

Jurnal Mahasiswa Aplikasi Teknologi Komputer dan Informasi

Vol. 8 No. 2 Thn 2026 Hal 243-248
E ISSN : 2685-6565

PREDIKSI CLICK-THROUGH RATE UNTUK OPTIMASI BIDDING IKLAN DIGITAL MENGGUNAKAN ALGORITMA

Valencia Anggelica Hadinoto[a], Erlin[b]

[a]*Institut Bisnis dan Teknologi Pelita Indonesia Jl. Jend. Ahmad Yani No. 78-88, vallencia.anggelica@student.pelitaindonesia.ac.id*
[b] *Institut Bisnis dan Teknologi Pelita Indonesia, Jl. Jend.Ahmad Yani No.77-88 Pekanbaru, erlin@lecturer.pelitaindonesia.id*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 11 Juli 2026
Revisi Akhir: 30 Juli 2026
Diterbitkan *Online*: 30 agustus 2026

KATA KUNCI

Click-through Rate, Machine Learning, Digital Advertising

KORESPONDENSI

E-mail:
vallencia.anggelica@student.pelitaindonesia.ac.id

ABSTRACT

Click-Through Rate (CTR) is a key metric in digital advertising used to evaluate campaign effectiveness and support bidding strategy decisions. Conventional bidding strategies are generally static and do not consider variations in click probability across ad impressions, leading to potential cost inefficiencies. This study aims to develop machine learning models to predict CTR probabilities and analyze their application in optimizing digital advertising bidding strategies. This research utilizes the publicly available Shopping Mall Paid Search Campaign Dataset, which is aggregate in nature. Several machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, are applied to model click probability based on pre-bidding features. Model performance is 1 evaluated using probabilistic metrics such as AUC ROC, Brier Score, and calibration curve analysis. Furthermore, predicted CTR values are employed in a threshold-based bidding simulation to assess advertising cost efficiency. The results indicate that the XGBoost model provides more stable and well-calibrated probability estimates compared to other models. The bidding simulation demonstrates that CTR prediction can significantly reduce advertising budget waste, although positive profitability has not yet been achieved. Therefore, this study serves as a proof-of-concept for applying machine learning to optimize digital advertising bidding strategies.

1. PENDAHULUAN

Perkembangan iklan digital mendorong kebutuhan akan pengelolaan anggaran yang lebih efisien dan berbasis data. Salah satu indikator utama keberhasilan kampanye adalah Click-Through Rate (CTR), yang menunjukkan rasio klik terhadap jumlah impresi.

Dalam praktiknya, strategi bidding masih banyak menggunakan pendekatan statis yang tidak mempertimbangkan probabilitas klik pada setiap impresi. Hal ini dapat menyebabkan pemborosan anggaran, terutama pada impresi dengan peluang klik rendah.

Machine learning memungkinkan prediksi probabilitas klik berdasarkan data historis, sehingga dapat digunakan sebagai dasar pengambilan keputusan bidding yang lebih selektif. Oleh karena itu, penelitian ini bertujuan untuk membangun model prediksi CTR dan menganalisis penggunaannya dalam optimasi strategi bidding.

2. TINJAUAN PUSTAKA

2.1. Click-Through Rate (CTR)

CTR adalah salah satu metrik digital marketing yang mengukur rasio total impresi terhadap klik pada mesin pencarian, iklan, email marketing, dan platform lainnya. Sederhananya, pada iklan online, CTR adalah persentase orang yang mengakses halaman website dengan mengklik iklan yang di internet. CTR dapat menjadi salah satu indikator keberhasilan pemasaran online yang dilakukan. (Qin, 2016).

2.1.1. Bidding Strategy

Strategi Bidding adalah fungsi/logika yang mengambil informasi dari kesempatan iklan (misalnya data pengguna, target iklan, KPI, volume lelang, anggaran, dan metrik performa) dan menentukan harga bid (Wang et al., 2017).

