Assessment of Material Supply Risks in Make-to-Order Manufacturing Using Machine Learning Methods*

Andrew Mrykhin[1,*], Svitlana Antoshchuk[1]

[1] *Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, Ukraine*

Abstract

The study investigates the use of machine learning to improve assessment of material supply risks in Make-to-Order (MTO) and High-Mix, Low-Volume (HMLV) manufacturing. Such production systems, characterized by product variety and low inventory buffers, are highly sensitive to delays in material deliveries. Traditional risk management frameworks, designed for stable mass production, often cannot respond quickly enough to the dynamics of MTO environments. The research proposes improvement of a prior supply risk evaluation model by introduction of ML-based supply delay prediction module that draws on information available in enterprise resource planning (ERP) systems. Several machine learning models were trained and compared, including logistic regression, gradient-boosted decision trees (XGBoost, LightGBM), and TabPFN, a recent transformer-based model for tabular data. Results show that while linear models offer interpretability, they lack sensitivity to minority delay cases that are most critical for operations. Gradient-boosted models significantly improved predictive quality and stability, with LightGBM providing the best trade-off between accuracy and explainability. The transformer-based TabPFN achieved the highest overall classification performance, confirming the growing potential of foundation models even on small industrial datasets. The study demonstrates that employing machine learning methods can enhance the precision and timeliness of supply risk evaluation, enabling near real-time monitoring and more informed procurement decisions.

Keywords

supply risk, machine learning, make-to-order manufacturing, XGBoost, LightGBM, TabPFN

1. Introduction

In recent decades, manufacturing supply chains have become increasingly complex and interdependent. Globalization, outsourcing, and product variety have brought efficiency gains but also new forms of vulnerability. Disruptions in logistics, materials, or suppliers now quickly propagate through production systems, exposing firms to operational and financial risks. As a result, the assessment and management of supply risks have become critical elements of modern industrial practice.

Classic supply chain risk management frameworks, developed for stable, high-volume production, rely on multi-stage processes and expert-driven evaluations. These approaches often struggle to keep pace with the volatility and data intensity of Make-to-Order (MTO) and High-Mix, Low-Volume (HMLV) manufacturing. In such environments, production schedules depend tightly on timely material deliveries, leaving little margin for delay and demanding faster, more adaptive forms of risk monitoring.

Advances in digital manufacturing and the availability of real-time data from ERP systems now open the way for more dynamic, data-based risk assessment. In this context, machine learning methods offer a promising path toward timely and automated evaluation of material supply risks—

Applied Information Systems and Technologies in the Digital Society (AISTDS-2025), October 01, 2025, Kyiv, Ukraine
* *Corresponding author.*
amrykhin@gmail.com (A. Mrykhin); asg@op.edu.ua (S. Antoshchuk)
0009-0009-1545-632X (A. Mrykhin); 0000-0002-9346-145X (S. Antoshchuk)

capable of capturing complex dependencies and improving decision support in procurement and production planning

2. Supply risks in manufacturing overview

Supply risk management traces its origins to the pioneering works of George Zsidisin and Christine Harland [1], [2], [3] at the turn of the millennium. During the following decade, the number of studies devoted to material supply risks and supply chain risks as a whole increased rapidly. In a 2021 review, Amulya Gurtu and Jestin John [4] identified 455 publications on this topic covering the ten-year period from 2010 to 2019.

By 2009 supply chain risk management - SCRM emerged as a separate discipline, officially recognized with the inclusion of Supply Chain Risk Management in the ISO 31000 standard [5]. Various approaches to risk assessment were proposed, both qualitative and quantitative, based on scores, ratings, weights, or probabilities [6], [7]. Different frameworks for managing supply chain risks have been developed.

Despite the variety of proposed systems, certain common features can be identified:

- multi-stage risk management process: identification, analysis, assessment, response, often presented in a cycle similar to the Deming cycle. This process can be quite laborious and time-consuming, requiring regular involvement of specialists and experts.
- adoption of the concept of risk realization as a consequence of relatively rare non-standard events (triggering events) that disrupt the normal functioning of business processes. Much effort is then put into identification, classification, tracing of these events, determining their probabilities and possible outcomes.
- the primary focus of SCRM on the impact of risks on the overall business operation, financial performance, or operational stability.

While the SCRM provides a well-developed foundation for supply risk management, its rigidity and focus on multi-stage procedures, creates limitations in case of Make-to-Order and, especially High-Mix, Low-Volume (HMLV) manufacturing.

Such production environments are defined by inherent instability and dynamism. MTO companies operate with a high variability of demand and, consequently, maintain limited safety stocks of specialized components [8].

