UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD DE INGENIERÍA
Ingeniería de Sistemas

Otto Group Product Classification Challenge, A Systematic Analysis

---

Systems Analysis & Design

Group 020-82

Workshop #1

Students

Juan Diego Lozada 20222020014
Juan Pablo Mosquera 20221020026
María Alejandra Ortiz Sánchez 20242020223
Jeison Felipe Cuenca: 20242020043

Professor: Carlos Andrés Sierra Virguez

Date: September 2025

Contents

1 Introduction

1.1 Context Summary

Otto Group is one of the largest e-Commerce companies in the world, having subsidiaries in more than 20 countries, including the USA with Crate & Barrel, Germany with Otto.de, and France with 3 Suisses. They sell millions of products worldwide every day, with several thousand products being added to their product line.

A consistent analysis of the performance of products is crucial. However, as a consequence of its diverse global infrastructure, many identical products are classified differently. Therefore, the quality of their product analysis depends on the ability to accurately cluster similar products. The better the classification, the more insight there can be generated about product range.

Submissions are evaluated using the multi-class logarithmic loss. Each product has been labeled with one true category. For each product, you must submit a set of predicted probabilities (one for every category). The formula is then:

$$\text{logloss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \tag{1}$$

1.2 General Context

An e-commerce is a business model where products or services are bought and sold through online platforms or websites; it allows customers to browse, compare, purchase, and pay digitally at any time, offering both physical goods like clothing or electronics and digital services such as subscriptions or software, with well-known examples including Amazon, Shein, and Temu.

It is important to know that an e-commerce (fig 1, system proposed for e-commerce challenge) is not just sell and buy products, is a complex system with several layers that interact between each other, there are others components that are included in the e-commerce that the sellers did not see such as:

- **Product Information Management (PIM):** Product attributes, descriptions, images, identifiers, this make the product more attractive for the people who is interested.
- **Catalog and Classification:** Categories, subcategories, and hierarchical structures that organize products, for find the best option.
- **Inventory and Logistics Management:** Stock levels, warehouses, suppliers, and delivery times. It is important to optimize the delivery time, some cases, the products could spend one month.
- **Sales and Transactions:** purchase records and dynamic pricing strategies.
- **Customer Data:** browsing patterns, cart activity, purchase history, and segmentation, know more about the user for show the correct product.

- **Marketing and Sales Channels:** mobile apps, external marketplaces, and digital campaigns like fb ads, are useful for make a direct interaction between the product and the buyer.
- **Analytics and Reporting:** – performance metrics (KPIs), clustering insights, sales trends, and profitability analysis.

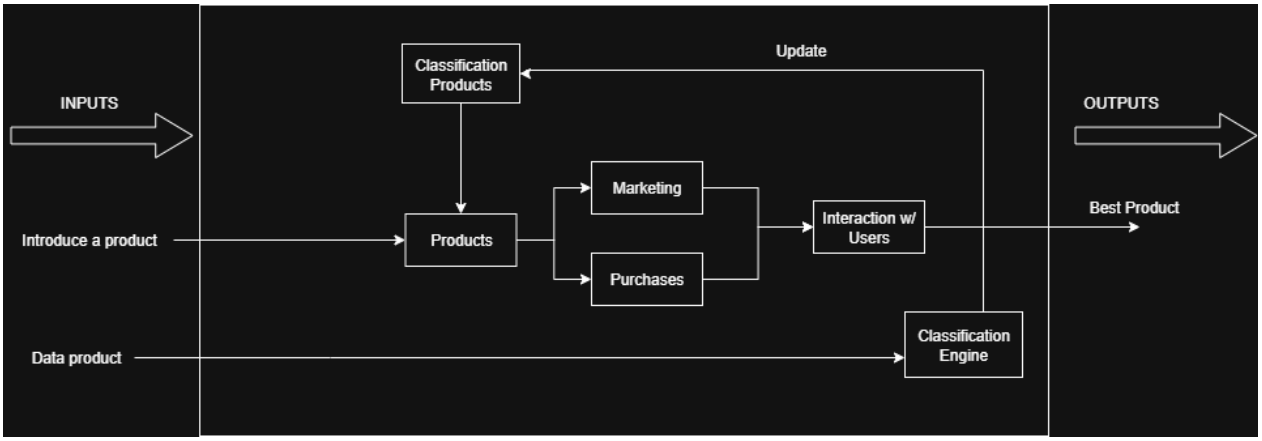

Figure 1: System About E-commerce - Kaggle

2 Methodology and Deliverables

2.1 Competence Summary

The challenge lies in building a classification model capable of correctly assigning products to 9 main product categories, given their numerical features.

