Hotel Bookings: Cancellation Prediction

Maxim Kazanov

October 14, 2025

Overview

01

Problem Statement

02

Data Overview

03

Exploratory Data Analysis (EDA)

04

Data Preprocessing

05

Predictive Modeling and Comparison

06

Conclusions

Problem Statement

CANCELLED BOOKINGS

01

Global distribution systems and online travel agencies open hotels to travellers from around the world. Same flexible tools allow guests to cancel or change their bookings in a matter of minutes.

LOST REVENUE

02

Cancellations often lead to vacancies that can't always be filled promptly, resulting in a direct loss of income for hotel owners.

INCREASED COSTS

03

Online advertising, staffing costs, maintenance are all fixed costs that will be incurred regardless of a guest deciding to stay.

GOAL

04

Leverage existing data to have a repeatable and and reliable prediction of a cancellation risk.
Common strategies:

- Overbooking
- Deposits
- Incentives
- Communications

Data Overview

2 hotels: City Hotel and Resort Hotel, located in Portugal

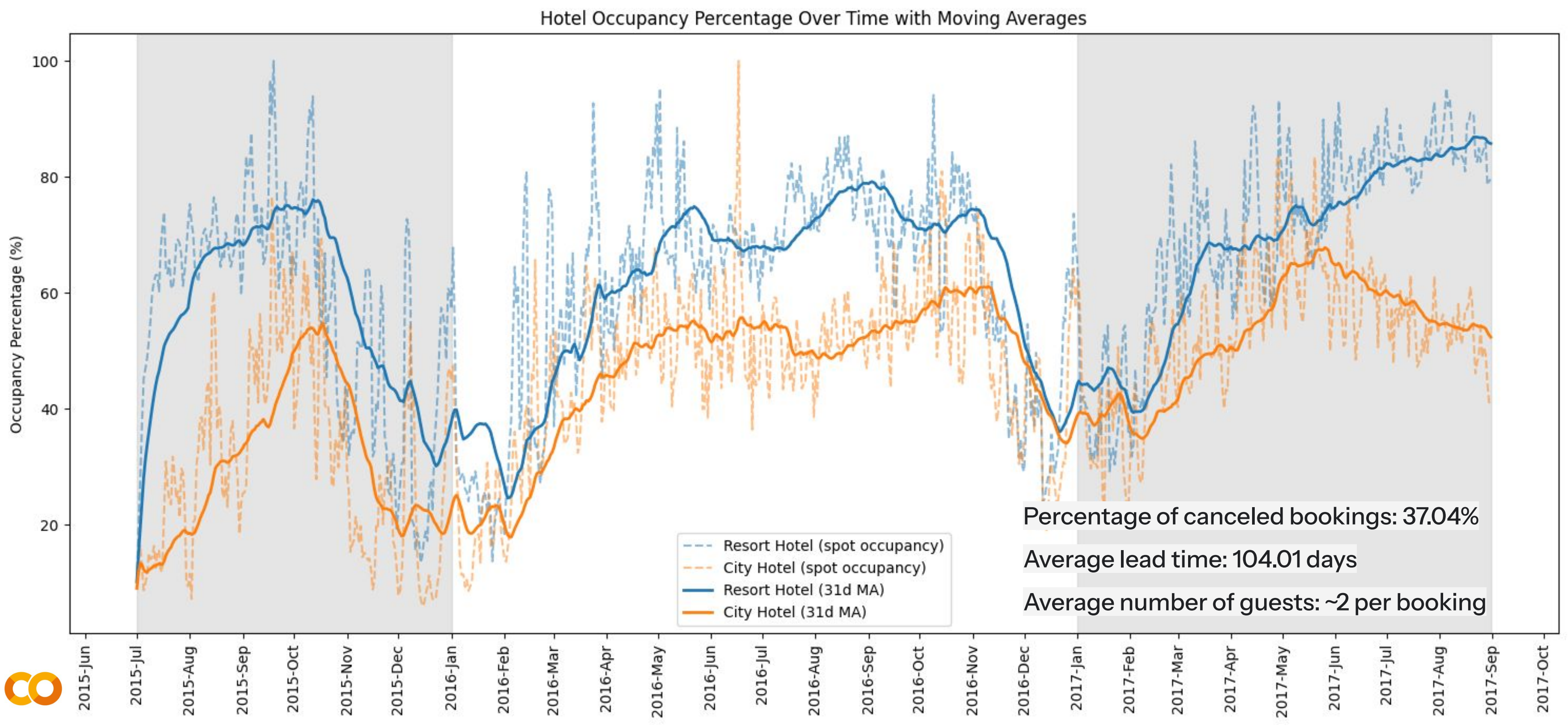
Over **119,000 bookings** from 2015–2017

