Predicting 30-Day Hospital Readmissions

Capstone Project | Data Analytics Bootcamp

General Assembly Singapore

👩‍💻 **Presenter: Myo Myint Aung Jimmy**

🗓️ **Date : 07 November 2025**

📊 **Engagement:** Contract-based Remote Data Analyst (30-day project)

🎯 **Objective: Build a recall-first machine learning model to predict hospital readmissions within 30 days and support post-discharge follow-up actions.**

🎯 Target Audience (Personas)

👩‍⚕️ Audrey Lim – Senior Care Manager (SGH)

🩺 Oversees discharge planning and patient follow-up

📋 Needs an easy daily list of high-risk patients for next-day calls

❤️ Prefers high recall (few misses), simple visuals, and interpretability

🔍 Uses dashboard insights to plan outreach schedules and allocate team effort

👨‍⚕️ Dr. Daniel Koh – Director, Quality & Operations

📈 Oversees hospital KPIs and readmission metrics

💰 Balances cost, workload, and intervention success rates

🧠 Needs clear evidence and feature importance for policy decisions

⚙️ Uses dashboard metrics to refine ward-specific thresholds

👨‍💻 Dennis Tan – Analytics Lead

🛠️ In charge of in-house data pipeline and reproducibility

🗒️ Cares about clean data exports and consistent model updates

💡 Prefers transparent modeling with documented code and reproducible output

💡 Problem Statement & Impact

The Challenge

💸 **30-day readmission** is a global healthcare challenge — costly, risky, and often preventable

🏥 Hospital penalties and patient dissatisfaction result from frequent readmissions

Approach

🎯 **Project goal:** Identify patients at risk of being readmitted within 30 days of discharge

⚙️ Use machine learning to provide a **recall-oriented, interpretable model**

📊 **Deliverables:** Reproducible code, documented insights, and Power BI dashboard for non-technical users

🗂️ Data Overview

🌐 **Dataset Source:** Kaggle — *"Hospital Readmissions"* (public, de-identified)

📊 **Rows:** 25,000 | **Columns:** 17 features

🧬 **Feature types:** demographics, encounter info, diagnoses, lab tests, medication counts, inpatient/outpatient visits, LOS, and medical specialty

🎯 **Target variable:** readmitted (binary 0/1, within 30 days)

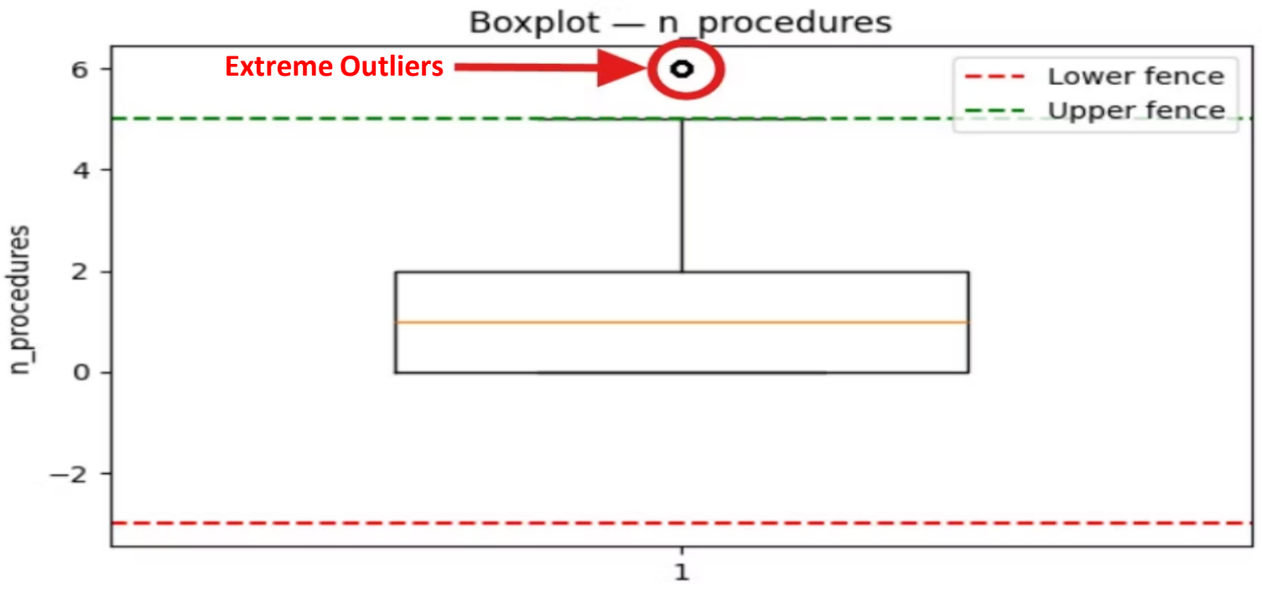
📦 Dataset downloaded via ZIP for better column alignment

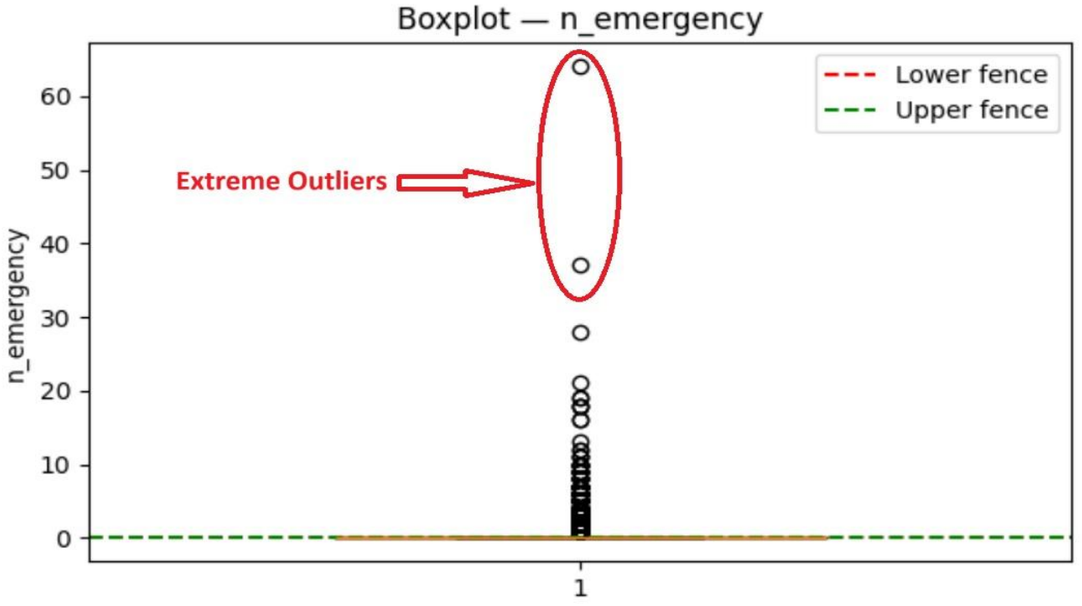
📁 Split into train/test using **stratified 80/20 split** to preserve imbalance