The metric used to evaluate the predictions is the *log loss* for multiple classes (referred to as multi-class logarithmic loss), which strongly penalizes incorrect predictions made with high confidence.

Available data:

More than 200,000 products (rows) are provided, each with 93 obfuscated numerical values. *features* (i.e., without direct interpretation). In the training set (*train*), the `target` column appears, which corresponds to the product category. In the test set (*test*), the `target` column is absent, and the task is to predict probabilities for each class.

In addition, each instance or product has an identifier `id`, which does not contribute to classification and is therefore discarded as a predictive variable.

The classes are not perfectly balanced: some categories have more examples than others.

Constraints:

The submission file had to include:

- Column `id` from the test set (`test`).
- Nine probability columns (one for each class).
- The sum of probabilities in each row had to be equal to 1.

Evaluation Metric: multi-class logarithmic loss.

Number of submissions: maximum 5 per day.

Leaderboard:

- **Public:** Only a portion of the `test` set.
- **Private:** Used for the final ranking.

Data:

- Work was restricted to the provided datasets (`train` and `test`).
- The use of external or domain-specific data was prohibited.
- The variables (93 *features*) were obfuscated.

Classes:

- There were 9 categories.
- The distribution among them was imbalanced.

2.2 System Analysis Report

Elements:

- **Products:** Each row (instance) corresponds to a different product offered by Otto Group.

  These are what must be classified. Business decisions depend on having an accurate classification: analysis, inventories, marketing, customer relations, etc.

- **Variables:** Numerical features of the products; they are "obfuscated" → their actual meaning is unknown.

  These are the signals that feed the model. Since their precise meaning is unknown, one relies on statistical correlation, feature engineering, variable selection, and exploratory analysis. They also influence model choice (e.g., models robust to independent variables, sparse values, correlations, etc.).

- **Class:** The category to which each product actually belongs.

  This is what is to be predicted. It is tied to the evaluation metric (multi-class log loss), to class imbalance, to model design, etc. In a business context, these categories represent the main product lines analyzed by Otto.

- **Training data:** The training set contains products with known categories; the test set contains only features, and probabilities must be predicted.

  This allows training models and then evaluating them on new data. The separation enforces consideration of generalization. The public/private leaderboard is based on partitions of the test set.

- **Metric:** Penalizes errors made with high confidence; requires calibrated probabilities for all classes, not only correct classification.

  This imposes that models must not only "predict the right class," but also produce reasonable probabilistic predictions. It influences decisions: algorithm choice, regularization, ensembling, calibration, etc. It also entails that errors on minority classes can weigh heavily if not modeled properly.

- **Constraints:** Technical/procedural limitations imposed by the competition.

  These serve to ensure fairness, avoid "overfitting" the leaderboard, and foster models that generalize. They influence strategy: cross-validation, masking of private test data, etc.

- **Business (Otto Group context):** Otto sells millions of products in many countries; they face issues of inconsistency in the classification of identical products, which affects analysis and decision-making.

  This justifies the challenge: improving the quality of product analysis for the company. With accurate classification, clustering, performance metrics, more reliable analysis, and improved customer experience can be achieved.

- **Correlations:** Technical processes applied by competitors to improve performance.

  These are directly related to the fact that variables lack explicit meaning, many may be correlated, and some irrelevant. Such steps help stabilize the model, avoid noise, and reduce the risk of overfitting.

- **Models and algorithms:** The machine learning methods used: trees, boosting, neural networks, ensembles, etc.

  Their selection depends on the data, the metric, the dataset size, and the restrictions. For example, models that provide probabilities for all classes are necessary, that can handle obfuscated features, and tolerate class imbalance. Hardware and available time are also determining factors.

