ISBN:
978-979-99168-1-5

PROSIDING
KNMSA 2015

Konferensi Nasional Matematika,
Sains dan Aplikasinya
Bandung, 26 Agustus 2015

Dipublikasikan oleh:
Fakultas Matematika & Ilmu Pengetahuan Alam
Universitas Islam Bandung

Daftar Isi

Prosiding KNMSA 2015
Fakultas MIPA Unisba, 26 Agustus 2015
ISBN: 978-979-99168-1-5

Model Credit Scoring Menggunakan Regresi Logistik Beserta Validasinya

Ade Irma Nurwahidah, Abdul Kudus, Suliadi

Program Studi Statistika Universitas Islam Bandung
e-mail: adeirmanurwahidah@gmail.com; akudusmilis@gmail.com; suliadi@gmail.com

Abstrak

Model credit scoring merupakan suatu alat dan teknik untuk meminimalkan risiko pada lembaga keuangan khususnya bank. Dari hasil model credit scoring akan didapat kartu skor hasil diskretisasi recursive partitioning dengan CART yaitu nilai untuk setiap kategori dari karakteristik calon debitur (variabel bebas) dengan menggunkan regresi logistik. Untuk mengukur seberapa baik model credit scoring dalam mengklasifikasikan nasabah yang baik dan buruk dan memperkuat ketetapan prediksi model maka dilakukan validasi terhadap model. Dalam skripsi ini, terdapat beberapa ukuran-ukuran untuk melakukan validasi model yaitu kurva Receiver Operating Characteristic (ROC), Statistik Kolmogorov-Smirnov (KS), Indeks Gini dan C-Statistik. Dalam mengaplikasikan penelitian ini, penulis menggunakan data kredit Jerman 1994. Setelah dilakukan analisis, bahwa model baik atau ideal sehingga model credit scoring dapat digunakan untuk penyeleksian nasabah yang akan menerima kredit.

Kata Kunci: Model Credit Scoring, Scorecard, Diskretisasi Recursive Partitoning, Regresi Logistik, Receiver Operating Characteristic (ROC), Statistik Kolmogorov-Smirnov (KS), Indeks Gini, C-Statistik.

1. Pendahuluan

Dalam menjalani kehidupan, setiap orang atau suatu lembaga menginginkan keuntungan bukan kerugian yang mengakibatkan risiko. Secara umum risiko dapat diartikan sebagai suatu keadaan yang dihadapi seseorang atau perusahaan dimana terdapat kemungkinan yang merugikan. Di zaman sekarang ini, banyak risiko yang terjadi di berbagai lingkup lembaga, misalnya saja pada lembaga keuangan yaitu bank. Istilah risiko yang terjadi di bidang tersebut yaitu risiko kredit. Risiko kredit terjadi karena ketidakmampuan nasabah atas kewajiban pembayaran utangnya baik utang pokok maupun bunganya atau kedunya (kredit macet). Faktor penyebab timbulnya kredit macet salah satunya yaitu penyelenggaraan credit scoring yang kurang mampu dilakukan oleh pihak bank (Sjafitri, 2011).

Untuk kebanyakan bank, risiko kredit merupakan risiko terbesar yang dihadapinya karena dapat menguras modal bank dengan cepat. Selain itu, peranan bank sebagai lembaga intermediasi tidak dapat berfungsi sehingga akan memperkecil kesempatan peluang bisnis, proyek baru, lapangan kerja baru, dan sebagainya. Terdapat cara untuk meminimalkan risiko ini yaitu dengan melakukan analisis risiko melalui model credit scoring.

Model credit scoring merupakan suatu alat dan teknik prediksi yang membantu lembaga keuangan dalam pemberian kredit (Rezac, 2011). Suwondo dan Santosa (2014) menyebutkan bahwa tujuan pembuatan model credit scoring untuk menganalisa dan membuat keputusan yang lebih cepat, tepat dan efisien terhadap penyeleksian nasabah yang akan menerima kredit. Model credit scoring akan menghasilkan scorecard yaitu nilai untuk setiap kategori dari karakteristik calon debitur (variabel bebas). Regresi logistik merupakan teknik yang umum digunakan untuk mengembangkan scorecard di sebagian lembaga keuangan, dimana variabel yang diprediksi adalah variabel kategori (Siddiqi, 2006). Dalam skripsi ini regresi logistik yang digunakan menggunakan regresi logistik biner.

