e-ISSN 2962-6129
Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya (SENAMIKA)
Jakarta-Indonesia, 20 Agustus 2022

Implementasi Penggunaan Algoritma Categorical Boosting (Catboost) Dengan Optimisasi Hiperparameter Dalam Memprediksi Pembatalan Pesanan Kamar Hotel

Johannes Christian[1], Iin Ernawati[2], Nurul Chamidah[3],
Informatika / Fakultas Ilmu Komputer
Universitas Pembangunan Nasional Veteran Jakarta
Jl. RS. Fatmawati , Pondok Labu, Jakarta Selatan, DKI Jakarta, 12450, Indonesia
johannesc@upnvj.ac.id[1], iinernawati@upnvj.ac.id[2], nurul.chamidah@upnvj.ac.id[3]

Abstrak. Online Travel Agent (OTA) yang tumbuh menjadi sebuah aplikasi untuk memudahkan masyarakat dalam melakukan pemesanan kamar hotel secara daring memberikan dampak yang signifikan terhadap manajemen hotel. Kemudahan tersebut membuat pengunjung melakukan multiple-booking yang mengakibatkan tingginya tingkat pembatalan pesanan kamar hotel. Strategi overbooking yang diterapkan oleh manajemen hotel memerlukan tingkat keakuratan yang tinggi dalam memperkirakan pengunjung yang melakukan pembatalan pesanan. Maka dari itu penelitian ini akan berfokus pada masalah tersebut menggunakan model pembelajaran mesin dengan algoritma CatBoost. Dalam proses pengklasifikasian data perlu dilakukan pembersihan data melalui proses data preparation. Setelah itu data akan dilakukan ekstraksi dan seleksi atribut sehingga data siap digunakan untuk melatih machine learning. Untuk meningkatkan performa dari model, optimisasi hiperparameter RandomizedSearchCV diterapkan terhadap model CatBoost. Hasil evaluasi dengan confusion matrix yaitu akurasi sebesar 88% dan precision sebesar 86%. Dengan menerapkan visualisasi SHAP pada CatBoost berhasil dihasilkan karakteristik – karakteristik pesanan kamar hotel yang akan berpeluang besar dibatalkan.

Kata Kunci: Hotel, Overbooking, CatBoost, Hyperparameter Tuning

1 Pendahuluan

Perkembangan teknologi mempengaruhi seluruh bidang industri tanpa terkecuali industri perhotelan. Sistem pemesanan hotel konvensional yang digunakan sebelumnya mulai tergantikan dengan adanya perusahaan – perusahaan teknologi inovatif yang disebut dengan Online Travel Agent (OTA) seperti Traveloka, Tiket.com, Pegipegi dan OTA lainnya.

Namun dengan mudahnya pengguna melakukan pemesanan hotel, muncul permasalahan yang dialami oleh penyedia hotel dimana tingginya tingkat pesanan yang dibatalkan mendekati hari reservasi / check-in. Penyebab tingginya pembatalan pesanan kamar hotel pada OTA dikarenakan banyak pengguna yang melakukan multiple-booking diawal untuk menyimpan penawaran – penawaran tiket hotel yang paling menarik seperti adanya diskon atau pemotongan harga, tetapi jika terdapat penawaran yang lebih baik, maka pengguna tersebut akan melakukan pembatalan pada pesanan tersebut[1].

Tingkat kerugian terbesar terdapat pada pengunjung yang melakukan pembatalan pesanan diakhir atau pengunjung yang tidak hadir sampai di hari reservasi[2]. Masalah ini mengakibatkan rendahnya tingkat okupansi ruangan dan menurunnya tingkat pendapatan yang diterima oleh hotel. Dengan permasalahan pembatalan pesanan tersebut, manajemen hotel harus mencari solusi untuk mencegah kerugian pendapatan yang lebih banyak lagi dengan menerapkan strategi overbooking pada kapasitas kamar hotel[3]. Penjualan kapasitas kamar hotel yang berlebih ini (overbooking) akan mengisi kekosongan ruangan yang diakibatkan dari pembatalan pesanan.

Berbagai studi telah dilakukan untuk membuat model prediksi yang terbaik dalam mengklasifikasikan pembatalan pesanan kamar hotel dengan berbagai macam metode pembelajaran mesin. Antonio dkk. mengembangkan model xgboost (XGB) untuk memprediksi pembatalan pesanan hotel menghasilkan akurasi sebesar 84%[4]. Tetapi metode xgboost (XGB) memiliki kelemahan pada hasilnya yang rentan terjadinya overfit apabila jumlah kombinasi tree yang diterapkan kurang tepat. Model klasifikasi dengan metode lain juga telah dikembangkan oleh Azhar dkk. (2021) dengan menggunakan metode Random Forest (RF) yang telah dioptimalisasi menggunakan hyperparameter-tuning, metode ini menghasilkan model klasifikasi dengan akurasi mencapai 87%[5]. Tetapi,

metode Random Forest juga memiliki kelemahan pada tidak konsistennya hasil akurasi yang dikeluarkan karena metode ini menggunakan fungsi acak dalam menentukan baris data dan kandidat atribut.

