Predicting Hotel Booking Cancellations with Machine Learning

Leveraging the Kaggle Hotel Booking Demand Dataset

By: Amanuel Agajjie Wasihun | 11,05,2025, Berlin

Project Description

Predicting Hotel Booking Cancellations with Machine Learning

Problem: Hotel booking cancellations and no-show lead to lost revenue and inefficiencies in resource planning, including overbooking, underbooking, and misallocated staff.

Solution: Using the Kaggle Hotel Booking Demand dataset, I built a classification model using Python and machine learning techniques.

The model identifies patterns in guest behavior, booking details, and timing to predict cancellations.

Outcome: The final tool deployed as a Streamlit app, predicts booking cancellations, providing actionable insights for better planning and operational efficiency.

Understanding the Data:

A Glimpse into the Kaggle Dataset

Period Covered: July 2015 – August 2017

Entries: 119,390 rows | 32 columns

Hotel Types: City and Resort Hotels

Data Types:

- Categorical: e.g. hotel, is_cancelled, arrival_date_month, reservation_status
- Integer: e.g. lead_time, adults, booking_changes
- Float: e.g. children, adr, agent

Target Variable: is_canceled (1 = canceled, 0 = not canceled)

```
print("--- Initial Data Exploration ---")
print(df.head())
print(df.info())
print("\n--- Missing Values (Initial) ---")
# Count the number of missing values
print(df.isnull().sum())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 32 columns):
 #   Column                          Non-Null Count   Dtype
---  ------                          --------------   -----
 0   hotel                           119390 non-null  object
 1   is_canceled                     119390 non-null  object
 2   lead_time                       119390 non-null  int64
 3   arrival_date_year               119390 non-null  int64
 4   arrival_date_month              119390 non-null  object
 5   arrival_date_week_number        119390 non-null  int64
 6   arrival_date_day_of_month       119390 non-null  int64
 7   stays_in_weekend_nights         119390 non-null  int64
 8   stays_in_week_nights            119390 non-null  int64
 9   adults                          119390 non-null  int64
 10  children                        119386 non-null  float64
 11  babies                          119390 non-null  int64
 12  meal                            119390 non-null  object
 13  country                         118902 non-null  object
 14  market_segment                  119390 non-null  object
 15  distribution_channel            119390 non-null  object
 16  is_repeated_guest               119390 non-null  int64
 17  previous_cancellations          119390 non-null  int64
 18  previous_bookings_not_canceled  119390 non-null  int64
 19  reserved_room_type              119390 non-null  object
 20  assigned_room_type              119390 non-null  object
 21  booking_changes                 119390 non-null  int64
 22  deposit_type                    119390 non-null  object
 23  agent                           103050 non-null  float64
 24  company                         6797 non-null    float64
 25  days_in_waiting_list            119390 non-null  int64
 26  customer_type                   119390 non-null  object
 27  adr                             119390 non-null  float64
 28  required_car_parking_spaces     119390 non-null  int64
 29  total_of_special_requests       119390 non-null  int64
 30  reservation_status              119390 non-null  object
 31  reservation_status_date         119390 non-null  object
dtypes: float64(4), int64(15), object(13)
memory usage: 29.1+ MB
```

Explanatory Data Analysis (EDA)

Cancellation Rate: 37% of bookings were canceled (is_canceled).

