Team 12 Report

Non-Technical Executive Summary

The questions our team trying to answer are, 1) what is BigSupply Co.'s overall shipping effectiveness, and how does it change based on other features? 2) What is a better algorithm to predict the possibility of late delivery?

To answer the first question, we did some exploratory data analysis. As can be seen in figure 1, about 100000 orders have been marked late delivery out of a total of 180519 orders, while the number of orders shipped earlier or on time is around 70000, taking up another 40% of all the orders. The others are those shipments being canceled.

Figure 1: Delivery Status for All Orders

It is interesting to find that shipping cancellation only occurred when the customers use transfer as the payment method from figure 2.

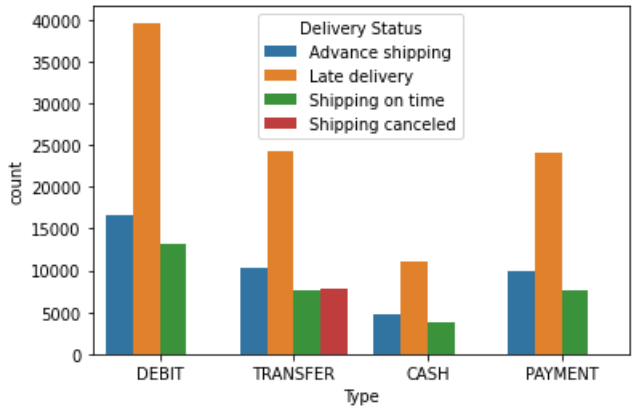

Figure 2: Delivery Status vs. Type

LATAM area has the largest amount of orders as well as its late shipment number, followed by Europe, which has a very similar share and pattern as the previous one. The African market has the smallest number of orders.

Figure 3: Delivery Status vs. Market

'Fan shop' has the largest amount of orders as well as its late shipment number, followed by 'Apparel' and 'Golf'. 'Health and Beauty' has the smallest share.

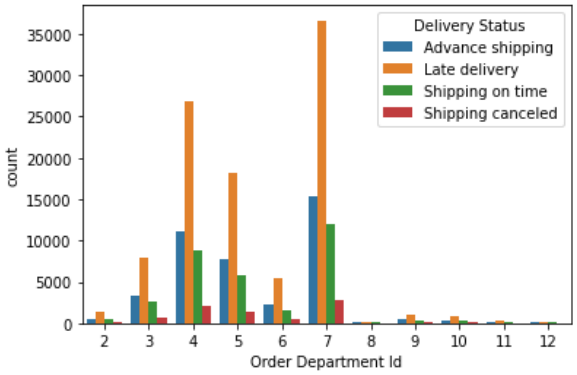

Figure 4: Delivery Status vs. Order Department

As can be seen in Figure 5, it is obvious that late delivery risk is another form to represent delivery status, where that late delivery is marked as 1 and others as 0. Therefore, either late delivery risk or delivery status needs to be dropped in a later investigation.

Figure 5: Delivery Status vs. Late Delivery Risk

We also discussed the impact of time on delivery status. However, there seems to be no clear pattern discovered except that a huge decline in shipping amounts is detected for

the last quarter of 2018, which may due to the lack of orders collection.

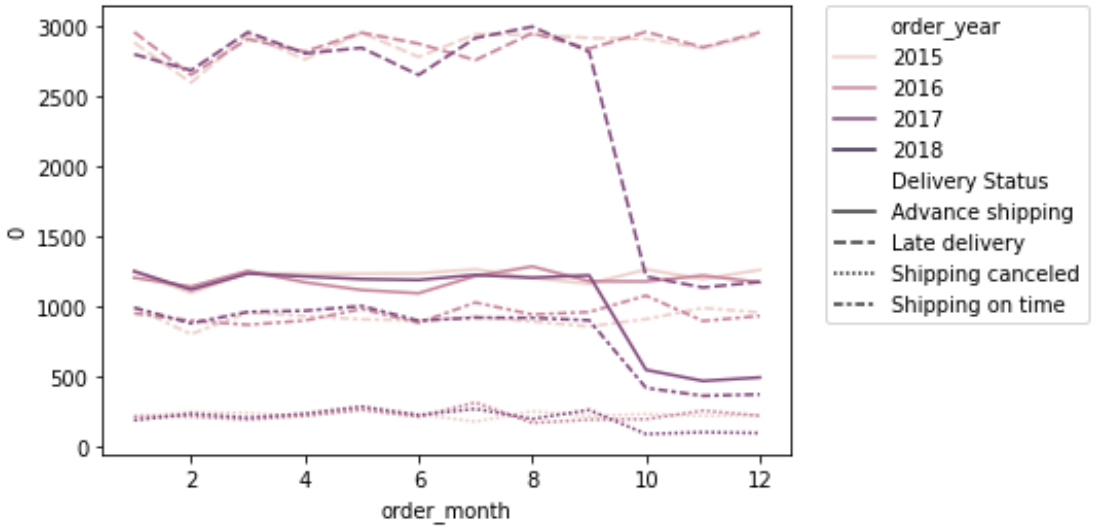

Figure 6: The Impact of Time on Delivery Status

Technical Exposition

To reply to the second question, we tried 4 models to predict delivery status if a new order is added. The models are K-Nearest Neighbors, Logistic Regression, Naïve Bayes, and Decision Tree. Meanwhile, delivery status and late delivery risk have been used as response variables separately.

Feature Selection

We selected Order Item Cardprod Id, Order Department Id, Market, Order Country, Order Region, Order Status, Order Item Discount Rate, Order Item Quantity, Order Profit, Type, Days for shipment (scheduled), Order year, month, week day, and hour as the features. Order Id, customer Id, Item Id, and order item total. have been dropped because they are unrelated to the delivery status. Order city and state have been dropped to avoid overfit and multicollinearity with other features such as order market. Order Zipcode contains a large amount of missing values. Order date has been divided into detailed features. Days for shipping (real) is a dependent variable instead of an explanatory variable. Order Item Discount and sales have been replaced by discount rate and profit.

Modeling

We first converted all the object variables into numerical ones, and we then divided the dataset into train and test sets (3:7). After scaling all the variables, we test their fidelity on different algorithms and response variables (delivery status/late delivery risk). And last, we did the cross-validation on each model and compared their performance. Overall, the cross-validation suggested high confidence in each model, and the models generally perform better when using delivery status as the dependent variables, with an around 20% increase of accuracy for all the models. Especially the classification of late delivery almost has no mistakes for all models. The difference in prediction accuracy

may be due to the existence of canceled shipments. Respectively, knn has an overall accuracy of 84.9%, 85.9% for logistic regression, 85.1% for naïve bayes, and 86.5% for decision tree. When using late delivery risk as the response variable, the best performance is made by decision tree as well, with an accuracy of around 70%. In conclusion, decision tree should be selected as a better algorithm to predict the classification of the delivery status.