HOTEL BOOKING CANCELLATION PREDICTION

MIS 451 - Machine Learning for Business

LECTURER: Ms. Huynh Gia Linh
LECTURER: Mr. Dang Thai Doan

Quarter 1, 2025-2026

TEAM MEMBERS

NGUYEN THANH DAT	2132300562
NGUYEN HOANG VINH	2132300522
NGUYEN QUANG TRUONG	2132309001

TABLE OF CONTENTS

I. Business Context & Objective

1. Business Context: The Cancellation Challenge

a. The Industry Challenge:

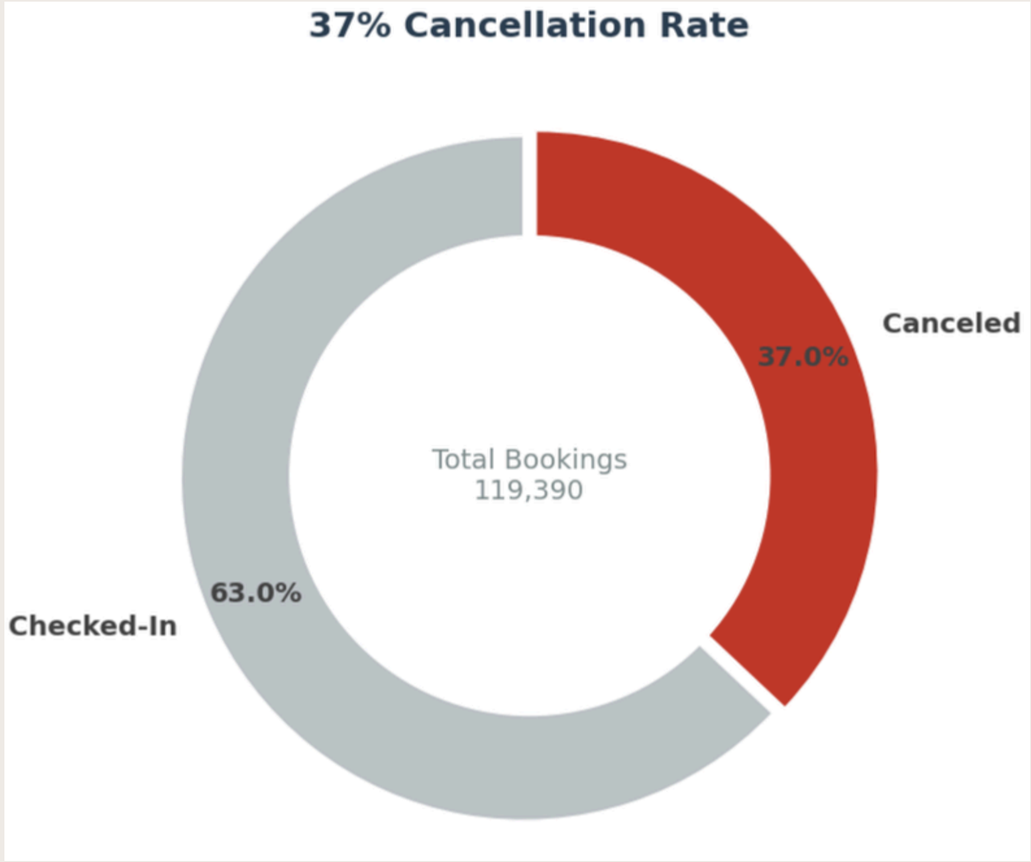
- Hospitality faces high volatility; cancellations lead to unsold rooms (revenue loss) or operational waste (overstaffing).
- Current State: Reactive. Treating all bookings equally until a cancellation happens.

b. The Core Problem

- ~37% Cancellation Rate in current dataset.
- This creates massive uncertainty for inventory management.

c. Stakeholder Impact:

- Revenue Managers: Struggle to optimize pricing and overbooking without accurate risk data.
- Operations: Cannot accurately plan daily staffing or food supplies.



I. Business Context & Objective

2. Objective: Predictive Modeling for Revenue Optimization

a. The Solution

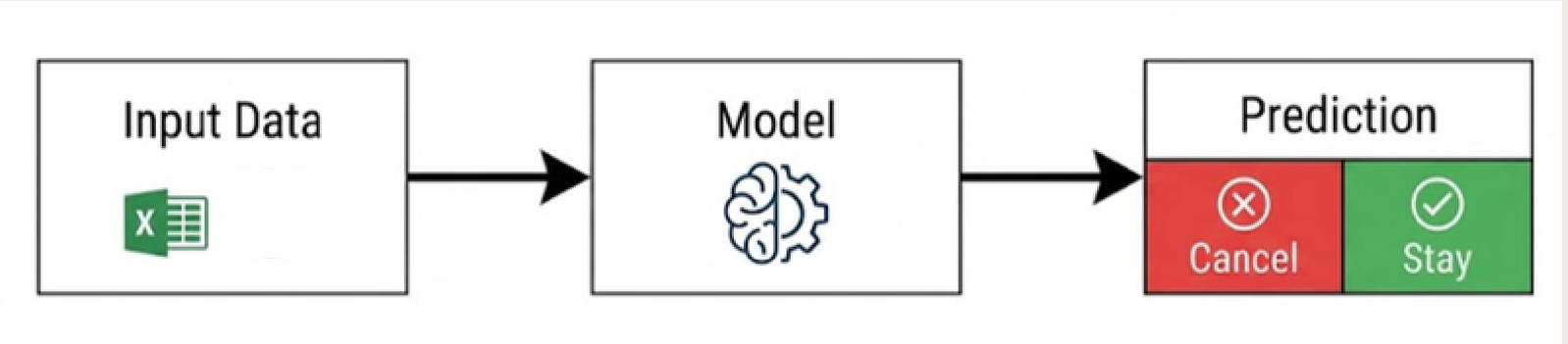
- **Develop Models:** Comparative analysis of Machine Learning vs Deep Learning approaches.
    - Target variable:
        - is_canceled (1 = Cancel, 0 = Non-canceled).
- **Goal:** Shift from Reactive to Proactive management.

b. Business Value:

Optimize Inventory: Enable calculated overbooking to fill rooms predicted to cancel.

Protect Revenue: Offer proactive incentives (e.g., discounts) to high-risk customers to secure commitments.

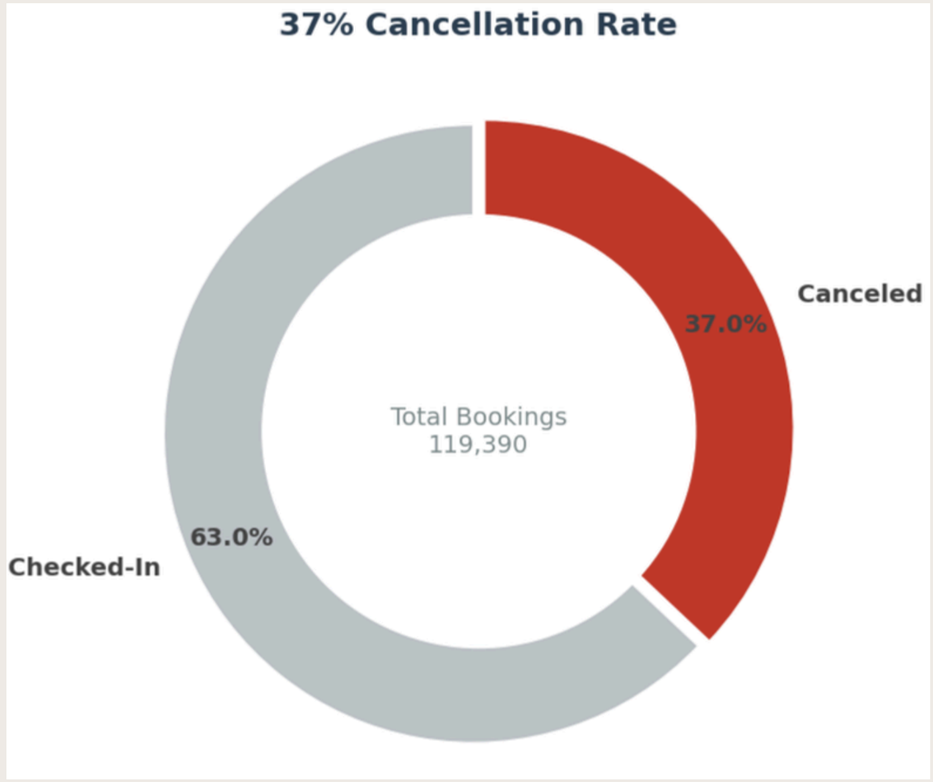
Efficient Operations: Align staffing and supplies with actual predicted arrivals, not just booked numbers.



