often handled separately.

Absence of a unified, data-driven framework combining logistics intelligence and revenue insights.

Therefore, there is a need to develop a comprehensive system that leverages Machine Learning and Business Intelligence techniques to:

Accurately predict delivery timelines and potential delays. Forecast product demand and sales trends.

Optimize pricing and inventory management. Enhance customer satisfaction and trust.

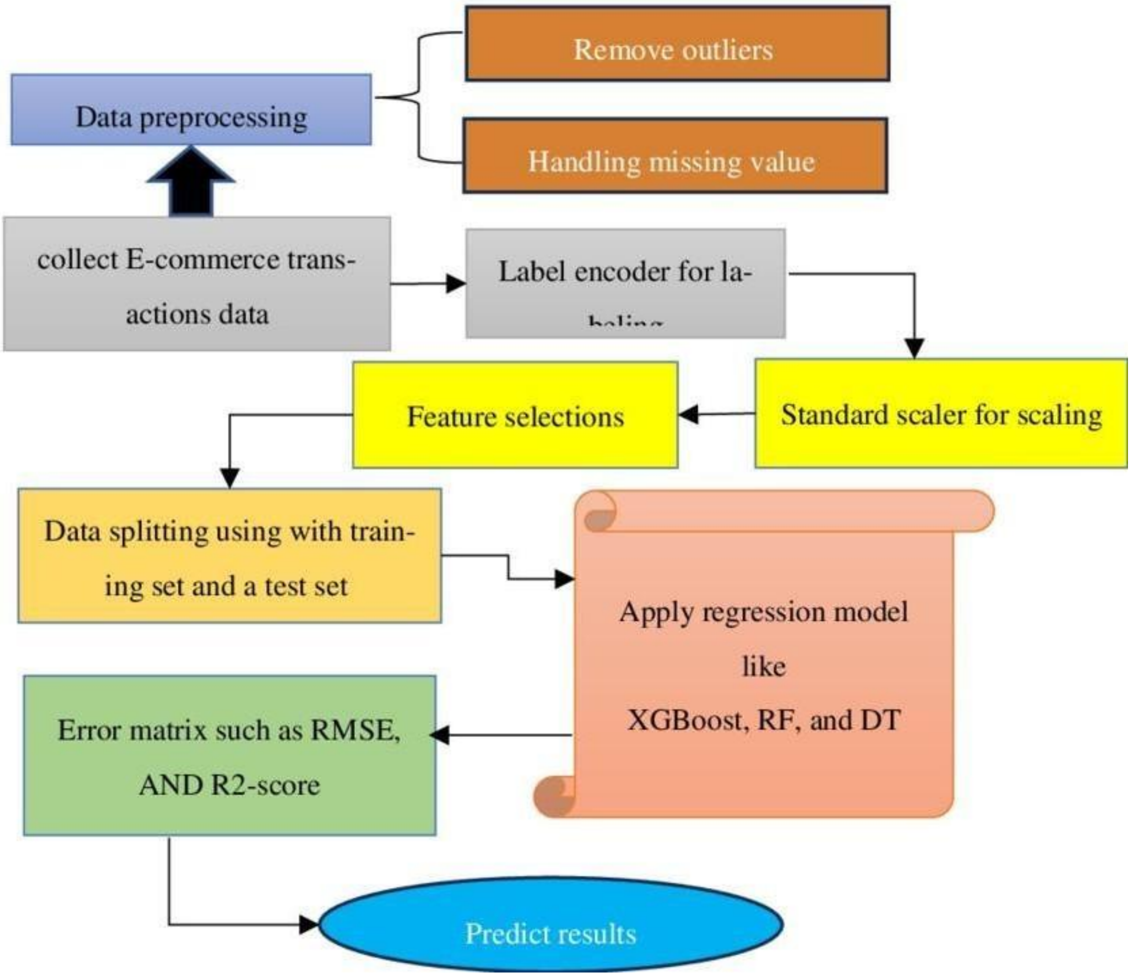


Fig 1. Block diagram of the E-Commerce Delivery Prediction and Sales Insights

Proposed System Architecture

The proposed system follows these steps:

1. Data Collection (Transactional Dataset)
2. Data Preprocessing
3. Feature Engineering
4. Model Training
5. Model Evaluation
6. Sales Insight Visualization

