Late Delivery Risk Analysis

Andhyka Cakrabuana Adhitama

Business Overview

Exposed to the wave of the Covid-19 pandemic, the marketplace industry experienced an exponential surge in sales. The hashtag #dirumahaja one of the triggers that ultimately encourages people to shop online, which ultimately brings benefits to certain lines of business, one of which is a third-party logistics service provider (Third Party Logistics). Competition among logistics service providers is heating up, new start-ups are emerging to answer the needs of order fulfillment services. After the pandemic subsided, the competition between logistics companies finally tried to offer various new benefits to support changes in business models for customers. B2B services to deal with customers who increasingly want to shop in offline stores, cooperation with several delivery service providers, and of course, **increase order fulfillment speed**. Through this analysis, the authors hope to exchange ideas regarding what are the factors that influence the delay in order fulfillment using machine learning models.

Problems

→ **Late Delivery Risk factor.**

What affects the level of order fulfillment speed, as well as the characteristics of orders that can trigger the emergence of late delivery risk.

→ **Recommendation**

What are the recommendations and improvised strategies that can increase the speed of order fulfillment and the preventive measures that can be taken to deal with delays.

Steps

→ **Data Preparation & Profiling**

→ **Exploratory Data Analysis**

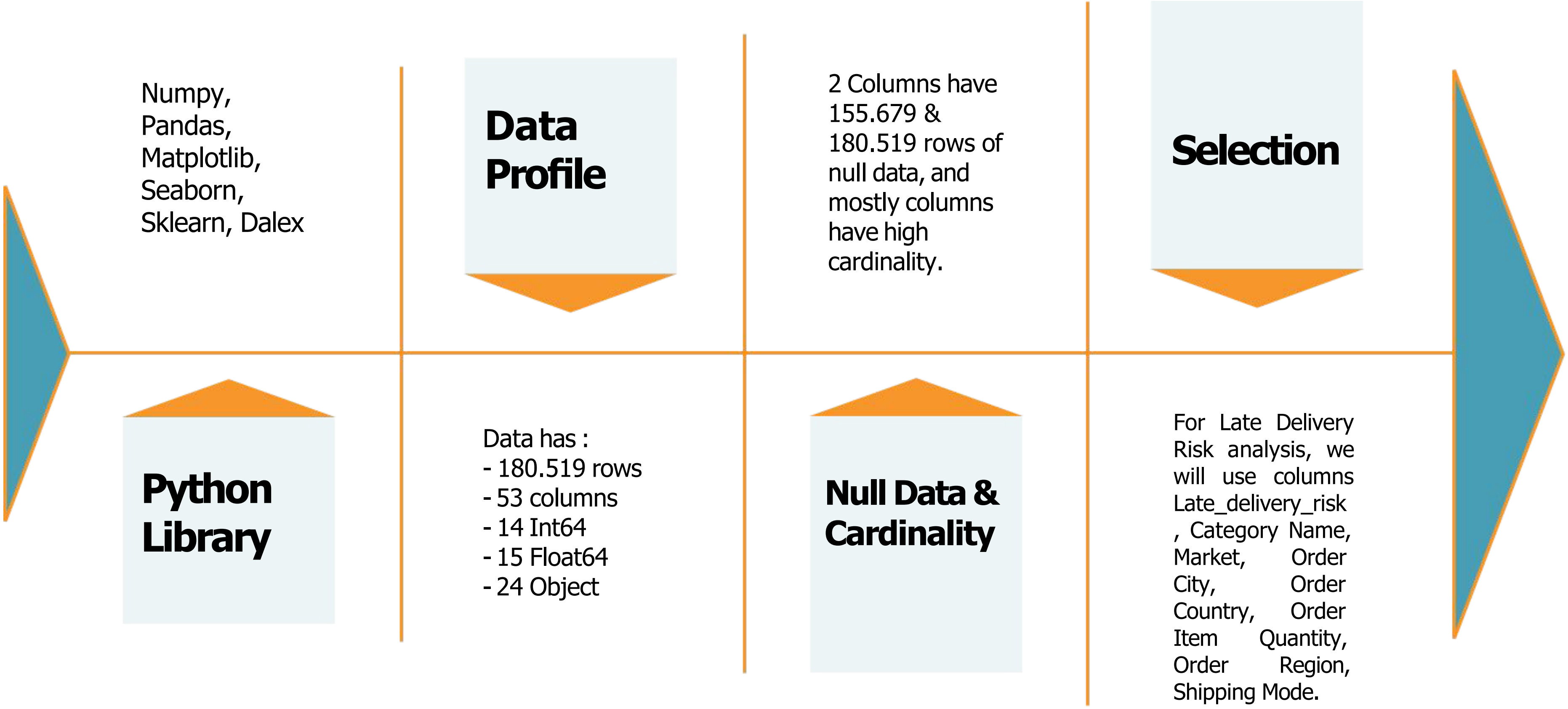
→ **Preprocessing**

→ **Modeling Preparation**

→ **Modelling**

→ **Recommendation**

DataCo Smart Supply Chain for Big Data Analytics



Python Library

Numpy,
Pandas,