Strategi bidding menentukan seberapa besar penjual bersedia membayar untuk setiap klik iklan. Optimalisasi bidding membantu memaksimalkan hasil iklan dengan biaya minimal (M. Liu et al., 2022).

2.1.2. *Machine Learning*

Machine Learning (ML) adalah cabang kecerdasan buatan yang memungkinkan sistem belajar dari data untuk membuat prediksi atau keputusan (Mitchell, 1997). Dalam konteks periklanan digital, ML banyak digunakan untuk memprediksi CTR dan mengoptimalkan strategi bidding. Algoritma yang digunakan pada penelitain ini adalah Logistic Regression, Random Forest, dan XGBoost.

Logistic regression adalah metodologi regresi adaptif yang mencoba membangun prediktor sebagai kombinasi Boolean dari kovariat biner (Ruczinski et al., 2003). Fungsi linier tersebut kemudian dipetakan ke dalam rentang 0–1 menggunakan fungsi sigmoid (Z.-P. Liu, 2013).

Random Forest merupakan algoritma ensemble learning berbasis bagging yang membangun sejumlah besar pohon keputusan (decision trees) secara independen menggunakan subset data dan fitur yang dipilih secara acak. Setiap pohon menghasilkan prediksi sendiri, dan prediksi akhir ditentukan melalui mekanisme agregasi, seperti majority voting untuk klasifikasi atau rata-rata untuk probabilitas.

Sementara itu, XGBoost (Extreme Gradient Boosting) adalah algoritma ensemble learning berbasis gradient boosting yang menggabungkan banyak model pohon keputusan (decision trees) untuk meningkatkan akurasi prediksi. Setiap pohon baru dibangun untuk memperbaiki kesalahan prediksi pohon sebelumnya. XGBoost memulai dengan model awal (bias sederhana). Kemudian, model baru (pohon keputusan) dibuat berdasarkan residual error dari prediksi sebelumnya (Chen & Guestrin, 2016). Proses ini diulang secara iteratif hingga error keseluruhan minimum. Prediksi akhir dihasilkan dari penjumlahan kontribusi seluruh pohon.

2.1.3. *Evaluation Metric*

Metric evaluasi adalah suatu fungsi yang mengukur kualitas prediksi model terhadap data yang diketahui (ground truth). Dari hasil prediksi dan label sebenarnya, dihitung nilai numerik seperti "berapa banyak prediksi benar", "berapa banyak prediksi salah", jarak antara nilai prediksi dan nilai nyata, dsb (Rainio et al., 2024). Metrik yang digunakan adalah Logloss dan AUC-ROC.
Logloss (*Logarithmic Loss*) adalah sebuah fungsi loss yang digunakan untuk mengevaluasi kualitas prediksi probabilistik pada model klasifikasi. Untuk setiap instance, logloss mengambil nilai negatif dari log probabilitas yang diberikan oleh model terhadap kelas yang benar (Wu et al., 2023).

AUC-ROC (Area Under the Curve - Receiver Operating Characteristic). AUC diartikan sebagai probabilitas bahwa, jika kita pilih satu sampel positif acak dan satu sampel negatif acak, model akan memberikan skor (prediksi probabilitas) lebih tinggi pada sampel positif daripada pada sampel negatif (Bosnic, 2015). Sementara itu, ROC (Receiver Operating Characteristic) adalah kurva yang menggambarkan trade-off antara True Positive Rate (TPR / sensitivitas) dan False Positive Rate (FPR / 1 − spesifisitas) ketika kita mengubah nilai threshold untuk klasifikasi biner (Park et al., 2004).

Brier Score adalah proper scoring rule yang mengukur selisih kuadrat antara probabilitas prediksi dan hasil aktual, namun juga menyoroti keterbatasannya dalam mengevaluasi "clinical utility" suatu model prediksi (Assel et al., 2017).

