Hotel Cancellation Prediction

Authors: Amit Kadane, Ghanshyam Burnwal, Nitish Kumar, Shalin Shah

Affiliation: Indian Institute of Science, Bangalore

Email Address:

amitkadane@iisc.ac.in, ghanshyamb@iisc.ac.in, nitishkumar@iisc.ac.in, shalinshah@iisc.ac.in

Problem Statement

Hotels struggle to distinguish reliable bookings from likely cancellations, hurting revenue and inventory planning. Traditional forecasting fails to capture dynamic pricing, customer behavior, and changing demand. Without accurate cancellation prediction, hotels cannot optimize overbooking or apply preventive measures. This uncertainty causes financial loss and operational inefficiency, highlighting the need for a data-driven solution.

Objective

This project aims to predict booking cancellation probability using historical data and machine learning. By analyzing booking patterns, customer types, seasonality, and payment details, the system estimates cancellation risk. These predictions enable proactive decisions—optimizing overbooking, pricing, deposits, and resource utilization to improve profitability.

Workflow

01

Data Collection

The dataset was sourced from the Hotel Booking Demand dataset available on Kaggle. It includes reservation information from both City and Resort hotels, containing approximately 119,390 booking records with 32 features.

02

Data Preparation and Exploration

The dataset was cleaned by handling missing values, duplicates, and categorical encoding. Exploratory Data Analysis (EDA) was performed to understand booking patterns, cancellation trends, correlations, and key behavioral insights.

03

Feature Engineering

It includes creating new variables like total stay length and grouping lead time into meaningful ranges. Categorical features were encoded, and numerical values were normalized or standardized where needed to improve model performance

04

Model Development

Multiple machine learning models such as Logistic Regression, Random Forest, Gradient Boosting and XGBoost were trained. Hyperparameter tuning and cross-validation were performed to optimize each model..

05

Model Evaluation & Selection

Models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Feature importance and confusion matrix analysis helped interpret model behavior and identify the best-performing model.

06

Deployment and Reporting

Results were visualized through plots and diagrams to support decision-making. Predictions enable proactive hotel strategies like overbooking adjustments, retention offers, and deposit policies