Matplotlib,
Seaborn,
Sklearn, Dalex

Data Profile

Data has :
- 180.519 rows
- 53 columns
- 14 Int64
- 15 Float64
- 24 Object

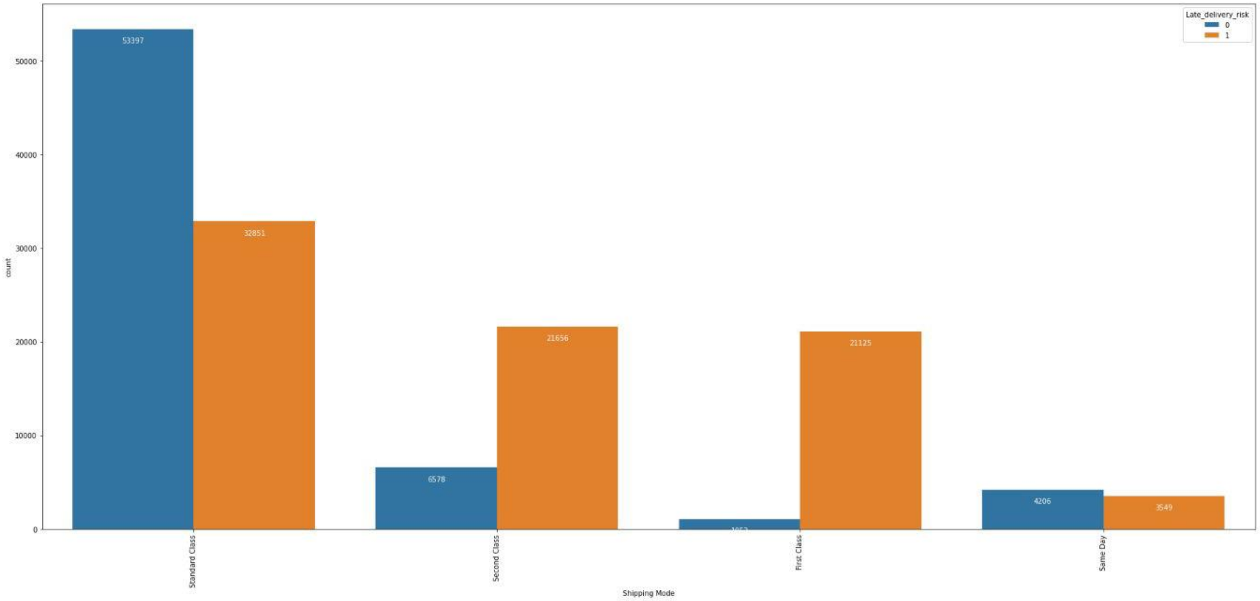
Null Data & Cardinality

2 Columns have 155.679 & 180.519 rows of null data, and mostly columns have high cardinality.

Selection

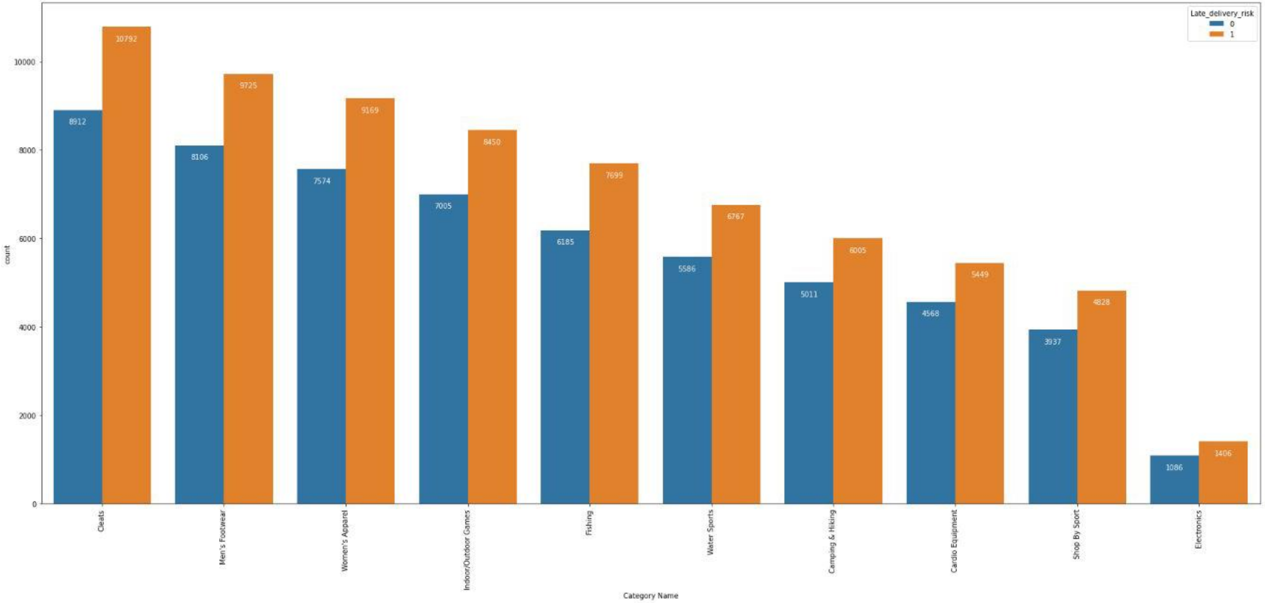
For Late Delivery Risk analysis, we will use columns Late_delivery_risk, Category Name, Market, Order City, Order Country, Order Item Quantity, Order Region, Shipping Mode.

Shipping Modes do matter!



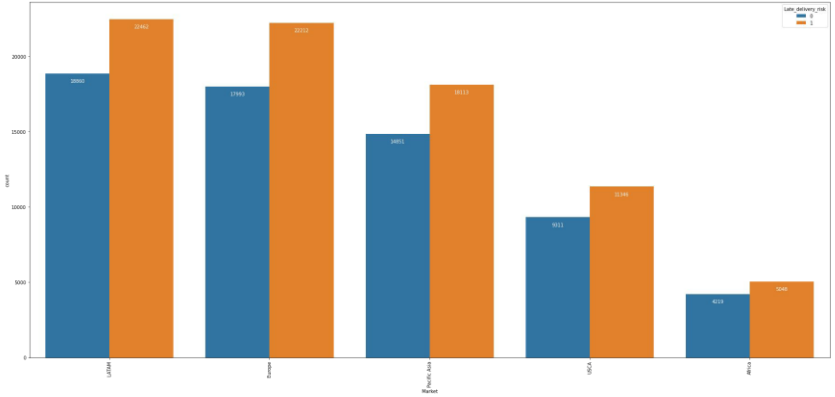
From the plot, we know that Shipping Modes can really determine in which late delivery risk more often to occur. As we can see the shipping mode that lowest chance to be over SLA is Standard Class and Same Day, however we could see that Same Day service have almost 50:50 chance for late delivery service.

