Intelligent Blood Supply Management

Demand Forecasting, Donor Segmentation & Recommender System for Inventory Optimization

Key Results Achieved

12.14%	0.41	+47%
FORECAST MAPE (Target: <20%)	SILHOUETTE SCORE (Target: >0.35)	RECOMMENDER IMPROVEMENT

1. Problem Background & Motivation

1.1 The Blood Supply Challenge

Blood transfusion is a critical life-saving medical intervention with no synthetic substitute available. Blood banks worldwide face a fundamental operational dilemma that impacts both patient outcomes and resource efficiency:

- **Shortage Risk:** Unpredictable demand spikes from accidents, surgeries, and medical emergencies, combined with variable donor availability, can lead to critical shortages that directly impact patient outcomes and potentially cause loss of life.
- **Wastage Risk:** Blood components are highly perishable—platelets expire in just 5 days, while packed red blood cells last 42 days. Overstocking leads to expensive wastage of a precious, voluntarily donated resource.

1.2 Scale of the Problem

According to the **2019 National Blood Collection and Utilization Survey (NBCUS)**:

- **13.02 million** donors presented for donation in the United States
- **10.8 million** RBC units were transfused to **4.2 million** recipients
- Wastage rates range from **5-15%** depending on component and facility
- Young donor participation (16-24 years) declined by 10.1%, raising sustainability concerns
- Donors aged 65+ increased by 10.5%, creating long-term supply challenges

The **Nepal BPKIHS study (2023)** documented **7.13% overall wastage** with 92.9% utilization rate, establishing a benchmark for inventory performance. Platelet concentrates showed the highest wastage rates due to their extremely short shelf life.

1.3 Project Objectives

Objective	Technique	Target Metric
Demand Forecasting	SARIMA, Prophet, Random Forest, Gradient Boosting	MAPE < 20%
Donor Segmentation	RFM Analysis + K-Means Clustering	Silhouette Score > 0.35
Recommender System	Hybrid Rule-Based Scoring	Outperform random selection
Anomaly Detection	Rolling Z-score	Detect demand spikes

2. Dataset Description

2.1 Data Sources Overview

Dataset	Records	Type	Source
UCI Blood Transfusion	748	Real	UCI ML Repository
Kaggle Blood Demand	-	Real	Kaggle
Kaggle Blood Donor	-	Real	Kaggle
Synthetic Donor Registry	10,000	Synthetic	NBCUS 2019 statistics
Demand Time-Series	4,380	Synthetic	3 years daily data
Supply/Inventory	4,380	Synthetic	Nepal BPKIHS study

2.2 UCI Blood Transfusion Dataset (Real Data)

This dataset from the Blood Transfusion Service Center in Taiwan contains donation behavior of 748 donors:

Feature	Description	Range
Recency	Months since last donation	0.03 - 74.4
Frequency	Total number of donations	1 - 50
Monetary	Total blood donated (cc)	250 - 12,500
Time	Months since first donation	2.27 - 98.3
Donated	Target: donated in March 2007	Binary (24% positive)

2.3 Synthetic Data Generation

Synthetic datasets were generated based on real-world statistics to ensure realistic patterns:

- **Blood Types:** WHO distribution (O+: 37%, A+: 28%, B+: 20%, AB+: 5%, O-: 6%, A-: 6%, B-: 2%, AB-: 1%)
- **Demographics:** NBCUS 2019 patterns (30% first-time donors, 70% repeat donors; 54% male, 46% female)
- **Deferrals:** 15% deferral rate (top reason: low hemoglobin at 43%)
- **Demand Patterns:** Weekly seasonality, winter peaks, holiday spikes, random emergency events

3. Methodology & Model Choice

3.1 Demand Forecasting Models

Model	Rationale	Configuration
SARIMA	Explicit seasonality modeling with confidence intervals	SARIMA(1,1,1)(1,1,1,7)
Prophet	Automatic seasonality detection, holiday handling	Weekly + yearly seasonality
Random Forest	Captures non-linear relationships, feature importance	100 trees, max_depth=10
Gradient Boosting	Sequential error correction, handles complex patterns	100 estimators, lr=0.1

Feature Engineering:

- **Time features:** day_of_week, month, quarter, is_weekend, is_holiday_season
- **Lag features:** lag_1, lag_7, lag_14, lag_30 (previous demand values)
- **Rolling statistics:** rolling_mean_7/14/30, rolling_std_7/14/30

3.2 Donor Segmentation Approach

RFM Framework is a proven marketing analytics technique adapted for blood donation:

- **R**ecency: Months since last donation (lower = better, more engaged)
- **F**requency: Total lifetime donations (higher = better, more committed)
- **M**onetary: Total blood volume donated in cc (higher = better, greater contribution)

Scoring: Quintile-based scoring (1-5 scale) for each RFM metric, combined to create donor value segments.

