EIU

TRƯỜNG ĐẠI HỌC
QUỐC TẾ
MIỀN ĐÔNG
EASTERN
INTERNATIONAL
UNIVERSITY

15 Year Anniversary 2010 - 2025

BECAMEX BUSINESS SCHOOL

Hotel Booking Cancellation Prediction

Course: MIS 451 - Machine Learning for Business

Lecturer: Ms. Huynh Gia Linh

Mr. Dang Thai Doan

Name	IRN
Nguyen Thanh Dat	2132300562
Nguyen Hoang Vinh	2132300522
Nguyen Quang Truong	2132309001

Semester: Quarter 1, Years 2025-2026

Table of contents

I. Business Context and Objectives

1. Business Context

Industry Overview: The hospitality industry faces a significant challenge with booking cancellations. Cancellations can lead to revenue loss due to unsold rooms (if not resold in time) or operational inefficiencies (overstaffing or under-preparedness).

The Stakeholders:

- Hotel Management: Interested in maximizing occupancy and Average Daily Rate (ADR).
- Revenue Managers: Need to decide on overbooking strategies and pricing adjustments based on cancellation probability.
- Front Desk/Operations: Need to plan for daily arrivals and room assignments.

The Core Problem: The hotels are experiencing a high rate of cancellations (approx. 37%), which creates uncertainty in inventory management.

2. Project Objective

Develop a classification machine learning and deep learning model to predict whether a specific hotel booking will be canceled or not. By predicting the likelihood of cancellation for each booking, the hotel can:

- Optimize Inventory: Implement calculated overbooking strategies to fill rooms that are predicted to be canceled.
- Revenue Protection: Offer targeted incentives (e.g., discounts for non-refundable deposits) to customers with a high probability of cancellation.
- Resource Planning: Better estimate the actual number of guests arriving to manage staffing and food supplies efficiently.

II. Data collection section

1. Data Source

The primary source of this dataset is the article titled *"Hotel booking demand datasets,"* authored by Nuno Antonio, Ana de Almeida, and Luis Nunes, and published in the journal *Data in Brief* in 2019. The data was extracted directly from the Property Management System

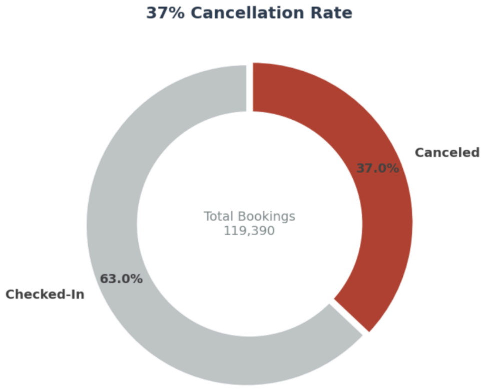
(PMS) databases of two distinct European hotels located in Portugal. One is a resort hotel (H1) likely situated in the Algarve region, and the other is a city hotel (H2) located in Lisbon.

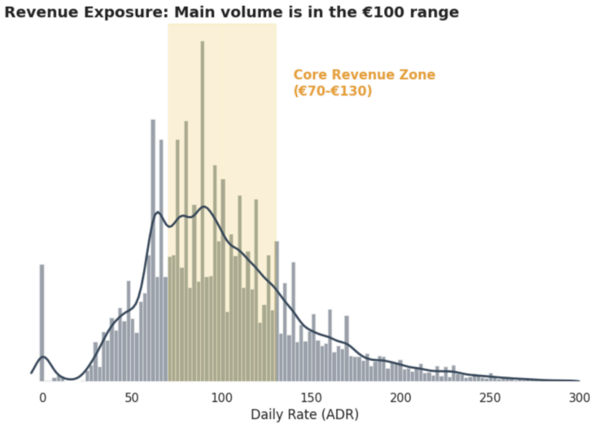
2. Data Description

The dataset serves as a comprehensive record of booking activities for the two hotels over a period of 26 months, from July 1, 2015, to August 31, 2017. It contains a total of 119,390 bookings. The data is structured into 32 variables (columns) that capture the entire lifecycle of a booking, from the initial reservation to the final status of check-out or cancellation. The variables include a mix of guest demographics, booking specifications, and transaction details. The target variable for classification is “is_canceled”.

III. Exploratory Data Analysis (EDA)

1. Current Business Landscape

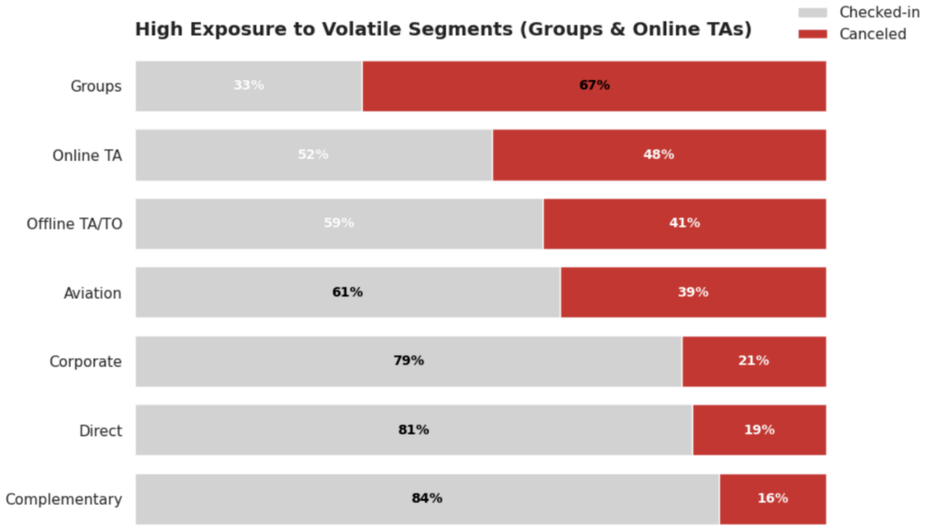




We identified a severe 37% cancellation rate that critically impacts operations. This volatility is concentrated within the €70–€130 ADR band, proving that revenue leakage is hitting the hotel's core business rather than just budget segments. This makes predictive mitigation a top strategic priority.

