IST 718: BIG DATA ANALYTICS
FINAL PROJECT REPORT

Supply Chain Analysis Using Spark

Group 6:

Elizabeth Jones
Tia Jones
Kishan Polekar
Prabin Raj Shrestha

The analysis of DataCo Global's supply chain dataset aims to mitigate late deliveries and optimize processes. Leveraging Spark, insights into provisioning, production, and distribution were gained. Notable findings include discrepancies in delivery time predictions and insights into order statuses and payment types. The analysis offers actionable insights for improving operational efficiency, customer satisfaction, and strategic decision-making in the company's supply chain

Project Goals

This report's objective is to conduct a comprehensive Delivery Risk Assessment, enabling accurate prediction of late deliveries. We aim to reduce these late delivery risks by identifying several factors that affect the scheduled deliveries and improving them so that the company does not bear huge losses. Using PySpark, we aim to gain insights into the supply chain dynamics, identify patterns, and optimize processes.

Data Overview

The DataCo dataset was sourced from Kaggle. It comprises 180,000 rows and 53 columns. It is a dataset of supply chain information and consists of order information including product and customer information as well as shipping and delivery details. Each row represents a single order, and orders span from 2015 to 2018. Below is a brief overview of our data:

1. Transaction Details:
    - Type: Type of transaction made.
    - Days for Shipping (Real): Actual shipping days of the purchased product.
    - Days for Shipment (Scheduled): Days of scheduled delivery of the purchased product.
2. Financial Information:
    - Benefit per Order: Earnings per order placed.
    - Sales per Customer: Total sales per customer made.
3. Customer Information: *[Customer City, Country, Email, Fname, Id, Lname, Password, Segment, State, Street, Zipcode] :*
4. Details of the customer making the purchase.
5. Store Information: *[Department Id, Name, Latitude, Longitude, Market]:*
6. Details of the store where the purchase is registered.
7. Order Details: *[Order City, Country, Customer Id, Date, Id, Item Cardprod Id, Item Discount, Discount Rate, Item Id, Item Product Price, Profit Ratio, Quantity, Sales, Total, Profit Per Order, Region, State, Status]:*
8. Information related to the order itself.
9. Product Information:*[Product Card Id, Category Id, Description, Image, Name, Price, Status]:*
10. Details of the product being purchased.
11. Shipping Information: *[Shipping Date, Mode]*

Before working with our dataset, we conducted some data cleanup. We used a combination of removal and imputation to handle missing zip code data. We also removed unnecessary columns and renamed columns to maintain uniformity in our dataset. Lastly, we parsed date columns and created several new columns including “late_days,” which is the

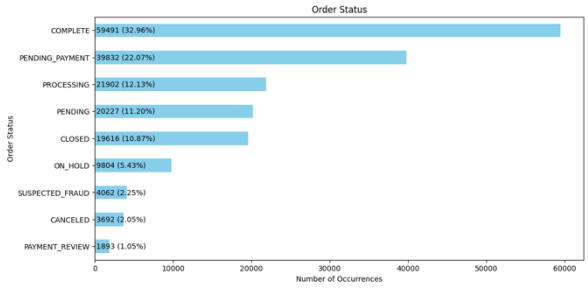
difference between the anticipated shipping days and actual shipping days, and “Late_Delivery”, a column indicating whether the delivery was late or not late.

Data Exploration

Our preliminary exploration revealed that DataCo’s delivery time predictions were suboptimal. While their estimates ranged from 0 to 4 days, with an average of 3 days for delivery time, actual delivery times varied from 0 to 6 days, with an average of 4 days, indicating a significant discrepancy between predicted and actual delivery times.

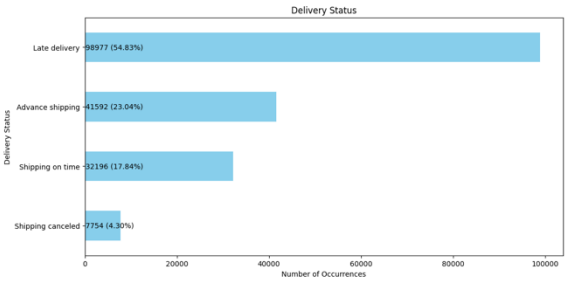
Order Status:

The distribution of order status across various categories revealed that the majority of orders, totaling 58,651 occurrences (51.96%), were marked as "COMPLETE." The "PENDING_PAYMENT" category followed closely with 30,822 occurrences (27.07%), while "PROCESSING" and "PENDING" accounted for 25,052 (22.13%) and 14,227 (11.20%) occurrences, respectively.



