BRAC University
Department of Computer Science and Engineering

CSE422 : Artificial Intelligence
Project Report
Fall 2025

Section 18

Group 11

Mahdi Rahman - 24141036
Tabia Sultana Ritu - 23201510

Table of Contents

Introduction

The main objective of this project is to apply a number of machine learning models to predict whether a customer will cancel his/her reservation of a hotel based on a hotel booking dataset. Through a number of preprocessing methods we addressed significant data quality issues by removing high-null features like company, handling missing values in country and agent, and correcting undefined categorical labels.By engineering new features such as total_guests and total_stay, and by applying One-Hot Encoding and Standard Scaling we have transformed the raw data into a format suitable for comparative modeling. We implemented and evaluated three distinct classification approaches like Logistic Regression, Neural Networks and Random Forest Classifier to identify the most accurate and interpretable method for identifying high-risk bookings.

Data Set Description

The project uses the Hotel booking dataset, which is a classic dataset for classification problems.

1. How many features?

The original dataset contains 31 features including hotel type, lead time, arrival date year, arrival date month, arrival date week number, and arrival date day of month, stays in weekend nights and stays in week nights, adults, children, and babies, meal, country, market segment, distribution channel, is repeated guest, previous cancellations, previous bookings not canceled, reserved room type, assigned room type, booking changes, deposit type, agent, company, days in waiting list, customer type, adr, required car parking spaces, total of special requests, reservation status, and reservation status date and the target variable is is_cancelled.

2. Classification or regression problem?Why do you think so?

This is a classification problem because the target variable, is_cancelled, is categorical and binary. The goal is to predict which of two discrete classes a booking falls into either "canceled" (1) or "not canceled" (0), rather than predicting a continuous numerical value.

3. How many data points? What kind of features are in your dataset?

The dataset consists of 119,390 data points and the dataset consists of both Quantitative(Lead time, stay duration, adults) and Categorical features(Hotel, meal, country).

4. Do we need to encode the categorical variables, why or why not?

Yes, we do need to encode categorical variables because the machine learning models we chose, including the Logistic Regression and Neural Network (MLP) models in our project, are built on mathematical equations that can only process numerical data. As most algorithms cannot interpret text strings like "Resort Hotel" or "July" directly. Transforming these into numbers (0s and 1s) allows the model to perform the necessary mathematical calculations for training and prediction.

Data Analysis

In this section, we analyze the dataset to understand patterns, distributions, and relationships between features and the target variable.

