CAPSTONE PROJECT: DATA SCIENCE AND ON-TIME DELIVERY (OTD)

FINAL REPORT

Lucia Zhang | Apr 8, 2023

The Problem Statement and Background

- Problem statement: Poor On-Time Delivery negatively
  - impact customer satisfaction, loyalty, and the company's bottom line.
  - negative review can do serious damage; to outperform competitors: 95% or above
- Business statement:
  - Using machine learning algorithms, take real-time data-driven insights
  - Optimize and facilitate timely delivery in most effective way

The rise of e-commerce and new consumer preferences for on-time delivery (OTD) and instant shipments have led to industry giants like Amazon and Walmart offering faster and punctual delivery options. With the current trend, it is reasonable to expect the 95% OTD requirement to become the minimum standard.

OTD is a critical metric for measuring supply chain efficiency and maintaining customer satisfaction, making it essential to focus on raising and maintaining consistent OTD. It is crucial for converting one-time customers into repeat customers and creating loyalty. By increasing customer retention rates by 5%, businesses can increase profits by 25% to 95%, making it cheaper to retain customers than to acquire new ones.

In an increasingly digital world, by using machine learning algorithms to take real-time data-driven insights, businesses can optimize and facilitate timely delivery in the most effective way possible. To avoid costly customer churn and improve business sustainability, this project aims to measure the number of days off the expected delivery. Additionally, this project will identify opportunities to improve sales while addressing the key performance indicator: OTD

About the Data:

The project dataset from Kaggle is maintained transparently with the Creative Commons 4.0 license by Fabian Constante, Fernando Silva, and António Pereira through the Mendeley data repository. It consists of roughly 180k transactions from DataCo Supply Chain during 2015-2018 with late delivery 54.83% to every market.

The company's major products include clothing & footwear, sports supplies as well as electronics. The dataset includes 3 files:1. Structured Data : DataCo Supply Chain.csv; 2. Unstructured Data : tokenized_access_logs.csv; 3. Data Dictionary: description of variables

With 180519 records, 53 features fall into the following categories: order details, product, market, delivery, payment, and finance info.

Summary of Data Cleaning and Exploratory Data Analysis:

1. The data cleaning process involved managing data types and formats, identifying and handling null and missing values, checking and removing duplicated and redundant columns, as well as dealing with correlation and multicollinearity. After cleaning, the data frame shape is 180519 by 32.
2. Further investigation of columns and subcategories, such as Second Class (2-day delivery), revealed that Shipping Mode may be a major feature causing delays, which was later confirmed by features and their importance values.

3. Feature selection, creation, and engineering is a challenging process in this project. The backward approach was taken, creating a new feature 'days_difference' = 'Days for shipping (real)'- 'Days for shipment (scheduled)', starting with all relative features. However, the model was overfitting with 99.7% accuracy even after dropping ‘scheduled days'. It was only after further dropping 'Days for shipping (real)' that the model produced a reasonably good result with balanced accuracy, mean squared error (MSE), and mean absolute error (MAE).
4. As there are many different variables with different ranges, the features were scaled to a range of 0 to 1 to help models that are sensitive to the scale of the input features, such as KNN. The models were trained with the dataset to predict how many days off the expected delivery days, going beyond simply predicting late or not, giving a much better perspective to understand the last mile performance. This enables the company to identify potential areas for improvement and develop actionable business strategies to achieve quick wins
5. Prior to regression modelling, ensured no high correlations between variables by both heatmap and pair plot.

Insights, Modeling and Results

- 3 popular regression models for this dataset: KNN, Decision Tree, and Random Forest as a higher model because it combines multiple decision trees and aggregates their predictions to make a final prediction. This helps to reduce the overfitting issue of decision trees and improves the model's performance, it can also provide a measure of feature importance.
- Before fitting models, data was scaled to account for the model's sensitivity to variety range of input features. Random state 42 a specific number of n_numbers to ensure consistent and reliable results during the modeling process.
- Random forest regression model is computationally expensive, a timer was inserted to track and compare the running time between different scalers and independent variable sets. The R-squared (R2), Mean Squared Error (MSE) and Mean Absolute Error (MAE) are 3 metrics to evaluate how well the models can predict the target variable.
- Hyperparameter tuning, different scaler and feature engineering are applied to improve model performance.

