UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD DE INGENIERÍA
Ingeniería de Sistemas

Otto Group Product Classification Challenge

Systems Analysis & Design

Group 020-82

Students

Juan Diego Lozada 20222020014
Juan Pablo Mosquera 20221020026
Alejandra
Jeison Cuenca: 20242020043

Professor: Carlos Andrés Sierra Virguez

Date: 2025

Contents

1 Introduction

The Otto Group is one of the world's biggest e-commerce companies, with subsidiaries in more than 20 countries, including Crate & Barrel (USA), Otto.de (Germany) and 3 Suisses (France). We are selling millions of products worldwide every day, with several thousand products being added to our product line.

A consistent analysis of the performance of our products is crucial. However, due to our diverse global infrastructure, many identical products get classified differently. Therefore, the quality of our product analysis depends heavily on the ability to accurately cluster similar products. The better the classification, the more insights we can generate about our product range.

2 Methodology and Deliverables

2.1 Summary About Competence

El reto consiste en construir un modelo de clasificación que pueda asignar correctamente productos a 9 categorías principales de productos, dadas sus características numéricas.

La métrica utilizada para evaluar las predicciones es el *log loss* multiclase (también llamado *multi-class logarithmic loss*), que penaliza fuertemente las predicciones incorrectas realizadas con alta confianza.

Datos disponibles:

Se proveen más de 200.000 productos (filas), cada uno con 93 características numéricas (*features*) obfuscadas (es decir, sin interpretación directa). En el conjunto de entrenamiento (*train*) aparece la columna `target`, que corresponde a la categoría de producto. En el conjunto de prueba (*test*) no aparece la columna `target`, y lo que se debe predecir son probabilidades para cada clase.

Además, cada instancia o producto tiene un identificador `id`, que no aporta a la clasificación y, por lo tanto, se descarta como variable predictiva.

Las clases no están perfectamente balanceadas: algunas categorías tienen más ejemplos que otras.

Restricciones:

El archivo de envío debía incluir:

- Columna `id` del conjunto de prueba (`test`).
- Nueve columnas de probabilidades (una por cada clase).
- La suma de las probabilidades en cada fila debía ser igual a 1.

Métrica de evaluación: multi-class logarithmic loss.
Número de envíos: máximo 5 por día.
Leaderboard:

- **Público:** solo una parte del conjunto `test`.
- **Privado:** usado para la clasificación final.

Datos:

- Se debía trabajar únicamente con los datos provistos (`train` y `test`).
- Estaba prohibido usar datos externos o de dominio específico.
- Las variables (93 *features*) estaban ofuscadas.

Clases:

- Existían 9 categorías.
- La distribución era desbalanceada entre ellas.

2.2 Informe del Análisis de Sistemas

Elementos:

- **Productos:** Cada fila (instancia) corresponde a un producto distinto ofrecido por Otto Group

  Son lo que hay que clasificar. Las decisiones de negocio dependen de tener una buena clasificación: análisis, inventarios, marketing, cliente, etc

- **Variables:** Características numéricas de los productos; están "obfuscadas" → no se conoce su significado real

  Son las señales que alimentan el modelo. Al no saber qué son exactamente, se depende de la correlación estadística, ingeniería de características, selección de variables, análisis exploratorio. También influyen en cómo elegir modelos (por ej., modelos robustos frente a variables independientes, valores esparsos, correlaciones, etc.)

- **Clase:** La categoría a la que realmente pertenece cada producto

  Es lo que se quiere predecir. Está ligado a la métrica de evaluación (log loss multiclase), al desbalance entre clases, al diseño del modelo, etc. En contexto de negocio, esas categorías serían las líneas principales de producto que Otto analiza

- **Datos de entrenamiento:** Entrenamiento tiene los productos con categoría conocida; test tiene solo características, se deben predecir probabilidades

  Permite entrenar modelos y luego evaluarlos en datos nuevos. La separación obliga a pensar en generalización. El leaderboard público/privado está basado en particiones de test

- **Métrica:** Penaliza errores con alta confianza; exige probabilidades calibradas para todas las clases, no solo clasificación correcta

  Impone que los modelos no solo "acierten" la clase, sino que sus predicciones probabilísticas sean razonables. Afecta decisiones: qué algoritmo, regularización, ensambles, calibración, etc. También conlleva que los errores sobre clases minoritarias puedan pesar mucho si no se modelan bien

- **Restricciones:** Limitaciones técnicas/procedimentales impuestas por la competencia

  Sirven para asegurar justicia, evitar "sobre-ajuste" del leaderboard, y fomentar modelos que generalicen. Influyen en estrategia: validación cruzada, enmascarar test privado, etc

- **Negocio (contexto de Otto Group):** Otto vende millones de productos en muchos países; tienen problemas de inconsistencia en la clasificación de productos idénticos, lo que afecta análisis y decisiones

  Justificación del reto: mejorar la calidad del análisis de productos para la empresa. A partir de una buena clasificación se pueden generar clustering, métricas de desempeño, análisis más fiables, mejor experiencia de cliente, etc

- **Correlaciones:** Procesos técnicos que los competidores aplicaron para mejorar desempeño

  Relacionan directamente con el hecho de que las variables no tienen significado explícito, muchas pueden estar correlacionadas, algunas irrelevantes. Estos pasos ayudan a estabilizar el modelo, evitar ruido, reducir riesgo de overfitting

- **Modelos y algoritmos:** Los métodos de machine learning que se usan: árboles, boosting, redes neuronales, ensambles, etc

  Dependen de los datos, de la métrica, del tamaño del dataset, y de las restricciones. Por ejemplo, modelos que proporcionen probabilidades para todas las clases son necesarios, que manejen features obfuscadas, que toleren desbalance de clases. También del hardware / tiempo disponible

2.3 Complejidad y Sensibilidad

Features Obfuscadas:

Puede suceder que al tener que procesar los productos en diferentes categorías, no se llegue a conocer el significado de las 93 vairables disponibles, lo que dificultaría hacer ingeniería de características basado en el conocimiento de dominio. Por lo mismo, obligaría a depender solo de patrones estadísticos y numéricos para la categorización de los productos

Alta Dimensionalidad:

3 Bibliografía

References

[1] Kaggle. (s. f.). *Otto Group Product Classification Challenge*. Disponible en: `https://www.kaggle.com/competitions/otto-group-product-classification-challenge/`. Accedido el 11 de septiembre de 2025.