FACULTAD DE CIENCIAS ECONÓMICAS

FAMAF Facultad de Matemática, Astronomía, Física y Computación

UNC Universidad Nacional de Córdoba

DIPLOMATURA
CIENCIA DE DATOS, INTELIGENCIA ARTIFICIAL Y SUS APLICACIONES EN ECONOMÍA Y NEGOCIOS

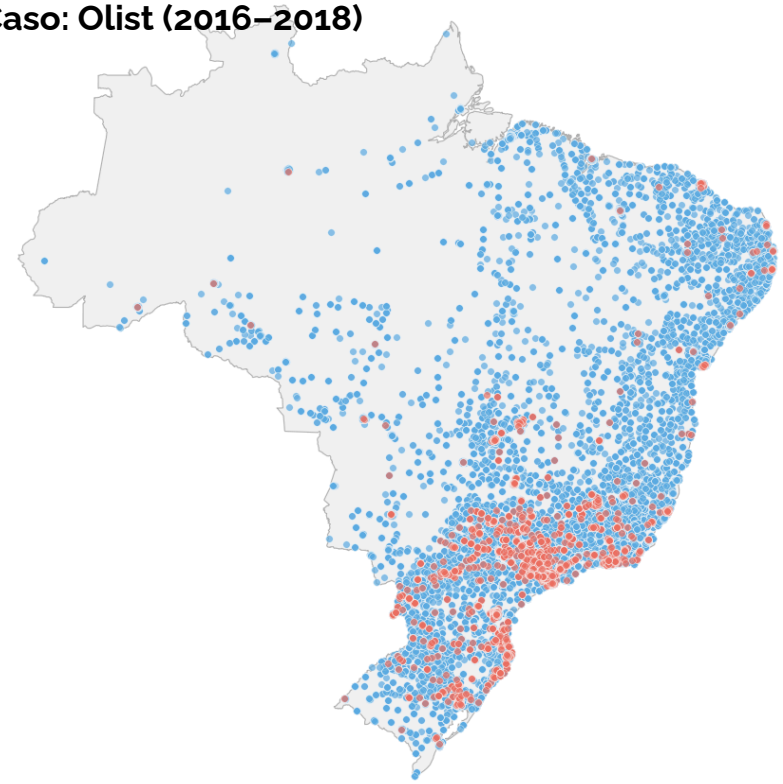
UNIVERSIDAD NACIONAL DE CÓRDOBA

FACULTAD DE CIENCIAS ECONÓMICAS – FAMAF

Diplomatura en Ciencia de Datos, Inteligencia Artificial
y sus Aplicaciones en Economía y Negocios

PREDICCIÓN DE VENTAS Y CALIDAD DE VENDEDORES

Caso: Olist (2016–2018)



Grupo 12

- **Herrera, Paula Alejandra**
- **López, Matías Fabián**
- **Rossi, Mauro Daniel**
- **Sibert, Maximiliano**

Tutor: Lic. Ignacio Fichetti

ÍNDICE

1. INTRODUCCIÓN **1**
2. METODOLOGÍA **2**
2.1. Datos y preparación 2
2.2 Análisis exploratorio y correlacional 2
3. INGENIERÍA DE VARIABLES **2**
3.1 Variables Creadas por Categoría 2
3.2 Prevención de Data Leakage 4
4. MODELOS PREDICTIVOS **4**
4.1 Modelos predictivos facturación 4
4.1.1 Variables analizadas 5
4.1.2 Resultados Principales 5
4.1.3 Validación y conclusiones 6
4.2 Modelo Predictivos Clasificación 7
4.2.1 Características 8
4.2.2 Modelo de regresión logística (Baseline) 8
4.2.3 Modelo de ML 8
4.2.3 Modelo Redes Neuronales 9
4.2.4 Resultados 10
4.2.5 .Conclusión 11
5. CONCLUSIÓN **11**

1. INTRODUCCIÓN

En el comercio electrónico, una sola reseña negativa puede reducir drásticamente la intención de compra de los consumidores. Detectar y anticipar estas reseñas es clave para preservar la reputación y la rentabilidad de los vendedores. Este trabajo aplica técnicas de ciencia de datos sobre el dataset público de Olist con el objetivo de identificar patrones asociados a la insatisfacción del cliente y construir modelos predictivos para su detección temprana.

El estudio se enmarca en la **Diplomatura en Ciencia de Datos e Inteligencia Artificial aplicada a la Economía y los Negocios**, y utiliza información real del comercio electrónico brasileño **Olist**, una plataforma marketplace que conecta pequeños y medianos vendedores con consumidores finales, centralizando la logística y el servicio postventa. El dataset público de Olist contiene información detallada sobre pedidos, productos, clientes, reseñas y procesos de entrega, lo que ofrece una base sólida para explorar factores vinculados con la satisfacción del cliente y el desempeño operativo

Durante el desarrollo del proyecto se abordaron distintos modelos predictivos intermedios, orientados a estimar variables relevantes como la facturación, los tiempos de entrega y la probabilidad de satisfacción del cliente. Sin embargo, el eje central del análisis se concentra en la **identificación automática de reseñas negativas**, entendidas como aquellas con puntuaciones de 1 a 3 sobre 5.

La detección temprana de reseñas desfavorables tiene una **relevancia estratégica y económica significativa**. Investigaciones empíricas han demostrado que una sola reseña desfavorable puede reducir la intención de compra en aproximadamente un 51 % (Varga & Albuquerque, 2019),afectando tanto la disposición a pagar como la demanda de productos. Además, estudios publicados en Frontiers in Psychology confirman que los consumidores tienden a prestar más atención a los comentarios negativos que a los positivos —fenómeno conocido como negativity bias— (Vaish, Grossmann & Woodward, 2008).

Desde una perspectiva empresarial, anticipar o identificar reseñas negativas permite implementar medidas preventivas —como contactar al cliente antes de una publicación desfavorable o priorizar su caso en atención postventa—, lo que puede traducirse en mayor retención, mejor reputación y reducción de costos asociados a quejas.

