sional-2021-Gaussian_Vol_10_No_4_Penerapan_Gradient_Boosting.pdf

by

Submission date: 12-Apr-2023 12:28PM (UTC+0700)
Submission ID: 2062296784
File name: sional-2021-Gaussian_Vol_10_No_4_Penerapan_Gradient_Boosting.pdf (260.4K)
Word count: 2681
Character count: 16512

ISSN: 2339-2541
JURNAL GAUSSIAN, Volume 10, Nomor 4, Tahun 2021, Halaman 617 - 623
Online di: https://ejournal3.undip.ac.id/index.php/gaussian/

PENERAPAN *GRADIENT BOOSTING* DENGAN *HYPEROPT* UNTUK MEMPREDIKSI KEBERHASILAN TELEMARKETING BANK

Silvia Elsa Suryana[1], Budi Warsito[2], Suparti[3]
[1,2,3] Departemen Statistika, Fakultas Sains dan Matematika, Universitas Diponegoro
silviaelsas12@gmail.com

ABSTRACT

Telemarketing is another form of marketing which is conducted via telephone. Bank can use telemarketing to offer its products such as term deposit. One of the most important strategy to the success of telemarketing is opting the potential customer to create effective telemarketing. Predicting the success of telemarketing can use machine learning. Gradient boosting is machine learning method with advanced decision tree. Gardient boosting involves many classification trees which are continually upgraded from previous tree. The optimal classification result cannot be separated from the role of the optimal hyperparameter. Hyperopt is Python library that can be used to tune hyperparameter effectively because it uses Bayesian optimization. Hyperopt uses hyperparameter prior distribution to find optimal hyperparameter. Data in this study including 20 independent variables and binary dependent variable which has 'yes' and 'no' classes. The study showed that gradient boosting reached classification accuracy up to 90,39%, precision 94,91%, and AUC 0,939. These values describe gradient boosting method is able to predict both classes 'yes' and 'no' relatively accurate.
Keywords: Telemarketing, Hyperopt, Gradient Boosting

1. PENDAHULUAN

Pemasaran adalah bagian yang tidak terpisahkan dari dunia bisnis. Terdapat berbagai macam jenis strategi pemasaran, salah satunya adalah perusahaan dapat menggunakan pemasaran langsung pada segmen target dengan cara menghubunginya. Seorang pendiri perusahaan yang bergerak di bidang telemarketing GSA Business Development, Jonathan Silverman menyusun beberapa strategi penting untuk kesuksesan telemarketing, salah satunya adalah menggunakan data yang akurat untuk memilih kontak yang tepat. Data dalam marketing dapat dimanfaatkan dalam mendukung pengambilan keputusan melalui *machine learning*. Salah satu metode *machine learning* adalah *boosting*. *Boosting* adalah strategi *ensemble* yang membagi data *training* menjadi beberapa bagian dan masing-masing bagian diterapkan model yang berbeda atau satu model dengan *setting* yang berbeda, kemudian hasilnya dikombinasikan berdasarkan 'suara' terbanyak (Essam, 2019).

Salah satu metode *boosting* adalah *gradient boosting*. *Gradient boosting* adalah teknik *supervised learning* berbasis *decision tree*. Algoritma dimulai dari menghasilkan pohon klasifikasi awal dan terus menyesuaikan pohon baru melalui minimalisasi fungsi kerugian (Natekin dan Knoll, 2013). *Gradient boosting* memerlukan konfigurasi *hyperparameter* pada tahap awal, kombinasi *hyperparameter* yang optimal mempengaruhi hasil prediksi *gradient boosting*. *Hyperopt* adalah *library* Python yang menunjang *tuning hyperparameter* dengan konsep optimasi Bayesian. Hyperopt menggunakan distribusi prior *hyperparameter* untuk mencari *hyperparameter* yang optimal. Berdasarkan penelitian Moro *et al*. (2014) mengenai aplikasi *machine learning* pada kegiatan telemarketing, maka salah satu teknik *machine learning* yaitu *gradient boosting* juga dapat digunakan untuk memprediksi klien yang berpotensi menerima tawaran langganan deposito dengan

hyperopt untuk *tuning hyperparameter*, sehingga hasil prediksinya dapat menunjang keputusan dalam proses pemilihan klien.

