Jurnal Teknik Informatika (JUTIF)
Vol. 5, No. 3, June 2024, pp. 661-667

DOI: https://doi.org/10.52436/1.jutif.2024.5.3.1284
p-ISSN: 2723-3863
e-ISSN: 2723-3871

ANALYSIS OF THE EFFECTIVENESS OF POLYNOMIAL FIT SMOTE MESH ON IMBALANCE DATASET FOR BANK CUSTOMER CHURN PREDICTION WITH XGBOOST AND BAYESIAN OPTIMIZATION

Jhiro Faran[1], Agung Triayudi[*2]

[1,2]Post Graduate Information Technology Department, Universitas Nasional, Indonesia
Email: [1]jhirofaran2022@student.unas.ac.id, [2]agung.triayudi@gmail.com

(Article received: Auguts 04, 2023; Revision: September 26, 2023; published: May 18, 2024)

Abstract

The case of churn in the banking industry, namely customers who leave or no longer use bank services, is a serious problem that requires an appropriate solution. The aim of this research is to predict churn and take appropriate preventive actions using machine learning. The dataset contains 10,000 bank customer data with 14 relevant features. Only about 20% of customers experience churn, creating a data imbalance problem in classification. To overcome data imbalances, the SMOTE oversampling technique was applied. Also introduced was the development of the SMOTE technique, namely, Polynomial Fit SMOTE Mesh (PFSM). PFSM works by combining each point in the data with a linear function and producing synthetic data at each connected distance. Experimental results show that the model developed using PFSM and optimized with Bayesian Optimization for the XGBoost algorithm achieved 86.1% accuracy, 70.87% precision, 53.81% recall, and 61.17% F-score. This indicates that the approach is successful in improving predictive capabilities and identifying potential customers for churn earlier. This research has significant relevance in the banking industry, helping banks to safeguard their customers and improve banking business performance..

Keywords: *Bayesian Optimization, Churn, Imbalance data, SMOTE, XGBoost.*

ANALISIS EFEKTIVITAS POLYNOMIAL FIT SMOTE MESH PADA IMBALANCE DATASET UNTUK CHURN PREDICTION CUSTOMER BANK DENGAN XGBOOST DAN BAYESIAN OPTIMIZATION

Abstrak

Kasus churn dalam industri perbankan, yaitu nasabah yang keluar atau tidak lagi menggunakan layanan bank, merupakan masalah serius yang memerlukan solusi yang tepat. Tujuan penelitian ini adalah untuk memprediksi churn dan mengambil tindakan preventif yang sesuai menggunakan machine learning. Dataset berisi 10.000 data nasabah bank dengan 14 fitur yang relevan. Hanya sekitar 20% dari nasabah yang mengalami churn, menciptakan masalah ketimpangan data dalam klasifikasi. Untuk mengatasi ketimpangan data, diterapkan teknik oversampling *SMOTE*. Juga diperkenalkan pengembangan teknik *SMOTE* yaitu, *Polynomial Fit SMOTE Mesh* (PFSM). PFSM bekerja dengan cara menggabungkan setiap titik pada data dengan fungsi linier dan menghasilkan data sintetis pada setiap jarak yang dihubungkan. Hasil eksperimen menunjukkan bahwa model yang dikembangkan menggunakan PFSM dan dioptimalkan dengan Bayesian Optimization untuk algoritma XGBoost mencapai akurasi 86,1%, presisi 70,87%, recall 53,81%, dan F-score 61,17%. Ini mengindikasikan bahwa pendekatan berhasil meningkatkan kemampuan prediksi dan mengidentifikasi nasabah yang potensial untuk churn lebih awal. Penelitian ini memiliki relevansi signifikan dalam industri perbankan, membantu bank-bank untuk menjaga nasabah mereka dan meningkatkan kinerja bisnis perbankan.

