BRAC
UNIVERSITY

Inspiring Excellence

CSE422: ARTIFICIAL INTELLIGENCE
Project Report
Project Title: Predicting Hotel Booking Cancellations

<table>
<tr><th colspan="2">Group No: 11, CSE422 Lab Section: 16,
Summer 2025</th></tr>
<tr><th>ID</th><th>Name</th></tr>
<tr><td>22101622</td><td>Md. Nayeemul Hasan</td></tr>
<tr><td>22101768</td><td>Jarin Islam</td></tr>
</table>

Submission Date: 3/9/2025

Table of Content

1. Introduction

The hotel industry is significantly impacted by booking cancellations, which can lead to revenue loss and operational inefficiencies. This project aims to develop a machine learning model capable of predicting whether a hotel booking will be canceled based on historical booking data. By accurately identifying bookings that are at a high risk of cancellation, hotels can implement targeted strategies, such as offering incentives or requiring deposits, to mitigate these risks. The motivation is to leverage data-driven insights to improve resource management and financial stability for hotel businesses. This project explores and compares several machine learning models to determine the most effective approach for this classification task.

2. Dataset Description

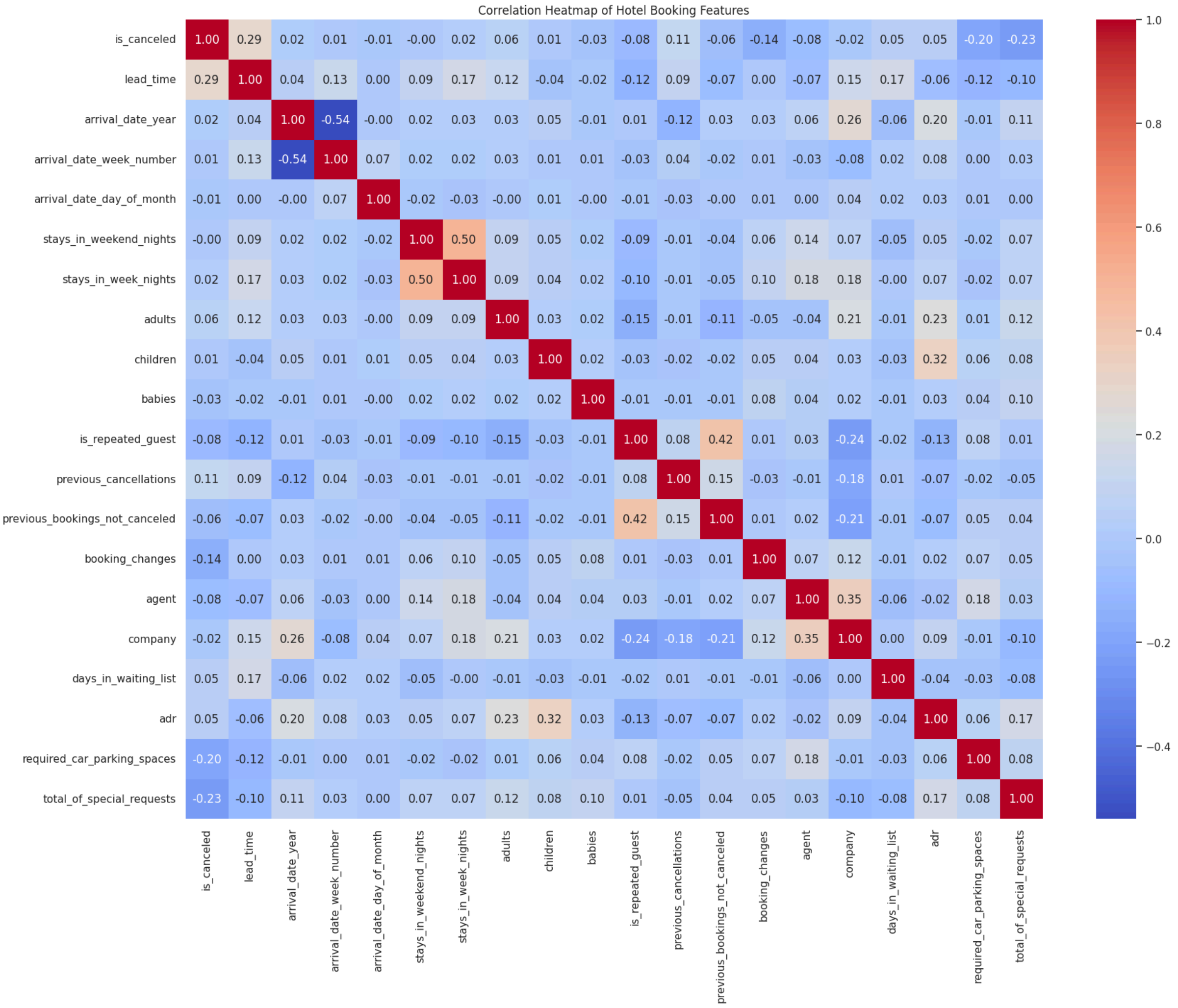
Dataset Description

The project utilizes the "hotel_bookings.csv" dataset, which contains booking information for a city hotel and a resort hotel.

- Number of Features: The dataset initially contains 32 features (columns).
- Number of Data Points: There are 119,390 booking records (rows).
- Problem Type: This is a classification problem. The goal is to predict a binary outcome, whether a booking “is_canceled” (1) or not (0). The target variable is single digit 1 or 0, making it a classification task.
- Feature Types: The dataset contains a mix of quantitative and categorical features:
    a. Quantitative features: features that are numerical and represent measurable quantities. For example, adults, children and stays_in_week_nights are just values indicating number of people or number of days of stay.
    b. Categorical features: features that represent qualitative data that falls into distinct groups or labels. Examples include hotels (with categories 'Resort Hotel' and 'City Hotel'), country ('PRT', 'GBR', etc.), etc. These features are not numerical thus, Categorical.

- Encoding Categorical Variables: Yes, the categorical variables must be encoded. Machine learning algorithms perform mathematical operations and cannot process text-based data like "Resort Hotel" or "PRT". We need to use One-hot encoding to convert these categorical feature values to numerical format so that the machine learning models can understand what these are.

