Predicting On-Time Delivery Using Machine Learning

A data-driven approach to optimizing logistics performance

Project by: Dennis Irimu

Objective: Build predictive models to forecast shipment delivery outcomes using customer analytics data, enabling proactive risk management and improved operational efficiency.

Approach: Logistic Regression and Decision Tree Classifier models trained on real-world shipping data.

Background & Motivation

Operational Impact

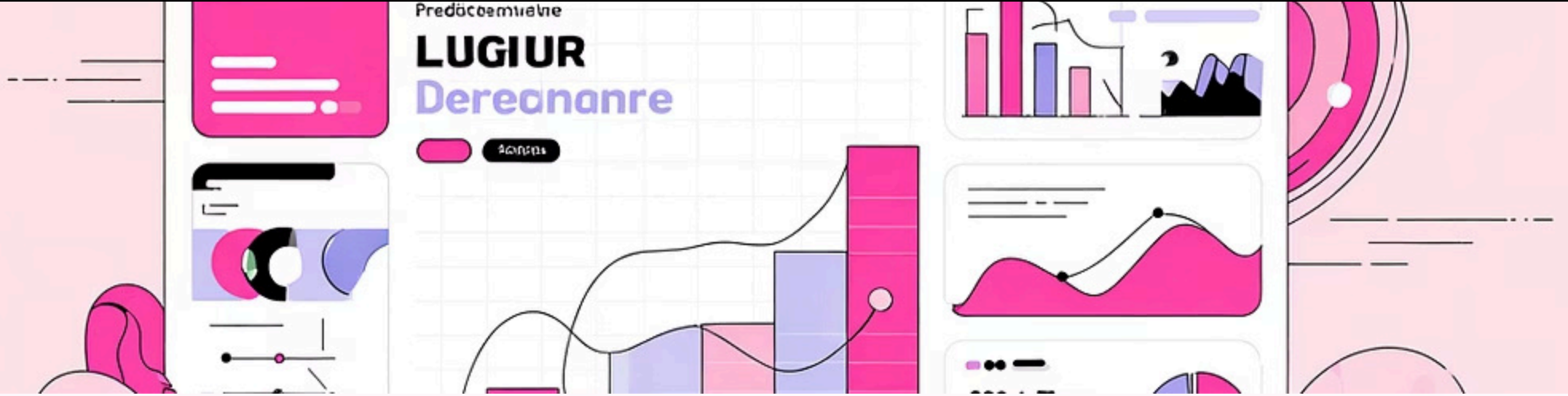
Timely delivery directly impacts customer satisfaction, retention, and brand reputation in an increasingly competitive market.

Complexity of Delays

Multiple factors influence delivery performance—shipment mode, warehouse efficiency, product prioritization, and distance all contribute to outcomes.

Data-Driven Solution

Predictive analytics transforms historical data into actionable intelligence, enabling logistics teams to anticipate and prevent delays before they occur.