3. METODOLOGI

Metode penelitian ini menggunakan pendekatan kuantitatif dengan eksperimen komputasional untuk memprediksi Click-Through Rate (CTR) iklan digital dan mensimulasikan strategi optimasi bidding. Data yang digunakan berasal dari dataset publik berjudul "Shopping Mall Paid Search Campaign Dataset" yang diperoleh melalui platform Kaggle.

Tahap awal penelitian adalah pra-pemrosesan data yang bertujuan untuk memastikan kualitas dan konsistensi data. Analisis data eksplorasi dilakukan secara statistik dan visual pada tahap ini, termasuk memeriksa distribusi data dan memvisualisasikan outlier menggunakan boxplot, untuk memastikan bahwa nilai-nilai yang digunakan masih berada dalam batas logika bisnis yang wajar. Dataset yang digunakan tidak memiliki nilai yang hilang secara signifikan, sehingga tidak perlu dilakukan imputasi lebih lanjut.

Dataset yang digunakan adalah data agregat, di mana setiap baris data mewakili ringkasan kinerja grup iklan (Ad Group) selama periode tertentu, seperti jumlah tayangan dan klik. Oleh karena itu, untuk memungkinkan pelatihan model prediksi CTR probabilistik yang valid pada tingkat tayangan, data ditransformasikan melalui proses ekspansi atau deagregasi data. Proses ekspansi atau deagregasi data ini melibatkan pemecahan data agregat menjadi titik data individual dengan label biner untuk peristiwa klik, di mana setiap baris mewakili satu peristiwa tayangan dengan label klik (1) atau tidak ada klik (0).

Fitur yang digunakan dalam model prediksi terbatas pada informasi yang sudah tersedia sebelum proses penawaran berlangsung. Variabel metrik pasca-klik seperti Pendapatan, Konversi, dan Biaya tidak disertakan sebagai fitur prediktor untuk mencegah kebocoran data. Fitur yang digunakan meliputi karakteristik kampanye, perangkat, dan jenis kata kunci yang relevan dengan proses penawaran.

Tahap selanjutnya adalah pemodelan, yang dilakukan menggunakan dua algoritma pembelajaran mesin, Regresi Logistik, Random Forest, dan XGBoost. Kedua model dilatih menggunakan data hasil agregasi untuk memprediksi probabilitas klik iklan. Proses pemodelan dilakukan menggunakan parameter standar sebagai dasar untuk membandingkan kinerja antar model. Evaluasi kinerja model dilakukan menggunakan beberapa metrik yang berfokus pada kualitas prediksi probabilitas. LogLoss digunakan sebagai metrik utama untuk menilai akurasi prediksi probabilitas model, sedangkan Brier Score digunakan untuk mengevaluasi tingkat kalibrasi probabilitas yang dihasilkan. Selain itu, Area Under the Receiver Operating Characteristic Curve (AUC ROC) digunakan untuk mengukur kemampuan model dalam memberi peringkat probabilitas klik. Evaluasi tambahan dilakukan melalui Precision Recall dan Kurva Kalibrasi untuk memberikan gambaran yang lebih komprehensif tentang stabilitas dan kualitas prediksi probabilitas. Model

dengan nilai LogLoss dan Brier Score yang lebih rendah serta nilai AUC yang stabil dipilih sebagai model terbaik..

4. HASIL DAN PEMBAHASAN

Hasil evaluasi kinerja model pada Tabel 4.1 menunjukkan bahwa ketiga model, yaitu Logistic Regression, Random Forest, dan XGBoost, memiliki performa yang relatif serupa. Nilai AUC-ROC yang berada pada kisaran 0.59 menunjukkan bahwa kemampuan model dalam membedakan antara kelas klik dan tidak klik masih tergolong terbatas, atau hanya sedikit lebih baik dibandingkan dengan tebakan acak. Hal ini mengindikasikan bahwa pola yang dapat dipelajari dari data belum cukup kuat untuk menghasilkan diskriminasi yang tinggi.