This variability is compounded by a complex supplier ecosystem, which includes a reliance on a wide base of partners, often involving niche suppliers for specific parts.

In MTO setting the firm's Master Production Schedule is tightly coupled to the supply schedule. Any delay in receiving materials due to an unforeseen supply risk may quickly lead to a production bottleneck.

Narrow window for effective risk intervention and quickly changing state of supply channels makes fast (ideally - real-time) updates of supply risks assessments crucial for keeping up with production schedule

3. Prior work

To address the aforementioned need of timely risk evaluations we proposed a model for automated assessment of small-to moderate deviations in supply channels (tactical risks) that was implemented as service in ERP system and relied on data available in the customers information systems [9].

The model developed automatically evaluates supply risks for each production order using data from the ERP system. It calculates the risk of delayed order fulfillment as the combined standard deviation of delivery times for all required materials. Each material's risk is derived from the deviations of its supply chain segments, updated dynamically from logistics data. The model outputs

quantitative risk measures, supporting real-time decision-making in production planning and procurement.

The model was deployed to production and demonstrated capability to produce useful risk predictions.

However a few limitations also emerged. Shortcomings discovered are mainly centered around following points:

- model relies on assumption that deviations follow normal distribution.
- model focus on time deviation may leave non-linear, categorical, or external factors uncaptured.
- complex interactions between features can't be modeled.
- data sparsity for some segments undermines stability of risk estimates.

These constraints indicate the need for more advanced analytical methods. In this regard, machine learning methods offers a promising direction [10], [11]:

- machine learning models are non-parametric, they do not require a rigid distributional assumption.
- ML models inherently handle non-linearity and can automatically quantify the contribution of numerous features to the final risk score.
- advanced ML models, i.e. gradient boosting based can effectively model.

4. Aims and objectives of the research

Having established the context of the problem, we can now delineate the specific aims and objectives of our research.

Our aim is to:

- explore application of machine learning methods for supply risk analysis in make-to-order and high-mix low-volume manufacturing.

Our tasks are:

- extraction and analysis of data on material supplies.
- data preprocessing.
- configuration and application of key tabular data classification models, including logistic regression, decision tree–based models and gradient boosting, as well as specialized implementations of neural networks.
- evaluation and comparison of model performance in terms of classification quality and result traceability.
- development of recommendations for integrating the studied models into an applied solution.

5. Enhanced model structure overview

The structure of the proposed model is presented in Figure 1.

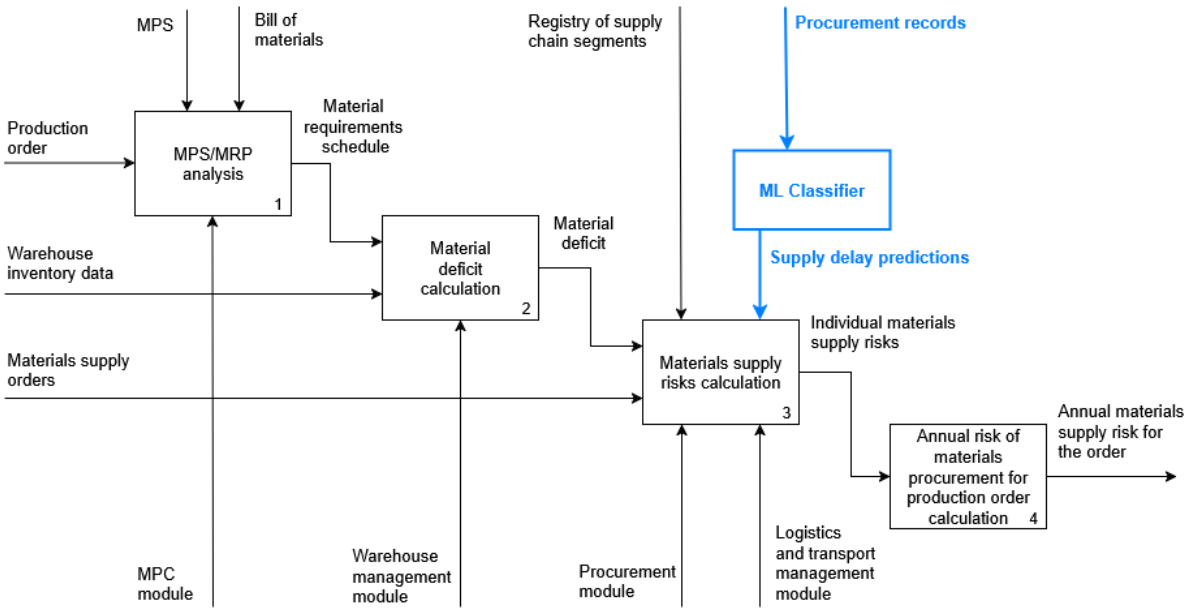

Figure 1: Proposed improved risk assessment model.

