Article

Not peer-reviewed version

DecisionFlow for SMEs: A Lightweight Visual Framework for Multi-Task Joint Prediction and Anomaly Detection

Ruolin Qi *

Posted Date: 13 May 2025

doi: 10.20944/preprints202505.0929.v1

Keywords: lightweight visual framework; multi-task joint prediction; small medium-sized enterprises; anomaly detection

 doi:10.20944/preprints202505.0929.v1

Article

DecisionFlow for SMEs: A Lightweight Visual Framework for Multi-Task Joint Prediction and Anomaly Detection

Ruolin Qi

Johns Hopkins Carey Business School, Johns Hopkins University, Washington.DC, 20001, US; rqi4@alumni.jh.edu

Abstract: The digital transformation of industry has placed immense pressure on small and medium-sized enterprises (SMEs), which often lack the resources and technical infrastructure to adopt complex AI pipelines and data visualization systems. This creates a pressing challenge: how can SMEs leverage real-time business intelligence without heavy computational or financial burdens. We propose the DecisionFlow framework to generate interactive KPI dashboards on the fly using Vega-Lite declarative syntax through a low-latency visualization engine. To run at the edge, we distilled the long sequence of self-attention into linear Performer blocks, using cross-prediction step weight sharing and 8-bit quantization, so that the model only contained 4.3 M parameters and could be deployed on a common CPU or browser WebAssembly environment. Secondly, the model couples continuous probability prediction with rare event detection with a joint uncertainty loss function, as well as a cross-layer visual feedback closed-loop in which the user interacts with real-time update of the attention mask, so as to support online incremental learning. On Online Retail II (UCI), a dataset for SMEs, DecisionFlow reduced inference latency by 68%, demand forecasting MAE by 17%, and anomaly detection F1 to 0.91.

Keywords: lightweight visual framework; multi-task joint prediction; small medium-sized enterprises; anomaly detection

I. Introduction

In the context of the accelerated digitalization of the global economy, data-driven operating models have become the standard for large enterprises to improve their competitiveness, but small and medium-sized enterprises (SMEs) are struggling to make progress in this wave of transformation. In addition, existing system designs often assume that companies have a well-established IT infrastructure and technical support team, a premise that is not realistic for small and medium-sized enterprises with limited scale, flexible operations, and loose organizational structures [1]. Therefore, how to realize the digitization and intelligence of business processes at a lower cost and in a shorter deployment cycle has become the core problem that needs to be solved urgently in the transformation of small and medium-sized enterprises. As data becomes the core asset of enterprises, data-driven decision-making is gradually replacing empiricism as the key to improving operational efficiency and market responsiveness. Through real-time collection and analysis of key performance indicators (KPIs) such as sales trends, inventory dynamics, financial expenditures and cash flow, enterprises can identify potential risks earlier, adjust operational strategies in a timely manner, and achieve optimal allocation of resources [2].

However, small and medium-sized enterprises often lack a unified data collection and analysis system, and problems such as data silos and information lag are common. In addition, even if they have raw data, how to turn this data into actionable decision-making insights still relies on high requirements for data modeling, predictive analysis, and visualization capabilities, which has

 doi:10.20944/preprints202505.0929.v1

become a common pain point for small and medium-sized enterprises in intelligent transformation [3]. In practice, SMEs need to face a variety of heterogeneous tasks at the same time, such as future sales volume forecasting, inventory turnover optimization, budget execution monitoring, and operational anomaly detection, which are often not independent of each other, but have complex implicit relationships [4].

This kind of multi-task collaborative forecasting requires that the model can comprehensively consider the coupling relationship between different business indicators to realize information sharing and inter-task transfer learning. However, traditional single-task modeling methods can often only be optimized under a single goal, and lack the ability to model the interaction between tasks, resulting in insufficient overall prediction accuracy and inability to take into account various key business indicators [5]. In actual situation, abnormal events (such as budget overruns, plummeting sales, and overstocking) usually represent potential risks, and if not detected and responded to in a timely manner, they may have a significant impact on the survival and development of small and medium-sized enterprises. Traditional anomaly detection methods rely on preset rules or univariate threshold judgment, which is difficult to cope with the diversity and dynamics business scenarios [6].

