Hotel Booking Cancellation Prediction

Author: Amit Kadane, Ghanshyam Burnwal, Nitish Kumar, Shalin Shah
Affiliation: Indian Institute of Science
Email Address: amitkadane@iisc.ac.in, ghanshyamb@iisc.ac.in, nitishkumar@iisc.ac.in, shalinshah@iisc.ac.in

Introduction

Hotel booking cancellations have become increasingly common due to flexible policies and online travel platforms. These cancellations disrupt occupancy forecasts and reduce operational efficiency, making planning unpredictable for hotels.

As a result, hotels often face lost revenue, vacant rooms, and misallocation of staff and resources. Understanding and predicting cancellation behavior has therefore become essential for improving business performance and planning accuracy.

Problem Statement

Current hotel operations struggle to differentiate confirmed bookings from those likely to cancel, leading to ineffective revenue and inventory planning. Traditional forecasting methods fail to account for dynamic pricing, customer variability, and fluctuating demand.

Without accurate cancellation prediction, hotels cannot adjust overbooking strategies, apply preventive measures, or manage resources efficiently. This uncertainty leads to financial losses and operational challenges that require a data-driven solution.

Objective

The objective of this project is to use historical booking data and machine learning models to predict cancellation probability at the time of reservation. By analyzing patterns in booking behavior, customer type, seasonality, and payment details, the system can estimate cancellation risks.

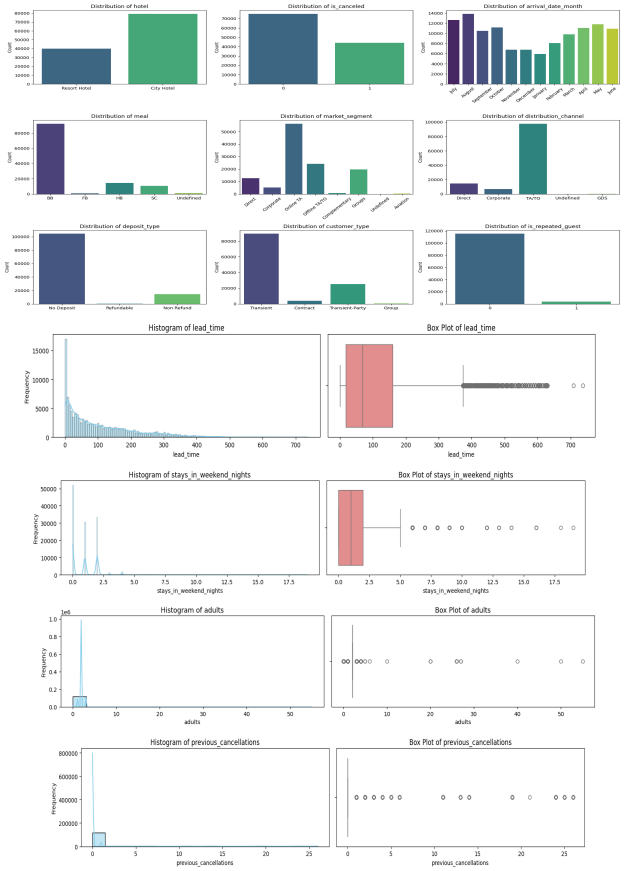
These predictions will help hotels make proactive decisions, such as adjusting overbooking levels, offering incentives, requesting deposits, and optimizing pricing. The goal is to enhance profitability, forecasting accuracy, and resource utilization.

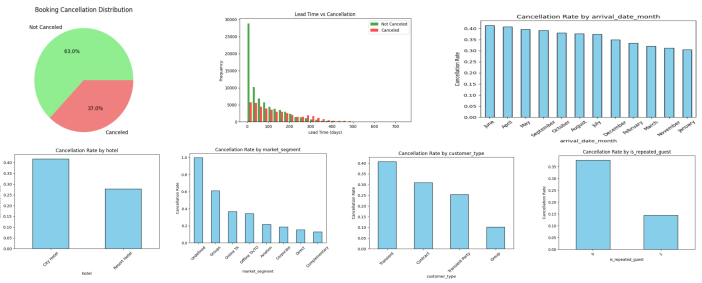
Dataset

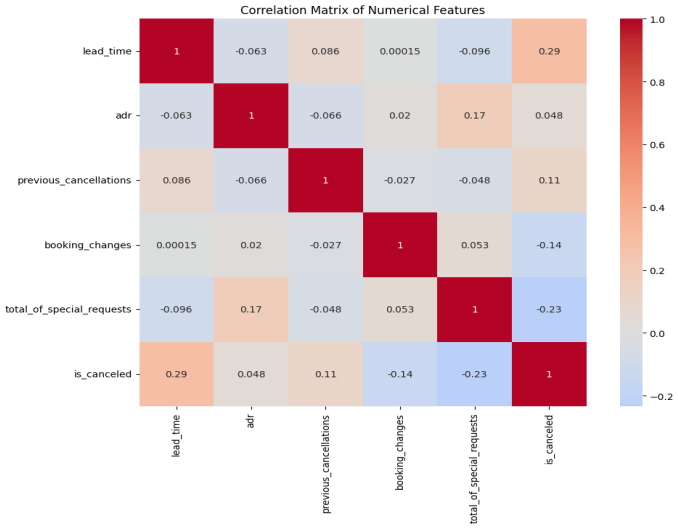
Dataset Name: Hotel Booking Demand
DatasetSource: Kaggle (collected from real hotel reservation systems and anonymized)
FileFormat:CSV (hotel_bookings.csv)
Size: ~119,390 booking records × 32 features
Source: https://www.kaggle.com/datasets/mojtaba142/hotel-booking
It contains booking information from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

EDA & Feature Engineering:

After checking overall shape of the data and verifying the data using info, describe etc. we performed detailed univariate, bivariate and multivariate EDA. Few graphs of the same are shown below.







