Predictive Modeling of Hotel Booking Cancellations: A Machine Learning Approach

DATA MINING AND MACHINE LEARNING PROJECT

Leonardo Ceccarelli

Master’s Degree in Artificial Intelligence
and Data Engineering

INTRODUCTION

Frequent Cancellations force hotels to rely on:

- **Overbooking**, which may lead to service denial, poor guest experience, and reputational damage.
- **Rigid cancellation policies**, which can deter customers and reduce booking volumes.

Consequences:

- Lower occupancy accuracy and revenue predictability
- Loss of customer trust and future bookings
- Inefficient room allocation and pricing decisions

Predicting cancellations in advance allows hotel managers to:

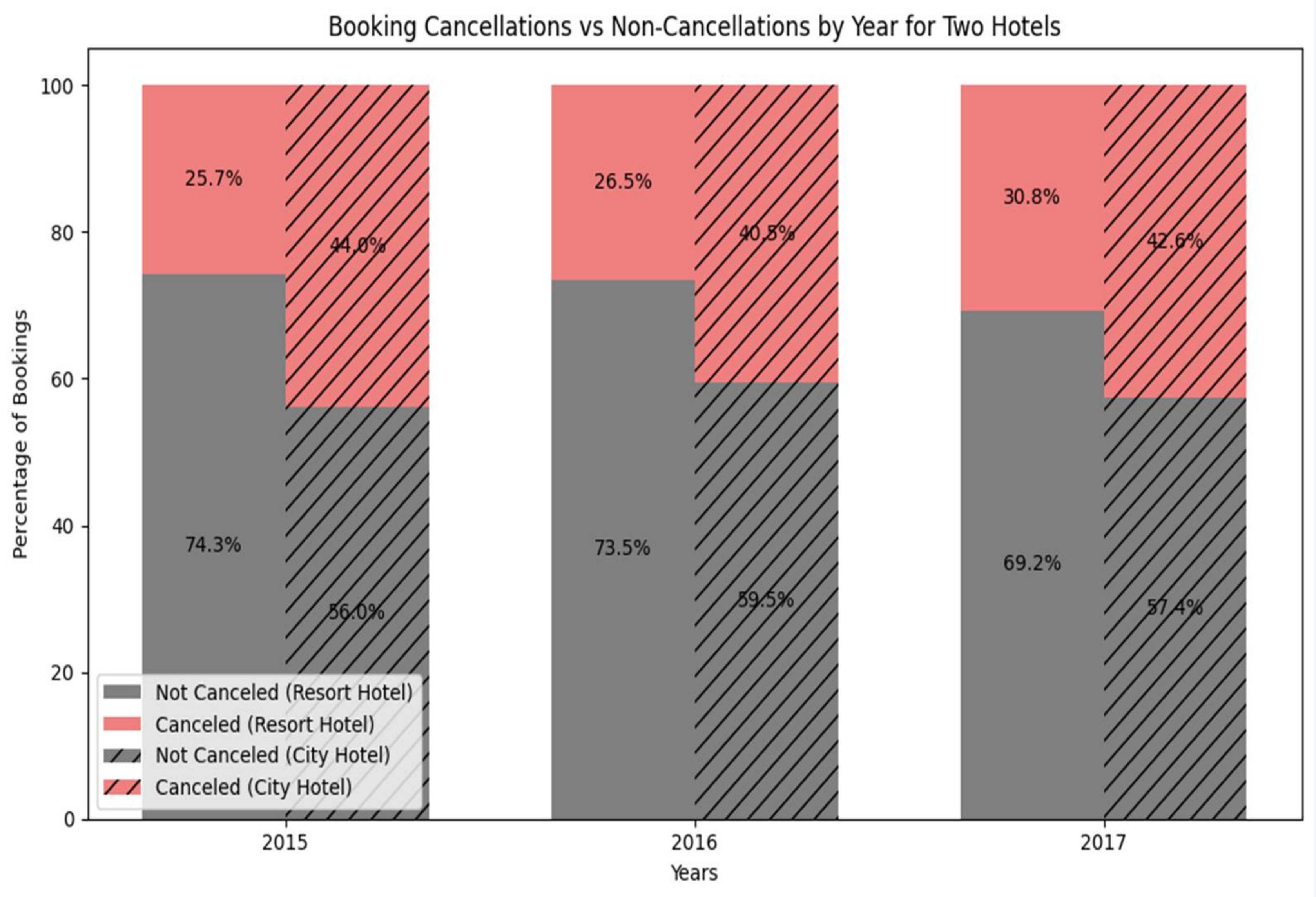
- Proactively mitigate losses with offers or upgrades
- Adjust pricing and overbooking strategies more precisely

