Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.1, February 2025, pp. 89~95
e-ISSN: 2548-6861

Evaluation of Telecommunication Customer Churn Classification with SMOTE Using Random Forest and XGBoost Algorithms

Lisa Nusrotul Wakhidah [1]*, Akhmad Khanif Zyen [2], Buang Budi Wahono [3]***
* Teknik Informatika, Universitas Islam Nahdlatul Ulama Jepara
lisanusrotulw@gmail.com [1], khanif.zyen@unisnu.ac.id [2], budihono@unisnu.ac.id [3]

Article Info

Article history:

Received 2024-10-24
Revised 2024-11-07
Accepted 2024-11-14

Keyword:

Customer Churn,
Telecommunication,
Random Forest,
SMOTE,
XGBoost.

ABSTRACT

Competition in the telecommunications industry, particularly among Internet Service Providers (ISPs), significantly influences customer churn, which negatively impacts revenue, profitability, and business sustainability. An effective approach to mitigate churn involves identifying potential churners early, enabling companies to implement strategic retention measures. However, predicting churn can be challenging due to the limited data available on churned customers. This study aims to predict customers likely to terminate or discontinue their subscriptions, focusing on addressing data imbalance using the Synthetic Minority Over-Sampling Technique (SMOTE). The dataset, sourced from Kaggle, comprises 21 attributes and 7,034 entries. The pre-processing phase includes data cleaning, feature encoding, and the implementation of Random Forest and XGBoost algorithms after data balancing with SMOTE. The findings reveal that the XGBoost algorithm achieves a prediction accuracy of 82%, outperforming Random Forest with 81%. Key factors influencing churn include Contract, TotalCharges, and tenure. The study concludes by emphasizing the significance of contract flexibility and the need to prioritize customers with high total costs or extended subscription periods to reduce churn rates. Future research is encouraged to investigate alternative methods for handling data imbalance and to explore advanced machine learning algorithms to further enhance prediction accuracy and the effectiveness of customer retention strategies.

I. PENDAHULUAN

Menurut Asosiasi Penyelenggara Jasa Internet Indonesia (APJII), jumlah pengguna internet di Indonesia diperkirakan mencapai 221,6 juta jiwa pada 2024 dari total populasi 278,7 juta jiwa pada 2023. Tingkat penetrasi internet juga naik menjadi 79,5%, meningkat 1,4% dari tahun sebelumnya. Pertumbuhan ini mencerminkan tingginya adopsi layanan internet dan ketergantungan masyarakat pada layanan digital, yang meningkatkan ekspektasi pelanggan terhadap kualitas layanan [1]. Persaingan ini mendorong operator untuk tidak hanya mengembangkan produk dan layanan, tetapi juga fokus pada pelanggan. Pelanggan kini bebas memilih operator sesuai kebutuhan dan dapat beralih atau churn sewaktu-waktu. Untuk menghadapi dinamika pasar dan tingginya ekspektasi pelanggan, operator perlu mengadopsi strategi yang berfokus tidak hanya pada akuisisi, tetapi juga pada retensi pelanggan [2].

Strategi peningkatan pendapatan dapat dicapai melalui perolehan pelanggan baru, peningkatan volume penjualan, serta perpanjangan periode retensi pelanggan. Diantara ketiga pendekatan ini strategi retensi pelanggan terbukti paling menguntungkan ketika dievaluasi berdasarkan Return on Investment (RoI). Hal ini disebabkan oleh rendahnya biaya yang diperlukan untuk mempertahankan pelanggan eksisting dibandingkan dengan biaya akuisisi pelanggan baru [3]. Salah satu upaya untuk meningkatkan retensi pelanggan adalah dengan mengurangi potensi churn, yaitu peralihan pelanggan dari satu perusahaan ke perusahaan lain dalam periode tertentu. Churn adalah istilah yang merujuk pada perpindahan pelanggan dari satu penyedia layanan ke layanan lain, yang bisa berdampak besar pada pendapatan perusahaan [4]. Untuk mengatasi hal ini, perusahaan perlu memprediksi risiko churn

dengan menganalisis data pelanggan dan menemukan pola perilaku yang relevan [5]. Salah satu pendekatan yang sering digunakan adalah data mining dan teknik klasifikasi, yang dapat membantu mengidentifikasi apakah seorang pelanggan berpotensi untuk churn atau tetap setia. Data mining sendiri berfungsi untuk mengekstraksi informasi dari data dalam jumlah besar, dengan tahapan seperti eksplorasi data, pemilihan metode yang sesuai, hingga pengembangan model prediksi yang optimal [6]. Dalam penelitian ini, teknik klasifikasi diterapkan untuk memproses data latih dan data uji, serta menilai akurasi model dalam memprediksi churn [7].

