**The Prediction of Late Delivery Risk Using Machine Learning Algorithms: Comparison of Performances of ML Algorithms**

**1. Research Objectives and Questions**

This project is to determine the factors related to late delivery risk by using Machine learning algorithms and discuss the performance of each ML algorithm. Specific research questions included as follows:

1. Is there any seasonality in late delivery?
2. Is there any regional disparity in late delivery?
3. What are the factors associated with late delivery risk?
4. Is there any course of action to increase late delivery risk?
5. Is there any difference in performance among machine learning algorithms in predicting late delivery risk?

**2. Methods**

**2.1 Data for the study**

Of a total of 180,519 cases in the dataset named “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS” from Kaggle site, this study utilized 91,197 after including order status of ‘Complete’, ‘On-hold’, and ‘Processing’ (excluding ‘suspected fraud’, ‘cancelled’, ‘pending’, etc.).

**2.2 Measures**

**2.2.1 Outcome variable:** Late delivery risk is dependent variable which was derived from delivery status (coded as 1 for late delivery; and 0 for advanced shipping, shipping cancelled and shipping on time).The final analytical dataset included *57.2% of late delivery risk and 42.8% of non-risk cases.*

**2.2.2 Predictor variables**

*Days till shipping from order* were derived from the functions of (shipping datetime – order datetime)/ 86400) and floor() to get only the largest integer.

*Order month/day/year* were derived from order datetime variable using functions of datepart(), month(), day(), and year().

*Region variable* was recoded from order state variable according to 10 Standard Federal Region which were established by OMB ([Office of Management and Budget](https://en.wikipedia.org/wiki/Office_of_Management_and_Budget)) Circular A-105, *"Standard Federal Regions",* in April 1974A screenshot of a video game

Description automatically generated with medium confidence as shown the map beside. We coded 11 the region for Samoa, Guam, etc. which were not included in the standard federal region.

Additionally, we included the following variables as independent variable in the analyses: Benefit\_per\_order, Department\_Name, Order\_Item\_Discount, Order\_Item\_Discount\_Rate, Order\_Item\_Product\_Price, Order\_Item\_Profit\_Ratio, Order\_Item\_Quantity, Order\_Item\_Total, and Shipping\_Mode

Table below shows the summary for interval independent variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | N | Mean | Std Dev | Minimum | Maximum |
| Benefit\_per\_order | 91197 | 22.20988 | 104.2115 | -4274.98 | 911.8 |
| Order\_Item\_Discount | 91197 | 20.72496 | 22.04732 | 0 | 500 |
| Order\_Item\_Discount\_Rate | 91197 | 0.101792 | 0.070435 | 0 | 0.25 |
| Order\_Item\_Product\_Price | 91197 | 141.2146 | 141.5466 | 9.99 | 1999.99 |
| Order\_Item\_Profit\_Ratio | 91197 | 0.120893 | 0.466061 | -2.75 | 0.5 |
| Order\_Item\_Quantity | 91197 | 2.128272 | 1.452689 | 1 | 5 |
| Order\_Item\_Total | 91197 | 183.0202 | 121.5661 | 7.49 | 1939.99 |
| days\_til\_ship | 91197 | 3.465344 | 1.663675 | 0 | 6 |

**2.3 Analysis Tools and Methods**

* SAS v9 was used to clean the data (i.e., sub-setting, derived variables, etc.).
* Tableau was used to graphically examine the seasonality and regional disparity in late delivery risk. Logistic regression was performed for statistical tests.
* Diagram

  Description automatically generatedEnterprise Miner was used to build and validate the predictive models of logistic regression, stepwise logistic, decision tree and neural network (iteration =100 and hidden layer=6). The figure below presents Enterprise Miner Diagram to build and validate the models.

**3 Results**

Chart, bar chart

Description automatically generated **3.1 Seasonality in Late Delivery**

For the most part, there does not seem to be a huge difference in late delivery risks by month. The only month that could stand out is a lower risk in June, followed by an increase in July. Risk was higher during the second half of the year (from July to December) than the first half of the year.

**3.2 Regional disparity in Late Delivery**

Chart, bar chart

Description automatically generatedFrom the beside chart, we can see that those regions 7 and 8 were significantly lower in late deliveries risk than the other regions. When we look at what states are in those regions, we can see why it makes more sense that there would be fewer late delivery risks in these areas. Region 7 consists of Kansas, Iowa, Nebraska, and Missouri. Region 8 consists of Colorado, Montana, North and South Dakota, Wyoming, and Utah. If we look at the other chart, it shows that in those areas, there are fewer sales, and some of the states in both Regions 7 and 8 fail to have any sales. This would decrease the opportunity for late deliveries