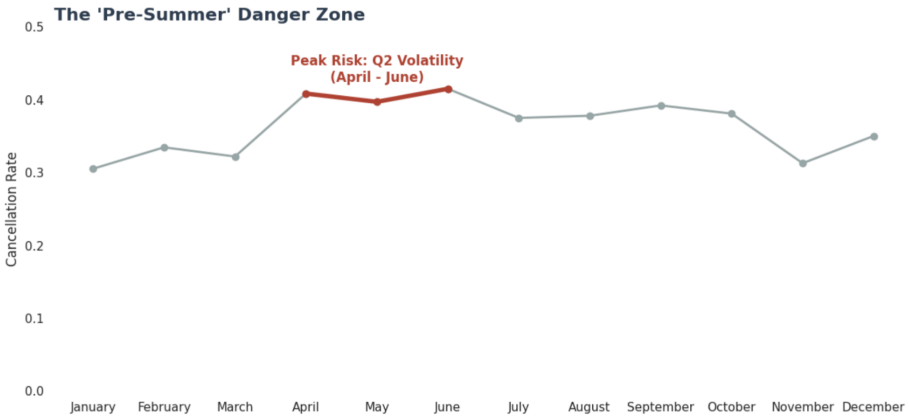
2. Drivers of Cancellation

A. Source Analysis



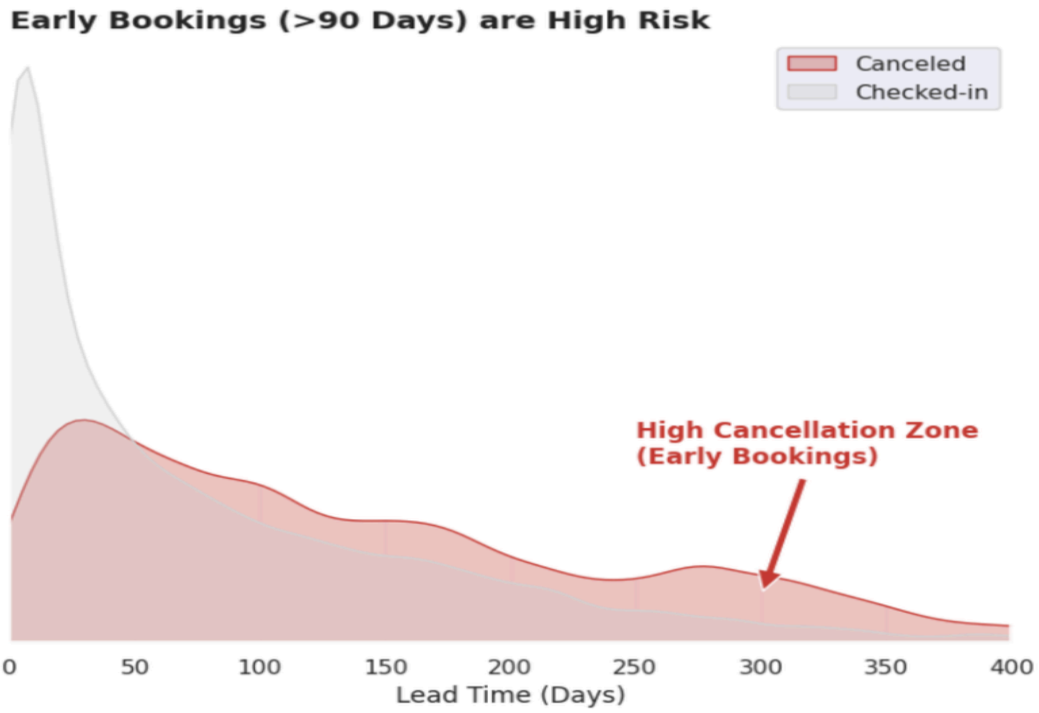
Decomposing the data by Market Segment exposes the structural origins of the cancellation problem. High-risk behavior is concentrated in the 'Groups' and 'OTA' segments, characterized by excessive rates and the volatility of bulk cancellations. Conversely, 'Direct' and 'Corporate' bookings show cancellation rates well below the 20% threshold.

