**Business Understanding**

As the big impact of COVID-19 pandemic, 75 percent of publicly listed apparel and fashion companies in North America could find themselves with negative EBITDA or untenable net debt-to-EBITDA ratios after three-month store closures in 2020. While consumer engagement with apparel may be up at this time—as more consumers find themselves at home, idly scrolling through social media—that traffic is not translating to conversion. Finding a way to encounter with the pandemic and make the transition from offline to online shop has become the main development direction of the apparel industry. Thus, launch of a new personalized online shopping service is essential.

“Just-in-time” service is a service that customers purchase at same time as they purchase the goods. By delivering this service, the company obtains the obligation to ensure that the goods are delivered within the specified time, otherwise, company needs to compensate the customers. Based on surveys of the similar services in other industry, 82% of customers claim that “Just-in-time” service is the reason influencing their orders. By purchasing these services, customers' anxiety about the length of time for shipment has been eased as their confidence in the "just-in-time" delivery effectiveness increased.

In order to estimate the revenue brought by the Service, we need to predict the proportion of late of delivery. The machine learning classifiers used in this project are Logistic Regression, Logistic Regression with interactions, and Classification tree to predict whether the delivery will be late which are compared with ROC and R-square.

**Data Understanding**

Market segmentation is the process of dividing consumers into different categories based on distinguishing characteristics. In our project, customer segment could be divided into customer and business, which are B2C and B2B. B2B covers Corporate and Home Office. Since different segmentations require different terms, we analyze these two segmentations separately (Figure 1). In figure 2, we also found that most regions mainly have B2B services. Also, in figure 2, the scope of business concentrated in United States. We could verify this in the figure 3.

Chart, treemap chart

Description automatically generatedMap

Description automatically generated

Figure 2

Figure 1

Based on the trends of figure 3, we could conclude Central America, Western Europe, and South America cover the most operation region for the company. We will only concentrate these 3 regions and use others to identify other regions. Figure 4 show the percentage of different markets for the company.

Chart, bar chart

Description automatically generatedChart, pie chart

Description automatically generated

Figure 4

Figure 3

Figure 5 show the different categories clothing. For different kinds of categories, the shipping time will change tremendously. We will focus on the dark red cycle, Cleats, Men’s Footwear, Women’s Apparel, and Indoor/Outdoor Games to analysis the possibility of late delivery.

Chart, bubble chart, treemap chart

Description automatically generated

Figure 5

After analyzing the individual variables, we decided to depict some visualizations to determine potential relationships between our remaining variables. Figures 6, 7, 8, and 9 depict the relationships between some of our variables.

Chart

Description automatically generated**Treemap chart

Description automatically generated with medium confidence**

Figure 7

Figure 6

**A picture containing diagram

Description automatically generatedCalendar

Description automatically generated with medium confidence**

Figure 9

Figure 8

**Data Preparation**

Step 1: Remove variables that do not provide values for prediction

Step 2: Check missing values and found no missing values in any variables

Step 3: Remove the duplicate records from the dataset

Step 4: Reduce variable categories

Step 5: Customer segmentation: B2C & B2B

Step 6: Construct dummy variables for non-numeric variables

**Modeling**

Model framework

Data: Order Characteristics (X), Number of late delivery (Q1), Number of not late delivery (Q0), Price of “Just-in-time” Service (P), Penalty of late delivery (PN)

Late\_delivery\_risk = the order is likely to be late

Cost-benefit Matrix:

|  |  |
| --- | --- |
|  | Customers buy “Just-in-time” Service |
| Not Late | Q0\* P |
| Late | Q1\* P- Q1\* PN |

Mathematical Model:

E[Gain | X,P, PN] = P(Not Late | X,P)Vnot late(X) + P(Late | X,P, PN) Vlate(X)

Decomposition:

* Assumptions

○ If the probability of late delivery > 0.5, we predict the order as late delivery

○ We ignore cost of establishing “Just-in-time” Service

○ Vnot late(X) = P, Vlate(X) = P - PN

* Final Equations

When all assumptions are taken:

* + E[Gain | X,P, PN] = Q0\* P + Q1\* P- Q1\* PN

Core Tasks & Data Mining Methods

* Using cleaned data which is available to predict numbers of Late Delivery & Not Late Delivery

**Evaluation**

To determine the best predictive model to use for our final model, we ran a k-fold cross validation with 5 iterations for each method. For B2B model, we determine that Logistic Interaction method returned the highest Out of Sample R-square value, which we used as our basis for evaluating the performance of each model. Figure 10 shows the comparison of R-square values for each of the 4 regression models, with the Logistic Interaction model having the highest R-square value of 0.05324561. For B2C model, we determine that Classification tree method returned the highest Out of Sample R-square value, which we used as our basis for evaluating the performance of each model. Figure 11 shows the comparison of R-square values for each of the 4 regression models, with the Classification tree model having the highest R-square value of 0.031586720.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 11—B2C R2

Figure 10—B2B R2

**Deployment**

Based on the eight prediction models above, For B2B model, Logistic Interaction model returned the highest Out of Sample R-square of 0.05. For B2C model, Classification tree method returned the highest Out of Sample R-square of 0.03. The lack of related variables may lead to low accuracy. With more variables with higher correlation with late delivery risk, for example, weather condition, we can achieve higher accuracy and fairness. However, when we improve our predicting model in the future, we should pay extra attention to not take genders or races into consideration. This prevents gender inequality and racial discrimination.

Also, with further analytics, companies can generate proper price for “Just-in-time” Service and for penalty, to achieve maximum profit.

Reference

<https://www.mckinsey.com/industries/retail/our-insights/perspectives-for-north-americas-fashion-industry-in-a-time-of-crisis>

<https://www.sohu.com/a/117860920_310877>