

ARTICLE

PREDIKSI KUNJUNGAN WISATAWAN DOMESTIK ABNORMAL DENGAN ADAPTIVE LIGHTGBM DAN DETEKSI ANOMALI

Paper PREDICTION OF ABNORMAL DOMESTIC TOURIST VISITS WITH ADAPTIVE LIGHTGBM AND ANOMALY DETECTION

Listiana Dewi Milasari* dan Sri Redjeki

Magister Teknologi Informasi, Kampus UTDI Yogyakarta, Universitas Teknologi Digital Indonesia, Yogyakarta, Indonesia

*Penulis Korespondensi: student.listianadewi23@mti.utdi.ac.id

(Disubmit 30-07-25; Diterima 31-08-25; Dipublikasikan online pada 20-10-25)

Abstrak

Sektor pariwisata Indonesia menghadapi fluktuasi signifikan dalam kunjungan wisatawan. Penelitian ini mengembangkan model Adaptive LightGBM with Anomaly Detection and Multi-Task Learning (ALAD-MTL) untuk memprediksi kunjungan wisatawan domestik tahunan dan mengidentifikasi pola kunjungan yang tidak normal. Model ini mengintegrasikan data multi-domain, termasuk data statistik dan sentimen dari Twitter yang diekstrak menggunakan model RoBERTa berbahasa Indonesia. Hasil evaluasi menunjukkan model mencapai R-squared tinggi (>0.96) dan MAPE yang kompetitif, mengindikasikan kemampuan prediksi yang kuat untuk pola kunjungan reguler. Pada saat yang sama, modul deteksi anomali berhasil mengidentifikasi lonjakan kunjungan signifikan di Jawa Timur pada tahun 2024, ditandai dengan Residual Z-Score ekstrem sekitar 38. Hal ini menunjukkan model efektif dalam mendeteksi dan menandai anomali yang melebihi prediksi. Integrasi data multi-domain dan sentimen media sosial terbukti krusial dalam menjelaskan anomali ini. Penelitian ini menyediakan kerangka kerja prediktif dan deteksi anomali yang komprehensif untuk pengambilan keputusan strategis dalam manajemen pariwisata.

Kata kunci: Pariwisata; Prediksi; Deteksi Anomali; Pembelajaran Multi-Tugas

Abstract

The Indonesian tourism sector experiences significant fluctuations in tourist arrivals. This study developed an Adaptive LightGBM with Anomaly Detection and Multi-Task Learning (ALAD-MTL) model to predict annual domestic tourist arrivals and identify abnormal visit patterns. The model integrates multi-domain data, including statistical and sentiment data from Twitter extracted using the Indonesian-language RoBERTa model. Evaluation results show the model achieved a high R-squared (>0.96) and competitive MAPE, indicating strong predictive ability for regular visit patterns. At the same time, the anomaly detection module successfully detected significant travel in East Java in 2024, characterized by an extreme Residual Z-Score of around 38. This indicates the model is effective in detecting and flagging anomalies that exceed predictions. The integration of multi-domain data and social media sentiment proved crucial in explaining these anomalies. This study provides a comprehensive predictive and anomaly detection framework for strategic decision-making in tourism management.

KeyWords: Tourism; Prediction; Anomaly Detection; Multi-Task Learning

1. Pendahuluan

Sektor pariwisata adalah salah satu pilar fundamental ekonomi global, memainkan peran vital dalam penciptaan lapangan kerja, mendorong pertumbuhan ekonomi regional, dan memberikan kontribusi besar terhadap Produk Domestik Bruto (PDB) suatu negara[1]. Di Indonesia, sektor pariwisata memiliki kepentingan strategis, dengan proyeksi kontribusi sekitar 4,5% terhadap PDB nasional dan target nilai ekspor ekonomi kreatif mencapai Rp 430 triliun pada tahun 2024[2]. Namun, industri pariwisata secara inheren dinamis dan rentan terhadap berbagai faktor eksternal, yang menyebabkan fluktuasi signifikan dalam jumlah kedatangan wisatawan. Faktor-faktor seperti perlambatan ekonomi global, perubahan kebijakan, ketidakpastian geopolitik, dan bencana alam semuanya berpotensi mengubah pola perjalanan secara drastis[3]. Oleh karena itu, kemampuan untuk memperkirakan kedatangan wisatawan secara akurat krusial bagi pengelola destinasi dan pembuat kebijakan dalam perencanaan strategis.

