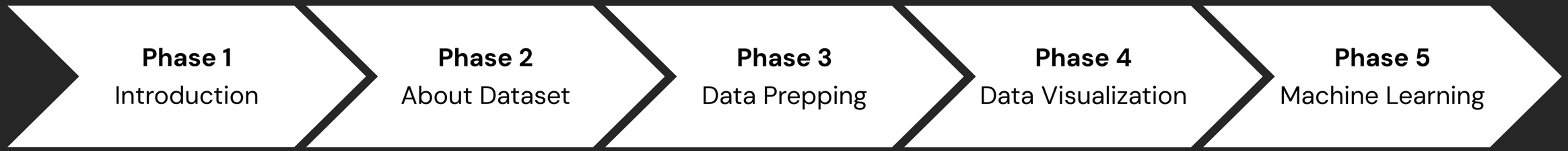
Hands on Machine Learning
Project Presentation.

Presented By: Bate Bita

Predicting Hotel Booking Cancellations.

Outline of Presentation



Source of Dataset:

Data for this project was purposefully collected by **Nuno Antonio, Ana de Almeida and Luis Nunes,** and published by **Elsevier Inc.** They obtained the data directly from 2 hotels' PMS database servers. Their findings are available on the **ScienceDirect** website.

The Dataset:

¬ Contains hotel data from two hotels in Portugal.

¬ Data was collected between the 1st of July, 2015 and the 31st of August, 2017.

¬ A total of 119,390 recorded observations of hotel bookings.

¬ 31 features are made available in the dataset.

“

Phase 1: Introduction

Project Objective:

Predicting whether or not a certain Customer will Cancel their Hotel Booking.

Project Characteristics:

¬ This is a **Binary Classification Problem**, given that our Target Variable in a Binary Format (0 or 1).

¬ Our dataset is **imbalanced**, with less people cancelling their hotel reservations.

¬ Due to the imbalanced nature of our dataset, Perfomance Metrics will be **Precision, Recall and F1 - Scores.**

¬ Models to be trained include:

=> Logistic Regression

=> K-Nearest Neighbours

=> Random Forests

=> Decision Trees

“

Phase 2: About Dataset

Purpose of Data Collection:

¬ To perform research in different problems such as booking cancellation prediction, customer segmentation, seasonality, etc.

¬ To evaluate the performance of different algorithms for solving similar types of problems (e.g classification, segmentation problems).

¬ To obtain statistics for data mining training.

¬To help Hotel Managers understand customer behaviours.

“

Phase 3: Data Prepping

Data Prepping Steps:

¬ Loading Data into Pandas DataFrame.

¬ Merging Datasets of the two Hotels.

¬ Understanding Features.

¬ Handling Missing Values.

¬ Splitting Data into Train and Test Splits.

Little Snippet into Dataset:

```
dataset
```

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth
0	0	342	2015	July
1	0	737	2015	July
2	0	7	2015	July
3	0	13	2015	July
4	0	14	2015	July
...	...	...	...	...
119385	0	23	2017	August
119386	0	102	2017	August
119387	0	34	2017	August
119388	0	109	2017	August
119389	0	205	2017	August

119390 rows × 31 columns

Data Features

Name of Feature	Description
ADR	Average Daily Rate as defined by the American Hotel and Lodging Association. It is calculated as the Total Revenue from Rooms Sold divided by Number of Rooms Sold.
Adults	Number of adults.
Agent	ID of the travel agency that made the booking.
ArrivalDateDayOfMonth	Day of the month of the arrival date.
ArrivalDateMonth	Month of arrival date with 12 categories: “January” to “December”.
ArrivalDateWeekNumber	Week number of the arrival date.
ArrivalDateYear	Year of arrival date.
AssignedRoomType	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request.
Babies	Number of babies.
BookingChanges	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation.
Children	Number of children.
Company	ID of the company/entity that made the booking or responsible for paying the booking.
Country	Country of origin.
CustomerType	Type of booking, assuming one of four categories: Contract, Group, Transient and Transient-party.
DaysInWaitingList	Number of days the booking was in the waiting list before it was confirmed to the customer.
DepositType	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: “No Deposit”, “Non Refund” and “Refundable”.

Data Features

Name of Feature	Description
DistributionChannel	Booking distribution channel. The term “TA” means “Travel Agents” and “TO” means “Tour Operators”.
IsCanceled	Value indicating if the booking was canceled (1) or not (0).
IsRepeatedGuest	Value indicating if the booking name was from a repeated guest (1) or not (0).
LeadTime	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date.
MarketSegment	Market segment designation.
Meal	Type of meal booked. Categories are presented in standard hospitality meal packages: BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch and dinner).
PreviousBookingsNotCanceled	Number of previous bookings not cancelled by the customer prior to the current booking.
PreviousCancellations	Number of previous bookings that were cancelled by the customer prior to the current booking.
RequiredCardParkingSpaces	Number of car parking spaces required by the customer.
ReservationStatus	Reservation last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show – customer did not check-in and did inform the hotel of the reason why.
ReservationStatusDate	Date at which the last status was set. This variable can be used in conjunction with the ReservationStatus to understand when was the booking canceled or when did the customer checked-out of the hotel.
ReservedRoomType	Code of room type reserved.
StaysInWeekendNights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel.
StaysInWeekNights	Number of week nights (Monday to Friday) the guest stayed or booked to stay at the hotel.
TotalOfSpecialRequests	Number of special requests made by the customer (e.g. twin bed or high floor).

Classification of Data Features:

Categorical	Numerical		Datetime
1. ArrivalDateMonth	1. IsCanceled	13. PreviousBookingsNotCanceled	1. ReservationStatusDate
2. Meal	2. LeadTime	14. BookingChanges	
3. Country	3. ArrivalDateYear	15. DaysInWaitingList	
4. MarketSegment	4. ArrivalDateWeekNumber	16. ADR	
5. DistributionChannel	5. ArrivalDateDayOfMonth	17. RequiredCarParkingSpaces	
6. ReservedRoomType	6. StaysInWeekendNights	18. TotalOfSpecialRequests	
7. AssignedRoomType	7. StaysInWeekNights		
8. DepositType	8. Adults		
9. Agent	9. Children		
10. Company	10. Babies		
11. CustomerType	11. IsRepeatedGuest		
12. ReservationStatus	12. PreviousCancellations		

Checking for Missing Values:

```
Columns with missing values:
Children       4
Country      488
dtype: int64
```

1) Handling 'Children' Missing Values:

¬ Mode of Column was calculated and imputed in the 4 missing entries.

2) Handling 'Country' Missing Values:

With the aim of ensuring fairness and impartiality in our Machine Learning Models, we will be dropping the 'Country' column entirely. This feature provides information about the customer's origin and including it in the model could introduce ethical concerns and potential biases, which is something I am committed to avoiding.

3) Handling Redundant Columns:

We will be dropping the '**ReservationStatus**' and '**ReservationStatusDate**' Columns for the following reasons:

The Reservation Status has a perfect correlation to our target feature, 'IsCanceled'. It has two values: 'Checked Out' and 'Canceled'. Including this feature in the model training will result in a 100% prediction accuracy, which is infact not the case.

The Reservation Status Date on the other hand, is the date for which the Reservation Status was recorded. It does not play any role in helping the model predict future cancellations.

“Phase 4: Data Visualization

Data Visuals:

¬ Target Feature - 'IsCancelled'.

¬ Train/Test Split.