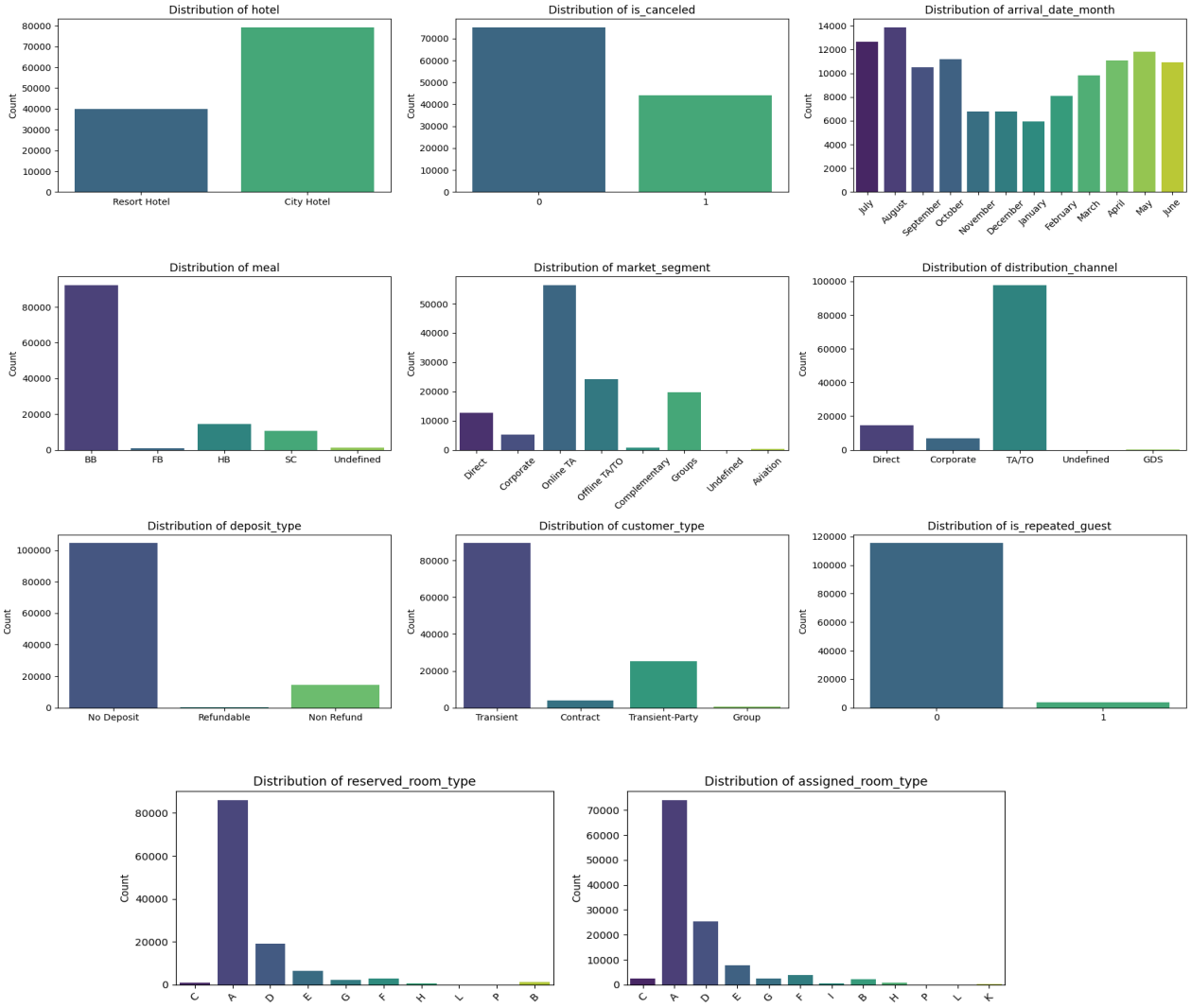
/04

Data Collection

- **Dataset Name**: Hotel Booking Demand
- **DatasetSource:** Kaggle (collected from real hotel reservation systems and anonymized)
- **FileFormat:**CSV (hotel_bookings.csv)
- **Size:** ~119,390 booking records × 32 features
- **Source:** https://www.kaggle.com/datasets/mojtaba142/hotel-booking
- It contains booking information from two types of hotels — City Hotel and Resort Hotel