B. Temporal Analysis



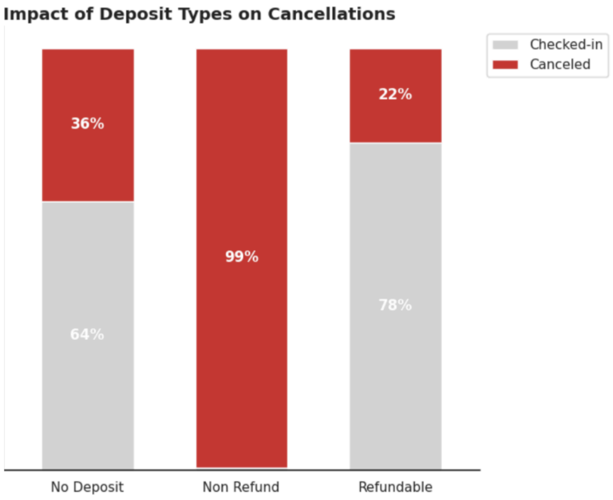
Analysis uncovers a 'Q2 Anomaly,' with risk peaking in April–June rather than the high season. This 'Pre-Summer Churn' indicates that spring bookings are often speculative placeholders, creating a false demand signal that evaporates before the core revenue months of July and August.

C. Lead Time Correlation



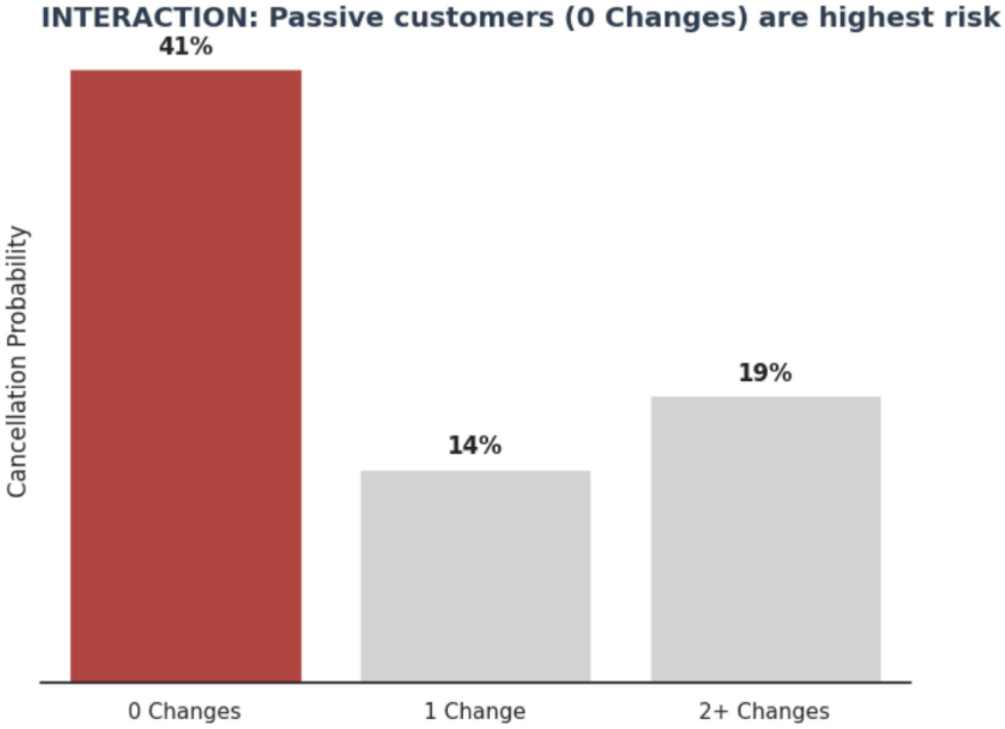
Risk increases linearly as the booking window extends. While bookings under 30 days are highly stable, those exceeding six months face 'Long-Term Instability,' where the probability of cancellation actually surpasses the probability of arrival. This identifies early bookings as 'soft' revenue compared to the firm commitment of short-term reservations.

D. The Deposit Anomaly



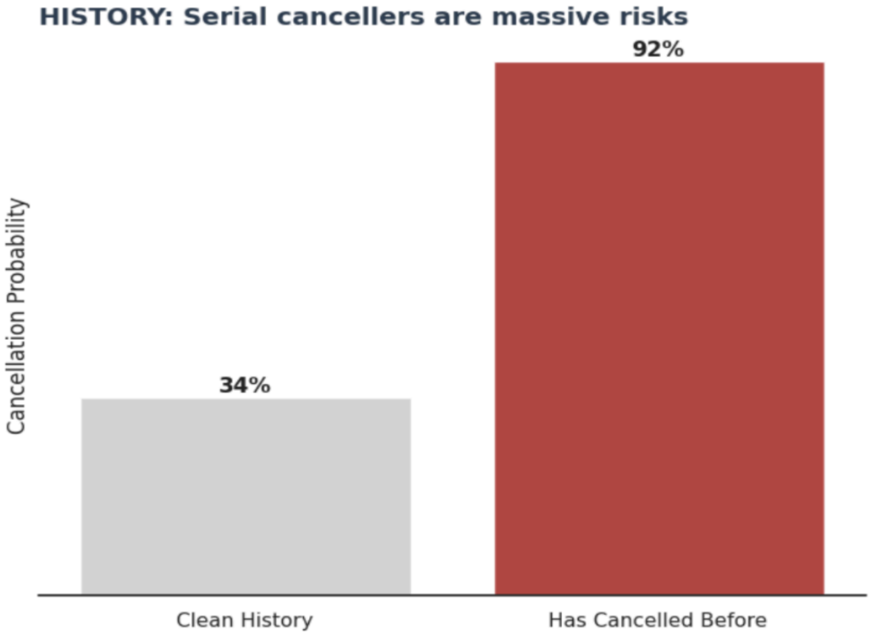
Contrary to expectations, the 'Non-Refundable' category shows a cancellation rate of nearly 99% in specific segments. Data suggests these are administrative voids or group placeholders rather than genuine prepaid reservations. This creates a dangerous 'false security' in the pipeline, where inventory marked as secure is actually the most volatile.