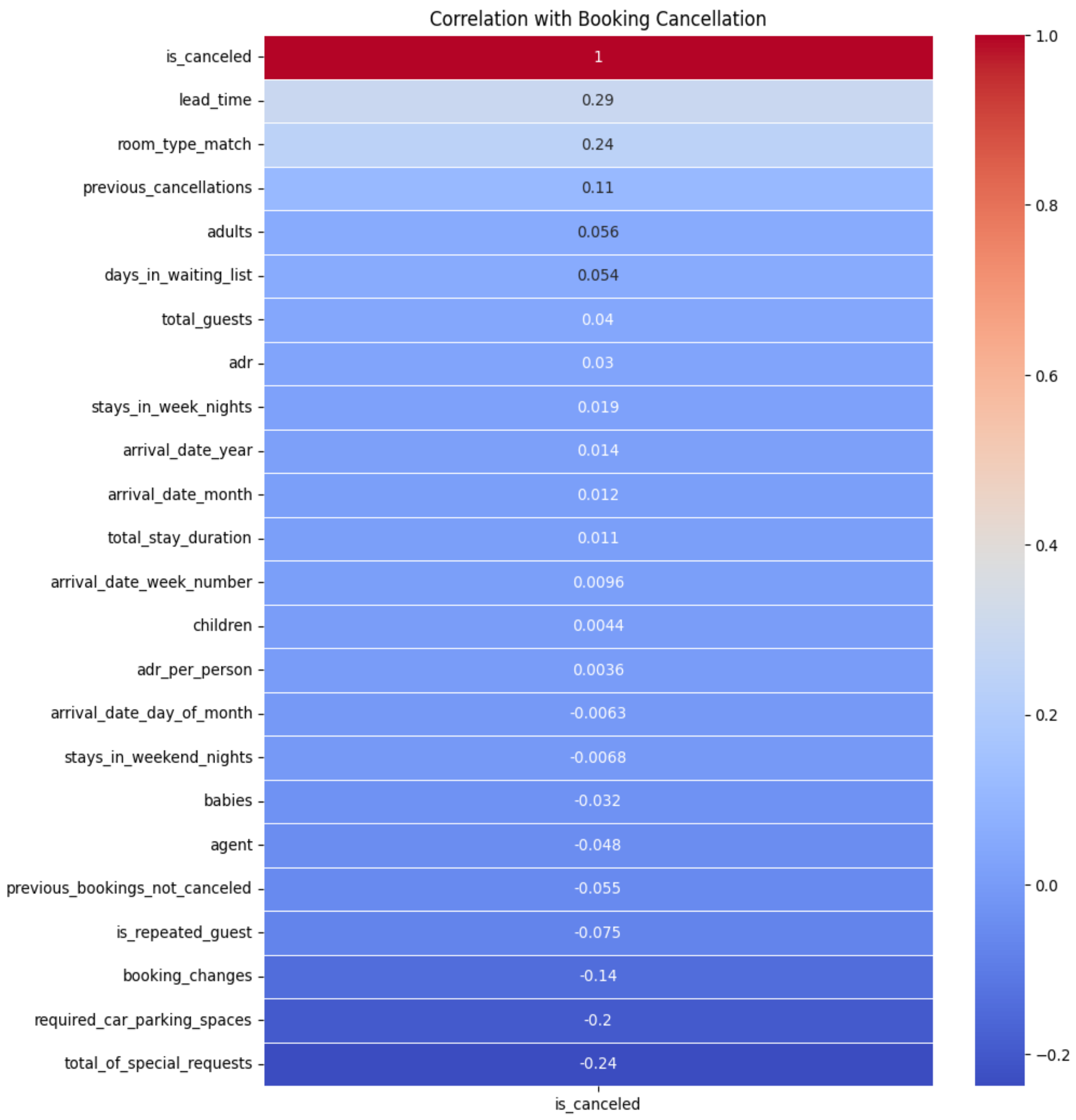
Lead Time: Positive correlation with cancellations (longer lead time → higher cancellation).

Special Requests & Parking: Negative correlation, suggesting more definite plans reduce cancellations.

Repeated Guests: Slightly less likely to cancel.

Booking Changes: Slightly less likely to cancel.

Weak Correlations: Many features have minimal linear impact on cancellations.