Data Exploration

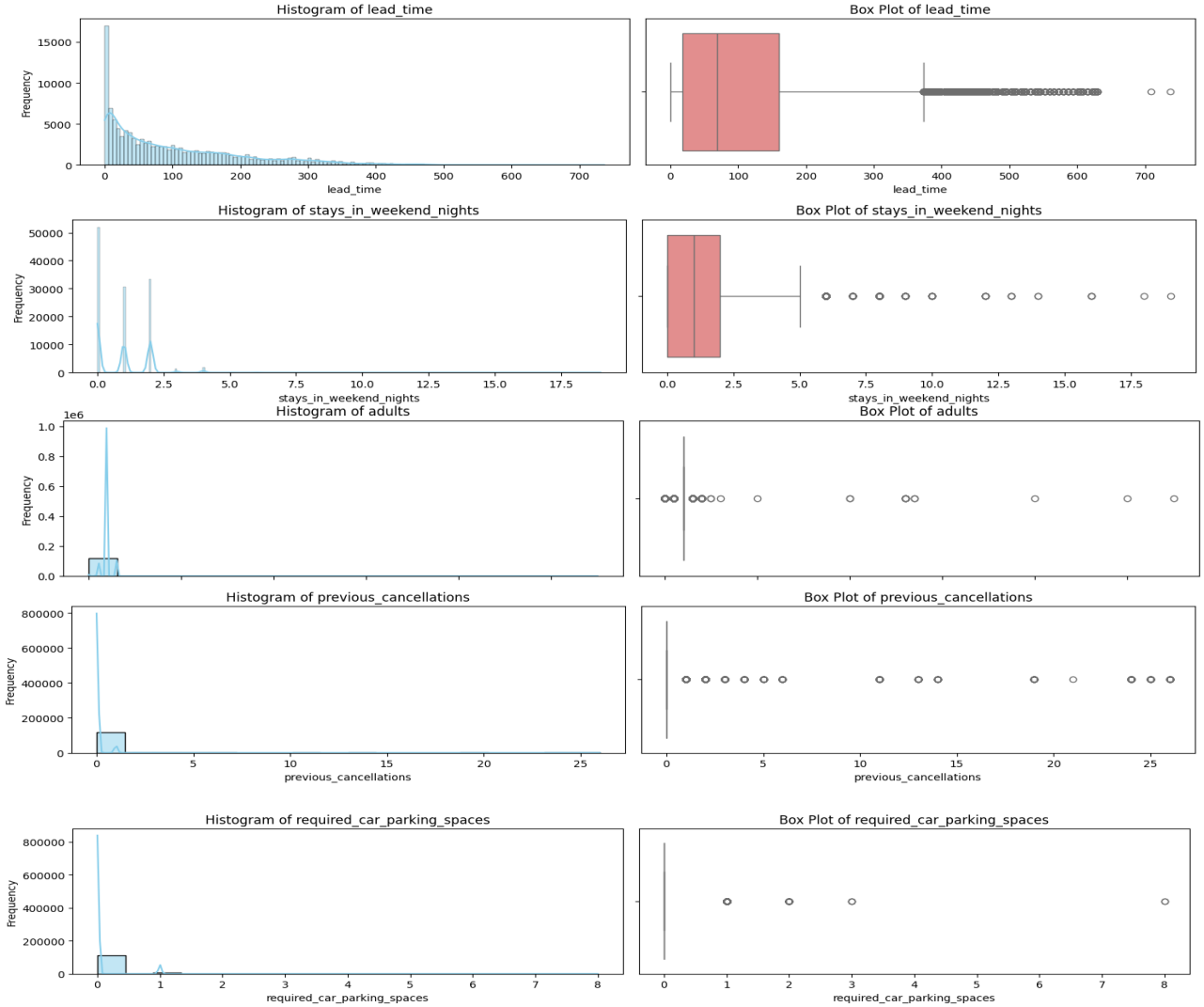
Univariate EDA – Categorical features



Key Univariate Data Distribution Insights

- **City hotels have more bookings than resort hotels**, suggesting higher demand for urban stays.
- **A large portion of bookings are non-cancelled**, but cancellations still make up a significant share, reflecting business risk.
- **Booking volume fluctuates by month**, with peaks around summer and early spring, indicating seasonal travel patterns.
- **Most guests book a Bed & Breakfast (BB) meal plan**, while other meal types are much less common.
- **Online Travel Agents (OTA) dominate booking channels**, showing that most customers rely on digital platforms rather than direct or corporate channels.
- **No-deposit bookings are overwhelmingly common**, which may encourage higher cancellation rates due to lower financial commitment.
- **The majority of customers are transient travelers**, while contract, group, or party bookings represent a smaller portion.
- **Most bookings are made by first-time guests**, with repeated guests being a very small segment.
- **Reserved and assigned room types follow a similar pattern**, but mismatches exist, indicating reassignment at check-in.

Univariate EDA – Numeric features



Key Univariate Data Distribution Insights

Lead Time
- **Distribution**: Highly right-skewed; most bookings occur within 0–100 days.
- **Outliers**: Extreme values beyond 400 days.
- **Insight**: Majority of customers book close to the stay date; very long lead times may indicate special cases or anomalies.

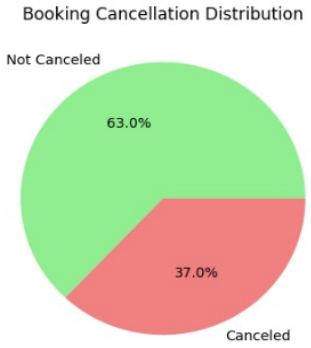
Stays in Weekend Nights
- **Distribution**: Most bookings have 0–2 weekend nights.
- **Outliers**: Rare cases up to 17 nights.
- **Insight**: Typical stays are short; long weekend stays are uncommon.

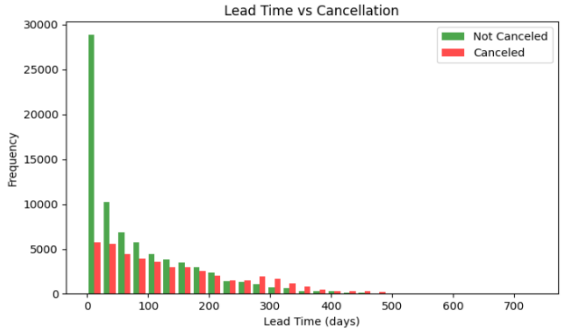
Adults
- **Distribution**: Dominated by 1–2 adults per booking.
- **Outliers**: Few cases with very high adult counts.
- **Insight**: Standard bookings involve small groups; extreme values may represent group bookings or data errors.

