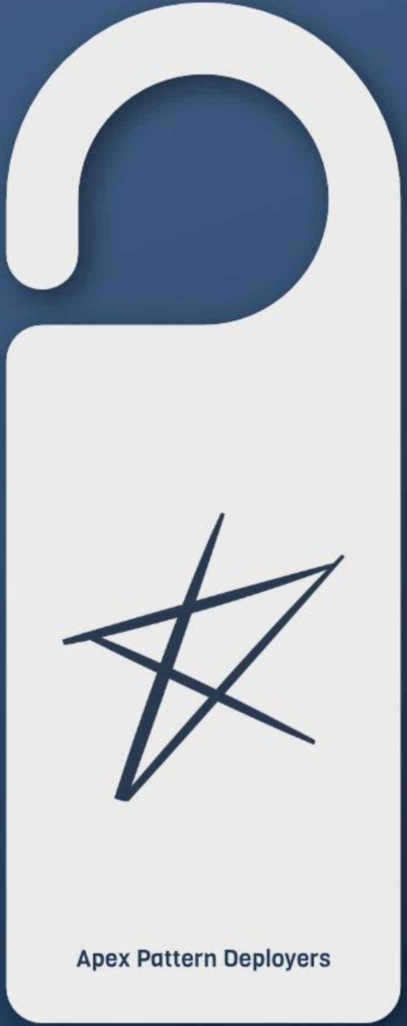
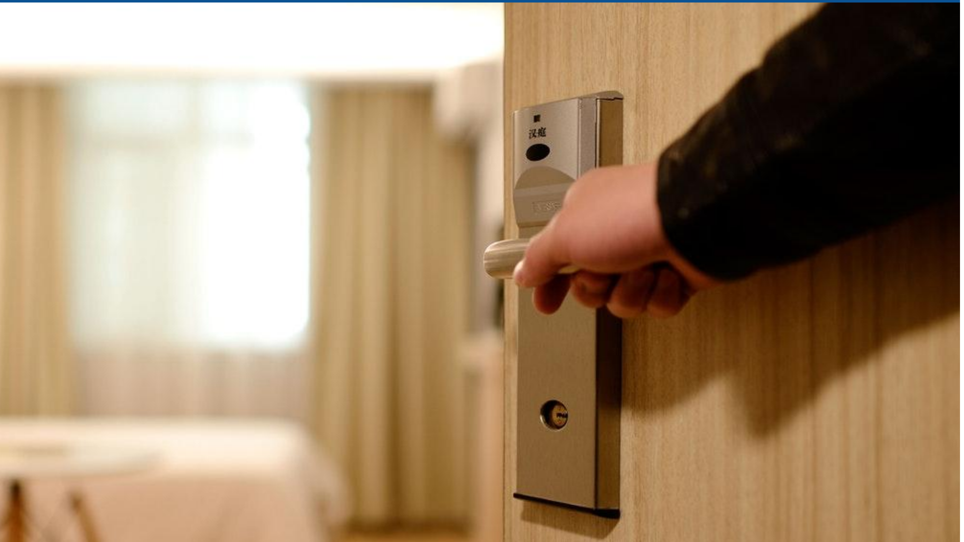


PREDICTING HOTEL BOOKING CANCELLATIONS

Today's Presentation



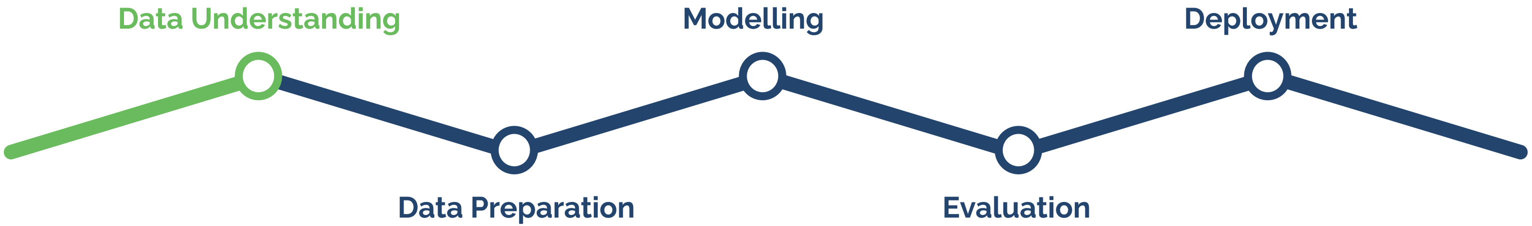
Introduction
Overview about the business needs;