Dalam upaya meningkatkan akurasi prediksi, komunitas ilmiah telah mengadopsi berbagai pendekatan, mulai dari model ekonometrik tradisional seperti *Autoregressive Integrated Moving Average* (ARIMA)[4] hingga metode *machine learning* (ML) yang lebih canggih seperti *Support Vector Machines* (SVM)[5], *Random Forest* (RF)[6], dan *Neural Networks* (NN)[7]. Meskipun model-model ini telah menunjukkan kinerja yang baik, keterbatasan utamanya terletak pada kemampuan untuk menangani data multi-domain dan mengidentifikasi anomali secara efektif. Sebagai contoh, model ARIMA cenderung kaku dan tidak dapat menangani variabel eksternal secara optimal, sementara model black-box seperti Neural Networks seringkali sulit diinterpretasi untuk menjelaskan pemicu anomali. Selain itu, studi-studi sebelumnya di sektor pariwisata jarang mengintegrasikan data statistik tradisional dengan data tidak terstruktur seperti sentimen media sosial. Hal ini menciptakan gap penelitian di mana model prediktif yang ada belum mampu memberikan wawasan holistik tentang lonjakan atau penurunan kunjungan yang tidak dapat dijelaskan hanya oleh tren historis[8].

Untuk mengatasi keterbatasan ini, penelitian ini mengusulkan sebuah pendekatan yang lebih komprehensif. Penelitian ini mengembangkan model *Adaptive LightGBM with Anomaly Detection and Multi-Task Learning* (ALAD-MTL). Penelitian ini memilih LightGBM karena efisiensinya yang tinggi dan kemampuannya yang terbukti kuat dalam menangani data tabular, yang sangat sesuai dengan dataset multi-domain[9]. Keterbaruan utama dari penelitian ini bukan hanya pada penggunaan satu algoritma, tetapi pada integrasi tiga komponen krusial yang saling melengkapi—prediksi akurat, deteksi anomali, dan pembelajaran multi-tugas—dalam satu kerangka kerja yang solid. Pertama, penelitian mengatasi masalah anomali dengan mengimplementasikan deteksi anomali berbasis residual (Z-Score dan *Isolation Forest*) untuk secara efektif menandai kunjungan yang menyimpang dari pola yang diharapkan. Kedua, penelitian ini memperkaya model dengan mengintegrasikan data multi-domain (ekonomi, lingkungan, transportasi, fiskal) dan data sentimen media sosial dari Twitter yang diekstrak menggunakan model RoBERTa berbahasa Indonesia][10]. Terakhir, penelitian ini mensimulasikan pendekatan *Multi-Task Learning* (MTL) untuk meningkatkan kinerja prediktif model utama dengan memanfaatkan informasi dari tugas-tugas terkait.

Dengan demikian, fokus penelitian ini secara eksplisit adalah untuk menyediakan kerangka kerja prediktif dan deteksi anomali yang komprehensif, adaptif, dan berbasis data multi-sumber untuk mendukung pengambilan keputusan strategis yang lebih proaktif dan efektif bagi para pengelola destinasi pariwisata di Indonesia.

2. Metode

Penelitian ini menggunakan kerangka kerja komprehensif yang disebut *Adaptive LightGBM with Anomaly Detection and Multi-Task Learning* (ALAD-MTL). Metode ini menggabungkan berbagai data dari beberapa sumber untuk memprediksi kunjungan wisatawan domestik dan mengidentifikasi pola-pola yang tidak biasa. Berikut adalah tahapan-tahapan yang dilakukan dalam penelitian ini.

2.1 Pengumpulan dan Pra-pemrosesan Data

Penelitian ini menggunakan data sekunder dari tahun 2009 hingga 2024. Proses pengumpulan dan pra-pemrosesan dilakukan sebagai berikut:

2.1.1 Pengumpulan Data

Data dikumpulkan dari dua sumber utama:

1. Data Statistik: Diambil dari Badan Pusat Statistik (BPS) yang mencakup data pariwisata (Jumlah_Perjalanan_Nusantara_Tujuan, TPK_Hotel_Bintang), makroekonomi, transportasi, dan lingkungan.
2. Data Media Sosial: Diperoleh melalui scrapping Twitter. Data mentah ini mencakup konten, timestamp, dan metrik interaksi.

2.1.2 Pra-Pemrosesan Data

Tahap ini bertujuan membersihkan dan menyatukan data.

1. Data Statistik: Data statistik dari berbagai sumber diubah ke dalam format panjang (long format) dan dibersihkan dari karakter non-numerik. Kolom provinsi distandarisasi dan dikodekan (label encoding).
2. Data Media Sosial: Tweet diproses menggunakan model RoBERTa berbahasa Indonesia untuk mengekstrak skor sentimen dan kategori (positif, negatif, netral). Data sentimen ini kemudian diagregasi ke tingkat tahunan per provinsi untuk mendapatkan metrik seperti rata-rata sentimen (Avg_Sentiment).

