

RESEARCH ARTICLE

Open Access

Predicting Hotel Booking Cancellations Using CatBoost and SHAP: An Explainable AI Approach Based on 2020–2026 Operational Data

I Made Sudana^{1*}, Endang Wahyu Pamungkas²

^{1,2} Master of Informatics, Universitas Muhammadiyah Surakarta, Sukoharjo Regency, Central Java Province, Indonesia.

*Correspondence email:
l208250008@student.ums.ac.id

Received: 26 July 2026.
Revised: 10 August 2026.
Accepted: 27 August 2026.
Published: 30 August 2026.

Full list of author information is
available at the end of the article.

Abstract

Hotel booking cancellations are a critical problem in hotel revenue management because they can cause operational inefficiencies and financial losses. This study develops an explainable cancellation prediction model using CatBoost integrated with SHAP, based on real-world Property Management System (PMS) operational data from a budget hotel in Central Java, Indonesia. The dataset spans 67 months (October 2020–April 2026) and, after preprocessing and data cleaning, yields 74,826 independent reservation records from 80,110 raw entries. The cancellation rate is extremely low (1.56%), creating a severe class imbalance challenge. Instead of synthetic oversampling, the proposed method applies cost-sensitive learning via CatBoost's `scale_pos_weight`, computed from the natural class ratio. Model performance is evaluated using hold-out validation (80% training, 20% testing). The proposed CatBoost model achieves an F1-score of 72.04%, with precision of 87.20% for the cancellation class and an AUC-ROC of 0.86, outperforming Random Forest and XGBoost baselines. SHAP analysis indicates that lead time, deposit type, and arrival month are the most influential features driving cancellation predictions. These findings support early warning decision-making for risk mitigation in hotel operations.

Keywords: Hotel Booking Cancellation; CatBoost; Class Imbalance; Explainable Artificial Intelligence (XAI); Shapley Additive exPlanations (SHAP).

Abstrak

Pembatalan pemesanan hotel merupakan tantangan penting dalam manajemen pendapatan karena dapat menyebabkan ketidakefisienan operasional dan kerugian finansial. Penelitian ini mengembangkan model prediksi pembatalan yang bersifat dapat dijelaskan (explainable AI) menggunakan CatBoost yang diintegrasikan dengan SHAP, berdasarkan data operasional nyata dari Property Management System (PMS) pada hotel budget di Jawa Tengah. Dataset mencakup 67 bulan, yaitu Oktober 2020 hingga April 2026. Setelah melalui prapemrosesan dan pembersihan data, diperoleh 74.826 catatan reservasi independen dari 80.110 data mentah. Rasio pembatalan yang sangat rendah (1,56%) menciptakan tantangan ketidakseimbangan kelas ekstrem. Untuk mengatasi hal tersebut, penelitian ini menerapkan cost-sensitive learning melalui parameter `scale_pos_weight` pada CatBoost, dengan nilai ditentukan berdasarkan rasio kelas, tanpa menggunakan oversampling sintesis. Model dievaluasi menggunakan hold-out validation (80% pelatihan dan 20% pengujian). Hasil menunjukkan bahwa CatBoost menghasilkan F1-score sebesar 72,04%, precision 87,20% pada kelas pembatalan, dan AUC-ROC sebesar 0,86, yang lebih baik dibandingkan baseline Random Forest dan XGBoost. Analisis SHAP mengindikasikan bahwa lead time, deposit type, serta arrival month merupakan fitur utama yang berkontribusi pada prediksi pembatalan, sehingga mendukung pengambilan keputusan berbasis peringatan dini.

Kata Kunci: Pembatalan Pemesanan Hotel; CatBoost; Ketidakseimbangan Kelas; Explainable Artificial Intelligence (XAI); Shapley Additive exPlanations (SHAP).



1. Pendahuluan

Sektor perhotelan modern beroperasi dalam lingkungan yang dinamis dengan tingkat ketidakpastian yang tinggi, terutama terkait perilaku pembatalan reservasi oleh tamu. Fenomena ini menjadi risiko operasional penting karena dapat mengganggu optimalisasi manajemen pendapatan (*revenue management*), menurunkan akurasi prakiraan tingkat hunian, dan menyebabkan hilangnya peluang penjualan kamar (*opportunity costs*) (Antonio *et al.*, 2019a). Dalam domain komputasi, pembatalan reservasi dapat diformulasikan sebagai masalah klasifikasi biner pada data tabular yang heterogen. Meskipun adopsi *Property Management System* (PMS) telah memfasilitasi pengumpulan data reservasi dalam jumlah besar, data tersebut sering kali masih terbatas digunakan untuk pelaporan administratif dan belum dimanfaatkan secara optimal untuk analisis prediktif. Pemanfaatan *machine learning* menjadi pendekatan yang relevan untuk mengidentifikasi pola operasional dan memperkirakan probabilitas pembatalan reservasi. Studi terdahulu telah menggunakan berbagai algoritma klasifikasi untuk tugas ini, termasuk model ensemble berbasis pohon seperti *Extreme Gradient Boosting* (XGBoost) dan Random Forest yang mampu menghasilkan akurasi global tinggi (Yoo *et al.*, 2023). Namun, masih terdapat beberapa celah penelitian. Pertama, sebagian besar studi masih bergantung pada dataset sekunder dari repositori publik, seperti dataset yang digunakan oleh Antonio *et al.* (2019b), sehingga belum sepenuhnya merepresentasikan perilaku konsumen lokal dan dinamika pasar pascapandemi yang dipengaruhi volatilitas makroekonomi (Yuan *et al.*, 2022). Kedua, algoritma boosting standar sering memerlukan *categorical encoding* manual, seperti *one-hot encoding*. Pada fitur perhotelan berkardinalitas tinggi, seperti kota asal tamu atau agen perjalanan daring (*online travel agent*/OTA), proses encoding tersebut dapat meningkatkan dimensi data secara substansial, menambah beban komputasi, dan meningkatkan risiko *overfitting*. Untuk menjawab celah tersebut, penelitian ini mengusulkan model prediksi pembatalan reservasi menggunakan CatBoost (*Categorical Boosting*) yang diintegrasikan dengan SHAP (*Shapley Additive exPlanations*). CatBoost menggunakan mekanisme *ordered boosting* yang dirancang untuk memitigasi *target leakage* dan *prediction shift*, sehingga sesuai untuk data operasional yang kompleks dan beragam (Prokhorenkova *et al.*, 2018). Algoritma ini juga dapat memproses fitur kategorikal secara *native* tanpa *one-hot encoding*, sehingga struktur informasi kategorikal dalam data dapat dipertahankan. Secara empiris, penelitian ini menggunakan dataset longitudinal primer yang diekstraksi dari PMS sebuah hotel budget di Jawa Tengah, mencakup periode 67 bulan dari Oktober 2020 hingga April 2026. Cakupan temporal ini penting karena merepresentasikan perubahan perilaku reservasi dari fase pemulihan pascapandemi pada 2020–2022 hingga fase normalisasi pasar pada 2023–2026. Tantangan utama dalam dataset ini adalah ketidakseimbangan kelas (*class imbalance*) yang ekstrem, dengan rasio pembatalan hanya sebesar 1,56%. Berbeda dari pendekatan yang mengubah distribusi data melalui *oversampling* sintetis seperti SMOTE, penelitian ini menerapkan *cost-sensitive learning* melalui parameter *scale_pos_weight* dalam arsitektur CatBoost. Parameter tersebut ditetapkan berdasarkan rasio kelas mayoritas terhadap kelas minoritas, sehingga model memberikan bobot penalti yang lebih besar terhadap kesalahan klasifikasi pada kelas pembatalan tanpa mengubah distribusi data empiris. Selain itu, untuk meningkatkan interpretabilitas model, penelitian ini mengintegrasikan SHAP sebagai pendekatan *Explainable Artificial Intelligence* (XAI) (Lundberg & Lee, 2017). SHAP menguraikan kontribusi setiap fitur terhadap output model, baik pada tingkat global maupun lokal. Dengan demikian, integrasi CatBoost dan SHAP diharapkan dapat menghasilkan model prediktif yang akurat sekaligus memberikan informasi yang dapat mendukung pengambilan keputusan manajerial dalam mitigasi risiko pembatalan reservasi hotel.

2. Method

Penelitian ini menggunakan kerangka kerja eksperimental berbasis *data science* untuk mengolah data operasional perhotelan mentah menjadi model prediktif yang dapat diinterpretasikan. Alur penelitian mengadaptasi *Cross-Industry Standard Process for Data Mining* (CRISP-DM), yang mencakup akuisisi data, prapemrosesan, rekayasa fitur, pelatihan model berbasis *cost-sensitive learning*, perbandingan model, dan interpretasi menggunakan *Explainable Artificial Intelligence* (XAI). Alur penelitian ditampilkan pada Figure 1.