📈 Data Preparation

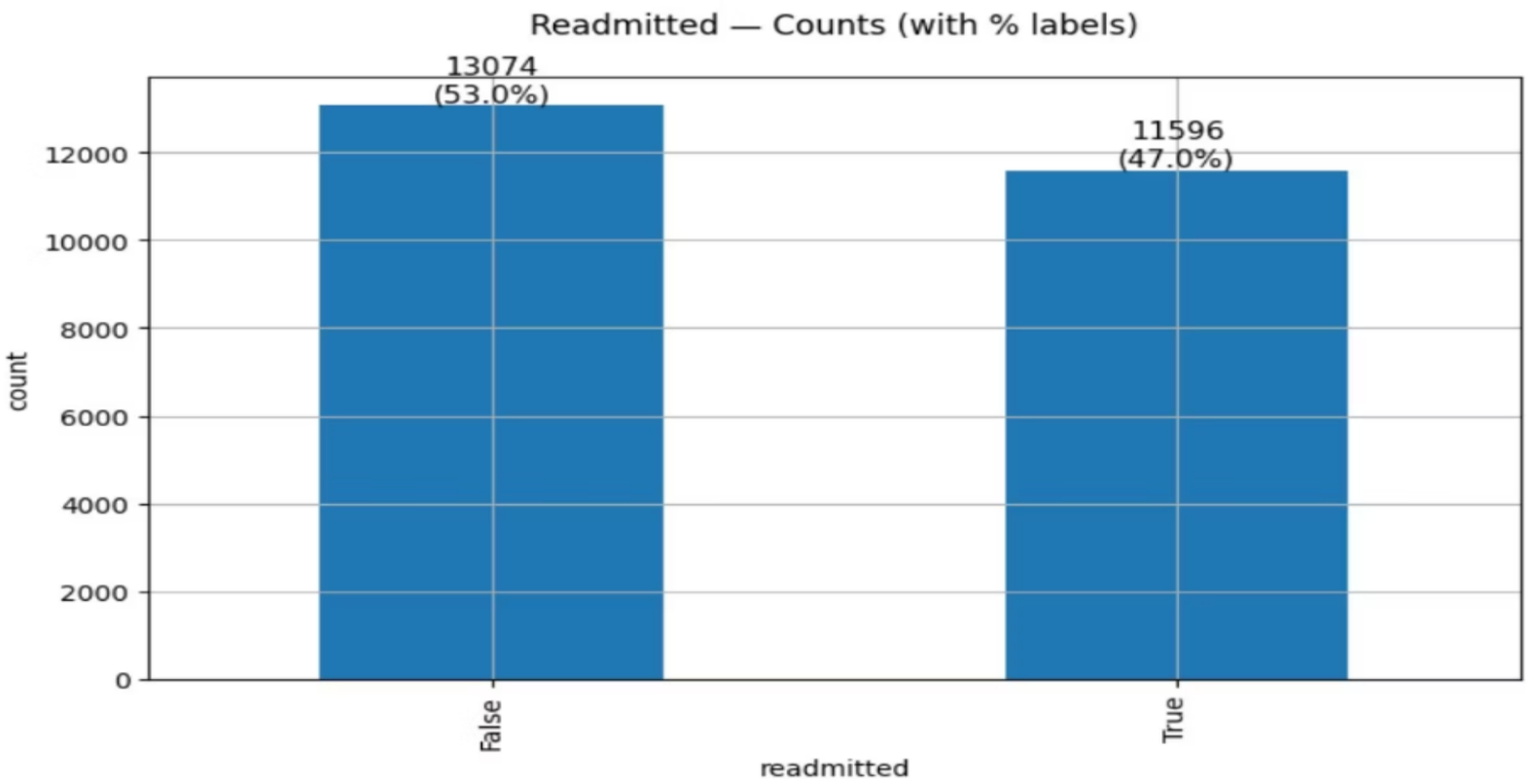
🍪 Data Cleaning Summary

- 🧹 Removed outlier rows (~ 0.2%) to avoid bias
- 🔠 Standardized "?" placeholders → **'Unknown'** for clarity
- ⚖️ Converted Yes/No fields into Boolean
- 📅 Verified date/time columns and normalized formats
- 📝 Checked nulls — mostly minimal, handled via replacement or indicator flags
- 📊 Used **IQR method** to remove extreme outliers (e.g., n_lab_procedures, n_emergency)





📊 Target Distribution



⚠️ Class slightly imbalanced: **47% readmitted** vs 53% not readmitted

🧭 Metric focus is still **F1**, **Recall**, and **ROC-AUC** (not just Accuracy).

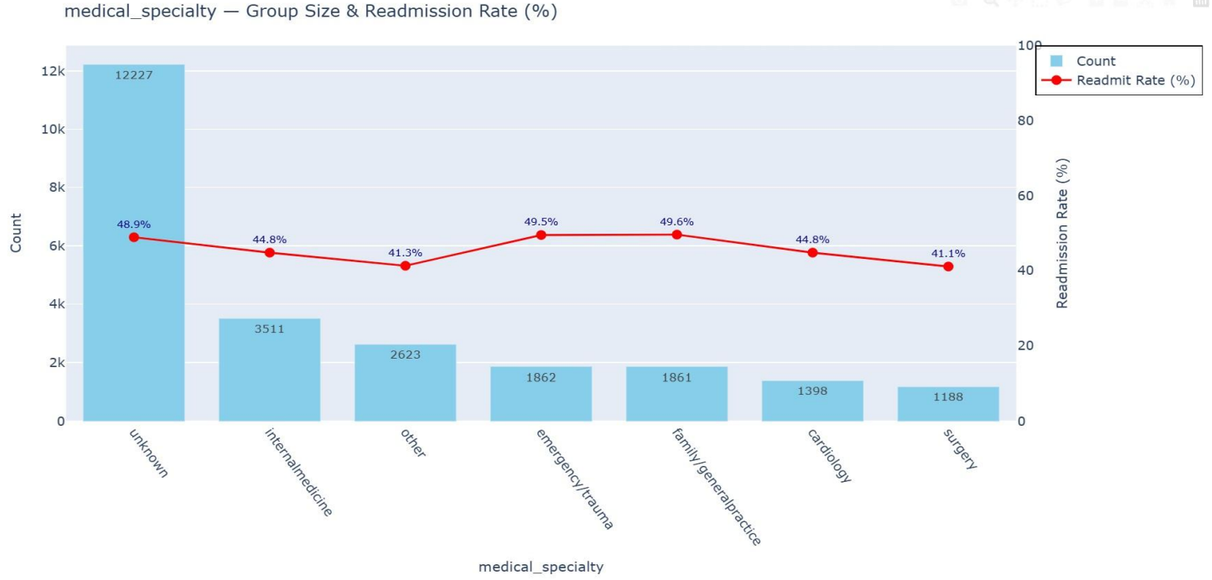
📉 Even with near balance, **missing true readmissions has higher cost**.

👩‍⚕️ Because missing a true readmission is costly, recall matters most for Audrey.

Categorical Feature Analysis

```
=== diag_1 ===
```


```
=== medical_specialty ===
```



Medical Specialty

Emergency/trauma, family/general practice, and unknown show significantly higher readmission rates

Primary Diagnosis

Diabetes, Respiratory, and Circulatory-related issues dominate the readmission landscape

Age Bands

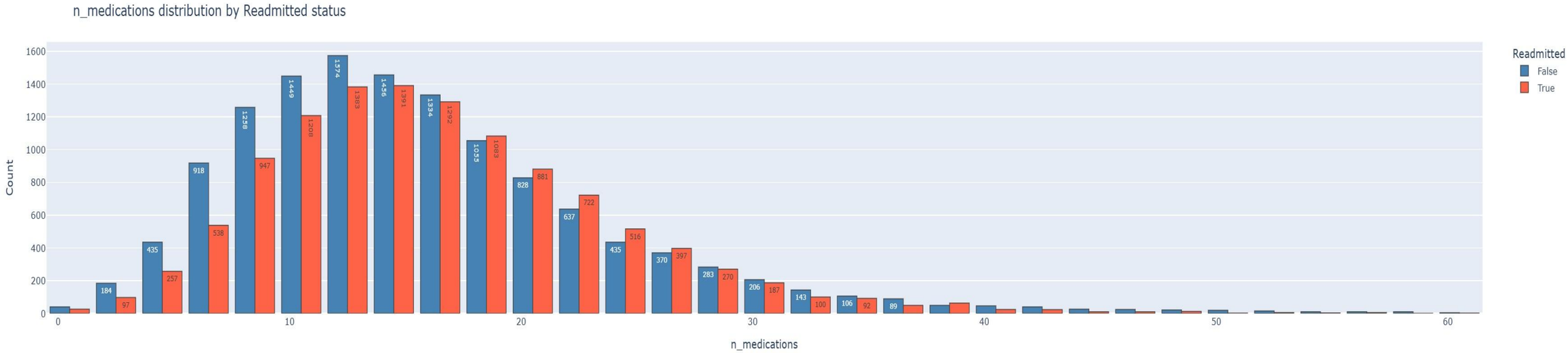
Older groups (60+) are considerably more likely to return within 30 days

