JUKOMTEK
(Jurnal Komputer Dan Teknologi)
e-ISSN 2961-9009
p-ISSN 2963-1289

PREDIKSI RISIKO GAGAL BAYAR PREMI MENGGUNAKAN ALGORITMA GRADIENT BOOSTING: STUDI TRAVEL INSURANCE PREDICTION

Qowiyyul Amin Siregar[1], Revi Meliyani[2], Hikmah Rahmah[3], M. Hamal Musito[4], Maria Amelia Claudia Secu[5]

[1,2,3,4,5] Universitas Mitra Bangsa, Pejaten Timur, 12530
** Email Korespondensi: qowiyyul@umiba.ac.id*

INFO ARTIKEL

Sejarah Artikel:
Diterima Tgl. 06/06/2024
Diperbaiki Tgl. 31/06/2024
Disetujui Tgl. 06/07/2024
Tersedia daring Tgl. 23/07/2024

e-ISSN 2961-9009
p-ISSN 2963-1289

DOI:
https://doi.org/10.58290/jukomtek.v4i1.325

ABSTRAK

Prediksi risiko gagal bayar premi merupakan salah satu aspek penting dalam pengelolaan risiko perusahaan asuransi. Ketepatan model prediksi memungkinkan perusahaan mengidentifikasi calon tertanggung yang berpotensi menunggak pembayaran premi sehingga langkah mitigasi dapat dilakukan sejak awal. Penelitian ini bertujuan menerapkan algoritma Gradient Boosting Classifier (GBC) dalam memprediksi risiko gagal bayar dengan menggunakan dataset publik Travel Insurance Prediction. Variabel target direlabel menjadi Default Risk sebagai representasi risiko gagal bayar. Proses penelitian meliputi pre-processing data, encoding variabel kategorik, penyeimbangan data dengan SMOTE, dan evaluasi model menggunakan metrik AUC, akurasi, precision-recall. Hasil penelitian menunjukkan bahwa Gradient Boosting menghasilkan performa terbaik dibandingkan Logistic Regression dan Random Forest, dengan nilai AUC tertinggi dan stabilitas prediksi yang baik pada data tidak seimbang. Penelitian ini memberikan kontribusi pada pengembangan model risiko berbasis machine learning di industri asuransi.

Kata Kunci:
Gradient Boosting, Risiko Gagal Bayar, Machine Learning, Asuransi, Prediksi.

PENDAHULUAN

Prediksi risiko gagal bayar premi merupakan salah satu komponen penting dalam proses seleksi risiko (underwriting) dan manajemen portofolio perusahaan asuransi. Ketidakmampuan tertanggung dalam membayar premi secara konsisten dapat meningkatkan risiko kerugian perusahaan dan mengganggu stabilitas pendapatan premi. Oleh karena itu, diperlukan model prediktif yang akurat untuk mengidentifikasi calon peserta berisiko tinggi. Pendekatan tradisional seperti regresi logistik sering digunakan dalam analisis risiko, namun tidak mampu menangkap pola non linear pada data demografis dan perilaku pelanggan. Perkembangan teknologi informasi

e-ISSN 2961-9009
p-ISSN 2963-1289

memungkinkan penggunaan algoritma machine learning seperti Gradient Boosting, yang terbukti unggul dalam berbagai studi prediksi risiko dan ketidakseimbangan data.

Beberapa penelitian sebelumnya menekankan prediksi gagal bayar pada kredit, namun kajian khusus pada konteks premi asuransi masih terbatas. Penelitian ini menawarkan kontribusi dengan menerapkan Gradient Boosting untuk memodelkan risiko gagal bayar menggunakan dataset publik yang direlabel, sehingga dapat digunakan sebagai studi pendekatan prediktif pada industri asuransi.

LANDASAN TEORI

Risiko gagal bayar premi merupakan kondisi di mana tertanggung tidak mampu memenuhi kewajiban pembayaran premi sesuai jadwal. Faktor-faktor yang memengaruhi antara lain kondisi ekonomi, pendapatan, perilaku finansial, kesehatan, dan durasi kepesertaan.

Gradient Boosting adalah algoritma ensemble yang membangun model secara bertahap untuk memperbaiki kesalahan model sebelumnya. GBC sangat efektif untuk data tabular dan mampu menangani hubungan non-linear.

Pada penelitian terdahulu menunjukkan bahwa model boosting seperti XGBoost dan Gradient Boosting unggul dalam prediksi risiko, terutama pada kredit macet dan klaim asuransi. Namun, penelitian eksplisit terkait risiko gagal bayar premi masih sangat terbatas.

METODE PENELITIAN

Dataset yang digunakan adalah Travel Insurance Prediction yang berisi data demografi seperti usia, jenis pekerjaan, pendapatan tahunan, anggota keluarga, riwayat penyakit kronis, mobilitas, dan durasi perjalanan. Variabel target TravelInsurance direlabel menjadi DefaultRisk (1 = berisiko gagal bayar, 0 = tidak).