Kata kunci: *Bayesian Optimization, Churn, Imbalance data, SMOTE, XGBoost.*

1. PENDAHULUAN

Customer churning merupakan sebutan bagi para pelanggan yang tidak ingin menggunakan lagi suatu produk atau keluar untuk berlangganan produk tersebut [1]. Untuk melakukan tindakan preventif perusahaan mencari cara untuk memprediksi nasabah yang mungkin akan melakukan *churn*. Sehingga perusahaan melakukan kombinasi dari analisis data, *data mining*, dan membangun model ML (*Machine Learning*) [2].

Data yang digunakan sering sekali ditemukan tidak seimbang atau imbalance. *Imbalance data* terjadi karena ada label data yang memiliki anggota lebih sedikit dibandingkan label lainnya. Sehingga disebut dengan kelas minoritas [3]. Alasan penanggulanan *imbalance* dataset pada karena teknik pembelajaran *machine learning* memiliki tendensius terhadap kelas mayoritas dan wilayah yang padat [4]. Untuk melakukan penanggulang CIP (*Class Imbalance Problem*) digunakannya DSTs (*Data Sampling Techniques*) [5]. Teknik seperti oversampling dan undersampling maupun kombinasi keduanya dapat menanggulangi masalah tersebut. Teknik oversampling digunakan untuk menghasilkan data sintetis dari kelas minoritas [6].

SMOTE merupakan salah satu teknik oversampling. Teknik ini meningkat jumlah kelas minoritas sehingga mendapatkan keseimbangan antar kelas [7]. SMOTE perlu menghasilkan data sintetis yang besar (±500%) untuk meningkatkan efektifitas model dalam melakukan klasifikasi kelas minoritas [8]. Dengan perkembangan SMOTE menghasilkan Polynomial Fit SMOTE (PFS) yang dapat memberikan hasil yang sangat baik dalam penanggulanang Imbalance Dataset [4], [9].

Extreme Gradient Boosting (XGBoost) merupakan salah satu *ensemble learning algorithm* yang memiliki keuntungan dalam fleksibilitas tinggi, prediksi kuat, generalisasi kuat, skalabilitas tinggi, efesien dan menghasilkan model yang kokoh [10]. Untuk meningkatkan akurasi dan mencegah model menjadi *overfitting* dilakukan reduksi fitur atau biasa disebut *dimensional reduction*. Dalam tahap *modeling* ML tahap reduksi ini berfungsi untuk mengurangi kompleksitas dan mempertahankan sebagian besar pola yang ada [11].

Selanjutnya Untuk meningkatkan performa model *machine learning* dapat menggunakan *Hyperparameter Tuning* [12]. *Bayesian Optimization* merupakan salah satu teknik optimasi parameter ML. Prosesnya lebih cepat dalam melakukan *tuning* parameter. Optimasi ini dilakukan untuk meningkatkan akurasi model serta mencegah model terjadi *overfitting* [13].

Stephane dan kawan-kawan melakukan *churn prediction* dengan menggunakan beberapa algoritma *machine learning* yang dikombinasikan dengan SMOTE dan Ensemble Method. Hasilnya SMOTE dan RF (*Random Forest*) mendapatkan hasil yang paling maksimal dengan nilai akurasi 86% [14].

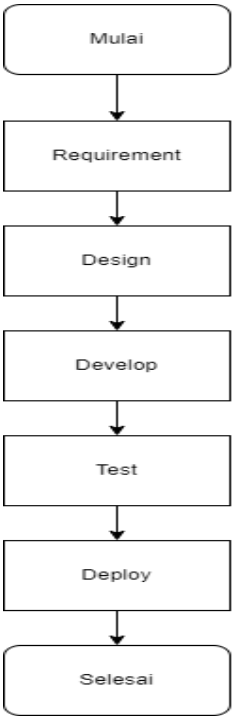
Praveen dan kawan-kawan melukan analisis *churn prediction* dengan pendekatan beberapa model ML machine learning. Penelitian ini membandingkan beberapa algoritma seperti XGBoost, SVM (*Suport Vector Machine*) dan RF. Hasilnya XGBoost mendapatkan hasil memuaskan dengan akurasi 80,8%, recall 82,2%, precision 81,2% dan AUC 82% [15].

