Data Exploration and Cleaning Project Report

Marco Teniente and Rani Misra

University of Texas at San Antonio

STA 4233-001

Professor Yao

4/23/2023

Table of Contents

Project Overview

Description

The COVID-19 pandemic has caused unprecedented disturbances in every industry in some capacity, most critically, by impacting their ability to manage supply and demand. ABC, a technological company in the business of desktops and laptops, operates in North and Latin America. Their perfected low-cost, low-risk model is threatened by a hybrid supply chain and the reliance on Asian and European manufacturing. Factories and manufacturers struggle to remain open during this time, and third-party logistics carriers are unable to provide reliable service regarding logistics or transportation on lead times. As demand continues to soar, the need to decrease lead times to gain the upper hand in the market is much more necessary.

Objectives

By using data preparation and exploration, we will analyze the influence of various explanatory variables in an effort to decrease in-transit lead times for third-party logistics carriers of ABC's consumer products and software.

Data Overview

Preparation

After reading the data and adding any required columns, all dates were converted to POSIXct format so that they could be used for future calculations. In order to assign Quarter and Year values to every observation, a for loop was created that read the Receipt Date for every observation. A match variable was then created to find the matching date ranges in the Calendar sheet. An if statement was then used to assign the Quarter and Year values for every match found in the observations. If a match was not found, an NA was recorded instead. The In-transit lead time was then calculated by subtracting the Ship Date from the Receipt Date. Similarly, the Manufacturing Lead Time was calculated by subtracting the PO Download Date from the Ship Date. These calculations yielded lead times in the unit of seconds. In order to convert these seconds to days, a function (named seconds_to_D) was created which divided the seconds by the numbers of seconds in a day. Using pipe operators, both In-transit Lead Time and Manufacturing Lead time were converted to numeric variables and the seconds_to_D function was applied.

Cleaning

In order to clean the data, all missing values were first identified. The missing lead times were first tackled by replacing the values with rounded mean lead times. As time cannot be negative, any lead times with negative values were removed. Outliers for both lead times were then found using boxplot statistics. The rounded mean for each of these outliers were then used as thresholds which determined which lead times were to be replaced. Any lead times that went over the threshold were replaced by the rounded mean lead times of the outliers.

Although missing lead times were imputed with new values, missing Start Dates and Receipt Dates still remained. In order to impute these values, Manufacturing Lead Time was

added to PO Download Date to find the Start Date. Similarly, In-transit Lead Time was added to Ship Date to find Receipt Date for the missing values. Another for loop and if statement were used to match Quarter and Year values from the Calendar sheet for all missing values. After the cleaning process, 8793 observations remained from the original 9124 observations to be analyzed. Thus, a total of 331 observations were removed due to the cleaning process.

Exploration

We first explored potential categorical/nominal variables and potential confounders via boxplot formatting, which may hold influence for In-transit Lead Time. A dummy-dataset was constructed for all categorical variables and dates, enabling for a numerical interpretation for correlation/regression analytical purposes. From the correlation plot *(Figure 1)* , the strongest associations for In-transit Lead Time were found in Ocean Ship Mode with a correlation coefficient (also known as R value) of 0.78657970, and in Ground Ship Mode with an R of -0.50332802.