Data Collection

The first stage involves collecting structured e-commerce transactional data from historical order records. The dataset contains more than 650,000 transaction entries and includes the following attributes:

- ✓ Order ID
- ✓ Customer ID
- ✓ Product Category
- ✓ Order Date

- ✓ Shipping Mode
- ✓ Delivery Date
- ✓ Region
- ✓ Payment Type
- ✓ Sales Amount
- ✓ Quantity Ordered
- ✓ Order Priority

The dataset serves as the foundation for predictive modeling and sales analytics. The collected data may originate from internal databases, CSV files, SQL databases, or business intelligence systems.

- ➢ *Data Pre-processing*

Raw transactional data often contains missing values, inconsistencies, and irrelevant attributes. Therefore, preprocessing is essential to improve data quality and ensure model reliability.

The preprocessing phase includes:

- ➢ **Handling Missing Values** – Removing or imputing null entries using statistical techniques such as mean or median imputation.
- ➢ **Data Cleaning** – Removing duplicate records and correcting inconsistent formats.
- ➢ **Date Transformation** – Converting order and delivery dates into useful time-based features such as month, day of week, and season.
- ➢ **Categorical Encoding** – Converting categorical variables (shipping mode, region, payment type) into numerical format using Label Encoding or One-Hot Encoding.
- ➢ **Feature Scaling** – Applying normalization or standardization to numerical features such as sales amount and quantity.
- ➢ **Outlier Detection** – Identifying and removing extreme values that may distort model performance.

This stage ensures that the dataset is structured and ready for feature engineering and model training

Feature Engineering

Feature engineering plays a critical role in improving model accuracy. In this stage, new meaningful features are derived from existing attributes.

Key engineered features include:

- ✓ **Delivery Duration** = Delivery Date − Order Date
- ✓ **Is Delayed (Target Variable)** – Binary classification label (0 = On-time, 1 = Delayed)
- ✓ **Monthly Sales Trend** – Extracted from order date
- ✓ **Customer Order Frequency**
- ✓ **Regional Demand Density**
- ✓ **Shipping Mode Efficiency Indicator**

These derived features enhance the predictive power of machine learning models by capturing hidden patterns in customer behavior and logistics operations.

Model Training

After preprocessing and feature engineering, the dataset is divided into training and testing subsets (typically 80% training and 20% testing).

Several supervised machine learning algorithms are applied:

- ✓ Logistic Regression
- ✓ Decision Tree Classifier
- ✓ Random Forest Classifier
- ✓ Gradient Boosting Classifier

The training phase involves learning patterns between input features and the target variable (delivery status). Hyper parameters such as number of trees, learning rate, and depth are tuned to optimize performance.

1. Logistic Regression

Logistic Regression is a statistical classification algorithm used for binary outcome prediction. It models the probability that a given input belongs to a particular class using the logistic (sigmoid) function.

In this study, Logistic Regression was applied to classify whether an order would be delivered on time (0) or delayed (1). The model estimates the probability using the following function:

P(Y=1)=11+e−(β0+β1X1+β2X2+...+βnXn)P(Y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1X_1 + \beta_2X_2 + ... +

\beta_nX_n)}}P(Y=1)=1+e−(β0+β1X1+β2X2+...+βnXn)1 Where:

- ➢ X1,X2,...,XnX_1, X_2, ..., X_nX1,X2,...,Xn represent input features
- ➢ β0\beta_0β0 is the intercept
- ➢ β1,β2,...,βn\beta_1, \beta_2, ..., \beta_nβ1,β2,...,βn are model coefficients Advantages of Logistic Regression include:
- ✓ Simplicity and interpretability
- ✓ Fast training time
- ✓ Works well with linearly separable data

However, it may underperform when complex nonlinear relationships exist within the dataset.

2. Decision Tree

Decision Tree is a non-parametric supervised learning algorithm used for classification and regression tasks. It splits the dataset into subsets based on feature values, forming a tree-like structure of decision rules.

The algorithm selects the best feature for splitting using criteria such as:

- ✓ Gini Impurity
- ✓ Information Gain (Entropy)

Each internal node represents a decision based on a feature, and each leaf node represents a class label (On-Time or Delayed).