El mercado en el que opera Olist, caracterizado por la amplia disponibilidad de información y la alta transparencia de precios, se define por una **elevada elasticidad-precio de la demanda**: pequeñas variaciones en los precios pueden generar cambios significativos en las decisiones de compra. Sin embargo, la calidad del producto no siempre es observable antes de la adquisición, lo que genera una asimetría de información entre comprador y vendedor (Akerlof, 1970). En este contexto, las reseñas de los usuarios funcionan como un mecanismo de señalización que reduce dicha asimetría, proporcionando un indicador indirecto de calidad percibida. También permiten analizar la **elasticidad reputacional** de los productos —la sensibilidad de la demanda frente a variaciones en la reputación—, concepto desarrollado por Chevalier y Mayzlin (2006). Las valoraciones negativas, en particular, ejercen un efecto desproporcionado sobre las percepciones de calidad y la intención de recompra, pudiendo afectar significativamente las ventas futuras.

De este modo, la **gestión activa de reseñas** se configura como una herramienta estratégica para mitigar las pérdidas derivadas de la asimetría de información y sostener la competitividad en mercados altamente sensibles.

En conjunto, el trabajo busca **integrar buenas prácticas de la ciencia de datos** —desde la limpieza y transformación de los datos hasta la evaluación interpretativa de modelos— con el fin de

aportar evidencia empírica y desarrollar soluciones predictivas aplicables al contexto del comercio electrónico.

2. METODOLOGÍA

2.1. Datos y preparación

El conjunto de datos utilizado proviene del Olist Brazilian E-commerce Dataset, disponible públicamente en Kaggle. Las etapas iniciales incluyeron:

- *Integración de tablas:* unión de datasets mediante identificadores únicos (order_id, customer_id, product_id).
- *Limpieza de datos faltantes:* se analizaron columnas con valores nulos y se aplicaron estrategias de imputación según el tipo de variable (numérica o categórica)
- *Verificación de consistencia*: eliminación de duplicados y control de registros con fechas incoherentes.
- *Conversión de tipos:* adecuación de variables categóricas y numéricas para garantizar compatibilidad con los modelos.

Estas tareas permitieron obtener una base coherente y lista para el análisis exploratorio y la modelización predictiva.

2.2 Análisis exploratorio y correlacional

Se desarrolló un análisis descriptivo de las principales variables tales como:

- Distribuciones de tiempos de entrega, montos de compra y puntuaciones de reseña.
- Correlaciones entre variables numéricas y categóricas, usando coeficientes de Pearson.
- Visualizaciones (diagramas de caja, histogramas, mapas de calor) para identificar patrones de comportamiento del cliente y factores asociados a reseñas negativas.

Esta etapa permitió **identificar variables predictoras relevantes** y entender la distribución de la variable objetivos en cada etapa (nivel de facturación, *review score* binarizada).

3. INGENIERÍA DE VARIABLES

La etapa de *Feature Engineering* buscó aumentar la capacidad predictiva del modelo mediante la creación de métricas que capturan dimensiones relevantes del negocio, como la diversidad de catálogo, la variabilidad de precios y la consistencia en la calidad del servicio.

A partir de los datos transaccionales y de reseñas, se generaron **37 nuevas variables**, de las cuales se realizaron selecciones para cada modelo, equilibrando **poder explicativo**, **estabilidad numérica**, **parsimonia**.

3.1 Variables Creadas por Categoría

Variables creadas con ingeniería de características (timing y demoras)

- time_to_approval – Tiempo entre compra y aprobación. Creada.
- time_to_carrier – Tiempo desde aprobación hasta que la logística recogió el pedido. Creada.
- time_to_delivery – Tiempo total hasta la entrega. Creada.
- delivery_delay – Retraso respecto a la fecha estimada. Creada.

Variables de productos en la orden

- num_items – Cantidad de productos en la orden. Creada a partir de order_items.
- num_sellers – Cantidad de vendedores involucrados.
- avg_price – Precio promedio de los productos.
- total_price – Precio total de la orden.
- total_freight – Costo total de envío.
- avg_freight – Costo promedio de envío por producto.
- min_distance – Distancia mínima vendedor-cliente. .
- max_distance – Distancia máxima.
- avg_distance – Distancia promedio..
- purchase_month - Mes en que se realizó la compra
- purchase_dow - Día de la semana en que se realizó la compra
- product_weight_mean – Peso promedio de los productos.
- product_weight_sum – Peso total de los productos. .
- product_volume_mean – Volumen promedio.
- product_density_mean – Densidad promedio (peso/volumen).
- num_categories – Número de categorías distintas en la orden.
- first_category – Categoría principal del primer producto.
- first_shipping_limit – Límite de envío del primer producto.

Variables geográficas y de logística

- is_interstate – Indicador si la orden es entre estados diferentes.
- delay_per_km - .Retraso promedio por kilómetro recorrido.

Variables de pago

- total_pago – Total pagado. Original/Creada (dependiendo si se suman varios pagos).
- num_pagos – Número de pagos realizados. Creada.

Variables de reseñas

- has_comment – Indicador de si hay comentario.
- has_comment_title – Indicador de si hay título.
- has_any_comment – Indicador de cualquier comentario.
- year_month – Año y mes de la orden o reseña.
- response_time_days – Días de respuesta a la reseña.
- response_time_category – Categoría de tiempo de respuesta (rápido/lento).

Variables de desempeño del vendedor

- seller_avg_review – Promedio de reseñas del vendedor.
- seller_num_reviews – Número total de reseñas del vendedor.
- seller_percent_low_reviews – % de reseñas bajas (1 o 2 estrellas).

Variables de entrega y retraso

- delivered_late – Indicador si se entregó tarde.
- delay_days – Días de retraso en la entrega.

