Pemilihan Variabel Prediktor Terbaik dalam Pemodelan Regresi Logistik untuk Data Pembayaran Kartu Kredit

Best Predictor Variables Selection in Logistic Regression Modeling for Credit Card Payment Data

[1]Fahmi Mohamad Alwi, [2]Abdul Kudus, [3]Aceng Komarudin Mutaqin
[1,2,3]*Prodi Statistika, FakultasMatematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Jl. Tamansari No.1 Bandung 40116*
email: [1]fahmimohamadalwi@gmail.com

Abstract. Credit scoring is a method used to evaluate credit risk in terms of loan applications from debtors. This method is used to determine the provision of credit limit from the debtor based on the payment behavior performed by the debtor. In the process because we are dealing with very large data either the number of variables or the number of observations, then we do the workmanship techniques based on the principle of data mining. Where the selection of best predictor variables will be selected based on the value of information value greater than 10% obtained using unsupervised learning techniques and modeling techniques in credit scoring is by using logistic regression which is a supervised learning technique in data mining. From the result of data transformation, correlation and modeling examination with backward elmination of 23 predictor variables obtained 5 best predictor variables modeled to determine the behavior of payments made by credit card customers in Taiwan, namely; Historical payment of the month of April 2005, the amount of the previous payment in June 2005, the amount of the previous payment in July 2005, the previous payment amount in August 2005, and the amount of previous payment in September 2005. Then after the best predictor variable is obtained we do logistic regression modeling, With the value of KSnya 50.33% and 60.84% and the value of GINI coefficient 53.83% and 62.56%, then the model can be said reliable in separating the debtor included in the good debtor and bad debtor.

Keywords: Credit Scoring, Data Mining, Logistic Regression.

Abstrak. *Credit scoring* adalah metode yang digunakan untuk mengevaluasi risiko kredit dalam hal permohonan pinjaman dari debitur. Metode ini digunakan untuk menentukan pemberian limit kredit dari debitur berdasarkan perilaku pembayaran yang dilakukan oleh debitur. Dalam pengerjaan karena kita berhadapan dengan data yang sangat besar baik itu banyaknya variabel maupun banyaknya observasi, maka kita melakukan teknik pengerjaan berdasarkan prinsip data *mining*. Dimana pemilihan variabel prediktor terbaik akan dipilih berdasarkan nilai *information value* lebih besar sama dengan 10% yang didapat menggunakan teknik *unsupervised* learning dan teknik pemodelan dalam *credit scoring* ini yaitu dengan menggunakan regresi logistik yang merupakan teknik *supervised learning* dalam data *mining*. Dari hasil transformasi data, pemeriksaan korelasi dan pemodelan dengan elminasi *backward* dari 23 variabel prediktor didapat 5 variabel prediktor terbaik yang dimodelkan untuk menentukan perilaku pembayaran yang dilakukan oleh nasabah kartu kredit di Taiwan, yaitu; historis pembayaran bulan april 2005, jumlah pembayaran sebelumnya dibulan juni 2005, jumlah pembayaran sebelumnya di bulan juli 2005, jumlah pembayaran sebelumnya di bulan agustus 2005, dan jumlah pembayaran sebelumnya di bulan September 2005. Kemudian setelah diperoleh variabel prediktor terbaiknya kita lakukan pemodelan regresi logistik, dengan nilai KSnya 50,33% dan 60,84% serta nilai koefisien GINInya 53,83% dan 62,56%, maka model dapat dikatakan handal dalam memisahkan debitur yang termasuk ke dalam *good* debitur dan *bad* debitur.

Kata Kunci: *Credit Scoring*, Data *Mining*, Regresi logistik.

A. Pendahuluan

Tingginya permintaan kredit ini tak lantas membuat bank akan dapat mengabulkan semua permohonan yang ada karena memberikan kredit kepada masyarakat memiliki risiko yang besar. Bank harus mengalokasikan dana operasional yang besar untuk menanggulanginya, sehingga diperlukan unit manajemen risiko kredit untuk meminimalisasi risiko kredit yang dihadapi bank. Proses tersebut dikenal sebagai proses penilaian atau *scoring* dengan menggunakan data historis maupun data pribadi dari debitur.

