**CAPSTONE PROJECT -TITLE**

LATE DELIVERY RISK PREDICTION IN SUPPLY CHAIN

MANAGEMENT  
USING MACHINE LEARNING

|  |  |
| --- | --- |
| **Batch details** | **PGPDSE BLR October22** |
| **Team members** | 1. Anand Kumar C S 2. Swastik Mohanty 3. Dhavala Narendra Kumar 4. Vaibhav S. Adhau |
| **Domain of Project** | Supply Chain Management |
| **Group Number** | 7 |
| **Mentor Name** | Mrs. Trupti Mathapati |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S.No | Topic | Page No |
| 1 | Business problem statement | 03 |
| 2 | Existing & Proposed Solution | 05 |
| 3 | Benefits of Using Model | 07 |
| 4 | Understanding the Data | 08 |
| 5 | Preprocessed Data Analysis | 10 |
| 6 | Sampling | 13 |
| 7 | Analysis of Sample | 18 |
| 8 | Base Model | 20 |
| 9 | Different Models Built and Their Results | 24 |
| 10 | Comparative Analysis of all the models | 26 |
| 11 | Business Interpretation | 35 |
| 12 | Limitation of Models | 36 |
|  |  |  |
|  |  |  |

# 

# 1. BUSINESS PROBLEM STATEMENT

## **1.1 Business problem understanding**

The business problem of late delivery risk prediction using machine learning is to identify orders that are at high risk of being delivered late to avoid potential customer dissatisfaction and revenue loss. By predicting late delivery risk, businesses can take preventive measures such as reallocating resources, optimizing delivery routes, and proactively managing customer expectations to ensure timely deliveries. This can improve customer satisfaction and retention while reducing the costs associated with late deliveries, such as shipping fees, order cancellations, and negative customer reviews.

Main challenges involved in Late Delivery Prediction are:

* Lack of visibility: Businesses may not have complete visibility into the supply chain, leading to delays and disruptions that can affect delivery times.
* Inaccurate demand forecasting: If businesses do not have an accurate forecast of demand, they may not have enough resources or inventory to fulfill orders on time.
* Traffic and weather conditions: External factors such as traffic congestion and severe weather can delay deliveries and make it difficult to meet customer expectations.
* Capacity constraints: If businesses do not have enough resources or capacity to fulfill orders, they may not be able to meet delivery deadlines.
* Poor communication: Inadequate communication between different departments or stakeholders involved in the delivery process can lead to delays and errors.
* Inefficient processes: Inefficient delivery processes such as manual data entry or inefficient routing can cause delays and increase the risk of late deliveries.
* Unexpected events: Unexpected events such as equipment failure or employee absence can disrupt the delivery process and lead to late deliveries.

## **1.2 Business Objective**

The main business objective of late delivery risk prediction using machine learning is to reduce the risk of late deliveries and improve customer satisfaction while minimizing the associated costs. By accurately predicting late delivery risk, businesses can take proactive measures to avoid delays, such as optimizing delivery routes, reallocating resources, and managing customer expectations. This can help businesses to improve customer satisfaction, reduce the number of order cancellations and returns, and ultimately increase revenue. The objective is to create a reliable and accurate predictive model that can be integrated into existing business processes to help businesses optimize their delivery operations and improve overall performance.



**2.EXISTING AND PROPOSED SOLUTION**

**2.1 Current solution to the problem:**

1. **Manual Forecasting:** One of the most common methods for predicting late delivery was to use manual forecasting methods, such as spreadsheets or paper-based records. This involved manually tracking shipments, analyzing delivery times and identifying potential risks. However, this method was often time-consuming, error-prone, and lacked the ability to capture complex patterns in the data.
2. **Business Intelligence Tools:** Some companies used Business Intelligence (BI) tools to help identify potential late delivery risks. These tools could analyze large amounts of data to identify patterns and trends, and generate reports and dashboards that could help supply chain managers make better decisions. However, these tools were often expensive, complex to implement and lacked the ability to predict future events with a high degree of accuracy.
3. **Rules-Based Systems:** Some companies developed rules-based systems that used a set of predefined rules to identify potential late delivery risks. For example, if a shipment was scheduled to arrive during a period of inclement weather, the system would flag it as high risk. However, these systems were limited in their ability to adapt to changing conditions and lacked the ability to learn from past experiences.
4. **Inventory Optimization:** Some companies focused on improving their inventory management practices to reduce the risk of late delivery. This involved using techniques such as just-in-time inventory management and safety stock optimization to ensure that inventory levels were optimized to meet demand while minimizing the risk of stockouts and late deliveries.

