PORTOFOLIO DATA SCIENCE

Zinedine Amalia N.M.R

Table of Contents

Introduction	03	Data Cleaning	12
Education & Training	04	Exploratory Data Analysis	13
Skills	05	Feature Engineering	21
Tools	06	Pemodelan	23
Work Experience	07	Feature Importance	26
Previous Project	08	Target Impact	27
Business Understanding	10	Conclusion	28
Data Information	11	Recomendation	30

Contact

31

My Name Is

ZINEDINE AMALIA N.M.R

“Observing how data has become central to decision-making in many companies has further strengthened my interest in the field of Data Science. With a background in Mathematics, I am eager to deepen my skills in processing, analyzing, and visualizing data. The rapid advancement of technology serves not only as a challenge but also as a motivation to continuously learn and innovate, in order to transform data into valuable insights for diverse needs”

Education

Jenderal Soedirman University

Mathematics

- Relevant Coursework: Mathematics, Statistics, Prediction Analysis.
- Journal: Peramalan Curah Hujan di Cilacap Menggunakan Seasonal Autoregressive Integrated Moving Average (https://repository.unsoed.ac.id/23598/)

Training

Dibimbing.id

Data Analyst & Data Science Boothcamp

- Studied Data Preprocessing, Database Management (SQL Query and Web Scraping), and the development of interactive dashboards with Tableau and Power BI, focusing on data storytelling.
- Explored machine learning models to understand the role of data in business analysis and its application in the fields of Data Analytics and Data Science.

English Boothcamp

- Relevant Course: Workplace English Communication, Professional Writing, Job Interview Preparation, Business Presentation, Meeting Communication.

Data Analysis

- Experienced in exploring and analyzing data to uncover patterns and business insights.
- Proficient in descriptive statistics, inferential statistics, and hypothesis testing.

Data Processing

- Proficient in cleaning, processing, and handling large datasets using Python.
- Expertise in data manipulation techniques using Pandas and SQL

Data Modeling

- Skilled in building predictive models using Machine Learning techniques such as regression, Random Forest, and XGBoost.
- Familiar with handling imbalanced data and model optimization strategies.

GColab
Postgresql
Jupyter
Ms.Excel

Visualization

Tableau
PowerBI
matplotlib
Seaborn

Framework

pandas

Programming Languages

phyton
mysql

2020

Administrative Staff Intern
Naval Hospital dr. Mintohardjo

- Supported administrative, operational, and personnel data management for over 80 employees, including performance analysis using SKP and Minitab to generate accurate insights.
- Prepared clear and concise performance reports to support evaluation and data-driven decision-making while maintaining a positive and supportive work environment.

2021

Mentor for the Statistics Subject
Jenderal Soedirman University

- Facilitated students' academic development through differentiated instruction, tailored strategies, and innovative teaching methods, while collaborating with fellow instructors to design comprehensive course materials and ensure full syllabus coverage.
- Conducted 12 tutorial sessions for over 130 students in Probability Theory and Mathematical Statistics, including the design of assessments and fair performance evaluations.

2025

Data Science Intern
Studyfirst

- Structured and optimized internal databases to ensure high data quality and readiness for analytical modeling, improving efficiency in data analysis workflows.
- Automated and streamlined CRM data pipelines to reduce manual processing, increase data reliability, and support downstream analytics.
- Conducted competitor CRM benchmarking to identify best practices and data features, supporting CRM improvement and data-driven product decisions.
- Developed automated KPI dashboards to monitor operational performance, identify trends, and support data-driven decision-making.

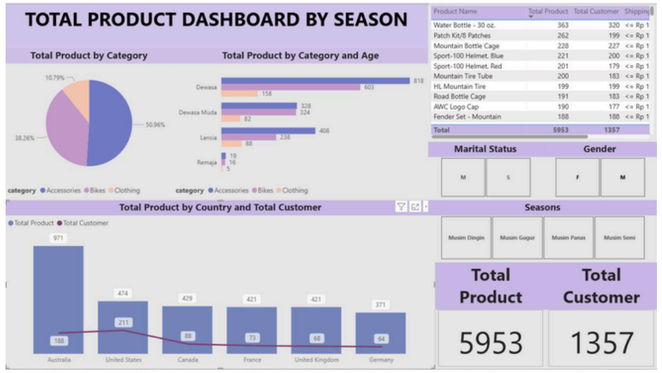
2026

English and Math Teacher
Eye Level Indonesia

- Conducted one-on-one coaching to strengthen literacy, numeracy, and problem-solving skills.
- Managed 5 students per week with personalized learning plans.
- Improved student mastery scores by 20% within 3 months through individualized learning plans and weekly assessments.

PREVIOUS PROJECTS

Sales Performance with Seasonal Insights

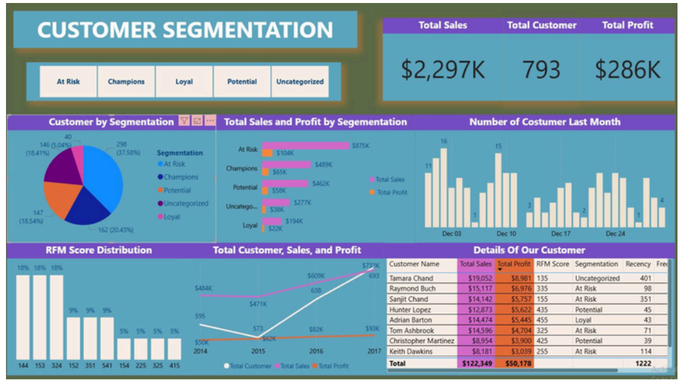


The goal of this project is to analyze the distribution and sales performance of products based on seasons. Therefore, it is essential to highlight key products that align with the four seasons: spring, summer, autumn, and winter.

- Approach: Seasonal Product & Customer Insight Analysis.
- Tools: Python and Power BI

Outcomes: With these seasonal insights, the marketing team can develop a more accurate target audiences, and key product priorities for each season.

Driving Business Growth with Customer Insights



The goal of this project is to evaluate customer segmentation in order to identify which segments contribute the most to sales and profitability, and to determine opportunities for retention, growth, and customer development.

