

# HOTEL BOOKING MANAGEMENT

## Analyzing Cancellations

# TEAM MEMBERS



OYUNDARI  
BATBAYAR

Business Analyst



JAYA  
CHATURVEDI

Data Analyst



VINAY  
GOVIAS

Data Scientist



KAZ  
HAYASHI

Marketing Analyst



REO PAUL  
JACKSON

Data Scientist



# TABLE OF CONTENTS

•01• Problem Statement

•02• Data Exploration

•03• Data Preparation

•04• Modeling

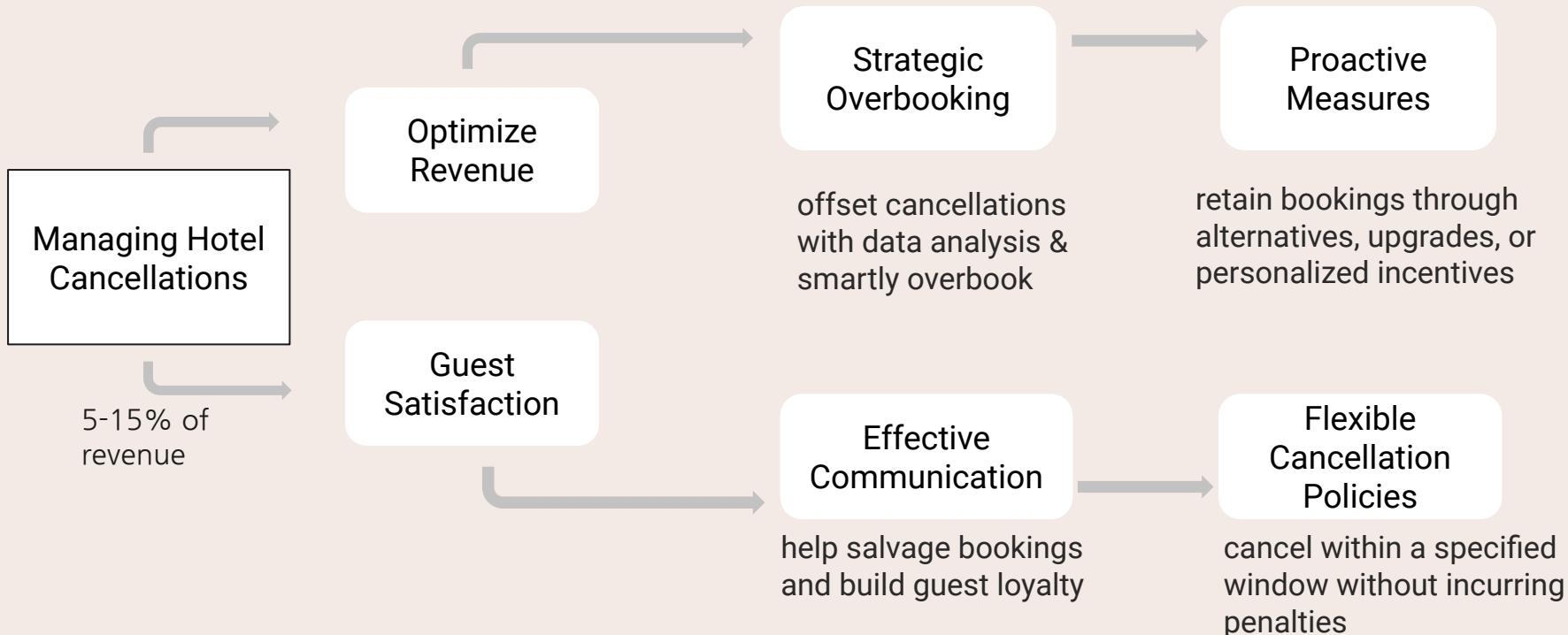
•05• Causal Inference

•06• Conclusions



# •01• PROBLEM STATEMENT

# 1.1 BUSINESS PROCESS FLOW



How can hotels effectively manage cancellations and enhance their competitive edge in the hospitality sector by leveraging ML tools and strategic approaches?

# 1.2 HYPOTHESIS

## MODELING FRAMEWORK

### CLASSIFICATION MODEL

*Predict if a customer would cancel a booking*

Predict Probability of  
Cancellation of each  
new booking

Identify features  
contributing to higher  
cancellations

### CAUSAL INFERENCE MODEL

*Causal Impact of Deposit on cancellations*

Identify customer  
groups sensitive to  
deposits

Create customized  
deposit policies for  
different customers

Based on the findings from these models, hotels can effectively design their reservation & cancellation policies, communication strategies and marketing campaigns to effectively handle cancellations and maximize revenue.





## •02• DATA EXPLORATION

## 2.1 OVERVIEW OF DATASET

**Data Source:** [Kaggle](#)

**Columns:** The Dataset contains a mix of data types - 36 Columns in total with 20 numeric columns and 16 categorical columns. The columns can be divided into 5 categories based on the information they contain:

1. **Customer Related Info:** Name, E-mail, Credit\_card, Phone\_number, Repeat Guests, Previous Bookings, Customer Type.
2. **Hotel Details:** Hotel name, Average Daily Rate
3. **Hotel Amenities:** Car Parking, Special Requests, Meals
4. **Booking Related Info:** Arrival Date, Stay Duration, Guest Composition(Adult, Children, Babies), Booking Changes, Booking Source, Waiting Time, Deposit Type, Reservation status
5. **Target Variable:** IsCancelled (Yes/No)

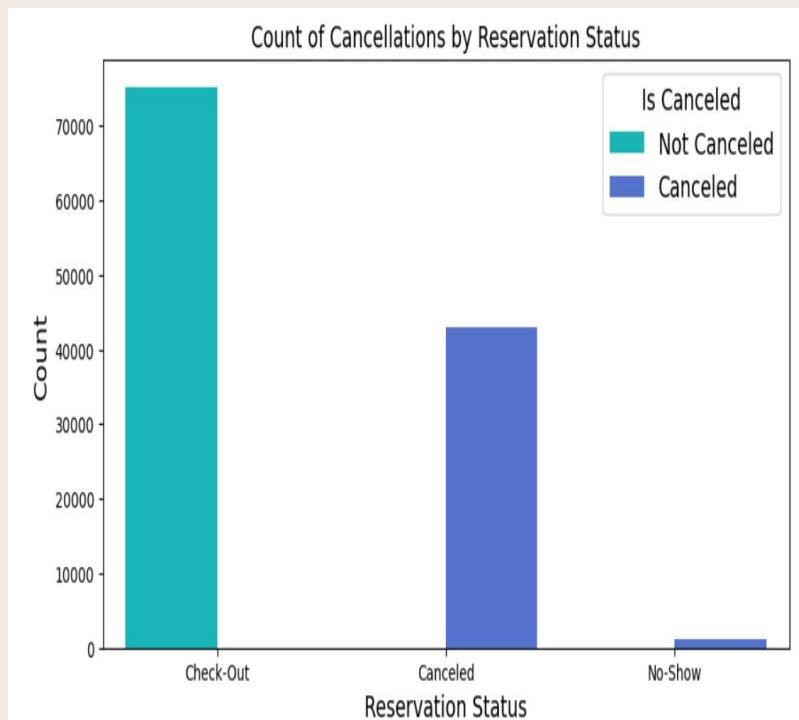


## 2.2 TARGET VARIABLE BY RESERVATION STATUS

The bar chart shows the count of hotel reservations by reservation status, with a distinction between those that were canceled and those that were not. From the chart, we can infer the following:

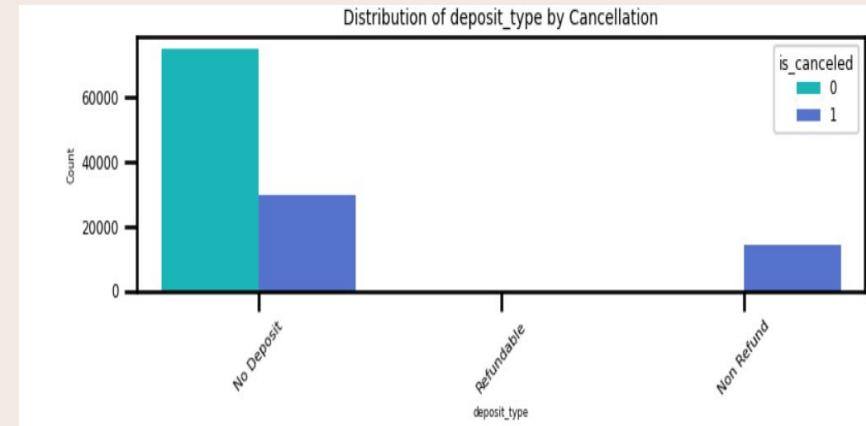
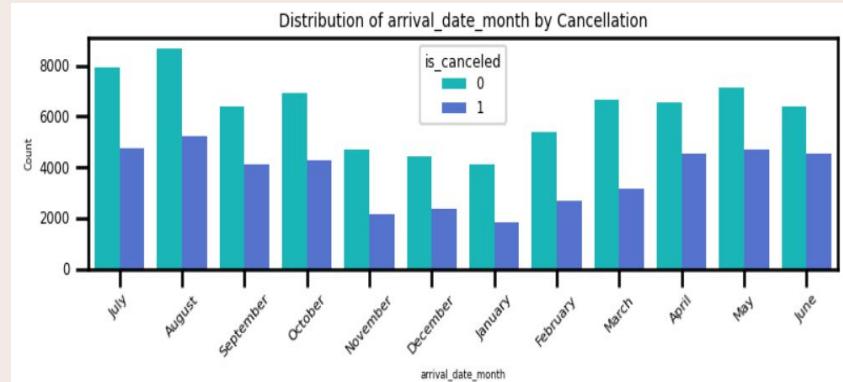
- 1. Check-Out:** A large number of reservations resulted in stays.
- 2. Canceled:** There is a significant count of reservations that were canceled.
- 3. No-Show:** For a smaller number of reservations, the guest did not arrive or did not formally cancel the booking.

The chart highlights that the majority of bookings in the dataset led to the guests checking out. This could indicate good conversion from bookings to actual stays for the hotel(s) in question. However, the count of cancellations could also suggest areas for improvement in customer retention or booking policies.



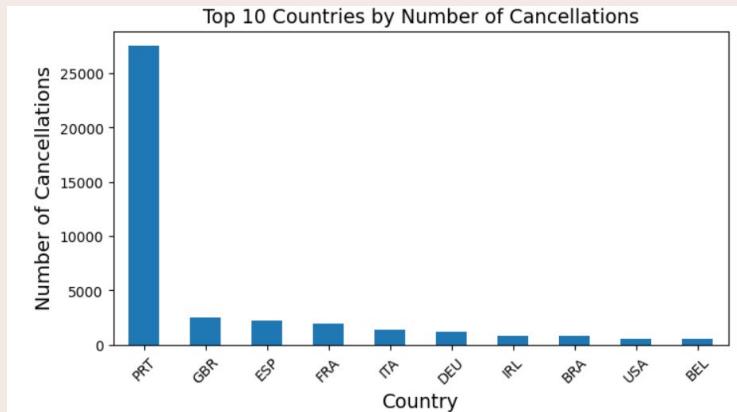
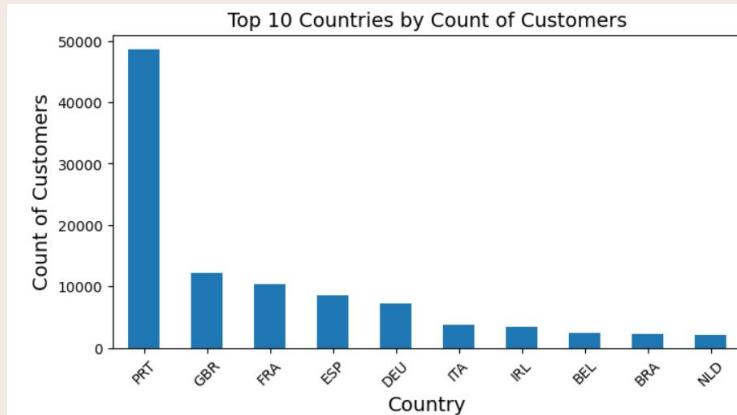
## 2.3 ANALYSIS OF BOOKINGS

- Both canceled and not canceled bookings are present in every month.
- The summer months (July and August) and spring months (March, April, May) show higher overall booking counts.
- Months with Higher Cancellations:** April, May, and June
- Months with Lower Cancellations:** September and October
- The majority of bookings are made with no deposit and have a lower cancellation rate compared to those with non-refundable deposits.
- Bookings made with refundable deposits have the least count. This could indicate that guests are less likely to cancel when they've made a financial commitment, especially if the deposit is non-refundable.



## 2.4 ANALYSIS OF CUSTOMERS

- The bar charts represent the top 10 countries by count of customers and number of cancellations at a hotel.
- There is a clear pattern that countries with higher booking counts also have higher cancellations.
- The country with the highest number of customers and cancellations is Portugal (PRT), followed by Great Britain (GBR), indicating a strong customer base from these countries.
- The presence of certain countries like the United States (USA) in the top 10 for cancellations but not for non-cancellations could indicate different booking behaviors or economic factors affecting travel decisions.



# •03• DATA PREPARATION



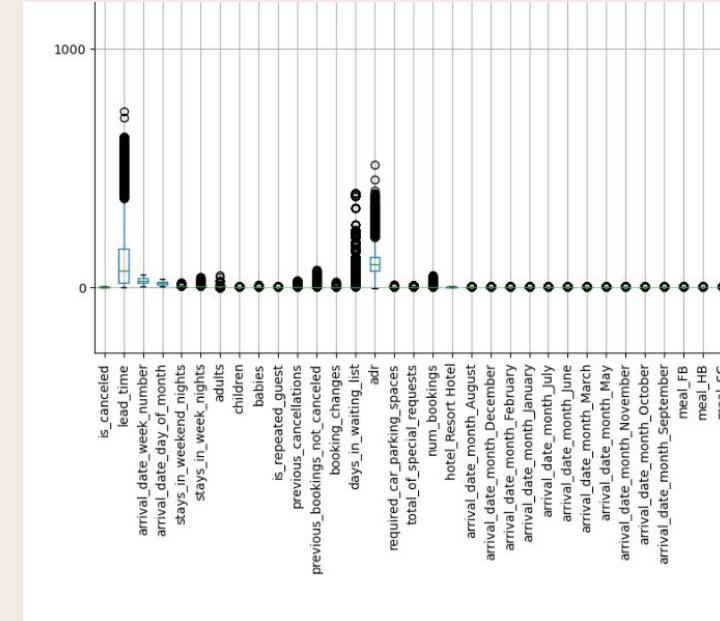
# 3.1 DATA CLEANING & FEATURE ENGINEERING

- **Prevent Data Leakage**
  - Drop `reservation_status_date`, `reservation_status`, and `assigned_room_type` columns, as they are updated post-cancellation.
- **Handle Sparse and Null Data**
  - Exclude `Company` and `Agent` columns due to a high number of unique values and nulls exceeding 10%.
- **Remove Sensitive Information**
  - Eliminate Personally Identifiable Information (PII) like name, phone number, email etc. to ensure data privacy.
- **Feature Engineering**
  - Create a new feature to track the number of bookings for each customer prior to their arrival date.
- **Modify Country Feature**
  - Retain only distinct country names with more than 1000 occurrences in the dataset for analysis.



## 3.2 OUTLIER REMOVAL AND FEATURE SELECTION

- Use Box Plots to show the spread of data and to identify outliers in various columns.
- Use Isolation Forest to remove the outliers detected (setting contamination to 2%)
- Using Random Forest feature selection, drop the 10 least important features like reserved\_room\_type\_L, market\_segment\_Undefined, distribution\_channel\_Undefined



## 3.3 BALANCING OF CLASSES IN THE DATASET

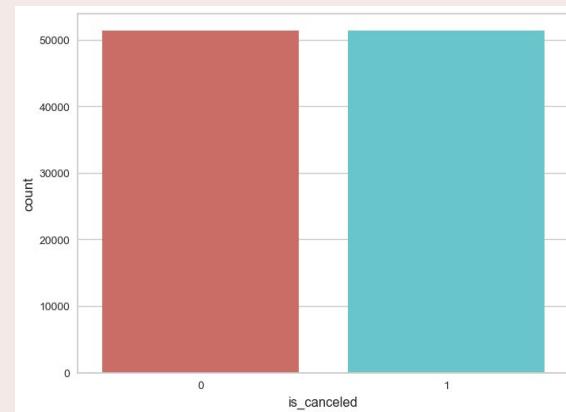
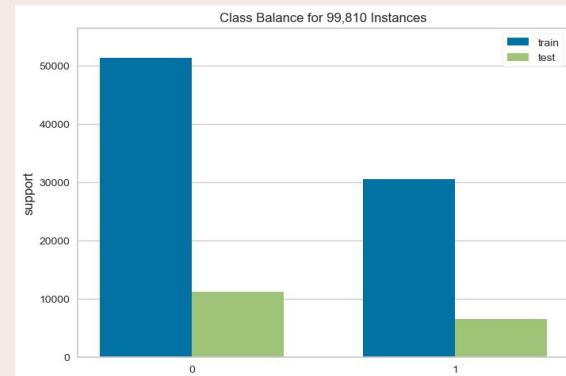
- **Random Oversampling :**

Over-sample the minority class(es) by picking samples at random with replacement.