Building the Predictive Model:

Our Machine Learning Pipeline

- Missing Values Imputed:
  `children`, `country`, `agent` → filled with median or mode
- Dropped Column:
  `company` (too many missing values)
- Data Type Conversion:
  `children` → integer, `reservation_status_date` → datetime
- Removed Invalid Entries:
  - Bookings with 0 **total guests**
  - Rows with ADR = 0
- Standardization & Filtering:
  - Standardized 'Undefined' in `meal`
  - Removed 'Undefined' from `market_segment` and `distribution_channel`

Data Loading → **Cleaning** → Encoding → Scaling → Modeling → Evaluation

Building the Predictive Model:
Our Machine Learning Pipeline

One-Hot Encoding *(Nominal)*:
`hotel`, `meal`, `market_segment`, `distribution_channel`, `deposit_type`, `customer_type`

Label Encoding *(Ordinal / Tree-friendly)*:
`arrival_date_month`, `reserved_room_type`, `assigned_room_type`

Standardized the numerical features using **`StandardScaler`** to ensure:

- Mean = 0,
- Standard Deviation = 1

Scaled Features

`lead_time`, `arrival_date_year`, `arrival_date_month`, `arrival_date_week_number`, `arrival_date_day_of_month`, `stays_in_weekend_nights`, `stays_in_week_nights`, `adults`, `children`, `babies`, `is_repeated_guest`, `previous_cancellations`, `previous_bookings_not_canceled`, `reserved_room_type`, `assigned_room_type`, `booking_changes`, `agent`, `days_in_waiting_list`, `adr`, `required_car_parking_spaces`, `total_of_special_requests`, `total_stay_duration`, `total_guests`, `adr_per_person`, `room_type_match`

Data Loading
Cleaning
Encoding
Scaling
Modeling
Evaluation

Building the Predictive Model:

Our Machine Learning Pipeline

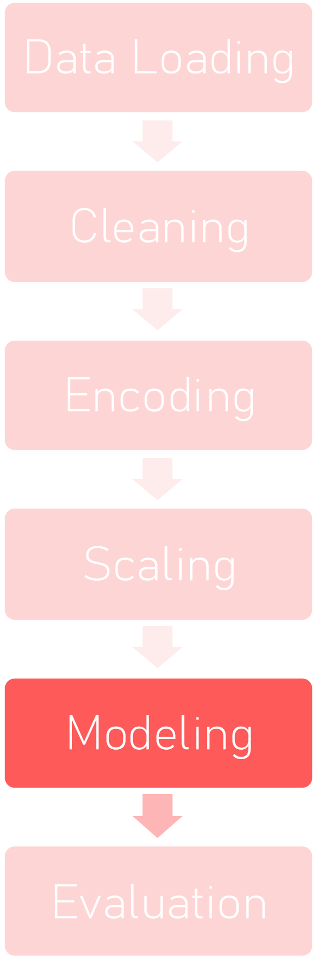
Data split:

- 80% total training data
    - 70% training, 10% validation
- 20% test set

Models Trained:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- XGBoost
- CatBoost

- Random Forest selected as best-performing model
  → Optimized using RandomizedSearchCV
- SMOTE used to handle class imbalance in training data



Building the Predictive Model:

Our Machine Learning Pipeline

Model	Accuracy	ROC AUC
Random Forest	0.8767	0.9387
Logistic Regression	0.7893	0.8690
KNN	0.8071	0.8819
Decision Tree	0.8275	0.8218
XGBoost	0.8599	0.9279
CatBoost	0.8637	0.9313

- Best Model: Random Forest with an accuracy of 0.8704 and ROC AUC of 0.9387 on the test set.
- Model Performance: Consistent across training, validation, and test sets, indicating strong generalization.
- Feature Importance: Identified key booking characteristics influencing cancellations.
- Evaluation Metrics: Accuracy, ROC AUC, and other metrics confirmed model reliability.