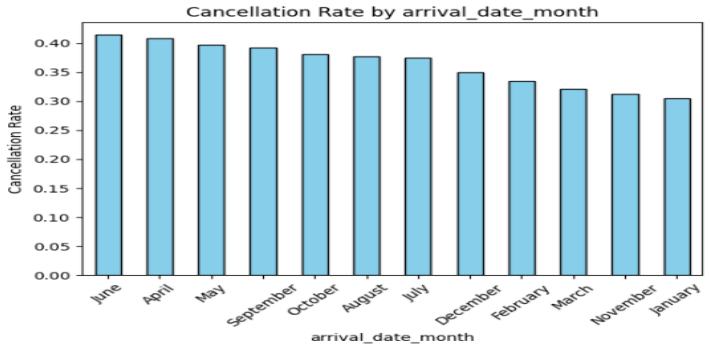
Previous Cancellations
- **Distribution**: Almost all customers have 0 previous cancellations.
- **Outliers**: Some customers have up to 25 cancellations.
- **Insight**: High previous cancellations strongly indicate future cancellation risk.

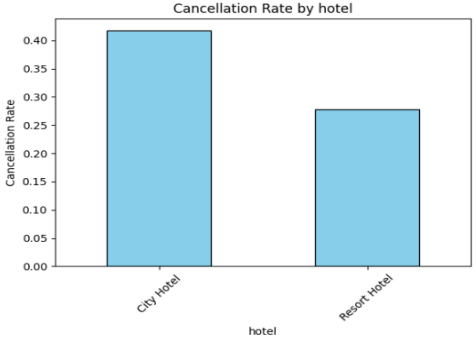
Required Car Parking Spaces
- **Distribution**: Most bookings require 0 spaces; very few need 1 or more.
- **Outliers**: Up to 8 spaces in rare cases.
- **Insight**: Parking demand is generally low; extreme values may represent special events or anomalies.

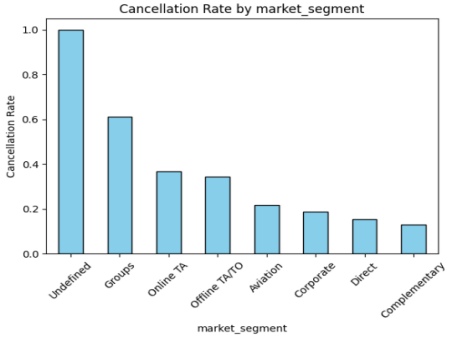
Hotel booking Cancellation Distribution and Rate