The bootstrap can be generated in a smoothed manner

- **SMOTE :**

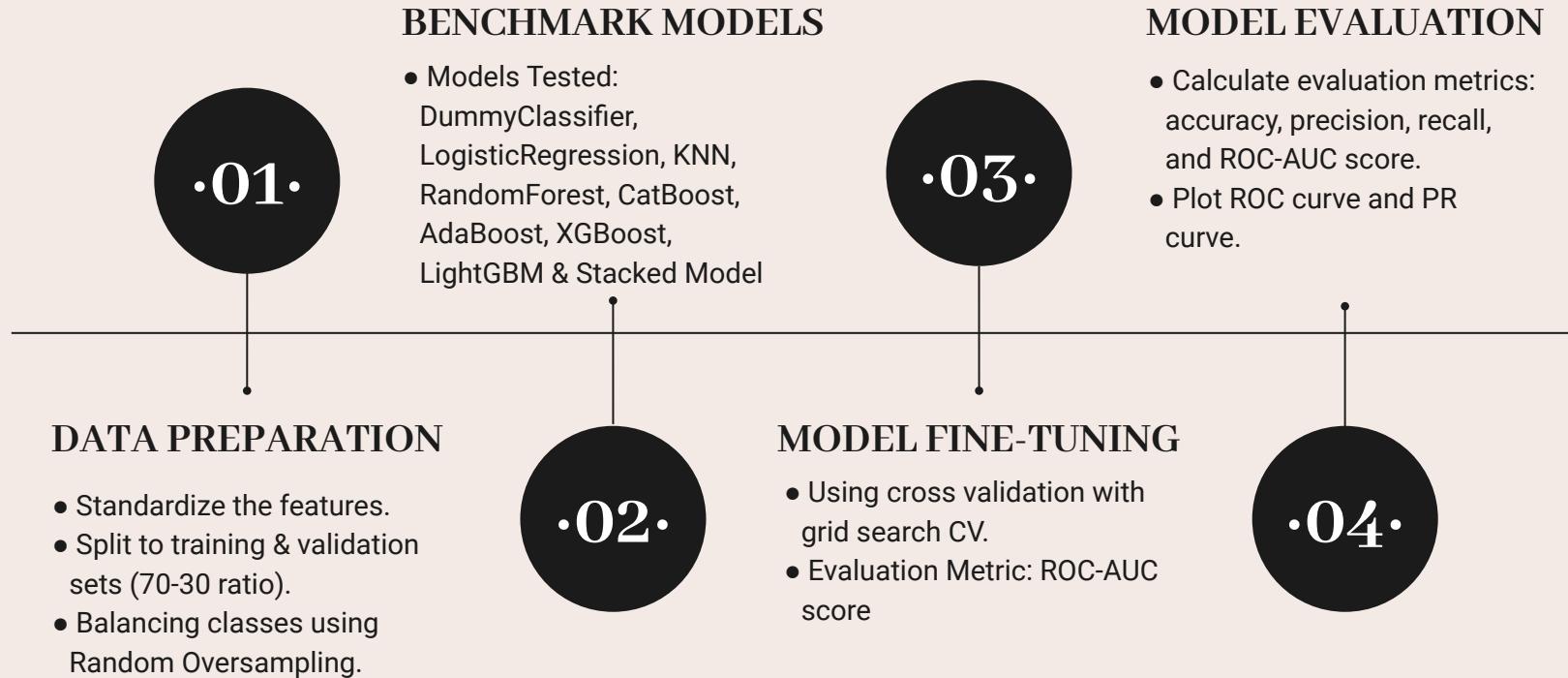
Balances the dataset by creating synthetic examples rather than by oversampling with replacement.



# •04• MODELING



# 4.1 MODELING PROCESS



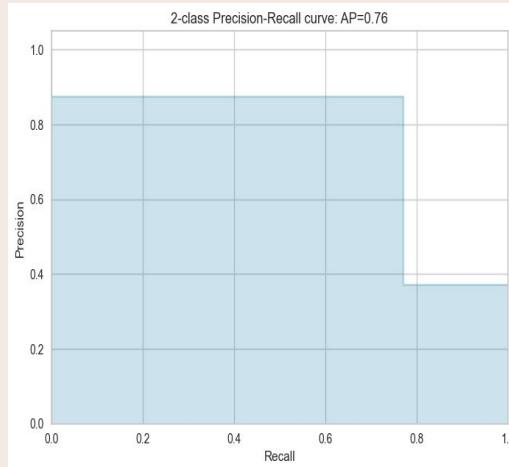
## 4.2 MODEL SELECTION (*using validation set*)

SI No	Model	Default Model		Fine-tuned Model	
		Accuracy	ROC-AUC	Accuracy	ROC-AUC
1	DummyClassifier	0.631	0.500		
2	LogisticRegression	0.798	0.786		
3	KNN	0.812	0.796		
4	<b>RandomForest</b>	<b>0.878</b>	<b>0.862</b>	<b>0.880</b>	<b>0.865</b>
5	XGBoost	0.733	0.763	0.769	0.785
6	AdaBoost	0.812	0.801	0.816	0.794
7	CatBoost	0.685	0.736	0.850	0.845
8	LightGBM	0.852	0.846	0.859	0.854
9	Stacked Model			0.876	0.852

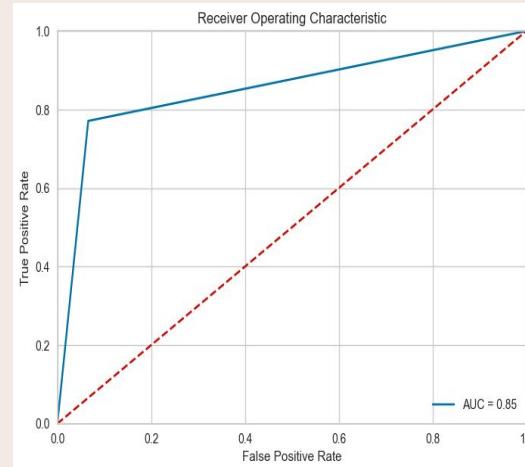


## 4.3 MODEL PERFORMANCE (*using best model on test set*)

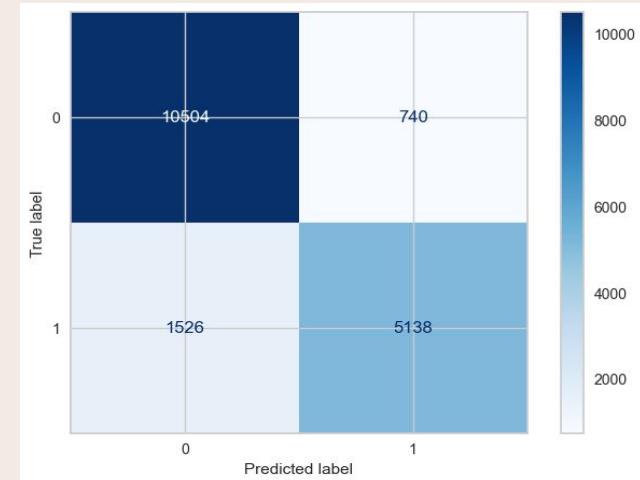
Accuracy	Precision	Recall	F1 score	ROC-AUC score
0.873	0.874	0.771	0.819	0.852



