Enhanced Delivery Performances with Risk Predictions

Mentor: Dr. Dongsheng Luo

Instructor: Dr. Ananda Mondal

JAYAVARDHAN REDDY SAMIDI

Abstract

In today's fast-paced world, ensuring on-time delivery has become a critical component for maintaining customer satisfaction and operational efficiency. However, unforeseen challenges such as logistical constraints, customer-specific variables, and resource allocation issues often lead to delays. This project investigates the problem of late deliveries by predicting delivery risks using a dataset encompassing shipping, order, and customer-related information[3]. The objective is to classify deliveries as either "on-time" or "late" while uncovering the underlying factors that contribute to delays.

A combination of feature engineering and machine learning algorithms was employed to create a predictive framework. The study not only identifies key variables affecting delivery timelines but also provides actionable insights for reducing delays. By addressing these factors, organizations can enhance their logistics performance, improve customer satisfaction, and streamline operations. The results demonstrate the practical value of predictive analytics in mitigating risks and optimizing supply chain efficiency.

Introduction

Delivering goods on time is a crucial factor in ensuring customer loyalty and maintaining the competitive edge of any business. In a global supply chain ecosystem, where speed and reliability are paramount, late deliveries can lead to dissatisfied customers, financial losses, and damaged reputations. Factors influencing delivery times can range from shipping delays and unforeseen weather conditions to order complexities and operational inefficiencies.

-These variables create a pressing need for data-driven solutions to predict and mitigate late delivery risks. This project focuses on developing a predictive model to address the risk of late deliveries by analyzing a rich dataset that includes shipping details, customer demographics, and order characteristics. The goal is to not only predict whether a delivery will be late but also identify the multiple factors that influence these delays[3]. By employing advanced data preprocessing, feature selection techniques, and machine learning algorithms, this study establishes a framework for better understanding delivery dynamics.

The insights gained from this project aim to equip businesses with actionable recommendations to proactively address risks. By understanding the root causes of late deliveries, organizations can optimize their processes, improve efficiency, and meet customer expectations more effectively. This work underscores the importance of predictive power in addressing real-world logsitics challenges and driving operational excellence.

Motivation

In Logistics and Operation Management, not on time deliveries are among the most common and impactful challenges, directly influencing customer satisfaction, operational efficiency, and financial performance. The frustration of missed delivery timelines not only damages customer trust but also incurs costs due to penalties, lost future sales, and increased operational adjustments. This project is driven by the pressing need to understand and mitigate late delivery risks, which are often caused by a complex interplay of factors such as shipping delays, inventory management issues, and customer-specific requirements.

The availability of rich datasets encompassing shipping, order, and customer information presents an opportunity to delve into these challenges through data-driven approaches. Predicting late delivery risks allows businesses to act proactively, enabling them to allocate resources effectively, optimize delivery schedules, and communicate transparently with customers[3]. The project was further inspired by the growing role of machine learning in logistics, demonstrating how advanced analytics can provide meaningful insights and actionable solutions to recurring delivery problems.

This initiative reflects a broader commitment to improving supply chain reliability by addressing one of its most critical pain points—delivery delays. The motivation lies in not just creating a predictive model but also offering insights into the underlying causes, empowering businesses to reduce late delivery occurrences, improve operational efficiency, and ultimately enhance customer satisfaction.

Project Applications

The late delivery risk prediction project has several practical applications has direct contact with machine learning model developed and main dataset used. These include:

1. The primary application is to classify orders as "on-time" or "late" using predictive analytics. The model leverages features such as shipping mode, order date, and delivery schedules to assess risks with accuracy.

2. By analyzing the relationships between input features and delivery outcomes, the project identifies significant factors—like shipping delays or customer-specific trends—that contribute to late deliveries.

3. Insights from the model can help optimize shipping processes by highlighting patterns in late deliveries, such as recurring issues with specific regions, modes, or times.

4. Accurate predictions allow businesses to take preemptive measures, such as notifying customers about potential delays or rerouting shipments to avoid disruptions.

5. The code developed provides a scalable framework that can be integrated into larger supply chain systems, enabling real-time risk monitoring.

Objectives and Deliverables

- Create machine learning models to ensure accurate predictions of late deliveries using available shipping, order, and customer-related data.
- Identify the most critical features affecting delivery risks and analyze their impact on the model's predictions.
- Compare the different machine learning models to determine the best-performing one for predicting risk of product delivering.
- Test impact of reducing 10% of the values for the top features on the performance of the top three models to assess how sensitive the predictions are to feature changes.
- Examine whether reducing specific features significantly alters risk prediction, allowing for more streamlined and efficient model development.
- Use the results to generate actionable insights and strategies for minimizing late delivery risks and improving overall supply chain operations.

Tools List

- Pandas
- NumPy
- Matplotlib
- Seaborn
- SciPy
- Scikit-learn
- LightGBM
- CatBoost
- XGBoost
- Jupyter Notebook

Dataset

This project utilizes three datasets, each contributing unique perspectives and insights into customer interactions, product details, and delivery operations. These datasets provide a multi-dimensional view of the factors influencing delivery performance:

1. **Logs Dataset**: This dataset captures customer interactions on the website, including data on product views, pages visited, and time spent on each page. It offers behavioral insights into customer preferences and engagement, which may indirectly influence order characteristics and delivery outcomes.

2. **Description Dataset**: This dataset contains metadata and descriptive details about the products, such as names, categories, and specifications. It complements the main dataset by adding context to the items being analyzed, enabling a more comprehensive understanding of product-level factors.

3. **Main Dataset**: The core dataset of the project, this contains detailed information on customer orders, shipping schedules, and product details. It is specifically chosen for its relevance to the project's goal of predicting late delivery risks.