Tabel 2 Hasil Pemodelan

Model	AUC-ROC	LogLoss	Brier Score
Logistic Regression	0.5969	0.6719	0.2408
Random Forest	0.5978	0.6249	0.2182
XGBoost	0.5978	0.6250	0.2183

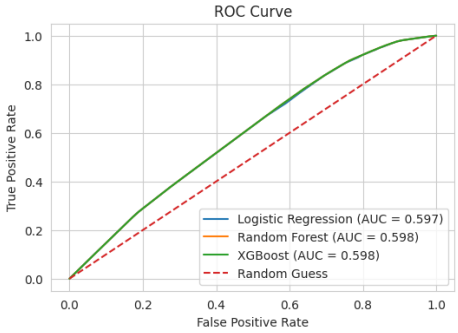
Random Forest dan XGBoost menghasilkan nilai AUC-ROC yang sama hingga empat angka desimal, yang menunjukkan bahwa kedua model memiliki kemampuan pemeringkatan (ranking ability) yang hampir identik. Sementara itu, Logistic Regression memiliki nilai AUC yang sedikit lebih rendah, namun selisihnya sangat kecil sehingga tidak menunjukkan perbedaan performa yang signifikan secara praktis.

Sebagai poin perbandingan tambahan, penggunaan Random Forest dalam penelitian ini bertujuan untuk mengevaluasi apakah pendekatan ensemble berbasis pohon keputusan dapat memberikan peningkatan performa yang berarti dibandingkan model linear. Namun, hasil yang diperoleh menunjukkan bahwa kinerja Random Forest tetap berada dalam kisaran yang sama dengan model lainnya. Hal ini mengindikasikan bahwa peningkatan kompleksitas model tidak secara otomatis menghasilkan peningkatan performa yang signifikan.

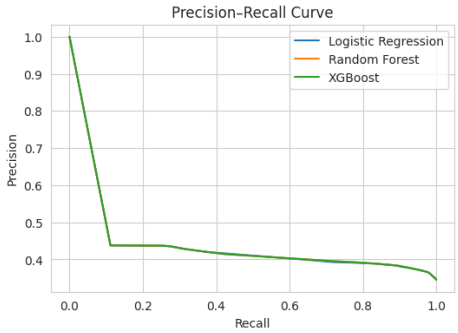
Temuan tersebut menunjukkan bahwa keterbatasan performa model kemungkinan besar lebih dipengaruhi oleh karakteristik data, seperti keterbatasan fitur pra-penawaran (pre-bid features) dan potensi ketidakseimbangan data, dibandingkan dengan pemilihan algoritma itu sendiri. Dengan kata lain, kualitas dan representativitas fitur menjadi faktor yang lebih dominan dalam menentukan kinerja model.

Dari perspektif kualitas probabilitas prediksi, Random Forest dan XGBoost menunjukkan nilai LogLoss dan Brier Score yang lebih rendah dibandingkan Logistic Regression. Hal ini menunjukkan bahwa kedua model tersebut mampu menghasilkan estimasi probabilitas yang lebih mendekati nilai aktual. XGBoost mencatat nilai LogLoss terendah, yang mengindikasikan keunggulan dalam meminimalkan kesalahan probabilistik, sedangkan Random Forest memiliki Brier Score terendah yang menunjukkan kalibrasi probabilitas yang sedikit lebih baik.

Meskipun demikian, selisih nilai antar model sangat kecil (kurang dari 0.001), sehingga tidak dapat disimpulkan adanya keunggulan yang dominan dari salah satu model. Kondisi ini memperkuat indikasi bahwa ketiga model menangkap pola yang serupa dari dataset yang digunakan.