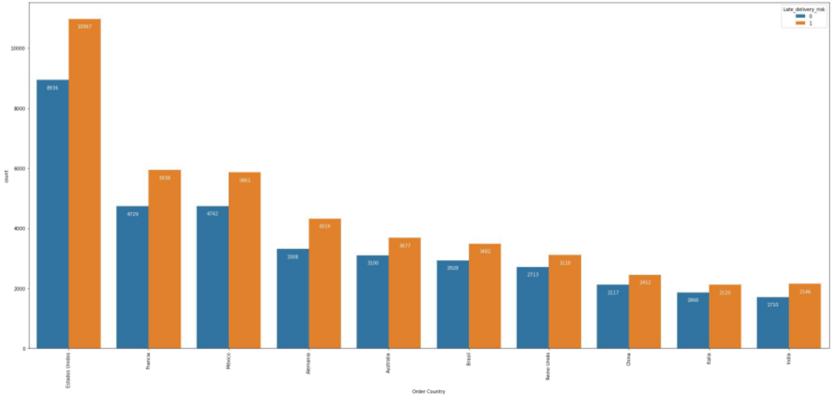
Category needs more deep analysis.

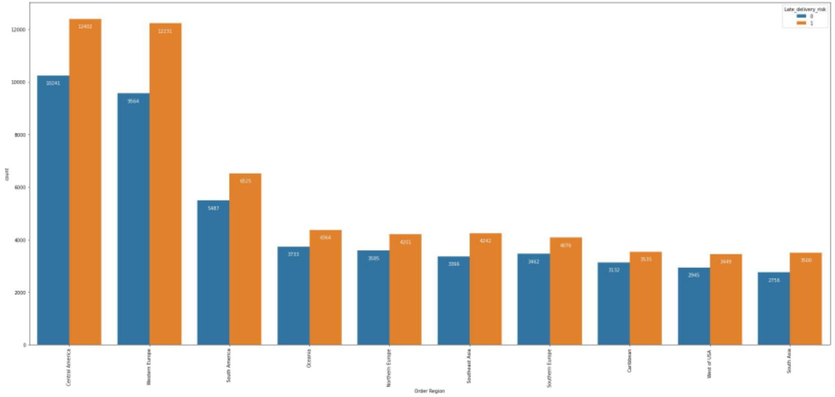


The Category didn’t seems very clear about determining late delivery risk. However, we know that some of category have different category movement. As we can see the more order for specific category is also more late delivery risk will show up.

The distance is not really a problem.







As we can see, there is no clear evidence that the late delivery is because of the distance between fulfillment center and the customers. However, we need further analysis is there any factors except distance that really make raise late delivery risk.

Lets Engineering our dataset!

01

Anomalies Handling, we eliminate the null values for 21 rows & duplicated values 7.908 rows.

02

Feature Selection, we drop the Order Id column since it’s only contain unique for every rows.

03

Encoding, we replace the high cardinality category columns with frequency encoding and binary columns with dummy encoding.

04

As we can observe from dataset so far, there is column like Order Quantity. For these, we using normalization in order to generalize all value into 0 ~1.

Choosing best model.

1 Decision Tree

F1 Score
0.6644

2 KNN

F1 Score
0.6803

3 XGB

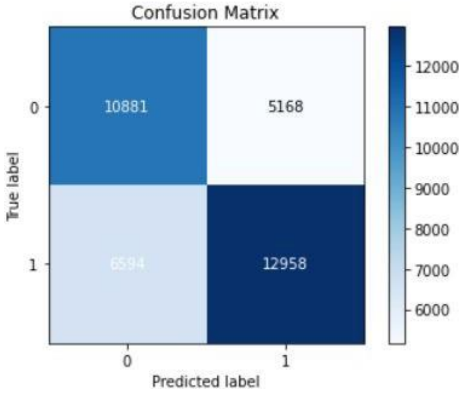
F1 Score
0.6633

4 Random Forest

F1 Score
0.6878

As we can see, the best model that best suit for our data set is **Random Forest** in terms of F1-Score.

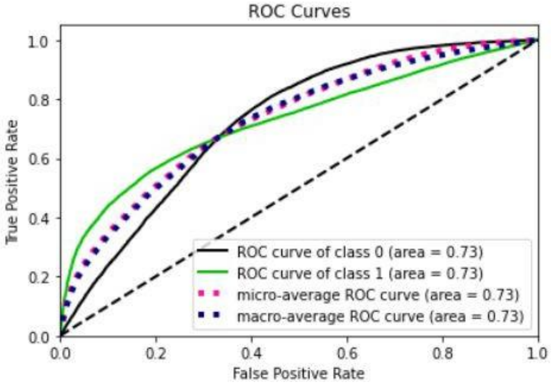
Let’s see how our model perform.