2.3 Complexity and Sensitivity

Obfuscated Features:

One of the main challenges lies in the fact that the 93 input variables are completely obfuscated, meaning that their real-world meaning is unknown. This implies that participants cannot rely on business domain knowledge to design new variables or discard irrelevant ones. For example, if a feature were "weight" or "color," it would be easier to interpret its

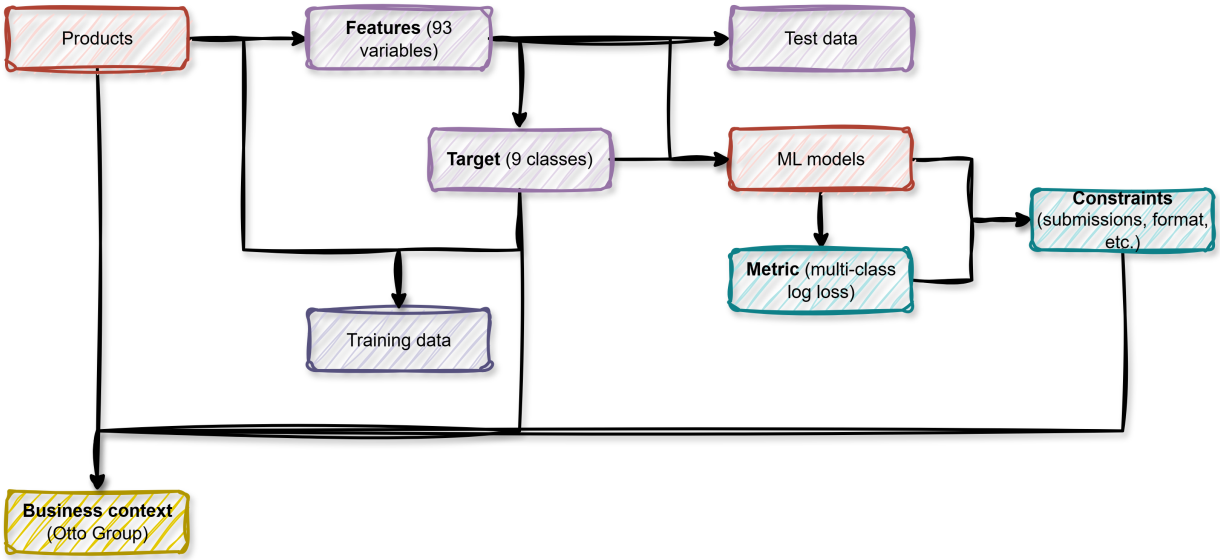
Products
Features (93 variables)
Test data
Target (9 classes)
ML models
Constraints (submissions, format, etc.)
Metric (multi-class log loss)
Training data
Business context (Otto Group)

Figure 2: Corelation Between Elements Diagram

relationship with the product, but here the information is given as numerical values without apparent semantics. This forces participants to work solely with statistical, mathematical, and correlation-exploration methods, greatly limiting intuitive feature engineering.

High Dimensionality:

The dataset contains 93 variables and more than 200,000 rows, which makes it a high-dimensional problem. This situation brings two challenges: on the one hand, the risk of overfitting, since a complex model may learn spurious patterns that do not generalize well; and on the other hand, computational costs, as training models on such a large space can be slow and resource-intensive. This necessitates the application of dimensionality reduction, variable selection, or regularization, as well as careful algorithm optimization to achieve a balance between accuracy and computational time.

Demanding Metric:

The evaluation metric is the multi-class log loss, which not only measures whether the most probable category is correctly predicted, but also the quality of the assigned probabilities. The major challenge is that it penalizes highly confident errors very severely: if the model predicts a probability close to 1 for an incorrect class, the loss is extremely high. This implies that the simple approach of “predicting the correct class” is not sufficient; instead, well-calibrated predictions must be produced, adding an additional layer of complexity to the task.

Proper Validation:

Designing an adequate validation strategy was another critical challenge. Since the public leaderboard used only a portion of the test set, there was a high risk of overfitting to the leaderboard; that is, optimizing models that seemed strong on the public portion but failed on the private one. If the internal validation (for example, a hold-out or cross-validation) did not accurately reflect the true distribution of the test data, the results obtained were misleading. Therefore, it was essential to use robust cross-validation and maintain discipline to avoid overfitting to what Kaggle's leaderboard displayed.

Model Selection:

The choice of algorithms also presented a challenge. Simple models, such as logistic regression, could be too weak to capture the complex relationships between the 93 features. In contrast, more powerful models such as XGBoost, neural networks, or ensembles offered great predictive capacity, but required fine-tuning of hyperparameters and longer training times. The difficulty lay in finding a balance between simplicity, power, and computational feasibility, especially because the dataset was large and the submission limit required prioritizing experiments with a higher chance of success.