Gambar 1. Kurva Perbandingan ROC



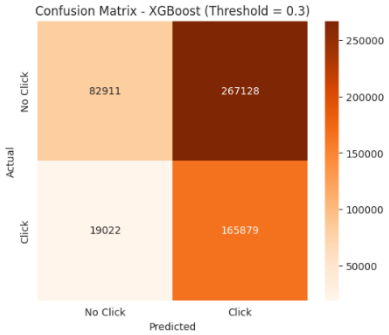
Gambar 2. Kurva Perbandingan Precision-Recall

Kurva ROC menunjukkan bahwa ketiga model memiliki kemampuan diskriminasi yang hampir sama. Namun, pada kasus data yang tidak seimbang seperti prediksi CTR, kurva Precision-Recall memberikan gambaran yang lebih informatif. Berdasarkan kurva Precision-Recall, Random Forest dan XGBoost menunjukkan kestabilan presisi yang relatif lebih baik pada berbagai tingkat recall dibandingkan Logistic Regression. Hal ini penting karena dalam konteks periklanan digital, kejadian klik merupakan peristiwa yang relatif jarang, sehingga model yang mampu menjaga presisi pada berbagai tingkat recall memiliki nilai praktis yang lebih tinggi.

Berdasarkan hasil evaluasi, XGBoost dipilih sebagai model terbaik dalam penelitian ini karena menghasilkan nilai LogLoss terendah, yang menunjukkan kemampuan terbaik dalam menghasilkan probabilitas prediksi yang akurat. Meskipun Random Forest memiliki nilai Brier Score yang sedikit lebih rendah, yang mengindikasikan kalibrasi probabilitas yang sedikit lebih baik, perbedaan tersebut sangat kecil dan tidak signifikan secara praktis. Selain itu, nilai AUC-ROC antara XGBoost dan Random Forest juga identik, sehingga kemampuan diskriminasi kedua model dapat dianggap setara. Dengan demikian, dalam konteks penelitian ini yang menekankan pada kualitas estimasi probabilitas untuk prediksi CTR, XGBoost dipandang memberikan keseimbangan performa yang paling optimal dibandingkan model lainnya.

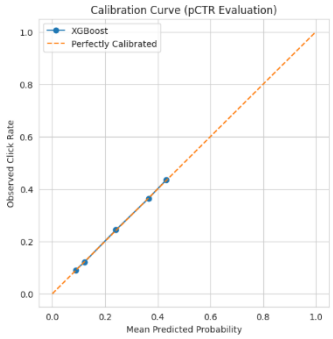
Sebagai analisis tambahan, evaluasi dilakukan menggunakan Matriks Konfusi pada model XGBoost, model berhasil mengurangi kesalahan positif palsu, yaitu kasus ketika model memprediksi klik padahal sebenarnya tidak terjadi klik. Dalam

konteks bisnis periklanan digital, kemampuan seperti itu sangat penting karena positif palsu dapat menyebabkan pemborosan anggaran iklan ketika strategi penawaran ditargetkan pada pengguna yang tidak mungkin mengklik.



Gambar 3. Confusion Matrix Heatmap XGBoost

Selain evaluasi berdasarkan kesalahan diskriminasi dan klasifikasi, analisis Kurva Kalibrasi yang menunjukkan bahwa sebagian besar titik prediksi terletak cukup dekat dengan garis diagonal (kalibrasi sempurna), yang menunjukkan bahwa probabilitas CTR yang diprediksi cukup konsisten dengan proporsi klik aktual. Temuan ini didukung oleh nilai Brier Score yang lebih rendah dibandingkan dengan Regresi Logistik, yang menunjukkan bahwa tingkat kesalahan probabilitas prediksi lebih kecil.



Gambar 4. Calibration Curve XGBoost

4.1. Analisis Pengaruh Fitur Terhadap Prediksi CTR

Hasil analisis menunjukkan bahwa Campaign Type merupakan fitur yang paling berpengaruh dalam prediksi CTR, diikuti oleh Keyword Type, sementara Device Type memiliki kontribusi yang relatif lebih kecil.