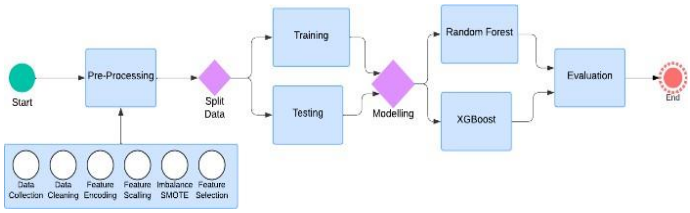
Penelitian terkait customer churn telah banyak dilakukan dengan berbagai pendekatan algoritma. Stevan Desena et al. menggunakan algoritma C4.5 untuk mengidentifikasi faktor-faktor yang memengaruhi churn di industri telekomunikasi, seperti kontrak, layanan internet, dan total biaya. Penelitian ini mencapai akurasi 79,53%, menunjukkan bahwa model C4.5 cukup efektif dalam membantu perusahaan mempertahankan pelanggan[8]. Kemudian penelitian oleh Muhammad Maulana Sidiq et al. menggunakan pendekatan Exploratory Data Analysis (EDA) dan beberapa model machine learning, termasuk Logistic Regression, Random Forest, SVM, Gradient Boosting, AdaBoost, dan XGBoost. Tantangan utama dalam penelitian ini adalah ketidakseimbangan kelas pada dataset, yang diatasi dengan teknik oversampling SMOTE. Hasil menunjukkan bahwa model XGBoost mencapai akurasi tertinggi sebesar 82,94%, mengungguli model lainnya[9]. Dalam konteks lain, penelitian oleh Agung Nugroho et al. menerapkan metode SMOTE pada algoritma Random Forest untuk memprediksi kebangkrutan perusahaan, dan berhasil meningkatkan akurasi sebesar 7,40%, mencapai 95,70%. Penelitian ini menunjukkan efektivitas SMOTE dalam meningkatkan akurasi prediksi pada dataset yang tidak seimbang[10].

Penanganan data yang tidak seimbang sangat penting dalam meningkatkan kinerja model prediksi churn di industri telekomunikasi. Dengan menerapkan teknik resampling seperti SMOTE dan menggunakan model prediktif yang canggih, perusahaan telekomunikasi dapat mengidentifikasi pelanggan yang berisiko churn dengan lebih akurat dan mengambil tindakan yang tepat untuk mempertahankan mereka. Untuk meningkatkan akurasi prediksi churn, segmentasi pelanggan berdasarkan perilaku dan pola penggunaan mereka sangat diperlukan. Penelitian telah membuktikan bahwa teknik seperti SMOTE (Synthetic Minority Over-sampling Technique) dapat menangani masalah dataset yang tidak seimbang[11], yang sering ditemukan dalam prediksi churn. Dengan demikian, adopsi teknik penyeimbangan data seperti SMOTE bersama algoritma pembelajaran mesin dapat menjadi strategi efektif untuk meningkatkan retensi pelanggan dalam industri telekomunikasi[12].

Penelitian ini melanjutkan studi-studi sebelumnya dengan mengevaluasi efektivitas dua algoritma machine learning, yakni Random Forest dan XGBoost, dalam memprediksi customer churn di industri telekomunikasi. Untuk menangani masalah ketidakseimbangan kelas pada dataset, diterapkan teknik SMOTE (Synthetic Minority Over- sampling Technique). Tujuan utama dari penelitian ini adalah mengidentifikasi model algoritma yang paling optimal dalam klasifikasi churn, sehingga perusahaan dapat mengenali karakteristik pelanggan yang berisiko churn dan merancang strategi retensi yang lebih efektif. Hasil penelitian ini diharapkan dapat memberikan wawasan baru terkait penerapan algoritma serta teknik penyeimbangan data, serta menjadi dasar pengembangan strategi untuk mengurangi churn, meningkatkan retensi pelanggan, dan mengurangi potensi kerugian perusahaan akibat kehilangan pelanggan..

II. METODE

Proses penelitian dimulai dengan pengumpulan data dari Kaggle, diikuti oleh preprocessing yang mencakup pembersihan, encoding, scaling, dan penyeimbangan data menggunakan SMOTE. Data kemudian dibagi untuk pelatihan dan pengujian, di mana algoritma Random Forest dan XGBoost diterapkan. Tahap akhir adalah evaluasi kinerja model. Uraian proses Teknik analisis data dijelaskan setelah pada Gambar 1.



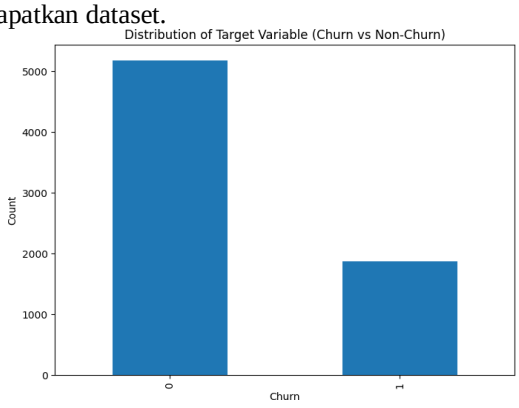
Gambar 1. Teknik Analisis Data

A. *Pre-Processing*

Pre-processing merupakan tahap awal dalam analisis data yang bertujuan untuk mempersiapkan data mentah agar siap digunakan dalam model. Tahapan ini terdiri dari beberapa langkah penting sebagai berikut.

1) *Data Collection*

Pengumpulan data adalah langkah pertama di mana data yang relevan dikumpulkan dari berbagai sumber. Dalam konteks penelitian ini, menggunakan Kaggle untuk mendapatkan dataset.



Gambar 2. Visualisasi Data Collection

Data dengan jumlah sebanyak 7043, dan 21 atribut, dan memiliki 1 label yaitu churn yang memiliki 2 kelas yaitu ‘1’ sebanyak 1869 dan ‘0’ sebanyak 5164. mencakup berbagai atribut seperti SeniorCitizen, Partner, Dependents, tenure, Contract, dan lainnya.