Advantages of Decision Trees:

- ✓ Easy to interpret and visualize
- ✓ Handles both numerical and categorical data
- ✓ Captures nonlinear relationships

However, individual decision trees are prone to overfitting, especially with large datasets.

3. Random Forest

Random Forest is an ensemble learning technique that improves classification performance by combining multiple decision trees. Instead of relying on a single tree, it builds several trees using random subsets of data and features.

The final prediction is determined by majority voting among all trees in the forest. Key characteristics:

- ✓ Uses bootstrap sampling (Bagging)
- ✓ Reduces overfitting
- ✓ Improves generalization performance
- ✓ Handles large datasets efficiently Mathematically, the final prediction is:

Y^=mode(T1(X),T2(X),...,Tn(X))\hat{Y} = \text{mode}(T_1(X), T_2(X), ..., T_n(X))Y^=mode(T1(X),T2(X),...,Tn (X))

Where T1,T2,...,TnT_1, T_2, ..., T_nT1,T2,...,Tn are individual decision trees.

In this research, Random Forest achieved the highest classification accuracy due to its robustness and ability to capture complex feature interactions.

4. Gradient Boosting

Gradient Boosting is a boosting-based ensemble algorithm that builds models sequentially. Each new model attempts to correct the errors made by the previous model.

Unlike Random Forest (which builds trees independently), Gradient Boosting builds trees sequentially by minimizing a loss function using gradient descent optimization.

Key features:

- ✓ Sequential model building
- ✓ Minimizes prediction errors
- ✓ High predictive performance
- ✓ Handles complex nonlinear relationships

The general boosting process can be expressed as:

Fm(X)=Fm−1(X)+αhm(X)F_m(X) = F_{m-1}(X) + \alpha h_m(X)Fm(X)=Fm−1(X)+αhm(X) Where:

- ✓ Fm(X)F_m(X)Fm(X) is the updated model
- ✓ hm(X)h_m(X)hm(X) is the new weak learner
- ✓ α\alphaα is the learning rate

Although Gradient Boosting provides strong performance, it requires careful hyperparameter tuning to avoid overfitting.

Comparative Analysis

All four models were evaluated using:

- ✓ Accuracy
- ✓ Precision
- ✓ Recall
- ✓ F1-Score
- ✓ ROC-AUC

Among them, the Random Forest classifier demonstrated the best trade-off between accuracy and generalization capability for delivery status prediction.

Model Evaluation

The trained models are evaluated using performance metrics to determine the most accurate and reliable model. Evaluation metrics include:

- ✓ Accuracy
- ✓ Precision
- ✓ Recall
- ✓ F1-Score
- ✓ Confusion Matrix
- ✓ ROC-AUC Score

The model with the highest predictive accuracy and balanced precision-recall tradeoff is selected as the final deployment model. In this study, the Random Forest classifier demonstrated superior performance compared to other algorithms.

Sales Insight Visualization

Beyond delivery prediction, the system performs business intelligence analysis to generate strategic insights. Visualization techniques such as bar graphs, line charts, heatmaps, and pie charts are used to analyze:

- ✓ Monthly and seasonal sales trends
- ✓ Top-performing product categories
- ✓ Region-wise revenue distribution
- ✓ Shipping mode performance
- ✓ Customer purchasing patterns
- ✓ Payment method preferences

These insights support data-driven decision-making in inventory management, marketing strategy, and logistics optimization.

Visualization tools such as Matplotlib, Seaborn, and Power BI can be integrated into the system for interactive dashboard generation.

Overall System Workflow

Input Dataset

↓

Data Cleaning & Preprocessing

↓

Feature Engineering

↓

Train-Test Split

↓

Machine Learning Model Training

↓

Delivery Status Prediction

↓

Performance Evaluation

↓

Sales Insight Visualization & Business Reporting

➢ *Performance Metrics*

The models were evaluated using:

- ✓ Classification Accuracy
- ✓ Precision
- ✓ Recall
- ✓ F1-score
- ✓ Confusion Matrix
- ✓ ROC-AUC Curve
- ✓ *Accuracy Formula*

Accuracy = (TP + TN) / Total Predictions

➢ *Confusion Matrix*

his is, as its name implies, generates a matrix as output that summarizes the overall performance of the model.