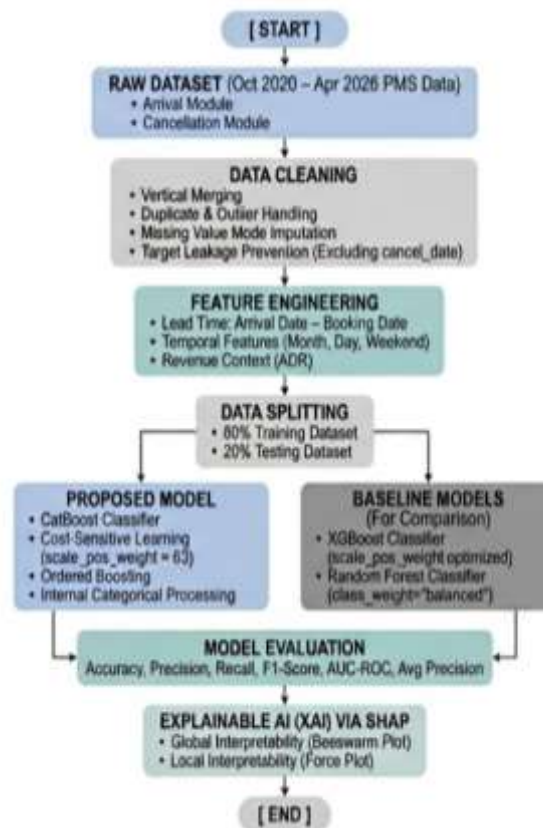


Figure 1. Proposed Research Workflow for Hotel Cancellation Prediction

Data primer diekstraksi dari *Property Management System* (PMS) sebuah hotel budget di Jawa Tengah. Dataset longitudinal ini mencakup periode 67 bulan, dari Oktober 2020 hingga April 2026. Cakupan temporal ini merepresentasikan dua periode utama, yaitu fase pemulihan pascapandemi pada 2020–2022 dan fase normalisasi pasar pada 2023–2026. Data mentah awal terdiri atas dua sumber utama, yaitu *Arrival module* dan *Cancellation module*. Kedua sumber data tersebut diintegrasikan ke dalam satu *dataframe* terpadu dengan variabel target biner `is_canceled`, yaitu 0 untuk reservasi yang terealisasi (*check-in*) dan 1 untuk reservasi yang dibatalkan. Prapemrosesan dilakukan untuk meningkatkan kualitas data sebelum pemodelan. Tahapan ini mencakup pembersihan data, imputasi nilai kosong, dan pencegahan *target leakage*. Pencarian (*outliers*) dihapus menggunakan metode Z-score dengan ambang batas $Z > 3$. Metode ini diterapkan pada fitur `adr` untuk mengeliminasi tarif kamar tidak rasional, termasuk nilai Rp0, serta pada fitur `adult`. Selain itu, nilai tidak standar seperti UNKNOWN pada kolom `guest_city` ditangani menggunakan imputasi modus berbasis kluster wilayah Jawa Tengah untuk mengurangi kehilangan data akibat penghapusan baris. Pencegahan *target leakage* dilakukan dengan memastikan variabel `deposit_type` merepresentasikan kebijakan deposit pada saat reservasi dibuat, bukan status akhir setelah pembatalan. Variabel yang hanya tersedia setelah pembatalan terjadi, seperti `cancel_date` dan `cancel_reason`, dihapus dari matriks fitur pelatihan. Setelah pembersihan dan standarisasi, jumlah data berkurang dari 80.110 rekaman mentah menjadi 74.826 catatan reservasi independen. Struktur fitur yang digunakan dalam penelitian ini disajikan pada Table 1.

Table 1. Feature Groupings and Variable Descriptions

Feature Group	Variable Name	Data Type	Description / Unique Values
Target Variable	<code>is_canceled</code>	Categorical (Binary)	The operational target for prediction: 0 (Check-In / Realized) or 1 (Canceled).
Temporal Characteristics	<code>lead_time</code>	Numeric	The number of days between the reservation date and the scheduled arrival date.
	<code>stay_duration</code>	Numeric	Total number of scheduled nights the guest booked to stay at the property.
Transactional Characteristics	<code>market_segment</code>	Categorical	The targeted market segment of the booking (e.g., Groups, Online TA, Corporate).

	deposit_type	Categorical	The transaction deposit policy applied to the booking (e.g., Non Refund, No Deposit).
	customer_type	Categorical	The profiling category of the customer (e.g., Transient, Group).
	distribution_channel	Categorical	The distribution channel through which the room reservation was secured.
Service Preferences	meal_plan	Categorical	The specific meal package or boarding arrangement selected by the guest.
	room_type	Categorical	The reserved room category, tier, or type defined by the hotel management.
	room_quantity	Numeric	The total number of room units reserved under a single booking record ID.
	adult	Numeric	The total number of registered adult guests included in the reservation.
Demographic Characteristics	guest_city	Categorical	The city of origin where the prospective customer officially resides.
	guest_region	Categorical	The regional province origin of the guest within Central Java.
	guest_country	Categorical	The country of origin of the guest (predominantly domestic customers).
Financial Metric	adr	Numeric	Average Daily Rate (The average conditional revenue generated per occupied room per day).

Rekayasa fitur dilakukan untuk memperkaya representasi variabel yang digunakan dalam pemodelan (Chen *et al.*, 2023). Fitur utama yang diekstraksi meliputi *lead_time*, fitur temporal, dan konteks pendapatan. *lead_time* dihitung sebagai selisih hari antara tanggal kedatangan dan tanggal pemesanan (Sánchez-Medina & C-Sánchez, 2020), sebagaimana ditunjukkan pada Persamaan (1).

$$\text{LeadTime} = \text{ArrivalDate} - \text{BookingDate} \quad (1)$$

Fitur temporal diekstraksi dari tanggal kedatangan, meliputi bulan kedatangan, hari kedatangan, dan indikator akhir pekan (*weekend*) untuk merepresentasikan pola musiman perilaku tamu (Yoo *et al.*, 2023). Selain itu, *Average Daily Rate* (ADR) digunakan sebagai fitur finansial untuk merepresentasikan konteks harga dalam prediksi pembatalan reservasi (Antonio *et al.*, 2019a). Dataset akhir terdiri atas 74.826 catatan reservasi, dengan distribusi 73.659 instans *check-in* (98,44%) dan 1.167 instans pembatalan (1,56%). Dataset kemudian dipisahkan menjadi 80% data pelatihan dan 20% data pengujian independen. Untuk menangani ketidakseimbangan kelas tanpa menggunakan *oversampling* sintesis pada model utama, penelitian ini menerapkan *cost-sensitive learning* dalam arsitektur CatBoost Classifier melalui parameter *scale_pos_weight*. Nilai *scale_pos_weight* ditetapkan sekitar 63 berdasarkan rasio kelas mayoritas terhadap kelas minoritas. Pembobotan ini memberikan penalti lebih besar terhadap kesalahan klasifikasi pada kelas minoritas, sehingga model lebih peka terhadap pola pembatalan selama proses pelatihan. Sebagai pembandingan, model CatBoost diuji terhadap XGBoost Classifier dan Random Forest Classifier. XGBoost dikonfigurasi dengan *learning rate* 0,05 dan *max_depth* = 6, sedangkan Random Forest menggunakan 500 estimator. Berbeda dengan CatBoost yang menggunakan pembobotan internal, kedua model baseline dikombinasikan dengan SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menangani ketidakseimbangan kelas secara eksternal. Setelah model terbaik diperoleh, SHAP digunakan untuk menginterpretasikan kontribusi fitur pada tingkat global dan lokal.

3. Results and Discussion

3.1 Results

Analisis data eksploratif atau *Exploratory Data Analysis* (EDA) dilakukan untuk mengidentifikasi karakteristik intrinsik data, memetakan distribusi variabel prediktor utama, serta mengamati pola awal yang berkaitan dengan perilaku pembatalan reservasi. Tahap ini penting untuk memberikan landasan kontekstual sebelum model *machine learning* diterapkan.

3.1.1 Analisis Tren Temporal Musiman

Analisis terhadap volume reservasi dan pembatalan bulanan selama periode Oktober 2020 hingga April 2026 ditampilkan pada Figure 2 dan Figure 3. Kedua visualisasi tersebut menunjukkan bahwa pola pembatalan tidak selalu bergerak linier dengan volume pemesanan. Dengan kata lain, bulan dengan jumlah reservasi tinggi tidak selalu menjadi bulan dengan jumlah pembatalan tertinggi.