Proses *credit scoring* ini tidak memungkinkan untuk dilakukan secara manual

karena data historis yang memiliki banyak variabel dan juga banyaknya calon debitur, oleh sebab itu kita memerlukan bantuan data *mining* dalam pengolahan data yang memiliki ukuran yang besar tersebut. Serta karena terlalu banyaknya variabel prediktor yang digunakan dalam pemodelan untuk menentukan peningkatan limit berdasarkan perilaku suatu calon debitur termasuk ke *good* atau *bad* akan mengakibatkan ketidakefisienan dari model yang dibuat, oleh sebab itu kita perlu memilih variabel-variabel prediktor mana saja yang memiliki kemampuan dalam memprediksi lebih baik dibanding variabel-variabel prediktor lainnya dimana setiap variabel prediktor dipilih berdasarkan nilai *information value* yang lebih besar dari 10%. Proses ini merupakan teknik *unsupervised learning* yaitu teknik yang mempelajari dan mencari pola-pola menarik pada data *input* yang diberikan, meskipun tidak disediakan *output* yang tepat secara eksplisit.

Sedangkan teknik *supervised learning* adalah model yang membuat fungsi untuk memetakan *input* ke *output* yang dikehendaki. Teknik ini digunakan dalam pembentukkan model menggunakan regresi logistik dengan aliminasi *backward* untuk semua variabel prediktor yang memiliki kemampuan paling baik dalam pemisahan suatu calon debitur yang akan dikategorikan ke dalam calon debitur yang *good* atau *bad*. Dimana hasil prediksi dari pemodelan regresi logistiknya dapat digunakan pihak bank dalam penentuan besar limit kredit yang harus diberikan oleh bank terhadap debitur untuk periode selanjutnya, berdasarkan perilaku pembayaran dari debitur yang bersangkutan. Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah dalam penelitian ini yaitu: "Variabel prediktor mana saja yang termasuk variabel prediktor terbaik dalam pemodelan *credit scoring* yang digunakan untuk penentuan pemberian limit kredit terhadap debitur untuk data pembayaran kartu kredit di Taiwan?". Selanjutnya, tujuan dalam penelitian ini yaitu: "Untuk mengetahui variabel prediktor mana saja yang termasuk variabel prediktor terbaik dalam pemodelan *credit scoring* yang digunakan untuk penentuan pemberian limit kredit terhadap debitur untuk data pembayaran kartu kredit di Taiwan".

B. Landasan Teori

Credit scoring adalah metode yang digunakan untuk mengevaluasi risiko kredit dalam hal permohonan pinjaman dari konsumen (Amaranth,2004). Metode ini digunakan untuk mengklasifikasikan konsumen yang mengajukan kredit termasuk ke dalam kelompok baik atau buruk. Untuk membuat suatu model *scoring* "*scorecard*" dalam menentukan karakteristik si peminjam, pengembangan analisis data dilakukan dengan melihat data historis konsumen kredit yang telah disetujui kreditnya atau tidak oleh pihak perusahaan. Hasil *scoring* ini akan berguna untuk memprediksi apakah calon konsumen dapat melaksanakan pinjaman dengan lancar atau tidak. Secara sederhana data mining adalah penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar (Davies,2004). Tujuan dari data mining yaitu: pertama *explanatory* adalah untuk menjelaskan beberapa kondisi penelitian, kedua *confirmatory* adalah untuk mempertegas hipotesis, dan *exploratory* adalah untuk menganalisa data yang memiliki hubungan yang baru. Pada pelaksanaanya terdapat dua teknik pengerjaan untuk data *mining* yaitu *Supervised Learning* dan *Unsupervised Learning*.

Data sangat penting bagi proses pengembangan *scorecard*. Persiapan data ini memakan waktu dan terdiri dari empat kegiatan utama yaitu sebagai berikut:

1. identifikasi karakteristik,
2. memilih data yang akan dijadikan sampel,
3. memeriksa kualitas data,

4. definisi *good/bad*.

Kemudian dalam transformasi data ini bertujuan untuk mengidentifikasi variabel prediktor mana saja yang dapat memisahkan calon debitur *good* dan *bad*, dimana variabel tersebut akan menjadi variabel prediktor dalam proses pembuatan model regresi logistiknya. Pada proses ini dibagi menjadi dua tahap yaitu *fine classing* dan *coarse classing*. Pada tahap *fine classing* pertama kita harus menghitung *good*/*bad odds* untuk setiap kategori dari variabel prediktor yang diperiksa, yang dihitung sebagai berikut:

$$Odds_i = \frac{\%Good_i}{\%Bad_i} \quad \ldots (1)$$

Dengan *good/bad odds*, kita dapat menghitung *weights* untuk semua kategori. *Weight* merupakan ukuran seberapa baik atau buruknya nasabah berada dalam kategori tertentu:

$$Weight\ of\ evidence_i = log(Odds_i) \quad \ldots (2)$$

Selanjutnya *weight* digunakan untuk menghitung *Information Value* (IV) dari suatu variabel. Pertama kita hitung terlebih dahulu IV untuk setiap kategori dari variabel prediktor yang diperiksa, dengan rumus sebagai berikut:

$$IV_i = Weight\ of\ evidence_i \times (\%Good_i - \%Bad_i) \quad \ldots (3)$$

kemudian kita jumlahkan nilai IV untuk setiap kategori yang hasilnya merupakan nilai *information value* untuk variabel prediktor yang diperiksa, dengan rumus sebagai berikut:

$$IV_{variabel} = \sum_{i=1}^{n} IV_i \quad \ldots (4)$$

Information value selalu lebih besar dari 0% tapi umumnya kurang dari 200% (Nguyen, 2014).

Berikut ini adalah salah satu aturan yang paling sering digunakan:

mengecualikan : IV < 10%,

memuaskan : 10% < IV < 100%,

sangat prediktif : IV > 100%.

Kemudian kita hitung nilai *bad rate* untuk setiap kategori dari variabel yang diperiksa dengan rumus sebagai berikut:

$$Bad\ Rate_i = \frac{Banyaknya\ Bad_i}{Banyak\ data\ pada\ kelas_i} \quad \ldots (5)$$

Selanjutnya dilakukan tahapan *coarse classing* yaitu menggabungkan kategori yang berdekatan dengan nilai *bad rate* yang hampir sama. Kemudian setelah diperoleh pengkategorian terakhir untuk semuavariabel prediktornya, selanjutnya diuji kuat hubungan atau korelasinya menggunakan koefisien cramers'v. Dimana kuat hubungannya harus rendah yaitu kurang dari 0,5. Formula koefisien cramer adalah sebagai berikut:

$$V = \sqrt{\frac{\chi^2}{N(k-1)}} \quad \ldots (6)$$

Setelah dipastikan korealsi anatar variabel prediktornya rendah kemudian

dialakukan pemodelan regresi logistik dimana variabel respon bersifat biner atau dikotomus. Variabel dikotomus adalah variabel yang hanya mempunyai dua kemungkinan nilai, misalnya macet $(y_i = 1)$ dan lancar $(y_i = 0)$.

Bentuk umum model peluang regresi logistiknya adalah sebagai berikut:

$$\pi(\mathbf{x}_i) = \frac{\exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_P x_{Pi})}{1+\exp(\beta_0+\beta_1 x_{1i}+\cdots+\beta_P x_{Pi})} \quad ...(7)$$

rasio antara $\pi(\mathbf{x}_i)$ dengan $1-\pi(\mathbf{x}_i)$, dapat dilakukan transformasi *logit*, yang didefinisikan sebagai:

$$g(\mathbf{x}_i) = log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_P x_{Pi} \quad \ldots(8)$$

Karena kita melakukan pengkategorian terhadap variabel-variabel prediktornya, maka kita membutuhkan bantuan dari variabel *dummy* dalam pemodelan serta interpretasinya. sehingga persamaan logitnya menjadi:

$$g(\mathbf{x}_i) = log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0 + \sum_{u=1}^{J-1}\beta_{1u}D_{1u} + \cdots + \sum_{u=1}^{J-1}\beta_{pu}D_{pu} \quad \ldots(9)$$

Kemudian menentukan apakah model yang telah kita buat itu sudah bagus atau tidak. Sebagaimana yang diungkapkan oleh (Nguyen, 2014) tentang ukuran-ukuran yang digunakan dalam mengukur kehandalan model yang telah dibuat, seperti:

1. Statistik *Kolmogorov Smirnov*

KS menghitung penyebaran terluas antara kumulatif *good* dan kumulatif *bad*. Dimana nilai KS yang semakin tinggi, semakin baik pula modelnya dalam membedakan antara calon debitur yang baik dan buruk. Rumus perhitungan KS adalah sebagai berikut:

$$D_{KS} = Max\{|cp_{B_i} - cp_{G_i}|\} \quad \ldots(10)$$

2. Koefisien GINI

Koefisien Gini adalah ukuran ketimpangan distribusi. Dalam kasus kredit koefisien Gini merupakan perhitungan kuantitatif dari seberapa bagus model dapat memisahkan antara calon debitur yang baik dan yang buruk. Rumus perhitungan koefisien Gininya sebagai berikut:

$$D_{Gini} = \sum_{i=1}^{n}\{(cp_{G_i} - cp_{G_{i-1}})(cp_{B_i} - cp_{B_{i-1}})\} - 1 \quad \ldots(11)$$

3. Distribusi skor

Nilai skor ini diperoleh dari Persamaan (2.11) dengan memasukkan variabel-variabel prediktornya kedalam model yang telah dibuat. Distribusi skornya dikatakan ideal jika tidak ada skor yang memiliki jumlah lebih dari 5% dari total akun didalamnya.