Precision-Recall curve

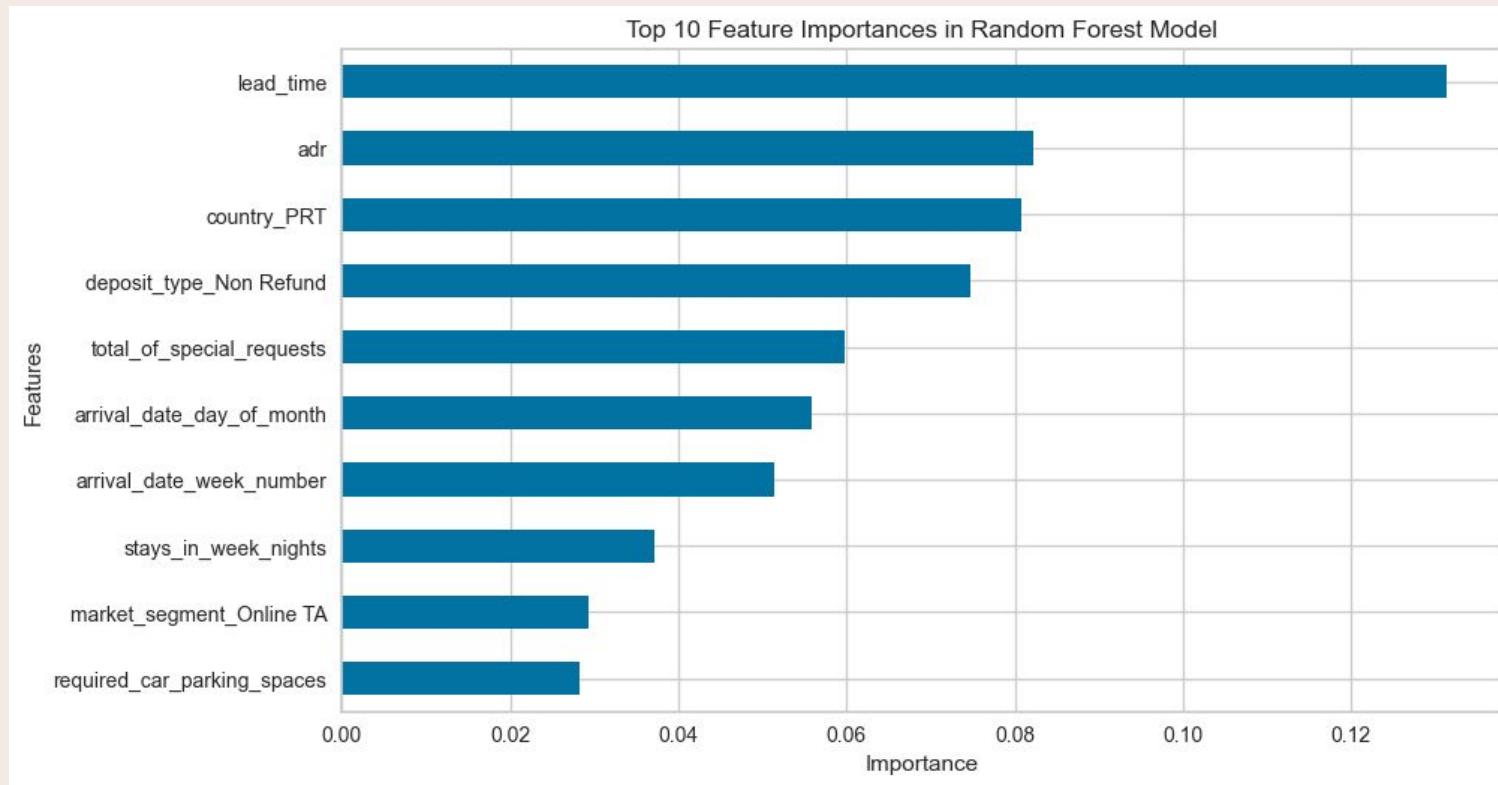


ROC curve



Confusion Matrix

## 4.4 FEATURE IMPORTANCE



^  
Home  
▼

## 4.4.1 INTERPRETATION



### LEAD TIME

Time span between booking and actual stay is the most important factor affecting cancellations.



### PRICE

Average Daily Rate of the room is the next important feature, suggesting that the pricing strategy of the hotel impacts cancellations.



### BACKGROUND

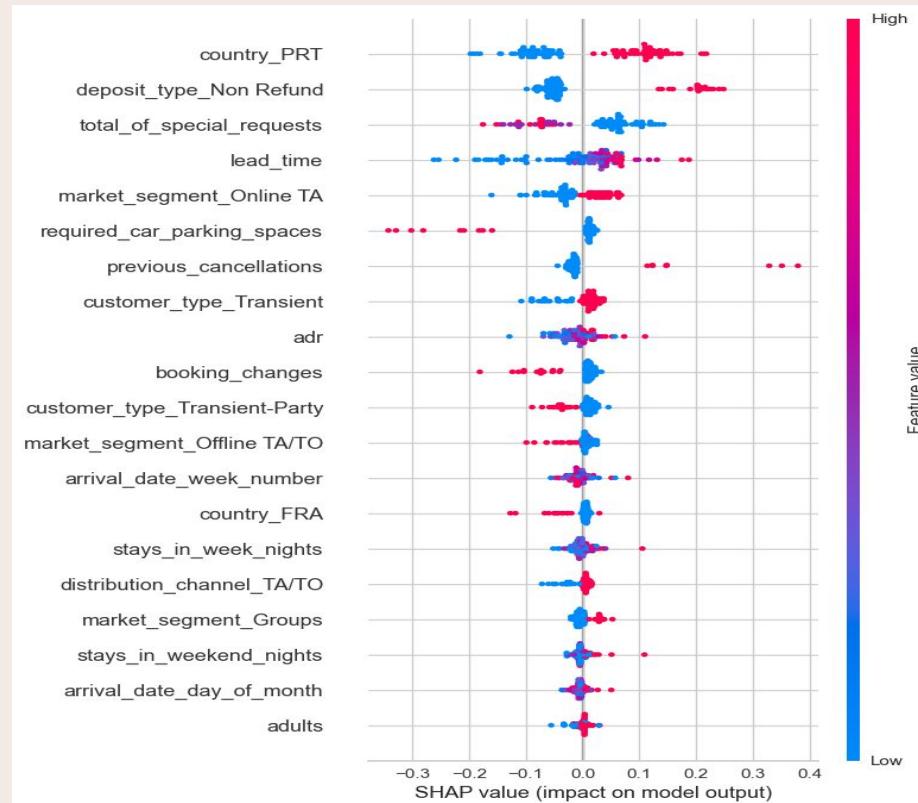
Country of Origin of customers impacts cancellation rates, possibly due to market-specific behavior or preferences.



### POLICIES

For example, the deposit type and special requests made by customers are important predictors affecting customer satisfaction and retention.

# 4.5 SHAP VALUES



## 4.5.1 INTERPRETATION



### CUSTOMER NATIONALITY

Example: Customers from Portugal are more likely to cancel



### DEPOSIT POLICY

Customers with non-refundable deposits are more likely to cancel



### SPECIAL REQUESTS

Customers who made special requests are more likely to be committed to the booking



### LEAD TIME

Longer time span between booking & check-in date leads to higher cancellation rates.



### BOOKING SOURCE

Customers who booked through online travel agencies are more likely to cancel



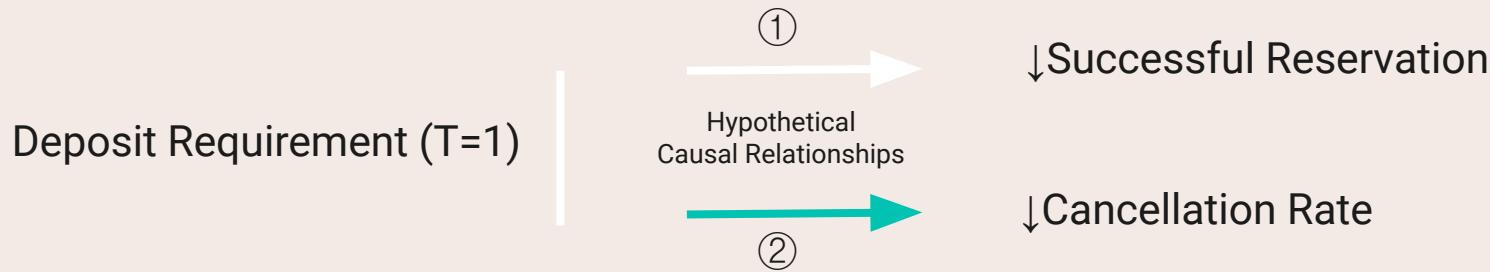
### AMENITIES

Bookings for hotels with additional amenities like parking space are less prone to cancellations.

# •05• CAUSAL INFERENCE



# 5.1 HYPOTHESIS - The Deposit



- Asking for deposit has **both positive and negative** impacts on the business
- For some customers, deposit might not matter or have smaller effect on ↓cancellation
- **If deposit does not reduce cancel rate, why would you ask for deposit?**
- So, we don't want to ask for deposit for every customer.

\*we will assume ① is true!

# 5.1 HYPOTHESIS - The Deposit

In this scenario, what is the goal of the hotel/Travel Agency?

