

Implementasi Metode Klasifikasi LightGBM dan Analisis Survival dalam Memprediksi Pelanggan Churn

Ibnu Zahy' Atha Illah^{1*}, Wahyu Syaifullah Jauharis Sapu², Aviolla Terza Damaliana³
^{1,2,3} Program Studi Sains Data, Universitas Pembangunan Nasional "Veteran" Jawa Timur
*email: 20083010016@student.upnjatim.ac.id

DOI: <https://doi.org/10.31603/komtika.v8i1.11194>

Received: 23-03-2024, Revised: 31-05-2024, Accepted: 01-06-2024

ABSTRACT

Increasingly tight competition in the business world causes every business sector to try to utilize relevant technology to maintain its market share. The success of a company is often measured by how strong the customer network they have. Loss of customers (customer churn) can cause a significant decrease in revenue and can even threaten the existence of the company itself. Therefore, predictive modelling and projection of customer churn is needed as a customer retention effort. This research involves the LightGBM classification algorithm for customer churn prediction and utilizes survival analysis for future projections. The results of the research can be used to prevent customer churn at companies, especially PT Kasir Pintar Internasional. LightGBM classification performance as measured by model evaluation reaches Accuracy, Precision, Recall, and F1-score values of 0.964, 0.971, 0.990, and 0.980 respectively. The LightGBM classification model also provides information on five important features that influence customer churn. Companies can use these five important features as material for designing customer retention strategies. Apart from that, the Cox Proportional Hazard survival model has a C-index evaluation value of 0.83, which means it is quite capable of projecting customer survival. The survival model also shows that currently non-churn customers have an average survival expectation of 15 months.

Keywords: Customer Churn, Lightgbm Classification, Cox Proportional Hazard Survival.

ABSTRAK

Persaingan dunia usaha yang semakin ketat menyebabkan setiap sektor usaha berupaya memanfaatkan teknologi relevan guna mempertahankan pangsa pasarnya. Keberhasilan suatu perusahaan ini seringkali diukur dari seberapa kuat jaringan pelanggan yang mereka miliki. Hilangnya pelanggan (*customer churn*) dapat menyebabkan penurunan pendapatan yang signifikan dan bahkan bisa mengancam eksistensi perusahaan itu sendiri. Oleh karena itu, diperlukan pemodelan prediksi dan proyeksi pelanggan *churn* sebagai upaya retensi pelanggan. Penelitian ini melibatkan algoritma klasifikasi LightGBM untuk prediksi pelanggan *churn* dan memanfaatkan analisis ketahanan hidup (*survival*) untuk proyeksi masa depan. Hasil dari penelitian dapat dimanfaatkan untuk mencegah terjadinya pelanggan *churn* pada perusahaan, utamanya PT Kasir Pintar Internasional. Performa klasifikasi LightGBM yang diukur dengan evaluasi model mencapai nilai *Accuracy*, *Precision*, *Recall*, dan *F1-score* yang masing-masing sebesar 0.964, 0.971, 0.990, dan 0.980. Model klasifikasi LightGBM juga memberikan informasi lima fitur penting yang mempengaruhi *churn* pelanggan. Kelima fitur penting tersebut dapat dimanfaatkan perusahaan sebagai bahan merancang strategi retensi pelanggan. Selain itu, model *survival Cox Proportional Hazard* memiliki nilai evaluasi C-index sebesar 0.83 yang berarti sudah cukup mumpuni untuk memproyeksikan ketahanan hidup pelanggan. Model *survival* juga menunjukkan bahwa pelanggan yang saat ini *non-churn* memiliki rata-rata ekspektasi *survival* selama 15 bulan.

Keywords: Pelanggan Churn, Klasifikasi LightGBM, Survival Cox Proportional Hazard.

PENDAHULUAN

Pada era inovasi teknologi yang pesat saat ini, terdapat dinamika yang signifikan dalam dunia usaha dan bisnis. Kemajuan teknologi digital yang serba otomatisasi sangat menunjang

keaktivitas manusia [1]. Kemajuan teknologi telah membuka pintu bagi munculnya berbagai ide usaha dan konsep bisnis yang revolusioner. Fenomena ini tidak terlepas dari transformasi digital yang memengaruhi hampir semua sektor industri. Berdasarkan semakin ketatnya persaingan yang ada, setiap sektor usaha berupaya untuk memanfaatkan teknologi relevan guna mempertahankan pangsa pasarnya. Indonesia memiliki 2.492 perusahaan rintisan dan menempati peringkat ke-6 terbesar dunia per tanggal 10 Mei 2023. Fakta ini juga menunjukkan bahwa Indonesia memimpin Asia Tenggara dengan jumlah perusahaan rintisan terbanyak, mengungguli Singapura, Filipina, dan Malaysia. Pada tingkat global, dominasi masih dipegang oleh Amerika Serikat, diikuti oleh India, Inggris, Kanada, dan Australia [2]. Hal ini semakin menegaskan bahwa persaingan perusahaan di Indonesia semakin kompetitif yang sekaligus perebutan pelanggannya pun juga semakin ketat. Dalam konteks ini, persaingan yang ketat memaksa perusahaan-perusahaan di Indonesia untuk terus meningkatkan kualitas produk dan layanan mereka guna memikat dan mempertahankan pelanggan setia.