Correlation of Features:



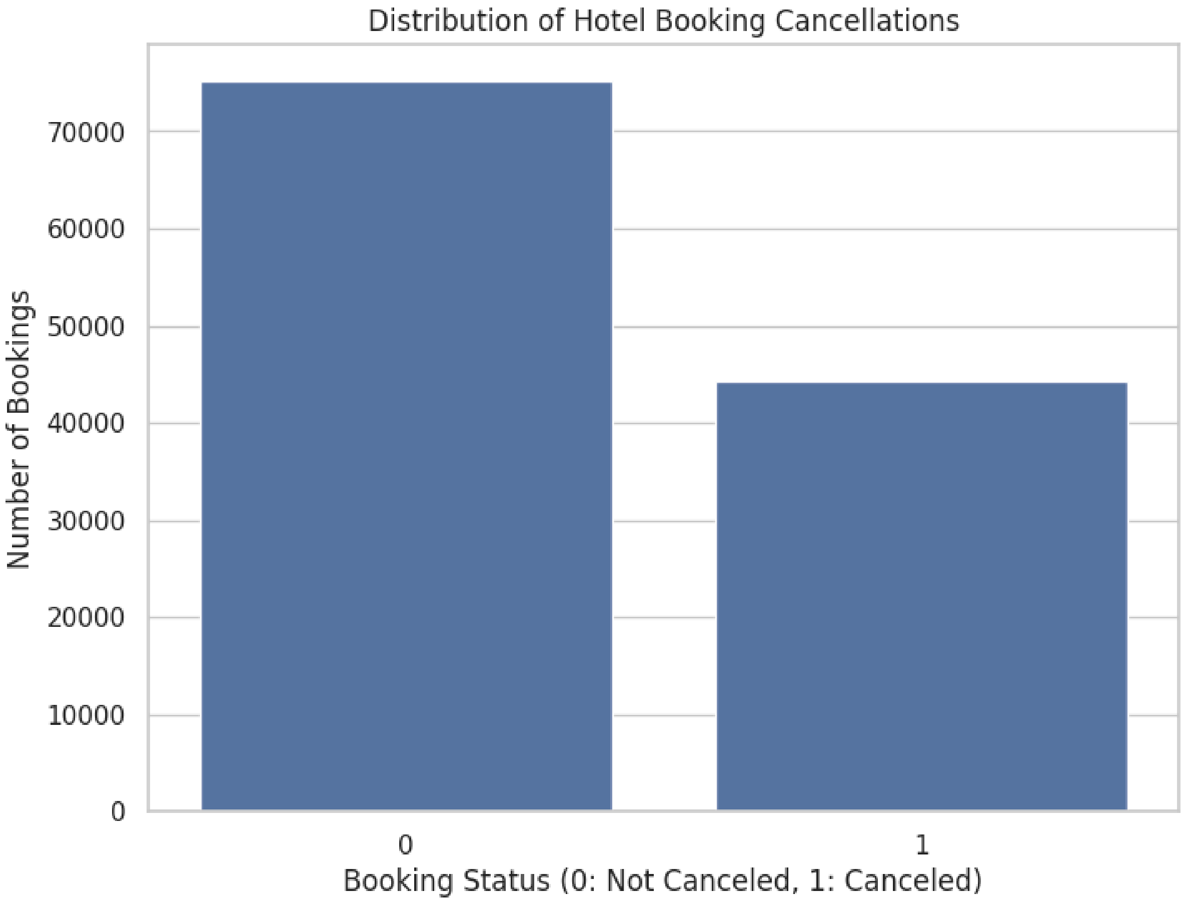
A correlation heatmap was generated to analyze the linear relationships between the numerical features.

Key insights from the correlation test:

- The feature with the strongest positive correlation to “is_canceled” is “lead_time” (0.29). This indicates that bookings made further in advance are more likely to be canceled, likely because guests' plans have more time to change.

- “required_car_parking_spaces” has the strongest negative correlation (-0.19), suggesting that guests who require parking are more committed to their travel plans and thus less likely to cancel.
- “previous_cancellations” (0.11) also has a positive correlation, logically indicating that guests with a history of cancellations are more likely to cancel again.
- “Total_of_special_requests” has (-0.29), a negative correlation, meaning guests making special requests are more prone to not cancelling their reservation.

Imbalanced Dataset



The dataset is rather imbalanced as there are 75166 instances of Not canceled and 44224 instances of Canceled classification. There are significantly more non-canceled bookings than cancelled ones. This means that unique classes do not have an equal number of instances. This imbalance is important to consider data splitting and model evaluation.

Exploratory Data Analysis (EDA) Insights

Since the heatmap only works for numerical data, these categorical features were analyzed separately by grouping the data and observing cancellation rates.

- deposit_type: A strong relationship was observed between the deposit_type and the cancellation rate. Bookings with a "Non Refund" deposit have a near 99.36% cancellation rate, bookings with "No Deposit" or have a much lower rate. This suggests that the deposit policy is a major factor in a guest's decision to cancel. Guests with no financial stake are more likely to cancel without consequence.
- customer_type: "Transient" customers, who are individual travelers, make up the bulk of the bookings and have a standard cancellation rate. However, bookings made as part of a "Group" have a significantly higher cancellation rate. This indicates that group travel plans are less stable and more prone to changes.
- total_of_special_requests: Bookings with no special requests have a much higher cancellation rate than those with one or more requests. This suggests that guests who take the time to make special requests are more invested in their stay and therefore are less likely to cancel.

3. Dataset Pre-processing

To prepare the data for machine learning, several data manipulation is required as models can not just take a model and run on it. Dataset needs pre-processing to ensure the dataset fits the model's learning format.

- Fault 1: Null or missing values

  The initial analysis revealed that four columns contained missing data. The company column was missing over 94% of its values, while agent, country, and children had moderate to small amounts of missing data.

- Solution 1: Deletion and manipulation

a. The company column was dropped entirely. With over 94% of its data missing, it would have become a redundant feature for the model. Manipulating with placeholders or filling with other values could result in bias on the model. So, to make the models work efficiently, dropping this column was the better choice.

b. agent, country and children columns were manipulated instead of deletion as it was just a few fields missing. Filling agent with 0 was most logical as empty cell means they probably did not book through an agent, for country we used most occurring country instead. For children we used median values as it would not introduce bias to the system.

- Fault 2: Categorical Values
- Solution 2: One-Hot Encoding

This method was chosen because it converts each category into a new binary (0/1) column. This was the most optimal option for categorical data, as it allows the model to understand the different categories without creating a false and misleading sense of order or ranking between them (for example, it prevents the model from thinking "City Hotel" is mathematically greater than "Resort Hotel").