Penelitian ini bertujuan untuk mengetahui hasil prediksi klasifikasi pada data pesanan kamar hotel di waktu yang akan datang. Hasil yang dikeluarkan berupa program yang dapat mengimplementasikan metode CatBoost sehingga dapat mengklasifikasikan pesanan yang berpeluang dibatalkan oleh pengunjung. Dengan adanya penelitian ini, diharapkan mampu mengatasi permasalahan yang dialami oleh manajemen hotel dalam menentukan kebijakan dari strategi overbooking yang akan diterapkan, sehingga mampu meningkatkan tingkat okupansi ruangan dan juga tingkat pendapatan manajemen hotel.

2 Landasan Teori

2.1 Overbooking

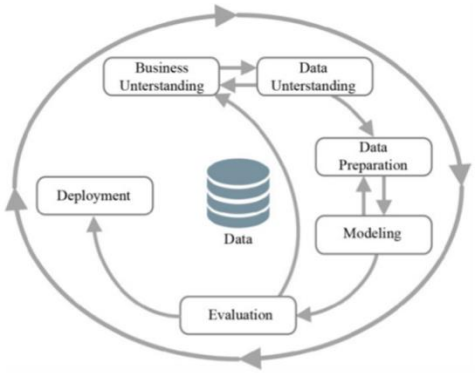
Overbooking adalah suatu strategi bisnis yang digunakan oleh suatu perusahaan untuk mengatasi permasalahan pembatalan pesanan yang dilakukan dengan cara menjual sejumlah barang fiktif diluar dari kapasitas barang aktual yang perusahaan tersebut miliki. Berikut adalah pendapat dari ahli mengenai pengertian dari overbooking :

1. Menurut Haynes dan Egan (2020), strategi overbooking adalah suatu proses yang muncul ketika ruangan yang direservasi atau dibooking melebihi dari kapasitas kamar hotel dan strategi ini diimplementasikan untuk melindungi penghasilan hotel dari permasalahan pengunjung yang membatalkan pesanan dan pengunjung yang tidak memberikan informasi hingga hari reservasi[6].
2. Strategi overbooking digambarkan sebagai sebuah kapasitas efektif yang digunakan sebagai parameter penjualan ruangan yang seringkali lebih besar daripada kapasitas aktual untuk mengatasi pembatalan reservasi dan ketidakhadiran pengunjung (Pimentel et al., 2021)[7].

Dalam industri perhotelan, overbooking didefinisikan sebagai suatu proses penjualan kamar hotel kepada calon pelanggan dengan menambahkan kamar fiktif sehingga jumlah kamar hotel yang dijual melebihi kapasitas kamar hotel sebenarnya.

2.2 Data Mining

Data mining yang berdasar pada prinsip statistik merupakan suatu proses eksplorasi dan analisis sejumlah data yang sangat besar untuk mendapatkan suatu pola atau informasi dari data tersebut. Algoritma akan digunakan dalam mencari pola dan hubungan didalam data dan selanjutnya informasi tersebut digunakan untuk melakukan peramalan atau perkiraan apa yang akan terjadi di masa yang akan dating[8]. Cross-Industry Standard Process untuk Data Mining atau disingkat CRISP-DM adalah suatu kerangka metode yang bertujuan untuk menerjemahkan permasalahan – permasalahan bisnis kedalam konteks data mining dan melaksanakan proyek data mining secara independen dari area aplikasi dan teknologi yang digunakan[9].



Gambar 1. Tahapan dari CRISP-DM yang dimulai dari Business Understanding hingga Deployment

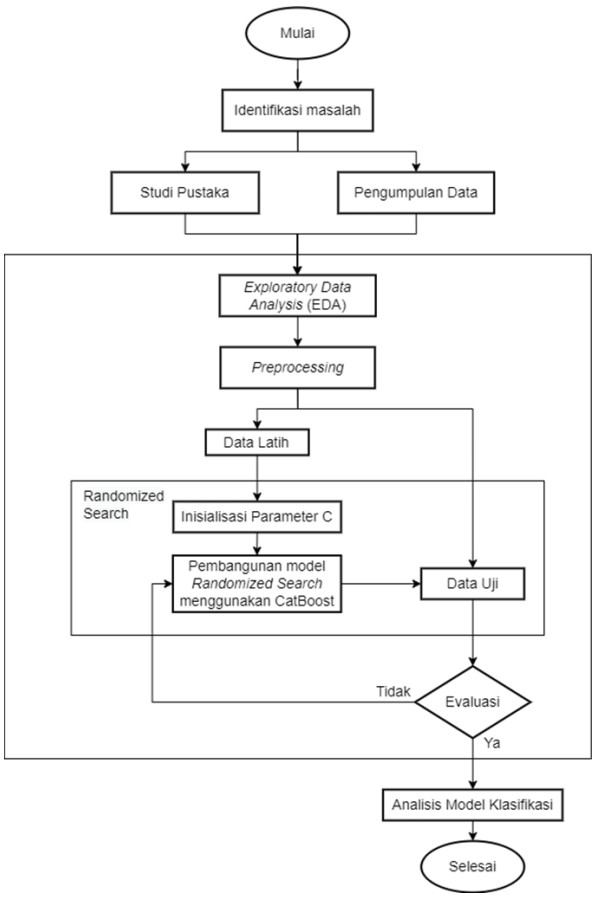
2.3 Exploratory Data Analysis (EDA)