Tak dapat dipungkiri, peran pelanggan dalam kelangsungan hidup sebuah perusahaan adalah tak tergantikan [3]. Jumlah pelanggan yang stabil dan bertumbuh secara sehat adalah fondasi dari pendapatan perusahaan. Dalam ekosistem bisnis yang kompetitif, keberhasilan suatu perusahaan seringkali diukur dari seberapa kuat jaringan pelanggan yang mereka miliki. Hilangnya pelanggan dapat menyebabkan penurunan pendapatan yang signifikan dan bahkan bisa mengancam eksistensi perusahaan itu sendiri [4]. Peristiwa beralihnya pelanggan dari suatu produk atau jasa ke penyedia lain biasa dikenal dengan istilah pelanggan *churn* [5]. Oleh karena itu, perlu untuk memprediksi serta memproyeksikan pelanggan *churn* sebagai langkah preventif sekaligus upaya retensi pelanggan.

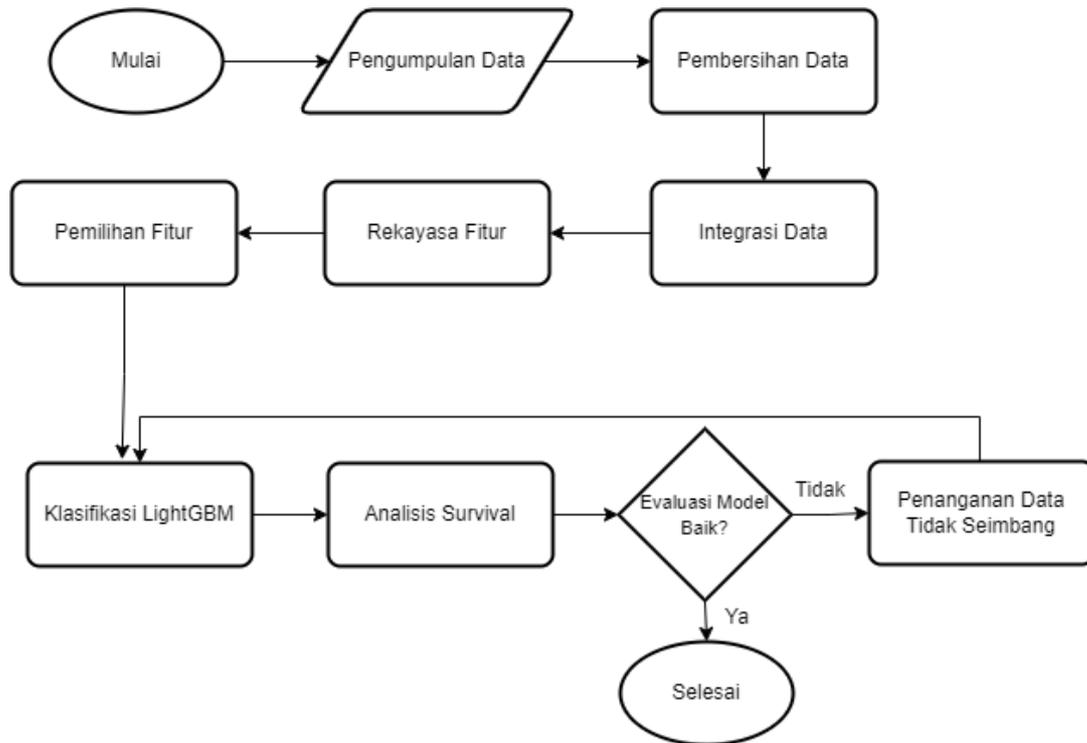
Terdapat beberapa penelitian mengenai pelanggan *churn* yang telah diulas sebelumnya, Swetha dkk yang menggunakan Fine-tuned XGBoost untuk memprediksi *churn* pada industri telekomunikasi. Fine-tuned XGBoost berhasil menangani persebaran dan *Imbalance Dataset* yang menghasilkan akurasi sebesar 99%. Penelitian lainnya dilakukan oleh Pamina dkk yang menggunakan pendekatan KNN, Random Forest, dan XGBoost untuk memprediksi *churn*. Hasil penelitian menunjukkan bahwa algoritma XGBoost memiliki kinerja yang lebih unggul dibandingkan dua algoritma lainnya dengan akurasi sebesar 79,80% [7]. Penelitian berikutnya juga dilakukan oleh Kohei Arai dkk yang menggunakan algoritma LightGBM. Penelitian tersebut menghasilkan nilai AUC sebesar 0.837 dan \log_loss sebesar 0.496 [8].

Berdasarkan pemanfaatan metode prediksi *churn* yang telah dibangun oleh peneliti sebelumnya, penelitian ini hendak melakukan hal yang serupa. Namun, tidak hanya terbatas pada pemanfaatan metode klasifikasi saja. Penelitian ini juga melibatkan analisis *survival* untuk memproyeksikan ketahanan pelanggan sehingga dapat memberikan *insight* yang lebih mendalam. Penelitian yang dilakukan juga menggunakan rekayasa fitur yang optimal sehingga dapat menghasilkan nilai evaluasi model yang tinggi. Penelitian ini dilakukan untuk dapat mencegah terjadinya *churn* pada perusahaan, utamanya PT Kasir Pintar Internasional. Dengan adanya model klasifikasi dan analisis ketahanan hidup (*survival*), perusahaan dapat melakukan prediksi dan mengatur strategi untuk mempertahankan pelanggannya.