→ Maximize reservation while minimizing cancellation by offering **dynamic deposit policy**



→ Logic: Dynamic Deposit Policy → More Successful reservations!

→ Method: Let's use uplift modeling to see how the effect deposit differs by ind/subgroups

## 5.2 CAUSAL INFERENCE MODELING

T-learner: the goal is to estimate CATE / ITE (subgroup or individual)

- ① Fit different model on T1 and T0 group (two models, so it is T-learner)

$$Y_i(T = 1)|X_i \quad Y_i(T = 0)|X_i$$

- ② The difference between these two expectations for each set of covariates.

$$\tau(X_i) = \mathbb{E}[Y_i(T = 1)|X_i] - \mathbb{E}[Y_i(T = 0)|X_i]$$

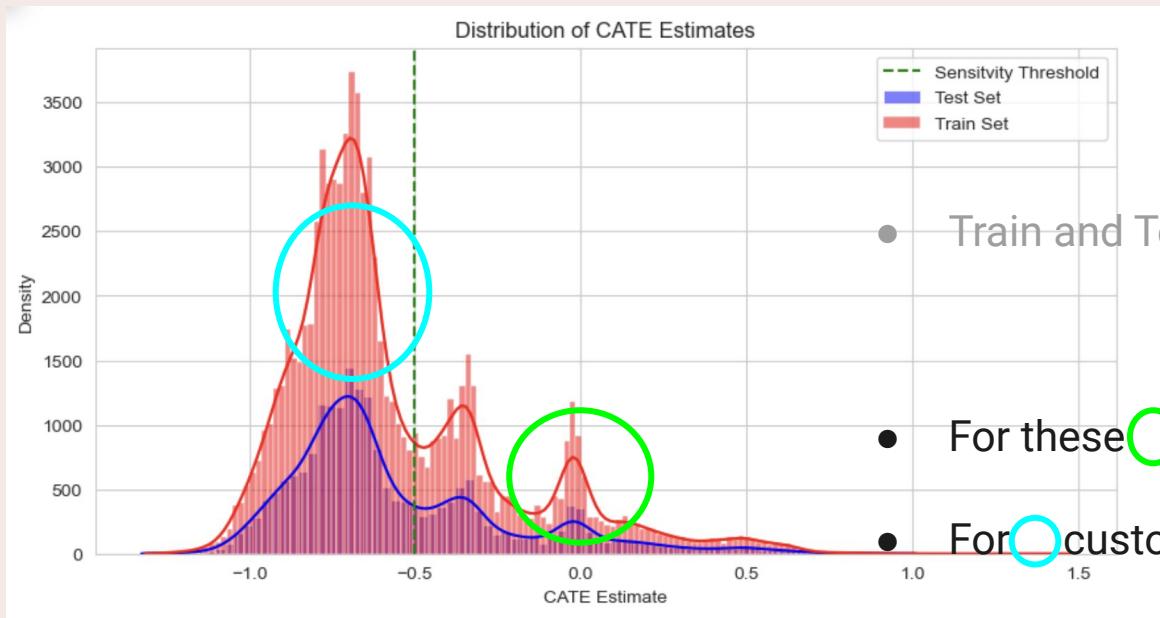


CATE/ITE for some set of X (from Japan, male, 23yo, etc.)

## 5.3 TREATMENT EFFECT

Average Treatment Effect (ATE) estimate: -0.56

- If you ask deposit, the chance of customers cancelling is lower



# 5.4 HETEROGENEITY

## Guest

Subgroup	CATE Estimate ( $\downarrow$ in cancellation)
Is Repeated Guest (0)	-0.56
Is Repeated Guest (1)	-0.33

First timers are more sensitive to deposit

## Distribution Channel

Subgroup	CATE Estimate
Distribution Channel - Direct	-0.63
Distribution Channel - Corporate	-0.44
Distribution Channel - TA/TO	-0.56
Distribution Channel - GDS	-0.71

GDS customers are often in large number. Thus, they might want to avoid losing deposit

## Market Segment

Subgroup	CATE Estimate
Market Segment - Groups	-0.50
Market Segment - Offline TA/TO	-0.69
Market Segment - Complementary	-0.78
Market Segment - Aviation	-0.56

If the trip is complimentary, you don't want to miss it nor lose the deposit

## 5.5 A SIMPLE SCENARIO: ROI ESTIMATION FOR FIRST-TIME GUESTS

$$ROI = \frac{Benefits - Costs}{Costs} \times 100 = 31.4\%$$

- Based on external data
- Our estimate

For every dollar spent on managing the deposit process, the hotel earns back \$31.40 in increased revenue and net gains from deposits.

Note this is where the simplification comes

Table. Assumptions for ROI Calculation

Parameter	Value
Average Revenue Per Booking (ARB)	\$200
Conditional Average Treatment Effect (CATE)	-0.56
Average Deposit Amount (ADA)	\$50
Number of First-Time Bookings (NFB)	1000
Costs Associated with Deposits (CAD) % of TDAC	10%

Table. Final Values from ROI Calculation

Metric	Value
Reduced Number of Cancellations (RNC)	560
Increased Revenue from Reduced Cancellations (IRRC)	\$112,000
Total Deposit Amount Collected (TDAC)	\$50,000
Costs Associated with Deposits (CAD)	\$5,000
Net Gain from Deposits (NGD)	\$45,000
ROI	31.4

- CATE is applied directly to estimate the reduction in cancellations due to the deposit policy. Note that this direct application is overly simplistic.
- **We must reconsider the rationale behind the deposit intervention, especially if it potentially reduces the overall number of successful bookings.**

## 5.6 FINDINGS

- The effect of deposit requirements on cancellation rates varies across different customer segments and booking conditions.
- Build policies that cater to the specific needs and sensitivities of different customer groups.
- By tailoring deposit requirements, hotels can optimize their booking to enhance customer satisfaction and revenue stability.



# 6. NEXT STEPS

•01•

## MODEL DEPLOYMENT

Create a user interface for the hotel management team to easily access and interpret model predictions.

•02•

## ADDITIONAL RESEARCH

Validating the impact of requiring a deposit on successful reservation rates requires further research since the dataset only includes data for completed reservations.

•03•

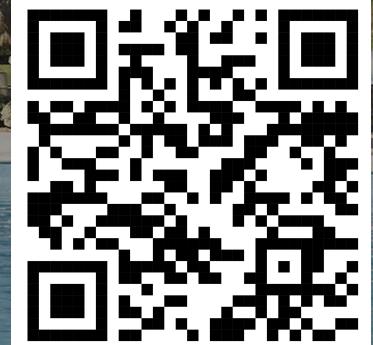
## OPTIMIZING DEPOSIT REQUIREMENT

With impact of deposit on cancellation and successful bookings, we formulate optimization problem to find the best mix of deposit policy for each customer



