FACULTAD DE CIENCIAS ECONÓMICAS

UNC Universidad Nacional de Córdoba

FAMAF Facultad de Matemática, Astronomía, Física y Computación

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

Título del trabajo: Estimación de días de entrega de una plataforma e-commerce

Grupo: 8

Integrantes:

- Sebastián Ignacio Lezama
- Alejandro Mezio
- Diego Nicolás Zallio
- Federico Churquina
- Ignacio Cordoba

Tutor: Lic. Ignacio Fichetti

Año: 2025

Introducción

El presente informe tiene como base el análisis de un dataset extraído de la empresa ecommerce Olist.

Olist, fundada en 2015 en Curitiba (estado de Paraná), Brasil, comenzó como una plataforma de marketplace enfocada en facilitar que comerciantes menos grandes pudieran acceder a los grandes marketplaces brasileños.

En los años más recientes (2023-2024) ha ampliado sus servicios hacia logística y capital (financiamiento para comerciantes), es decir, ya no se limita sólo a ser un integrador de marketplaces, sino busca ofrecer un ecosistema completo para los comercios permitiéndoles vender en múltiples canales, gestionar logística, finanzas y operaciones de venta de forma integrada.

Acercamiento al Dataset

El dataset de Olist cuenta con 8 tablas originales, con ventana temporal de Septiembre de 2016 a Agosto de 2018, la cual complementamos con una tabla de datos de geolocalización utilizada a los fines de profundizar nuestro análisis. (*Ver Gráfico 1 - Diagrama entidad- relación en anexo I*)

En conjunto, este esquema relacional permite reconstruir el ciclo completo de una transacción: desde el cliente y el pedido, pasando por el producto y el vendedor, hasta el pago, la entrega y la valoración final.

Entre sus tablas podemos encontrar **olist_customers_dataset** que contiene los datos identificatorios y ubicación de los clientes, mientras que **olist_sellers_dataset** almacena la información de los vendedores, lo que permite analizar su distribución geográfica. La tabla **olist_orders_dataset** registra los pedidos realizados y sus etapas logísticas, y se relaciona con **olist_order_items_dataset**, donde se detalla el contenido de cada orden junto con los productos y vendedores involucrados.

La información de los productos se encuentra en **olist_products_dataset**, y sus categorías se complementan con **product_category_name_translation**. Los métodos y montos de pago se registran en **olist_order_payments_dataset**, mientras que la satisfacción del cliente se analiza a través de **olist_order_reviews_dataset**. Finalmente, la tabla **olist_geolocation_dataset** aporta los prefijos postales y coordenadas necesarios para vincular espacialmente clientes y vendedores.

Oportunidad de mejora y objetivo del trabajo

Al analizar el modelo de negocio de Olist y revisar en detalle cada tabla del dataset, prestamos especial atención a las variables temporales relacionadas con el proceso de entrega, como la fecha de compra, despacho y recepción. Durante esta exploración detectamos una oportunidad de mejora: en muchos casos, los plazos de entrega estimados por la plataforma resultaban significativamente mayores que los tiempos de entrega reales.

A partir de esta observación definimos el objetivo principal del trabajo:

- **Desarrollar un modelo de machine learning capaz de predecir con mayor precisión el tiempo estimado de entrega de un pedido**, considerando el intervalo entre la fecha de compra y la recepción final por parte del cliente. (*Ver gráfico 2 en Anexo I*)

Análisis Exploratorio de Datos y limpieza

En la etapa de **EDA inicial y limpieza**, realizamos un análisis exploratorio independiente de cada tabla, verificando consistencia estructural, tipos de variables y presencia de valores faltantes. Eliminamos duplicados y depuramos registros incompletos. Analizamos las variables temporales y su correlación, y filtramos únicamente las órdenes con estatus **"delivered"**, por ser las que permiten calcular tiempos reales de entrega. Además, detectamos meses con bajo volumen de ventas durante el inicio del histórico, por lo que acotamos el período de estudio a **febrero 2017 - agosto 2018**, garantizando una distribución temporal estable. (*Ver Gráfico 3 en Anexo I*)