Delivery Status:

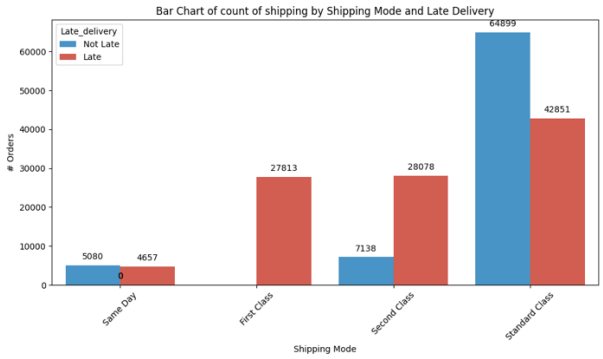
The "DEBIT" payment type emerged as the most common, with 59,295 occurrences (38.59% of the total). "TRANSFER" and "PAYMENT" closely followed with 48,883 (27.63%) and 41,725 (23.11%) occurrences, respectively, while "CASH" was the least frequent type, with 19,816 occurrences (10.87% of the total).



Shipping Mode:

An examination of shipping modes revealed valuable insights. For non-late deliveries, the median real shipping days effectively remained at 0, indicating minimal delays and no significant outliers. However, late standard mail deliveries exhibited considerable delays, with a median of 4-5 days and some outliers extending beyond 5 days, suggesting substantial delays for late deliveries via this shipping mode. The box plot analysis further highlighted that late deliveries generally experienced longer real shipping days (delays) compared to non-late deliveries, with the magnitude of delay varying across different shipping modes. Standard mail and first-class shipments demonstrated minimal delays for non-late deliveries, suggesting efficient operations. Conversely, courier services (both

ground and standard) tended to have slightly longer de lays, even for non-late deliveries, indicating potential areas for improvement in delivery efficiency.



Correlation Matrix:

Scheduled Days for Shipping and Late Delivery Risk:

- A moderate positive correlation (0.37) exists between scheduled days for shipping and late delivery risk.
- Longer scheduled shipping times are associated with a lower risk of late delivery, indicating potential prioritization of orders with longer shipping durations to ensure timely delivery.

Shipping Mode and Late Delivery:

- A moderate positive correlation (0.35) is observed between shipping mode and late delivery.
- Certain shipping modes may demonstrate a lower risk of late delivery, highlighting the importance of selecting appropriate shipping methods to minimize the likelihood of delays.

Modeling

We constructed both Random Forest and Logistic Regression models for predicting late deliveries. Logistic regression, being a widely used algorithm for binary classification tasks, was deemed suitable for our prediction objective.

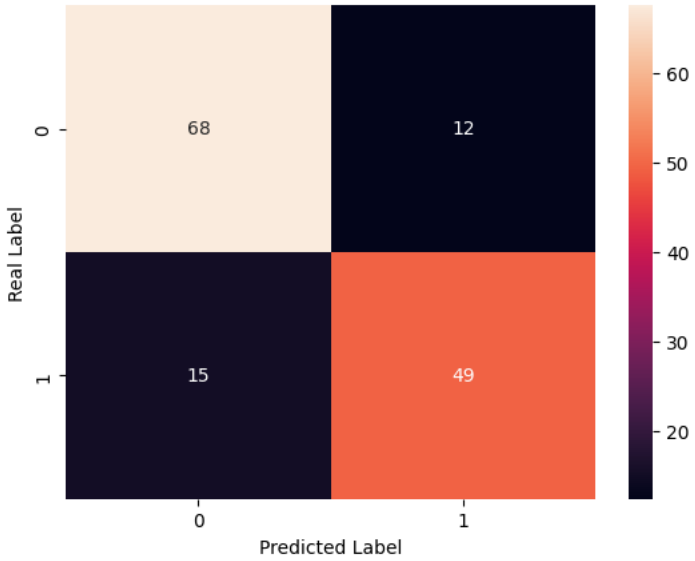
The models input variables encompass essential aspects of the supply chain and customer-related data, including scheduled shipping times, shipping modes, detailed order information (such as item category, discount, price, and profit), customer details (city, country, segment), and product specifics (category). The target variable, late delivery, serves as a critical metric for assessing operational efficiency and customer satisfaction, providing insights into the effectiveness of the supply chain management processes.