METODE

Metode yang digunakan dalam penelitian “Implementasi Metode Klasifikasi LightGBM dan Analisis *Survival* dalam Memprediksi Pelanggan *Churn*” disajikan dalam bentuk diagram

seperti yang disajikan pada Gambar 1. Alur penelitian ini dimulai dari pengumpulan data, pembersihan data, integrasi data, rekayasa fitur, pemilihan fitur, hingga pemodelan klasifikasi LightGBM dan analisis *survival*. Terdapat tahap penanganan data tidak seimbang yang akan dilakukan jika dibutuhkan. Proses tersebut dilakukan ketika hasil evaluasi model dirasa kurang mumpuni.



Gambar 1. Diagram Alur Penelitian Pelanggan *Churn*

Berikut merupakan penjelasan dari masing-masing diagram alur penelitian pada Gambar 1. Diagram alur ini akan sangat berfungsi untuk pemetaan langkah-langkah yang harus dilakukan selama penelitian.

Pengumpulan Data

Pada konteks prediksi pelanggan *churn* di PT Kasir Pintar Internasional, tahap pengumpulan data melibatkan akuisisi informasi terkait perilaku dan kebiasaan pelanggan. Data ini mencakup tabel laporan, tabel biling, tabel toko, tabel *nps_user*, tabel *promo_user*, dan tabel *promo*. Pemilihan tabel tersebut didasarkan pada intuisi dan wawasan data yang penting untuk mendukung pemikiran cepat dan pemahaman yang mendalam dalam mengambil keputusan [9]. Pada Tabel 1 berikut disajikan penjelasan dari masing-masing fitur yang akan digunakan dalam penelitian.

Tabel 1. Fitur Pada Penelitian

No	Fitur	Tipe Data	Keterangan
1	id_pelanggan	Integer	Kode id pelanggan
2	tanggal	String	Tanggal transaksi
3	jumlah_transaksi	Integer	Jumlah transaksi
4	gmv	Integer	Pendapatan transaksi
5	waktu_akhir	String	Waktu terakhir berlangganan
6	harga	Integer	Harga berlangganan
7	jenis	Integer	Jenis transaksi
8	status	Integer	Label pelanggan
9	kota	Integer	Kode id kota pelanggan
10	bidang	Integer	Kode id bidang pelanggan
11	skor_rekomendasi	Integer	Skor kepuasan pelanggan
12	tahun	Integer	Tahun pemberian skor rekomendasi
13	id_promo	Integer	Kode id promo yang diikuti
14	created_at	Integer	Waktu beli promo
15	name	String	Nama promo
16	coin	Integer	Jumlah koin perolehan dari promo
17	price	Integer	Harga promo
18	start_date	String	Tanggal dimulai promo
19	end_date	String	Tanggal terakhir promo

Pembersihan Data

Pada tahap pembersihan data dilakukan proses identifikasi dan perbaikan masalah dalam data. Hal ini meliputi penanganan nilai yang hilang, penanganan nilai yang tidak akurat, dan pembersihan data dari duplikasi. Data yang mengalami masalah tersebut akan dihapus dan tidak digunakan dalam penelitian. Tujuan dari tahap ini adalah memastikan bahwa data yang digunakan untuk analisis adalah berkualitas tinggi dan dapat diandalkan [10].

Integrasi Data

Pada tahap integrasi data melibatkan penggabungan data dari berbagai sumber atau tabel ke dalam tempat penyimpanan data tunggal [11]. Tujuannya adalah untuk menciptakan tampilan yang holistik dan terpadu dari informasi yang terkandung di dalam data tersebut. Pada penelitian ini, data dari berbagai tabel tersebut dijadikan satu *data frame* dalam proses integrasi data.

Rekayasa Fitur

Pada tahap rekayasa fitur, proses yang dilakukan yaitu pembuatan atau modifikasi variabel (fitur) dari data mentah untuk meningkatkan kinerja model. Hal ini dapat mencakup transformasi variabel, pembuatan variabel baru, atau penggabungan variabel yang ada untuk memperoleh informasi tambahan yang berguna [12]. Setelah melewati proses rekayasa fitur, penelitian ini mencakup beragam variabel, seperti *id_user*, *gmv*, *jumlah_transaksi*, *total_promo*, *total_price_promo*, *avg_skor_rekomendasi*, *total_harga_langganan*,

waktu_akhir_langganan, gmv_10, jumlah_transaksi_10, gmv_20, jumlah_transaksi_20, gmv_30, jumlah_transaksi_30, skor_rekomendasi_min, skor_rekomendasi_max, tenure_month, kota, bidang, dan status_churn.