Analisis lanjutan terhadap rata-rata nilai prediksi probabilitas klik (predicted CTR / pCTR) pada masing-masing kategori memperkuat temuan tersebut. Pada tingkat kampanye, tipe 1:1 menghasilkan rata-rata pCTR tertinggi, diikuti oleh tipe Exact dan Phrase. Dari sisi perangkat, iklan yang ditayangkan pada Mobile menunjukkan rata-rata pCTR yang lebih tinggi dibandingkan Desktop. Sementara itu, pada level kata kunci, kategori yang mengandung unsur promosi langsung seperti coupon code, promo code, dan discount code mendominasi daftar keyword dengan nilai pCTR tertinggi. Hal ini mencerminkan kecenderungan pengguna untuk lebih responsif terhadap iklan yang menawarkan insentif eksplisit.

Secara keseluruhan, hasil analisis ini menunjukkan bahwa meskipun performa model secara keseluruhan masih bersifat proof-of-concept, pola-pola prediksi yang dihasilkan tetap konsisten dengan logika bisnis periklanan digital. Oleh karena itu, hasil interpretasi model ini dapat dimanfaatkan sebagai dasar pendukung dalam perumusan strategi bidding yang lebih terarah dan efisien.

4.2. Simulasi Strategi Bidding Prediksi CTR

Simulasi dilakukan menggunakan sepuluh skenario threshold, mulai dari 0.10 hingga 0.90. Pada setiap skenario, dihitung total jumlah bidding yang dilakukan (Total Bids), total biaya iklan yang dikeluarkan (Total Cost), total pendapatan yang dihasilkan (Total Revenue), Net Profit, serta Return on Investment (ROI). Perhitungan Cost dan Revenue didasarkan pada nilai aktual pada data uji, sehingga simulasi mencerminkan kondisi operasional yang realistis.

Tabel 3. Simulasi Bidding

Threshold (Confidence)	Total Bids	Total Cost ($)	Total Revenue ($)	Net Profit ($)	ROI
0.10	534940	127073.26	112391.55	-14681.71	-11.55
0.19	498574	124920.85	111485.75	-13435.1	-10.75
0.28	493180	124002.65	110804.72	-13197.93	-10.64
0.37	470733	120445.97	108059.81	-12386.16	-10.28
0.46	431387	112746.83	101459.06	-11287.77	-10.01
0.54	116903	33017.98	30421.72	-2596.26	-7.86
0.63	0	0.0	0.0	0.0	0.0
0.72	0	0.0	0.0	0.0	0.0
0.81	0	0.0	0.0	0.0	0.0
0.90	0	0.0	0.0	0.0	0.0

Berdasarkan hasil simulasi, terlihat adanya trade-off yang jelas antara agresivitas bidding dan efisiensi biaya. Pada threshold rendah (0.10–0.28), sistem cenderung melakukan bidding secara agresif dengan jumlah impresi yang sangat besar. Meskipun menghasilkan total pendapatan yang relatif tinggi, biaya iklan yang dikeluarkan juga meningkat secara signifikan sehingga Net Profit masih menunjukkan kerugian yang cukup besar.

Seiring dengan meningkatnya nilai threshold, jumlah bidding yang dilakukan menurun secara substansial, yang berdampak langsung pada penurunan total biaya iklan. Threshold 0.54

menunjukkan hasil yang paling efisien, dengan penurunan kerugian lebih dari 80% dibandingkan threshold terendah serta peningkatan nilai ROI menjadi –7.86%. Hasil ini menunjukkan bahwa model prediksi CTR mampu berfungsi sebagai filter awal yang efektif untuk mengurangi pemborosan anggaran iklan dengan menghindari impresi yang memiliki probabilitas klik rendah.