Di samping itu, untuk mengukur seberapa baik model credit scoring dalam mengklasifikasikan nasabah yang baik dan buruk dan memperkuat ketetapan prediksi model maka dilakukan validasi terhadap model. Model yang baik akan berdampak terhadap penyeleksian calon nasabah yang akan

menerima pinjaman secara akurat. Terdapat beberapa ukuran yang dapat digunakan untuk melakukan validasi model dan pihak lembaga dapat memilih ukuran yang untuk melakukan validasi model. Tujuan dari penelitian ini yaitu mempraktikkan beberapa ukuran validasi model credit scoring diantaranya kurva Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), Statistik Kolmogorov-Smirnov (KS), Indeks Gini, dan C-Statistik.

2. Kajian Pustaka

2.1. Kredit

Menurut undang-undang Nomor 14 tahun 1967 tentang Pokok-Pokok Perbankan, yang dimaksud dengan kredit adalah: "Penyediaan uang atau tagihan-tagihan yang dapat disamakan dengan itu berdasarkan persetujuan pinjam-meminjam antara bank dengan pihak lain dalam hal mana pihak peminjam berkewajiban melunasi utangnya setelah jangka waktu tertentu dengan jumlah bunga yang telah ditentukan".

2.2 Diskritisasi

Diskritisasi merupakan proses transformasi data kuantitatif menjadi pengkategorian yang berguna untuk proses scorecard. Menurut Kotsiantis dan Kanellopoulos (2006) terdapat empat tahapan diskritisasi, yaitu:

1. Mengurutkan nilai kontinu yang akan didiskritisasi.
2. Mengevaluasi titik potong sebagai pemisah selang atau penggabung selang yang berdekatan.
3. Berdasarkan kriteria tertentu dilakukan pemisahan atau penyatuan selang nilai.
4. Menghentikan proses pada titik tertentu.

Salah satu metode diskritisasi yaitu metode tersupervisi dan tidak tersupervisi. Dalam penelitian ini menggunakan metode diskritisasi tersupervisi yaitu recursive partitioning dengan CART (Classification and Regression Trees). CART adalah salah satu metode atau algoritma dari salah satu teknik eksplorasi data yaitu teknik keputusan. Algoritma pembentukan pohon klasifikasi terdiri dari empat (Kardiana dkk, 2006), yaitu pemilihan pemilih, penentuan simpul terminal, penandaan label kelas, dan pentuan pohon dengan ukuran tepat.

1. Pemilihan pemilah

Pada tahap ini dicari pemilih dari setiap simpul yang menghasilkan penurunan tingkat kehetrogenan paling tinggi. Keheterogenan yaitu simpul diukur berdasarkan nilai impurity-nya. Fungsi impuritas yang dapat digunakan adalah indeks Gini. Bila impuritas suatu simpul semakin besar maka semakin heterogen simpul tersebut (Brieman dkk, 1993).

Nilai impuritas menggunakan indeks Gini pada simpul t, i(t), dapat ditulis sebagai berikut:

$$i(t) = 1 - \Sigma_j p^2(j \mid t) \qquad ...(2.1)$$

dimana $p(p \mid t)$ adalah peluang unit pengamatan dalam kelas ke-j dari simpul t yang dinyatakan sebagai berikut:

$$p(j \mid t) = \frac{\pi_j N_j(t)/N_j}{\Sigma_j \pi_j N_j(t)/N_j} \qquad ...(2.2)$$

Dengan π_j adalah peluang awal kelas ke-j, N_j adalah banyaknya unit pengamatan dalam ke-j, dan $N_j(t)$ adalah banyaknya unit pengamatan yang termasuk ke dalam kelas ke-j pada simpul t.

2. Penentuan simpul terminal

Suatu simpul t akan menjadi simpul terminal atau tidak akan dipilih kembali, jika jumlah pengamatannya kurang dari jumlah minimum. Umumnya jumlah pengamatan minimum pada

simpul besar 5 dan terkadang berjumlah 1 (Brieman dkk, 1993). Maka selanjutnya t tidak dipilih lagi tetapi dijadikan simpul terminal dan hentikan pembuatan pohon.

3. Penandaan label kelas

Label kelas dri simpul terminal ditentukan berdasarkan aturan jumlah terbanyak, yaitu jika $P(j_0 \mid t) = \max_j P(j \mid t)$, maka label kelas untuk terminal t adalah j_0 (Brieman dkk,1993).