Main Dataset

The size of Dataset is 180519 and 52 dependent variables, encompassing a mix of 24 categorical variables, 28 continuous variables, and a target variable. Key attributes in this dataset include:

- Shipping Details: Features like mode, days for real shipping shipping dates provide insights into logistical timelines and variations.
- Order Characteristics: Fields such as order dates, order status, and product prices help analyze patterns in customer orders and their potential influence on delivery performance.
- Customer Information: Details like customer city, state, and segment (e.g., Consumer, Corporate) offer demographic insights that may correlate with delivery outcomes.
- Independent Variable: The Late_delivery_risk variable serves as the classification target, indicating whether a delivery was late or expected time.

Data Preprocessing

Data Cleaning: In the data cleaning phase, we examined the dataset for missing values. Four columns were identified with null values:

- **Customer Lname**: 8
- **Customer Zipcode**: 3
- **Order Zipcode**: 155,679
- **Product Description**: 180,519

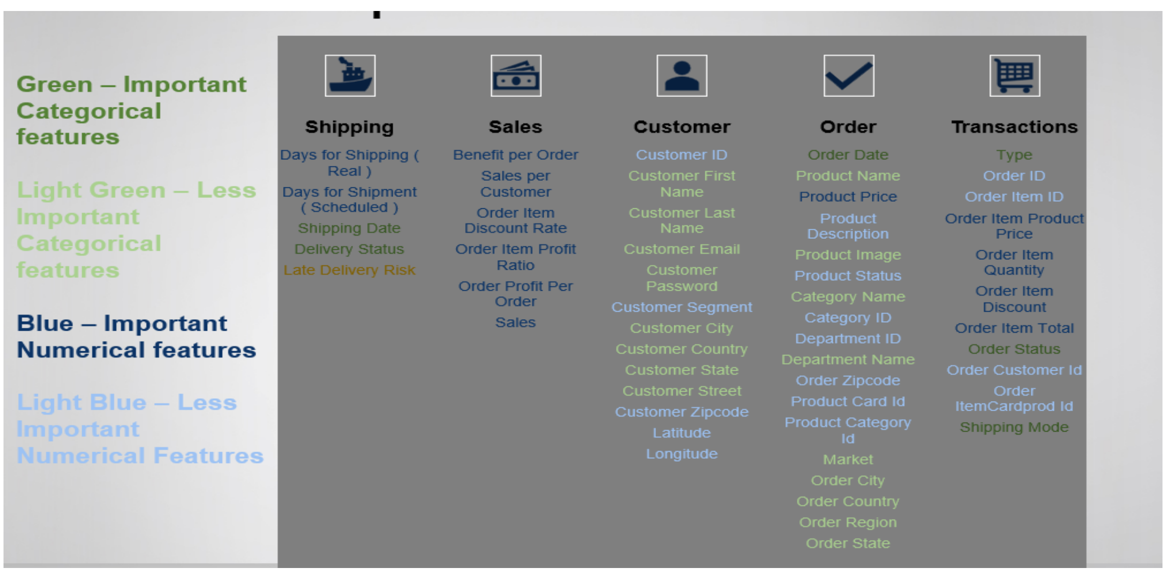


Figure 1: Key Features for Predicting Late Delivery Risk

These columns are not considered critical for predicting the target variable , based on their low relevance to delivery performance. Hence, they were excluded from further analysis to maintain focus on the most impactful variables.

Selection of Relevant Variables: Out of the 52 variables, the following were identified as important features for predicting late delivery risk:

1. **Days for Shipping (Real)** and **Days for Shipment (Scheduled)**: These variables directly measure the actual and planned shipping timelines. Any significant gap between them can indicate potential delays, making them crucial for understanding delivery risks.

2. **Benefit per Order** and **Sales per Customer**: These variables reflect the financial dynamics of each order. Higher benefit or sales values might indicate higher-priority orders that could influence delivery speed.

3. **Order Item Discount Rate** and **Order Item Profit Ratio**: Discounts and profit ratios provide insights into order-level prioritization. For example, highly discounted items might be deprioritized, contributing to delays.

4. **Order Profit per Order** and **Sales**: These metrics indicate the revenue generated per order and overall sales value, which can correlate with operational priorities and impact delivery schedules.

5. **Order Item Product Price**, **Order Item Quantity**, and **Order Item Total**: These factors provide granular details about the size and complexity of orders. Larger or bulkier orders might experience delays due to handling and shipping constraints.

6. **Shipping Date** and **Delivery Status**: These directly track the logistics timeline and outcomes of the delivery process, making them essential for predicting delays.

7. **Order Date**, **Type**, and **Order Status**:These variables help capture temporal and transactional patterns, offering insights into delivery timing and order processing efficiency.

8. **Shipping Mode**: Choice of this variablehas a direct impact on delivery timelines and risk.

Why Other Variables Are Less Important:

Other features, such as **Customer City**, **Customer Country**, **Latitude**, and **Longitude**, provide geographic context but have minimal direct influence on delivery timing once the shipping mode and timeline are factored in. Similarly, variables like **Product Description**, **Customer Lname**, and **Customer Zipcode** are unlikely to affect delivery performance, as they are either descriptive or redundant for this analysis.

By focusing on the selected **important variables**, we ensure that the predictive model is built using the most relevant information, reducing noise and improving accuracy. This streamlined approach allows the model to identify meaningful patterns and make more precise predictions about late delivery risks.

Exploratory Data Analysis of Categorical Features:

Type vs. Target Variable: Here, it shows the analysis between transaction type and late delivery risk reveals an interesting pattern. Among the types—Debit, Cash, Transfer, and Payment—Transfer consistently shows a lower risk. This suggests type of transaction may influence delivery reliability and accuracy.