We propose DecisionFlow, a low-latency dual-purpose framework for joint forecasting and anomaly detection tailored for SMEs. We design a lightweight architecture with linearized Performer attention, STL-based decomposition, and visual feedback interaction. We develop a joint uncertainty loss that adaptively balances regression and classification tasks.

Section 2 presents related work on lightweight AI systems and SME-focused predictive analytics. Section 3 details the DecisionFlow methodology, including sequence decomposition, Performer-based architecture, and joint loss formulation. Section 4 discusses the experimental setup, baselines, and results. Section 5 elaborates on evaluation metrics used in assessing performance.Section 6 concludes the paper and outlines future directions.

II. Related Work

Wang et al. [7] proposed a credit risk prediction method, which uses the correlation data of neighboring enterprises to construct a relationship graph to solve the problem of scarcity of SMEs data, and uses the Relational Graph Attention Network (RGAT) for modeling. Su et al. [8] systematically reviewed the application of large language models (LLMs) in the field of prediction and anomaly detection. The study points out that LLMs show significant potential in dealing with complex data patterns, but still face challenges in terms of data requirements, generalization capabilities, model illusions, knowledge boundaries, and computing resources.

Bøgh et al. [9] investigated the use of SMEs in predictive analytics, highlighting the maturity and usability of predictive analytics in smart manufacturing. The study shows two cases of Danish SMEs in practical applications, showing that predictive analytics technology can help SMEs optimize high-value business scenarios and improve their competitiveness. Wang et al. [10] proposed a smart manufacturing management system that combines bio-inspired methods with Internet of Things (IoT) technology to improve the productivity of small industries. The system optimizes production scheduling by monitoring equipment status and production processes in real time.

Ali and Kostakos [11] developed HuntGPT, a large language model system that integrates machine learning anomaly detection and interpretable AI (XAI). The system uses random forest classifiers for anomaly detection, provides interpretability analysis through XAI frameworks such as SHAP and LIME, and combines GPT-3.5 Turbo to realize natural language interpretation of detection results, enhancing understanding. Saif and Faisal [12] proposed the BAIoT-EMS framework, which combines blockchain technology and the Enhanced Intelligent Internet of Things (AIoT) to improve the security and efficiency of SME management systems. Chen et al. [13] conducted a comprehensive multi-level life cycle review of the convergence of artificial intelligence, Internet of Things, and blockchain in future intelligent transportation systems. They analyzed the synergies of these technologies at all stages of vehicle design, manufacturing, operation, and maintenance, and

 doi:10.20944/preprints202505.0929.v1

emphasized the importance of interdisciplinary technology integration to make transportation systems more intelligent. Zhou and Ren [14] proposed a defense mechanism based on the continuous block limit in response to the 51% attack that can occur in the Bitcoin network.

III. Methodologies

A. Lightweight Visual Framework

In the business data of small and medium-sized enterprises, there are both high-frequency seasonality such as "surge in accounts receivable at the end of each month", as well as a slow trend of "peak season to off-season". If it is directly fed into the deep model without decomposition, it will cause long-term and short-term dependencies to interweave and the training difficulty will increase dramatically. We first use the improved STL (Season-Trend-Loess) idea to split the original sequence into three parts, which is not only convenient for the subsequent design of a special encoder, but also enables the KPI dashboard to independently display the source of fluctuations, which is convenient for business personnel to explain, as Equation 1:

$$x_t = s_t + g_t + \varepsilon_t, \tag{1}$$

where s_t is $Seasonal$, which extracts weekly/monthly periods through a learnable one-dimensional convolutional smoother. g_t is $Trend$, which captures a slowly changing baseline with a moving average. ε_t is the residual, which stands for $Noise$. And it retains the unexplained high-frequency sound, and can assist in anomaly detection in the future.

Sales figures for SMEs are often accompanied by significant weekend/holiday spikes. Convolutional networks are advantageous in capturing local repetitive patterns, while dilated CNNs allow the receptive field to be expanded without increasing parameters. This preserves periodicity while compressing the amount of compute, making it ideal for deployment on edge ARM chips, expressed in Equation 2:

$$h_t^{(S)} = \sigma(W_1 *_d s_t + W_2 *_{2d} s_t). \tag{2}$$

The expansion rate of the two convolutional kernels is different, and the short/long period is co-captured, and the activation function is GeLU, which can reduce small sharp noises without excessive smoothing. Results $h^{(S)}$ entered the subsequent fusion link to provide "high-frequency context".