4. Penentuan pohon optimum

Menurut Brieman dkk. (1993), salah satu cara mendapatkan pohon optimum yaitu dengan pemangkasan (pruning). Pemangkas berturut-turut memangkas pohon bagian yang kurang penting. Tingkat kepentingan sebuah pohon bagian diukur berdasarkan ukuran biaya kompleksitas (cost-complexity). Persamaanya adalah:

$$R_\alpha(T_k) = R(T_k) + \alpha\left|\tilde{T}_k\right| \qquad ...(2.3)$$

Hasil proses pemangkasan berupa sederet pohon klasifikasi T_k dan dengan validasi saling (cross-validation sample) dapat ditentukan pohon optimum T_{k0} sebagai berikut:

$$R^{CV}(T_{ko}) = \min_k\left(R^{CV}(T_k)\right)$$

2.3. Weight of Evidence (WoE)

Weight of Evidence (WoE) adalah suatu nilai yang digunakan untuk mengukur kekuatan setiap kategori dari variabel bebas setelah didiskritisasi (Siddiqi, 2006). Perhitungan nilai WoE dilakukan untuk setiap kategori di variabel bebas hasil diskritisasi. WoE dapat dirumuskan sebagai berikut:

$$WoE_i = \left[\ln\left(\frac{\text{DistrGood}_i}{\text{DistrBad}_i}\right)\right] \text{x } 100 \qquad ...(2.4)$$

Dimana:

Distr Good_i : Presentase jumlah nasabah kategori ke-i pada kelompok nasabah yang baik terhadap jumlah nasabah kelompok baik.

Distr Bad_i : Presentase jumlah nasabah kategori ke-i pada kelompok nasabah yang buruk terhadap jumlah nasabah kelompok buruk.

i : 1, 2, 3, ..., k

k : Banyaknya kategori untuk variabel bebas tertentu.

2.4. Information Value (InV)

Information Value (InV) merupakan suatu nilai yang digunakan untuk mengukur kekuatan dari variabel bebas setelah didiskritisasi (Siddiqi, 2006). Nilai InV digunakan untuk menyeleksi variabel bebas mana saja yang akan masuk kedalam model regresi logistik berdasarkan nilai batas tingkat prediksi tertentu. Berdasarkan SAS Institute (2012) jika nilai InV kurang dari 0,02 maka variabel bebas tidak dimasukkan kedalam model. Information Value dapat dirumuskan sebagai berikut:

$$InV = \sum_{i=1}^{k}\left(\text{DistrGood}_i - \text{DistrBad}_i\right) * \ln\left(\frac{\text{DistrGood}_i}{\text{DistrBad}_i}\right) \qquad ...(2.5)$$

Menurut Siddiqi (2006) tingkat prediksi InV dibagi kedalam beberapa kategori, yaitu:

1. Nilai InV < 0,02 : variabel bebas tidak prediktif
2. Nilai 0,02 < InV < 0,1 : variabel bebas memiliki tingkat prediktif lemah
3. Nilai 0,1 < InV < 0,3 : variabel bebas memiliki tingkat prediktif medium
4. Nilai InV $\geq$ 0,3 : variabel bebas memiliki tingkat prediktif kuat

Berdasarkan SAS Institute Inc (2012) jika nilai *InV* kurang dari 0,02 maka variabel bebas dikatakan tidak prediktif sehingga variabel bebas tersebut tidak dimasukkan kedalam model.

2.5. Regresi Logistik

Regresi logistik merupakan metode analisis statistika yang digunakan untuk menganalisis hubungan antara variabel tak bebas yang bersifat biner atau dikotomus dengan satu atau lebih variabel bebas (Hosmer dan Lemeshow, 2000). Pada regresi logistik, variabel tak bebas berskala kategorik. Variabel tak bebas yang dinotasikan dengan y bersifat biner atau dikotomus yang mempunyai dua nilai yaitu 0 dan 1. Dalam keadaan demikian, variabel y mengikuti distribusi Bernoulli untuk setiap observasi tunggal. Fungsi probabilitas untuk setiap observasi diberikan sebagai berikut:

$$f(y) = \pi^{y}(1-\pi)^{1-y} \qquad y = 0,1 \qquad \text{...(2.6)}$$

sehingga diperoleh:

Jika $y = 0$ maka $f(y) = \pi^{0}(1-\pi)^{1-0} = 1-\pi$

Jika $y = 1$ maka $f(y) = \pi^{1}(1-\pi)^{1-1} = \pi$

Fungsi regresi logistiknya dapat dituliskan sebagai berikut:

$$f(y) = \frac{1}{1+e^{-y}} \qquad \text{atau} \qquad f(y) = \frac{e^{y}}{1+e^{y}} \qquad \text{...(2.7)}$$

dimana $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ dengan p = banyak variabel predikator. Nilai y antara $-\infty$ dan $+\infty$ sehingga nilai $f(y)$ terletak antara 0 dan 1 untuk setiap nilai y yang diberikan. Hal tersebut menunjukkan bahwa model logistik sebenarnya menggambarkan probabilitas atau risiko dari suatu objek. Model regresi logistik adalah sebagai berikut.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}, \text{ dimana } p \text{ = banyaknya variabel prediktor.} \qquad \text{...(2.8)}$$

Fungsi $\pi(x)$ di atas berbentuk non linear sehingga untuk membuatnya menjadi fungsi linier harus dilakukan transformasi logit sebagai berikut:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad \text{...(2.9)}$$

2.5.1. Pendugaan Parameter Regresi Logistik

Pendugaan parameter yang digunakan regresi logistik adalah metode kemungkinan maksimum (maximum likelihood, ML). Prinsip dasar dari metode kemungkinan maksimum adalah memilih suatu penaksir parameter sedemikian rupa sehingga dapat memaksimumkan fungsi peluang yang diamati. Jika xi dan yi adalah pasangan variabel prediktor dan terikat pada pengamatan ke – i dan diasumsikan bahwa setiap pasangan pengamatan saling bebas dengan pasangan pengamatan lainnya, i = 1, 2, ... , n, maka fungsi probabilitas untuk setiap pasangan adalah sebagai berikut:

$$f(x_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad ; y_i = 0, 1 \qquad \text{... (2.10)}$$

dengan,

$$\pi_i = \frac{e^{\left(\sum_{j=0}^{p}\beta_j x_j\right)}}{1+e^{\left(\sum_{j=0}^{p}\beta_j x_j\right)}} \qquad \text{... (2.11)}$$

dimana ketika $j = 0$ maka nilai $x_{ij} = x_{i0} = 1$. Fungsi likelihood untuk distribusi Bernoulli adalah :

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{yi} \left[1 - \pi(x_i)\right]^{1-yi}$$

$$= \prod_{i}^{n} \left(\frac{\pi_i}{1-\pi_i}\right)^{yi} (1-\pi_i)$$

dengan persamaan logistik $\pi_i = \frac{e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}}{1+e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}} = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$,

$$1-\pi_i = 1 - \frac{e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}}{1+e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip})}} = 1 - \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}} = (1+e^{x_i^T\beta})^{-1}$$

Dari persamaan diatas maka diperoleh $\frac{\pi_i}{1-\pi_i} = e^{x_i^T\beta}$. Sehingga fungsi *likelihood* dapat diperoleh menjadi :

$$l(\boldsymbol{\beta}) = \prod_{i}^{n} e^{x_i^T\beta y_i} (1+e^{x_i^T\beta})^{-1}$$

...(2.12)

Setelah fungsi *likelihood* didapat, langkah selanjutnya yaitu memperoleh nilai log- *likelihood* yang dapat dinyatakan sebagai berikut :

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta})$$

$$= \sum_{i=1}^{n} \left[x_i^T \beta y_i - \ln(1+e^{x_i^T\beta}) \right] \quad \text{...(2.13)}$$

Untuk mendapatkan nilai penaksiran koefisien regresi logistik $(\hat{\beta})$ dilakukan dengan penurunan $L(\beta)$ terhadap β dan disamakan dengan 0. Turunan pertama dari $x_i^T\beta$ terhadap β_j adalah x_{ij}, sehingga penaksir β di hitung menggunakan rumus:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \quad \text{...(2.14)}$$

Kita ulangi proses tersebut sampai dengan konvergen, artinya nilai $\beta^{(t+1)}$ sangat mendekati nilai $\beta^{(t)}$. Demikian juga pada saat konvergen, maka $\left(X^T W^{(t)} X\right)^{-1}$ merupakan penaksir matriks kovarians bagi $\hat{\beta}$.