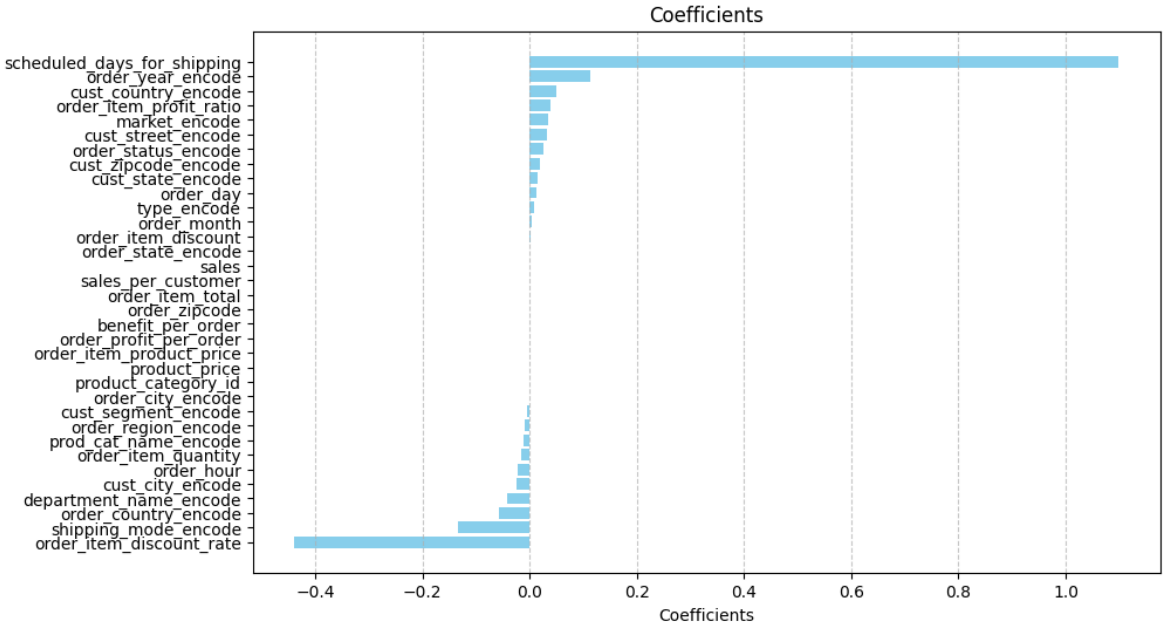
Results

The logistic regression model exhibited a much stronger performance metrics than the decision tree with high accuracy, precision, recall, and F1 score.

Logistic Regression Scores:

- Accuracy: 0.81
- Precision: 0.80
- Recall: 0.76
- F1 Score: 0.78
- AUC: 0.81





Top Features:

The top features contributing to the predictive performance of the logistic regression model are as follows:

1. Scheduled days for shipping
2. Order item discount rate

3. Shipping mode
4. Order year
5. Order Country
6. Customer Country
7. Department Name
8. Order item Profit Ratio
9. Market
10. Customer street

The top features identified by the logistic regression model offer valuable insights into the factors influencing delivery performance. Scheduled days for shipping, representing the planned duration for order shipment, serve as a crucial determinant of delivery timelines. Longer scheduled times may signal potential delays in processing, packaging, or transportation, directly impacting customer satisfaction.

Moreover, the order item discount rate emerges as a significant predictor, as discounts can influence consumer behavior and order volumes, potentially straining logistics operations. Variations in shipping mode, the third key feature, directly affect delivery timelines and service quality.

Interesting/Surprising Results

The analysis revealed an intriguing observation - despite orders being canceled, the company appears to generate an average revenue of $20 per canceled order. This prompts further investigation into the mechanisms driving profitability from canceled transactions. Potential factors contributing to this phenomenon may include cancellation fees, retention of partial or full payment amounts, and effective inventory management strategies. However, a complete understanding of this profitability hinges on a detailed examination of the company's specific policies, processes, and contractual agreements governing order cancellations. While existing research offers insights into tracking, reducing, and recovering canceled orders, it does not directly address the policies that could clarify the profitability from canceled orders.

Another interesting observation was found when viewing orders shipping first class. Although 100% of these orders were expected to ship out in 1 day, they took 2 days to ship. When we looked closer at the data, we discovered two things. That most shipments were ordered on Sunday (which would explain the delay), and that after Sunday there is a small spike on Thursdays that slowly decreases to the end of week. It is likely that staffing is not matching the volume of orders shipping out each day and that additional staff or overtime is needed. In essence, while our analysis sheds light on these intriguing finding, a thorough understanding requires access to internal company policies and practices, which falls outside the scope of this project.

Conclusion:

The analysis and modeling efforts undertaken have provided insights into the dataset containing information related to transactions and orders. Through a systematic data processing pipeline and the utilization of logistic regression modeling, several key findings were uncovered.

The top features identified by the logistic regression model shed light on the factors influencing late deliveries, including scheduled days for shipping, order item discount rate, shipping mode, and others. These features underscore the importance of efficient management of scheduled ship times, consideration of discount impacts, and selection of optimal shipping modes in minimizing late deliveries and enhancing overall customer experience.

The combination of effective data processing, robust modeling techniques, and insightful analysis has enabled a comprehensive understanding of delivery operations and late delivery prediction within the organization. By leveraging these findings, stakeholders can make informed decisions to drive operational efficiency, improve customer experiences, and achieve strategic objectives.

Reference

- Cleaning Data with PySpark Python (2023) https://www.geeksforgeeks.org/cleaning-data-with-pyspark-python
- PySpark Documentation
    - https://spark.apache.org/docs/latest/api/python/index.html
    - https://spark.apache.org/docs/latest/sql-programming-guide.html
    - https://pypi.org/project/pyspark/
    - https://spark.apache.org/docs/latest/ml-pipeline.html
- Jagdeesh, (2023) PySpark OneHot Encoding https://www.machinelearningplus.com/pyspark/pyspark-onehot-encoding/
- Nutan, (2023) Role of StringIndexer and Pipelines in PySpark ML Feature https://medium.com/@nutanbhogendrasharma/role-of-stringindexer-and-pipelines-in-pyspark-ml-feature-b79085bb8a6c