DATASET

- This project uses a real-world dataset of hotel bookings from two distinct hotel types:
- **Resort Hotel** (H1): 40,060 records
- **City Hotel** (H2): 79,330 records
- Covers bookings scheduled to arrive between July 2015 and August 2017
- Includes both confirmed and canceled reservations

Feature Composition (31 Variables):

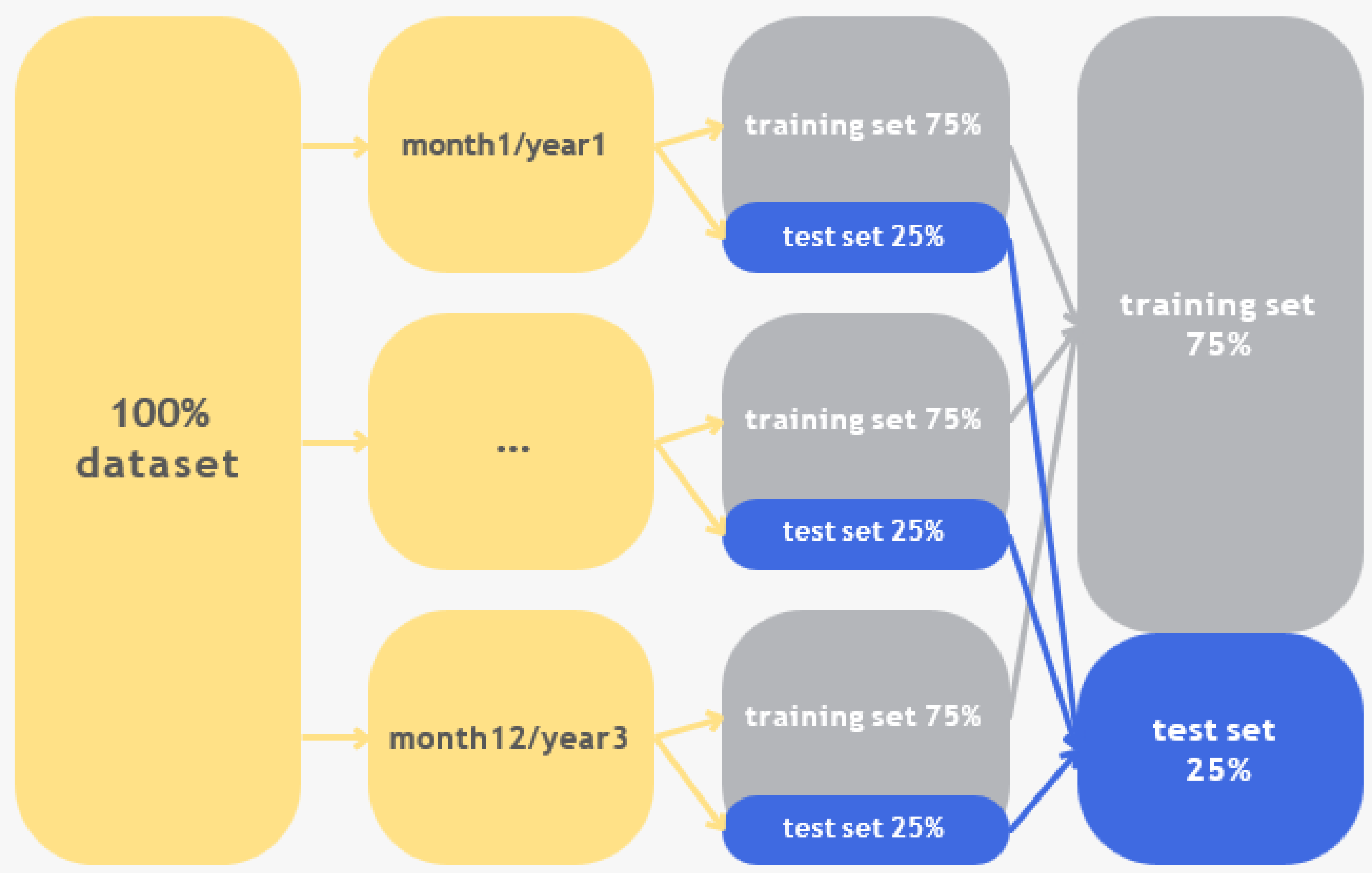
- **Numerical**: Guests, stay duration, lead time, daily rate (ADR)
- **Categorical**: Customer type, deposit policy, booking channel, room type, country
- **Temporal**: Year, month, week, day of week of arrival
- **Target**: IsCanceled: Binary cancellation flag



DATASET SPLITTING

Convenience Splitting

- Data was sorted chronologically to keep the temporal structure of the dataset.
- Each block is split: **75% training, 25% testing** based on **month-year blocks.**
- It ensures the model is trained on training data and **evaluated on unseen data**, providing a **more realistic performance assessment.**



