Project Title: Forecasting Next Month Booked Hours per Cost Center using Machine Learning

1. Problem Statement

In the manufacturing organization, hours are manually booked daily against cost centers in ERP. Finance and operations want to know how many hours each cost center will likely consume next month for budgeting, cost control, and capacity planning. The ML goal is to predict total booked hours for each cost center for the next month using historical data.

2. Business to ML Conversion

This is a regression problem where the input is historical monthly data and the output is a numeric value representing next month's booked hours.

3. Data Description

Source: ERP booking data

Columns: date, plant, cost_center, booked_hours

Nature: Manual, noisy, human-entered, highly variable, represents organizational behavior rather than machine physics.

4. Data Preparation

Daily data is aggregated to monthly data per plant and cost center. This is done because business planning happens monthly and daily data is too noisy. Monthly aggregation smooths random fluctuations and highlights real trends and seasonality.

5. Feature Engineering

Lag features: last_month_hours, 3_months_ago_hours

Rolling features: last_3_month_avg, last_6_month_avg

Seasonality: month_number (1–12)

Target: next_month_total_hours

6. Train-Test Strategy

Time-based split is used: older months for training and recent months for testing. This avoids data leakage and matches real-world forecasting behavior.

7. Model Choice

RandomForest Regressor is used because the data is noisy, non-linear, and tabular. RandomForest is robust, handles outliers well, does not overfit easily, and requires minimal preprocessing.

8. Training and Learning

The model learns patterns like rising trends and seasonality effects based on historical behavior.

9. Evaluation

Metrics used: MAE and RMSE. Also compared against a naive baseline where next month equals last month. If ML is not better than naive, it has no business value.

10. Results and Observations

Observed high variance in data. Standard deviation is almost as large as the mean. Prediction error is around 30–40%. This indicates the underlying process is highly volatile.

11. Key Business Insight

The main limitation is not the model, but the instability of the business process itself. Manual booking, changing priorities, and special jobs make the process weakly predictable.

12. What Was Achieved

Built full ML pipeline including data aggregation, feature engineering, model training, evaluation, and prediction. Also derived an important business insight about process instability.

13. Interview-Ready Summary

“I built a time-series forecasting system for cost center workload using historical ERP booking data. The project included data aggregation, feature engineering, RandomForest regression, and error analysis. The key insight was that the underlying process is highly volatile, which limits predictability — an important business finding in itself.”

14. Key Learnings as an AI Engineer

1. Data quality and stability matter more than algorithms.

2. ML cannot create patterns if they do not exist.

3. Feature engineering is more important than model choice.

4. Always compare against a baseline.

5. Sometimes the right conclusion is that the process is not predictable enough.

15. Final Conclusion

This project demonstrates real-world AI engineering: not just building models, but understanding data, business processes, and their limitations.