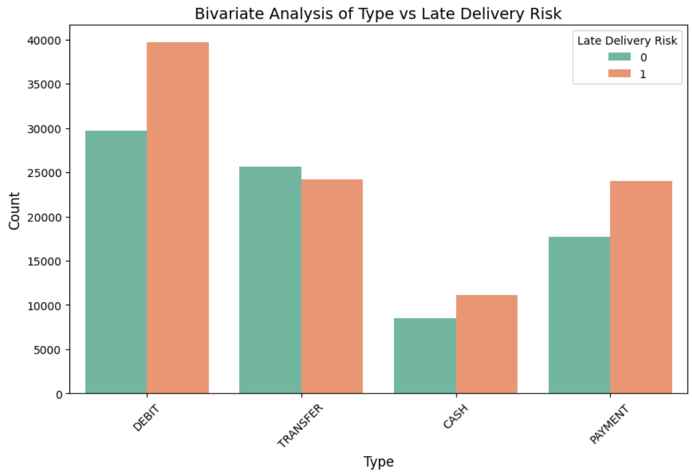


Figure 2

Customer Segment vs. Late Delivery Risk: When exploring this column which has three field values shown in Figure 3, there are observable differences in late delivery risk. Further validation using statistical tests, such as p-values, would help confirm whether these segments significantly impact delivery risk.

Figure 3

Order Status vs. Late Delivery Risk: All categories of Order Status exhibit consistently high late delivery risk. This uniformity suggests that order status alone might not provide meaningful differentiation for predicting late deliveries.

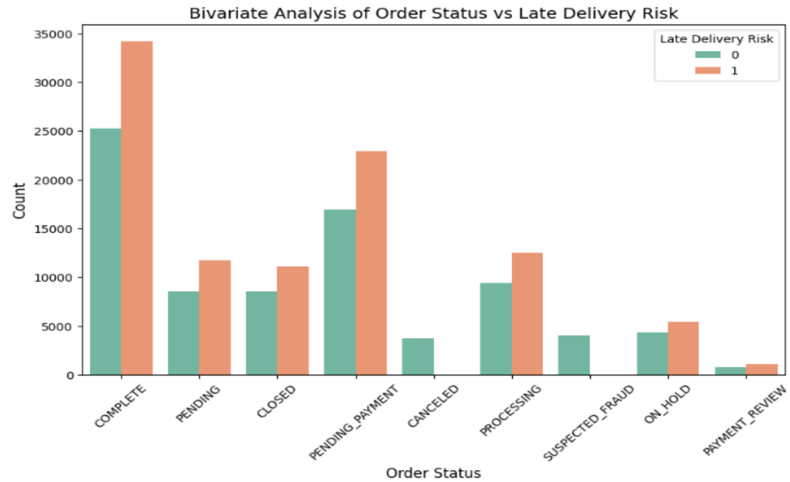


Figure 4

Shipping Mode vs. Late Delivery Risk: Among the different shipping modes, Standard Class and Same Day options demonstrate noticeably lower late delivery risks. These modes appear to be more reliable compared to others. In contrast, shipping options other than these, such as Second Class and First Class, show higher rates of late deliveries. This suggests that shipping mode plays a significant role in delivery performance, with certain modes being more dependable for on-time shipments. Focusing on the promotion or strategic use of Standard Class and Same Day shipping could help reduce overall late delivery risks.

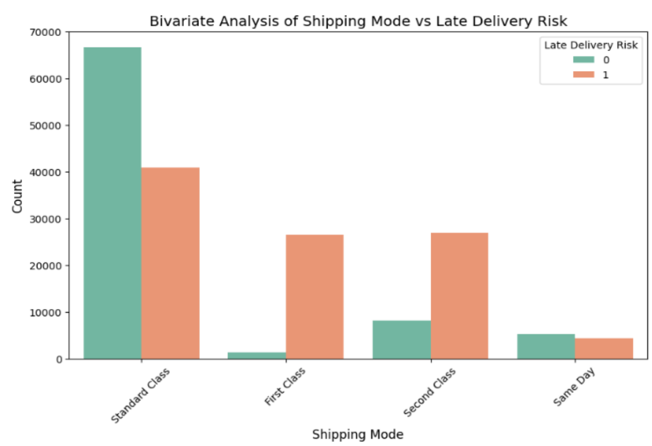


Figure 5

Delivery Status vs. Late Delivery Risk: Delivery statuses such as Advanced Shipping, Shipping on Time, and Shipping Canceled show no late delivery occurrences, maintaining a 0% late delivery risk. On the other hand, statuses marked as Late Delivery are associated with a 100% late delivery risk. This makes delivery status a critical factor in predicting late delivery outcomes.

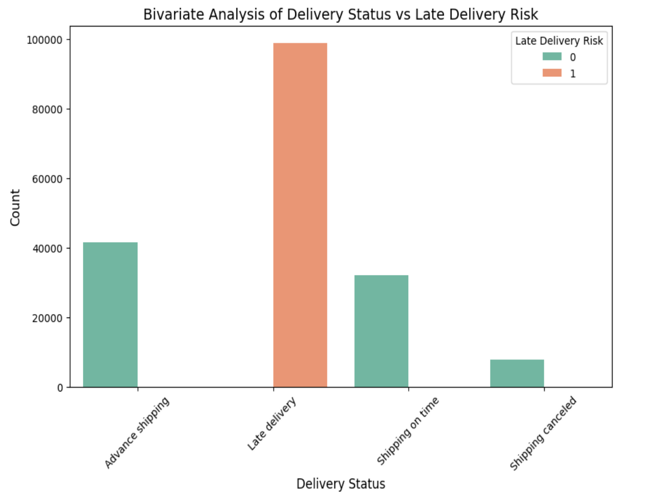


Figure 6

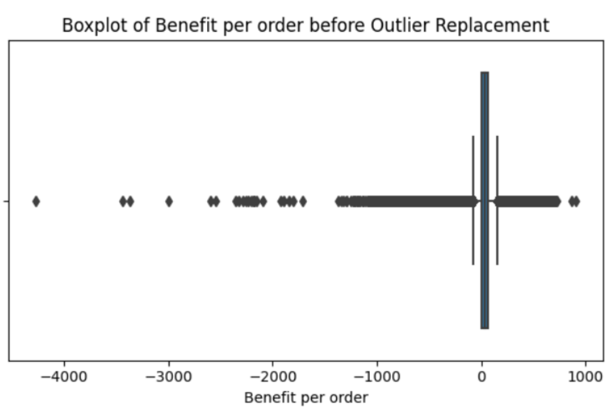
Outlier Detection and Handing in Numerical Features:

Outlier detection and handling is a critical step to ensure the uncertainity in the data which affects model performance. For this analysis, Z-score method preferred. The Z-score formula, [$Z=(X-\mu)/\sigma$]measures how far a data point is from the mean in terms of standard deviation. Values outside the range of −3-to 3 were classified as outliers [3]. Once identified, these outliers were replaced with the median of the respective feature to maintain the data's integrity.