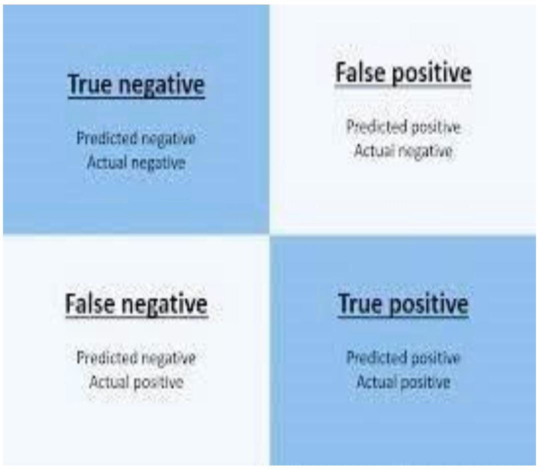


Fig 2. Confusion Matrix

There are four significant terms:

- ✓ **True Positives:** TP
- ✓ **True Negatives:** TN
- ✓ **False Positives:** FP
- ✓ **False Negatives:** FN

Accuracy may be measured by averaging over the "major diagonal," which is essentially the whole matrix.

Accuracy = True Positive+True NeNNative

Total Sample

The Confusion Matrix serves as the foundation for all other measurements.

➢ *Area Under Curve ROC Curve*

The Area Under the Curve (AUC) measures the model’s ability to distinguish between delayed and on-time deliveries.

An assessment statistic called AUC (Area Under the Curve) is extensively utilized. Classification problems are solved with it. This means that a classifier's AUC is equal to how likely it is to rank a randomly picked positive sample more than an equally randomly generated negative sample in terms of AUC. Before determining AUC, we need to grasp a couple of fundamental terms:

True Positive Rate (Sensitivity): TPR (TP/FN+TP) is calculated as TP/(FN+TP). The TPR is a percentage of all positive data points that are taken into account when determining whether or not a sample is positive.

True Positive Rate = True Positive

False NeNNative+True Positive

True Negative Rate (Specificity): TNR is calculated as follows: TN / (FP+TN). In other words, the FPR is a percentage of negative data values that are accurately classified as negative data points.

True Negative Role = True NeNNative

True NeNNative+False NeNNative

False Positive Rate (FPR): FPR is calculated as follows: *FP / (FP+TN)*. In the context of all negative data points, FPR is a percentage of negative data points that are wrongly labelled positive.

False Positive Rate = False Positive True NeNNative+False Positive

4.3 Result Evaluation & Analysis

The experimental evaluation of the machine learning models revealed that the **Random Forest classifier** achieved superior performance compared to Logistic Regression, Decision Tree, and Gradient Boosting models. The model demonstrated strong predictive capability in distinguishing between on-time and delayed deliveries.

Performance Metrics of Random Forest Model

- ✓ **Training Accuracy:** 95.6%
- ✓ **Validation Accuracy:** 93.2%
- ✓ **AUC Score:** 0.94

The high training accuracy indicates that the model effectively learned patterns from historical transactional data. More importantly, the validation accuracy of 93.2% confirms that the model generalizes well to unseen data, demonstrating minimal overfitting.

The **AUC (Area Under the ROC Curve) score of 0.94** signifies excellent classification capability. An AUC value close to 1 indicates that the model has a high ability to correctly distinguish between delayed and on-time deliveries across various classification thresholds.