# THANK YOU!

## Questions?



# APPENDIX

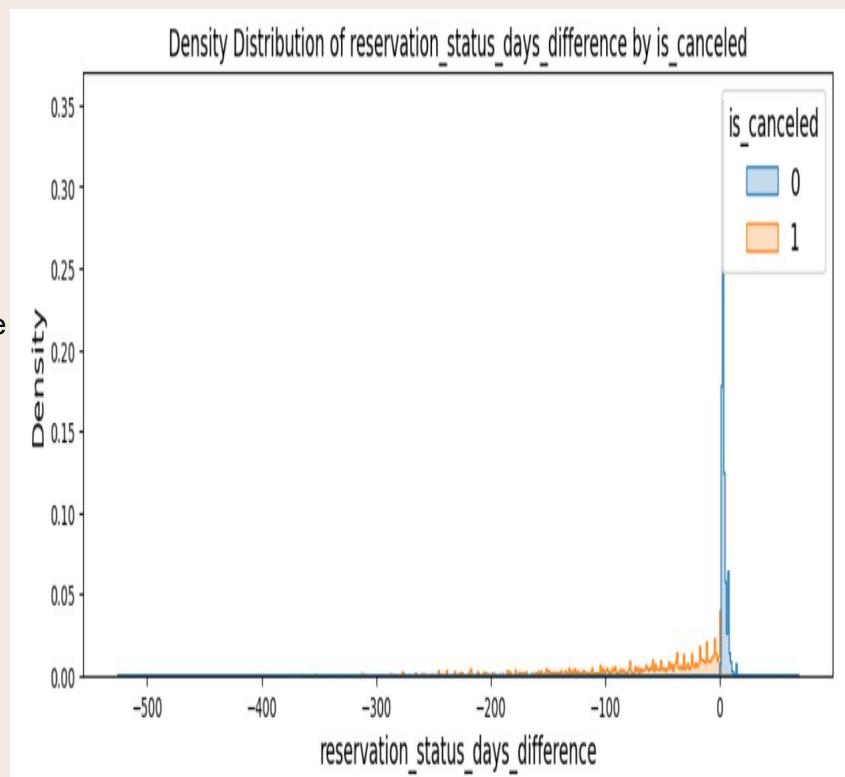


# ADDITIONAL INSIGHTS FROM EDA



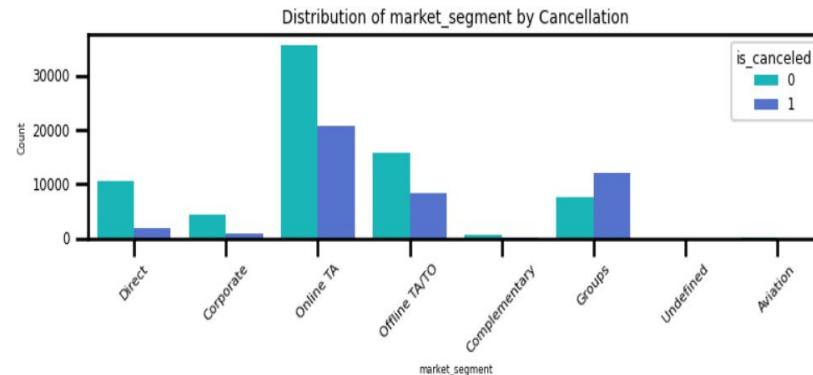
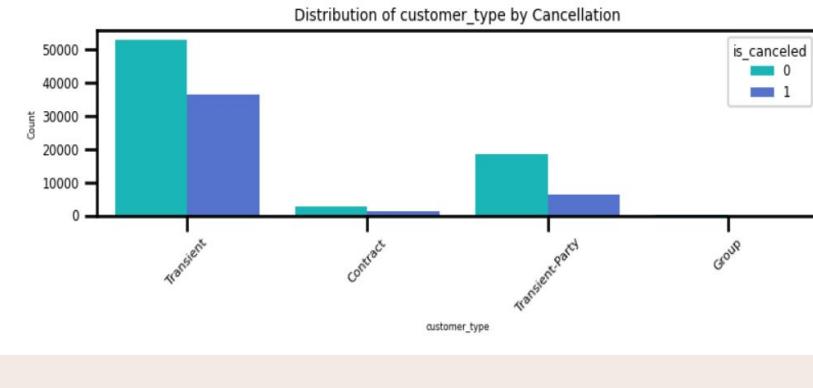
# ANALYSIS OF CANCELLATION TIME

- Bookings with a positive 'reservation\_status\_days\_difference' are not canceled. They represent cases where the booking was used and the status was updated after the stay.
- Bookings with a negative 'reservation\_status\_days\_difference' are all canceled, which indicate that the reservation was canceled a certain number of days before the expected arrival.
- This shows a clear relationship between the 'reservation\_status\_days\_difference' and the 'is\_canceled' feature. If a booking is canceled, the reservation status is usually updated before the arrival date. Conversely, if a booking is not canceled, the reservation status is usually updated after the arrival date.



# INSIGHTS

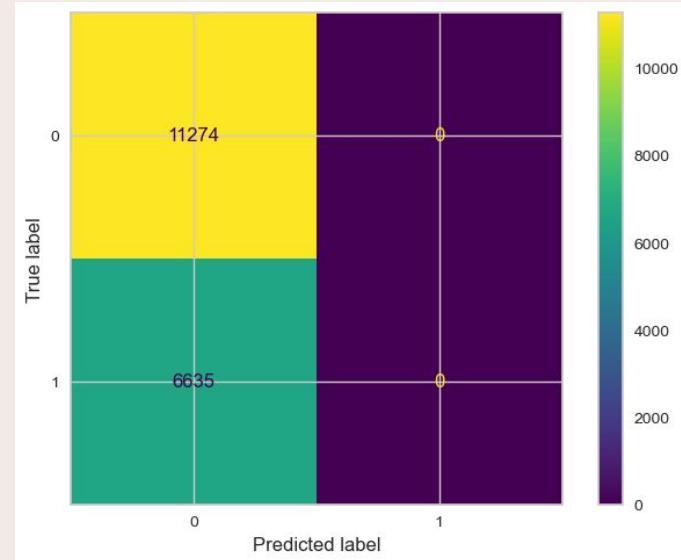
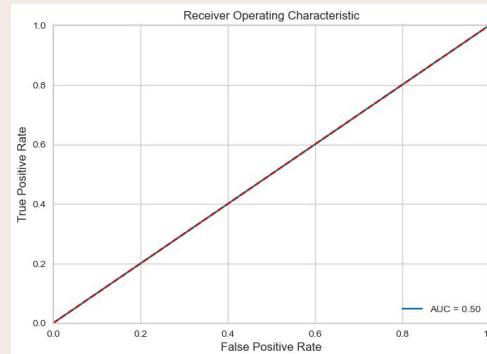
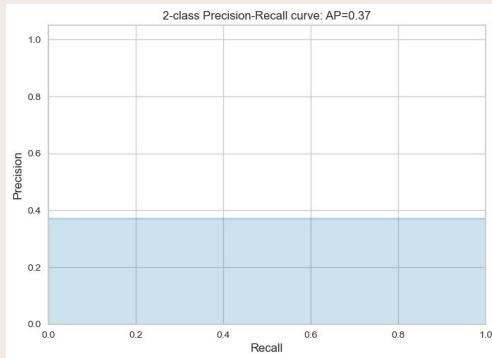
- Transient customers—the largest segment—have a substantial number of both canceled and non-canceled bookings, indicating high turnover.
- The Transient-Party segment has a relatively balanced distribution of cancellations, while the Group segment shows the least number of bookings.
- The Online TA (Travel Agent) segment has the highest overall count of bookings and a significant proportion of these bookings were canceled.
- The Offline TA/TO segment also has a substantial number of bookings with a relatively high cancellation rate. Direct bookings appear to have a lower cancellation rate compared to the TA/TO segments.
- This distribution suggests that the market segment is a considerable factor in the likelihood of cancellations, with online bookings showing a tendency for higher cancellations.



# MODEL EVALUATION



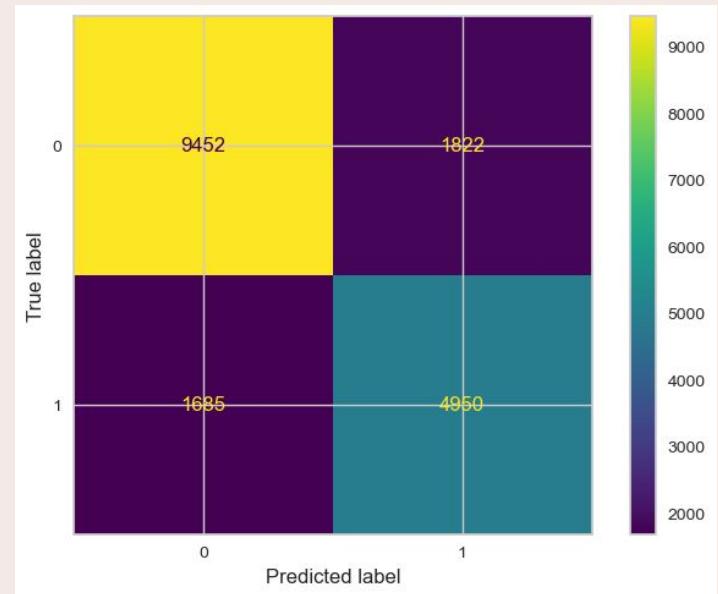
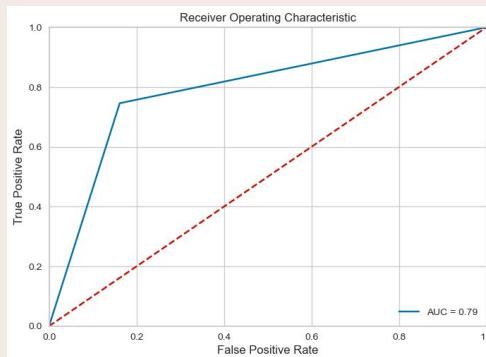
# Model Evaluation - Dummy Classifier



## Classification Report:

	precision	recall	f1-score	support
0	0.63	1.00	0.77	11274
1	0.00	0.00	0.00	6635
accuracy			0.63	17909
macro avg	0.31	0.50	0.39	17909
weighted avg	0.40	0.63	0.49	17909