- Approach: Analysis of customer behavior using RFM scoring and segmentation.
- Tools: Python, Power BI (dashboard visualization), Streamlit.

Outcomes: This project delivers strategic insights to improve customer retention, maximize sales growth, and support data driven decision making.

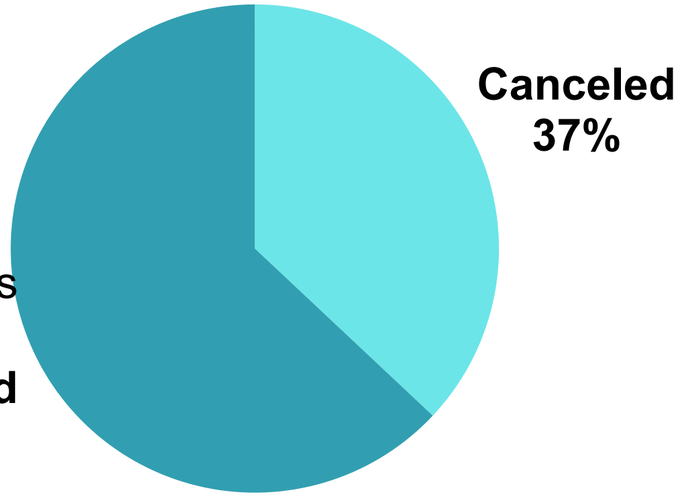
MAIN PROJECTS

Reducing Hotel Cancellation Loss with Risk-Based Targeting

BUSINESS UNDERSTANDING

Business Problem

The hotel is facing high uncertainty due to the increasing cancellation rate, which has **now exceeded 25%**.



This increase impacts revenue: from 2017 to 2018, revenue loss rose by approximately threefold.

2017	2018
$1,586,588	$5,027,374

Therefore, it is necessary to reduce cancellation-related losses by building a model that provides the probability of cancellation for each booking.

Objective

- Identify the main drivers behind the high cancellation rate.
- Build an accurate prediction model that can be used to mitigate cancellations.

DATA UNDERSTANDING

The dataset was sourced from Kaggle and contains 83,293 rows and 33 columns. It includes hotel booking records such as customer information, length of stay, booking status, pricing, and distribution channels. The objective of this analysis is to understand customer behavior and booking trends, with a focus on analyzing cancellations.

- Hotel Info: hotel
- Booking Status: **is_canceled**
- Booking Timing: **lead_time**
- Arrival Date: **arrival_date_year, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month**
- **Stay Details: stays_in_weekend_nights, stays_in_week_night**
- Guest Info: adults, children, babies, is_repeated_guest, previous_cancellations, previous_bookings_not_canceled, customer_typ
- Booking Details: meal
- Room Details: reserved_room_type, assigned_room_type
- Booking Changes: **booking_changes**
- Payment Info: deposit_type
- Agency Info: agent, company
- Waiting Info: days_in_waiting_list
- Customer Info: country, **market_segment,** distribution_channel
- Pricing: **adr**

83,293 Baris dan 33 Kolom

DATA CLEANING

Handle Incorrect dtypes & Anomaly

Incorrect dtpyes:

- Reservation_date converted datetime
- Agent menjadi category
- Company converted category

Anomaly:

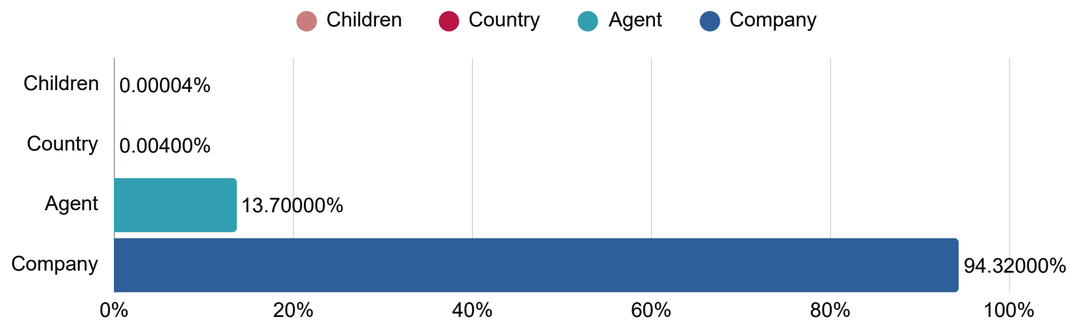
- Rows with total_guests == 0 were removed, as they are considered data anomalies.

Duplicate

None

Missing Values

There are missing values in the following columns:



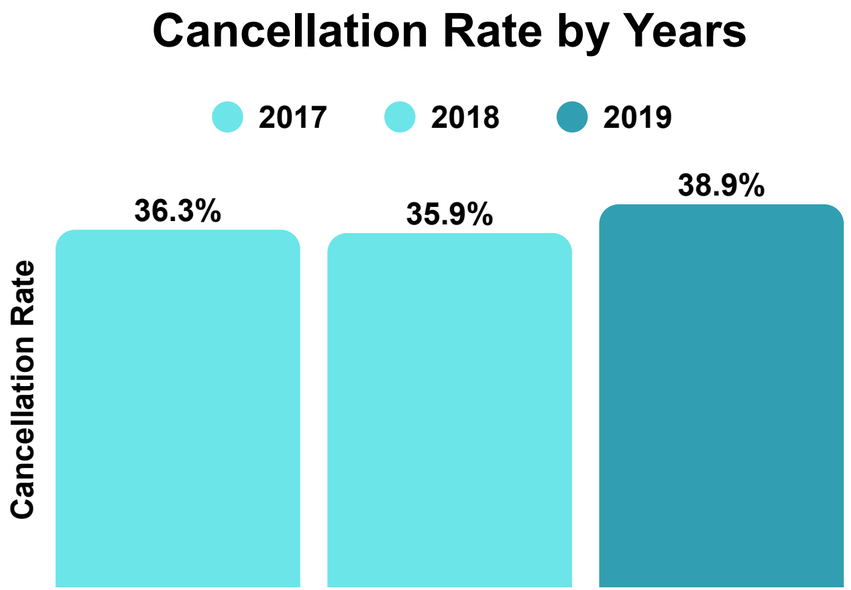
They were handled as follows:

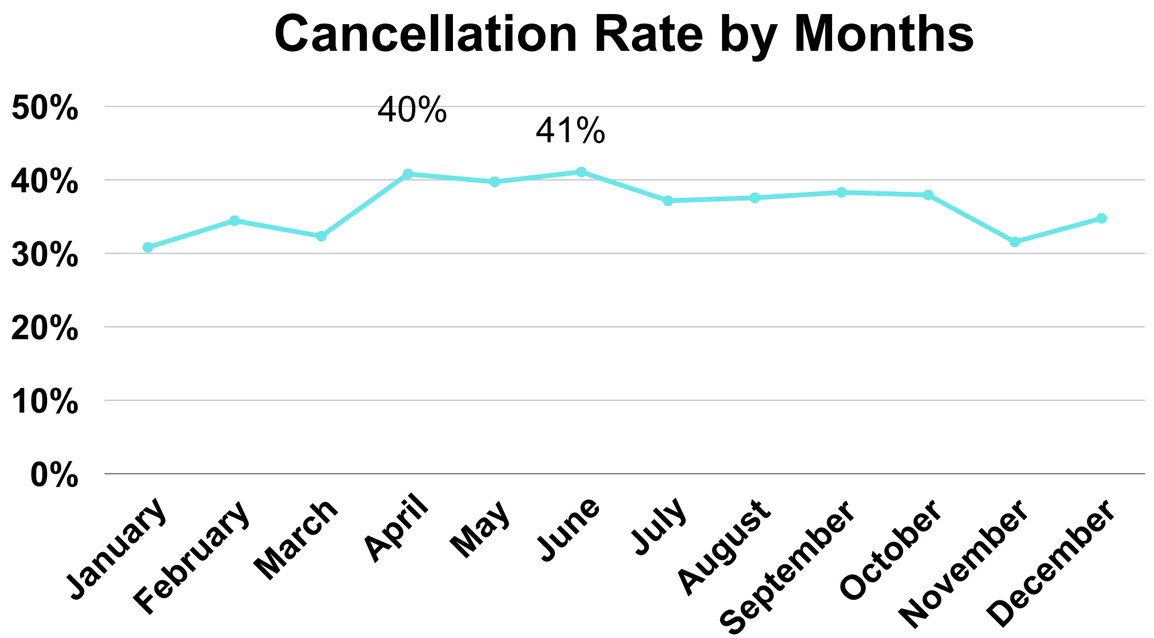
- children filled with 0
- country filled with "Unknown"
- agent replaced using a has_agent
- company replaced using a has_company

Outlier

Outliers were identified, but since the values are still within plausible ranges, no outlier treatment was applied (using the IQR method).

EXPLORATORY DATA ANALYSIS



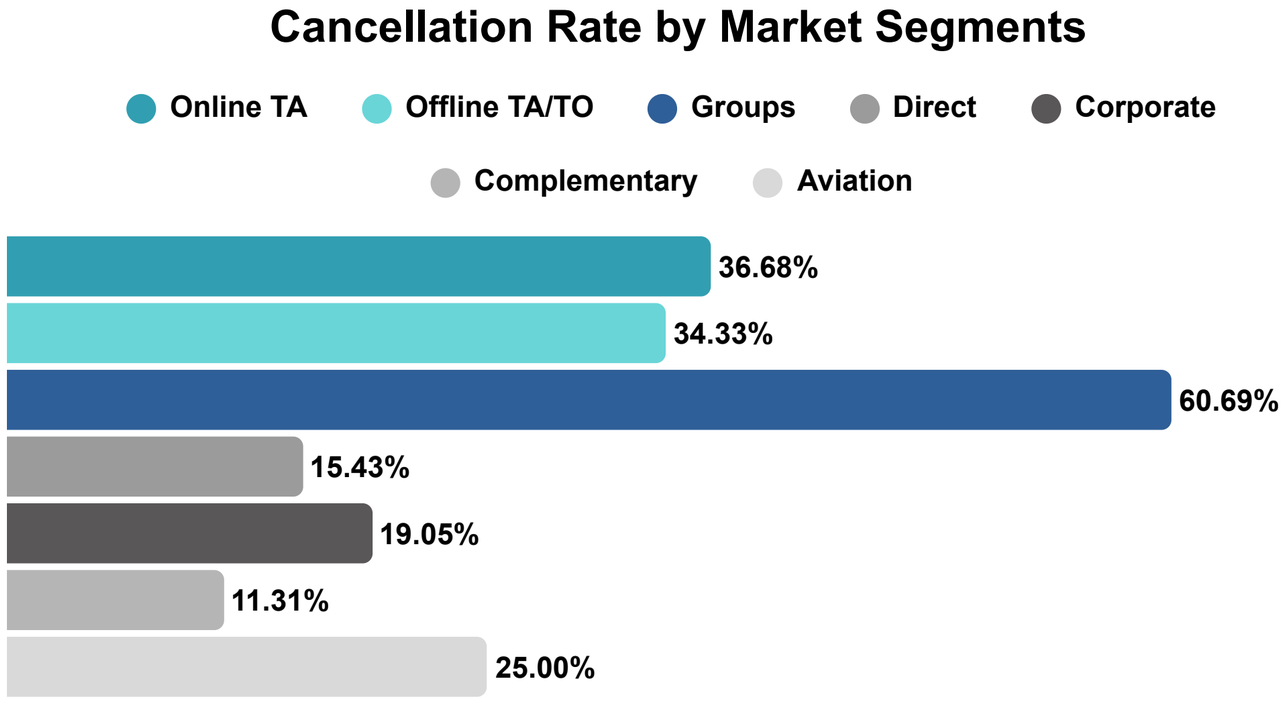


“Cancellations peak in mid-year, not evenly spread across months”

The highest cancellation pressure appears around April–June (~40–41%), suggesting seasonality effects and the need for targeted policies during these months.