Namun, pada threshold yang terlalu tinggi (≥0.63), sistem menjadi sangat selektif sehingga tidak melakukan bidding sama sekali. Kondisi ini mengindikasikan terjadinya opportunity loss, di mana potensi pendapatan hilang karena iklan tidak ditayangkan. Secara keseluruhan, simulasi ini menunjukkan bahwa penerapan prediksi CTR dalam strategi bidding dapat meningkatkan efisiensi biaya iklan, meskipun keterbatasan performa model dan struktur data agregat masih menjadi faktor pembatas untuk mencapai profitabilitas penuh.

5. KESIMPULAN DAN SARAN

5.1. Kesimpulan

1. Pembangunan model machine learning untuk prediksi CTR.

Model machine learning untuk memprediksi CTR berhasil dibangun melalui serangkaian tahapan, yaitu preprocessing data, transformasi fitur menggunakan one-hot encoding, serta pelatihan model menggunakan algoritma Logistic Regression, Random Forest, dan XGBoost. Evaluasi dilakukan menggunakan metrik AUC-ROC, LogLoss, dan Brier Score untuk menilai kemampuan model dalam menghasilkan prediksi probabilitas. Hasil pengujian menunjukkan bahwa model mampu menghasilkan prediksi CTR, meskipun tingkat kemampuan diskriminasi masih tergolong sedang dengan nilai AUC sekitar 0.59. Hal ini menunjukkan bahwa model telah mampu menangkap pola dari data, namun masih terdapat keterbatasan dalam kualitas fitur yang digunakan.

2. Perbandingan kinerja algoritma

Hasil perbandingan menunjukkan bahwa ketiga model memiliki performa yang relatif serupa. Random Forest dan XGBoost menghasilkan nilai AUC yang sama, sedangkan Logistic Regression sedikit lebih rendah. Dari sisi kualitas probabilitas, XGBoost menghasilkan nilai LogLoss terendah, yang menunjukkan kemampuan terbaik dalam meminimalkan kesalahan prediksi probabilitas. Sementara itu, Random Forest memiliki nilai Brier Score terendah yang menunjukkan kalibrasi probabilitas yang sedikit lebih baik. Namun, perbedaan antar model sangat kecil sehingga tidak menunjukkan keunggulan yang signifikan secara praktis. Dengan mempertimbangkan keseluruhan metrik, XGBoost dipilih sebagai model terbaik karena memberikan keseimbangan performa yang paling optimal.

3. Pemanfaatan prediksi CTR untuk optimasi strategi bidding

Hasil prediksi CTR terbukti dapat dimanfaatkan sebagai dasar dalam mendukung optimasi strategi bidding iklan digital melalui simulasi berbasis threshold probabilitas. Penerapan threshold memungkinkan sistem untuk menyaring impresi dengan probabilitas klik rendah, sehingga dapat mengurangi pemborosan anggaran iklan. Hasil simulasi menunjukkan bahwa penggunaan threshold tertentu mampu menurunkan kerugian secara signifikan serta meningkatkan efisiensi biaya, meskipun pada threshold yang terlalu tinggi dapat menyebabkan hilangnya peluang (opportunity loss).

5.2. Saran

1. Penggunaan metode penanganan ketidakseimbangan data yang lebih komprehensif.

Penelitian di masa mendatang disarankan untuk mengeksplorasi teknik resampling yang lebih canggih seperti SMOTE (Synthetic Minority Over-sampling Technique) untuk menghasilkan sampel sintetis pada kelas minoritas. Metode ini berpotensi meningkatkan kinerja model, khususnya dalam meningkatkan presisi tanpa mengorbankan kemampuan model dalam mendeteksi kelas positif.

2. Integrasi metode interpretabilitas model.

Penelitian selanjutnya dapat mengintegrasikan metode interpretabilitas seperti SHAP (SHapley Additive exPlanations) untuk menganalisis kontribusi masing-masing fitur terhadap prediksi yang dihasilkan model. Pendekatan ini dapat memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang memengaruhi probabilitas klik iklan serta meningkatkan transparansi dan kepercayaan terhadap hasil prediksi model.