Data Loading → Cleaning → Encoding → Scaling → Modeling → **Evaluation**

Conclusion and Looking Ahead

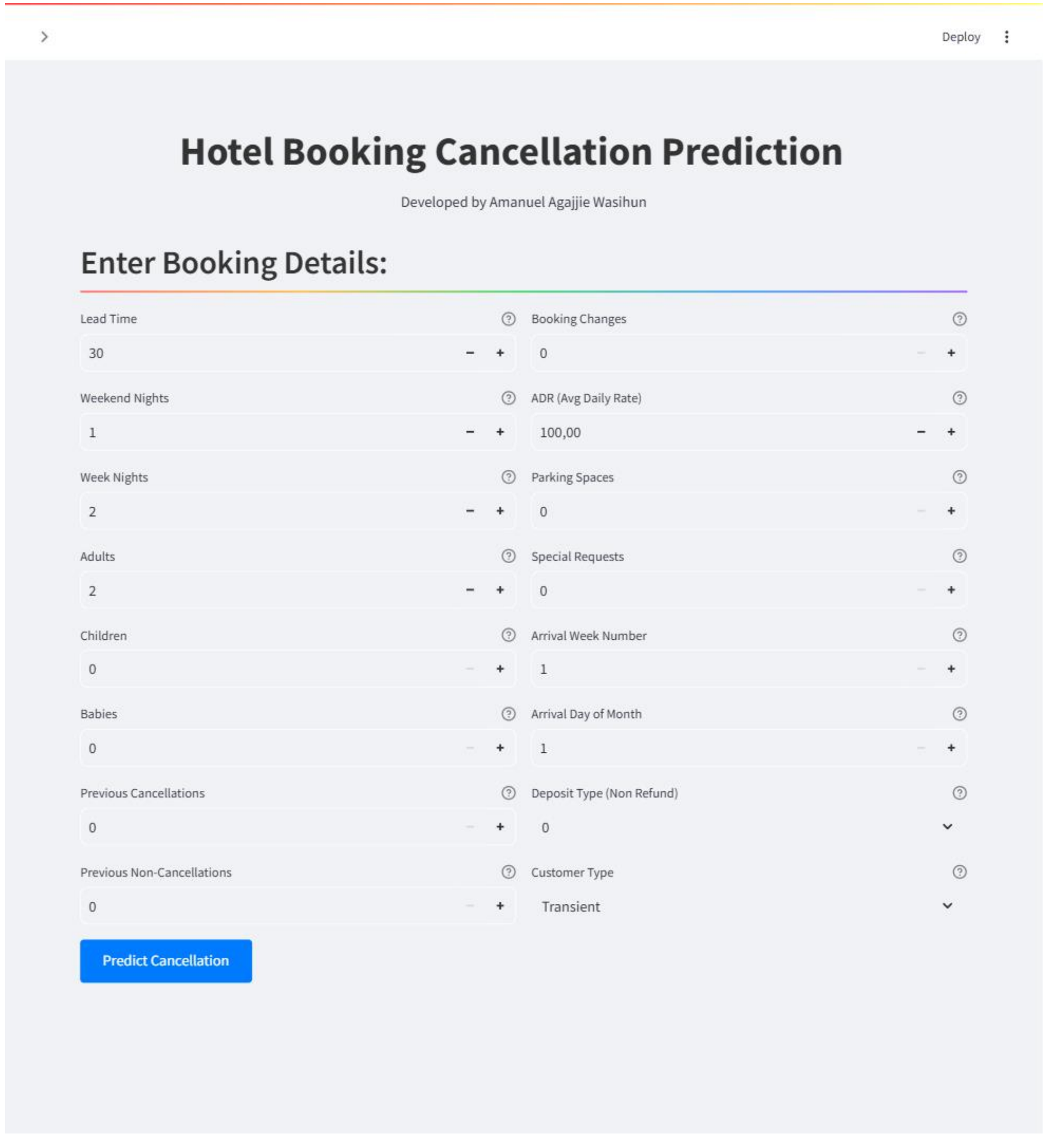
A Random Forest model was developed to predict hotel booking cancellations with high accuracy. The model is deployed in a Streamlit app for real-time cancellation predictions.

Limitations:

- Model performance relies on historical data and may degrade with significant shifts in booking patterns.
- Limited by lack of external factors like weather, competitor pricing, and real-time events.
- Predicts probabilities, not definitive outcomes; human judgment still needed.

Future Directions:

- Incorporate additional real-time data sources (weather, reviews, etc.)
- Explore more advanced techniques (ensemble methods, deep learning)
- Streamlit app developed for real-time cancellation prediction and interactive use



Thank you!

Happy to answer any questions or hear your thoughts.