Penelitian ini akan melakukan penanggulangan imbalance dataset menggunakan PFSM (*Polynomial Fit SMOTE Mesh*) dan melakukan optimasi parameter XGBoost dengan BO (*Bayesian Optimization*) pada model ML. Dataset yang digunakan oleh penulis sebanyak 10000 baris dan 14 fitur yang diambil dari himpunan data https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling dan hasil evaluasi matriks yang digunakan adalah *accuracy*, *precision*, *recall* dan *F-score*. Hasil penelitian ini berfokus pada efektifitas PFSM dan BO pada XGBoost.

Pentingnya penelitian ini untuk perbankan harus memahami dan mengatasi ketimpangan data yang sering terjadi dalam masalah churn. Hal ini karena hanya sekitar 20% dari nasabah yang mengalami churn, menciptakan ketidakseimbangan dalam data yang dapat mengganggu kinerja model klasifikasi.

2. METODE PENELITIAN

Berikut desain penelitian yang akan dilakukan agar mencapai tujuan penelitian.



Gambar 1. Desain penelitian

Berdasarkan gambar 1 dapat dilihat desain penelitian terdiri dari *requirement* menyiapkan *tools* (perangkat), *design* membuat sketsa atau *mock up* sistem yang ingin dibuat, *develop* membangun sistem berdasarkan *design* yang telah dibuat, *test* melakukan uji coba pada sistem yang telah dibangun, dan *deploy* tahap ini merupakan hasil analisis sistem yang telah dibangun.

2.1. *Requirements*

Pada tahap ini penulis melakukan pengumpulan data dan *tools* yang akan digunakan dalam penelitian sehingga penelitian dapat dilakukan dan mencapai

tujuan penelitian. Berikut perangkat yang digunakan penulis.

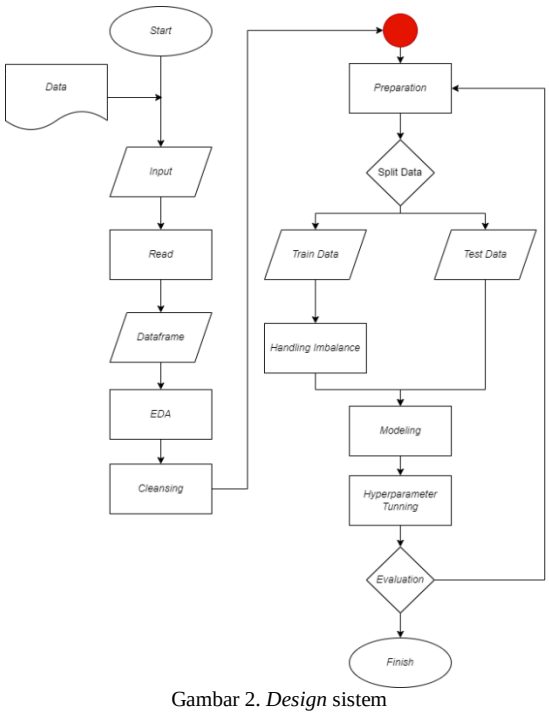
Tabel 1. Perangkat sistem

Perangkat	Keterangan
Kaggle	Sumber data
Python	Bahasa pemograman
Colab	Tempat menulis baris code
Pandas	Manipulasi data
Numpy	Array
Seaborn, matplotlib	Visualisasi data
Sklearn	*Preparation, modelling*, dan evaluasi
Smote Varians	*Handling imbalance data*

Berdasarkan tabel 1 dapat diketahui perangkat apa saja yang digunakan untuk mencapai tujuan penelitian.