2. TINJAUAN PUSTAKA

2.1. Telemarketing

Secara umum telemarketing adalah kegiatan menjangkau langsung calon pelanggan lewat pembicaraan telepon. Menurut Stone (1992) ada beberapa hal dari telepon yang bisa menjadikannya sebagai perangkat efektif kegiatan pemasaran, diantaranya adalah komunikasi langsung 2 arah, seperti pertemuan 4 mata, dapat memperoleh jawaban secara langsung, fleksibel dan murah. Jonathan Silverman pendiri GSA Business Development, usaha berbasis layanan telemarketing di Inggris, menyusun beberapa strategi penting untuk menunjang kesuksesan kegiatan telemarketing. Strategi pertama adalah menentukan objek yang realistis, objek harus spesifik, terukur, terjangkau, dan realistis. Strategi kedua adalah mendefinisikan target pasar. Salah satu strategi penting dalam telemarketing adalah pemanfaatan data calon pelanggan. Teknik kecerdasan buatan yang dapat mendukung keputusan dalam manajerial disebut *Decision Support Systems* (DSSs). Salah satu konsep DSSs adalah *Business Intelligence* (BI) yang memasukkan teknologi informasi seperti *warehouses* dan *data mining* untuk mendukung pembuatan keputusan menggunakan data bisnis.

2.2. Data *Pre-processing*

Data *preprocessing* adalah teknik data mining yang mentransformasi data mentah kedalam bentuk yang mudah dimengerti oleh algoritma *machine learning*. Van der Heijden *et al* (2006) mengembangkan metode untuk menangani *missing values* pada teknik *machine learning* Metode yang paling umum untuk menangani *missing value* adalah *Complete Case Analysis* yaitu menghapus baris data yang mengandung *missing values*. Selain itu juga ada metode lain seperti *Single Unconditional Mean Imputation*, dan *Single Conditional Mean Imputation*. Tahap selanjutnya adalah pendeteksian *outlier* yang dapat dilakukan menggunakan nilai IQR. IQR adalah selisih kuartil atas dan kuartil bawah. Apabila data tidak terletak pada rentang antara (Q1 – 1.5 IQR) dan (Q3 + 1.5 IQR) maka disebut *outlier*. Metode lain yang umum adalah *z-score* yang mentransformasi data ke dalam nilai *z-score*. Pada data juga sering ditemui bermacam label kategorik dalam satu kolom, *label encoding* mengarah pada mengubah label data kategorik menjadi bentuk numerik sehingga menjadi mudah dibaca oleh algoritma *machine learning*.

2.3. Validasi Model

Proses validasi tidak dapat secara langsung menunjukkan masalah apa yang terjadi dalam model, tapi proses validasi dapat menunjukkan jika terdapat masalah dengan stabilitas model. Metode validasi paling dasar adalah *holdout validation*. Metode ini membagi data menjadi 2 bagian dengan proporsi tertentu. Proporsi yang umum digunakan adalah 60/40, 70/30, 80/20 (Raschka, 2018). Metode lain adalah K-fold *cross validation,* metode ini membagi data latih dan data uji sebanyak "K" kelompok, sehingga proses pelatihan akan menjadi sebanyak "K". Performa dari model adalah rata-rata dari semua proses pelatihan.

2.4. SMOTE-NC

SMOTE atau *Synthetic Minority Over-Sampling Technique* adalah pendekatan *over-sampling* pada kelas minoritas dengan cara menciptakan "sintetis" untuk menangani masalah kelas tidak seimbang. Teknik SMOTE didasarkan atas *nearest neighbors* oleh

jarak Euclidean antara titik data dalam *feature space* atau ruang variabel. SMOTE-NC atau *Synthetic Minority Over-Sampling Technique-Nominal Continous* adalah teknik khusus yang dapat menangani dataset bertipe kontinyu dan juga kategorik. Misalkan akan dihitung jarak antara x_i dan y_i, maka jarak x_i dan y_i adalah:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

Pembangkitan observasi sintesis dilakukan dengan menggunakan rumus:

$$x_{syn} = x_i + (x_{knn} - x_i)\gamma$$

x_{syn} adalah pengamatan baru hasil pembangkitan, x_i adalah pengamatan ke-i, x_{knn} adalah x terdekat dari x_i, γ adalah bilangan acak antara 0 dan 1. Pada data kategorik data sintetis adalah modus antara vektor amatan dengan vektor *k-nearest neighbor,* apabila terdapat nilai yang sama maka dipilih secara acak (Chawla *et al*., 2002).

