Jurnal Algoritma, Logika dan Komputasi *Vol.VII (No. 01) : 634 - 642. Th. 2024*
Versi Online: https://journal.ubm.ac.id/index.php/alu *p-ISSN: 2620-620X*
DOI: http://dx.doi.org/10.30813/j-alu.v2i2.6030 *e-ISSN: 2621-9840*

GRADIENT BOOSTING TREES UNTUK PEMODELAN DAN PREDIKSI BIAYA KERUGIAN ASURANSI MOBIL

Gradient Boosting Trees for Auto Insurance Loss Cost Modeling and Prediction

Eric Fammaldo, ericfammaldo@gmail.com[1)], Merryana Lestari, mlestari@bundamulia.ac.id[2)*], Chandra Hermawan, l0744@lecturer.ubm.ac.id[3)]
[1),3)]Program Studi Informatika/Fakultas Teknologi dan Desain, Universitas Bunda Mulia, Jakarta
[2)*]Program Studi Sistem Informasi/Fakultas Teknologi dan Desain, Universitas Bunda Mulia, Jakarta

Diterima 26 Juli 2024 / Disetujui 31 Juli 2024

ABSTRACT

Gradient Boosting is a machine learning algorithm that combines several simple parameter functions that aim to predict a fairly accurate information from existing data. In contrast to statistical methods in general, this Gradient boosting provides interpretable information, while requiring little data preprocessing and tuning of parameters. Boosting Gradient can be applied to classify or regress data, complex interaction is modeled simply and minimizes loss of information while in predictor management, so this algorithm is good enough to be used for modeling the cost of insurance loss. This paper presents the GB theory and its application to the problem of predicting " at-fault " accidents on auto loss costs using data from Canadian insurance companies. The predictive accuracy of the model is compared to the conventional Generalized Linear Model (GLM) approach.
***Keywords**: Gradient Boosting, Generalized Linear Model, Cost of insurance loss*

ABSTRAK

Gradient Boosting (GB) adalah sebuah algoritma *machine learning* yang menggabungkan beberapa fungsi parameter sederhana yang bertujuan untuk memprediksi sebuah informasi yang cukup akurat dari data-data yang ada. Berbeda dengan metode statistika pada umumnya, *Gradient boosting* ini memberikan informasi yang dapat diinterpretasi, sementara membutuhkan sedikit data *preprocessing* dan *tuning* dari *parameter*. *Gradient Boosting* dapat diterapkan untuk melakukan klasifikasi maupun regresi pada data-data, Interaksi kompleks dimodelkan secara sederhana dan meminimalisir kehilangan informasi saat dalam pengelolaan prediktor, sehingga algoritma ini cukup baik digunakan untuk pemodelan biaya asuransi kerugian. Penelitian ini menyajikan teori GB dan aplikasinya untuk masalah memprediksi kecelakaan *''at-fault''* pada biaya kerugian mobil menggunakan data dari perusahaan asuransi Kanada. Akurasi prediksi model GB dibandingkan terhadap pendekatan *Generalized Linear Model* (GLM) konvensional.
Kata Kunci: Gradient Boosting, Generalized Linear Model, Biaya asuransi kerugian

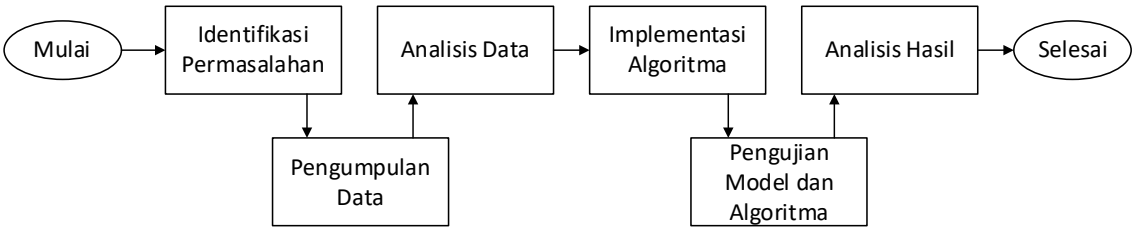
PENDAHULUAN

Generalized Linear Models (GLMs) merupakan sebuah model yang banyak digunakan dalam penetapan harga asuransi, model ini didasarkan pada pendekatan konvensional pemodelan statistik yang dimulai dengan mengasumsikan bahwa data dihasilkan dengan model statistik bertentu [1]. Dari asumsi tersebut dihasilkan parameter yang dapat diinterpretasikan dan dikombinasikan dengan cara multiplikatif untuk mendapatkan perkiraan biaya kerugian. Banyak percobaan pada beberapa dekade terakhir ini yang menciptakan beberapa pengembangan dalam mengolah data, namun berbeda dengan pemodelan data, model algoritma tidak menganggap beberapa model spesifik sebagai data, namun sebagai data yang tidak dikenal [2][3]. Sehingga mereka lebih efisien dalam menangani data yang besar dan komplek serta data yang bersifat *non-*