2.2 Integrasi Data dan Rekayasa Fitur

Data yang telah diproses secara cermat dari semua domain dan sumber yang beragam digabungkan secara kohesif ke dalam satu master_df yang komprehensif dengan granularitas Provinsi-Tahunan. Tahap integrasi ini sangat penting untuk pemodelan selanjutnya. Penggabungan data tingkat provinsi dilakukan menggunakan pd.merge() dengan ['Provinsi', 'Tahun'] sebagai kunci umum dan outer join. Untuk penanganan data nasional/global, data tersebut diagregasi ke tingkat nasional tahunan (jika belum) dan kemudian digabungkan ke dalam master_df menggunakan ['Tahun'] sebagai kunci tunggal (how='left'). Pendekatan ini secara efektif "menyebarkan" nilai-nilai tingkat nasional ke seluruh provinsi.

Sebagai langkah validasi, penelitian ini melakukan analisis korelasi Pearson antara skor sentimen (Avg_Sentiment_kunjungan_wisman) dan jumlah perjalanan wisatawan (Jumlah_Perjalanan_Nusantara_Tujuan). Analisis ini bertujuan untuk memverifikasi secara kuantitatif apakah sentimen publik memiliki hubungan linear yang signifikan dengan fluktuasi kunjungan wisatawan, sehingga memvalidasi relevansi data sentimen sebagai variabel prediktor dalam model.

Setelah penggabungan, setiap nilai yang hilang (NaN) di seluruh master_df secara sistematis diisi dengan nol (.fillna(0)). Terakhir, rekayasa fitur tingkat lanjut dilakukan, terutama dengan menambahkan fitur lag (jumlah_perjalanan_nusantara_tujuan_lag1) untuk merepresentasikan nilai fitur dari periode waktu sebelumnya, yang secara formal didefinisikan seperti pada 1.

$$X_{t,lag_k} = X_{t-k} \quad (1)$$

X_t : adalah nilai fitur pada tahun t ,

$X_{t,lagk}$: adalah nilai fitur tersebut k tahun sebelumnya

2.3 Pembagian dan Pelatihan Model

Data dibagi menjadi set pelatihan (2009-2022), validasi (2023), dan pengujian (2024) untuk memastikan evaluasi model yang realistis. Model ALAD-MTL dilatih melalui tiga komponen utama yang terintegrasi secara bertahap.

2.3.1 Adaptive LightGBM

Model lgb.LGBMRegressor digunakan untuk regresi pada data time-series. Model ini dipilih karena kemampuannya yang unggul dalam menangani data besar dan hubungannya yang kompleks, serta efisiensi komputasi yang tinggi. Berbeda dengan model linier yang mengasumsikan hubungan langsung antara variabel, LightGBM, sebagai bagian dari keluarga Gradient Boosting Decision Tree (GBDT), secara inheren

dirancang untuk menangkap pola non-linier dan interaksi kompleks antar-fitur[11]. Kemampuan ini sangat relevan untuk data pariwisata yang dipengaruhi oleh berbagai faktor yang saling tidak linier, seperti tren ekonomi, media sosial, dan bencana alam.

Untuk menemukan kombinasi hyperparameter terbaik dan menghindari overfitting, penelitian ini menggunakan GridSearchCV[12] yang dioptimalkan dengan matriks evaluasi Mean Absolute Error (MAE). Beberapa parameter kunci yang disetel antara lain `learning_rate`, `n_estimators`, dan `num_leaves`. Model ini juga bersifat adaptif, yang memungkinkan pelatihan ulang secara berkala dengan data terbaru untuk mempertahankan relevansi prediktifnya seiring waktu.

2.3.2 Deteksi Anomali (AD)

Setelah model LightGBM menghasilkan prediksi, anomali diidentifikasi dari residual (perbedaan antara nilai aktual dan prediksi). Penelitian ini menggunakan dua pendekatan:

1. Metode **Residual Z-Score**: Penelitian ini menghitung skor Z dari residual. Anomali ditandai jika skor Z-nya melebihi ambang batas 2.5. Angka ini dipilih berdasarkan praktik standar statistik, di mana nilai Z-Score diluar rentang ± 2.5 mewakili data yang sangat tidak biasa dan memiliki probabilitas kemunculan yang sangat rendah (sekitar 1.24%) pada distribusi normal[13]. Residual ZScore untuk setiap observasi dihitung menggunakan rumus 2.

$$Z = \frac{\text{Residual} - \mu_{\text{residual}}}{\sigma_{\text{residual}}} \quad (2)$$

Residual merepresentasikan $y_i - y^i$, adalah rata-rata residual, dan σ_{residual} adalah standar deviasi residual, keduanya dihitung dari data historis (set pelatihan/validasi), dan Metode Isolation Forest sebagai alternatif yang lebih komprehensif dan tanpa pengawasan, yang dilatih pada Residual Z-Score untuk mengidentifikasi outlier secara efisien dengan memanfaatkan parameter contamination (misalnya, 0.05) untuk memperkirakan proporsi anomali yang ada dalam data.