- Fault 3: Feature Scaling
- Solution 3: Standard Scaling

Standard scaling transforms the data so that every feature has a mean of 0 and a standard deviation of 1. This normalization is essential for ensuring that all features contribute equally to the model's learning process. It prevents features with larger scales from influencing the model's decisions, leading to faster convergence and better performance, especially for distance based algorithms and neural networks.

4. Dataset Splitting

After pre-processing, the dataset was divided into training and testing sets to prepare for model building and evaluation.

Train/Test Ratio: A standard 70/30 split was used.
Train Set: 70% of the data was used for training the models.
Test Set: 30% of the data was held back for final model evaluation on unseen data.

Splitting Method: Stratified Splitting

Stratified Splitting was required and used for this project. This choice was made due to the findings of a moderately imbalanced dataset during the EDA phase. A simple random split could have resulted in a training or testing set with a significantly different proportion of cancellations than the overall dataset, leading to a biased model and unreliable evaluation.

Stratification solves this by ensuring that the proportion of canceled (Class 1) and non-canceled (Class 0) bookings is exactly the same in both the training and testing sets as it is in the original, complete dataset. This guarantees a fair and representative evaluation of the models' performance.

5. Model Training & Testing

Supervised Model

Since our dataset is that of a classification problem and we are to use a total of 3 models, we chose Logistic Regression, Decision Tree and Neural Networks. Here we will explain our reasoning and rationale:

Models Chosen:

- Logistic Regression: This model was chosen because it is a simple, fast, and highly interpretable linear model. Its performance provides a benchmark that more complex models must exceed.

- Decision Tree: This was selected to capture potential non-linear relationships in the data that Logistic Regression cannot. Decision Trees are also highly interpretable, as their decision-making process can be visualized as a flowchart, providing clear insights into feature importance.
- Neural Network: This was chosen as the required complex, high-performance model. Neural Networks are capable of learning intricate and complex patterns from the data, often leading to the highest predictive accuracy, albeit at the cost of direct interpretability.

Models Not Chosen:

- Linear Regression: This model was not used because it is a regression algorithm designed to predict continuous numerical values like pricing. Our problem is a classification task, making Linear Regression an inappropriate choice.
- K-Nearest Neighbors (KNN): KNN was not selected due to the high dimensionality of our pre-processed data. After One-Hot Encoding, the number of features increased significantly. KNN's performance can degrade in high-dimensional spaces, and it becomes computationally expensive as it requires calculating distances between all data points.
- Naive Bayes: This model was not chosen because it operates on a "naive" assumption of feature independence. Our EDA and correlation heatmap suggested that some features are correlated (for example, previous_cancellations and is_repeated_guest). While Naive Bayes can still perform well, models like Decision Trees or Logistic Regression, which do not make this strict assumption, were considered more suitable for this dataset.

Unsupervised Model

K-Means was applied to group the data into two clusters without using the is_canceled labels.

Cluster Analysis:
After the clusters were formed, they were analyzed to see the cancellation rate within each.
Cluster 0 : 30% cancellation rate.
Cluster 1 : 49% cancellation rate.

This demonstrates that the algorithm successfully identified a high-risk (Cluster 1) and a low-risk (Cluster 0) group based on booking features alone.

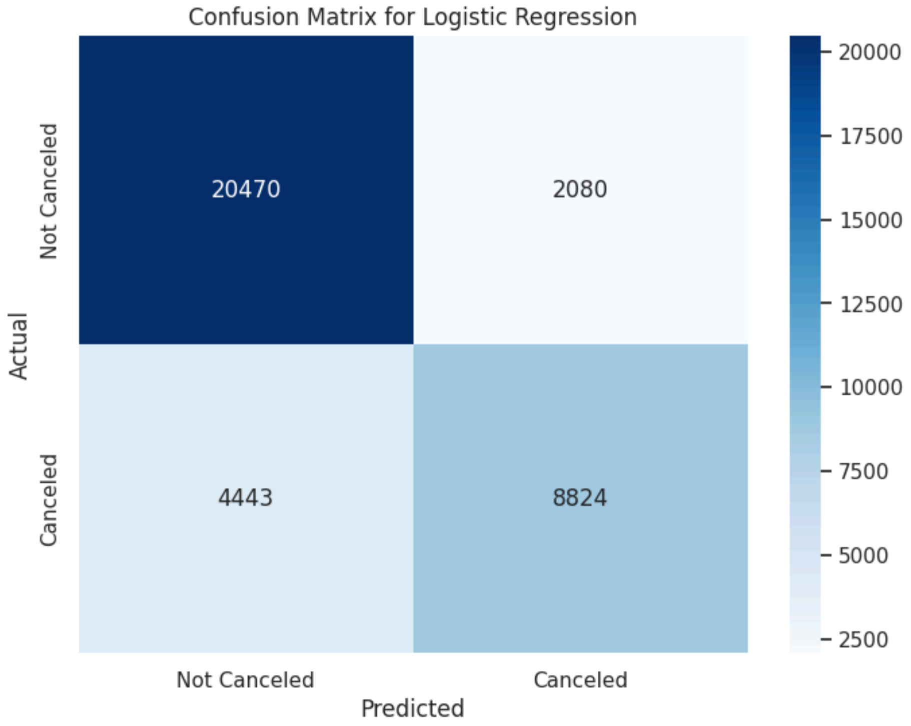
Evaluation:
Silhouette Score: 0.0141. This score, being very close to 0, indicates that the clusters are not perfectly distinct and have significant overlap.