Unknown Categories

"Unknown" specialties remain meaningful — often representing untagged general cases that still carry risk

Numeric Feature Analysis

```
=== n_medications ===
```



Medication Count

💊 Higher **n_medications** indicates more chronic patients with elevated risk profiles

Lab Procedures

🧪 More **lab procedures** suggest complex monitoring needs and underlying health complexity

Emergency Visits

🚨 More **emergency visits** correlate strongly with higher readmission probability

Length of Stay

🏥 **Longer stays** indicate sicker patients or more unstable discharges requiring closer follow-up

Key numeric indicators reveal patterns in patient complexity and resource utilization that correlate with readmission risk.

🛠️ Feature Engineering

	glucose_test	glucose_done	glucose_high	A1Ctest	A1C_done	A1C_high
0	no	0	0	no	0	0
1	no	0	0	no	0	0
2	no	0	0	no	0	0
3	no	0	0	no	0	0
4	no	0	0	no	0	0

- ⚙️ Engineered **A1C_test_flag** and **glucose_test_flag** → diabetic monitoring markers
- 🧮 Applied **one-hot encoding** to categorical fields
- ***Key categorical features:*** *medical_specialty, diag_1, diag_2, diag_3, age, change, diabetes_med*

🧭 Modeling Roadmap

01

🧱 Dummy (stratified) → baseline

02

➕ Logistic Regression (baseline)

03

⚖️ Logistic Regression (class_weight='balanced')

04

🌳 Random Forest (baseline)

05

🌳 ⚖️ Random Forest (balanced)

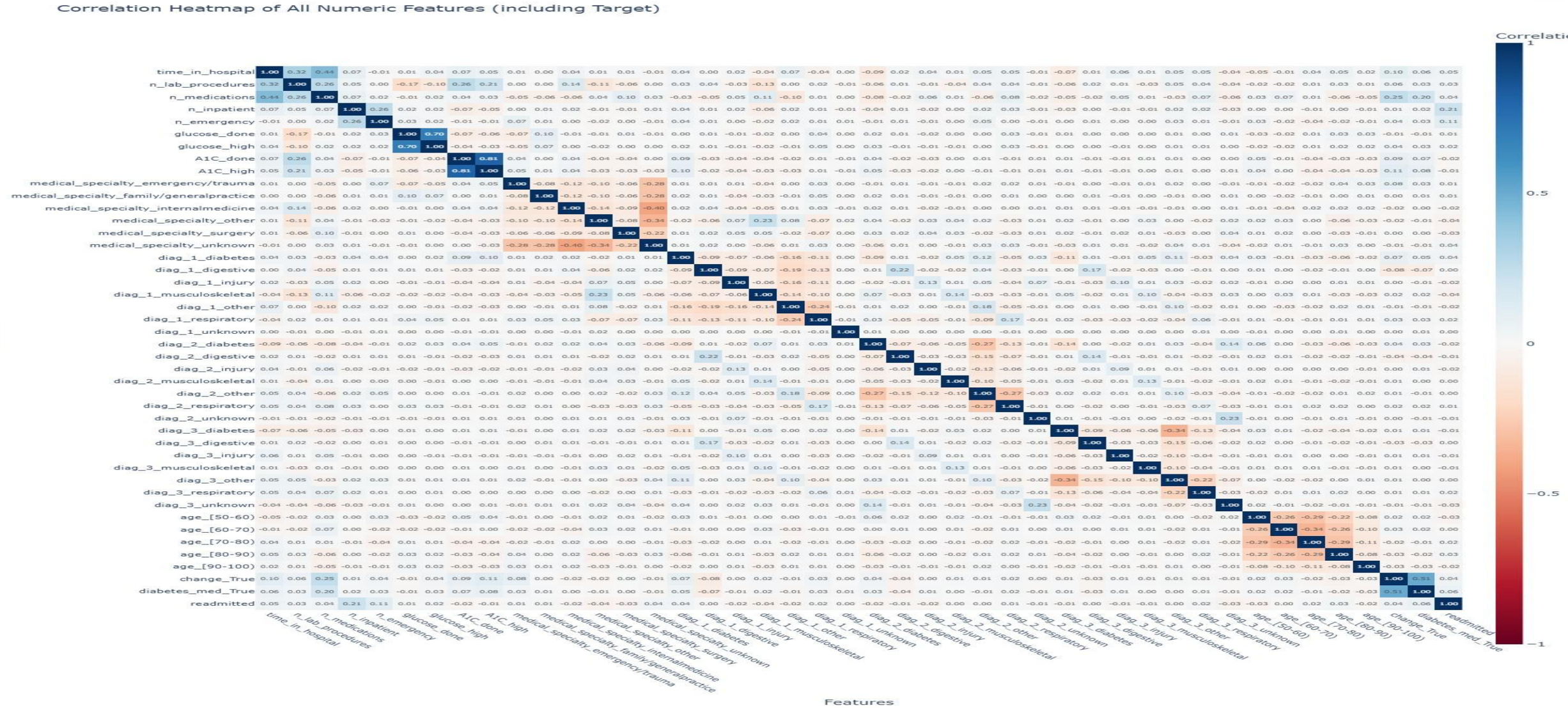
06

🔧 Random Forest (tuned via RandomizedSearchCV)

07

🎚️ **RF_best with threshold = 0.373 (chosen for F1/recall mix)**

Correlation & Redundancy Check



🔍 Key Findings

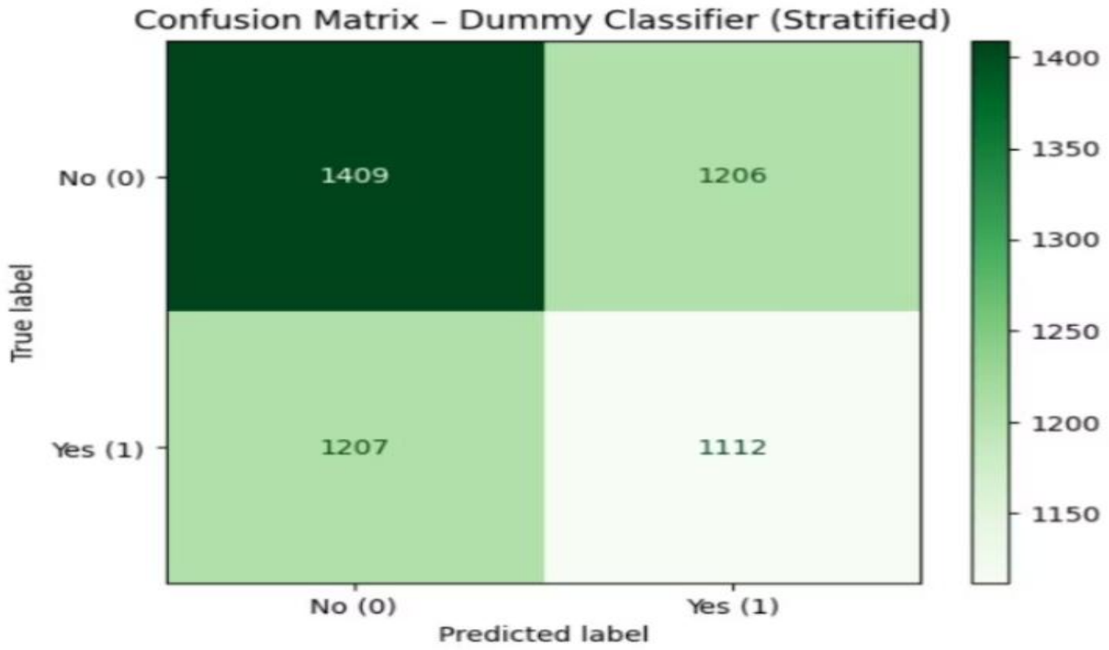
🔥 Some expected clusters emerged (medications, labs, tests, visits)

🧯 **No harmful multicollinearity** with |r|>0.8 across retained features

🌳 Trees tolerate correlation; still reviewed to keep interpretation clear

The correlation analysis confirmed that our feature set maintains statistical independence while capturing meaningful clinical relationships.