*Korespondensi Penulis:
E-mail: **mlestari@bundamulia.ac.id**

linear. Banyak pengembangan yang dilakukan demi mengembangkan GLM seperti Regresi Poisson, Regresi Gamma dan Regresi Logistik [4]. Dalam pengunaan *Gradient Boost* untuk saat ini masih belum di dokumentasi dalam penentuan harga asuransi. Tujuan dari penelitian ini adalah untuk mempresentasikan teori *Gradient Boost* dan pengunaannya dalam menganalisa pemodelan biaya kerugian dengan mengunakan data dari perusahaan asuransi Kanada dan mejelaskan detail mengenai *Gradient Boost* dari pembelajaran statistik perspektif, serta menjelaskan pengaplikasinya dalam analisis asuransi kerugian kecelakaan *"at-fault"*.

METODOLOGI PENELITIAN



Gambar 1. Tahapan Penelitian

Tahapan penelitian ini dapat dilihat pada Gambar 1, dimulai dengan identifikasi permasalahan yang akan menjadi objek penelitian, kemudian dilakukan proses pengumpulan data, data yang menjadi objek penelitian adalah data dari perusahaan asuransi dan didalam proses analisis data dilakukan data *preprocessing* sebelumnya supaya memastikan bahwa data yang digunakan dalam penelitian sesuai [5]. Setelah dilakukan proses analisis data, dilanjutkan pada proses implementasi algoritma sehingga dapat dilakukan pengujian model dan algoritma yang paling sesuai, kemudian dilakukan proses analisis hasil untuk didapatkan kesimpulan dari penelitian ini.

Predictive Learning and Boosting

Masalah *predictive learning* dapat dicirikan dengan vektor input atau variabel prediktor x = {x1, ..., xp} dan output atau target variabel y. Dalam aplikasi penelitian ini, variabel input diwakili oleh kumpulan atribut kuantitatif dan kualitatif dari kendaraan dan diasuransikan, dan output adalah biaya kerugian yang sebenarnya. Diberikan sebuah koleksi M {(yi, xi); i = 1, ..., M} dari nilai-nilai yang dikenal (y, x), tujuan menggunakan data ini untuk mendapatkan dan memperkirakan fungsi yang memetakan vektor input x ke dalam nilai-nilai dari output y. Fungsi ini kemudian dapat digunakan untuk membuat prediksi pada *instance* di mana hanya nilai x yang diamati. Secara formal, penelitian ini mempelajari fungsi prediksi yang meminimalkan harapan dari beberapa fungsi kerugian L (y, f) atas distribusi bersama dari semua nilai (y, x) dapat dilihat pada rumus (1).

$$\hat{f}(\mathbf{x}) \quad \text{argmin}_{f(\mathbf{x})} E_{y,\mathbf{x}} L(y, f(\mathbf{x})) \tag{1}$$

Metode *Boosting* adalah metode yang berdasarkan *intuitive* yang mengabungkan banyak aturan *"weak"* yang menghasilkan model klasifikasi dan regresi dengan pengembangan prediksi dalam performanya [6]. Pengabungan dari banyak aturan tersebut dapat menciptakan sebuah pemodelan yang akurat, Ide ini dikenal sebagai *"the Strength of weak learnability"* [7]. Banyak metode *boosting* yang ada seperti *AdaBoost* yang populer dikarenakan Freund dan Schapire (1996) merupakan salah metode yang menerapkan prinsip pengabungan weak rules, metode memang bagus ini namun memiliki keterbatasan (*unOverfit*) sehingga tidak semua analisis data dapat mengunakan metode ini, sedangkan *Gradient Boost* lebih bersifat meningkatkan performa, sehingga *Gradient boost* ini bisa digunakan pada *AdaBoost* atau pemodelan lainnya (*OverFit*) [8].