EDA adalah suatu langkah fundamental awal yang dilakukan setelah pengumpulan data dengan melakukan visualisasi, plotting, dan manipulasi dengan tujuan untuk memahami kualitas dari data yang akan digunakan untuk membangun model[10]. Hasil yang dikeluarkan oleh proses EDA dapat digunakan untuk memahami permasalahan dari data yang selanjutnya akan dilakukan penanganan pada praproses data.

2.4 CatBoost (Categorical Boosting)

CatBoost adalah algoritma pembelajaran mesing yang masih tergabung dalam keluarga Gradient Boosted Decision Trees (GBDT) yang berada dalam lingkupan ensemble learning. CatBoost adalah suatu algoritma yang dibuka secara umum untuk terus dikembangkan dalam lingkup Supervised ML yang membawa 2 inovasi, yaitu: Ordered Target Statistics dan Ordered Boosting[11]. Dibeberapa kajian, penerapan gradient boosting menemukan beberapa masalah statistik. Model yang dihasilkan oleh gradient boosting adalah model yang didapatkan setelah beberapa langkah boosting dilakukan sehingga bergantung dari semua data latih. Proses boosting yang dilakukan berulang kali mengakibatkan terjadinya pergeseran distribusi dari model yang dipelajari. Selain itu, masalah lainnya yang dihadapi adalah algoritma dalam penanganan atribut data yang memiliki sifat kategorik. Teknik ordering boosting adalah teknik yang merupakan modifikasi dari algoritma gradient boosting standar dan akan digunakan untuk menghindari terjadinya kebocoran target. Selain itu terdapat algoritma baru yang berguna untuk melakukan pemrosesan data kategorik. Kombinasi dari keduanya disebut dengan algoritma CatBoost (Categorical Boosting)

3 Metode Penelitian



Gambar 2. Bagan Alur Penelitian

3.1 Identifikasi Masalah

Identifikasi masalah dilakukan agar peneliti fokus terhadap titik permasalahan yang akan diselesaikan. Pada kasus ini, peneliti menemukan masalah terkait pengklasifikasian pesanan kamar hotel untuk kelas dibatalkan atau tidak dibatalkan, sehingga hasil model klasifikasi tersebut dapat dijadikan landasan dalam menentukan parameter strategi overbooking yang akan digunakan oleh manajemen hotel.

3.2 Studi Pustaka

Sebagai sumber pustaka dalam penelitian ini, peneliti mempelajari kajian – kajian penelitian sebelumnya dengan mengumpulkan artikel, jurnal dan buku mengenai strategi overbooking pada industri perhotelan, pra-proses data dan algoritma klasifikasi catboost yang akan dijadikan referensi dalam penyelesaian permasalahan penelitian ini.

3.3 Metode Pengumpulan Data

Data sekunder merupakan jenis data penelitian yang tidak diperoleh secara langsung (sumber utama) oleh peneliti, melainkan diperoleh melalui media perantara yang sifatnya tidak langsung seperti: catatan tertulis, buku, website, dokumen terstruktur dan tidak terstruktur, atau arsip yang telah dipublikasikan maupun yang tidak dipublikasikan secara umum. Dalam kasus ini, sumber objek penelitian diperoleh dari website www.kaggle.com yang diunggah oleh Mojtaba pada tahun 2021 dengan judul Hotel Booking.

3.4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis yang selanjutnya disingkat EDA merupakan proses eksplorasi data yang tujuannya adalah memahami karakteristik data yang akan kita lakukan analisis kedepannya. Informasi yang dihasilkan pada tahapan EDA akan digunakan sebagai dasar dalam penentuan proses yang akan dilakukan pada preprocessing data hingga pembuatan model klasifikasi.

3.5 Preprocessing

Preprocessing adalah tahapan yang sangat penting dalam pembuatan model prediksi karena data yang diterima dari realtime database seringkali memiliki data yang tidak lengkap, tidak seragam dan tidak konsisten sehingga mengakibatkan hasil prediksi yang tidak tepat dan kurang akurat.

Pembersihan data dilakukan untuk mengatasi permasalahan – permasalahan yang terdapat pada data kotor. Pembersihan awal dilakukan dengan memeriksa dan memperbaiki seluruh kelengkapan atribut apakah terdapat data yang hilang atau biasa disebut dengan missing data. Pembersihan selanjutnya dilakukan untuk memperbaiki tipe data yang kurang tepat serta menghapus baris data yang nilainya tidak valid. Permasalahan yang seringkali muncul pada data kotor dan harus diatasi adalah data outliers. Data outliers adalah data yang memiliki nilai ekstrem. Data yang bernilai ekstrem ini harus diatasi dengan tujuan untuk mencegah terjadinya permasalahan pada model prediksi yang disebut dengan overfit.