The Process
How we performed our analyses

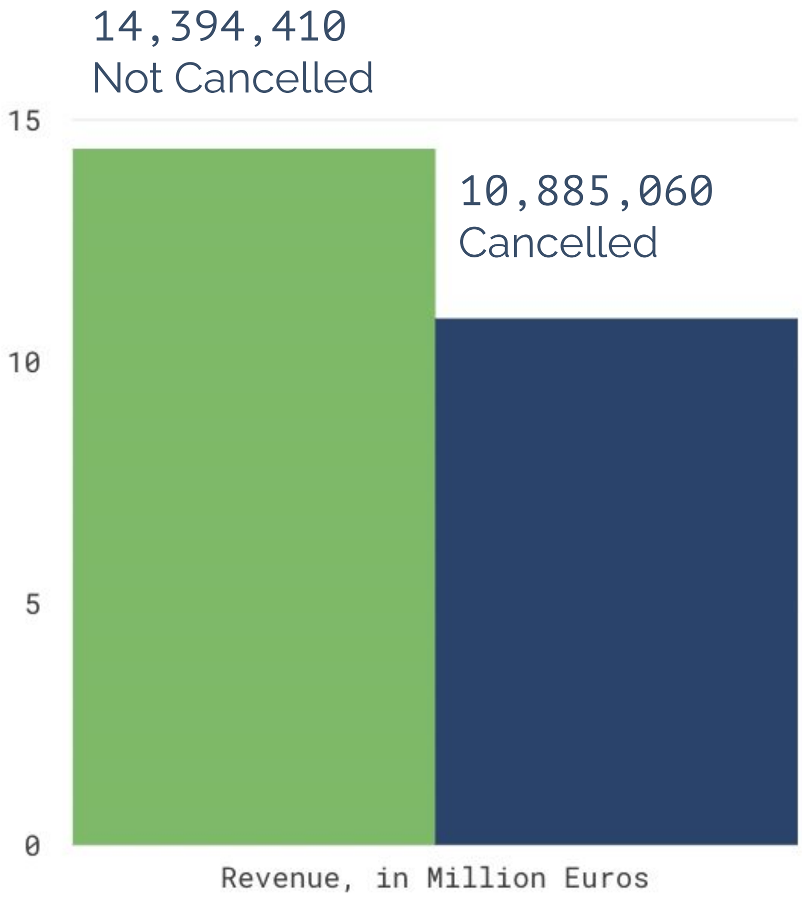
The Models
Which model we chose and why

The Insights
What the results mean for the business

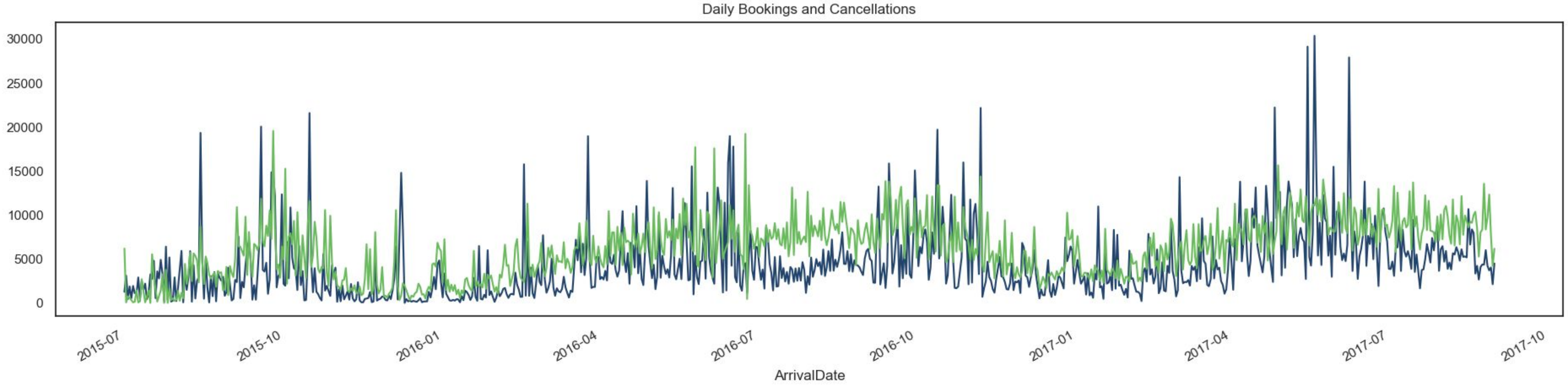
Introduction


Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

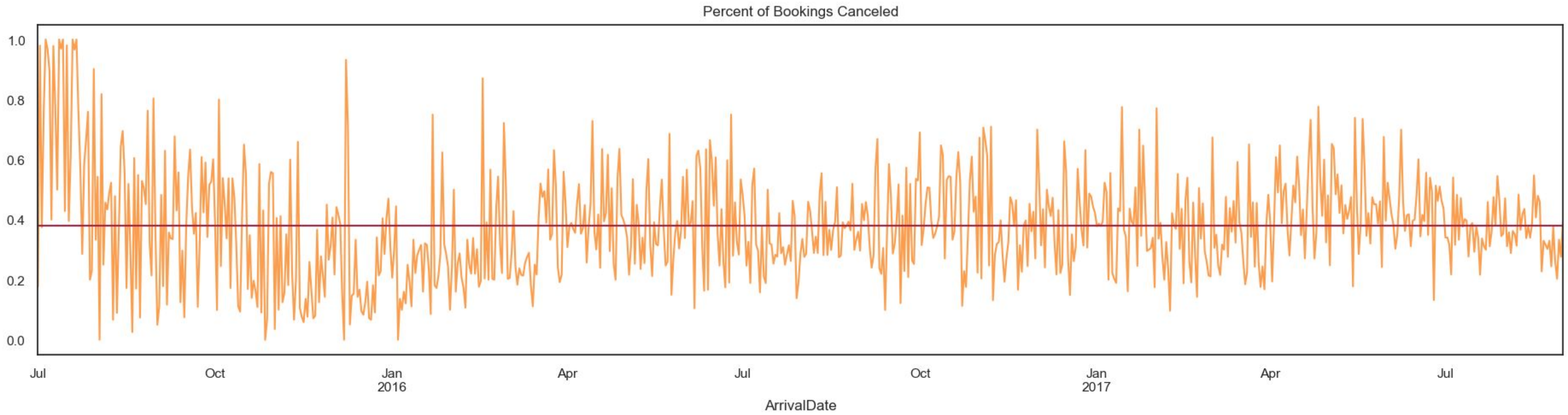
Room Revenue *in* Euros



Daily Bookings *and* Cancellations



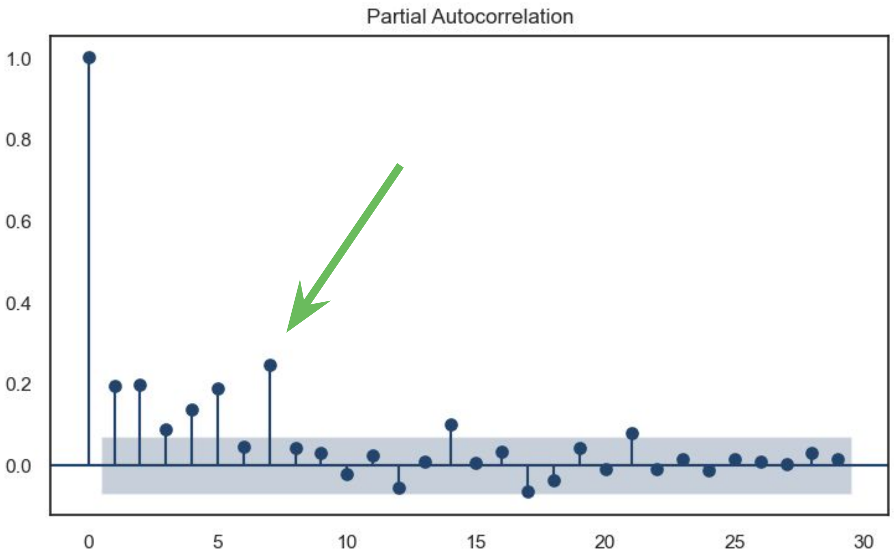
Percent *of* Daily Bookings Cancelled



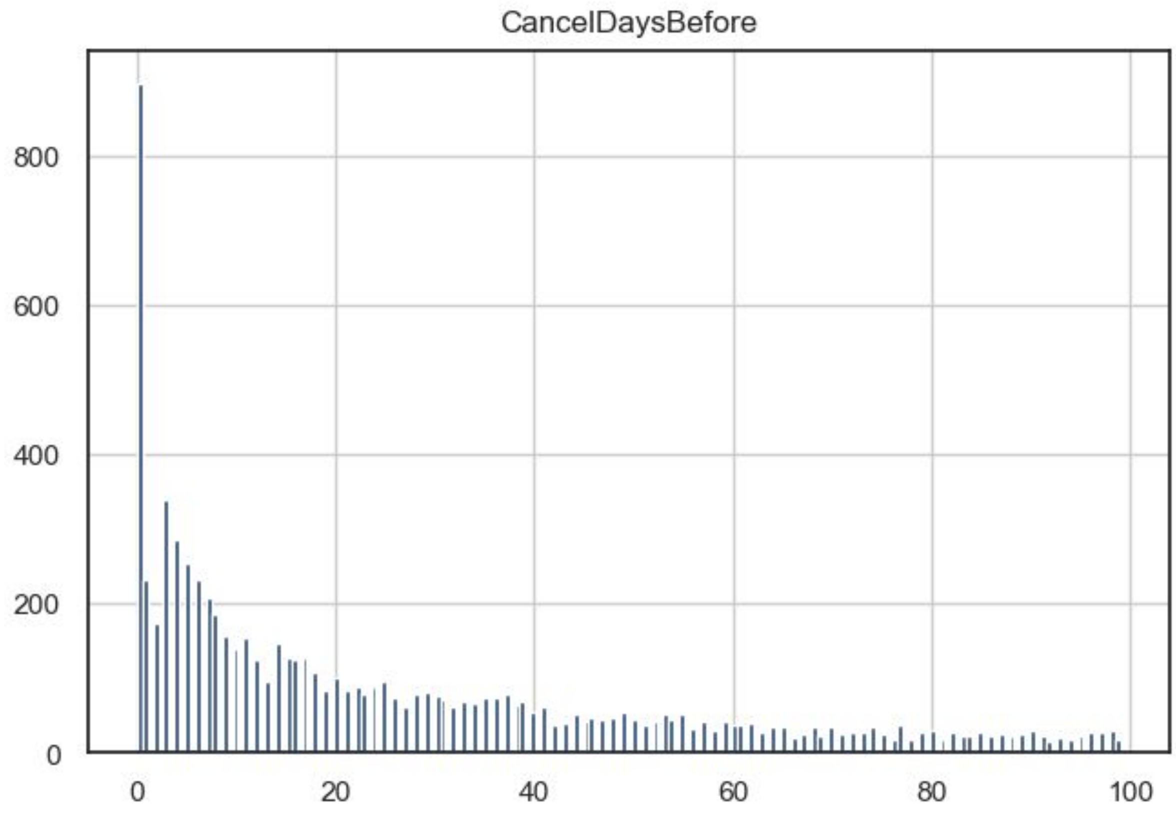
Time Series Analysis

7-day seasonality

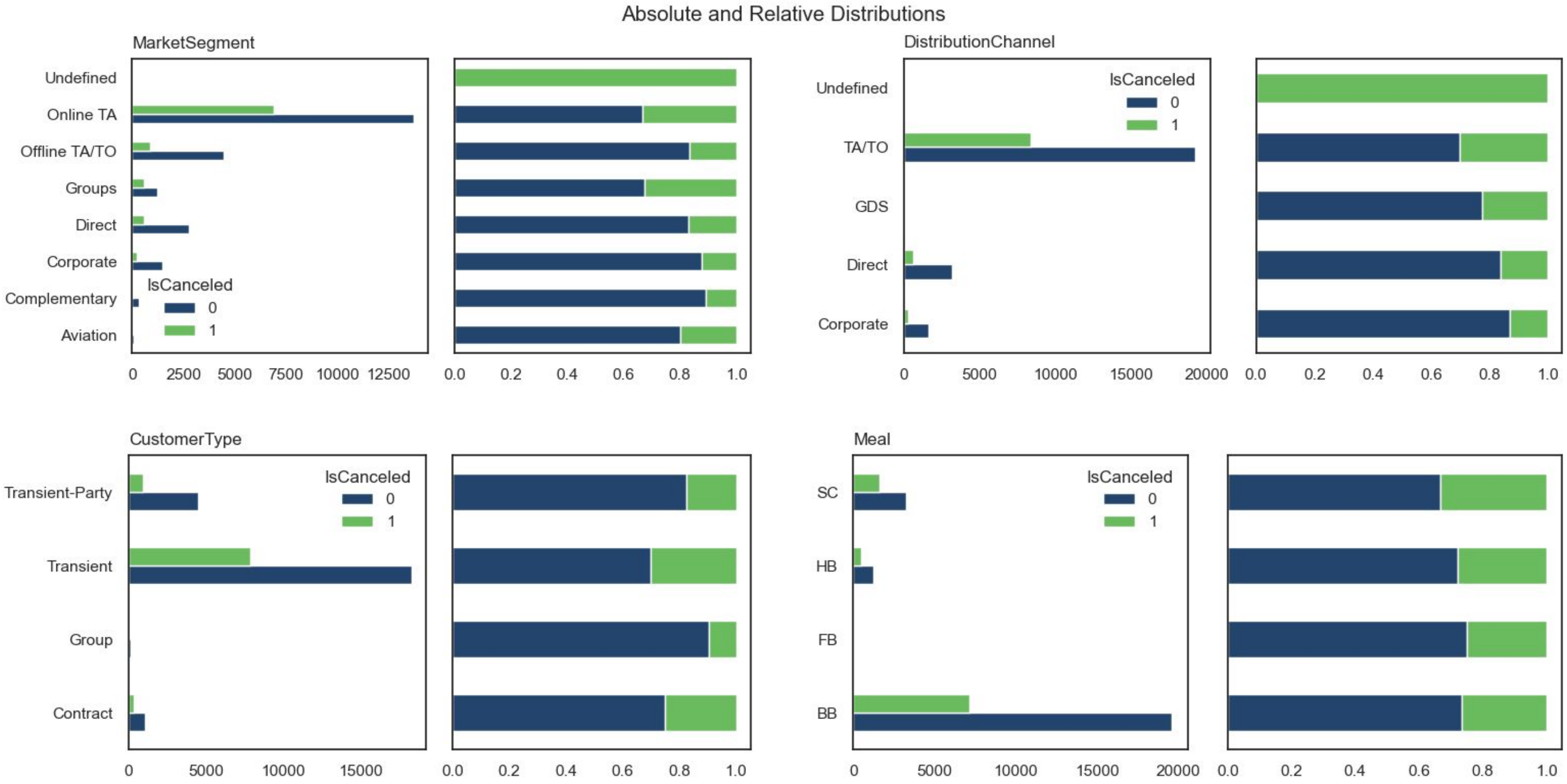
Percent of bookings cancelled



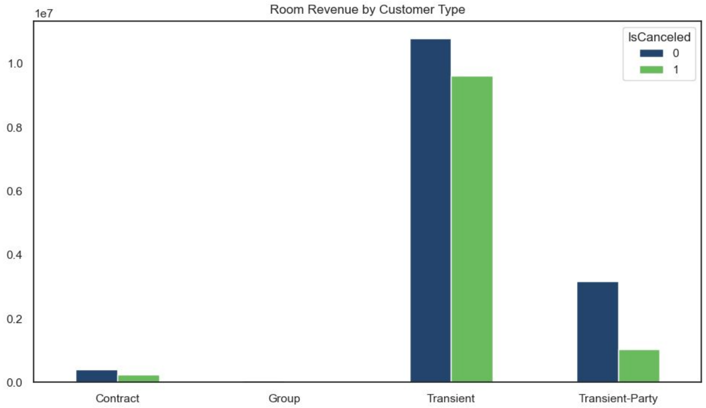
Distribution of Cancellation Days Before Arrival Date