N.B: the problem remains a **classification problem**, not time-series forecasting problem

DATASET PREPROCESSING

Handling Missing & Undefined Values

- Replaced NaN in Children with 0 (assumed no children);
- Converted categorical months to numerical 1-12;
- Replaced SC (Self Catering) with "Undefined" in the Meal column;
- Dropped rows that have missing distribution_channel, market_segment feature;
- Removed bookings with no guests (Adults + Children + Babies = 0);

Preventing Data Leakage

- Removed Country (e.g., default value “Portugal” often updated only after check-in);
- Removed AssignedRoomType, ReservationStatus, ReservationStatusDate, RequiredCarParkingSpaces(revealed post check-in);

Feature Engineering

- Created ADRThirdQuartileDeviation to capture ADR variability;

$$\frac{ADR}{Q3_{ADR}}$$

Encoding & Scaling

- Used Standard Scaler for all numerical variables to ensure uniform scale;
- Applied One-Hot Encoding to all non-binary categorical features;
- Applied Logit-Odds Encoding (for high-cardinality categorical features like Agent,Company)

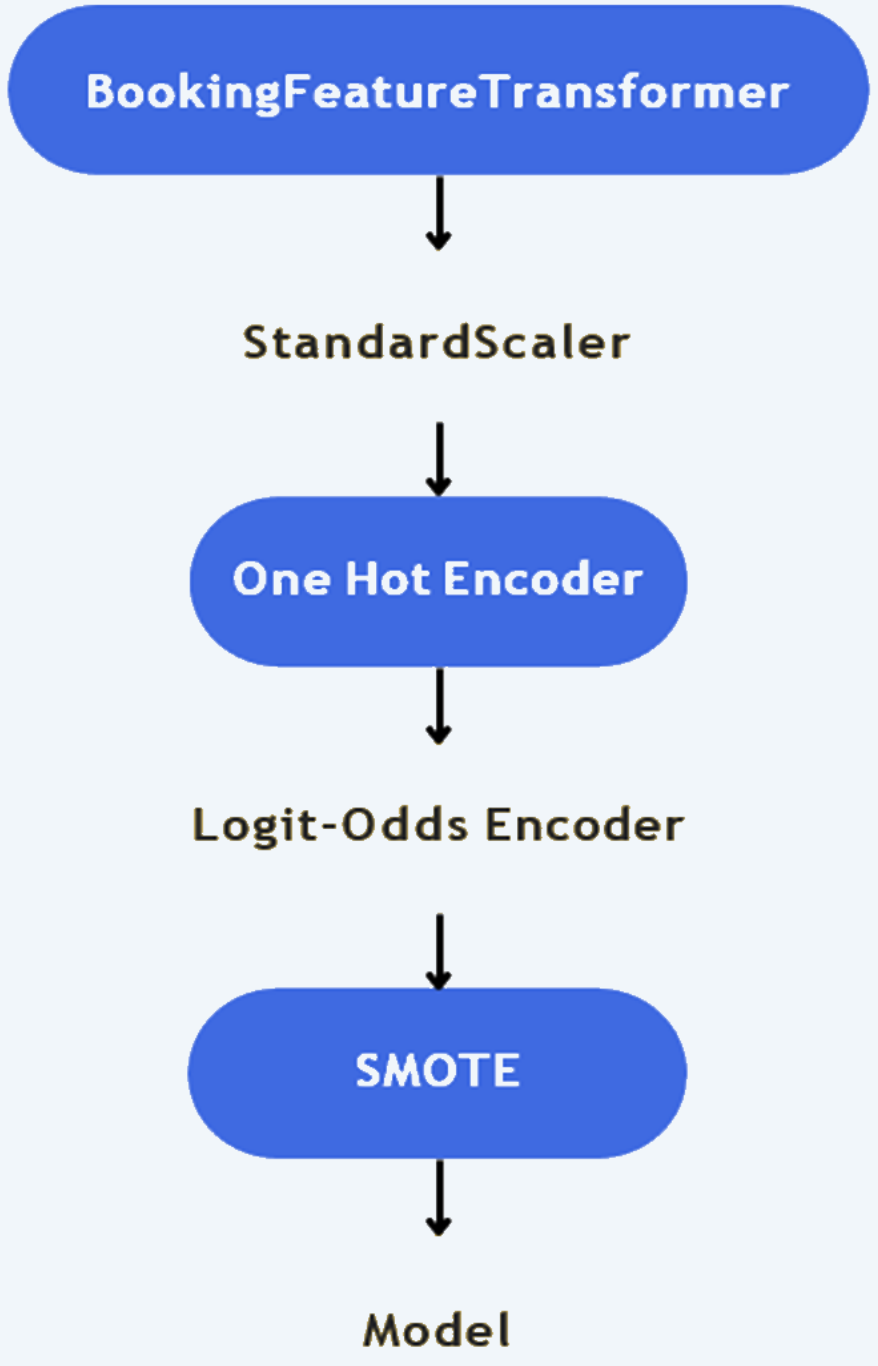
MODEL TRAINING: PIPELINES BUILDING

Feature engineering: BookingFeaturesTransformer computes the engineered feature "ADRThirdQuartileDeviation";

