Apu Datta | MS Business Analytics, Baruch College

**ERP Spend Forecasting:**

AI-Driven Forecasting with Real-Time Market Insights

**Business Problem Definition and Importance:**

As Factory CFO, I lived through constant stress every month, our procurement costs would spike unexpectedly. Our SAP system was great at recording what happened, but terrible at telling us what would happen next. The simple math we relied on (quantity times price) completely ignored the messy realities of doing business - market swings, seasonal rushes, and supply chain hiccups that always seemed to catch us off guard. When these surprises hit, I'd have to scramble for last-minute approvals from our Head Office, which slowed everything down and made us look unprepared. I built this AI system because I was tired of playing defense. Instead of constantly reacting to cost shocks, we need tools that can see them coming and help us plan smarter.

**Data source and preprocessing steps:**

I worked with procurement dataset from Kaggle containing 800+ ERP transaction records including purchase orders, suppliers, item categories, dates, quantities, and prices. My preprocessing approach focused on practical data quality improvements: removing **invalid entries with zero** or **negative values**, applying **Z-score outlier detection** and using **category-specific mean** imputation to handle missing data intelligently. I engineered meaningful business features including monthly spend calculations (quantity × price), lead time extraction from delivery intervals, and realistic department assignments across Production, Logistics, Procurement, Finance, and Maintenance using reproducible random allocation. The breakthrough enhancement was integrating real-time market data through the Federal Reserve Economic Data (FRED) API, where I mapped our ten procurement categories to specific Producer Price Index.

Data Source: <https://www.kaggle.com/datasets/shahriarkabir/procurement-kpi-analysis-dataset>

**Model Selection Process and Results:**

I tested five different algorithms to find the best predictor: Linear Regression as my baseline, plus Ridge, Lasso, Random Forest, and XGBoost. I split the data 80/20 for training and testing, then used 5-fold cross-validation to make sure each model was robust. Since linear models work better with normalized data, I applied Standard Scaler only to Linear, Ridge, and Lasso - letting the tree-based models work with raw features as they prefer. I focused on R² score as my main success metric, backed up by MAE and RMSE to get the full picture of performance. **RandomForest emerged as the clear winner with R² = 0.9966, MAE = $1,781, and RMSE = $2,734**. Cross-validation showed excellent stability with CV R² = 0.9684, confirming the model's reliability across different data splits. Feature importance analysis revealed quantity as the dominant cost driver (approximately 50%), followed by negotiated price (35%), PPI market conditions (10%), and lead time (5%).

**Key Insights and Recommendations:**

My analysis revealed why our procurement budgets kept missing targets - real costs deviate 8-12% from simple calculations due to market realities we weren't tracking. **Key findings:** PPI swings over ±3% signal incoming cost spikes requiring immediate supplier validation; rush orders under 7 days cost us 5-8% premiums while 90+ day planning saves 2-3%; seasonal patterns hit hard with Electronics spiking 5% in Q4 and Raw Materials climbing 3% in spring/fall; volatile categories like Metals and Chemicals need 15-20% contingency buffers. **My recommendations:** Integrate this AI system into existing ERP dashboards for real-time alerts, set automated risk thresholds by category, build supplier scorecards tracking lead time performance, and switch from static budgets to dynamic models adjusting for seasonal patterns. Time contract negotiations during low-volatility periods, maintain backup suppliers for high-risk categories, and use the system's confidence intervals for scenario planning instead of hoping for the best.

**Limitations and Future Improvements:**

Current limitations include dependency on historical patterns that may not capture unprecedented market disruptions, limited supplier-specific performance data, and potential model degradation during extreme economic events. The model assumes consistent supplier behavior and may not fully account for geopolitical factors affecting global supply chains. Future enhancements should incorporate: supplier performance metrics including quality scores, delivery reliability, and financial stability indicators; geopolitical risk indices and commodity futures data for enhanced market intelligence; advanced time-series forecasting using LSTM/GRU networks for long-term trend prediction; multi-objective optimization balancing cost, quality, and delivery reliability; integration with additional external data sources including weather patterns for agricultural commodities and shipping rates for international procurement. Additionally, implementing online learning capabilities would enable continuous model adaptation to evolving market conditions, while expanding the dataset to include supplier-specific historical performance would improve prediction accuracy and enable more granular risk assessment at the vendor level.