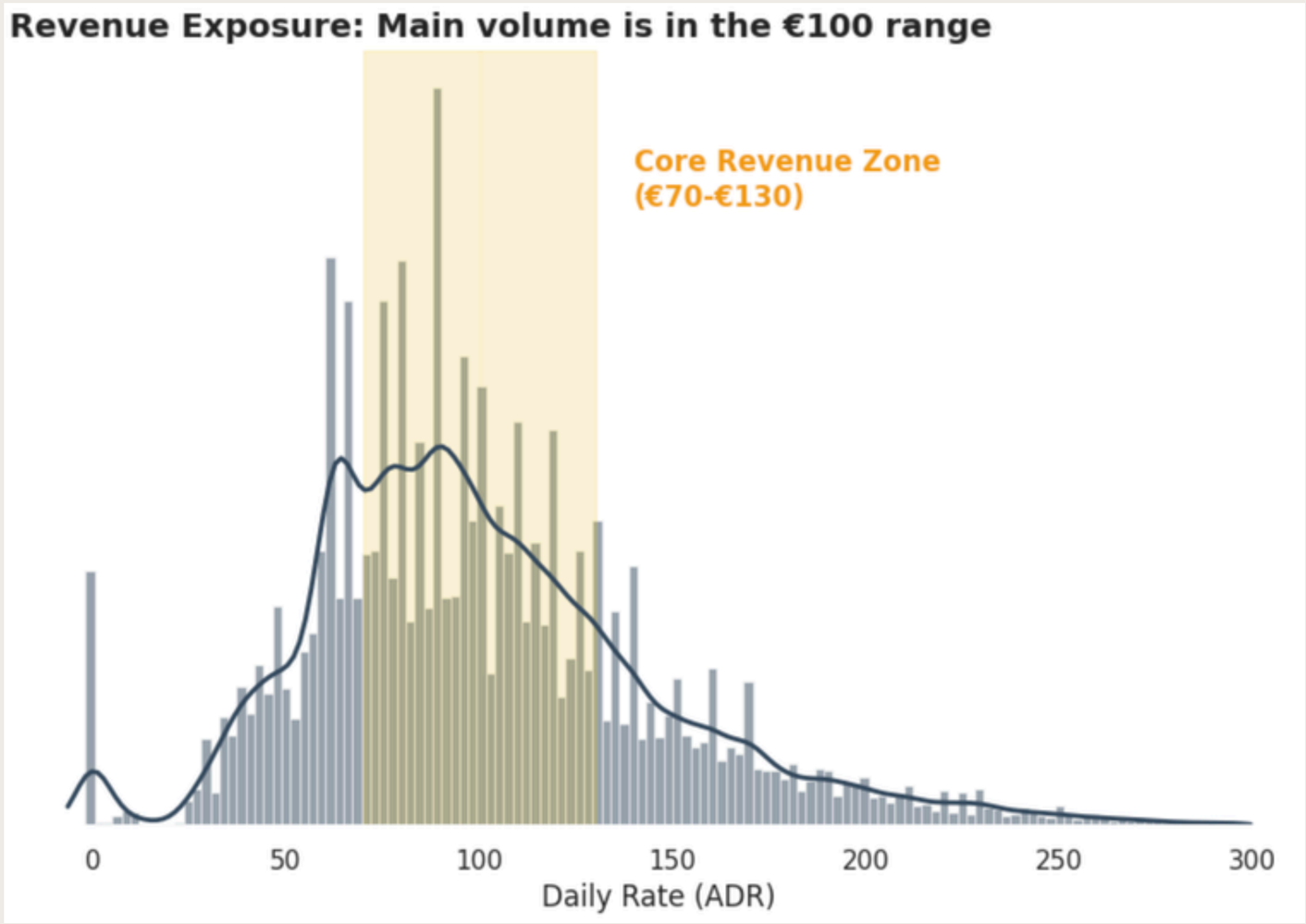
II. Exploratory Data Analysis

1. Current Business Landscape

Project Context: 37% of bookings cancel, putting our core revenue band (€70-€130) at risk.



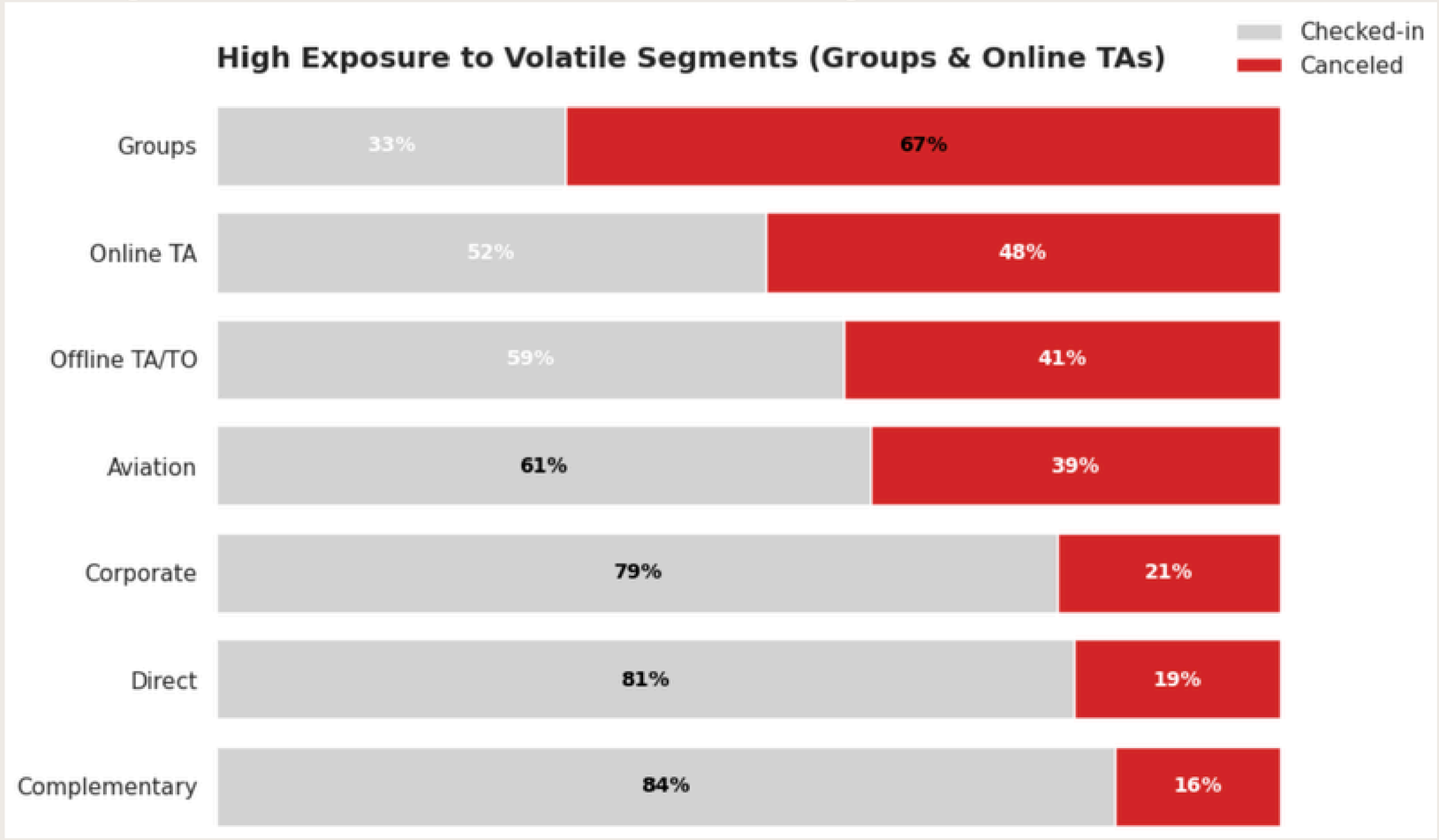
37% cancellations ≈ €4.5 Million in lost potential revenue