Additive Model and Boosting

Penelitian ini akan berfokus pada masalah regresi, dimana y kuantitatif dan tujuannya adalah untuk mendapatkan estimasi rata-rata E(y|x)-f(x). Pada umumnya, linear regresi model sebuah bentuk linear dapat dilihat pada rumus (2).

$$E(y|\mathbf{x}) \quad f(\mathbf{x}) \quad \sum_{j-1}^{p} \beta_j x_j \tag{2}$$

Dengan adanya penambahan model *additive*, akan adanya perubahan komponen pada rumus bentuk linear menjadi rumus (3).

$$E(y|\mathbf{x}) \quad f(\mathbf{x}) \quad \sum_{j-1}^{p} f_j(x_j), \tag{3}$$

Model ini bisa di kembangkan lagi dengan melakukan pertimbangan pada model *additive* dengan fungsi ft(x),t-{1,...,T} dari kemungkinan semua input *variable* dapat dilihat pada rumus (4).

$$f(\mathbf{x}) \quad \sum_{t-1}^{T} f_t(\mathbf{x}) \quad \sum_{t-1}^{T} \beta_t h(\mathbf{x}; \mathbf{a}_t), \tag{4}$$

Pada konteks *Boosting*, βth(x;at) mewakili *weak leaner* dan f(x) mewakili pertimbangan mayoritas suara individu *weak learners* [9]. Sehingga bisa di estimasi *parameter* dalam pemecahannya yang dapat dilihat pada rumus (5).

$$\min_{\{\beta_t, a_t\}_1^T} \sum_{i-1}^{M} L\left(y_i, \sum_{t-1}^{T} \beta_t h(\mathbf{x}_i; \mathbf{a}_t)\right), \tag{5}$$

Dengan menggunakan Algoritma *Forward Stagewise Additive Modeling* sebagai berikut, diperoleh alur *pseudocode* seperti pada Gambar 2.

1: Initialize $f_0(\mathbf{x}) = 0$
2: **for** $t = 1$ to T **do**
3: Obtain estimates β_t and a_t by minimizing $\sum_{i-1}^{M} L(y_i, f_{t\ 1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a}))$
4: Update $f_t(\mathbf{x}) = f_{t-1}(\mathbf{x}) + \beta_t h(\mathbf{x}; \mathbf{a}_t)$
5: **end for**
6: Output $\hat{f}(\mathbf{x}) = f_T(\mathbf{x})$

Gambar 2. *Pseudocode* Algoritma *Forward Stagewise Additive Modeling*

Jika *squared-error* digunakan sebagai fungsi kerugiannya maka pada baris ke 3 (tiga) akan berubah menjadi rumus (6)

$$L(y_i, f_{t\ 1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a})) - (y_i - f_{t\ 1}(\mathbf{x}_i) - \beta h(\mathbf{x}_i; \mathbf{a}))^2 - (r_{it} - \beta h(\mathbf{x}_i; \mathbf{a}))^2, \tag{6}$$

Gradient Boost

Squared-error dan *exponential error* adalah fungsi kerugian yang masuk akal yang umumnya digunakan untuk masalah regresi dan klasifikasi [10]. Namun, mungkin ada situasi di mana fungsi kerugian lainnya lebih tepat. Misalnya, penyimpangan binomial masih jauh lebih kuat daripada kehilangan eksponensial dalam pengaturan yang berisik di mana tingkat kesalahan Bayes tidak mendekati nol, atau dalam situasi di mana kelas target salah diberi label. Demikian pula, kinerja kesalahan-kuadrat secara signifikan terdegradasi untuk distribusi kesalahan berekor panjang atau kehadiran *"outliers"* dalam data [11]. Dalam situasi seperti itu, fungsi lain seperti kesalahan absolut atau Huber loss lebih tepat.

1: Initialize $f_0(\mathbf{x})$ to be a constant, $f_0(\mathbf{x}) = \text{argmin}_\beta \sum_{i-1}^{M} L(y_i, \beta)$
2: **for** $t = 1$ to T **do**
3: Compute the negative gradient as the working response

$$r_i - -\left[\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)}\right]_{f(\mathbf{x})-f_{t\ 1}(\mathbf{x})}, i - \{1, \ldots, M\}$$

4: Fit a regression model to r_i by least-squares using the input $\mathbf{x}_i$ and get the estimate $\mathbf{a}_t$ of $\beta h(\mathbf{x}; \mathbf{a})$
5: Get the estimate β_t by minimizing $L(y_i, f_{t-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a}_t))$
6: Update $f_t(\mathbf{x}) = f_{t-1}(\mathbf{x}) + \beta_t h(\mathbf{x}; \mathbf{a}_t)$
7: **end for**
8: Output $\hat{f}(\mathbf{x}) = f_T(\mathbf{x})$

Gambar 3. *Pseudocode* Algoritma *Gradient Boosting*

Di bawah spesifikasi alternatif untuk fungsi kerugian dan untuk *weak learner* tertentu, solusi untuk baris ke 3 dalam Algoritma *Additive Modeling* sulit diperoleh. Algoritma peningkatan *gradient* memecahkan masalah menggunakan prosedur dua langkah yang dapat diterapkan pada fungsi kerugian terdiferensiasi. Langkah pertama yaitu memperkirakan dengan menyesuaikan *weak learner* h (x; a) ke *gradient* negatif dari fungsi kerugian yaitu *"pseudo-residuals"* menggunakan kuadrat terkecil. Pada langkah kedua, nilai βt optimal ditentukan h (x; at). Prosedur ditunjukan dalam Algoritma *Gradient Boosting* pada Gambar 3. Untuk masalah *squared-error loss*, pada baris ke 3 di algoritma diatas, negative berfungsi untuk mengurangi standar *least-squares boosting*. Dengan hilangnya kesalahan absolut, *gradient* negatif adalah tanda residual. *Least-squares* digunakan pada baris ke 4 terlepas dari fungsi kerugian yang dipilih.