Business Objective

1 Classify Delivery Outcomes

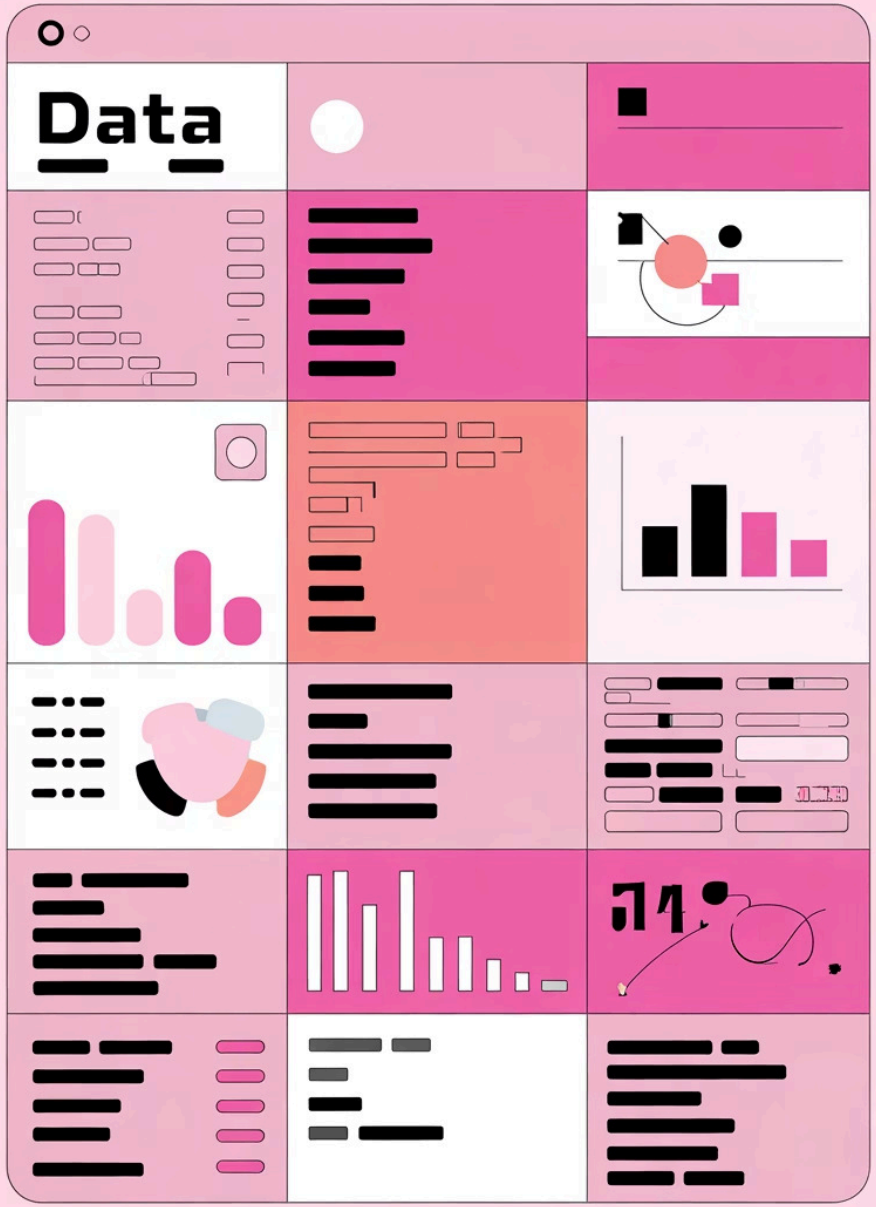
Predict whether each shipment will arrive on time (1) or experience delays (0) using shipment and product characteristics.

2 Enable Proactive Management

Identify at-risk shipments early, allowing logistics teams to allocate resources strategically and manage customer expectations.

3 Improve Reliability

Increase on-time delivery rates through targeted operational improvements and data-backed decision-making across the supply chain.



Dataset Overview

Customer Analytics dataset from Kaggle containing comprehensive shipping records and operational metrics.

10,999

Total Records

Shipment transactions analyzed

9

Core Features

Warehouse, shipment mode, product attributes

Key Variables: Warehouse block, shipment mode (Ship/Flight/Road), product importance rating, customer gender, shipment weight, delivery distance, product cost, and binary target: on-time arrival status.

Data Preparation & Processing

01

Missing Value Handling

Identified and appropriately treated incomplete records to ensure data quality and model reliability.

02

Categorical Encoding

Applied One-Hot Encoding to convert categorical variables (warehouse blocks, shipment modes) into numerical format for model compatibility.

03

Train-Test Split

Divided data using 80/20 ratio to ensure robust model evaluation on unseen data.

04

Feature Scaling

Normalized numerical features using standardization to optimize logistic regression performance and convergence.