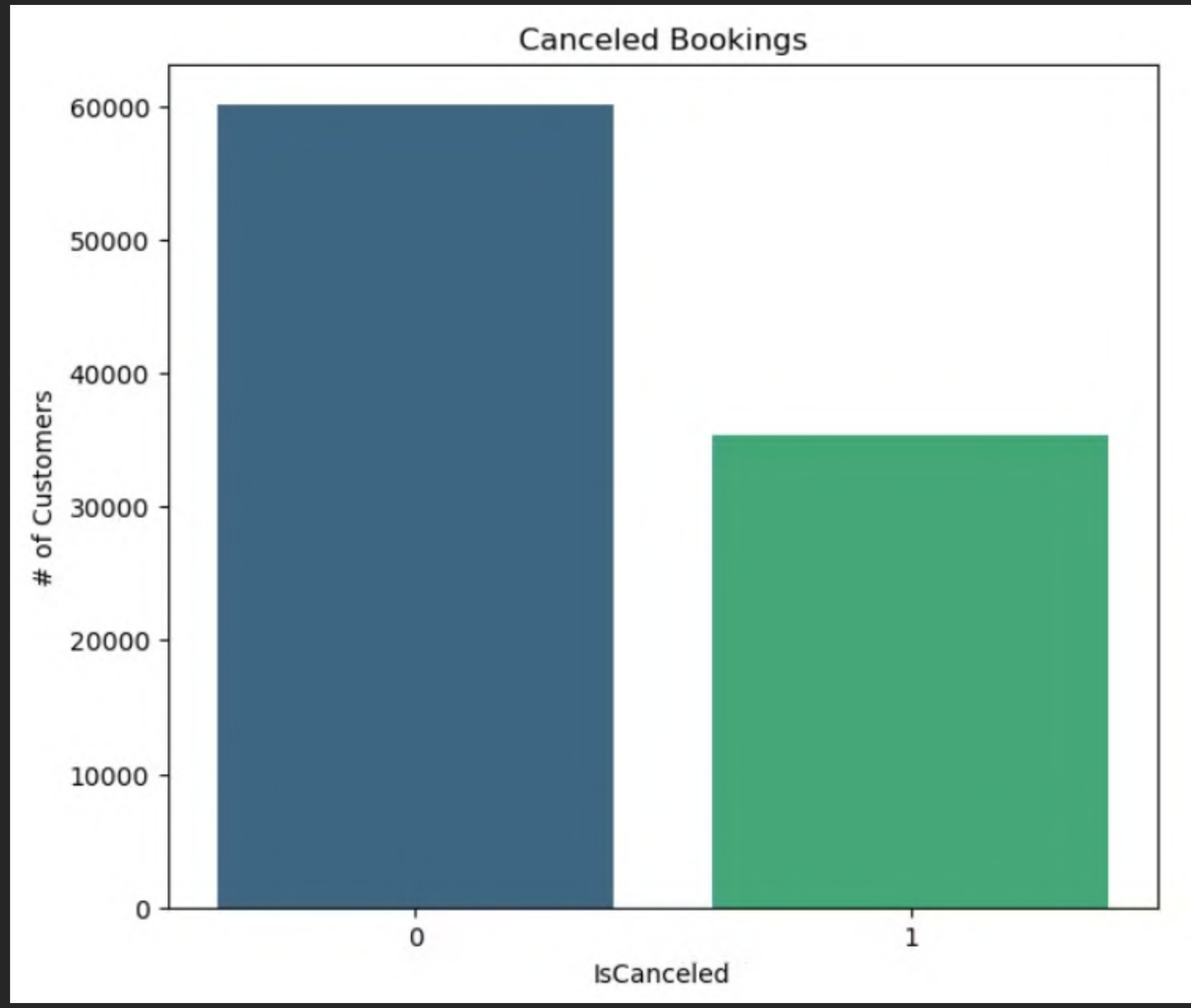
¬ Number of Cancellations by:

=> Cutomer Type

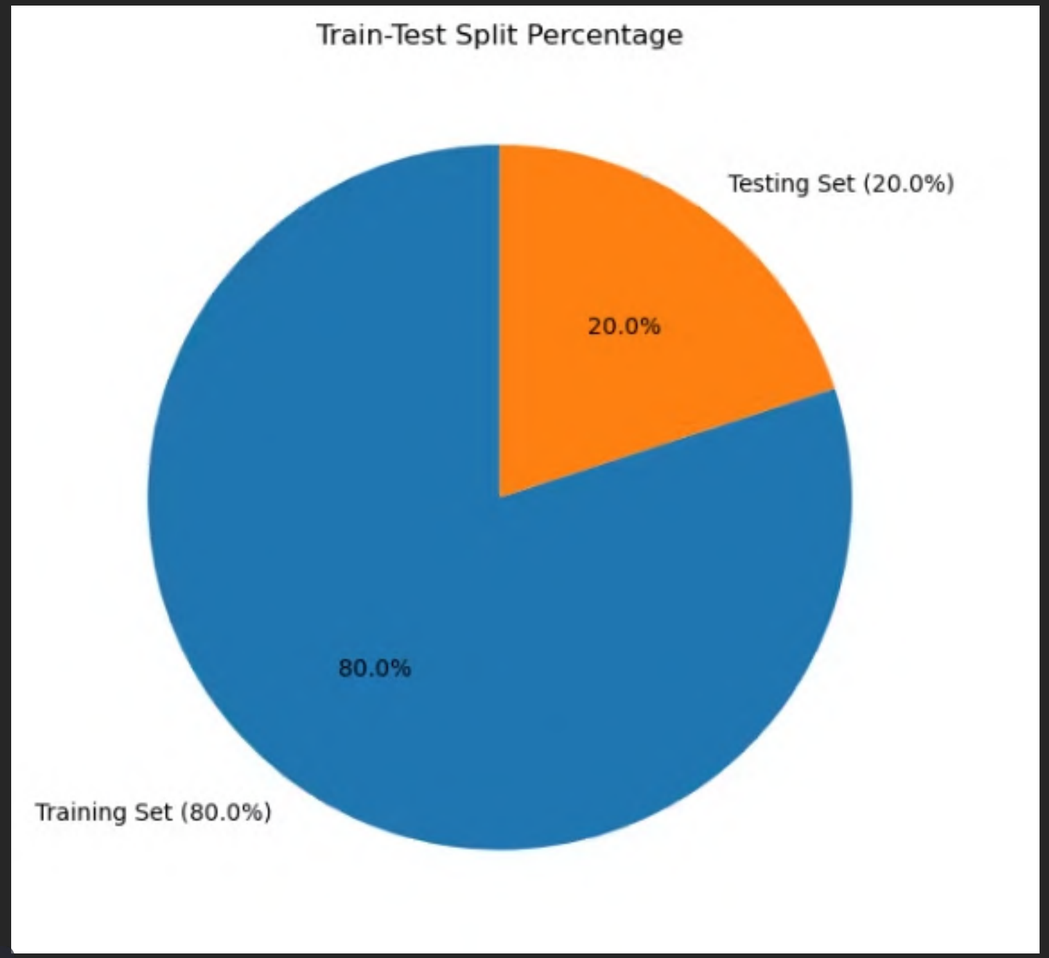
=> Deposit Type

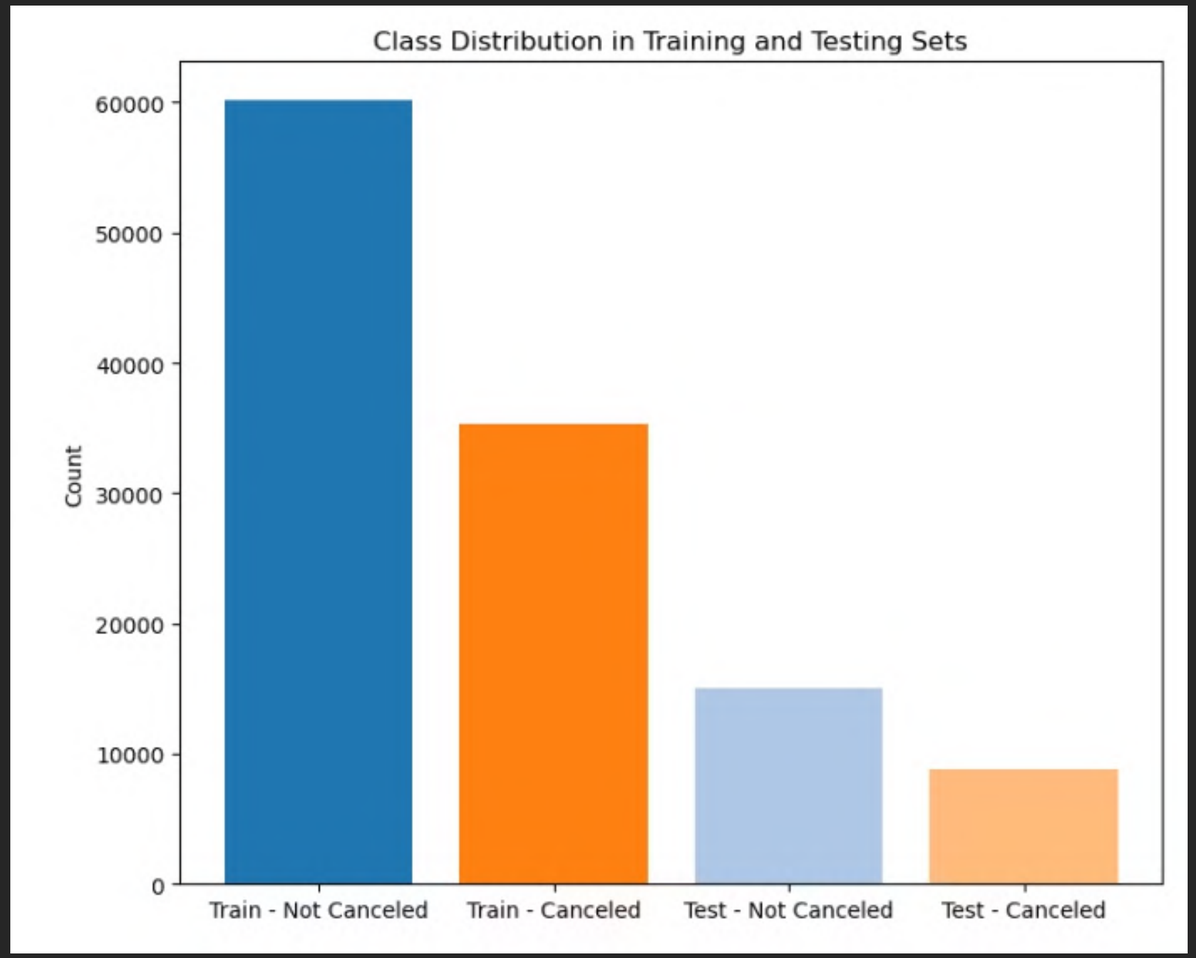
=> Market Segment.