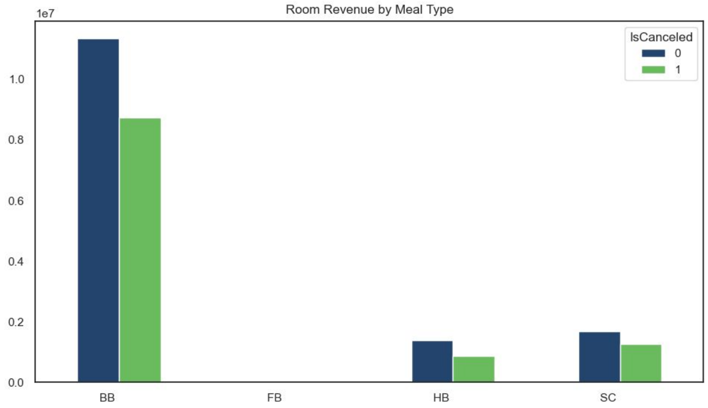


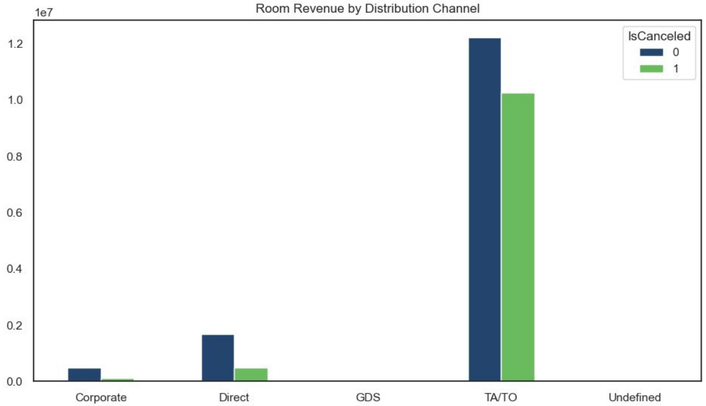
Who are cancelling bookings?

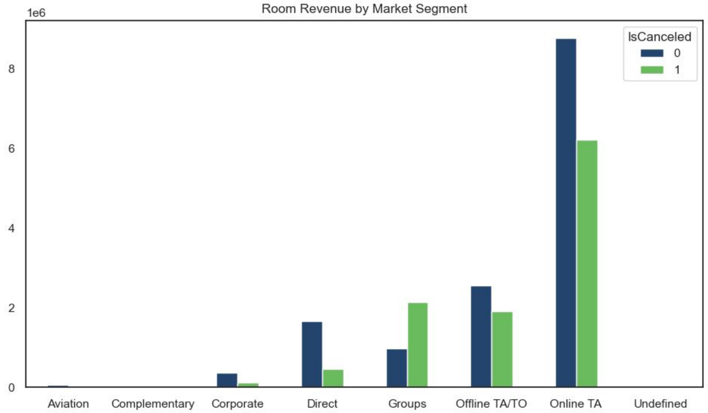


Who are cancelling bookings?

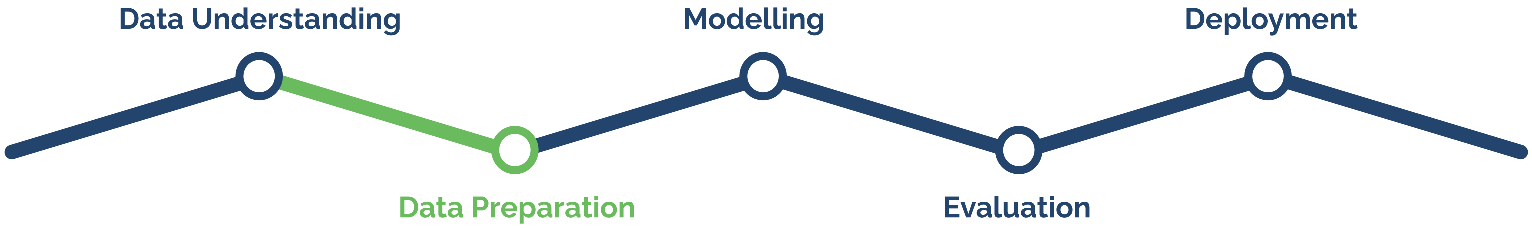




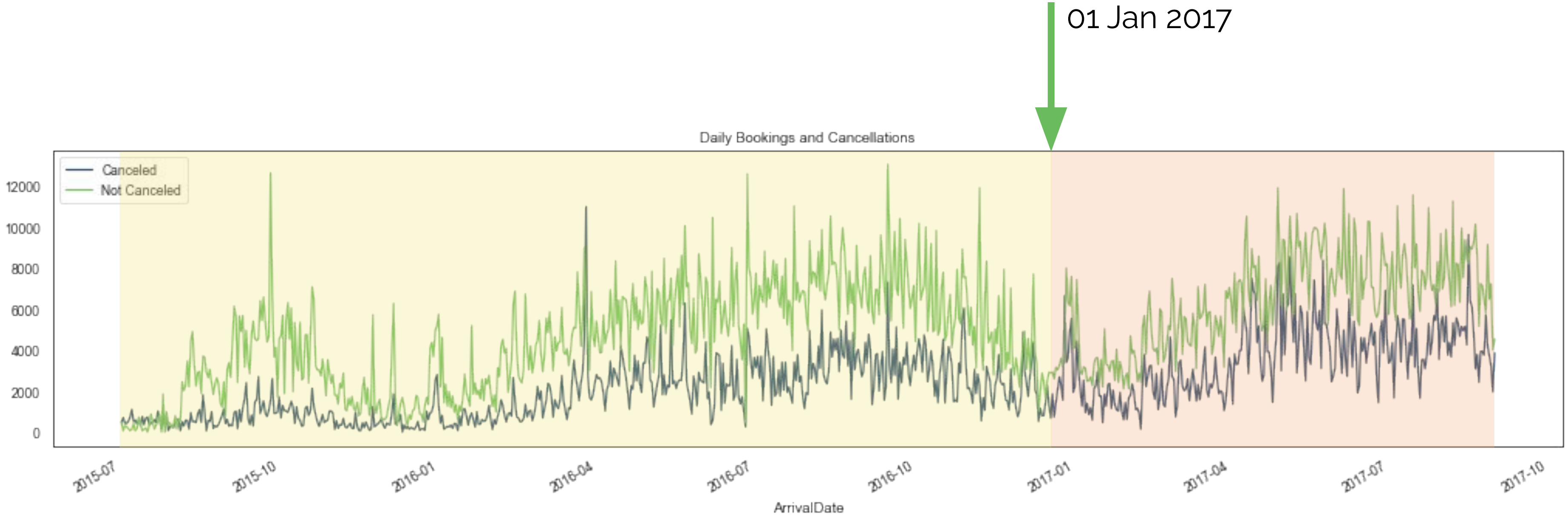




Data Preparation


Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Splitting the Dataset



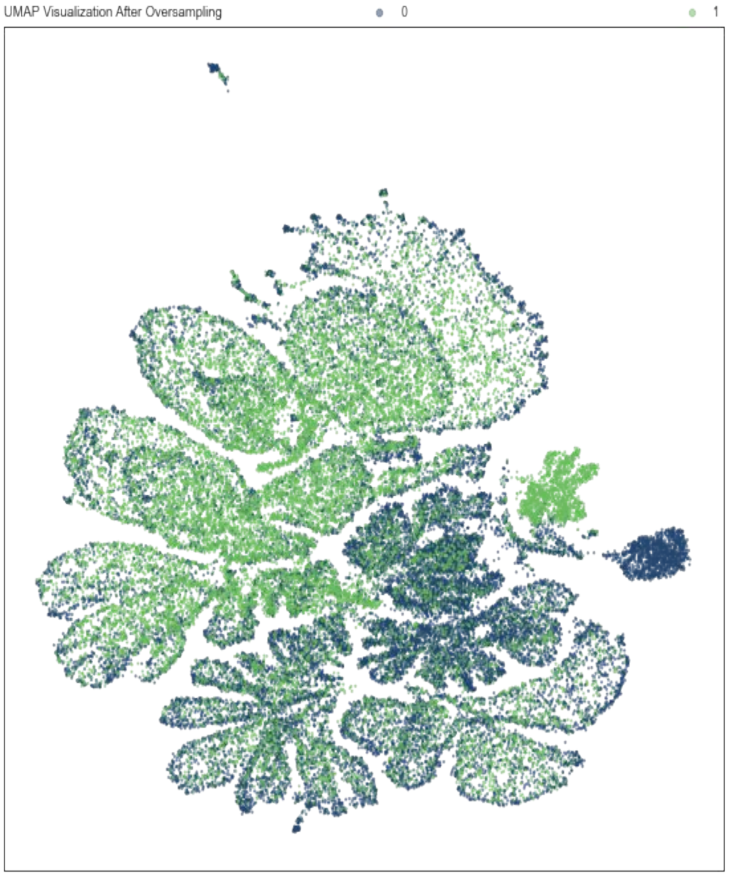
Key Decisions During Data Preparation

Duplicates	Redundant Features	Expendable Features	Imbalanced Data
Duplicates in training data removed to avoid adding weight to identical data points	Features corresponding to outcome class removed as info would not be available until after cancellation	Unnecessary, inaccurate, or likely to change features removed	SMOTENC used to balance the dataset as it works for both numerical and categorical features