Overall, these steps helped to ensure a robust and accurate analysis of the data. The models accuracy score as below:

	KNN	Decision Tree	Random Forest
R2 score	0.3997	0.6667	0.7454
Mean Squared Error	1.3379	0.7428	0.5673
Mean Absolute Error	0.8427	0.3136	0.5273

By evaluating the models based on these metrics, we can see Random Forest is the best model for our specific problem, the R2 score of 0.7452 can explain 74.52% the variance in the target variable, while MSE and MAE around 0.5 help us to understand that the model is performing very well.

Findings from Model and Importance Values:

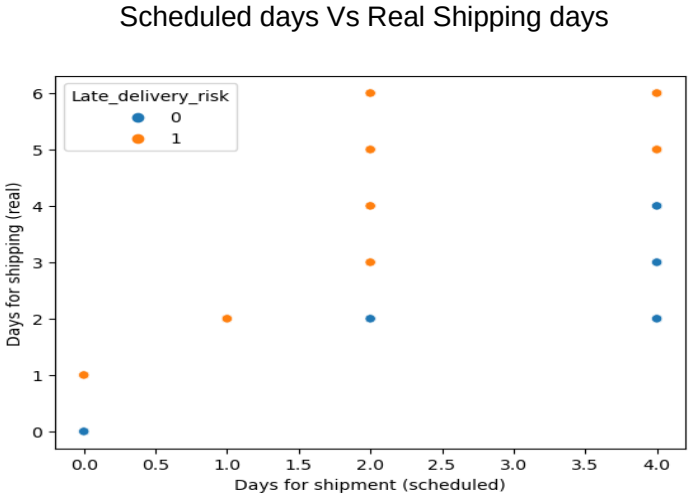
- The "Shipping Mode" feature, specifically the "Second Class" subcategory, is the most important feature to predict the target. This feature has the highest importance value of (0.220524), indicating it has a strong impact on the target variable.

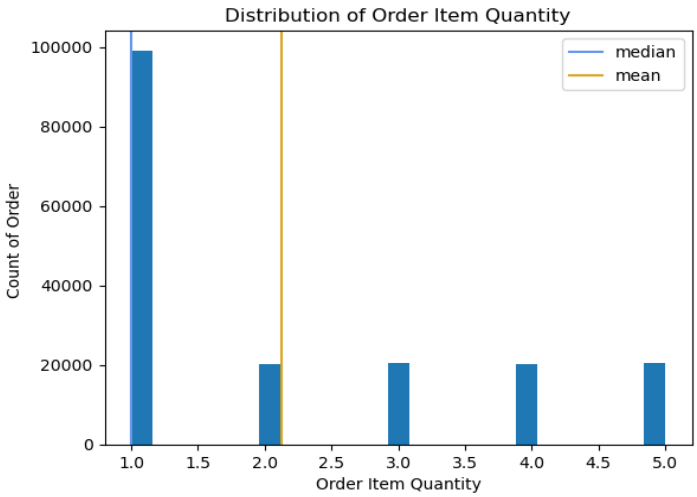
- Latitude and Longitude are the next, indicating customer's location also plays a role in determining the target variable.

- Followed by order time, which is understandable that during certain time of the year, on-time delivery can be affected by holidays, special occasions/situations like strikes, pandemic, etc. Jan: busiest shipping month for DataCo.
- Order Profit and Sales per customer: it is possible that higher profit and higher sales orders are given a higher priority in terms of shipping and delivery, as they may be more valuable to the company. But it's important to note that the choice of shipping mode is influenced by various factors, including cost, delivery time, customer preferences, and more.
- Order Status and Payment Type are also somewhat important for predicting the target variable. Though less so than the features above, it suggests that order with pending status may slow down the fulfillment process and has a negative impact on the target variable.
    1. both factors are related to the company's operational process
    2. Delays or slow processing of orders may lead to longer delivery times, affecting the target variable and on-time delivery.
    3. Therefore, it is important for the company to optimize its order processing and payment procedures to improve its delivery performance.