Ekstraksi atribut dilakukan dengan mengambil ciri atau inti sari dari satu atau lebih atribut. Hasil dari ekstraksi atribut dapat berupa atribut baru yang diekstrak dari satu atau lebih atribut data sehingga dapat digunakan sebagai atribut tambahan sehingga model mampu mempelajari lebih banyak data. Ekstraksi Atribut bertujuan untuk mengurangi pengurangan informasi data awal berupa atribut yang tidak dapat diproses oleh model pembelajaran mesin.

Seleksi Atribut merupakan suatu proses pengurangan jumlah atribut yang akan digunakan sebagai input atau masukkan dari model pembelajaran mesin yang akan dibangun. Seleksi atribut bertujuan untuk mengurangi biaya komputasi dari model prediksi dan dibeberapa kasus dapat meningkatkan performa dari model.

3.6 Pemodelan Data dan Optimasi Hiperparameter

Dalam pembangunan model pembelajaran mesin, terdapat komponen model yang disebut dengan hyperparameter

yang dijadikan sebagai dasar aturan bagaimana mesin akan belajar. Proses menentukan nilai parameter ini disebut dengan optimisasi. Proses optimisasi dilakukan dengan melakukan beberapa eksperimen kombinasi dari berbagai parameter hingga mendapatkan metrik evaluasi terbaik. Metode yang akan digunakan pada penelitian ini adalah metode Randomized Search. Untuk metode Randomized Search parameter n_iter secara default berjumlah 10 iterasi, artinya terdapat 10 jenis model dengan kombinasi parameter yang berbeda dan diambil secara acak pada parameter yang telah diinisialisasi. Parameter pada algoritma CatBoost yang dilakukan eksperimen ditunjukkan pada **Tabel 1.**

Tabel 1. Parameter yang akan diuji berulang

Parameter	Penjelasan
iterations	Jumlah maksimum pohon yang dapat dibuat
depth	Jumlah kedalaman pohon
learning_rate	Besaran tingkat pembelajaran
l2_leaf_reg	Koefisien L2 sebagai koefisien regularisasi model

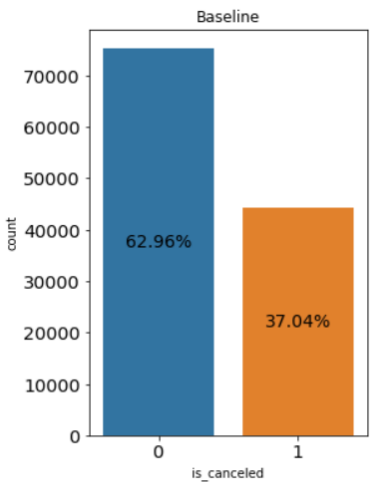
Tahap pemodelan data merupakan proses penerapan Algoritma CatBoost pada objek penelitian dalam memprediksi apakah suatu pesanan kamar hotel berpeluang besar dibatalkan atau tidak. Tahapan pemodelan dilakukan sebagai berikut:

1. Melakukan pengacakan dengan permutasi setiap S baris data untuk membuat set data baru. Proses permutasi dilakukan sebanyak S kali hingga terbentuk S set data latih yang berbeda
2. Melakukan inisialisasi matriks
3. Melakukan pengambilan secara acak satu set data latih hasil permutasi acak, berikut langkah – langkah selanjutnya:
4. Menerapkan Ordered Target Statistics (OTS) sebagai metode pengubahan data kategorik ke data numerik.
5. Membuat Ordered Boosting Tree T baru. Pohon tersebut digunakan untuk memperkirakan gradient atau residual dari setiap set S_r dengan menggunakan matriks M(r, i).
6. Dengan menggunakan Ordered Boosting Tree T untuk memprediksi semua set permutasi S1, S2, … , Ss dan melakukan pembaharuan M terus menerus untuk setiap set permutasi.
7. Berdasarkan I pohon, prediksi dilakukan dengan menghitung rata – rata dari semua I pohon prediksi.

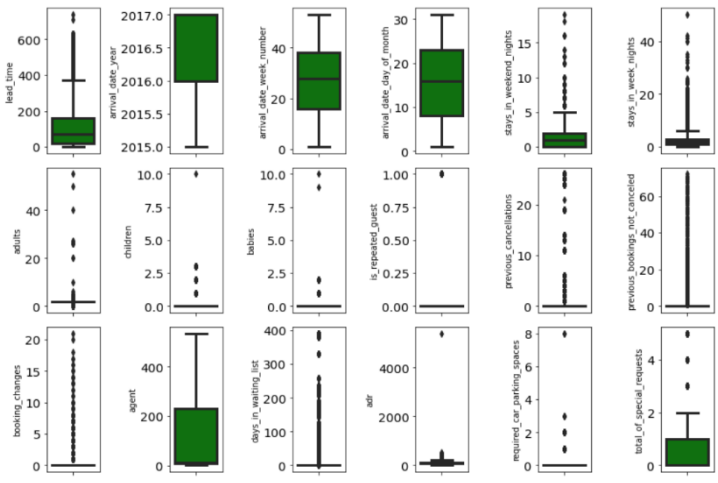
4 Hasil dan Pembahasan

Data collection akan dilakukan dengan mengunduh dokumen ‘hotel booking.csv’ dari website kaggle dengan tautan https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand.

