

[Home](#)[Analytics](#)[Conclusion](#)

# Hotel Booking Cancellation Prediction

Leveraging the power of Machine Learning to accurately predict hotel booking cancellations, enabling improved operational efficiency and maximising revenue in the hospitality industry







Home

Analytics

Conclusion

# Business Overview







## Business Background

- Hotel booking cancellations lead to **lost revenue**, **operational inefficiency**, and **poor demand forecasting**.
- **Predictive analytics** can help forecast which **bookings** are likely to be canceled.
- **Enables proactive strategies:** overbooking, dynamic pricing, and targeted guest engagement.

## ! Problem Statement

- **High cancellation rates** reduce occupancy and revenue.
- **Unpredictable cancellations** make forecasting difficult.
- No targeted approach for **high-risk bookings** without prediction.

[Home](#)[Analytics](#)[Conclusion](#)

## Risk Cost

- **False Positive (FP):**
  - Predicts cancellation, but guest shows up.
  - Potential overbooking, guest dissatisfaction, **\$100 cost per booking**.
- **False Negative (FN):**
  - Predicts guest will come, but booking is canceled.
  - Empty room, direct revenue loss, **\$500 cost per booking**.

## Goals Statement

- Focus on recall by **optimising F2-score** to reduce financial loss from unexpected cancellations
- **Minimise False Negative** cases to avoid empty rooms and revenue losses.
- Support smarter **business recommendations** as booking risk management.



# Dataset Overview



## 1. Guest & Booking Information

- country: Guest's country of origin
- customer\_type: Booking type (Transient, Contract, Group, Transient-Party)
- market\_segment: Booking channel (Online TA, Offline TA/TO, Direct, etc.)

## 2. Booking Behaviour & History

- previous\_cancellations: Number of past cancellations by the guest
- booking\_changes: Number of changes made to the booking

## 3. Reservation Details

- deposit\_type: Deposit guarantee type (No Deposit, Non Refund, Refundable)
- reserved\_room\_type: Reserved room category (anonymized code)
- days\_in\_waiting\_list: Days spent on the waiting list before confirmation/cancellation
- required\_car\_parking\_spaces: Number of parking spaces requested
- total\_of\_special\_requests: Number of special requests (e.g., twin beds, high floor)

## 4. Target Variable

- is\_canceled: Indicates if the booking was canceled (1) or not (0)