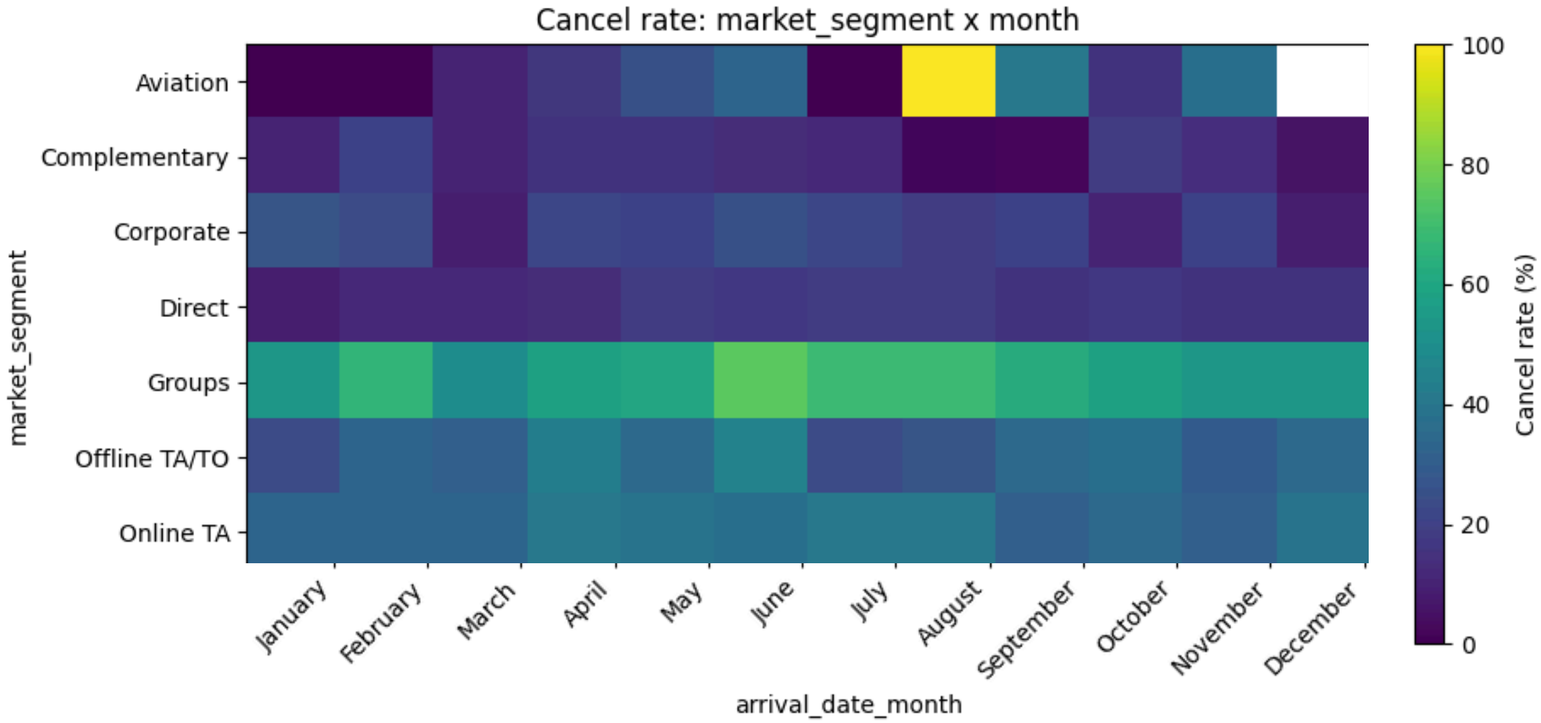
“Revenue loss is driven by high-volume segments, not only high cancellation rates”

Revenue Loss by Market Segments	
Market Segment	**Revenue Loss**
Online TA	$7,121,217
Groups	$1,948,497
Offline TA/TO	$1,731,548
Direct	$706,583



Online TA has a relatively high but still “reasonable” cancellation rate (~36.7%). However, because it has the largest volume, it contributes the biggest revenue loss (~$7.12M). Meanwhile, Groups has the most extreme cancellation rate (~60.7%) but a smaller total loss (~$1.95M) due to lower volume.

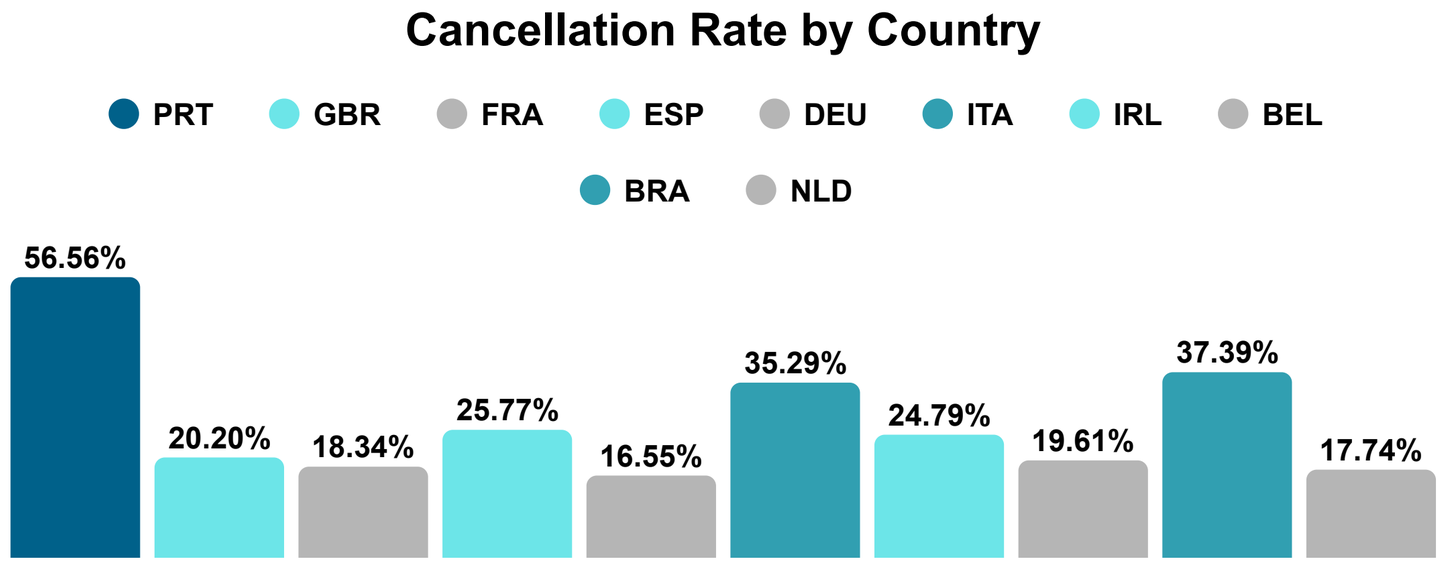
“Groups shows consistently elevated cancellations across most months”



The heatmap shows that each market segment has different high-risk months: Groups tends to remain high across many mon
(a consistently high pattern), while Online TA and Offline TA/TO are more moderate but steady throughout the year. This helps
explain why the overall monthly trend fluctuates—changes in segment mix by month affect the total cancellation rate.