2) Data Cleaning

Data Cleaning adalah proses membersihkan data dengan menghapus atau memperbaiki nilai yang hilang, menghilangkan duplikasi, dan menghapus data yang tidak relevan[13]. Tujuannya memastikan bahwa dataset bersih dari kesalahan sehingga dapat memberikan hasil yang akurat pada model. Proses ini, kolom 'customerID' dihapus karena tidak berpengaruh pada analisis

3) Feature Encoding

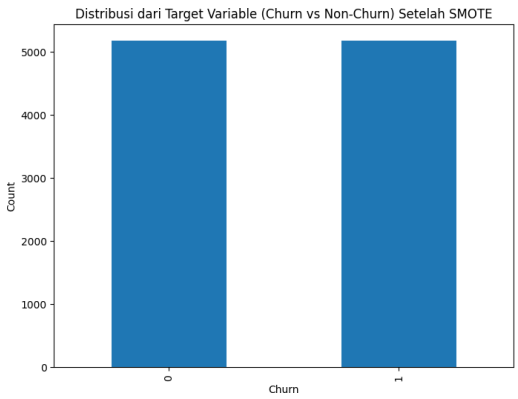
Feature Encoding bertujuan untuk membuat data yang bersifat kategorikal (teks) menjadi angka sehingga algoritma machine learning dapat memahami dan mengolah data tersebut[14]. Teknik encoding pada penelitian ini yang digunakan adalah one-hot encoding atau label encoding. Nilai 'Yes' pada kolom 'Churn' diubah menjadi 1 dan 'No' diubah menjadi 0.

4) Feature Scalling

Untuk memastikan bahwa semua fitur memiliki skala yang sama, langkah scaling diterapkan[14]. Dalam penelitian ini, digunakan metode StandardScaler yang bertujuan untuk menstandarisasi fitur seperti 'TotalCharges' dan 'tenure'. Tanpa scaling, model dapat bias terhadap fitur dengan rentang nilai yang lebih besar, sehingga menyulitkan algoritma dalam belajar dari data tersebut

5) Imbalance handling dengan SMOTE

SMOTE (synthetic Minority Over-Sampling Technique) digunakan untuk menangani ketidakseimbangan kelas dengan membuat sampel sintesis dari kelas minoritas[15]. proses ini melibatkan pemilihan sampel minoritas, menemukan tetangga terdekatnya dan menginterpolasi nilai fitur untuk menghasilkan sampel baru. dengan menggunakan SMOTE, distribusi kelas menjadi seimbang, sehingga model dapat lebih akurat dalam memprediksi kelas minoritas[16].



Gambar 3. Visualisasi setelah SMOTE

6) Feature Selection

Setelah melakukan encoding dan scaling, Langkah feature selection dilakukan untuk memilih fitur-fitur yang paling relevan terhadap variable target[17].

```
[8] # Select relevant features
    selected_features = ['Contract', 'InternetService', 'TotalCharges', 'tenure', 'PaperlessBilling', 'MultipleLines', 'StreamingMovies']
    X_selected = X_resampled[:, [X.columns.get_loc(feature) for feature in selected_features]]
```

Gambar 4. Feature Selection

Pada penelitian fitur yang dipilih mencakup fitur seperti ‘Contract’, TotalCharges’, ‘tenure’, . pemilihan fitur ini bertujuan untuk mengurangi komplesitas model dan meningkatkan akurasi dengan hanya mempertahankan fitur yang siginifikan.

B. Split Data

Setelah pre-processing selesai, data dibagi menjadi dua bagian training set dan test set. Training set digunakan untuk melatih model agar model dapat belajar dan menemukan pola dari data tersebut. Sementara itu, test set digunakan untuk menguji performa model pada data yang belum pernah dilihat sebelumnya. Pembagian data ini penting agar kita bisa mengevaluasi seberapa baik model bekerja pada data baru. Dalam skrip ini, pembagian dilakukan dengan train_test_split, di mana 80% data digunakan untuk pelatihan dan 20% untuk pengujian[12]. Pendekatan ini bertujuan untuk mencegah overfitting dan memastikan bahwa model memiliki generalisasi yang baik saat dihadapkan pada data yang belum dikenal.

C. Modelling

Algoritma Random Forest dan XGBoost dipilih dalam penelitian ini karena keduanya memiliki keunggulan dalam menangani data churn yang kompleks dan tidak seimbang di industri telekomunikasi.

Random Forest merupakan algoritma berbasis ensemble yang menghasilkan prediksi dengan menggabungkan banyak pohon keputusan (decision trees), sehingga memberikan hasil yang lebih akurat dan stabil. Kemampuan algoritma ini untuk mengurangi risiko overfitting menjadikannya pilihan yang tepat untuk dataset dengan banyak fitur, yang sering kali menjadi tantangan dalam memprediksi churn pelanggan[18]. Selain itu, Random Forest memiliki fleksibilitas dalam menangani missing values dan tidak sensitif terhadap perbedaan skala antar fitur, sehingga cocok untuk diterapkan pada berbagai jenis data.