At early stages of procurement order execution an existing statistical risk computation algorithm is employed. However, as more transaction details become available, completing the features set required by the ML classification model, the classifier output is introduced to supply risk calculation.

Addition of ML model predictions into the risk evaluation pipeline allows to overcome statistic model rigidity and ensures higher precision of risk assessments.

6. Supply delay prediction task

To develop and train supply delays prediction model we explored the ERP system of the customer for relevant data. The dataset consisting of materials procurement records was extracted from the ERP database.

The total number of extracted samples is 3381.

Overall, 39 features were extracted including data about supplied material, quantity, currency, price, sum, supplier, country, delivery method, dates of order, shipment, delivery etc., payments, documents issued and persons involved. A notable characteristic of the extracted dataset is the predominance of categorical features.

Some samples contained missing values in different features including some key columns such as shipping or delivery dates. We strived to fill them as much as possible at the extraction stage using indirect data such as ERP logs, information from coupled modules, etc.

Baseline post-extraction preprocessing was performed including type conversions and correction of identified errors in data. We also calculated delivery deviations and shipping time frames from transition dates available in the dataset.

Our prediction target is deviation of actual delivery time from planned. Target values distribution is both skewed and extremely high-peaked, see histograms in Figure 2.

At first, we tried to implement a regression model, predicting the deviation of the actual delivery date from the planned one. However, regression results proved to be unstable, so we moved to multiclass classification and separated deviations into four target classes based on business-significant thresholds, as presented in Table 1.

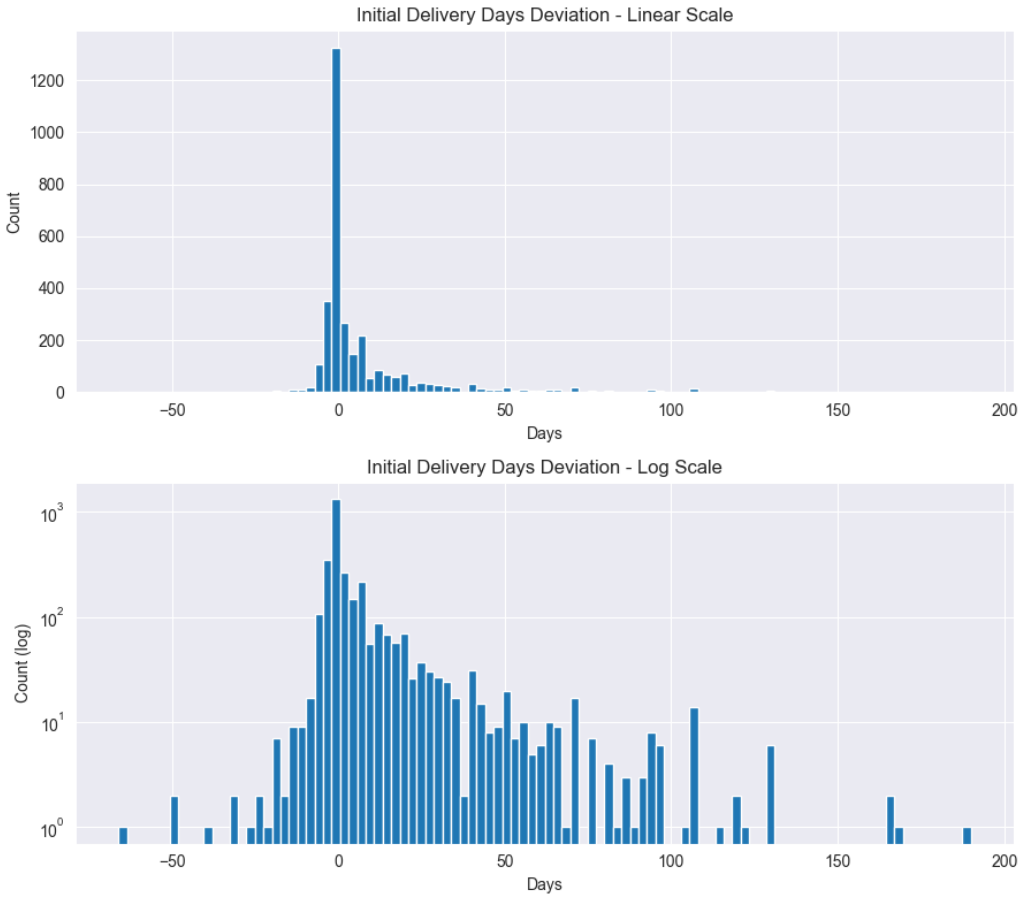

Figure 2: Target values distribution. Top – linear scale, bottom – log scale.

