

## INTRODUCTION

- In a globalized market, the continuous shipment of goods is something routinary. Economies depend on the import and export of goods not only for commercial reasons but also to sustain the growing market demands of agricultural goods and raw materials. Maritime companies and air shipping companies face constant and steady growth. According to statistics by IATA ( <https://tinyurl.com/IATA-Cargo-Analysis>) the air cargo industry closed 2023 with “strong momentum”. The demand for air cargo reached 22.8 billion cargo ton kilometers. On the other hand, the Global maritime trade increased 2.4% to 12.3 billion tons in 2023, according to the Review of Maritime Transport 2024 published by the United Nations Trade and Development Department (UNCTAD, <https://unctad.org/publication/review-maritime-transport-2024>). In many cases the demand is exceeding the available supply of air and maritime shipment, this is caused not only by the constant growth but by the constraints and bottlenecks that some routes represent, such as the Panama or the Suez Canal.
- Companies and individuals are looking at options, not only to protect their cargo against damage or loss, but also against delays. Logistic delay insurance might represent a lucrative business for insurance companies, as there is a constant growth in the export and import of goods, but for companies to truly obtain benefits, the policies and pricing of these policies must be reviewed.
- We will perform such analysis of insurance premium and freight delay risk utilizing machine learning models to predict the likelihood of a company paying for late delivery claims, with only information available at the time of purchase.

## Data Preparation

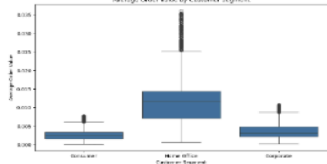
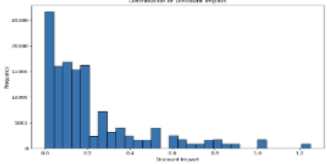
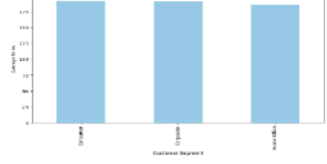
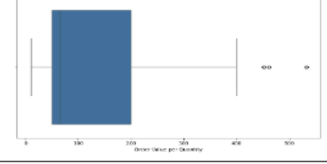
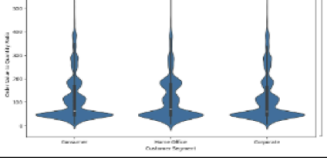
Once the data was loaded, we proceeded with data preprocessing.

The process began with data exploration , emphasizing examination of the categorical(object) and numerical columns. Within exploration, two lists titled “remove\_cat” and “remove\_cts” were created for the respective removal of redundant categorical and continuous features. Redundancy removal prepared the model for data cleaning which involved checks for missing values using `isna().sum()`, duplication observation using `duplicated().any()`, and removal of the previously identified features. The cleaned dataset returned with 180,519 rows and 21 columns.

To easily identify influential factors and relationships between the numerical variables we visually employed correlation analysis. The final step of data preparation consisted of applying one-hot encoding to the categorical variables “ Type” ,and “Shipping Mode”. The encoded columns were then converted into 0s and 1s to facilitate efficient integration in the model process. A box plot then visualized the distribution of the contiguous variables and highlighted the potential outliers. Two methods were involved to detect outliers: “Z-Score Method” and “IQR”. These outlier detection methods served to address any anomalies that could potentially negatively affect analysis and replaced the identified outliers with the respective median columns value.

## ENGINEERED FEATURES

- After the initial data cleaning and performing Exploratory Data Analysis. We are left with 27 features. It is important to note that to continue our analysis we will exclude all features that are not available at the time of purchase, such as: days for shipping (real), delivery status, late delivery risk and shipping date.
- With the available features, we proceed to the creation of new engineered features that will help us capture factors that might influence delays.

ENGINEERED FEATURE	FORMULA	IMPORTANCE	GRAPH	NOTES
<b>AVERAGE CUSTOMER ORDER VALUE</b>	$\text{Average\_Order\_Value} = \text{Sales per customer} / \text{Total Orders per Customer Segment}$	This feature represents the average value of orders by customer segment. This can represent how valuable is each customer segment for the company. Orders with higher values or certain customer segments can receive priority among orders to avoid delays.		From the boxplot we can easily view the distribution of order values in different customer segments. We can see that the home office segment places higher value orders.
<b>DISCOUNT IMPACT</b>	$\text{Discount Impact} = \text{Order Item Discount Rate} * \text{Order Item Quantity}$	Through these feature we want to evaluate if orders with a high discount experiences delays, due to lack of priority. As discounted orders amount for smaller profit margins companies might treat these orders as "last" priority.		The histogram shows us the distribution of Discount Impact, this helps us visualize if higher discounts are something constant or occasional and if there is a correlation between this feature and delay on orders. The majority of discounts are between 0.4, this suggest that most orders have a low discount.
<b>AVERAGE SALES PER CUSTOMER SEGMENT</b>	$\text{Avg\_Sales\_Segment} = \text{Average of Sales per Customer Segment}$	This engineered feature shows the revenue for each customer segment. It can be possible that segments that represent a higher revenue are prioritized over segments that signify a smaller revenue.		This bar plot provides an easy way to compare sales per customer segment. The graph shows no relevant difference between segments, this suggests that segment prioritization is less probable.
<b>ORDER VALUE PER QUANTITY</b>	$\text{Order\_Value\_per\_Quantity} = \text{Sales} / \text{Order Item Quantity}$	Through this feature, we are calculating the average value of each item for an order. As we have explained in the previous features, higher "ticket" might lead an organization to prioritize an order, decreasing the late delivery risk.		From the boxplot we can see that most values are clustered below 200, although there are some outliers at 500. The outliers represent orders with a high value per item. This means that these type of orders might get prioritized as they represent higher revenue.
<b>ORDER VALUE TO QUANTITY RATIO</b>	$\text{Order\_Value\_to\_Quantity\_Ratio} = \text{Order Item Total} / \text{Order Item Quantity}$	This feature measures the order value by the amount of items in an order. It can help us distinguish if orders with a high value but with low quantity items are giving a higher priority because they are "easier" to handle and represent higher revenue.		As we can see in the violin plot, most values are graphed below 100. We can see that some ratios in the corporate and consumer segments extend beyond 100, this might signify an increased priority level for high value and low quantity orders.