* 1. **Factors Associated with Late Delivery Risk**

As seen in the table below, the multiple logistic regression model without any selection criteria presented that days between order and shipping, region, and order day were significantly associated with late delivery risk (p-value <0.05). Order month was marginally associated with it (0.05 < p-value<0.1). However, days between order and shipping and shipping modes were significant in the stepwise logistic regression model.

|  |  |  |  |
| --- | --- | --- | --- |
| Effect | DF | Chi-Square | Pr > ChiSq |
| Benefit\_per\_order | 1 | 0.5766 | 0.4477 |
| Department\_Name | 10 | 13.5872 | 0.1927 |
| Order\_Item\_Discount | 1 | 1.423 | 0.2329 |
| Order\_Item\_Discount\_Rate | 1 | 0.4702 | 0.4929 |
| Order\_Item\_Product\_Price | 1 | 3.5776 | 0.0586 |
| Order\_Item\_Profit\_Ratio | 1 | 1.2031 | 0.2727 |
| Order\_Item\_Quantity | 1 | 2.4267 | 0.1193 |
| Order\_Item\_Total | 1 | 2.1089 | 0.1464 |
| Shipping\_Mode | 2 | 1.5965 | 0.4501 |
| days\_til\_ship | 1 | 15.368 | <.0001 |
| order\_d | 30 | 96.9209 | <.0001 |
| order\_m | 11 | 17.5356 | 0.093 |
| region | 10 | 19.8371 | 0.0308 |

* 1. A picture containing timeline

     Description automatically generated**Course of Action to Increase Late Delivery Risk**

As seen in the decision tree, the risk pf late delivery was heightened in the case that 1) days between order and ship was greater or equal to 4.5 days, or 2) shipping mode was standard classes or missing even though days between order and shipping less than 4.5 days.

**3.5 Model Comparison**

Model comparison in which selection criteria was set at average square error revealed that the average squared error and misclassification rate for validation data were the lowest in non-stepwise logistic regression, implying that select logistic regression model performed the best in predicting risk of late delivery. However, all four models did not notably differ each other in model performance.

**Calendar

Description automatically generated with low confidence**

**4. Discussion of Findings**

The study intended to determine factors impacting risk of late delivery using publicly available data on supply chain. Our analyses suggested some notable findings as follows:

1. We found some seasonality in risk of late delivery. We assumed a higher risk would be during the holiday season (Independent days, Thanksgiving and Christmas days). Our additional analysis showed that odds of being late delivery risk was 1.05 times higher during holiday season than non-holiday season (p=0.0021). Risk was higher during the second half of the year (from July to December) than the first half of the year (p=0.0003).
2. There was a regional disparity in risk of late delivery. Risk of late delivery was higher in the regions located in US coastal area than inland regions. US coastal states usually have a high population density and metropolitan cities, implying the heavy traffic which is more prone to populated cities may impact the risk of late delivery.
3. Days from order to shipping was the most important factor to predict the risk of delayed delivery. Our additional analysis showed that odds of being late delivery risk was 4.1 times higher when 4 or more days were needed for shipping after receiving order than less than 4 days. This may imply that enough stock in sellers is the crucial point in on-time delivery in that they may need additional time for stock-up when they have lack in inventory. Further study is needed to determine the association of risk of late delivery and inventory.
4. Shipping mode was also important factor in predicting the risk of late delivery. Decision tree showed that the risk became higher when shipping mode was other than standard classes or missing even though days between order and shipping were less than 4.5 days.
5. The performance of four machine learning models in predicting the risk of late delivery was excellent, in that misclassification rate was <0.025, saying that less than 3 % of validation data were misclassified by the trained models.

Despite these notable findings, this study is subject to ***few caveats***. Since this study used the secondary data, it was impossible to determine the causes of late delivery. The findings from this study were just associational, not cause-effect. Another limitation is that the study data included a limited information, especially on potential factors associated with the risk. Actually, current inventory, relationship with partner, road condition, weather, traffic, so on may be more important in predicting the risk. However, that information was not provided by the dataset.

**6. Conclusion**

Despite caveats listed above, the study revealed that there was seasonality and regional disparity in the risk of late delivery, although causes were not known from the analyses. Days between order and shipping, and shipping modes were the most important factors in predicting the risk. Predictive machine learning models were excellent in predicting the risk of the late delivery. However, further study is needed after including more qualitative factors (relationship with suppling companies, road conditions, etc.). After conducting this research, we have found answers to our various research questions and can apply practical implications. Our research is relevant to help companies see where the issues lie in their shipping processes. It can help determine how to better fix that process. To summarize our research finding, SAS was able to help us see that days until shipped is an important factor in if the product will reach its destination in the desired time. Moving forward, we highlight this information as a key issue to focus on. We have also discovered that there was generally no regional disparity in late deliveries. Finally, we found that there does not seem to be seasonal disparity in late deliveries either.