Below are the details of the outliers detected for each numerical feature:

Benefit per Order: Outliers Detected are 3,608 (Figure 7 and Figure 8)

- Outliers : [−425.58,−783.67,595.35,415.80,−459.00,−447.05,−459.67,−540.79,−652.43,−790.42]



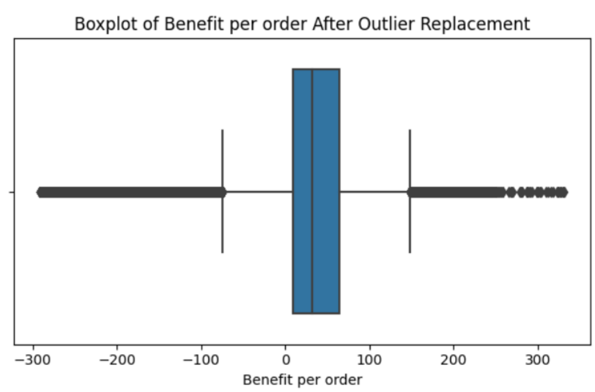


Figure 7

Figure 8

Sales per Customer: Outliers Detected are 477 (Figure 9 and Figure 10)

- Outliers: [1417.50,1395.00,1365.00,1200.00,989.99,1230.00,1275.00,1245.00,1125.00,1417.50]

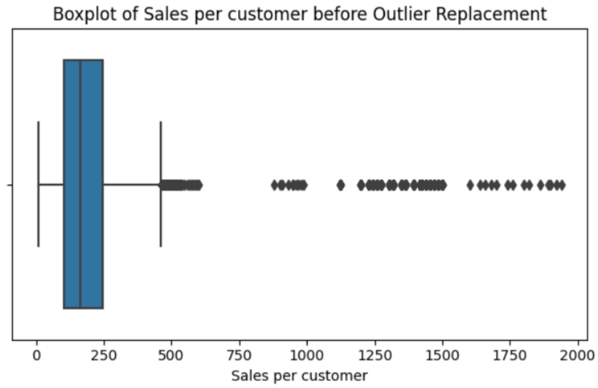


Figure 9

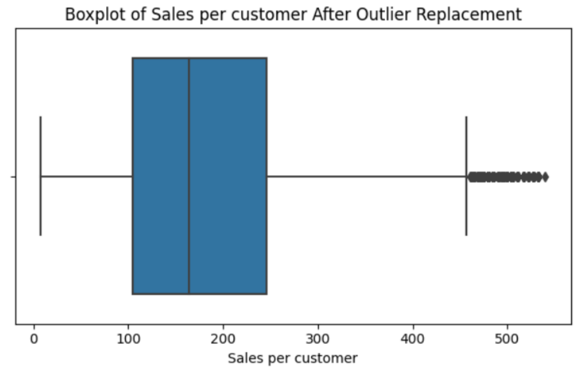


Figure 10

Order Item Discount: Outliers Detected are 2,106 (Figure 11 and Figure 12)

- Outliers: [112.49,105.00,135.00,300.00,113.01,124.99,99.99,89.99,89.99,89.99]

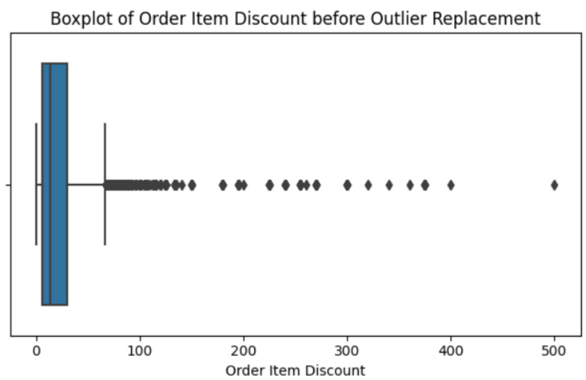


Figure 11

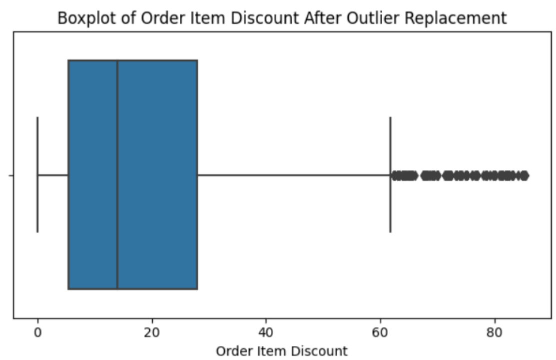


Figure 12

Order Item Product Price: Outliers Detected are 488 (Figure 13 and Figure 14)

- Outliers:[1500.00,1500.00,1500.00,1500.00,999.99,1500.00,1500.00,1500.00,1500.00,1500.00

