

DATA SCIENCE

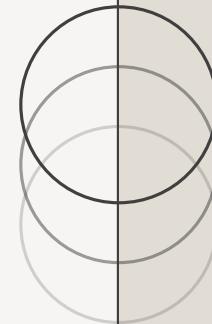
OCT 2025



# PREDICTION OF TICKET TRAVEL CANCELLATION

BY LAMZAHHERA BERINPALLA

PORTOFOLIO



# Lamzahhera Berinpalla

## EDUCATION

- Dibimbing
- Trisakti School of Management

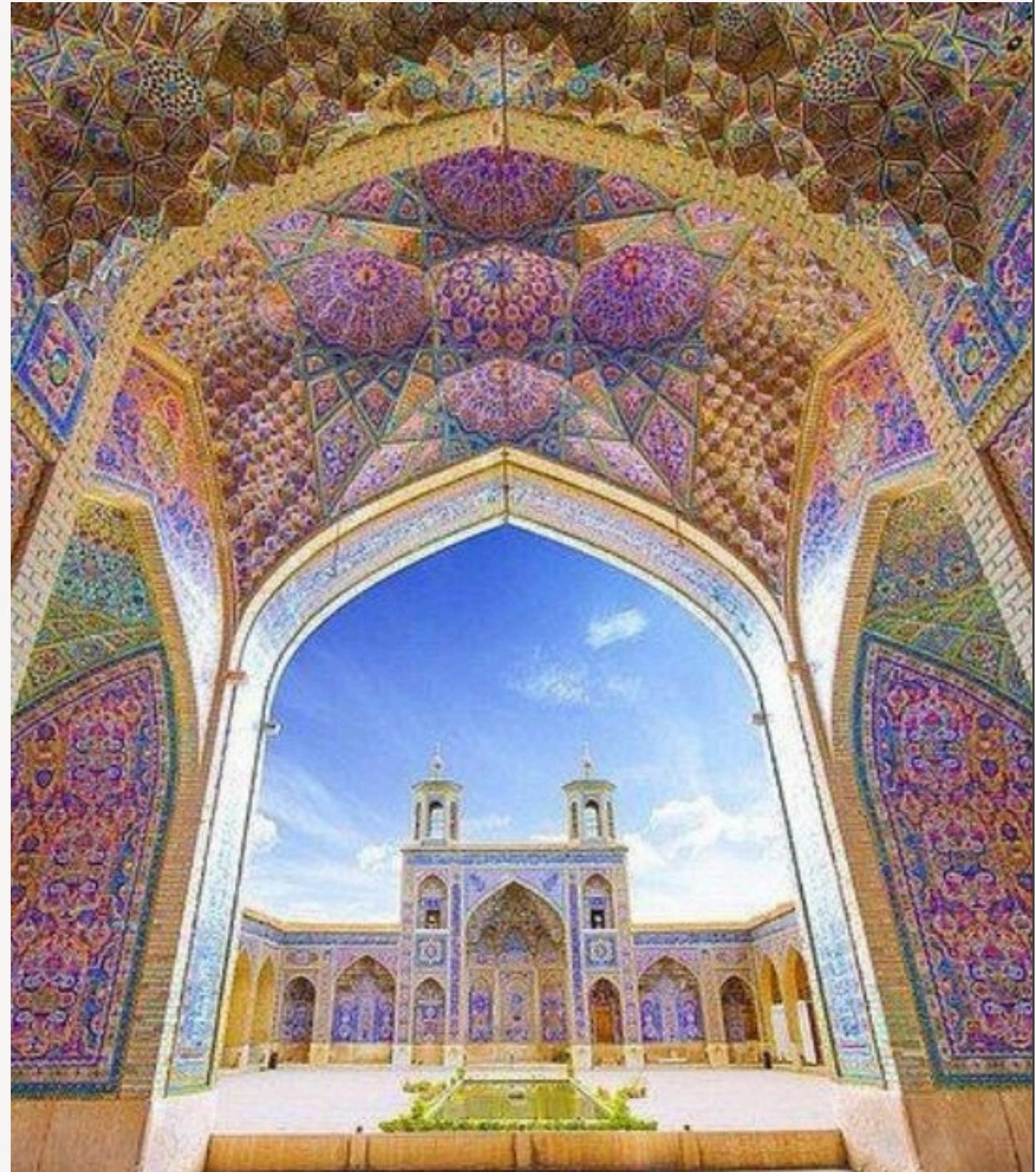
## WORK EXPERIENCE

- Ma Little Be
- EY Indonesia

# Other Project

## Data Analyst & Data Science

- Beecycle Customer Segmentation
- A/B Testing
- Customer Satisfaction
- People Analytics
- Marketing Channel Analysis
- Churn Analysis on Telco Company
- etc,



# SUMMARY

## Background:

Pembatalan tiket travel menyebabkan kerugian finansial dan operasional bagi agen perjalanan. Memahami pola pembatalan dapat membantu dalam mengambil tindakan pencegahan dan meningkatkan strategi pemasaran.

## Problem Statement:

Tingginya angka pembatalan tiket mengakibatkan pendapatan tidak optimal dan inefisiensi operasional.

## Goals & Objectives:

- Mengidentifikasi faktor-faktor yang memengaruhi pembatalan tiket.
- Membangun model prediktif untuk memperkirakan kemungkinan pembatalan tiket.
- 

## Key Insights:

- Tujuan perjalanan (purpose) dan durasi perjalanan memiliki pengaruh signifikan terhadap pembatalan.
- Pelanggan dengan riwayat pembatalan sebelumnya cenderung membatalkan lagi.