4.1 Exploratory Data Analysis



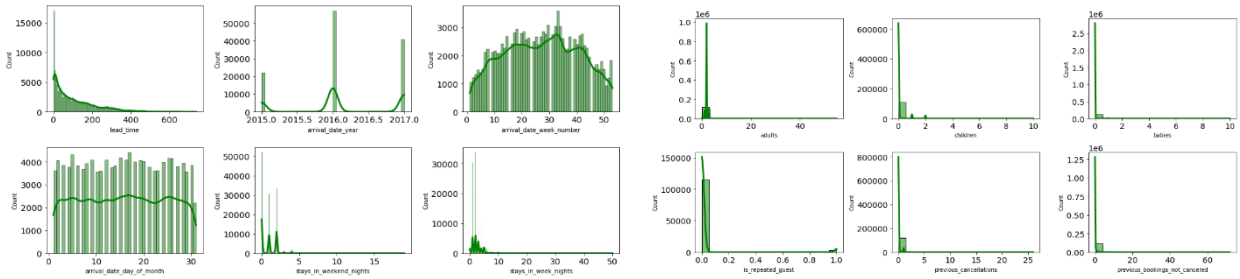
Gambar 3. Proporsi label



Gambar 4. Visualisasi boxplot

Gambar 3 menunjukkan proporsi pengunjung yang tidak membatalkan pesanan memiliki persentase 62,96% lebih besar daripada pengunjung yang membatalkan pesanan sebesar 37.04%. Proporsi tersebut masih masuk kedalam kategori normal sehingga tidak perlu dilakukan treatment imbalance data.

Sedangkan pada Gambar 4 Menunjukkan mayoritas fitur memiliki nilai outlier kecuali fitur arrival_date_year, arrival_date_week_number, arrival_date_day_of_month, dan agent. Data didominasi dengan fitur – fitur yang memiliki distribusi didaerah 0 sehingga perlu kita lakukan transformasi logaritma sebelum dilakukan penghilangan nilai outlier.



Gambar 5. Visualisasi Histogram

Gambar 5 diatas menunjukkan bentuk distribusi dari setiap fitur yang akan dianalisis. Visualisasi tersebut menunjukkan fitur lead_time dan total_of_special_requests memiliki bentuk distribusi positively skewed sedangkan Fitur adults, children, babies, booking_changes, days_in_waiting_list, adr, required_car_parking_spaces memiliki bentuk distribusi long tail.

4.2 Data Preprocessing

Pada penelitian ini, pembersihan data diawali dengan pemeriksaan jumlah nilai NAN yang terkandung pada setiap atribut. Gambar 18 menunjukkan jumlah dan persentase missing data pada atribut – atribut yang memiliki nilai nan.

	features	total_nan	percentage(%)
0	company	112593	94.31
1	agent	16340	13.69
2	country	488	0.41
3	children	4	0.00

Gambar 6. Jumlah dan persentase nilai NAN

Pembersihan data dari missing data dilakukan dengan melakukan drop pada baris data yang memiliki nilai nan. Jumlah data yang di-drop sebanyak 16,494 baris data. Sehingga data yang tersisa setelah dibersihkan dari missing data sebanyak 102,896.

Selain itu perlu juga dilakukan handling terhadap data – data yang memiliki nilai yang tidak valid. Dari 102,896 baris data tersebut terdapat 170 data yang atribut adult, children dan babies bernilai 0, sehingga perlu dilakukan drop karena nilai invalid tersebut mengartikan bahwa tidak ada pax / orang yang akan menginap. Sehingga data yang tersisa setelah dibersihkan dari invalid values sebanyak 102,776.

4.3 Ekstraksi Fitur / Atribut

Proses ekstraksi atribut dilakukan dengan mengambil intisari dari satu atau lebih atribut. Atribut – Atribut yang diekstrak pada penelitian ini diantaranya:

1. total_people: Merupakan atribut jumlah pax yang akan menginap. Hasil ekstraksi dari atribut babies, adult, children.
2. total_stay_nights: Merupakan jumlah lama menginap baik dihari kerja ataupun diakhir pekan. Hasil ekstraksi dari atribut stays_in_weekend_nights dan stays_in_week_nights.
3. cancellation_time: Merupakan lama waktu pelanggan membatalkan pesanan terhitung dari waktu

kedatangan. Hasil ekstraksi dari atribut arrival_date_day, arrival_date_month, arrival_date_year dan reservation_status_date.

4. was_in_waiting_list: Merupakan atribut yang menyimpan data apakah seorang pengunjung pernah masuk kedalam waiting list kamar atau tidak. Hasil ekstraksi dari atribut days_in_waiting_list.