Es importante destacar el siguiente tratamiento de datos que nos permitió la creación de las features usadas en el modelo:

- Siguiendo los estándares de la IATA construimos la variable peso volumétrico o peso dimensional, que busca ponderar en iguales condiciones peso y volumen de un paquete. Ésta se calcula dividiendo el volumen del paquete por 6000, obteniendo así una cantidad en kg que puede compararse con el peso real del paquete. El volumen se calcula en base a las medidas proporcionadas en la tabla de productos (largo x alto x ancho). Finalmente, tomamos el mayor valor entre el peso real y el peso volumétrico en la variable *"product_chosen_weight".*
- Creamos una nueva variable numérica continua *"distance_km"* que mide la distancia en km entre vendedor y comprador para cada órden de la base de datos. Para su construcción nos basamos en el prefijo de código postal de vendedores y compradores, y lo relacionamos con coordenadas de latitud y longitud que se encuentran disponibles en la tabla de geolocalización.
- Definimos que la fecha inicial para estimar los días de entrega sea el momento de compra *"order_purchase_date"*.

  Previamente evaluamos como posible fecha el momento en que se aprueba el pago de la órden, pero encontramos casos en que esta fecha no guarda relación con el despacho del paquete por parte del vendedor.

  Finalmente, considerando la lógica de funcionamiento de los modelos de e-commerce actuales, resulta natural que la fecha inicial sea la fecha de compra.

- Armamos la variable *"days_to_delivered"* que mide los días transcurridos entre la compra y la entrega, que es nuestra variable objetivo. También armamos la variable análoga *"days_estimated"* con la fecha de entrega estimada, a los fines de comparar con el modelo de Olist.

- Creamos la variable categórica *"rutas"*, que clasifica cada orden según la combinación de estado de origen (vendedor) y destino (comprador), resumiendo así el flujo logístico. Luego la ordenamos por el volumen de órdenes y utilizamos las 10 primeras para entrenar el modelo.
- Definimos una variable categórica *"product_venta"* que mide el volumen vendido por categoría de producto y lo clasifica en bajo, medio o alto.
- Finalmente, construimos la variable *"sales_same_state"*, que indica de manera binaria si la venta se realizó entre comprador y vendedor dentro del mismo estado, lo cual permite evaluar diferencias entre envíos locales y de larga distancia

Selección y definición de variables del modelo

Objetivo (target).
La variable objetivo del modelo es *"days_to_delivered"*: número de días transcurridos entre la fecha de compra *(order_purchase_date)* y la fecha de entrega al cliente *(order_delivered_customer_date)*. Esta variable cuantifica el tiempo real de entrega y es la que queremos predecir.

Features (variables explicativas).
Seleccionamos variables con soporte operativo (disponibles antes del despacho) y con relevancia logística:

- **Numéricas:**
  - *distance_km*: distancia Haversine entre vendedor y comprador (km).
  - *price*: precio del producto.
  - *freight_value*: costo de envío.
  - *product_chosen_weight*: peso escogido (máximo entre peso real y peso volumétrico).
  - *sales_same_state* (binaria): pedido en mismo estado del vendedor.

  Aplicamos normalización con StandardScaler para estas variables numéricas.

- **Categóricas:**
  - *product_venta*: categoría según volumen de órdenes por producto.
  - *rutas*: combinación de estado de vendedor y comprador (top 10 rutas + "Otra ruta").

  Estas variables fueron codificadas luego con OneHotEncoder.

División en datos de entrenamiento y test:

Realizamos una división aleatoria train/test (80% y 20%) con random_state fijo para reproducibilidad.
Para evaluación estable utilizamos validación cruzada (subrutina *GridSearchCV*) para la optimización de hiperparámetros.