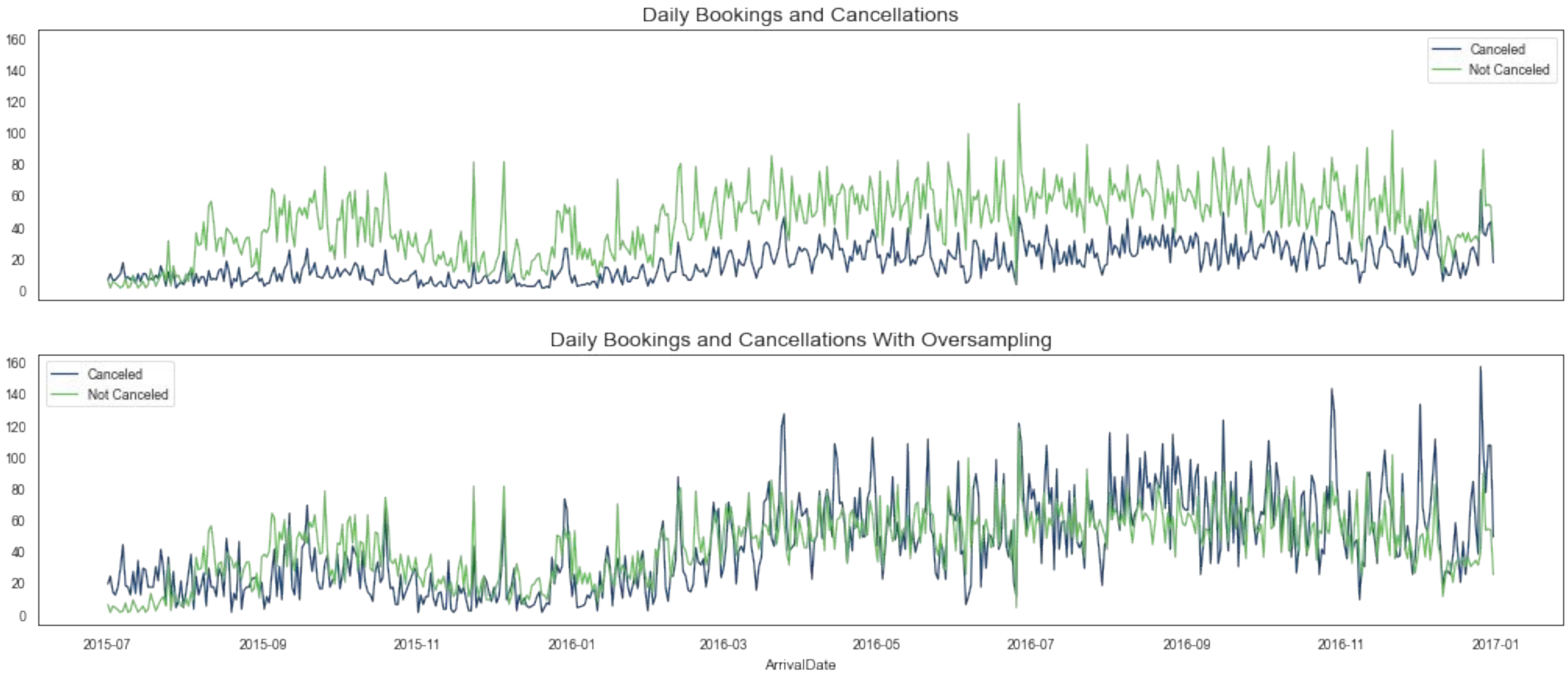
Preparing the Data

01	**Remove Duplicates**	• 18 514 out of 51 822 rows in the training/testing dataset
02	**Remove Null Values**	• 75 641 null values 'Company' column • 4 null values 'Children' • 24 null values 'Country'
03	**Too Many Category Values**	• Agent
04	**Remove Unreliable Data**	• Adults, Children, Babies, Meal, Country and AssignedRoomType
05	**Remove Non-Relevant**	• ReservationStatusDate, ReservationStatus

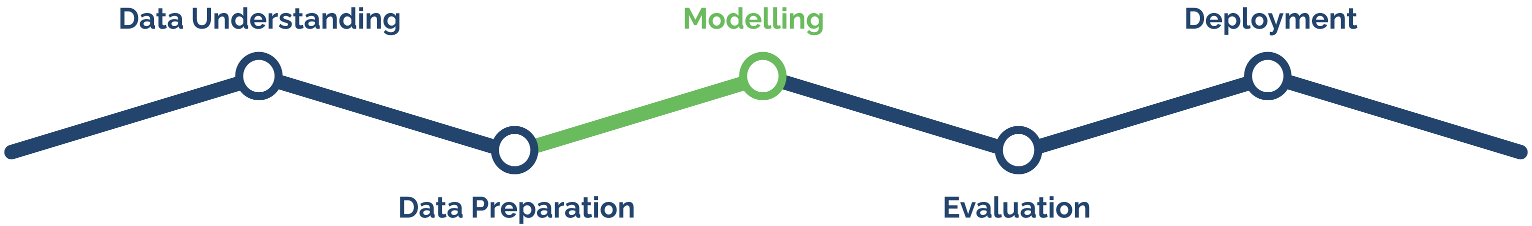
Balance the Dataset



Balance the Dataset



Modeling and Evaluation

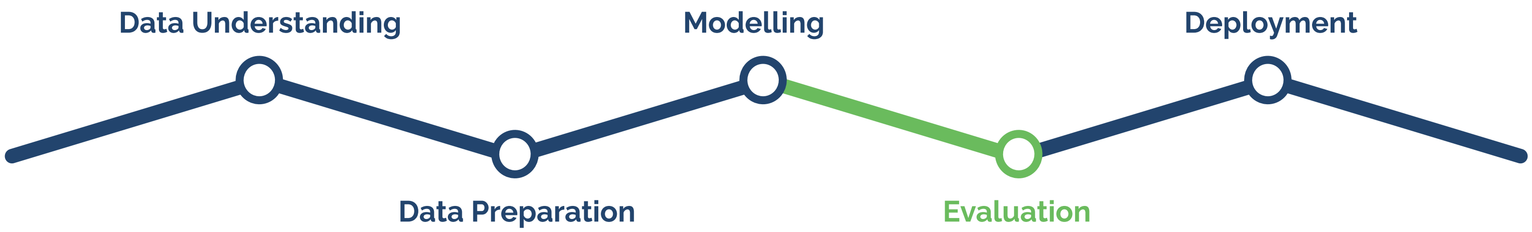

Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Modeling

- Gradient Boosting/ XG Boost
- MLP Classifier
- Gaussian Naive Bayes
- Decision Trees
- Random Forest
- K Neighbors
- Logistic Regression
- Voting Classifier

Classification Algorithms

Provide class membership probabilities, leaf probabilities, or neighborhood voting proportions.


Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Evaluation Measures

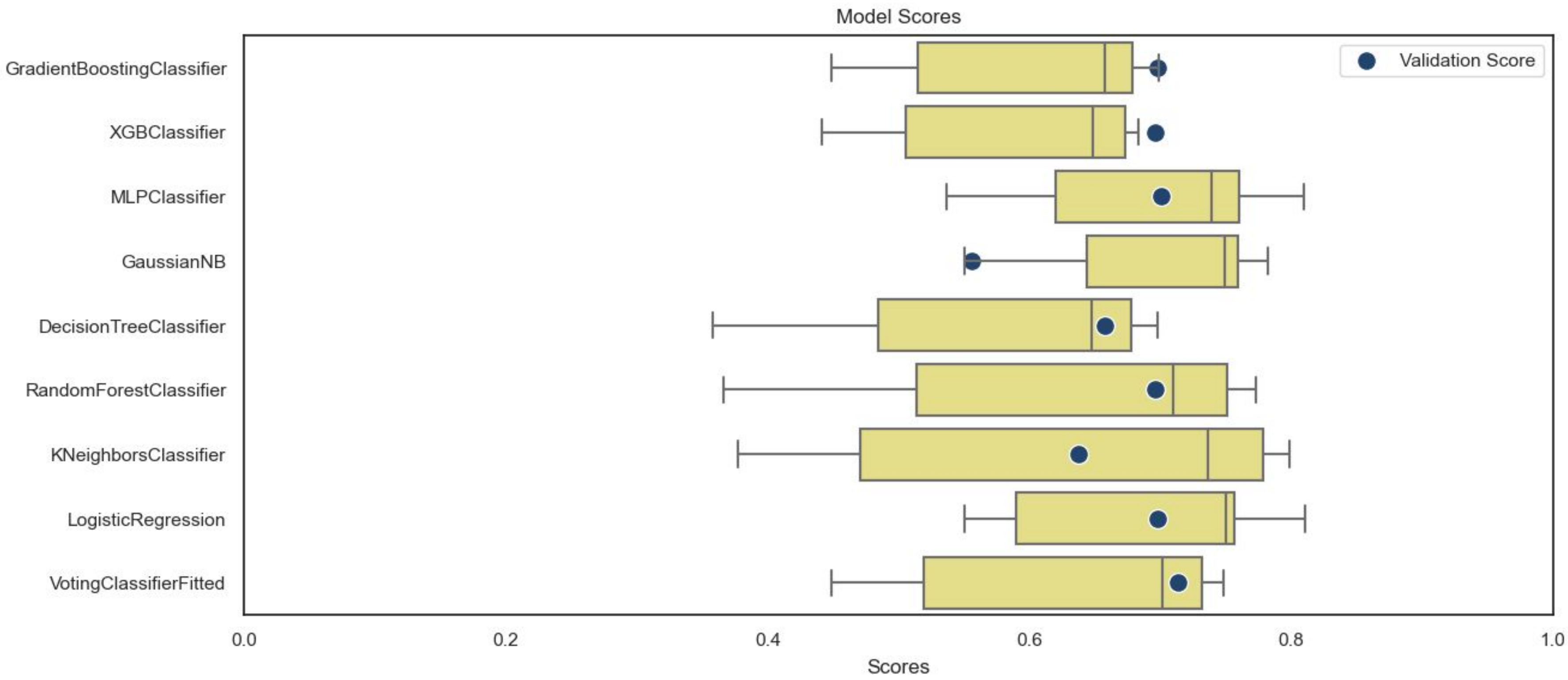
F1 = harmonic mean of precision and recall

$$\frac{\text{True Positives}}{\text{True Positives} + \frac{1}{2}(\text{False Positives} + \text{False Negatives})}$$