Home

Analytics

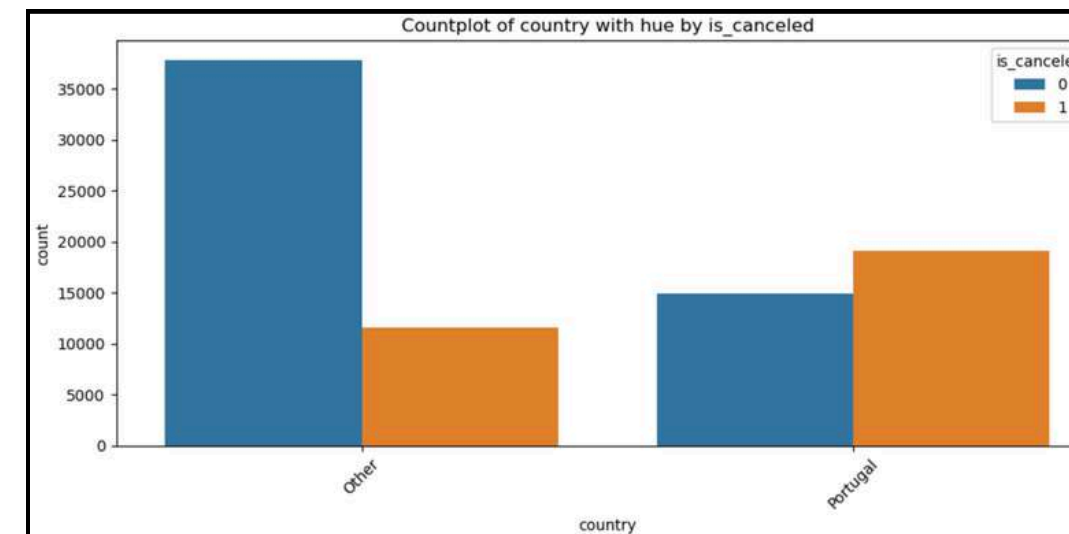
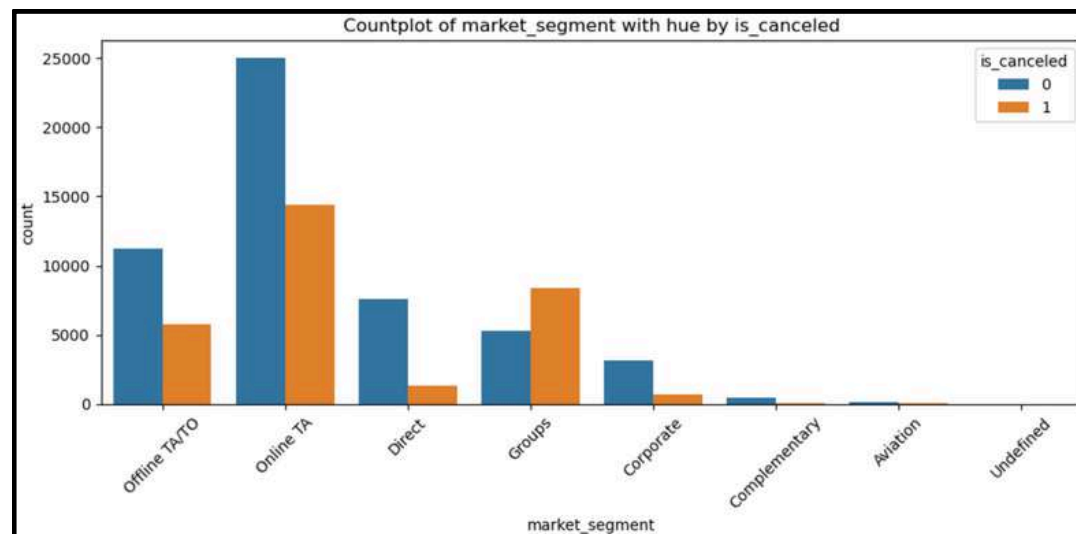
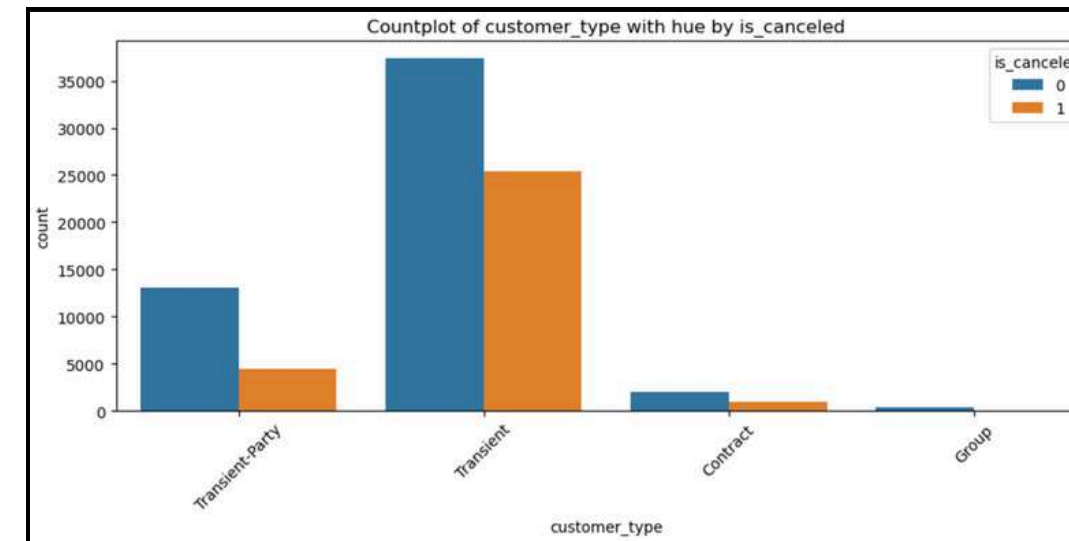
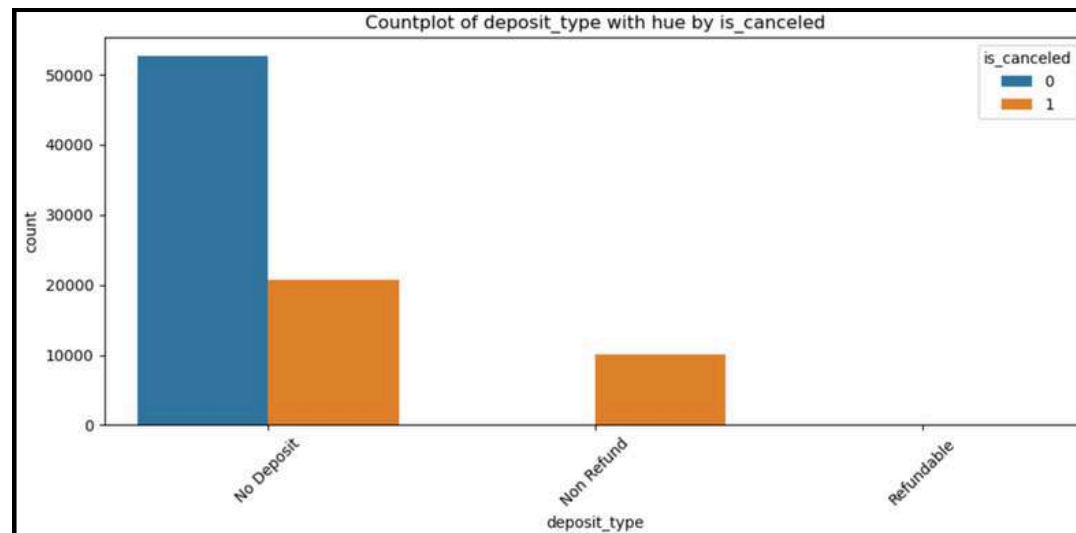
Conclusion

# Exploratory Data Analysis

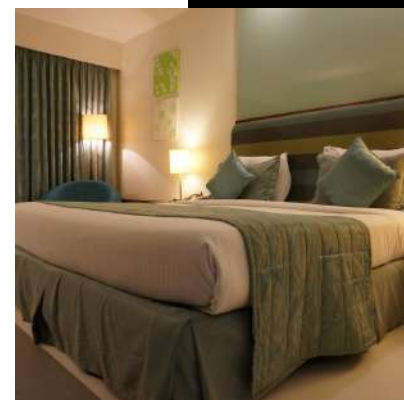
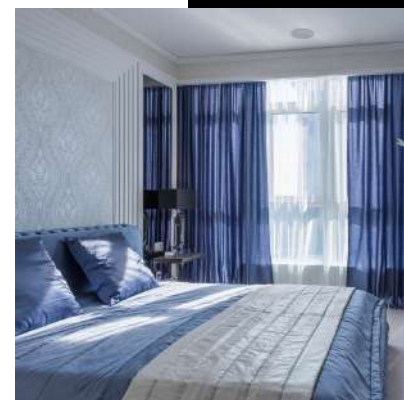




# Exploratory Categorical Data



- Deposit Type:
  - Most cancellations occur with "Non Refund" deposit type, showing policy impact on guest behavior.
- Market Segment:
  - "Online TA" and "Groups" segments have the highest cancellation rates.
- Customer Type:
  - "Transient" customers cancel more often than other customer types.
- Country:
  - More booking cancelations come from local guests (Portugal) than from other countries

