**NAME :NAVYASHREE J**

**SRN:PES1UG23AM180**

**SECTION: C**

**SCM Optimization using AI: Delivery Delay Prediction Report**

This report details the predictive modeling effort to forecast shipment delivery delays for a logistics company using historical data, based on the "SCM optimization using AI - Banana Problem" and updated with new performance metrics.

**Project Objectives**

The primary goals of this data analysis task were to:

**Predict** whether a shipment will be **delayed (1)** or **on time (0)**1.

**Identify key factors** (e.g., distance, vehicle type, weather, traffic, driver experience) contributing to delays2.

Provide insights to **improve logistics efficiency** and route planning3.

**Methodology and Model**

**Data Preprocessing**

The "Shipment Delivery Dataset" was preprocessed before model training4:

**Missing Values:** Missing numeric values in the 'Driver\_Experience\_Years' column were imputed using the **median**.

**Categorical Encoding:** Categorical features such as 'Origin', 'Destination', 'Vehicle\_Type', and 'Weather' were converted into a numerical format using **Label Encoding**.

**Model Selection and Training**

**Model:** A **Logistic Regression** model was chosen for this binary classification task.

**Target Variable:** The target variable was **'Delayed'** (0 = On-time, 1 = Delayed)5.

**Model Performance Evaluation (UPDATED)**

The Logistic Regression model was evaluated on the test set, and the performance metrics have been updated as requested.

**Performance Metrics Table (Updated)**

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| --- | --- | --- | --- | --- | --- |
| **Logistic Regression** | **0.96** | **1.00** | **0.80** | **0.89** | **0.90** |

**Detailed Explanation of Accuracies and Metrics**

The updated metrics show the model is **highly accurate** and **extremely precise** in identifying delays.

**Accuracy (96%)**: The model correctly predicted the outcome for **96%** of all shipments (both on-time and delayed) in the test set. This indicates excellent overall performance.

**Precision (100%)** for the **'Delayed' class (1)**: This is a perfect score. It means that **every single shipment the model predicted to be delayed was actually delayed**. In a business context, this is valuable because it guarantees that any operational intervention based on the model's prediction will not be a **false alarm** (i.e., minimal wasted resources).

**Recall (80%)** for the **'Delayed' class (1)**: This means the model successfully identified **80%** of the shipments that *actually* ended up being delayed. While excellent, the 20% of actual delays that the model missed (false negatives) represent opportunities for improvement in identifying *all* at-risk shipments.

**F1-Score (89%)**: This score balances the Precision and Recall. An F1-Score of 0.89 confirms a **strong classification performance**, driven by the high Precision.

**ROC-AUC (90%)**: The Area Under the Curve is 0.90, indicating the model has an **excellent ability to distinguish** between delayed and on-time shipments across different probability thresholds.

**Classification Report (Detailed Breakdown)**

| Class | Precision | Recall | F1-Score | Support |
| --- | --- | --- | --- | --- |
| **0 (On-time)** | 0.95 | 1.00 | 0.98 | 159 |
| **1 (Delayed)** | 1.00 | 0.80 | 0.89 | 41 |
| **accuracy** |  |  | 0.96 | 200 |
| **macro avg** | 0.98 | 0.90 | 0.93 | 200 |
| **weighted avg** | 0.96 | 0.96 | 0.96 | 200 |

The report highlights the model's strength in identifying the "On-time" class (Recall of 1.00) and its ability to ensure predictions of "Delayed" are correct (Precision of 1.00).

**Key Factors Contributing to Delays**

Based on the feature importance analysis (model coefficients) from the Logistic Regression model, the factors most influencing shipment delays are:

**Actual\_Delivery\_Hours**: This feature has the strongest positive correlation, confirming that longer actual transit time is the primary indicator of a delay.

**Distance\_km**: Longer shipment distances are a significant factor, increasing the probability of a delay.

**Scheduled\_Delivery\_Hours**: Shipments with longer planned delivery times are inherently more likely to face delays.

**Weather**: Different weather conditions have a substantial impact on delay risk.

**Route\_Traffic\_Index**: High traffic congestion is a key external factor contributing to delays.

**Prediction Output**

The final predictions were generated for all shipments 6and saved in the required format7.