4.4 Seleksi Fitur / Atribut

Proses seleksi atribut dilakukan dengan memilih atribut – atribut yang bisa digunakan oleh machine learning. Atribut – atribut yang dipilih pada penelitian ini diantaranya: hotel, is_canceled, lead_time, arrival_date_month, arrival_date_week_number, arrival_date_day_of_month, meal, country, market_segment, distribution_channel, is_repeated_guest, previous_cancellations, previous_bookings_not_canceled, reserved_room_type, assigned_room_type, booking_changes, deposit_type, agent, days_in_waiting_list, customer_type, adr, required_car_parking_spaces, total_of_special_requests, total_people, total_stays_night, was_in_waiting_list. Hasil dari proses seleksi atribut didapatkan 26 atribut yang akan digunakan sebagai bahan pembelajaran oleh machine learning dengan algoritma categorical boosting atau catboost.

4.5 Pembagian Data

Pada penelitian ini, pembagian data dilakukan dengan menggunakan library sklearn yaitu dengan metode train_test_split. Untuk memastikan bahwa sampel data yang dilatih memiliki proporsi yang sama dengan data populasi, maka dari itu akan dilakukan pembagian data dengan penambahan atribut stratify terhadap variabel target. Pembagian data akan dilakukan dengan metode hold-out-estimation dengan proporsi 75% data sebagai data latih dan 25% sebagai data uji seperti yang ditunjukkan pada Tabel 2.

Tabel 2. Hasil Pembagian Data Latih dan Uji

Data Latih							
hotel	**lead_time**	**customer_type**	**adr**	**…**	**total _people**	**total_stays _night**	**jumlah data**
Resort Hotel	13	Transient	75	…	1	1	77,082
Resort Hotel	14	Transient	98	…	2	2	
Resort Hotel	14	Transient	98	…	2	2	
…	…	…	…	…	…	…	
Resort Hotel	85	Transient	82	…	2	3	
Data Uji							
hotel	**lead_time**	**customer_type**	**adr**	**…**	**total _people**	**total_stays _night**	**jumlah data**
…	…	…	…	…	…	…	25,694
City Hotel	32	Transient	23	…	1	3	

4.6 Hyperparameter Tuning

Pada penelitian ini, optimisasi hiperparameter dilakukan menggunakan metode Randomized Search CV yang merupakan metode yang disediakan oleh library scikit learn. Pada penelitian ini, CatBoost menggunakan Gradient Boosting sebagai Teknik boosting-nya. Pada algoritma CatBoost terdapat banyak parameter yang bisa dilakukan optimasi, tetapi pada penelitian ini akan digunakan 4 parameter, diantaranya: iterations, depth, learning_rate, dan l2_leaf_reg.

Tabel 3. Variasi Parameter

Atribut	Nilai Parameter
iterations	500, 1000, 2000
depth	1, 3, 6, 10
learning_rate	0.012, 0.055, 0.064, 0.1
l2_leaf_reg	1, 3, 5

Secara default Randomized Search CV memiliki n_iter sebanyak 10 iterasi, yang artinya akan dibuat 10 model

dengan 10 variasi parameter. Kesepuluh model tersebut akan dilatih dan performanya akan dibandingkan sesuai dengan metrik yang ingin dioptimalkan, pada penelitian ini metrik yang digunakan adalah metrik akurasi. Parameter yang memiliki akurasi terbaik akan digunakan sebagai parameter pembangunan model.

4.7 Pembangunan Model

Pada penelitian ini, pemodelan data akan menggunakan metode CatBoost. Metode CatBoost memiliki kemampuan yang sangat baik dalam mengubah data kategorikal menjadi data numerikal.

hotel	arrival_date_month	meal	country	...	assigned_room_type	deposit_type	agent	customer_type
Resort Hotel	July	BB	GBR	...	A	No Deposit	304	Transient
Resort Hotel	July	BB	GBR	...	A	No Deposit	240	Transient
Resort Hotel	July	BB	GBR	...	A	No Deposit	240	Transient
Resort Hotel	July	FB	PRT	...	C	No Deposit	303	Transient
Resort Hotel	July	BB	PRT	...	A	No Deposit	240	Transient

hotel	arrival_date_month	meal	country	...	assigned_room_type	deposit_type	agent	customer_type
0.390688	0.390688	0.390688	0.390688	...	0.390688	0.390688	0.390688	0.390688
0.195344	0.195344	0.195344	0.195344	...	0.195344	0.195344	0.390688	0.195344
0.130229	0.130229	0.130229	0.130229	...	0.130229	0.130229	0.195344	0.130229
0.097672	0.097672	0.390688	0.390688	...	0.390688	0.097672	0.390688	0.097672
0.078138	0.078138	0.097672	0.195344	...	0.097672	0.078138	0.130229	0.078138

Gambar 7. Fitur kategorik sebelum dan sesudah dilakukan Catboost Encoding

Setelah seluruh atribut kategorik sudah diubah kedalam bentuk numerik, maka selanjutnya akan dilakukan pelatihan pembelajaran mesin CatBoost terhadap data latih untuk didapatkannya model klasifikasi. Pembentukan model berulang akan diterapkan oleh RandomizedSearchCV hingga mendapatkan model dengan parameter terbaik. Model yang dibangun secara berulang oleh RandomizedSearchCV dibangun menggunakan library dari catboost. Model yang dibangun akan berjumlah 10 model dan akan dibandingkan skor akurasi terbaiknya.