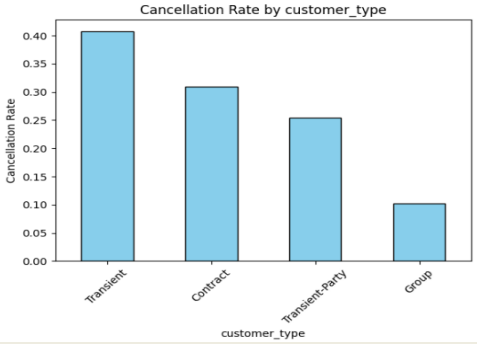


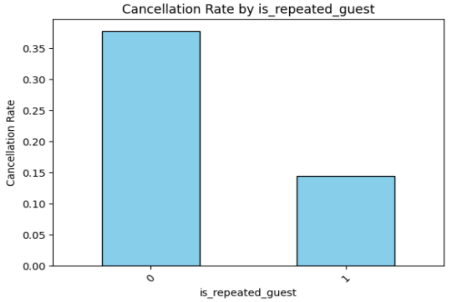






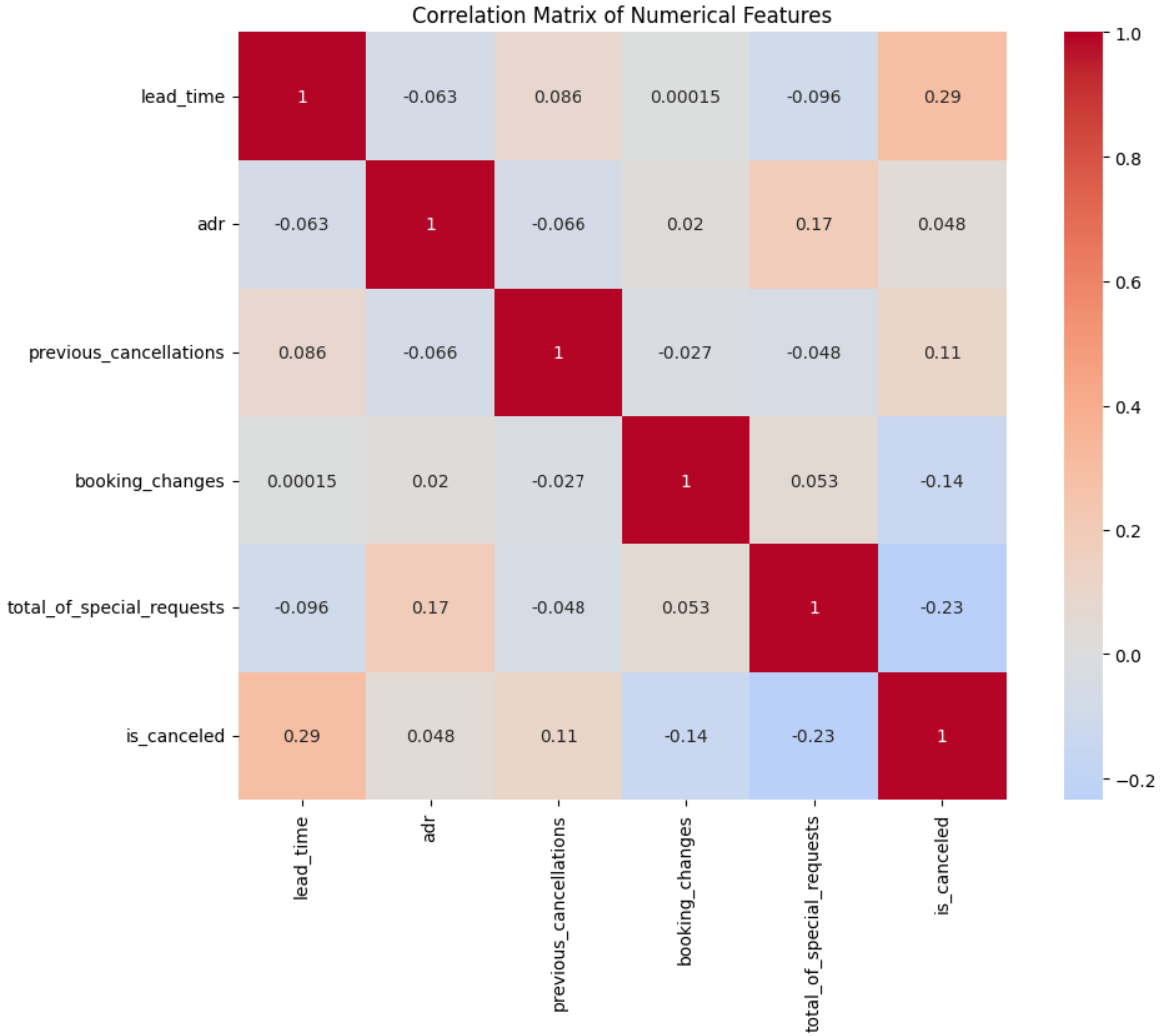






- **Around 37% of all bookings result in cancellations,** which highlights a significant challenge for forecasting occupancy and managing operational planning
- **Bookings made far in advance tend to have a much higher cancellation probability,** indicating that longer lead times create uncertainty in final stay confirmations
- **City Hotels experience noticeably higher cancellation rates compared to Resort Hotels,** suggesting that business or short-stay travel plans are more volatile
- **Bookings coming through online travel agents are the most likely to be cancelled,** whereas direct and corporate bookings tend to be more reliable and stable
- **Transient travelers cancel far more frequently than contract or group customers,** showing that individual or flexible travelers are less committed to their reservations
- **First-time hotel guests are significantly more likely to cancel their booking** compared to returning guests, implying loyalty plays a key role in booking reliability
- **Cancellation activity is highest during the peak summer months and declines in off-season periods like January and November,** indicating clear seasonal trends in booking behavior

Correlation Matrix of Numeric Features



Lead time shows the strongest positive correlation with cancellations (0.29), indicating that bookings made far in advance are more likely to be cancelled

Previous cancellations also show a mild positive correlation (0.11) with cancellation likelihood, suggesting customers with a history of cancellations are more prone to cancel again

ADR (Average Daily Rate) has a weak positive correlation (0.048) with cancellation behavior, meaning price does not significantly influence cancellation decisions

Total special requests show a slight negative relationship (-0.23) with cancellations, implying guests who make special requests are generally more committed and less likely to cancel