E. The "Passive Guest" Profile



Guest interaction is a strong predictor of stability. While bookings with zero changes carry the highest risk, those with just one modification show a drastic drop in cancellation probability. This confirms that 'dormant' bookings are volatile, while active engagement signals firm commitment.

F. History of Cancellation



Past actions strongly predict future risk. A single prior cancellation correlates with an exponential rise in churn probability compared to the baseline. This confirms that cancellation is often a repetitive behavior, making guest history a critical filter for assessing revenue quality.

IV. Data Preprocessing

1. Checking duplicates

The first step of our data preprocessing was cleaning the dataset to ensure that the information used for model training was reliable and consistent. We began by checking duplicated rows and found 31,994 duplicates. Since duplicates can negatively affect model learning by repeating the same patterns, we decided to remove all of them. After removing duplicates, the dataset was reduced to 87,396 rows.

2. Checking missing values

Attribute	Missing Values
children	4
country	452

agent	12,193
company	82,137

Next, we examined missing values. The attribute “Company” contained 82,137 missing rows, accounting for around 94% of the entire column. Because the proportion of missing data was too large, we decided to drop this column completely. The “Children” column only had 4 missing rows, so we used the mean value to fill these gaps since the number was very small and unlikely to affect the overall distribution. For the “Country” column, which had 452 missing rows (about 0.5%), we used a Random Forest model to impute the missing values. This method was chosen because it can learn from other variables and provide more accurate predictions than simple imputation. For the “Agent” column, we found 333 unique agent codes, while many rows had null values. These null values represent bookings with no agent. To simplify the dataset and reduce unnecessary complexity, we transformed every non-null value into a single category called “agent”.

3. Checking outliers

attributes	num_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

We then checked for outliers across the dataset. There were nine attributes with outliers, but after exploring the data, we realized that some columns such as “Adults,” “Children,” “Babies,” and “adr” had only a very small number of extreme values. Because these extreme values were most likely typing errors rather than real observations, we removed these rows to keep the dataset clean and valid. Outliers in the remaining variables were kept because they still represented realistic values and could be important for model learning.

4. Encoding categorical variables

After cleaning, we encoded the categorical variables. There were ten nominal categorical columns in the dataset. We used OneHotEncoder to create dummy variables for each category because machine learning models cannot work directly with text values. Once the encoding was completed, we added the newly created dummy columns to the dataset and removed the original unencoded categorical columns. As a result, the dataset expanded to 231 independent variables.

5. Splitting the dataset

Next, we prepared the dataset for training and testing. The final dataset included 87,379 rows and 231 independent columns. We split the data into two parts: 80% for training the model and 20% for testing. This split helps evaluate the model’s ability to generalize and perform well on unseen data.

6. Checking imbalance target variable

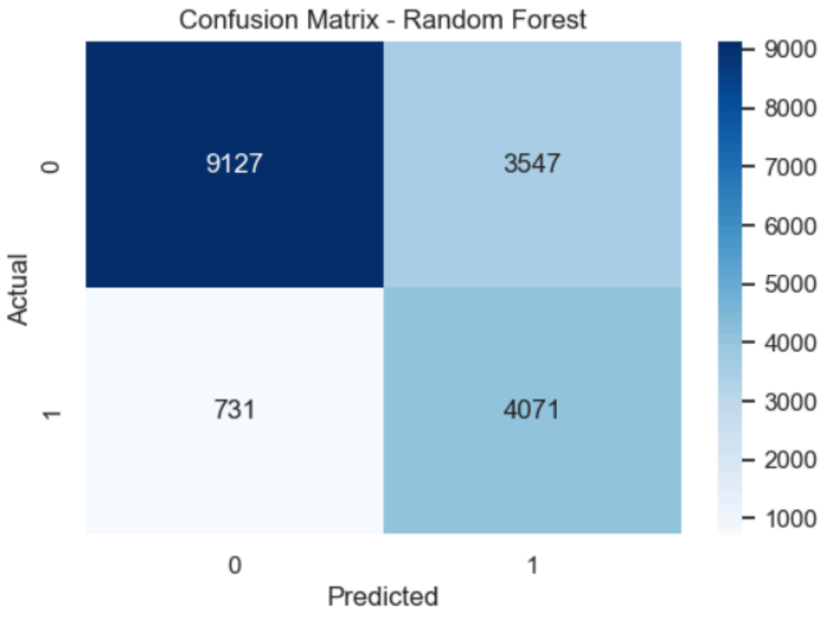
We also examined the class distribution of the target variable. The class labeled “0” made up around 72.5% of the dataset, which means the data is imbalanced. Imbalanced data can cause the model to focus too much on the majority class and ignore the minority class. Because of this, we applied specific techniques to help the model learn more fairly from both classes.