- Metode pembayaran dan jarak waktu pemesanan juga memengaruhi keputusan pembatalan.



# Business Understanding

## Problem

Agen travel kesulitan memprediksi pelanggan yang berisiko membatalkan tiket.

## Objective

Mengembangkan model untuk memprediksi probabilitas pembatalan tiket berdasarkan data pelanggan dan riwayat pemesanan.

## Goal

Memahami profil pelanggan dan pola pembatalan tiket.

## Metrics

Precision, Recall, F1-Score, ROC-AUC

# ABOUT DATASET

Dataset Ticket Travel Cancellation dari Kaggle memberikan informasi komprehensif tentang pemesanan travel dan pola pembatalan.

Raw  
Dataset

101.017 Rows  
22 Features



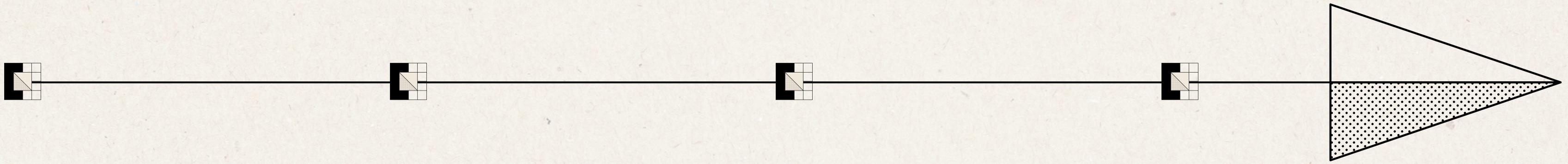
Cleaned  
Dataset

101.013 Rows  
16 Features

# Data Cleaning

for a better result we need to clean the data first

BY Lamzahhera Berinpalla



**Data  
Understanding**

**Missing  
Value**

**Duplicated  
Data**

**Outlier Data**

By Lamzahhera Berinpalla



# DATA INFO

## EDA

### 14 FEATURES

- Created
- DepartureTime
- TicketID
- ReserveStatus
- Male
- Price
- CouponDiscount
- From.
- To
- Domestic
- Vehicle
- TripReason
- PriceCategory
- CouponCategory

## Machine Learning

### 10 FEATURES

- ReserveStatus
- Male
- Price
- CouponDiscount
- From.
- To
- Domestic
- Vehicle
- TripReason
- Days\_Difference