Booking changes also have a small negative correlation (-0.14) with cancellation probability, which suggests guests who modify their booking tend to follow through rather than cancel

Overall**, no feature shows a very strong linear correlation,** indicating that cancellations are influenced by multiple factors rather than a single dominant variable

These findings highlight the importance of using **machine learning models rather than relying solely on correlation-based rules** to predict cancellations effectively

Feature Engineering

Handling missing values:

- 'Children', 'Agent', 'Company' set to 0
- 'Country' set to 'Unknown'

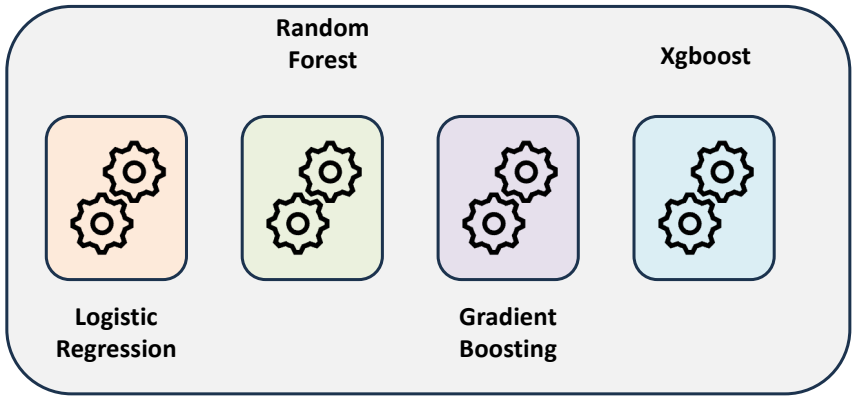
Creating new features:

- Total guests – by adding adults, children and babies
- Total stay duration – by adding weekend and weekday stays
- Family booking indicator – by checking if booking has children or babies
- Booking season – creating 4 seasons across a year
- Special requests per stay – by checking total special requests and total stay
- High lead time indicator – if lead time > 100
- Has previous cancellations – if previous cancellations > 0

Categorical variables encoding

Numerical features scaling

Model Development




```python
models = {
    'logistic_regression': LogisticRegression(random_state=42, max_iter=1000),
    'random_forest': RandomForestClassifier(random_state=42, n_estimators=100, n_jobs=-1),
    'gradient_boosting': GradientBoostingClassifier(random_state=42),
    'xgboost': xgb.XGBClassifier(random_state=42, eval_metric='logloss', n_jobs=-1)
}
```

```
... Training Multiple Models...
================================================

Training logistic_regression...
logistic_regression - ROC-AUC: 0.8565

Training random_forest...
  Best parameters: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
random_forest - ROC-AUC: 0.9574

Training gradient_boosting...
gradient_boosting - ROC-AUC: 0.9255

Training xgboost...
  Best parameters: {'learning_rate': 0.2, 'max_depth': 9, 'n_estimators': 200}
xgboost - ROC-AUC: 0.9592

Best model: xgboost with ROC-AUC: 0.9592

Model Comparison:
                model  roc_auc
              xgboost 0.959184
        random_forest 0.957389
    gradient_boosting 0.925504
  logistic_regression 0.856539
```

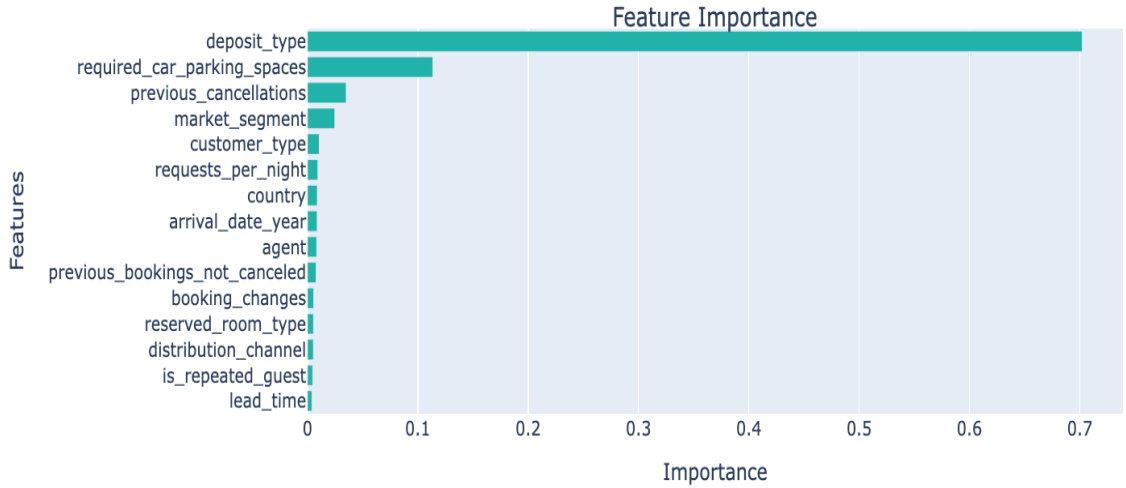
```python
# Parameter grids for hyperparameter tuning
param_grids = {
    'random_forest': {
        'n_estimators': [100, 200],
        'max_depth': [10, 20, None],
        'min_samples_split': [2, 5]
    },
    'xgboost': {
        'n_estimators': [100, 200],
        'max_depth': [3, 6, 9],
        'learning_rate': [0.01, 0.1, 0.2]
    }
}
```