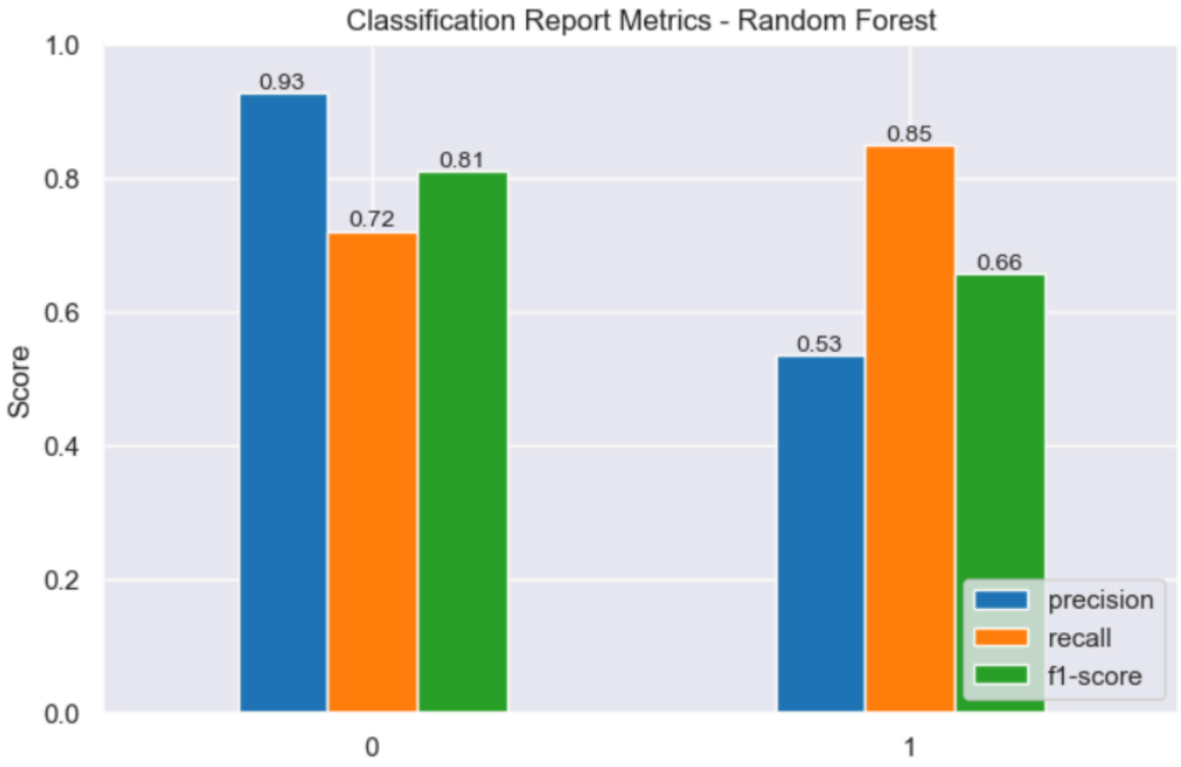
7. Scaling data

Finally, we performed data scaling before training certain machine learning models. The purpose of scaling is to ensure that all numerical features have a similar range of values. Many ML models, such as Logistic Regression, KNN, and SVM, can be affected when variables have very different scales. For example, a feature with large values may dominate another feature with smaller values, even if both are equally important. Scaling helps improve model accuracy, training stability, and the overall performance of distance-based or gradient-based algorithms.