🧱 Dummy Baseline (Stratified)

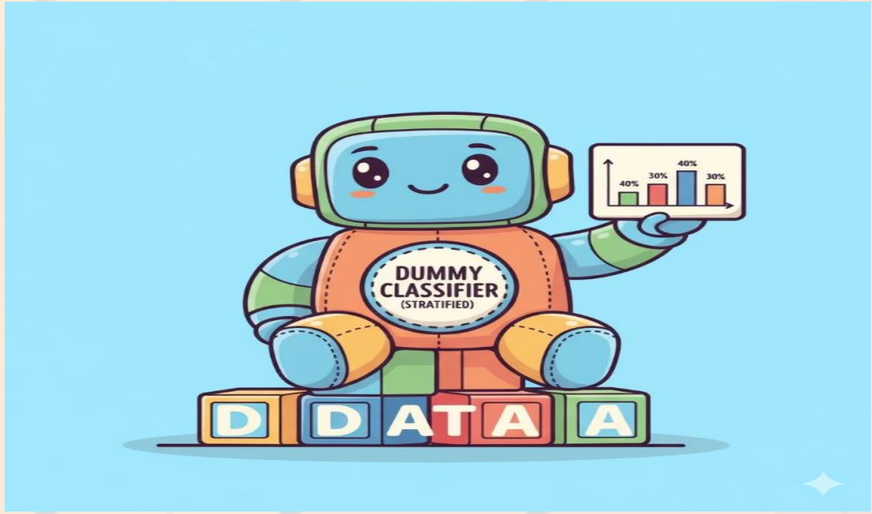


Accuracy	Precision	Recall	F1	ROC_AUC
0.5109	0.4797	0.4795	0.4796	0.5092

True Negative : 1409	False Positive : 1206
False Negative : 1207	True Positive : 1112

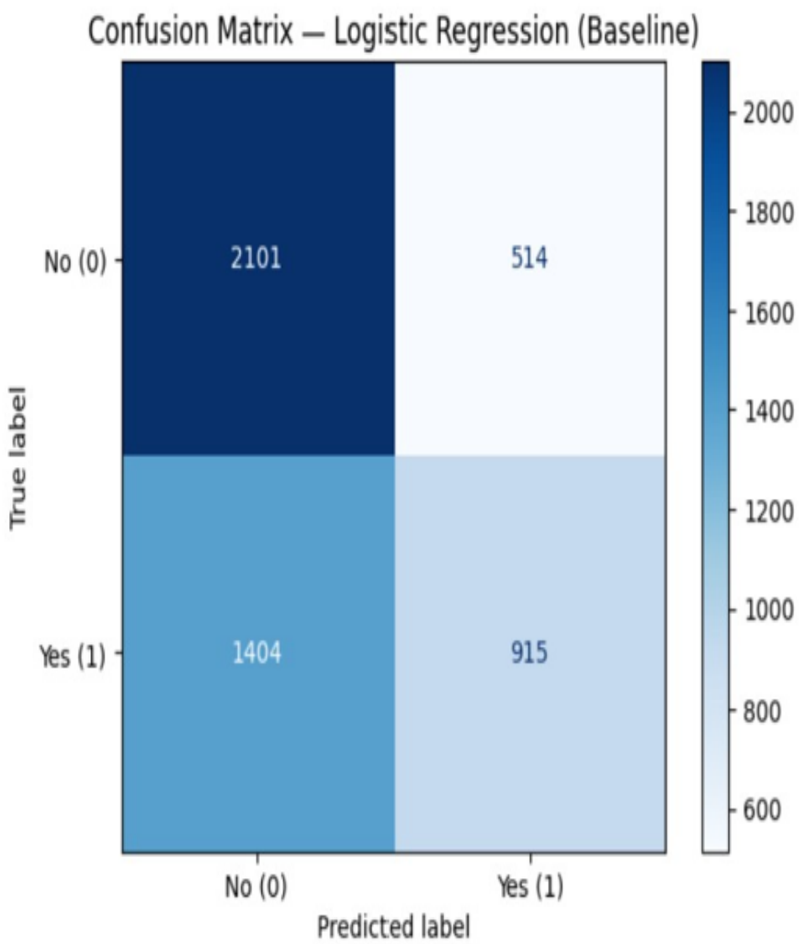
📋 Baseline Performance

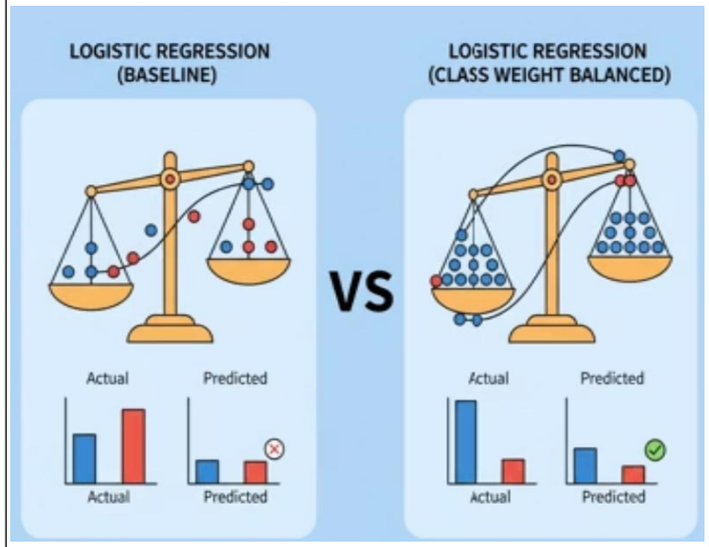
- 🪙 Predicts by class proportion (~ 47% / 53%)
- 📉 ROC-AUC ~ 0.51; very poor discriminative power
- 📉 Recall ~ 0.48; very low recall score
- 🎯 Purpose: establishes the floor benchmark