Clustering: K-Means algorithm with optimal K selection via elbow method and silhouette analysis. Target segments: Champions, Loyal, Potential, At Risk, Hibernating.

3.3 Hybrid Recommender System

A rule-based hybrid system that matches donors to blood donation campaigns using weighted scoring:

Component Details:

- **RFM Score (40%):** Normalized donor value score from segmentation (0-1)
- **Availability (20%):** Eligibility based on last donation date (56-day minimum gap)

- **Segment Match (20%):** Campaign-segment alignment (e.g., reactivation campaigns → At Risk donors)
- **Blood Match (10%):** Priority for needed blood types (O- for emergencies)
- **Urgency Boost (10%):** Higher weight for critical shortage campaigns

3.4 Anomaly Detection

Method: Rolling Z-score with 30-day window and threshold of 2.5σ

This method detects unusual demand spikes from emergencies, seasonal surges, or operational disruptions, enabling proactive inventory management.

4. Experiments & Results

4.1 Demand Forecasting Performance

Test Setup: 90-day holdout period (last 3 months of data)

Model	MAE	RMSE	MAPE	R^2
Random Forest	**19.68**	**29.61**	**12.14%**	**0.45**
SARIMA	22.19	35.65	12.40%	0.20
Gradient Boosting	22.11	33.91	13.75%	0.28

> **Key Finding:** All models achieved MAPE below the 20% target. Random Forest performed best with 12.14% MAPE and highest R^2 of 0.45, indicating it captures demand patterns most effectively.

Top 5 Important Features:

1. lag_1 (0.42) - Previous day demand is the strongest predictor
2. rolling_mean_7 (0.18) - Weekly average captures short-term trends
3. lag_7 (0.12) - Same day last week shows weekly patterns
4. day_of_week (0.08) - Weekend vs weekday effect
5. rolling_std_7 (0.05) - Volatility indicator

Time Series Patterns Identified:

- **Weekly cycle:** 30% lower demand on weekends (reduced elective surgeries)
- **Seasonal trend:** +15% demand in winter months (December-February)
- **Holiday spikes:** +20-25% during major holidays (Diwali, Christmas)
- **Random spikes:** Emergency events visible in residuals

4.2 Donor Segmentation Results

Clustering Performance: Silhouette Score = **0.41** (Target: >0.35) ✓

Segment	Count	%	Avg Recency	Avg Frequency	Strategy
Champions	1,151	11.5%	1.9 months	8.2	Retain & Reward
Loyal	2,234	22.3%	3.8 months	5.7	Referral Programs
Potential	2,847	28.5%	4.2 months	2.1	Nurture & Convert
At Risk	1,956	19.6%	8.4 months	4.3	Urgent Reactivation

Hibernating	1,812	18.1%	12.1 months	1.8	Last-chance Campaign

4.3 Recommender System Results

Campaigns Generated:

Campaign	Target Segment	Blood Priority	Urgency
Winter Blood Drive	Champions, Loyal	All types	High
O-Negative Emergency	All eligible O- donors	O-	Critical
Platelet Donation Week	Champions	A+, O+	High
Summer Stock Building	Potential	All types	Medium
Lapsed Donor Reactivation	At Risk, Hibernating	All types	Medium

Simulation Results (100 donors, 5 campaigns, 10 simulations):

Strategy	Avg Donations	Improvement
Random Selection	~45	baseline
RFM-based Selection	~58	+29%
Hybrid Recommender	**~66**	**+47%**

> **Key Finding:** The hybrid recommender system achieved **47% improvement** over random donor selection, demonstrating the value of intelligent donor-campaign matching.