Table 1
Classification target classes and samples distribution

Class	Delay, days	Samples no
IN_TIME	<= 1	2343
SHORT_DELAY	2 - 5	289
MODERATE_DELAY	6 - 15	281
LONG_DELAY	> 15	260

The division is grounded on expert assessments provided by the customer's manufacturing managers, reflecting the fact that changes in delay duration correspond to different managerial responses:

- SHORT_DELAY: 1–5 days — typically manageable using existing inventory or minor adjustments to the production schedule without affecting the final order commitment date;
- MODERATE_DELAY: 5–15 days — represents a risk that may or may not impact the final order depending on the material's criticality and current buffer stock, requiring managerial intervention and additional assessment;
- LONG_DELAY : > 15 days — delays beyond this threshold generally lead to expedited shipping costs, alternative procurement actions, or a failure to meet the promised order completion date.

The resulting target classes are highly imbalanced, with the majority of samples falling into the IN_TIME category.

During the exploratory data analysis we utilized both visual and numerical tools to uncover patterns and relationships in our data and identify features most useful for our prediction task. Overall 7 features were chosen as most promising predictors, a list of selected predictors is presented in Table 2.

Table 2
Features selected for modeling

Feature	Type	Cardinality
SUMM	Float	-
ID_TYPE_DELIVERY	Category	6
ID_CURRENCY	Category	3
ID_COUNTRY	Category	29
ID_SELLER	Category	158
ID_ELEM	Category	698
DAYS_TO_SHIP	Integer	-

Among features selected for modeling are 5 categorials with cardinality from 3 to 698. As some of the models used in classification require encoding of categorial features we performed one-hot encoding of ID_CURRENCY, ID_TYPE_DELIVERY, ID_COUNTRY features and target encoding of ID_SELLER, ID_ELEM features while separately keeping original for use with models that handle categorials natively.

7. Modeling

For comparative evaluation on our classification task we selected four models:

- Logistic regression - essential baseline model, simple, efficient and interpretable.
- XGBoost - mature and widespread implementation of gradient boosting machines, the dominant choice for tabular data analysis for years.
- LightGBM - modern optimized boosting framework capable of efficient native handling of categorial features.
- TabPFN - a transformer-based foundation model that can deliver highly accurate predictions on small to medium-sized datasets with minimal preprocessing.

7.1. Logistic Regression

Logistic Regression is a classical statistical classification method that models the probability of a categorical outcome using a logistic (sigmoid) function. Despite its simplicity, it remains a widely used baseline in supervised learning due to its interpretability and efficiency. Its main advantages are transparency, low computational cost, and ease of regularization, which makes it well suited for initial benchmarking and feature relevance assessment before applying more complex models.

Logistic Regression doesn't support categorial features natively, so we have to use one-hot and target encodings.

We will employ regularization and cross-validation to control overfitting. And we use class weighting to counter target classes imbalance. We also tested SMOTE, but it produced worse results than weighting.

The model fitting process has converged in 61 iterations and performed consistently between cross-validation folds. This indicates model stability and ability to capture significant risk features. However, the model performed poorly at classifying the critical minority risk classes. Overall accuracy achieved is 67%, macro-F1 score - 52%. Low recall in delay classes is unacceptable for risk management because the model frequently predicts in-time delivery in cases of actual delays.