C. Hasil Penelitian dan Pembahasan

Pemilihan Variabel Prediktor Terbaik

Pada tahap ini dari 23 variabel prediktor yang ada akan dipilih variabel prediktor terbaiknya, pada tahapan ini akan dilakukan pembentukan *fine claasing* dan *coarse classing* untuk menghitung nilai *information value* dari setiap variabel prediktor, dimana variabel prediktor tersebut dikatakan memiliki kemampuan yang

baik dalam memisahkan debitur yang baik dari debitur yang buruk jika nilai *information value* nya lebih besar sama dengan 10%. Dari pembentukan *fine classing* dan *coarse classing* untuk semua variable prediktor kemudian diperoleh ringkasan dari *information value* nya yang diurutkan dari nilai yang terbesar ke nilai yang terkecil dengan nilai korelasi yang rendah, yaitu:

Tabel 1. Variabel Prediktor dengan Korelasi Rendah

No.	Variabel Prediktor Terbaik	Keterangan
1.	**X1**	Limit Kredit yang diberikan
2.	**X6**	Historis Pembayaran Bulan April 2005
3.	**X18**	Jumlah Pembayaran Sebelumnya di Bulan April 2005
4.	**X19**	Jumlah Pembayaran Sebelumnya di Bulan Mei 2005
5.	**X20**	Jumlah Pembayaran Sebelumnya di Bulan Juni 2005
6.	**X21**	Jumlah Pembayaran Sebelumnya di Bulan Juli 2005
7.	**X22**	Jumlah Pembayaran Sebelumnya di Bulan Agustus 2005
8.	**X23**	Jumlah Pembayaran Sebelumnya di Bulan September 2005

Setelah dipastikan semua variabel prediktor yang memiliki *information value* lebih besar sama dengan 10% memiliki korelasi yang rendah untuk setiap kombinasi dua dari variabel prediktornya, kemudian dilakukan pembagian data menjadi *development* dan *validation sample* dimana *development sample* merupakan 80% dari banyaknya data yaitu sebanyak 3225 data dan *validation sample* 20% dari abanyaknya data yaitu sebanyak 805 data. Dari hasil pembagian tersebut kita gunakan data *development* untuk pemodelan regresi logistik menggunakan eliminasi *backward* dan diperoleh persamaan regresi logistiknya sebagai berikut:

$$g(\mathbf{x}_i) = 0{,}82346 - 0{,}14613D_{1;1} - 0{,}09939D_{1;2} + 0{,}29870D_{1;3} + 0{,}29870D_{1;4} - 2{,}04485D_{6;1} - 0{,}26126D_{20;1} + 0{,}06382D_{20;2} + 0{,}28303D_{20;3} + 0{,}08633D_{20;4} + 0{,}07613D_{21;1} + 0{,}70986D_{21,2} + 0{,}27297D_{21,3} + 13906D_{22;1} + 0{,}38297D_{22,2} + 0{,}27701D_{22,3} + 0{,}25795D_{23;1} + 0{,}01314D_{23;2} + 0{,}35546D_{23,3}$$

Ukuran Kehandalan Model

1. **Statistik *Kolmogorov Smirnov* (KS)**

Tabel 2. Nilai Statistik *Kolmogorov Smirnov*

Development Sample	Validation Sample
50,33	60,84

Dari nilai KS diatas dapat kita ketahui jika KS yang diperoleh baik dari *development sample* maupun *validation sampel*, keduanya memiliki nilai yang lebih besar dari 35%, yang artinya kemampuan model dalam memisahkan debitur yang baik dan yang buruk bisa dikatakan handal.