This visualization provides a simple overview of how many customers canceled their bookings and how many didn't. **0** means that the customer **'Did Not Cancel'** their booking and **1** means the customer **'Did Infact Cancel'** their hotel booking. This plot tells us that our data is imbalanced, with less number of people actually cancelling their hotel booking.

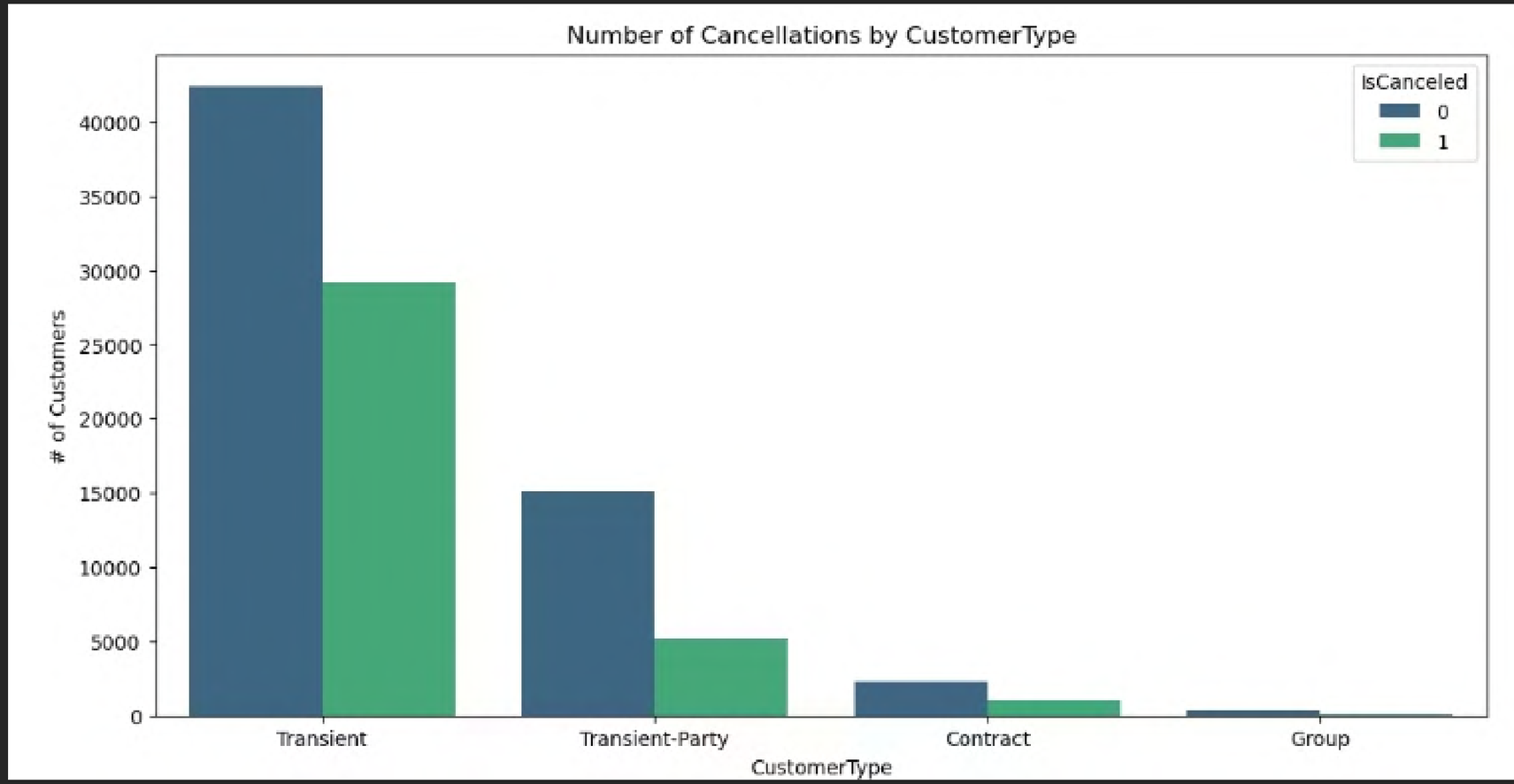


Visualizing Train and Test Sets.