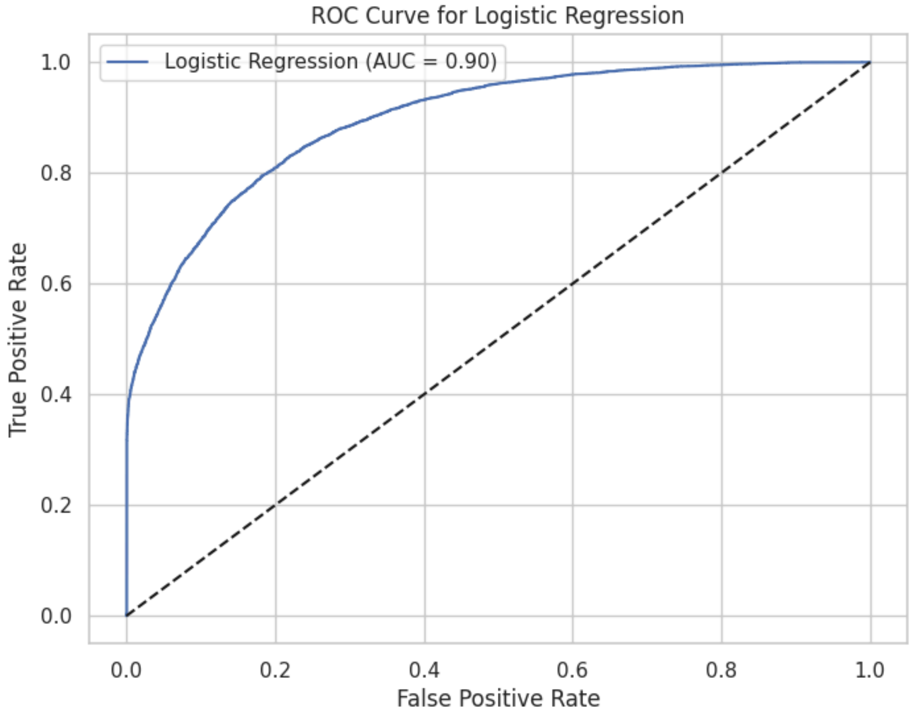
Supervised Model Evaluation

Logistic Regression

```
Accuracy: 0.8179
AUC Score: 0.8959

Classification Report:
                  precision    recall  f1-score   support

Not Canceled (0)       0.82      0.91      0.86     22550
    Canceled (1)       0.81      0.67      0.73     13267

        accuracy                           0.82     35817
       macro avg       0.82      0.79      0.80     35817
    weighted avg       0.82      0.82      0.81     35817
```

