

Predicting Procurement Compliance Using KPI-Driven Machine Learning Models

Brittany M.D. Dowdle

Northwest Missouri State University
School of Computer Science and Information Systems
Maryville MO 64468, USA
S574362@nwmissouri.edu

Abstract. Keywords: procurement · compliance · prediction · product

1 Introduction

This project explores procurement performance using real-world data; the goal is to predict whether a purchase order will result in supplier compliance. Order compliance is critical in performance metrics. It reflects whether supplier contracts are meeting agreed-upon delivery schedules, product quality, and pricing terms. By building a predictive model based on attributes such as quantity, price deviation, or defect rates, the objective was to provide procurement teams with data-driven insights. This is useful to predict risk, improve supplier relationships, and enhance operational efficiency.

1.1 Goal of This Project

The final deliverables from this project include the following:

- ✓ Machine learning model predicting order compliance with supporting visualizations [4]
- ✓ A PDF report written in LaTeX via Overleaf
- ✓ A fully documented GitHub repository with Jupyter notebooks including EDA [4]

1.2 Project Resources

- **GitHub Repository:** <https://github.com/Bdowdle4/DowdleAnalyticsCapstone>
- **Overleaf Report:** <https://www.overleaf.com/read/bszyhdxsnrsf>
- **Pro Analytics 01:** <https://github.com/denisecase/pro-analytics-01>
The guide used to follow a repeatable workflow for professional python projects [1]

2 Collect and Describe Data

The data set analyzed consists of 777 purchase order records with 11 original columns. Each row in the data set represents a unique transaction with attributes that relate to the identity of the supplier, the characteristics of the order, the pricing, the defects, and whether or not the supplier was compliant. The data set was publicly accessible on Kaggle; the link is in Section 2.2 Dataset Resource.

2.1 Dataset Overview

- ◆ **Data Type:** Structured (Avg of 10 key features per PO)
- ◆ **Source:** Kaggle
- ◆ **Size:** 68 KB
- ◆ **Rows:** 777
- ◆ **Columns:** 11
- ◆ **File Extension:** .csv
- ◆ **Tool for Ingestion:** pandas in Python

The data set was downloaded from Kaggle [2] and moved to the project folder. Then it was added to the GitHub project repository in the "Data" folder. Finally, it was read into two Jupyter notebooks using the pandas library (Cleaning and Modeling).

2.2 Data Attribute Dictionary

| Column Name | Description | Data Type | Example |
|------------------|--|-----------|---------------|
| po_id | Unique identifier for the purchase order | String | PO-10231 |
| supplier | Supplier name or identifier | String | Supplier_A |
| order_date | Date the order was placed | Date | 2024-01-03 |
| delivery_date | Date the order was delivered | Date | 2024-01-11 |
| item_category | Category or type of item ordered | String | Raw Materials |
| order_status | Status of the order (e.g., Delivered, Pending) | String | Delivered |
| quantity | Quantity of units ordered | Integer | 500 |
| unit_price | Price per unit paid | Float | 12.75 |
| negotiated_price | Contractually agreed price per unit | Float | 12.00 |
| defective_units | Number of defective units in the delivery | Integer | 5 |
| compliance | Binary indicator of contract compliance | Integer | 1 |

Table 1. Original data attributes and examples.

2.3 Domain and Professional Description

Domain: Business Operations

Subdomain: Procurement / Supply Chain

This project would be important to:

1. Supply Chain Analysts - to identify patterns in KPI metrics
2. Procurement Managers - for supplier scorecards and vendor decisions
3. Chief Purchasing Officer (CPO) - to support strategic sourcing and policy decisions

This data set falls within the field of procurement analytics. Procurement professionals use analytics to track KPIs such as on-time deliveries, cost savings, and defect rates to measure supplier performance. When a supplier consistently does not meet the expected performance level, they are a financial and operational risk. Suppliers can be considered non-compliant for late deliveries, defective products, and violations of negotiated pricing. By identifying patterns in procurement data and predicting compliance outcomes, organizations can optimize sourcing strategies, negotiate better contracts, and reduce supply-side risk. [3]

2.4 Dataset Resource

Procurement KPI Analysis Dataset:

<https://www.kaggle.com/datasets/shahriarkabir/procurement-kpi-analysis-dataset>

As mentioned in the Kaggle Data Card, this data set is anonymized to protect company and supplier identities and provides real-world transactions of 5 different suppliers from 2022-2023. This data set reflects challenges such as supplier delays, compliance gaps, defects, and inflationary price trends over time. It is not expected to be updated. [2]

3 Clean and Prepare Data

The raw procurement data set must be cleaned and preprocessed to ensure consistency, accuracy, and usability. This section outlines the data cleaning and preparation steps completed in the Jupyter notebook named `cleaning.ipynb`. The goal was to ensure that it was ready for EDA in the next section.

3.1 Data Formatting and Standardization

To ensure consistency in data types and naming conventions:

- `order_date` and `delivery_date` were converted to datetime objects using `pd.to_datetime()`

- All column names were standardized to lowercase letters and any white space was removed using
`str.strip().str.lower()`

3.2 Handling Missing Values

Missing data was identified and addressed with conditional logic using

`df.info`

- `defective_units` missing values were assumed to have no defects and replaced with 0 using
`.fillna(0)`
- Missing `delivery_date` effected 87 rows and that seemed like a significant loss of data. So, instead, the rows that had an `order_status` of "Delivered" were kept and used a combined median imputation and a flag column to maintain data integrity.
- Each supplier's median `lead_time_days` was calculated using
`.dropna(subset=['lead_time_days']).groupby('supplier')['lead_time_days'].median()`
- The column to flag missing delivery date was created using
`df['delivery_date'].isnull() & (df['order_status'].str.lower() == 'delivered')`
- The column to flag missing delivery date was converted to an integer using
`.astype(int)`
- Imputed `delivery_date` used `order_date` + median lead time per supplier.
- 19 rows still had missing `delivery_date` after this and had an `order_status` of "cancelled", "pending", or "partially delivered". They were removed because they represent incomplete transactions and could introduce bias or noise into the model.

