AHEAD OF EXPIRY: MACHINE LEARNING DRIVEN ALERTS FOR SMARTER PROCUREMENT IN THE PETROCHEMICAL SECTOR

Capstone Project Public Report

By
Márton Nagy

Submitted to Central European University - Private University
Department of Economics

In partial fulfilment of the requirements for the degree of Master of Science in Business Analytics

Vienna, Austria
2025

COPYRIGHT NOTICE

[1]

For bibliographic and reference purposes this thesis should be referred to as: Nagy, M. 2025. Ahead of Expiry: Machine Learning Driven Alerts for Smarter Procurement in the Petrochemical Sector. Capstone Project Public Report. MSc Capstone Project Public Report, Department of Economics, Central European University, Vienna.

[1] Icon by Font Awesome.

AUTHOR'S DECLARATION

I, the undersigned, **Márton Nagy**, candidate for the MSc degree in Business Analytics declare herewith that the present thesis titled "Ahead of Expiry: Machine Learning Driven Alerts for Smarter Procurement in the Petrochemical Sector" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography.
I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 08 June 2025

Márton Nagy

TABLE OF CONTENTS

1. Introduction

Procurement plays a crucial role in ensuring the smooth operation of large-scale industrial companies. However, the time it takes to renew or renegotiate procurement contracts (that is, the lead time of the process) can be extremely variable and it depends on multiple factors. As a result, it may not be correctly predicted solely based on expert judgement. However, if a new contract is not in place by the time the previous one expires or is exhausted, companies face the risk of rushed decisions, higher costs, and limited supplier options. Therefore, being able to correctly predict the lead times of the contract renegotiation processes is a highly beneficial ability for companies (Ross, 2015).

This project was developed in collaboration with a major international petrochemical company (name omitted for confidentiality, further referred to as the Client[2]) to tackle this challenge. The goal was simple yet impactful: develop an early-alert platform powered by machine learning models to replace the current system of relying excessively on expert judgement and to help procurement professionals better time when to start the renegotiation process for expiring contracts (Client, personal communication, 31 January 2025)[3].

2. Project Goals and Approach

The primary goal of the project was to build a smart alert system that ranks contracts based on the ideal start dates to begin the renegotiation process, therefore helping procurement professionals of the Client identify which contracts to prioritize. To achieve this, the Client's carefully collected database (Client, 2025) about past contracts and their lead times was leveraged. This included roughly 16 thousand contracts, out of which around 7 thousand could be used for model training (as most of the rest did not have lead times associated with them (around 8 thousand contracts) or was excluded during data cleaning (around 1 thousand contracts)). In addition, the dataset contained attributes about each contract such as the people and units associated with it at the Client, start and end date, value and some information about the partners the contract has been signed with.

Before modeling, to make the models more aligned with business needs, I designed a custom evaluation method (that is, an asymmetric and convex loss function) that prioritized overestimating lead times rather than underestimating them (Békés & Kézdi, 2021, Ch. 13). This was needed as the Client heavily emphasized that it is far less risky to start a negotiation too early than too late (Client, personal communication, 28 February 2025).

[2] For confidentiality reasons, sources authored by the Client are also referenced as stating the "Client" as the author instead of the actual company name.
[3] Note that based on the APA style guide, personal communications are not included in the references.

The modelling process was two-fold. First, I applied time series methods[4] to account for trends and seasonal patterns. This was especially important, as at times of disruptions (like COVID-19) lead time patterns were substantially different, and as subsequent machine learning approaches cannot reliably extrapolate trends (Malistov & Trushin, 2019). Then, I used machine learning models[5] (mostly different boosting algorithms, as these tend to be the most performant on tabular data – see e.g. Shwartz-Ziv and Armon (2021)) to capture more complex relationships in the data that has now been cleaned of time-based patterns. Lastly, a post-prediction binning adjustment was performed, to further prioritize overpredictions, following a similar logic to that of finding the most suitable classification threshold in the case of classification tasks (Békés & Kézdi, 2021, Ch. 17). During all modeling steps, hyperparameters were tuned[6] and the best models were selected via rigorous cross-validation.

The final product was a user-friendly dashboard-platform[7]. There, procurement professionals can easily filter the database to find relevant contracts, view and interpret predictions in detail, explore explanations behind them, and even generate what-if scenarios by adjusting the inputs for existing predictions, or simulating predictions for completely new contracts.