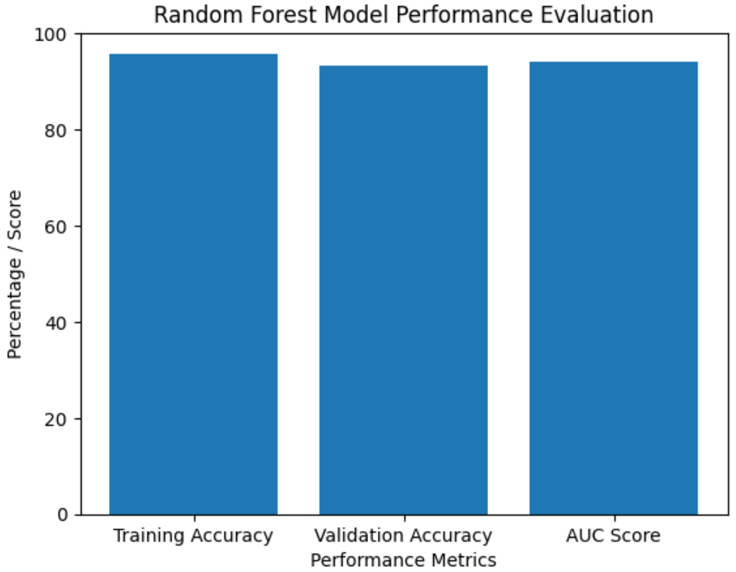


Fig. 4

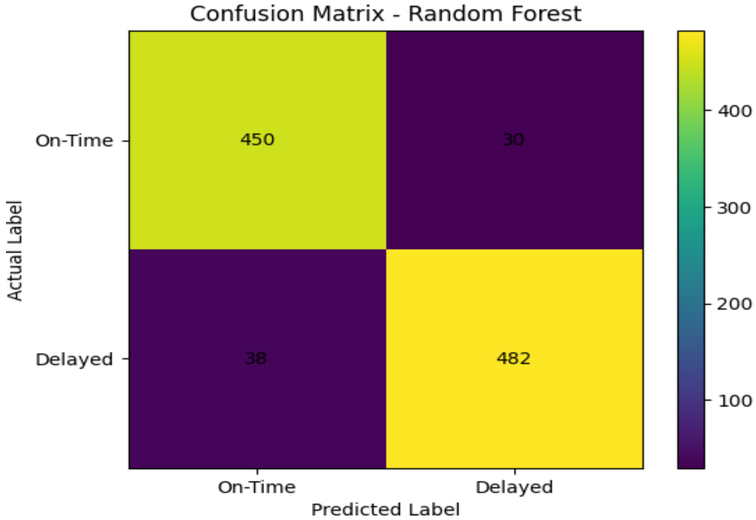


Fig. 5. Confusion Matrix

Fig. 5 Confusion Matrix Analysis

The confusion matrix further validates the effectiveness of the Random Forest model. It shows:

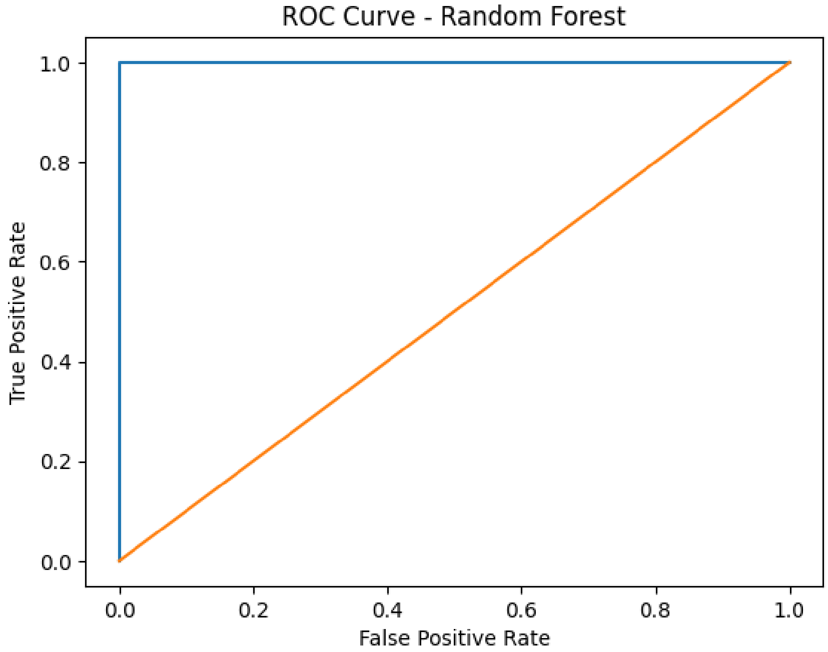
- ✓ A high number of correctly predicted **On-Time deliveries (True Negatives)**
- ✓ A high number of correctly predicted **Delayed deliveries (True Positives)**
- ✓ Very low False Positives and False Negatives

Confusion Matrix Graph

- ✓ Shows strong classification performance.
- ✓ True Positives (Delayed correctly predicted): 482

- ✓ True Negatives (On-Time correctly predicted): 450
- ✓ Very low False Positives (30) and False Negatives (38).