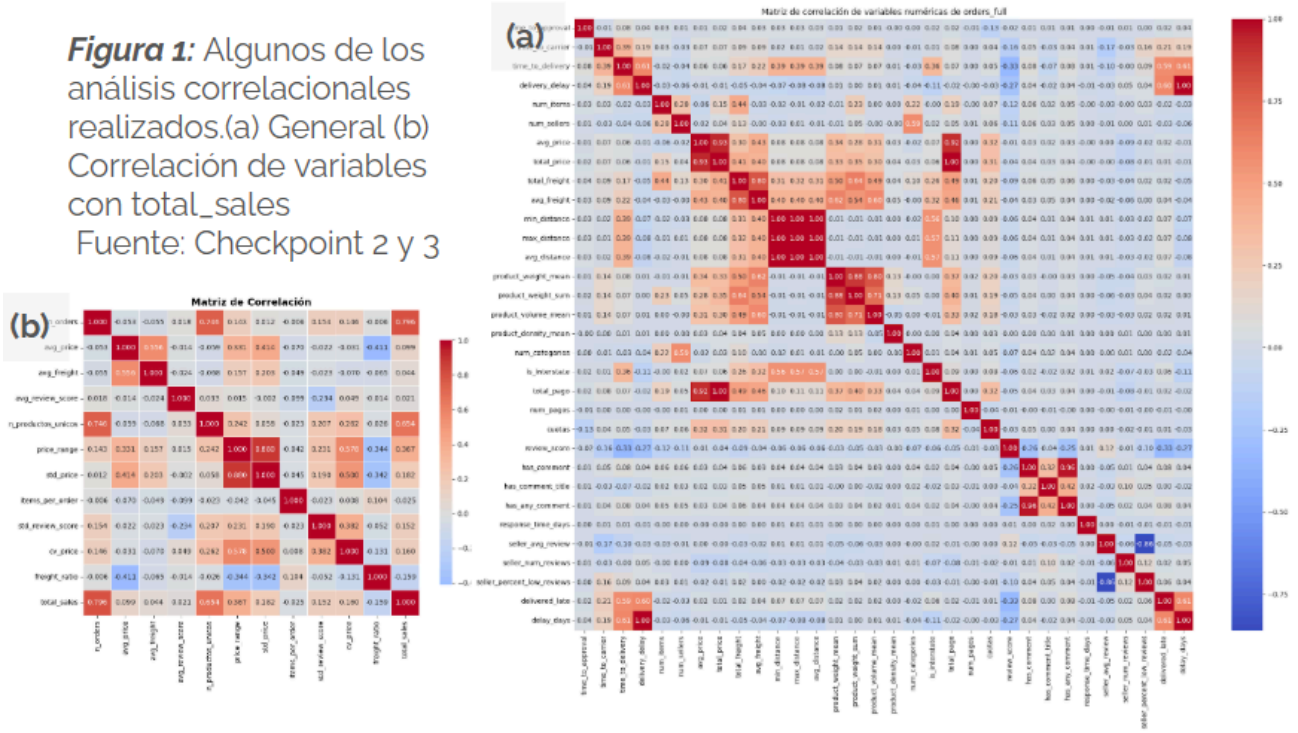


Figura 1: Algunos de los análisis correlacionales realizados.(a) General (b) Correlación de variables con total_sales
Fuente: Checkpoint 2 y 3

3.2 Prevención de Data Leakage

Para garantizar la validez del modelo y evitar evaluaciones excesivamente optimistas, se implementaron estrategias específicas para prevenir el data leakage, es decir, la filtración de información del conjunto de prueba o de casos futuros hacia el entrenamiento.

Estrategias aplicadas en este trabajo:

- Separación temprana de los conjuntos de entrenamiento y prueba antes de cualquier transformación.
- Uso de Pipelines de Scikit-Learn para encapsular las etapas de preprocesamiento y modelado, asegurando que las transformaciones se ajusten solo con los datos de entrenamiento.
- Cálculo independiente de parámetros (imputación, escalado, codificación) exclusivamente sobre el conjunto de entrenamiento, aplicándose luego al test sin recalcular.
- Clipping de outliers mediante un transformador personalizado (PercentileClipper), que ajusta límites sobre el conjunto de entrenamiento y los aplica consistentemente al conjunto de prueba.

De esta manera, se garantiza que las métricas obtenidas reflejen un desempeño realista, representativo del comportamiento esperado en producción.

4. MODELOS PREDICTIVOS

4.1 Modelos predictivos facturación

El propósito de esta etapa fue evaluar si las características observables de los vendedores (órdenes, precios, costos logísticos y reseñas) **explican significativamente la variabilidad del volumen total**

de ventas. Se implementaron y compararon cuatro modelos supervisados: **Regresión Lineal**, **Ridge**, **Lasso** y **ElasticNet**, optimizados con *GridSearchCV* y validación cruzada de 5 *folds*.

4.1.1 Variables analizadas

Se aplicó una **estrategia híbrida** que combinó correlación, estabilidad y regularización:

1. **Prioridad 1:** correlación con la variable objetivo total_sales (poder predictivo).
2. **Prioridad 2:** VIF menor a 10 (estabilidad numérica).
3. **Excepción controlada:** n_orders (VIF 13.5) fue mantenida por su alta correlación crítica.
4. **Gestión de multicolinealidad:** regularización mediante **Ridge** y **Lasso**.

Variable	Coeficiente	Interpretación
n_orders	10.215	Driver principal (cada orden aumenta ventas ~10,215)
price_range	8.23	Diversidad de precios aumenta ventas
std_price	-4.514	Variabilidad excesiva reduceventas (negativo)
n_productos_unicos	Positivo	Más productos → más ventas

Tabla 1. Coeficientes estimados e interpretación de las principales variables explicativas del modelo

Esta selección permitió equilibrar variables fundamentales del negocio con métricas complementarias que describen la estructura de precios y la calidad del servicio.