4.4 Anomaly Detection Results

The rolling Z-score method detected **12 significant anomalies** over the 3-year period, corresponding to:

- Winter demand surges (December peaks)
- Holiday-related spikes (festival seasons)
- Simulated emergency events

5. Insights & Business Interpretation

5.1 Demand Pattern Insights

Pattern	Finding	Business Action
Weekly	30% lower demand on weekends	Reduce weekend collection drives, optimize staff scheduling
Winter Peak	+15% demand (Dec-Feb)	Pre-stock 15-20% buffer before November
Holiday Spikes	+20-25% during festivals	Build emergency reserves, organize pre-holiday drives
Summer Dip	-8% (Jun-Aug)	Intensify donor drives, target corporate donors

5.2 Donor Engagement Strategy

Segment	Strategy	Channel	Expected Yield
Champions	Retain & Reward	Personal calls, VIP events	90%+ retention
Loyal	Referral programs	Email, SMS campaigns	70-80% engagement
Potential	Nurture & Convert	Education campaigns	40-50% activation
At Risk	Urgent reactivation	Personal outreach	30-40% win-back
Hibernating	Last-chance campaign	Direct mail	10-20% reactivation

5.3 Inventory Optimization Recommendations

Component	Current Wastage	Shelf Life	Recommended Strategy
Platelets	12.3%	5 days	Just-in-time ordering, max 2-day advance
Packed RBC	5.2%	42 days	Maintain 7-day rolling stock buffer
Fresh Frozen Plasma	6.4%	1 year	Moderate buffer acceptable
O-negative	-	-	Priority collection (universal donor, only 6%)

5.4 Projected Business Impact

Initiative	Current State	Target	Projected Impact
Wastage Rate	8.8%	5%	~$150K/year savings

Donor Outreach	Random selection	Recommender system	+47% donation yield
Forecast Accuracy	None	MAPE <15%	Reduced emergency procurement
Donor Retention	~70%	85%	15% more reliable supply

6. Limitations & Future Scope

6.1 Limitations

- **Synthetic Data:** While based on real statistics (NBCUS, BPKIHS, WHO), synthetic data may not capture all real-world complexities. Validation with production hospital data is required before deployment.
- **Unpredictable Events:** Emergency events (natural disasters, mass casualty incidents) cannot be forecast by any statistical model.
- **External Factors:** Pandemics, policy changes, economic conditions, and social factors are not modeled.
- **Implementation Challenges:** Integration with existing hospital information systems, staff training requirements, and regulatory compliance (HIPAA, local regulations) present practical challenges.

6.2 Future Scope

Area	Opportunity
Deep Learning	LSTM/Transformer models for improved sequence modeling
Reinforcement Learning	Dynamic inventory policies that adapt to changing conditions
Real-Time Systems	IoT cold-chain monitoring, hospital EMR integration
Personalization	Individual donor scheduling optimization, mobile app engagement
Multi-Center Network	Regional network optimization across multiple blood banks

7. Conclusion

This project demonstrates the successful application of advanced machine learning techniques to blood supply management:

Objective	Target	Achieved	Status
Demand Forecasting	MAPE < 20%	**12.14%**	✓ Exceeded
Donor Segmentation	Silhouette > 0.35	**0.41**	✓ Exceeded

Recommender System	Beat random	**+47%**	✓ Exceeded

Key Takeaways:

- Random Forest achieved best forecasting performance with 12.14% MAPE
- RFM-based clustering successfully identified 5 actionable donor segments
- Hybrid recommender improved donor outreach yield by 47% over random selection
- Weekly (30% weekend dip) and seasonal (15% winter peak) patterns enable proactive inventory planning
- Platelets require just-in-time management due to 5-day shelf life

> *"Advanced ML is not about prediction accuracy alone. It is about discovering structure, extracting insights, and supporting decision-making in complex, unlabeled, or temporal data."*

References

1. NBCUS (2019). National Blood Collection and Utilization Survey. US Department of Health and Human Services.
2. Singh, M., Pradhan, A., & Poudyal, P. (2023). Blood component usage and wastage at BPKIHS. Journal of Pathology of Nepal, 13(1).
3. UCI Machine Learning Repository. Blood Transfusion Service Center Dataset. https://archive.ics.uci.edu/dataset/176
4. World Health Organization. Blood Safety and Availability Fact Sheet.
5. Kaggle. Blood Demand Dataset & Blood Donor Dataset.

Code: 5 Jupyter notebooks | Visualizations: report/ folder | Data: Google Drive