GT Buzz Sporting Goods

Summer 2022

MGT6203 – Team 70 Final Report

Background

The supply chain industry is the engine that runs every economy in the world

Supply chain disruptions in major industries have caused shortages that have had downstream effects

These disruptions have impacted all areas of industry and lifestyles specially during the Covid –19 pandemic

Problem Statement

Businesses require an advanced approach for analyses and prediction of the future state of any supply chain operation

The modern supply chain must evolve to meet increasing demands and supply chain managers need to plan to keep everything flowing smoothly

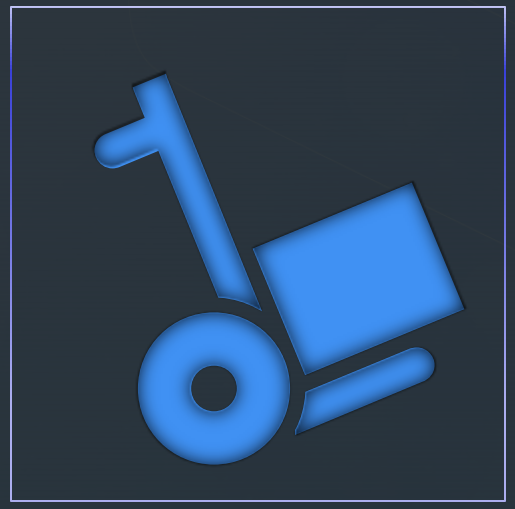
By understanding, predicting, and optimizing supply chain operations, businesses can reduce costs, improve their growth potential, and ultimately maintain happy customers

We will look at the various factors that can be used to explain whether an order for goods and services is delivered on time

Data Preparation and Feature Engineering

- Dataset: “DataCo Smart Supply Chain for Big Data Analysis” (2019)
- Translated order and customer addresses from Spanish to English
- Roughly 180K rows in dataset
- Creation of Shipping Distance and Order Delay features