Confusion Matrix

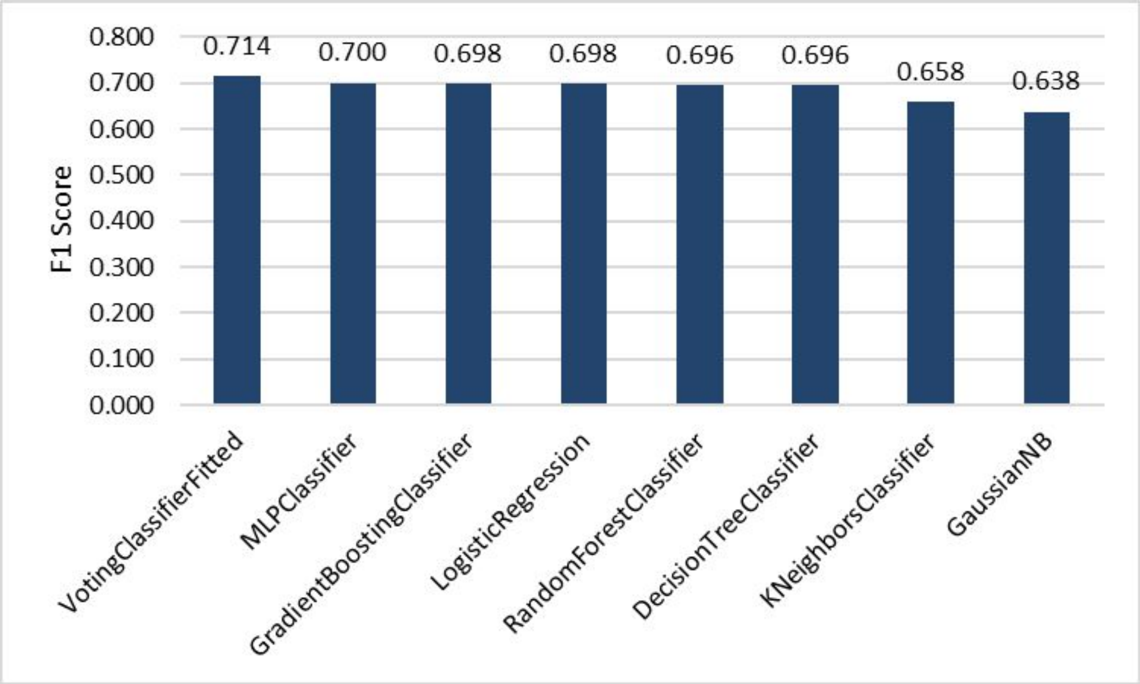
TRUE CLASS \ PREDICTED CLASS		
	True Positive (predicted cancel, canceled)	False Negative (predicted arrival, canceled
	False Positive (predicted cancel, arrived	True Negative (predicted arrival, arrived)

Model Performance

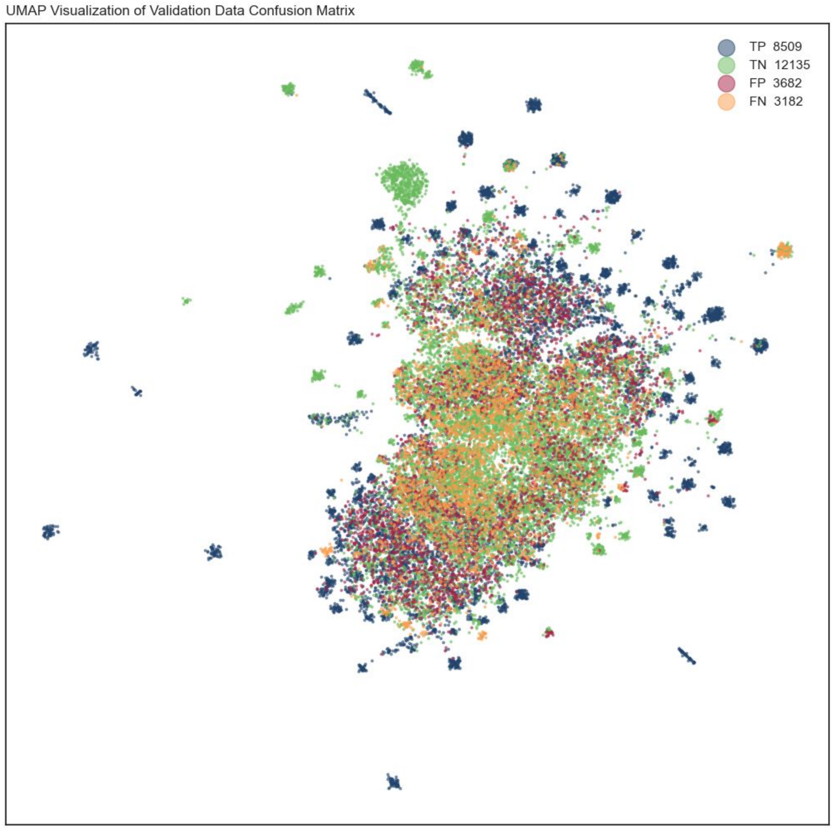


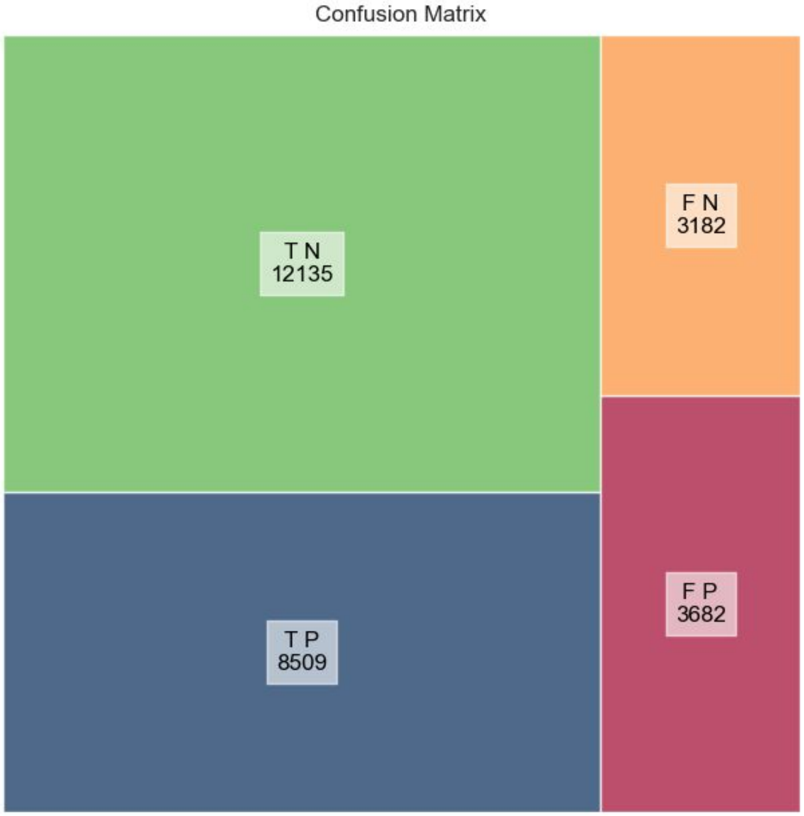
Model Results

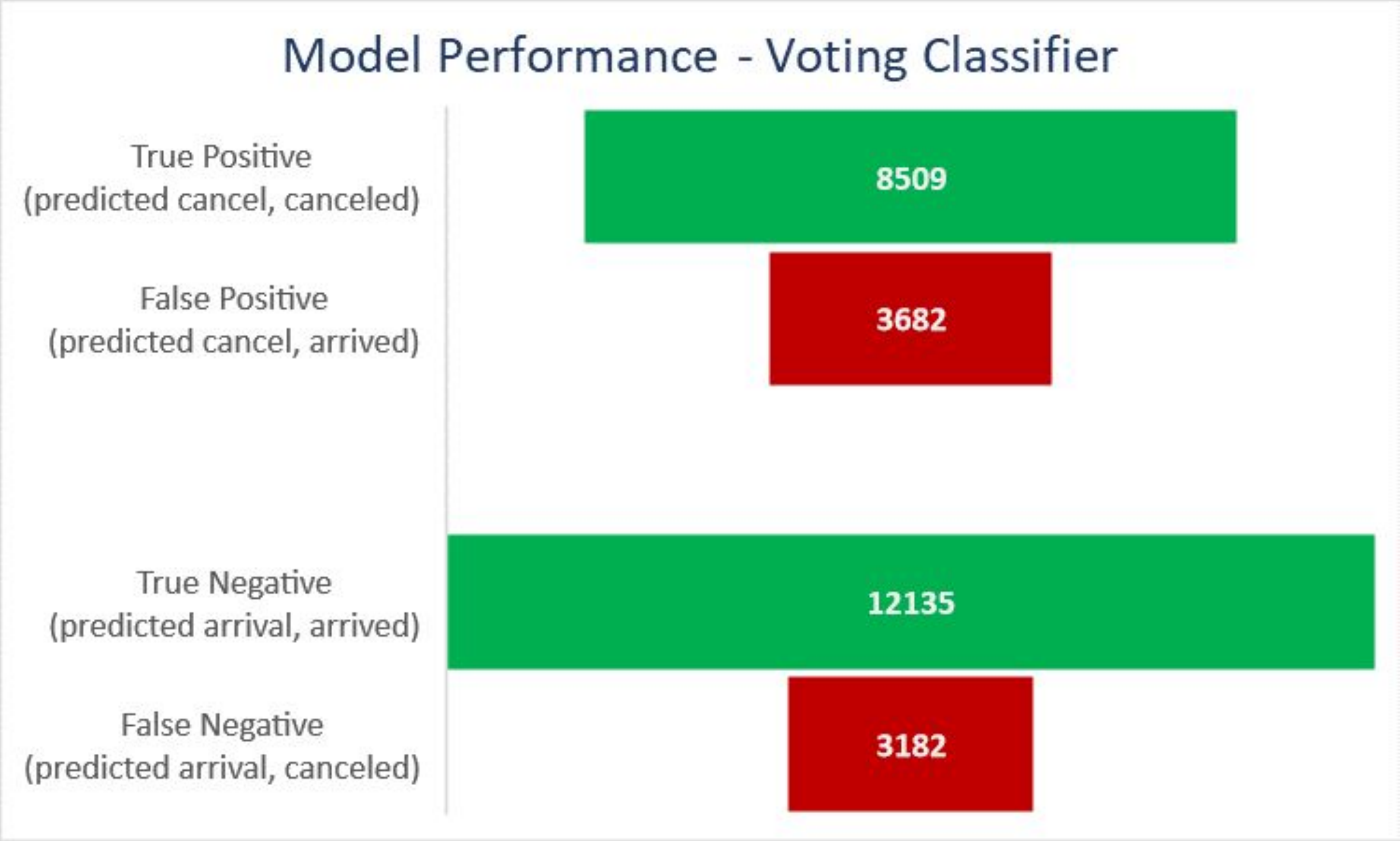
	Scores
VotingClassifierFitted	0.7138
MLPClassifier	0.7004
GradientBoostingClassifier	0.6983
LogisticRegression	0.6979
RandomForestClassifier	0.6963
XGBClassifier	0.6960
DecisionTreeClassifier	0.6578
KNeighborsClassifier	0.6377
GaussianNB	0.5556