3. Key Results and Benefits

The resulting platform performed substantially better than the Client's current approach, which relied mostly on expert judgement (Client, personal communication, 10 April 2025). Though the exact gain in accuracy could not be quantified, as the Client could not provide information about the current approach's performance, it was still deemed substantial by the Client's experts (Client, personal communication, 10 April 2025). The implemented platform can efficiently help procurement teams plan ahead more confidently, reduce rushed decisions, and potentially save on costs by enabling more competitive supplier selection.

Although predictions for some contract types, like sole source contracts, were less reliable, the tool still represents a major leap forward in data-driven decision-making at the Client. In addition, predictions for tender contracts proved to be quite reliable, which was prioritized much higher by the Client (Client, personal communication, 28 February 2025). This is because

[4] Different OLS models, a PLS model, a PLS model with automatic breakpoint detection, and an exponential smoothing model on monthly aggregated lead times. For OLS and simple PLS models, the scikit-learn implementation by Pedregosa et al. (2011) was used. For the automatic breakpoint detection, I used the ruptures package (Truong et al, 2019). The exponential smoothing model used the statsmodels implementation by Seabold and Perktold (2010).

[5] For OLS, LASSO, elastic net, random forest, ADABoost and support vector regression, the scikit-learn implementation of Pedregosa et al. (2011) was used. For XGBoost, I relied on the implementation of Chen and Guestrin (2016). For LightGBM, the package set forth by Ke et al. (2017) was used.

[6] Using Optuna – see Akiba et al. (2019).

[7] Implemented as a Streamlit application (Snowflake Inc., 2025).

in general, tender renegotiations take longer but can potentially yield more efficient results, thus if time allows, most contracts should be renegotiated in a tender process.

Beyond technical accuracy, a major benefit of the solution is its accessibility. Even non-technical users can understand and interact with the predictions, thanks to the intuitive user interface, built-in explanations, and user guides. The entire process is also reproducible, meaning that new data can be fed into the system, and the models can be retrained with minimal effort. In fact, if the shipped solution is deployed, periodic (e.g. annual or semi-annual) retraining of the models will be needed to account for changes in trends and other patterns.

4. Lessons Learned and Personal Development

This project was a major learning experience for me, both technically and professionally. On the technical side, I gained deeper expertise in how to combine time series modelling with more advanced machine learning approaches, and machine learning explanation techniques such as SHAP-values (Lundberg & Lee, 2017) or permutation-based feature importance (Breiman, 2001).

I also had to think critically about aligning model objectives with real business priorities. It was an enlightening experience to uncover how the Client's business criteria could be translated into loss functions or post-prediction adjustments. On the flip side, I also gained valuable insight into how to pose technical questions in such a way that is also understandable for the business. Perhaps most importantly, I learned the value of interpretability and usability. Building a model and making predictions is one thing, making sure people can actually use and trust it is another. This required me to think about design, user experience, and communication just as much as model performance.

5. Conclusion and Outlook

This project showed how machine learning can be applied even in such business units, like procurement, where approaches utilizing artificial intelligence are still scarce, in a way that is both practical and impactful for the business. The early-alert system for procurement lead times not only improves how contract renegotiations are managed but also opens the door for more data-driven decision-making in other parts of the Client's organization.

While the solution is already usable, there is still room for future development. As mentioned earlier, the predictions could still be improved, especially for sole source contracts. I think that the key to this lies in feature engineering based on deeper domain knowledge which was not accessible as an external consultant. Still, I believe the foundations of the presented approach are solid, and the system is ready to deliver ongoing value with minimal maintenance.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Békés, G., & Kézdi, G. (2021). *Data analysis for business, economics, and policy* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781108591102

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Client. (2025). *Anonymized dataset of procurement contracts* [Private dataset as Microsoft Excel table].

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*. https://hal.science/hal-03953007

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*. https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Malistov, A., & Trushin, A. (2019). Gradient Boosted Trees with Extrapolation. *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. https://doi.org/10.1109/icmla.2019.00138

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). SciKit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*. https://doi.org/10.5555/1953048.2078195

Ross, D. F. (2015). Procurement and supplier management. In *Distribution Planning and Control* (pp. 531–604). Springer. https://doi.org/10.1007/978-1-4899-7578-2_11

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the Python in Science Conferences*, 92–96. https://doi.org/10.25080/majora-92bf1922-011

Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011

Snowflake Inc. (2025). *Streamlit: A faster way to build and share data apps*. Streamlit. https://streamlit.io/

Truong, C., Oudre, L., & Vayatis, N. (2019). Selective review of offline change point detection methods. *Signal Processing*, *167*, 107299. https://doi.org/10.1016/j.sigpro.2019.107299