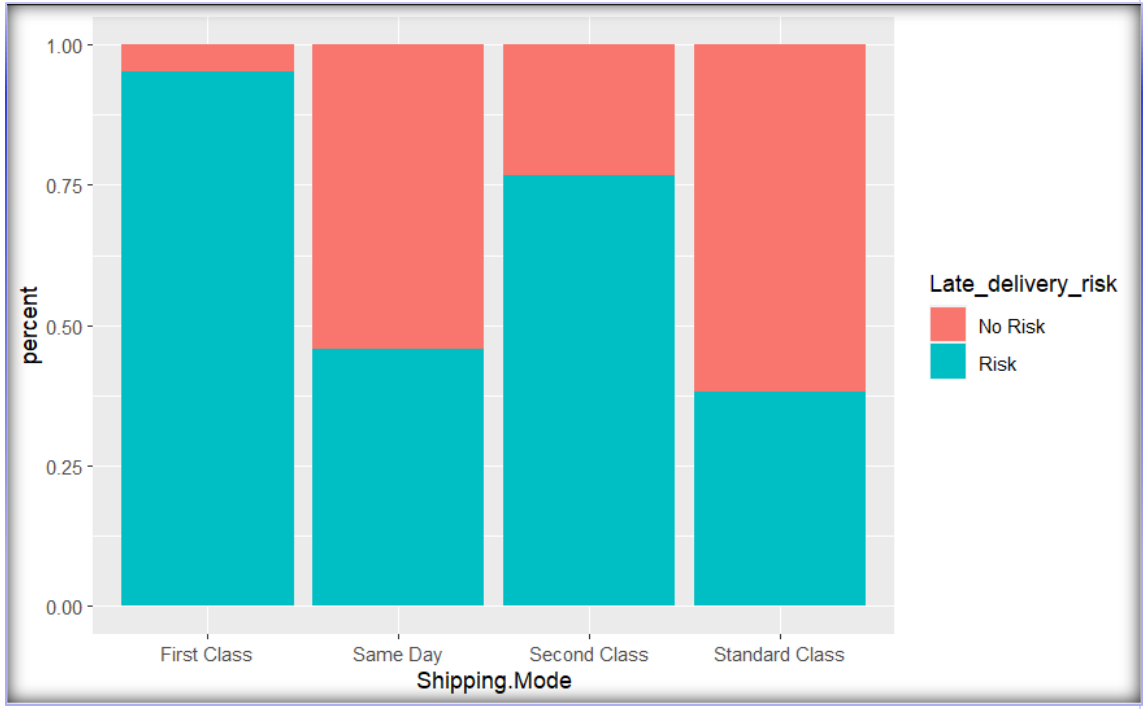
Hypothesis



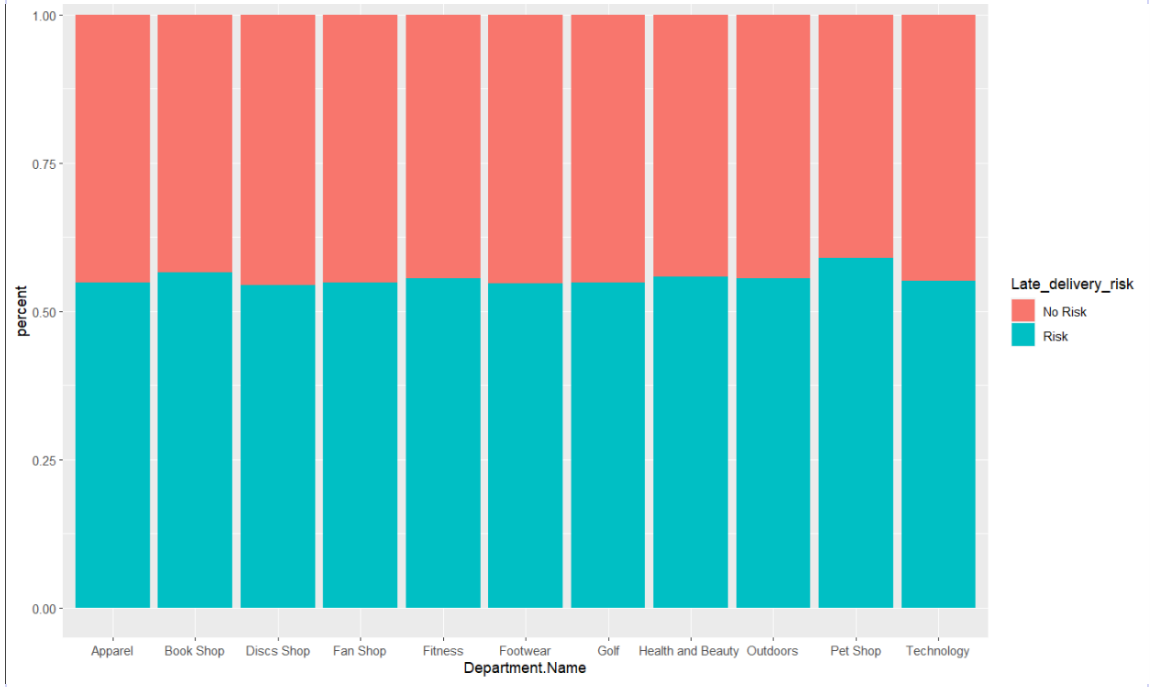
The delivery location, distance, time of year/season, and shipping method are the best predictors of whether an order will be delivered late or on time.