II. Exploratory Data Analysis

2. Drivers of Cancellation

High Exposure to Volatile Market Segments

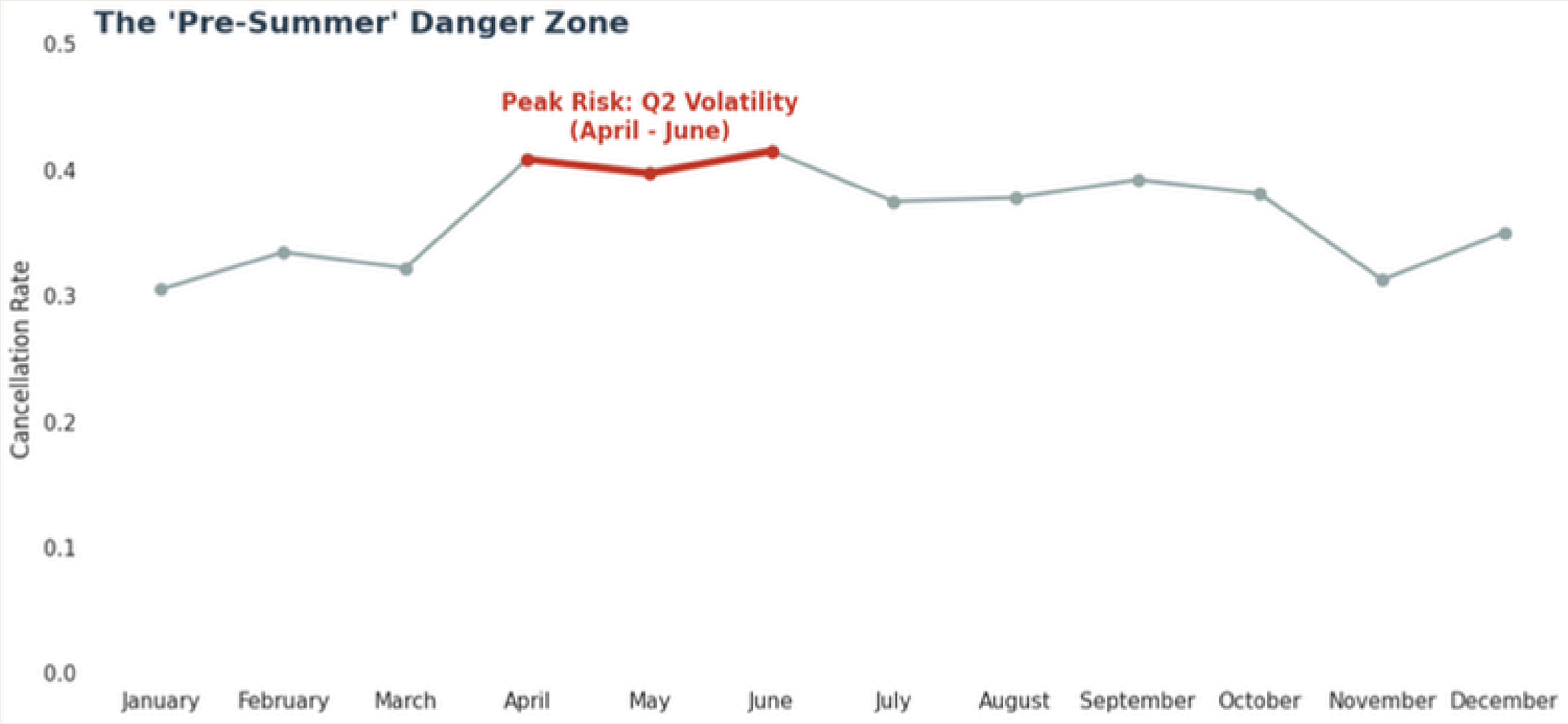


Online TAs and Groups drive the majority of cancellations. Direct and Corporate bookings show significantly higher commitment.

II. Exploratory Data Analysis

2. Drivers of Cancellation

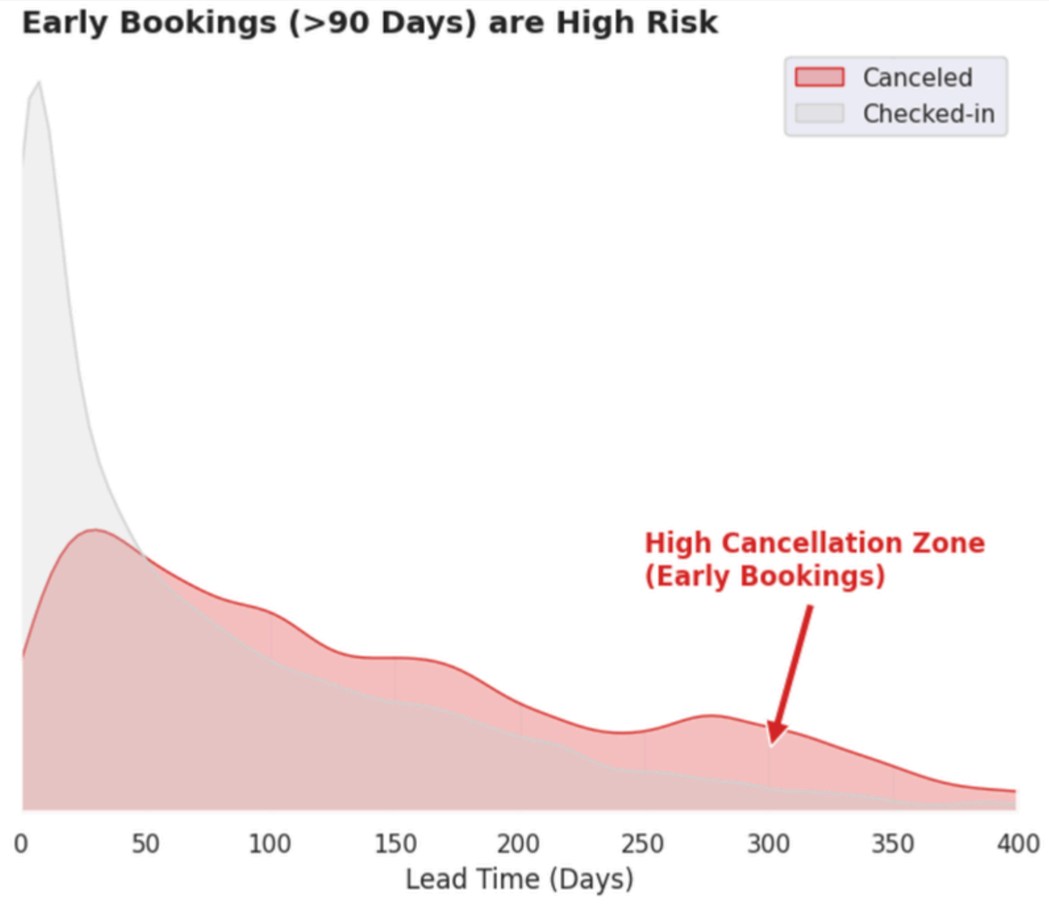
The 'Spring Spike' (April-June) is our most dangerous window