2.2. *Design*

Tahap selanjut adalah melakukan *design* sistem. Dalam penelitian ini sistem yang dibangun merupakan model machine learning. Berikut sistem yang akan dibangun oleh penulis.



Gambar 2. *Design* sistem

Berdasarkan gambar 2 sistem yang dikembangkan akan mengacu pada *design* sistem tersebut. Tahap awal dilakukan dalam membaca data dan melakukan *Exploratory Data Analysis* (EDA). Tahap selanjutnya adalah membersihkan data, *data preparation*, pembagian data, dan melakukan *handling outlier*. Tahap terakhir adalah membangun model serta melakukan optimasi parameter pada model dan melakukan evaluasi model.

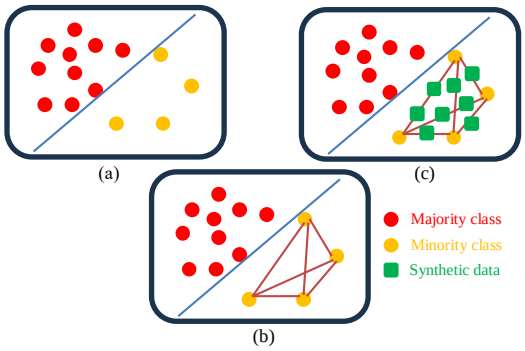
2.3. *Development*

EDA – mencari *insight* dari dataset untuk mendapatkan pola pada dataset, mendapatkan data secara deskriptif, melakukan uji korelasi baik dari data kategorikal dan rasio,dan melakukan visualisasi data. Secara garis besar proses dimana memahami dataset yang akan digunakan.

Cleansing – membersihkan data dari duplikasi data, menaggulangi *missing value*, dan *handling outlier*. Tahap ini dilakukan agar terhindar dari inkonsistensi data.

Preparation – menyiapkan data untuk melakukan analisis prediksi seperti *label encoding* atau *target encoding* untuk data categorical, *feature selection*, dan *scaling data* menggunakan *minmax scaller*.

Split Data – proses membagi data menjadi dua satu untuk pembelajaran model lainnya digunakan menjadi data uji pada model yang telah dibuat. Pembagian dataset pada penelitian ini menggunakan perbandingan 80:20, 80 untuk *data train* dan 20 untuk *data test*.



Gambar 3. (a) Penyebaran data (b) setiap node kelas minoritas dihubungkan (c) data sintetis

Handling Imbalance – berdasarkan gambar 3 pennaggulangan *imbalance data* denan PFSM. PFSM mirip dengan polynom PFSB (Polynom Fit SMOTE Bus) dimana setiap node titik dihubungkan. Hubungan tersebut menggunakan fungsi linier.

$$f(x) = ax + b \qquad (1)$$

Selanjutnya dengan menentukan koefisiensi "a" dan "b" untuk menyamakan dengan jarak "f(x)" antar sample. Setelahnya melakukan generasi k secara linier pada value x_k $(k \in [-1,1])$ mengacu pada tingkat oversampling.

Modelling – pada penelitian ini penulis menggunakan XGBoost sebagai teknik modeling yang digunakan. XGBoost memiliki kecepatan dan kinerja tinggi [16]. XGboost mampu mengelola data kategorikal dan numerical. Secara garis besar XGBoost menggunakan sistem *ensemble learning* atau menggabungkan hasil prediksi beberapa model untuk menghasilkan keputusan prediksi terbaik.