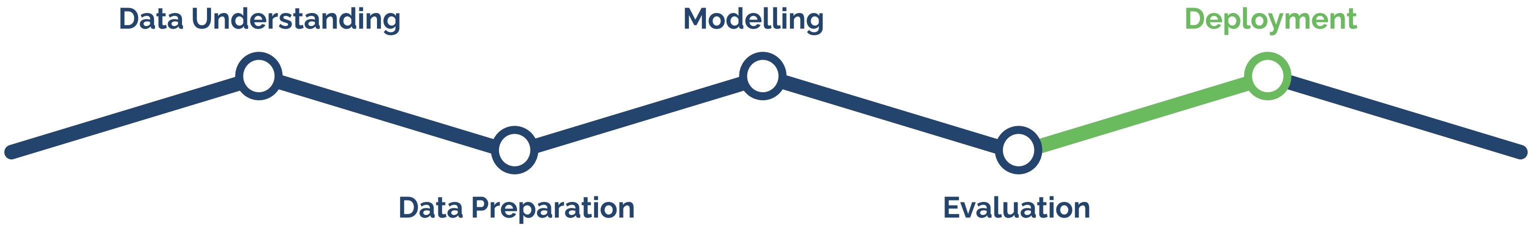
Confusion Matrix for VotingClassifier



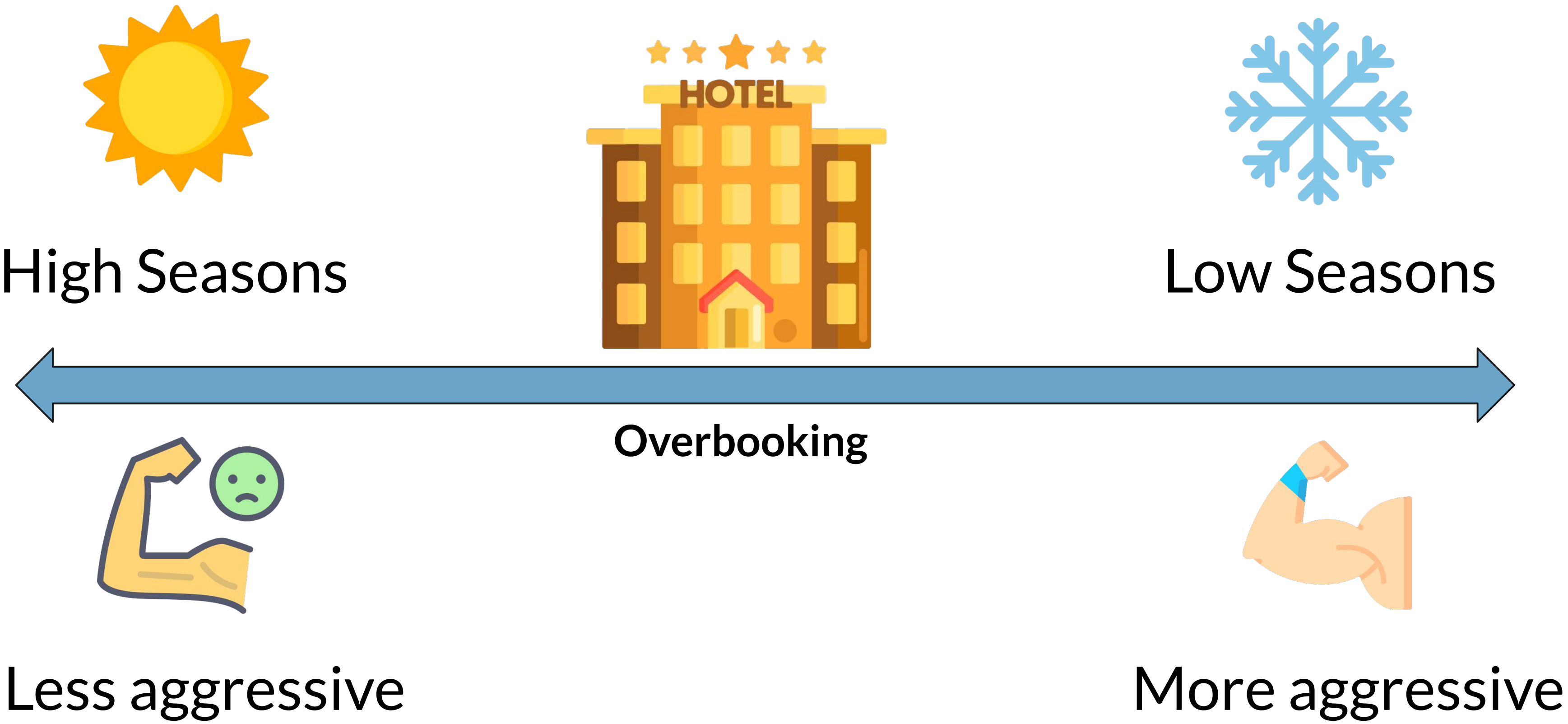



Model Performance - Voting Classifier
True Positive
(predicted cancel, canceled)
8509
False Positive
(predicted cancel, arrived)
3682
True Negative
(predicted arrival, arrived)
12135
False Negative
(predicted arrival, canceled)
3182

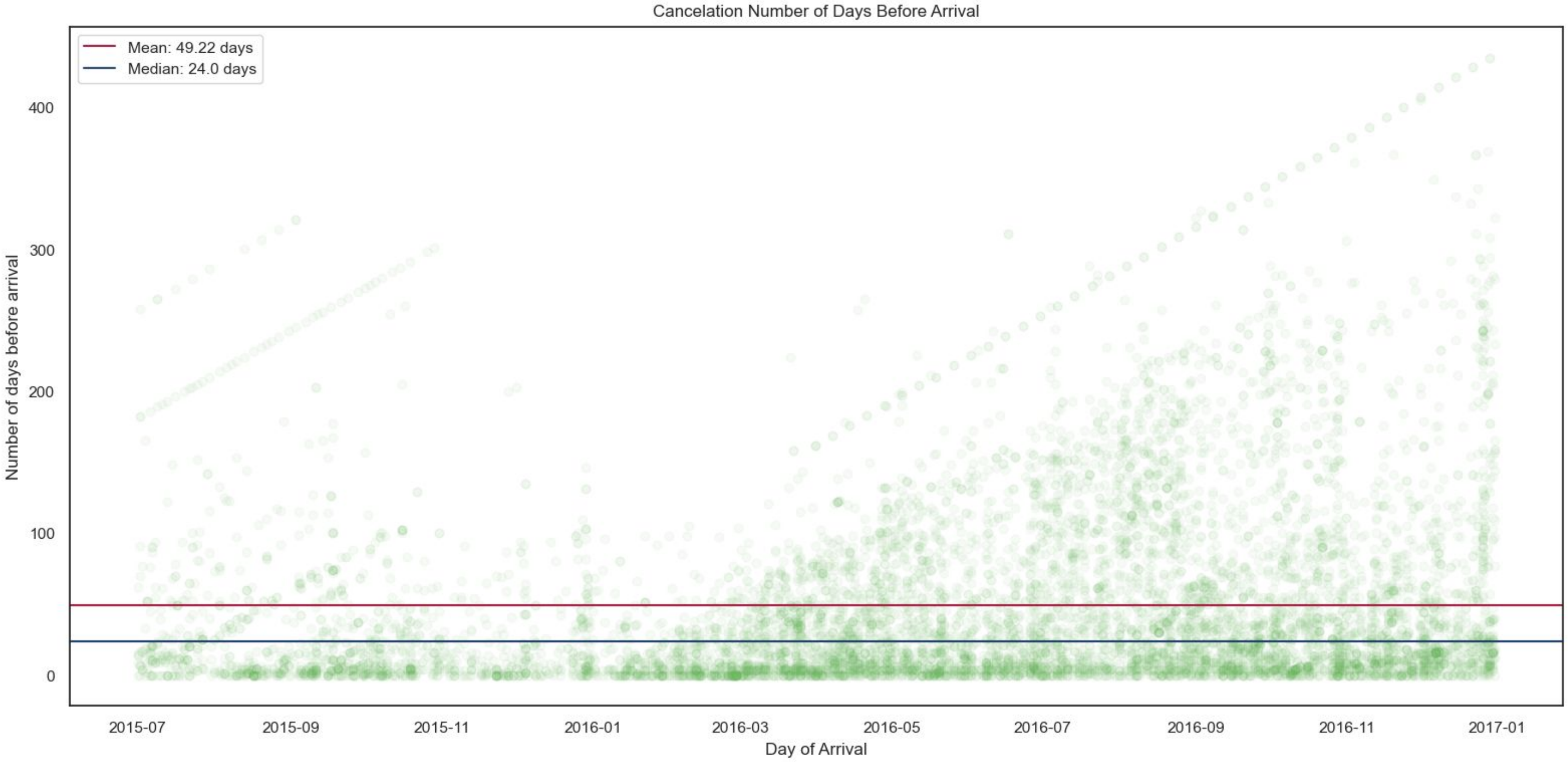
Deployment


Data Understanding
Data Preparation
Modelling
Evaluation
Deployment

Overall Strategy



When are bookings cancelled?



Insights

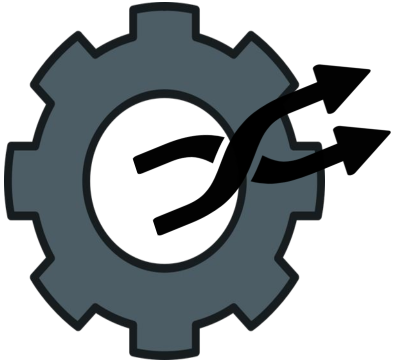
- Cancellation prevention

TRADEOFF

Customer incentives?

Which ones to contact?