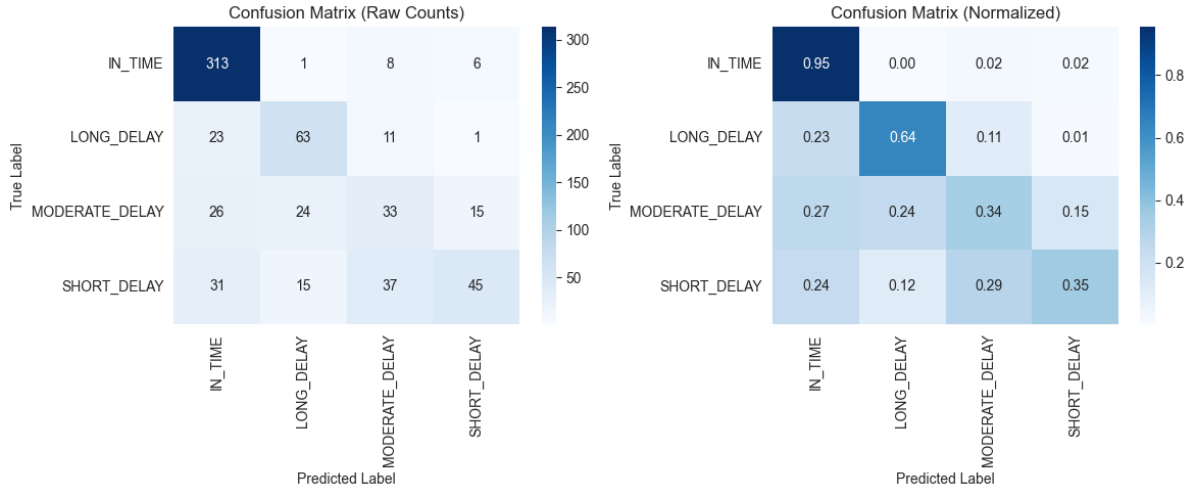

Figure 3: Confusion matrix for Logistic Regression output. Left – raw counts, right – normalized.

7.2. XGBoost

XGBoost is a powerful ensemble machine learning algorithm built on the principles of Gradient Boosting Machines (GBM) and is widely regarded as a standard method for classifying structured (tabular) data. Developed by Tianqi Chen in 2016 [12], XGBoost (Extreme Gradient Boosting) represents a specialized and highly optimized implementation of gradient boosting that quickly gained prominence due to its balance of accuracy, computational efficiency, and interpretability.

Initially XGBoost had no support for categorial features and encoding was required, however in later versions such support was added as experimental. We tried both one-hot plus target encodings and native handling with later delivering the better results.

We have configured early stopping to control overfitting. And we use class weighting to counter target classes imbalance. We also tested SMOTE, but it produced worse results than weighting.

The XGBoost model not only achieved a higher overall Accuracy (73% vs 67%) but, more critically, delivered a 9.2 percentage point improvement in Macro-F1 Score. This demonstrates its superior ability to correctly identify and differentiate the complex minority delay risks.

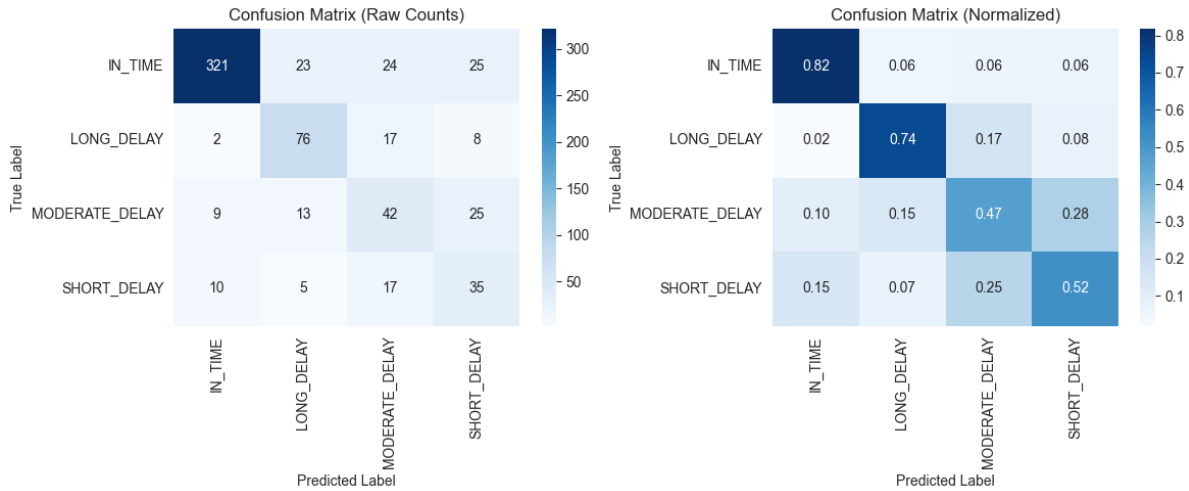

Figure 4: Confusion matrix for XGBoost output. Left – raw counts, right – normalized.

7.3. LightGBM

LightGBM (Light Gradient Boosting Machine) is a high-performance framework for gradient boosting developed by Microsoft Research in 2017 [13]. It was designed to improve both the speed and scalability of tree-based ensemble learning, addressing some of the computational bottlenecks

found in earlier implementations such as XGBoost. LightGBM introduces several algorithmic innovations, including histogram-based decision tree construction and the use of leaf-wise tree growth with depth constraints, which together enable faster training and better accuracy on large datasets.