Hyperparameter Tunning – teknik ini digunakan untuk memperbaiki model dan mencegah model agar tidak terjadi overfitting dan meningkatkan akurasi model ML. Penelitian ini menggunakan *BO* sebagai optimasi parameter. Cara kerja optimasi ini mencari fungsi target, memodelkan perkiraan fungsi,

dan melakukan akuisisi atau menjelajah ke wilayah baru.

$$D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\} \quad (2)$$

Pengamatan awal x merupakan parameter yang yang akan dilakukan optimasi dan y merupakan fungsi yang ingin dioptimasi $y = f(x)$.

$$M(x) \sim GP(\mu(x), \sigma^2(x)) \quad (3)$$

Langkah selanjut adalah pembaruan model Surrogate dimana $\mu(x)$ adalah nilai rata-rata dan $\sigma^2(x)$ adalah varians dari model surrogate pada titik x.

$$x_{next} = argmax_x A(x; D.M(x)) \quad (4)$$

Langkah akuisisi ini untuk menentukan titik x yang akan dievaluasi selanjutnya. $argmax_x A$ merupakan pencarian titik x dengan memaksimalkan nilai dari fungsi A(x).

$$y_{next} = f(x_{next}) \quad (5)$$

Tahap melakukan evaluasi pada fungsi target. Selanjutnya adalah menambahkan data (x_{next}, y_{next}) dalam data obersevasi dan mengulangi sampai menyelesaikan iterasi.

Evaluasi – pada tahap ini melakukan evaluasi dengan menggunakan *confusion matrix* dan menghitung nilai *accuracy*, *precision*, *recall* dan *F-score*.

2.4. *Test*

Setelah mengembangkan model, model akan dilakukan test dengan data test. Apakah model dapat melakukan prediski dengan baik dan tidak mengalami overfitting.

2.5. *Deploy*

Proses dimana melakukan analisis terhadap model yang telah dibuat. Penelitian ini akan membandingkan model yang tidak melakukan oversampling, model melakukan sampling tanpa optimasi, dan model yang mengkombinasikan oversampling dan optimasi.

3. HASIL

Hasil dalam penelitian ini hanya membandingkan efektivitas oversampling data menggungakan PFSM dan BO pada algoritma XGBoost.

3.1. Dataset

Setelah melakukan penarikan data dan melakukan obeservas pada data. Data terdiri dari 10000 baris dan 14 kolom. Berikut hasil observasi pada data.

Tabel 2. Dataset cutomer bank

Nama	Tipe Data	*Missing Value*	Data
RowNumber	Integer	0	Rasio
CustomerId	Integer	0	Unik ID
Surname	Object	0	Nominal
CreditScore	Integer	0	Rasio
Geography	Object	0	Kategori
Gender	Object	0	Kategori
Age	Integer	0	Rasio
Tenure	Integer	0	Rasio
Balance	Float	0	Rasio
NumofProduct	Integer	0	Rasio
HasCrCard	Integer	0	Kategori
IsActiveMember	Integer	0	Kategori
EstimatedSalary	Float	0	Rasio
Exited	Integer	0	Kategori

Berdasarkan tabel 2 kolom "*RowNumber*" "*CustomerId*", dan "*Surname*" akan dieleminasi karena tidak memiliki keterkaitan pada modeling ML, Target prediksi atau label prediksi akan dilakukan pada kolom "*Exited*".

3.2. *Exploratory Data Analysist*

Data deskriptif akan dilakukan hanya kepada tipe data rasio. Berikut hasil dari data deskriptif dari dataset.

Berdasarkan tabel 3 didapatkan nilai *mean*, minimal, kuartil-1, kuartil-2, kuartil-3, dan maksimal. Berdasarkan nilai-nilai tersebut kita dapat mengetahui penyebaran pada setiap data dan dapat membantu untuk mencari *outlier*.

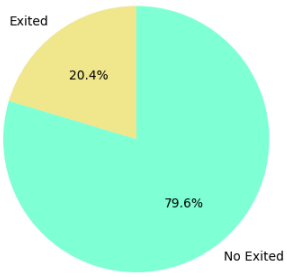
Proses selanjutnya mencari informasi atau *insight* dari data dengan cara melakukan visualisasi data. Berikut beberapa visualisasi yang digunakan.