**2.2 Proposed solution to the problem:**

1. **Naive Bayes :** Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
2. **Decision Tree:** Decision trees are a type of machine learning model that uses a tree-like structure to make decisions. Each node in the tree represents a decision based on a specific feature, such as the carrier or transportation mode, and each leaf node represents a predicted outcome, such as whether a shipment will be delivered late or on time.
3. **Random Forest:** Random Forest is an ensemble model that combines multiple decision trees to improve accuracy and reduce overfitting. It randomly samples the data and features used to build each tree, and then aggregates the predictions of each tree to create a final prediction.
4. **Gradient Boosting:** Gradient boosting is another ensemble model that combines multiple weak models to create a stronger model. It trains a series of decision trees, each one learning from the errors of the previous tree, to make better and more accurate predictions.
5. **XGBoost:** XGBoost is an optimized implementation of gradient boosting that uses a combination of regularization, parallel processing, and tree pruning techniques to improve accuracy and speed. It is particularly useful for large datasets with many features and can handle both regression and classification problem

**3.BENEFITS OF USING THE MODEL**

1. **Improved delivery performance:** The model can identify orders at risk of being delivered late, allowing for proactive intervention to improve delivery times and customer satisfaction.
2. **Optimized resource allocation:** By identifying the root cause of late deliveries, the model can help companies allocate resources more effectively and efficiently, reducing costs and improving delivery times.
3. **Reduced manual effort:** The model can automate the process of identifying and prioritizing late delivery risks, reducing the need for manual intervention and saving time and effort.

## **4. UNDERSTANDING THE DATA**

### 

### **4.1** **Data Dictionary:**

|  |  |
| --- | --- |
| **Fields** | **Description** |
| Type | Type of transaction made |
| Days for shipping (real) | Actual shipping days of the purchased product |
| Days for shipment (scheduled) | Days of scheduled delivery of the purchased product |
| Benefit per order | Earnings per order placed |
| Sales per customer | Total sales per customer made per customer |
| Delivery Status | Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time |
| Late\_delivery\_risk | Categorical variable that indicates if sending is late (1), it is not late (0). |
| Category Id | Product category code |
| Category Name | Description of the product category |
| Customer City | City where the customer made the purchase |
| Customer Country | Country where the customer made the purchase |
| Customer Email | Customer's email |
| Customer Fname | Customer name |
| Customer Id | Customer ID |
| Customer Lname | Customer last name |
| Customer Password | Masked customer key |
| Customer Segment | Types of Customers: Consumer, Corporate, Home Office |
| Customer State | State to which the store where the purchase is registered belongs |
| Customer Street | Street to which the store where the purchase is registered belongs |
| Customer Zipcode | Customer Zipcode |
| Department Id | Department code of store |
| Department Name | Department name of store |
| Latitude | Latitude corresponding to location of store |
| Longitude | Longitude corresponding to location of store |
| Market | Market to where the order is delivered: Africa, Europe, LATAM , Pacific Asia , USCA |
| Order City | Destination city of the order |
| Order Country | Destination country of the order |
| Order Customer Id | Customer order code |
| order date (Date Orders) | Date on which the order is made |
| Order Id | Order code |
| Order Item Card prod Id | Product code generated through the RFID reader |
| Order Item Discount | Order item discount value |
| Order Item Discount Rate | Order item discount percentage |
| Order Item Id | Order item code |
| Order Item Product Price | Price of products without discount |
| Order Item Profit Ratio | Order Item Profit Ratio |
| Order Item Quantity | Number of products per order |
| Sales | Value in sales |
| Order Item Total | Total amount per order |
| Order Profit Per Order | Order Profit Per Order |
| Order Region | Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern Asia, West Asia , West of USA , US Center , West Africa, Central Africa ,North Africa ,Western Europe ,Northern , Caribbean , South America ,East Africa ,Southern Europe , East of USA ,Canada ,Southern Africa , Central Asia , Europe , Central America, Eastern Europe , South of USA |
| Order State | State of the region where the order is delivered |
| Order Status | Order Status : Complete , Pending , Closed , Pending payment ,Cancelled, Processing ,Suspected fraud ,On\_hold,Payment\_review |
| Product Card Id | Product code |
| Product Category Id | Product category code |
| Product Description | Product Description |
| Product Image | Link of visit and purchase of the product |
| Product Name | Product Name |
| Product Price | Product Price |
| Product Status | Status of the product stock: If it is 1 not available, 0 the product is available |
| Shipping date (Date Orders) | Exact date and time of shipment |
| Shipping Mode | The following shipping modes are presented : Standard Class , First Class , Second Class , Same Day |