1.Bar Chart

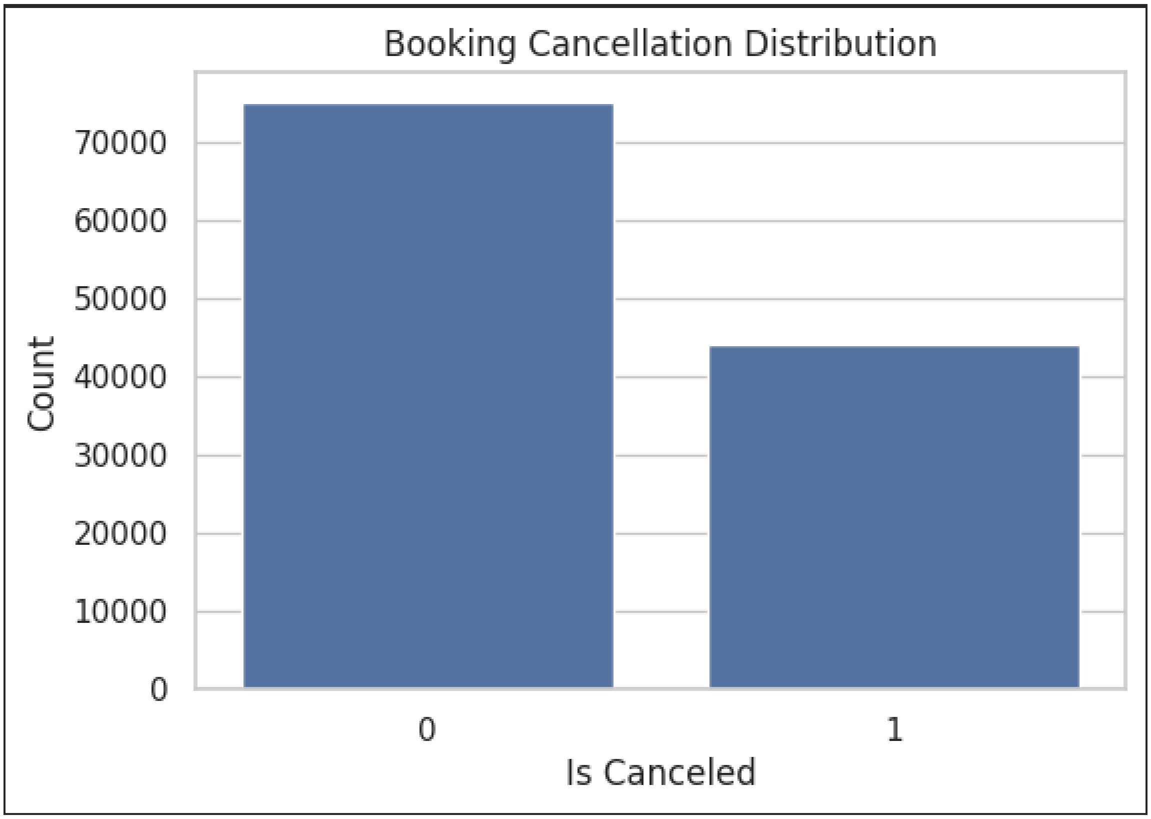


Fig 01: This this the bar chart for not cancelled and is cancelled

2. Histogram

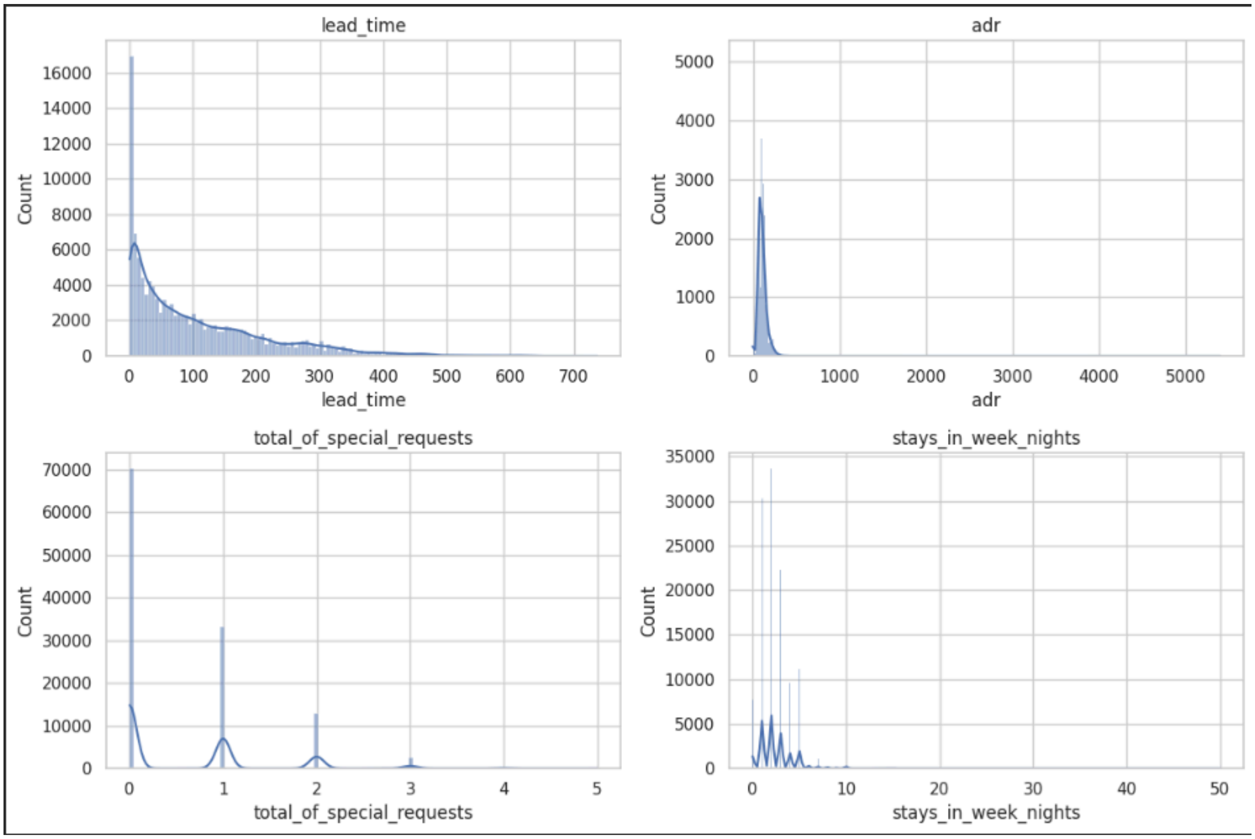


Fig 02: Histogram to visualise the spread between features

3. Correlation Heatmap

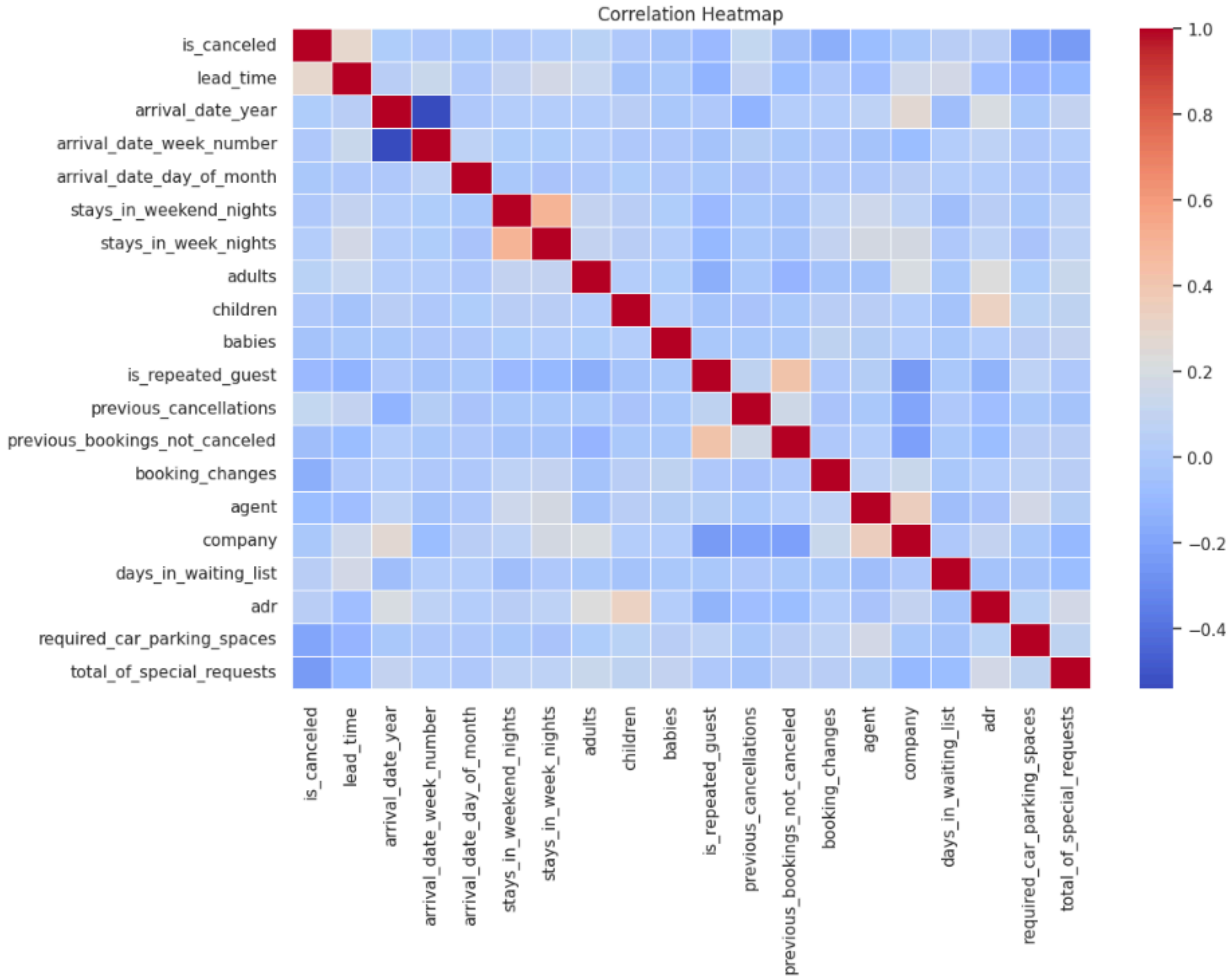


Fig 03: A visual summary of how different booking variables relate to each other and the target variable.

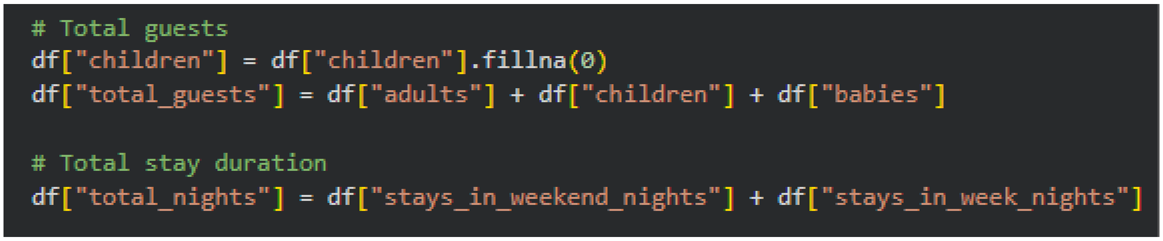
Dataset Preprocessing

The raw dataset contains several faults that need to be addressed before training the models.

1.Null / Missing values: One column that has the most null values is the company which consists of 112,593 missing values(approximately 94% of the data). There are others such as agents with 16,340 missing values and countries which have 488 missing values.