V. Machine Learning Model

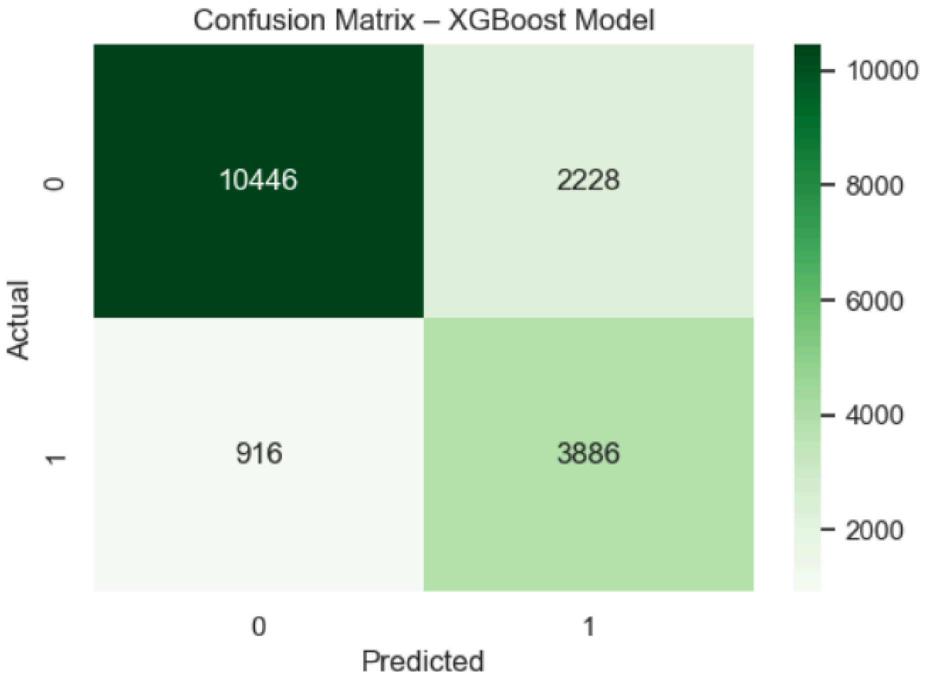
1. Random Forest

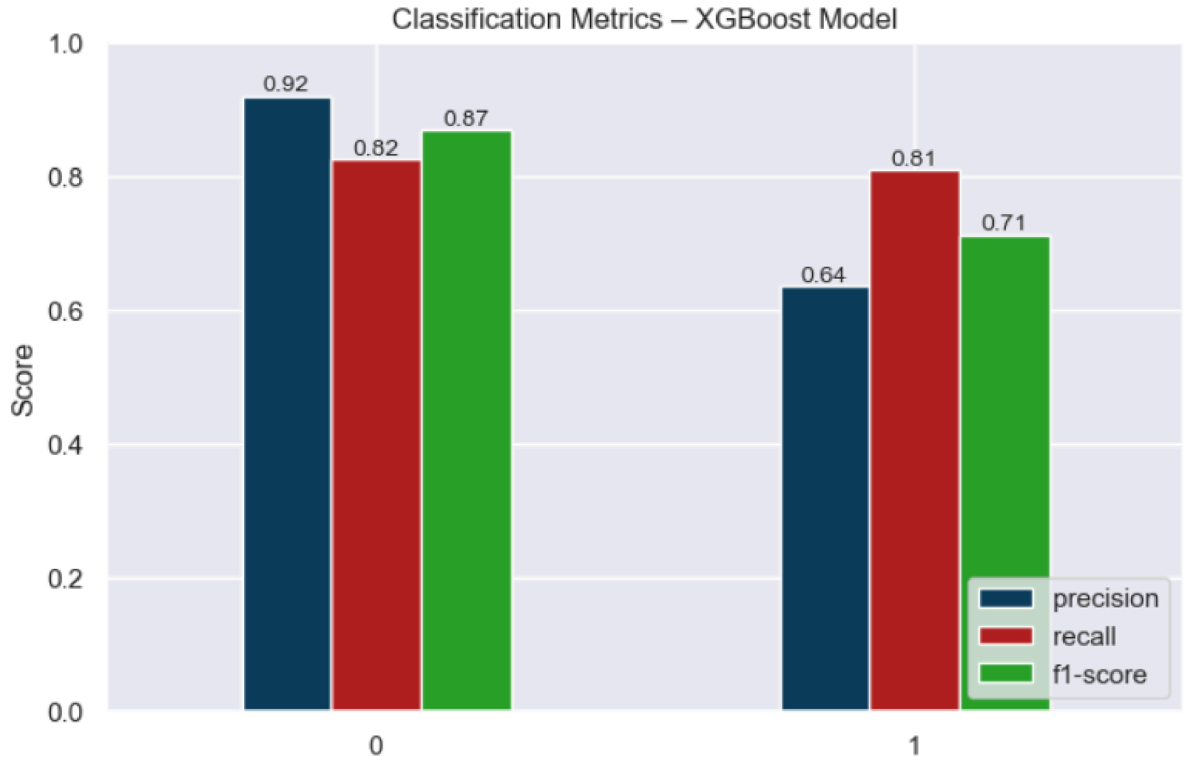




The Random Forest model shows an overall accuracy of 76%, which means it performs moderately well in predicting both classes. The model predicts class “0” with very high precision (0.93), but its recall is lower (0.73), suggesting that it correctly identifies most true class “0” cases but still misses some of them. For class “1,” the model has a lower precision of 0.54 but a high recall of 0.85. This means it can capture most canceled bookings (class “1”), although it also produces many false positives. Overall, the model is strong at detecting cancellations but struggles with precision, which reflects the impact of the imbalanced dataset.