### **4.2 Shape of The Dataset:**

Graphical user interface, text, application

Description automatically generated

### **4.3 Variable Categorization:**

|  |  |
| --- | --- |
| **Variables** | **Count** |
| Numerical Variables | 28 |
| Categorical Variables | 24 |
| Target Variables | 1 |
| Total Variables (Columns) | 53 |

**5.PRE-PROCESSED DATA ANALYSIS**

A picture containing graphical user interface

Description automatically generated

**5.1 Variables:**

The above picture shows Datatype, unique values count, count of null values of each and every column.

**Graphical user interface, application, table, Excel

Description automatically generated5.2 Summary of Data:**

## **5.3 Dropping Redundant Features:**

|  |  |
| --- | --- |
| **Feature Dropped** | **Reason for Dropping** |
| **Product Description** | It contained **100% null values**. This column did not provide any useful information for analysis and was therefore deemed irrelevant. |
| **Order Zipcode** | It contained **86.23% null values**. This column did not provide any useful information for analysis and was therefore deemed irrelevant. |
| **Customer Email, Customer Password, Customer Fname & Customer Lname** | They were not expected to have a significant impact on predicting the target variable 'late delivery risk'. These columns were deemed irrelevant to the analysis. |
| **Benefit per order** | It is giving the same information as "order profit per order". |
| **Category Id and Category Name** | It is giving the same information about "Product Category ID". |
| **Order Customer ID** | It is giving the same information as "Customer Id" |
| **Sales per customer** | It is giving the same information as "Order Item Total" |
| **Department Name** | It is giving the same information as "Department ID" |
| **Customer Street** | The information it contained was redundant with other columns such as 'Customer Zip code', 'Customer City', 'Customer State', and 'Customer Country'. |
| **Product Images** | They do not provide any direct information related to the delivery process |
| **Order date and Shipping date** | They contained redundant information with the columns'Days for shipping(real)' and'Days for shipment(scheduled)'.The columns 'Days for shipping (real)' and 'Days for shipment (scheduled)' for analysis were retained as they provided more relevant information about the shipping and delivery process. |
| **Product Status** | All values in "Product Status" are zero so given feature is not giving any information. |
| **Order Id, Order Item Id, & Product Card Id** | They were not expected to have a significant impact on predicting the target variable 'late delivery risk'. These columns might have been redundant or irrelevant to the analysis and could also have had data quality issues. |

## **6.DATA PREPROCESSING**

### **6.1 Null Values:**

Table

Description automatically generated

From the above table we can able to infer that the column product Description and order Zipcode are filled with 100 and 86percent null values so we are dropping that columns.

### **6.2 Outliers Treatment:**

**Outliers:** Outliers is an observation which deviates so much from the other observations, that it become suspicious that it was generated by different mechanism or simply by error.

### **Reason for Outliers:**

1. Variability in the data
2. Experimental measurement errors

### **Impact of Outliers on Dataset:**

1. Causes problems during statistical analysis
2. Effects mean and standard deviation

### **6.3 Univariate Analysis for Numerical Variables:**

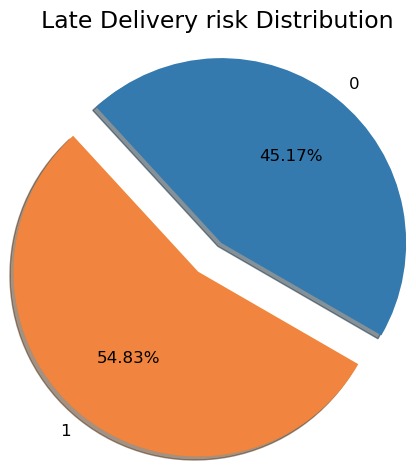
### Graphical user interface, histogram Description automatically generated

Graphical user interface, chart, histogram

Description automatically generated

All the **Numerical Columns**which we have in our Dataset is **Fully Skewed either right or left** . We are going with statistical tests to verify whether these columns are giving information about Predicting the target. After that we need to Remove the Outliers by IQR method, capping or based on the Business understanding.