## Target

- Cancel

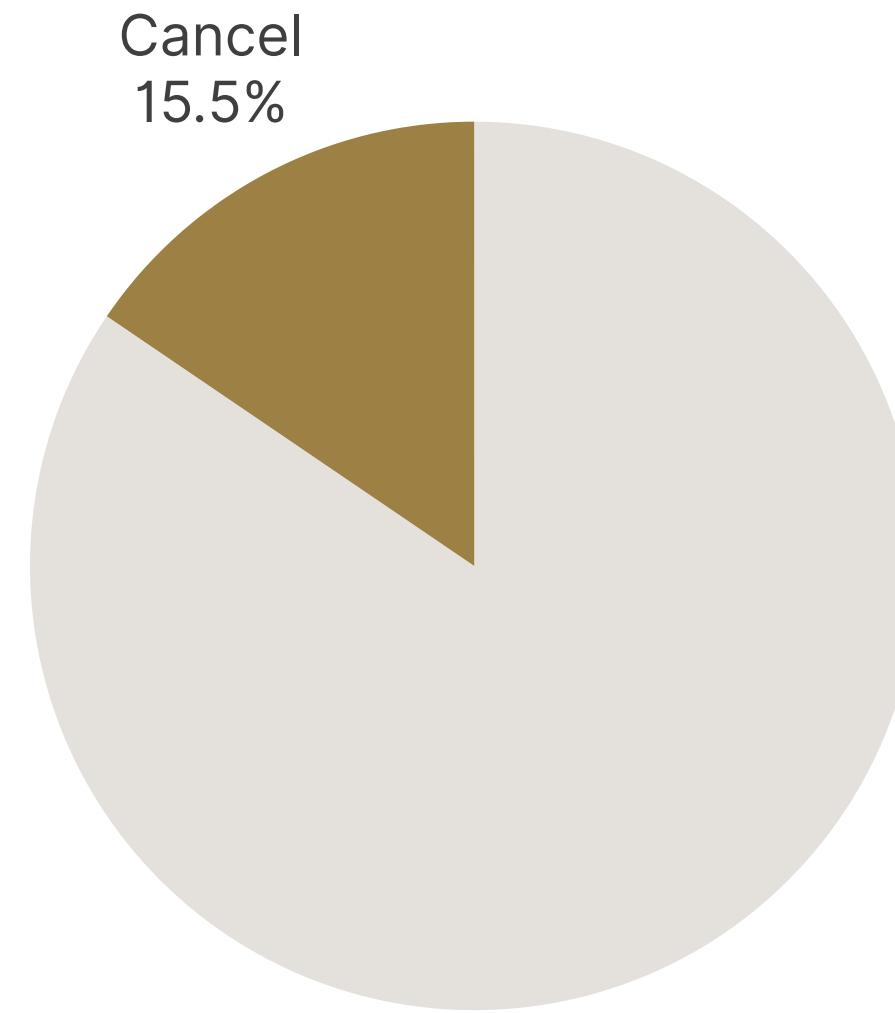
*strongly believe - strongly believe*



# Berapa tingkat pembatalan keseluruhan tiket travel?

oooooooooooo

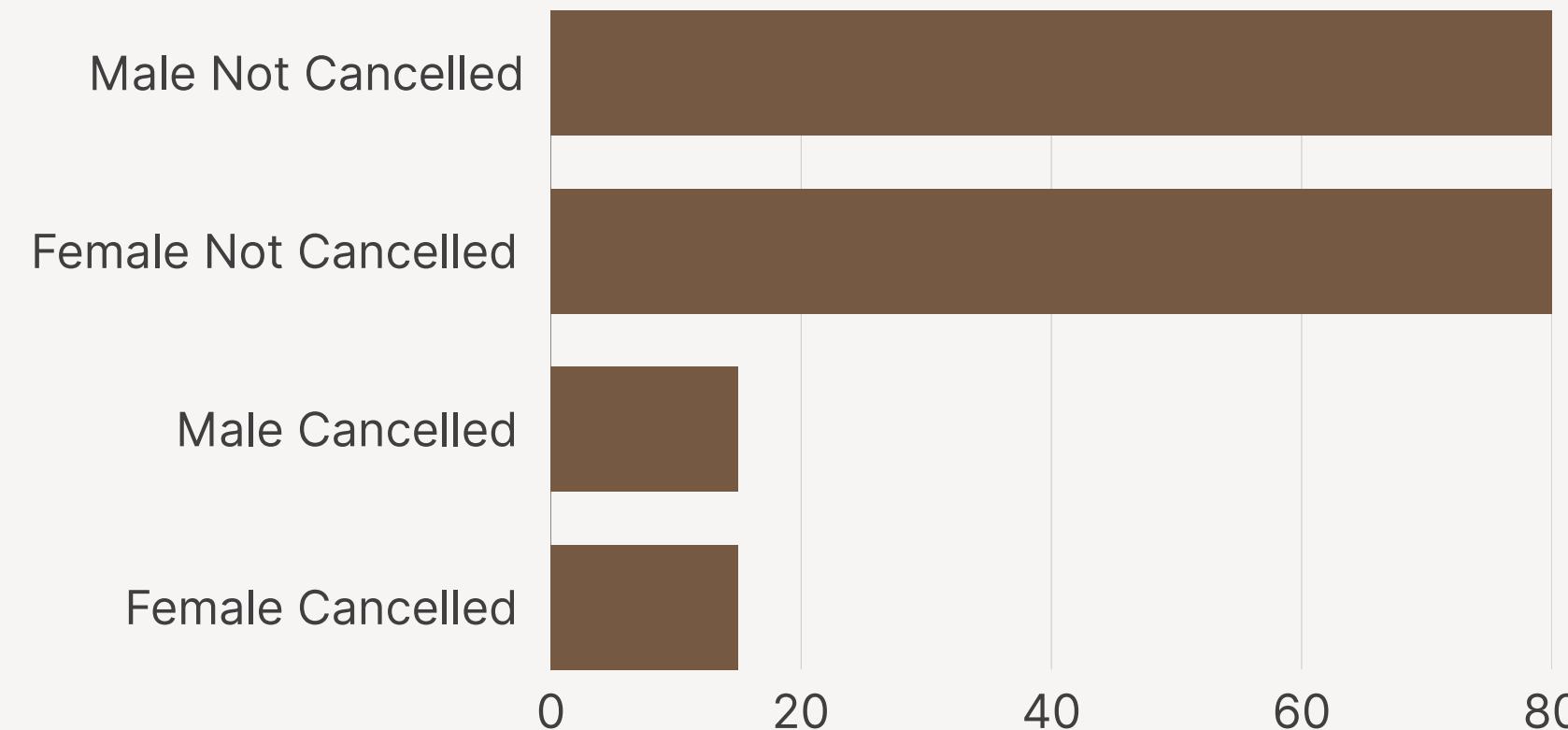
PROPORTION OF CANCELLED TRAVEL TICKET



Tingkat pembatalan tiket perjalanan sebesar 15.5% merupakan angka yang signifikan. Angka ini menunjukkan bahwa dari seluruh tiket yang dipesan, hampir 1 dari setiap 6 tiket akhirnya dibatalkan. Ini adalah metrik operasional dan finansial yang penting untuk ditelusuri lebih lanjut.