Injecting randomness and regularization

Dalam mengunakan algoritma pada *Gradient Boosting* untuk mencegah *"overfitting"* digunakanlah metode *regularization* dimana bertujuan untuk membatasi parameter dalam *Gradient Boosting* guna mengontrol sejumlah iterasi yang akan terjadi saat dalam proses, sehingga pada baris ke 6 di algoritma *Gradient Boosting* terjadi perubahan menjadi Rumus (7).

$$f_t(\mathbf{x}) - f_{t\ 1}(\mathbf{x}) + \tau \cdot \beta_t h(\mathbf{x}; \mathbf{a}_t) \tag{7}$$

Perubahan kedua adalah pada pengunaan metode *randomness* pada prosedur yang bertujuan untuk mengurangi permintaan komputasi pada baris ke 4 di algoritma *Gradient* untuk melakukan penyesuaian pada data *weak learner* sehingga variasi pada *weak learner* pada setiap iterasi akan meningkat, tetapi korelasi antara estimasi pada iterasi yang berbeda akan menurun.

HASIL DAN PEMBAHASAN

Data yang digunakan dalam penelitian ini adalah data dari asuransi di Kanada, yang didalamnya tercantum peraturan dan informasi di tingkat kendaraan individu dan pada tiap beberapa waktu akan dilakukan pengecekan mengenai masalah *"at-fault"*. Data ini termasuk 426.838 eksposur yang diterima dan dilihat dari kendaraannya pertahun dari bulan januari 2006 sampai bulan juni 2009, dan tercatat sebanyak 14,984 kali yang terjadi pada periode yang sama, dengan kerugian berdasarkan estimasi cadangan terbaik pada bulan desember 2009. Variable yang diinput diukur dari mulainya masa eksposur dan diwakili dengan kualitas serta kuantitas pada kendaraan. Output yang keluar adalah biaya kerugian yang di hitung berdasarkan rasio dari total kerugian pada eksposur.

Pada percobaan ini, 70% data *training* digunakan untuk penyeleksian dan pelatihan pemodelan, sedangkan 30% data uji digunakan untuk pengujian pada memprediksi ketepatan *gradient boost* dengan *generalized* linear model dimana digunakan perbandingan pada keduanya untuk menentukan biaya kerugian. Biaya kerugian biasanya dibagi menjadi 2, frekuensi pengambilan yang dihitung berdasarkan rasio pengambilan dalam mendapatkan eksposur dan pengambilan berdasarkan tingkat keparahan yaitu perhitungannya bedasarkan rasio total kerugian dalam pengambilannya. Tabel 1 menunjukkan varible input pada data yang dibutuhkan.

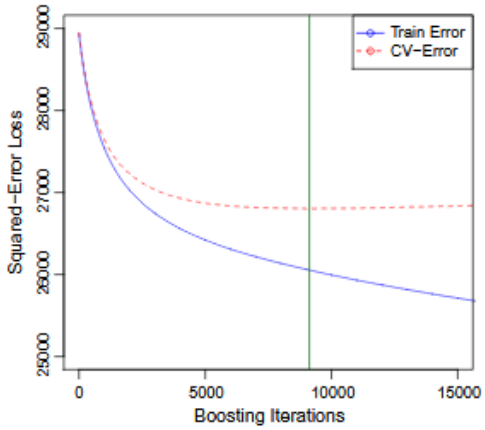
Tabel 1. Variabel Input

Overview of loss cost predictors.