Métricas y criterio de selección

En relación a las métricas utilizadas para evaluar los resultados y performance de los modelos barajamos inicialmente tres:

- **MSE**: El error cuadrático medio (o su raíz cuadrada) penaliza más los errores más grandes, siendo más sensible a outliers. Su interpretación no es tan directa.

- **MAE**: El error absoluto medio es más robusto y menos sensible a valores outliers. Su interpretación es más directa, ya que es el promedio de los errores.

Elegimos MAE como métrica para el entrenamiento de los modelos, ya que tenemos más interpretabilidad, aún cuando no amplificamos los errores más grandes.

RMSE quedó como métrica complementaria para comprender la varianza y los errores extremos.

Modelos utilizados (cronológicamente)

Para la predicción de días de entrega reales se probaron diversos modelos supervisados. El punto de partida fue una **Regresión Lineal sin regularización**, utilizada como modelo base de comparación. Este modelo asume una relación lineal entre variables explicativas y el objetivo.

Luego, evaluamos variantes con regularización:

- **Lasso Regression**: aplica penalización L1, favoreciendo la selección automática de variables al reducir a cero los coeficientes menos relevantes.

- **Ridge Regression**: aplica penalización L2, reduciendo el impacto de la multicolinealidad y estabilizando los coeficientes.

- **Regresión Polinómica con Ridge**: extiende la capacidad del modelo para capturar relaciones no lineales a través de términos polinómicos, manteniendo la regularización para evitar sobreajuste.

Posteriormente, incorporamos modelos basados en árboles:

- **Árbol de Decisión**: divide los datos en regiones a través de reglas jerárquicas, permitiendo capturar relaciones no lineales. Es fácil de interpretar, pero puede sobreajustar si no se controla la profundidad.

- **Random Forest**: combina múltiples árboles de decisión construidos sobre distintos subconjuntos de datos y variables, promediando los resultados. Este enfoque reduce la varianza y mejora significativamente la capacidad predictiva. Este modelo demoraba mucho tiempo en entrenamiento, por eso usamos la subrutina *RandomizedSearchCV*.

Selección del mejor modelo y su rendimiento

Tras comparar las métricas de evaluación (MAE y RMSE), el **Random Forest** mostró el mejor equilibrio entre precisión y generalización.

En la siguiente tabla plasmamos los resultados de cada modelo según las métricas antes mencionadas:

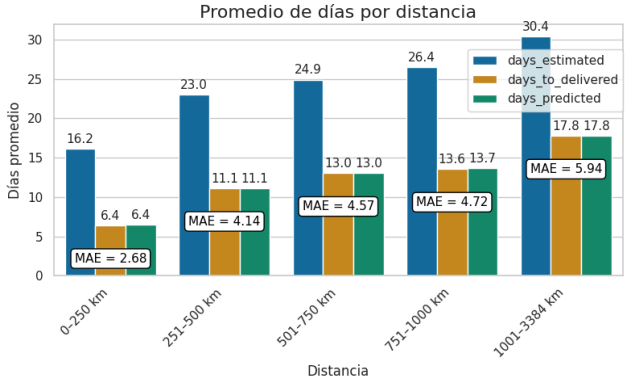
Metrica/Modelo	linreg_base	lasso	ridge	poly	tree	forest
MAE	4,84	4,84	4,84	4,80	4,76	4,67
RMSE	7,18	7,18	7,18	7,15	7,12	7,03

Luego de seleccionar el random forest afirmamos que distancia, precio de flete y si el envío es dentro del mismo estado son las variables más importantes, como se ve en la siguiente tabla:

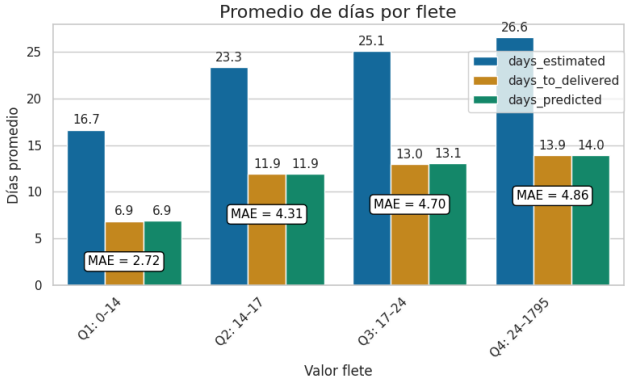
Feature	Importance
num__distance_km	32,8%
num__freight_value	16,6%
num__sales_same_state	13,6%
num__product_chosen_weight	9,9%
num__price	9,8%
cat__rutas_SP-SP	7,5%
cat__rutas_otra_ruta	2,3%
cat__rutas_SP-RJ	1,6%
cat__rutas_SP-BA	1,3%
cat__rutas_PR-SP	0,8%
cat__rutas_MG-SP	0,8%

La categorización por volumen de ventas, `product_venta`, resultó irrelevante a la hora de predecir con nuestro modelo. Puede no incluirse en futuros reentrenamientos.

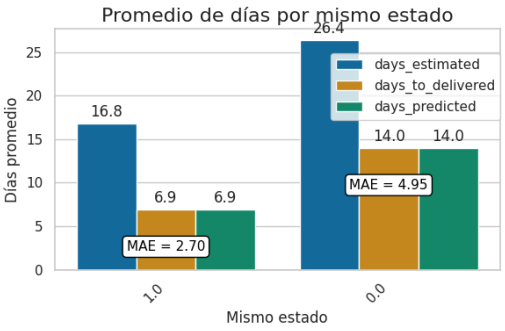
Luego procedimos a analizar la performance de cada variable del modelo vs los tiempos reales de entrega y los tiempos estimados originalmente por Olist. Podemos destacar lo siguiente:



- La variable distancia es muy explicativa de la variación del target, ya que a mayor distancia aumenta el valor estimado. Además podemos destacar que ante distancias cortas (menores a 250 km) la performance aumenta sustancialmente.



- Algo similar se observa con el precio del flete, donde los envíos que duran más tiempo suelen ser más costosos. También los envíos de menor costo son los mejores estimados, como puede verse con el bajo MAE.



- Destacamos que la variable que mide el mismo estado está correlacionada con la distancia y, podemos inferir, indirectamente con los precios de flete bajos. Ya que un tercio del total de las órdenes se producen dentro del estado de São Paulo, resulta lógico que las tres variables relevantes tengan una alta performance para las distancias cortas, fletes económicos y pedidos intra-estado.

Conclusión

El trabajo se orientó a mejorar la estimación del tiempo real de entrega de pedidos dentro de la plataforma, reemplazando o complementando la estimación original provista en el dataset. A partir del proceso de selección de variables, construcción de nuevas características logísticas y comparación de distintos modelos de machine learning, el modelo que presentó el mejor desempeño fue **Random Forest**, principalmente por su capacidad para capturar relaciones no lineales y efectos combinados entre variables.

El análisis de la importancia de variables y las curvas de dependencia parcial muestra que la **precisión en la predicción de los días de entrega** se ve explicada principalmente por:

- La **distancia en kilómetros** entre comprador y vendedor.
- El hecho de que la compra y la venta se encuentren **dentro de un mismo estado**.
- El **valor del flete**, que actúa como proxy del peso, volumen y distancia.
- La **ruta comprador-vendedor** (aunque con menor incidencia, ya que su efecto se encuentra parcialmente representado en la variable de distancia y en la relación dentro/diferente estado).

En conjunto, estas variables sintetizan componentes clave del proceso logístico: ubicación geográfica, complejidad de transporte y características del paquete.

Posibles pasos y mejoras a futuro

- **Incorporar información de reviews y desempeño del vendedor**, para evaluar si las calificaciones y tiempos de despacho influyen en los días de entrega y así crear una variable de “score del vendedor”. Tener en cuenta el caso de un vendedor nuevo de la plataforma.

- **Acotar el período de análisis**, priorizando años con mayor estabilidad operativa (por ejemplo, 2018) para mejorar la precisión del modelo y evitar variaciones asociadas a etapas tempranas del crecimiento de la plataforma.