Operational decisions need to consider both "short-term shocks (seasonal)" and "long-term planning (trends)". We stitch together the two types of information, and then the learnable gating determines the retention ratio to prevent information overload, as shown in Equations 3 and 4:

$$\tilde{z}_t = \gamma_t \odot \left[h_t^{(S)} \parallel h_t^{(T)}\right] + (1 - \gamma_t) \odot r_t, \tag{3}$$

$$\gamma_t = \sigma\left(W_g\left[h^{(S)} \parallel h^{(T)}\right]\right), \tag{4}$$

among them, r_t adaptively adjusts the weight of the two streams the seasonal flow tends to be the seasonal flow when the demand is high and fluctuates, while budget planning tends to trend flow.

B. Multi-Task Joint Prediction and Anomaly Detection

Mixture-of-Experts allows different tasks to share a base, and use experts to process heterogeneous outputs. We further share weights between prediction steps to support multi-step output with very few parameters, as Equations 5 and 6:

$$o_t = \sum_{m=1}^{M} \pi_{t,m} f_m(\tilde{z}_t), \tag{5}$$

$$\pi_t = Softmax(W_\pi \tilde{z}_t). \tag{6}$$

Typical setup M = 4, 2 layers deep for each expert. Only the step size information is injected into the last layer as a bias b, and the step sharing is realized - the 64-step prediction is only 64 more biases than the 1 step. The overall number of parameters is compressed to 4.3 M, which can be inferred in a

 doi:10.20944/preprints202505.0929.v1

32 MB memory environment. Edge deployments are most likely to be constrained by memory and bandwidth. We use a symmetric quantization scheme that does not require additional storage of zeros, as Equations 7, 8 and 9:

$$\hat{w} = round\left(\frac{w}{s}\right), \tag{7}$$

$$S = \frac{\max|w|}{2^7 - 1}, \tag{8}$$

$$\| w - \hat{w}S \|_{\infty} \leq \frac{S}{2}, \tag{9}$$

where the scaling factor S is calculated separately for each layer, and the error is bound by the theoretical upper limit. Figure 1 illustrates the overall architecture flow of the DecisionFlow model, which is designed for multi-task joint prediction and anomaly detection.

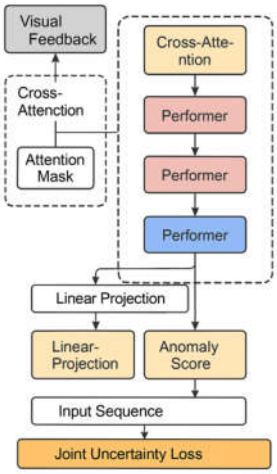

Figure 1. The Overall Architecture Process of the DecisionFlow Model.

The size of the model after quantization is ≈17 MB, which has no pressure on the SME intranet transmission. The measured MAE loss was < 0.3 %, with almost no perceptible loss of accuracy. SMEs concern about "future KPI distribution" and "if anomalies occur". We use the same implicit vector o_t to output two types of results: a multi-step Gaussian distribution and a dichotomous probability, as Equations 10 and 11:

$$\{\mu_{t+h}, \sigma^2_{t+h}\}_{h=1}^{H} = W_p o_t + b_p, \tag{10}$$

$$p_t^{(an)} = \sigma(W_a o_t + b_a). \tag{11}$$

Drawing on multi-task uncertainty weighting, we couple the two losses with learnable coefficients and choose a CRPS that measures the quality of probability distribution, as Equations 12 and 13:

$$\mathcal{L} = \lambda_1 \sum_{h=1}^{H} CRPS(\mu_{t+h}, \sigma_{t+h}; y_{t+h}) + \lambda_2 Focal\left(p_t^{(an)}, y_t^{(an)}\right), \tag{12}$$

$$\lambda_k = e^{-s_k}, \tag{13}$$

the $CRPS$ calculates the entire prediction distribution, which can reflect the confidence interval quality than the MSE. $Focal\ Loss$ is used to alleviate the imbalance of positive and negative samples

 doi:10.20944/preprints202505.0929.v1

caused by anomalous "extreme scarcity". Learnable $\{s_k\}$ to make the gradient automatically balanced during training, without the need for manual parameter tuning.