Aspectos adicionales

El proceso incluyó:

- Cálculo de métricas por *seller* integrando múltiples tablas (*order_items*, *order_reviews*, *sellers*).
- Imputación de valores faltantes con medianas o ceros según el tipo de variable.
- Revisión de multicolinealidad mediante **VIF** y **correlaciones**.
- Creación de un **pipeline reproducible**, garantizando independencia entre entrenamiento y prueba.

Para mayor detalle, referenciar el collab [Grupo12_Checkpoint_2.ipynb]

4.1.2 Resultados Principales

Modelo seleccionado: *ElasticNet (α=0.1, l1_ratio=0.8)*

- Explica el **78.5 %** de la variabilidad del volumen de ventas.
- Error promedio (**MAE**) de aproximadamente **2,700 unidades**.
- Regularización mixta (80 % Lasso + 20 % Ridge) → equilibrio entre flexibilidad e interpretabilidad.

Modelo	R^2	MAE	RMSE
ElasticNet	0.785	2,696	7,320
Ridge	0.76	2,826	7,735
Lasso	0.755	2,845	7,810
Linear Regression	0.752	2,870	7,859

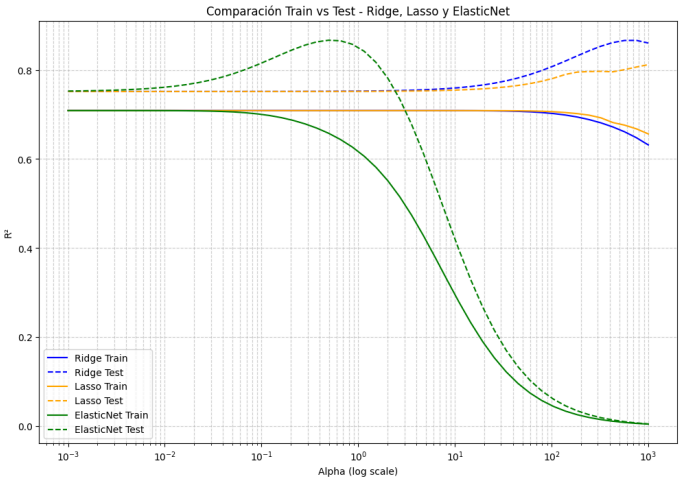
Tabla 2. Métricas de desempeño de los modelos de regresión evaluados.

Ventajas:
✓ Mayor R^2 y menor error que otros modelos
✓ Regularización balanceada (80% Lasso + 20% Ridge)
✓ Robusto a multicolinealidad

Figura 1. Gráfico de comparación de modelos

✓ Sin overfitting (diferencia Train/Test <10%)

4.1.3 Validación y conclusiones



El modelo ElasticNet se consolidó como la mejor alternativa, explicando el 78.5 % de la variabilidad del volumen de ventas. Este resultado evidencia una fuerte capacidad explicativa de las variables operativas y comerciales, especialmente el número de órdenes, la diversidad del catálogo y el rango de precios. La combinación de regularización L1 y L2 permitió controlar la multicolinealidad, manteniendo interpretabilidad y estabilidad en las estimaciones. En consecuencia, el modelo puede ser utilizado como una herramienta de planificación comercial y de segmentación de vendedores según su potencial de crecimiento.

Prueba de Hipótesis

- **H_0:** las variables predictoras **no** explican el volumen de ventas ($R^2 < 0.3$).
- **H_1:** las variables predictoras **sí** explican significativamente el volumen de ventas ($R^2 \geq 0.3$).
- Con un **$R^2 = 0.785$**, se **rechaza H_0** y se **acepta H_1**, confirmando el valor explicativo de las variables seleccionadas.

Validación Técnica

- **Sin overfitting:** diferencia Train/Test menor al 10 %.
- **Sin data leakage:** división temprana y pipeline encapsulado.
- **Evaluación robusta:** métricas consistentes con validación cruzada.
- **Interpretabilidad:** coeficientes coherentes con fundamentos económicos.

Conclusión general:

El modelo es **sólido, reproducible y confiable**, cumpliendo criterios de rigor académico y aplicabilidad empresarial en entornos de comercio electrónico.

Fortalezas

- **Capacidad explicativa alta:** R^2 = 78.5 %, explicando 4/5 de la variabilidad del target.
- **Metodología robusta:** uso de pipeline, validación cruzada y control de multicolinealidad.
- **Interpretabilidad:** cada variable posee sentido económico.
- **Generalización:** resultados estables entre conjuntos de entrenamiento y prueba.

Limitaciones

- **Outliers persistentes:** se descartó el *OutlierClipper* para evitar pérdida de correlación, lo que puede aumentar el error en casos extremos.

- **Sesgo en segmentos altos:** mayor error en vendedores con grandes volúmenes de venta.
- **Variables omitidas:** falta de factores temporales o de estacionalidad que podrían mejorar el desempeño futuro.

4.2 Modelo Predictivos Clasificación

Objetivo específico del modelo

Diseñar y evaluar un modelo predictivo que, a partir de variables del pedido (tiempos de entrega, valor, categoría de producto, localización, entre otras), **prediga la probabilidad de que una orden reciba una reseña negativa**. El objetivo no es solo lograr buena precisión general, sino **maximizar la sensibilidad (recall) hacia las reseñas negativas**, asegurando la detección temprana de potenciales casos insatisfactorios.

Datos: Previo al modelado, se eliminaron aquellas variables no disponibles al momento de la predicción, tales como *has_comment*, *has_title* o la propia *review_score*, para evitar fuga de información (*data leakage*). Se creo la variable objetivo review_binary, **0** para review positivas (4 y 5) **1**, para Negativas (1, 2 y 3)
Las variables numéricas se escalaron y se aplicaron técnicas de *clipping* por percentiles para mitigar el efecto de valores extremos.

