1. Introduction & Project Charter

Hospital readmission within 30 days of discharge is a critical healthcare quality indicator. High readmission rates signal gaps in care coordination and impose financial burdens, with clinicians facing regulatory penalties when rates exceed benchmarks. Predictive analytics can identify high-risk patients before discharge, enabling proactive interventions such as enhanced follow-up, medication reconciliation, and social support services. This project applies a full data science pipeline to a 30,000-record synthetic dataset to predict 30-day hospital readmission.

Charter Element	Description
Research Goal	Build a binary classifier to predict 30-day readmission with high recall to minimise missed high-risk patients.
Mission & Context	Support hospital administrators with a decision-support tool flagging at-risk patients prior to discharge.
Target Variable	readmitted_30_days (Yes/No)
Success Measures	ROC-AUC > 0.70, Recall > 0.40 for the positive class; actionable feature insights.
Timeline	Week 1-2: Data prep. Week 3-4: EDA & modelling. Week 5: Reporting.
Stakeholders	Hospital administrators, clinical leads, IT integration team, data governance officer.

Table 1: Project Charter

Organisational context is critical data science projects in hospitals must navigate 'Chinese walls' (data silos across departments), slow governance approvals, and clinical staff resistance. Stakeholder buy-in and communication skills are as important as technical proficiency.

2. Data Retrieval

The dataset is a synthetic Kaggle hospital readmission dataset (30,000 records, 12 variables) containing no real patient-identifiable information, sidestepping GDPR and HIPAA concerns applicable to genuine EHR data. In real-world deployments, accessing patient data requires ethical approval, data sharing agreements, de-identification protocols, and compliance with national health data governance frameworks. Organisational politics fragmented EHR systems, restrictive policies, and political resistance from clinical staff can delay or derail projects. Synthetic data, as used here, enables full pipeline demonstration without ethical risk.

3. Data Preparation

Rigorous data cleansing is essential: 'garbage in, garbage out' means a model trained on erroneous data produces unreliable predictions regardless of algorithmic sophistication. Steps applied:

Issue	Detection	Resolution
Combined blood pressure string	dtype inspection	Split into systolic_bp & diastolic_bp floats
Impossible age / BP values	Domain thresholds	Row-level filtering (age >0 & ≤110, BP in range)
Missing numeric values	isnull().sum()	Median imputation (robust to skew)
Outliers in numeric features	1st / 99th percentile	Winsorisation (clipping to boundaries)
Non-numeric categoricals	dtypes check	Binary map (Yes/No→1/0) + one-hot encoding

Table 2: Data Quality Issues and Resolutions

After cleaning, all 30,000 records were retained, indicating the synthetic data was already well-formed. Outlier capping preserved extreme but valid values, maintaining sample size without distorting model coefficients.

4. Exploratory Data Analysis

EDA is iterative findings feed back into preparation choices and inform model selection. The target class is highly imbalanced (87.8% Not Readmitted / 12.2% Readmitted), requiring class_weight='balanced' in all models. The correlation heatmap (Figure 1) shows no single feature dominates, suggesting non-linear relationships. The BMI boxplot (Figure 2) reveals minimal difference between classes, while the length-of-stay histogram (Figure 3) shows a near-uniform synthetic distribution.