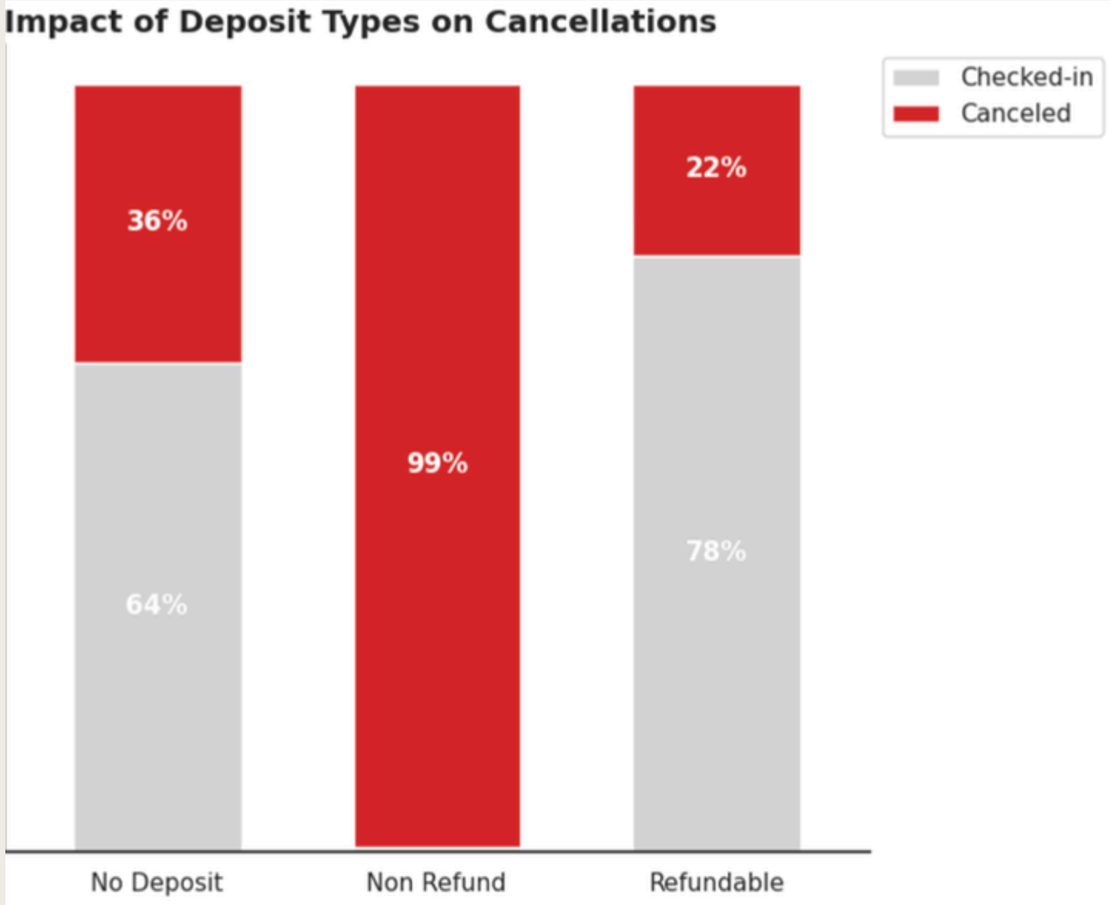


II. Exploratory Data Analysis

2. Drivers of Cancellation

Early bookings and 'Non-Refundable' groups create uncertainty

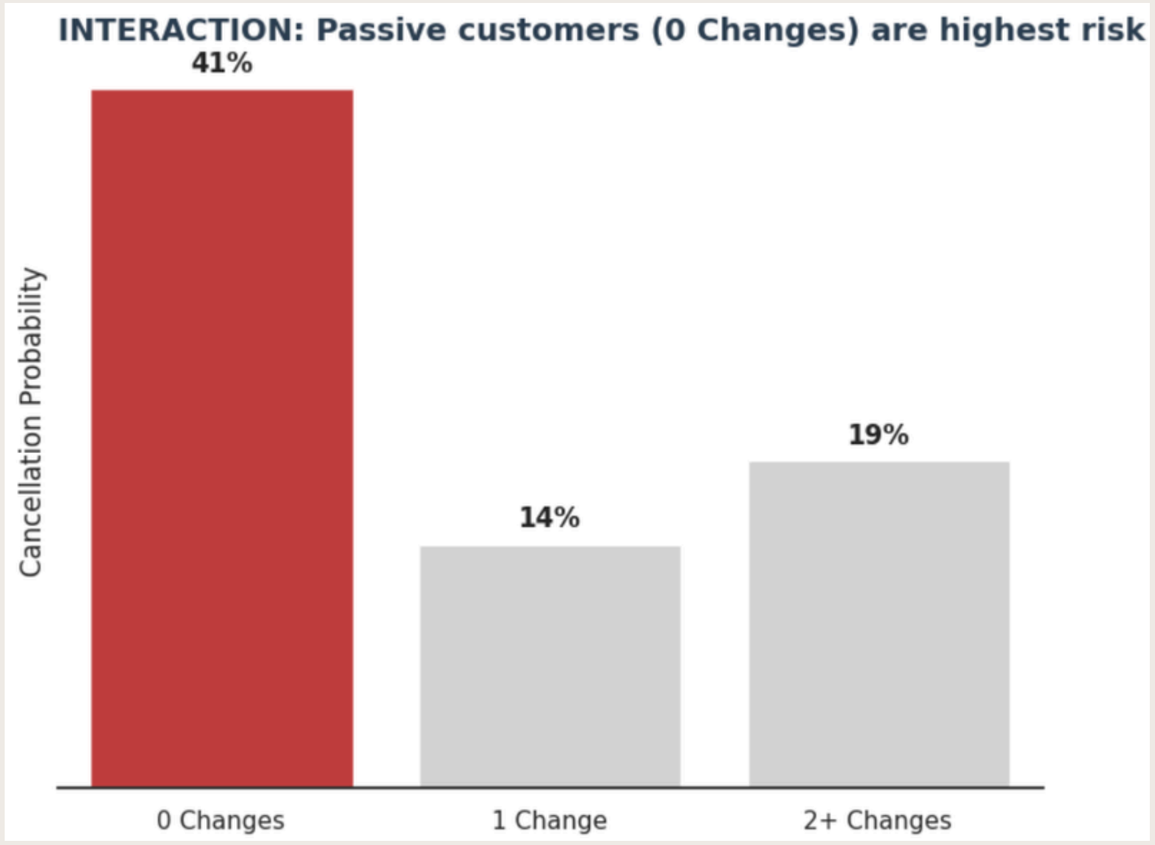




II. Exploratory Data Analysis

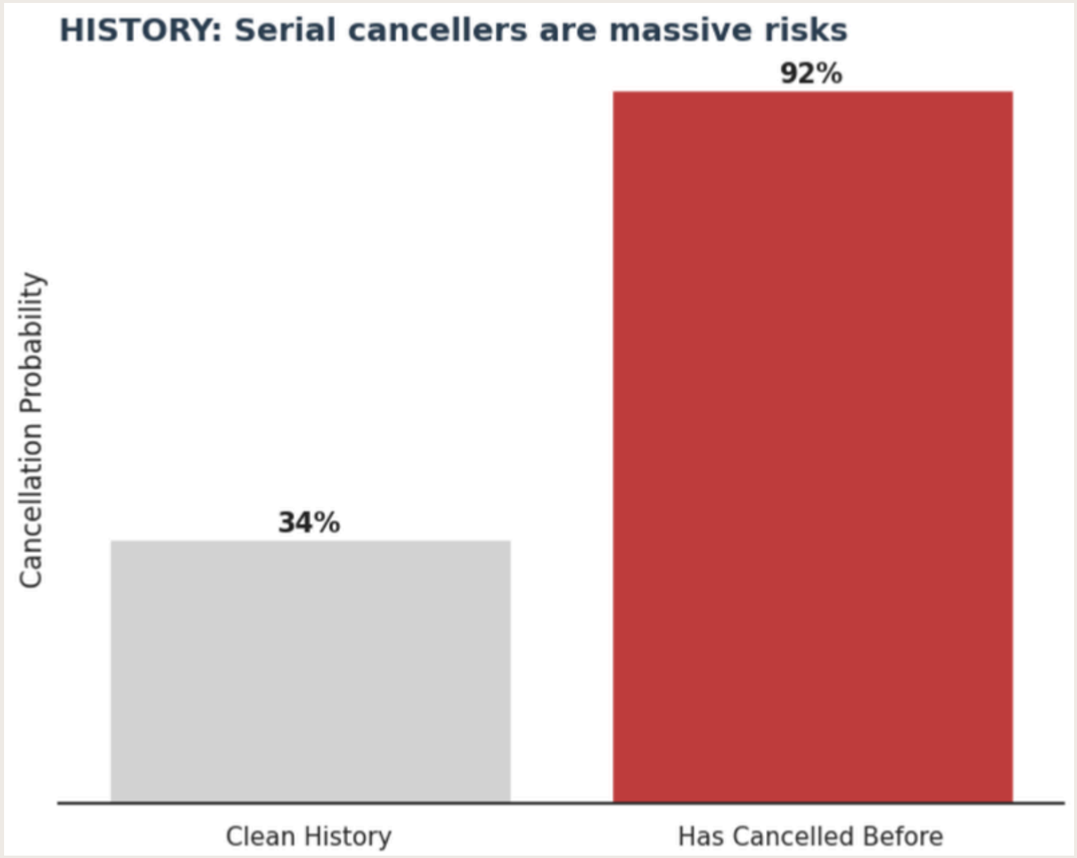
2. Drivers of Cancellation

Customer behavior (changes) and history are stronger predictors than demographics.



The "Silence" Signal

- Passive bookings (0 changes) carry the highest risk. Every modification reduces the probability of cancellation by half.



The "Repeat Offender" Signal

- Cancellation is a habit. Previous history is the single strongest indicator of unreliability.

