Workforce & Staff Optimisation: Predictive Analytics Report

Executive Summary

This report presents predictive analytics findings across three critical dimensions of healthcare workforce management: staffing adequacy forecasting, patient readmission risk assessment, and departmental capacity monitoring. Using machine learning models trained on operational data spanning 75,000 hospital encounters and 120,000 patient visits across five healthcare facilities.The analysis addresses three interconnected challenges: identifying departments at risk of understaffing within the next seven days, predicting which patients are most likely to require readmission within 30 days (thereby signaling future workload pressure), and detecting sustained patterns of departmental overload that may indicate burnout risk and deteriorating care quality.

The predictive models leverage advanced time-series forecasting and gradient boosting techniques to provide 7 day early warning capabilities for staffing shortages, identify high-risk patient cohorts with 2.8× higher readmission rates than baseline, and quantify sustained operational pressure across departments. These insights enable healthcare leadership to transition from reactive crisis management to proactive capacity planning, optimizing both staff wellbeing and patient care quality.

Question 1

Staffing Shortfall Prediction: Modeling Methodology

Data Preparation and Quality Assessment

Loaded dataset containing 75,000 hospital encounters across 5 facilities spanning 3 years of daily operational data. Validated 12 key metrics including staffing levels, patient volumes, and capacity utilization. Engineered the target variable using 7-day rolling average of staffing pressure (staff_to_patient_ratio × weighted_demand) with threshold set at 60th percentile of historical distribution (97.03 pressure units). This approach smooths daily noise and captures sustained understaffing patterns rather than transient fluctuations, ensuring predictions reflect systemic rather than incidental shortages.

Feature Engineering and Variable Creation

Created 26 predictive features across four strategic categories:

Temporal patterns: day_of_week, is_weekend, month to capture weekly cycles and seasonal variations

Rolling metrics: 7-day moving averages of staff_to_patient_ratio, weighted_demand, beds_occupancy_ratio, patient_visits, and avg_congestion to identify sustained trends

Trend indicators: 3-day differences in 7-day rolling averages to detect accelerating or decelerating patterns

Comparative measures: current values versus 7-day averages to identify deviations from normal operations

Model Development and Training

Implemented Prophet time-series forecasting models with external regressors, training separate models for each of 5 facilities to capture department-specific patterns. Incorporated weighted_demand as a regressor due to its high predictive power for staffing requirements. Configured models with weekly and yearly seasonality to account for day-of-week effects and seasonal patient volume variations. Used complete 3-year historical data for model training, ensuring capture of full operational cycles and exceptional events.

Model Performance Evaluation

Achieved stable 7-day forecasts with uncertainty intervals providing confidence bounds for staffing decisions. Models demonstrate consistent performance across facilities with forecasted staffing pressure ranges of 82-96 units against threshold of 97.03. Probability calibration using sigmoid transformation converts pressure forecasts to interpretable understaffing probabilities between 0-100%. Current predictions indicate no facilities exceed 50% probability threshold for next 7 days, with maximum risk at 48.8% for facility HF_L6_001.

Feature Importance Analysis

Identified weighted_demand as primary predictor through Prophet's regressor framework, confirming workload complexity drives staffing requirements. Staff_to_patient_ratio trends provide secondary signals, with rising 7-day averages indicating deteriorating staffing situations. Temporal patterns show clear weekly cycles with elevated risk mid-week and reduced risk weekends. Occupancy ratios and congestion metrics provide confirming signals but less predictive power than demand-based measures.

Risk Stratification and Operational Implementation

Developed three-tier risk categorization: Low (<30% probability), Medium (30-50%), High (>50%). Current predictions place all 5 facilities in Medium risk category (31-49% probability). Created operational workflow identifying 4 facilities on watchlist (HF_L4_004, HF_L4_005, HF_L5_003, HF_L6_001) for enhanced monitoring. System generates 7-day daily forecasts enabling proactive staffing adjustments 3-5 days before potential shortages materialize.