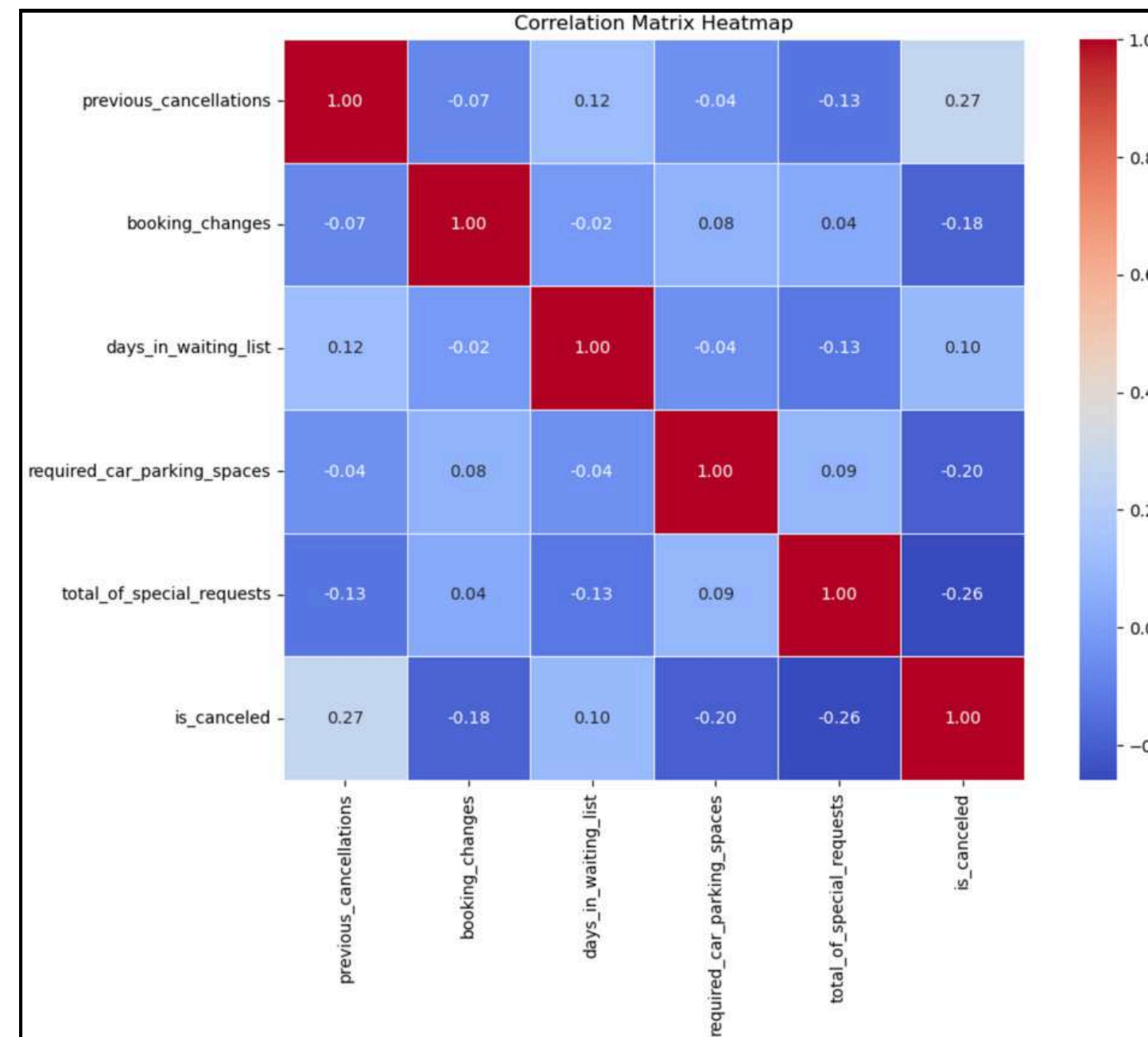






# Exploratory Numerical Data

- Previous cancellations show a positive correlation (0.27) with cancellations, meaning guests with a history of cancellations are more likely to cancel again.
- Booking changes are negatively correlated (-0.18) with cancellations, suggesting that bookings with more modifications are less likely to be canceled.
- Required car parking spaces (-0.2) and total special requests (-0.26) both have negative correlations with cancellations, indicating that guests with more specific needs tend to keep their bookings.
- Days in waiting list has only a weak positive correlation (0.1) with cancellations, so its impact is minimal.
- Overall, correlations between features are low, meaning each feature provides unique information for predicting cancellations.







Home

Analytics

Conclusion

# Model Benchmarking







# Data Preprocessing



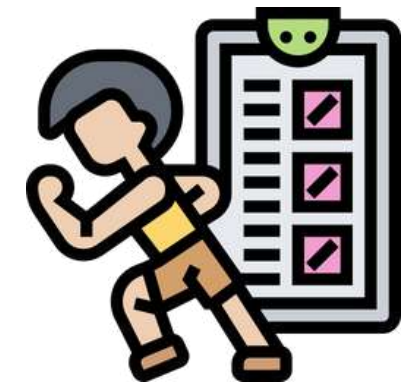
Feature engineering  
1. Grouping country data by 2 groups: Portugal and Other  
2. Onehot and Ordinal Encoding