III. Data Preprocessing

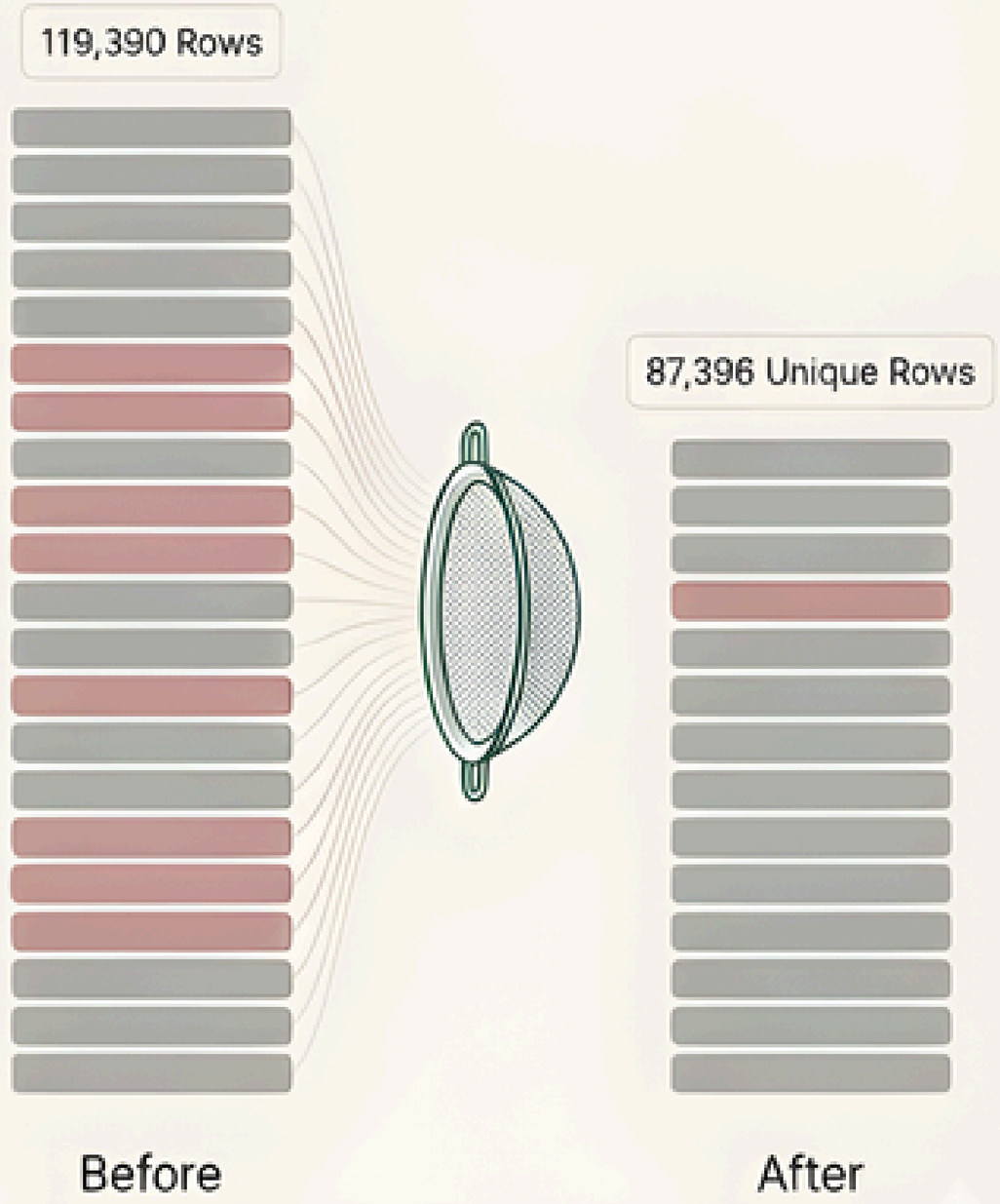
1. Checking Duplicates

- **The Problem:** Duplicates can negatively affect model learning by artificially inflating the importance of repeated patterns.
- **Analysis:** An initial check identified 31,994 duplicated rows within the dataset.
- **Action Taken:** All duplicate rows were removed to ensure the integrity of the training data.

Quantified Impact

119,390 → 87,396

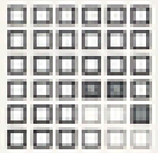
A reduction of over 25% in total rows.



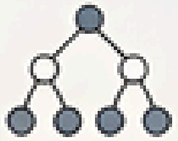
2. Checking Missing Values

The Challenge: Four key attributes contained missing values, each requiring a distinct, context-aware solution.

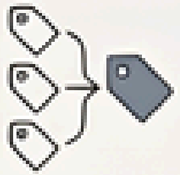
Company: With **~94%** of values missing (82,137 rows), the column offered little predictive power.
Action: Column dropped completely.



Children: Only 4 rows were missing.
Action: Imputed with the column's mean value to preserve data integrity with minimal distortion.



Country: With 452 missing rows (~0.5%), a more sophisticated approach was warranted.
Action: A Random Forest model was used for impuitation, leveraging other variables to make accurate predictions.



Agent: Bookings with 'null' simply meant no agent was used.
Action: All 333 unique agent codes were consolidated into a single 'agent' category, and nulls were treated as 'no agent', simplifying the feature space.

3. Detecting Outliers

Inter Regular, charcoal (#212121)

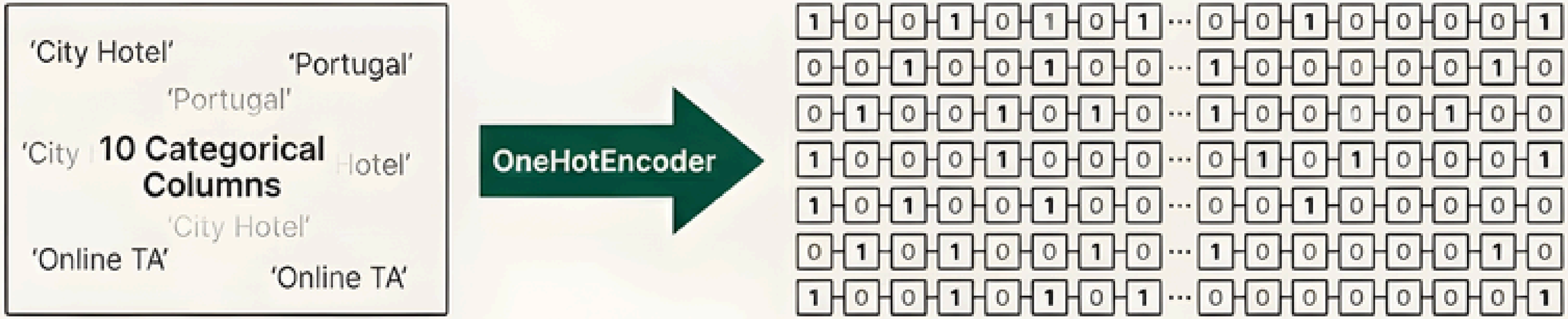
The Rationale: Outliers can skew model training. However, it's crucial to distinguish between genuine data and entry errors.
Analysis: Nine attributes were found to contain outliers.
Action & Justification:

- **Removed**: Rows with extreme values in 'Adults', 'Children', 'Babies', and 'adr' were removed. These were deemed highly improbable and likely the result of typing errors.
- **Retained**: Outliers in the remaining variables were kept, as they represented realistic, albeit infrequent, scenarios that could contain valuable information for the model.

Outlier Analysis

Attribute	Number of Outliers
lead_time	2396
stays_in_weekend_nights	220
stays_in_week_nights	1531
adults	22899
children	8368
babies	914
adr	2490
required_car_parking_spaces	7313
total_of_special_requests	2673

4. Encoding Categorical Variables



The Result:

- The original ten categorical columns were removed.
- The newly created dummy columns were added.
- The dataset's feature space expanded significantly to **231 independent variables**, making it ready for model training.

Other Steps Before Building Model

Size

87,379 rows | 231 columns

A robust dataset of unique records and numerically encoded features.

Data Split

80% / 20%

A standard split was used to create training and testing sets for robust model evaluation.

Target Variable Imbalance

72.5% Non-Cancellations

The class imbalance was noted and accounted for during the modelling process.

