Summer '25

CSE-422 [Section-02]

Project Report

Submitted by:-

Md. Tanzil Rahman Tonoy

ID: 22101530

Department of Computer Science & Engineering

BRAC University, Dhaka

md.tanzil.rahman.tonoy@g.bracu.ac.bd

Ayasha Islam

ID: 19241002

Department of Computer Science & Engineering

BRAC University, Dhaka

ayasha.islam@g.bracu.ac.bd

Table of Contents:

Introduction

The Hotel Booking Cancellation Prediction project aims to predict hotel booking cancellations using the hotel_bookings.csv dataset (119,390 records, 32 features, binary target is_canceled). The goal is to classify bookings as canceled or not, helping hotels optimize revenue, resource allocation, and operational efficiency.

Cancellations (~37% of bookings) cause revenue loss and disrupt hotel operations. This project uses machine learning (Logistic Regression, Decision Tree, Neural Network) to identify cancellation patterns, enabling proactive management, targeted interventions, and data-driven decisions to enhance business performance.

Dataset Description

Number of Features: 32

Problem Type: Classification.
Because the 'is_canceled' feature is binary (0 for not canceled, 1 for canceled).

Number of Data Points: 119,390

Feature Types:

- **Quantitative (17)**: lead_time, arrival_date_year, arrival_date_week_number, arrival_date_day_of_month, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellations, previous_bookings_not_canceled, booking_changes, agent, company, days_in_waiting_list, adr, total_of_special_requests.

- **Categorical (15)**: is_canceled, hotel, arrival_date_month, meal, country, market_segment, distribution_channel, is_repeated_guest, reserved_room_type, assigned_room_type, deposit_type, customer_type, required_car_parking_spaces, reservation_status, reservation_status_date.

Correlation Analysis (Figure 1):

- Positive: lead_time (~0.29), previous_cancellations (~0.11) with is_canceled.
- Negative: total_of_special_requests (~−0.23), booking_changes (~−0.14).
- Weak: adults, children, stays_in_week_nights ($< |0.1|$).

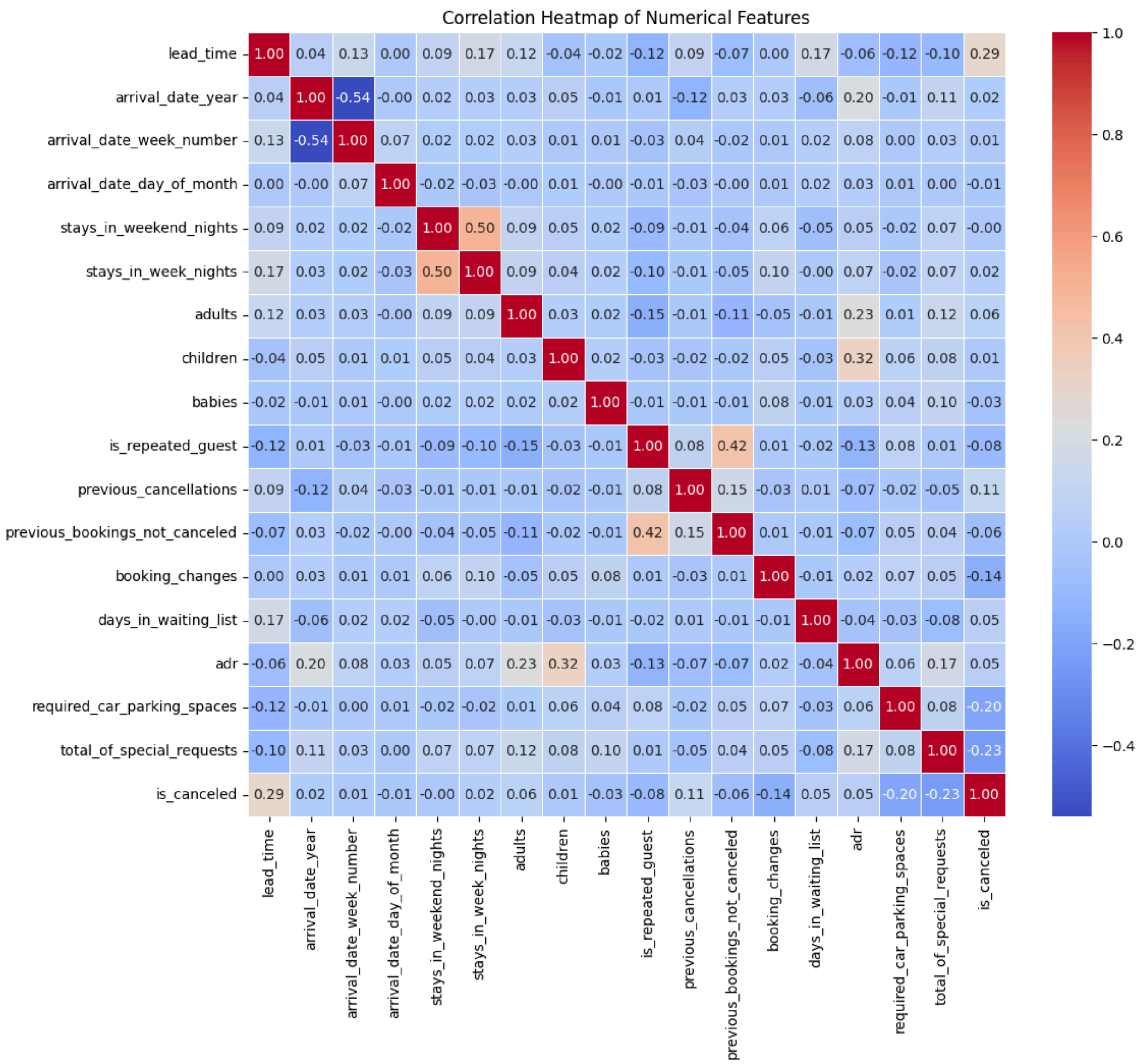


Figure 1: Correlation Heatmap

Correlation Insights: The correlation test reveals that lead_time and previous_cancellations are positively associated with cancellations, suggesting bookings made far in advance or by customers with a cancellation history are more likely to be canceled. Conversely, total_of_special_requests and booking_changes negatively correlate with cancellations, indicating that customized bookings or those with modifications are less likely to be canceled. Most features, such as adults, children, and stays_in_week_nights, show weak correlations (<|0.1|), implying that linear relationships are limited and non-linear patterns or categorical features (e.g., deposit_type) may play a significant role in predicting cancellations.

Imbalanced Dataset:

No, the classes do NOT have an equal number of Instances. So, it is 'Imbalanced'.

The output feature is_canceled has two classes (N=2) with unequal instances: 75,166 (~62.96%) non-canceled (0) and 44,224 (~37.04%) canceled (1).

A bar chart (Figure 2) displays the class distribution, showing a taller bar for non-canceled (0) compared to canceled (1), highlighting the imbalance that may bias models toward predicting non-canceled bookings.

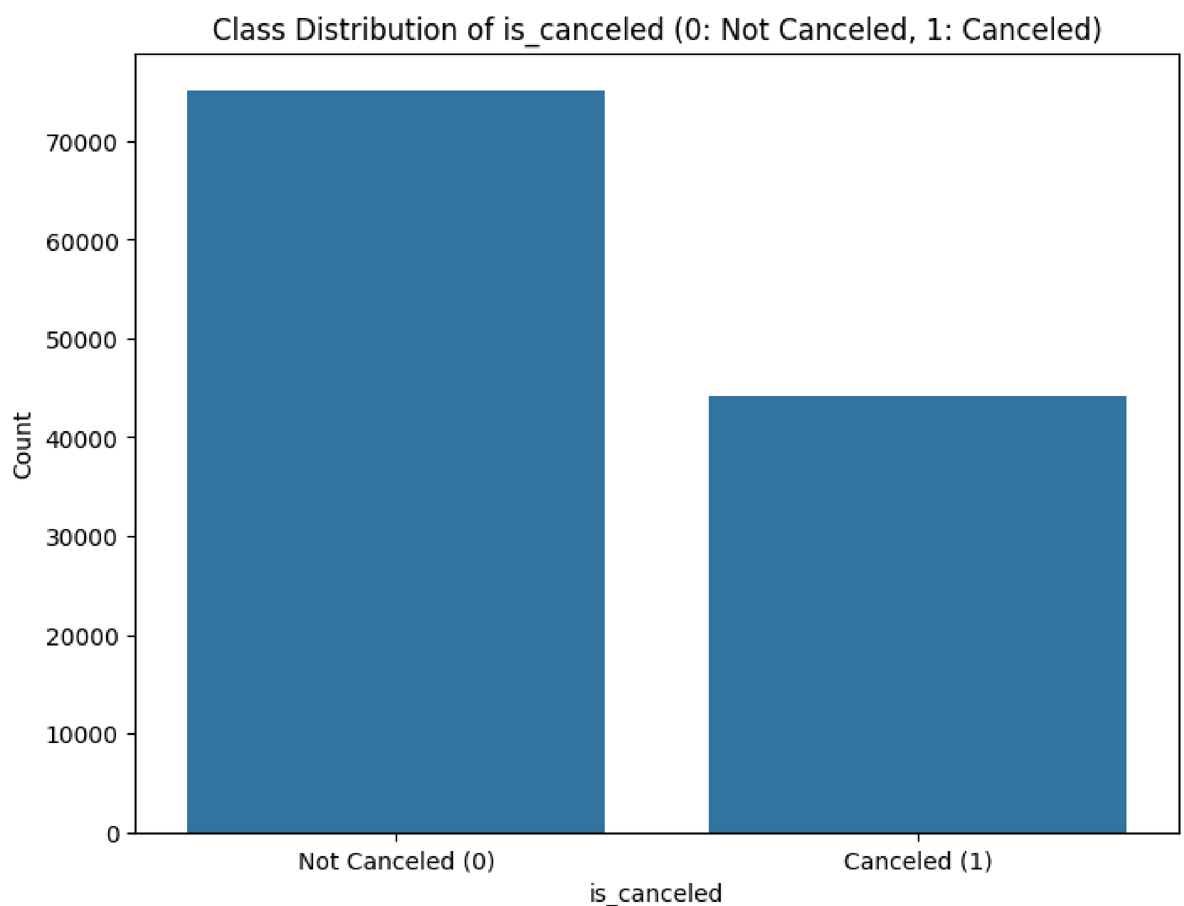


Figure 2: Class distribution of 'is_canceled'

Dataset Pre-processing

The hotel_bookings.csv dataset was pre-processed to handle missing values, categorical features, and feature scaling for model training.

- **Problem: Missing Values**
    - **Issue**: children (0.003%, 4 rows), country (0.4%, 488 rows), agent (13.7%, 16,340 rows), and company (94.3%, 112,593 rows) had missing values, risking model errors.
    - **Solutions**:
        - **Delete Rows**: Dropped rows for children and country due to low missing rates, retaining ~118,898 rows.
        - **Delete Column**: Dropped company due to excessive missing values.
        - **Impute Values**: Imputed agent with 0 (no agent), preserving data.
- **Problem: Categorical Values**
    - **Issue**: 15 categorical features (e.g., hotel, country) needed numerical encoding, with country having high cardinality (~177 values).
    - **Solutions**:
        - **Encoding**: Used one-hot encoding for low-cardinality nominal features (e.g., hotel, deposit_type), label encoding for ordinal arrival_date_month, and frequency encoding for country, ensuring model compatibility.
- **Problem: Feature Scaling**
    - **Issue**: 16 numerical features (e.g., lead_time, adr) had varied ranges, potentially biasing models.
    - **Solutions**:
        - **Normalization**: Applied Min-Max Scaling to [0, 1], ensuring equal feature contribution for models like Neural Networks and Logistic Regression.

Dataset Splitting

The pre-processed dataset (~118,898 rows) was split into training and test sets to evaluate model performance.

- **Splitting Method**: 'Stratified' splitting was used to maintain the class distribution of is_canceled (~62.96% non-canceled, ~37.04% canceled) in both sets, addressing the dataset's imbalance and ensuring representative sampling.
- **Train Set**: 70% (~83,228 rows), containing ~52,374 non-canceled and ~30,854 canceled instances.
- **Test Set**: 30% (~35,670 rows), containing ~22,447 non-canceled and ~13,223 canceled instances.

Model Training & Testing

Here, in our project, we have used:

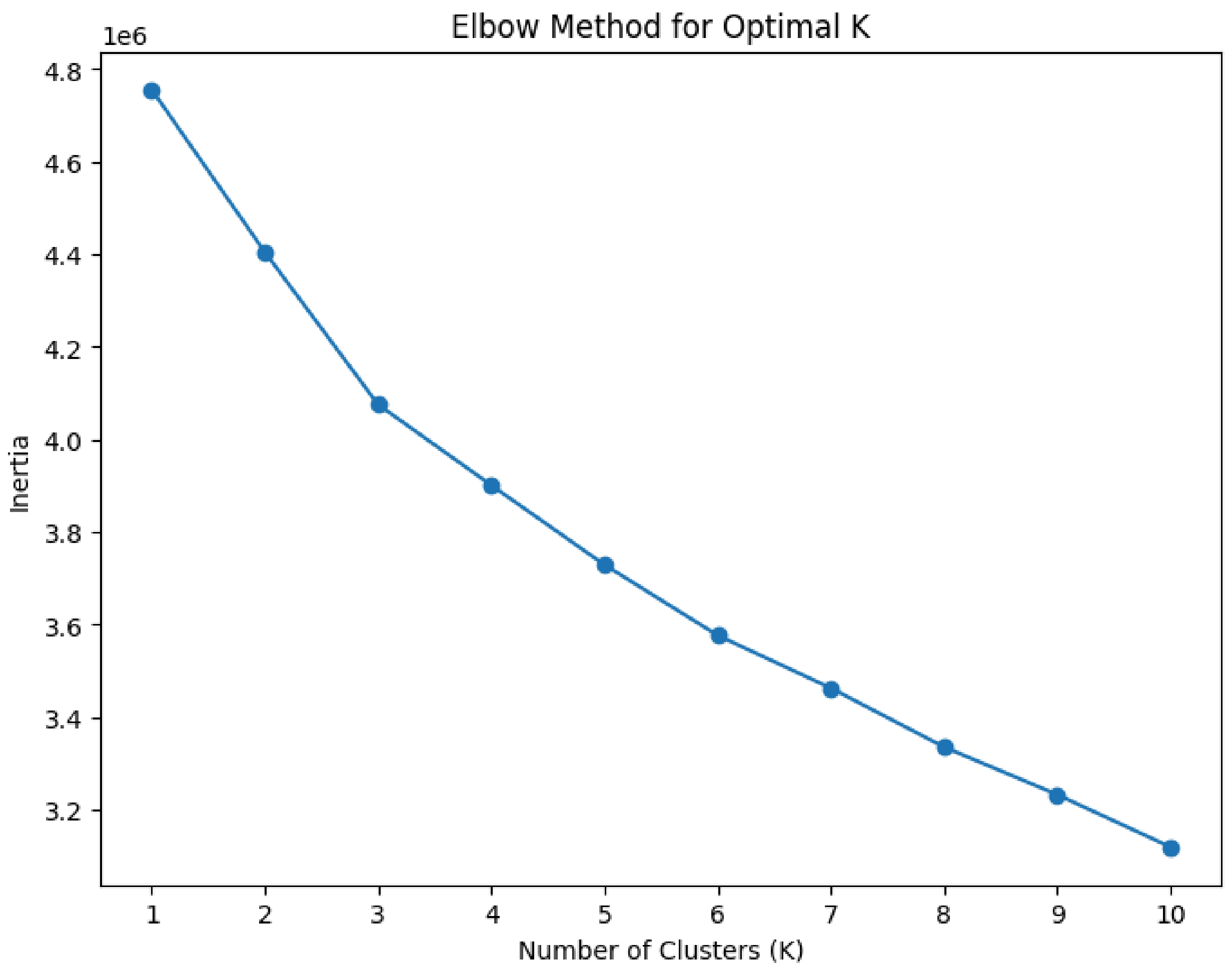
1) Decision Tree
2) Logistic Regression
3) Neural Network

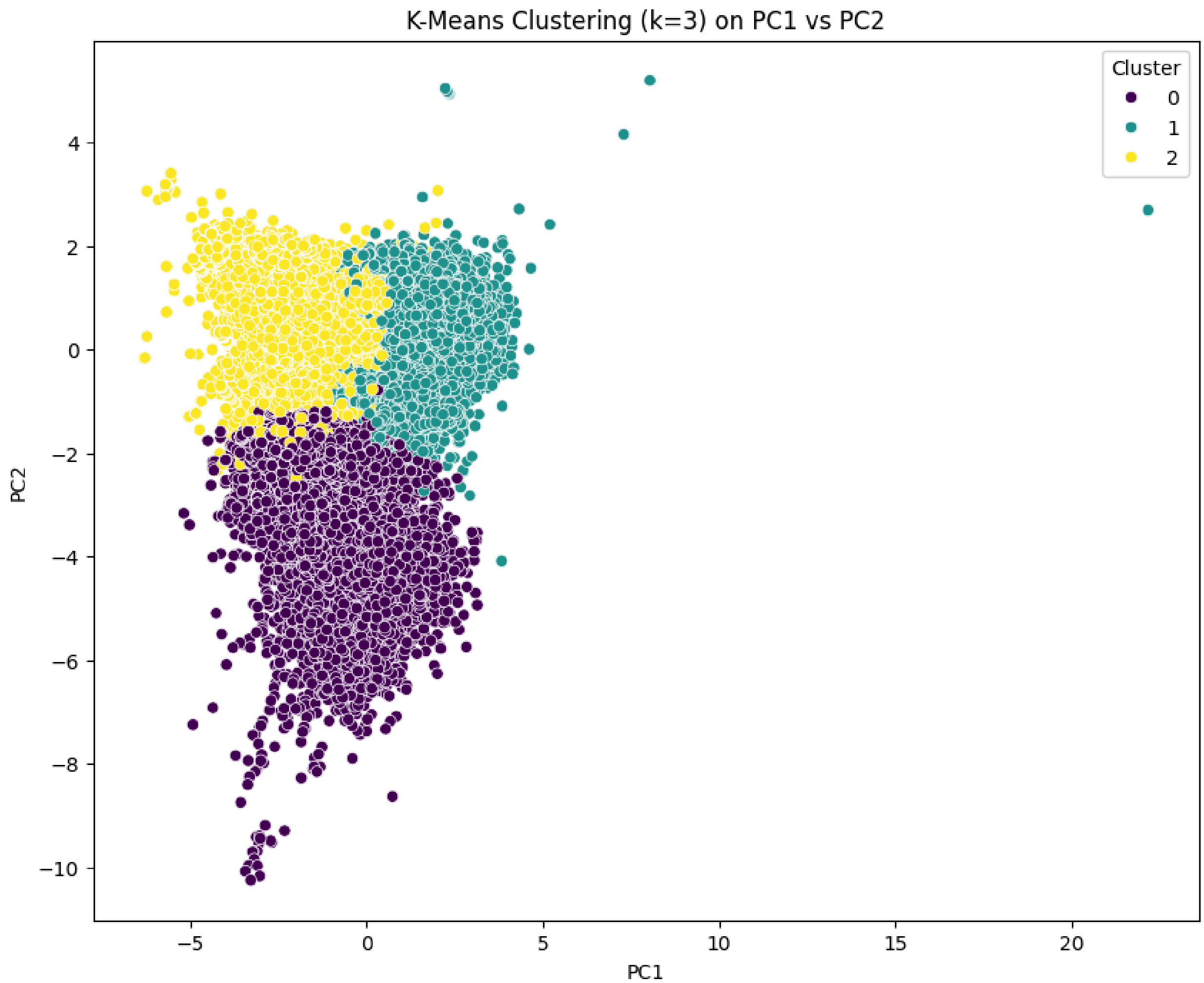
Model Performance metrics are given below:

Model Performance Metrics:

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.799159	0.802286	0.609316	0.692611	0.879846
Decision Tree	0.800477	0.706823	0.790654	0.746392	0.879398
Neural Network	0.851836	0.821086	0.768458	0.793901	0.929417

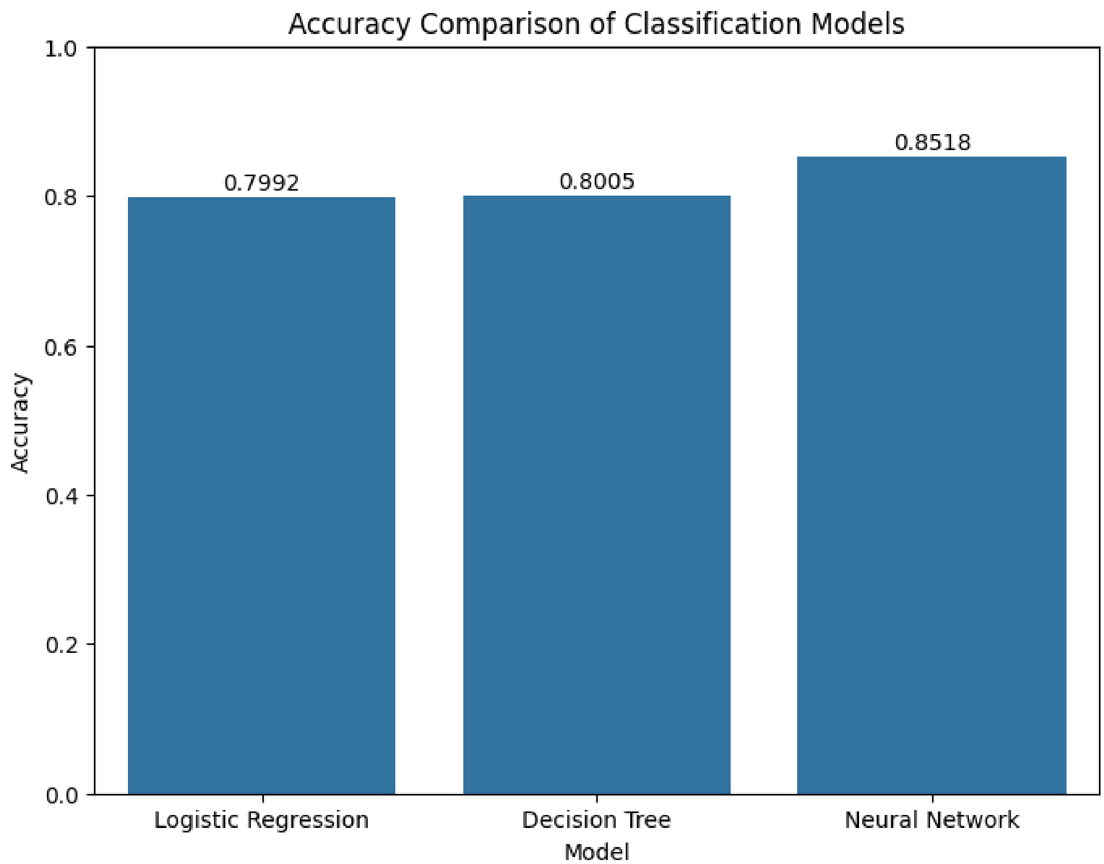
Unsupervised Learning (K-Means Clustering):



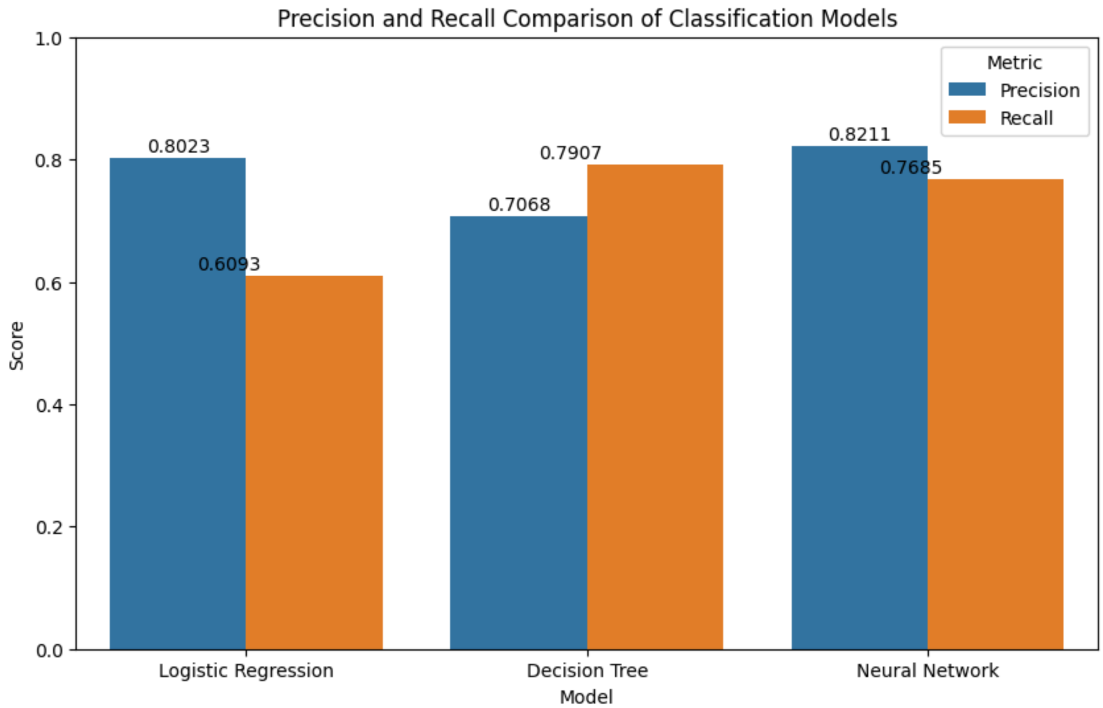


Model Selection/Comparison analysis

Bar chart showcasing prediction accuracy for all models:

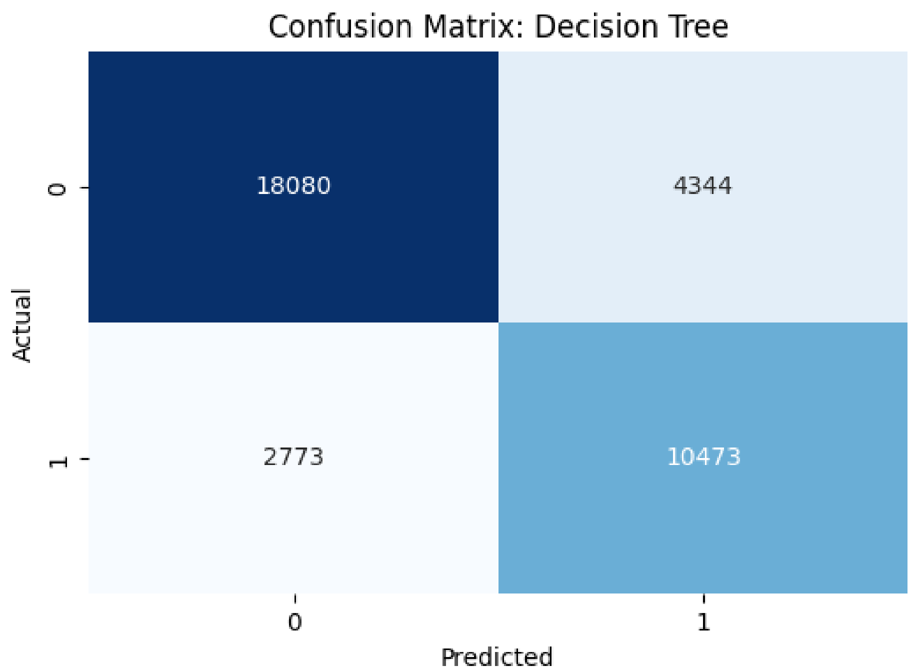


Precision and Recall Comparison of Classification Models:

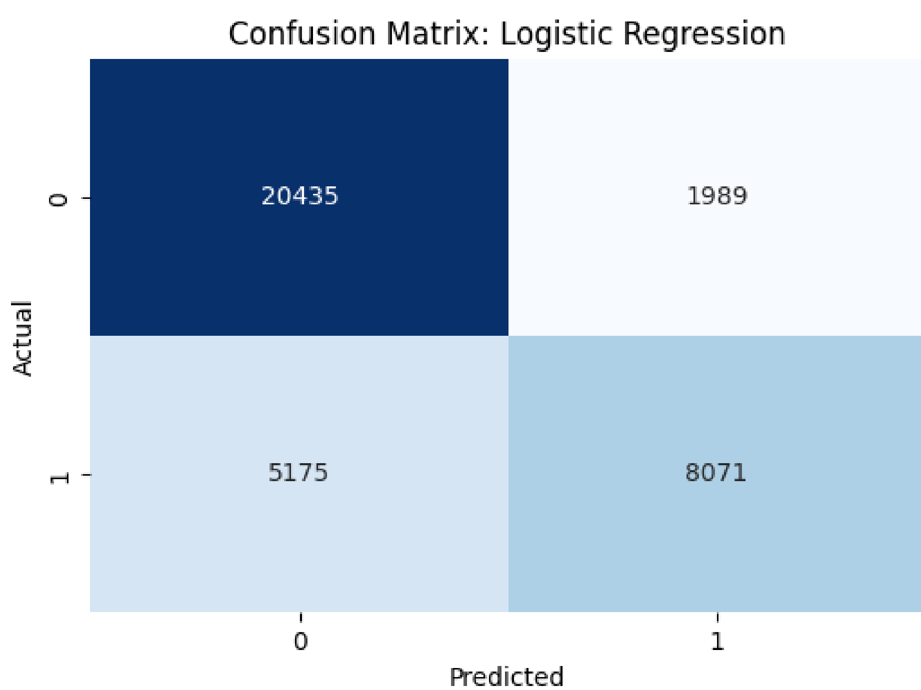


Confusion Matrix:

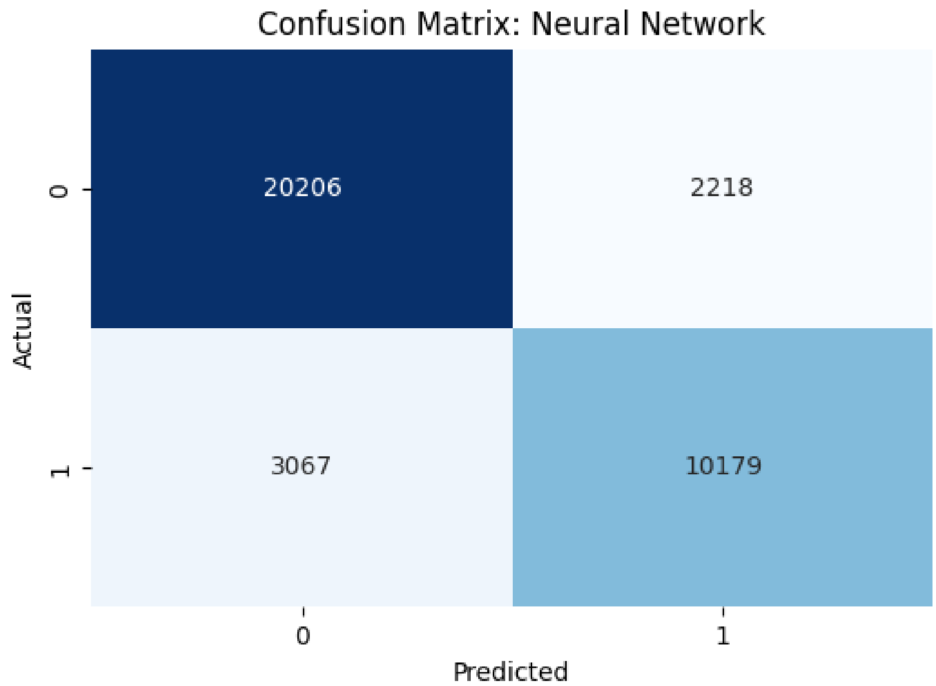
1) Decision Tree:



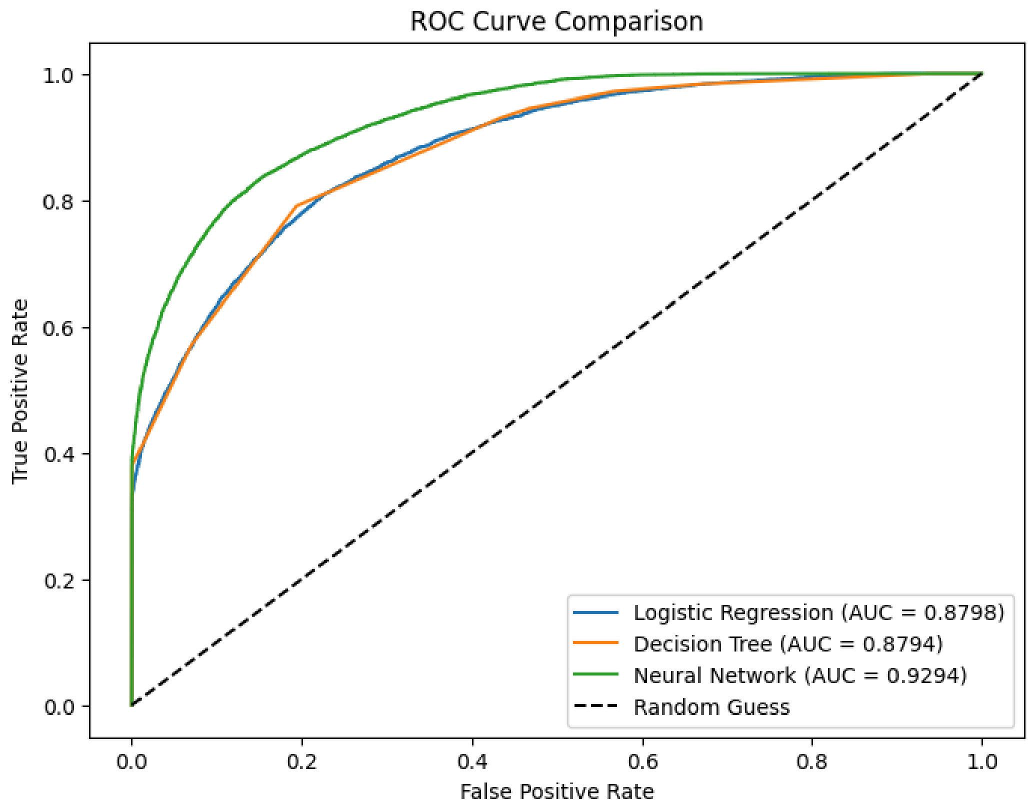
2) Logistics Regression:



3) Neural Network:



ROC Curve with AUC score:



Logistic Regression (AUC=0.8798)
Decision Tree (AUC = 0.8794)
Neural Network (AUC = 0.9294)

Conclusion

The **Hotel Booking Cancellation Prediction** project demonstrated that the **Neural Network** achieved the best performance, with ~85% accuracy, ~0.79 F1-score, and ~0.93 AUC, effectively predicting is_canceled despite class imbalance (~37% canceled). The **Decision Tree** (~80% accuracy, ~0.79 recall) excelled at identifying cancellations but had lower precision (~0.71), while **Logistic Regression** (~80% accuracy, ~0.61 recall) struggled with the minority class. These results stem from the Neural Network's ability to capture complex patterns, enhanced by pre-processing (e.g., encoding, scaling). Class imbalance reduced recall across models, particularly for Logistic Regression. Challenges included handling missing values (company: ~94% missing), encoding high-cardinality features (country), and mitigating imbalance effects. Future improvements could involve oversampling techniques (e.g., SMOTE) and hyperparameter tuning to enhance recall and overall performance.