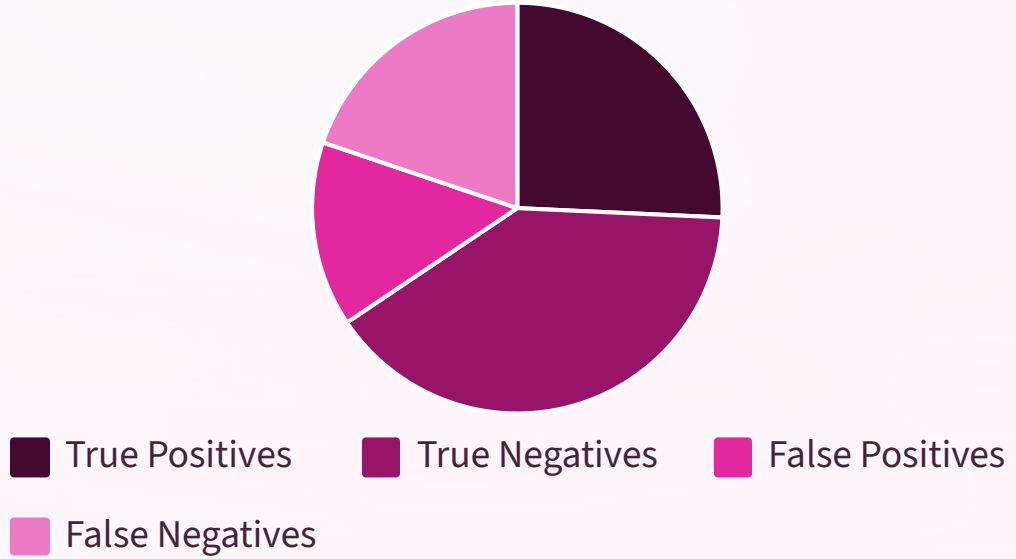
Result: Clean, balanced dataset ready for model training and validation.

Model 1: Logistic Regression

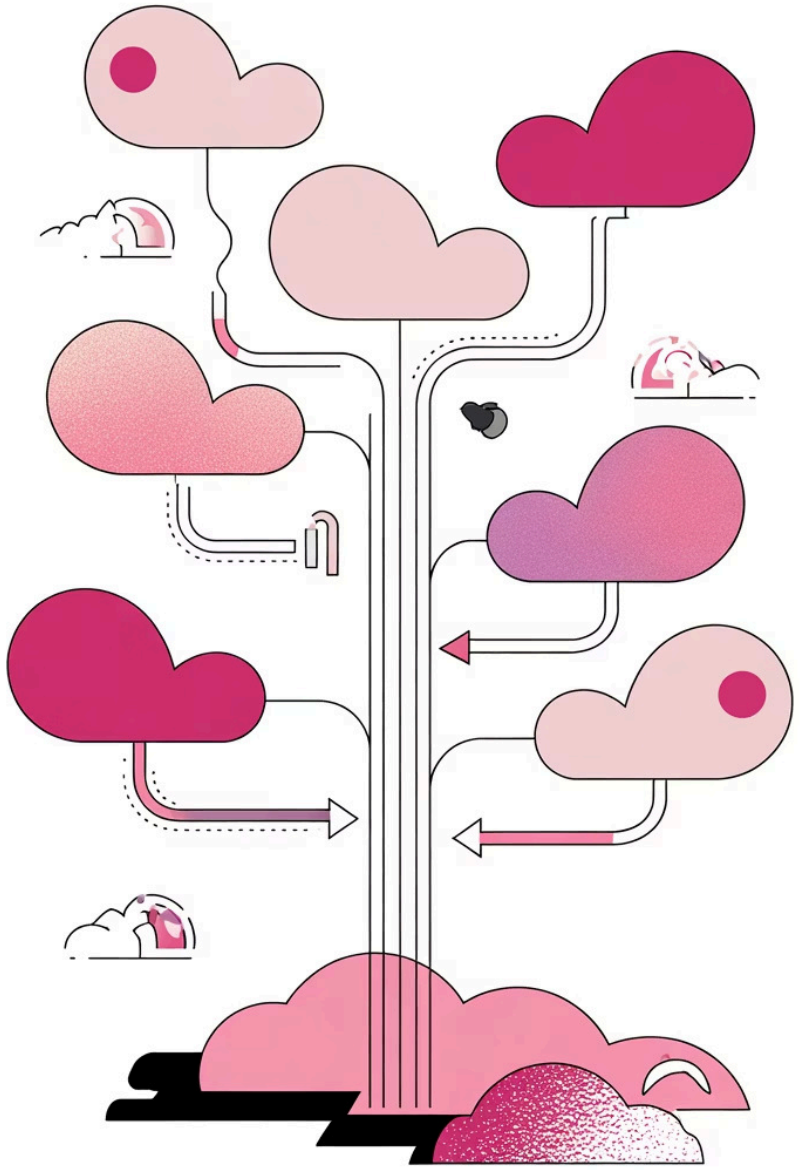
Baseline Classification Model

Linear model establishes performance baseline, providing interpretable coefficients and fast predictions for real-time deployment scenarios.