Scaling numerical data with RobustScaling method



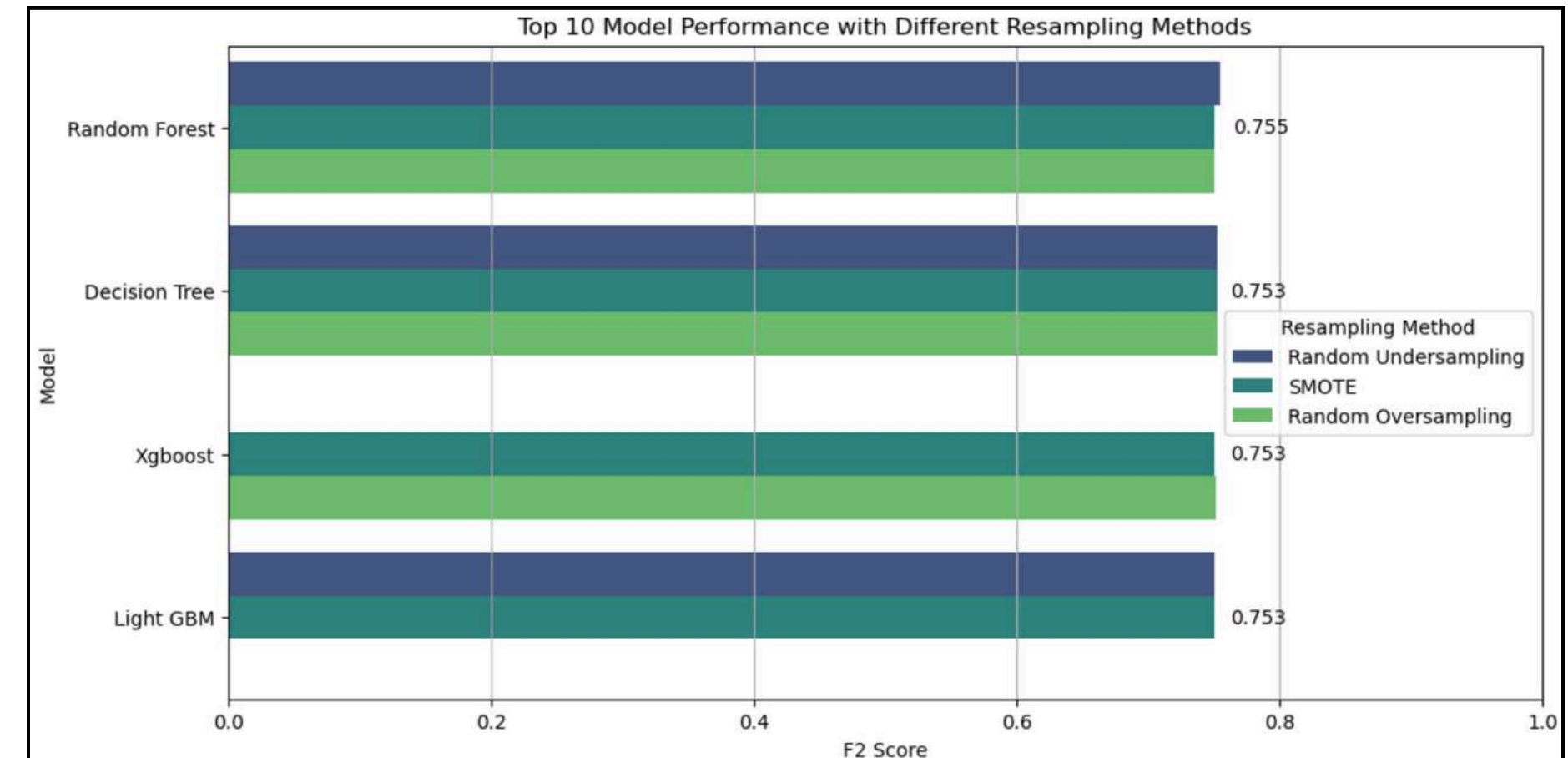
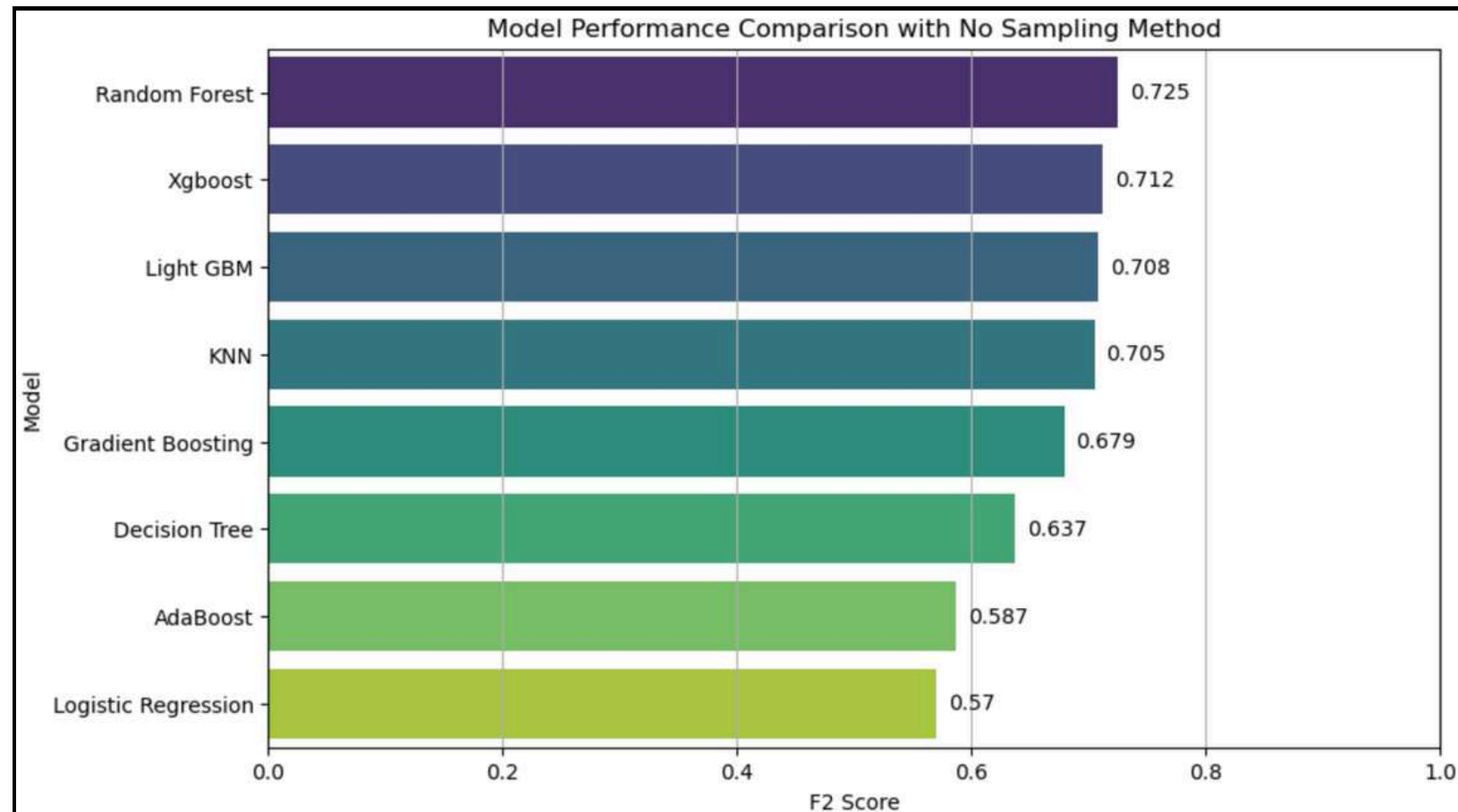
Define Cancellation Status is the Target, and other fields as the features