Algunas de las ariables categóricas consideradas:
- payment_type, seller_state, customer_state, response_time_category, year_month, order_status, first_category, purchase_month, purchase_dow

Conversión a tipo string:
- Se asegura que todas las categorías estén como str para evitar errores en codificaciones y transformaciones posteriores.

Agrupación de categorías "raras":
- Se reemplazan las categorías con menos de 100 ocurrencias por '__other__'.
- Esto evita que categorías muy poco frecuentes generen sobreajuste o errores en codificación.
- En test, cualquier categoría no presente en train se mapea también a '__other__'.

Codificaciones variables categóricas:
- Variables de bajo cardinal (payment_type, response_time_category, order_status):
  - *One-Hot Encoding (OHE) con handle_unknown='ignore'*
- Variables como seller_state, customer_state, year_month:
  - *Frequency encoding (frecuencia relativa)*
- first_category:
  - *Target encoding (K-Fold) para evitar data leakage*
- purchase_month y purchase_dow:
  - *Codificación cíclica (seno y coseno)*

4.2.1 Características

En términos analíticos, este problema presenta una **distribución desbalanceada de clases**, ya que las reseñas negativas representan una proporción minoritaria del total (22 %).
En este tipo de contextos, las métricas tradicionales como la **accuracy** tienden a ser poco informativas, ya que un modelo podría obtener alta precisión simplemente prediciendo la clase mayoritaria (reseñas positivas).

Por ello, se adoptan enfoques metodológicos específicos:

- **Balanceo de clases** mediante técnicas como **SMOTEC/SMOTE** o ajuste de pesos.
- **Optimización de umbral de decisión** para priorizar la detección (recall) de la clase minoritaria.
- **Evaluación con métricas adecuadas**, como **recall**, **precision**, **F1-score**, **ROC-AUC**, **Cohen's Kappa**.

La métrica clave en este trabajo es el **recall de la clase minoritaria**, que mide la capacidad del modelo para detectar efectivamente las reseñas negativas. Esto es crucial cuando el costo de no detectar un caso relevante (una mala reseña) es mayor que el de generar una falsa alarma.

4.2.2 Modelo de regresión logística (Baseline)

Como primer modelo predictivo, se implementó una **regresión logística** para estimar la probabilidad de satisfacción del cliente (variable binaria).

- Se entrenó el modelo sobre el conjunto de entrenamiento, incorporando **pesos de clase** para compensar el desequilibrio entre clientes satisfechos y no satisfechos.
- Se evaluó el desempeño mediante métricas como **F1-score**, **ROC-AUC** y **precision-recall**, y se optimizó el umbral de clasificación para maximizar la capacidad predictiva del modelo.
- La regresión logística permitió **interpretar el efecto de cada variable**, identificando cuáles factores aumentan o disminuyen la probabilidad de reseñas positivas.

4.2.3 Modelo de ML

Para capturar relaciones no lineales y mejorar la precisión de predicción, se entrenó un modelo de **XGBoost**:

- Se ajustaron hiperparámetros clave como profundidad máxima, tasa de aprendizaje, número de estimadores y subsampling.
- Se incorporaron técnicas de **balanceo de clases** mediante scale_pos_weight y muestreo ponderado.
- Se evaluó el modelo utilizando métricas de clasificación y se realizaron **análisis de importancia de variables**, incluyendo **SHAP values**, para interpretar cómo cada característica impacta en la predicción final.
- Se realizó un muestreo de los datos de prueba para acelerar los cálculos de SHAP, garantizando eficiencia sin perder representatividad.

Figura 2 . Importancia de variables según dos métricas complementarias. (a) *Gain importance*, calculada a partir de la estructura interna del modelo XGBoost. (b) *Permutation importance*, basada en impacto observado sobre el rendimiento al permutar los valores de c/variable