Figure 2. Monthly Room Booking Volume Trends, 2020–2026



Figure 3. Monthly Room Cancellation Trends, 2020–2026

Berdasarkan Figure 2, Desember merupakan bulan dengan volume reservasi tertinggi, yaitu 7.814 transaksi. Sebaliknya, November mencatat volume reservasi yang lebih rendah, yaitu 5.637 transaksi. Namun, Figure 3 menunjukkan bahwa jumlah pembatalan justru terjadi pada November dengan 199 kasus, lebih tinggi dibandingkan Desember yang mencatat 176 kasus. Pola ini mengindikasikan bahwa risiko pembatalan dapat meningkat menjelang periode puncak perjalanan akhir tahun. Fenomena tersebut dapat dikaitkan dengan perilaku pemesanan lebih awal untuk mengamankan ketersediaan kamar pada periode liburan akhir tahun. Menjelang bulan November, calon tamu berpotensi melakukan evaluasi ulang terhadap agenda perjalanan, kepastian cuti, anggaran, atau kewajiban pembayaran deposit. Kondisi ini dapat memicu peningkatan pembatalan sebelum memasuki periode puncak kunjungan pada bulan Desember. Temuan ini sejalan dengan Yoo *et al.* (2023) dan Yang & Miao (2024), yang menunjukkan bahwa pola musiman dan dinamika kalender dapat memengaruhi stabilitas reservasi hotel. Pola lain terlihat pada bulan Agustus, yang mencatat 134 kasus pembatalan meskipun volume reservasi pada bulan tersebut relatif lebih rendah, yaitu 5.452 reservasi. Kondisi ini dapat mengindikasikan adanya pembatalan menjelang tanggal kedatangan yang berkaitan dengan perubahan prioritas perjalanan, kebutuhan rumah tangga, atau penyesuaian anggaran keluarga. Namun, interpretasi ini perlu dibaca secara hati-hati karena dataset tidak secara langsung memuat alasan subjektif pembatalan tamu. Sebaliknya, jumlah pembatalan relatif rendah ditemukan pada kuartal pertama, khususnya Maret dengan 35 kasus dan April dengan 57 kasus. Rendahnya pembatalan pada periode ini dapat berkaitan dengan pola perjalanan bisnis, perjalanan dinas, atau reservasi dengan jadwal yang lebih terencana. Secara umum, pola temporal tersebut menunjukkan bahwa fitur waktu, seperti *arrival_month*, memiliki potensi kontribusi penting dalam membedakan periode dengan risiko pembatalan yang lebih tinggi dan lebih rendah.

3.1.2 Analisis Distribusi Variabel Waktu Tunggu

Variabel *lead_time* merepresentasikan selisih hari antara tanggal pemesanan dan tanggal kedatangan. Distribusi variabel ini dibandingkan antara kelas Check-In dan Canceled menggunakan kombinasi box plot dan Kernel Density Estimate (KDE), sebagaimana ditampilkan pada Figure 4. Untuk memperkuat temuan distribusi *lead_time*, perlu dicatat bahwa variabel ini merepresentasikan *jendela keputusan* tamu sejak reservasi dibuat hingga tanggal kedatangan. Semakin panjang *lead_time*, semakin besar peluang terjadinya perubahan rencana karena tamu memiliki waktu lebih untuk mengevaluasi alternatif akomodasi, menyesuaikan jadwal perjalanan, maupun mempertimbangkan kondisi finansial. Implikasinya, *lead_time* tidak hanya berfungsi sebagai penanda jarak waktu, tetapi juga sebagai proksi terhadap tingkat fleksibilitas keputusan konsumen. Selain itu, sebaran *lead_time* yang lebih melebar pada kelas *Canceled* mengindikasikan bahwa pembatalan dapat terjadi pada berbagai fase waktu tunggu, baik jauh sebelum maupun mendekati tanggal kedatangan. Kondisi ini menegaskan bahwa risiko pembatalan bersifat dinamis dan tidak terbatas pada satu periode tertentu. Dengan demikian, keberadaan pola tersebut memberi dukungan empiris bahwa model memerlukan fitur waktu untuk dapat membedakan reservasi berisiko tinggi

dan berisiko rendah secara lebih akurat.

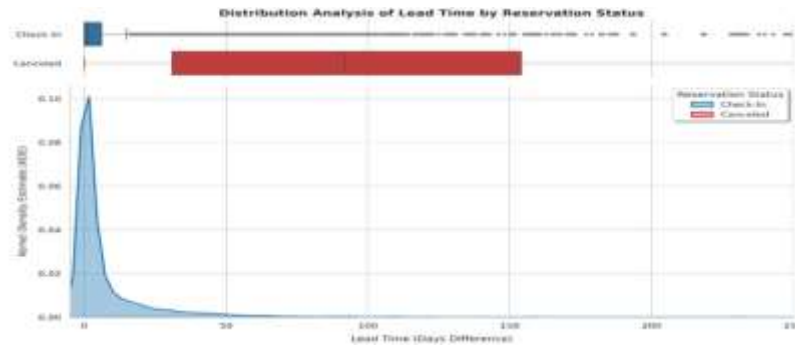


Figure 4. Distribution Analysis of Lead Time Across Reservation Status Using Box Plot and KDE

Figure 4 menunjukkan bahwa kelas Check-In memiliki distribusi yang terkonsentrasi pada rentang lead_time pendek, terutama 0 hingga 14 hari. Kurva KDE kelas Check-In membentuk puncak yang tajam di area kiri grafik, sedangkan box plot menunjukkan bahwa nilai kuartil bawah, median, dan kuartil atas berada dekat dengan nilai rendah. Hal ini mengindikasikan bahwa reservasi yang terealisasi cenderung berasal dari pemesanan dengan jarak waktu pendek sebelum tanggal kedatangan. Sebaliknya, kelas Canceled memiliki distribusi yang lebih melebar. Median lead_time pada kelas ini berada sekitar 92 hari, dengan rentang interkuartil sekitar 33 hingga 155 hari. Pola KDE yang lebih landai menunjukkan bahwa pembatalan tersebar pada rentang waktu tunggu yang lebih panjang. Secara konseptual, semakin panjang jendela pemesanan, semakin besar peluang terjadinya perubahan agenda, kendala transportasi, perubahan kondisi finansial, atau evaluasi ulang pilihan akomodasi. Rentang lead_time yang panjang juga dapat memberi ruang bagi konsumen untuk membandingkan harga melalui berbagai platform online travel agent (OTA). Interpretasi ini selaras dengan hasil SHAP pada bagian berikutnya, yang menunjukkan bahwa lead_time menjadi fitur dengan kontribusi terbesar terhadap prediksi pembatalan. Meskipun demikian, perilaku price-shopping perlu dipahami sebagai kemungkinan penjelasan, bukan kesimpulan kausal langsung, karena aktivitas perbandingan harga tidak diamati secara eksplisit dalam dataset. Dengan demikian, hasil EDA menunjukkan bahwa lead_time merupakan fitur penting dalam membedakan reservasi yang terealisasi dan dibatalkan. Temuan ini konsisten dengan Antonio *et al.* (2017), yang menyatakan bahwa lead time panjang dapat meningkatkan fleksibilitas konsumen untuk membatalkan reservasi.

3.1.3 Analisis Parameter Finansial: Average Daily Rate

Average Daily Rate (ADR) merepresentasikan rata-rata tarif harian kamar. Distribusi ADR dianalisis menggunakan kombinasi box plot dan KDE, sebagaimana ditampilkan pada Figure 5.

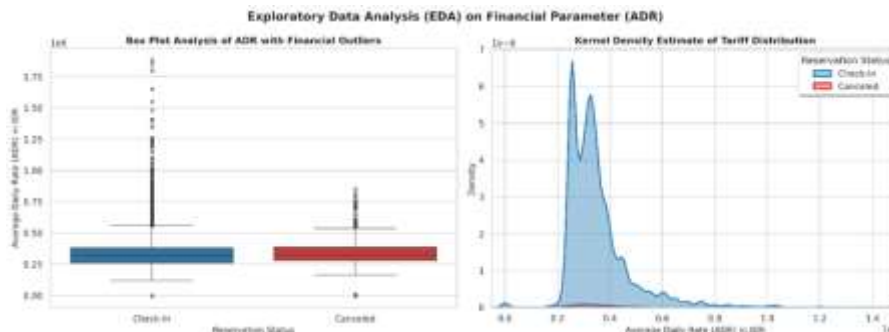


Figure 5. Exploratory Data Analysis on Average Daily Rate

Figure 5 menunjukkan bahwa sebagian besar data ADR terkonsentrasi pada rentang Rp250.000 hingga Rp400.000. Rentang ini sesuai dengan karakteristik objek penelitian sebagai hotel budget yang melayani segmen konsumen sensitif harga. Pada rentang tarif reguler, distribusi ADR antara kelas Check-In dan Canceled tampak relatif mirip. Hal ini menunjukkan bahwa pembatalan tidak dapat dijelaskan hanya melalui nilai tarif awal, tetapi kemungkinan dipengaruhi oleh interaksi antara ADR, lead_time, kanal pemesanan, dan kebijakan deposit. Visualisasi juga menunjukkan adanya nilai ADR sebesar Rp0. Nilai ini dapat merepresentasikan transaksi non-komersial, seperti kamar complimentary, promosi internal, kompensasi