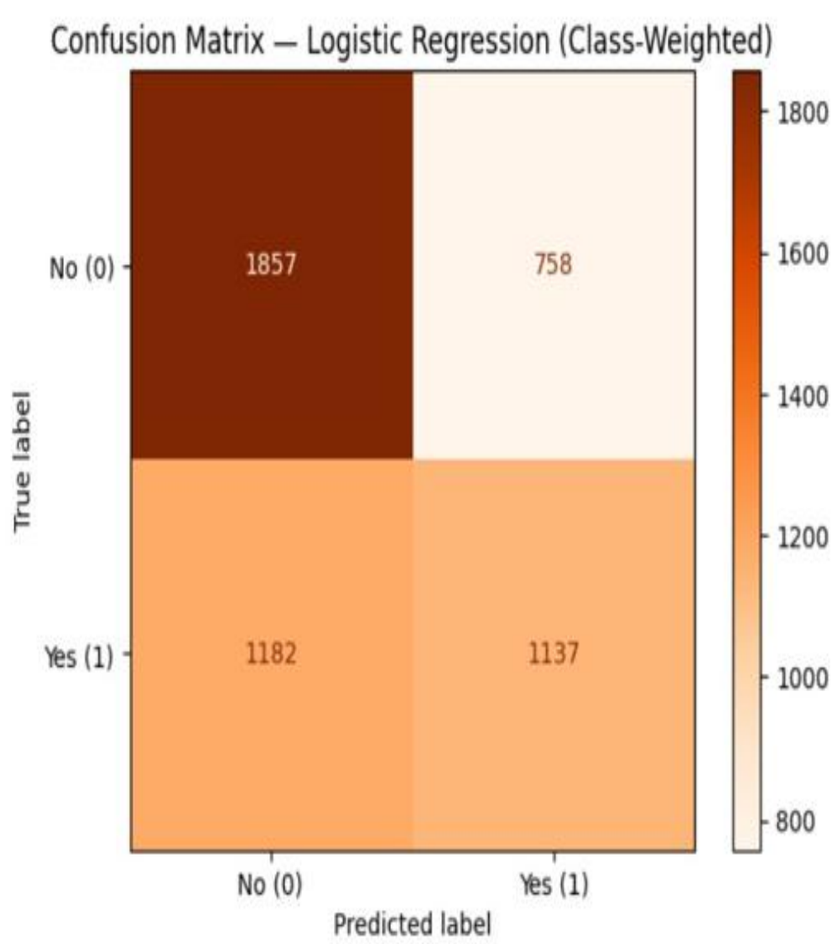


🧠 **This naive model provides our baseline — any meaningful model must significantly outperform this threshold.**

Logistic Regression: Baseline vs Class Weight Balanced







Confusion Matrix – Logistic Regression (Baseline)	
True Negative : 2101	False Positive: 514
False Negative: 1404	True Positive: 915

Confusion Matrix – Logistic Regression (Balanced)	
True Negative : 1857	False Positive: 758
False Negative: 1182	True Positive: 1137

Logistic Regression: Baseline vs Class Weight Balanced

Confusion Matrix – Logistic Regression (Baseline)	
True Negative : 2101	False Positive: 514
False Negative: 1404	True Positive: 915

Confusion Matrix – Logistic Regression (Balanced)	
True Negative : 1857	False Positive: 758
False Negative: 1182	True Positive: 1137

Model	Recall	ROC_AUC	F1	Precision	Accuracy
Dummy (Stratified)	0.4795	0.5092	0.4796	0.4797	0.5109
LogReg (Baseline)	0.3946	0.6410	0.4883	0.6403	0.6113
LogReg (Balanced)	0.4903	0.6410	0.5396	0.6000	0.6068

Baseline Logistic Regression

⚙️ Linear and interpretable approach

📈 Moderate improvement vs dummy; **ROC-AUC ~ mid 0.6**

🔎 Still misses many true readmissions compared to Random Forest

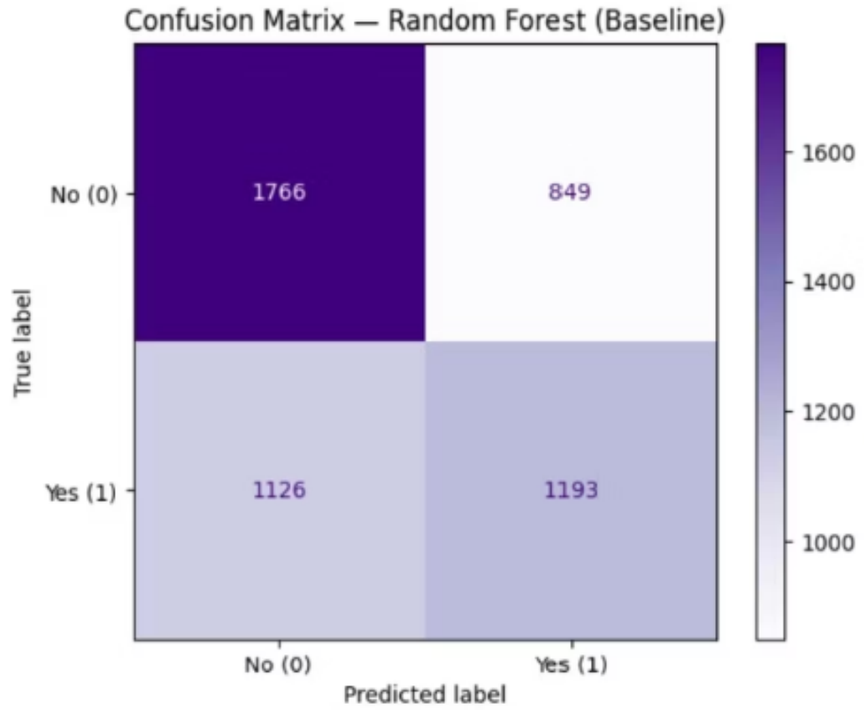
Class Weight Balanced

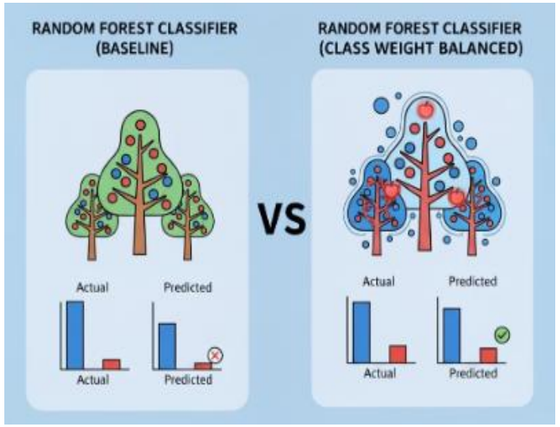
⚖️ With class_weight='balanced' — encourages sensitivity despite near-balanced classes

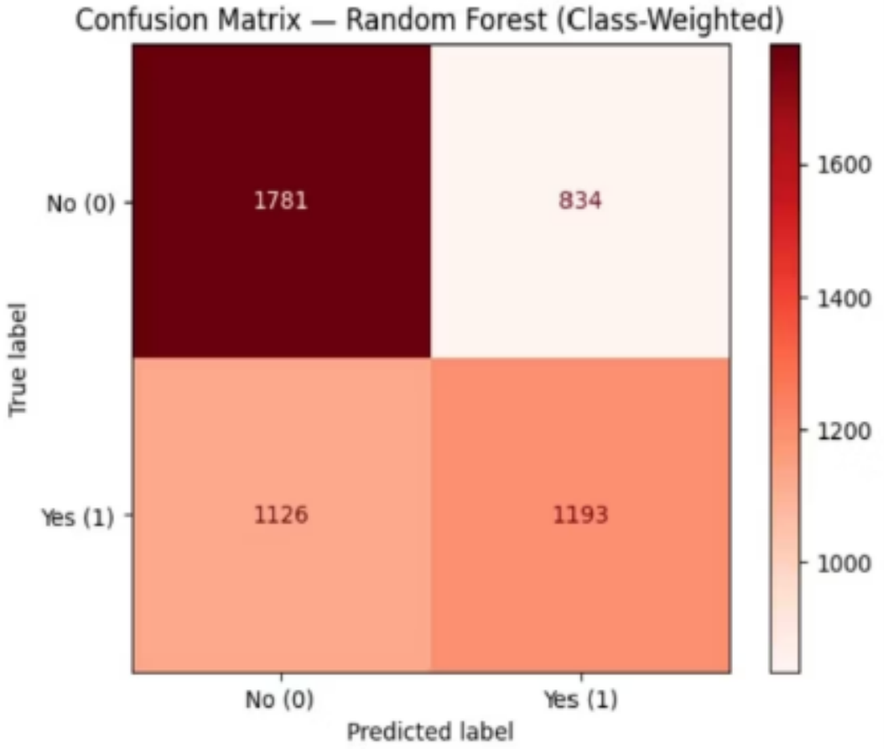
📈 **Recall improves**; F1 slightly better; ROC-AUC similar (~0.62)