Handling Imbalanced Data: : SMOTE;

Preprocessing: StandardScaler + One Hot Encoder + Logit-Odds Encoder;

Classifier: Logistic Regression, Random Forest, XGBoost, AdaBoost, Bagging, Naive Bayesian, Decision Tree, KNN;



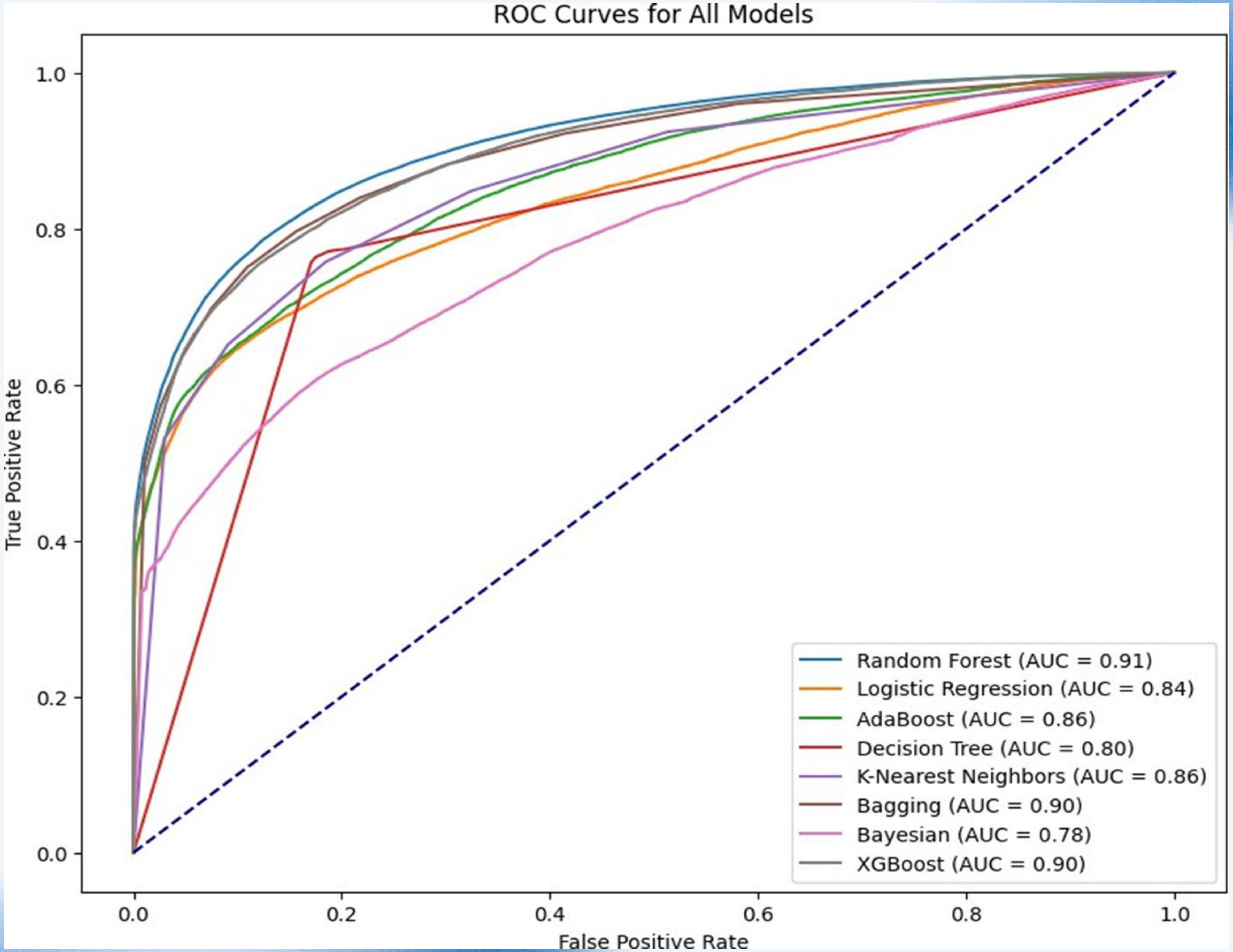
Models Comparison:

To evaluate the performance of various machine learning models for predicting hotel booking cancellations, several models were cross-validated using a 5-fold StratifiedKFold.