# Gender pelanggan mana yang menunjukkan risiko pembatalan tertinggi?

## CANCELLED TRAVEL TICKET BASED ON GENDER

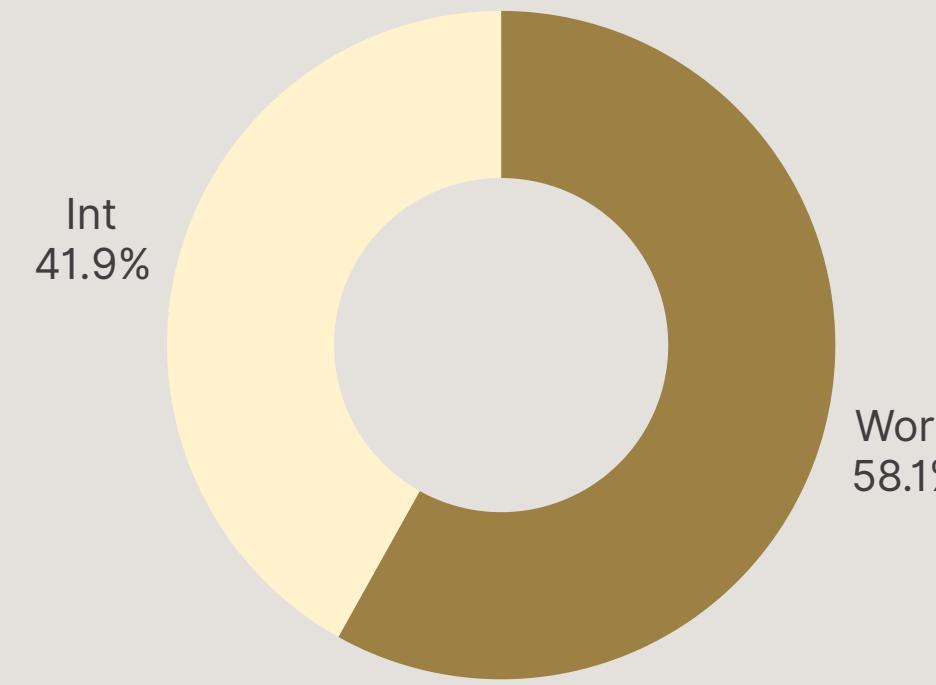


Gender tidak menunjukkan perbedaan dalam perilaku pembatalan. Pelanggan pria dan wanita membatalkan dengan tingkat yang hampir sama, menunjukkan bahwa strategi pemasaran yang spesifik gender mungkin tidak efektif untuk mengurangi pembatalan.