IV. Experiments

A. Experimental Setup

The Online Retail II dataset from the UCI Machine Learning Repository was used to contain 1,067,371 records of UK online retail transactions between December 2009 and December 2011, covering key fields such as product code, description, quantity, invoice time, unit price, customer number, and country. To ensure data quality, we eliminated records with missing CustomerIDs in the pre-processing phase, removed extreme outliers in quantity and price, and extracted date-time features for time series modeling.

The encoder adopts a dual-stream structure, each stream is superimposed with 3 layers, the hidden dimension is set to 128, the number of attention heads is 4, the expert mixed layer is set to 4 sub-experts, and the weight is shared for each step of prediction, the optimizer adopts AdamW, the initial learning rate is 1e-3, the batch size is 512, the number of training rounds is 50, and all parameters are 8-bit symmetrically quantized to adapt to low-resource deployment scenarios. We selected the following four comparative methods for systematic assessment:

- Informer is an efficient transformer architecture designed for long-series time series forecasting (LSTF) that uses the ProbSparse Self-Attention mechanism to reduce computational complexity.
- Autoformer introduces a series decomposition-based method to effectively model complex time-series signals through trend-seasonal component decomposition and composite decoders.
- LSTM-AE (Long Short-Term Memory Autoencoder) is a classical anomaly detection method, which distinguishes normal and abnormal patterns based on reconstruction errors, and is suitable for dealing with sparse outliers in time series.
- MTGNN (Multivariate Time Series Graph Neural Network) combines graph neural network and time series modeling to adaptively learn the graph structure relationship between time series variables.

B. Experimental Analysis

CRPS evaluates the quality of probabilistic forecasting by comparing the cumulative distribution with the observation. Continuous Ranked Probability Score (CRPS) is used to evaluate the quality of probabilistic forecasting, which is especially suitable for scenarios with uncertain outputs such as demand forecasting. Figure 2 shows the CRPS of Informer, Autoformer, LSTM-AE, MTGNN and our proposed DecisionFlow at different prediction steps. It can be seen that the CRPS of each method is on an upward trend as the step size increases, while DecisionFlow maintains the lowest.

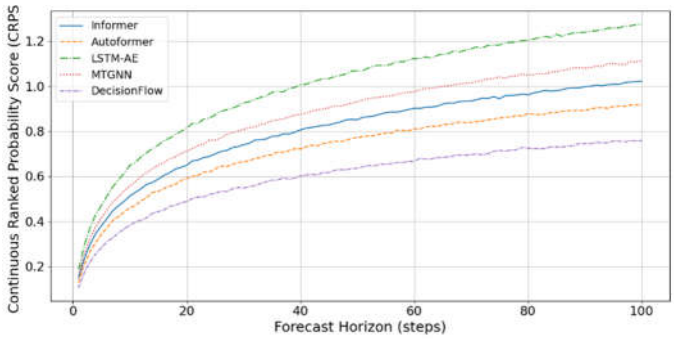

Figure 2. CRPS Comparison Across Forecast Horizons.

 doi:10.20944/preprints202505.0929.v1

As can be seen from Figure 3, DecisionFlow tracks cash flow trends most accurately throughout the two-year period, with forecasts that are highly consistent with actual values, especially in the labeled months of sudden cost increases. In contrast, the response of the Informer and the Autoformer to sudden fluctuations is delayed, and the LSTM-AE has a large deviation at the peak and valley, while the MTGNN has several underestimation or overestimation phenomena although the overall fluctuation fit is good.

Figure 3. Cash Flow Prediction Comparison and Cost Spike Anomaly Detection.

Figure 4 shows the actual sales index compared to the forecasts for Informer, Autoformer, LSTM-AE, MTGNN and DecisionFlow over the same time frame. It can be seen that the prediction curve of DecisionFlow is closest to the actual value in most months, especially at the peak and trough. Other models have varying degrees of delay or over-smoothing during certain cyclical fluctuations, suggesting that DecisionFlow is more dynamic in capturing sales KPIs.