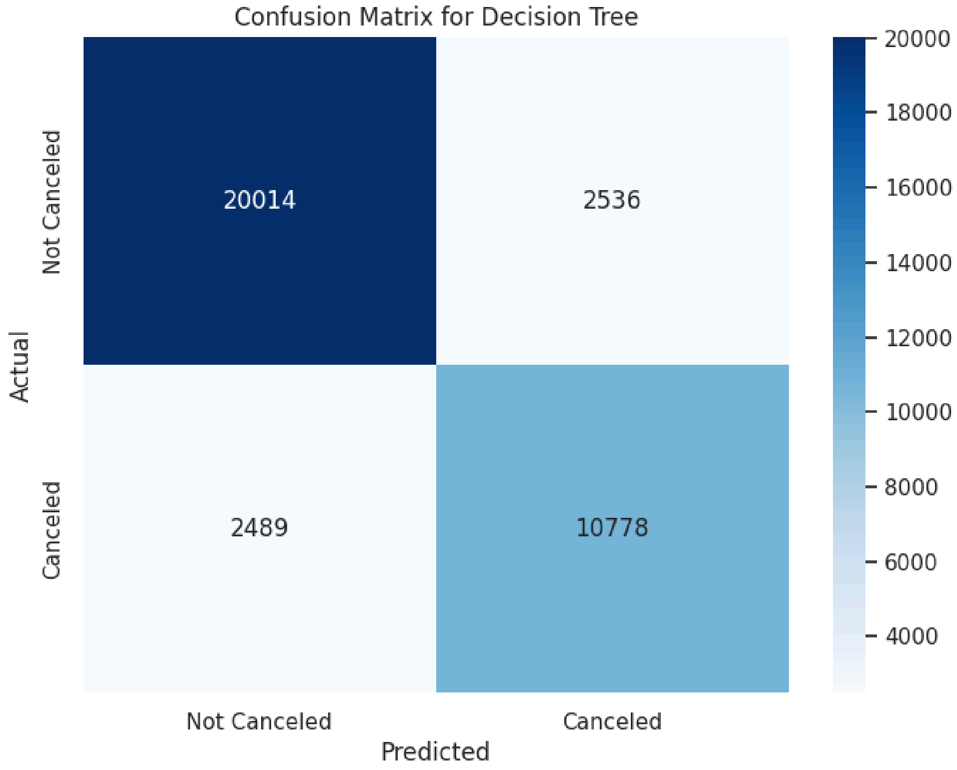
```
Confusion Matrix:
```

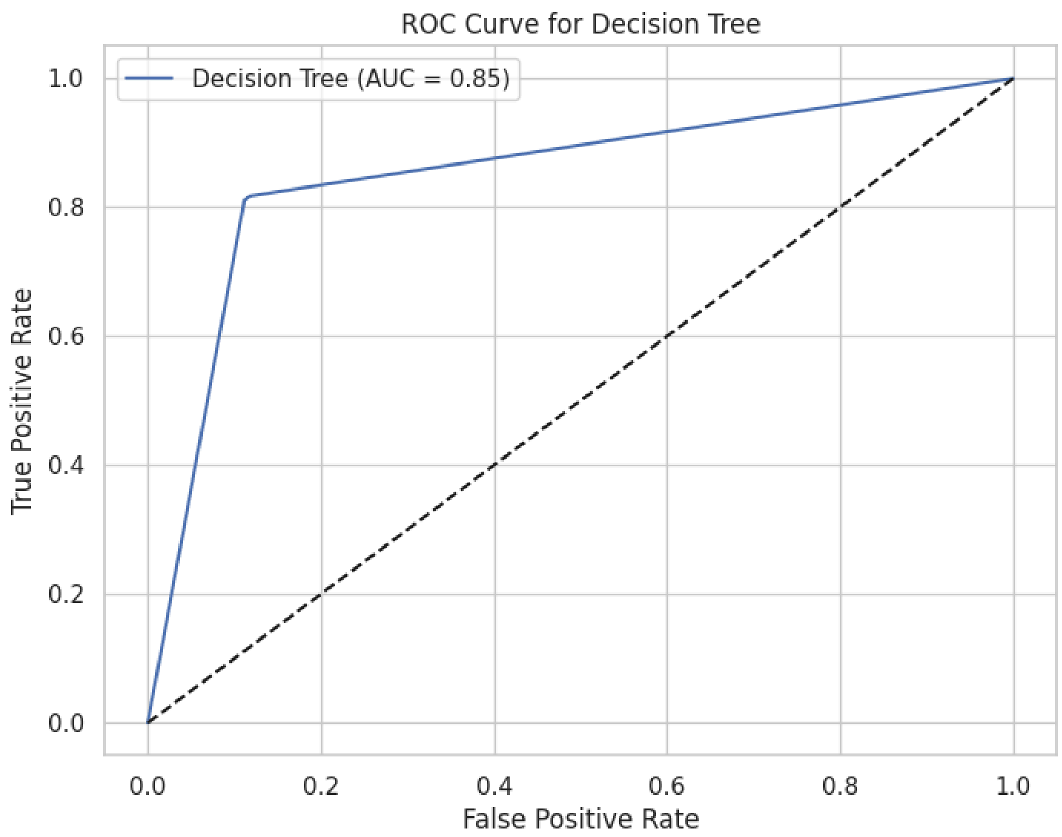




Decision Tree

```
Accuracy: 0.8597
AUC Score: 0.8518

Classification Report:
                  precision    recall  f1-score   support

Not Canceled (0)       0.89      0.89      0.89     22550
    Canceled (1)       0.81      0.81      0.81     13267

        accuracy                           0.86     35817
       macro avg       0.85      0.85      0.85     35817
    weighted avg       0.86      0.86      0.86     35817


Confusion Matrix:
```