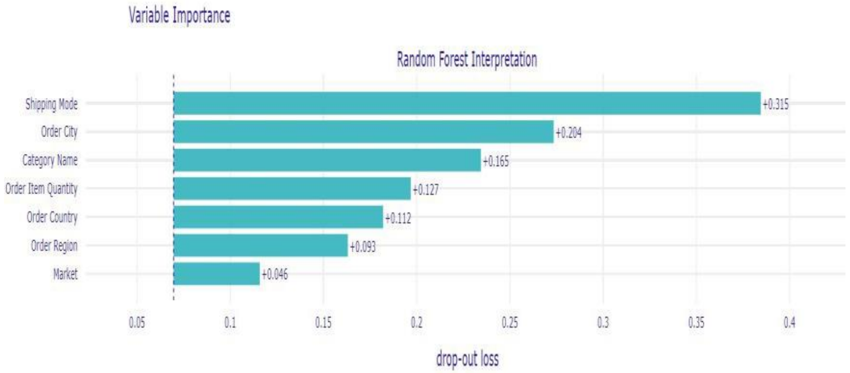
The confusion matrix show :

- The model predicted delivery will be late and actually late is 12958. (True Positive)
- The model predicted delivery will not late and actually late is 6394. (False Negative)
- The model predicted delivery will late and actually not is 5168. (False Positive)
- The model predicted delivery will not late and actually not late is 10881. (True Negative)

ROC Curvers show that our model success rate in distinct True Positive and True Negative is 73%. With the F1-Score (the harmony between precision and recall) is 68.7%.



Now let’s talk with our machine learning model.

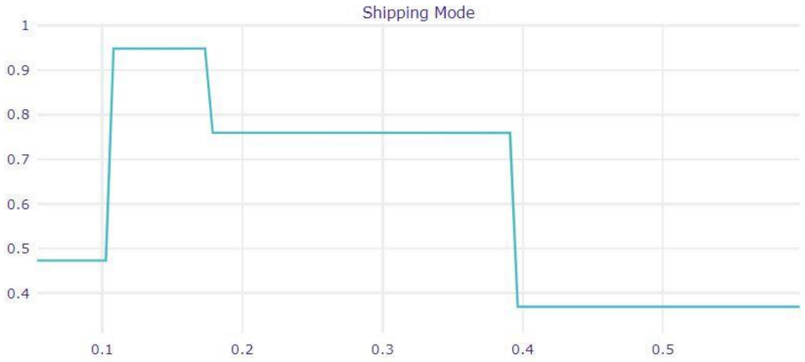


From the feature importance, we know that the highest four importance variable that help our model to determine late delivery risk is :

- Shipping Mode
- Order City
- Category Name
- Order Item Quantity

All of the following result really make sense. However, we also noticed that Order Item Quantity is also determine late delivery risk in which we didn’t found on first EDA. We need to do more analysis for this point.

From the partial dependence plot, we know that even there are many factor that impact late delivery, the higher effect comes from shipping mode. We know that if the more shipping mode get used by the company, the higher chance that the delivery will be on time.



Last but not least, improvement step!

From the EDA and maching learning modelling, can formulate recommendation like :

- **Check the possibility of closing few of Shipping mode. We found that the Standard Class have best performance amongst all, so we can recommend this shipping mode for customer. We can also opening partnership with third-party courier with more sorting hub since we know where the potential orders comes from.**
- **Doing more analysis for characteristic of the customer city. From the modelling we know that city more important than country and region in terms of predicting late delivery risk. Therefore, there is possibility that the distance is not so important. How about the road condition? Weather climate? or anything else that can determine late delivery risk on potential city.**
- **Initiate movement grouping for all Category. From the modelling we know that Category is also have higher importance in determine SLA. Therefore, we can make movement grouping to FAST, MED, and SLOW movement for the Category that company handle. So we can prepare the more effective layout for outbound staging to minimize the lead time for order processing.**
- **Calculate manpower more efficiently. As we can see, the order quantity is also one of important feature. The company can calculate several point and analyze the possibility where the most needed for additional manpower. Also we can add the facility to maximize the productivity like conveyor belt so the picker could deliver the goods to packing station without actually walking there.**

Let's connect to discuss!

Full Python Code