32 features, including guest and booking details

	Column	Parameter Type	Number of Unique Values	% of Null	Min	Max	Average	Top 5 Categories	% of > 2 St.Dev.
0	hotel	object	2	0.000000	NaN	NaN	NaN	[City Hotel, Resort Hotel]	NaN
1	is_canceled	int64	2	0.000000	0.00	1.0	0.370416	None	0.000000
2	lead_time	int64	479	0.000000	0.00	737.0	104.011416	None	5.083340
3	arrival_date_year	int64	3	0.000000	2015.00	2017.0	2016.156554	None	0.000000
4	arrival_date_month	object	12	0.000000	NaN	NaN	NaN	[August, July, May, October, April]	NaN
5	arrival_date_week_number	int64	53	0.000000	1.00	53.0	27.165173	None	0.000000
6	arrival_date_day_of_month	int64	31	0.000000	1.00	31.0	15.798241	None	0.000000
7	stays_in_weekend_nights	int64	17	0.000000	0.00	19.0	0.927599	None	2.896390
8	stays_in_week_nights	int64	35	0.000000	0.00	50.0	2.500302	None	2.809281
9	adults	int64	14	0.000000	0.00	55.0	1.856403	None	0.402881
10	children	float64	5	0.000034	0.00	10.0	0.103890	None	7.194907
11	babies	int64	5	0.000000	0.00	10.0	0.007949	None	0.768071
12	meal	object	5	0.000000	NaN	NaN	NaN	[BB, HB, SC, Undefined, FB]	NaN

Dataset: https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand
Description: https://www.sciencedirect.com/science/article/pii/S2352340918315191#bib5

Exploratory Data Analysis (EDA)



Percentage of canceled bookings: 37.04%

Average lead time: 104.01 days

Average number of guests: ~2 per booking

Data Preprocessing

Lead time & Seasonality

Lead Time: # of days before arrival
Week Number: 1 thru 53
ADR (normalized by stdev): daily $ rate
Occupancy %: occ. rooms / est. capacity
Days in waiting list: days before confirmed

Booking details

Segment (one hot): Direct, OTA, Corp
Room Type (one hot)
Meal: # of meal per day, 0-3
Parking: # of spaces
Deposit (one hot): Yes/No/Refundable
Special requests: # of requests

Guest Information

Number of guests: Adults & kids
Kids: Yes/No
Country: Domestic or not
Length of Stay: # of days
Weekends: # of days

Previous behavior

Repeated Guest: Yes/No
Previous Cancellations: count
Booking Changes: count

Predictive Modeling and Comparison

Modeling objective

Model accuracy was the primary objective. **Interpretability** would be desirable but **not critical.** The intent is not to identify relationship and be able to influence the outcome but rather to a) estimate cancelation rates to modulate the scale of response and b) identify bookings likely to cancel to target them with individual counter-measures.

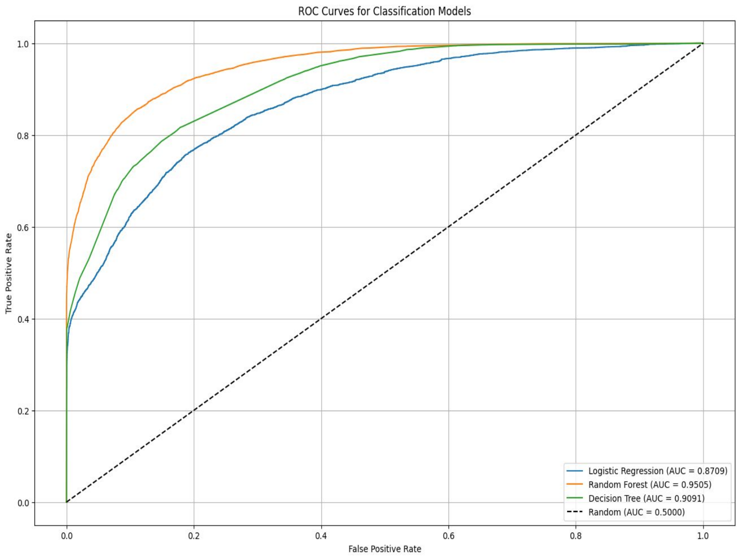
Hyperparameter Tuning

Logistic Regression:
- C: 0.001, 0.01, 0.1, 1, 10, 100*

Random Forest:
- n_estimators: 100, 200*
- max_depth: 10, 20, None*

Decision Tree:
- max_depth: 10*, 20, None
- min_samples_split: 2, 5, 10*

*Note: * best performing parameters*



	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.795376	0.798058	0.596710	0.682851
Random Forest	0.881020	0.867224	0.800227	0.832379
Decision Tree	0.834618	0.804049	0.729892	0.765178

Conclusions

Best model
Random Forest proved to be a superior model for the cancelation risk prediction. Accuracy of around 88% provides a sizable improvement over random guess. With additional cycles of squashing outliers, feature engineering and parameter tuning even higher accuracy is achievable. Accuracy of c. 95% is desirable.

Is this model practical?
Being an ensemble model Random Forest is hard to interpret. If the premise of the project was to find a relationship between booking features and cancelation risk, then the use of the model would be scrutinized further. However, the stated objective was to achieve accurate prediction and identify high risk bookings. Therefore the selected model is practical and may be used by hotel owners and operators.