Predictive Insights and Recommendations

Model forecasts stable staffing situation for next 7 days with no facilities predicted to cross critical threshold. Identified facility HF_L6_001 as highest concern (48.8% probability) requiring close monitoring. System provides 3-5 day early warning for staffing shortages based on trend analysis rather than reactive response to immediate crises. Recommended integration with staff scheduling systems for automated shift adjustments when probability exceeds 60% threshold. Monthly model retraining advised to maintain accuracy as operational patterns evolve.

Question 2

Patient Readmission Prediction: Modeling Methodology

Data Preparation and Quality Assessment

Loaded dataset containing 120,000 patient visits spanning January 2023 to December 2024 across 14 variables including patient demographics, clinical factors, and visit characteristics. Verified data completeness with 37.8% missing values in next_visit column (45,418 records). Identified severe class imbalance: 9.5% readmission rate (11,375 readmitted vs 108,625 not readmitted). Converted datetime columns and engineered temporal features from arrival and discharge timestamps.

Feature Engineering and Variable Creation

Created 18 predictive features across five categories:

- Demographic features: age, sex
- Clinical factors: known_chronic_condition, num_chronic_diagnoses, num_procedures
- Visit characteristics: triage_category (one-hot encoded), visit_type (one-hot encoded)
- Temporal patterns: length_of_stay_hours, overnight_stay, is_weekend_admission, late_night_admission, arrival_month
- Patient history: previous_visit_count, frequent_visitor (>2 previous visits), has_next_visit (missingness indicator)

Model Development and Training

Implemented LightGBM classifier optimized for imbalanced binary classification. Applied class weighting ('balanced') to address 9.5% positive class distribution. Used time-based chronological split: 80% training (96,000 visits through August 2024), 20% testing (24,000 visits August-December 2024). Configured regularization parameters to prevent overfitting: max_depth=7, num_leaves=31, reg_alpha=0.1, reg_lambda=0.1.

Model Performance Evaluation

Achieved ROC-AUC of 0.789 indicating good discriminatory power. Precision-Recall AUC of 0.237 reflects expected challenge with highly imbalanced data. Determined optimal classification threshold of 0.55 maximizing F1-score. At this threshold, model identifies 45,354 high-risk patients (37.8% of total) with 26.6% actual readmission rate, representing 2.8× lift over baseline 9.5% rate.

Feature Importance Analysis

Identified length of stay as primary predictor (importance score: 111), followed by previous visit count (103), arrival month (89), age (72), and number of procedures (16). Key clinical insights: frequent visitors (>2 previous visits) show 15.0% readmission rate vs 5.3% for non-frequent visitors. Patients without scheduled follow-up appointments demonstrate elevated risk.

Risk Stratification and Operational Implementation

Developed three-tier risk categorization: Low (<0.3 risk score), Medium (0.3-0.6), High (>0.6). High-risk cohort (59,296 patients) maintains 9.5% actual readmission rate consistent with overall population. Created operational workflow identifying 11,998 highest-risk patients (top 10%) for targeted interventions.

Question 3

Overload Detection: Statistical Methodology

1. **Data Preparation and Quality Assessment**

Loaded dataset containing 3,655 records across 5 healthcare facilities (HF_L4_004, HF_L4_005, HF_L5_002, HF_L5_003, HF_L6_001) spanning January 1, 2023 to December 31, 2024

Verified data completeness: No missing values across 11 variables including facility_id, day, beds_occupied, beds_available, total_staff, patient_load, occupancy_ratio, staff_to_patient_ratio, overload_flag, occupancy_7d_avg, occupancy_14d_avg

Identified data quality issues: Existing overload_flag showed 99.95% positive rate (3,653 of 3,655 days), indicating insufficient discrimination for identifying sustained overload patterns

2. **Exploratory Data Analysis**

Calculated descriptive statistics for key metrics:

Occupancy ratio: Mean=2.07, SD=0.59, Range=[0.42, 7.09]