“Portugal (PRT) is the dominant driver of cancellation loss—both highest rate and highest loss”

Portugal (PRT) has the highest **cancellation rate (56.56%)** and also contributes the largest revenue loss (~$5.99M), far exceedin other countries. This indicates that PRT is not only “frequently canceling,” but also “high impact” financially



Revenue Loss by Country	
Country	**Revenue Loss**
Portugal (PRT)	$5,987,648
Britania Raya (GBR)	$807,702
Spanyol (ESP)	$734,587
France (FRA)	$640,441

“A few hotels repeatedly appear as high-risk”

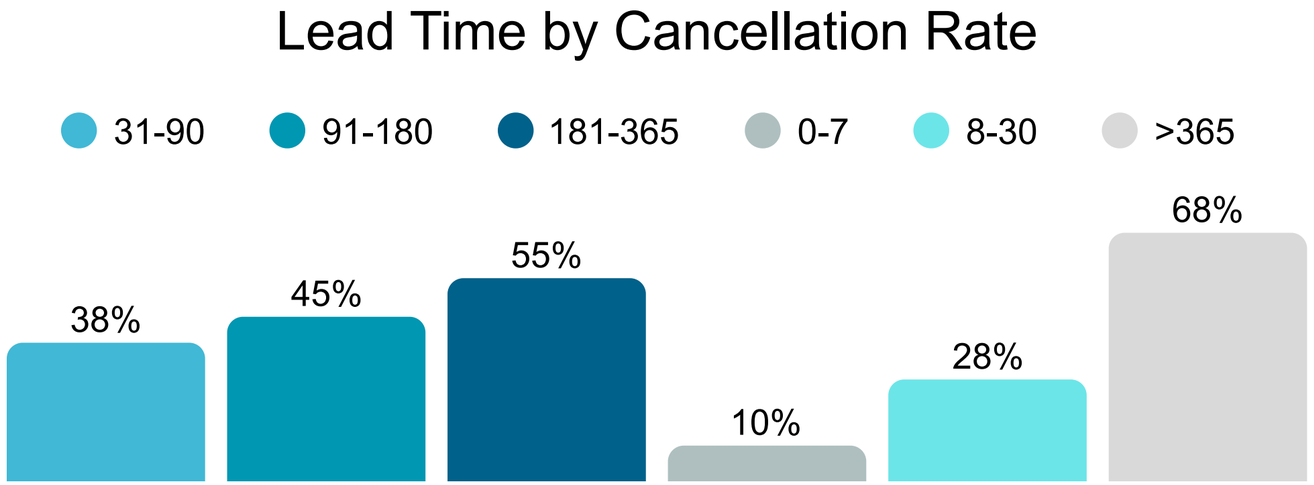
There is an important overlap: hotels such as **The Ritz-Carlton Berlin and Renaissance New York Times Square appear in both the high cancellation-rate and high revenue-loss contexts**, indicating a combination of “frequent cancellations” and “expensive impact” for certain properties.

Hotel Name	Cancellation Rate
Sheraton Lima Hotel & Convention Center Lima, Peru	38.91%
Renaissance New York Times Square Hotel New York, NY	37.61%
Heidelberg Marriott Hotel Heidelberg, Germany	37.57%
The Ritz-Carlton, Berlin Berlin, Germany	37.41%
Protea Hotel by Marriott Midrand Midrand, South Africa	36.27%

Hotel Name	Revenue Loss
Renaissance Santo Domingo Jaragua Hotel & Casino Santo Domingo, Dominican Republic	$206,451
Protea Hotel by Marriott Midrand Midrand, South Africa	$203,205
Sheraton Lima Hotel & Convention Center Lima, Peru	$199,248
Renaissance New York Times Square Hotel New York, NY	$195,208
The Ritz-Carlton, Berlin Berlin, Germany	$193,620

“Cancellation risk rises sharply with longer lead time, but revenue loss concentrates in mid-to-long”

Cancellation rate increases with lead time—from ~10% (0–7 days) up to ~68% (>365 days). However, the financial impact does not follow the same pattern—**revenue loss is concentrated in the 31–365 day** (≈$3.05M–$3.67M), not in the extreme >365-day group.