Driver characteristics	Accident/conviction history	Policy characteristics	Vehicle characteristics
DC1. Age of principal operator	AC1. Number of chargeable accidents (last 1–3 years)	PC1. Years since policy inception	VC1. Vehicle make
DC2. Years licensed	AC2. Number of chargeable accidents (last 4–6 years)	PC2. Presence of multi-vehicle	VC2. Vehicle purchased new or used
DC3. Age licensed	AC3. Number of non-chargeable accidents (last 1–3 years)	PC3. Collision deductible	VC3. Vehicle leased
DC4. License class	AC4. Number of non-chargeable accidents (last 4–6 years)	PC4. Billing type	VC4. Horse power to weight ratio
DC5. Gender	AC5. Number of driving convictions (last 1–3 years)	PC5. Billing status	VC5. Vehicle age
DC6. Marital status	AC6. Prior examination costs from accident-benefit claims	PC6. Rating territory	VC6. Vehicle price
DC7. Prior facility association		PC7. Presence of occasional driver under 25	
DC8. Postal code risk score		PC8. Presence of occasional driver over 25	
DC9. Insurance lapses		PC9. Group business	
DC10. Insurance suspensions		PC10. Business origin	
		PC11. Dwelling unit type	

Pembangunan Model

Pertama yang dilakukan untuk pemodelan adalah memilih sebuah *loss function* yang tepat L(y,f(x)) atau bisa dibilang *weak leaners*. Pengunaan *square-error loss* dan *bernoulli deviance* digunakan untuk mendefinisi *error* pada prediksi untuk tingkat keseringan atau *frequency models* dan tingkat keparahan atau *severity models*, serta butuh melakukan pemilihan dalam tiap pohon dan sub samplenya. Yang pertama ditetapkan sebagai nilai tetap 0.001 dan yang setelahnya 50%. Kemudian ukuran pohon individu S dan jumlah iterasi *boosting* T butuh diseleksi. Ukuran tiap pohon diseleksi dengan meningkatkan kedalaman interaksi pohon secara berurutan, dimulai dengan sebuah *additive* model dan diikuti dengan *two-ways interactions* dan *six-ways interaction*. Ini dilakukan secara bergantian untuk *frequency* dan *severity models*. Setiap model tersebut diuji cobakan sebanyak 20.000 iterasi *boosting* mengunakan data *training*. Seperti yang ditunjukkan pada Gambar 4.

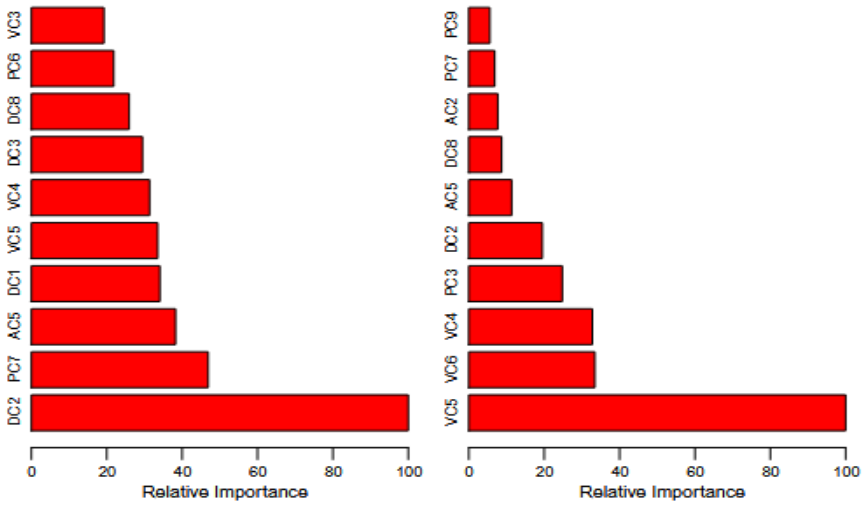


Gambar 4. Iterasi *boosting*

Dari data grafik pada gambar 4 menunjukan relasi antara *Train Error* dan *Cross Validation error* dan garis biru menunjukan jumlah iterasi *boosting* yang optimal.