Train-Test Split with proportion 70:30



# Model Benchmarking



1. Random Forest improved from 0.725 to 0.755 with Random Under Sampling is the highest performance.
2. Decision Tree improved from 0.637 to 0.753 with Random Under Sampling.
3. Light GBM improved from 0.708 to 0.753 with Random Under Sampling.
4. Xgboost improved from 0.712 to 0.753 with SMOTE.
5. All models improved with resampling methods compared to no sampling.



# Best Model for Hyper-parameter Tuning



🧠 Random Forest is a machine learning algorithm based on an ensemble of decision trees, where multiple trees are trained on different subsets of the data and their predictions are combined to improve overall accuracy and reduce overfitting.



⚙️ Its key characteristics include the ability to handle non-linear data, robustness to noise and overfitting due to ensemble averaging, and automatic handling of missing values. It is well-suited for complex datasets and delivers stable and reliable predictions.



📊 Random Forest excels in performance due to its high accuracy, flexibility in modeling various types of data, strong generalization capability, and efficiency in handling large datasets with minimal preprocessing.



Home

Analytics

Conclusion

# Hyperparameter Tuning







# Hyper-parameter Tuning

- During training, SMOTE achieved a higher F2 Score (0.787) compared to Random Under Sampling(0.781), indicating better handling of class imbalance through synthetic oversampling.

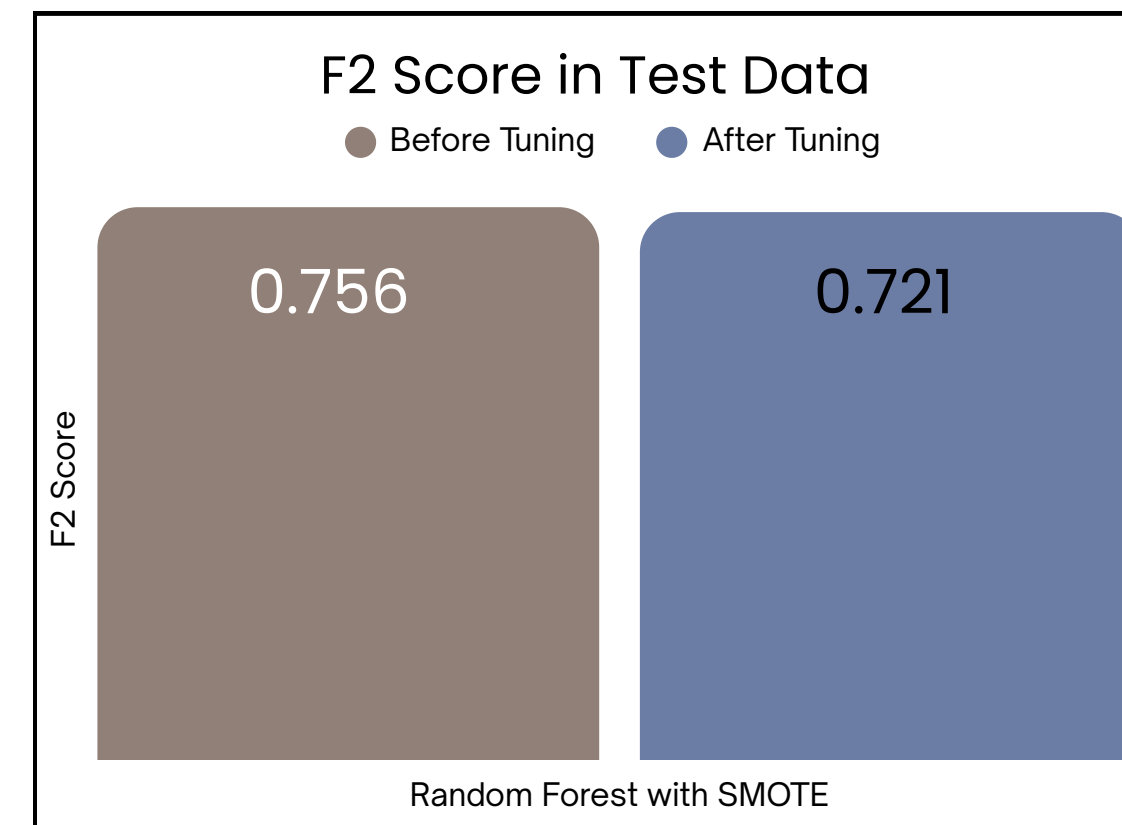
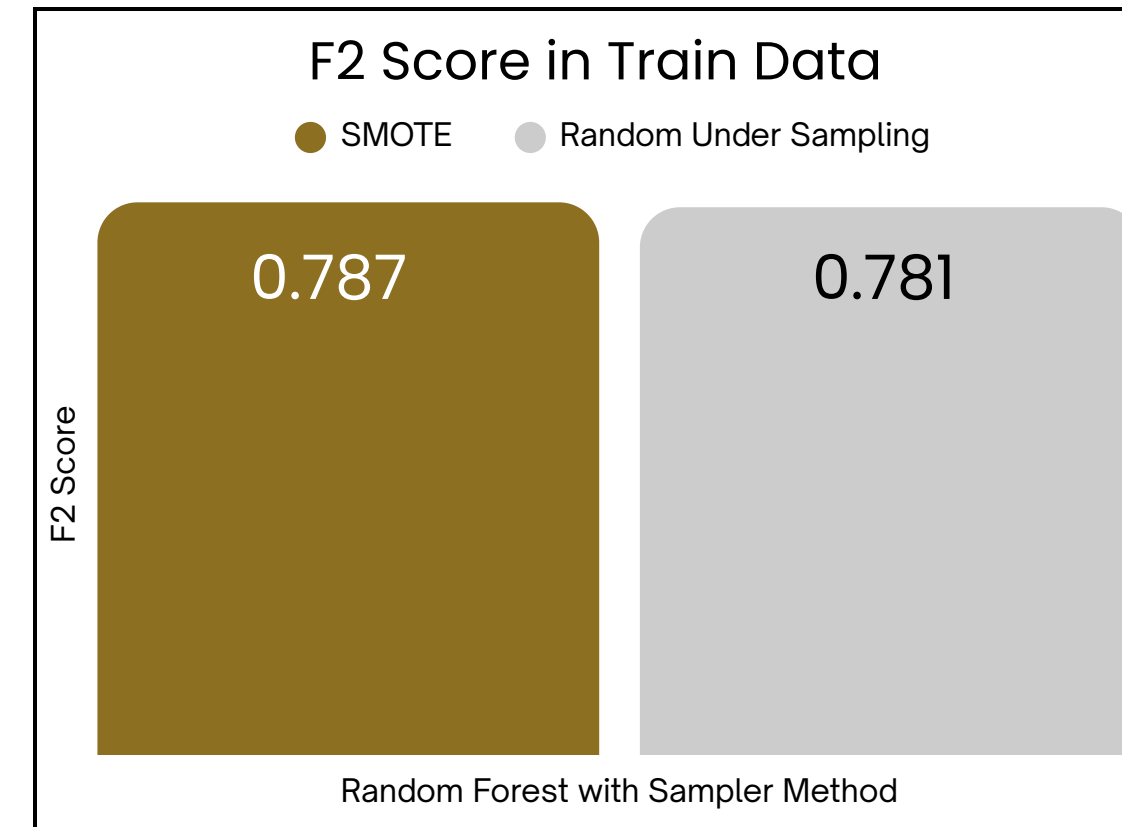
Based on this result, SMOTE was selected for further testing.