2.5. *Hyperopt*

Hyperopt menyediakan algoritma dan infrastruktur *software* untuk menjalankan optimasi *hyperparameter* pada algoritma *machine learning* (Bergstra *et al*., 2015). *Tree-structured Parzen Estimator* adalah bentuk strategi yang dipakai *Hyperopt*. TPE mendefinisikan *p*(*x*|*y*) atau probabilitas *hyperparameter* menggunakan 2 densitas berikut :

$$p(x|y) = \begin{cases} l(x) & if\ y < y^* \\ g(x) & if\ y \geq y^*, \end{cases}$$

l(*x*) adalah densitas yang terbentuk dari observasi $\{x^{(i)}\}$ yang memiliki *loss* $f(x^{(i)})$ kurang dari y^*, dan $g(x)$ adalah densitas yang terbentuk dari observasi sisanya. Algoritma TPE memilih *y** untuk menjadi kuantil γ dari nilai observasi *y*, sehingga $p(y < y^*) = \gamma$. Kombinasi *hyperparameter* yang menghasilkan nilai tertinggi di bawah $l(x)/g(x)$ akan dievaluasi pada fungsi obyektif.

2.6. *Gradient Boosting*

Gradient boosting termasuk *supervised learning* berbasis *decision tree* yang dapat digunakan untuk klasifikasi. Algoritma *gradient boosting* bekerja secara sekuensial menambahkan prediktor sebelumnya yang kurang cocok dengan prediksi ke *ensemble*, memastikan kesalahan yang dibuat sebelumnya diperbaiki. Penggambaran sederhana konsep *ensemble* adalah keputusan-keputusan dari berbagai mesin pembelajaran digabungkan, kemudian untuk kelas yang menerima mayoritas 'suara' adalah kelas yang akan diprediksi oleh keseluruhan *ensemble*. *Gradient boosting* dimulai dengan menghasilkan pohon klasifikasi awal dan terus menyesuaikan pohon baru melalui minimalisasi fungsi kerugian (Natekin dan Knoll, 2013):

$$-\log L1 = -\sum_{i=1}^{N} y_i \log(odds) + \log(1 + e^{\log(odds)})$$

2.7. Pengukuran Hasil Prediksi

Pengukuran hasil klasifikasi dapat menggunakan akurasi yang dihitung dengan rumus:

$$\text{Akurasi} = \frac{Banyak\ prediksi\ benar}{Banyak\ total\ prediksi}$$

Ukuran kebaikan hasil klasifikasi lain adalah *confusion matrix*. Matriks ini berisi 4 istilah penting yaitu *True Positive* adalah jumlah observasi kelas positif yang tepat diklasifikasikan pada kelas positif, *True Negatives* adalah observasi kelas negatif yang tepat diklasifikasikan pada kelas negatif, *False Positive* adalah jumlah observasi negatif yangs alah diklasifikasikan sebagai kelas positif, sedangkan *False Negative* adalah jumlah observasi positif yang salah diklasifikasikan sebagaik kelas negatif (Chawla *et al*., 2002). Ukuran kebaikan hasil klasifikasi lain adalah AUC yang menunjukkan kemampuan metode klasifikasi dalam membedakan antar kelas. Semakin tinggi AUC, semakin baik model memprediksi dengan tepat pada masing-masing kelas.

Precision menunjukkan ukuran keakuratan model dalam memprediksi kelas positif, atau dengan kata lain untuk setiap baris yang diprediksi sebagai kelas positif, seberapa persen yang benar. Berikut adalah rumus *precision:*

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

3. METODE PENELITIAN

Data yang digunakan dalam penelitian adalah data status keberhasilan telemarketing produk bank deposito berjangka berdasarkan profil calon pelanggan dan beberapa tambahan variabel berkaitan dengan keadaan sosial ekonomi Data terdiri dari 21 variabel dan 41188 baris. Variabel dependen berupa status keberhasilan kegiatan telemarketing produk bank deposito berjangka. Sementara itu variabel independen sebanyak 20.

Pengolahan data dilakukan menggunakan *Google Colaboratory* (Colab). Colab adalah pengembangan *environment* Python yang dijalankan dalam *browser* menggunakan *Google Cloud*. Data yang telah diperoleh kemudian dianalisis sebagai berikut:

1. Memasukkan data ke dalam Colab dari *Google Drive*.
2. Melakukan pengecekan *missing value*. Apabila pada baris data observasi terdapat *missing value,* maka data observasi dihapus.
3. Melakukan pengecekan *outlier*. Jika terdapat *outlier* maka ditangani dengan mempertimbangkan penyebabnya. Jika sebabnya karena *human error,* maka dihapus, jika murni karena sifat data maka tidak dibuang.
4. Melakukan pembagian data latih dan data uji dengan *holdout validation* dengan proporsi 80:20.
5. Pengecekan *imbalance class* pada data latih. Apabila mengalami *imbalance class* maka diterapkan metode SMOTE-NC untuk menyeimbangkan kelas data.
6. Menerapkan *label encoding* pada data latih dan data uji untuk mengubah variabel kategorik huruf menjadi kategorik numerik.
7. Melakukan *tuning hyperparameter* dengan *hyperopt* pada data latih.
8. Menerapkan metode *gradient boosting* untuk membentuk pohon klasifikasi.
9. Melakukan evaluasi hasil klasifikasi menggunakan akurasi, *confusion matrix,* AUC dan *precision*.
10. Melakukan prediksi data baru menggunakan model *gradient boosting* yang sudah terbentuk.