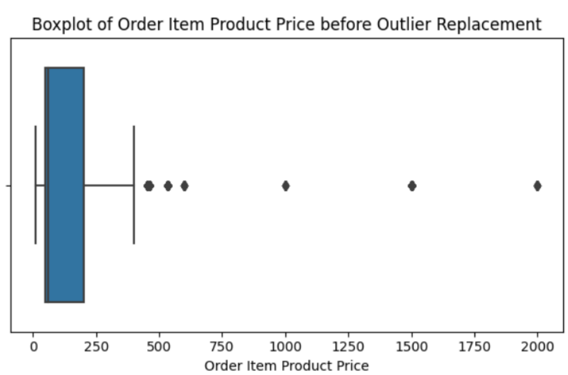


Figure 13

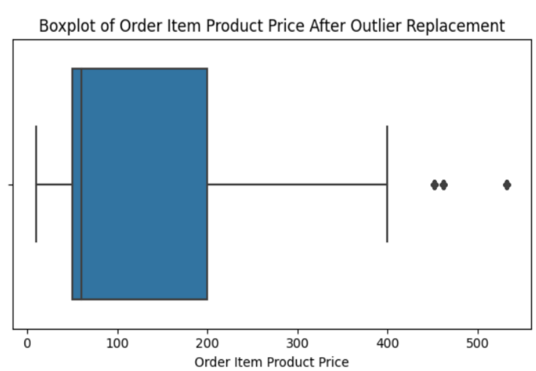


Figure 14

Order Item Profit Ratio: (Figure 15 and Figure 16)

- Outliers Detected: 6,013
- Outliers: [−1.33,−1.55,−1.60,−1.65,−1.70,−2.70,−1.55,−1.50,−1.55,−1.60]

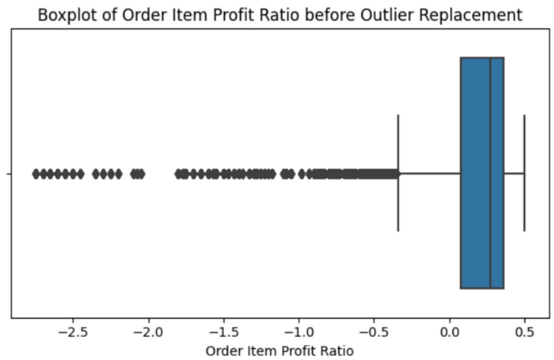


Figure 15

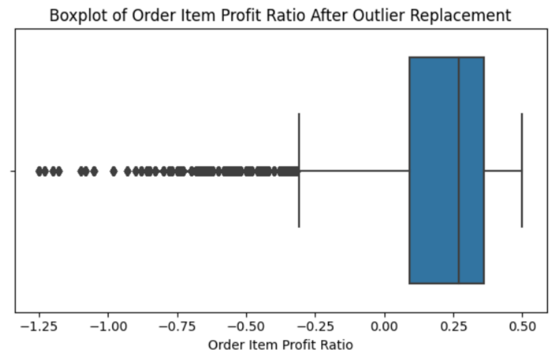


Figure 16

Sales: (Figure 17 and Figure 18)

- Outliers Detected: 467
- Outliers: [1000.00,1500.00,2000.00,1500.00,999.99,1500.00,1000.00,2000.00,2000.00,1500.00]

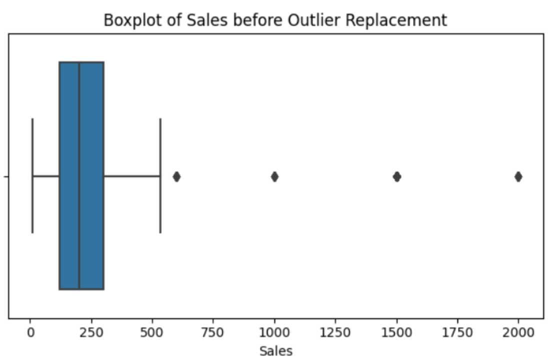


Figure 17

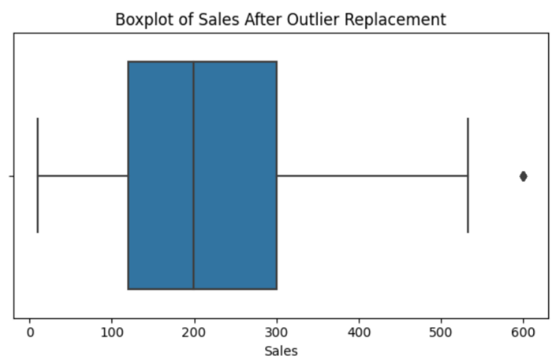


Figure 18

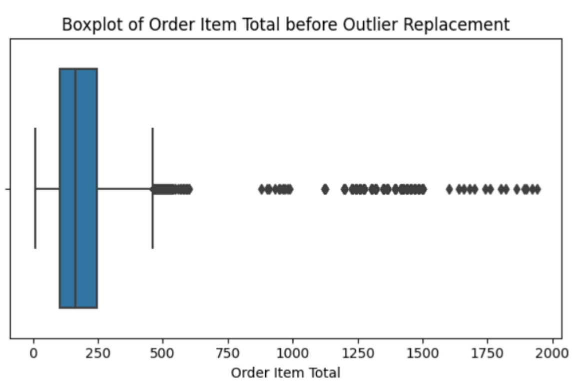


Figure 19

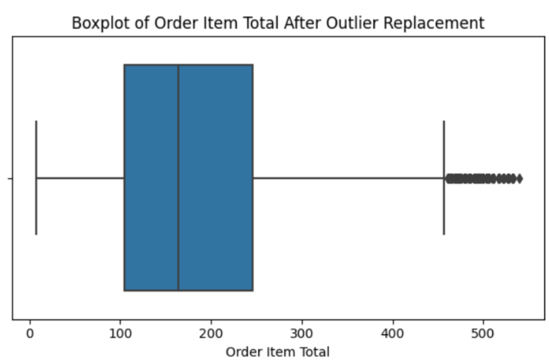


Figure 20

Order Item Total: Outliers Detected are 477 (Figure 19and Figure 20)

- Outliers: [1417.50,1395.00,1365.00,1200.00,989.99,1230.00,1275.00,1245.00,1125.00,1417.50]

Order Profit per Order: (Figure 21 and Figure 22)

- Outliers Detected: 3,608
- Outliers: [−425.58,−783.67,595.35,415.80,−459.00,−447.05,−459.67,−540.79,−652.43,−790.42]