2. Categorical Garbage: The meal column which contains 1169 undefined values. The market segment and distribution channel also has very little undefined values.

3. Feature Engineering: We can create more meaningful features that often have a higher correlation with cancellations than the raw columns. Such as:



```
# Total guests
df["children"] = df["children"].fillna(0)
df["total_guests"] = df["adults"] + df["children"] + df["babies"]

# Total stay duration
df["total_nights"] = df["stays_in_weekend_nights"] + df["stays_in_week_nights"]
```

4. Target Separation: This is where we separate the target from the feature so that the data can be used for training and testing.

5. Pre-processing pipeline: Here we first separate the numerical feature from the categorical features. For numerical features, missing values are handled using a median strategy which replaces missing values with the median of each column. For the missing values in categorical features we filled them with the most frequent category. Then we perform One Hot Encoding on the categorical columns to turn them into binary columns and then finally we combine the two pipelines.

Dataset Splitting

The pre-processed dataset was split into training and testing sets to evaluate the models' performance on unseen data. We used a standard train_test_split with a test size of 20%, resulting in:
Train set: 80% of the data, used to train the models.
Test set: 20% of the data, used to test the models' performance.

Model Training & Testing

For Model training we used 3 key Machine Learning models -
1. Logistic regression- Logistic regression is a linear classification algorithm used for binary classification problems.
Logistic Regression Performance
Accuracy : 81.08%
Precision: 81.22%
Recall : 63.63%
F1-score : 71.36%