layanan, atau transaksi internal manajemen. Dalam konteks pemodelan prediktif, nilai tersebut perlu ditangani secara hati-hati karena tidak merepresentasikan tarif pasar reguler. Selain itu, terdapat pencilan atas pada rentang sekitar Rp1.500.000 hingga Rp1.850.000. Nilai tersebut tidak lazim untuk kategori hotel budget dan dapat berasal dari kesalahan input, penggabungan biaya fasilitas non-kamar, atau transaksi khusus. Temuan ini mendukung perlunya pembersihan pencilan pada tahap prapemrosesan menggunakan metode Z-score. Dengan pembersihan pencilan tersebut, model diharapkan dapat menangkap pola tarif yang lebih representatif terhadap kondisi pasar sebenarnya dan mengurangi risiko distorsi pada proses pembelajaran algoritma. Secara umum, hasil analisis ADR menunjukkan bahwa variabel harga tetap memiliki relevansi dalam prediksi pembatalan, tetapi pengaruhnya tidak berdiri sendiri. Pada konteks hotel budget, keputusan pembatalan lebih tepat dipahami sebagai hasil interaksi antara harga, waktu tunggu, kanal pemesanan, dan fleksibilitas kebijakan transaksi.

3.1.4 Karakteristik Transaksional

Analisis karakteristik transaksional dilakukan untuk melihat hubungan antara segmen pasar, kebijakan deposit, dan status pembatalan. Ringkasan distribusi historis kedua fitur tersebut disajikan pada Table 2.

Table 2. Bivariate Cross-Tabulation Analysis of Booking Characteristics and Cancellation Rates

Feature Category	Variable Attribute	Total Bookings	Check-In	Cancellations	Cancellation Ratio
Market Segment	Online TA	51,137	50,425	712	1.39%
Market Segment	Direct	14,842	14,590	252	1.70%
Market Segment	Corporate	4,295	4,217	78	1.82%
Market Segment	Government	2,260	2,223	37	1.64%
Market Segment	Group	1,441	1,407	34	2.36%
Market Segment	Offline TA	593	568	25	4.22%
Market Segment	Compliment	258	229	29	11.24%
Total Market Segment		74,826	73,659	1,167	1.56%
Deposit Type	Deposit	64,310	63,903	407	0.63%
Deposit Type	Refundable	7,487	6,881	606	8.09%
Deposit Type	No Deposit	3,008	2,875	133	4.42%
Deposit Type	Non-Refund	21	0	21	100.00%
Total Deposit Type		74,826	73,659	1,167	1.56%

Berdasarkan Table 2, reservasi didominasi oleh segmen Online Travel Agent atau Online TA dengan 51.137 reservasi, diikuti oleh kanal Direct sebanyak 14.842 reservasi. Secara keseluruhan, dataset terdiri atas 74.826 catatan reservasi, dengan 73.659 reservasi terealisasi sebagai Check-In dan 1.167 reservasi dibatalkan. Dengan demikian, rasio pembatalan keseluruhan adalah 1,56%. Pada kategori *market segment*, rasio pembatalan tertinggi ditemukan pada segmen Compliment, yaitu 11,24%, meskipun jumlah sampelnya relatif kecil, yaitu 258 reservasi. Hal ini menunjukkan bahwa reservasi tanpa beban biaya langsung berpotensi memiliki komitmen kedatangan yang lebih rendah. Namun, karena jumlah observasi pada kategori ini terbatas, interpretasi terhadap segmen Compliment perlu dilakukan secara hati-hati. Segmen Offline TA juga menunjukkan rasio pembatalan relatif tinggi, yaitu 4,22%, diikuti oleh Group sebesar 2,36%. Sementara itu, segmen Government memiliki rasio pembatalan sebesar 1,64%, yang relatif dekat dengan rasio pembatalan keseluruhan dataset. Pola ini menunjukkan bahwa karakteristik kanal pemesanan dapat memberikan informasi tambahan bagi model dalam membedakan reservasi yang berisiko batal dan yang cenderung terealisasi. Pada fitur kebijakan deposit, kategori Deposit memiliki jumlah reservasi terbesar, yaitu 64.310 reservasi, dengan rasio pembatalan paling rendah sebesar 0,63%. Temuan ini menunjukkan bahwa pembayaran uang muka berkaitan dengan tingkat komitmen reservasi yang lebih tinggi. Sebaliknya, kategori Refundable dan No Deposit menunjukkan rasio pembatalan lebih tinggi, masing-masing sebesar 8,09% dan 4,42%. Kategori Non-Refund mencatat 21 reservasi dan seluruhnya berada pada status pembatalan. Karena jumlah sampel pada kategori ini sangat kecil, rasio 100,00% tersebut tidak dapat digeneralisasi secara luas. Namun, kategori ini tetap memberikan sinyal yang perlu diperhatikan dalam pemodelan karena menunjukkan pola transaksi yang berbeda dari kategori deposit lainnya.

3.1.5 Pengujian Komparatif Performa Model

Pengujian model dilakukan menggunakan skema *hold-out validation* dengan proporsi 80% data pelatihan dan 20% data pengujian. Model CatBoost dibandingkan dengan dua baseline, yaitu Random Forest

+ SMOTE dan XGBoost + SMOTE. CatBoost dikonfigurasi dengan 500 iterasi, *learning rate* 0,05, dan kedalaman pohon 6. Hasil evaluasi ditampilkan pada Table 3.

Table 3. Evaluation Metrics Comparison of Predictive Model Performance

Evaluation Metrics	Random Forest + SMOTE	XGBoost + SMOTE	CatBoost + Weighting	Delta CatBoost vs XGBoost
Accuracy	95.86%	97.09%	99.25%	+2.16%
Precision (Canceled)	24.83%	32.80%	87.20%	+54.40%
Recall (Canceled)	79.83%	79.83%	61.37%	-18.46%
F1-score (Canceled)	37.88%	46.50%	72.04%	+25.54%
AUC-ROC	0.82	0.84	0.86	+0.02

Table 3 menunjukkan bahwa model baseline yang menggunakan SMOTE menghasilkan *recall* tinggi sebesar 79,83% pada kelas pembatalan. Hal ini menunjukkan bahwa Random Forest + SMOTE dan XGBoost + SMOTE mampu menangkap sebagian besar kasus pembatalan aktual. Namun, nilai *precision* kedua baseline relatif rendah, yaitu 24,83% pada Random Forest dan 32,80% pada XGBoost. Kondisi ini menunjukkan bahwa model baseline cenderung menghasilkan lebih banyak *false positive*, yaitu reservasi yang diprediksi batal tetapi sebenarnya terealisasi sebagai Check-In. Model CatBoost dengan pembobotan internal menghasilkan *precision* tertinggi sebesar 87,20% dan F1-score sebesar 72,04% pada kelas pembatalan. Meskipun *recall* CatBoost lebih rendah daripada XGBoost, yaitu 61,37%, nilai F1-score yang lebih tinggi menunjukkan keseimbangan yang lebih baik antara *precision* dan *recall*. Dengan demikian, CatBoost lebih mampu menjaga akurasi prediksi pada kelas pembatalan tanpa meningkatkan jumlah salah alarm secara berlebihan. Dalam konteks operasional hotel, *precision* yang tinggi penting untuk mengurangi risiko salah alarm ketika prediksi digunakan sebagai dasar tindakan seperti konfirmasi ulang, penyesuaian deposit, atau pengelolaan ketersediaan kamar. Jika model terlalu banyak menghasilkan *false positive*, manajemen berisiko menganggap tamu yang sebenarnya akan datang sebagai tamu yang akan membatalkan reservasi. Hal ini dapat berdampak pada keputusan operasional yang kurang tepat, termasuk risiko pengelolaan inventori kamar yang terlalu agresif. Sebaliknya, konfigurasi CatBoost menunjukkan pendekatan yang lebih konservatif dan sesuai untuk konteks hotel, karena mampu menekan *false positive* sambil tetap mempertahankan kemampuan deteksi pembatalan yang cukup baik. Hasil ini menunjukkan bahwa *cost-sensitive learning* melalui parameter *scale_pos_weight* dapat menjadi alternatif yang efektif dibandingkan penggunaan *oversampling* sintesis pada data pembatalan yang sangat tidak seimbang.