XGBoost merupakan algoritma boosting yang efisien dan efektif dalam meningkatkan akurasi prediksi melalui pendekatan iteratif. Algoritma ini memberikan bobot lebih besar pada kesalahan klasifikasi di setiap iterasi, sehingga model dapat memperbaiki prediksi secara bertahap. Pendekatan ini sangat relevan untuk menangani dataset churn yang tidak seimbang, di mana kelas minoritas sering kali kurang terwakili[19]. Selain efisiensi komputasi, kemampuan optimasi XGBoost memungkinkan pemrosesan dataset berukuran besar dengan cepat. Dengan menggabungkan kedua algoritma ini, penelitian ini bertujuan untuk

memaksimalkan akurasi prediksi churn sekaligus mengatasi tantangan ketidakseimbangan data, sehingga menghasilkan model yang lebih baik dalam mengenali pola churn dan mendeteksi kelas minoritas.

D. Evaluation

Proses evaluasi model melibatkan penggunaan Confusion Matrix dan perhitungan beberapa metrik penting seperti akurasi, precision, recall, F1-score, dan ROC AUC Score. Evaluasi ini bertujuan untuk menilai performa model dalam memprediksi churn pelanggan. Akurasi dihitung sebagai persentase prediksi yang benar terhadap total data, sementara precision mengukur ketepatan model dalam memprediksi pelanggan churn. Recall menunjukkan seberapa baik model dalam mendeteksi pelanggan yang churn, dan F1-score adalah metrik yang menyeimbangkan antara precision dan recall, yang sangat penting untuk kasus data yang tidak seimbang. Selain itu, ROC AUC Score digunakan untuk menilai kemampuan model dalam membedakan antara kelas churn dan non-churn di berbagai ambang batas prediksi nilai AUC yang lebih tinggi mengindikasikan kinerja model yang lebih baik.

TABEL I
CONFUSION MATRIX

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Confusion Matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan menunjukkan jumlah prediksi benar dan salah yang dibuat oleh model [20]. Confusion Matrix mencakup empat komponen utama: True Positives (TP), yaitu jumlah kasus di mana model memprediksi churn dengan benar; True Negatives (TN), yaitu jumlah kasus di mana model memprediksi tidak churn dengan benar; False Positives (FP), yaitu jumlah kasus di mana model memprediksi churn tetapi sebenarnya tidak churn dan False Negatives (FN), yaitu jumlah kasus di mana model memprediksi tidak churn tetapi sebenarnya churn. Hasil dari Confusion Matrix disajikan dalam Tabel 1 untuk memberikan gambaran tentang kemampuan model dalam mengklasifikasikan churn dan non-churn secara akurat.

III. HASIL DAN PEMBAHASAN

Pada tahap preprocessing, data diolah dan dilakukan penyeimbangan kelas menggunakan metode SMOTE (Synthetic Minority Over-sampling Technique). SMOTE dipilih karena kemampuannya yang terbukti dalam meningkatkan representasi kelas minoritas tanpa mengurangi informasi pada kelas mayoritas, berbeda dengan metode penghapusan data yang berisiko menghilangkan informasi penting. Dengan menerapkan SMOTE, distribusi kelas minoritas yang sebelumnya tidak seimbang menjadi lebih proporsional, yang pada gilirannya meningkatkan akurasi dan kemampuan generalisasi model. Teknik ini sangat krusial dalam skenario churn di mana data ketidakseimbangan kelas signifikan dan cenderung menyebabkan model berpihak pada kelas mayoritas. Setelah proses penyeimbangan, fitur-fitur yang relevan dipilih melalui feature selection, dan data kemudian dibagi menjadi dua bagian, yaitu 80% untuk data latih dan 20% untuk data uji.

Model Random Forest dan XGBoost dipilih karena kemampuannya dalam menangani data yang kompleks dan fitur yang saling berinteraksi, serta ketahanannya terhadap overfitting. Kedua algoritma ini juga dikenal memiliki performa yang baik pada data yang tidak seimbang, yang menjadikannya kandidat ideal untuk prediksi churn. Dalam proses evaluasi model, berbagai metrik digunakan, termasuk akurasi, precision, recall, F1-score, dan ROC AUC score. Metrik akurasi saja tidak cukup untuk data tidak seimbang karena bisa menyesatkan bila salah satu kelas mendominasi prediksi model. Oleh karena itu, metrik precision dan recall memberikan informasi tambahan tentang ketepatan dan sensitivitas model dalam mendeteksi kelas churn, sedangkan F1-score menawarkan keseimbangan antara precision dan recall. Metrik ROC AUC, di sisi lain, memberikan gambaran keseluruhan kemampuan model dalam membedakan kelas churn dan non-churn di berbagai ambang batas prediksi. Penggunaan metrik-metrik ini membantu memastikan bahwa model tidak hanya akurat secara keseluruhan tetapi juga efektif dalam mendeteksi pelanggan yang cenderung churn.

	precision	recall	f1-score	support
No Churn	0.79	0.82	0.81	1021
Churn	0.82	0.79	0.81	1049
accuracy			0.81	2070
macro avg	0.81	0.81	0.81	2070
weighted avg	0.81	0.81	0.81	2070