Feature Scaling

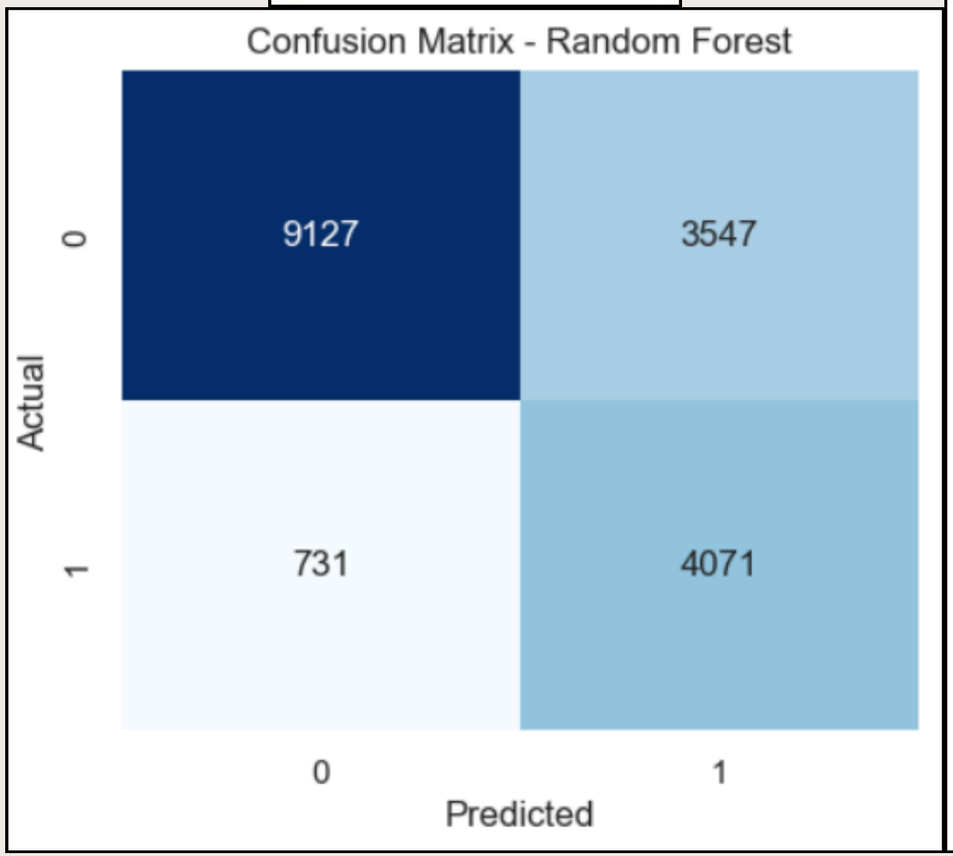
Standardized Range

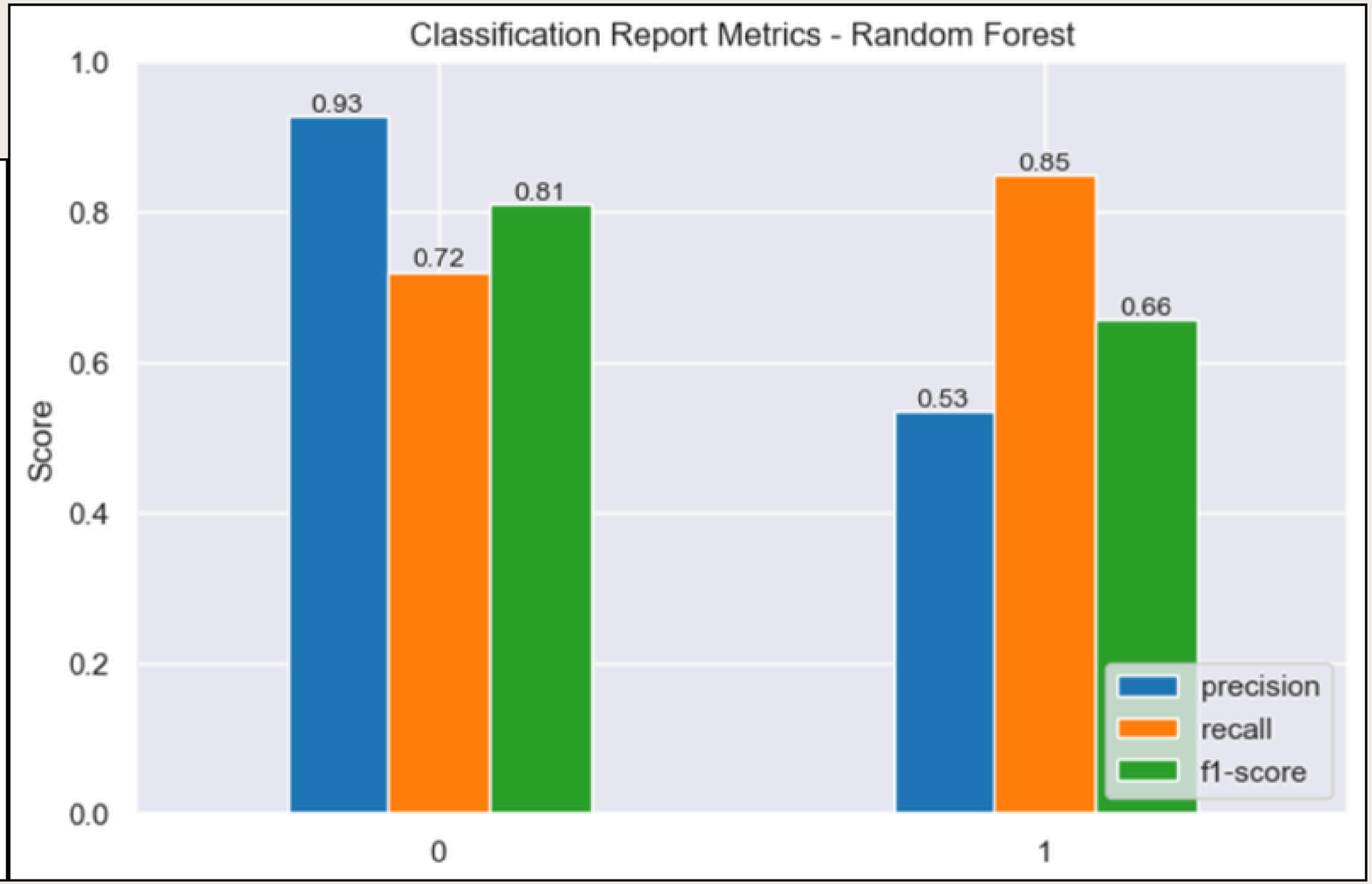
Numerical features were scaled to ensure stable and accurate algorithm performance.