2. Koefisien GINI

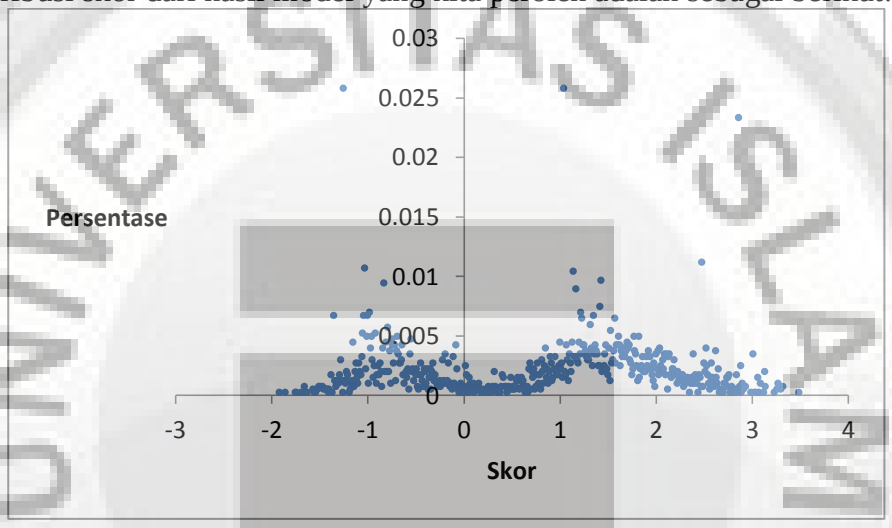
Tabel 3. Nilai Koefisien GINI

Development Sample	Validation Sample
53,83	62,56

Dari nilai GINI diatas dapat kita ketahui bahwa GINI yang diperoleh baik dari development sample maupun validation sampel, keduanya memiliki nilai yang lebih besar dari 40%, yang artinya kemampuan model dalam memisahkan debitur yang baik dan yang buruk bisa dikatakan handal.

3. Distribusi skor

Distribusi skor dari hasil model yang kita peroleh adalah sebagai berikut:



Gambar 1. Distribusi Skor

Dari plot diatas dapat dikatakan jika skor dari model yang diperoleh telah ideal karena dari setiap nilai skor yang ada tidak ada skor yang jumlahnya melebihi 5% dari banyaknya skor yang ada

D. Kesimpulan dan Saran

Kesimpulan

Berdasarkan hasil perhitungan di maka dapat disimpulkan dari 23 variabel prediktor terpilih sebanyak 5 variabel prediktor terbaik yang memiliki kemampuan dalam memisahkan debitur yang baik dari yang buruk untuk data pembayaran kartu kredit di Taiwan dan memiliki nilai korelasinya yang rendah serta hasil dari eliminasi *backward*, yaitu:

1. X_6 = Historis pembayaran bulan april 2005,
2. X_{20} = Jumlah pembayaan sebelumnya di bulan juni 2005,
3. X_{21} = Jumlah pembayaan sebelumnya di bulan juli 2005,
4. X_{22} = Jumlah pembayaan sebelumnya di bulan agustus 2005,
5. X_{23} = Jumlah pembayaan sebelumnya di bulan september 2005.

Berdasarkan ukuran kehandalan modelnya yaitu nilai KS sebesar 50,33% untuk *development sample* dan 60,84% untuk *validation sample* dimana keduanya melebihi 35% serta nilai koefisien GINInya sebesar 53,83% untuk *development sample* dan 62,56% untuk *validation sample* dimana keduanya melebihi 40% maka model

penentuan perilaku nasabah kartu kredit di Taiwan ini dapat dikatakan handal dalam membedakan nasabah yang baik dari nasabah yang buruk. Dan untuk distribusi dari skor dapat dikatakan ideal karena tidak ada nilai skor yang banyaknya melebihi 5%.

Saran

Adapun saran yang dapat dikemukakan dalam penuslisan skripsi ini adalah:

1. Dalam pembentukan model kredit sebaiknya, data dalam pemodelan harus ditambah dengan data pribadi dari nasabah seperti pekerjaan, lamanya bekerja, status kepemilikan rumah, dan lain-lain.
2. Untuk penelitian selanjutnya, disarankan untuk mencari referensi-referensi bisa melakukan perhitungan dalam penentuan *credit rating* sehingga bisa ditentukan pengelompokan terhadap debitur.
3. Untuk penelitian selanjutnya juga sebaiknya mencoba metode lain selain regresi logistik dalam memprediksi perilaku pembayaran dari nasabah, seperti menggunakan metode *Classification and Regression Tree (CART)*, analisis diskriminan, analisis probit, *decision tree*, dan sebagainya untuk mengetahui metode yang paling baik dalam analisis untuk data kredit.

Daftar Pustaka

Amaranth, K.N. (2004).*Statistical Methods in Consumer Credit Scoring.* Cranes Software Internasional Ltd . Product Abalyst .

Davies, B. (2004). *Database Systems 3rd Edition*. Palgrave. Basingstoke. UK.

Hajarisman, N. (2009). *Analisis Data Kategorik*. Bandung: Program Studi Statistika, Universitas Islam Bandung.

Naeem S. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons.

Nguyen H. (2014). Default Predictors in Credit Scoring – Evidence from France's Retail Banking Institution. *Journal of Economic*, 26, 1-20