3.1.6 Analisis Kurva ROC dan Precision–Recall

Evaluasi model CatBoost diperluas menggunakan kurva Receiver Operating Characteristic (ROC) dan Precision–Recall (PR). Evaluasi ini digunakan untuk menilai kemampuan diskriminasi model pada berbagai ambang klasifikasi. Hasilnya ditampilkan pada Figure 6.

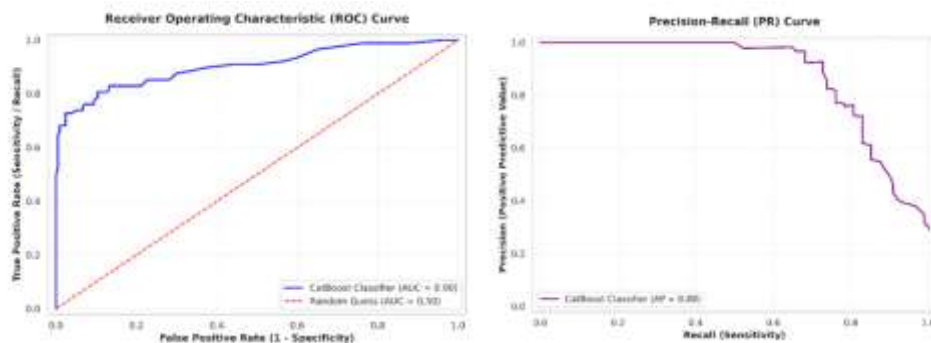


Figure 6. Receiver Operating Characteristic and Precision–Recall Curves

Figure 6 menunjukkan bahwa CatBoost memperoleh nilai AUC-ROC sebesar 0,86. Nilai ini menunjukkan bahwa model memiliki kemampuan diskriminasi yang baik dalam membedakan reservasi yang terealisasi dan dibatalkan. Kurva ROC menggambarkan hubungan antara true positive rate dan false positive rate pada berbagai ambang klasifikasi, sehingga memberikan gambaran umum mengenai stabilitas performa model. Selain ROC, kurva Precision–Recall digunakan karena lebih informatif pada kondisi data tidak seimbang, terutama ketika fokus evaluasi berada pada kelas minoritas pembatalan. Dalam dataset penelitian ini, kelas

Canceled hanya mencakup 1,56% dari total data, sehingga evaluasi berbasis akurasi atau ROC saja belum cukup untuk menilai performa model secara menyeluruh. Nilai Average Precision (AP) sebesar 0,77 menunjukkan bahwa model mampu mempertahankan performa yang cukup baik pada kelas minoritas. Nilai AP yang lebih rendah daripada AUC-ROC merupakan hal yang umum pada data tidak seimbang karena kurva PR lebih sensitif terhadap performa pada kelas positif dan tidak memasukkan true negative dalam perhitungannya. Dengan demikian, kombinasi AUC-ROC sebesar 0,86 dan AP sebesar 0,77 menunjukkan bahwa CatBoost memiliki kemampuan diskriminasi yang baik sekaligus tetap relevan untuk mendeteksi kelas pembatalan yang jumlahnya sangat kecil.

3.1.7 Interpretasi Global Fitur Menggunakan SHAP

Untuk meningkatkan interpretabilitas model, penelitian ini menggunakan SHAP (Shapley Additive exPlanations). SHAP digunakan untuk menguraikan kontribusi setiap fitur terhadap prediksi model. Visualisasi global ditampilkan melalui SHAP beeswarm summary plot pada Figure 7.

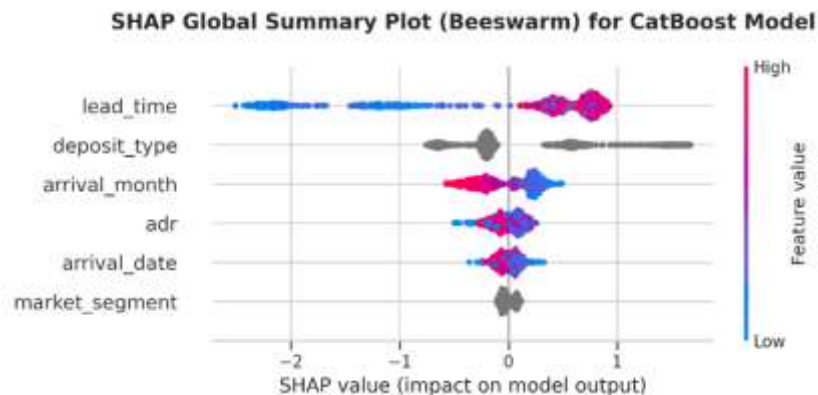


Figure 7. SHAP Global Summary Plot for CatBoost Model Interpretation

Berdasarkan Figure 7, *lead_time* menjadi fitur dengan kontribusi terbesar terhadap prediksi pembatalan. Nilai *lead_time* yang tinggi cenderung berada pada sisi SHAP positif, yang menunjukkan peningkatan kontribusi terhadap prediksi kelas Canceled. Sebaliknya, nilai *lead_time* yang rendah cenderung berada pada sisi SHAP negatif, yang menunjukkan kontribusi terhadap prediksi Check-In. Temuan ini memperkuat hasil EDA sebelumnya bahwa jendela pemesanan yang lebih panjang berkaitan dengan risiko pembatalan yang lebih tinggi. Fitur *arrival_month* juga memberikan kontribusi penting, tetapi dengan pola yang lebih non-linear. Hal ini menunjukkan bahwa bulan kedatangan berperan sebagai konteks temporal yang berinteraksi dengan fitur lain, bukan sebagai penentu tunggal pembatalan. Peran *arrival_month* sejalan dengan pola musiman yang terlihat pada analisis tren temporal, terutama adanya peningkatan risiko pembatalan pada bulan-bulan tertentu menjelang periode puncak perjalanan. Fitur *deposit_type* menunjukkan kontribusi terhadap prediksi pembatalan pada kategori tertentu, terutama ketika reservasi tidak disertai jaminan pembayaran awal atau memiliki fleksibilitas pengembalian. Hal ini konsisten dengan hasil tabulasi karakteristik transaksional yang menunjukkan bahwa kategori Refundable dan No Deposit memiliki rasio pembatalan lebih tinggi dibandingkan kategori Deposit. Sementara itu, fitur *adr* menunjukkan bahwa variasi tarif dapat berkontribusi terhadap prediksi model, meskipun interpretasinya perlu mempertimbangkan interaksi dengan *lead_time*, kanal pemesanan, dan kebijakan deposit. Pada konteks hotel budget, tarif kamar dapat berkaitan dengan sensitivitas harga konsumen, tetapi pengaruhnya tidak dapat dipisahkan dari kondisi transaksi dan waktu pemesanan. Dengan demikian, SHAP membantu menjelaskan fitur yang paling berkontribusi terhadap prediksi model. Namun, interpretasi SHAP tetap merupakan atribusi terhadap output model, bukan bukti kausal langsung terhadap keputusan pembatalan konsumen.

3.1.8 Validasi Granular Melalui Confusion Matrix

Evaluasi kesalahan klasifikasi model CatBoost divisualisasikan melalui confusion matrix pada Figure 8. Visualisasi ini digunakan untuk menunjukkan distribusi prediksi benar dan salah pada kelas Check-In dan Canceled. Confusion matrix pada Gambar 8 juga mengindikasikan bahwa mayoritas kesalahan klasifikasi model berasal dari kelas pembatalan (*Canceled*) yang terprediksi sebagai *Check-In* (false negative = 451). Pola ini memperjelas bahwa, meskipun CatBoost mampu menekan *false positive* sehingga precision kelas *Canceled* tinggi (87,20%), model masih belum sepenuhnya menangkap seluruh variasi karakteristik

reservasi yang akhirnya dibatalkan. Hal ini wajar pada situasi ketidakseimbangan kelas ekstrem (1,56%), karena model harus belajar dari sedikit sampel pembatalan, sehingga keputusan batas (decision boundary) cenderung lebih konservatif: lebih banyak kasus pembatalan “terlewat” dibandingkan menghasilkan alarm pembatalan yang keliru.