2.6. Model Credit Scoring

Model credit scoring merupakan suatu metode untuk mengevaluasi kelayakan kredit seseorag berdasarkan rumus tertentu atau suatu aturan tertentu. Dalam hal ini, model credit scoring menghasilkan kartu skor yaitu nilai skor setiap kategori di variabel bebas. Teknik ini mengacu pada jangkauan dan format skor dalam kartu skor (Siddiqi, 2006).

Perhitungan skor untuk setiap kategori pada satu variabel bebas, disajikan sebagai berikut :

$$Score_i = \sum_{j=1}^{p} \left(\left(WoE_i * \hat{\beta}_j + \frac{\hat{a}}{p} \right) * factor + \frac{offset}{p} \right) \quad \text{...(2.15)}$$

Dimana :

WoE_i : Nilai *weight of evidence* untuk setiap kategori

$\hat{\beta}_j$: Nilai dugaan koefisien parameter setiap variabel bebas

$\hat{a}$: Nilai *intercept* dari hasil regresi logistik

p : banyaknya variabel bebas

j : indeks untuk variabel bebas ; j =1, 2, ..., p

2.7. Ukuran-Ukuran Validasi Model

Beberapa ukuran yang dapat dilakukan untuk validasi yaitu Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), Statistik Kolmogorov-Smirnov (KS), Indeks Gini, dan C-Statistik.

2.7.1 Receiver Operating Characteristic (ROC)

Kurva ROC adalah plot kombinasi nilai sensitivitas dengan nilai 1-spesivisitas dengan berbagai *cut off* yang mungkin. Suatu model yang dikatakan baik jika mendekati 100% sebaliknya model yang tidak baik mendekati 50%. Kurva ROC merupakan hasil dari tabel klasifikasi. Untuk memperoleh tabel klasifikasi, kita harus menetapkan *cutpoint c* misalnya berdasarkan desil. Setelah nilai desil diperoleh, kemudian membandingkan nilai desil dengan setiap skor sehingga diperoleh $\hat{y}_i$.

$$\hat{y}_i \begin{cases} 1, \text{ jika} s_i > c \\ 0, \text{ jika} s_i \leq c \end{cases}$$

Kemudian nilai-nilai tak bebas y_i yang sebenarnya dirangkum ke dalam tabel kontingensi bersama nilai-nilai prediksinya $\hat{y}_i$. Hasil tabulasi silang tersebut disebut *confusion matrix*. Bentuk dari *confusion matrix* diperlihatkan pada Tabel 2.1.

Tabel 2.1 Bentuk dari *confusion matrix* untuk *cutpoint c* tertentu (Fawcett, 2003)

		Kelas Sebenarnya: Benar	Kelas Sebenarnya: Salah
Kelas Prediksi	Benar	Benar Positif (*True Positives* (TP))	Salah Positif (*False Positives* (FP))
	Salah	Salah Negatif (*False Negatives*(FN))	Benar Negatif (*True Negatives* (TN))
Jumlah Total Kolom:		P	N

Beberapa parameter pengukur kinerja ditunjukkan dengan Persamaan (2.16) sampai dengan Persamaan (2.20).

$$\text{FP } rate = \frac{FP}{N} \quad \text{...(2.16)}$$

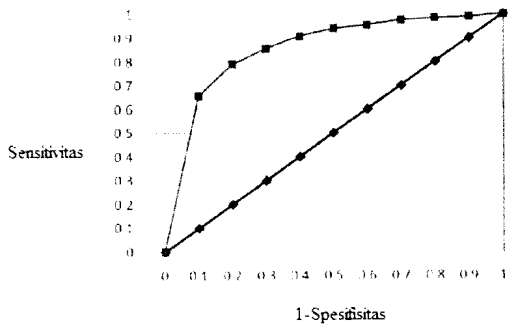
$$\text{TP } rate = \frac{TP}{P} = Recall \quad \text{...(2.17)}$$

$$\text{Sensitivitas} = Recall \quad \text{...(2.18)}$$

$$\text{Spesifisitas} = \frac{TN}{FP+TN} = 1 - \text{FP } rate \quad \text{...(2.19)}$$

$$1 - \text{Spesifisitas} = 1 - (1 - \text{FP } rate) = \text{FP } rate \quad \text{...(2.20)}$$

Kurva Receiver Operating Characteristic (ROC) dibentuk berdasar pasangan koordinat 1-spesifisitas dan sensitivitas untuk berbagai nilai c, seperti pada gambar 2.3.