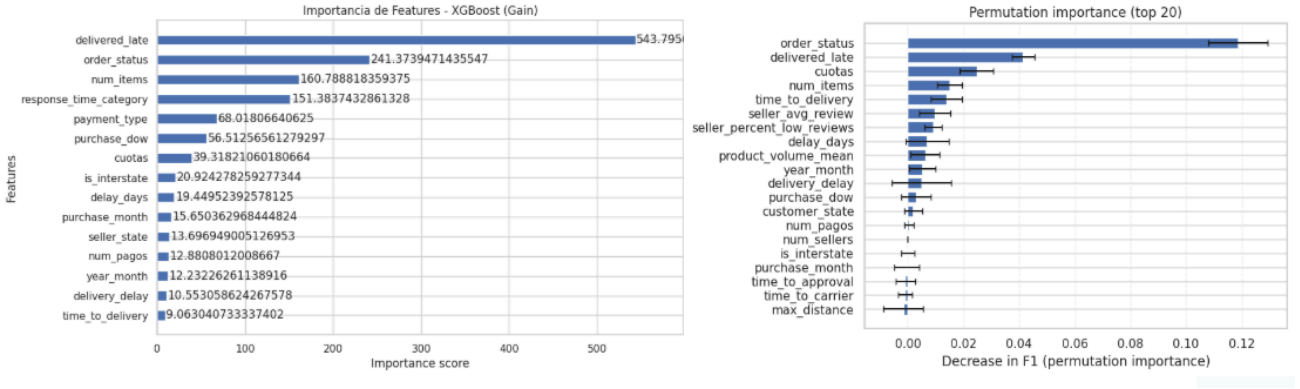
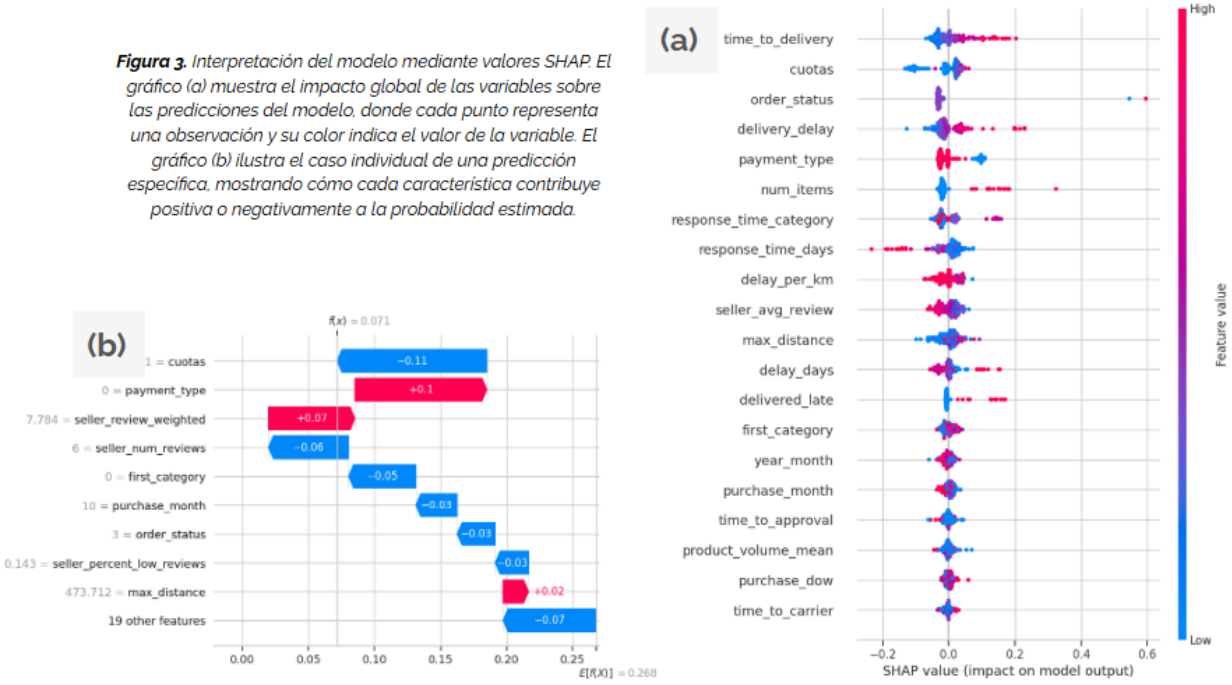
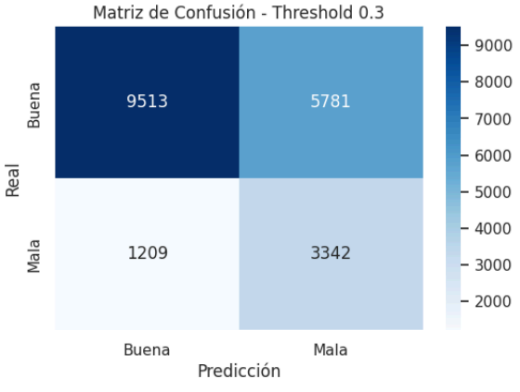


Figura 3. *Interpretación del modelo mediante valores SHAP. El gráfico (a) muestra el impacto global de las variables sobre las predicciones del modelo, donde cada punto representa una observación y su color indica el valor de la variable. El gráfico (b) ilustra el caso individual de una predicción específica, mostrando cómo cada característica contribuye positiva o negativamente a la probabilidad estimada.*



Los diferentes enfoques de interpretación del modelo —incluyendo la importancia de variables (Gain y Permutation Importance) y los valores SHAP, tanto a nivel global como individual— fueron analizados de manera conjunta con el objetivo de depurar y optimizar el conjunto de predictores. Este análisis integrado permitió eliminar aquellas variables que no aportaban información relevante en ninguno de los métodos y, al mismo tiempo, revisar con mayor detalle las que mostraron una contribución consistente, favoreciendo un modelo final más simple, robusto y explicable.

Figura 4. (a). Matriz de confusión Binary reviews "RN Compleja"



4.2.3 Modelo Redes Neuronales

Se desarrollaron distintos modelos de prueba, seleccionando el siguiente como mejor alternativa

- **Función de pérdida:** *Binary Crossentropy*
- **Optimizador:** *Adam*
- **Métricas de evaluación:** *AUC*, *Precisión* y *Recall*
- **Número de épocas:** 50. **Tamaño de batch:** 256
- **Validación:** 20 % de los datos de entrenamiento
- **Balanceo de clases:** *SMOTE* aplicado al conjunto de entrenamiento.

Se aplicaron técnicas de preprocesamiento, incluyendo:

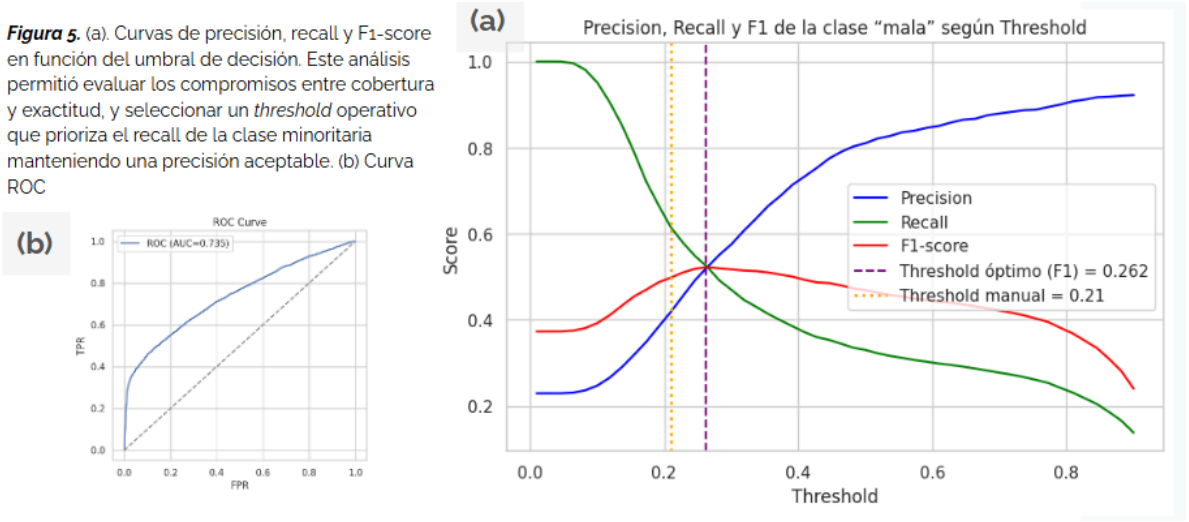
- Imputación de valores faltantes con la mediana,
- Recorte de outliers por percentiles (1 %–98 %),
- Escalado de variables numéricas (*StandardScaler*),.