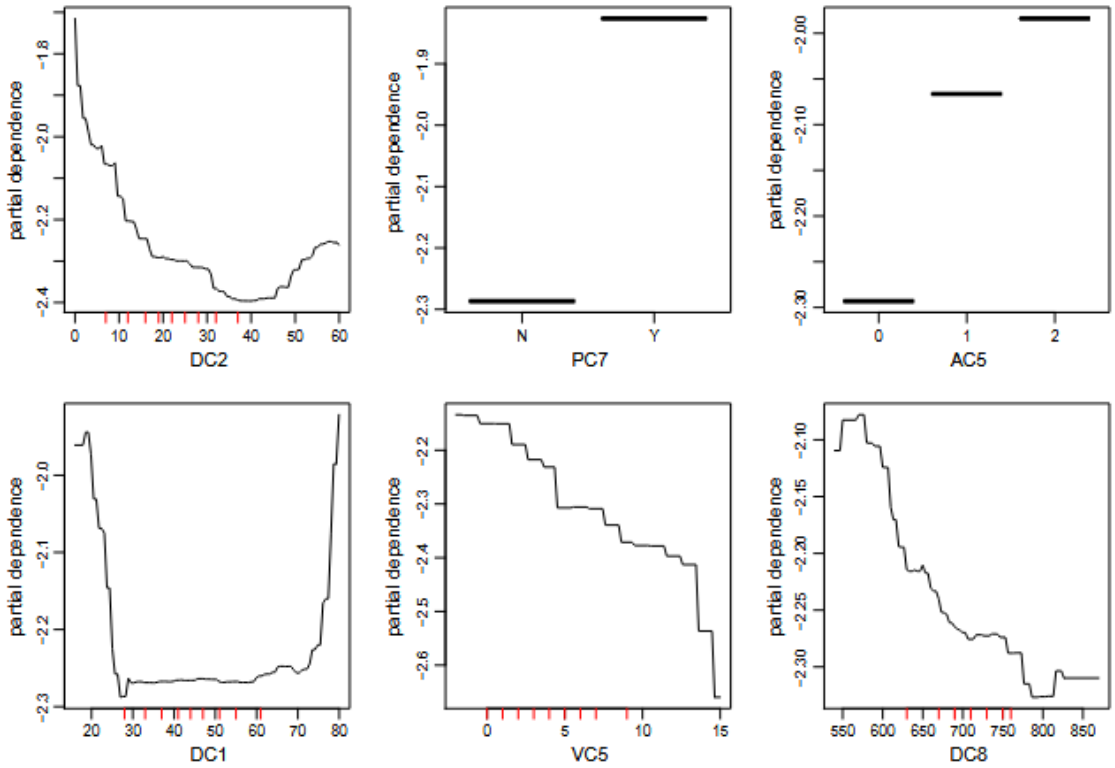
Analisis Hasil

Pada grafik *relative Importance* pada Gambar 5 menunjukan 10 (sepuluh) data penting yang dibutuhkan dalam varible prediksi yang disebelah kiri untuk model *frequency* dan sebelah kanan untuk model *severity*. Karena pengukuran ini relatif, maka 100 data yang digunakan untuk prediksi yang penting dan sisanya mengikuti. Pada kedua model tersebut sangat jelas tampak perbedaannya, pertama pada jumlah tahun lisensi prinsip operator pada kendaraan pada *frequency* model lebih relevan prediksinya berbeda dengan *severity*, selain itu bisa dilihat juga usia kendaraan pada *severity* model lebih mendominasi dibanding *frequency*, begitu juga data yang lainnya. Pada grafik *partial dependence* (DC2,PC7,AC5,DC1,VC5,DC8). Pada *frequency* model dibagian vertikal menunjukan *log odds* dan tanda *hash* menunjukan desil distibusi *variable* yang sesuai. Pada *frequency* mempunyai *partial dependence* yang *non-monotonic* pada bagian tahun lisensi dan terus menurun dan naik saat diujung data. Data yang ada pada grafik tersebut juga ada yang mengalami penurunan dan kenaikan yang dapat dilihat pada Gambar 6.