Gambar 2.1 Kurva ROC

2.7.2. Area Under the Curve (AUC)

Dari prosedur ROC akan sekaligus mendapatkan nilai Area Under the Curve (AUC) yang letaknya dibawah kurva ROC. Nilai AUC akan mudah didapat dengan menghitung luas daerah dibawah kurva ROC. Untuk suatu kurva ROC yang memadai, maka letaknya selalu berada di daerah sebelah atas dari garis diagonal (0,0) dan (1,1). Nilai AUC 1 merupakan nilai AUC terbaik sehingga model akan semakin baik ketika nilai AUC mendekati 1.

2.7.3. Statistik Kolmogorov-Smirnov

Statistik Kolmogorov-Smirnov (KS) digunakan untuk melihat seberapa besar model untuk memisahkan nasabah baik dan buruk. Nilai KS memuat nilai antara 0 sampai 1. Jika nilai statistik KS mendekati nilai 0 maka model semakin tidak dapat membedakan nasabah baik dan buruk begitupun sebaliknya jika nilai statistik KS mendekati nilai 1 maka model semakin ideal dalam membedakan nasabah baik dan buruk (Rezac, 2011). Asumsikan bahwa untuk setiap nasabah terdapat informasi mengenai nilai skor *s* dan keterangan sebagai berikut:

$$D_k = \begin{cases} 1, & nasabah yang baik \\ 0, & sebaliknya \end{cases}$$

Sehingga KS dapat didefinisikan sebagai:

$$KS = \max_{c \in [L,H]} \left| F_{m.BAD}(c) - F_{n.GOOD}(c) \right| \qquad ...(2.21)$$

2.7.4. Indeks Gini

Indeks ini menggambarkan kualitas global dari fungsi skor yang nilainya berkisar antara -1 dan 1 (Rezac, 2011). Model yang ideal memiliki indeks Gini mendekati dengan nilai 1, yaitu fungsi skor yang sempurna memisahkan nasabah yang baik dan buruk. Di sisi lain, model yang memberikan fungsi scoring acak kepada nasabah akan memiliki indeks Gini sama dengan 0. Nilai negatif mempunyai makna fungsi skor yang terbalik yaitu jika skor semakin kecil maka peluang prediksi nasabah yang baiknya semakin tinggi. Indeks Gini dapat dirumuskan sebagai beikut:

$$Gini = 1 - \sum_{k=2}^{n+m} \left| \left(F_{m.BAD_k} - F_{m.BAD_{k-1}} \right) . \left(F_{n.GOOD_k} + F_{n.GOOD_{k-1}} \right) \right| \qquad ...(2.22)$$

dimana :

$F_{m.BAD_k}$ $(F_{n.GOOD_k})$: nilai vektor ke-*k* dari fungsi distribusi empirik nasabah buruk (baik).

2.7.5. C-Statistik

Ukuran lain dari kualitas model adalah c-statistik. Nilai c-statistik diantara 0,5 sampai 1, dimana nilai c-statistik sebesar 0,5 menunjukkan nilai kualitas model acak dan 1 menunjukkan nilai kualitas model yang ideal (Rezac, 2011).

Sehubungan dengan indeks Gini, c-statistik (Siddiqi, 2006) dirumuskan sebagai:

$$c - statistik = \frac{1 + Gini}{2} \quad ...(2.23)$$

3. Data dan Hasil

Sumber data yang digunakan merupakan data sekunder nasabah yang terdapat pada data kredit di Jerman yang terdiri 1000 nasabah yaitu 700 nasabah baik dan 300 nasabah buruk. Data kredit yang akan digunakan terdapat 1 variabel tak bebas dan 20 variabel bebas yang terdiri dari 7 variabel bebas dengan skala pengukuran numerik dan 13 lainnya dengan skala pengukuran kategori . Variabel tak bebas dalam penelitian ini adalah status kolektibilitas nasabah berupa status baik (Y=1) dan buruk (Y=0). Pada tabel 3.1 disajikan contoh data yang memuat variabel respon dan variabel prediktor.