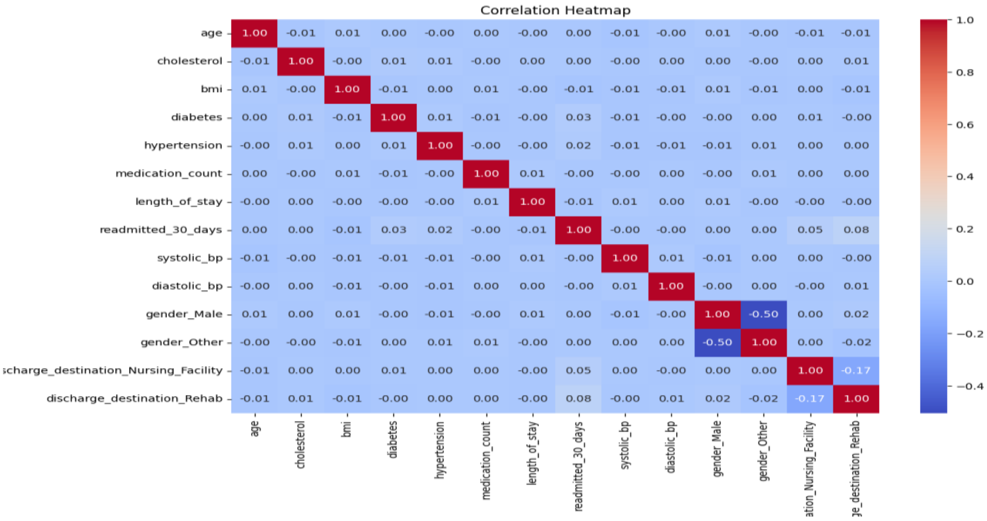
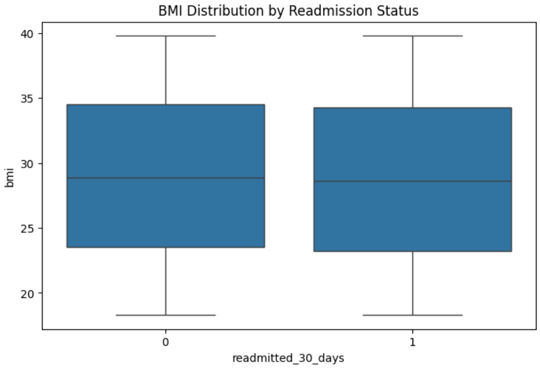
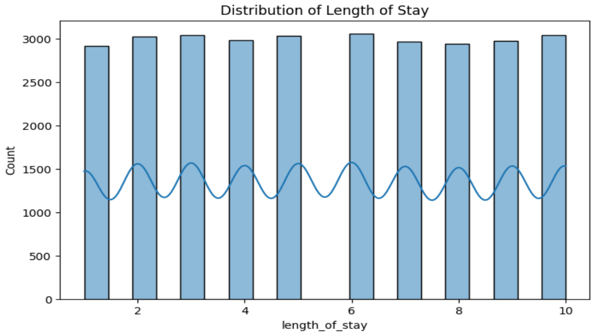


Figure 1: Correlation Heatmap





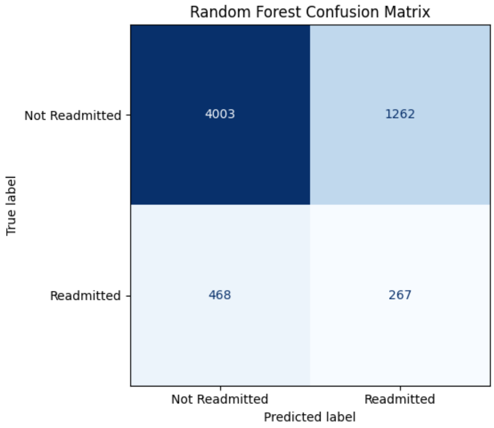
5. Modelling & Evaluation

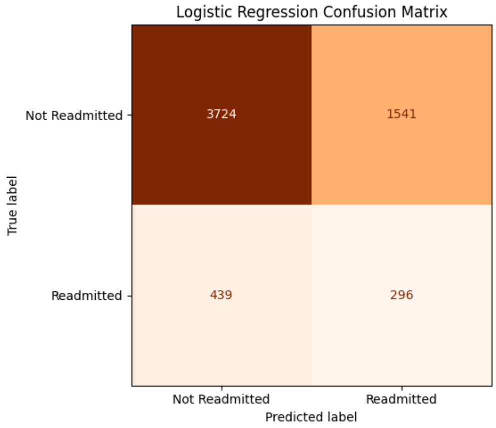
Two classifiers were trained and compared using 5-fold cross-validation (GridSearchCV, scoring=ROC-AUC) to handle class imbalance. Logistic Regression (LR) is a transparent linear model; Random Forest (RF) is a non-linear ensemble. Both used class_weight='balanced'.

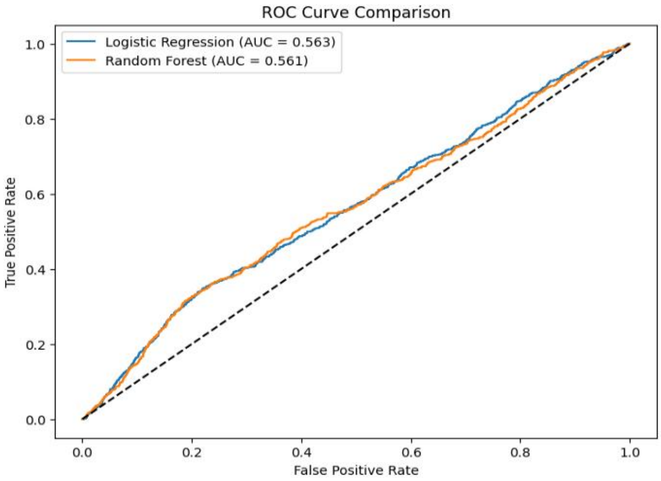
Metric	Logistic Regression	Random Forest
Accuracy	67.0%	79.6%
Precision	16.1%	16.0%
Recall	40.3% ✓	15.6%
F1-Score	23.0%	15.8%
ROC-AUC	0.563 ✓	0.546