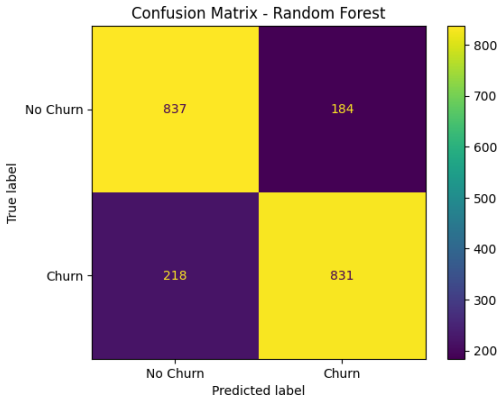
Gambar 4. Hasil Evaluasi Random Forest

	precision	recall	f1-score	support
No Churn	0.83	0.80	0.82	1021
Churn	0.81	0.84	0.83	1049
accuracy			0.82	2070
macro avg	0.82	0.82	0.82	2070
weighted avg	0.82	0.82	0.82	2070

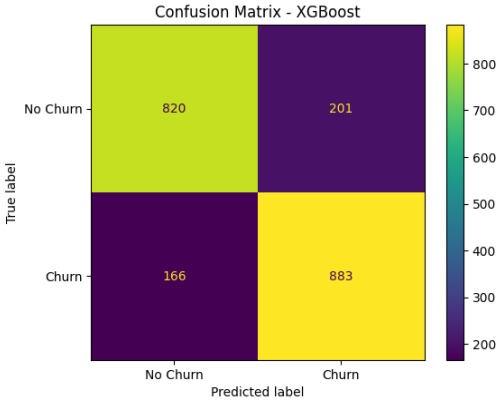
Gambar 5. Hasil Evaluasi XGBoost

Hasil evaluasi yang ditunjukkan pada gambar 4 dan gambar 5 mengilustrasikan peran penting SMOTE dalam menyeimbangkan kelas minoritas dan mayoritas. Penyeimbangan ini terbukti meningkatkan akurasi dan generalisasi model. Model Random Forest memberikan performa dengan akurasi 81%, precision 81%, recall 79%, dan F1-score 81%, sedangkan nilai AUC ROC-nya mencapai 0.89. Hasil ini menunjukkan bahwa model memiliki kemampuan cukup baik dalam membedakan antara kelas churn dan non-churn. Model XGBoost sedikit lebih unggul dengan akurasi 82%, precision 81%, recall 84%, dan F1-score 83%. AUC ROC XGBoost mencapai 0.91, yang mendekati nilai sempurna. Peningkatan recall pada model XGBoost menunjukkan kemampuan yang lebih baik dalam mendeteksi pelanggan churn, yang menjadi kelas minoritas dalam dataset. Hal ini penting dalam konteks prediksi churn karena kemampuan model untuk mendeteksi sebanyak mungkin

pelanggan yang akan churn (recall tinggi) lebih diutamakan daripada hanya sekedar akurasi keseluruhan. Perbedaan performa antara Random Forest dan XGBoost menunjukkan bahwa XGBoost lebih responsif dalam mengklasifikasikan pelanggan churn, menjadikannya model yang lebih andal untuk strategi retensi pelanggan di industri telekomunikasi.

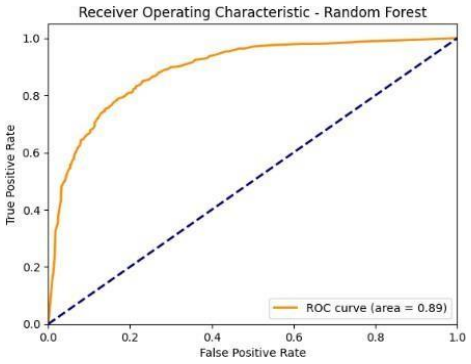


Gambar 6. Confusion Matrix Random Forest

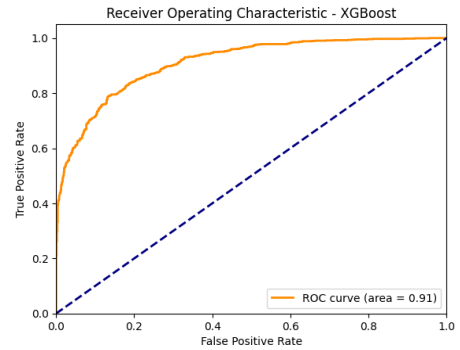


Gambar 7. Confusion Matrix XGBoost

Confusion Matrix yang disajikan pada Gambar 6 dan Gambar 7 memberikan informasi rinci mengenai jumlah prediksi benar dan salah yang dilakukan oleh model untuk setiap kelas, Confusion Matrix membantu kita memahami distribusi prediksi model, terutama dalam konteks data yang tidak seimbang seperti churn di industri telekomunikasi. Dalam kasus ini, True Positives (TP) menunjukkan kemampuan model dalam mendeteksi pelanggan yang benar-benar churn, sedangkan False Negatives (FN) menunjukkan pelanggan churn yang gagal terdeteksi oleh model. Tingginya jumlah FN akan berdampak signifikan pada strategi retensi pelanggan, karena pelanggan yang tidak terdeteksi berisiko tidak mendapatkan tindakan pencegahan dari perusahaan. Demikian pula, True Negatives (TN) dan False Positives (FP) penting dalam menilai efisiensi model, karena FP dapat menyebabkan alokasi sumber daya yang kurang efektif untuk pelanggan yang sebenarnya tidak churn Metrik ROC AUC digunakan untuk mengevaluasi kemampuan model dalam membedakan kelas churn dan non-churn. Kurva ini menunjukkan trade-off antara True Positive Rate dan False Positive Rate pada berbagai ambang batas. Gambar 8 dan 9 menampilkan kurva ROC AUC untuk model Random Forest dan XGBoost, memberikan gambaran visual mengenai performa prediktif kedua model.