🧭 Good explainability; still outperformed by Random Forest

🌳 Random Forest: Baseline vs Class Weight Balanced







Random Forest (Baseline)	
True Negative : 1766	**False Positive: 849**
False Negative: 1126	True Positive: 1193

Random Forest (Class Weight Balanced)	
True Negative : 1781	**False Positive: 834**
False Negative: 1126	True Positive: 1193

🌳 Random Forest: Baseline vs Class Weight Balanced

Random Forest (Baseline)	
True Negative : 1766	**False Positive: 849**
False Negative: 1126	True Positive: 1193

Random Forest (Class Weight Balanced)	
True Negative : 1781	**False Positive: 834**
False Negative: 1126	True Positive: 1193

Model	Recall	ROC_AUC	F1	Precision	Accuracy
Dummy (Stratified)	0.4795	0.5092	0.4796	0.4797	0.5109
LogReg (Baseline)	0.3946	0.6410	0.4883	0.6403	0.6113
LogReg (Balanced)	0.4903	0.6410	0.5396	0.6000	0.6068
Random Forest (Baseline)	0.5144	0.6371	0.5471	0.5842	0.5997
Random Forest (Balanced)	0.5144	0.6354	0.5490	0.5886	0.6028

Random Forest (Baseline)

- 🎄 Captures non-linearities and feature interactions
- 📈 **Recall rises** notably better than Logistic Regression
- 💉 Robust to mixed data types; **ROC-AUC ~0.63**

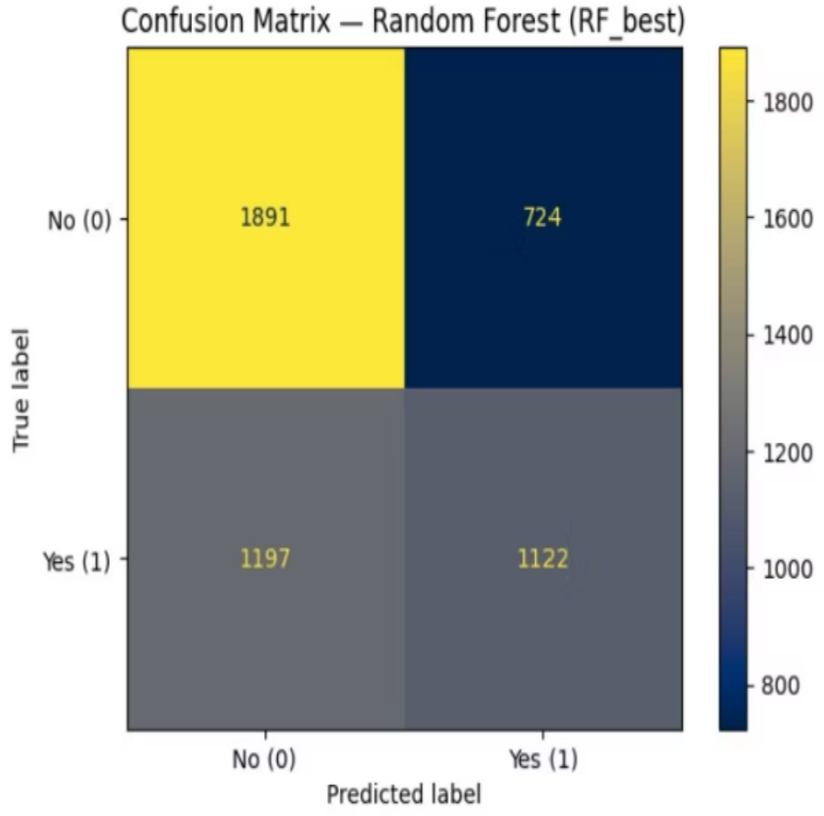
Random Forest (Class Weight Balanced)

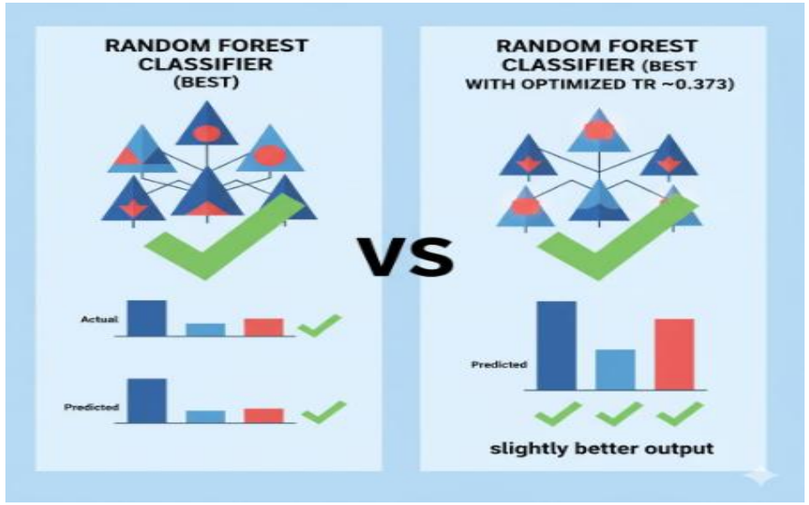
- ⚖️ class_weight='balanced' nudges trees toward sensitivity
- 📈 **F1 increases further**; precision dips slightly
- 🧭 Better alignment with care-manager need (fewer misses)

👩‍⚕️ Closer to Audrey's need: don't miss true readmissions

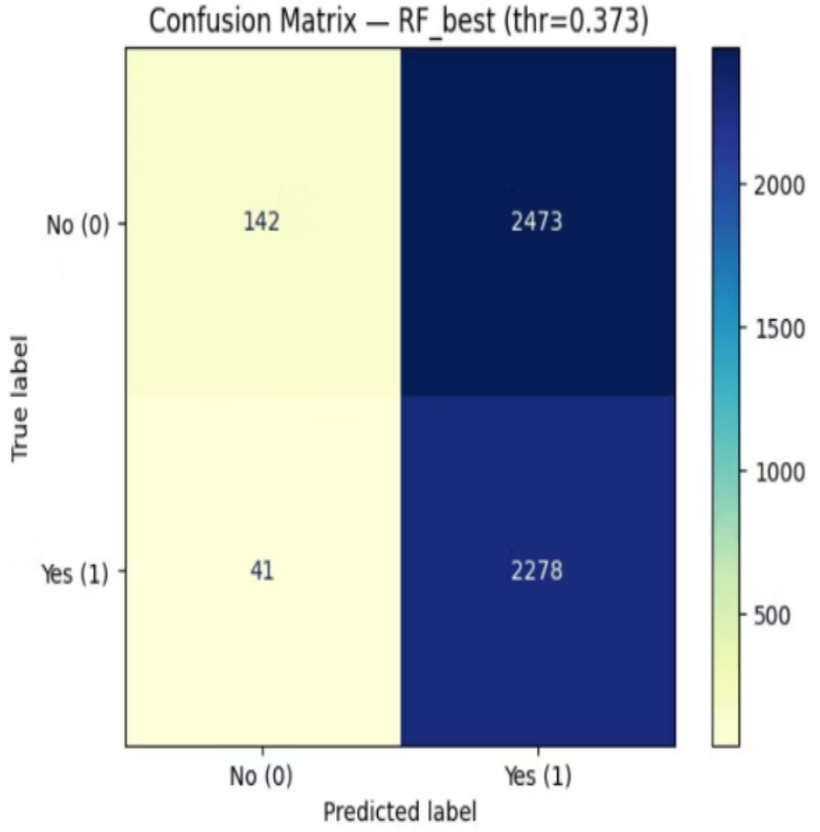
Random Forest: Tuned vs Tuned with Threshold