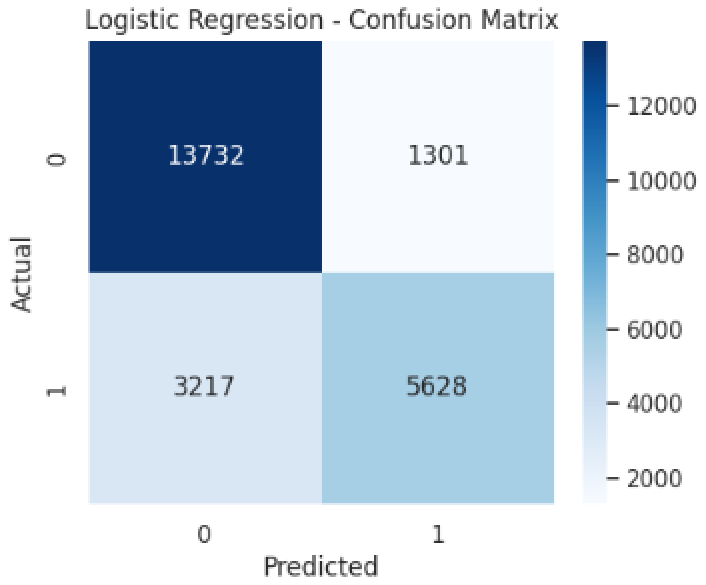


Fig 04: Confusion matrix of logistic regression

2. Neural Network: A multi-layer Perceptron is a feedforward artificial neural network that can model complex non linear relations.

Neural Network Performance
Accuracy : 86.79%
Precision: 83.44%
Recall : 80.26%
F1-score : 81.82%

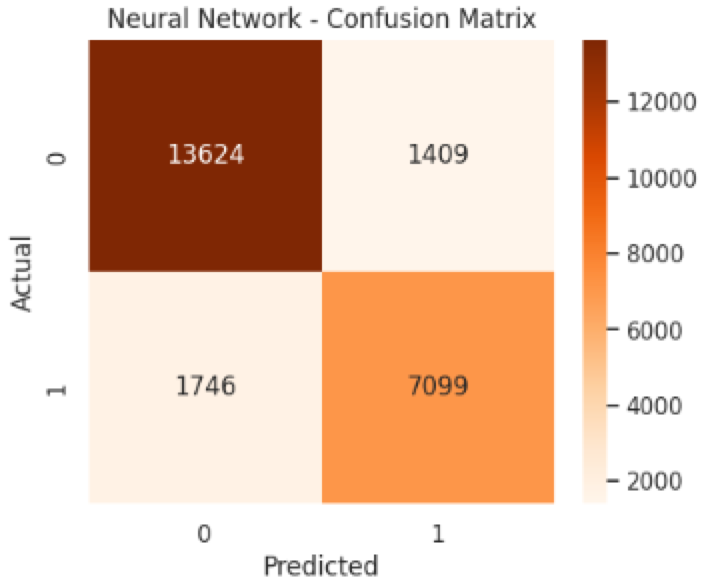


Fig 05: Confusion Matrix of neural network

3. Random forest classifier: Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting.

Random Forest Performance

Accuracy : 89.42%

Precision: 88.95%

Recall : 81.56%

F1-score : 85.10%

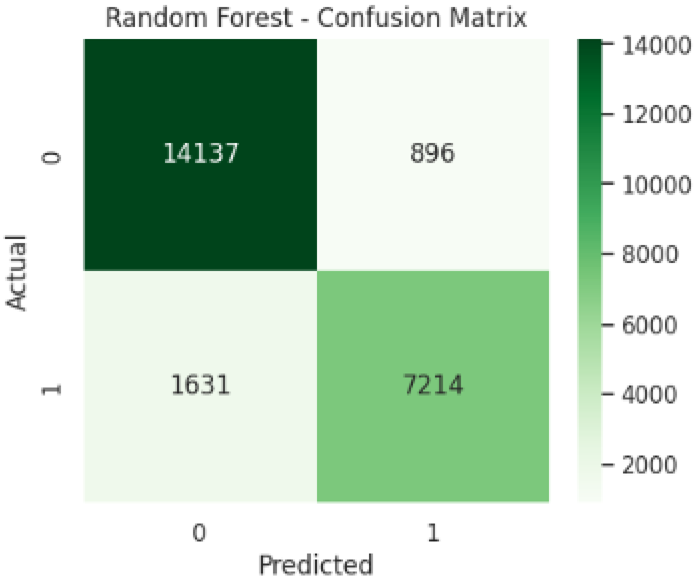


Fig 06: Confusion Matrix of Random Forest.