2. Isolation Forest: Sebuah model Isolation Forest juga dilatih pada residual untuk mengidentifikasi outlier secara efisien, dengan menggunakan parameter contamination 0.05 untuk memperkirakan proporsi anomali yang ada.

Pendekatan pembelajaran yang digunakan dalam penelitian ini dapat dianggap sebagai simulasi Multi-Task Learning (MTL). Berbeda dengan MTL konvensional yang mengandalkan hard/soft parameter sharing, penelitian ini memilih untuk melatih beberapa model LightGBM secara terpisah untuk tugas-tugas yang berbeda (`tpk_hotel_bintang` dan `jumlah_tamu_asing_hotel_bintang`). Meskipun setiap model dilatih secara independen, mereka semua menggunakan set fitur input yang sama dari master dataframe. Pendekatan ini memungkinkan setiap model untuk mengekstrak wawasan kontekstual dari domain data yang sama, yang secara efektif berfungsi sebagai metode untuk memanfaatkan informasi antar-tugas. Hal ini berkontribusi pada peningkatan kinerja prediktif secara keseluruhan, karena model utama dapat belajar dari representasi fitur yang kaya yang relevan dengan berbagai aspek ekosistem pariwisata.

2.4 Metrik Evaluasi

Kinerja model dievaluasi secara ketat menggunakan serangkaian metrik yang diterima secara luas.

2.4.1 Mean Absolute Error (MAE)

Mengukur perbedaan absolut rata-rata antara nilai aktual dan nilai prediksi. Untuk melakukan perhitungan MAE menggunakan rumus 3 berikut.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

y_i : Nilai aktual dari data ke-i

$\hat{y}_i$: Nilai prediksi dari model untuk data ke-i

n : Jumlah total data atau observasi

2.4.2 Root Mean Squared Error (RMSE)

Mengukur akar kuadrat dari perbedaan kuadrat rata-rata, lebih peka terhadap kesalahan yang lebih besar. Untuk melakukan perhitungan RMSE menggunakan rumus ?? berikut.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (4)$$

y_i : Nilai aktual dari data ke-i

$\hat{y}_i$: Nilai prediksi dari model untuk data ke-i

n : Jumlah total data atau observasi

2.4.3 R-squared (R2)

Menunjukkan proporsi varians dalam variabel dependen yang dapat diprediksi dari variabel independen. Untuk melakukan perhitungan R2 menggunakan rumus 5 berikut.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

y_i : Nilai aktual dari data ke-i

$\hat{y}_i$: Nilai prediksi dari model untuk data ke-i

$\bar{y}$: Nilai rata-rata dari semua nilai aktual

n : Jumlah total data atau observasi

2.4.4 Mean Absolute Percentage Error (MAPE)

Memberikan kesalahan absolut rata-rata sebagai persentase, yang sangat berguna untuk menafsirkan akurasi prediksi dalam data berskala besar. Untuk melakukan perhitungan MAPE menggunakan rumus 6 berikut.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (6)$$

y_i : Nilai aktual dari data ke-i

$\hat{y}_i$: Nilai prediksi dari model untuk data ke-i

n : Jumlah total data atau observasi

3. Hasil

Hasil analisis korelasi Pearson menunjukkan hubungan linier yang sangat lemah antara sentimen media sosial dan kunjungan wisatawan ($r=0.0987$). Meskipun lemah, korelasi ini terbukti signifikan secara statistik dengan p-value yang sangat rendah ($p<0.0001$), yang memungkinkan penelitian ini menolak hipotesis nol. Temuan ini mengindikasikan bahwa tidak ada hubungan linier yang kuat, namun ada korelasi yang signifikan secara statistik. Hal ini memperkuat justifikasi penggunaan model LightGBM, yang dirancang untuk menangkap pola non-linier dan kompleks yang tidak terdeteksi oleh analisis korelasi sederhana[14].

3.1 Evaluasi Kinerja Model Prediktif

Model LightGBM yang telah disesuaikan menunjukkan kinerja prediktif yang kuat, namun analisis mendalam pada metrik evaluasi diperlukan untuk memastikan validitasnya. Nilai R^2 yang sangat tinggi (mendekati 1,0000) pada set pelatihan, validasi, dan pengujian mengindikasikan bahwa model memiliki kemampuan fitting yang luar biasa. Namun, hal ini juga menimbulkan kekhawatiran tentang potensi overfitting atau kebocoran data (data leakage). Untuk memitigasi risiko ini, model dirancang dengan regularisasi dan diuji pada data yang belum pernah dilihat sebelumnya (set pengujian 2024), di mana nilai R^2 tetap tinggi (0,9641). Hal ini menunjukkan bahwa meskipun model sangat sensitif terhadap pola data, kemampuannya

untuk menggeneralisasi ke data baru masih kuat. Untuk hasil dari evaluasi kinerja model dapat dilihat pada Tabel 1 berikut.