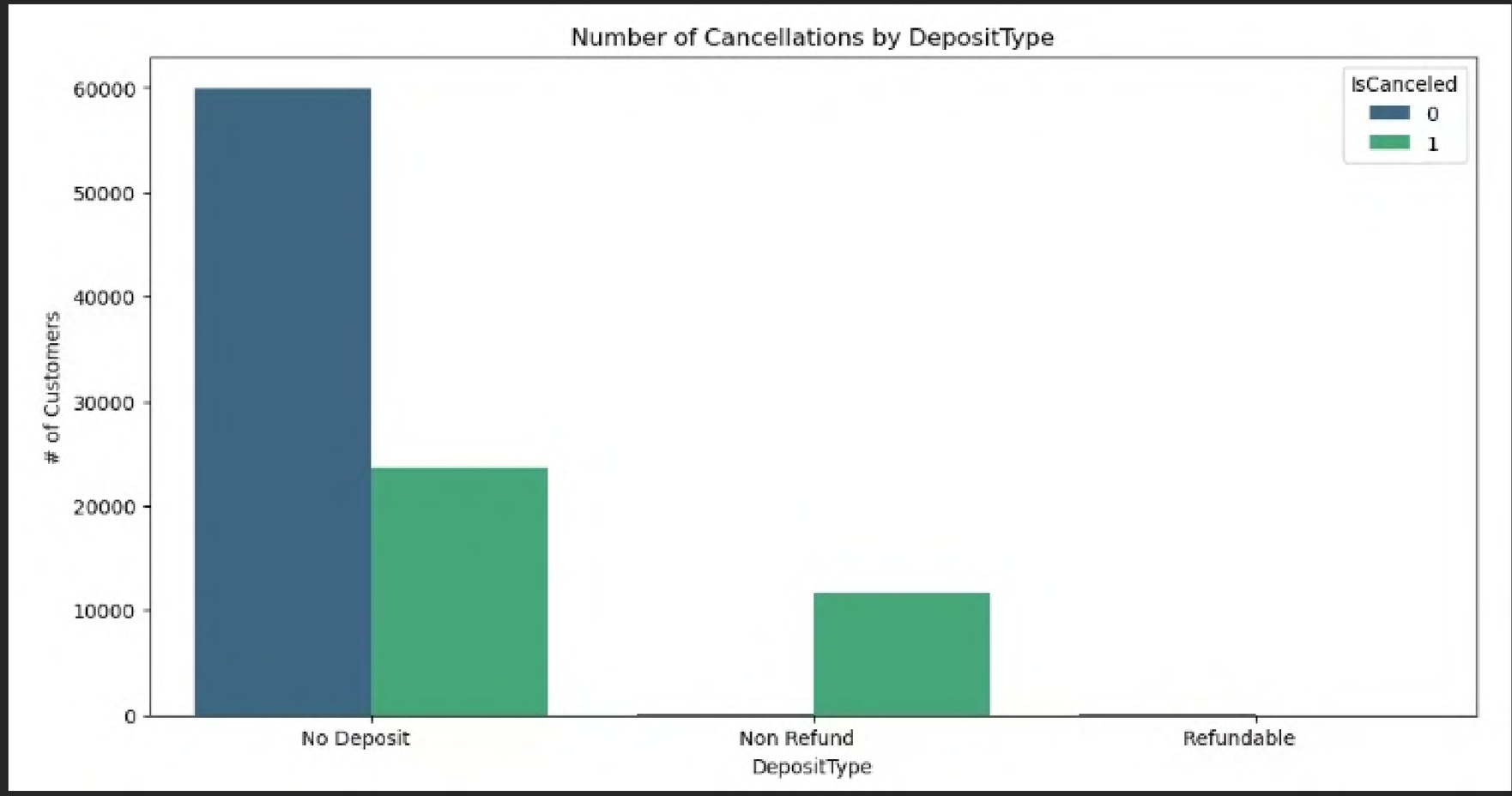




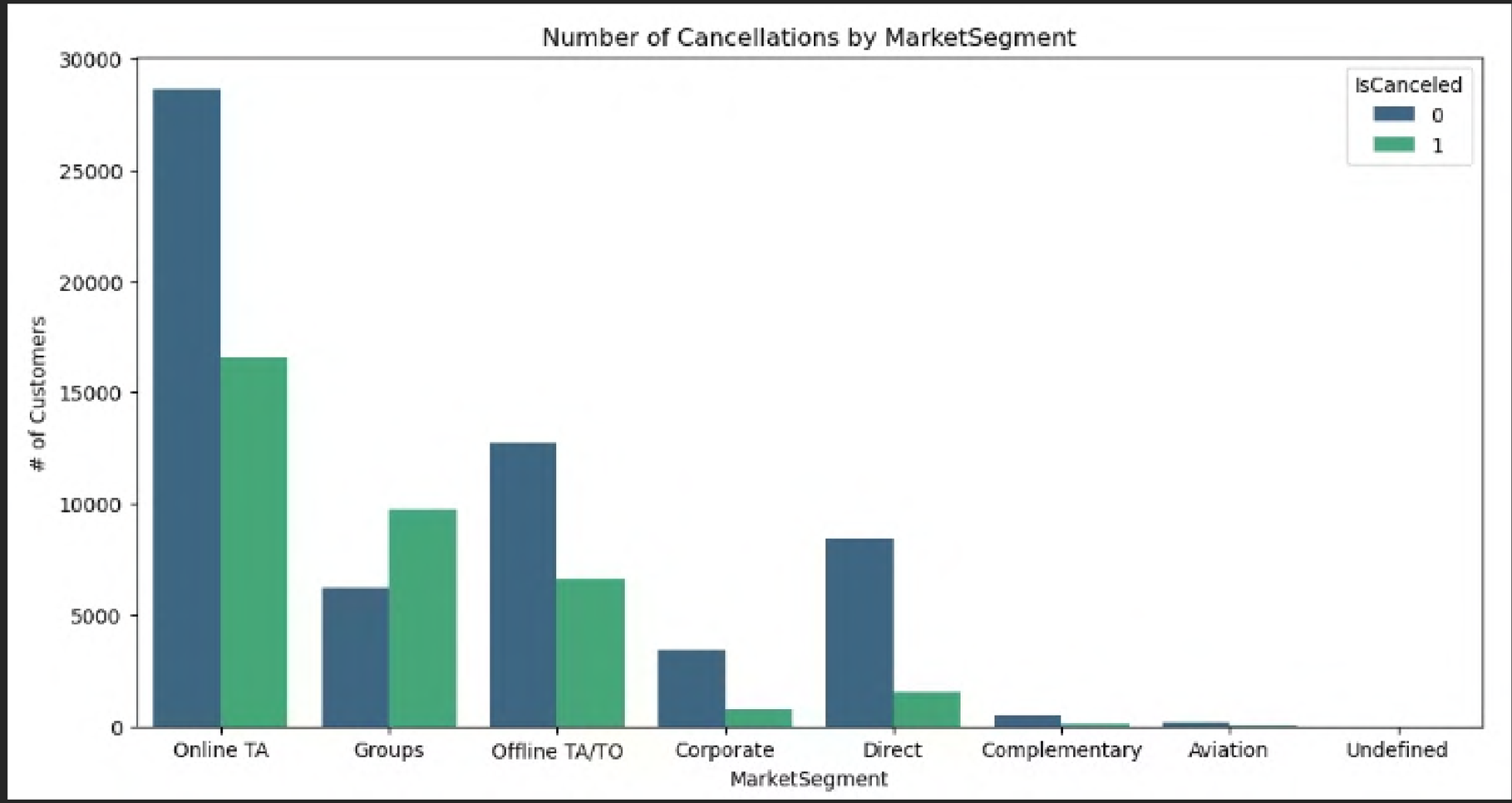
This count plot shows the comparison between different customer types and their cancellation behaviours.



This plot shows the relationship between deposit types (No Deposit, Non Refund, Refundable) and their cancellation behaviours. Ironically, clients who paid a Non-Refundable Deposit, have demonstrated a higher cancellation rate.



This graph shows the cancellation behaviours of the different Market Segments. It reveals which market segments are more prone to cancellations.
“TA” means “Travel Agents”
“TO” means “Tour Operators”



“

Phase 5: Machine Learning

Machine Learning Process.

Importing Libraries

Import libraries like Pandas, NumPy, Scikit-learn, etc

Preprocessing

¬ Train/Test Split
¬ Scaling
¬ Encoding
¬ Pipelining

Model Training

¬ Fit model into training data.
¬ Hyperparameter tuning.
¬ Cross-Validation.
¬ Threshold Adjustment.

Model Evaluation

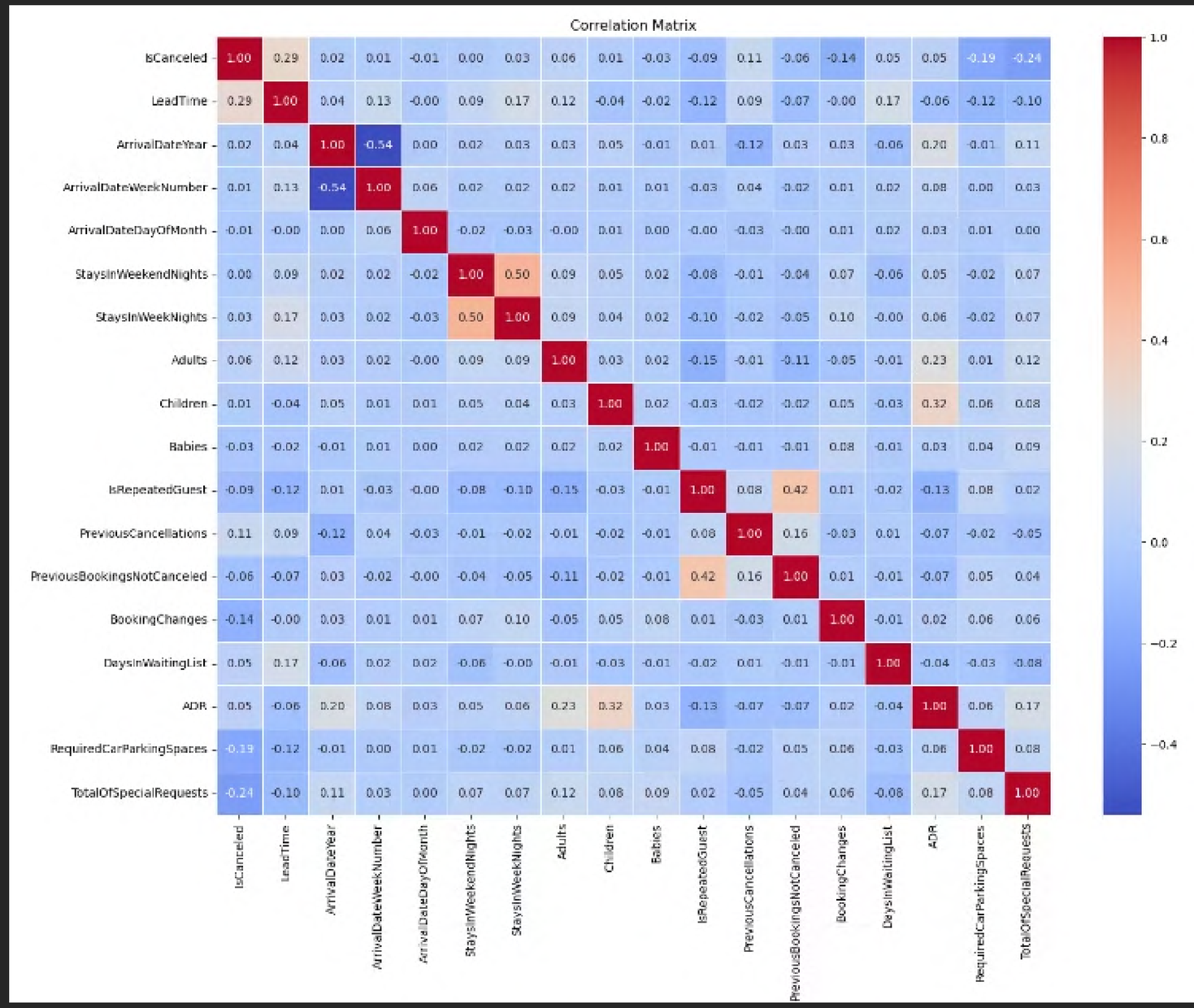
¬ Evaluate trained model on Test set.

“

Since the cost of False Negatives in our Data Set is higher, our goal will be to minimize this as much as possible. As such, we have to keep a close look at the Recall score, which shows us how well we are minimizing our False Negatives.

The correlation matrix heatmap visualizes the relationships between numerical features in the dataset, with color intensity indicating the strength of correlations. Red represents strong positive correlations, while blue signifies strong negative correlations. This plot helps identify which features have significant relationships with each other and the target variable 'IsCanceled'.

Our top five, most correlated features, based on our Correlation Matrix are: LeadTime, PreviousCancellations, BookingChanges, RequiredCarParkingSpaces, TotalOfSpecialRequests.

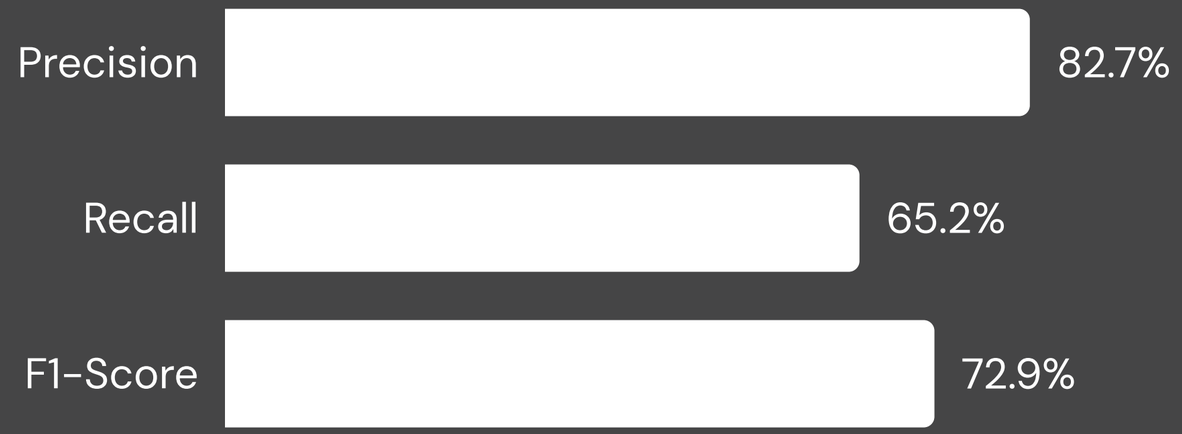


Logistic Regression.