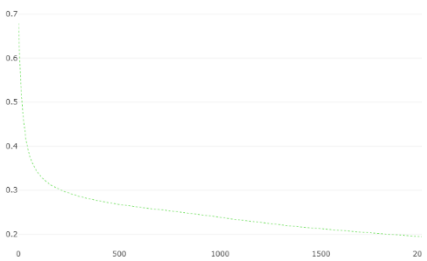
Tabel 4. Kombinasi Hyperparameter dan Score Akurasi

learning_rate	l2_leaf_reg	iterations	depth	Accuracy	rank_score
0.012	1	2000	10	0.8794	1
0.055	5	1000	10	0.8731	2
0.055	1	2000	6	0.8705	3
0.064	5	500	6	0.8683	4
0.064	3	2000	3	0.8674	5
0.055	3	2000	3	0.8653	6
0.012	5	1000	3	0.8522	7
0.1	3	1000	1	0.8283	8
0.064	3	1000	1	0.8260	9
0.012	1	1000	1	0.8121	10

Tabel 4 menunjukkan 10 model yang telah dibangun dengan 10 kombinasi hyperparameter yang berbeda serta skor akurasi yang didapatkan dari model tersebut. Hyperparameter terbaik ditunjukkan pada model dengan rank_score = 1. Model tersebut menghasilkan akurasi tertinggi, yaitu: 0,8794 atau dalam persentase sebesar 87.94%.

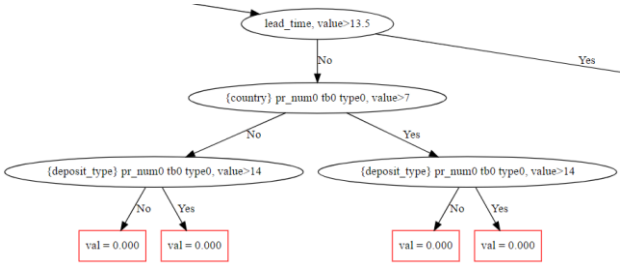
```
Model Tuned = CatBoostClassifier(iterations = 2000, depth = 10, learning_rate = 0.012, l2_leaf_reg = 1,
              random_state = 42)
```

Selanjutnya model yang sudah dioptimasi akan di latih terhadap data latih dan diuji terhadap data uji untuk didapatkan performanya.



Gambar 8. Visualisasi Log-loss training CatBoost

Gambar 8 menunjukkan proses pembelajaran mesin terhadap data latih mulai dari iterasi ke-0 hingga iterasi ke-2000. Error log loss terlihat berkurang seiring bertambahnya iterasi, artinya pembelajaran mesin belajar dengan baik.

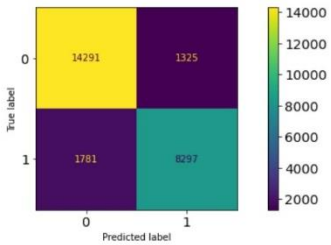


Gambar 9. Sampel Visualisasi Tree dari Model

Gambar 9 menunjukkan sampel dari visualisasi tree yang berbentuk simetris yang dihasilkan oleh metode CatBoost. Visualisasi tree tersebut yang merupakan bentuk interpertasi dari model yang dibuat.

4.8 Evaluasi

Setelah model sudah terbangun, model akan dievaluasi performa dalam melakukan prediksi. Pada penelitian ini, model akan dievaluasi menggunakan confussion matrix. Pada confussion matrix, dihasilkan True Positive (TP), True negative (TN), False Positive (FP) dan False Negative (FN).



Gambar 10. Confussion Matrix Data Uji

Gambar 10 menunjukkan visualisasi dari confussion matrix hasil dari prediksi terhadap data uji. Model menghasilkan TP sebesar 8.297 data berhasil diprediksi benar positif, TN sebesar 14.291 data berhasil diprediksi benar negatif, FP sebesar 1.325 data gagal diprediksi negatif dan FN sebesar 1.781 data gagal diprediksi positif.

Hasil perhitungan diatas menunjukkan bahwa model yang dibangun memiliki performa yang baik dalam memprediksi apakah pesanan kamar hotel dibatalkan atau tidak dengan akurasi sebesar 88% dan presisi sebesar 86%. Evaluasi selanjutnya yang akan dilakukan adalah pemeriksaan terhadap kekonsistenan model dalam memprediksi pesanan kamar hotel. Seberapa konsisten model dalam memprediksi, baik memprediksi data latih ataupun memprediksi data uji.