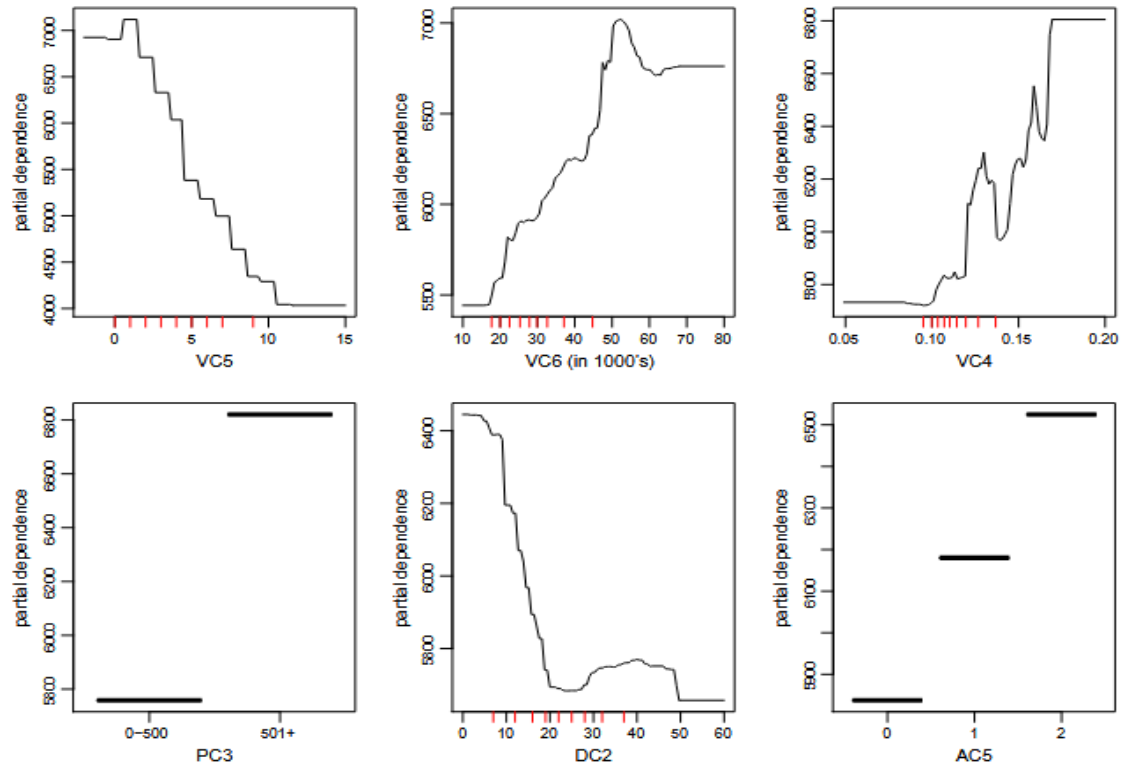


Gambar 5. *Relative Importance*

Vol.VII (No. 01) : 634 - 642. Th. 2024
p-ISSN: 2620-620X
e-ISSN: 2621-9840

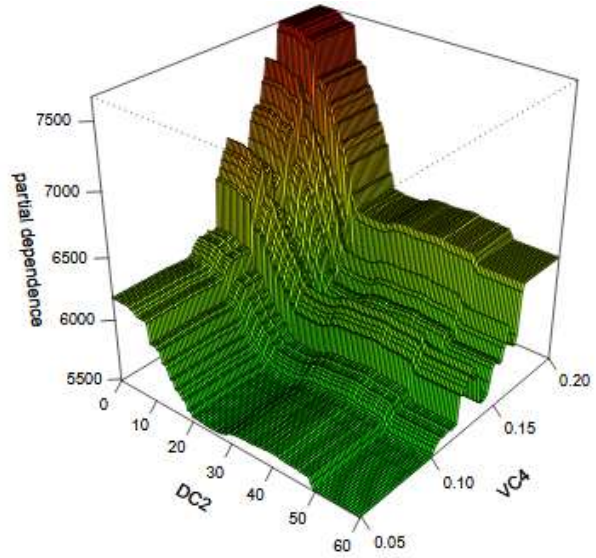


Gambar 6. ***Frequency Model***



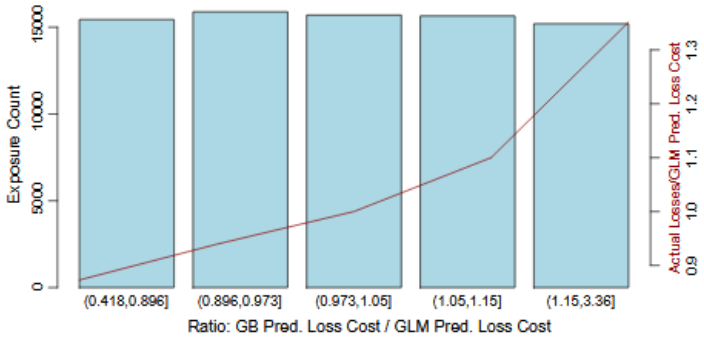
Gambar 7. ***Severity Model***

Pada Grafik di gambar 7 terdapat *severity models* (VC5, VC6, VC4, PC3, DC2, AC5) menunjukan bawah ketergantungan umur kendaraan dan harganya bergantung pada biaya perbaikian untuk mobil yang lebih mahal dan lebih baru. Bentuk kurva pada grafiknya cukup linear dari banyak data-data tersebut. Pada tiap grafiknya menunjukan ada variable ketergantungan yang mempengaruhi *severity* model.



Gambar 8. Grafik DC2-VC4

Pada Gambar 8, grafik (DC2, VC4) menunjukan bahwa ketergantungan antara tahun lisensi dan tenaga kuda ke rasio berat pada *severity* model. Dimana tingkat rasio berat tenaga kuda lebih tinggi dari nilai tahun lisensi. Selanjutnya perbandingan ketepatan prediksi antara *Generalized Linear Models* (GLM) dan *Gradient Boosting* dengan data sample dengan menghitung tingkat rasionya. Tercatat bahwa GLM rasio kerugian meningkat ketikan *Gradient Boost* model melakukan *charge* yang *relative* tinggi ke GLM. Kecenderungan ke atas dalam kurva rasio GLM-*loss* menunjukkan kinerja prediktif yang lebih tinggi dari *Gradient Boosting* relatif untuk GLM. Gambar 9 adalah grafik tingkat rasio perbandingannya.