Tabel 3.1 Data Kredit Jerman Tahun 1994

X_1	X_2	X_3	...	X_{18}	X_{19}	X_{20}	Y
A11	6	A34	...	1	A192	A201	0
A12	48	A32	...	1	A191	A201	1
A14	12	A34	...	2	A191	A201	0
A12	45	A34	...	1	A191	A201	0

Sumber: https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

3.1. Hasil Diskritisasi, Weight of Evidance (WoE) dan Information Value (InV)

Dalam pembuatan model credit scoring, hal yang harus dilakukan terlebih dahulu yaitu melaukukan diskritisasi terhadap varaabel yang bertipe numerik, melakukan perhitungan WoE dan InV. Dibawah ini merupakan hasil dari diskritisasi, WoE, dan InV yang disajikan dalam tabel 3.2.

Tabel 3.2 Hasil Diskritisasi, WoE, dan InV

Variabel	Kategori	WoE	InV	Tingkat Prediksi	Kesimpulan
X_1	A11	-0,7409	0,5581	Kuat	Masuk ke dalam model
	A12	-0,4080			
	A13	0,4987			
	A14	1,0242			
X_2	0 < lamanya kredit ≤ 12	0,3905	0,2226	Medium	Masuk ke dalam model
	12 < lamanya kredit ≤ 15	0,8782			
	15 < lamanya kredit ≤ 24	-0,0203			
	24 < lamanya kredit ≤ 30	-0,0403			
	lamanya kredit > 30	-0,8135			
...	...	...	...	...	...
X_{20}	A201	-0,0367	0,0463	Lemah	Tidak masuk ke dalam model
	A202	1,2659			

Dengan melihat tabel 3.1 dapat disimpulkan bahwa yang masuk kedalam model regresi logistik terdapat 15 variabel bebas dan sisanya yaitu 5 variabel bebas tidak masuk kedalam kedalam model regresi logistik yang selanjutnya akan digunakan untuk pembuatan skor beserta validasinya.

3.2. Membentuk Model Credit Scoring

Sebelum melakukan validasi terlebih dahulu membentuk model credit scoring yang berisikan skor untuk setiap kategori dalam suatu variabel bebas. Pembuatan skor memuat nilai taksiran parameter dari model regresi logistik. Di bawah ini hasil taksiran model regresi logistik berdasarkan persamaan 2.14 dengan bantuan software R 3.2.1.

Tabel 3.3 Taksiran Parameter untuk Model Regresi Logistik

Y	Parameter	Taksiran
Kolektibilitas Nasbah	Intercept	0,8104
	X_1	0,7629
	X_2	0,6859
	...	...
	X_{15}	0,3236
	X_{20}	0,8725

Setelah nilai taksiran parameter diperoleh langkah selanjutnya pembuatan kartu skor berdasarkan persamaan 2.15. Di bawah ini akan dipaparkan hasil kartu skor untuk setiap kategori dalam suatu variabel bebas terpilih yang disajikan dalam tabel 3.4.

Tabel 3.4 Kartu Skor untuk Setiap Kategori

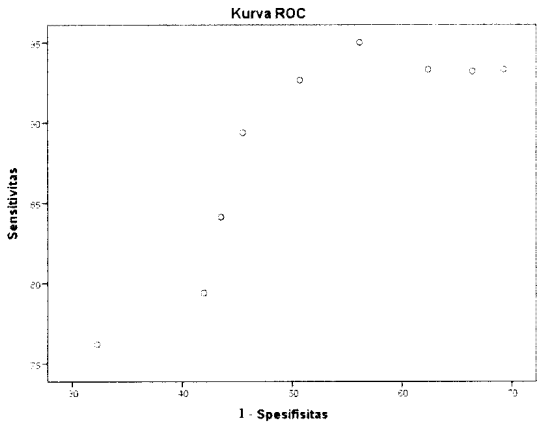
Variabel	Kategori	Skor
X_1	A11	18
	A12	25
	A13	45
	A14	57
X_2	0 < lamanya kredit ≤ 12	42
	12 < lamanya kredit ≤ 15	51
	15 < lamanya kredit ≤ 24	34
	24 < lamanya kredit ≤ 30	33
	lamanya kredit > 30	18
...	...	...
X_{20}	A201	33

3.1. Validasi Model

Setelah dilakukan pembuatan model credit scoring, perlu dilakukan validasi model terhadap data testing untuk memperkuat ketetapan prediksi dan melihat seberapa baik model credit scoring dalam mengklasifikasikan nasabah yang baik dan buruk. Beberapa ukuran yang dapat dilakukan untuk validasi yaitu Receiver Operating Characteristic (ROC), Area Under the Curve (AUC), Statistik Kolmogorov-Smirnov (KS), Indeks Gini, dan C-Statistik.