4. HASIL DAN PEMBAHASAN

4.1. Data *Pre-processing*

Tahap awal data *pre-processing* dapat dilakukan pendeteksian *missing value*, setelah diamati maka diketahui bahwa tidak ada nilai yang hilang sehingga dilanjutkan dengan pendeteksian *outlier*. Apabila data tidak terletak pada rentang antara (Q1 – 1,5 IQR) dan (Q3 + 1,5 IQR) maka disebut *outlier*. Pemeriksaan menunjukkan terdapat *outlier* pada beberapa atribut, namun nilai *outlier* yang muncul bukan karena kesalahan namun memang nilai asli variabel tersebut yang berada di luar distribusi normal. Oleh karena itu peneliti memutuskan untuk membiarkan nilai *outlier* yang ada untuk diolah dalam analisis selanjutnya. Namun untuk variabel *pdays* yang adalah jumlah hari sejak klien terakhir kali dihubungi terdapat rentang data yang sangat ekstrim, terdapat banyak nilai 999 merupakan simbol untuk klien yang sudah sangat lama tidak dihubungi sehingga tidak dicatat jumlah harinya. Masalah ini dapat menganggu kinerja algoritma sehingga akan dilakukan pembagian kategori pada variabel *pdays*. Kategori akan ada 2 yaitu periode lama (jumlah hari ≥100) dan periode pendek (jumlah hari < 100).

4.2. SMOTE-NC

Dalam penelitian ini banyak data yang berada di kelas ‘no’ atau menolak penawaran adalah sebesar 29182 sedangkan data yang berada di kelas ‘yes’ atau menerima penawaran adalah sebesar 3768. Tidak ada ukuran tertentu sebuah dataset dikatakan memiliki kelas tidak seimbang, namun jika melihat pada kasus penelitian ini maka bisa dikatakan kondisi kelas tidak seimbang karena selisih yang sangat besar pada jumlah data di kedua kelas. Maka dari itu digunakan SMOTE-NC pada data latih untuk menghindari algoritma hanya belajar dari kelas dominan. Jumlah data pasca *oversampling* SMOTE-NC adalah 29222 baris untuk setiap kelas.

4.3. *Tuning Hyperparameter* dengan *Hyperopt*

Penentuan kombinasi parameter terbaik menggunakan *hyperopt* dengan spesifikasi masing-masing parameter sebagai berikut:

Tabel 1. *Search Space* Parameter *Gradient Boosting*

Parameter	Ekspresi parameter
learning_rate	*hp.loguniform('learning_rate',0,2)*
n_estimators	*hp.randint('n_estimators',500)*
max_depth	*hp.quniform('max_depth', 1, 10, 1)*
min_samples_split	*hp.uniform('min_samples_split', 0, 0.5)*
min_samples_leaf	*hp.uniform('min_samples_leaf', 0, 0.5)*

Penentuan parameter terbaik yang meminimalkan fungsi obyektif negatif log-likelihood dilakukan dalam 100 iterasi. Parameter hasil *tuning hyperparameter* tersedia pada Tabel 2.

Tabel 2. Parameter Terbaik Hasil Tuning

Nama parameter	Nilai parameter
learning_rate	0,5882
n_estimators	405
max_depth	7
min_samples_split	0,0924
min_samples_leaf	0,0113

4.4. *Gradient Boosting*

Model pohon klasifikasi yang dibentuk menggunakan metode *gradient boosting* memiliki *learning rate* 0,5882, kedalaman pohon 7, sampel minimal di *leaf node* 0,0113, sampel minimal untuk *split node* 0,0924, banyak pohon 405. Hasil klasifikasi dengan metode *gradient boosting* tersedia pada Tabel 3.