Accuracy: 65.6%



Interpretation: The model captures moderate predictive power but exhibits room for improvement in recall and F1 score, particularly in identifying late deliveries (false negatives).



Model 2: Decision Tree Classifier

Non-Linear Capability

Captures complex interactions and thresholds in delivery data that linear models miss.

Hyperparameter Tuning

GridSearchCV optimization of max_depth and min_samples_split parameters for balanced accuracy and interpretability.

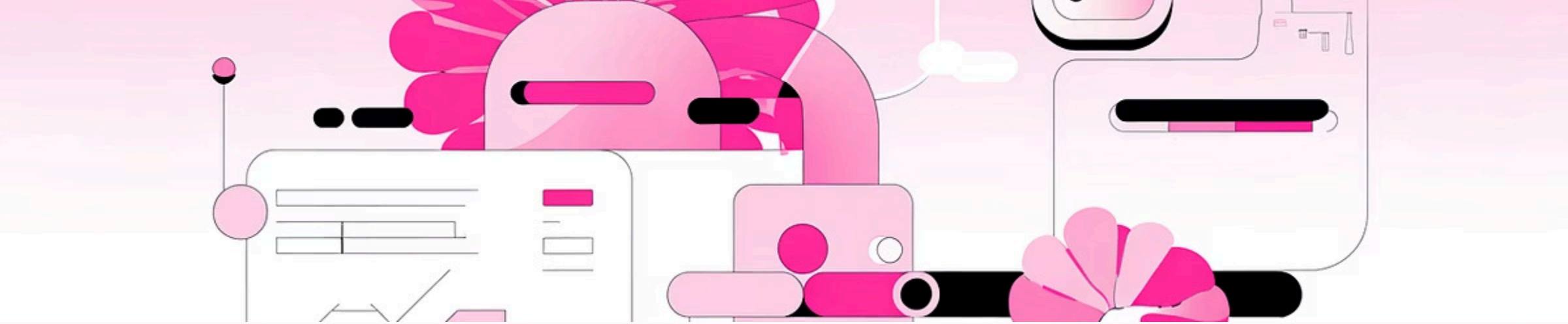
Improved Performance

Achieved better accuracy through tuning, with enhanced ability to identify key decision thresholds influencing on-time delivery.

Model Comparison

Dimension	Logistic Regression	Decision Tree
Complexity	Simple, linear	Flexible, non-linear
Speed	Fast inference	Moderate speed
Interpretability	Coefficients	Rule-based paths
Feature Interactions	Limited	Captures complex patterns
Accuracy	65.6%	Improved via tuning

Selection Criteria: Decision Tree selected based on superior cross-validation performance, explainability for stakeholder communication, and ability to identify actionable operational levers.



Key Insights & Feature Importance

Shipment Mode

Air shipments demonstrated significantly higher on-time probability compared to maritime and road options, indicating faster transit with fewer delays.

Warehouse Block

Specific warehouse locations consistently influenced delivery outcomes. Some blocks exhibited superior processing efficiency and fulfillment speed.

Product Importance

Low-priority products correlated with slower fulfillment times, suggesting resource allocation favors high-value shipments in operational scheduling.

Recommendations & Next Steps

Operational Improvements

Conduct detailed analysis of underperforming warehouse blocks. Implement targeted process improvements, staff training, and equipment upgrades to eliminate bottlenecks.

Resource Prioritization

Allocate premium logistics resources to critical and high-value product shipments. Establish priority queues and expedited handling protocols.

Dashboard Integration

Embed model predictions into real-time logistics dashboards. Enable proactive monitoring and early intervention for at-risk shipments.

Model Enhancement

Explore ensemble methods like Random Forests and Gradient Boosting to capture additional patterns and further improve prediction accuracy.