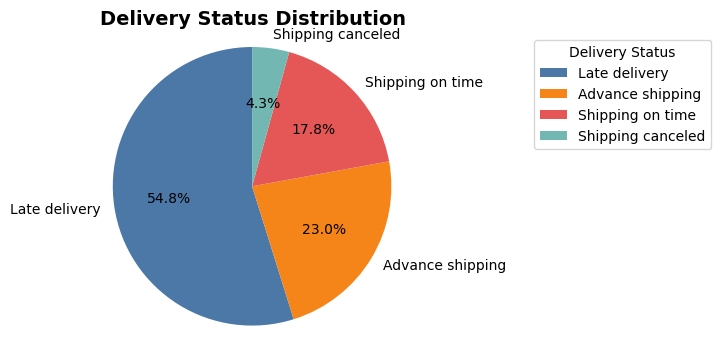
### **6.4 Univariate Analysis for Categorical Variables:**



Class 1 (Late delivery risk) - 54.8%

Class 0 (Late delivery risk) - 45.2%

Our **target variable**is balance it ensures that the model is exposed to a similar number of examples from each class during training, and this can help prevent biases towards one class and improve the overall performance of the model.



Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

**6.5 Distribution of Top 20 and least 10 values for**

**Customer city and Order City:**

Chart, histogram

Description automatically generated

**Inferences:**

* Most of the customers are situated in **Caguas City** who ordered more and customers which situated in **CA** who are order less.
* Most of the customers order from **Estados Unidos** and very less customers are order form **Serbia and Burundi .**

**6.SAMPLING**

**6.1 Simple Random Sampling:**

Simple random sampling is a type of probability sampling in which the samples will be randomly select from population.

**6.2 Mann Whitney U-hypothesis test :**[**¶**](http://localhost:8888/notebooks/Desktop/population_data2.ipynb#MannWhitney-U-hypotheisis-test-:)

In numerical analysis for population we see that all the numerical variable are skewed i.e. **data is not normal.**so we are not able to use parametric test like t-test two independent sample test so we are using **MannWhitney-Utest**for checking is the population mean is equal to sample mean

* **H0:** Population mean = Sample mean
* **H1:** Population mean != Sample mean

**6.3 Chi-square goodness of fit test:**

For categorical feature we have to check the distribution of all types in each categorical feature in sample and population will be same so for that we are using the Chi Square goodness of fit

* **H0:** There is no significance difference between expected values(from population) and observed values (from sample)
* **H1:** There is significance difference between expected values(from population) and observed values (from sample)

Expected values : Distribution of the categorical data in Population.

Observed values : Distribution of the categorical data in Sample.

By using the Simple Random sampling technique we are generating 30 different random sample for each sample size(5k,10k,50k….) and by doing the **‘Mann Whitney U’** test and **‘chi square goodness of fit’** for each generated sample. And calculating the average of no of columns passing both the test for every sample size for **30 iterations**. From this we make a plot and we can clearly able to see there is too much hike between the range of 5k to 20k and again from 20k to 45k. After that we can able to see there is no hike in the average which is indicates that all the columns are passing the both test at each iteration.

Chart, line chart

Description automatically generated

**6.4 Population:**

Graphical user interface

Description automatically generated

**6.5 Sample:**

Table

Description automatically generated with medium confidence

**7.ANALYSIS OF SAMPLE**

**7.1Bivariate Analysis:**

Bivariate analysis is carried out between categorical variables and target Variable.

**7.1.1 City vs Late\_delivery\_risk:**

**Chart, bar chart

Description automatically generated**

* When the order comming from the **New York City and Santo Dommingo**most of the delivery are late.
* Anyhow most of the Customer from the **Caguas city** so at that city late dilvery risk is high.