Seleksi Fitur

Seleksi fitur adalah proses memilih subset dari variabel atau fitur yang paling relevan dan berpengaruh dalam memprediksi atau menjelaskan fenomena yang sedang diamati. Tujuannya adalah untuk mengurangi kompleksitas model, meningkatkan interpretabilitas, dan menghindari *overfitting* [13]. Hasil seleksi fitur yang digunakan dalam penelitian ini yaitu gmv, jumlah_transaksi, total_promo, avg_skor_rekomendasi, total_harga_langganan, gmv_10, jumlah_transaksi_10, gmv_20, jumlah_transaksi_20, gmv_30, jumlah_transaksi_30, tenure_month, dan kota. Terdapat tambahan variabel penting lain yang harus dimasukkan dalam penelitian yaitu id_user dan status_churn.

Penanganan Data Tidak Seimbang

Pada konteks prediksi pelanggan *churn* di PT Kasir Pintar Internasional, terjadi ketidakseimbangan antara jumlah pelanggan *churn* dan *non-churn*. Hal ini dapat mempengaruhi kinerja model karena model cenderung lebih condong ke kelas mayoritas [14]. Namun, pada penelitian ini model klasifikasi LightGBM cukup mumpuni untuk memodelkan data tanpa penanganan khusus pada data tidak seimbang. Oleh karena itu, penulis tidak mengimplementasikan tahap penanganan data tidak seimbang pada data pelanggan di PT Kasir Pintar Internasional.

Klasifikasi LightGBM

Metode klasifikasi LightGBM digunakan untuk membangun model prediksi pelanggan *churn*. LightGBM adalah algoritma Machine Learning yang efisien dan mampu mengatasi masalah klasifikasi dengan komputasi yang minimal [15]. Dengan memanfaatkan teknik *boosting*, LightGBM dapat memprediksi apakah seorang pelanggan cenderung melakukan *churn* berdasarkan fitur-fitur yang telah dipilih.

Analisis Survival

Analisis ketahanan hidup (*survival*) adalah metode yang digunakan untuk menggambarkan informasi dari data mengenai ketahanan hidup seperti kematian, kesembuhan, kekambuhan, ataupun kerusakan alat [16]. Pada konteks prediksi pelanggan *churn*, analisis ketahanan hidup (*survival*) digunakan untuk memahami faktor-faktor yang mempengaruhi waktu sampai pelanggan melakukan *churn*. Hal ini memungkinkan PT Kasir Pintar Internasional untuk mengidentifikasi pelanggan yang berisiko tinggi untuk *churn* dan mengambil tindakan pencegahan atau strategi retensi yang sesuai. Metode yang digunakan dalam penelitian ini yaitu Regresi *Cox Proportional Hazard* yang terkenal dan sering digunakan dalam penelitian analisis *survival*. Metode ini memungkinkan penanganan data tanpa mengharuskan adanya asumsi tertentu mengenai distribusi [17].

Evaluasi Model

Evaluasi merupakan proses yang dilakukan untuk mengukur sejauh mana model yang diterapkan mampu memberikan kinerja yang optimal dan memuaskan dalam konteks analisis atau prediksi yang dilakukan [18]. Setelah model prediksi pelanggan *churn* dibangun, perlu dilakukan evaluasi untuk menentukan seberapa baik performa model tersebut. Hal ini melibatkan penggunaan metrik-metrik evaluasi seperti akurasi, presisi, *recall*, *f1-score*, nilai AUC, dan nilai *binary log loss* untuk menilai sejauh mana model mampu memprediksi pelanggan *churn* dengan akurat. Sedangkan, metode evaluasi yang digunakan untuk analisis ketahanan hidup (*survival*) yaitu nilai C-index. Berikut merupakan beberapa rumus yang digunakan dalam evaluasi model [19].

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$precision = \frac{TP}{(TP + FP)} \quad (2)$$

$$recall = \frac{TP}{(TP + FN)} \quad (3)$$

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

HASIL DAN PEMBAHASAN

Hasil Klasifikasi LightGBM

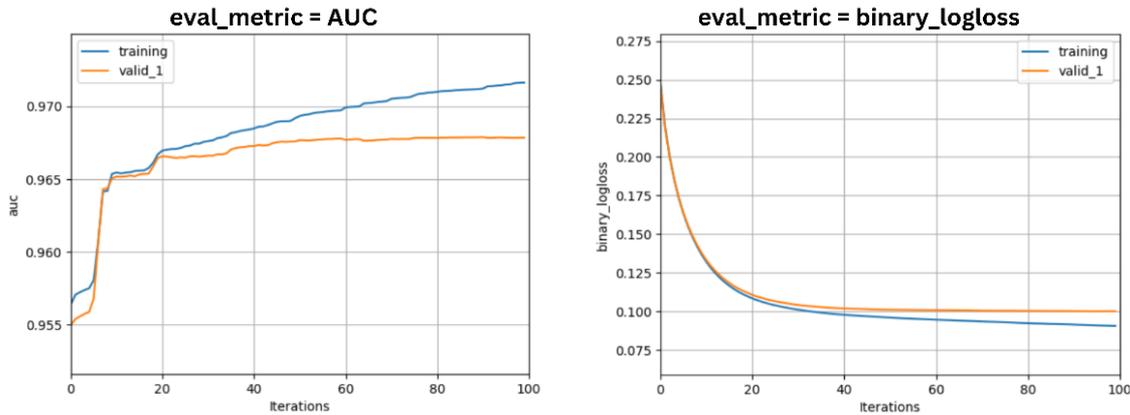
Model Klasifikasi LightGBM dibuat tanpa perubahan parameter dengan *learning rate* sebesar 0.1 dan melatih model selama 100 iterasi. Fitur yang digunakan sebagai variabel prediktor yaitu 'gmv', 'jumlah_transaksi', 'total_promo', 'avg_skor_rekomendasi', 'total_harga_langganan', 'gmv_10', 'jumlah_transaksi_10', 'gmv_20', 'jumlah_transaksi_20', 'gmv_30', 'jumlah_transaksi_30', 'tenure_month', dan 'kota'. Sedangkan, fitur yang digunakan sebagai variabel target adalah 'status_churn' dengan 185.367 untuk kelas *churn* dan 19.158 untuk kelas *non-churn*. Dataset yang digunakan dibagi menjadi tiga bagian, yaitu data latih (70%), data validasi (15%), dan data uji (15%). Pembagian data ini dilakukan untuk memastikan keberagaman data selama proses pengembangan dan pengujian model [20]. Pada Tabel 2 ditampilkan hasil evaluasi model pada nilai Akurasi, Presisi, *Recall*, dan *F1-score*.