The models tested included:

- Random Forest
- Logistic Regression
- Decision Tree
- K-Nearest Neighbors
- AdaBoost
- XGBoost
- Bagging
- Naive Bayesian

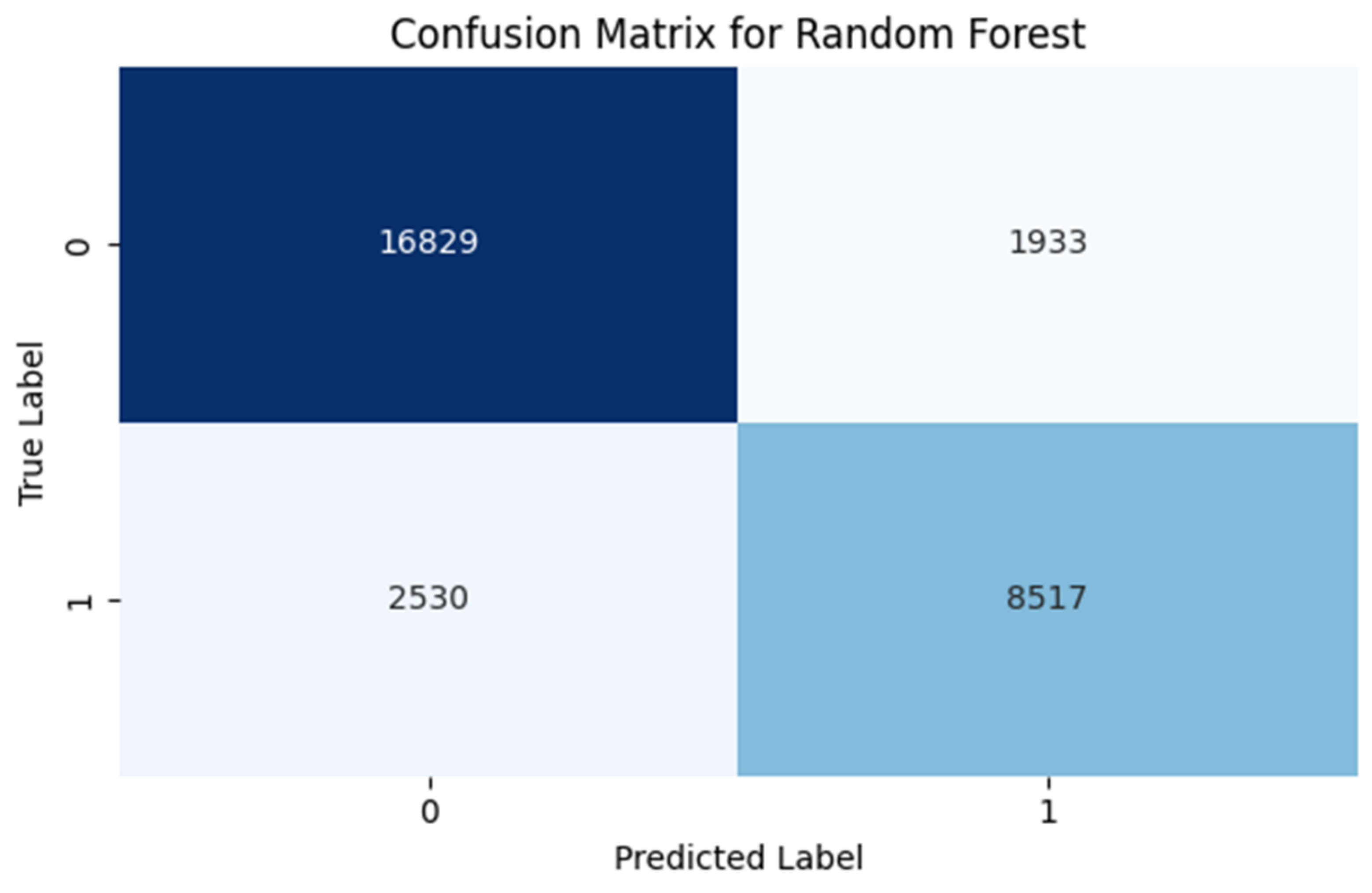
MODELS COMPARISON

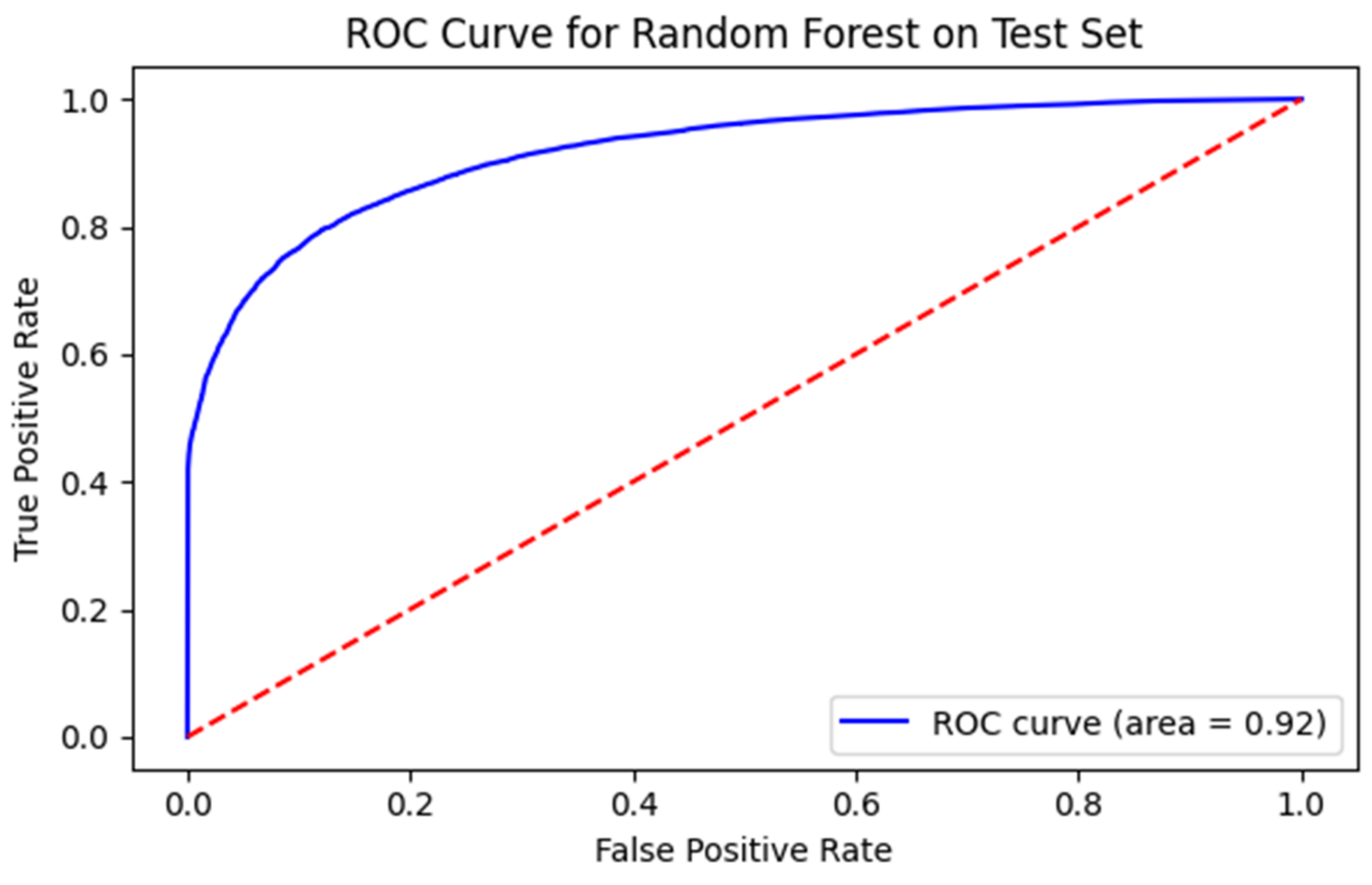


MODELS COMPARISON

	Accuracy	Precision	Recall	F1	ROC AUC
Random Forest	0.8457	0.8095	0.7635	0.7858	0.9110
Logistic Regression	0.7855	0.7153	0.7001	0.7076	0.8397
AdaBoost	0.7924	0.7262	0.7063	0.7161	0.8607
Decision Tree	0.8022	0.7198	0.7639	0.7412	0.7962
K-NN	0.7939	0.7073	0.7576	0.7316	0.8580
Bagging	0.8379	0.7982	0.7531	0.7750	0.8953
Bayesian	0.6578	0.5260	0.7765	0.6272	0.7787
XGBoost	0.8348	0.7958	0.7456	0.7699	0.9000