- 🎯 However, on the test set, the F2 Score dropped from 0.756 (before tuning) to 0.721 (after tuning).
- 📌 Next Step: While SMOTE improved training performance, tuning did not enhance test results. To improve generalisation and maximise F2 Score on unseen data, threshold optimisation is highly recommended as the last resort.

Home

Analytics

Conclusion





# Threshold Optimisation

✓ The optimal threshold is found at 0.18, where the F2 Score reaches its highest point. This threshold balances recall and precision with a heavier emphasis on recall — ideal for imbalanced classification tasks.

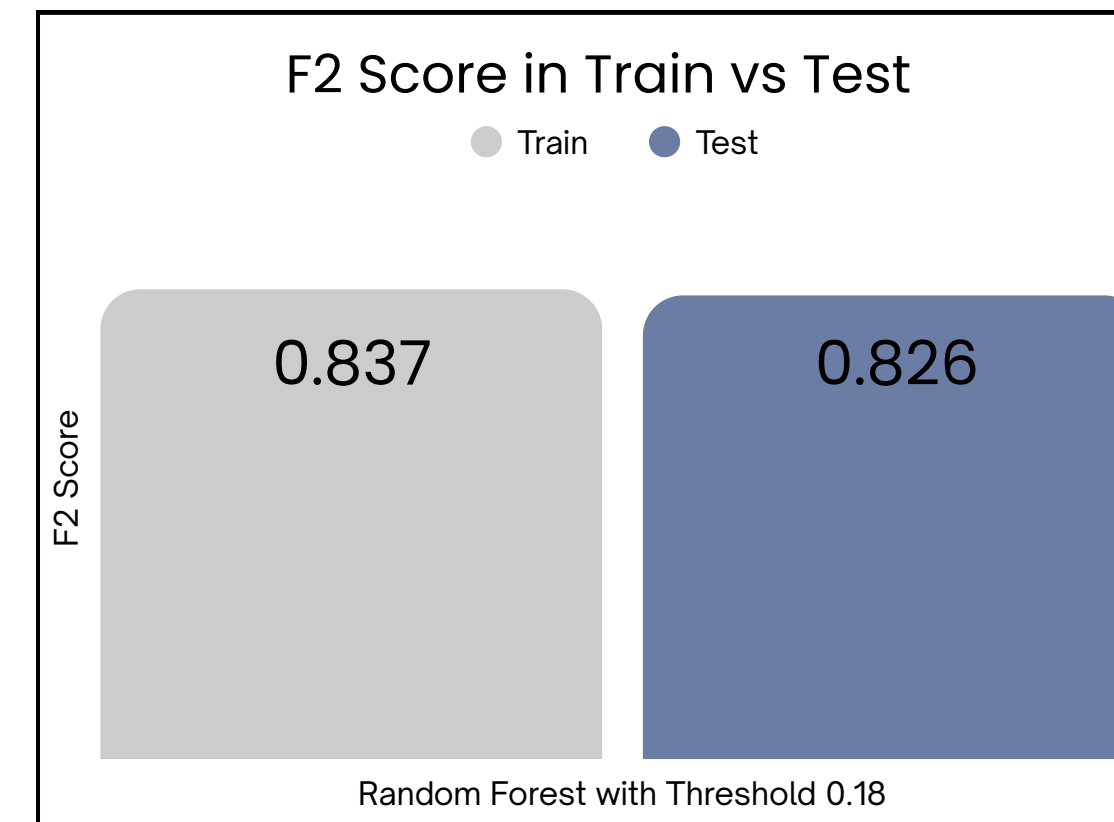
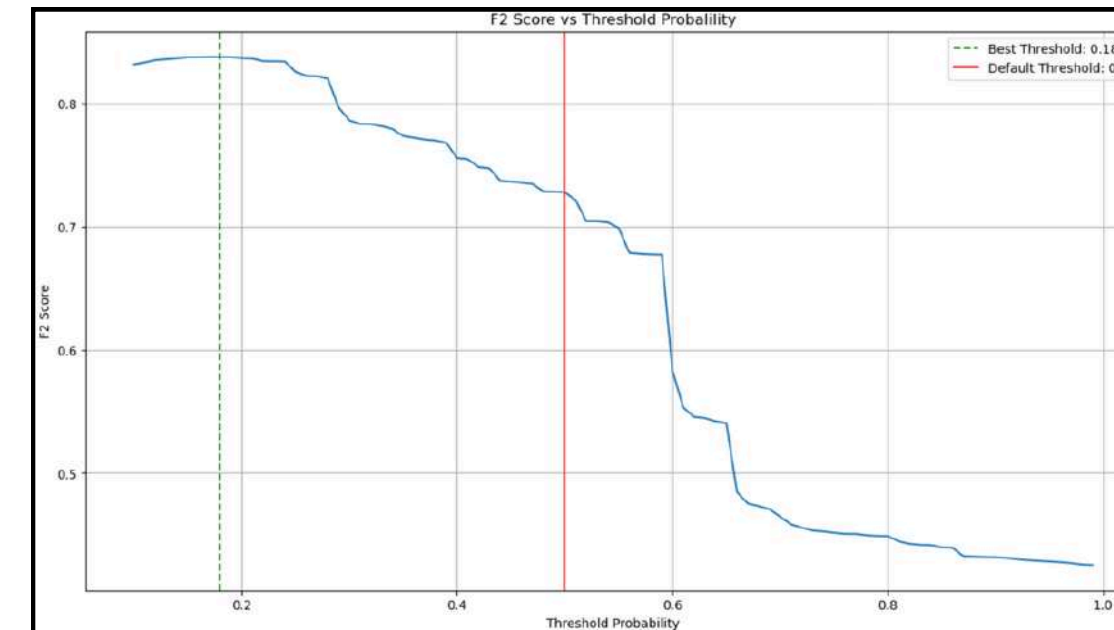
📈 When applying this threshold to the Random Forest model:

- Train F2 Score: 0.837
- Test F2 Score: 0.826

➡ This indicates excellent generalization performance with minimal overfitting.

🚀 Hyper-Parameter Tuning Conclusion:

The Random Forest model with a threshold of 0.18 delivers the best F2 performance on both training and unseen data. This model is therefore highly recommended as the final model for deployment







[Home](#)

[Analytics](#)

[Conclusion](#)

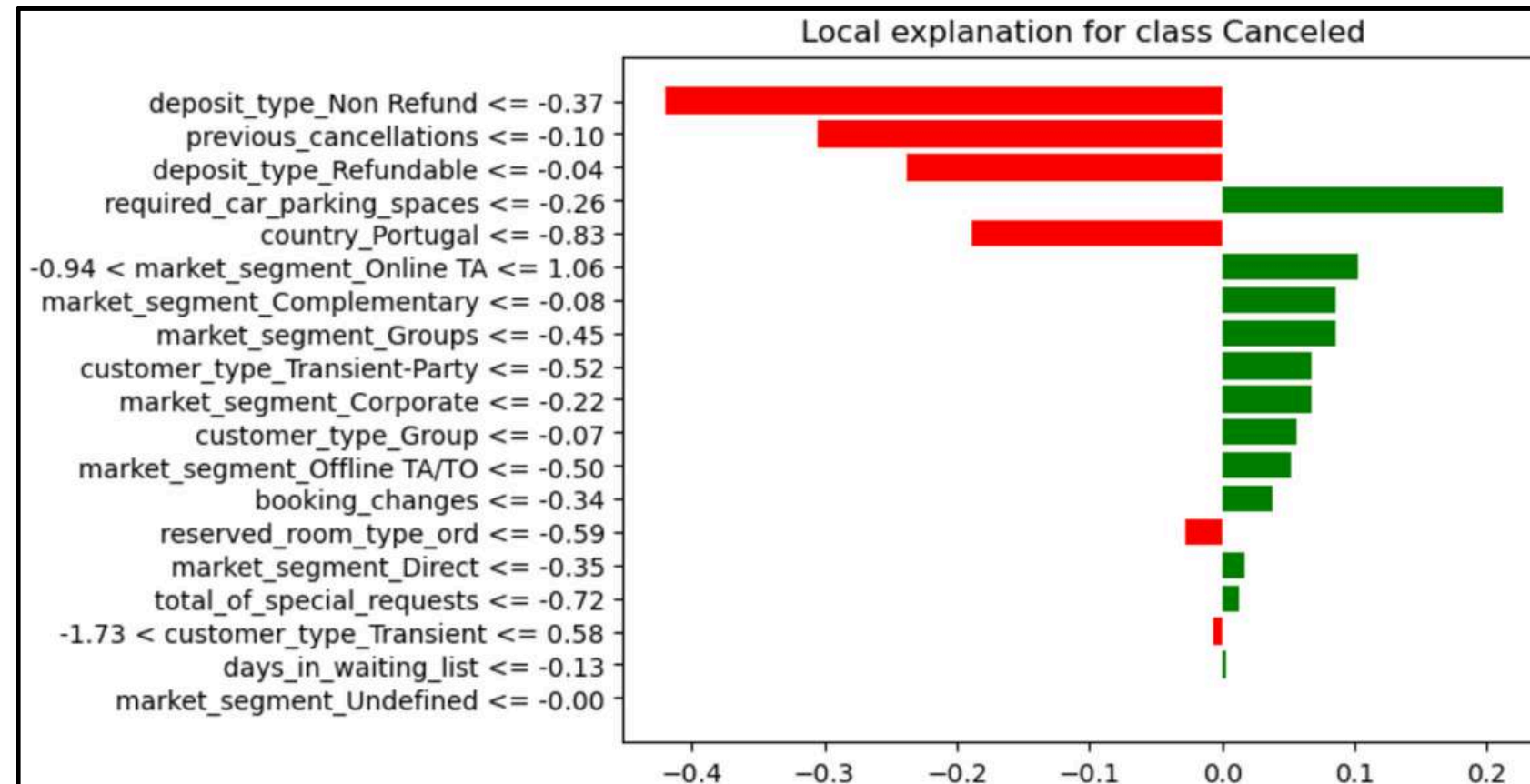
# Model Interpretation & Business Implementation







# Model Interpretation



## Strong Negative Impact (Reducing Cancellations)

- 🇵🇹 Portuguese guests are least likely to cancel → A reliable market to target with special offers.
- 💰 Non-refundable deposits significantly deter cancellations → Use tiered refund policies to encourage commitment.
- 🚗 Parking requirements signal well-planned stays → Promote “parking-included” packages to attract stable bookers.