**7.1.2 Shipping Mode vs Late\_delivery\_risk:**

Chart, bar chart

Description automatically generated

* **Standard Class** is the most popular shipping mode
* Late delivery is even observed in **first class** shipping mode.

**7.1.3 Customer Segment vs Late\_delivery\_risk:**

**Chart, bar chart

Description automatically generated**

* **28.3%**late delivery to Consumer and **23.5%**are not late delivery to Consumer
* **16.6%**late delivery to Corporate Customer Segment and **13.8%** are not late delivery to Corporate Customer Segment
* **10.0%**late delivery to Home Office and **7.9**are not late delivery to Home Office

**7.1.4 Market vs Late\_delivery\_risk:**

**Chart, bar chart

Description automatically generated**

* Same number of percentage **(approx. 15.5%)**of late delivery from EUROPE and LATAM
* Less number of percentage **3.5%**of late delivery from Africa because anyhow less number of delivery are coming from Africa.

**7.1.5 Customer Country vs Late\_delivery\_risk:**

**Chart, bar chart

Description automatically generated**

* **33.8%**late delivery going to EE.UU country and **27.8%**are not late delivery.
* **21.0%**late delivery going to Puerto Rico and **17.4%**are not late delivery.

**7.2 Statistical Hypothesis test with target variable**

**7.2.1 Chi2 Contingency test :**

* **H0 : X and Y(Late delivery risk) are independent.**
* **H1 : X and Y (Late delivery risk) are dependent.**
* Using above test we compare the catogorical feature vs target variable and above test giving that which feature is dependent on target variable.

**7.2.2 Mann–Whitney U test:**

Here we used the Mann-Whitney U test to compare the mean of Numerical columns with respective target variable (“Late\_delivery\_risk”)If mean are same then we are able to say that the numerical columns is Insignificant with target variable.

* H\_0: The numerical mean ( 0 ) is equal to numerical mean ( 1 )
* H\_1: The numerical mean ( 0 ) is not equal to numerical mean ( 1 )

Graphical user interface, application

Description automatically generated

### **Checking For Class Imbalance:**

**Target Variable: Late Delivery Risk**

Chart, bar chart

Description automatically generated

Here we are able to see that our target variable is not very much imbalanced.

* **class 1 (Late delivery risk) - 54.8%**
* **class 0 (Late delivery risk) - 45.2%**

Our **target variable**is nearly balanced it ensures that the model is exposed to a similar number of examples from each class during training, and this can help prevent biases towards one class and improve the overall performance of the model.

**8.BASE MODEL**

### **8.1 Encoding Before Building of Base Model:**

* Target encoding (probability encoding all except type and shipping mode as dummy)
* Models implemented: Logit, Naive bayes ,Decision tree and Random forest

**8.2 Logistic Regression as Base Model :**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated with low confidence

Chart, line chart

Description automatically generated

Possibly complete quasi-separation: A fraction 0.11 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

From Business point of view we need to get a good recall score because when we predict its an late delivery risk and its not then not a problem but when we say its not a late delivery risk and its actually a late delivery then its a problem.

So, we try to focus mainly on recall score and not on accuracy

**9.DIFFERENT MODELS BUILT AND THEIR RESULTS**

**9.1 Naive Bayes:**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

**9.2 Decision Tree:**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

**9.3 Random Forest**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

**9.5 AdaBoost:**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

**9.6 Gradient Boost:**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

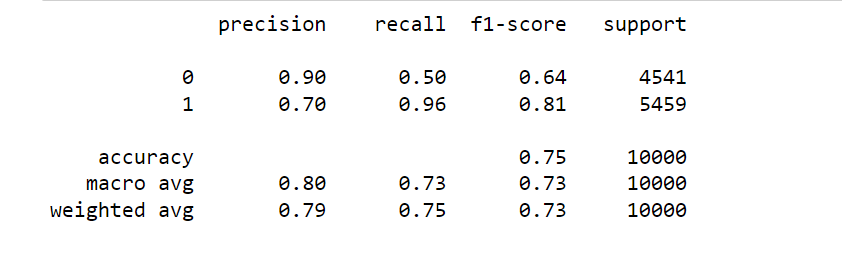
Chart, line chart

Description automatically generated

**9.7 XG Boost:**

Chart, treemap chart

Description automatically generated



Chart, line chart

Description automatically generated

**10.COMPARITIVE ANALYSIS OF ALL MODELS**

The following table shows the comparative analysis of results of all the models built.