Model Comparison

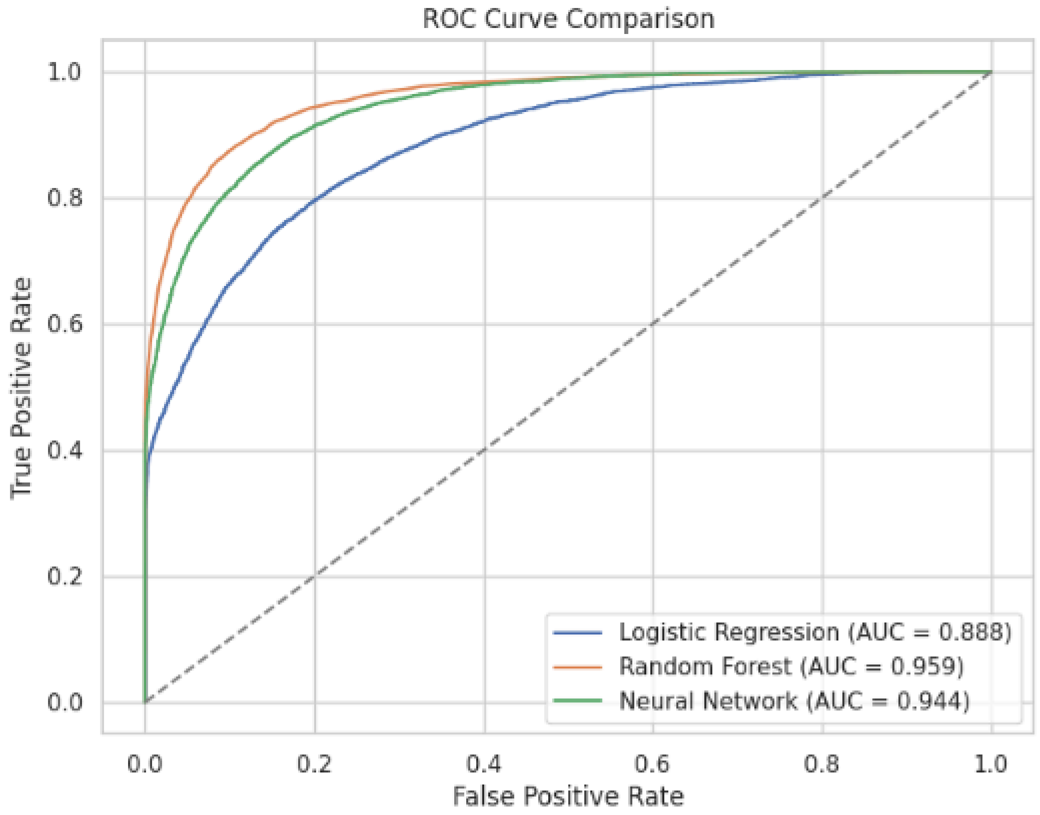


Fig 07: ROC curve of the 3 models.

Unsupervised Learning

In addition to supervised learning, we also used K-Means clustering for the unsupervised part. To perform the k means we first need to choose an optimal number of clusters. To find the optimal k value we test multiple values on a sample then we use the elbow method and the silhouette score plot to determine the reasonable K value. After that we applied K-Means to the preprocessed training data with the optimal number of clusters(K = 4). The model was trained on the scaled and encoded features, and cluster labels were assigned to both the training and testing datasets. These clusters were then used for interpretability by mapping them back to the original training data. In parallel, the performance of three supervised models—Logistic Regression, Random Forest, and a neural network was evaluated using accuracy, precision, recall, and F1-score, and the results were summarized in a comparison table. To better understand customer behavior, the cluster labels were combined with the target variable is_cancelled to compute the cancellation rate for each cluster. The resulting bar chart shows clear differences in cancellation tendencies across clusters, indicating that the clustering successfully grouped bookings with distinct cancellation patterns, which provides useful insights beyond pure classification performance.