Tabel 2. Hasil Evaluasi Model Klasifikasi LightGBM

Evaluasi Model			
Akurasi	Presisi	Recall	F1-Score
Train: 0.969			
Val: 0.966	Val: 0.973	Val: 0.990	Val: 0.981
Test: 0.964	Test: 0.971	Test: 0.990	Test: 0.980

Hasil evaluasi pada nilai Akurasi, Presisi, *Recall*, dan *F1-score* menunjukkan bahwa model cukup mumpuni untuk mengklasifikasikan antara pelanggan *churn* dan *non-churn*. Seluruh nilai tersebut telah berada di angka yang tinggi baik pada saat proses pelatihan, validasi, maupun pengujian. Namun, untuk lebih memastikan kualitas model diperlukan juga memeriksa

grafik selama proses pelatihan dan validasi guna mendeteksi kemungkinan *overfitting* seperti disajikan pada Gambar 2. Terlihat bahwa nilai akhir AUC pada data validasi sebesar 0.95. Sedangkan, nilai akhir *binary log loss* pada data validasi sebesar 0.10.



Gambar 2. Grafik Selama Proses Pelatihan dan Validasi Model

Pada Gambar 2 terlihat grafik pelatihan dan validasi model klasifikasi LightGBM dengan evaluasi metrik yaitu AUC dan *binary_logloss*. Grafik tersebut menunjukkan bahwa model sudah cukup baik dan terhindar dari *overfitting* selama proses pelatihan dan validasi. Penanda utama bahwa model telah terbebas dari *overfitting* yaitu grafik tersebut mengikuti pola yang serupa dan tidak terlalu jauh terpisah satu sama lain. Hal ini menunjukkan bahwa model tidak hanya mengingat data pelatihan tetapi juga mampu menggeneralisasi pola ke data yang belum pernah dilihat sebelumnya. Sampai tahap ini sudah bisa terlihat bahwa algoritma LightGBM cukup baik untuk memodelkan data dengan proporsi yang tidak seimbang. Di samping itu, hasil Confusion Matrix pada saat validasi juga disajikan pada Gambar 3.

```

Validation Confusion Matrix:
[[ 2063  767]
 [  277 27572]]
Validation Classification Report:

```

	precision	recall	f1-score	support
0	0.88	0.73	0.80	2830
1	0.97	0.99	0.98	27849
accuracy			0.97	30679
macro avg	0.93	0.86	0.89	30679
weighted avg	0.96	0.97	0.96	30679

Gambar 3. Confusion Matrix Validasi

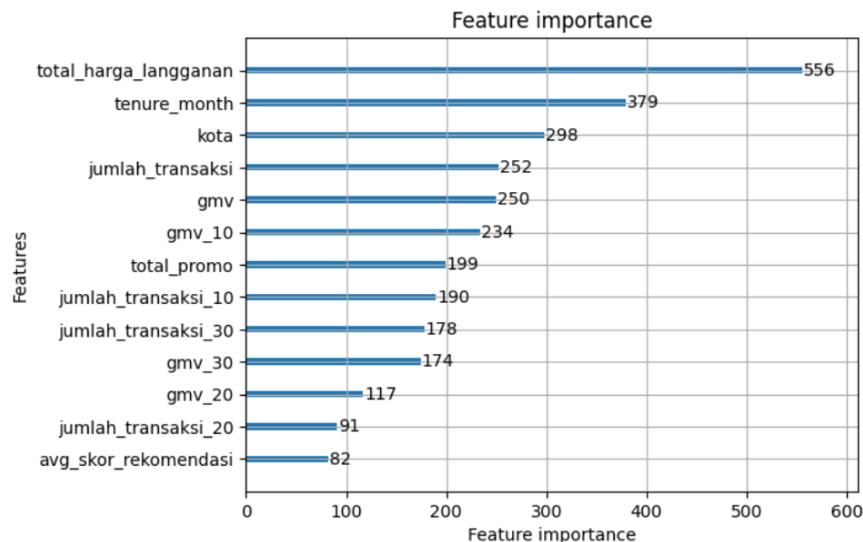
Pada Gambar 3 terlihat nilai TP (*True positive*) sebanyak 2063, TN (*True negative*) sebanyak 27572, FP (*False Positive*) sebanyak 767, dan FN (*False Negative*) sebanyak 277. Nilai-nilai tersebut mengindikasikan bahwa model telah baik karena mampu menghasilkan sejumlah besar prediksi yang benar (TP dan TN tinggi) sambil meminimalkan kesalahan (FP dan FN rendah). Selanjutnya, pada Gambar 4 ditampilkan hasil Confusion Matrix saat proses pengujian. Nilai Confusion Matrix pada saat validasi dan pengujian tidak memiliki perbedaan