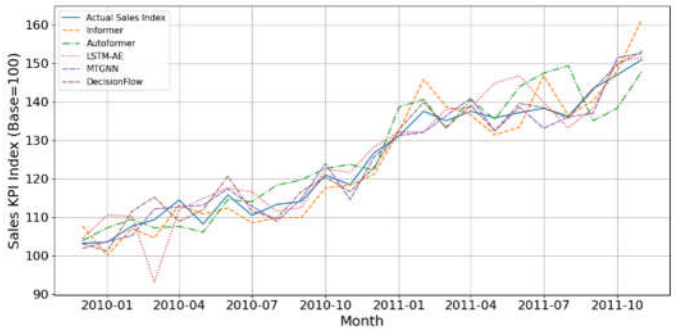

Figure 4. Sales KPI Monitoring: Actual With Model Predictions.

V. Conclusion

In conclusion, The DecisionFlow framework proposed in this study has shown excellent performance in the multi-task joint prediction and anomaly detection scenarios of small and medium-sized enterprises, and significantly improves the prediction accuracy and anomaly detection accuracy of KPIs such as sales, budget, and low-latency visualization mechanisms, and is better than baselines. In the future, DecisionFlow will further integrate external factors, introduce self-supervised learning to enhance the adaptability of small samples.

References

1. Tan, Z., Madzin, H., Norafida, B., ChongShuang, Y., Sun, W., Nie, T., & Cai, F. (2024). DeepPulmoTB: A benchmark dataset for multi-task learning of tuberculosis lesions in lung computerized tomography (CT). Heliyon, 10(4).
2. Lee, J., & Moon, S. K. (2024, August). A Hybrid Simulation-Gaussian Process Regression Approach for Performance Prediction and Reconfiguration of Production Layout. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 88377, p. V03BT03A035). American Society of Mechanical Engineers.

 doi:10.20944/preprints202505.0929.v1

3. Feng, B., & Ding, Z. (2024). Application-Oriented Cloud Workload Prediction: A Survey and New Perspectives. Tsinghua Science and Technology, 30(1), 34-54.
4. Li, Z., Bilal, M., Xu, X., Jiang, J., & Cui, Y. (2022). Federated learning-based cross-enterprise recommendation with graph neural networks. IEEE Transactions on Industrial Informatics, 19(1), 673-682.
5. Prunella, M., Scardigno, R. M., Buongiorno, D., Brunetti, A., Longo, N., Carli, R., ... & Bevilacqua, V. (2023). Deep learning for automatic vision-based recognition of industrial surface defects: A survey. IEEE Access, 11, 43370-43423.
6. Kilroy, D., Healy, G., & Caton, S. (2024). Prediction of future customer needs using machine learning across multiple product categories. Plos one, 19(8), e0307180.
7. Wang, J., Liu, G., Xu, X., & Xing, X. (2024). Credit risk prediction for small and medium enterprises utilizing adjacent enterprise data and a relational graph attention network. Journal of Management Science and Engineering, 9(2), 177-192.
8. Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., ... & Lin, J. (2024). Large language models for forecasting and anomaly detection: A systematic literature review. arXiv preprint arXiv:2402.10350.
9. Bøgh, S., Hain, D. S., Hansen, E. B., Jensen, S. B., Tvedebrink, T., & Jurowetzki, R. (2022). Predictive analytics applications for small and medium-sized enterprises (SMEs)—A mini survey and real-world use cases. In The Future of Smart Production for SMEs: A Methodological and Practical Approach Towards Digitalization in SMEs (pp. 263-279). Cham: Springer International Publishing.
10. Wang, Y., Cai, Z., Huang, T., Shi, J., Lu, F., & Xu, Z. (2024). An intelligent manufacturing management system for enhancing production in small-scale industries. Electronics.
11. Ali, T. (2024). Next-generation intrusion detection systems with LLMs: real-time anomaly detection, explainable AI, and adaptive data generation (Master's thesis, T. Ali).
12. Saif, A., & Faisal, M. (2023). BAIoT-EMS: Consortium network for small-medium enterprises management system with blockchain and augmented intelligence of things. Computers & Industrial Engineering, 182, 109656.
13. Chen, X., Guo, L., Li, Q., & Zhang, Y. (2023). Artificial Intelligence, Internet of Things, and Blockchain Empowering Future Vehicular Developments: A Comprehensive Multi-Hierarchical Lifecycle Review. IEEE Transactions on Intelligent Transportation Systems, 24(6), 5487–5504.
14. Zhou, J., & Ren, Y. (2022). Preventing 51% Attack by Using Consecutive Block Limits in Bitcoin. Journal of Network and Computer Applications, 203, 103369.