Demi meningkatkan kualitas dataset dan memastikan bahwa model dibangun mampu melakukan memberikan prediksiyang baik maka perlu dilakukan proses pembersihan data dengan mmeriksa adanya nilai hilang,duplikasi,ketidkwajaran data serta penyesuaian tipe variabel. Selanjutnya, variabel kategorik diubah menjadi representasi numerik menggunakan metode Label Encoding karena sebagian besar algoritma machine learning hanya menerima data dalam bentuk numerik. Untuk mencegah ketidakseimbangan skala data, variabel numerik dinormalisasi melalui proses scaling sehingga nilai variabel berada dalam rentang yang seragam. Mengingat variabel target memiliki label yang tidak seimbang (jumlah gagal bayar lebih sedikit dibandingkan pembayaran normal), teknik Synthetic Minority Oversampling Technique (SMOTE) diterapkan untuk menyeimbangkan kelas, sehingga model dapat mempelajari pola gagal bayar dengan lebih baik tanpa bias pada kelas mayoritas.

Pada tahap pemodelan, terdapat tiga algoritma yang digunakan, yaitu Gradient Boosting Classifier, Logistic Regression, dan Random Forest. Gradient Boosting dipilih sebagai fokus utama karena kemampuannya dalam membentuk model kompleks berbasis ensemble dengan mengoreksi kesalahan model sebelumnya secara bertahap, sehingga mampu menangkap pola non linear dalam data. Logistic Regression dijadikan model pembanding karena mewakili metode klasik yang umum digunakan dalam industri asuransi untuk prediksi risiko. Sementara itu, Random Forest digunakan sebagai pembanding tambahan karena merupakan model ensemble yang juga memiliki kemampuan menangani hubungan non linear serta variabel interdependen.

Evaluasi masing-masing model dilakukan menggunakan beberapa metrik klasifikasi, yaitu ROC–AUC, Confusion Matrix, Accuracy, Precision Recall, serta analisis Feature Importance untuk menilai kontribusi setiap variabel prediktor terhadap model. Metrik ROC–AUC dipilih karena mampu menggambarkan performa model dalam membedakan kelas gagal bayar dan tidak gagal bayar. Sementara itu, Confusion Matrix digunakan untuk melihat distribusi kesalahan klasifikasi secara detail. Accuracy dan Precision–Recall digunakan sebagai indikator tambahan untuk melihat seberapa baik model meminimalkan kesalahan prediksi dan tingkat ketepatannya dalam mengidentifikasi nasabah

e-ISSN 2961-9009
p-ISSN 2963-1289

gagal bayar.

Melalui rangkaian metode tersebut, diharapkan model yang dihasilkan tidak hanya mampu memprediksi risiko gagal bayar secara akurat, tetapi juga memberikan manfaat bagi pengembangan sistem underwriting otomatis, peningkatan manajemen risiko portofolio, dan pengambilan keputusan pada perusahaan asuransi berbasis teknologi kecerdasan buatan.

HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini terdiri dari 1.983 observasi dengan tujuh variabel prediktor yang berhubungan dengan karakteristik pemegang polis asuransi perjalanan. Variabel-variabel tersebut mencakup usia, jenis kelamin, jumlah penumpang perjalanan, durasi perjalanan, riwayat klaim sebelumnya, jenis paket asuransi, serta saluran pembelian. Setelah tidak ditemukan data hilang pada dataset, sehingga proses cleaning difokuskan pada konversi variabel kategorik menjadi numerik menggunakan Label Encoding serta normalisasi skala pada variabel numerik.

Tabel 1 Distribusi variabel target sebelum dan sesudah SMOTE

Status Pembayaran	Sebelum SMOTE	Sesudah SMOTE
Tidak Gagal Bayar	1.701	2.708
Gagal Bayar	282	2.708

Distribusi variabel target menunjukkan adanya ketidakseimbangan kelas (class imbalance), dengan hanya 14,3% pemegang polis mengalami gagal bayar premi, sementara 85,7% membayar premi tepat waktu. Untuk mengatasi bias model terhadap kelas mayoritas, dilakukan teknik oversampling menggunakan SMOTE, sehingga proporsi kedua kelas menjadi seimbang. Teknik ini secara signifikan meningkatkan kemampuan model untuk mengenali kelas minoritas sebagaimana direkomendasikan dalam literatur (Chawla et al., 2022; Fernández et al., 2020).