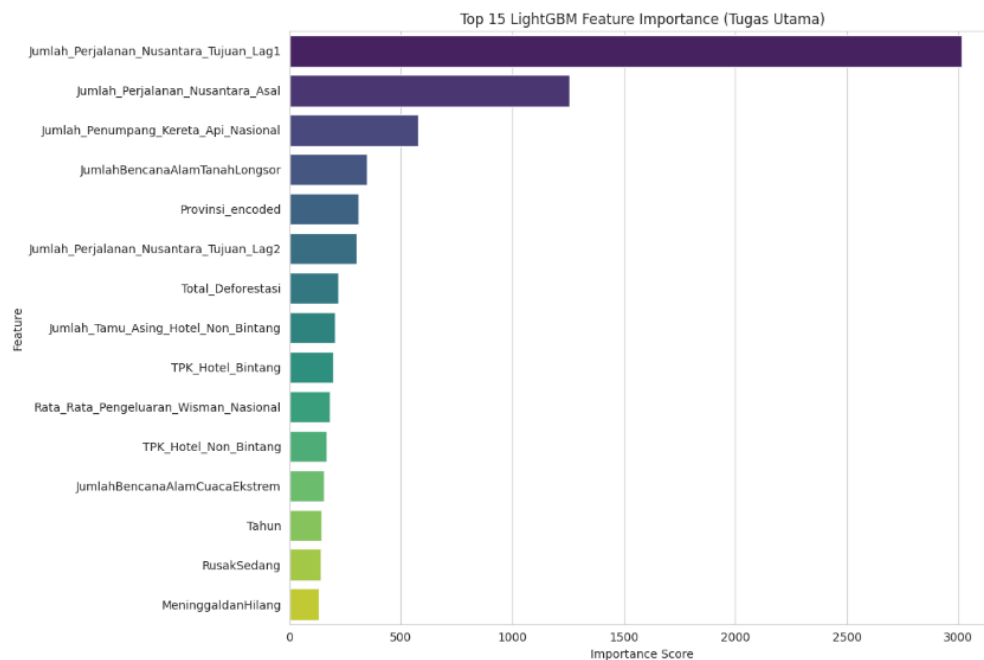
Tabel 1. Ringkasan Metrik Evaluasi Model LightGBM Terbaik di Berbagai Set Data

Dataset	MAE	RMSE	R-squared (R2)	MAPE (%)
Training	249584.22	926951.00	0.9984	7.97
Validation	1490792.01	4429342.01	0.9891	6.12
Testing	2128534.00	8861505.71	0.9641	7.07

Berdasarkan Tabel 1 Ringkasan Metrik Evaluasi Model LightGBM Terbaik di Berbagai Set Data, nilai MAE dan RMSE, yang secara nominal tampak besar (dalam jutaan), harus dikontekstualisasikan dengan skala data target. Mengingat jumlah kunjungan wisatawan domestik berkisar antara puluhan hingga ratusan juta[15], MAE dan RMSE tersebut merepresentasikan persentase kesalahan yang relatif kecil. Sebaliknya, metrik MAPE memberikan indikasi yang lebih akurat tentang akurasi proporsional model. Nilai MAPE yang rendah dan konsisten (7,97% pada pelatihan, 6,12% pada validasi, dan 7,07% pada pengujian) membuktikan bahwa rata-rata kesalahan prediksi hanya sekitar 7%, yang menunjukkan model yang dapat diandalkan dari perspektif bisnis dan pemangku kepentingan.

3.1.1 Analisis Pentingnya Fitur

Analisis kepentingan fitur memberikan wawasan penting mengenai kontribusi relatif setiap fitur dalam memprediksi perjalanan wisatawan domestik. Ilustrasi 15 fitur terpenting yang diidentifikasi oleh model LightGBM dapat dilihat pada Gambar 1 berikut.