yang cukup jauh. Hal ini menunjukkan bahwa model telah cukup konsisten dan optimal untuk memprediksi data diluar pelatihan.

```
Test Confusion Matrix:  
[[ 2094  830]  
 [ 269 27486]]  
Test Classification Report:  
              precision    recall  f1-score   support  
  
    0           0.89       0.72       0.79       2924  
    1           0.97       0.99       0.98       27755  
  
 accuracy              0.96       30679  
 macro avg           0.93       0.85       0.89       30679  
 weighted avg       0.96       0.96       0.96       30679
```

Gambar 4. Confusion Matrix Pengujian

Pada Gambar 4 terlihat nilai TP (*True positive*) sebanyak 2094, TN (*True negative*) sebanyak 27486, FP (*False Positive*) sebanyak 830, dan FN (*False Negative*) sebanyak 269. Nilai ini berguna untuk menentukan hasil evaluasi model yang telah disajikan pada Tabel 2 sebelumnya. Selanjutnya disajikan hasil fitur penting pada model klasifikasi seperti pada Gambar 5. Terlihat bahwa fitur ‘total_harga_langganan’, ‘tenure_month’, ‘kota’, ‘jumlah_transaksi’, dan ‘gmv’ merupakan lima fitur penting tertinggi yang mempengaruhi status *churn* pelanggan. Fitur tersebut dapat dianalisis lebih lanjut dan dapat digunakan sebagai bahan retensi pelanggan. Perusahaan dapat melihat lebih jauh terkait perilaku pelanggan terhadap kelima fitur tersebut. Hal ini dilakukan sebagai upaya merancang strategi sehingga dapat melakukan retensi pelanggan.



Gambar 5. Fitur Penting Model Klasifikasi LightGBM

Hasil Analisis Survival

Hasil model analisis ketahanan hidup (*survival*) memiliki nilai *Concordance* yang mengukur seberapa baik model dapat memprediksi peristiwa dalam data. Nilai *Concordance* berkisar antara 0 dan 1, di mana nilai yang lebih tinggi menunjukkan kinerja model yang lebih

baik. Pada Gambar 6 ditunjukkan nilai *Concordance* (C-index) sebesar 0.83 yang menandakan bahwa model memiliki kemampuan yang baik untuk memprediksi urutan relatif dari waktu kejadian yang diamati.

Concordance	0.83
Partial AIC	4158276.64
log-likelihood ratio test	84363.18 on 12 df
-log2(p) of ll-ratio test	inf

Gambar 6. Evaluasi Model Survival Cox Proportional Hazard

Selanjutnya, akan ditampilkan hasil prediksi peluang pelanggan yang *non-churn* untuk tetap menggunakan layanan perusahaan. Fokus proyeksi ini terletak pada pelanggan *non-churn* karena metode analisis ketahanan hidup (*survival*) terbukti efektif dalam menganalisis lamanya masa berlangganan (bertahan hidup). Tujuan penelitian ini adalah untuk memperkirakan berapa lama pelanggan yang *non-churn* dapat tetap menggunakan layanan perusahaan berdasarkan data yang ada. Gambar 7 berikut memperlihatkan hasil probabilitas pelanggan *non-churn* untuk bertahan di perusahaan dengan baris pertama menunjukkan id pelanggan dan kolom pertama menunjukkan jumlah bulan.