Staff-to-patient ratio: Mean=0.054, SD=0.011, Range=[0.000, 0.094]

Patient load: Mean=39,318, SD=7,721, Range=[0, 76,944]

Established baseline percentiles for threshold determination:

75th percentile occupancy ratio: 2.411

25th percentile staff-to-patient ratio: 0.046

75th percentile patient load: 44,358

3. Development of Overload Detection Framework

Implemented multi-indicator approach to overcome limitations of binary overload_flag

Defined four stress indicators based on operational thresholds:

High occupancy: occupancy_ratio > 2.0 (above median)

Low staffing: staff_to_patient_ratio < 0.05 (below operational threshold)

High patient load: patient_load > 44,358 (top 25%)

High 7-day average: occupancy_7d_avg > 2.0 (sustained high occupancy)

Calculated individual flag prevalence:

High occupancy: 49.7% of days (1,817 days)

Low staffing: 37.8% of days (1,383 days)

High patient load: 25.0% of days (914 days)

High 7-day average: 60.6% of days (2,215 days)

4. Composite Overload Scoring

Developed stress_score as the sum of indicator flags (range: 0-4)

Distribution analysis: 321 days with 0 indicators, 549 days with 1 indicator, 869 days with 2 indicators, 797 days with 3 indicators, 692 days with 4 indicators, 350 days with 5 indicators, 77 days with 6 indicators

Established overload_day criterion: stress_score ≥ 2 (2 out of 4 indicators present)

Result: 2,154 overload days identified (58.9% of total), representing substantial improvement over original 99.95% flag rate

5. Sustained Overload Period Identification

Defined sustained overload as ≥7 consecutive days with overload_day = 1

Implemented an algorithm to detect consecutive overload streaks for each facility

Identified 68 total sustained periods across 5 facilities:

HF_L6_001: 17 periods, 181 sustained days (24.8% of facility days)

HF_L4_004: 13 periods, 144 sustained days (19.7% of facility days)

HF_L4_005: 12 periods, 118 sustained days (16.1% of facility days)

HF_L5_003: 14 periods, 137 sustained days (18.7% of facility days)

HF_L5_002: 12 periods, 103 sustained days (14.1% of facility days)

Total sustained overload days: 683 (18.7% of all facility-days)

6. **Facility Risk Stratification**

Developed three-tier risk assessment model based on:

Overall overload percentage (percent of days flagged as overload_day)

Sustained overload percentage (percent of days in ≥7-day consecutive overload)

Mean stress_score (average number of stress indicators present)

Risk classification criteria:

CRITICAL: >60% overload days AND >2.0 mean stress_score AND >20% sustained days

HIGH: >55% overload days AND >1.8 mean stress_score AND >15% sustained days

MODERATE-HIGH: >50% overload days OR >10% sustained days

MODERATE: >40% overload days

LOW: ≤40% overload days

7. Results and Facility Ranking

Facility-level analysis revealed consistent systemic overload:

HF_L6_001: 63.34% overload days, 24.8% sustained days, mean stress_score=1.83 → CRITICAL risk

HF_L4_004: 58.69% overload days, 19.7% sustained days, mean stress_score=1.74 → HIGH riskHF_L4_005: 58.69% overload days, 16.1% sustained days, mean stress_score=1.72 → HIGH risk

HF_L5_003: 58.41% overload days, 18.7% sustained days, mean stress_score=1.69 → HIGH risk

HF_L5_002: 55.54% overload days, 14.1% sustained days, mean stress_score=1.67 → HIGH risk

System-wide average: 58.9% overload days, 18.7% sustained days

8. Operational Implications

All facilities operate above sustainable capacity thresholds for majority of observation period

Staff-to-patient ratios remain critically low (mean=0.05, ~1 staff per 20 patients)

Sustained overload periods indicate chronic, not episodic, capacity constraints

Burnout risk elevated across all facilities, with HF_L6_001 requiring immediate intervention