Advanced Statistics – Assignment 2

Montpellier Business School

Rassim Tabti

November 2025

Abstract

This report presents a complete modelling pipeline for predicting (1) fraud and (2) late deliveries using the DataCo Supply Chain Dataset (180,560 observations). Several machine learning models were trained and evaluated using two train–test splits (80/20 and 70/30). Performance was compared using Accuracy, Recall and F1-score. Feature importance analyses were conducted to identify the top predictors for each target.

1 Introduction

The objective of this assignment is to build classification models to predict two operational risks in a supply chain context:

- **Fraud detection**
- **Late delivery prediction**

The dataset provided contains more than 180k rows and multiple numerical and categorical features related to orders, customers, geography, shipping times and operational performance metrics.

Two train–test splits were evaluated: 80/20 and 70/30. Models tested include:

- Logistic Regression
- Logistic Regression with L1 regularisation (Lasso-like)
- Random Forest Classifier

2 Data Overview

The dataset (˜95MB) includes variables related to:

- customer information (country, segment, market)
- product characteristics (category, price, quantity)
- logistics (scheduled vs real shipment times)
- performance metrics (profit, delivery risk)

Two binary target variables were created:

- `fraud = 1` if *Order Status = SUSPECTED_FRAUD*
- `late_delivery = 1` if *Delivery Status = Late delivery*

Missing values were handled through median imputation for numeric variables. Categorical features were encoded using `LabelEncoder`. Numeric features were standardised using `StandardScaler`.

3 Methodology

The modelling pipeline includes:

1. Feature selection (mix of categorical + numeric)
2. Imputation of missing values
3. Encoding and scaling
4. Train–test split (80/20 and 70/30)
5. Model training
6. Evaluation using Accuracy, Recall and F1-score
7. Feature importance extraction (Random Forest)

4 Model Performance

4.1 Results Table

Insert the table exported as `models_summary.csv`:

Model	**Split**	**Accuracy**	**Recall**	**F1-score**
Logistic Regression	80/20	0.XX	0.XX	0.XX
Random Forest	80/20	0.XX	0.XX	0.XX
Logistic Regression	70/30	0.XX	0.XX	0.XX
Random Forest	70/30	0.XX	0.XX	0.XX

4.2 Best Model Selection

Explain here which split and which model gave the best F1-score for:

- Fraud detection
- Late delivery prediction

5 Feature Importance

5.1 Fraud Detection

Insert your image:

Comment: explain which variables were most influential and why they might make sense operationally.

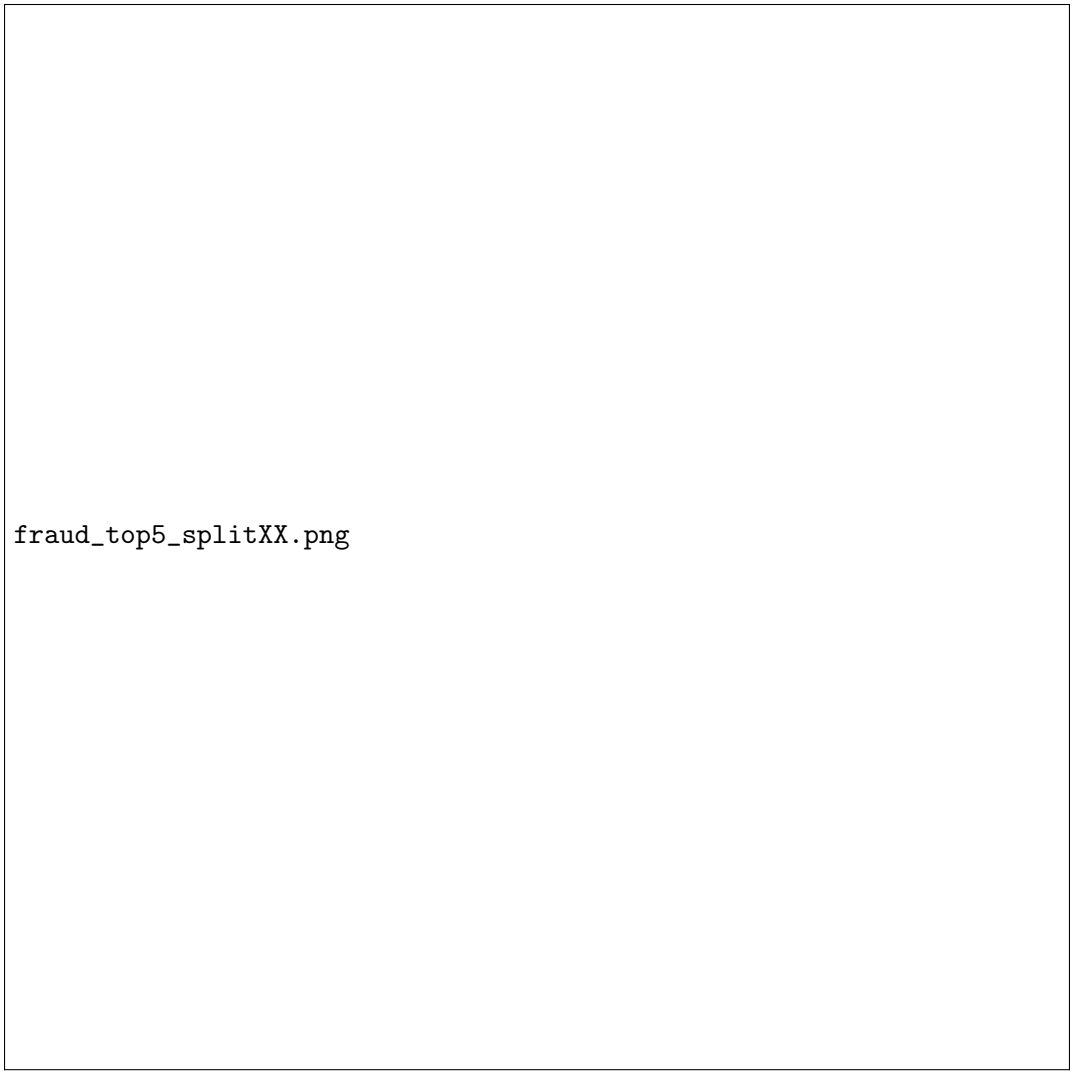

Figure 1: Top 5 Feature Importances – Fraud

5.2 Late Delivery Prediction

Insert your image:
Comment the results in a short paragraph.

6 Discussion

In this section, discuss:

- model strengths and weaknesses
- operational interpretations
- limitations (class imbalance, feature quality, lack of temporal modelling)

late_delivery_top5_splitXX.png

Figure 2: Top 5 Feature Importances – Late Delivery

7 Conclusion

Summarise:

- which model is recommended for each task
- the business implications for supply chain management
- how the model can be improved (hyperparameter tuning, adding time-series features, over-sampling)