- ***Receiver Operating Characteristic* (ROC)**

Dibawah ini merupakan kurva ROC hasil tabulasi silang atau confunsion matrix yang memuat nilai sensitivitas dan 1-spesifisitas.



Gambar 3.1 Kurva ROC Data Kredit Jerman

Pada kurva ROC diatas menunjukkan bahwa kurva mendekati kurava 100% yang dapat diinterpretasikan bahwa model baik atau akurat dalam penyeleksian yang akan menerima kredit.

Tabel 3.5 Hasil Validasi AUC, Statistik K-S, Indeks Gini, dan C-Statistik

Area Under the Curve (AUC)	Statistik Kolmogorov-Smirnov (K-S)	Indeks Gini	C-Statistik
0,5	0,5207	0,6443	0,8321

Pada tabel 3.5 nilai AUC, statistic K-S, indeks Gini, dan c-statistik mendekati nilai 1 yang dapat diinterpretasikan bahwa model baik atau kualitas model ideal dalam memsisahkan nasabah baik dan buruk.

4. Kesimpulan

Berdasarkan analisis yang dilakukan terhadap data kredit Jerman tahun 1994 dapat disimpulkan bahwa model credit scoring yang dibuat menggunakan regresi logistik baik atau ideal dan dapat mengklasifikasikan nasabah baik dan buruk, itu terlihat dari hasil validasi model. Dari hasil validasi model yang pertama yaitu Receiver Operating Characteristic (ROC) dapat disimpulkan bahwa model baik atau akurat dalam penyeleksian yang akan menerima kredit dengan berbagai cutt-point berdasarkan desil dan berdasarkan Area Under the Curve (AUC) model mendekati nilai 1 sehingga dapat disimpulkan bahwa model baik dalam penyeleksian yang akan menerima kredit. Selanjutnya ukuran validasi model dengan statistik K-S model hanya dapat memisahkan nasabah baik dan buruk sebesar 0,5171 sehingga model efektif dalam penyeleksian yang akan menerima kredit. Begitupun dengan hasil ukuran validasi model indeks Gini dan c-statistik yang menunjukkan bahwa fungsi scoring yang sempurna dan kualitas model yang ideal sehingga model credit scoring dapat digunakan untuk penyeleksian nasabah yang akan menerima kredit.

Daftar Pustaka

Ardiati, I. D. 2014. Perbandingan Metode Diskretisasi dalam Model Regresi Logistik (Studi Kasus: Pembentukan Model Penskoran Kredit Bank X) . Skripsi. Bogor : Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor.

Brieman, dkk. 1993. Classification and Regression Trees. New York: Champan and Hall.

Fawcett, T. 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. HP Laboratories Working Paper, 861-874.

Hajarisman, Nusar. 2009. Buku Ajar Analisis Data Kategorik. Bandung: Program Studi Statistika Universitas Islam Bandung.

Hosmer, D. W., dan Lemeshow. 2000. Applied Logistic Regression. New York: John Wiley and Sons.

Kardiana, dkk. 2006. Metode Klasifikasi Berstruktrur Pohon Biner: Kasus Perkiraan Sifat Huja Bulanan dib Bogor. Yogyakarta, 17 Juni 2006. Seminar Nsional Aplikasi Teknologi Informasi (SNATI): G21-G25.

Kotsiantis S, Kannellpoulus D. 2006. Discretization Technique: A recent survey International Transaction On Computer Science and Engineering, 32 : 47-58.

Rezac, M. 2011. How to Measure the Quality of Credit Scoring Models. Journal of Economics and Finance, 5 : 486-507.

Siddiqi, N. 2006. Credit Risk Scorecard Developing and Implementing Intelligent Credit Scoring. New Jersey (US) : John Willey & Sons.

Sjafitri, Henny. 2011. Faktor-Faktor yang Mempengaruhi Kualitas Kredit dalam Dunia Perbankan. Jurnal Manajemen dan Kewirausahaan, 2: 160-120.

Suwondono dan Santosa, Stefanus. 2014. Credit Scoring Menggunakan Metode Support Vector Machine dengan Teknik Seleksi Atribut Berbasis Chi Squared Statistic dan Particle Swarm Optimization, 10 : 1-18.

Suyatno, Thomas., dkk. 2010. Dasar-Dasar Perkreditan Edisi Keempat. Raja Grapindo Persada.