## Mixed Impact (Context Matters)

- 🌐 Online Travel Agents (OTA) have dual effects — convenient yet prone to impulsive cancellations → Apply OTA-specific rules like higher deposits or loyalty perks.
- 👤 Transient guests show varied behavior depending on season or flexibility → Consider dynamic cancellation risk modeling.

## Low Influence

- Features like Group type or Undefined segments show negligible impact → Not worth focusing resources here.







# Business Implementation Scenario

## Without Machine Learning

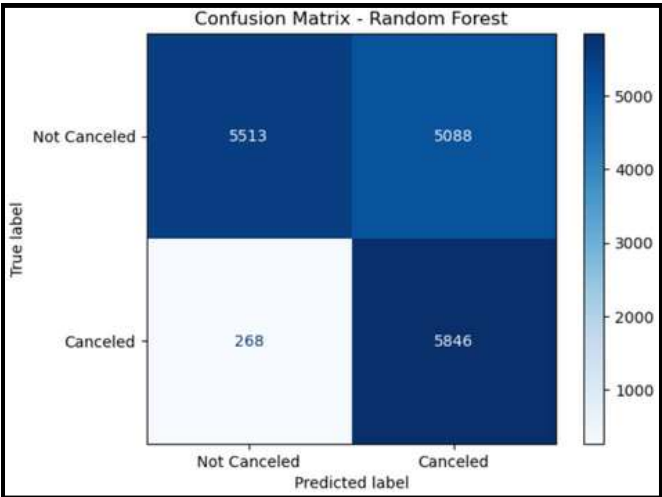
To avoid the costly risk of FN, the business assumes **all bookings will cancel**.

🔍 Prediction Outcome & Business Impact		
Metric	Value	Description
False Positives (FP)	10,601	Not canceled but predicted as canceled
True Positives (TP)	6,114	Correctly predicted canceled bookings
Resold Bookings (70% TP)	4,280	Canceled bookings successfully resold
Unresold Bookings (30% TP)	1,834	Canceled bookings that could not be resold

💰 Estimated Cost Impact		
Cost Component	Formula	Amount
Loss from FP	$10,601 \times \$100$	\$1,060,100
Loss from TP Unresold	$1,834 \times \$500$	\$917,000
✅ Total Estimated Loss		\$1,977,100

## With Machine Learning

ML model predicts cancellations with the following **confusion matrix**



🔍 Prediction Outcome & Business Impact (with ML Model)		
Prediction Metric	Value	Description
True Positives (TP)	5,846	Canceled bookings correctly predicted
False Positives (FP)	5,088	Not canceled but wrongly predicted as canceled
False Negatives (FN)	268	Canceled but missed by model
True Negatives (TN)	5,513	Not canceled and correctly predicted
Resold Bookings (70% of TP)	4,092	Canceled bookings successfully resold
Unresold Bookings (30% of TP)	1,754	Canceled bookings that couldn't be resold

💰 Estimated Cost Breakdown		
Cost Component	Formula	Amount
Loss from FP	$5,088 \times \$100$	\$508,800
Loss from FN	$268 \times \$500$	\$134,000
Loss from TP Unresold	$1,754 \times \$500$	\$877,000
✅ Total Estimated Loss		\$1,519,800

📊 Cost Comparison Summary	
Scenario	Total Loss
Without ML	\$1,977,100
With ML	\$1,519,800
💡 Savings	\$457,300



Home

Analytics

Conclusion

# Conclusion and Recommendation







# Conclusion and Recommendations

## Conclusion

### Model Performance

- Best model: Random Forest Classifier with SMOTE and Threshold Optimisation can provide F2-score: 0.826 (prioritises recall).
- Top predictors:
  - deposit\_type\_Non Refund
  - previous\_cancellations
  - required\_car\_parking\_spaces

### Business Impact

- False Negative cost: \$500/booking (lost revenue)
- False Positive cost: \$100/booking (operational inefficiency)
- Model saves up to \$457K via excellent machine learning decision-making implementation

## Business Recommendation

### Operations











- Dynamic Overbooking
  - Groups: +8–12%, Online TA: +5–7%
- Deposit Policy Tuning
  - Incentivise refundable options for high-risk guests
  - Restrict non-refundable to corporate/repeat guests

### Data-Driven Actions

- Early Warning System
  - Flag bookings with >65% cancellation risk
- Policy Tweaks
  - Limit booking lead times & stay durations for risk segments
- Model Maintenance
  - Retrain regularly to adapt to behavioral shifts



# Model Limitations

1.  Missing Booking Details
  - No booking date, room count, or cancellation reason
  -  Limits detection of last-minute cancellations or event-driven trends
2.  No Room-Level Information
  - Missing room type, pricing, and availability
  -  Model can't differentiate guest preferences or revenue potential
3.  Incomplete Guest Data
  - Missing number of guests, guest type (solo, family), and length of stay
  -  Hinders behavioral insights and personalization opportunities
4.  Duplicates & Outliers
  - High volume of duplicates and outliers in lead\_time, ADR, etc.
  -  Risk of data leakage and biased learning
5.  No Cancellation Reason Field
  - No column capturing why bookings were canceled
  -  Model relies on indirect proxies, not real causes





Home

Analytics

Conclusion



# Thank You