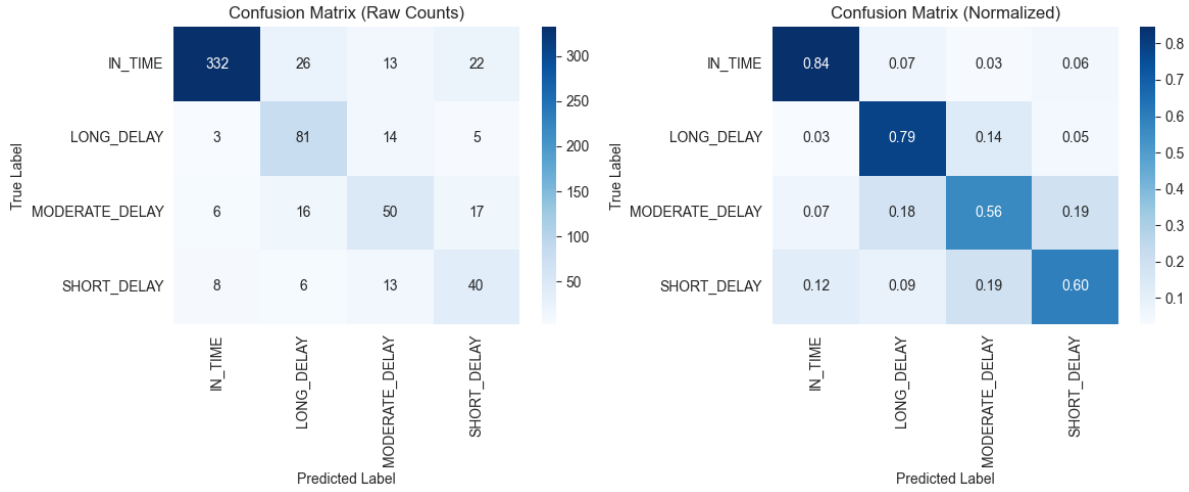

Figure 5: Confusion matrix for LightGBM output. Left – raw counts, right – normalized.

LightGBM has native efficient handling of categorical data.

To deal with overfitting and class imbalance we employed same techniques as with XGBoost: early stopping and class weighting.

Among models tested up to now The LightGBM model yielded the highest scores, particularly on the critical minority classes. The model showed a significant 6 percentage point Macro-F1 improvement over XGBoost, reaching a stable 67% Macro-F1 Score.

7.4. TabPFN

TabPFN (Tabular Prior-Fitted Network) is a recent transformer-based model designed specifically for tabular data classification [14], [15]. We are using version 2 of the model, developed by researchers at the Technical University of Munich and introduced in 2025. The model represents a novel paradigm that applies the principles of large pre-trained foundation models—common in natural language processing—to structured datasets. TabPFN implements the idea of in-context learning (ICL), similar to what we see in large language models. Instead of learning a fixed mapping from features to outcomes for one dataset, it learns how to learn from examples.

TabPFN is pre-trained on a vast, synthetically generated collection of millions of small tabular problems. This pre-training enables the model to internalize a wide range of statistical patterns and relationships, allowing it to generalize effectively to new datasets with minimal additional training.

To better use the tabular structure, TabPFN authors proposed an architecture that uses a two-way attention mechanism, with each cell attending to the other features in its row (that is, its sample) and then attending to the same feature across its column (that is, all other samples). This design enables the architecture to be invariant to the order of both samples and features and enables more efficient training and extrapolation to larger tables than those encountered during training.

TabPFN also addresses inherent to transformer-based ICL algorithms problem of repeating computations on the training set for each test sample in a fit-predict setting. The model can separate the inference on the training and test samples. This allows to perform ICL on the training set once, save the resulting state and reuse it for multiple test set inferences.

The model works remarkably well on datasets up to around ten thousand samples and a few hundred features, often outperforming widely used tree-based methods without dataset-specific training or tuning.

It can handle categorial features and target classes imbalance natively.