Figure 21

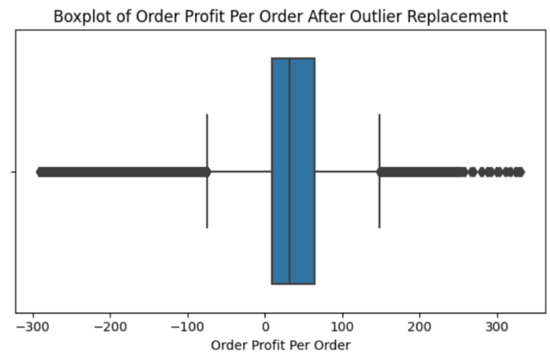


Figure 22

Product Price: (Figure 23 and Figure 24)

- Outliers Detected: 489
- Outliers: [450.80,500.00,1500.00,1500.00,999.99,1500.00,1000.00,1500.00,1500.00,2000.00]

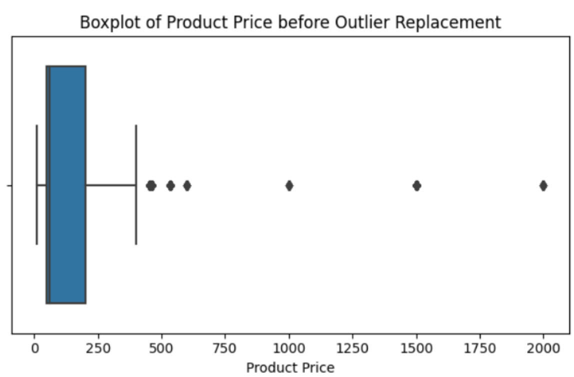


Figure 23

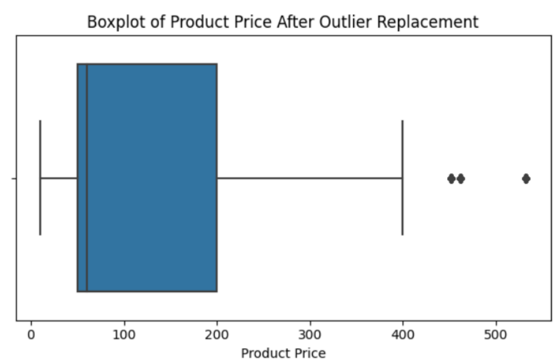


Figure 24

Order Item Discount Rate and Delivery Time:

- Outliers Detected: 0
- No outliers were found in this feature.

The replacement of outliers with the median ensures that extreme values do not distort the data while preserving the overall distribution [5]. This preprocessing step helps improve the robustness of subsequent analyses and model prediction

Feature Engineering of Categorical Variables

To improve the power of dataset in predicting target variable and extract meaningful patterns, several new features were engineered using the Order Date and Shipping Date[5]. These transformations aimed to provide insights into time-related factors that could influence delivery performance. Below is a summary of the feature engineering performed:

1. Date Components Extraction: Key components were extracted from both Order Date and Shipping Date to capture temporal patterns:

    - Year: In which year, the order placed
    - Month: In which month, the transaction made.
    - Day: The specific day of the month.
    - Weekday: Indicates the day of the week to account for potential operational differences.
    - Hour: The time the order or shipment was processed, providing insights into hourly trends.

2. Delivery Time Calculation: A new variable, Delivery Time (in hours), was calculated to measure the duration between the Order Date and Shipping Date. This metric is critical for understanding the time taken for order processing and shipping.

3. Delivery Time Categories: Delivery times were grouped into categories to simplify interpretation and highlight potential risk zones:

    - Very Short (0): Less than 20 hours.
    - Short (1): Between 20 and 50 hours.
    - Medium (2): Between 51 and 80 hours.
    - Long (3): More than 80 hours.

4. Time of Day Categorization: The time of day for both Order Date and Shipping Date was categorized into meaningful groups:

    - Morning (0): From 5:00AM to 11:59AM
    - Afternoon (1): From 12:00 PM to 4:59 PM.
    - Evening (2): From 5:00 PM to 8:59 PM.
    - Night (3): From 9:00 PM to 4:59 AM.

    These categories help understand whether the time of order or shipment affects delivery outcomes.

5. Weekend Flag: A binary variable was created to indicate whether the order was placed on a weekend (1) or a weekday (0). This feature helps assess whether operational or logistical constraints during weekends impact delivery performance.

These engineered features provide a richer representation of temporal data, enabling the model to identify time-related factors that influence the likelihood of late deliveries [5]. By incorporating these features, we aim to improve the overall accuracy and interpretability of the predictive model.

Statistical Analysis of Categorical Variables

To identify variables with a significant impact on late delivery risk, statistical tests are performed. A p-value less than 0.05 was used to know statistical significance relationship.

Methodology: We utilized the Chi-Square Test to examine the relationship between categorical variables and the target variable. For each variable, a contingency table was created, and the chi-square value, p-value, and degrees of freedom were calculated[4]. This method helped identify whether there was a meaningful association between the variable and late delivery risk.

Key Findings: Variables like Shipping Mode, Type, Shipping Hour, Order Hour, Shipping Time of Day, and Order Time of Day showed p-values below 0.05, confirming their significance in predicting late delivery risk.

These variables will be included in the predictive model to enhance its performance and reliability.

Categorical Variables	P-value
Type	5.128e-239
Customer Segment	0.59907
Shipping Mode	0.0001e-53
Shipping Weekday	0.755
Shipping Hour	5.8219e-76
Shipping Time of Day	0.00022
Order Weekday	0.633
Order Hour	5.5523e-89
Order Time of Day	2.589e-10
Order Year	0.354
Order Month	0.271
Order Day	0.388
Weekend	0.689
Shipping Year	0.849
Shipping Month	0.117
Shipping Day	0.869

Figure 25

Hot Encoding and Normalization Process

One-Hot Encoding: One-hot encoding was applied to convert them into numerical format for categorical variables. This transformation ensures compatibility with machine learning models:

- **Type**: Created new columns: Type_CASH, Type_DEBIT, and Type_TRANSFER to represent the categories.
- **Shipping Mode**: Created new columns: Shipping_Mode_Standard Class, Shipping_Mode_First Class, Shipping_Mode_Second Class, and Shipping_Mode_Same Day.