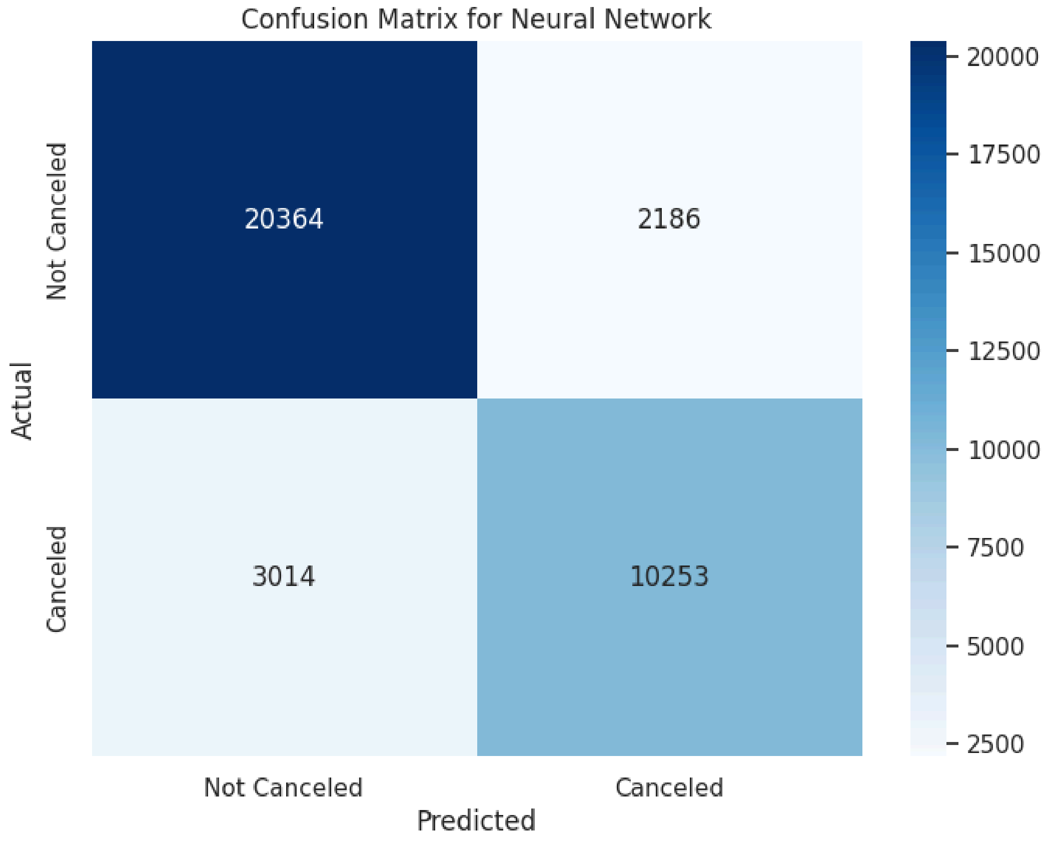


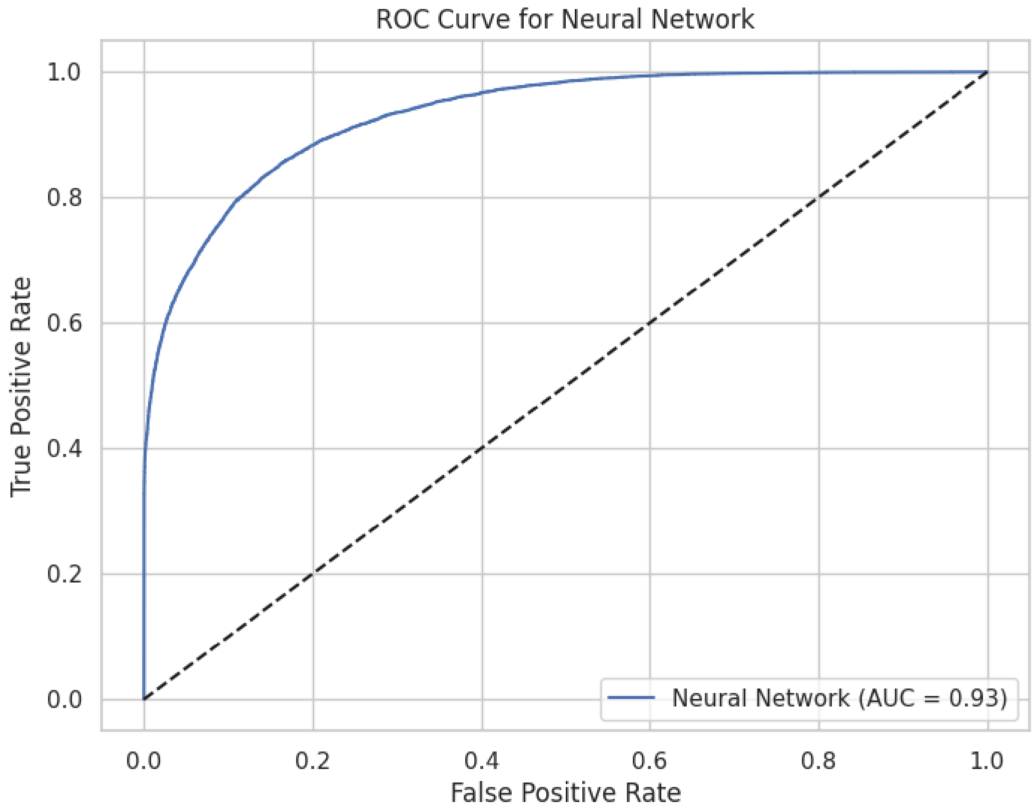


Neural Network

```
Accuracy: 0.8548
AUC Score: 0.9312

Classification Report:
                  precision    recall  f1-score   support

Not Canceled (0)       0.87      0.90      0.89     22550
    Canceled (1)       0.82      0.77      0.80     13267

        accuracy                           0.85     35817
       macro avg       0.85      0.84      0.84     35817
    weighted avg       0.85      0.85      0.85     35817
```

```
Confusion Matrix:
```