3. Eksplorasi arsitektur Deep Learning untuk prediksi CTR.

Eksperimen dapat diperluas dengan menerapkan arsitektur Deep Learning yang secara khusus dirancang untuk prediksi CTR, seperti DeepFM atau Wide & Deep Learning. Model-model tersebut memiliki kemampuan yang lebih baik dalam menangkap interaksi fitur yang kompleks dibandingkan model berbasis pohon yang digunakan dalam penelitian ini.

4. Penggunaan data mentah dengan atribut yang lebih rinci.

Selain pengembangan model, penggunaan data mentah pada tingkat impresi yang mencakup atribut pengguna yang lebih detail, seperti karakteristik demografis atau riwayat penelusuran, juga sangat disarankan. Ketersediaan data yang lebih granular diharapkan dapat meningkatkan kualitas fitur yang digunakan dalam pemodelan serta berpotensi meningkatkan akurasi prediksi secara lebih signifikan.Berdasarkan temuan dan keterbatasan tersebut, pengembangan penelitian selanjutnya sangat disarankan untuk berfokus pada aspek teknis penanganan data dan eksplorasi model lanjutan. Terkait ketidakseimbangan data, penelitian masa depan sebaiknya mengeksplorasi teknik resampling yang lebih canggih untuk mensintesis data baru pada kelas minoritas. Selain itu, untuk meningkatkan transparansi dan kepercayaan terhadap keputusan model, penerapan analisis SHAP (SHapley Additive exPlanations) sangat direkomendasikan guna mengidentifikasi kontribusi spesifik setiap fitur terhadap prediksi klik.

DAFTAR PUSTAKA

Assel, M., Sjoberg, D. D., & Vickers, A. J. (2017). The

Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic Research*, *1*(1), 1–7. https://doi.org/10.1186/s41512-017-0020-3

Bosnic, Z. (2015). *ROC analysis of classifiers in machine learning : A survey ROC Analysis of Classifiers in Machine Learning : A Survey Technical report MM-1 / 2011*. *July*. https://doi.org/10.3233/IDA-130592

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Liu, M., Hu, Z., Lai, Z., Zheng, D., & Nie, X. (2022). *Real-time Bidding Strategy in Display Advertising: An Empirical Analysis*. http://arxiv.org/abs/2212.02222

Liu, Z.-P. (2013). Logistic Regression. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1142–1143). Springer New York. https://doi.org/10.1007/978-1-4419-9863-7_396

Park, S. H., Goo, J. M., & Jo, C. H. (2004). Receiver operating characteristic (ROC) curve: Practical review for radiologists. *Korean Journal of Radiology*, *5*(1), 11–18. https://doi.org/10.3348/kjr.2004.5.1.11

Qin, C. (2016). Empirical Research on How Keyword Advertising Linked Item Characteristics Affect CTR. *International Journal of Hybrid Information Technology*, *9*(5), 419–430. https://doi.org/10.14257/ijhit.2016.9.5.36

Rainio, O., Teuho, J., & Klen, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 1–14. https://doi.org/10.1038/s41598-024-56706-x

Ruczinski, I., Kooperberg, C., & LeBlanc, M. (2003). *Logic Regression — Methods and Software*. *1024*, 333–343. https://doi.org/10.1007/978-0-387-21579-2_21

Wang, J., Zhang, W., & Yuan, S. (2017). Display advertising with real-time bidding (RTB) and behavioural targeting. *Foundations and Trends in Information Retrieval*, *11*(4–5), 297–435. https://doi.org/10.1561/1500000049

Wu, C., Heidari, M., Grama, A., & Szpankowski, W. (2023). *Regret Bounds for Log-loss via Bayesian Algorithms*.