**Output File:** Predicted\_Delays.csv

**Columns:** Shipment\_ID, Predicted\_Delay

**Example Predictions:**

| Shipment\_ID | Predicted\_Delay |
| --- | --- |
| S006 | 1 |
| S007 | 0 |
| S008 | 1 |

**Prediction CSV File: Predicted\_Delays.csv**

**Description:**

This CSV file contains predictions from the trained model for every shipment record.

It includes two columns:

Shipment\_ID: Unique identifier for each shipment.

Predicted\_Delay: Model’s binary prediction where **1 indicates a delay** and **0 means on-time delivery**.

**Purpose and Usage:**

Enables logistics teams to identify shipments at high risk of delay before dispatch.

Supports proactive decision-making to reroute, reschedule, or allocate resources efficiently.

Can be integrated into tracking dashboards or customer notification systems.

**Feature Importance File: Feature\_Importances.csv**

**Description:**

Lists all model input features along with their corresponding importance scores derived from the logistic regression coefficients.

**Positive values** indicate features that increase the likelihood of delay.

**Negative values** indicate features that reduce delay risk.

**Key Features Typically Ranked:**

**Actual\_Delivery\_Hours:** Higher values correlate strongly with delay likelihood.

**Scheduled\_Delivery\_Hours:** Negative influence, indicating that allowing more planned time reduces delays.

**Vehicle\_Type:** Certain transport modes (e.g., Trucks) may have higher delay risks.

**Route\_Traffic\_Index:** Higher traffic tends to increase delay probability.

**Driver\_Experience\_Years:** More experienced drivers tend to reduce delays.

**Weather Conditions:** Severe weather (Rainy, Stormy, Foggy) increases risk.

**Importance:**

Enables root cause analysis to target operational improvements and reduce delays.

Helps prioritize intervention areas such as route planning, vehicle assignments, and driver training.

**Model Performance Metrics Table & Classification Report**

**Metrics Included:**

**Accuracy:** Overall correctness of model predictions.

**Precision:** Percentage of predicted delays that were actual delays (minimizing false alarms).

**Recall:** Percentage of actual delays correctly identified (minimizing missed delays).

**F1-Score:** Harmonic mean of Precision and Recall; balances false positives and false negatives.

**ROC-AUC:** Ability of the model to discriminate between delayed and on-time shipments across threshold variations.

**Sample Evaluation Results:**

| **Metric** | **Score** |
| --- | --- |
| Accuracy | 0.96 |
| Precision | 1.00 |
| Recall | 0.80 |
| F1-Score | 0.89 |
| ROC-AUC | 0.90 |

**Interpretation:**

High precision (1.00) shows the model rarely mislabels on-time shipments as delayed.

Good recall (0.80) shows most actual delays are caught, though some are missed.

Strong F1 and ROC-AUC represent balanced and reliable classification.

**Usage:**

Model tuning and validation to ensure reliability before deployment.

Monitoring model drift or performance degradation over time.

**Python Script: shipment\_delay\_prediction.py**

**Contents:**

Complete, reproducible pipeline including:

Data loading and preprocessing (handling missing data, encoding categorical variables).

Model training using logistic regression.

Model evaluation with accuracy and classification metrics.

Generation of prediction outputs.

Extraction and saving of feature importances.

**Importance:**

Enables full transparency and reproducibility of the model.

Facilitates updates, improvements, or change of models in the future.

Serves as documentation and technical reference.

**Linking Outputs to Project Goals**

| **Output** | **How it Supports Goals** |
| --- | --- |
| Prediction CSV | Predict which shipments will delay; plan operations accordingly. |
| Feature Importance | Identify and understand key causes; focus on areas to reduce delays. |
| Performance Metrics | Validate models to ensure trustworthy predictions. |
| Python Script | Document and replicate entire process; enable future innovation. |
| **🚚 FourKites: Transforming Delay Prediction into Proactive Logistics**  The logistics company can significantly enhance its Supply Chain Management (SCM) optimization efforts and move beyond the current predictive model by adopting a Real-Time Transportation Visibility Platform (RTTVP) such as **FourKites**. While the developed Logistic Regression model achieves high prediction accuracy (96%) and perfect precision (1.00) in identifying delays, its reliance on historical data makes it primarily reactive. FourKites directly addresses the core challenge of multi-modal logistics by providing highly accurate **Dynamic ETAs®** across road, rail, and air, using **real-time data** from over 150 dynamic factors, including live weather and traffic conditions. This capability operationalizes the model. |  |