Tabel 3. Confusion Matrix Hasil Klasifikasi Gradient Boosting pada Data Latih dan Uji

Data Latih			Data Uji		
Aktual	Prediksi		Aktual	Prediksi	
	Ya	Tidak		Ya	Tidak
Ya	28293	946	Ya	6886	423
Tidak	720	28519	Tidak	369	560

Tabel 3 menunjukkan bahwa metode *gradient boosting* melakukan klasifikasi pada kedua kelas dengan relatif baik pada data latih. Hal ini terlihat dari jumlah kesalahan prediksi yang jauh lebih sedikit daripada yang benar diprediksi. Sedangkan pada data uji, model terlihat sangat baik memprediksi kelas positif, namun selisih kesalahan prediksi dengan yang benar pada kelas negatif relatif kecil, capaian ini masih kurang dibandingkan pada data latih.

Tabel 4. Ukuran Kebaikan Model Gradient Boosting

Ukuran Kebaikan Model	Data	
	Latih	Uji
Akurasi	97,15%	90,39%
Precision	97,52%	94,91%
AUC	0,997	0,939

Akurasi data latih menunjukkan bahwa *gradient boosting* dapat memprediksi dengan benar sebesar 97,15% dari total jumlah data. Namun pada data uji, akurasinya menurun hingga menjadi 90,39%, meskipun angka ini relatif tergolong baik namun bisa dikatakan dalam pelatihan data terindikasi *overfitting*. Nilai *precision* menjelaskan tentang jumlah yang diprediksi benar dari kelas positif. Pada data latih sebanyak 97,52% data benar diprediksi dari kelas positif. Sedangkan pada data uji sebanyak 94,91% data benar diprediksi dari kelas positif. Nilai yang diperoleh pada kedua dataset menunjukkan bahwa *gradient boosting* sanggup memprediksi dengan baik calon target pelanggan yang hendak menerima tawaran produk telemarketing bank.

Nilai AUC menunjukkan kebaikan metode klasifikasi dalam membedakan kelas positif dan negatif. Nilai AUC pada data latih mencapai 0,997 yang mana skor ini mendekati nilai sempurna 1. AUC pada data uji adalah sebesar 0,939, nilai ini juga masih tergolong relatif baik sehingga dapat dikatakan metode *gradient boosting* dapat membedakan kelas positif dan negatif dengan baik.

5. KESIMPULAN

Berdasarkan hasil dan pembahasan maka dapat diketahui bahwa *hyperopt* dapat digunakan untuk *tuning hyperparameter* dengan *search space* yang luas. Metode *gradient boosting* yang diterapkan untuk memprediksi keberhasilan telemarketing menggunakan

hyperparameter hasil *tuning* menghasilkan akurasi klasifikasi 90,39%, nilai *precision* sebesar 94,91%, dan nilai AUC sebesar 0,939 pada data uji. Pencapaian ini relatif sangat baik sehingga dapat disimpulkan bahwa model dapat mengklasifikasi dan memprediksi kedua kelas dengan baik.

DAFTAR PUSTAKA

Bergstra, J., Komer, B., Eliasmith, C., Yasmins, D., Cox, D. D. 2015. *Hyperopt: A Python library for model selection and hyperparameter optimization*. Computational Science and Discovery Vol. 8, No. 1.

Chawla, N. V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. 2002. *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research Vol 16, Hal. 321–357.

Essam, A. D. 2019. *Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset*. International Journal of Computer and Information Engineering Vol. 13, No.1 : Hal. 6–10.

Moro, S., Cortez, P., Rita, P. 2014. *A data-driven approach to predict the success of bank telemarketing*. Decision Support Systems Vol 62, Hal 22–31.

Natekin, A. Knoll, A. 2013. *Gradient boosting machines, a tutorial*. Frontiers in Neurorobotics.

Raschka, S. 2018. *Model evaluation, model selection, and algorithm selection in machine learning*. arXiv.

Rust, R. T., Moorman, C. and Bhalla, G. 2010. *Rethinking marketing*, Harvard Business Review.

Stone, B. 1992. *Successful Telemarketing by Bob Stone, John Wyman*. New York: McGraw-Hill Professional.

Van der Heijden, G. J. M. G., Donders, A, R, T., Stijnen, T., Moons, K.G.M. 2006. *Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example*. Journal of Clinical Epidemiology Vol 59, No. 10 : Hal. 1102–1109.

sional-2021-Gaussian_Vol_10_No_4_Penerapan_Gradient_Boosting.pdf

ORIGINALITY REPORT

12%	9%	3%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Tadulako Student Paper	4%
2	exsight.id Internet Source	2%
3	download.garuda.ristekdikti.go.id Internet Source	1%
4	Submitted to Universitas Brawijaya Student Paper	1%
5	join.if.uinsgd.ac.id Internet Source	1%
6	doku.pub Internet Source	1%
7	lab_adrk.ub.ac.id Internet Source	1%
8	mdpi.com Internet Source	1%
9	journal.lembagakita.org Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		