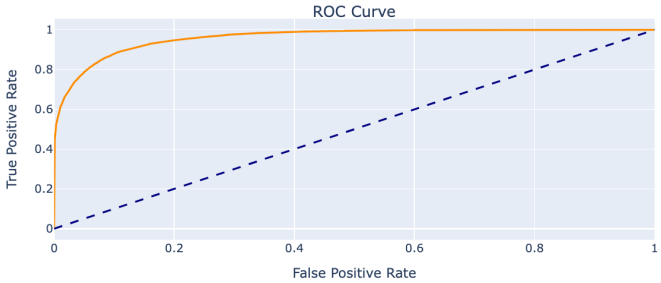
```python
if name in param_grids:
    # Perform grid search for models with parameter grids
    grid_search = GridSearchCV(model, param_grids[name],
                               cv=5, scoring='roc_auc', n_jobs=-1, verbose=0)
    grid_search.fit(X_train, y_train)
    models[name] = grid_search.best_estimator_
    score = grid_search.best_score_
    print(f"Best parameters: {grid_search.best_params_}")
else:
    # Use cross-validation for other models
    model.fit(X_train, y_train)
    models[name] = model
    scores = cross_val_score(model, X_train, y_train,
                             cv=5, scoring='roc_auc')
    score = np.mean(scores)
```

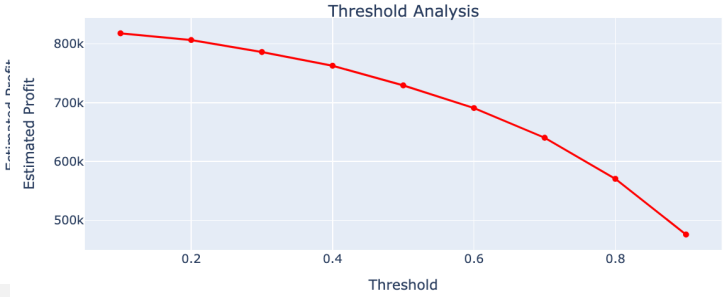
Model Evaluation & Business Recommendations

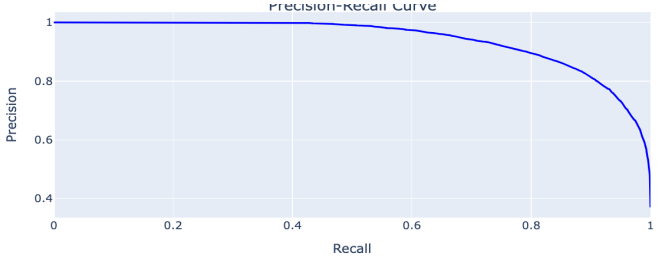
Xgboost - Evaluation

```
Model Performance Metrics:
  ROC-AUC Score: 0.9613
  PR-AUC Score: 0.9429
  Accuracy: 0.8938
  Precision: 0.8712
  Recall: 0.8370
  F1-Score: 0.8538
Generating Evaluation Plots...
```









Model Performance Insights

•The model demonstrates **strong predictive performance**, with an AUC score of **0.9613**, indicating excellent ability to distinguish between cancelled and non-cancelled bookings.
•An overall **accuracy of 89.38%** and **F1-score of 0.8538** show a well-balanced performance between precision and recall, making the model reliable for real-world hotel booking scenarios.
•The **precision (0.8712)** and **recall (0.8370)** values suggest the model effectively flags high-risk cancellations while keeping false positives minimal.