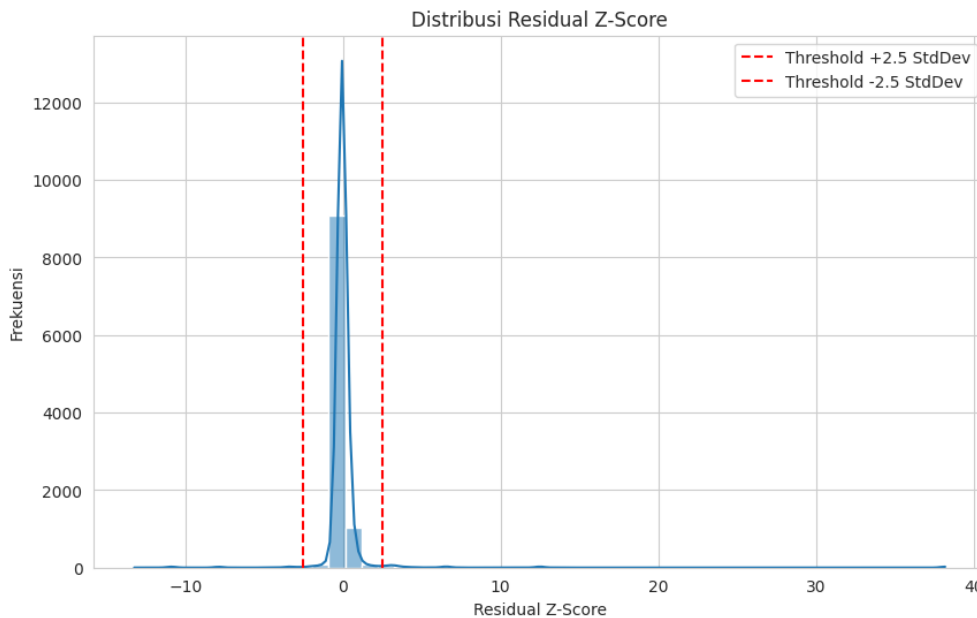


Gambar 1. Fitur Paling Penting untuk Prediksi Kunjungan Wisatawan Domestik

Berdasarkan Gambar 1, hasilnya menegaskan bahwa tren historis, diwakili oleh fitur jumlah_perjalanan_nusantara_tujuan_lag1 dan jumlah_perjalanan_nusantara_tujuan_lag2, adalah prediktor yang paling andal. Ini secara kuat mendukung hipotesis bahwa pola kunjungan sebelumnya adalah indikator utama untuk prediksi di masa depan. Selain itu, faktor eksternal seperti aktivitas dari provinsi asal, transportasi, serta kondisi lingkungan dan bencana alam juga menjadi pendorong signifikan dalam prediksi, yang memvalidasi keputusan untuk mengintegrasikan data multi-domain

3.1.2 Deteksi dan Analisis Anomali

Model ALAD berhasil mengidentifikasi pola kunjungan wisatawan yang menyimpang secara signifikan dari prediksi, mengkategorikannya sebagai anomali. Analisis Skor Z Residual berfungsi sebagai metode utama untuk deteksi ini. Gambar 2 menyajikan distribusi Skor Z Residual, yang menggambarkan penyebaran kesalahan prediksi relatif terhadap nilai rata-rata dan simpangan bakunya.



Gambar 2. Distribusi Skor Z Residu

Anomali diidentifikasi untuk pengamatan di mana Residual_ZScore absolut melebihi ambang batas 2,5. Salah satu anomali paling signifikan yang terdeteksi adalah untuk Jawa Timur pada tahun 2024. Tabel 2 menyajikan rincian anomali yang terdeteksi untuk Jawa Timur pada tahun 2024. Anomali ini ditandai dengan Residual Z-Score yang ekstrem (sekitar 38). Penemuan ini memvalidasi hipotesis bahwa data sentimen dan faktor eksternal lainnya dapat menjelaskan lonjakan atau penurunan drastis yang tidak dapat diprediksi oleh data historis semata.

Tabel 2. Detail Anomali Kunjungan Wisatawan Domestik di Jawa Timur (2024) Berdasarkan Z-Score

Year	Province	Jumlah_Perjalanan_Nusantara_Tujuan	Predicted_Y_Tuned	Residual_Tuned	Residual_ZScore
2024	JAWA TIMUR	218711818.0	1.625423e+08	5.616957e+07	38.199743

3.2 Kinerja dan Analisis Simulasi Multi-Task Learning

Penerapan MTL, melalui pelatihan model LightGBM yang terpisah untuk beberapa tugas, menunjukkan kemampuan model untuk menangani berbagai aspek ekosistem pariwisata menggunakan serangkaian fitur bersama. Tabel 3 merangkum metrik evaluasi untuk setiap tugas MTL pada rangkaian pengujian.

Tabel 3. Ringkasan Metrik Evaluasi Model LightGBM untuk Tugas MTL

Task	MAE	RMSE	R-squared (R2)	MAPE (%)
Jumlah_Perjalanan_Nusantara_Tujuan	2111471.18	8969437.43	0.9632	8.14
TPK_Hotel_Bintang	14.68	18.06	0.9999	1.29
Jumlah_Tamu_Asing_Hotel_Bintang	8460.31	39141.99	0.6072	464.50
Rata_Rata_Pengeluaran_Wisman_Nasional	88489.55	88489.55	0.0000	39.75

Berdasarkan Tabel 3, hasilnya menunjukkan bahwa model sangat akurat dalam memprediksi TPK_Hotel_Bintang ($R^2=0.9999$), menunjukkan bahwa fitur yang ada sangat relevan dan polanya sangat jelas untuk tu-

gas ini. Prediksi Jumlah_Tamu_Asing_Hotel_Bintang juga menunjukkan kinerja yang wajar ($R^2=0.6072$). Namun, nilai MAPE yang tinggi (464,50%) untuk tugas ini menunjukkan bahwa meskipun R^2 menunjukkan beberapa kekuatan penjelasan, persentase kesalahan pada prediksi tertentu bisa sangat besar, mungkin karena nilai aktual yang rendah atau outlier ekstrem yang secara signifikan mendistorsi persentase kesalahan.