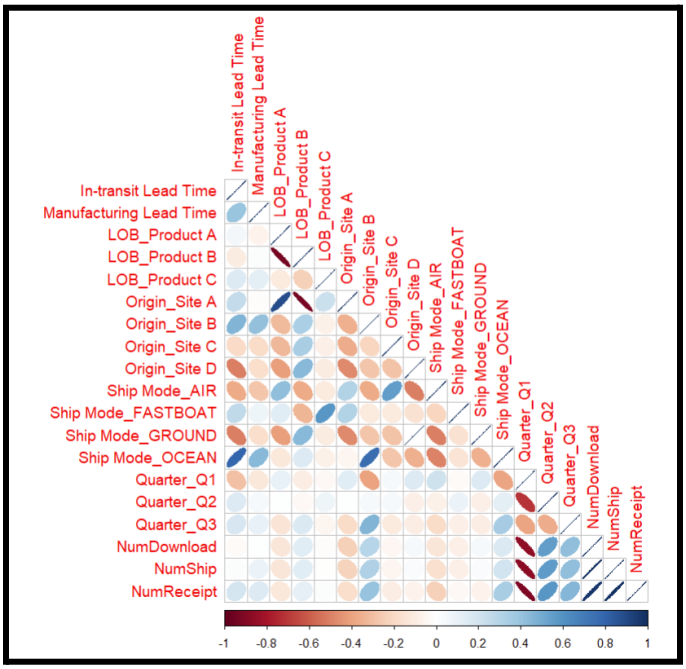

Figure 1: Correlation Plot

Finally, linear models were utilized with the aforementioned dummy-data to assess potential predictors (excluding Receipt Date and Ship Date) for In-transit Lead Time. For the tested regressions, 3 specific lines were constructed, per each categorical dummy variable which held high influence in correlation (Ship Mode, Origin Site, LOB). Given the R^2 represents the proportion of variability from the dependent variable explained by the independent variable, we utilized this as an applicable effect size measure. From the R^2 (both adjusted and multiple), Ship Mode held the most influence in model predictability as the Adjusted R^2 was 0.7919 *(Figure 2)*. Origin Site followed with an Adjusted R^2 of 0.4358 *(Figure 3)*. Lastly, LOB held the least amount of influence with an Adjusted R^2 of 0.02702 *(Figure 4)*. Origin_Site A and LOB_Product A were not included in their respective linear models as a perfect linear relationship existed between 2 or more variables (singularities).

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.27672    0.08857   93.45   <2e-16 ***
`Ship Mode_FASTBOAT` 17.64382    0.25230   69.93   <2e-16 ***
`Ship Mode_GROUND`   -4.65407    0.14195  -32.79   <2e-16 ***
`Ship Mode_OCEAN`    20.88175    0.14336  145.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.366 on 8789 degrees of freedom
Multiple R-squared:  0.792,     Adjusted R-squared:  0.7919
F-statistic: 1.115e+04 on 3 and 8789 DF,  p-value: < 2.2e-16
```

Figure 2: Linear Model for Ship Mode

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      17.2551     0.1528  112.90   <2e-16 ***
`Origin_Site B`   8.1619     0.2735   29.84   <2e-16 ***
`Origin_Site C`  -8.7342     0.2691  -32.46   <2e-16 ***
`Origin_Site D` -13.6324     0.2382  -57.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.836 on 8789 degrees of freedom
Multiple R-squared:  0.436,     Adjusted R-squared:  0.4358
F-statistic:  2265 on 3 and 8789 DF,  p-value: < 2.2e-16
```

Figure 3: Linear Model for Origin

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      14.3026     0.2180  65.616  < 2e-16 ***
`LOB_Product B`  -1.7293     0.2667  -6.484 9.41e-11 ***
`LOB_Product C`   9.3214     0.7546  12.353  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.6 on 8790 degrees of freedom
Multiple R-squared:  0.02724,   Adjusted R-squared:  0.02702
F-statistic: 123.1 on 2 and 8790 DF,  p-value: < 2.2e-16
```

Figure 4: Linear Model for LOB

All numeric variables were associated with the calculation of In-transit lead time. Thus, they were not explored statistically.

Conclusion

Based on our quantitative findings, while most variables tested as significant when run through a regression analysis, Ship Mode and Origin held the most influence in terms of predictive power for In-transit Lead Time. More specifically, the Ocean Ship Mode and Origin Site B demonstrate a strong positive correlation with In-transit Lead Time, while the Ground Ship Mode and Origin Site D are negatively correlated with In-transit Lead Time.

In other words, to minimize lead times and maximize efficiency of third-party logistics services for Asian and European manufacturing branches, ABC must rely on the Ground Ship Mode method of logistics transportation from Origin Site D. Avoiding the Ocean Ship Mode and Origin Site B is also crucial, given their association with high logistics transit times.