Effective for binary outcomes like booking cancellations, offering interpretable results and identifying key factors influencing the likelihood of cancellations.

Hyperparameters:

¬ Classifiers: Used to control the inverse of regularization strength, prevent overfitting, and enhance the model's ability to predict cancellations accurately.

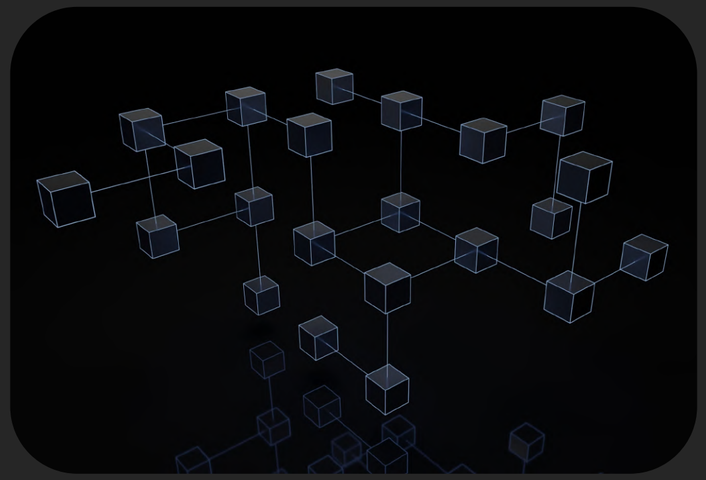


Techniques:

¬ GridSearchCV: Used to find the optimal hyperparameters, improving the model’s performance in predicting cancellations.

¬ OneHotEncoder: Converts categorical variables into a numerical format.

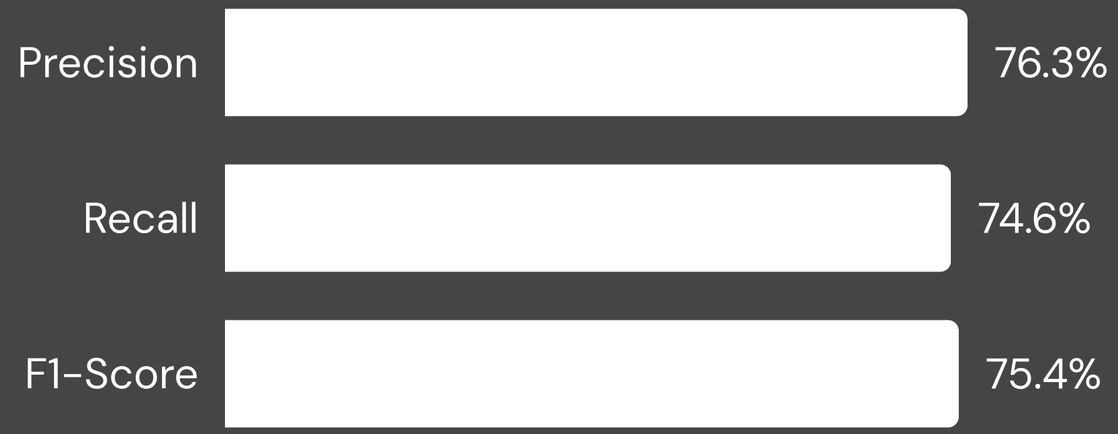
K-Nearest Neighbours.



Captures non-linear relationships, improving accuracy for cancellation predictions by comparing each booking to similar past customer behaviors.

Hyperparameters:

¬ Classifiers: Specifies the number of neighbors to consider when making predictions thereby improving accuracy for cancellation predictions.

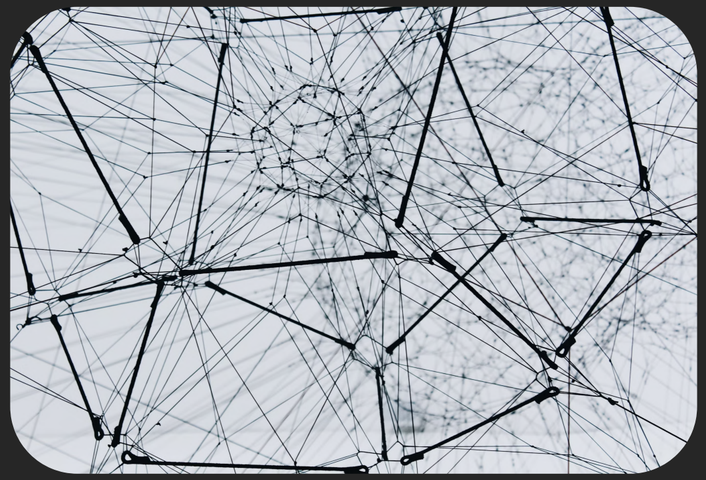


Techniques:

¬ GridSearchCV: Helps in selecting the best number of neighbors.

¬ OneHotEncoder: Used to encode categorical features.

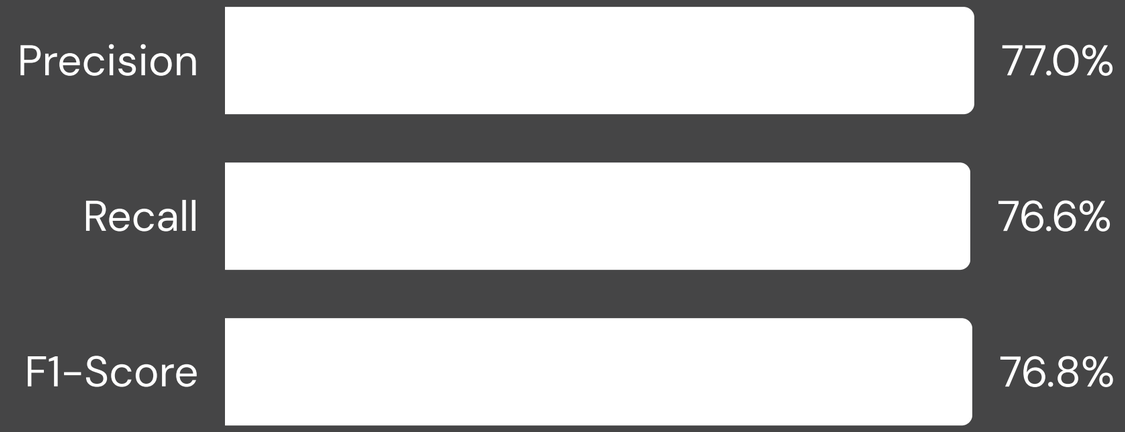
Decision Trees.



Provides clear, interpretable rules and handles non-linear relationships, helping to understand key factors leading to booking cancellations.

Hyperparameters:

¬ Classifiers: Used to prevent overfitting, prevent tree from becoming too complex, and enhancing prediction accuracy.



Techniques:

¬ GridSearchCV: Used to find the optimal hyperparameters.

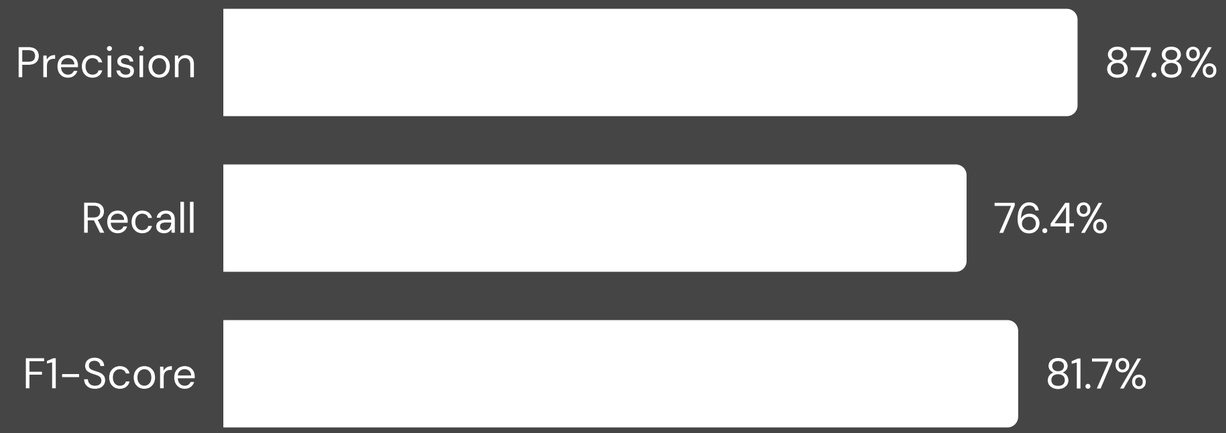
¬ OrdinalEncoder: Converts categorical features to ordinal values.

Random Forests.

Handles complex datasets, improving prediction accuracy and robustness against overfitting by aggregating decisions from multiple decision trees for better cancellation prediction.

Hyperparameters:

¬ Classifiers: Used to optimize prediction accuracy, prevent overfitting and improve prediction robustness.