# Bagaimana tujuan travel memengaruhi perilaku pembatalan?

oooooooooo

TRIP REASON DISTRIBUTION



CANCELLATION BASED ON TRIP REASON

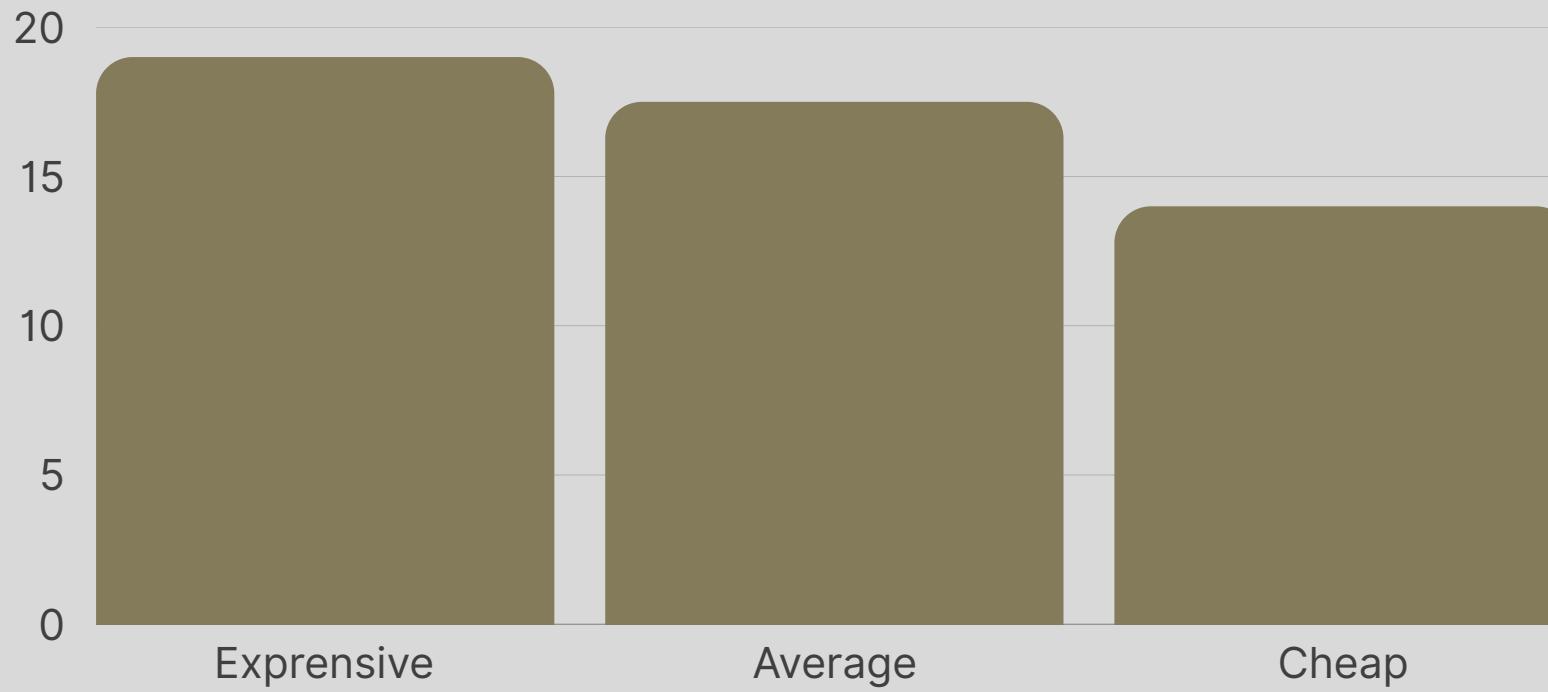


Trip bisnis (58,1%) menunjukkan distribusi yang jauh lebih tinggi dibandingkan trip intentional (41,9%). Namun, trip intentional (16%) memiliki peluang lebih tinggi untuk membatalkan perjalanan dibandingkan trip bisnis (17%). Hal ini menunjukkan bahwa perjalanan bisnis lebih terjadwal sehingga lebih sulit untuk melakukan modifikasi rencana di menit-menit terakhir.

# Apa peran harga dalam keputusan pembatalan?

oooooooooo

## CANCELLATION BASED ON PRICE GROUP



Segmen harga premium menunjukkan tingkat pembatalan tertinggi, yang menegaskan bahwa pelanggan yang membayar harga premium tidak lagi membuat mereka lebih serius dalam komitmen perjalanan mereka. Sebaliknya, hal ini justru membawa beban ekspektasi dan kompleksitas yang lebih besar.

# Bulan apa yang memiliki rate pembatalan tertinggi?

oooooooooooo

## CANCELLATION BASED ON SEASONAL



Pola musiman yang jelas muncul, dengan 25% pelanggan pada bulan Oktober (kemungkinan musim liburan atau periode puncak perjalanan) menunjukkan perilaku pembatalan berbeda yang dapat menginformasikan strategi inventaris dan harga.

# TOP 4 ROUTES WITH HIGHEST CANCELLATION

BANDAR ABBAS - FASA



BEHABAD - KERMAN



BEHABAD - MASHAD



BAJISTAN - TEHRAN



# PRE-PROCESSING DATA FOR MACHINE LEARNING MODEL

## Data Cleaning

handling duplicated  
and missing data

## Feature Selection & Engineering

encoding  
categorical  
variables & Checking  
multicollinearity

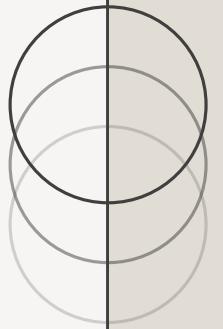
## Train-Test Split

80% Train: 20% Test

## Outlier Handling

IQR Method

# Model Selection



00000000

MODEL	DATA	METRICS			
		Precision	Recall	F1-Score	Roc-Auc
<b>Decission Tree</b>	<b>Train</b>	99.84%	97.25%	98.53%	99.98%
	<b>Test</b>	93.20%	91.84%	92.52%	95.76%
<b>Random Forest</b>	<b>Train</b>	99.54%	97.55%	98.53%	99.98%
	<b>Test</b>	99.14%	89.80%	94.24%	99.42%
<b>XGBoost</b>	<b>Train</b>	99.86%	91.03%	95.24%	99.81%
	<b>Test</b>	99.10%	89.45%	94.03%	99.53%
<b>LightGBM</b>	<b>Train</b>	99.93%	90.02%	94.71%	99.79
	<b>Test</b>	99.74%	89.13%	94.13%	99.52%

Meskipun XGBoost memiliki Skor F1 tertinggi kedua pada set pengujian (94%) dengan tetap mempertahankan skor ROC AUC yang tinggi (99%). XGBoost memiliki selisih terkecil antara proses Train dan Test dibandingkan model lainnya. Hal ini menunjukkan keseimbangan presisi-recall yang kuat dan kemampuan klasifikasi keseluruhan yang sangat baik.