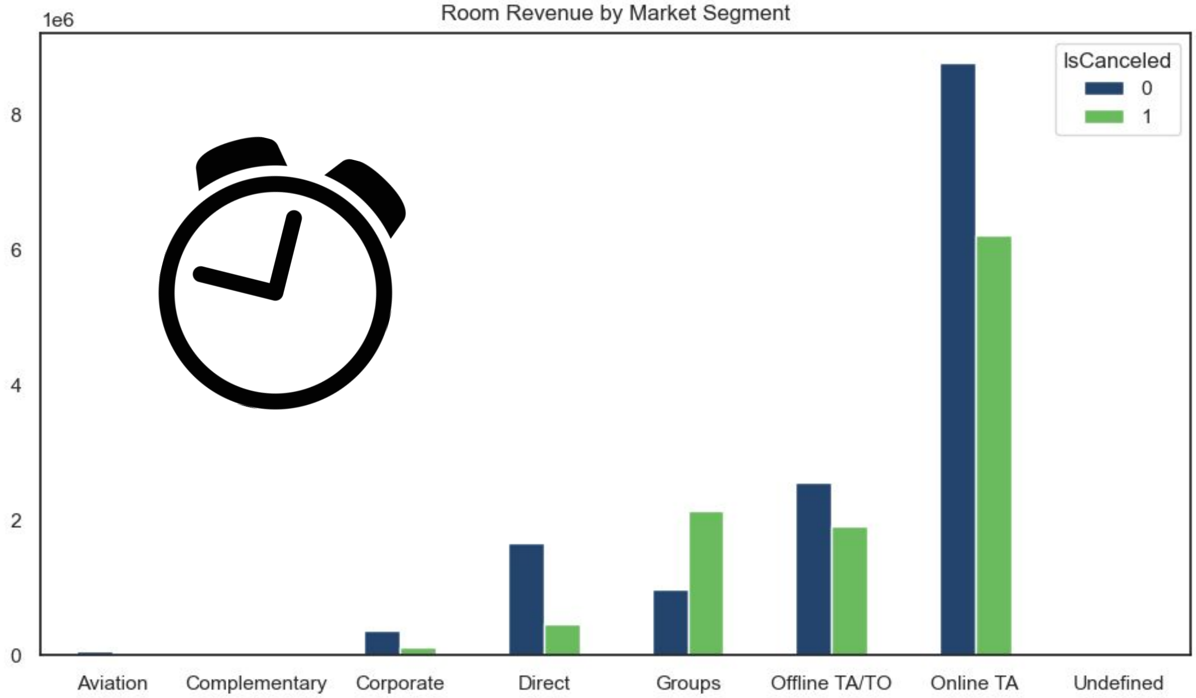
- Adjusting model threshold



Higher likelihood of customer relocation

Waive potential revenue

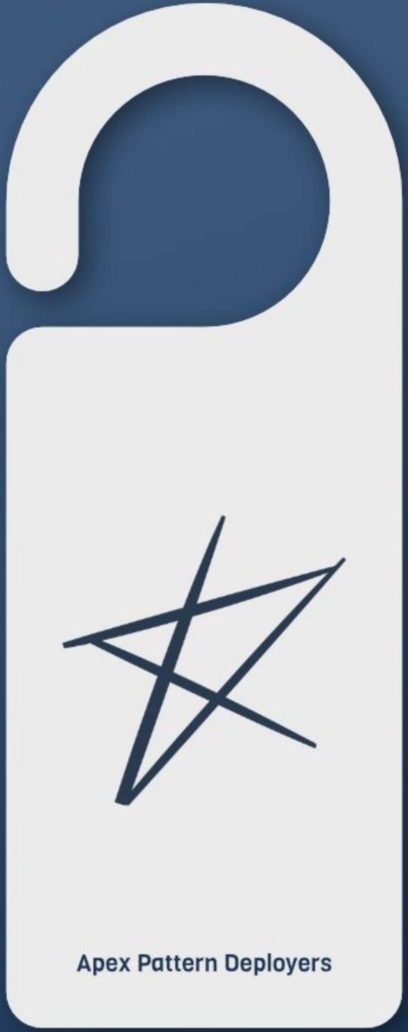
Cancellation Policies



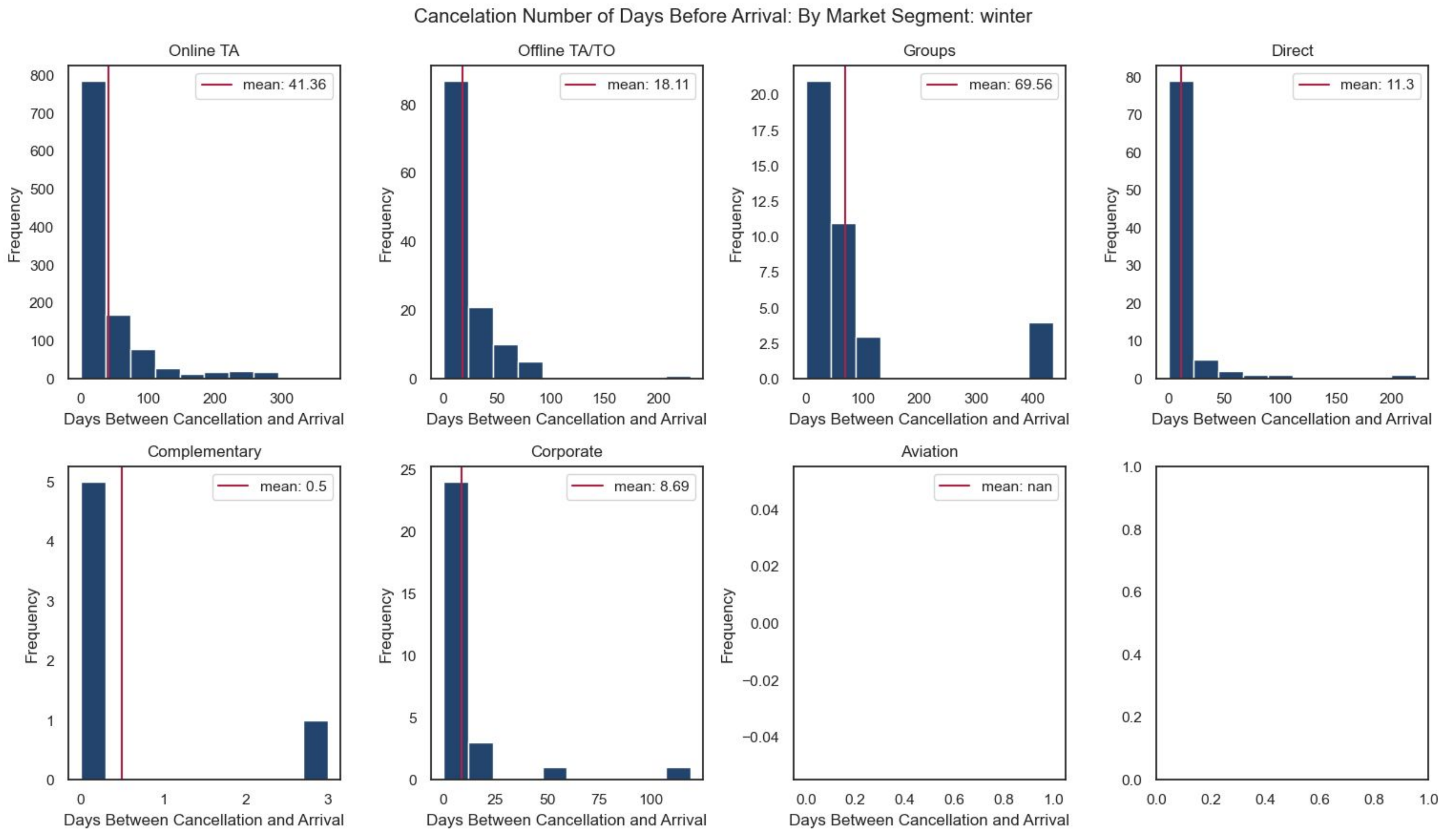
Revenue Estimation

8 416 Predictions → 50% → 4 208 Calls/Offers → 50% → 2 104 Accept

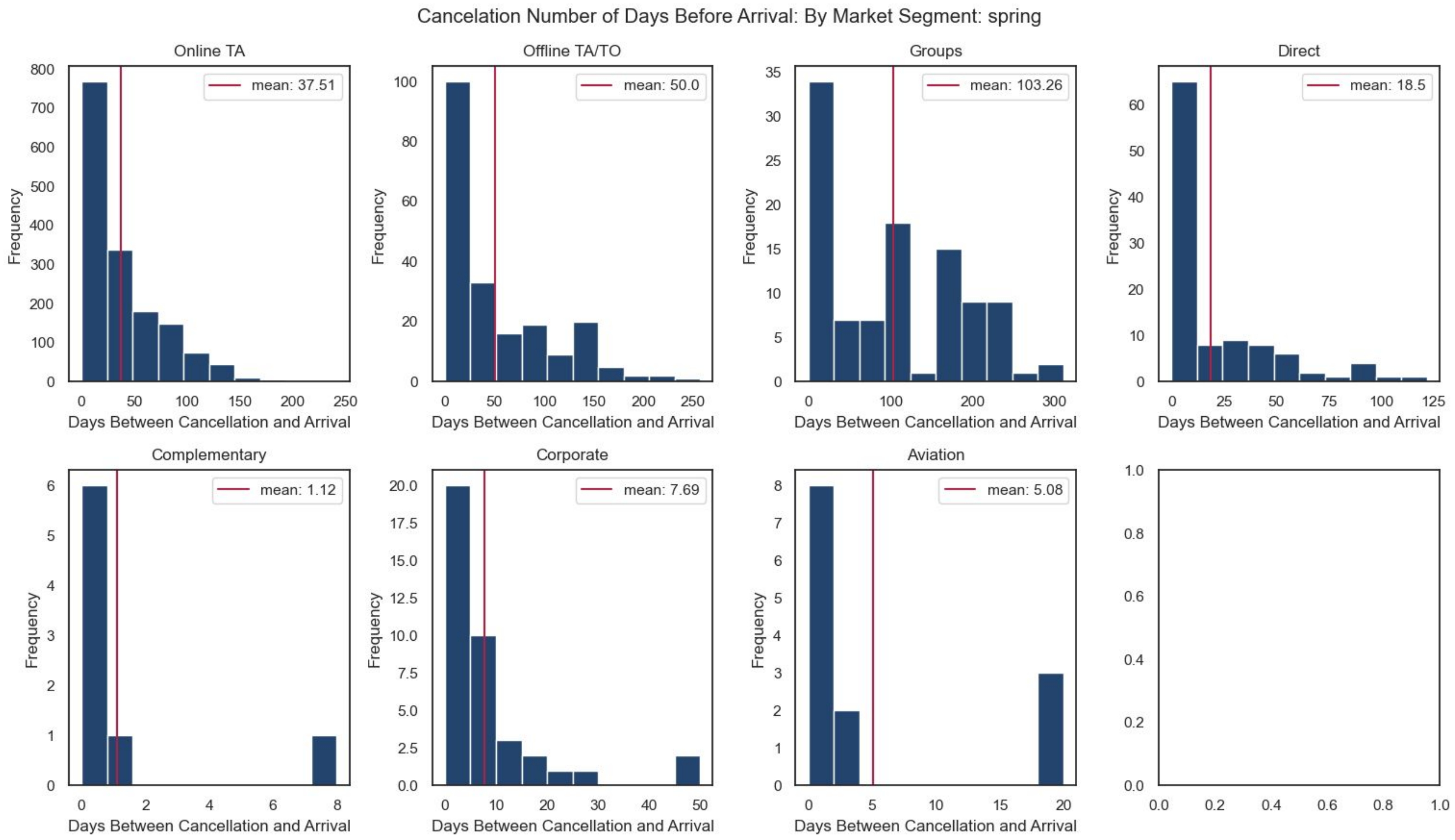
2 104 Accept × Average Booking 318 € = **669 072 €**