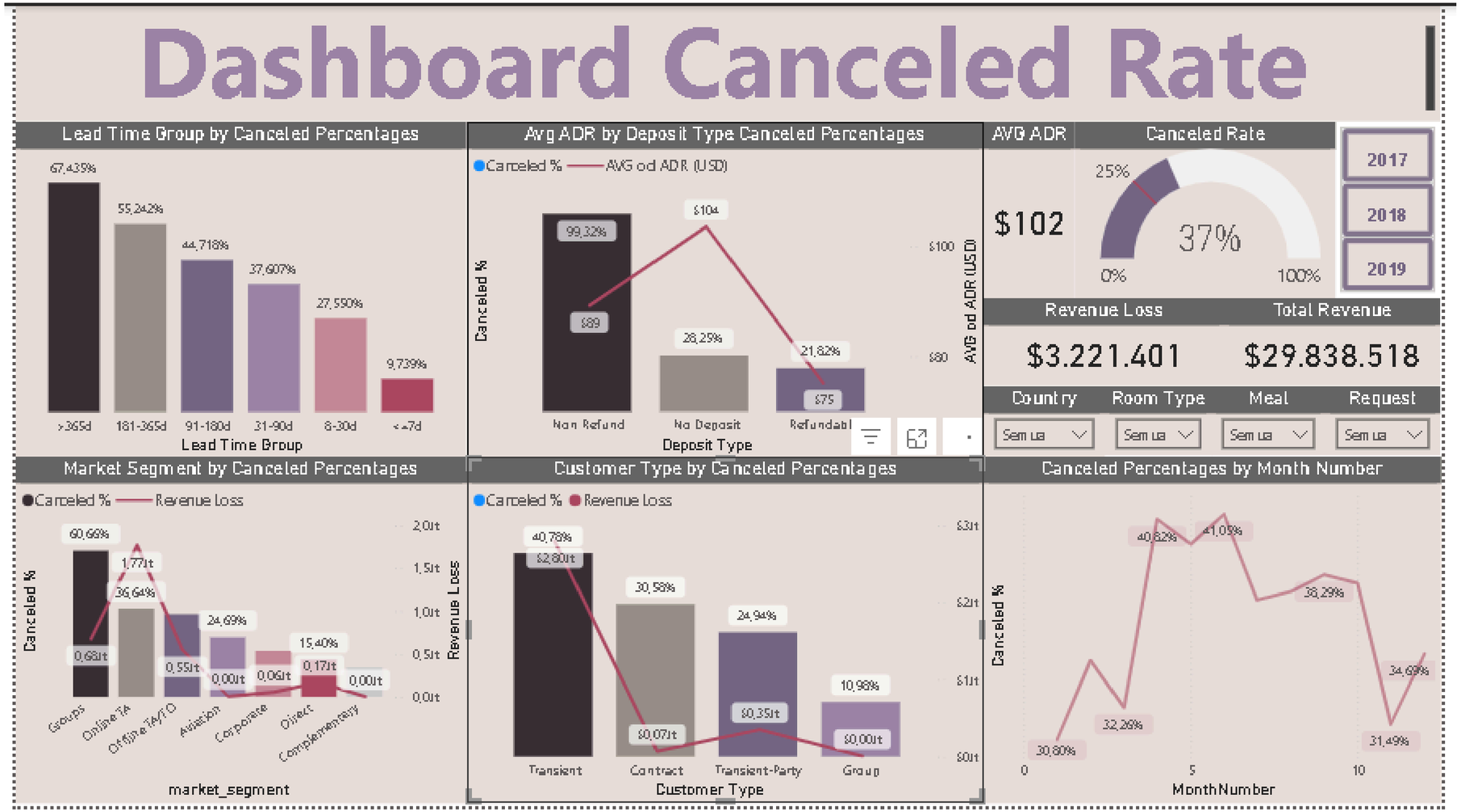
Lead Time	Revenue Loss
0-7	$312,863
8-30	$1,409,364
31-90	$3,050,827
91-180	$3,563,962
181-365	$3,066,877
>365	$253,074

“Across all four hotels, cancellations are segment-driven: Groups is consistently the highest”

Across the four hotels, average lead time is relatively similar (around ~100 days), yet Groups consistently has the highest cancellation rates (≈56%–70%), far above Corporate/Direct (≈12%–19%). This **suggests cancellation risk is driven more by segment** type than by hotel differences alone.

hotel	Cancellation Rate					AVG Lead Time
	Corporate	Direct	Groups	Offline TA/TO	Online TA	
Protea Hotel by Marriott Midrand Midrand, South Africa	12.50%	12.58%	59.91%	33.85%	37.03%	110
Renaissance New York Times Square Hotel New York, NY	15.62%	13.77%	70.22%	32.63%	36.21%	106
Sheraton Lima Hotel & Convention Center Lima, Peru	18.03%	17.89%	62.13%	33.82%	39.25%	105
The Ritz-Carlton, Berlin Berlin, Germany	19.30%	13.56%	55.95%	37.96%	39.15%	102

DASHBOARD


Dashboard Canceled Rate
Lead Time Group by Canceled Percentages
67,43%
55,24%
44,71%
37,60%
27,55%
9,73%
>365d
181-365d
91-180d
31-90d
8-30d
<=7d
Lead Time Group
Avg ADR by Deposit Type Canceled Percentages
Canceled %
AVG od ADR (USD)
99,32%
$89
$104
28,25%
21,82%
$75
$100
$80
Non Refund
No Deposit
Refundable
Deposit Type
AVG ADR
$102
Canceled Rate
25%
37%
0%
100%
2017
2018
2019
Revenue Loss
$3.221.401
Total Revenue
$29.838.518
Country
Room Type
Meal
Request
Sem us
Market Segment by Canceled Percentages
Canceled %
Revenue Loss
60,66%
1,77k
36,64%
24,69%
15,40%
0,68k
0,55k
0,00k
0,06k
0,17k
0,00k
2,0k
1,5k
1,0k
0,5k
0,0k
Groups
Online TA
Offline TA/TO
Aviation
Corporate
Direct
Complementary
market_segment
Customer Type by Canceled Percentages
Canceled %
Revenue Loss
40,78%
$2,80k
30,58%
$0,07k
24,94%
$0,35k
10,98%
$0,00k
$3k
$2k
$1k
$0k
Transient
Contract
Transient-Party
Group
Customer Type
Canceled Percentages by Month Number
Canceled %
30,80%
32,29%
40,62%
41,05%
38,29%
31,49%
34,69%
0
5
10
MonthNumber

FEATURE ENGINEERING

Split Data

80% Train, 20% Test

Scaller and Encoding

- Scaller dengan **Standar Scaller**
- Encoding dengan **One Hot Encoder**

Check Multicollinarity

Check VIF score

Multikolinearitas

Feature	VIF
reserved_room_type	13.54962
distribution_channel	13.166355
assigned_room_type	11.869413
market_segment	9.700938
country_grp	2.449124
arrival_date_year	1.780776
...	

drop coloumn that has highest VIF score:

- assigned_room_type
- distribution_channel

Feature	VIF
reserved_room_type	4.421894
market_segment	3.519526
country_grp	2.276439
arrival_date_year	1.762776
arrival_date_week_number	1.611162
adr	1.518227
...	