Tabel 3. Data deskriptif

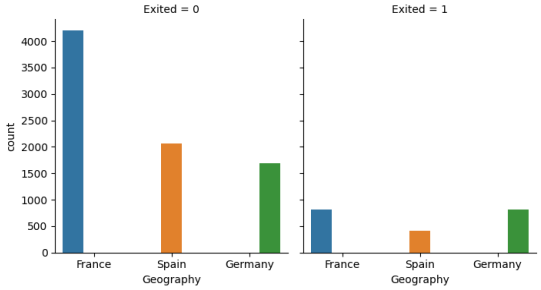
Fitur	Mean	Min	Q1	Q2	Q3	Max
Credit Score	650,5	350	584	652	718	850
Age	38,9	18	32	37	44	92
Tenure	5,01	0	3	5	7	10
Balance	76486,889	0	0	97198,54	127644,24	250889,09
Numof Product	1,53	1	1	1	2	4
Estimated Salary	100090,239	11,58	51002,11	100193,91	149388,247	199992,48

Berdasarkan gambar 4 *customer* yang melakukan *churn* sebanyak 20,4% dari jumlah data dan sisanya masih menggunakan produk tersebut. Hasil tersebut memberikan informasi terjadi ketidak seimbangan antar label atau *imbalance data*.

Gambar 5 memberikan informasi tentang perbandingan *customer* yang melakukan *churn* dari segi "*Geography*". Dalam segi "*Geography*" terjadinya *churn* terbanyak berada pada negara Jerman dan Perancis.



Gambar 4. Perbandingan jumlah label

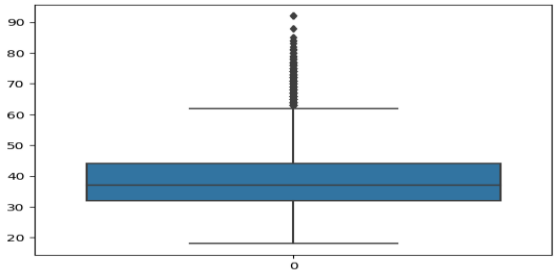


Gambar 5. Perbandingan jumlah orang berdasarkan label dan *geography*

Berdasarkan tabel 4 persentase *customer* yang melakukan churn pada setiap negara. Tabel ini akan menggantikan *geography* saat pengolahan data untuk *modelling*.

Tabel 4. Persentase *customer* melakukan churn dari setiap negara

Geography	**Persentase Exit**
Germany	32,44%
France	16,15%
Spain	16,67%



Gambar 6. Visualisasi outlier pada fitur "*Age*"

Berdasarkan pengecekan fitur "*Age*" pada gambar 6 mempunyai nilai *outlier*. Nilai *outlier* tersebut hanya ada disisi *upper*. Penanggulangan *outlier* dilakukan dengan cara memanipulasi nilai tersebut.

Tabel 5. Chi square independensi antara fitur kategorikal terhadap label

Fitur kategorikal	**P-Value**
Gender	2,248e-26
Geography	3,830e-66
HasCrCard	0,471
IsActiveMember	8.785e-55

Berdasarkan tabel 5 dilakukan uji *chi square* idenpendensi, ini dilakukan untuk mengetahui keterkaitan varibel kategorikal dengan kelas label. Hasilnya fitur "HasCrCard" tidak memiliki keterkaitan dengan label karena *P-value* > 0,05.

Tabel 6. Uji pearson dan speaman pada fitur rasio terhadap label

Fitur Rasio	**Pearson**	**Spearman**
Age	0,29	0,32
CreditScore	-0,027	-0,023
Balance	0,12	0,11
Tenure	-0,014	-0,014
Estimated Salary	0,012	0,012
NumOfProduct	-0,048	-0,13

Berdasarkan data yang didapatkan dari tabel 6 fitur-fitur rasio memiliki korelasi yang rendah pada label. Hanya fitur "*Age*" dan "*Balance*" tidak yang memiliki korelasi positif pada label.