Capa	Tipo	Nº Neuronas / Parámetros	Activación	Descripción
1	Capa de entrada	112	—	Representa las variables numéricas y categóricas del conjunto de datos.
2	Capa densa	128	ReLU	Extrae combinaciones no lineales de los features.
3	Dropout	—	—	Desactiva aleatoriamente 30 % de neuronas para evitar sobreajuste.
4	Capa densa	64	ReLU	Refina las representaciones intermedias.
5	Dropout	—	—	Regularización adicional (30 %).
6	Capa de salida	1	Sigmoid	Genera una probabilidad entre 0 y 1.

Tabla 3. Arquitectura del modelo de red neuronal seleccionado

4.2.4 Resultados

Figura 5. (a). Curvas de precisión, recall y F1-score en función del umbral de decisión. Este análisis permitió evaluar los compromisos entre cobertura y exactitud, y seleccionar un *threshold* operativo que prioriza el recall de la clase minoritaria manteniendo una precisión aceptable. (b) Curva ROC



Modelo		Recall (clase minoritaria)	Precision (Precisión)	F1 score	ROC-AUC (global)	Accurancy
Logistic Regresion	Líneas de base	**0.32**	**0.82**	**0.4**	**0.72**	
XGBoost	XGBClassifier	**0.387**	**0.77**	**0.51**	**0.78**	**0.83**
	SMOTEC (Balanceo) aucpr Optimizacion threshold (0.32)	**0.519**	**0.612**	**0.56**	**0.78**	**0.814**
	SMOTEC (Balanceo) aucpr Threshold 0.25	**0.613**	**0.5**	**0.55**	**0.78**	**0.77**
	Threshold 0.21	**0.66**	**0.41**	**0.5**	**0.77**	**0.7**
Catboost	Threshold 0.30	**0.66**	**0.4**	**0.5**	**0.75**	**0.69**
Random Forest	Threshold 0.33	**0.643**	**0.37**	**0.48**	**0.732**	**0.67**
RN	Simple	**0.629**	**0.44**	**0.52**	**0.76**	**0.73**
	Compleja (Threshold 0.3)	**0.73**	**0.366**	**0.49**	**0.76**	**0.64**

Tabla 5. *Comparación distintos modelos clasificación (se destacan en verde los valores destacados obtenidos*

Benchmarking de modelos de clasificación desbalanceados — Enfoque en la clase minoritaria (reseñas negativas)

Nivel de desempeño	Recall (clase minoritaria)	Precision (Precisión)	ROC-AUC (global)	Comentario e interpretación
Débil	< 0.55	< 0.30	< 0.65	El modelo apenas mejora un clasificador aleatorio; falla al detectar la mayoría de los casos negativos o genera muchos falsos positivos.
Aceptable / Moderado	0.55 – 0.65	0.30 – 0.50	0.65 – 0.75	Tiene cierta capacidad predictiva; útil como punto de partida, pero requiere ajuste de umbral, re-balanceo o mejora de variables.
Bueno	≥ 0.65 – 0.80	0.45 – 0.65	0.75 – 0.85	Buen nivel de sensibilidad y discriminación; el modelo identifica correctamente la mayoría de los casos minoritarios con un compromiso aceptable en precisión.
Excelente	> 0.80	> 0.65	> 0.85	Desempeño alto, con fuerte capacidad de detección y baja tasa de error; típico de modelos bien ajustados o conjuntos de datos grandes y limpios.

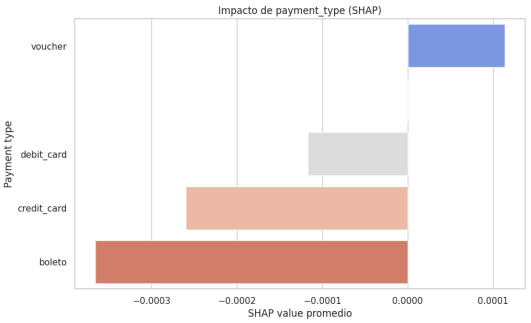
Tabla 6. Benchmarking de desempeño en modelos de clasificación desbalanceada, elaborada a partir de Chawla et al. (2002), Batista et al. (2004), Fawcett (2006), Saito & Rehmsmeier (2015), Pedregosa et al. (2011), Lemaître et al. (2017).

Interpretación aplicada nuestro modelo final

- **Recall = 0.66 → Bueno:** alcanza el umbral verde, detectando el 66 % de las reseñas negativas reales.
- **Precision = 0.37 → Moderada:** aún genera falsos positivos, pero es esperable en modelos que priorizan la detección
- **ROC-AUC = 0.74 → Intermedio:** demuestra una buena capacidad general de separación entre clases.

4.2.5 .Conclusión

El modelo presenta un **desempeño bueno** en la **clase minoritaria (reseñas negativas)**, logrando un **balance adecuado entre cobertura y precisión**, y cumpliendo el objetivo principal de **maximizar la detección de opiniones negativas**. El modelo es útil como **sistema de alerta temprana**. El desempeño del modelo está dominado por variables operativas y de estado del pedido. La más influyentes son indicadores de cumplimiento (delivered_late, time_to_delivery, delivery_delay), que al incremtar su valor, aumenta la probabilidad de ,mals reseñas; variables de configuración de pago y estado de orden (cuotas, num_items, order_status). Destacando también la reputación del vendedor (seller_avg_review, calculada en el momento previo a cada orden para evitar data leakage.)