# Optimizing XGBoost Parameters



oooooooooooo

Model	Metrics			
	Precision	Recall	F1-Score	Roc-Auc
Base XGBoost	99.01%	89.45%	94.03%	99.52%
Tuned XGBoost (Without SMOTE)	99.60%	89.29%	94.16%	99.53%
Tuned XGBoost With SMOTE	56.99%	100%	72.60%	92.58%

**Metrik yang diimprove:** XGBoost yang Dituned mengungguli based model di metrik Precision, F1-score, dan Roc-Auc.

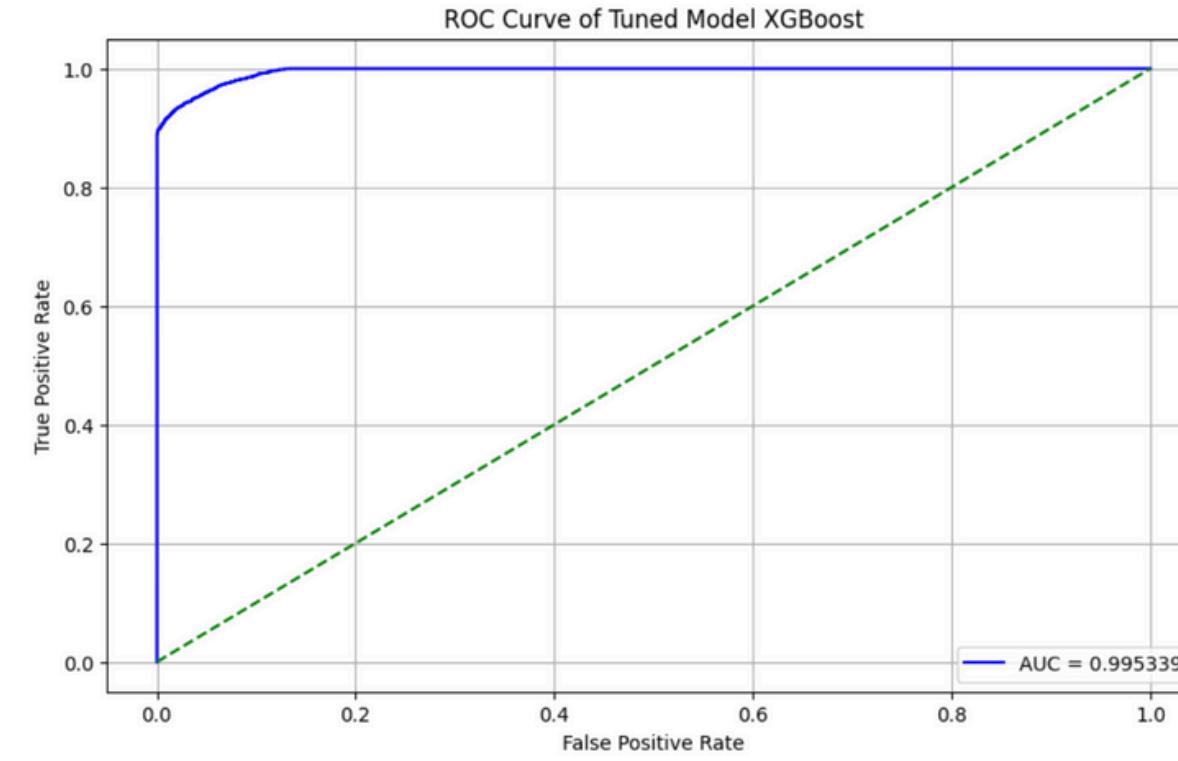
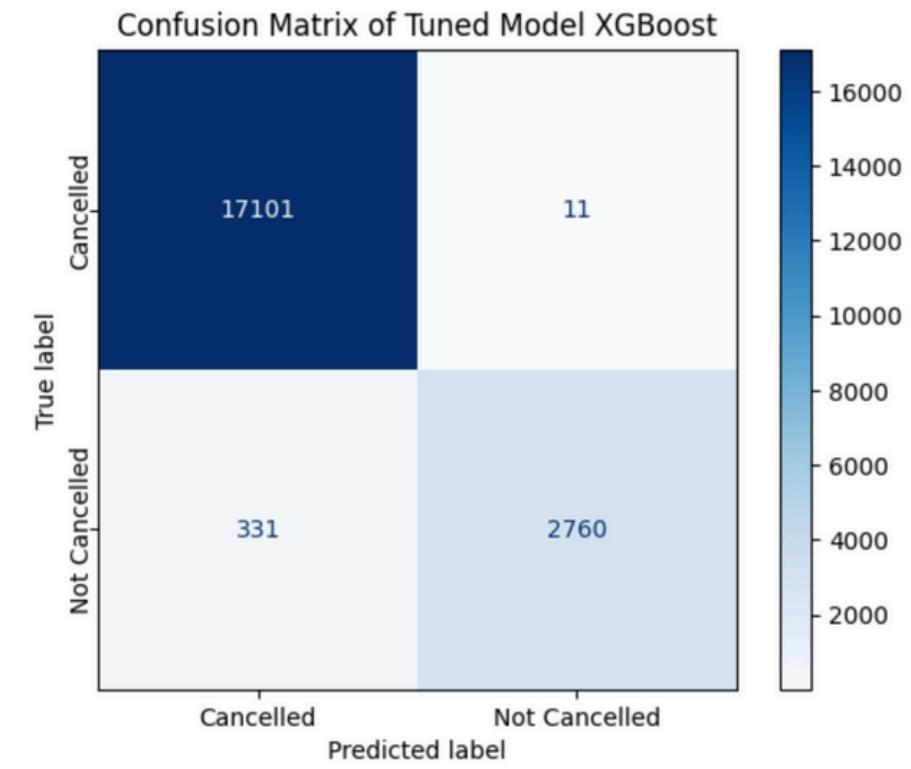
**Dampak SMOTE:** SMOTE kemungkinan menimbulkan pengurangan pada performa model.