Random Forest Classification

Overall Accuracy

76%

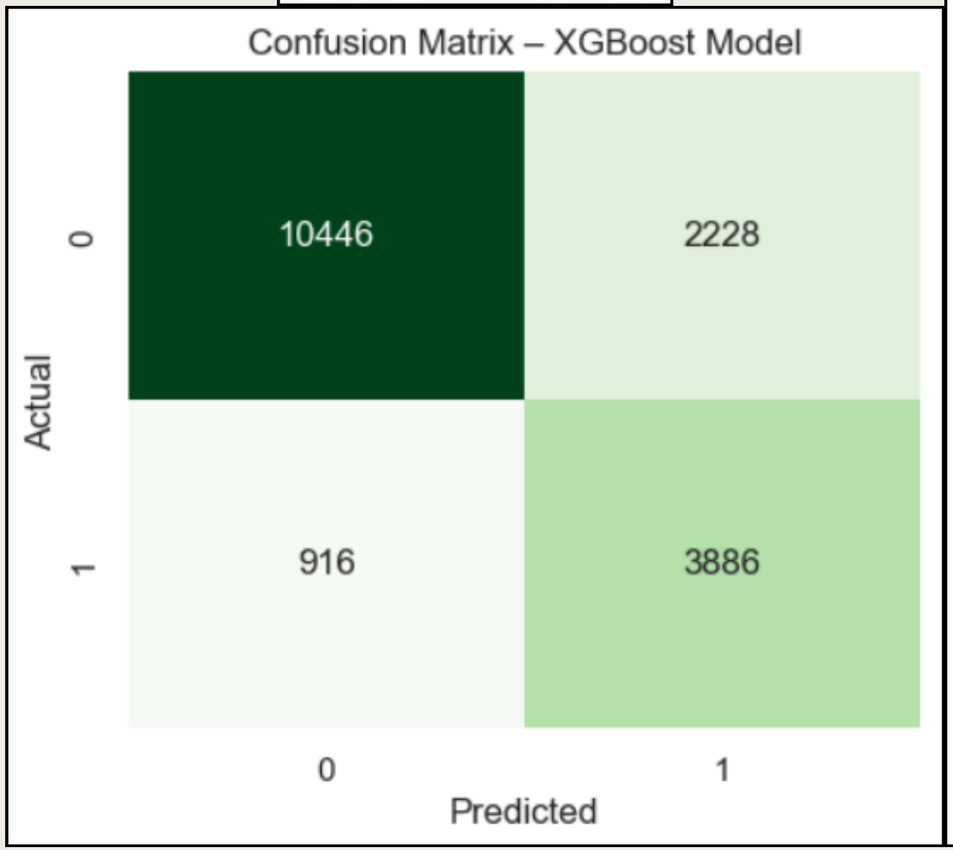


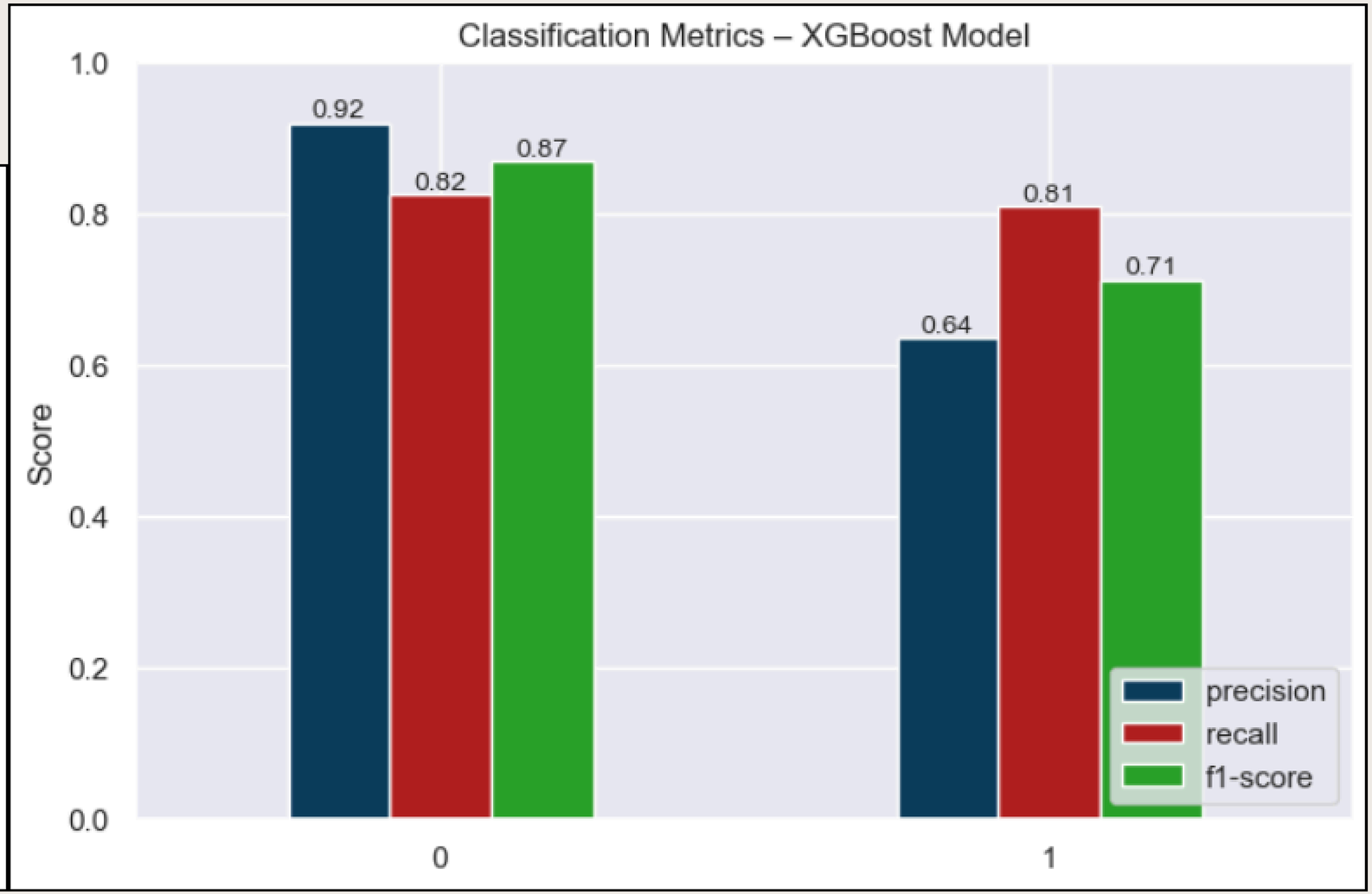


XGBoost Classification

Overall Accuracy

82%

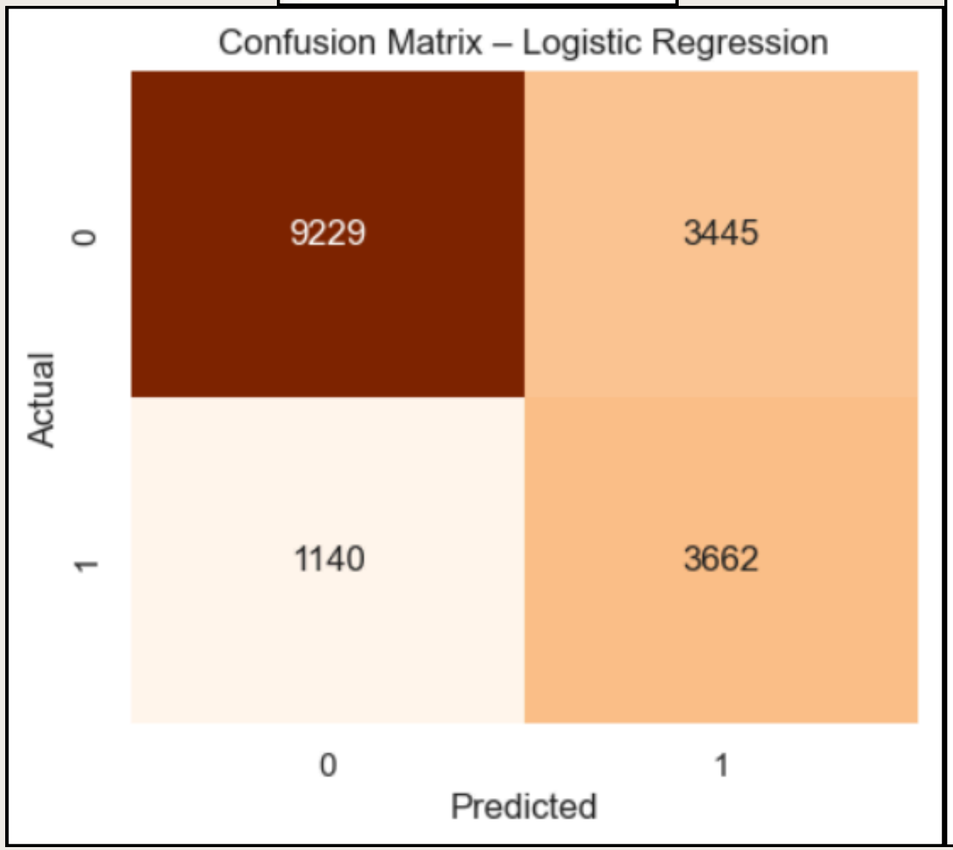


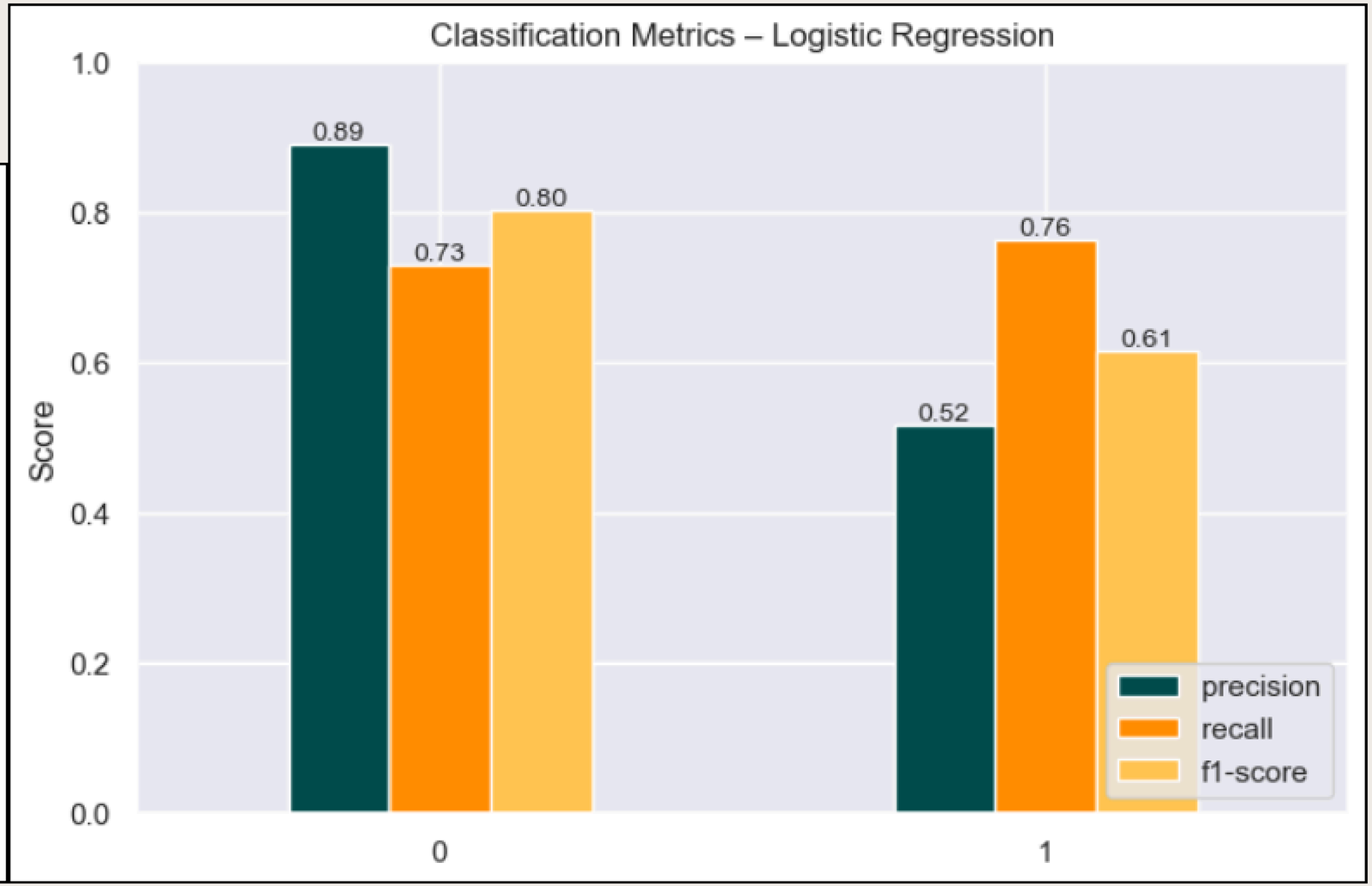


Logistic Regression Classification

Overall Accuracy

74%





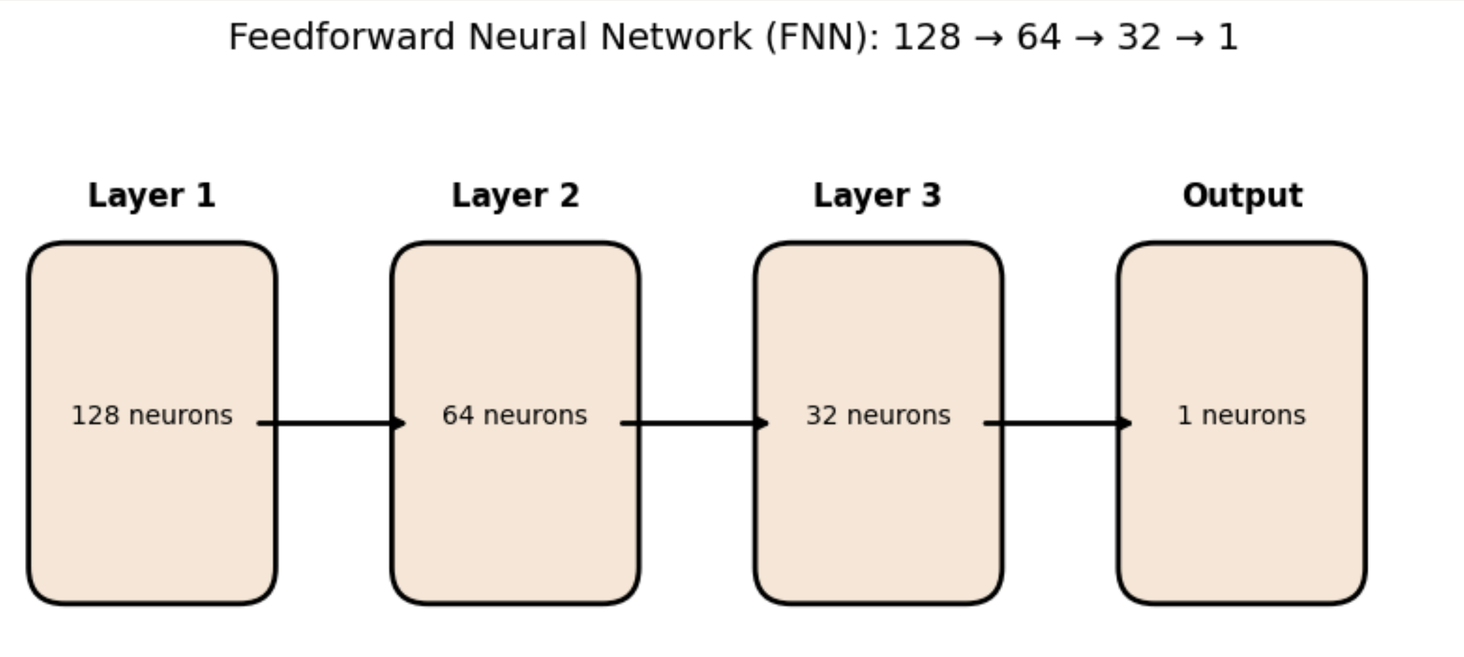
The Model Comparision

Focus: Evaluating the models on their ability to predict the minority class (Class 1: Cancelled), which is often the primary business objective.

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
XGBoost	**0.818**	**0.631**	**0.811**	**0.710**
Random Forest	0.761	0.541	0.853	0.662
Logistic Regression	0.741	0.519	0.766	0.619

VI. Deep Learning & Business Interpretation

1. Build and train the Deep Learning model



📌 **Model Architecture:**

- Feedforward Neural Network (FNN)
- 4 layers: 128 → 64 → 32 → 1
- Designed for binary classification

📌 **Regularization:**

- 30% Dropout applied to hidden layers to reduce overfitting

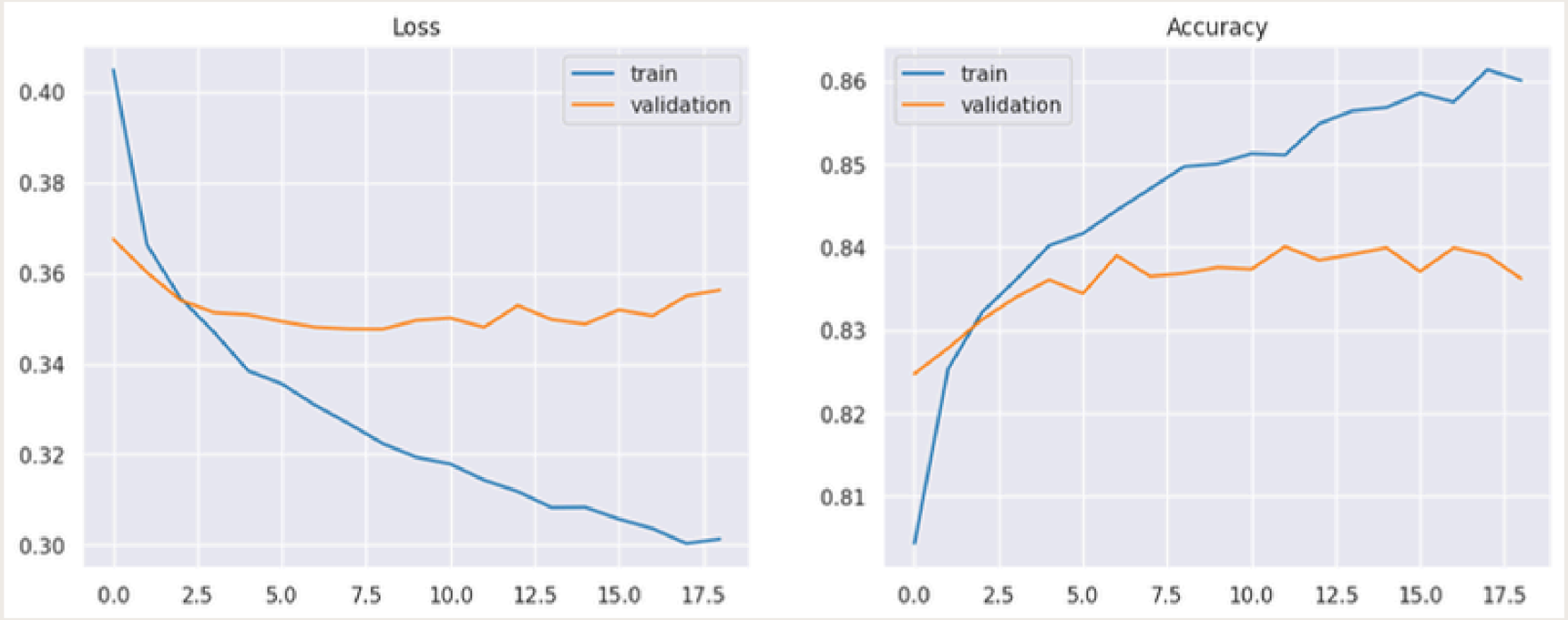
📌 **Training Setup**

- Optimizer: Adam
- Loss Function: Binary Cross-Entropy
- Evaluation Metric: Accuracy

📌 **Early Stopping**

- Monitors validation performance
- Stops training after 5 epochs with no improvement
- Ensures optimal and well-generalized model selection

2. Model evaluation & comparison



📌 **Training & Validation Performance**

- Training loss decreased consistently, indicating effective learning
- Validation accuracy stabilized around ~84% after 5–10 epochs
- Validation loss plateaued → model reached performance saturation

📌 **Overfitting Observation**

- Slight gap between training and validation curves → minor overfitting
- Early Stopping successfully prevented further degradation.
- Best model weights were restored automatically.

Metric	Value
Test Accuracy	83.15%
Test Loss	0.3534

→ Results are consistent with validation performance → good generalization

	Precision	Recall	f1-score	Support
0	0.87	0.9	0.89	12674
1	0.71	0.65	0.68	4801
Accuracy			0.83	17475
Macro avg	0.79	0.77	0.78	17475
Weighted avg	0.83	0.83	0.83	17475

📌 **Overall Model Performance**