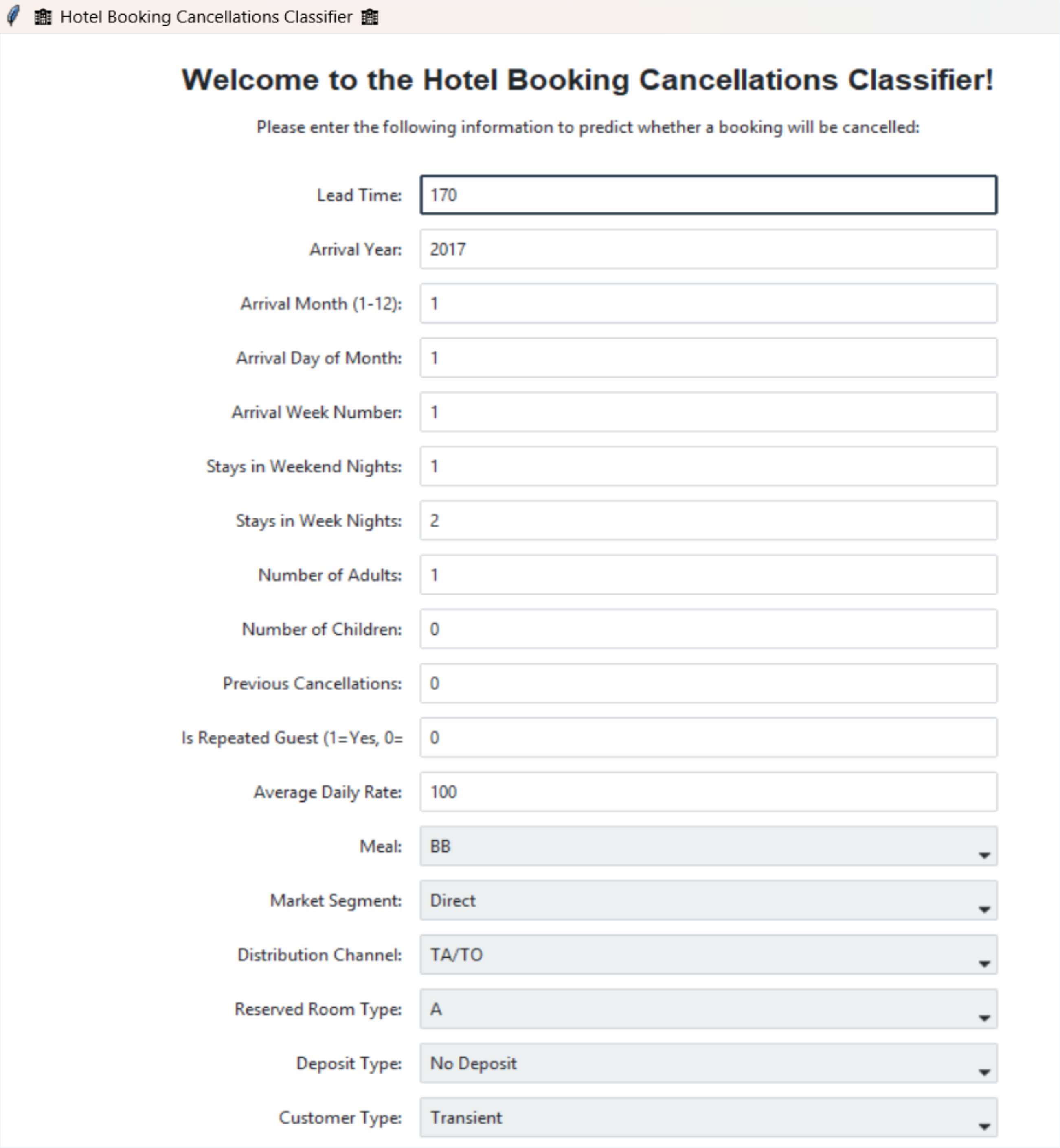
Model (RF) Evaluation

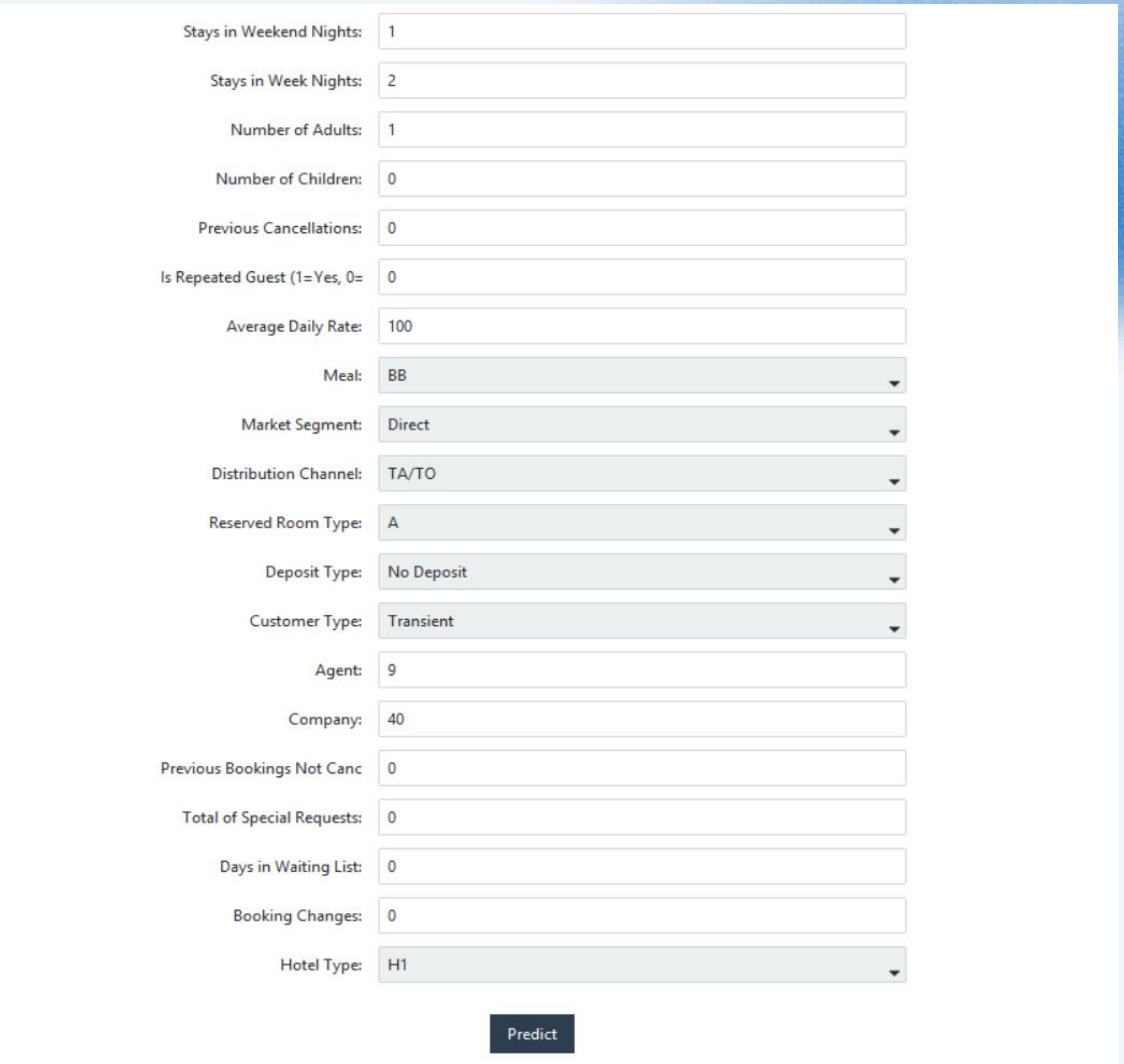


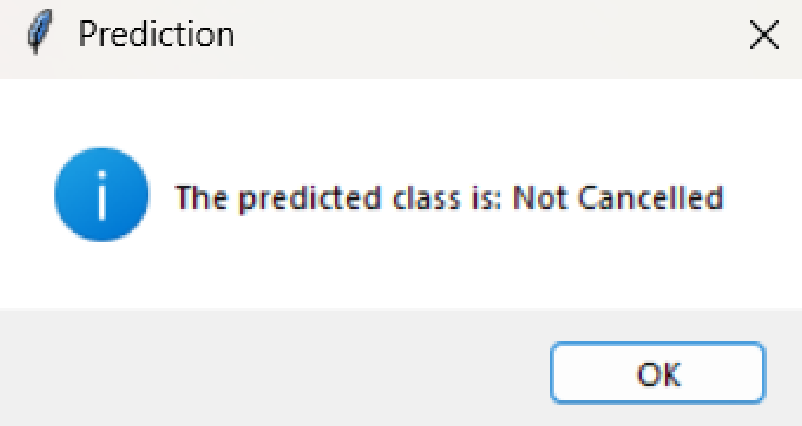


	Accuracy	Precision	Recall	F1
Random Forest	0.8503	0.8150	0.7710	0.7924

GRAPHIC USER INTERFACE







REFERENCES

- **Dataset**:Nuno Antonio, Ana de Almeida, and Luis Nunes. Hotel Booking Demand Datasets. Published in Data in Brief, Volume 22, Pages 41–49.
  DOI: https://doi.org/10.1016/j.dib.2018.11.126

- Nuno Antonio, Ana Maria De Almeida, and Luís Nunes. Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. 2017 IEEE International Conference on Machine Learning and Applications (ICMLA).
  DOI: 10.1109/ICMLA.2017.00-11
- Zharfan Akbar Andriawan, Ricko, Feri Wijayanto, Satriawan Rasyid Purnama, Adi Wibowo, Adam Sukma Darmawan, and Aris Sugiharto. Prediction of Hotel Booking Cancellation using CRISP-DM. 2020 4th International Conference on Informatics and Computational Sciences (ICICoS). DOI: 10.1109/ICICoS51170.2020.9299011

THANK YOU FOR THE ATTENTION