4.9 Analisis Model Klasifikasi

Setelah model sudah terbangun, selanjutnya model klasifikasi akan dianalisis melalui visualisasi – visualisasi yang dapat menunjukkan karakteristik dari pengunjung yang membatalkan pesanan. Gambar 26 menunjukkan urutan dari atribut – atribut yang penting dalam menentukan apakah seorang pengunjung akan membatalkan pesanannya atau tidak. Pada penelitian ini atribut yang dianggap penting adalah atribut yang memiliki nilai SHAP Values diatas +0.1, sehingga atribut yang paling berpengaruh dalam menentukan apakah suatu booking dibatalkan atau tidak adalah: country, deposit_type, agent, total_of_special_requests, customer_type, lead_time, previous_cancellations, required_car_parking_spaces, assigned_room_type, reserved_room_type, market_segment, arrival_date_month, booking_changes, adr.

Gambar 11 menunjukkan urutan dari atribut – atribut yang penting dalam menentukan apakah seorang pengunjung akan membatalkan pesanannya atau tidak. Pada penelitian ini atribut yang paling berpengaruh dalam menentukan apakah suatu booking dibatalkan atau tidak adalah: country, deposit_type, agent, total_of_special_requests, customer_type, lead_time, previous_cancellations, required_car_parking_spaces, assigned_room_type, reserved_room_type, market_segment, arrival_date_month, booking_changes, adr.

Gambar 11. Features Importance

5 Kesimpulan

Setelah dilakukannya penelitian terhadap prediksi apakah suatu pesanan hotel akan dibatalkan atau tidak menggunakan algoritma CatBoost dan mengetahui karakteristik dari pesanan yang dibatalkan, maka dapat disimpulkan sebagai berikut:

1. Hasil evaluasi yang dihasilkan menggunakan algoritma CatBoost dengan optimisasi hiperparameter mampu menghasilkan performa yang lebih baik dibandingkan dengan algoritma yang digunakan pada penelitian sebelumnya, yaitu Random Forest dengan optimisasi hiperparameter.
2. Dengan melakukan optimisasi hiperparameter didapatkan parameter – parameter terbaik dalam mengklasifikasikan pesanan kamar hotel menggunakan algoritma CatBoost diantaranya: depth = 10, iterations = 2000, l2_leaf_reg = 1, learning_rate = 0.012. Hiperparameter tersebut dapat menghasilkan model klasifikasi dengan akurasi yang dibulatkan sebesar 0.88 atau dalam persentase sebesar 88% dan nilai presisi sebesar 0.86 atau dalam persentase sebesar 86%.
3. Dihasilkan atribut yang paling berpengaruh dalam menentukan apakah suatu booking dibatalkan atau tidak adalah: country, deposit_type, agent, total_of_special_requests, customer_type, lead_time, previous_cancellations, required_car_parking_spaces, assigned_room_type, reserved_room_type, market_segment, arrival_date_month, booking_changes, adr.

Referensi

[1] Chen, C. C., & Xie, K. (2013). Differentiation of cancellation policies in the U.S. hotel industry. International Journal of Hospitality Management, 34(1), 66–72. https://doi.org/10.1016/j.ijhm.2013.02.007

[2] Chen, C. C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. International Journal of Hospitality Management, 30(1), 129–135. https://doi.org/10.1016/j.ijhm.2010.03.010

[3] Phumchusri, N., & Maneesophon, P. (2014). Optimal overbooking decision for hotel rooms revenue management. Journal of Hospitality and Tourism Technology, 5(3), 261–277. https://doi.org/10.1108/JHTT-03-2014-0006.

[4] Antonio, N., De Almeida, A., & Nunes, L. (2017). Predicting hotel bookings cancellation with a machine learning classification model. Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 2017-Decem, 1049–1054. https://doi.org/10.1109/ICMLA.2017.00-11

[5] Azhar, Y., Mahesa, G. A., & Mustaqim, M. C. (2021). Prediction of hotel bookings cancellation using hyperparameter optimization on Random Forest algorithm. Jurnal Teknologi Dan Sistem Komputer, 9(1), 15–21. https://doi.org/10.14710/jtsiskom.2020.13790

[6] Haynes, N., & Egan, D. (2020). The perceptions of frontline employees towards hotel overbooking practices: exploring ethical challenges. Journal of Revenue and Pricing Management, 19(2), 119–128. https://doi.org/10.1057/s41272-019-00226-1

[7] Pimentel, V., Eziz, A., & Baker, T. (2021). Patterns in Hotel Revenue Management Forecasting Systems: Improved Sample Sizes, Frozen Intervals, Horizon Lengths, and Accuracy Measures. Mathematics and Computer Science, 6(1), 8. https://doi.org/10.11648/j.mcs.20210601.12

[8] Hurwitz, J., & Kirsch, D. (2018). Machine Learning For Dummies®, IBM Limited Edition Published (C. A. Burchfield (ed.)). John Wiley & Sons, Inc.
[9] Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. Procedia CIRP, 79, 403–408. https://doi.org/10.1016/j.procir.2019.02.106
[10] Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Secondary Analysis of Electronic Health Records. Secondary Analysis of Electronic Health Records, October, 1–427. https://doi.org/10.1007/978-3-319-43742-2
[11] Hancock, J. T., & Khoshgoftaar, T. M. (2020b). CatBoost for big data: an interdisciplinary review. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00369-8