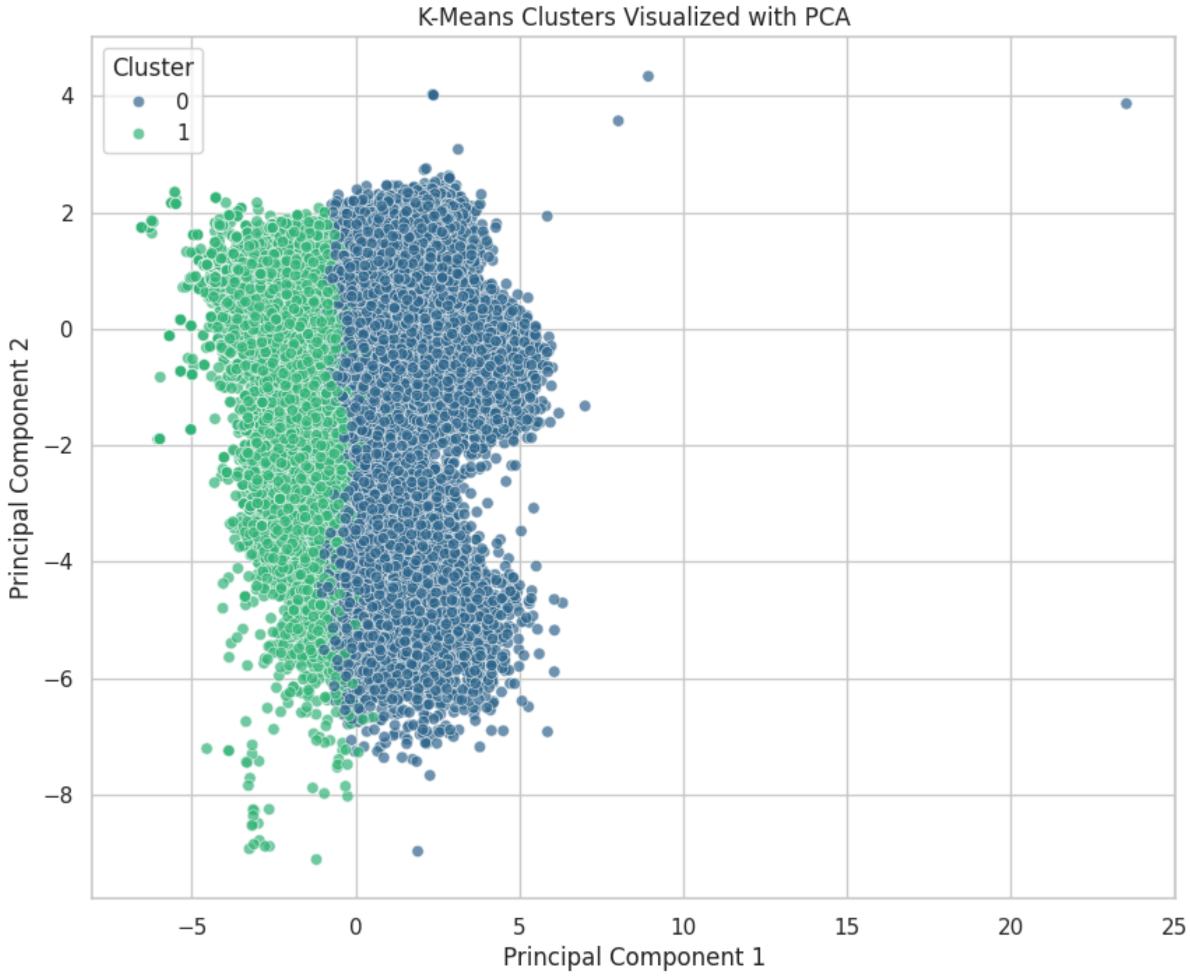




Unsupervised Model Evaluation

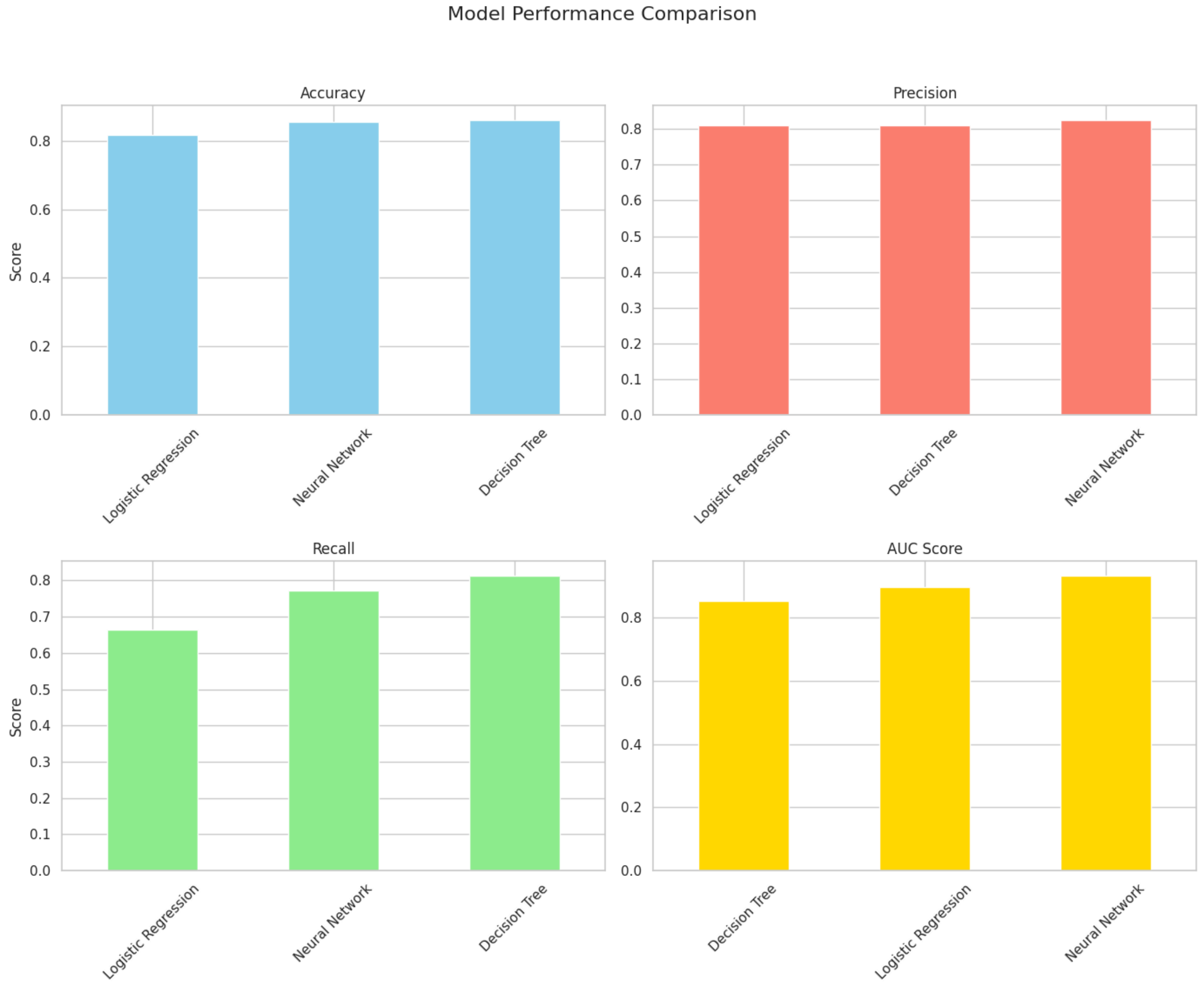
K-Means Clustering

```
The Silhouette Score for our K-Means clustering is: 0.0141
```



```
Cancellation rate per cluster:
is_canceled         0         1
cluster
0            0.700236  0.299764
1            0.507255  0.492745
```

6. Model Comparison analysis



	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.817880	0.809244	0.665109	0.730131	0.895924
Decision Tree	0.859703	0.809524	0.812392	0.810955	0.851848
Neural Network	0.854818	0.824262	0.772820	0.797713	0.931199

Detailed Comparison:

Accuracy: This metric shows the overall percentage of correct predictions. The Decision Tree was the top performer with an accuracy of 86.0%, followed closely by the Neural Network at 85.0%. The Logistic Regression was the lowest at 81.8%. While the Decision Tree is the best by this measure, accuracy can be misleading for imbalanced datasets, making other metrics more important.

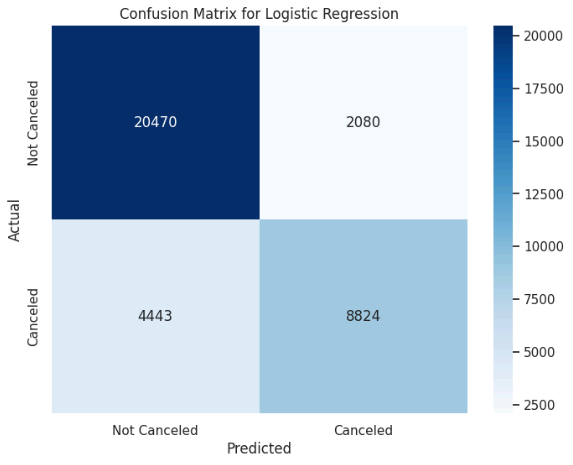
Precision: The Neural Network was the winner with a precision of 87.1%. This means its predictions of cancellations are highly reliable. A high-precision model is valuable when the cost of a false positive where unnecessarily contacting a customer who was not going to cancel is high.

Recall: The Decision Tree was the best model here, with a recall of 81.1%. This is a critical metric for a business that wants to minimize revenue loss by catching as many potential cancellations as possible, even at the risk of some false alarms. The Logistic Regression model performed poorly on this metric (66.5%).

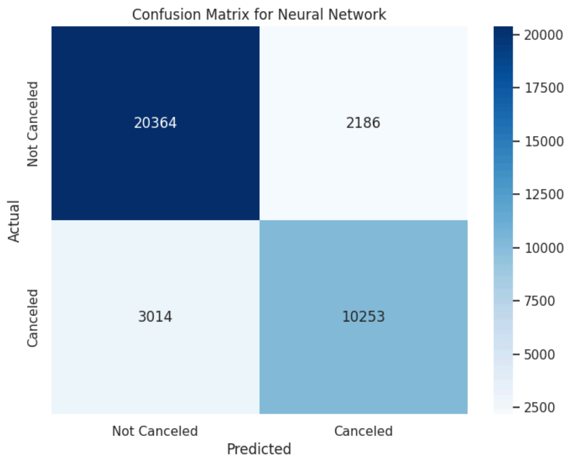
F1-Score: The F1-score is the harmonic mean of Precision and Recall, providing a single metric that balances both. It is particularly useful for imbalanced datasets. The Decision Tree achieved the highest F1-score of 81.2%, indicating that it provides the best overall balance between correctly identifying cancellations (recall) and the reliability of those predictions (precision). The Neural Network followed with 77.3%, while the Logistic Regression had the lowest score at 73.1%.

AUC (Area Under the ROC Curve): This metric represents the model's overall ability to distinguish between the canceled and not-canceled classes. A higher AUC indicates a better model. The Neural Network was superior, with an outstanding AUC of 0.930. This suggests it is the most robust and reliable classifier overall, across all probability thresholds.