Feature Engineering:

Although the features were obfuscated, feature engineering remained key. Creating new derived variables, such as counts of non-zero values, attribute sums, logarithmic transformations, or binarizations, could significantly improve performance. The challenge was identifying which of these transformations truly contributed to the improvement and which introduced unnecessary noise. Without a semantic guideline, a large number of combinations had to be empirically experimented with, which increased the effort and the risk of falling into irrelevant patterns.

Kaggle Restrictions:

Kaggle's own rules added strategic obstacles. The limit of five submissions per day required careful planning of which models to test on the leaderboard. Furthermore, the submission format required that the nine probabilities sum to one per row, necessitating strict control over output processing. Finally, the prohibition on using external data limited creativity, forcing everyone to work solely with the provided dataset. These restrictions sought fairness, but at the same time demanded discipline and strategy in participation.

Calibration Of Possibilities:

Many classification algorithms do not naturally deliver well-calibrated probabilities. For example, Random Forest or Gradient Boosting can generate outputs that misrepresent true confidence. This is a major problem in a competition whose metric is log loss, because the evaluation depends on probabilities accurately reflecting uncertainty. Addressing this issue required applying calibration techniques such as Platt scaling or isotonic regression, or using cross-validation to adjust probabilities. The additional effort to achieve this calibration was one of the most critical tasks.

Time and Resources Management:

Finally, time and computational resource management were a determining factor. Testing multiple models, ensembles, and hyperparameter configurations was computationally demanding. Furthermore, the competition window required prioritizing which experiments to run first. Poor time management could lead to mediocre models, while a good experiment organization and prioritization strategy made the difference between an average and a competitive result.

2.4 Chaos and Randomness

In addition to complexity and sensitivity, the Otto Group Product Classification Challenge also exhibits elements that can be interpreted through the lens of chaos theory and randomness.

Unpredictable Feature Interactions: The dataset consists of 93 obfuscated numerical features, with no explicit semantic meaning. This creates potential for hidden, nonlinear interactions between variables. Some features may strongly correlate under certain subsets of data but appear independent in others, introducing unpredictable effects in the model's performance.

Feedback Loops with the Leaderboard: The Kaggle leaderboard itself introduces a feedback dynamic. Competitors adjust their models based on public leaderboard results, which represent only a partial view of the test set. This iterative adjustment may create unpredictable oscillations in strategy, where small changes lead to disproportionate effects in private leaderboard ranking.

Business and System-Level Uncertainty: At the business level, inconsistent classification of identical products demonstrates chaotic dynamics. Small differences in product description or feature encoding can lead to entirely different classifications, which in turn affects inventories, customer recommendations, and strategic decisions.

Overall, these aspects highlight that the competition system is not purely deterministic. It involves stochastic training dynamics, unpredictable variable interactions, and systemic feedback loops, all of which align with principles of chaos and randomness in complex systems.

2.5 Conclusion

- The analysis of the Otto Group Product Classification Challenge highlights the complexity of building robust classification systems under strong constraints.
- The competition involves high-dimensional and obfuscated data, making domain-driven feature engineering infeasible and forcing reliance on statistical and computational methods.
- The demanding evaluation metric (multi-class log loss) requires not only correct classifications but also well-calibrated probabilities, significantly increasing the modeling challenge.

- There are multiple interdependence such as datasets, variables, business context, evaluation rules, and strategic constraints, whose interact in nonlinear and sometimes unpredictable ways, sensitivity and chaotic dynamics in model outcomes.
- The strengths of the system lie in its ability to simulate real-world classification uncertainties and encourage the use of advanced machine learning techniques. However, weaknesses include the lack of semantic interpretability of features and the strong susceptibility to overfitting, both to the data and to the public leaderboard.
- the Otto challenge serves not only as a machine learning competition but also as a case study in systems analysis. This makes it an ideal scenario to apply systems engineering principles, bridging data science practice with systemic thinking.
-

The strengths of the system lie in its ability to simulate real-world classification uncertainties and encourage the use of advanced machine learning techniques. However, weaknesses include the lack of semantic interpretability of features and the strong susceptibility to overfitting, both to the data and to the public leaderboard.

In conclusion, the Otto challenge serves not only as a machine learning competition but also as a case study in systems analysis. This makes it an ideal scenario to apply systems engineering principles, bridging data science practice with systemic thinking.

3 Bibliography

References

[1] Kaggle. (s. f.). *Otto Group Product Classification Challenge*. Available at: `https://www.kaggle.com/competitions/otto-group-product-classification-challenge/` Accessed September 11, 2025.