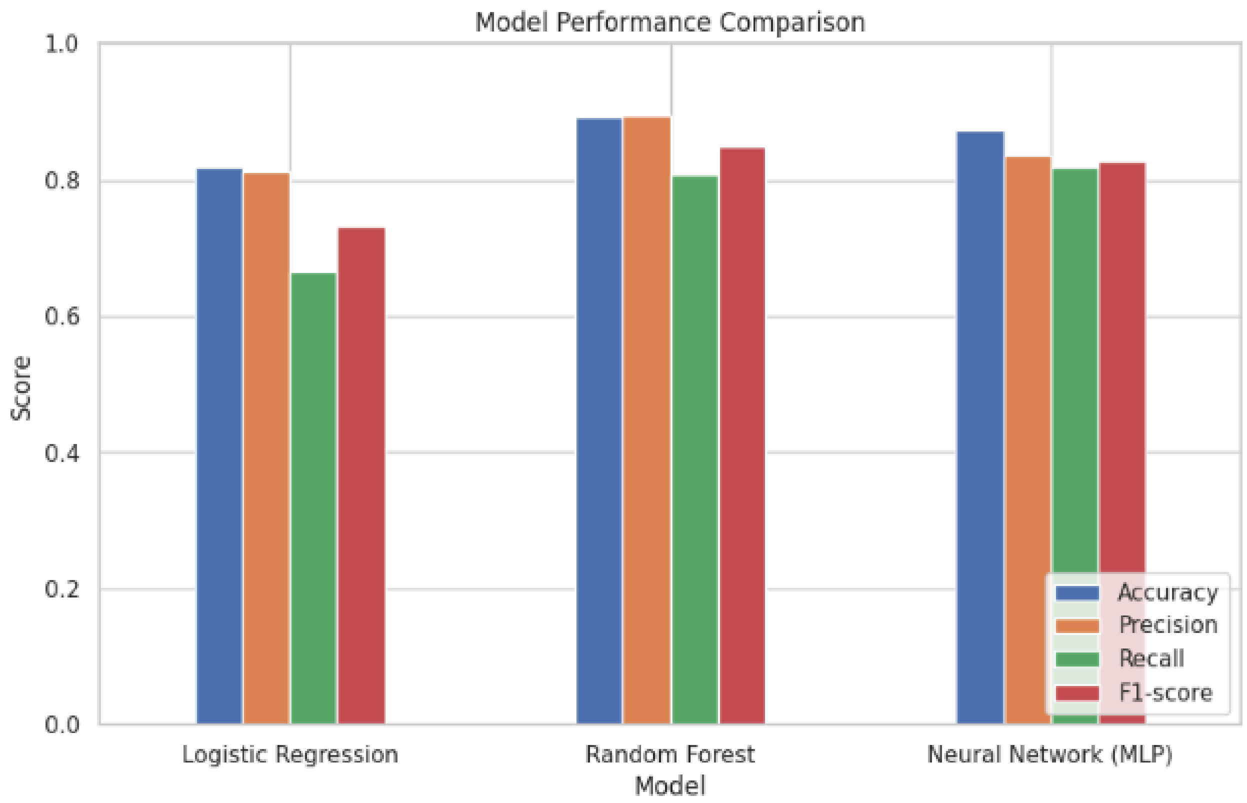


Fig 08: Model performance table.

Conclusion

From the results we can see that Random forest has the best overall performances as it has a good accuracy of 89.32% and the best F1 score of 84.86% while also demonstrating a strong balance between precision and recall. Its accuracy, precision, and recall scores were consistently higher than those of Logistic Regression and Neural Network which is shown in its ROC AUC score. Neural Networks however also demonstrated its efficiency in identifying canceled reservations and capturing non-linear patterns with an accuracy of 87.33% and a high recall of 81.83%. Logistic regression, on the other hand, performed very poorly in recall (66.55%). This indicates that even though it had a good precision, it missed more real cancellations. These suggest that models, such Random Forest and Neural Networks, are more appropriate for this issue than linear models, with Random Forest offering the most consistent and well-rounded performance out of the three. Furthermore, the use of K-Means clustering highlighted underlying patterns in the data that are not immediately visible through classification alone. Further revealing separate customer segments with varying cancellation tendencies. This integrated method not only increases prediction accuracy but also provides useful insights into various booking characteristics, facilitating improved decision-making and focused cancellation reduction tactics.