**10.1 Summary For All Models Built(Models-1):**

Table

Description automatically generated

### **10.2 Encoding :**

**Target** **encoding**with features was done and then plugged probability of positive class to each type in category.

**10.3 Analysis of Models after Encoding(Models-2):**

**Table

Description automatically generated**

### **10.4 Encoding :**

Encoding similar to previous one was did with feature engineering with K-Mens clustering

And models’ results are again analysed.

**10.5 Analysis of Models after further Encoding(Models-3):**

Chart, line chart

Description automatically generated

* 1. **Insights:**

### Models- 1 gives better results than Models 2 and Models 3

### In Models 1, **Xg boost**is best model and generalised model.

* 1. **Hyperparameter Tunning :**

Results for hyper parameter tuning are as follows:

**Best Hyperparameters:**

1. 'subsample': 0.9500000000000004
2. 'n\_estimators': 400
3. 'max\_depth': 4
4. 'gamma': 1
5. Best Score: 0.8615
   1. **XGBoost before Hyper Parameter Tuning:**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated with medium confidence

Chart, line chart

Description automatically generated

From above model we got to know that Xg boost is consistent and generalized model and to reduce variance we used first random forest but after hyper parameter tunning the variance not decrease so we did hyper parameter tunning on Xgboost and the model given the generalized model and in given model recall score also 86% to predict late delivery class.

* 1. **XGBoost After Hyper Parameter Tuning:**

Chart, treemap chart

Description automatically generated

Table

Description automatically generated

**Chart, line chart

Description automatically generated**

Based on the evaluation metrics, it can be concluded that the **XGBoost** model has performed well in predicting the target variable **'Late\_delivery\_risk'**. The model has achieved an overall accuracy of **85%** on the **test data**, which indicates that **85%** of the predictions made by the model were correct.

The precision of the model is high for both the classes, which means that the model has a low false positive rate. This is important as predicting that a delivery will be late when it is not, can result in unnecessary costs and damage to the reputation of the business.

The recall of the model is also good for both the classes, with a higher recall for the class 1, which indicates that the model has a low false negative rate. This is important as predicting that a delivery will not be late when it actually is, can result in dissatisfied customers and damage to the reputation of the business

**11. Business Interpretation :**

The target variable in this model is 'Late\_delivery\_risk', which is an important metric for businesses involved in the delivery of goods to customers. Late deliveries can have a negative impact on customer satisfaction, which can ultimately harm the reputation of the business and result in lost revenue.

With the help of this model, businesses can identify the factors that are contributing to late deliveries and take corrective measures to improve the delivery process. The model can also be used to predict the risk of a delivery being late, which can help businesses prioritize their resources and take proactive steps to prevent delays.

For instance, based on the features included in the model, businesses can identify the customer segments that are most likely to experience late deliveries, the market segments that are most prone to delays, and the shipping modes that are most likely to result in late deliveries. This information can be used to optimize the delivery process and reduce the risk of late deliveries.

In addition, the model can also be used to identify the regions or states where late deliveries are most common. This can help businesses to focus their efforts on these areas and take steps to improve the delivery process, such as increasing the number of delivery personnel or improving the transportation infrastructure.

Overall, the model can provide valuable insights to businesses involved in the delivery of goods and help them improve the efficiency of their delivery process, which can ultimately lead to increased customer satisfaction and improved business performance.

**12. Limitations of the model:**

**Limited feature set:** The model uses a limited set of features to predict the target variable 'Late\_delivery\_risk'. There may be other important factors that contribute to the risk of late delivery, which are not included in the model. Therefore, the model may not capture the full complexity of the problem.

**Lack of temporal data:** The model does not take into account the temporal nature of the data. Delivery patterns and customer behavior may change over time, and the model may not be able to capture these changes.

**Model Interpretability:** XGBoost is a complex model, which can make it difficult to interpret the results and understand how the model is making its predictions. This can make it challenging to identify the root causes of late deliveries and take corrective actions.

**Lack of external factors**: The model does not take into account external factors that may impact the delivery process, such as weather conditions, traffic congestion, or transportation disruptions. Therefore, the model may not be able to fully capture the complexity of the problem and may lead to inaccurate predictions.