MODEL

Model	Split	Precision	Recall	F1-score
Logistic Regression	Train	0.72	0.76	0.74
	Test	0.72	0.76	0.74
Random Forest	Train	1	1	1
	Test	0.87	0.78	0.83
XGBoost	Train	0.76	0.85	0.8
	Test	0.76	**0.84**	0.8

“XGBoost is the most reliable model for cancellation prediction, balancing strong recall with minimal overfitting”

Compared to Logistic Regression (stable but weaker at capturing cancellations) and Random Forest (perfect train scores indicating overfitting), XGBoost achieves high test recall (0.84) and solid F1 (0.80) with a small train and test gap, making it the safest choice for risk-based targeting.

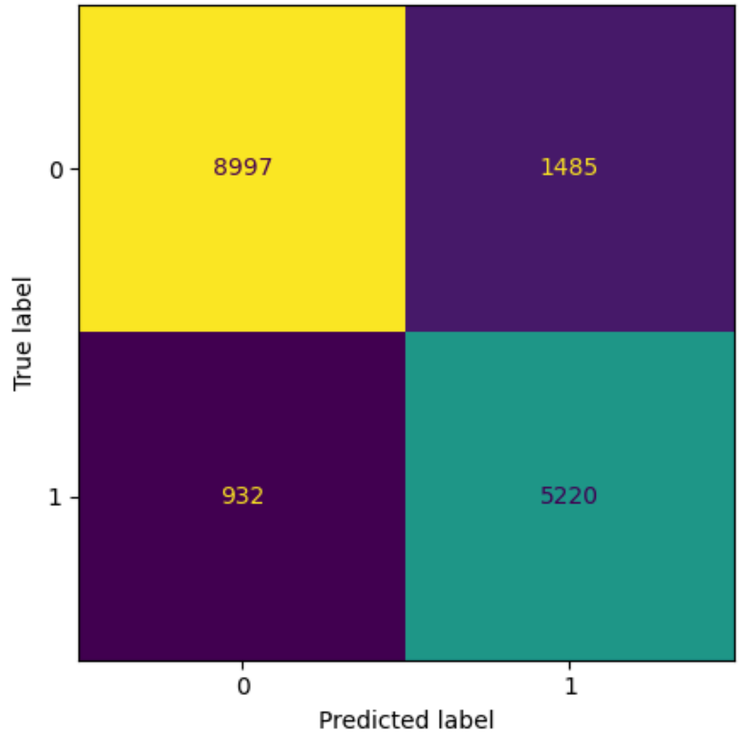
MODEL

Model	Split	Precision	Recall	F1-score	ROC-AUC	PR-AUC
XGBoost after Hypertuning	Train	0.83	0.9	0.86	0.96	0.95
	Test	0.78	**0.85**	0.85	0.94	0.91

“Tuned XGBoost achieves strong test performance with a small generalization gap, making it reliable for risk scoring”

On the test set, it reaches Recall 0.85 and F1 0.85 with ROC-AUC 0.94 and PR-AUC 0.91, while train vs test remains close (PR-AUC in train 0.95 and test: 0.91), indicating good discrimination without severe overfitting.

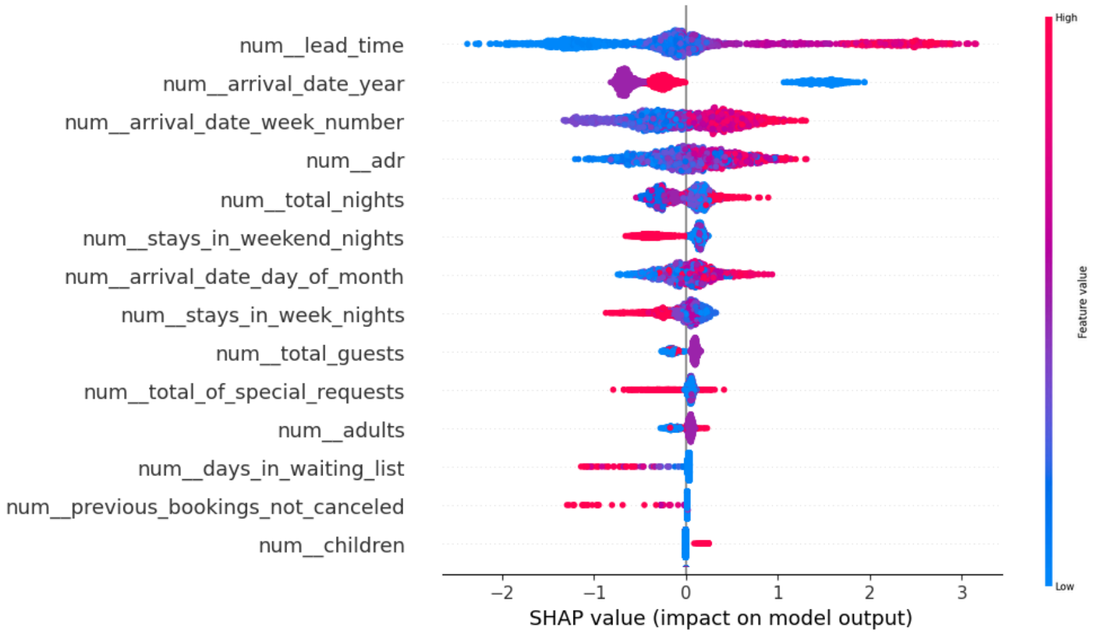
CONFUSION MATRIX



From the confusion matrix of the tuned XGBoost model:

- The **model correctly captures 5,220 cancellations** (TP: True Positives).
- It still **misses 932 cancellations** (FN: False Negatives).
- It **flags 1,485 bookings as cancellations even though they are not** (FP: False Positives).