Gambar 8. ROC Curve Random Forest

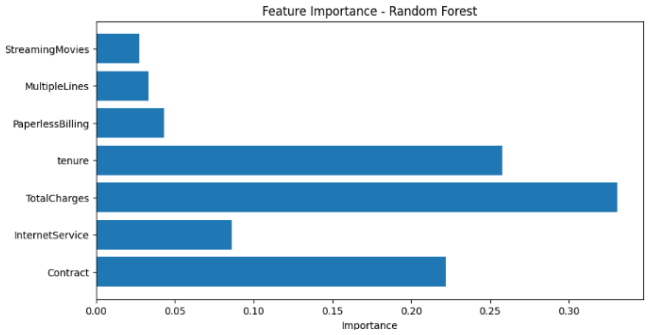


Gambar 9. ROC Curve XGBoost

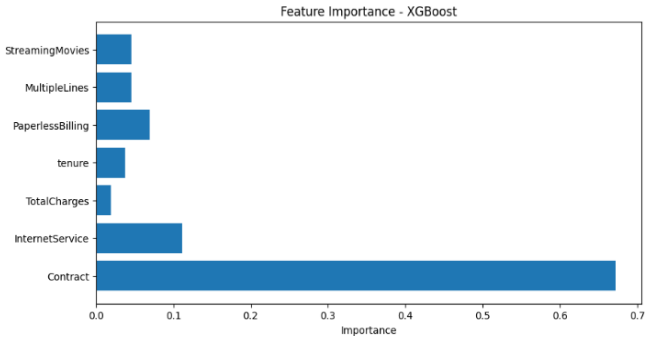
Metrik ROC AUC dipilih karena memberikan pandangan menyeluruh tentang kemampuan model dalam membedakan antara kelas churn dan non-churn pada berbagai ambang batas prediksi. Dalam industri telekomunikasi, kemampuan untuk mengidentifikasi pelanggan yang berpotensi churn sangat penting karena perusahaan perlu mengalokasikan sumber daya untuk strategi retensi dengan tepat. AUC yang tinggi mendekati 1.0 menunjukkan bahwa model memiliki kemampuan yang baik dalam memisahkan kelas, sehingga perusahaan dapat lebih yakin dalam menggunakan prediksi ini untuk pengambilan keputusan bisnis yang strategis.

Random Forest dan XGBoost dipilih karena keduanya merupakan algoritma yang andal untuk klasifikasi dengan data yang mungkin memiliki hubungan non-linear, seperti data churn pelanggan. Random Forest, sebagai metode ensemble, membantu dalam mengurangi overfitting dan meningkatkan stabilitas model. Di sisi lain, XGBoost yang berbasis gradient boosting umumnya lebih efektif dalam menangani ketidakseimbangan data, karena algoritma ini dapat memberikan bobot yang lebih besar pada kelas minoritas selama pelatihan. Dalam konteks ini, XGBoost sedikit lebih unggul karena mampu memberikan akurasi lebih

tinggi pada kelas churn tanpa terlalu banyak meningkatkan False Positives.



Gambar 10. BarPlot Feature Importance Random Forest



Gambar 11. BarPlot Feature Importance XGBoost

Gambar 10 dan 11 menunjukkan BarPlot Feature Importance Random Forest dan XGBoost dimana algoritma Random Forest dan XGBoost mengungkap tiga fitur utama yang paling relevan dalam memprediksi churn pelanggan di industri telekomunikasi, yaitu TotalCharges, tenure, dan Contract. Masing-masing fitur ini berperan penting dalam memahami pola perilaku pelanggan, yang pada gilirannya berkontribusi pada pengembangan strategi retensi yang lebih efektif dan berbasis data.

1) TotalCharges sebagai prediktor utama dalam model Random Forest mengindikasikan bahwa pelanggan dengan biaya akumulatif yang tinggi cenderung memiliki ekspektasi tinggi terhadap layanan. Ketika ekspektasi ini tidak terpenuhi, risiko churn meningkat. TotalCharges berfungsi sebagai indikator komprehensif, mencerminkan interaksi pelanggan yang lebih dalam dengan layanan yang mereka gunakan. Dari perspektif bisnis, pelanggan dengan TotalCharges yang tinggi adalah segmen bernilai tinggi yang berpotensi meningkatkan pendapatan jangka panjang, tetapi mereka juga lebih rentan churn jika mereka merasa tidak mendapatkan manfaat yang sebanding. Oleh karena itu, perusahaan dapat memantau pelanggan yang mendekati atau melebihi tingkat biaya tertentu dan menawarkan diskon atau insentif untuk mempertahankan mereka.

2) Tenure, atau lama langganan, juga memiliki kontribusi signifikan dalam model. Pelanggan dengan durasi layanan yang pendek atau ekstrem, sangat pendek maupun sangat panjang, menunjukkan kecenderungan churn yang lebih tinggi. Pelanggan baru mungkin merasa kurang terikat pada layanan, sementara pelanggan jangka panjang, meskipun loyal, dapat menunjukkan kejenuhan atau kekecewaan jika tidak ada inovasi atau peningkatan dalam layanan. Oleh karena itu, pendekatan retensi yang ditargetkan sesuai dengan fase hubungan pelanggan dengan perusahaan dapat meningkatkan kepuasan pelanggan secara keseluruhan. Misalnya, memberikan penawaran spesial pada tahun pertama atau memberikan insentif loyalitas pada pelanggan lama dapat membantu menurunkan tingkat churn.