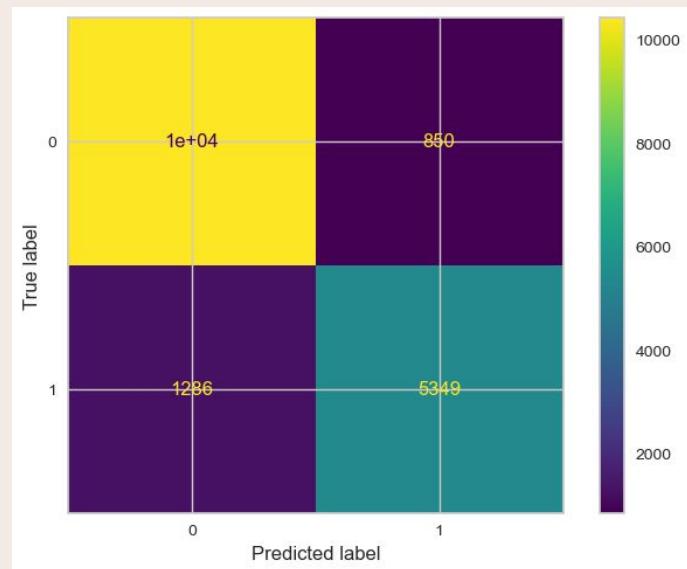
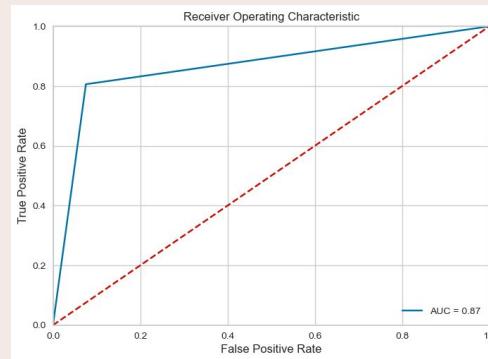
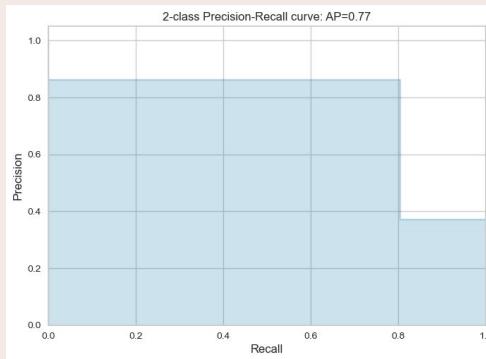
# Model Evaluation - Logistic Regression



Classification Report:				
	precision	recall	f1-score	support
0	0.85	0.84	0.84	11274
1	0.73	0.75	0.74	6635
accuracy			0.80	17909
macro avg	0.79	0.79	0.79	17909
weighted avg	0.81	0.80	0.80	17909

**accuracy score: 0.8042**

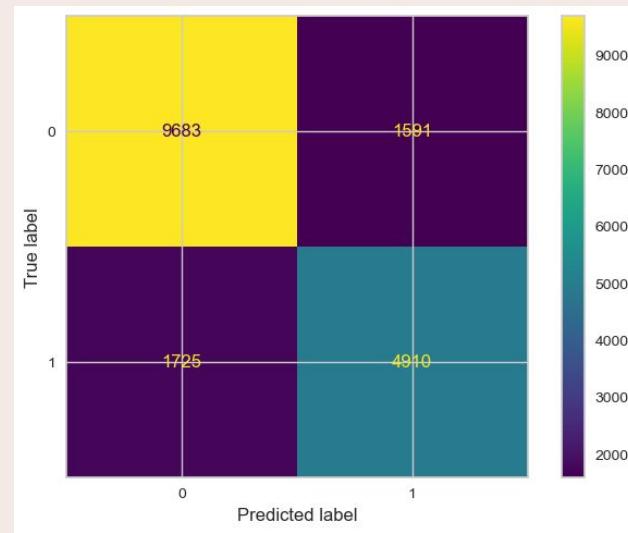
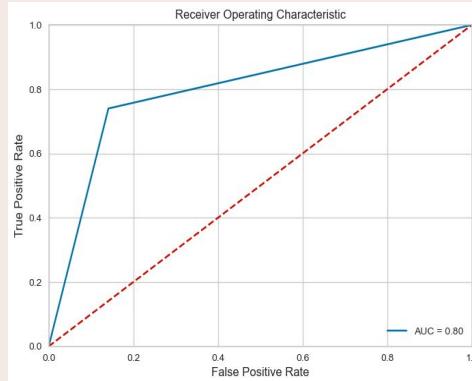
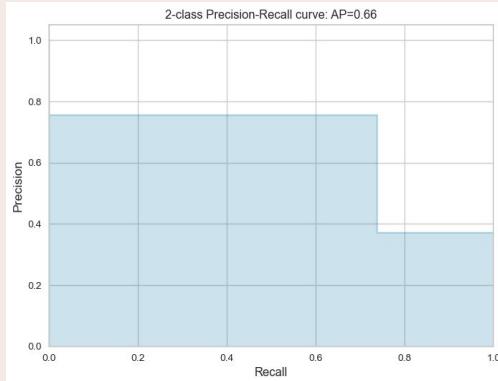
# Model Evaluation - Random Forest



Classification Report:				
	precision	recall	f1-score	support
0	0.89	0.92	0.91	11274
1	0.86	0.81	0.83	6635
accuracy			0.88	17909
macro avg	0.88	0.87	0.87	17909
weighted avg	0.88	0.88	0.88	17909

**accuracy score: 0.8807**

# Model Evaluation - KNN

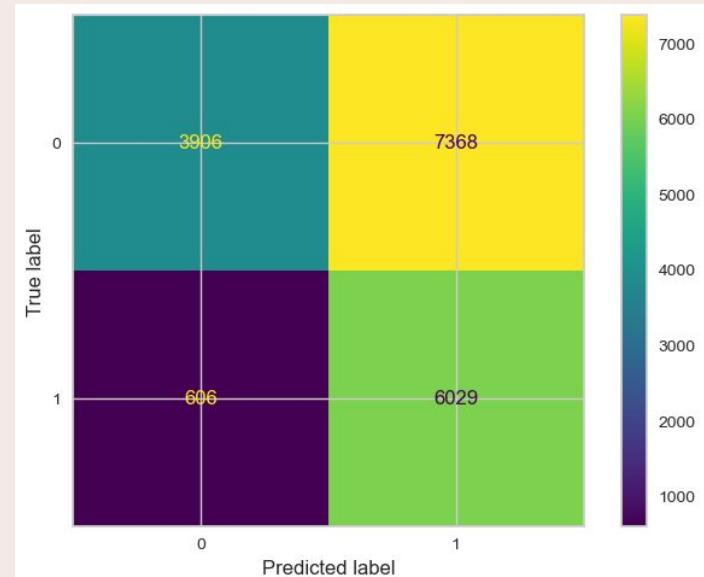
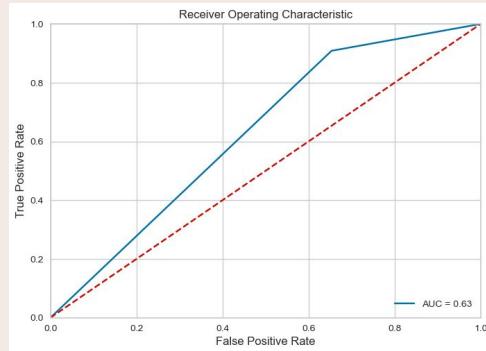
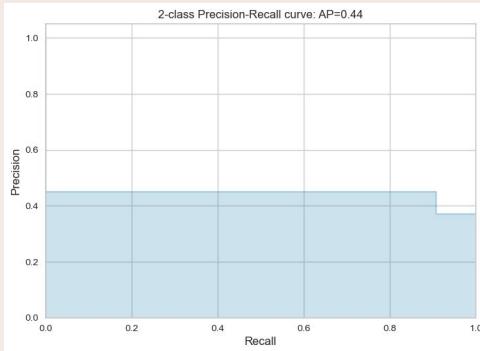


## Classification Report:

	precision	recall	f1-score	support
0	0.85	0.86	0.85	11274
1	0.76	0.74	0.75	6635
accuracy			0.81	17909
macro avg	0.80	0.80	0.80	17909
weighted avg	0.81	0.81	0.81	17909

