UNIVERSIDAD CATOLICA BOLIVIANA ¨SAN PABLO¨

Caso 6 — E-Commerce Logistics Optimization (Logística/Retail

INTEGRANTES:

Jean Marco Fernandez Silva

Sergio Alejandro Arias Mayta

Marvin Larry Mollo Ramirez

Jaime Ignacio Huaycho Clavel

Sergio Alexander Mendoza Choque

MATERIA: Machine Learning

AÑO: 2026

1. Descripción del Problema y Contexto del Negocio

1.1. Contexto Operativo

La base de datos histórica disponible, compuesta por 10,999 registros de envíos, revela patrones críticos sobre el comportamiento logístico y la satisfacción del cliente. Variables como el número de llamadas a atención al cliente (Customer_care_calls) y la calificación del servicio (Customer_rating) sugieren una correlación directa entre la eficiencia de la entrega y la percepción de marca. En un mercado donde la inmediatez es la norma, la capacidad de anticipar fallos en la entrega y entender la estructura de costos de los productos enviados es vital para mantener la competitividad y la rentabilidad operativa.

Actualmente, la empresa opera bajo un modelo reactivo: los problemas se gestionan una vez que el retraso ha ocurrido. El objetivo de este proyecto es transicionar hacia un modelo predictivo y proactivo mediante técnicas de Machine Learning, permitiendo a la gerencia tomar decisiones basadas en datos antes de que los incidentes impacten al cliente final.

1.2. Definición de los Objetivos del Proyecto

El proyecto se divide en dos flujos de trabajo analíticos complementarios, diseñados para abordar tanto la gestión financiera del inventario como la eficiencia operativa de los envíos:

A. Objetivo de Regresión: Estimación del Valor del Inventario El primer objetivo consiste en modelar y estimar la variable continua Cost_of_the_Product (Costo del Producto).

B. Objetivo de Clasificación: Predicción de Puntualidad (Critical Path) El objetivo central y más crítico del negocio es la predicción de la variable binaria Reached.on.Time_Y.N.

Es imperativo notar la codificación específica del dataset para este problema:

1. Clase 1 (Positiva): El producto NO llegó a tiempo (Retraso).
2. Clase 0 (Negativa): El producto llegó a tiempo.

2. Preprocesamiento de Datos
En esta fase se preparó el dataset para garantizar la calidad y compatibilidad con los algoritmos de Machine Learning. El pipeline de transformación incluyó los siguientes pasos críticos:

2.1. Limpieza e Inspección Inicial

-Carga del Dataset: Se ingirió el archivo Train.csv conteniendo 10,999 observaciones y 12 variables.

-Eliminación de Ruido: Se procedió a eliminar la columna ID (Identificador del cliente).

-Verificación de Calidad: Se realizaron comprobaciones de valores nulos (isnull().sum()) y duplicados (duplicated().sum()) para asegurar la integridad de los datos antes del modelado.

2.1. Codificación de Variables Categóricas

Se utilizaron dos estrategias distintas para transformar las variables categóricas en representaciones numéricas, preservando la naturaleza de los datos:

-Codificación Nominal (OneHotEncoder): Se aplicó a variables sin orden intrínseco para evitar que el modelo asume jerarquías falsas.

> *Variables:* Warehouse_block (Bloques de almacén A-E), Mode_of_Shipment (Marítimo, Aéreo, Terrestre) y Gender.
>
> *Resultado:* Conversión a vectores binarios dispersos (dummy variables).

-Codificación Ordinal (OrdinalEncoder): Se aplicó específicamente a la variable Product_importance para capturar la jerarquía de negocio.

> *Mapeo:* low (0) < medium (1) < high (2).
>
> *Justificación:* Esta codificación permite al modelo entender que una importancia "alta" tiene mayor peso o magnitud que una "baja".

3. Modelado

En esta etapa se implementaron algoritmos de aprendizaje supervisado para abordar los dos objetivos estratégicos del negocio: la estimación de costos (regresión) y la predicción de puntualidad (clasificación).

