**Apu Datta**

ERP Spend Forecasting Using ML

AI Driven Forecasting with Real Time Market Insights

**Data Mining and Visualization**

In this project, I built a comprehensive machine learning system to predict future procurement spending based on quantity, negotiated prices, lead times, and real time market conditions. I used historical ERP procurement data integrated with Federal Reserve Economic Data (FRED) API to capture live market indicators including Producer Price Index (PPI), volatility measures, and economic uncertainty factors.

**Methodology:**

For this project, I have developed an AI powered forecasting system to predict procurement costs by blending machine learning with real world business intelligence. The process began with data cleaning and preparation, including outlier removal using the Z score method, filling missing values based on item categories, and enriching the dataset with real time inflation data (PPI) from **10 relevant FRED API** series mapped to different procurement categories. I have split the data into training **(80%)** and testing **(20%)** sets and tested five regression models: **Linear Regression, Ridge, Lasso, Random Forest, and XGBoost**. To ensure a fair comparison, I standardized features for the linear models while keeping raw values for tree based models. Model performance was evaluated using **5 fold cross validation**, and results were compared based on **R², MAE,** and **RMSE** to select the best performing model.

**Model Evaluation:**

I have conducted 5 fold cross validation across all regression algorithms to measure performance. The Random Forest Regressor outperformed all others, explaining 99.66% of the variation in procurement costs (R² = 0.9966) with an exceptionally low MAE of $1,781 and RMSE of $2,734. Cross validation confirmed its stability (CV R² = 0.9684 ± 0.0283), making it highly reliable. Feature importance analysis showed that Quantity (58.61%) and Negotiated Price (40.84%) were the dominant cost drivers, while PPI (0.32%) and Lead Time (0.23%) had minimal direct impact. This suggests procurement costs in this dataset are primarily driven by order volume and negotiated pricing.

The finalized Random Forest model was serialized using Joblib for deployment and paired with additional business logic to incorporate market risk analysis, seasonal trend adjustments, and forecast uncertainty ranges. This allows decision makers to interpret results with confidence intervals at 68%, 95%, and 99%, enabling more informed procurement strategies.

**Deployment:**

The final system is live on Streamlit Cloud, offering 24/7 access through a user friendly web app. It fetches real time data from the FRED API, provides interactive predictions, and features business dashboards with model insights, market risk analysis, and seasonal trends.

<https://erp-spend-forecasting-app-with-app-and-ml-pipeline-p9zrfxmjz44.streamlit.app/>

**Real World Application:**

This system helps procure teams forecast budgets with ±12% accuracy, optimize order timing, and anticipate cost or supply risks. It supports smarter negotiations, real time cost planning, and seasonal budgeting driving strategic, data informed decisions.

**Limitations and Disclaimers:**

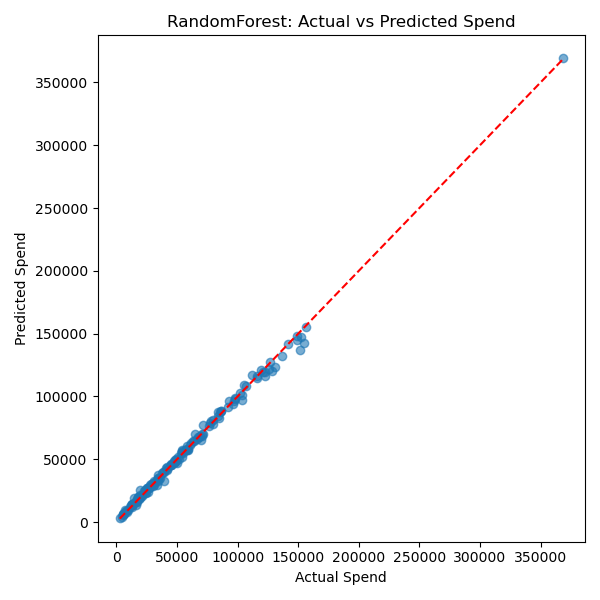
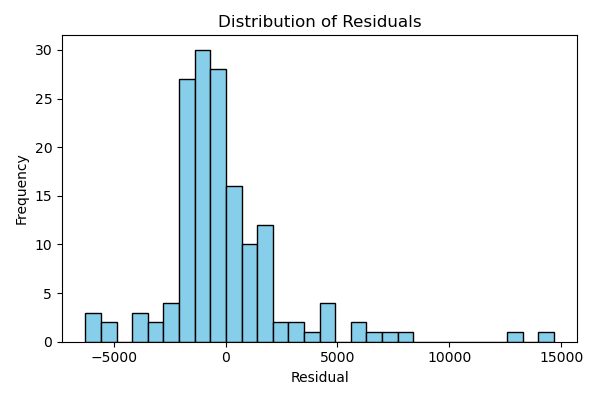
**Limited Training Scope:** Model trained on less than 800 transactions; maybe less accurate for rare items or unusual order sizes. **Optimal Range:** Best results for orders between 50 to5,000 units. Larger orders need manual review. **Market Data Dependency:** Uses real time PPI from FRED API forecasts may vary if the API is unavailable. **Learned Volume Discounts:** Model assumes 20 to 35% discounts for bulk orders based on past data, which may not apply to current vendor deals. **Hybrid Logic:** Forecast combines ML with rule based business logic (e.g., risk, seasonality). **Use with Caution:** Flag high price, emergency, or off pattern orders for manual validation.

**Appendix: Visualizations**

1. **Feature\_importance\_plot.png:**

A graph with a bar and text

AI-generated content may be incorrect.

1. **Predicted\_vs\_actual.png:**
2. **Residuals\_histogram.png**
3. **Residuals\_vs\_actual**