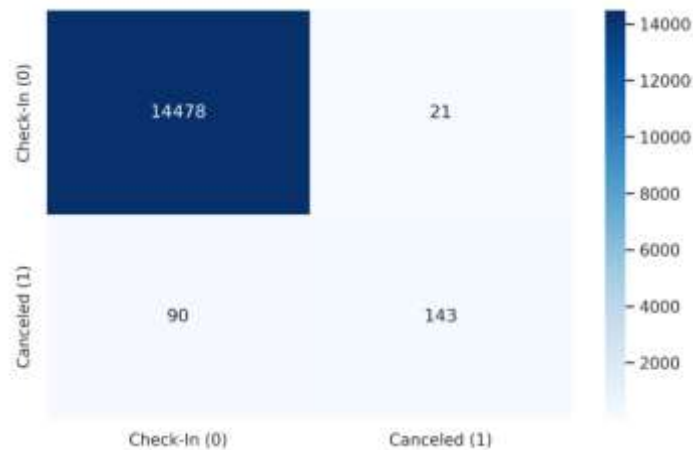


Figure 8. Confusion Matrix of the Final CatBoost Classifier Model

Confusion matrix model final disusun berdasarkan keseluruhan data akhir sebanyak 74.826 catatan reservasi yang telah melalui proses prapemrosesan. Hasil confusion matrix menunjukkan bahwa model mencatat true negative sebanyak 73.554, yaitu reservasi yang diprediksi sebagai Check-In dan benar-benar terealisasi. Nilai ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengenali kelas mayoritas. Model juga mencatat true positive sebanyak 716, yaitu reservasi yang diprediksi batal dan benar-benar dibatalkan. Jumlah ini menunjukkan bahwa CatBoost mampu menangkap sebagian pola pembatalan meskipun kelas Canceled hanya mencakup 1,56% dari total dataset. Pada sisi kesalahan klasifikasi, terdapat false positive sebanyak 105 kasus. Artinya, terdapat 105 reservasi yang diprediksi batal oleh model, tetapi pada kondisi aktual tetap terealisasi sebagai Check-In. Jumlah false positive yang rendah ini mendukung capaian precision model sebesar 87,20% pada kelas pembatalan. Selain itu, terdapat false negative sebanyak 451 kasus, yaitu reservasi yang sebenarnya dibatalkan tetapi diprediksi sebagai Check-In. Nilai ini menjelaskan mengapa recall model berada pada 61,37%. Dengan kata lain, model masih melewatkan sebagian kasus pembatalan aktual. Namun, dalam konteks operasional hotel, konfigurasi ini dapat dipandang lebih konservatif karena model lebih menekan risiko salah alarm dibandingkan mengejar deteksi seluruh pembatalan dengan konsekuensi meningkatnya false positive. Secara keseluruhan, confusion matrix ini konsisten dengan performa model pada Table 3.2, yaitu accuracy sebesar 99,25%, precision kelas Canceled sebesar 87,20%, recall sebesar 61,37%, dan F1-score sebesar 72,04%.

3.1.9 Implikasi Manajerial

Hasil model CatBoost menunjukkan potensi penerapan prediksi pembatalan dalam mendukung kebijakan revenue management pada hotel budget di Jawa Tengah. Dengan precision yang tinggi pada kelas pembatalan, model dapat digunakan sebagai sistem pendukung keputusan untuk menandai reservasi berisiko tinggi, terutama berdasarkan kombinasi fitur seperti lead_time, arrival_month, deposit_type, dan adr. Pertama, model dapat diintegrasikan ke dalam PMS atau channel manager sebagai sistem peringatan dini pembatalan. Reservasi dengan kombinasi fitur berisiko tinggi dapat diberi penanda risiko untuk ditindaklanjuti oleh staf hotel. Tindakan yang dapat dipertimbangkan meliputi konfirmasi ulang, pengingat pembayaran, atau penyesuaian kebijakan deposit. Dengan demikian, hotel dapat melakukan mitigasi sebelum pembatalan benar-benar terjadi. Kedua, nilai false positive yang rendah dapat membantu mengurangi komunikasi yang tidak perlu kepada tamu yang kemungkinan besar tetap datang. Manajemen dapat memprioritaskan intervensi pada reservasi yang memiliki skor risiko tinggi, sehingga sumber daya operasional lebih efisien. Pendekatan ini juga dapat mengurangi potensi gangguan komunikasi kepada pelanggan yang tidak berisiko. Ketiga, hasil prediksi dapat menjadi salah satu masukan dalam strategi controlled overbooking. Namun, penerapan strategi ini perlu dilakukan secara bertahap dan hati-hati. Keputusan membuka kembali kamar atau menyesuaikan inventori sebaiknya tidak hanya didasarkan pada prediksi model, tetapi juga mempertimbangkan kapasitas hotel, musim kunjungan, nilai probabilitas prediksi,

serta evaluasi manual dari manajemen. Dengan demikian, model CatBoost tidak hanya berfungsi sebagai alat klasifikasi, tetapi juga sebagai sistem pendukung keputusan yang dapat membantu hotel mengelola risiko pembatalan secara lebih terukur dan berbasis data.

3.2 Pembahasan

Temuan empiris dalam penelitian ini memberikan kontribusi terhadap literatur *predictive analytics* dan *revenue management* pada industri perhotelan. Secara umum, hasil penelitian menunjukkan bahwa CatBoost dengan pendekatan *cost-sensitive learning* mampu menangani data pembatalan reservasi yang sangat tidak seimbang dan menghasilkan performa prediksi yang lebih seimbang dibandingkan model baseline berbasis SMOTE. Pertama, eksperimen ini menunjukkan bahwa CatBoost menghasilkan F1-score sebesar 72,04% dan *precision* sebesar 87,20% pada kelas pembatalan. Capaian ini lebih tinggi dibandingkan Random Forest + SMOTE dengan F1-score 37,88% dan XGBoost + SMOTE dengan F1-score 46,50%. Meskipun kedua model baseline menghasilkan *recall* yang lebih tinggi, nilai *precision* yang rendah menunjukkan adanya kecenderungan menghasilkan banyak *false positive*. Dalam konteks operasional hotel, kondisi tersebut dapat menimbulkan risiko salah alarm dan keputusan inventori yang kurang tepat. Keunggulan CatBoost dapat dikaitkan dengan kemampuannya dalam menangani fitur kategorikal secara *native* serta penggunaan mekanisme *ordered boosting* yang dirancang untuk mengurangi *target leakage* dan *prediction shift* (Prokhorenkova et al., 2018). Selain itu, penggunaan *scale_pos_weight* memungkinkan model memberikan bobot lebih besar terhadap kelas minoritas tanpa harus mengubah distribusi data melalui *oversampling* sintesis. Temuan ini menunjukkan bahwa *cost-sensitive learning* dapat menjadi pendekatan yang lebih sesuai ketika data pembatalan sangat tidak seimbang dan ketika pengurangan *false positive* menjadi prioritas operasional. Kedua, penggunaan dataset primer dengan cakupan temporal Oktober 2020 hingga April 2026 memberikan konteks empiris yang lebih mutakhir dibandingkan studi yang menggunakan dataset publik. Pola temporal yang ditemukan menunjukkan bahwa risiko pembatalan berkaitan dengan dinamika waktu pemesanan, musim perjalanan, dan karakteristik transaksi. Temuan mengenai kontribusi *lead_time* sejalan dengan studi sebelumnya yang menunjukkan bahwa semakin panjang jarak antara pemesanan dan kedatangan, semakin besar peluang terjadinya pembatalan karena konsumen memiliki ruang waktu lebih luas untuk mengubah rencana. Selain itu, pola pembatalan yang meningkat pada bulan tertentu, seperti November, menunjukkan bahwa *arrival_month* dapat menangkap konteks musiman yang relevan.

Namun, pola musiman tersebut tidak bersifat linear dan perlu dipahami sebagai bagian dari interaksi antarfitur. Oleh karena itu, penggunaan model berbasis pohon seperti CatBoost menjadi relevan karena mampu menangkap hubungan non-linear antarvariabel. Ketiga, integrasi SHAP memberikan nilai tambah pada aspek interpretabilitas model. Hasil SHAP menunjukkan bahwa *lead_time*, *arrival_month*, *deposit_type*, dan *adr* merupakan fitur penting dalam prediksi pembatalan. Dengan demikian, model tidak hanya menghasilkan output klasifikasi, tetapi juga memberikan informasi mengenai faktor-faktor yang paling berkontribusi terhadap prediksi. Hal ini penting karena model prediktif dalam konteks bisnis tidak cukup hanya akurat, tetapi juga perlu dapat dijelaskan agar dapat dipercaya dan digunakan dalam proses pengambilan keputusan. Hasil interpretasi SHAP juga menunjukkan bahwa pembatalan reservasi tidak dapat dijelaskan oleh satu faktor tunggal. Risiko pembatalan muncul dari kombinasi antara panjangnya jendela pemesanan, konteks bulan kedatangan, fleksibilitas kebijakan deposit, dan sensitivitas terhadap harga. Dengan demikian, strategi mitigasi risiko pembatalan perlu dirancang secara tersegmentasi, bukan menggunakan pendekatan yang sama untuk seluruh reservasi. Secara praktis, hasil penelitian ini memberikan dasar bagi hotel budget untuk mengembangkan sistem peringatan dini berbasis data. Reservasi dengan risiko tinggi dapat diprioritaskan untuk tindakan mitigasi, seperti konfirmasi ulang, permintaan deposit, atau pemantauan inventori yang lebih ketat. Namun, penerapan model tetap perlu memperhatikan aspek operasional dan pengalaman pelanggan. Prediksi model sebaiknya digunakan sebagai alat bantu keputusan, bukan sebagai mekanisme otomatis yang sepenuhnya menggantikan pertimbangan manajerial. Secara keseluruhan, penelitian ini menunjukkan bahwa kombinasi CatBoost dan SHAP mampu menghasilkan model prediksi pembatalan yang tidak hanya kompetitif secara performa, tetapi juga dapat diinterpretasikan. Pendekatan ini relevan untuk mendukung pengambilan keputusan berbasis data dalam manajemen pendapatan hotel, terutama pada kondisi data operasional yang kompleks, heterogen, dan sangat tidak seimbang.