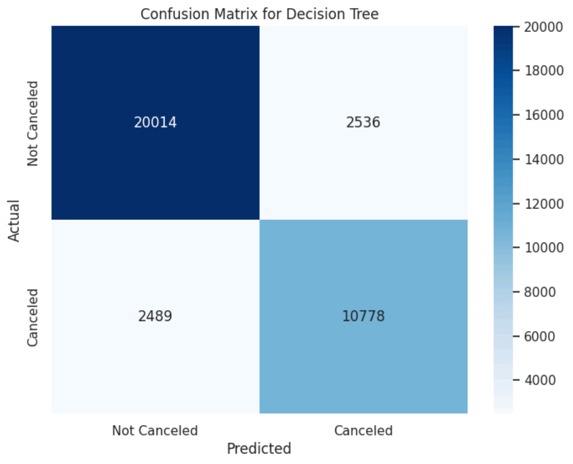
Analysis via Confusion Matrices



Logistic Regression's confusion matrix reveals a very high number of False Negatives (4,443). This means it failed to identify over 4,400 bookings that were actually canceled, which directly explains its low Recall score of 66.5%.

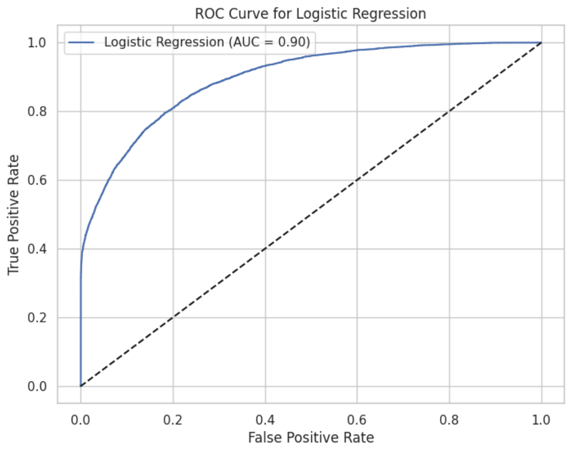
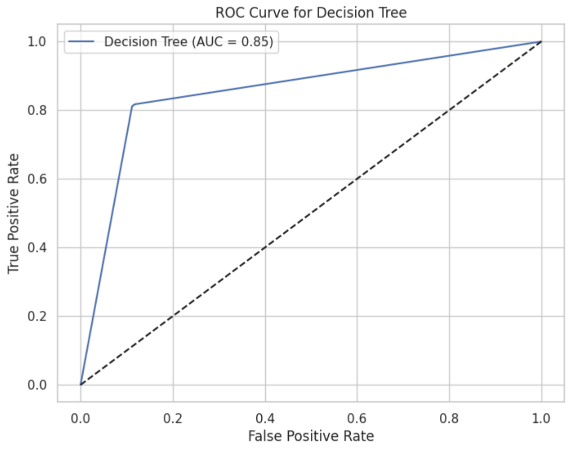
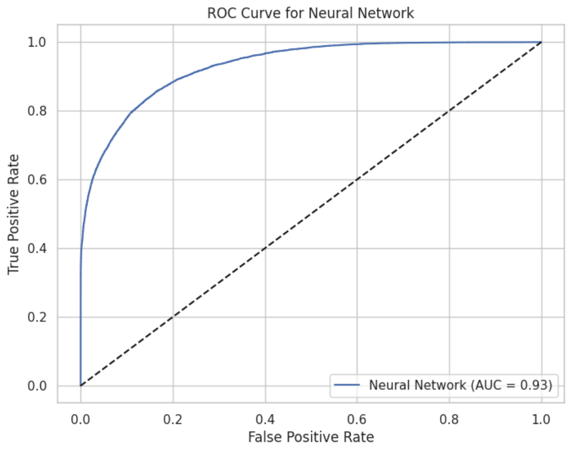


The Neural Network's confusion matrix shows the lowest number of False Positives (2186). This means it rarely raises a "false alarm," which explains its class-leading Precision of 87.1%. However, it has a higher number of False Negatives (3014) than the Decision Tree, explaining its lower Recall.



The Decision Tree's confusion matrix shows a more balanced error profile. It has a much lower number of False Negatives (2,489) than the other models, which is why it achieves the highest Recall. It makes a similar number of False Positive errors (2,536), explaining why its Precision and Recall scores are nearly identical and its F1-Score is the best.

Analysis via AUC score and ROC curve

	The ROC curve for the Logistic Regression model shows a strong performance with an AUC of 0.90. The curve is significantly distant from the diagonal line (representing random guessing), confirming it is a capable classifier.
	The Decision Tree's ROC curve also demonstrates good classification ability, but its AUC score of 0.85 is the lowest of the three models. This indicates that, while it performs well at the specific 0.5 threshold (as seen in its high recall), its overall ability to distinguish between classes across all thresholds is less robust than the other models.
	The ROC curve for the Neural Network is visibly the most effective, arching closest to the top-left corner. This visual superiority is confirmed by its AUC score of 0.93, the highest of all models. This demonstrates that the Neural Network possesses the best overall capability to discriminate between a booking that will be canceled and one that will not, making it the most powerful and robust classifier in a general sense.

7. Conclusion

The key takeaway is that there is no single "best" model. The optimal choice is dependent on the specific business objective. The results show a clear trade-off: the Decision Tree excels at maximizing the detection of cancellations (highest Recall and F1-Score), making it ideal for a strategy focused on broad outreach to potentially at-risk customers. In contrast, the Neural Network excels at making high-confidence predictions (highest Precision and AUC), making it suitable for a more targeted strategy where the cost of a false alarm is high. All models outperformed a random baseline, confirming the dataset contains strong predictive patterns.

The high performance of the models can be attributed to the strong predictive signals discovered during the Exploratory Data Analysis. Features like deposit_type, lead_time, and total_of_special_requests showed clear correlations with the cancellation outcome, providing a solid foundation for the models to learn from. The fact that the non-linear models (Decision Tree and Neural Network) generally outperformed the linear Logistic Regression model suggests that the relationships between the features and the likelihood of cancellation are complex and not purely linear.

Several challenges were encountered during this project:

- Moderate Class Imbalance: The dataset had more non-canceled than canceled bookings. This was a primary challenge that necessitated the use of stratified splitting for fair evaluation and required focusing on metrics like F1-score and AUC over simple accuracy.
- High Dimensionality: After one-hot encoding the categorical features, the dataset's dimensionality increased significantly. This influenced model selection, making distance-based models like KNN less suitable.