Exploratory Data Analysis

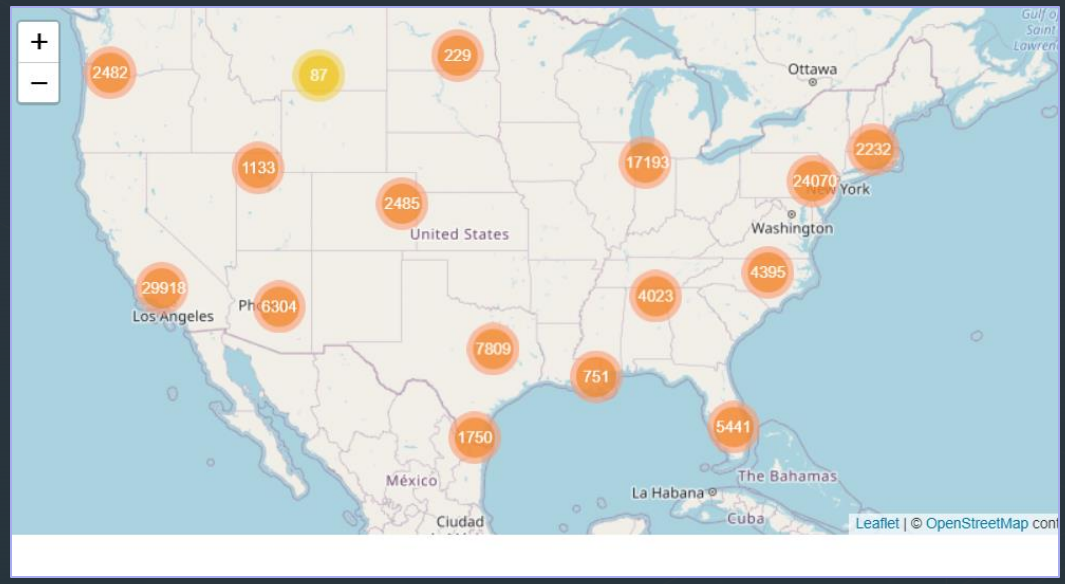
Late Delivery Risk and Shipping Mode



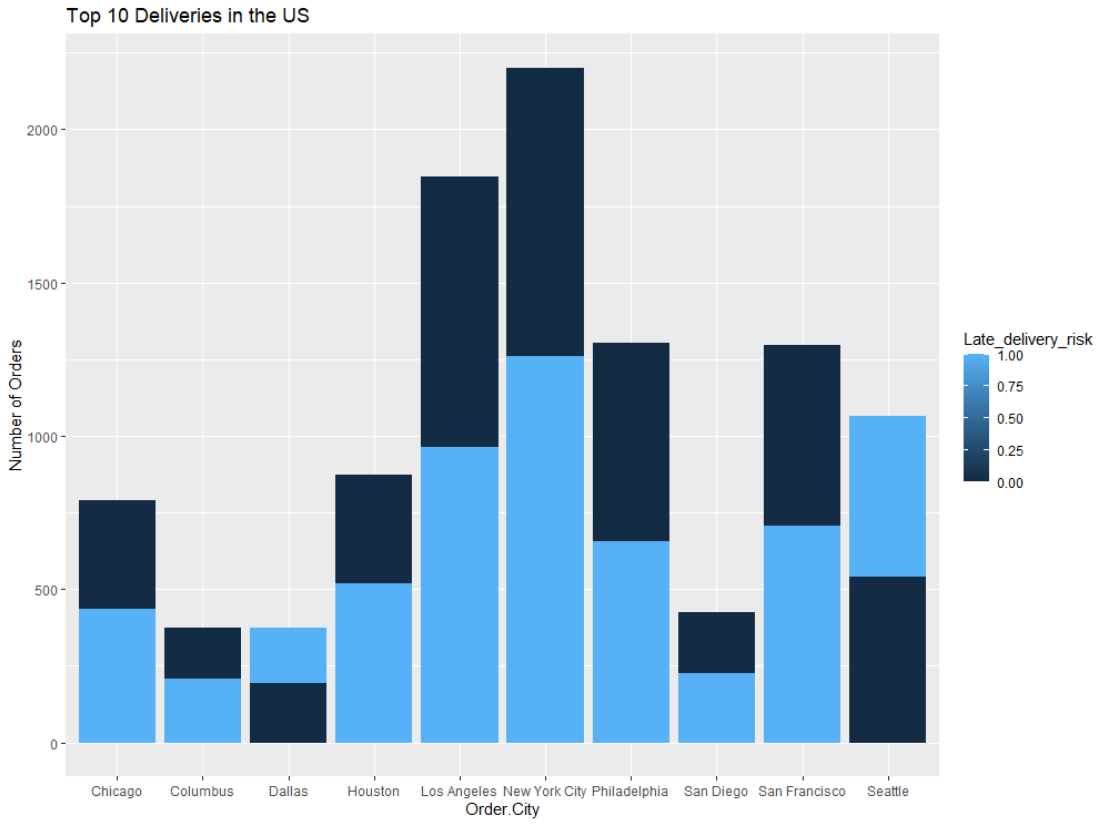
Late Delivery Risk and Department



Order Locations

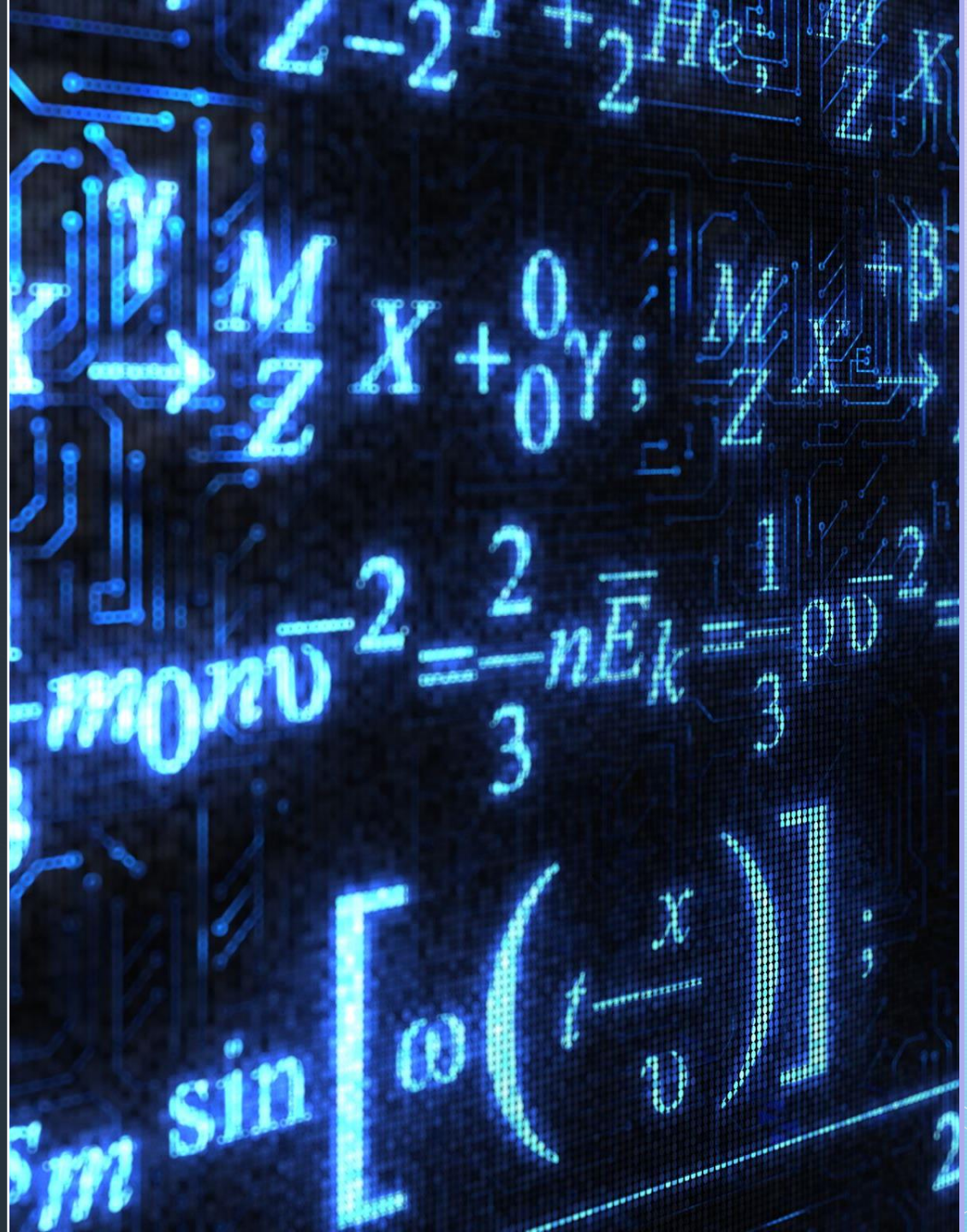


Late Delivery Risk by US City

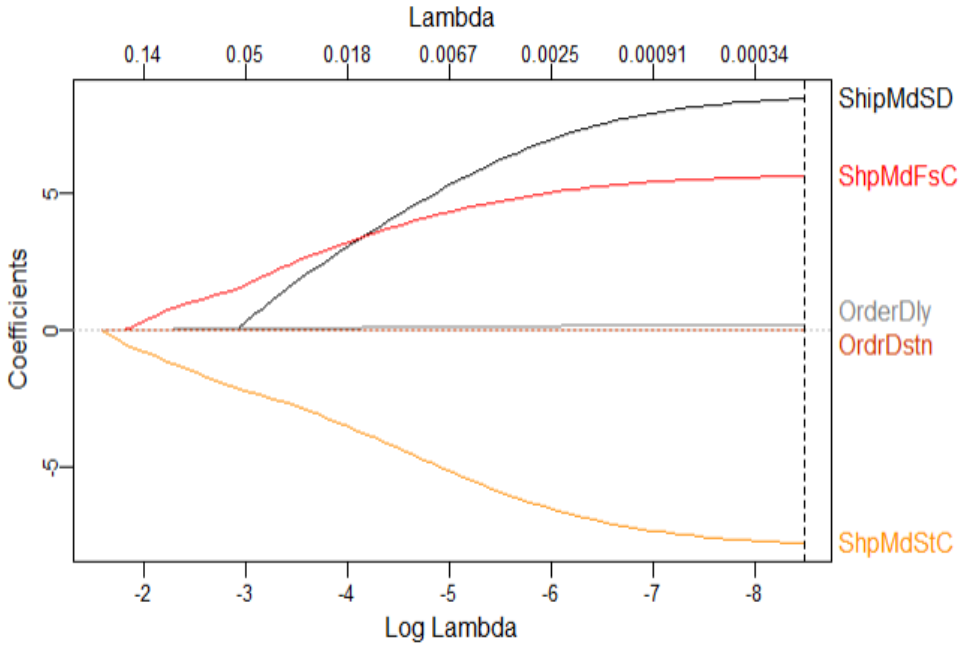


Modeling

1. Before building and testing our models, we performed an 80/20 train/test split