4. Kesimpulan

Berdasarkan rangkaian eksperimen, analisis fitur, evaluasi model, dan interpretasi hasil yang telah dilakukan, penelitian ini menghasilkan beberapa kesimpulan utama. Pertama, tahapan prapemrosesan data berhasil menghasilkan dataset akhir sebanyak 74.826 catatan reservasi independen dari data mentah yang diekstraksi langsung dari *Property Management System* (PMS) sebuah hotel budget di Jawa Tengah. Dataset tersebut mencakup periode longitudinal selama 67 bulan, dari Oktober 2020 hingga April 2026. Melalui pembersihan data, penanganan nilai tidak valid, imputasi nilai kosong, serta pencegahan *target leakage*, data operasional dapat disiapkan menjadi basis pemodelan yang lebih representatif dan layak digunakan untuk prediksi pembatalan reservasi. Kedua, implementasi algoritma CatBoost Classifier dengan pendekatan *cost-sensitive learning* melalui parameter *scale_pos_weight* menunjukkan performa yang lebih baik dibandingkan model baseline berbasis SMOTE. Pada kelas pembatalan, model CatBoost menghasilkan *precision* sebesar 87,20%, *recall* sebesar 61,37%, dan F1-score sebesar 72,04%, dengan nilai AUC-ROC sebesar 0,86. Dibandingkan Random Forest + SMOTE dan XGBoost + SMOTE, CatBoost mampu mempertahankan keseimbangan yang lebih baik antara ketepatan prediksi dan kemampuan deteksi kelas minoritas. Meskipun model baseline menghasilkan *recall* lebih tinggi, nilai *precision* yang rendah menunjukkan kecenderungan menghasilkan lebih banyak *false positive*. Dalam konteks operasional hotel, kondisi tersebut berisiko memicu salah alarm dan dapat berdampak pada keputusan manajemen inventori kamar yang kurang tepat. Ketiga, integrasi SHAP (*Shapley Additive exPlanations*) berhasil meningkatkan interpretabilitas model. Hasil interpretasi menunjukkan bahwa *lead_time* merupakan fitur paling dominan dalam prediksi pembatalan. Reservasi dengan waktu tunggu yang lebih panjang cenderung memiliki risiko pembatalan lebih tinggi karena adanya ruang waktu yang lebih besar bagi tamu untuk mengubah rencana atau mengevaluasi alternatif akomodasi. Selain *lead_time*, fitur *arrival_month*, *deposit_type*, dan *adr* juga memberikan kontribusi penting terhadap prediksi model. Temuan ini menunjukkan bahwa risiko pembatalan tidak ditentukan oleh satu variabel tunggal, melainkan oleh interaksi antara dimensi temporal, transaksional, dan finansial.

Secara manajerial, hasil penelitian ini menunjukkan bahwa model CatBoost dapat digunakan sebagai dasar pengembangan sistem peringatan dini pembatalan reservasi. Integrasi model ke dalam PMS atau *channel manager* dapat membantu hotel menandai reservasi berisiko tinggi sejak awal. Prediksi tersebut dapat digunakan untuk mendukung tindakan mitigasi seperti konfirmasi ulang, pengingat pembayaran, atau penyesuaian kebijakan deposit, khususnya pada reservasi dengan *lead_time* panjang dan risiko pembatalan tinggi. Dengan nilai *false positive* yang relatif rendah, yaitu 105 kasus, model juga berpotensi membantu manajemen mengurangi intervensi yang tidak perlu terhadap tamu yang sebenarnya tetap akan melakukan Check-In. Untuk pengembangan penelitian selanjutnya, beberapa arah dapat dipertimbangkan. Pertama, optimasi hiperparameter yang lebih sistematis, seperti menggunakan Optuna atau *genetic algorithm*, dapat dilakukan untuk mencari keseimbangan terbaik antara *precision* dan *recall*. Kedua, penelitian mendatang dapat menambahkan variabel eksternal seperti kalender hari libur nasional, cuaca lokal, agenda acara daerah, indikator ekonomi pariwisata, atau sentimen ulasan digital pada platform OTA. Ketiga, perbandingan dengan arsitektur *deep learning* untuk data tabular, seperti TabNet atau Neural Oblivious Decision Ensembles (NODE), dapat dilakukan untuk menilai apakah pendekatan tersebut mampu memberikan peningkatan performa dibandingkan model berbasis *gradient boosting*. Selain itu, validasi lintas properti atau lintas wilayah juga penting dilakukan agar generalisasi model dapat diuji pada konteks hotel dan pasar yang lebih beragam. Secara keseluruhan, penelitian ini menunjukkan bahwa kombinasi CatBoost dan SHAP mampu menghasilkan model prediksi pembatalan reservasi yang akurat, cukup seimbang, dan dapat diinterpretasikan. Pendekatan ini relevan untuk mendukung pengambilan keputusan berbasis data dalam manajemen pendapatan hotel, terutama pada kondisi data operasional yang heterogen dan sangat tidak seimbang.

References

- Adil, M., Ansari, M. F., Alahmadi, A., Wu, J. Z., & Chakraborty, R. K. (2021). Solving the problem of class imbalance in the prediction of hotel cancellations: A hybridized machine learning approach. *Processes*, 9(10), Article 1713. <https://doi.org/10.3390/pr9101713>

- Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. *Journal of Business Research*, 144, 201–216. <https://doi.org/10.1016/j.jbusres.2022.01.083>
- Andriawan, Z. A., Purnama, S. R., Darmawan, A. S., Wibowo, A., Sugiharto, A., & Wijayanto, F. (2020). Prediction of hotel booking cancellation using CRISP-DM. *Proceedings of the 4th International Conference on Informatics and Computational Sciences (ICICoS 2020)*, 1–6. <https://doi.org/10.1109/ICICoS51170.2020.9299011>
- Antonio, N., de Almeida, A., & Nunes, L. (2017a). Predicting hotel bookings cancellation with a machine learning classification model. *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 1049–1054. <https://doi.org/10.1109/ICMLA.2017.00-11>
- Antonio, N., de Almeida, A., & Nunes, L. (2017b). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, 13(2), 25–39. <https://doi.org/10.18089/tms.2017.13203>
- Antonio, N., de Almeida, A., & Nunes, L. (2019a). An automated machine learning based decision support system to predict hotel booking cancellations. *Data Science Journal*, 18(1), 1–12. <https://doi.org/10.5334/dsj-2019-032>
- Antonio, N., de Almeida, A., & Nunes, L. (2019b). Big data in hotel revenue management: Exploring cancellation drivers to gain insights into booking cancellation behavior. *Cornell Hospitality Quarterly*, 60(4), 298–319. <https://doi.org/10.1177/1938965519851466>
- Ben Ahmed, S., Elaoud, A., & Ben Hafaiedh, I. (2026). A model-based approach for guided parameter exploration in machine learning classifiers. *Knowledge and Information Systems*, 68(1), 115–138. <https://doi.org/10.1007/s10115-026-02720-6>
- Camilleri, M. A. (2026). Opening the black box: Operational principles, tools and frameworks that advance explainable artificial intelligence (XAI) models. *Technological Forecasting and Social Change*, 229, Article 124710. <https://doi.org/10.1016/j.techfore.2026.124710>
- Cepi, G., Umbara, F. R., & Ashaury, H. (2025). Prediksi pembatalan reservasi hotel menggunakan metode XGBoost. *Jurnal Global Ilmiah*, 2(12), 1420–1432. <https://doi.org/10.55324/jgi.v2i12.270>
- Chen, I. F., & Lu, C. J. (2021). Demand forecasting for multichannel fashion retailers by integrating clustering and machine learning algorithms. *Processes*, 9(9), Article 1578. <https://doi.org/10.3390/pr9091578>
- Chen, S., Ngai, E. W. T., Ku, Y., Xu, Z., Gou, X., & Zhang, C. (2023). Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction. *Decision Support Systems*, 170, Article 113959. <https://doi.org/10.1016/j.dss.2023.113959>
- Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management*, 24(3), 323–330. [https://doi.org/10.1016/S0261-5177\(02\)00068-7](https://doi.org/10.1016/S0261-5177(02)00068-7)
- Chu, F. L. (2009). Forecasting tourism demand with ARMA-based methods. *Tourism Management*, 30(5), 740–751. <https://doi.org/10.1016/j.tourman.2008.10.016>
- Deng, C., Liu, X., Zhang, J., Mo, Y., Li, P., Liang, X., & Li, N. (2025). Prediction of retail commodity hot-spots: A machine learning approach. *Data Science and Management*, 8(4), 414–422. <https://doi.org/10.1016/j.dsm.2025.02.003>
- Falk, M., & Vieru, M. (2018). Modelling the cancellation behaviour of hotel guests. *International Journal of Contemporary Hospitality Management*, 30(10), 3100–3116. <https://doi.org/10.1108/IJCHM-08-2017-0509>