Suggestions: do more by doing less

1. It indicates that 'Days for shipment (scheduled)' as 2 days has higher risk of late delivery, it actually takes 3,4,5,6 days to arrive; shipment (scheduled) as 4 days actually takes 2,3,4 days to arrive early and on time, takes 5,6 days to be late for 1 day or 2 days.
2. It looks like 'Days for shipment (scheduled)' needs to be adjusted to reflect the actual days for delivery, or needs a better way to buffer and set realistic on time delivery goals.
3. It is worth the effort to investigate the carrier's performance together with internal order-fulfilment and shipping process, find the root cause, reduce wastes, and add values
4. Negotiate for better and faster shipping services. Optimize order and payment procedures to improve performance.
5. Top 5 order delivery cities are Santo Domingo, New York, Los Angeles, Tegucigalpa, Managua, provide a guide line for logistics/supply chain planning where to negotiate a 3rd party logistic center to improve the delivery efficiency and shipping cost savings
6. 38.39% orders come from the top 5 countries: United States, France, Mexico, Germany, Australia, it would be most effective and efficient to prioritize delivery and marketing plans/strategies in these 5 top countries,
7. Top 6 most ordered categories contributed ~80% of the total orders: Cleats, Men's Footwear, Women's Apparel, Indoor/Outdoor Games, Fishing, Water Sports. The company can expedite these categories and achieve easy wins by applying 20-80 rules: do more by doing less.
8. Opportunities to improve the sales while addressing OTD: majority orders with 1 item only, average 2 items, and max. 5, 59.69% of the shipment are by standard class, therefore the company could take advantage of recommender system, encourage shoppers to add more items, in return they can either save shipping cost or upgrade to a faster shipping mode.





Next Steps:

- Due to tight deadline and computationally expensive model, I'll continue Hyperparameter Tuning process afterwards.
- Will conduct recommender system analysis, so that the business can take opportunity to increase sales while improve on-time delivery performance.

Data Source and References:

- The dataset can be downloaded from Kaggle: https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis
- AWS last mile solution for faster delivery, lower costs, and a better customer experience, by Chen Wu and Manuel Baeuml | on 01 MAR 2023 https://aws.amazon.com/blogs/supply-chain/aws-last-mile-solution-for-faster-delivery-lower-costs-and-a-better-customer-experience/
- Retailers Double Down On Costly Last-Mile Arms Race- Forbes https://www.forbes.com/sites/gregpetro/2023/03/08/retailers-double-down-on-costly-last-mile-arms-race/?sh=458b03066e12
- Amazon Expands Same-Day Delivery, With Fees, While Battling Slow Growth-The Wall Street Journal https://www.wsj.com/articles/amazon-expands-same-day-delivery-with-fees-while-battling-slow-growth-344bd3a6?st=t2cek4jrellhnxo&reflink=desktopwebshare_permalink
- The "Matthew Effect" and Market Concentration: Jes´us Fern´andez-Villaverde University of Pennsylvania Federico Mandelman Federal Reserve Bank of Atlanta Yang Yu Shanghai University of Finance and Economics Francesco Zanetti∗ University of Oxford February 8, 2021 https://www.sas.upenn.edu/~jesusfv/Matthew_Effect.pdf
- The Pareto Principle- 80/20 Rule – Do more by doing less - Supply Chain Today https://www.supplychaintoday.com/the-pareto-principle-80-20-rule-do-more-by-doing-less/
- Explaining Feature Importance by example of a Random Forest: Learn the most popular methods of determining feature importance in Python https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e