	1035000139	1035000168	1035000181	1035000494	1035000672	1035000701	1035000793	1035000949	1035001224	1035001285	...	1036999664
1.0	0.995113	0.995548	0.420728	1.000000	0.999918	0.999039	0.980641	1.000000	0.385364	0.999999	...	0.639843
2.0	0.993391	0.993980	0.309835	1.000000	0.999889	0.998700	0.973890	1.000000	0.275123	0.999999	...	0.546442
3.0	0.992146	0.992844	0.248218	1.000000	0.999868	0.998454	0.969026	1.000000	0.215508	0.999999	...	0.487390
4.0	0.991124	0.991914	0.206918	1.000000	0.999851	0.998252	0.965052	1.000000	0.176366	0.999998	...	0.443725
5.0	0.990171	0.991045	0.174569	1.000000	0.999835	0.998064	0.961354	1.000000	0.146249	0.999998	...	0.406475
6.0	0.989258	0.990212	0.148291	1.000000	0.999819	0.997883	0.957820	1.000000	0.122196	0.999998	...	0.373673
7.0	0.988262	0.989304	0.124111	1.000000	0.999802	0.997686	0.953978	1.000000	0.100441	0.999998	...	0.340894
8.0	0.987324	0.988449	0.104937	1.000000	0.999787	0.997500	0.950370	1.000000	0.083492	0.999998	...	0.312631
9.0	0.986414	0.987619	0.089161	1.000000	0.999771	0.997320	0.946880	1.000000	0.069777	0.999997	...	0.287434
10.0	0.985553	0.986834	0.076412	1.000000	0.999757	0.997149	0.943586	1.000000	0.058871	0.999997	...	0.265446
11.0	0.984567	0.985935	0.064022	1.000000	0.999740	0.996953	0.939825	1.000000	0.048449	0.999997	...	0.242298
12.0	0.983533	0.984992	0.053177	1.000000	0.999722	0.996747	0.935894	1.000000	0.039491	0.999997	...	0.220178
13.0	0.982364	0.983926	0.043098	1.000000	0.999702	0.996515	0.931464	1.000000	0.031331	0.999997	...	0.197562
14.0	0.981302	0.982957	0.035596	1.000000	0.999684	0.996303	0.927450	1.000000	0.025380	0.999996	...	0.179005
15.0	0.979487	0.981301	0.029666	1.000000	0.999663	0.995942	0.920626	1.000000	0.017703	0.999996	...	0.151218
16.0	0.978163	0.980093	0.020209	1.000000	0.999631	0.995677	0.915670	1.000000	0.013605	0.999996	...	0.133677
17.0	0.976814	0.978862	0.015835	1.000000	0.999608	0.995408	0.910642	1.000000	0.010401	0.999996	...	0.117878
18.0	0.975529	0.977689	0.012549	1.000000	0.999585	0.995151	0.905871	1.000000	0.008050	0.999995	...	0.104552
19.0	0.974100	0.976385	0.009685	1.000000	0.999561	0.994864	0.900588	1.000000	0.006052	0.999995	...	0.091476
20.0	0.972633	0.975045	0.007420	1.000000	0.999536	0.994570	0.895187	1.000000	0.004513	0.999995	...	0.079733
21.0	0.971038	0.973589	0.005552	1.000000	0.999508	0.994250	0.889344	1.000000	0.003279	0.999994	...	0.068855
22.0	0.969535	0.972217	0.004223	1.000000	0.999482	0.993948	0.883864	1.000000	0.002425	0.999994	...	0.059616
23.0	0.967942	0.970762	0.003158	1.000000	0.999455	0.993627	0.878086	1.000000	0.001761	0.999994	...	0.051321
24.0	0.966227	0.969196	0.002309	1.000000	0.999425	0.993282	0.871894	1.000000	0.001247	0.999993	...	0.043662
25.0	0.964311	0.966641	0.001383	1.000000	0.999377	0.992717	0.861869	1.000000	0.000710	0.999993	...	0.033526

25 rows × 19158 columns

Gambar 7. Hasil Proyeksi Pelanggan *non-churn*

Selanjutnya disajikan hasil statistika deskriptif dari proyeksi pelanggan *non-churn* untuk bertahan hidup di perusahaan seperti pada Gambar 8. Terlihat bahwa rata-rata ekspektasi *survival* dari pelanggan *non-churn* untuk bersama dengan perusahaan adalah 15 bulan. Sementara ekspektasi *survival* terendah pelanggan adalah 1 bulan dan yang tertinggi sebesar 24 bulan.

```
count    19158.000000  
mean     15.648125  
std      8.478723  
min      1.508163  
25%      6.968115  
50%      19.893583  
75%      23.087027  
max      24.000000  
dtype: float64
```

Gambar 8. Statistika Deskriptif Ekspetasi *Survival*

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa model klasifikasi LightGBM berhasil secara optimal dalam menangani prediksi pelanggan *churn* dengan data yang tidak seimbang. Model tersebut menghasilkan nilai evaluasi yang baik tanpa perlu penyesuaian parameter maupun perlakuan khusus tambahan. Hasil evaluasi model ini mencapai nilai *Accuracy*, *Precision*, *Recall*, dan *F1-score* masing-masing sebesar 0.964, 0.971, 0.990, dan 0.980. Selain itu, model survival Cox Proportional Hazard mampu secara optimal memproyeksikan pelanggan *non-churn*. Model tersebut menghasilkan nilai evaluasi C-index sebesar 0.83 yang berarti sudah cukup mumpuni untuk memproyeksikan ketahanan hidup (*survival*) pelanggan. Selain itu, model ini memberikan informasi bahwa pelanggan yang saat ini *non-churn* memiliki rata-rata ekspetasi *survival* selama 15 bulan.