Key EDA 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.
- 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.
- 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.

Feature Engineering:

- Handled missing values of children, country, agent, company features.
- Created new engineered features – total_guests, total_stay, is_family, arrival_date, arrival_season, requests_per_night, high_lead_time, has_previous_cancellations etc.
- Encoded categorical features.
- Used standard scaler for numeric features.

Models used:

We have used 4 models along with the mentioned configurations. We have also tuned hyperparameters using GridSearchCV (for random forest and xgboost) and used cross-validation for other models.

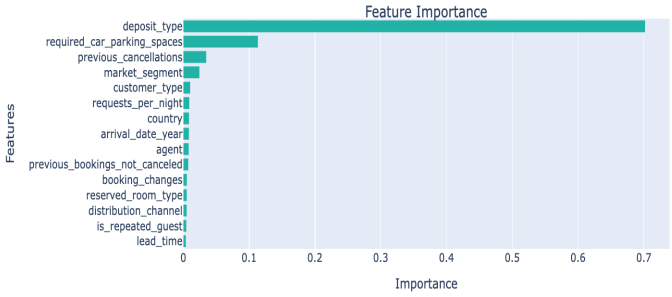
```
models = {
    'logistic_regression': LogisticRegression(random_state=42, max_iter=1000),
    'random_forest': RandomForestClassifier(random_state=42, n_estimators=100, n_jobs=-1),
    'gradient_boosting': GradientBoostingClassifier(random_state=42),
    'xgboost': xgb.XGBClassifier(random_state=42,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

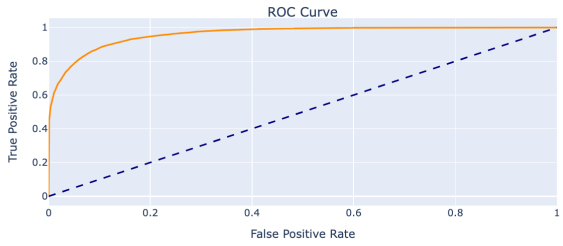
```
# Parameter grids for hyperparameter tuning
param_grids = {
    'random_forest': {
        'n_estimators': [100, 200],
        'max_depth': [10, 20, None],
        'min_samples_split': [2, 5]
    },
    'xgboost': {
        'n_estimators': [100, 200],
        'max_depth': [3, 6, 9],
        'learning_rate':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

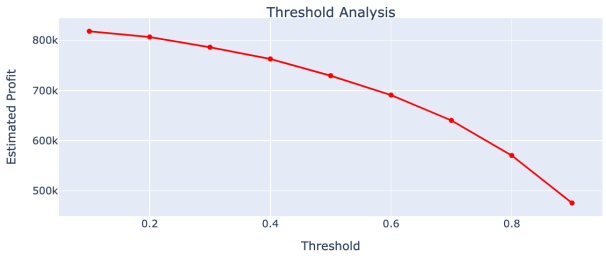
```
if name in param_grids:
    # Perform grid search for models with parameter grids
    grid_search = GridSearchCV(model, param_grids[name],
                               cv=5, scoring='roc_auc', n_jobs=-1, verbose=0)
    grid_search.fit(X_train, y_train)
    models[name] = grid_search.best_estimator_
    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)
```

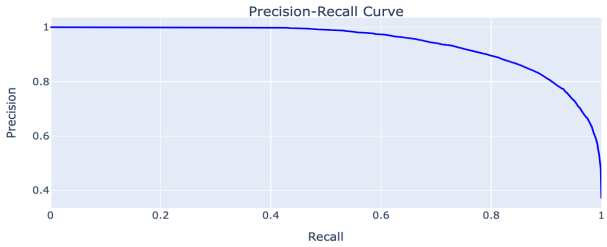
```
Training Multiple Models...
==================================================
Training logistic_regression...
logistic_regression - ROC-AUC: 0.8565
Training random_forest...
Best parameters: {'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
random_forest - ROC-AUC: 0.9575
Training 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 0.959185
        random_forest 0.957479
    gradient_boosting 0.925504
  logistic_regression 0.856473
```

Model Evaluation (Xgboost) & Business Recommendations:









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:

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

TEAM CONTRIBUTION	
Project topic selection and project proposal submission	ALL
Data collection and pre-processing	ALL
EDA – Univariate, Bivariate and Multivariate analysis	Shalin, reviewed by others
Logistic Regression	Shalin, reviewed by others
Random Forest	Ghanshyam, reviewed by others
Gradient Boosting	Amit, reviewed by others
Xgboost	Nitish, reviewed by others
Model Evaluation (Xgboost)	ALL
Business Insights and Actionable Recommendations	ALL
Project Report and Presentation	ALL