2. During the model validation phase, we utilized the caret package to help us perform a 10-fold cross validation for each model.
3. To do **variable selection** we used Lasso and **Elastic Net**, a combination of L1 and L2 regularization techniques, to choose the best variables and account for overfitting.
    1. We chose the best model based on the best cross validated ROC value.
4. Finally, we used the test data on the best model to test its performance.



Elastic Net Regression

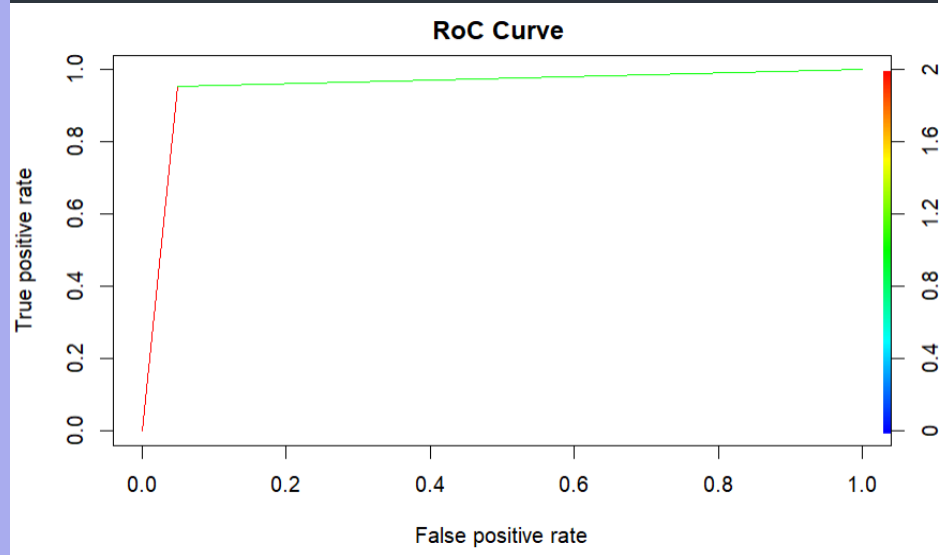



```
6 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)   -1.052930e+01
ShipModeSD     8.436925e+00
ShipModeFstC   5.603538e+00
ShipModeStdC  -7.783560e+00
OrderDelay     1.666001e-01
OrderDistance -3.167493e-08
```

Final Model Training

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0826  -0.0639   0.0155   0.3151   1.2487