Random Forest (Tuned)





Random Forest Tuned (Threshold = 0.373)



Best Parameters for Best CV ROC_AUC : 0.6462 :

Rf_n_estimators	300
Rf_min_samples_split	10
Rf_min_samples_leaf	8
Rf_max_features	sqrt
Rf_max_depth	8
Rf_class_weight	Balanced_subsample
Rf_bootstrap	True

Random Forest (Tuned)	
True Negative : 1891	**False Positive: 724**
False Negative: 1197	True Positive: 1122

Random Forest Tuned (Threshold = 0.373)	
True Negative : 142	**False Positive: 2473**
False Negative: 41	True Positive: 2278

🌳 🔧 Random Forest: Tuned vs Tuned with Threshold

Random Forest (Tuned)	
True Negative : 1891	False Positive: 724
False Negative: 1197	True Positive: 1122

Random Forest Tuned (Threshold = 0.373)	
True Negative : 142	False Positive: 2473
False Negative: 41	True Positive: 2278

Model	Recall	ROC_AUC	F1	Precision	Accuracy
Dummy (Stratified)	0.4795	0.5092	0.4796	0.4797	0.5109
LogReg (Baseline)	0.3946	0.6410	0.4883	0.6403	0.6113
LogReg (Balanced)	0.4903	0.6410	0.5396	0.6000	0.6068
Random Forest (Baseline)	0.5144	0.6371	0.5471	0.5842	0.5997
Random Forest (Balanced)	0.5144	0.6354	0.5490	0.5886	0.6028
Random Forest (Tuned)	0.4838	0.6527	0.5388	0.6078	0.6107
Random Forest (Tuned, Threshold = 0.37)	0.9823	0.6527	0.6444	0.4795	0.4905

Random Forest (Tuned)

🔨 **RandomizedSearchCV** over n_estimators, max_depth, min_samples_split, min_samples_leaf

🏅 Best validation F1 ~0.54; ROC-AUC ~0.65, but Recall is Lower @ ~0.48

🧭 Strong candidate pre-threshold

🤦 Meets Dennis & Dr. Koh's expectations but not Audrey's

RF Tuned (Threshold = 0.373)

🎯 Threshold tuned to **maximize F1** while **preserving high recall**

🧪 **Test metrics:** ROC-AUC ~0.65, F1 ~0.64, **Recall ~0.98**, Precision ~0.48, Accuracy ~0.49

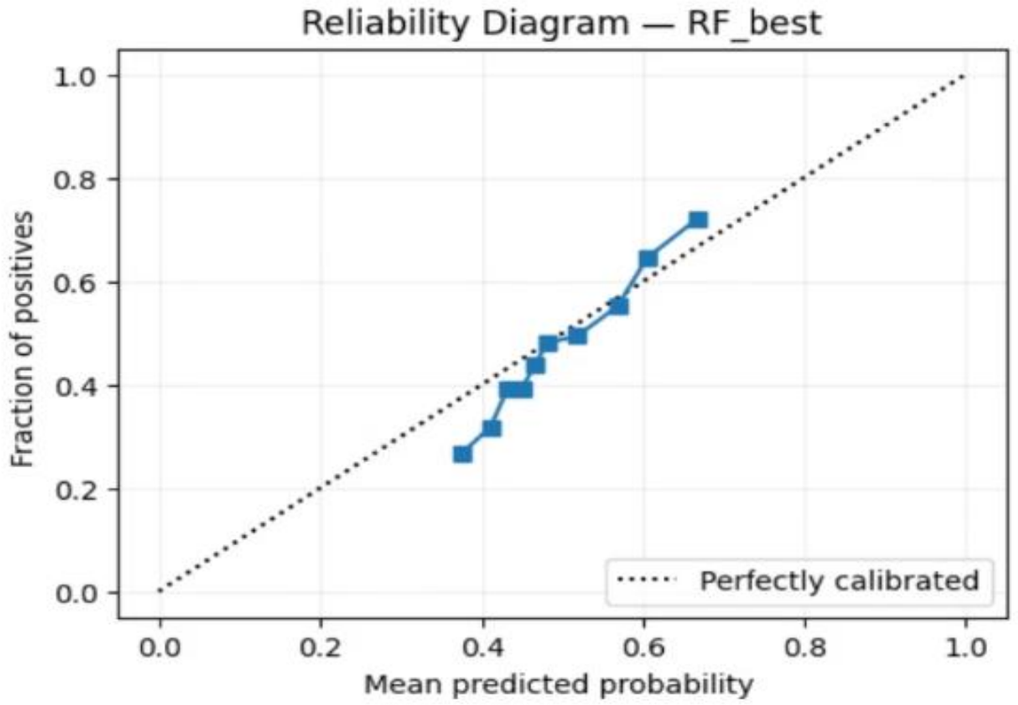
💡 **Operational fit**: very few missed readmissions; manageable false alarms

👩‍⚕️ Optimal for Audrey's workflow and Audrey is satisfied now

Model Validation & Insights

📐 Calibration / Reliability

Brier Score: 0.2341



📈 Decile / Ranking Insight

decile	n	positives	avg_proba	positives_rate	lift_vs_overall
D10	494	132	0.375255	0.267206	0.568520
D9	493	157	0.410993	0.318458	0.677565
D8	493	193	0.431531	0.391481	0.832931
D7	494	194	0.449717	0.392713	0.835551
D6	493	216	0.465582	0.438134	0.932192
D5	493	237	0.481056	0.480730	1.022821
D4	494	244	0.517957	0.493927	1.050900
D3	493	272	0.568140	0.551724	1.173871
D2	493	318	0.604371	0.645030	1.372393
D1	494	356	0.666609	0.720648	1.533280

- 📊 Reliability curve acceptable; probabilities usable for ranking
- 🧮 Brier score in acceptable range for this problem
- 📈 Supports **probability-based triage** (top-N lists)

🏅 This gives leadership confidence in using the scores for policy decisions

- 🔟 Sort by predicted probability → deciles
- 🏆 Top deciles concentrate majority of true readmissions
- 📞 Enables **daily prioritized call lists** for Audrey's team

🏁 Final Model Summary

	model	roc_auc	brier	precision@0.373	recall@0.373	f1@0.373
0	RF_best (raw)	0.652746	0.234057	0.479588	0.982751	0.644605
1	RF_best + sigmoid calibration	0.652247	0.231411	0.523035	0.832255	0.642370

0.65
ROC-AUC

0.64
F1 Score

0.98
Recall

0.48
Precision

🌲 Final Choice: RF_best (threshold = 0.373)

🎯 Chosen for near-balanced classes which suited best for Audrey's needs and also it was strongly supported by Dr Koh and Dennis