3) Contract dalam model XGBoost ditunjukkan sebagai variabel paling dominan, menegaskan pentingnya jenis kontrak dalam mempengaruhi keputusan churn. Pelanggan dengan kontrak bulanan cenderung lebih bebas untuk berpindah layanan dibandingkan dengan pelanggan yang terikat kontrak tahunan atau jangka panjang, yang memiliki lebih banyak komitmen. Perusahaan telekomunikasi dapat memanfaatkan informasi ini dengan menyediakan opsi kontrak yang fleksibel, Oleh karena itu, strategi utama yang dapat diterapkan adalah mendorong pelanggan bulanan untuk berpindah ke kontrak yang lebih lama dengan menawarkan insentif tertentu, seperti diskon bulanan atau peningkatan layanan. Langkah ini dapat mengurangi churn pada segmen pelanggan kontrak bulanan yang lebih fluktuatif.

Model XGBoost sedikit lebih unggul dalam akurasi dibandingkan Random Forest, terutama dalam mengidentifikasi pelanggan yang lebih rentan terhadap churn berdasarkan fleksibilitas kontrak. Penerapan SMOTE dalam penyeimbangan data turut meningkatkan performa prediksi kedua model, yang berarti bahwa pengelolaan data yang lebih seimbang penting dalam memaksimalkan keandalan model prediktif pada dataset dengan kelas yang tidak seimbang.
Secara keseluruhan, hasil feature importance ini memberikan wawasan strategis yang mendalam, menunjukkan bahwa fokus pada pengelolaan biaya pelanggan, durasi layanan, dan fleksibilitas kontrak merupakan langkah utama dalam memperkuat loyalitas dan meminimalkan churn. Dengan memahami prioritas setiap fitur dalam prediksi churn, perusahaan telekomunikasi dapat mengimplementasikan intervensi yang lebih tepat sasaran untuk meningkatkan kepuasan pelanggan dan mengurangi tingkat churn, sehingga memaksimalkan retensi pelanggan dan pendapatan jangka panjang.

V. KESIMPULAN

Berdasarkan penelitian ini, dapat disimpulkan bahwa penanganan ketidakseimbangan data menggunakan metode SMOTE (Synthetic Minority Over-sampling Technique) merupakan langkah penting untuk meningkatkan akurasi prediksi churn pelanggan di industri telekomunikasi. Analisis yang dilakukan dengan algoritma Random Forest dan

XGBoost menunjukkan bahwa kedua model efektif dalam memprediksi churn, dengan XGBoost memberikan performa yang sedikit lebih baik dengan akurasi sebesar 82% dan nilai ROC AUC 0,91, dibandingkan Random Forest yang memiliki akurasi 81% dan nilai ROC AUC 0,89.

Fitur-fitur utama yang berperan signifikan dalam prediksi churn adalah TotalCharges, tenure, dan Contract. Secara khusus, XGBoost mengidentifikasi jenis kontrak, terutama kontrak bulanan, sebagai faktor yang paling terkait dengan risiko churn yang tinggi, karena fleksibilitasnya memungkinkan pelanggan lebih mudah berhenti berlangganan. Di sisi lain, Random Forest menyoroti bahwa pelanggan dengan total biaya yang tinggi dan durasi langganan tertentu memiliki kecenderungan churn yang lebih besar.

Implikasi praktis dari hasil penelitian ini adalah perlunya perusahaan telekomunikasi mengembangkan strategi retensi yang fleksibel tetapi menarik, seperti memberikan insentif untuk kontrak jangka panjang guna mengurangi risiko churn pada pelanggan bulanan. Selain itu, perhatian khusus harus diberikan kepada pelanggan dengan total biaya tinggi atau durasi langganan tertentu untuk menekan risiko churn. Dengan menerapkan strategi ini, perusahaan dapat merancang langkah-langkah proaktif yang lebih tepat sasaran dalam mempertahankan pelanggan sekaligus mengurangi tingkat churn secara signifikan.

DAFTAR PUSTAKA

[1] survei.apjii.or.id, "Survei Internet APJII 2024," survei.apjii.or.id. Accessed: Oct. 29, 2024. [Online]. Available: https://survei.apjii.or.id/

[2] A. Wicaksono, A. Anita, and T. N. Padilah, "Uji Performa Teknik Klasifikasi untuk Memprediksi Customer Churn," Bianglala Inform., vol. 9, no. 1, pp. 37–45, 2021, doi: 10.31294/bi.v9i1.9992.

[3] Siti Alvi Sholikhatin Khairunnisak Nur Isnaini, "Faculty of Sains and Technology, Ibrahimy University," Ilm. Inform., vol. 6, no. 1, pp. 43–49, 2021, [Online]. Available: Siti Alvi Sholikhatin1), Khairunnisak Nur Isnaini2)

[4] V. Kavitha, G. Hemanth Kumar, S. V Mohan Kumar, and M. Harish, "Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms," Int. J. Eng. Res., vol. V9, no. 05, pp. 181– 184, 2020, doi: 10.17577/ijertv9is050022.