Coefficients:
                      Estimate       Std. Error  z value            Pr(>|z|)
(Intercept)    -10.359836065909   0.089511757169 -115.737 <0.0000000000000002 ***
OrderDelay       0.161421470580   0.001237246504  130.468 <0.0000000000000002 ***
OrderDistance    0.000000001787   0.000000002625    0.681                0.496
ShipModeSD       8.256682078899   0.077866582777  106.036 <0.0000000000000002 ***
ShipModeFstC     5.576684135358   0.051813922539  107.629 <0.0000000000000002 ***
ShipModeStdC    -7.473552091167   0.066320670734 -112.688 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 155078  on 112663  degrees of freedom
Residual deviance:  44053  on 112658  degrees of freedom
AIC: 44065

Number of Fisher Scoring iterations: 7
```

```
##   parameter       ROC      Sens      Spec      ROCSD      SensSD      SpecSD
## 1      none 0.9711444 0.9474701 0.9536277 0.00173703 0.003023774 0.002261609
```



Final Model Testing

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    X0    X1
##         X0 12058   631
##         X1   684 14794
##
##                Accuracy : 0.9533
##                  95% CI : (0.9508, 0.9557)
##     No Information Rate : 0.5476
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9057
##
##  Mcnemar's Test P-Value : 0.1516
##
##             Sensitivity : 0.9591
##             Specificity : 0.9463
##          Pos Pred Value : 0.9558
##          Neg Pred Value : 0.9503
##              Prevalence : 0.5476
##          Detection Rate : 0.5252
##    Detection Prevalence : 0.5495
##       Balanced Accuracy : 0.9527
##
##        'Positive' Class : X1
##
##          Model       Auc
## 1 final.Model 0.9733868
```

Prediction Sample Data

	OrderDelay <dbl>	OrderDistance <dbl>	ShipModeSD <dbl>	ShipModeFstC <dbl>	ShipModeStdC <dbl>
1	5	100	1	0	0
2	10	200	0	1	0
3	100	300	0	0	0
4	20	400	0	0	1

	X0 <dbl>	X1 <dbl>
1	0.785168856	0.2148311439320
2	0.959648437	0.0403515633301
3	0.003072117	0.9969278829857
4	0.999999546	0.0000004540954

X0 = Probability of order being on time
X1 = Probability of order being late

Conclusion

Model Overview

The final model includes features Order Delay, Order Distance, and Shipping Mode.

The Order Distance is kept in the final model because it was considered useful for explaining the model.

Reducing the model via variable selection is particularly important for reducing the impact of overfitting in predictions and of course for the simplicity of the model.

Model Features and Interpretation

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-10.359836065909	0.089511757169	-115.737	<0.0000000000000002	***
OrderDelay	0.161421470580	0.001237246504	130.468	<0.0000000000000002	***
OrderDistance	0.000000001787	0.000000002625	0.681	0.496	
ShipModeSD	8.256682078899	0.077866582777	106.036	<0.0000000000000002	***
ShipModeFstC	5.576684135358	0.051813922539	107.629	<0.0000000000000002	***
ShipModeStdC	-7.473552091167	0.066320670734	-112.688	<0.0000000000000002	***

Default (Intercept) = Second Class
OrderDelay = Days between order date and shipping date
OrderDistance = Distance between order store and customer address
ShipModeSD = Same day
ShipModeFstc = First class
ShipModeStdc = Standard class

Recommendations

Minimize time between Order Placement and Order Shipment

- Investigate order pick rate
- Investigate picking, packaging, shipping operations
- Further simulation studies maybe required

Subscription model or loyalty program for expedited shipping

- Investigate subscription model vs loyalty program
- Conduct cost benefit analysis

Deploy production models to catch potential delays

- Update model periodically with most recent data
- Implement order delay control charts

Works Cited

Constante, Fabian; Silva, Fernando; Pereira, António (2019), “DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS”, Mendeley Data, V5, doi: 10.17632/8gx2fvg2k6.5 https://data.mendeley.com/datasets/8gx2fvg2k6/5

Schork, J. (2022, March 21). *R geospatial distance between 2 points: Geographical latitude/longitude*. Statistics Globe. Retrieved June 29, 2022, from https://statisticsglobe.com/geospatial-distance-between-two-points-in-r

THANK YOU!