After encoding, original columns for **Type** and **Shipping Mode** were dropped to avoid redundancy and ensure model[2] compatibility.

Normalization of Numerical Variables: Once outliers were replaced with the median, the next step was to scale the numerical features for better model performance. Min-Max Scaling was used to transform values into a standardized range of [0, 1]. The Important numerical features which are there in outliers detection and handling in numerical features section[2]. This normalization process helps the model to train better .

Modeling and Insights

All seven models below in the Figure 26 were trained and used to predict late delivery risk using a 70-30 train-test split. The results show strong performance across the board, with ensemble methods generally outperforming simpler models.

Gradient Boosting, **LightGBM**, and **CatBoost** are top 3 more accuracies models. These models effectively handle complex data patterns, making them ideal candidates for further optimization. Simpler models like **Logistic Regression** and **Decision Tree** also delivered solid results[1], but their performance was slightly lower compared to the advanced ensemble techniques.

Model	Accuracy
Logistic Regression	0.9708
Decision Tree	0.9567
Random Forest	0.9740
Gradient Boosting	0.9752
XGBoost	0.9747
CatBoost	0.9749
LightGBM	0.9751

Figure 26

Top 10 Important Features of all Models:

To identify the most significant predictors, feature importance was calculated for each model. For **Logistic Regression**, coefficients were used, while for tree-based models, the feature_importances_ attribute was utilized. The top 10 features for each model were ranked based on their importance scores.

Common Features Across Models

Several features consistently appeared among the top predictors across all models, highlighting their critical role in finding delivery risk of product. The top 5 recurring features were:

- **Days_for_shipment_(scheduled)**
- **Days_for_shipping_(real)**

- **Type_TRANSFER**
- **Order_Hour**
- **Order_Item_Discount**

These features were identified as strong predictors, demonstrating their significant impact on model accuracy. Their repeated occurrence across models confirms their relevance and importance in the dataset.

Logistic Regression	Decision Tree	Random Forest	Gradient Boosting	XGBoost	CatBoost	LightGBM
Days_for_shipment_(scheduled)	Days_for_shipment_(scheduled)	Days_for_shipping_(real)	Days_for_shipment_(scheduled)	Days_for_shipment_(scheduled)	Days_for_shipping_(real)	Benefit_per_order
Type_Transfer	Delivery_Time	Delivery_Time	Days_for_shipping_(real)	Days_for_shipping_(real)	Delivery_Time	Days_for_shipping_(real)
Shipping_Mode_First_Class	Days_for_shipping_(real)	Days_for_shipment_(scheduled)	Delivery_Time	Type_TRANSFER	Days_for_shipment_(scheduled)	Sales_per_customer
Shipping_Mode_Second_Class	Benefit_per_order	Shipping_Mode_Standard_Class	Shipping_Mode_First_Class	Order_Item_Discount_Rate	Shipping_Mode_Standard_Class	Order_Hour
Shipping_Mode_Standard_Class	Type_TRANSFER	Shipping_Mode_First_Class	Shipping_Mode_Standard_Class	Shipping_Time_of_Day	Type_TRANSFER	Order_Item_Discount
Shipping_Mode_Same_Day	Order_Profit_Per_Order	Shipping_Mode_Second_Class	Shipping_Mode_Second_Class	Benefit_per_order	Shipping_Mode_Second_Class	Order_Item_Profit_Ratio
Shipping_Hour	Order_Item_Profit_Ratio	Order_Hour	Type_TRANSFER	Order_Item_Discount	Shipping_Mode_First_Class	Days_for_shipment_(scheduled)
Order_Hour	Order_Item_Discount	Shipping_Hour	Shipping_Mode_Same_Day	Shipping_Hour	Order_Hour	Order_Item_Product_Price
Order_Item_Profit	Order_Item_Total	Order_Profit_Per_Order	Order_Hour	Order_Item_Profit_Ratio	Shipping_Hour	Type_TRANSFER
Type_CASH	Sales_per_customer	Benefit_per_order	Order_Item_Discount	Order_Item_Product_Price	Type_PAYMENT	Order_Item_Discount_Rate

Figure 27

Impact of Removing Top 5 Common Features on Model Accuracy:

The graph highlights the impact of removing the most common features shared across all models, specifically Days_for_shipment_(scheduled), Days_for_shipping_(real), Type_TRANSFER, Order_Hour, and Order_Item_Discount. Removing the first feature results in only a slight drop in accuracy, particularly for robust ensemble models like Gradient Boosting, XGBoost, and LightGBM. However, as more common features are removed, especially the top three, a sharp decline in accuracy is evident, with simpler models such as Logistic Regression and Decision Tree being more affected.

Ensemble models, while more resilient, also experience performance degradation when more than three of these features are removed, demonstrating the importance of these common predictors. These results highlight the critical role of features like Days_for_shipment_(scheduled) and Type_TRANSFER in maintaining model accuracy, as their removal consistently impacts prediction performance across all models.