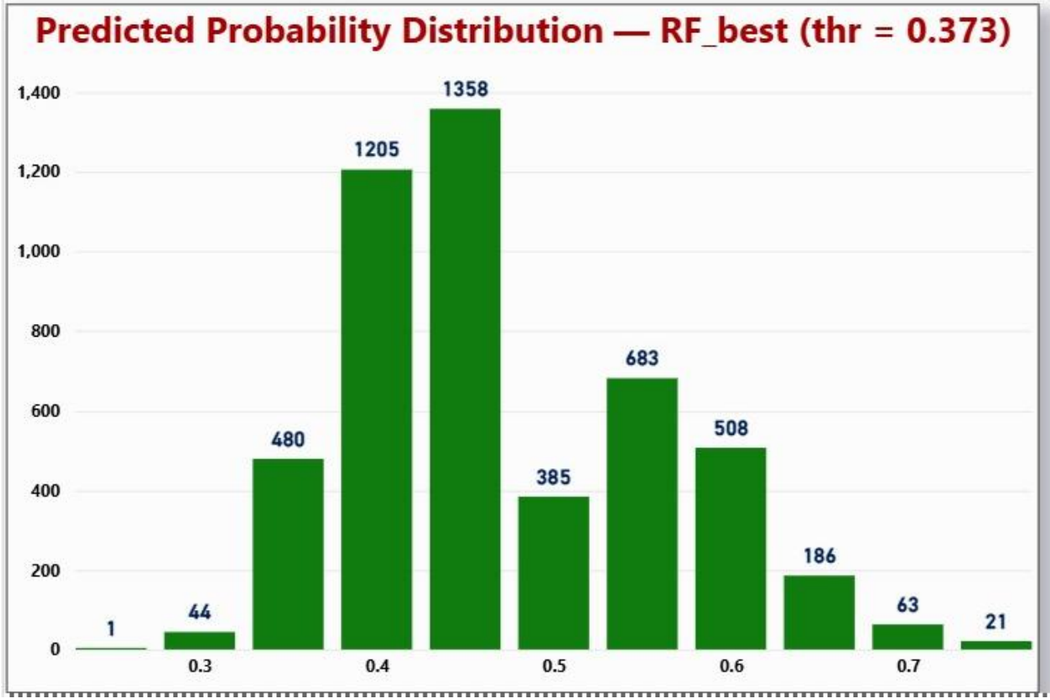
Power BI Dashboard & Key Insights

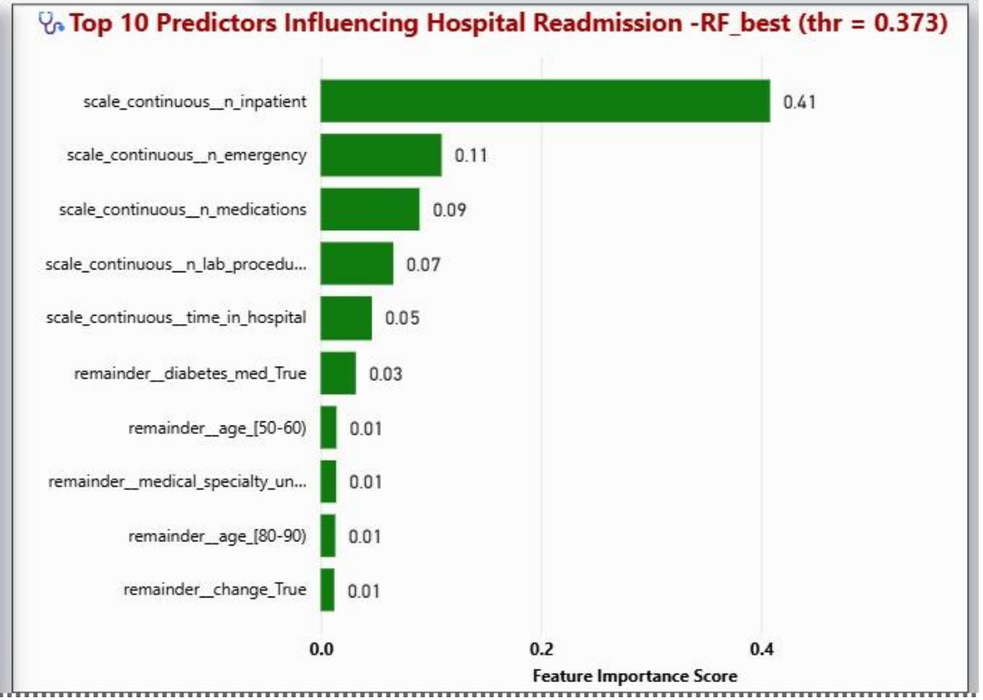
Total Patients	Predicted Positives	Actual Readmissions
4934	4752	2319

Precision	Recall	F1	Accuracy	ROC_AUC
47.96%	98.28%	64.46%	49.07%	65.27%

Confusion Matrix — RF_best (thr = 0.373)

y_true	0	1	Total
1	40	2279	2319
0	142	2473	2615
Total	182	4752	4934





Dashboard Components

- KPI cards: Accuracy, Precision, Recall, F1, ROC-AUC (as %)
- Confusion Matrix visual
- Probability Histogram (cut at 0.373)
- Top-10 Predictors (from RF importances CSV)

Key Insights

- **Recall ~98%** → almost all readmissions flagged
- Precision ~48% → manageable follow-up load
- ROC-AUC ~65% → decent ranking across near-balanced classes

Top Drivers

- Prior inpatient & emergency visits, medications, lab procedures, length of stay, diabetes medications, and age
- Class base rates: nearly 47/53, so performance gains are **from signal**, not imbalance artifacts

🧭 Persona-Based Actions

👩‍⚕️ Audrey (Care Manager)

- 📋 Use **top deciles** list daily
- 🩺 Add nurse triage and short checklist
- 📉 Track outcomes weekly to refine approach

👩‍🔬 Dr. Koh (Quality & Ops)

- 🚦 Pilot thresholds by ward
- 🔎 Monitor precision/recall drift monthly
- 🔧 Adjust operational protocols based on data

👩‍💻 Dennis (Analytics)

- 🛠️ Automate monthly retrain process
- ⚙️ Maintain stable CSVs (hosp_dffinal.csv, rf_feature_importances.csv)
- 📅 Schedule Power BI refresh and monitoring

⚠️ Limitations

📄 Public, de-identified dataset; missing richer clinical and social determinants features

⚙️ ROC-AUC ~0.65 → moderate performance; real-world uplift expected with better features

🔒 No patient IDs → no longitudinal re-linking possible in current dataset

🚀 Next Steps

🎯 Future Enhancements

Enrich Features

🔁 Add comorbidity indices, medication classes, and procedure details

Explainability

🧠 Add **SHAP** for patient-level explanations and clinical transparency

MLOps Pipeline

🏭 Monthly retrain, dashboard refresh, and alert governance

Advanced Models

🧪 Explore **XGBoost/CatBoost** for incremental AUC gains; compare cost curves

✅ Conclusion

❤️ Near-balanced classes (47/53) → improvements reflect **true signal**

🌳 **RF_best (threshold 0.373)** achieves **very high recall** with workable precision

📊 Dashboard operationalizes the model for care teams and leadership

🎯 Ready for pilot deployment with continuous monitoring

🙏 Thank You

Questions ❓ ❓ ❓

📋 References

Kaggle Dataset – *Hospital Readmissions (Diabetic Inpatients)*

https://www.kaggle.com/datasets/dubradave/hospital-readmissions

Scikit-learn Documentation – *Dummy Classifier*

https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html

Scikit-learn Documentation – *Logistic Regression*

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Scikit-learn Documentation – *Random Forest Classifier/Regressor*

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

DataCamp Tutorial – *Random Forests Classifier in Python*

https://www.datacamp.com/tutorial/random-forests-classifier-python

Kaggle Notebook – *Predicting Hospital Re-Admissions* by Jeyapal

https://www.kaggle.com/code/jeyapal/predicting-hospital-re-admissions

Kaggle Notebook – *Hospital Readmission EDA and ML (61-49)* by Raphael Marconato

https://www.kaggle.com/code/raphaelmarconato/hospital-readmission-eda-and-ml-61-49

BMJ Open Diabetes Research & Care Journal (2020) – *Predictors of Hospital Readmission Among Patients with Diabetes*

https://drc.bmj.com/content/8/1/e001227