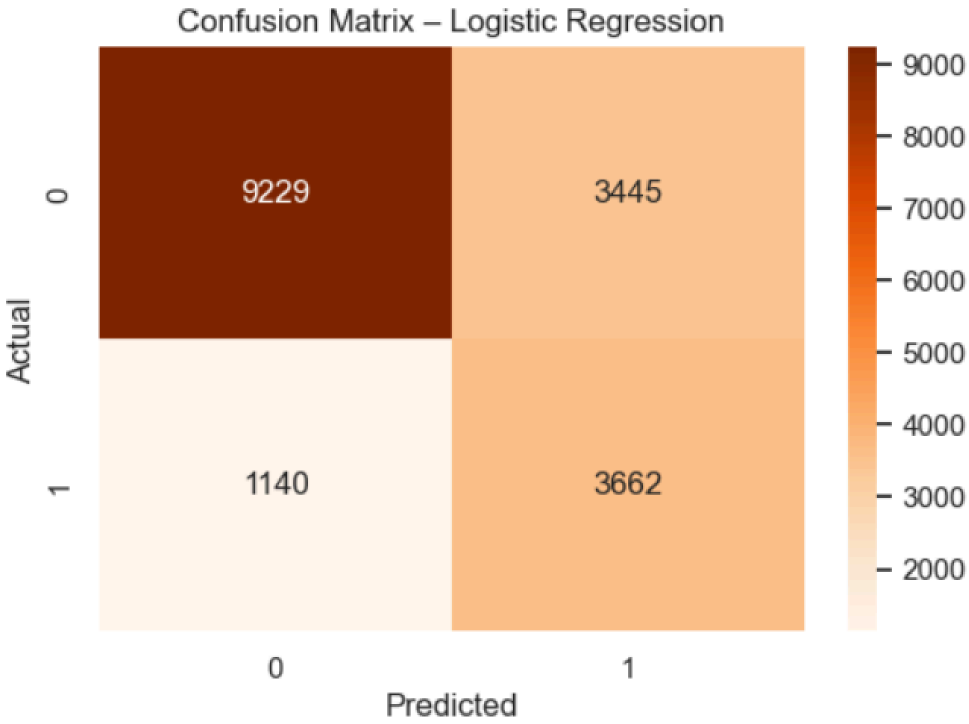
2. **XGBoost**

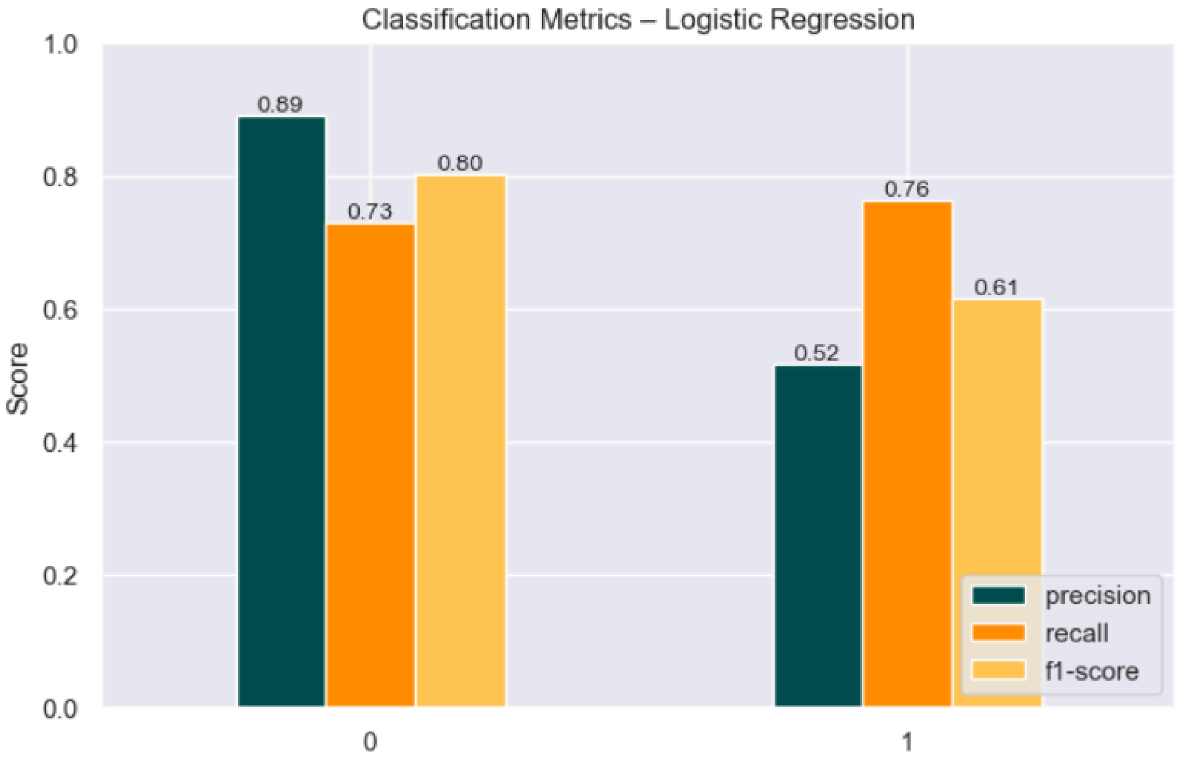




The XGBoost model reaches an accuracy of 82%, showing that it predicts the target variable quite well. For class “0,” the model has a precision of 0.92 and a recall of 0.82, meaning it correctly identifies most non-canceled bookings and keeps the number of incorrect predictions low. For class “1,” the precision is 0.63 and the recall is 0.81, which shows that the model can detect a large portion of canceled bookings, although some predictions still fall into the wrong class. Overall, the results suggest that XGBoost performs well in recognizing both classes and provides balanced and reliable predictions.

3. Logistic Regression

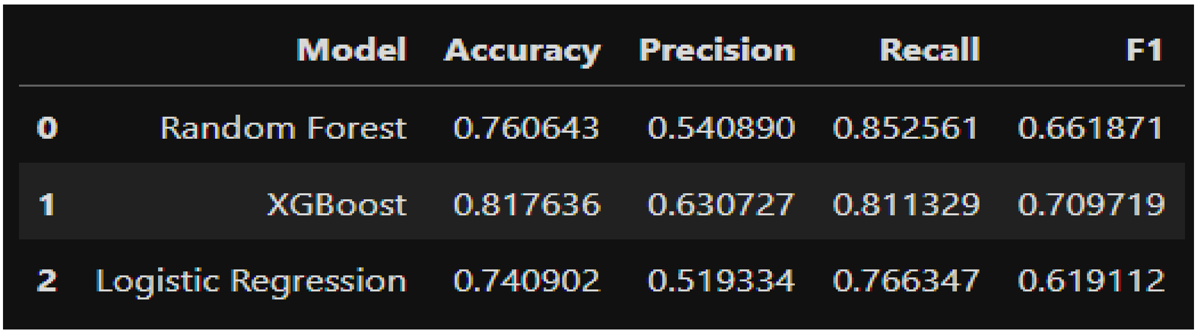




The Logistic Regression model achieves an accuracy of 74%, showing a moderate level of performance in predicting the target variable. For class “0,” the model has a precision of 0.89

and a recall of 0.73, which means it correctly identifies most non-canceled bookings but still misses some true cases. For class “1,” the precision is 0.52 and the recall is 0.77, indicating that the model detects a high number of canceled bookings but also produces many false positives. Overall, the results suggest that the model performs better at identifying class “0” but can still capture a good portion of class “1” cases.