- The model achieves a strong Overall Accuracy of 83%, indicating good predictive performance on the test set.
- Weighted Avg Precision, Recall, and F1-score all ≈ 0.83, showing consistent and balanced behavior across the whole dataset.

📌 **Class-Specific Insights:**

Class 0 (Majority Class)

- Precision: 0.87 | Recall: 0.90 | F1: 0.89
- The model performs very well at correctly identifying Class 0.
- High recall means it captures most Class 0 samples (low false negatives).

Class 1 (Minority Class)

- Precision: 0.71 | Recall: 0.65 | F1: 0.68
- Performance is noticeably lower for Class 1.
- Lower recall indicates the model misses some true Class 1 cases, likely due to class imbalance.
- Precision remains acceptable, meaning predictions of Class 1 are reasonably reliable.

📌 **Macro vs. Weighted Performance**

- Macro Avg F1: 0.78 → shows moderate performance when treating both classes equally.
- Weighted Avg F1: 0.83 → boosted by the majority class (Class 0), reflecting dataset imbalance.

3. Conclusion & Business Recommendation

📌 **Conclusion**

- The model achieves a solid 83.15% accuracy, making it reliable enough to support data-driven decision-making.

- However, performance on Class 1 (canceled bookings) reveals important business implications:
  - Recall = 65% → the model misses 35% of actual cancellations, leading to lost revenue opportunities because the company cannot intervene in time.
  - Precision = 71% → 29% of predicted cancellations are incorrect, causing unnecessary retention efforts and additional operational costs.

📌 **Business Impact:**

- False Negatives (missed cancellations) = potential revenue leakage

- False Positives (incorrect alerts) = wasted marketing and manpower

- This imbalance shows that improving Class 1 detection is critical for maximizing financial impact.

📌 **Recommendation:**

- Prioritize increasing Recall for Class 1 (target ≥ 80%).

- A higher Recall ensures the company identifies more at-risk bookings, allowing timely intervention to:
  - Reduce preventable cancellations
  - Protect revenue
  - Optimize retention resource allocation

Reference

Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. Data in Brief, 22, 41–49. https://doi.org/10.1016/j.dib.2018.11.126

Thank You.