3.1. Estrategia de Regresión (Target: Cost_of_the_Product)

Se desarrollaron y compararon dos modelos lineales para establecer una línea base y evaluar la necesidad de regularización ante la posible multicolinealidad introducida por el OneHotEncoder.

3.1.1 Modelo Baseline: Regresión Lineal Simple (OLS)

-Configuración: Se utilizó el algoritmo de Mínimos Cuadrados Ordinarios sin penalización.

-Propósito: Servir como referencia de desempeño (Benchmark). Si un modelo más complejo no supera significativamente a este baseline, se prefiere la simplicidad del modelo lineal por su interpretabilidad.

3.1.2 Modelo Regularizado: Regresión Ridge (L2)

Se aplicó una penalización L2 a los coeficientes del modelo.

-Justificación Técnica: Dado que las variables dummies (generadas por Warehouse_block o Mode_of_Shipment) pueden estar altamente correlacionadas, la Regresión Ridge reduce la varianza del modelo contrayendo los coeficientes hacia cero (sin anularlos). Esto mejora la capacidad de generalización y mitiga el riesgo de *overfitting* en dimensiones altas.

3.2. Estrategia de Clasificación (Target: Reached.on.Time_Y.N)

Para la predicción binaria de retrasos, se seleccionó un modelo probabilístico robusto.

Algoritmo: Regresión Logística

-Configuración: Se empleó el optimizador estándar para maximizar la verosimilitud de la función Sigmoide.

4. Evaluación del Modelo

La evaluación se realizó por separado para regresión (estimación del costo) y clasificación (predicción de puntualidad), utilizando una división 80/20 entre entrenamiento y prueba para garantizar capacidad de generalización.

4.1 Evaluación del Modelo de Regresión

Target: Cost_of_the_Product

Se evaluaron tres modelos utilizando MAE, RMSE y R^2.

Modelo	**MAE**	**RMSE**
Linear Regression	0.7684	0.9271
Ridge	0.7684	0.9271
Lasso	0.8366	0.9797

4.2 Evaluación del Modelo de Clasificación

Target: Reached.on.Time_Y.N

Se utilizó división estratificada y un Árbol de Decisión.

Accuracy

- Accuracy = 0.6409 (64.09%)

	Predicción 0	**Predicción 1**

Real 0	471	416
Real 1	374	939

5. Interpretación y Decisiones

5.1 Interpretación Técnica

- El modelo de regresión presenta baja capacidad explicativa ($R^2 \approx 0.10$).
- La regularización no mejora el desempeño; la limitación es informativa.
- El modelo de clasificación tiene desempeño moderado.
- El recall de 0.72 en retrasos indica buena capacidad para detectar incumplimientos.

5.2 Implicaciones de los Errores

- Falsos Negativos (374): retrasos no detectados; mayor riesgo reputacional.
- Falsos Positivos (416): alertas innecesarias; costo operativo menor.

5.3 Decisiones Estratégicas

Respecto a la regresión:

- No se recomienda implementación productiva actual.
- Se requieren variables comerciales adicionales.
- Se pueden evaluar modelos no lineales como Random Forest o Gradient Boosting.

Respecto a la clasificación:

- El modelo tiene valor estratégico para anticipar retrasos.
- Se recomienda priorizar el recall de la clase 1 si se busca proteger la experiencia del cliente.
- Puede ajustarse el umbral para aumentar la detección, aceptando más falsas alarmas.

En conclusión, el mayor valor del proyecto se encuentra en el modelo de clasificación, ya que permite anticipar retrasos antes de que impacten al cliente, transformando la gestión

logística de un enfoque reactivo a uno predictivo. Con un recall del 72% en la clase de retrasos, el sistema puede identificar la mayoría de los incumplimientos y activar acciones preventivas como priorización de envíos, reasignación de recursos o comunicación anticipada con el cliente. Esto reduce el riesgo reputacional, mejora la experiencia del usuario y optimiza la toma de decisiones operativas basada en datos.