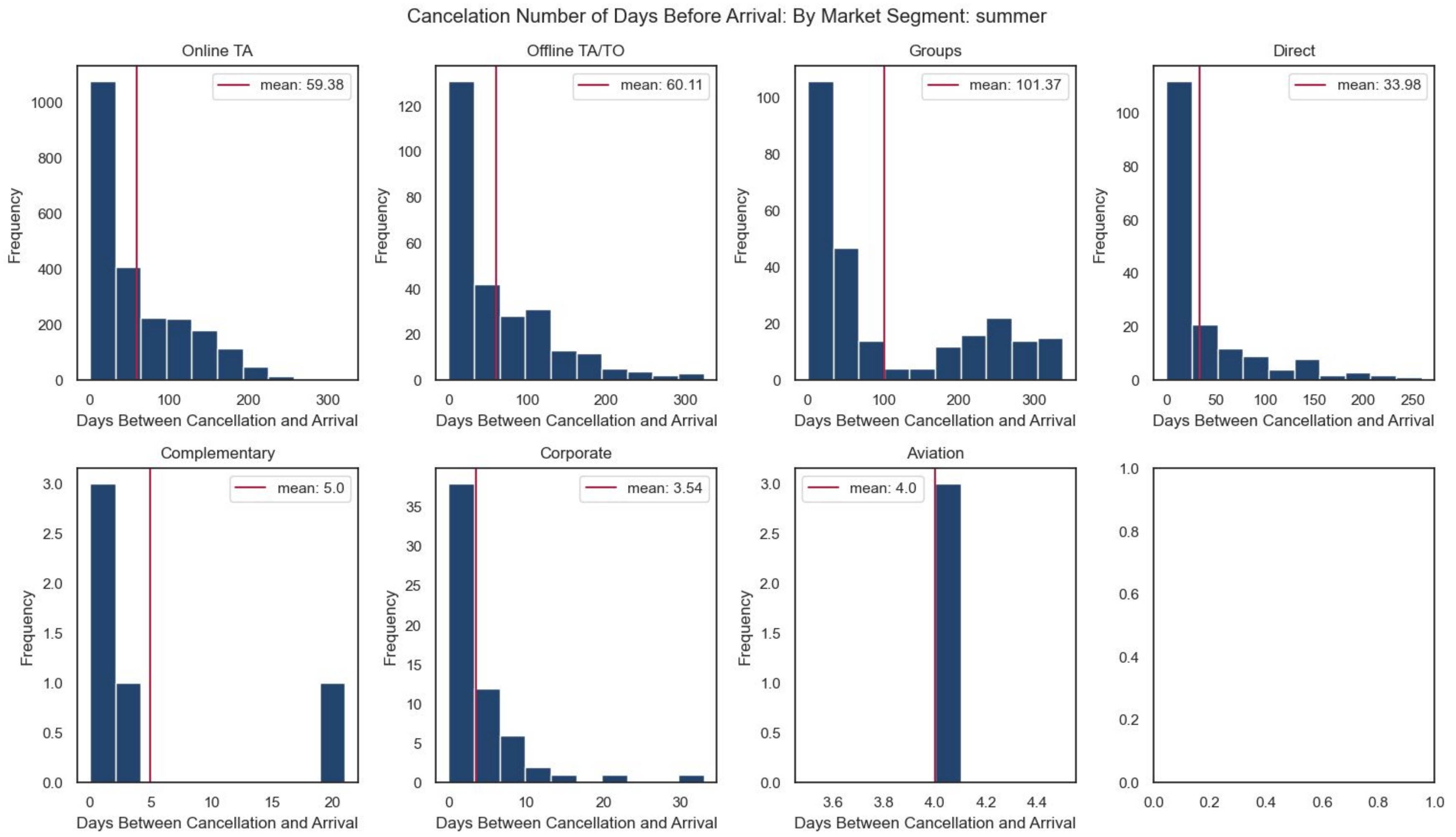
Thank you.

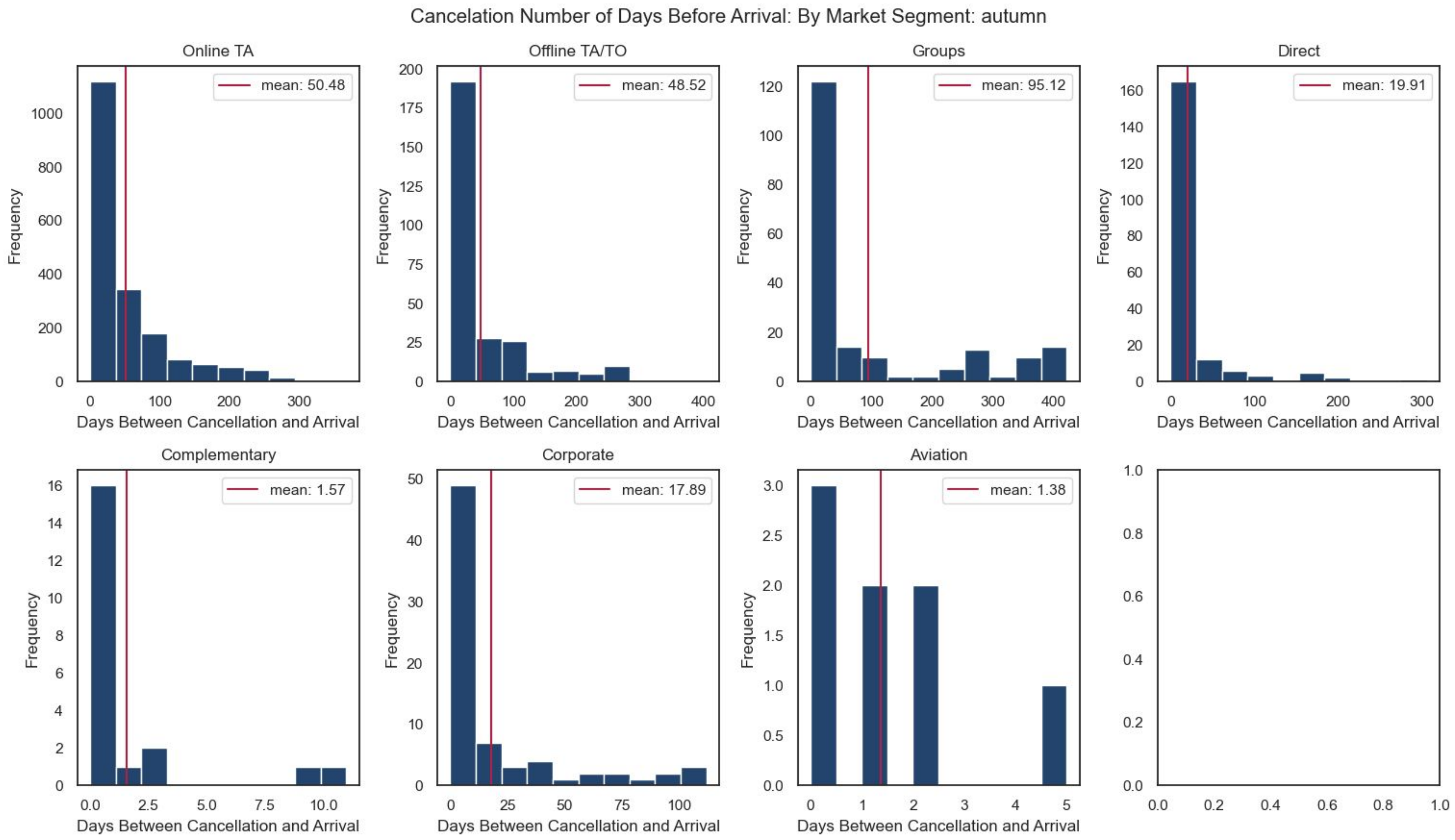


Cancelation Number of Days Before Arrival: By Market Segment: winter
Online TA
mean: 41.36
Offline TA/TO
mean: 18.11
Groups
mean: 69.56
Direct
mean: 11.3
Complementary
mean: 0.5
Corporate
mean: 8.69
Aviation
mean: nan
Frequency
Days Between Cancellation and Arrival



Cancelation Number of Days Before Arrival: By Market Segment: spring
Online TA
mean: 37.51
Offline TA/TO
mean: 50.0
Groups
mean: 103.26
Direct
mean: 18.5
Complementary
mean: 1.12
Corporate
mean: 7.69
Aviation
mean: 5.08
Frequency
Days Between Cancellation and Arrival


Cancelation Number of Days Before Arrival: By Market Segment: summer
Online TA
mean: 59.38
Offline TA/TO
mean: 60.11
Groups
mean: 101.37
Direct
mean: 33.98
Complementary
mean: 5.0
Corporate
mean: 3.54
Aviation
mean: 4.0
Frequency
Days Between Cancellation and Arrival


Cancelation Number of Days Before Arrival: By Market Segment: autumn
Online TA
mean: 50.48
Offline TA/TO
mean: 48.52
Groups
mean: 95.12
Direct
mean: 19.91
Complementary
mean: 1.57
Corporate
mean: 17.89
Aviation
mean: 1.38
Frequency
Days Between Cancellation and Arrival