**accuracy score: 0.8148**

# Model Evaluation - XGBoost

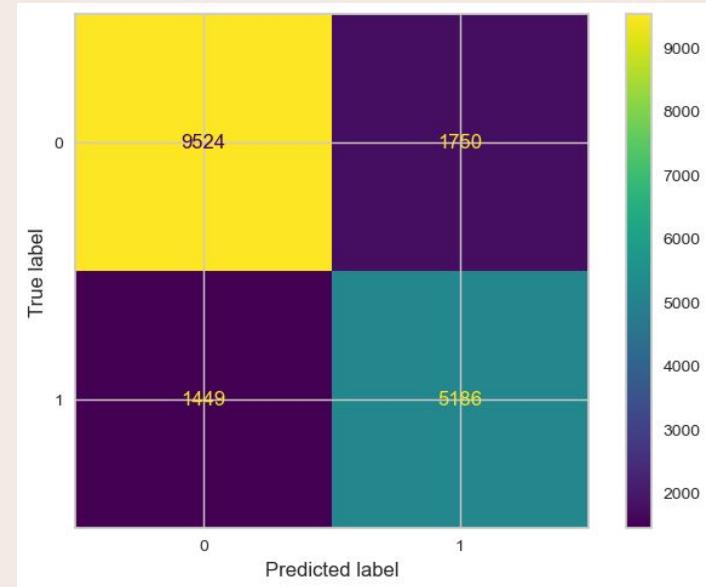
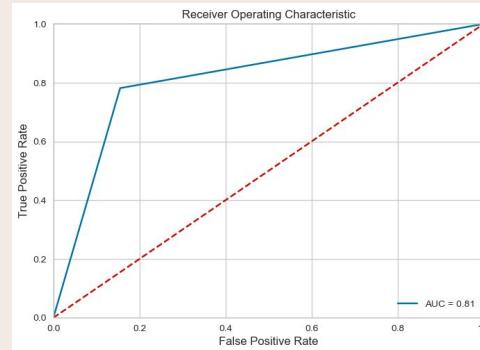
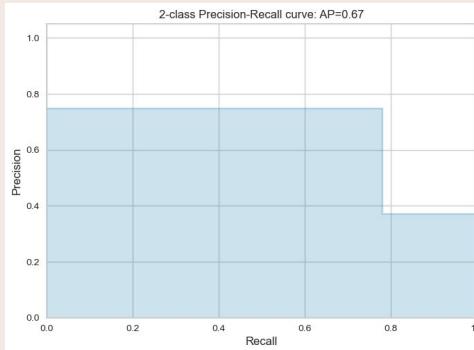


Classification Report:

	precision	recall	f1-score	support
0	0.87	0.35	0.49	11274
1	0.45	0.91	0.60	6635
accuracy			0.55	17909
macro avg	0.66	0.63	0.55	17909
weighted avg	0.71	0.55	0.53	17909

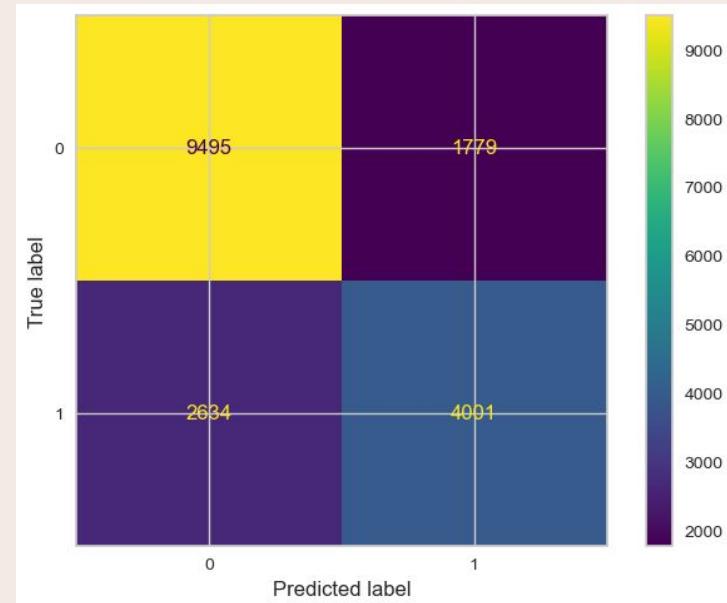
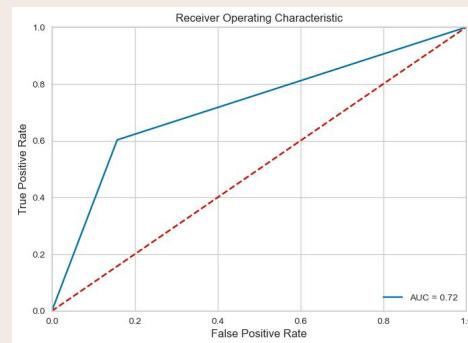
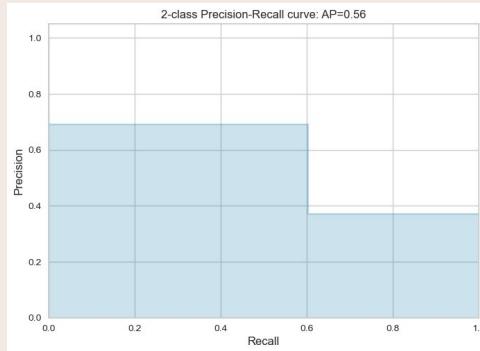
accuracy score: 0.5547

# Model Evaluation - AdaBoost



accuracy score: 0.8214

# Model Evaluation - CatBoost

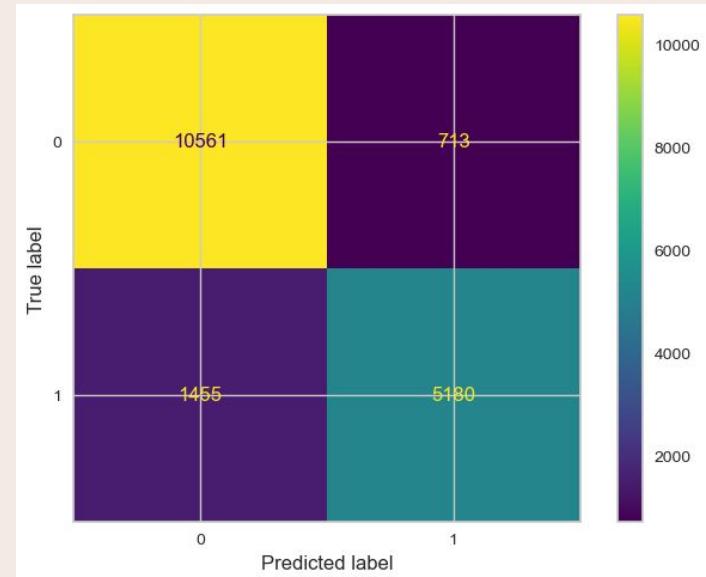
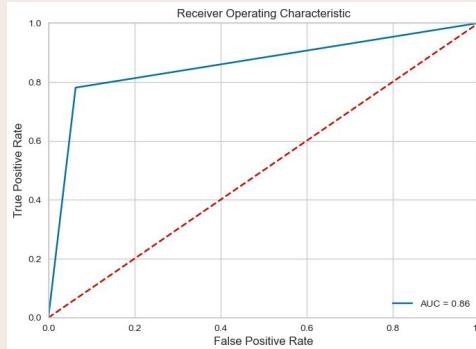
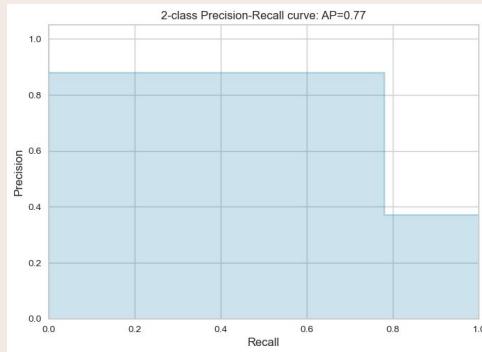


Classification Report:

	precision	recall	f1-score	support
0	0.78	0.84	0.81	11274
1	0.69	0.60	0.64	6635
accuracy			0.75	17909
macro avg	0.74	0.72	0.73	17909
weighted avg	0.75	0.75	0.75	17909

accuracy score: 0.7536

# Model Evaluation - Stacked Model



## Classification Report:

	precision	recall	f1-score	support
0	0.88	0.94	0.91	11274
1	0.88	0.78	0.83	6635
accuracy			0.88	17909
macro avg	0.88	0.86	0.87	17909
weighted avg	0.88	0.88	0.88	17909

accuracy score: 0.8789