# **DataCo Global / Shipping Times Predictions**

*A year’s worth of worldwide retail sales and shipping information provided by Kaggle.com is the basis of this evaluation. The data consists of over 150,000 sales events which tracks customers, products ordered, destination, the status of orders, and the status of delivery amongst numerous demographic and payments information. The most obvious problem with operations is that over 50% of all product deliveries are late. The initial focus of this analysis is on predicting which orders need special attention in regards to potential late delivery.*

## **1. Data**

The data set is found in a downloadable csv file on the Kaggle website at the below link:

<https://www.kaggle.com/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis>

## **2. Data Cleaning**

Evaluating the data set found that nearly every field was populated for every row. There were some conspicuous anomalies in the Product Description, which had zero entries and an obvious shortcoming considering that the database is for a global concern, the number customer and order zipcode fields were only populated for orders from and for the US.

There were a number of categorical columns, such as shipping mode, shipping status, order status, late delivery risk, category\_name, customer state, order state, customer region, order region, customer country, order country, department name, product name, product category, product status.

There were no duplicated rows.

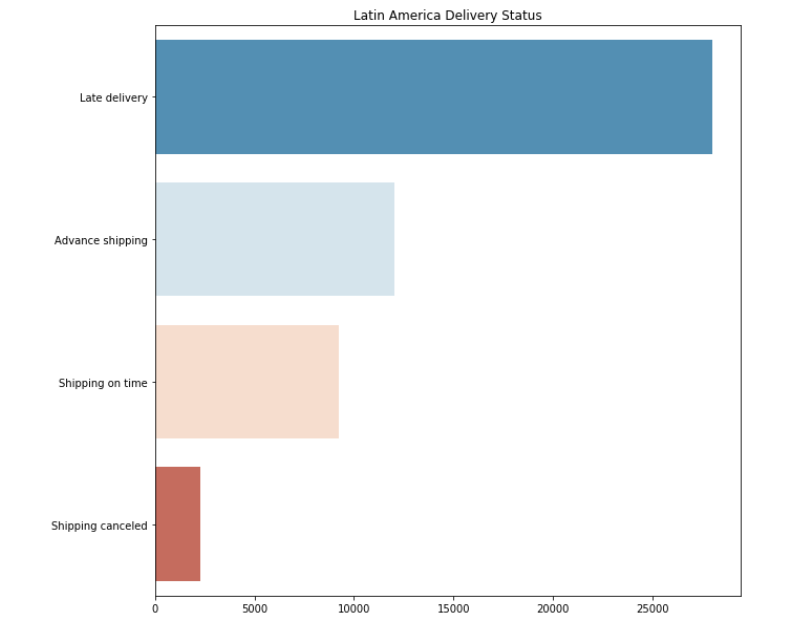
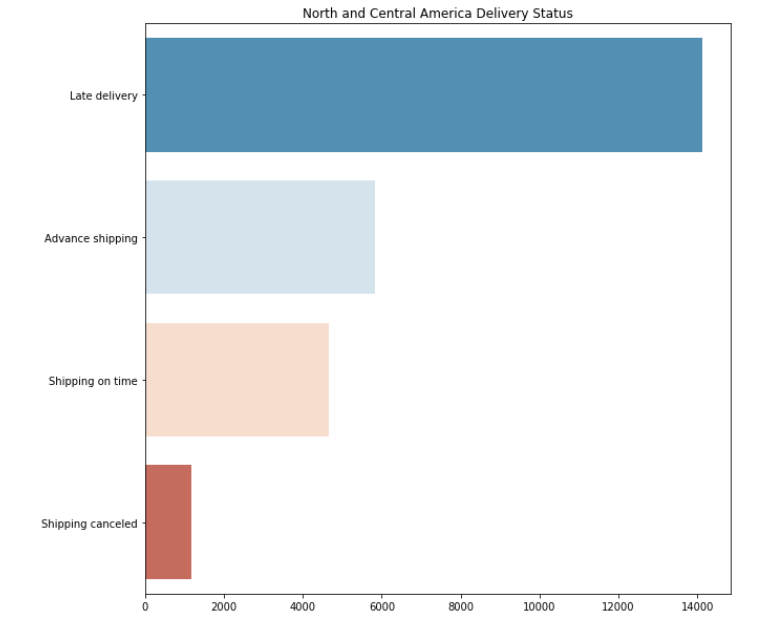
The only NaNs or Nulls were in the Product Description column. it was completely blank. Product Description column was removed. The customer zipcode and order zipcode columns were only populated for US-based orders. These columns were mostly empty.