ANEXO I

Gráfico 1.- Diagrama Entidad Relación

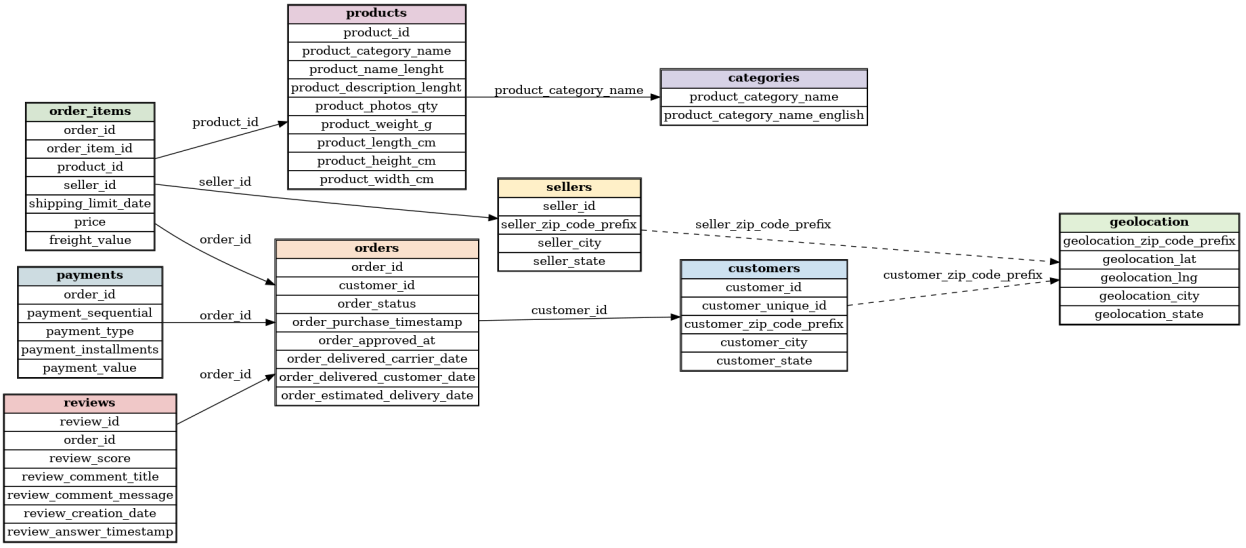


Gráfico 2.- Evolutivo mensual de Días reales y estimados promedios de entrega

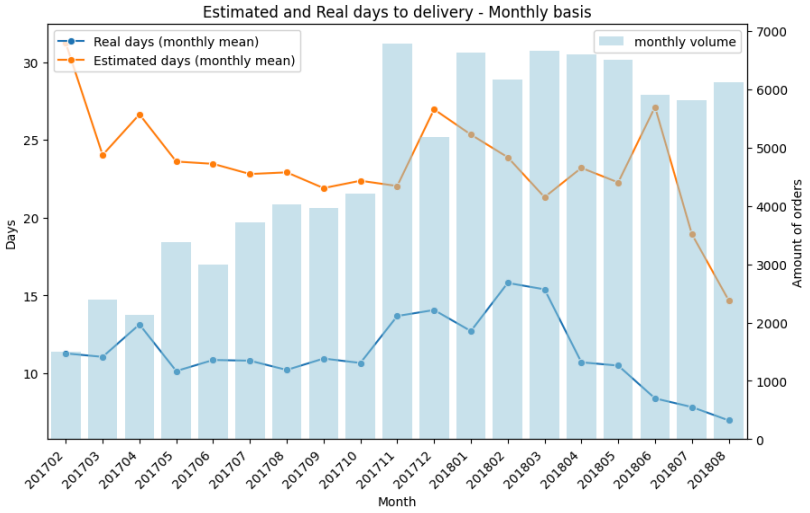


Gráfico 3.- Evolución del total de Órdenes por mes