3.3 Feature Engineering

Two more attributes were created to allow for a deeper analysis of procurement efficiency and quality because they will support more meaningful comparisons between suppliers and orders.

- `price_diff` measures the difference between the `negotiated_price` and the `unit_price`.
- `defect_rate` calculates the proportion of `defective_units` relative to the `quantity` ordered.

3.4 Outlier Detection and Treatment

Outliers were identified using the IQR method for the following 4 attributes: `quantity`, `unit_price`, `price_diff`, and `defect_rate`. Outliers were identified in 15 of 777 rows. This was equal to 1.93% and was assumed to be a natural variance. The treatment decided for these rows was to leave as is.

3.5 Exporting Cleaned Dataset

The cleaned data set contained 15 attributes and 758 records. It was exported as a CSV file using

```
df.to_csv()
```

It is saved as `cleaned_procurement_data.csv` in the project repository Data folder. The figure below is an image of the Jupyter notebook output that showed the first few rows using

```
print(df.head())
```

| | po_id | supplier | order_date | delivery_date | item_category | \ |
|---|--------------|-----------------|----------------------------|------------------|-----------------|---|
| 0 | PO-00001 | Alpha_Inc | 2023-10-17 | 2023-10-25 | Office Supplies | |
| 1 | PO-00002 | Delta_Logistics | 2022-04-25 | 2022-05-05 | Office Supplies | |
| 2 | PO-00003 | Gamma_Co | 2022-01-26 | 2022-02-15 | MRO | |
| 3 | PO-00004 | Beta_Supplies | 2022-10-09 | 2022-10-28 | Packaging | |
| 4 | PO-00005 | Delta_Logistics | 2022-09-08 | 2022-09-20 | Raw Materials | |
| | order_status | quantity | unit_price | negotiated_price | defective_units | \ |
| 0 | Cancelled | 1176 | 20.13 | 17.81 | 0.0 | |
| 1 | Delivered | 1509 | 39.32 | 37.34 | 235.0 | |
| 2 | Delivered | 910 | 95.51 | 92.26 | 41.0 | |
| 3 | Delivered | 1344 | 99.85 | 95.52 | 112.0 | |
| 4 | Delivered | 1180 | 64.07 | 60.53 | 171.0 | |
| | compliance | lead_time_days | delivery_date_missing_flag | price_diff | \ | |
| 0 | Yes | 8.0 | | 0 | 2.32 | |
| 1 | Yes | 10.0 | | 0 | 1.98 | |
| 2 | Yes | 20.0 | | 0 | 3.25 | |
| 3 | Yes | 19.0 | | 0 | 4.33 | |
| 4 | No | 12.0 | | 0 | 3.54 | |
| | defect_rate | | | | | |
| 0 | 0.000000 | | | | | |
| 1 | 0.155732 | | | | | |
| 2 | 0.045055 | | | | | |
| 3 | 0.083333 | | | | | |
| 4 | 0.144915 | | | | | |

Fig. 1. Cleaned dataset head.

4 Exploratory Data Analysis (EDA)

4.1 Trend Line Charts

4.2 Correlation Matrix

4.3 Vendor Comparisons

5 Model and Generate Insights

5.1 Chosen Models

5.2 Parameters

5.3 Metrics

6 Present Results

6.1 Summarize KPIs

6.2 Plot Model Predictions

6.3 Maybe Dashboard Style Visuals?

7 Finalize Deliverables

7.1 GitHub Repo Completeness

7.2 Overleaf Report Completeness

7.3 Limitations

7.4 Future Work

8 Giving Credit Where It's Due

References

1. Case, D.: Git add, commit, push guide. [bluehttps://github.com/denisecase/pro-analytics-01/blob/main/03-repeatable-workflow/06-git-add-commit-push.md](https://github.com/denisecase/pro-analytics-01/blob/main/03-repeatable-workflow/06-git-add-commit-push.md) (2023), accessed: 2025-06-29
2. Kabir, S.: Procurement kpi analysis dataset. [bluehttps://www.kaggle.com/datasets/shahriarkabir/procurement-kpi-analysis-dataset](https://www.kaggle.com/datasets/shahriarkabir/procurement-kpi-analysis-dataset) (2022), accessed: 2025-06-23
3. Team, T.E.: Why procurement is a good career: Key reasons to consider this path. [bluehttps://www.techneeds.com/2025/04/05/why-procurement-is-a-good-career-key-reasons-to-consider-this-path/](https://www.techneeds.com/2025/04/05/why-procurement-is-a-good-career-key-reasons-to-consider-this-path/) (April 2025), accessed: 2025-06-29
4. Team, U.E.: Top 100 data science project ideas for beginners. [bluehttps://www.upgrad.com/blog/data-science-project-ideas-topics-beginners/](https://www.upgrad.com/blog/data-science-project-ideas-topics-beginners/) (2024), accessed: 2025-06-23