## **3. EDA**

There was a lot of data to evaluate. Quite a bit that didn’t seem to have any relationship or impact on shipping and delivery times. There are missing connections between customer and order addresses and the actual location or warehouse items ship from would be critical in the proper evaluation of what impacts the on-time delivery of the products.

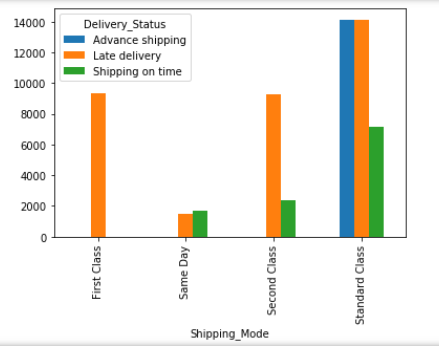
Evaluating the order status, which could be pending payment, complete, canceled, suspected fraud, payment review, on hold, closed, and processing. This provided approximately 58,000 completed orders. It makes sense to evaluate the completed orders since the other orders are not shipped but pending payments, processing, or checked for fraud.

Geographic evaluation hinged on three opportunities. The largest geographical unit was the Market consisting of 5 continent-like areas, then region which consisted of 23 regions around the globe. The next largest geographic unit was the country with 164, then city with 3597, then latitude and longitude which was large. Let’s start with the largest unit, the Market. Close to 55% or more of all deliveries in any market was equally late. In the chart below 4 of the markets are displayed and have remarkably similar shapes and late delivery rates of 54 to 55+ percent.



|  |  |
| --- | --- |
|  |  |

There were a few places where additional or more granular data could have an impact but were not present. This additional information could be the actual shipping companies involved. If DataCo is only using one vendor, I’d recommend utilizing others to find more competent or consistent delivery companies or methods.

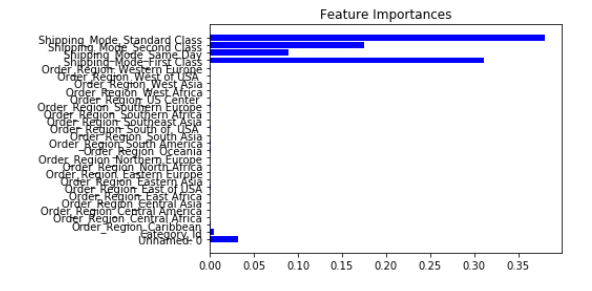
Drilling down further into the region shipping capability proved to provide similar and unremarkable results. Regional delivery stats follow in lock step with the larger Market to a tee. The Central America, Northern and Western Europe, and Oceania all have late delivery rates of 54 and 55%. 

Renamed the columns to include ‘\_’ between the words in the column headings. Removed all Pending, Processing, Pending\_payment, Closed, on\_hold, Suspected Fraud, canceled and payment\_review shipments and focused on orders that have completed the cycle. That reduced the number of orders from 180,000 to just under 60,000 records. Basically, analyzing orders that were actually delivered.

Evaluated Shipping methods with results: Essentially, First and Second Class delivery were mostly late. Half of the Same Day deliveries were late and a good percentage of the Standard Class were delivered late.

Finally, decided to run modeling on Regions, product, and methods of shipping to evaluate whether an order will be late. I’m using Regions just to verify if there is a geographical component to the shipping delivery issue. Removing any date-related data and focus on decision tree models.

Implementing a feature importance plot confirmed that there is no correlation with geographic location in determining delivery issues with this dataset. The below chart shows the four main features: Shipping\_Mode\_Standard\_Class, Shipping\_Mode\_Second\_Class, Shipping\_Mode\_First\_Class, Shipping\_Mode\_Same\_Day.



Since we are trying to predict the late deliveries before they happen the Delivery\_Status column has three components, late, on-time and advanced or early. Trimming this column to a binary result of late or happy (on-time) would have an impact on predictive powers of the algorithms.

## **4. Algorithms & Modeling**

K-Means test essentially confirmed the grouping of the dataset into 3 groups. The groups were Late, On-Time and Advanced delivery status feature. To utilize decision tree based models, encoding of the ‘delivery status’ feature was conducted to a binomial state, Late = 0 while On-Time and Advanced were converted to 1. Who cares if the customer received their package early, except for the impressed and happier customer.