## MODEL ASSUMPTIONS

The machine learning model we used to predict the likelihood of a company paying for late delivery claims was a random forest. To do so, we made some assumptions as follows:

- Observations are independent of one another
- The dataset is accurate

## MODEL BUILDING

- For the creation of our model, we decided to include the following variables:
- **Original Features:** 'Days for shipment (scheduled)', 'Benefit per order', 'Sales per customer', 'Order Item Discount', 'Order Item Profit Ratio', 'Order Item Quantity', 'Sales', 'Order Item Total', 'Order Profit Per Order'.
- **Engineered Features:** 'Average Order Value', 'Discount\_Impact', 'Avg\_Sales\_Segment', 'Order\_Value\_per\_Quantity', 'Order\_Value\_to\_Quantity\_Ratio'.
- The chosen features can lead to a good model, as they are focused not only on financial performance factors but also on logistic factors, both of which can lead to risk delay.
- Order prioritization is also addressed through features like 'sales per customer', 'average sales segment', and 'Benefit per order'.
- Additionally engineered features like 'Discount Impact' and 'Order\_Value\_to\_Quantity\_Ratio', will provide our model with new insights and depth that did not appear in the original features.

## RANDOM FOREST CLASSIFIER

- For our model we decided to use Random Forest Classifier, we choose this model because we are assuming that all clients buy the policy and because a claim and subsequent payment occur no matter the amount of days a shipment is delayed (1 or many). This classifier algorithm is a good choice because it can: handle mixed data types, capture complex nonlinear interactions, handle outliers and redundant features.
- Our model performs as follows:

Classification Report				
	precision	recall	f1-score	support
0	0.57	0.59	0.58	10899
1	0.69	0.67	0.68	15220
accuracy			0.64	26119
macro avg	0.63	0.63	0.63	26119
weighted avg	0.64	0.64	0.64	26119
Overall Accuracy:		0.637547		

**PRECISION**, The model is better at predicting delayed orders, this is helpful as the company will pay premiums on delayed orders and having better predictions will allow the company to manage payments better.

**RECALL**, our model is better at identifying delayed orders (as it shows a higher recall number for class 1 than class 0).

**ACCURACY** percentage shows that the model can correctly classify orders delay or non-delayed 63.75% of the time.

**CALCULATING EXPECTED CLAIMS.** Using our model, we proceed to calculate claims for the testing dataset:

- **Claims: \$55.84**

**Formula for claims = Probability of late delivery \* 100**

This value represents the average amount the insurance company expects to pay out in claims per policyholder due to late deliveries. So, on average, for every insurance policy, the company anticipates having to pay out about \$55.84 when considering the likelihood of claims.

- **Calculated Premium: \$69.13**

**Formula for premium = (expected claims + fixed expense) \* (1+profit premium).**

This is the premium that the insurance company should charge each policyholder to cover the expected claims, fixed expenses, and profit loading.

The calculation include:

- Expected Claims (\$55.84): The anticipated payout for each policyholder.
- Fixed Expense (\$10): A constant cost incurred for every policy sold, regardless of whether a claim is made.
- Profit Premium (5%): An additional charge added to ensure the company makes a profit.

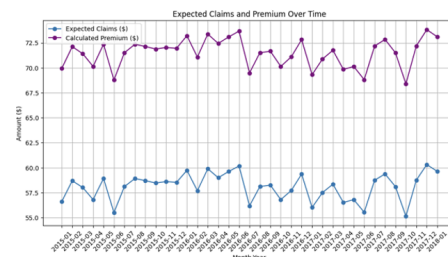
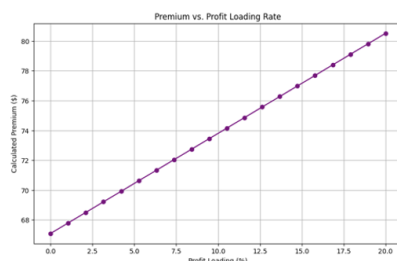
The calculated premium of \$69.13 is higher than the expected claims of \$55.84. This shows that the insurance company is pricing the policy appropriately to cover potential payouts while also making a profit.

#### Expected Claims and Premiums over Time

- This line chart illustrates premium changes with different loading rates, ranging from 0% to 20% profit premium. We can see a linear increase between these two variables.

#### Adding expected claims and premium as example values for each month

- This graph shows us the expected claims and the premium calculations through time.



## CONCLUSIONS

Through this project our goal was to build a model that predicted freight delay risk, which would allow for personalized insurance premium for customers based on likelihood of delay of shipments.

A random forest classifier (with original and engineered features), allowed us to build a model that classifies custom freight delay risk and leads to individual premium calculations.

The model and formula applications allows us as actuaries to set premiums rates that:

1. Generate financial profitability for our company.
2. Price personalization for customers based on their risk.