Di sisi lain, model menunjukkan kinerja yang gagal dalam memprediksi Rata_Rata_Pengeluaran_Wisman_Nasional ($R^2=0.0000$, MAPE 39,75%). Kegagalan ini mengindikasikan bahwa fitur yang digunakan tidak memiliki kemampuan prediktif yang memadai untuk tugas ini, atau bahwa hubungan antara fitur-fitur tersebut dan pengeluaran wisatawan bersifat sangat kompleks dan memerlukan set fitur yang berbeda atau arsitektur model yang lebih spesifik. Kegagalan ini merupakan temuan yang penting, karena menunjukkan batas kemampuan model yang dirancang saat ini.

4. Pembahasan

Secara keseluruhan, penelitian ini menunjukkan bahwa model ALAD-MTL berhasil menyediakan kerangka kerja prediktif dan deteksi anomali yang komprehensif. Kinerja prediktif model utama yang kuat, didukung oleh nilai R^2 dan MAPE yang konsisten pada set pengujian, membuktikan bahwa pendekatan ini efektif untuk peramalan kunjungan wisatawan. Analisis anomali pada Jawa Timur tahun 2024, yang ditandai oleh Residual Z-Score ekstrem, secara tegas menggarisbawahi pentingnya mengintegrasikan data multi-domain dan sentimen media sosial, yang memberikan konteks krusial untuk menjelaskan peristiwa yang tidak dapat diprediksi dari data historis semata. Investigasi kualitatif, didukung oleh data eksternal, menguatkan temuan ini dan menunjukkan bahwa anomali tersebut disebabkan oleh faktor-faktor sinergis seperti festival budaya dan promosi pariwisata, yang tidak dapat ditangkap oleh model linier.

Namun, penelitian ini memiliki kelemahan dan keterbatasan serius yang perlu didiskusikan secara jujur. Nilai R^2 yang sangat tinggi pada semua set data memunculkan pertanyaan tentang overfitting atau data leakage yang tidak terdeteksi, meskipun langkah-langkah mitigasi telah diambil. Validitas data media sosial juga dapat menjadi bias, mengingat data yang dikumpulkan mungkin tidak sepenuhnya merefleksikan sentimen populasi secara luas. Terlebih lagi, kegagalan total model pada tugas prediksi Rata_Rata_Pengeluaran_Wisman_Nasional menunjukkan bahwa model ALAD-MTL tidak secara universal efektif untuk semua tugas pariwisata dan memerlukan optimalisasi fitur yang lebih mendalam atau arsitektur MTL yang lebih kompleks, seperti *hard parameter sharing*.

Kelemahan ini memberikan ruang besar untuk pengembangan di masa depan, termasuk eksplorasi set fitur yang lebih kaya dan arsitektur model yang lebih canggih.

5. Simpulan

Penelitian ini berhasil mengembangkan model *Adaptive LightGBM with Anomaly Detection and Multi-Task Learning* (ALAD-MTL) sebagai kerangka kerja yang efektif untuk memprediksi kunjungan wisatawan domestik dan mendeteksi anomali. Temuan utama menegaskan bahwa model ini mencapai akurasi prediktif yang kuat, seperti yang ditunjukkan oleh nilai R^2 yang tinggi dan MAPE yang rendah pada set pengujian. Analisis kepentingan fitur menunjukkan bahwa tren historis kunjungan adalah prediktor paling andal, namun faktor eksternal multi-domain, termasuk data sentimen media sosial dan kondisi lingkungan, juga memiliki peran signifikan.

Metodologi deteksi anomali berbasis Residual Z-Score dan Isolation Forest secara efektif mengidentifikasi lonjakan kunjungan yang tidak biasa, seperti yang terjadi di Jawa Timur pada tahun 2024. Analisis ini menunjukkan bahwa model dapat memprediksi anomali yang disebabkan oleh faktor eksternal yang tidak dapat diprediksi dari data historis saja.