- Febrian, Y. Y., Wijaya, D. R., & Ervina, E. (2024). Hotel reservation cancellation prediction using boosting model. *Proceedings of the 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)*, 138–143. <https://doi.org/10.1109/ICoSEIT60086.2024.10497479>
- Gomez-Talal, I., Azizsoltani, M., Talon-Ballester, P., & Singh, A. (2025). Machine learning in hospitality: Interpretable forecasting of booking cancellations. *IEEE Access*, 13, 26622–26638. <https://doi.org/10.1109/ACCESS.2025.3536094>
- Jiang, P., Liu, Z., Zhang, L., & Wang, J. (2023). Hybrid model for profit-driven churn prediction based on cost minimization and return maximization. *Expert Systems with Applications*, 228, Article 120354. <https://doi.org/10.1016/j.eswa.2023.120354>
- Kumar, A., Prasad, U., Tiwari, R. K., & Pandey, V. (2023). Data preprocessing using machine learning for prediction of booking cancellations. *Communications in Computer and Information Science*, 1822, 164–182. https://doi.org/10.1007/978-3-031-37303-9_13
- Kundu, S., Roy, S., Shukla, A., & Mateen, A. (2025). Unveiling cancellation dynamics: A two-stage model for predictive analytics. *Data & Knowledge Engineering*, 160, Article 102467. <https://doi.org/10.1016/j.datak.2025.102467>
- Lecheheb, S., Boulehouache, S., & Brahimi, S. (2026). Comparison of AutoML frameworks for effective analysis in MAPE-K self-adaptive systems. *Discover Artificial Intelligence*, 6(1), 45–58. <https://doi.org/10.1007/s44163-026-01083-9>
- Li, X., Pan, B., Law, R., & Huang, X. (2017). Forecasting tourism demand with composite search index. *Tourism Management*, 59, 57–66. <https://doi.org/10.1016/j.tourman.2016.07.005>
- Li, X., & Zong, Q. (2026). How does explainable AI affect service innovation of frontline employees? A moderated chain mediation model. *International Journal of Contemporary Hospitality Management*, 38(4), 1172–1191. <https://doi.org/10.1108/IJCHM-01-2025-0164>
- Liao, B., & Yu, C. (2025). Enhanced predictive analytics for hotel booking cancellations: A fusion approach integrating artificial neural networks and stacking techniques. *Proceedings of the 7th International Conference on Information Science, Electrical and Automation Engineering (ISEAE)*, 632–636. <https://doi.org/10.1109/ISEAE64934.2025.11042049>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774. <https://doi.org/10.48550/arXiv.1705.07874>
- Noorul Ameen, J., Irshad Ahamed, M., Senthil Mahesh, P. C., & Banu Ahamed, S. (2026). Machine learning-based adaptive optimization algorithm for Lipschitz continuity in biosignal telemetry. *International Journal of Data Science and Analytics*, 22(1), 89–104. <https://doi.org/10.1007/s41060-026-01092-y>
- Pramanik, P., Jana, R. K., & Ghosh, I. (2024). AI readiness enablers in developed and developing economies: Findings from the XGBoost regression and explainable AI framework. *Technological Forecasting and Social Change*, 205, Article 123482. <https://doi.org/10.1016/j.techfore.2024.123482>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 6638–6648.
- Sánchez-Medina, A. J., & C-Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management*, 89, Article 102546. <https://doi.org/10.1016/j.ijhm.2020.102546>
- Saputro, P. H., & Nanang, H. (2021). Exploratory data analysis & booking cancellation prediction on hotel booking demands datasets. *Journal of Applied Data Sciences*, 2(1), 40–56. <https://doi.org/10.47738/jads.v2i1.20>

- Satu, M. S., Ahammed, K., & Abedin, M. Z. (2020). Performance analysis of machine learning techniques to predict hotel booking cancellations in hospitality industry. *Proceedings of the 23rd International Conference on Computer and Information Technology (ICCIT)*, 1–6. <https://doi.org/10.1109/ICCIT51783.2020.9392648>
- Sekhon, G., & Ahuja, S. (2023). Review machine learning models for managing hotel cancellations in the tourism industry. *Proceedings of the 3rd International Conference on Intelligent Technologies (CONIT)*, 1–5. <https://doi.org/10.1109/CONIT59222.2023.10205827>
- Shirisha, N., Anusha, K., Kiran, A., & Buavani, Y. T. S. (2023). Prediction of hotel booking & cancellation using machine learning algorithms. *Proceedings of the International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. <https://doi.org/10.1109/ICCCI56745.2023.10128484>
- Silvestre, P., & António, N. (2025). Predicting hotel booking cancellations during high-volatility times. *Springer Proceedings in Business and Economics*, 363–373. https://doi.org/10.1007/978-3-031-83705-0_30
- Silvestre, P., Antonio, N., & Carrasco, P. (2026). Navigating uncertainty: Enhancing hotel cancellation predictions with adaptive machine learning. *Information Technology & Tourism*, 28(1), 44–69. <https://doi.org/10.1007/s40558-025-00349-9>
- Song, H., Qiu, R. T. R., & Park, J. (2019). A review of research on tourism demand forecasting. *Annals of Tourism Research*, 75, 338–362. <https://doi.org/10.1016/j.annals.2018.12.001>
- Sun, S., Wei, Y., Tsui, K. L., & Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management*, 70, 1–10. <https://doi.org/10.1016/j.tourman.2018.07.010>
- Theodorakopoulos, L., Kalliampakou, I., Ntantou, A., & Halkiopoulos, C. (2025). Leveraging machine learning for sustainable hotel management: Predicting booking cancellations to optimize operations. *Springer Proceedings in Business and Economics*, 135–178. https://doi.org/10.1007/978-3-031-78471-2_6
- Wardley, L. J., Rajabi, E., Amin, S. H., & Ramesh, M. (2024). A machine learning approach feature to forecast the future performance of the universities in Canada. *Machine Learning with Applications*, 16, Article 100548. <https://doi.org/10.1016/j.mlwa.2024.100548>
- Yang, D., & Miao, X. (2024). Predicting hotel booking cancellations using tree-based neural network. *PeerJ Computer Science*, 10, Article e2473. <https://doi.org/10.7717/peerj-cs.2473>
- Yoo, M., Singh, A. K., & Loewy, N. (2023). Predicting hotel booking cancelation with machine learning techniques. *Journal of Hospitality and Tourism Technology*, 15(1), 54–69. <https://doi.org/10.1108/JHTT-07-2022-0227>
- Yuan, X., Wu, D., & Gayle, D. D. B. (2022). Technology adoption and COVID-19 pandemic. *Data and Information Management*, 6(2), Article 100016. <https://doi.org/10.1016/j.dim.2022.100016>
- Zhai, Q., Tian, Y., Luo, J., & Zhou, J. (2023). Hotel overbooking based on no-show probability forecasts. *Computers & Industrial Engineering*, 180, Article 109226. <https://doi.org/10.1016/j.cie.2023.109226>
- Zhang, J. (2025). Model-aware preprocessing: How imputation and feature selection uniquely interact with DNNs and LSTMs for hotel cancellation prediction. *Proceedings of the 8th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI)*. <https://doi.org/10.1109/ACAI68217.2025.11406664>
- Zhu, R. (2026). Machine learning-driven dynamic matching of industry-education demands and course generation algorithms. *Discover Artificial Intelligence*, 6(1). <https://doi.org/10.1007/s44163-025-00781-0>

How Cites

Sudana, I. M., & Pamungkas, E. W. (2026). Predicting Hotel Booking Cancellations Using CatBoost and SHAP: An Explainable AI Approach Based on 2020–2026 Operational Data. *Computer Journal*, 328–343. <https://journal.ypmma.org/cj/article/view/511>.

Publisher's Note

Yayasan Pendidikan Mitra Mandiri Aceh (YPPMA) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Submit your manuscript to YPMMA Journal and *benefit* from: <https://journal.ypmma.org/index.php/cj>.