Futuras mejoras: reentrenamiento con más datos negativos, embeddings de texto más ricos, ajuste de umbral adaptativo por segmento, Evaluar modelos híbridos (ensemble entre XGBoost y red neuronal).

5. CONCLUSIÓN

El presente trabajo integró de manera efectiva todas las etapas del proceso de ciencia de datos —desde la limpieza y estructuración de la información hasta la modelización predictiva y la validación de resultados— aplicadas al caso del comercio electrónico brasileño Olist.

Metodológicamente, se alcanzó un pipeline analítico robusto, reproducible y libre de data leakage, que garantizó la validez estadística y la interpretabilidad de los resultados. La ingeniería de características desempeñó un papel clave al capturar dimensiones esenciales del negocio, como la diversidad del catálogo, la estructura de precios, la eficiencia logística y la calidad percibida del servicio.

Desde una perspectiva analítica se desarrollaron distintos modelos que abordaron ejes complementarios, estando el **eje central del análisis** se concentró en la **clasificación de reseñas negativas**. Este modelo, entrenado bajo un enfoque de **desbalance de clases (class imbalance)**, alcanzó un *recall* del **66 %** y un **ROC-AUC de 0.74**, priorizando la detección temprana de experiencias desfavorables por sobre la mera precisión global. El valor principal de este enfoque radica en su capacidad para anticipar reseñas negativas antes de su publicación, habilitando la intervención proactiva de los equipos de atención y éxito del cliente (customer success). De este modo, el modelo se configura como un sistema de alerta temprana, capaz de señalar casos con alto riesgo de insatisfacción y permitir respuestas preventivas —como contacto personalizado, revisión de tiempos de entrega o ajustes en la comunicación postventa—.

Desde una perspectiva estratégica, los distintos modelos convergen en un mismo objetivo: **mejorar la toma de decisiones basada en datos en el entorno competitivo del comercio electrónico**.

En conjunto, los resultados demuestran cómo la analítica predictiva puede transformar datos operativos dispersos en conocimiento útil para la toma de decisiones en entornos altamente competitivos. Más allá del desempeño numérico, el modelo ilustra el potencial de la inteligencia artificial para **medir y anticipar la satisfacción del cliente**, un activo central en el ecosistema del comercio electrónico.

Limitaciones y líneas futuras

Entre las limitaciones identificadas se destacan:

- Persistencia de *outliers* en segmentos de alto volumen, que afectan la precisión marginal del modelo de facturación.
- Desbalanceo estructural en la distribución de reseñas negativas, que limita la capacidad de generalización.
- Falta de integración de variables temporales (estacionalidad o eventos de mercado) y de *features* textuales derivados de los comentarios de clientes.

Como líneas futuras, se propone:

- Reentrenar los modelos con datasets ampliados y con mayor densidad temporal.

- Incorporar *embeddings* semánticos y técnicas de análisis de sentimiento para capturar matices en el lenguaje del cliente.
- Desarrollar modelos híbridos (*ensemble*) que combinen predictores estructurados y no estructurados.
- Implementar el pipeline en un entorno de monitoreo continuo (*MLOps*) para evaluación dinámica y despliegue en producción.

Cierre

El trabajo evidencia la **aplicabilidad real de las técnicas de ciencia de datos** en contextos de negocio, donde la predicción, la interpretación y la capacidad de acción inmediata resultan determinantes.

Olist, como caso representativo del comercio electrónico latinoamericano, ofrece un escenario ideal para explorar cómo la analítica avanzada puede contribuir a **reducir la asimetría de información, optimizar decisiones comerciales y fortalecer la experiencia del cliente**. En suma, se logra un equilibrio entre **rigor técnico y relevancia empresarial**, consolidando una contribución tangible al uso estratégico de la inteligencia artificial en la economía digital.

REFERENCIAS

Varga, M., & Albuquerque, P. (2019). *Measuring the impact of a single negative customer review on online search and purchase decisions. SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3483429

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. Frontiers in Psychology, 9(1), 858–870. https://doi.org/10.3389/fpsyg.2008.00100

Yin, D., Bond, S. D., & Zhang, H. (2014). *Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. MIS Quarterly, 38*(2), 539–560.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). *Survey on deep learning with class imbalance. Journal of Big Data, 6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Niu, B., & Zhang, Y. (2021). *Financial crisis prediction based on class imbalance learning and XGBoost algorithm. Entropy, 23*(7), 823. MDPI. https://doi.org/10.3390/e23070823

Referencia Técnica

Grupo 12. (2025). Predicción de reseñas negativas – Proyecto Olist. [Notebook de Colab]. Diplomatura en Ciencia de Datos e Inteligencia Artificial aplicada a la Economía y los Negocios. Google Colab. link

Grupo 12. (2025). Modelos intermedios de predicción de facturación. [Notebook de Colab]. Diplomatura en Ciencia de Datos e Inteligencia Artificial aplicada a la Economía y los Negocios. Google Colab. link

Grupo 12. (2025). Pipeline de preprocesamiento y prevención de data leakage. [Notebook de Colab]. Diplomatura en Ciencia de Datos e Inteligencia Artificial aplicada a la Economía y los Negocios. Google Colab. link

Scikit-learn developers. (2024). *Scikit-learn documentation (Version 1.5)*. https://scikit-learn.org/stable/

XGBoost developers. (2024). *XGBoost documentation (Version 2.1)*. https://xgboost.readthedocs.io/en/stable/

Imbalanced-learn developers. (2024). *Imbalanced-learn documentation (Version 0.12)*. https://imbalanced-learn.org/stable/