30-DAY HOSPITAL READMISSION PREDICTION USING MACHINE LEARNING IN R FOR A U.S. HEALTHCARE PROVIDER

1. Background

A public hospital in Chicago faced financial and quality penalties under Medicare due to high 30-day patient readmission rates. Clinical teams lacked predictive tools to identify which discharged patients were most likely to return.

We were engaged to build an interpretable machine learning model in R that could predict patient readmission risk at discharge using EHR data. The goal was to support care coordination, improve post-discharge follow-up, and reduce penalties tied to preventable readmissions.

2. Objective

- To build a supervised classification model in R that predicts 30-day readmissions
- To identify key clinical and demographic predictors driving readmission
- To help clinicians triage discharge plans for high-risk patients

3. Data Used

Source: Anonymized hospital EHR system (2022 discharges)

Dataset Details:

- 9,380 adult inpatient discharge records
- Fields included:
    - Readmitted_30Days (target: 1 = Yes, 0 = No)
    - Age, Gender, Comorbidity_Index, Length_of_Stay, Discharge_Disposition, Primary_Diagnosis_Code, Prior_Readmissions, Insurance_Type, ICU_Stay_Flag
- Data preprocessing:
    - Imputed missing values using median (numeric) and mode (categorical)
    - Encoded categorical variables using caret::dummyVars
    - Scaled numerical variables using scale() from base

4. Methodology

4.1 Model Training

- Split dataset (80% train, 20% test)
- Algorithms implemented in R:
  - **Logistic Regression** using glm()
  - **Random Forest** using randomForest package
  - **Cross-validation** (5-fold) using caret::trainControl

4.2 Model Evaluation Metrics

- Accuracy
- Sensitivity (Recall for predicting positive cases)
- Specificity
- Area Under ROC Curve (AUC)

4.3 Feature Importance

- Ranked features based on varImp() from caret and Gini index from random forest

5. Model Results

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	78.2%	61.4%	83.5%	0.74
Random Forest	82.5%	72.9%	86.1%	0.84

Top 5 Predictors (Random Forest):

1. Prior_Readmissions
2. Comorbidity_Index
3. Length_of_Stay
4. ICU_Stay_Flag
5. Age

6. Interpretation and Strategy

- Patients with a history of readmissions and high comorbidity are the strongest risk candidates
- ICU stays and longer admissions increased readmission odds
- Random forest offered better generalization without overfitting

Strategic Recommendations:

- Integrate model outputs into discharge planning system
- Flag top 25% high-risk patients for extra follow-up (calls, home visits)
- Tailor post-discharge instructions based on comorbidity profiles
- Assign case managers to patients with >2 prior readmissions

7. Reporting Output

- **R Markdown Report (PDF, 21 pages)**:
  - Cleaned dataset summary
  - Model training and ROC curves
  - Confusion matrix and feature importance plots
  - Strategic recommendations for hospital operations
- **Interactive Script (Optional)**:
  - Provided app.R using shiny to test readmission predictions in-browser
  - Form interface for inputting patient features and getting risk prediction
- **Excel Output**:
  - Patient_ID, prediction probabilities, and risk category (High / Medium / Low)

8. Business Impact

- **Readmission rate dropped by 12.4%** in Q1 after model deployment
- Model used in **80% of discharge decisions** by care coordination teams
- Helped avoid **$280,000+ in annual Medicare penalties**
- Enabled development of a long-term digital discharge decision-support system