4. Model Comparison



	Model	Accuracy	Precision	Recall	F1
0	Random Forest	0.760643	0.540890	0.852561	0.661871
1	XGBoost	0.817636	0.630727	0.811329	0.709719
2	Logistic Regression	0.740902	0.519334	0.766347	0.619112

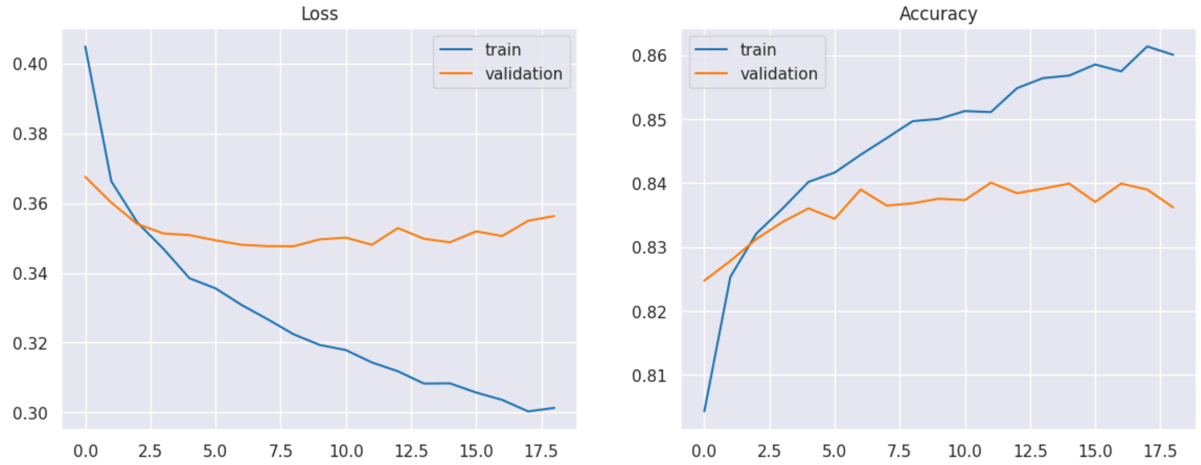
The comparison of the three models shows that XGBoost delivers the strongest overall performance, achieving the highest accuracy, precision, and F1-score among the models. It maintains a good balance between correctly predicting both classes while reducing errors. Random Forest performs moderately well, with high recall but lower precision, which means it identifies many true positive cases but also produces more false positives. Logistic Regression has the lowest accuracy and F1-score, showing that it captures general patterns but is less effective in producing accurate predictions compared to the tree-based models. Overall, XGBoost stands out as the most reliable model for this classification task based on the evaluation metrics.

VI. Deep Learning Model

1. Build and train the Deep Learning model

The final model is a Feedforward Neural Network (FNN) consisting of four layers, specifically designed for binary classification tasks (such as customer churn prediction or transaction categorization). To ensure the model learns robust patterns and avoids over-relying on noise in the training data, we incorporated Dropout (30% random deactivation), a crucial technique for preventing overfitting. The training process utilized the Adam optimizer and the standard Cross-entropy loss function. Critically, we implemented Early Stopping to monitor the performance on the validation set and halt training if the performance failed to improve for five consecutive epochs. This mechanism guarantees that we select the most generalized and optimal version of the model.

2. Model evaluation & comparison



The analysis of the Loss and Accuracy plots confirms effective learning: the training loss consistently decreased, while the validation loss and accuracy stabilized around 84% after roughly 5-10 epochs. The slight divergence between the training and validation curves indicates minor overfitting, but the Early Stopping mechanism intervened effectively, restoring the best-performing version of the model. When evaluated on the independent test dataset, the model achieved an overall Test Accuracy of 83.15% and a Test Loss of 0.3534.

	Precision	**Recall**	**f1-score**	**Support**
0	0.87	0.90	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

For a more comprehensive evaluation, which is vital for potentially imbalanced datasets, we examined the classification report. The model performed exceptionally well in identifying Class 0 (with 87% Precision and 90% Recall). However, performance on Class 1 (the minority or critical target class, e.g., customer churn) was moderate: Precision was 71% and Recall was 65%. This means that while the model is reasonably accurate when it predicts a sample is Class 1 (71% of those predictions are correct), it misses 35% of the actual Class 1 samples.

3. Conclusion & business recommendation

With an overall accuracy of 83.15%, the model is robust enough to support data-driven business decisions.If Class 1 represents the booking is canceled, the 65% Recall is a key area of concern. Missing 35% of the bookings that were actually canceled (False Negatives) translates directly to lost revenue opportunities as the company fails to intervene and attempt to save those bookings. Conversely, the 71% Precision means that 29% of the bookings predicted to be canceled (False Positives) will receive unnecessary retention offers or interventions, resulting in wasted operational and marketing expenditure.

While the current model is good, to maximize business impact, the focus should be shifted to improving the Recall for Class 1 to a higher threshold (e.g., above 80%). Achieving this will enable the company to identify and potentially save a significantly larger portion of at-risk bookings, effectively minimizing revenue leakage caused by cancellations.

VII. 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