[5] J. -, S. Usman, and F. Aziz, "Analisis Perilaku Pelanggan menggunakan Metode Ensemble Logistic Regression," J. Teknol. Dan Ilmu Komput. Prima, vol. 6, no. 2, pp. 90–97, 2023, doi: 10.34012/jutikomp.v6i2.4258.

[6] I. M. Latief, A. Subekti, and W. Gata, "Prediksi Tingkat Pelanggan Churn Pada Perusahaan Telekomunikasi Dengan Algoritma Adaboost," J. Inform., vol. 21, no. 1, pp. 34–43, 2021, doi: 10.30873/ji.v21i1.2867.

[7] J. Penelitian Ilmu Komputer, M. Dahlan Kurnia, D. Universitas Pamulang, and P. Brin, "Klasifikasi Customer Relationship Management Perusahaan Telekomunikasi Seluler Dengan Metode Machine Learning," vol. 1, no. 4, pp. 63–76, 2023, [Online]. Available: https://mypublikasi.com/

[8] S. D. Damanik and M. I. Jambak, "Klasifikasi Customer Churn pada Telekomunikasi Industri Untuk Retensi Pelanggan Menggunakan Algoritma C4.5," KLIK Kaji. Ilm. Inform. dan Komput., vol. 3, no. 6, pp. 1303–1309, 2023, doi: 10.30865/klik.v3i6.829.

[9] M. M. S. Nurhidayat and Dyah Anggraini, "Analysis and Classification of Customer Churn Using Machine Learning Models," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 7, no. 6, pp. 1253–1259, 2023, doi: 10.29207/resti.v7i6.4933.

[10] A. Nugroho and E. Rilvani, "Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan Application of the SMOTE Oversampling Method to the Random Forest Algorithm for Predicting Company Bankruptcy," Februari, vol. 22, no. 1, pp. 207–214, 2024.

[11] Cosmas Haryawan and Yosef Muria Kusuma Ardhana, "Analisa Perbandingan Teknik Oversampling Smote Pada Imbalanced Data," J. Inform. dan Rekayasa Elektron., vol. 6, no. 1, pp. 73–78, 2023, doi: 10.36595/jire.v6i1.834.

[12] A. N. Kasanah, M. Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 3, no. 2, pp. 196–201, 2019, doi: 10.29207/resti.v3i2.945.

[13] B. Ramadhan, D. Firdaus, and A. R. Rafi, "MIND (Multimedia Artificial Intelligent Networking Database Teknik SMOTE Sebagai Solusi Imbalance Class dalam Model Deteksi Intrusi DDoS dengan Metode PCA- Random Forest," J. MIND J. | ISSN, vol. 8, no. 1, pp. 52–64, 2023, [Online]. Available: https://doi.org/10.26760/mindjournal.v8i1.52-64

[14] A. Y. W. Chong, K. W. Khaw, W. C. Yeong, and W. X. Chuah, "Customer Churn Prediction of Telecom Company Using Machine Learning Algorithms," J. Soft Comput. Data Min., vol. 4, no. 2, pp. 1–22, 2023, doi: 10.30880/jscdm.2023.04.02.001.

[15] N. Suryana, P. Pratiwi, and R. T. Prasetio, "Penanganan Ketidakseimbangan Data pada Prediksi Customer Churn Menggunakan Kombinasi SMOTE dan Boosting," IJCIT (Indonesian J. Comput. Inf. Technol., vol. 6, no. 1, May 2021, doi: 10.31294/ijcit.v6i1.9545.

[16] Anis Fitri Nur Masruriyah, Hilda Yulia Novita, Cici Emilia Sukmawati, Siti Novianti Nuraini Arif, Angga Ramda Ramadhan, and P. Studi Informatika, "Evaluasi Algoritma Pembelajaran Terbimbing terhadap Dataset Penyakit Jantung yang telah Dilakukan Oversampling," J. MIND J. | ISSN, vol. 8, no. 2, pp. 242–253, 2023, [Online]. Available: https://doi.org/10.26760/mindjournal.v8i2.242-253

[17] F. E. P. Nadya, M. F. I. Ferdiansyah, V. R. S. Nastiti, and C. S. K. Aditya, "Implementation of Feature Selection Strategies to Enhance Classification Using XGBoost and Decision Tree," Sci. J. Informatics, vol. 11, no. 1, pp. 18–194, 2024, doi: 10.15294/sji.v11i1.48145.

[18] Yoga Religia, Agung Nugroho, and Wahyu Hadikristanto, "Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 1, pp. 187–192, 2021, doi: 10.29207/resti.v5i1.2813.

[19] B. Alnur, Mulyono, Fitri Amillia, and S. Sutoyo, "JITE (Journal of Informatics andTelecommunication Engineering)," J. Informatics Telecommun. Eng., vol. 7, no. 1, pp. 102–111, 2023, [Online]. Available: https://www.researchgate.net/publication/335117624_Malang_City_Polytechnic_Web_Based_Student_Attendance_Information_System_Telecommunications_Engineering_Study_Program_Using_Fingerprint/fulltext/5d515fe34585153e594ef214/Malang-City-Polytechnic-Web-Based-S

[20] M. M. S. Nurhidayat and Dyah Anggraini, "Analysis and Classification of Customer Churn Using Machine Learning Models," J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 7, no. 6, pp. 1253–1259, Nov. 2023, doi: 10.29207/resti.v7i6.4933.