Gambar 9. Tingkat Rasio

Pada penelitian ini dijelaskan mengenai *Gradient Boost* dan pengaplikasian analisisnya terhadap kasus yang dibahas, dimana *Gradient Boost* ini diwakili sebagai *additive* model yang secara berturut-turut disesuaikan sebagai *weak learner* ke residual saat ini dengan *least-squares*. Dan berdasarkan data yang sudah dicoba dengan pengunaan *Gradient Boosting* pada analisa data ini, menunjukan pendekatan *Gradient Boost* relatif lebih tinggi dibanding GLM. Ini tidak mengherankan karena GLM adalah, pada dasarnya, model linier yang relatif sederhana dan dengan demikian mereka dibatasi oleh kelas fungsi yang mereka dapat perkiraan. Kedua,

dibandingkan dengan metode pembelajaran statistik *non-linear* lainnya seperti jaringan saraf dan mesin pendukung vektor, GB memberikan hasil yang dapat diinterpretasikan melalui pengaruh relatif dari variabel *input* dan plot ketergantungan parsial mereka [12][13][14]. Ini adalah aspek penting untuk dipertimbangkan dalam lingkungan bisnis, di mana model biasanya harus disetujui oleh pembuat keputusan yang tidak memiliki statistik terlatih yang perlu memahami bagaimana output dari "kotak hitam" sedang diproduksi. Ketiga, GB membutuhkan sedikit *data preprocessing* yang merupakan salah satu kegiatan yang paling memakan waktu dalam proyek *data mining*. Terakhir, perlunya analisis pemilihan model dilakukan sebagai bagian integral dari prosedur GB. Singkatnya, *Gradient Boosting* adalah metode alternatif yang baik untuk *Generalized Linear Models* dalam membangun model biaya kerugian asuransi.

SIMPULAN

Berdasarkan hasil penelitian yang dilakukan dapat disimpulkan bahwa pengguna *Gradient Boosting* pada biaya kerugian dibanding dengan GLM menunjukan bahwa pengunaan *Gradient Boosting* bisa dijadikan alternatif dalam melakukan prediksi biaya kerugian karena lebih cepat dan akurasi/ketepatannya pun tidak kalah dengan mengunakan metode konvensional GLM.

DAFTAR PUSTAKA

[1] McCullagh, P., & Nelder, J. (1989).*Generalized linear models(2nd ed.)*. Chapman and Hall.
[2] Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D. Schirmacher, E., & Thandi, N.(2007). *A practitioner's guide to generalized linear models*. Casualty Actuarial Society (CAS), Syllabus Year: 2010, Exam Number: 9, 1–116.
[3] Haberman, S., & Renshaw, A. (1996). *Generalized linear models and actuarial science*. Journal of the Royal Statistical Society, Series D, 45, 407–436.
[4] J. A. Ginting, "Data Mining Untuk Analisa Pengajuan Kredit Dengan Menggunakan Metode Logistik Regresi," *J. Algoritm. Log. dan Komputasi*, vol. 2, no. 2, pp. 164–169, 2019, doi: 10.30813/j-alu.v2i2.1845.
[5] B. Hakim, "Analisa Sentimen Data Text Preprocessing Pada Data Mining Dengan Menggunakan Machine Learning," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 2, pp. 16–22, 2021, doi: 10.30813/jbase.v4i2.3000.
[6] Brieman, L., Friedman, J., Olshen, R., & Stone, C. (1984).*Classification and regression trees*. CRC Press.
[7] Schapire, R. (1990). *The strength of weak learnability. Machine Learning*, 5, 197–227.
[8] Freund, Y., & Schapire, R. (1996). *Experiments with a new boosting algorithm*. Proceedings of the International Conference on Machine Learning, 13(pp. 148–156).
[9] Friedman, J., Hastie, T., & Tibshirani, R. (2000). *Additive logistic regression: A statistical view of boosting*.The Annals of Statistics, 28, 337–407.
[10] C. Herdian, A. Kamila, and I. G. Agung Musa Budidarma, "Studi Kasus Feature Engineering Untuk Data Teks: Perbandingan Label Encoding dan One-Hot Encoding Pada Metode Linear Regresi," *Technol. J. Ilm.*, vol. 15, no. 1, p. 93, 2024, doi: 10.31602/tji.v15i1.13457.
[11] Friedman, J. (2001). *Greedy function approximation: A gradient boosting machine*. The Annals of Statistics, 29, 1189–1232.
[12] Brieman, L. (2001). *Statistical modeling: The two cultures.Statistical Science*, 16,199–231.
[13] Francis, L. (2001). *Neural networks demystified*. Casualty Actuarial Society Forum, Winter, 2001, 252–319.
[14] Wikipedia. 2018. *Gradient Boosting*. Diakses pada 23 Mei 2018, dari https://en.wikipedia.org/wiki/Gradient_boosting