Feature Influence Insights

•**Deposit type is the most influential factor** driving cancellations, meaning customers required to pay upfront are far less likely to cancel.
•**Required parking spaces and previous cancellations** also significantly impact prediction, suggesting behavioral and commitment-related factors play a major role.
•Market-related variables like **market segment, customer type, and agent** contribute meaningfully, showing booking source strongly aligns with cancellation trends.

Threshold and Business Interpretation

•The **threshold analysis shows profit decreases as the decision threshold increases**, meaning stricter cancellation classification leads to lost opportunity revenue.
•The **precision-recall curve is strong**, indicating that even at lower thresholds, the model maintains high accuracy in identifying genuine cancellation risks.

Business Insights and Recommendations

Business Insights

High-Risk Booking Profile

- A total of **4,776 high-risk bookings** were detected, representing **20% of all bookings**.
- These bookings have an **average lead time of only 0.8 days**.
- Guests in this group previously cancelled **0.38 times on average**.

Revenue Impact

- With an average room rate of **$100**, high-risk bookings put **$143,280** of revenue at risk.
- Interventions aimed at reducing cancellations could save an estimated **$71,640**

Actionable Recommendations Management

Proactive Risk Management

- Require deposits when:
  - Lead time > 100 days
  - Customer has past cancellation history
  - Non-refundable deposit types
- Implement **real-time alerts** for bookings with **>70% cancellation probability**

Targeted Communication

- Send reminders and special offers to high-risk customers 2 weeks before arrival
- Implement **double-confirmation** for risky bookings

Revenue Optimization

- Adjust **overbooking levels** using predicted cancellation rates
- **Dynamic pricing:** Offer discounts to high-risk customers ~30 days before arrival

Operational Efficiency

- Adjust **staffing** based on predicted arrivals vs cancellations
- Optimize **room allocation** and maintenance scheduling

Data Science Canvas

Project: Hotel Booking Cancellation Prediction

Team: Amit, Ghanshyam, Nitish, Shalin

Problem Statement

Business Case & Value Added

Hotel booking cancellations significantly impact revenue, occupancy forecasting, and resource planning. With flexible booking options and online travel platforms, cancellations have become frequent and unpredictable, leading to lost revenue and inefficient room utilization.

Data Landscape

Dataset used : Hotel Booking Demand Dataset from Kaggle.

Source: https://www.kaggle.com/datasets/mojtaba142/hotel-booking

Model Selection

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. XGBoost

Model Requirements

Supervised classification, Simple interpretable models, For higher accuracy and non-linear relationships, avoid overfitting, Optimize model performance, Reliable performance estimation

Software & Libraries

- pandas, numpy – data handling,
- scikit-learn – modeling and evaluation
- matplotlib, seaborn – visualization
- xgboost – advanced boosting models

Skills

- Python knowledge,
- Panda,
- Numpy,
- Scikit learn
- matplotlib, seaborn
- Understanding of Data cleaning, EDA concepts
- Knowledge of different regression and classification algorithms.

Execution & Evaluation

Model Evaluation

ROC-AUC – 0.9613
PR-AUC – 0.9429
Accuracy – 0.8938
Precision – 0.8712
Recall – 0.8370
F1-Score – 0.8538

Data Storytelling

Hotel Management - High-level summary, visuals, dashboards

Marketing & Sales Teams - Charts, heatmaps, and actionable recommendations

Operations Team - Simple dashboards with daily/weekly cancellation predictions

Data Science / IT Team - Technical report, model performance graphs, SHAP feature explanations

Data Collection & Preparation

Data Selection & Cleansing

- **Relevant features**: Feature engineering to be done post EDA analysis
- **Cleaning required**: Yes — we'll clean for missing data, outliers, and possibly drop or transform features with low utility.

Data Integration

Not applicable for our project.

Data Collection

- No separate data collection activity.
- Data will be collected from standard hotel booking and cancellation data sources

Not in scope of project work. We are using ready-made dataset from Kaggle.

Explorative Data Analysis

Detailed Univariate, Bivariate and Multivariate analysis done.

- **Encoding done for categorical features.**
- **Scaling done for numeric features.**
- **Added few more features as part of Feature Engineering.**
- **Handled missing values for 4 features.**

Adopted from: https://github.com/tomalytics/datasciencecanvas

THANK YOU