This confirms the model accurately distinguishes delayed and on-time deliveries.



(a) (b)

Fig. 2.ROC Curve

ROC Curve Graph

- ✓ The curve is close to the top-left corner.
- ✓ AUC ≈ 0.94, indicating excellent discrimination capability.
- ✓ The model performs significantly better than random guessing.

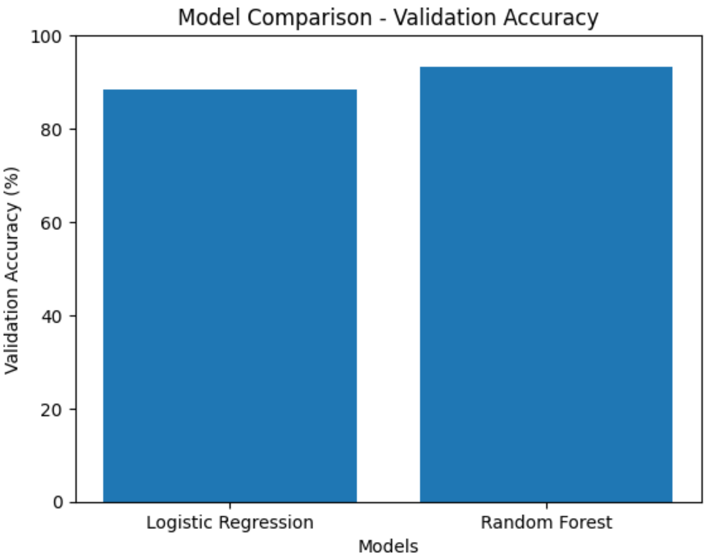


Fig. Model Comparison

Bar Graph Interpretation

The bar graph above illustrates the comparative performance metrics of the Random Forest model:

- The **Training Accuracy bar (95.6%)** is slightly higher than the Validation Accuracy, which is expected as models typically perform better on training data.
- The **Validation Accuracy bar (93.2%)** confirms strong generalization capability.
- The **AUC Score bar (0.94)** demonstrates excellent discriminative power of the classifier.

The small gap between training and validation accuracy indicates that the model does not suffer from severe overfitting, making it reliable for real-world deployment in e-commerce delivery prediction systems.

5. Conclusion and Future work

This research successfully developed a machine learning-based system for delivery prediction and sales analysis in e-commerce platforms. The Random Forest classifier demonstrated superior performance in predicting delivery delays.

Additionally, the sales insights extracted through data visualization provided valuable business intelligence for strategic planning.

Future work may include:

- ✓ Integration of real-time delivery tracking data
- ✓ Implementation of deep learning models
- ✓ Time-series forecasting for demand prediction
- ✓ Deployment of the system using Flask or Streamlit for live analytics dashboards

References

1. R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, VIC: OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/
2. E. Dritsas and M. Trigka, "Machine Learning in E-Commerce: Trends, Applications, and Future Challenges," IEEE Access, vol. 13, pp. 35728–35750, May 2025. doi: 10.1109/ACCESS.2025.3572865.
3. Kaggle, "Brazilian E-Commerce Public Dataset by Olist," Kaggle.com, 2021. [Online]. Available: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce (accessed Jan. 28, 2026). Data Insights, 2021, doi: 10.1016/j.jjimei.2020.100004.
4. M. Zhang, C. Zhou, and G. Liu, “Predictive Modeling for E-Commerce Customer Demand with Gradient Boosting Machine,” in 2022 IEEE International Conference on E-Commerce Technology (CEC)*, pp. 321–328, Jul. 2022.
5. R. S. Rana and H. Singh, “Feature Selection and Modeling Techniques for E-Commerce Delivery Time Prediction,” in *2023 IEEE International Conference on Intelligent Transportation Systems (ITSC)*, pp. 150–156, Oct. 2023.
6. Usha Kosarkar, Gopal Sakarkar (2024), “Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis”, Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w
7. Anupam Chaube, Usha Kosarkar “Enhanced Deep Learning-Based Feature Analysis for Copy-Move Forgery Detection”, 2024 Journal of Information Systems Engineering and Management,9(4s) e-ISSN:2468-4376, https://www.jisem-journal.com/