**Model Terbaik:** Model XGBoost yang Dituned (tanpa SMOTE) memberikan kinerja tertinggi dan paling seimbang



PREPARED BY LAMZAHHERA

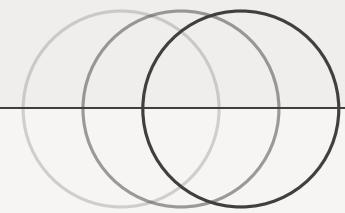
# MODEL ANALYSIS



- **True Positive (TP):** 2760 customers were correctly identified
- **False Positives (FP):** 11 customers were predicted to cancelled but turn out not cancel
- **True Negatives (TN):** 17101 customers were correctly predicted to cancelled
- **False Negatives (FN):** 331 customers were predicted to cancelled but actually did not cancelled

Model tersebut memiliki AUC sebesar 99,53% yang menunjukkan bahwa model tersebut secara efektif membedakan antara pelanggan yang cenderung membatalkan dan mereka yang tidak.

# Recommendations



## Manajemen Harga Dinamis:

Untuk rute dengan risiko pembatalan tinggi, pertimbangkan untuk menaikkan harga sedikit untuk mengkompensasi risiko, atau menawarkan harga non-refundable yang lebih murah.

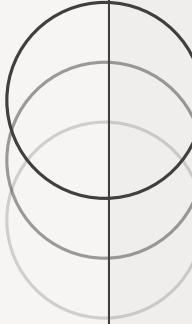
## Optimalkan Kebijakan Pembatalan:

Berikan kebijakan pembatalan yang lebih ketat untuk pelanggan yang diprediksi memiliki risiko pembatalan tinggi, misalnya dengan mengurangi refund atau menerapkan biaya pembatalan yang lebih tinggi.

Untuk pelanggan dengan risiko rendah, tawarkan kebijakan pembatalan yang fleksibel sebagai nilai tambah.

## Pengembangan Program Loyalitas:

Berikan reward kepada pelanggan yang setia (tidak pernah membatalkan) untuk meningkatkan retensi.



**THANKYOU**