DAFTAR PUSTAKA

- [1] A. Tahar, P. B. Setiadi, dan S. Rahayu, “Strategi Pengembangan Sumber Daya Manusia dalam Menghadapi Era Revolusi Industri 4.0 Menuju Era Society 5.0,” *J. Pendidik. Tambusai*, vol. 6, no. 2, hal. 12380–12394, 2022.
- [2] U. Alrasyid, “Infografis Startup Indonesia Terbesar Ke-6 Dunia,” iNews.id. Diakses: 29 September 2023. [Daring]. Tersedia pada: <https://www.inews.id/multimedia/infografis/infografis-startup-indonesia-terbesar-ke-6-dunia>
- [3] F. F. Maulyan, D. Yuniati Drajat, R. Y. Angliawati, dan D. Sandini, “Pengaruh Service Excellent Terhadap Citra Perusahaan Dan Loyalitas Pelanggan: Theoretical Review,” *J. Sains Manaj.*, vol. 4, no. 1, hal. 8–17, 2022, doi: 10.51977/jsm.v4i1.660.
- [4] L. B. Asmaradana dan M. D. Satyawan, “Analisis Financial Distress dengan Model Altman, Grover, Springate, Zmijewski, dan Ohlson pada Perusahaan Subsektor Jasa Konsumen yang terdaftar di BEI,” *J. Ilmu Komputer, Ekon. dan Manaj.*, vol. 2, no. 1, hal. 1325–1341, 2022.
- [5] H. N. Irmanda, R. Astriratma, dan S. Afrizal, “Perbandingan Metode Jaringan Syaraf Tiruan Dan Pohon Keputusan Untuk Prediksi Churn,” *JSI J. Sist. Inf.*, vol. 11, no. 2, hal. 1817–1825, 2019, doi: 10.36706/jsi.v11i2.9286.
- [6] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, dan S. W. Kim, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,” *IEEE Access*, vol. 7, hal. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
- [7] J. Pamina *dkk.*, “An effective classifier for predicting churn in telecommunication,” *J. Adv. Res. Dyn. Control Syst.*, vol. 11, no. 1 Special Issue, hal. 221–229, 2019.

- [8] K. Arai, I. Fujikawa, Y. Nakagawa, R. Momozaki, dan S. Ogawa, “Churn Customer Estimation Method based on LightGBM for Improving Sales,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, hal. 119–125, 2023, doi: 10.14569/IJACSA.2023.0140215.
- [9] T. M. Fahrudin dan I. Z. A. Illah, “SkinMate: Mobile-Based Application for Detecting Multi-Class Skin Diseases Classification Using Pre-Trained MobileNetV2 on CNN Architecture,” *Proceeding - IEEE 9th Inf. Technol. Int. Semin. ITIS 2023*, 2023, doi: 10.1109/ITIS59651.2023.10420370.
- [10] T. Informasi, V. No, dan M. A. D. Yuda, “Transformasi Data Solarman Untuk Pengungkapan Informasi dan Pola PLTS dengan Metode Semi-Supervised Learning,” vol. 1, no. 2, hal. 100–110.
- [11] C. Zai, “Implementasi Data Mining Sebagai Pengolahan Data,” vol. 2, no. 3, hal. 1–12, 2022.
- [12] M. N. Faiz, O. Somantri, dan A. W. Muhammad, “Rekayasa Fitur Berbasis Machine Learning untuk Mendeteksi Serangan DDoS,” vol. 11, no. 3, hal. 176–182, 2022.
- [13] Btd, “Simplifying Complexity: Feature Selection with Recursive Feature Elimination,” medium. Diakses: 2 Februari 2024. [Daring]. Tersedia pada: <https://readmedium.com/en/https://baotramduong.medium.com/machine-learning-feature-selection-recursive-feature-elimination-88750767c05f>
- [14] E. Y. Puspaningrum, Y. V. Via, C. Nisa, H. Maulana, dan W. S. J. Saputra, “Oversampled-Based Approach to Overcome Imbalance Data in the Classification of Apple Leaf Disease with SMOTE,” *Rom. J. Ofapplied Sci. Technol.*, vol. 16, no. 3, hal. 112–117, 2023.
- [15] F. I. Kurniadi dan P. D. Larasati, “Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke,” *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 6, no. 1, hal. 67–72, 2022, doi: 10.47970/siskom-kb.v6i1.328.
- [16] D. A. I. Maruddani, Tarno, A. Hoyyi, R. Rahmawati, dan Y. Wilandari, *Survival analysis*. UNDIP Press Semarang, 2021. doi: 10.1007/978-1-4614-4322-3_1.
- [17] J. Soedirman, “Analisis Bibliometrik Penggunaan Analisis Survival Model Regresi Cox,” *J. Ris. Akunt. Soedirman*, vol. 2, no. 2, hal. 111–120, 2023, doi: 10.32424/1.jras.2023.2.2.10806.
- [18] P. Purwono, A. Wirasto, dan K. Nisa, “Comparison of Machine Learning Algorithms for Classification of Drug Groups,” *Sisfotenika*, vol. 11, no. 2, hal. 196, 2021, doi: 10.30700/jst.v11i2.1134.
- [19] A. T. Damaliana, T. Trimono, and D. A. Prasetya, “Ensemble Tree untuk Memprediksi Level Resiko Maternal Mortality di Bangladesh,” *Prosiding Seminar Nasional Sains Data*, vol. 2, no. 1, pp. 24–29, Nov. 2022, Accessed: Jun. 08, 2024.
- [20] D. A. Prawinata, A. D. R. Rahajoe, dan I. G. S. M. Diyasa, “Analisis Sentimen Kendaraan Listrik Pada Twitter Menggunakan Metode Long Short Term Memory,” *SABER J. Tek. Inform. Sains dan Ilmu Komun.*, vol. 2, no. 1, hal. 300–313, 2024.