FEATURE IMPORTANCE



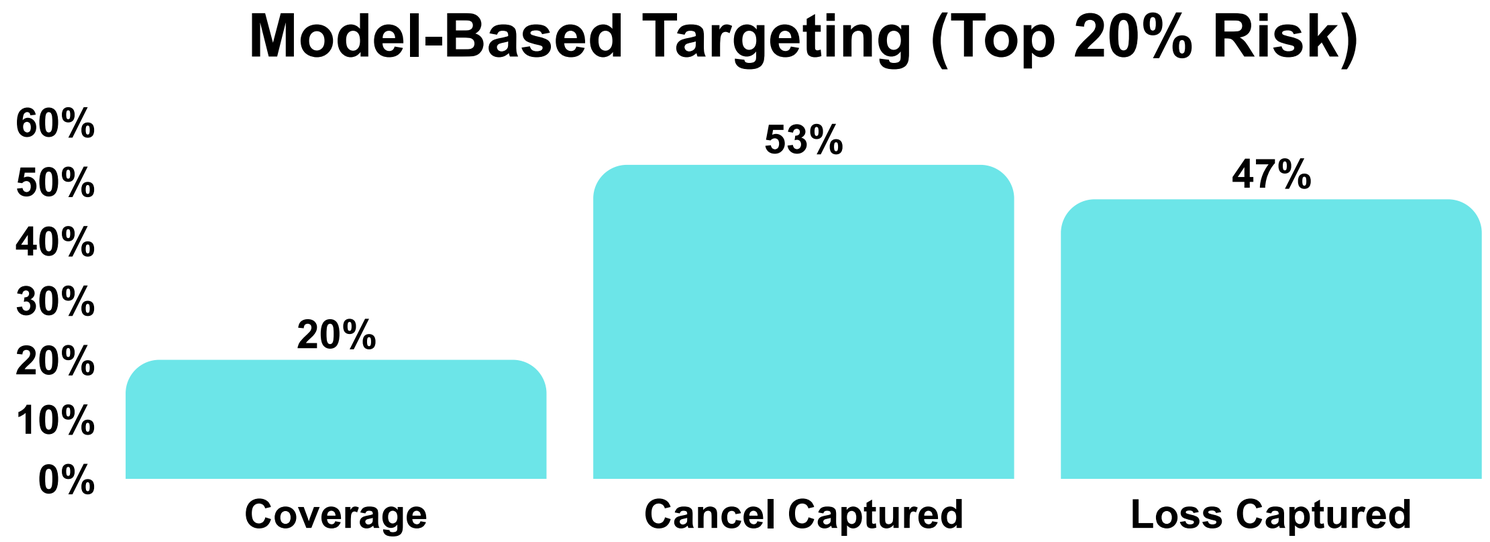
From the SHAP interpretation:

- **lead_time is the most dominant feature:** higher lead_time values (red points) tend to push predictions to the right, meaning a higher likelihood of cancellation.
- **arrival_date_year / week_number / day_of_month** also matter, **showing time/seasonality patterns** influence cancellation risk.
- **adr and total_nights** contribute as well—higher/lower values can shift the prediction, but their impact is below lead_time.
- Features such as special_requests, adults, children, previous bookings have smaller influence (most points are near zero).

TARGETING STRATEGY

If we focus on a small subset of bookings, how much impact can we capture?

- Total booking = 83,166, Top 20% = 16,465 bookings (out of 83,166)
- Risk score from tuned XGBoost with PR-AUC 0.91and ROC-AUC 0.94



Captured = share of total cancellations/loss that fall inside the Top 20% risk group (ranked by XGBoost probability).

IMPACT SIMULATION

TOTAL_LOSS = $11,656,969
LOSS_TOP20 = 47.230511% × TOTAL_LOSS ≈ $5.506M
saved_amount = reduction_% × LOSS_TOP20

Targeting	Reduction	Saved Amount	Saved of Total Loss	Saved of Total Cancel
0	10%	$550,564	4.72%	5.30%
1	20%	$1,101,129	9.45%	10.60%
2	30%	$1,651,693	14.17%	15.90%

Reduction = assumed policy effectiveness within Top 20% group.

The model selects the Top 20% most at-risk bookings, which contain 47.23% of total cancellation loss. If an intervention reduces loss within this group by 10–30%, the expected total savings are estimated at 4.72–14.17% of total loss ($0.55M–$1.65M).

CONCLUSION

- **Cancellations are persistently high** (≈36–39% yearly) with clear seasonality, **peaking around April–June** (~40–41%).
- Loss is highly concentrated: Online TA drives the largest loss due to volume, **Portugal/PRT dominates both cancel rate (56.56%) and loss (~$5.99M)**, and a few hotels repeatedly appear as high-impact hotspots.
- **Cancellation risk rises strongly with lead time** (up to ~55% for 181-365 days), but the largest revenue loss concentrates in 31–365 days (≈$3.05M–$3.67M per bin), not the extreme >365 group.
- **The tuned XGBoost model provides reliable risk scoring** (PR-AUC 0.91, ROC-AUC 0.94); SHAP highlights lead_time and arrival-time features as the main drivers of cancellation risk.
- **Model-based targeting is efficient: the Top 20% highest-risk bookings** capture ~52.77% of cancellations and ~47.23% of total loss, **implying estimated savings of ~4.72–14.17% ($0.55M–$1.65M)** under 10–30% policy effectiveness assumptions.

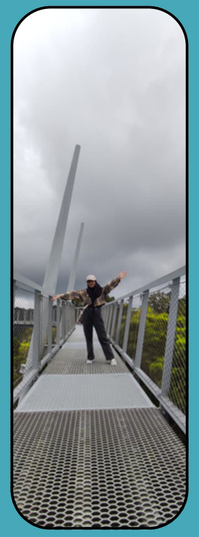
RECOMENDATION

- **Deploy model-based targeting**: Apply interventions only to Top 20% risk bookings (XGB score) with action auto-reconfirmation (T-14/T-7/T-3) with “one-click confirm” and reminder multi-channel..
- **PRT-specific tightening**: For Portugal (PRT) high-risk bookings, require earlier deposit and monitor the KPI.
- **Segment rules (different levers):**

1. Groups: contract cutoff and staged/non-refundable deposit.
2. Online TA / Offline TA/TO: keep policy light for low-risk, tighten only for high-risk.

- **Lead-time focus (31–365 days):** Prioritize reconfirmation and commitment nudges for 31–365 day lead time.

LET'S TOGETHER



WORK

GET IN TOUCH

+62 8577-4805-287

GitHub

zinedineamalianoor.com