Utilizing the Decision Tree model to evaluate the binary late and on-time results improves the predictive by approximately 12 to 17% from the mid 50’s to upper 60’s. Results are decent, or at least better than just guessing with the accuracy stats noted in the table below.

|  |  |  |
| --- | --- | --- |
| **Parameters (‘Gini’, Max\_Depth = 2)** | **Multi-Class Results** | **Binary Results** |
| Accuracy | 0.577 | 0.692 |
| Precision | 0.603 | 0.717 |
| Recall | 0.564 | 0.714 |

|  |  |  |
| --- | --- | --- |
| **Parameters (‘Gini’, Max\_Depth = 3)** | **Multi-Class Results** | **Binary Results** |
| Accuracy | 0.577 | 0.698 |
| Precision | 0.603 | 0.741 |
| Recall | 0.564 | 0.725 |

Increasing max\_depth to ‘None’ made insignificant changes in the predictive powers of the model. Switching to an ‘entropy’ decision tree reduced the values slightly as well.

|  |  |  |
| --- | --- | --- |
| **Parameters (‘Gini’, Max\_Depth = None)** | **Multi-Class Results** | **Binary Results** |
| Accuracy | 0.577 | 0.697 |
| Precision | 0.603 | 0.730 |
| Recall | 0.564 | 0.720 |

|  |  |  |
| --- | --- | --- |
| **Parameters (‘entropy’, Max\_Depth = 3)** | **Multi-Class Results** | **Binary Results** |
| Accuracy | 0.577 | 0.697 |
| Precision | 0.599 | 0.738 |
| Recall | 0.562 | 0.724 |

Moving to a Random Forest model provided a similar result.

|  |  |  |
| --- | --- | --- |
| **Parameters (RandomForest, Max\_Depth = 3)** | **Multi-Class Results** | **Binary Results** |
| Accuracy | 0.576 | 0.695 |
| Precision | 0.331 | 0.734 |
| Recall | 0.576 | 0.695 |

Increasing the depth of the multi-class model to ‘None’ provided an increase in precision to 0.56, while the Accuracy and Recall values remained nearly unchanged.

There appears to be no linear or logistic correlation among this data as well.

Feature importance of the various models focused on the shipping methods. Clearly, simple charts at the top of the report shows that the two shipping modes that produced the largest number of late deliveries were First\_Class and Standard\_Class. This is easily confirmed in the raw numbers in the dataset.

|  |  |
| --- | --- |
| **Shipping Method** | **Feature Importance Value** |
| First\_Class | 0.381 |
| Same\_Day | 0.034 |
| Second\_Class | 0.124 |
| Standard\_Class | 0.438 |

The delivery methods seem to reflect that if you provide the same day delivery, the product will be delivered timely. As is the case with Second\_Class methods, which the customer is clearly not invested in a timely delivery, will provide satisfactory results. However, the First\_Class and Standard\_Class methods provide less than stellar performance in delivery in a projected timely manner.

## **5. Predictions**

There are other issues involved in the delivery issues for Big Mart data. Further information may provide the missing link or links that would help build a prediction model that is a bit more helpful. There isn’t enough volume of traffic to warrant further investment into solidifying the predictors to shipment delivery and would probably be just as easily resolved by adding an extra day to the expected delivery date to the First and Standard Class delivery methods.

## **6. Future Improvements**

The Big Mart Data is quite diverse. The initial issue that jumped out of the data was the number of slow deliveries of their products, which is over 50%. Creating a process that could predict whether a shipment was going to be delivered in a normal delivery window could be helpful in adjusting customer’s expectations before they are waiting a day or two or more for their product and reduce the number of customer service calls because of slow delivery. However, there are numerous other opportunities to extract insight from this dataset. Some of those opportunities include further exploration into the following;

* Delivery Method performance by comparing and contrasting the various methods of shipping and results.
  + This would provide hard data on whether different shipping methods should be abandoned or preferred down.
    - Ship by the preferred method by Product.
    - Ship by the preferred method to specific Regions.
    - Ship by the preferred method to specific City.
  + Whether changing shipping companies would alleviate the issues.
* Evaluate Fraud
  + Flag possible orders for Fraud.
    - Save a significant time and money handling fraudulent orders.
    - Build or reach out to new customers who may have been targets of fraud.
* Evaluate returns
  + Possibly pinpoint faulty or poorly made products of low value
    - Should these products be
      * Discontinued
      * Upgraded
      * Repriced
* Predict Suspected Fraudulent orders
  + Reduce wasted time handling fraudulent orders
* Predict and evaluate Cancelled Orders
  + Reduce wasted time handling orders that
* Predict and process Pending payments
  + Identify slow-pay methods and adjust according to company policy

If there was any data associated with distribution centers for their products would be a huge boon in providing clarity whether there were internal structural process problems that could be addressed to improve the delivery process.

## **7. Credits**