Techniques:

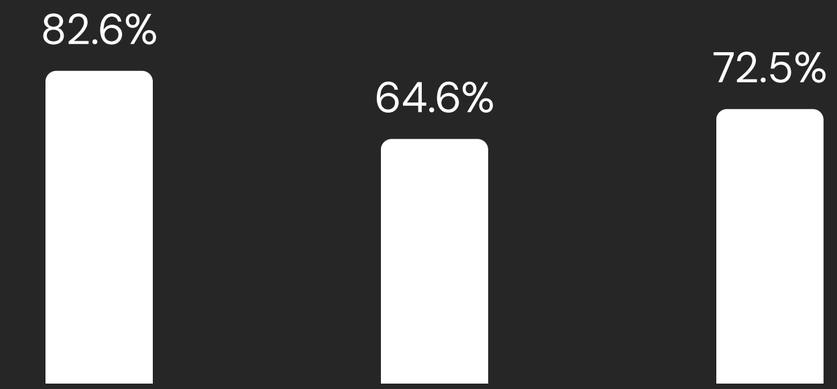
¬ GridSearchCV: Used to optimize the hyperparameters, ensuring the best performance in predicting cancellations.

¬ OrdinalEncoder: Encodes categorical features as numerical features.

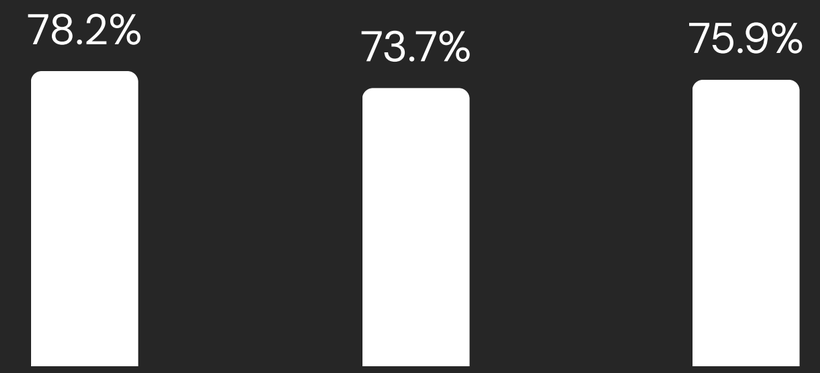
“Evaluating Models on Test Set

Test Scores Summary.

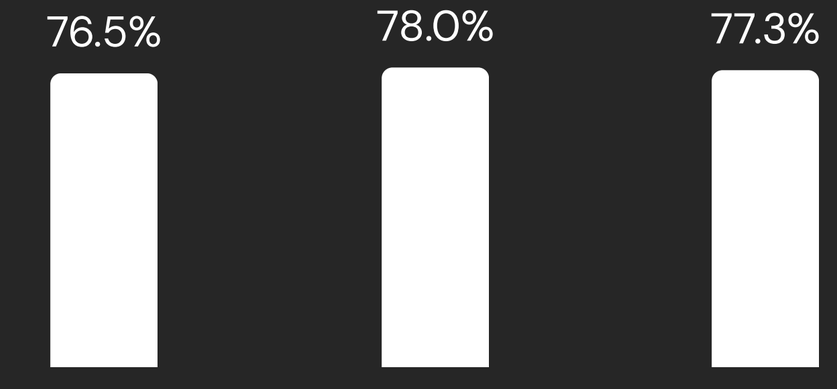
Logistic Regression



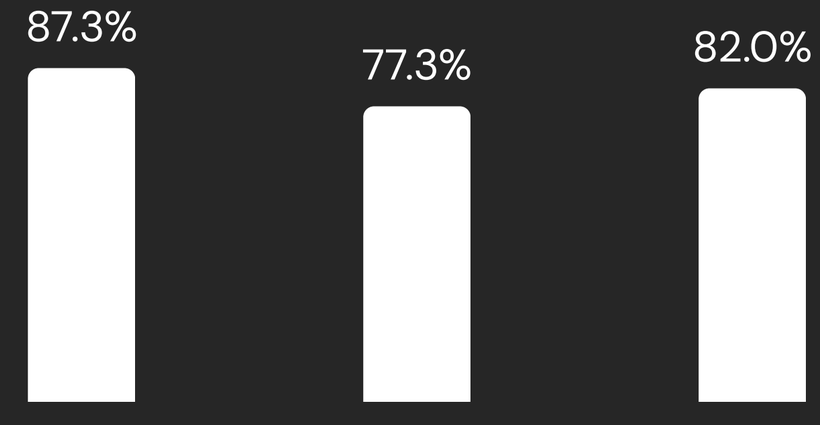
K-Nearest Neighbours



Decision Trees



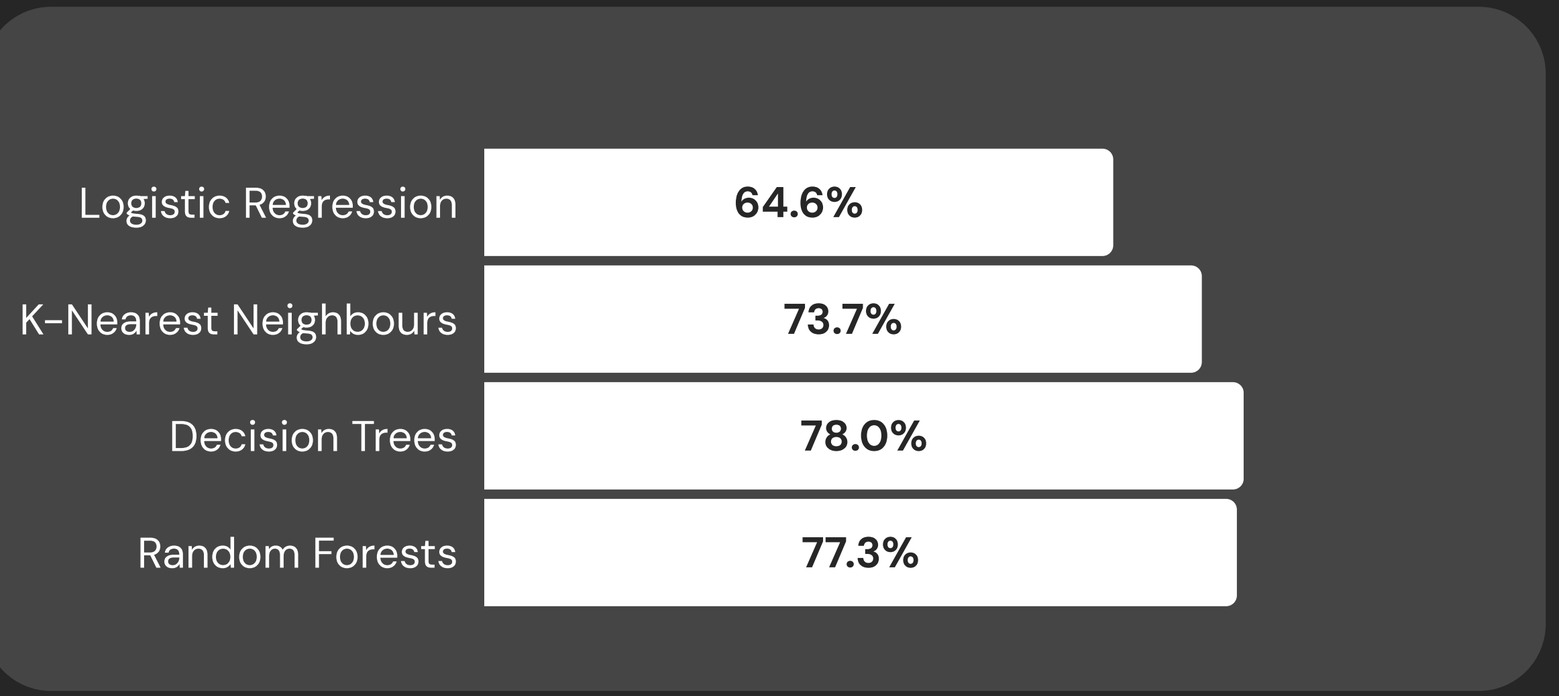
Random Forests

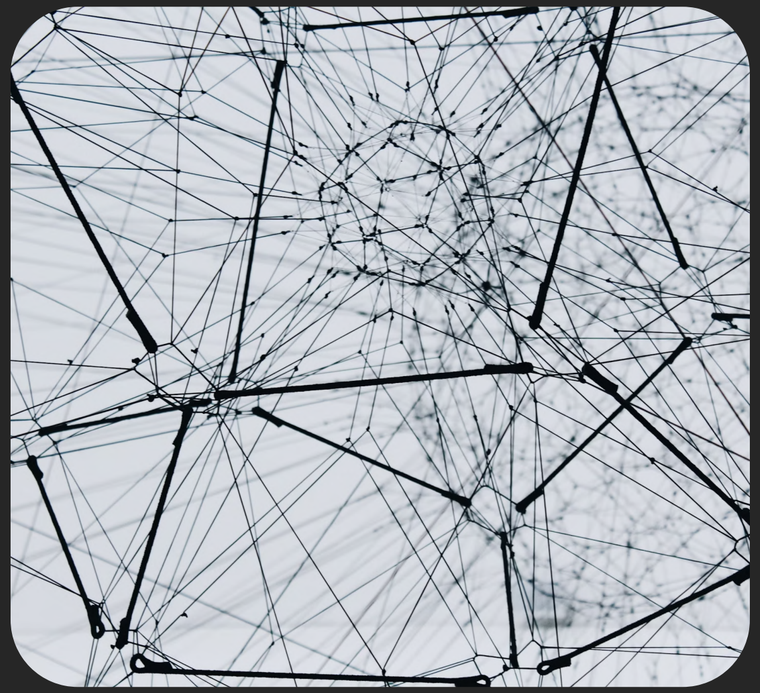


“Award Ceremony”