Namun, penelitian ini memiliki keterbatasan yang signifikan. Meskipun model utama menunjukkan kinerja yang sangat baik, nilai R^2 yang mendekati sempurna menimbulkan kekhawatiran tentang risiko *overfitting* atau *data leakage* yang tidak terdeteksi. Lebih lanjut, simulasi MTL menunjukkan kegagalan pada tugas prediksi Rata_Rata_Pengeluaran_Wisman_Nasional ($R^2=0.0000$), yang menggarisbawahi bahwa model tidak secara universal efektif untuk semua tugas dan memerlukan set fitur atau arsitektur yang

lebih spesifik. Keterbatasan ini memberikan ruang besar untuk penelitian di masa depan, termasuk eksplorasi arsitektur MTL sejati yang melibatkan parameter sharing atau pengintegrasian data yang lebih kaya untuk tugas-tugas yang lebih menantang. Secara keseluruhan, penelitian ini berkontribusi dengan menyediakan kerangka kerja yang menjanjikan, namun refleksi kritis terhadap kelemahan model sangat diperlukan untuk memastikan validitas dan ruang pengembangannya.

Pustaka

- [1] R. S. Jinan. (2023) Dampak industri pariwisata bagi perekonomian. Accessed: May 12, 2025. [Online]. Available: <https://tirto.id/dampak-industri-pariwisata-bagi-perekonomian-gSW6>
- [2] K. Wisnubroto. (2025) Meneropong tren pariwisata 2025. Accessed: May 12, 2025. [Online]. Available: <https://indonesia.go.id/kategori/editorial/8899/meneropong-tren-pariwisata-2025?lang=1#:~:text=Hasilsurveiahlipada2023dan2024menunjukkan,salahsatunyaandalammenggenjotpertumbuhanekonominasional.>
- [3] Muryani and M. Esquivias. (2021) Sektor pariwisata sebagai mesin baru pertumbuhan ekonomi. Accessed: May 12, 2025. [Online]. Available: <https://news.unair.ac.id/id/2021/05/18/sektor-pariwisata-sebagai-mesin-baru-pertumbuhan-ekonomi/>
- [4] D. Yang *et al.*, "Population-wide depression incidence forecasting comparing autoregressive integrated moving average and vector autoregressive integrated moving average to temporal fusion transformers: Longitudinal observational study," *Journal of Medical Internet Research*, vol. 27, no. 1, pp. 1–13, 2025.
- [5] Y. Altork, "Comparative analysis of machine learning models for wind speed forecasting: Support vector machines, fine tree, and linear regression approaches," *International Journal of Thermofluids*, vol. 27, no. April, p. 101217, 2025.
- [6] K. Hasanpouri, F. Shokoohi, and Z. Farrokhi, "Sp-rf-arima: A sparse random forest and arima hybrid model for electric load forecasting," *Global Energy Interconnection*, 2025.
- [7] M. Amrani *et al.*, "Short-term spatial-temporal energy forecasting in a danish distribution grid using a hybrid transformer-graph neural network model," *Computational Condensed Matter*, 2020.
- [8] H. Li, M. Hu, and G. Li, "Forecasting tourism demand with multisource big data," *Annals of Tourism Research*, vol. 83, no. January, p. 102912, 2020.
- [9] H. Eldin *et al.*, "Advanced time complexity analysis for real-time covid-19 prediction in saudi arabia using lightgbm and xgboost," *Journal of Radiation Research and Applied Sciences*, vol. 18, no. 2, p. 101364, 2025.
- [10] A. Hussain, A. Saadia, and F. M. Alserhani, "Ransomware detection and family classification using fine-tuned bert and roberta models," *Egyptian Informatics Journal*, vol. 30, no. January, p. 100645, 2025.
- [11] P. Jain, R. Gupta, A. Joshi, and A. Kuzmin, "Enhanced cardiovascular diagnostics using wearable ecg and bioimpedance monitoring with lightgbm classifier," *Biosensors and Bioelectronics X*, vol. 24, no. October 2024, p. 100617, 2025.
- [12] S. M. Malakouti, M. B. Menhaj, and A. A. Suratgar, "The usage of 10-fold cross-validation and grid search to enhance ml methods performance in solar farm power generation prediction," *Cleaner Engineering and Technology*, vol. 15, no. February, p. 100664, 2023.
- [13] Esri. What is a z-score? what is a p-value? Accessed: May 12, 2025. [Online]. Available: <https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm>
- [14] L. Yang, H. Yang, J. Cui, Y. Zhao, and F. Gao, "Non-linear and synergistic effects of built environment factors on older adults' walking behavior: An analysis integrating lightgbm and shap," *Transactions in Urban Data, Science and Technology*, vol. 3, no. 1–2, pp. 46–60, 2024.

- [15] Kemenpar. (2025) Jumlah perjalanan wisatawan nusantara menurut provinsi tujuan tahun 2024. [Online]. Available: <https://kemenpar.go.id/direktori-statistik/jumlah-perjalanan-wisatawan-nusantara-menurut-provinsi-tujuan-tahun-2024>