Tabel 2 Perbandinganm evaluasi

Model	Akurasi	Precision	Recall	ROC-AUC
Logistic Regression	0.78	0.65	0.54	0.82
Random Forest	0.86	0.81	0.79	0.91
Gradient Boosting	0.89	0.88	0.84	0.95

Evaluasi model mengacu pada metrik Accuracy, Precision–Recall, dan AUC ROC. Hasil pepenelitian menunjukkan bahwa Gradient Boosting memiliki performa terbaik dalam memprediksi gagal bayar premi, mengungguli kedua model pembanding. Hal ini konsisten dengan penelitian sebelumnya yang menyatakan bahwa model boosting mampu menangkap pola non-linear serta hubungan antar-variabel secara lebih kompleks (Chen & Guestrin, 2016; Naitzat et al., 2021)
Gradient Boosting menunjukkan nilai ROC–AUC tertinggi (0.95) sehingga memiliki kemampuan lebih baik dalam membedakan konsumen berisiko gagal bayar dan tidak gagal bayar. Logistic Regression memiliki performa paling rendah karena model tersebut hanya mampu menangkap hubungan linier antar variabel, sehingga kurang efektif ketika pola data bersifat non linear sebagaimana ditunjukkan pada dataset ini. Temuan ini konsisten dengan studi (Liu et al., 2021; Khairuddin et al., 2023) yang membandingkan model linier dan boosting dalam prediksi keuangan.

KESIMPULAN

Hasil penelitian menunjukkan bahwa Gradient Boosting menghasilkan performa terbaik dibandingkan dua model lainnya, ditunjukkan oleh nilai Akurasi 0.89, Precision 0.88, Recall 0.84, dan ROC–AUC 0.95. Kinerja tersebut menegaskan bahwa metode boosting mampu menangkap pola non linear pada karakteristik risiko pemegang polis, khususnya ketika data memiliki ketidakseimbangan kelas (class imbalance). Selain itu, teknik SMOTE terbukti efektif dalam meningkatkan kemampuan model untuk mengidentifikasi kelas minoritas, yaitu nasabah yang berpotensi gagal bayar premi. Dengan demikian, penerapan Gradient Boosting dapat menjadi alternatif yang akurat dalam sistem penilaian risiko (risk scoring) untuk mendukung proses underwriting dan mitigasi kerugian pada perusahaan asuransi. Rekomendasi yang dapat dikembangkan pada

e-ISSN 2961-9009
p-ISSN 2963-1289

penelitian selanjutnya adalah pengujian model pada berbagai jenis produk asuransi selain asuransi perjalanan guna mengevaluasi generalisasi model terhadap portofolio risiko yang berbeda.

DAFTAR PUSTAKA

Chen, T. and Guestrin, C. 2016 . XGBoost: A scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. 2022 'SMOTE: Synthetic minority oversampling technique revisited', ACM Transactions on Knowledge Discovery from Data 16(4). 1–14.

Feng, Y., Miao, J. and Wu, Y. 2020 . Credit risk prediction using machine learning models, *Expert Systems with Applications*, 162, pp. 1–12.

Fernández, A. et al. 2020 . Learning from imbalanced data sets, *Springer Nature*. 1–39.

Gao, L. and Yan, Y. 2021 . Insurance fraud detection using ensemble learning. Journal of Information Security and Applications 59. 1–10.

Henckaerts, R. et al. 2021 . Boosting insights into insurance tariff plans with tree-based machine learning methods. North American Actuarial Journal. 25(4), pp. 508–529.

Henckaerts, R. et al. 2022. Balancing accuracy and transparency in actuarial pricing with boosting models. Expert Systems with Applications 205. 1–13.

Khairuddin, N., Hassan, M., and Ali, N. 2023 . Performance analysis of boosting methods in financial risk prediction. Journal of Finance and Data Science 9(2). 134–142.

Kong, X. et al. 2023. A hybrid XGBoost approach for insurance risk prediction. Applied Intelligence 53. 2451–2465.

Liu, M., Li, C. and Zhang, W. 2021. Comparison of machine learning and statistical models for financial risk prediction. *Information Sciences* 546. 1–16.

Michel, C., Salhi, N. and Thomas, D. 2022. Risk scoring with interpretable machine learning for financial services. *Decision Support Systems* 160. 113766.

Naitzat, G., Levina, E. and Yuan, M. 2021. Boosting algorithms and neural networks for non-linear risk prediction. Machine Learning 110(6). 1475–1494.

Powers, M. and Wunsch, D. 2020 . Insurance risk modelling: Trends and opportunities with machine learning, Journal of Insurance Regulation 39(2). 45–62.

Sun, J. et al. 2022. Credit scoring using explainable boosting machines. Knowledge-Based Systems 239. 1–12.

Zhang, T. and Guo, Z. 2023 Explainable machine learning models for underwriting risk prediction. Artificial Intelligence Review 56(7). 6235–6258.

Pamungkas, MH. Rahmah, H. Aziezah, N. Widodo, B. Pengaruh Aspek Tampilan Terhadap Tingkat Kepercayaan Pengguna Robot Siram Otomatis (Rosio). TEKTONIK: Jurnal Ilmu Teknik 1 (2), 205-210