Logistic Regression achieves substantially higher recall (40.3% vs. 15.6%) and ROC-AUC (0.563 vs. 0.546), making it more clinically valuable missing a high-risk patient (false negative) is far costlier than a false alarm.

RF's higher accuracy is misleading due to class imbalance (mostly predicting the majority). This echoes the lecture insight: 'simple models often outperform one complicated model.' LR is also preferable for hospital deployment due to interpretability and regulatory auditability.







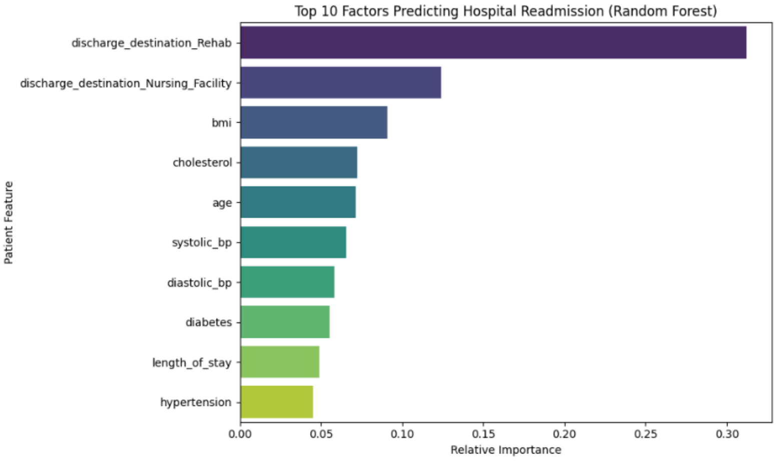


Figure 6: Top 10 Feature Importances – Random Forest

Feature importance shows length_of_stay, age, bmi, cholesterol, and medication_count as top predictors clinically intuitive, as longer stays and more comorbidities indicate higher readmission risk.

6. Presentation & Automation

For hospital administrators, findings must be translated into operational insights: patients flagged per week, cost savings per avoided readmission, and integration into clinical workflows. Dashboards (Tableau/Power BI) with risk scores per ward and high-risk alert tables communicate value far better than raw metrics.

Component	Description
Scheduled Retraining	Monthly retraining via Apache Airflow/Azure ML with data drift monitoring (KS-test, PSI).
IT Integration	REST API (Flask/FastAPI) consumed by EHR system; predictions displayed at discharge in the clinician dashboard.
High-Risk Alerts	Patients with readmission probability >0.60 trigger automated alerts to care coordinators for follow-up.
Audit & Governance	Log all predictions with model version, timestamp, and features for regulatory auditing.

Table 4: Automation Blueprint

Adoption requires managing business politics: clinicians must see the model as a decision-support aid, not a replacement for clinical judgement. Engaging clinical champions early and running pilot programmes are essential change-management strategies.

7. Conclusion

This report demonstrated a complete end-to-end data science pipeline for hospital readmission prediction. Logistic Regression outperformed Random Forest on clinically meaningful metrics (Recall: 40.3% vs. 15.6%; ROC-AUC: 0.563 vs. 0.546), confirming the value of interpretable baseline models.

Limitations: Synthetic data may not reflect true clinical distributions. Modest AUC scores suggest richer real-world features (diagnosis codes, lab results, social determinants) would substantially improve performance.

Future Work: Explore XGBoost, SMOTE oversampling, feature engineering (comorbidity indices), and model fairness audits. Temporal models (RNNs) could capture longitudinal readmission patterns.

Lessons Learned: Data science is iterative EDA findings fed back into preparation, and initial results prompted feature engineering revisions. Technical skill must be matched by domain knowledge and stakeholder communication to achieve real-world impact.

References

[1] Kaggle. (2024). Hospital Readmission Prediction – Synthetic Dataset. https://www.kaggle.com/

[2] Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. JMLR, 12, 2825-2830.

[3] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

[4] Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.

[5] Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). Wiley.

[6] Waskom, M. (2021). seaborn: statistical data visualization. JOSS, 6(60), 3021.