Best Recall Score:

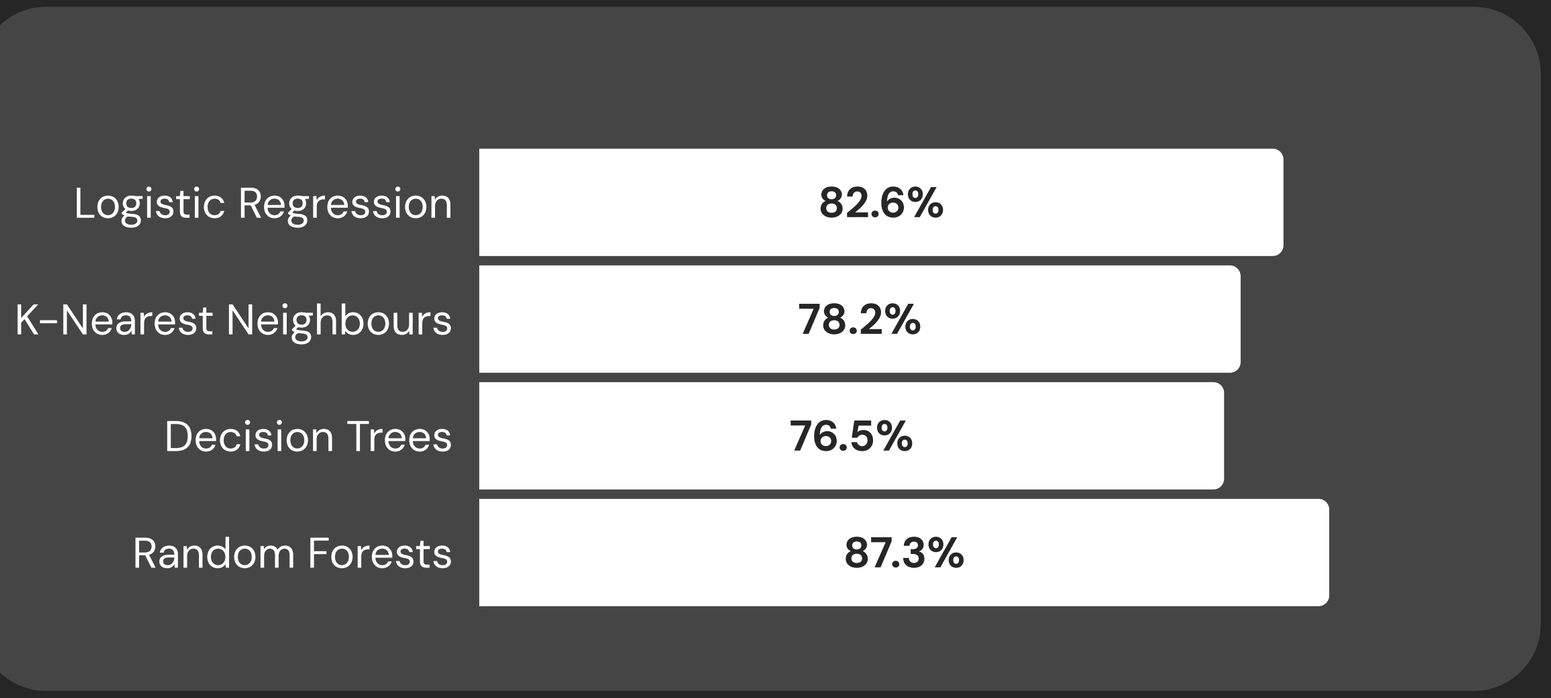
Decision Trees.





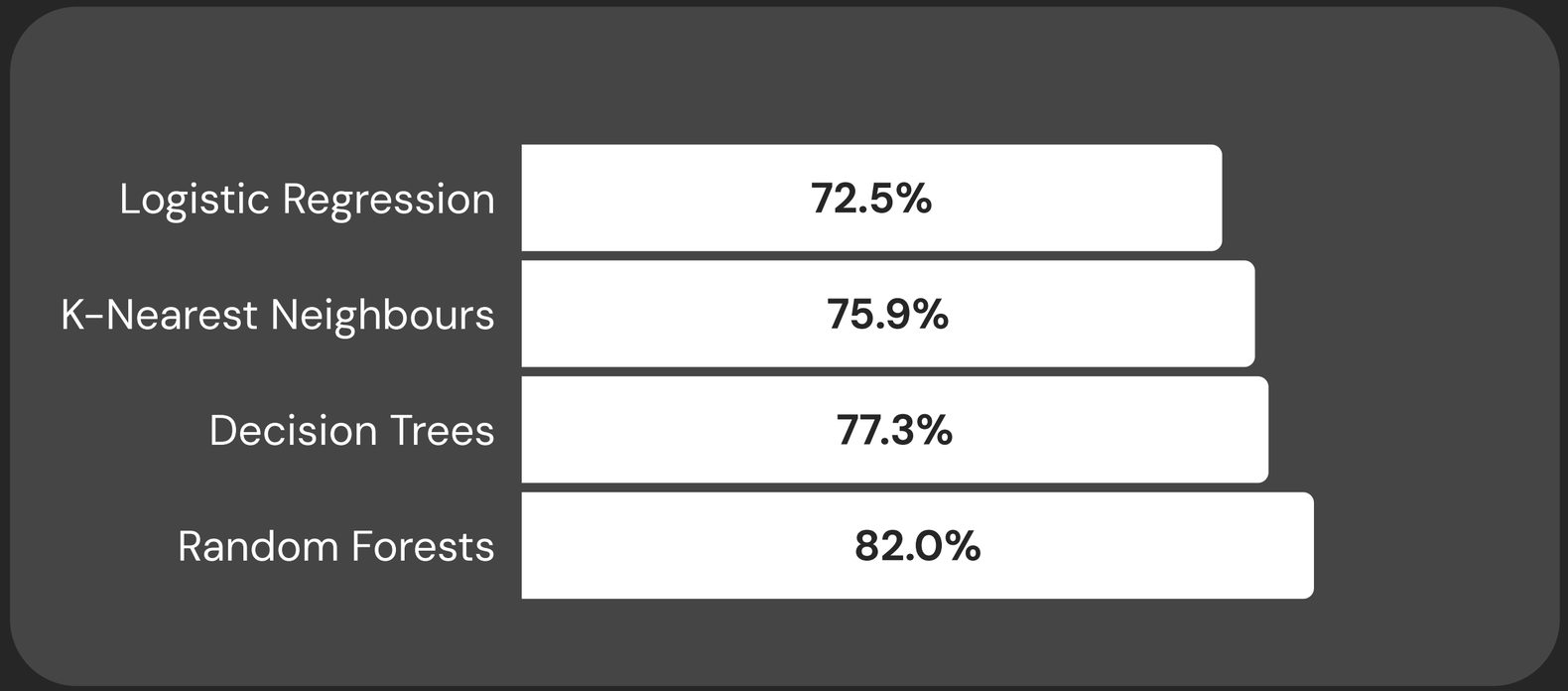
Best Precision Score:

Random Forests.

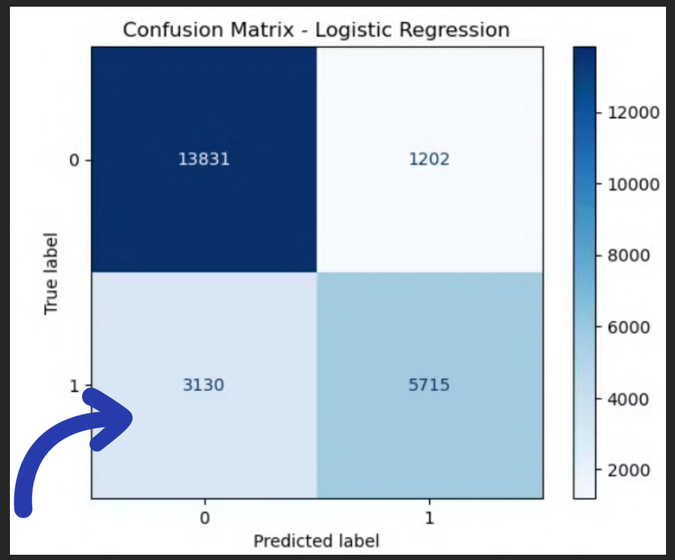


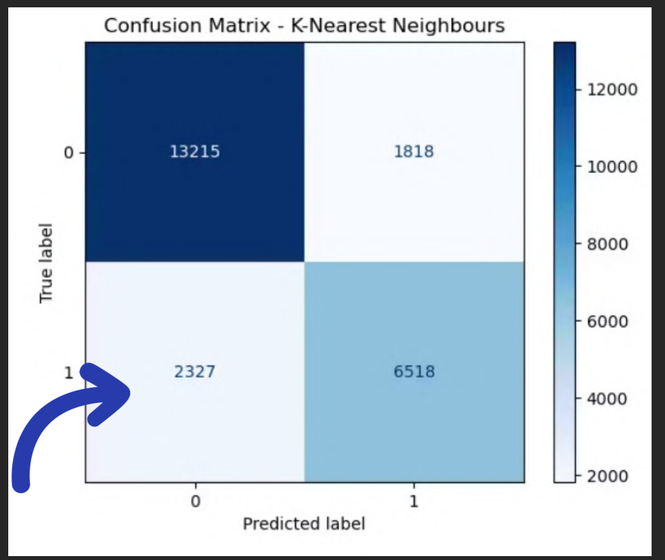
Best F1-Score:

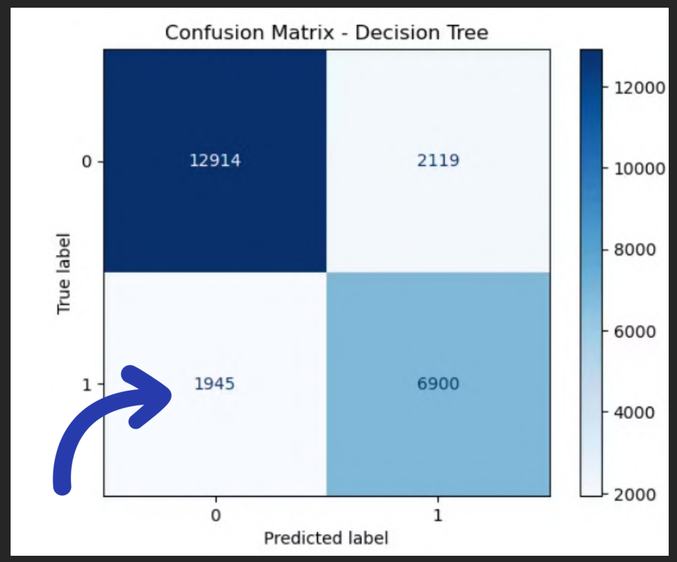
Random Forests.

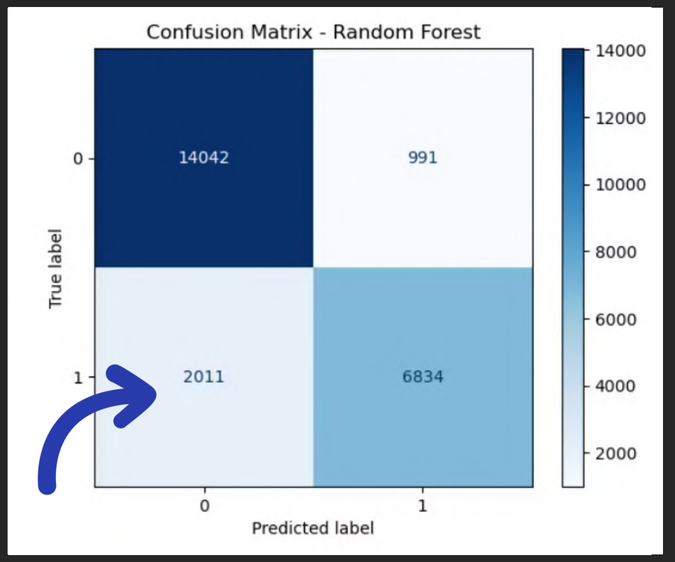


Comparing Confusion Matrices of Various Models on Test Set.

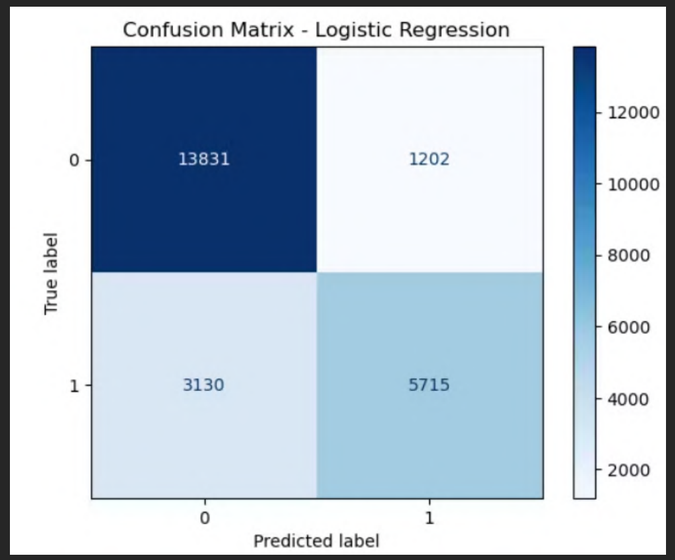


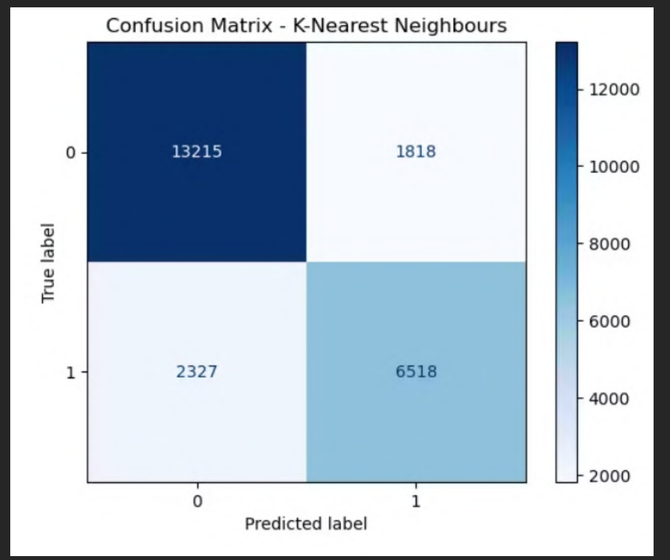


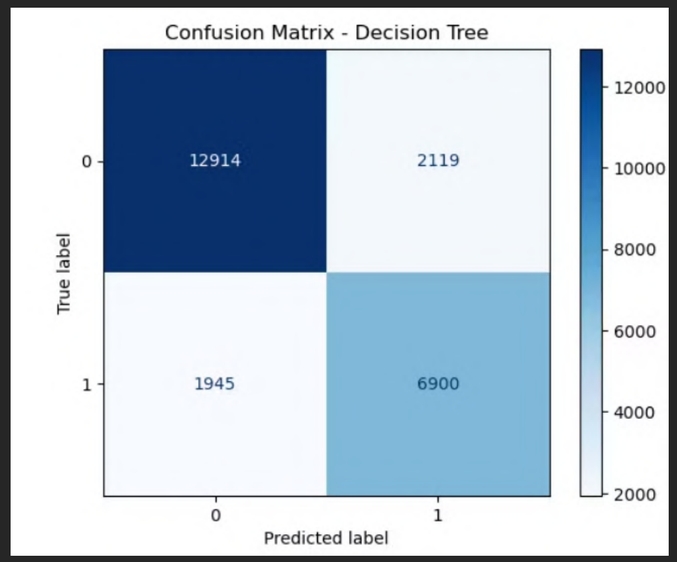


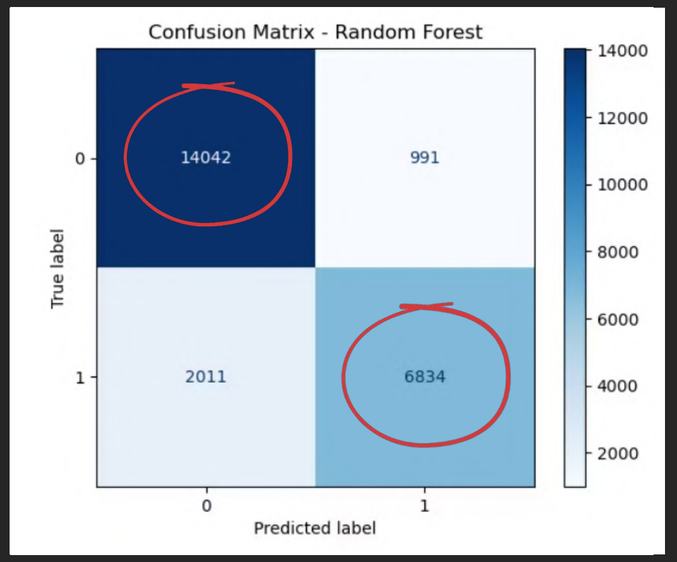


Comparing Confusion Matrices of Various Models on Test Set.









Remarks;

1. Decision Trees achieved the highest recall score of 78% on the test set, indicating its superior ability to identify possible cancellations.

2. Random Forests exhibited the highest F1 - Score of 82%, representing the best balance between precision and recall.

3. Logistic Regression does really well in Precision but very poorly in Recall and is therefore not a suitable model for this case.

4. Overall, while the Decision Tree model offers the best assurance against missed cancellations, Random Forests provide a robust performance across both recall and precision metrics, making it a well-rounded choice for predicting hotel booking cancellations in the real world.

Vielen Dank!!