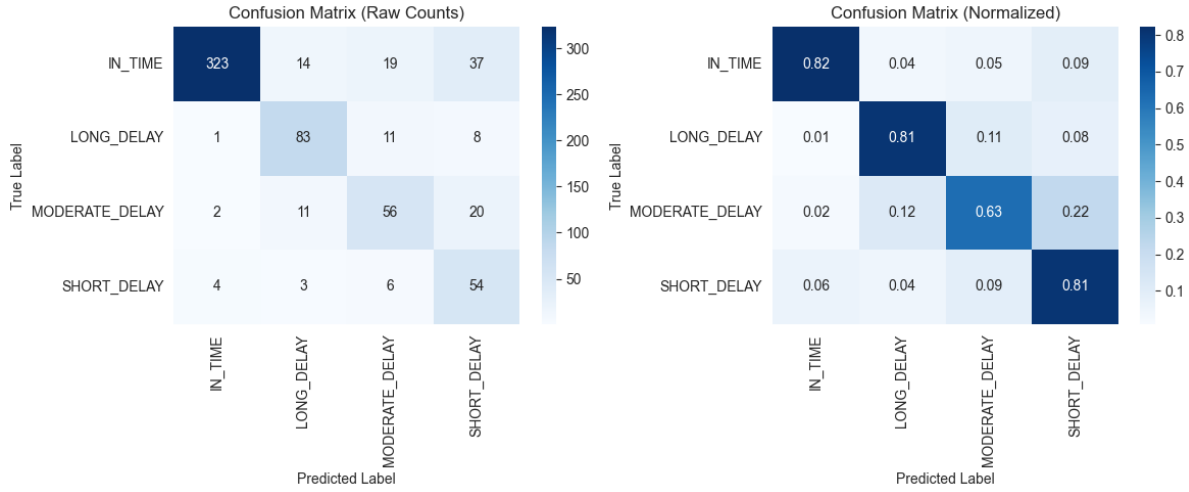

Figure 6: Confusion matrix for TabPFN output. Left – raw counts, right – normalized.

TabPFN achieved a 79% Accuracy and a new high of 72% Macro-F1 Score. These are the best classification results of all our models.

7.5. Modeling results summary

Let's review classification results focusing on balanced performance across all target classes (macro-f1 metric), and the critical 'long delay' outcome—our worst-case scenario, see Table 3.

Table 3
Classification results comparison

Model	Accuracy	Macro-F1	Recall on LONG_DELAY
Logistic Regression	67%	51.8%	54%
XGBoost	73%	61.0%	74%
LightGBM	77%	67.0%	79%
TabPFN	79%	72.0%	81%

We have two top-performers. The novel TabPFN model achieved the best classification quality outperforming established leaders in the class - the gradient-boosted decision trees models (GBDTs). However, because computational cost of its transformer attention mechanism grows quadratically with number of samples, performance may degrade as dataset size grows. Also in terms of explainability, TabPFN currently offers limited interpretability compared to tree-based models.

The second performer, LightGBM, combines strong classification accuracy with excellent interpretability and scalability. Its built-in feature importance metrics and compatibility with SHAP analysis provide clear insights into model behavior. This makes LightGBM particularly suitable for integration into applied industrial systems where both predictive performance and transparency are essential.

8. Limitations and Directions for Further Research

The limitations of this study are primarily due to the fact that the work was conducted within the production environment and based on the data extracted from the information systems of a single enterprise. While this ensured the use of real-case-based data, managerial practices, and operational conditions, it also introduces several constraints.

First of all, the size of the dataset employed is limited by the historical information available in the customer's ERP system, where the volume of records naturally depends on the scale of materials procurement. Although sufficient for methodological demonstration, such data may not fully capture the diversity of supplier behaviors, material types, and logistics conditions typical of broader industrial ecosystems.

Secondly, the model's close alignment with the business practices and production processes of a specific manufacturer may complicate its direct integration into risk management systems of other enterprises, especially those with different operational structures or data models.

Some issues mentioned in this study also require further elaboration. A systematic examination of model explainability techniques—such as SHAP or other feature attribution methods—was not performed. This omission may limit the applicability of the model in organizations where interpretability is required due to internal policies or regulatory constraints. The question of model scalability for larger and more heterogeneous datasets was also only briefly explored.

In future work, the authors plan to focus primarily on long-term monitoring of model performance, comparative analysis with existing and alternative risk assessment solutions, and evaluation of the model's impact on business processes and its resulting economic benefits. Further attention will also be given to questions of interpretability and statistical significance testing of model outcomes.

A particularly important direction for continued research is the integration of model outputs into the company's overall risk management workflow, ensuring that the predicted risk indicators align with managerial decision-making and planning.

Another key area of future development is the expansion and enrichment of training datasets. This may be achieved by incorporating additional information already present within the client's IT systems or becoming available through the ongoing adoption of digital manufacturing and logistics technologies. It is also promising to supplement internal enterprise data with external inputs received from partners and suppliers via electronic data exchange systems, where permitted.