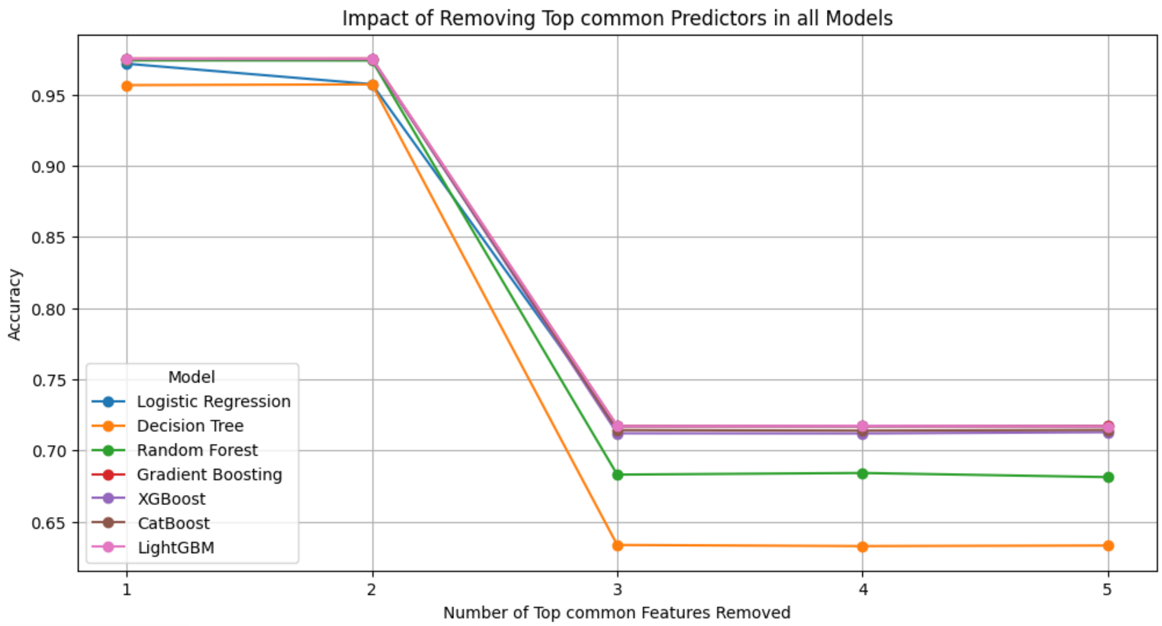


Figure 28

Impact of Predictor Adjustments on Late Delivery Risk Reduction:

This analysis focuses on evaluating the impact of modifying specific predictors by 10% on the late delivery risk for the top three performing models: Gradient Boosting, CatBoost, and LightGBM.

The baseline late delivery risk reflects the average predicted probability of late deliveries across the test data before any changes. After a 10% adjustment to key predictors, the models predict a new average risk, allowing us to measure the risk reduction resulting from the changes.

From the table, significant insights emerge.

- For Gradient Boosting, modifying Days_for_shipping_(real) reduced the late delivery risk by 12.54%, while changing Delivery_Time achieved a 10.04% reduction. Combining these two predictors led to an even greater risk reduction of 20.56%, highlighting the compounded effect of adjusting multiple predictors.
- Similarly, in CatBoost, adjusting both Days_for_shipping_(real) and Delivery_Time together resulted in the largest risk reduction of 20.59%, compared to smaller reductions when adjusted individually.
- For LightGBM, altering Days_for_shipping_(real) produced a significant risk reduction of 20.58%, while the combined adjustment with Benefit_per_order achieved the highest reduction of 20.74%. These findings demonstrate the pivotal role of certain predictors, such as Days_for_shipping_(real) and Delivery_Time, in minimizing late delivery risk and improving delivery performance predictions.

Impact of 10% Predictor Reduction on Late Delivery Risk

Model	Predictor	Change(%)	Baseline Risk(%)	Risk After Change(%)	Risk Reduction(%)
Gradient Boosting	Delivery_Time	10	54.94	49.31	10.04
Gradient Boosting	Days_for_Shipping(Real)	10	54.94	48.05	12.54
Gradient Boosting	Delivery_Time & Days_for_shipping_(real)	10	54.94	43.65	20.56
CatBoost	Days_for_shipping_(real)	10	55.00	44.08	19.87
CatBoost	Delivery_Time	10	55.00	53.24	3.20
CatBoost	Days_for_shipping_(real) & Delivery_Time	10	55.00	43.68	20.59
LightGBM	Benefit_per_order	10	54.96	54.84	0.21
LightGBM	Days_for_shipping_(real)	10	54.96	43.65	20.58
LightGBM	Benefit_per_order & Days_for_shipping_(real)	10	54.96	43.56	20.74

Figure 29

Conclusion

- **Key Predictors Drive Performance**: Features like **Days_for_shipping_(real)** and **Delivery_Time** consistently demonstrated significant influence on reducing late delivery risk across models.
- **Combining Predictors Amplifies Impact**: Adjusting multiple predictors together showed a compounded effect, leading to higher reductions in late delivery risk compared to changing them individually.
- **Model Resilience to Feature Loss**: Ensemble models such as **Gradient Boosting**, **CatBoost**, and **LightGBM** maintained accuracy better than simpler models when critical features were removed.
- **Optimization Potential**: Fine-tuning key predictors by small percentages (e.g., 10%) can lead to significant improvements in delivery performance, emphasizing the importance of predictor optimization.
- **Ensemble Models Are Most Effective**: Advanced models outperformed simpler ones, highlighting their ability to handle complex relationships between predictors and late delivery risk.

Future Work

- **Real-Time Data Integration**: Utilize live shipping and delivery data to enable dynamic updates to the model, enhancing real-time risk prediction and improving decision-making.
- **Incorporation of External Factors**: Include additional features like weather, traffic, and regional demand patterns to account for external factors that impact delivery timelines.
- **Scenario-Based Modeling**: Develop models capable of simulating disruptions or demand spikes to proactively identify high-risk situations and mitigate potential issues.

- **Adaptive Risk Thresholds**: Design flexible models with adjustable risk thresholds aligned with business objectives, such as optimizing costs or enhancing customer satisfaction.

DataSource Link: Link

Github Link: Github

References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from https://www.researchgate.net/publication/51969319_Scikit-learn_Machine_Learning_in_Python
2. Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Wiley-Interscience. Retrieved from https://www.researchgate.net/publication/268861577_An_Introduction_to_Categorical_Data_Analysis_Second_Edition
3. Shahbaz, M. S., Rasi, R. Z. R. M., & Ahmad, M. F. (2019). A novel classification of supply chain risks: Scale development and validation. *Journal of Industrial Engineering and Management (JIEM)*, 12(1), 201–218. https://doi.org/10.3926/jiem.2792
4. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. Springer Science and Business Media. Retrieved from https://www.researchgate.net/publication/354866394_An_introduction_to_statistical_learning_with_applications_in_R_
5. Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann. Retrieved from https://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems