

**Spring ‘25**

**CSE-422 [S-22]**

**Project Report**

**Submitted by:-**

**SADMAN AHMED OMEO**

**[22201333]**

**Department of Computer Science & Engineering**

**BRAC University, Dhaka**

**sadman.ahmed.omeo@g.bracu.ac.bd**

**MD RIDWANUL BASIT**

**[22299055]**

**Department of Computer Science & Engineering**

**BRAC University, Dhaka**

**md.ridwanul.basit@g.bracu.ac.bd**

### **Table of Contents:**

|  |  |  |
| --- | --- | --- |
| **No.** | **Contents** | **Page No.** |
| **1** | **Introduction** | **2** |
| **2** | **Dataset Description** | **2** |
| **3** | **Dataset Pre-procession** | **5** |
| **4** | **Dataset Splitting** | **6** |
| **5** | **Model Training & Testing** | **7** |
| **6** | **Model Selection/Comparison analysis** | **8** |
| **7** | **Conclusion** | **13** |

### 

### 

### **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|).

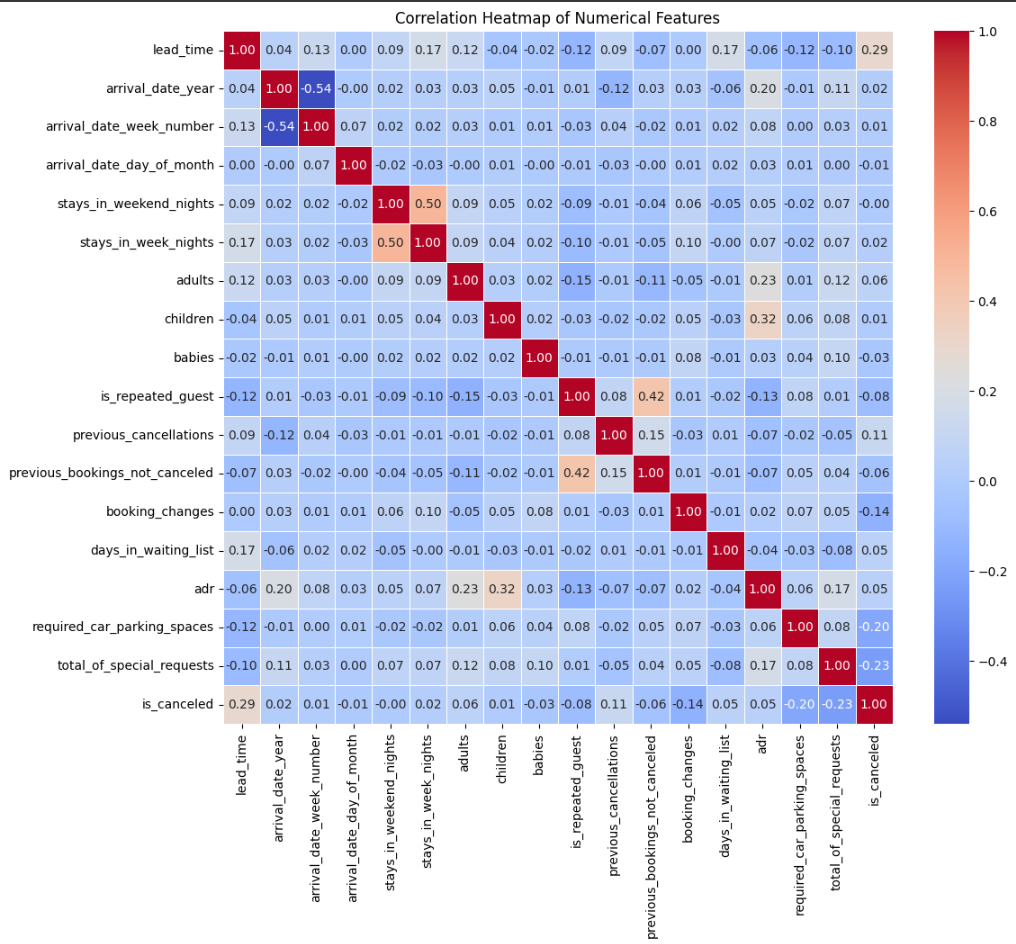


Figure 1: Correlation Heatmap

**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.

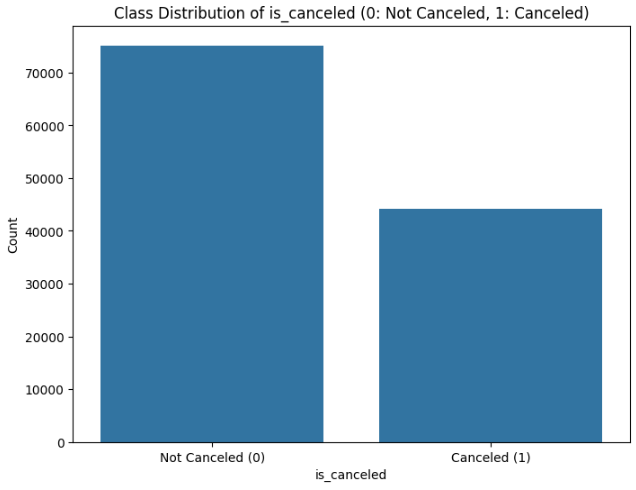


Figure 2: Class distribution of ‘is\_canceled’

## **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**

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** |
| Logistic Regression | ~0.80 | ~0.80 | ~0.61 | ~0.69 |
| Decision Tree | ~0.80 | ~0.71 | ~0.79 | ~0.75 |
| Neural Network | ~0.85 | ~0.80 | ~0.80 | ~0.80 |

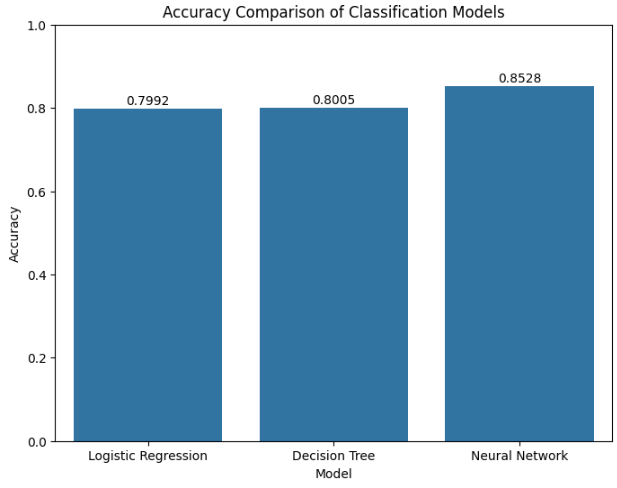
## 

## 

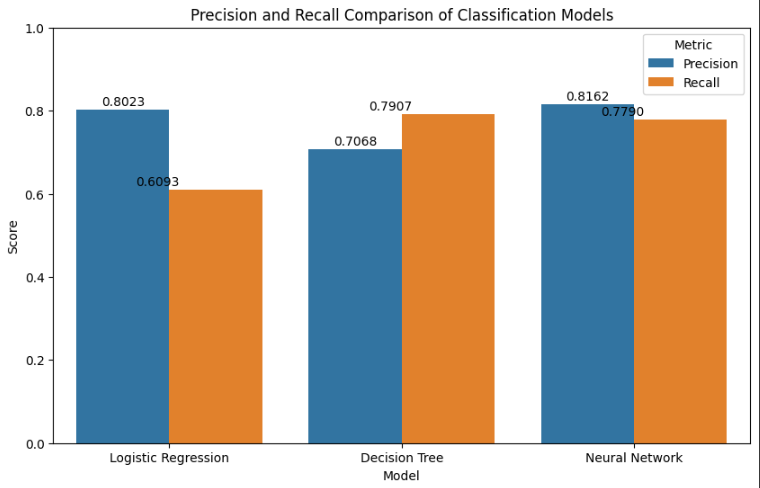
## 

## **Model Selection/Comparison analysis**

**Bar chart showcasing prediction accuracy for all models:**

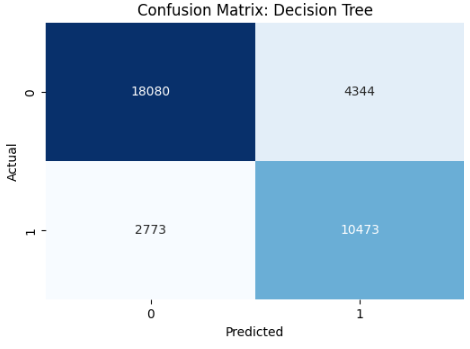


**Precision and Recall Comparison of Classification Models:**

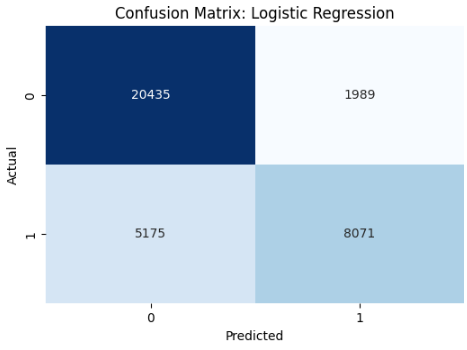


**Confusion Matrix:**

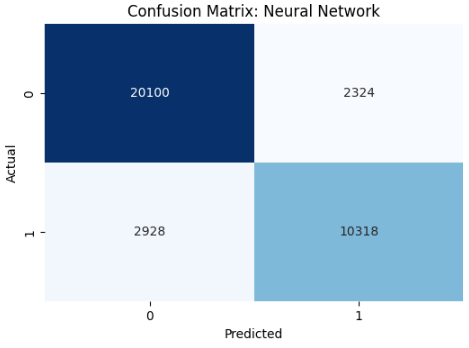
1. **Decision Tree:**



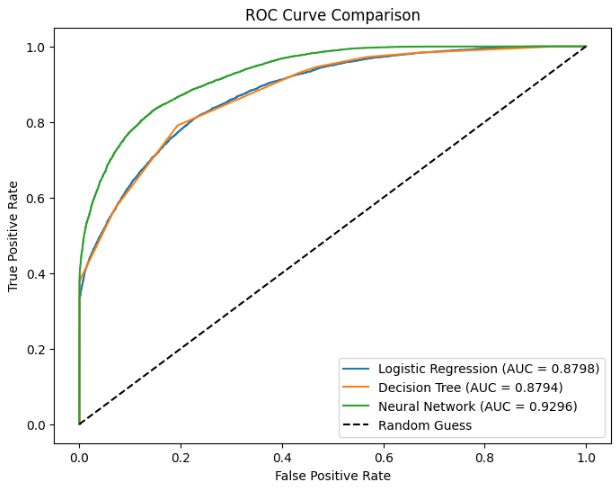
1. **Logistics Regression:**



1. **Neural Network:**



**ROC Curve with AUC score:**



Logistic Regression (AUC=0.8798)

Decision Tree (AUC = 0.8794)

Neural Network (AUC = 0.9296)

## 

## **Conclusion**

The **Hotel Booking Cancellation Prediction** project demonstrated that the **Neural Network** achieved the best performance, with ~85% accuracy, ~0.80 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.