The authors also intend to explore the use of non-tabular data—such as time series and geospatial information—and the corresponding extension of analytical tools, both by integrating these modalities into tabular representations (e.g., through embeddings) and by analyzing them separately using appropriate types of neural network architectures.

Finally, the authors strive to expand further research beyond the limits of the business environment of a single company. Efforts will be made to involve other enterprises employing MTO and HMVL manufacturing models, and we are open to cooperation on this topic.

9. Conclusion

We demonstrated the capability of ML classification models, particularly ensemble and deep learning techniques, to yield high-quality predictions even on challenging tabular datasets characterized by a limited number of samples, numerous categorical features, and the absence of strong, isolated linear predictors.

A notable finding is the declining need for complex, manually engineered categorical feature encodings. Modern classification models proved their ability to effectively handle categorials natively without sacrificing predictive power

The modeling demonstrated the exceptional predictive capabilities of the Transformer-based TabPFN classifier on tabular data. This Prior-Fitting Network achieved the highest classification quality on our dataset, challenging the current dominance of tree-based ensemble methods in this domain, especially for smaller data samples.

The integration of these advanced ML tools into the risk assessment framework allows us to significantly improve both the precision and the stability of risk evaluations. This leads to better managerial decisions regarding inventory buffering, supplier selection, and proactive production schedule adjustments.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] G. Zsidisin, A. Panelli, and R. Upton, "Purchasing organization involvement in risk assessments, contingency plans, and risk management: An exploratory study," *Supply Chain Management: An International Journal*, vol. 5, pp. 187–198, 2000, doi: 10.1108/13598540010347307.

[2] C. Harland, R. Brenchley, and H. Walker, "Risk in supply networks," *Journal of Purchasing and Supply Management*, vol. 9, pp. 51–62, 2003, doi: 10.1016/S1478-4092(03)00004-9.

[3] G. Zsidisin, "A grounded definition of supply risk," *Journal of Purchasing and Supply Management*, vol. 9, pp. 217–224, 2003, doi: 10.1016/j.pursup.2003.07.002.

[4] [A. Gurtu and J. Johny, "Supply chain risk management: Literature review," *Risks*, vol. 9, no. 1, 2021, doi: 10.3390/risks9010016.

[5] G. Schlegel and R. Trent, Supply Chain Risk Management: An Emerging Discipline. CRC Press, 2014, doi: 10.1201/b17531.

[6] V. Lytvynenko, S. Antoshuk, A. Mrykhin, N. Savina, and O. Marchuk, "Applying the Monte Carlo Method for Modeling Order Fulfillment with Consideration of Supply Risk," in *Proc. Modeling, Control and Information Technologies: International Scientific and Practical Conference*, vol. 7, pp. 251–257, 2025, doi: 10.31713/MCIT.2024.078.

[7] I. Heckmann, T. Comes, and S. Nickel, "A critical review on supply chain risk – definition, measure and modeling," *Omega*, vol. 52, 2014, doi: 10.1016/j.omega.2014.10.004.

[8] [8] Z. L. Gan, S. N. Musa, and H. J. Yap, "A review of the high-mix, low-volume manufacturing industry," *Applied Sciences*, vol. 13, no. 3, 2023, doi: 10.3390/app13031687.

[9] A. L. Mrykhin and S. G. Antoshchuk, "Information Model for Assessing the Impact of Tactical Material Procurement Risks on Order Fulfillment in Make-to-Order Manufacturing," *Herald of Advanced Information Technology*, vol. 7, pp. 243–252, 2024, doi: 10.15276/hait.07.2024.16.

[10] [A. Aljohani, "Predictive Analytics and Machine Learning for Real-Time Supply Chain Risk Mitigation and Agility," *Sustainability*, vol. 15, no. 20, 2023, doi: 10.3390/su152015088.

[11] N. Rezki and M. Mansouri, "Machine Learning for Proactive Supply Chain Risk Management: Predicting Delays and Enhancing Operational Efficiency," *Management Systems in Production Engineering*, vol. 32, pp. 345–356, 2024, doi: 10.2478/mspe-2024-0033.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, pp. 785–794, 2016, doi: 10.48550/arXiv.1603.02754.

[13] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 30, pp. 3146–3154, 2017, doi: 10.5555/3294996.3295074.

[14] N. Hollmann, S. Müller, L. Purucker, et al., "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, pp. 319–326, 2025, doi: 10.1038/s41586-024-08328-6.

[15] Q. Zhang, Y. S. Tan, Q. Tian, and P. Li, "TabPFN: One Model to Rule Them All?," *arXiv preprint arXiv:2505.20003*, 2025, doi: 10.48550/arXiv.2505.20003.