3.3. *Cleansing Data*

Tahap ini hanya melakukan penanggulangan pada *outlier* karena tidak terdapat *missin value*. Fitur "*Age*" memiliki outlier berdasarkan visualisasi boxplot gambar 6. Outlier tersebut akan dimanipulasi nilainya menjadi kuartil-3 (Q_3).

3.4. *Data Preparation*

Berdasarkan *EDA* yang telah dilakukan proses selanjutnya menyeleksi fitur yang akan digunakan pada permodelan. Fitur yang digunakan dalam penelitian ini "*Gender*", "*Geography*", "*Age*", "*Balance*", "*NumOfProduct*", "*IsActiveMember*", dan "*Exited*" (target label).

Tabel 7. Penanggulangan data kategorikal

Fitur	**Metode**
Gender	*One Hot Encoding*
Geography	*Target Encoding*

Dari tabel diatas fitur "*Geography*" akan menggunakan *Target Encoding*. Metode ini memanipulasi dengang menggantikan *feature value* dengan persentase atau peluang *customer* melakukan churn berdasarkan fitur *"Geography"*. Persentase dapat dilihat pada tabel 4.

Proses selanjutnya merupakan proses *scaling* data. *Scaling data* pada penelitian ini menggunakan metode *minmax scaler*. Fitur yang diseleksi "*Balance*" dan "*Age*" karena memiliki rentang yang besar dengan kolom lainnya.

Tabel 8. Pembagian data *train* dan *test*

Data	**Total**	***Not_churn***	***Churn***
Train	8000	6370	1630
Test	2000	1593	407

Berdasarkan tabel 8 baik data train dan test mengalami *imbalance data*, sehingga model akan cenderung atau lebih baik memprediksi kelas mayoritas (*Not_churn*). Data *train* selanjutnya akan menggunakan PFSM untuk menanggulangi *imbalance data* yang terjadi.

Tabel 9. Data train setelah dilakukan oversampling menggunakan PFSM

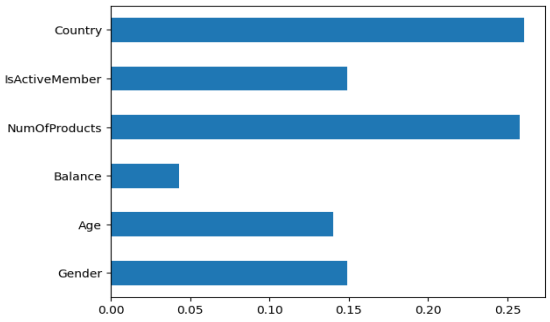
Data	**Total**	***Not_churn***	***Churn***
Train_PFSM	12740	6370	6370
Test	2000	1593	407

Dari tabel 9 setelah dilakukan oversampling label *churn* naik 290,47% dari data awal. Setelah melakukan oversampling penulis melakukan permodelan dengan kondisi; (1) Model menggunakan data *train-test* dengan parameter XGBoost *default* (2) Model menggunakan data *train-test* yang telah dilakukan *oversampling* dengan parameter XGBoost *default* (3) Model menggunakan data *train-test* yang telah dilakukan *oversampling* dengan parameter XGBoost menggunakan BO.

Tabel 10. Perbandingan evaluasi model

Model	*Accuracy*	*Recall*	*Precision*	*F-score*
Model 1	85,1%	50,12%	68,23%	57,79%
Model 2	84,85%	52,58%	66,05%	58,55%
Model 3	86,1%	53,81%	70,87%	61,17%

Dari tabel 10 model 3 mempunyai hasil terbaik dari segala aspek evaluasi. Penanggulangan *imbalance data* meningkatkan persentase pada evaluasi *recall*. Nilai *F-score* meningkat menandakan penanggulangan *imbalance dataset* mempengaruhi keseimbangan antara *recall* dan *precision*.



Gambar 7. *Feature importance* model 3

Berdasarkan gambar 7 fitur yang paling berpengaruh adalah *Country* (kategori) dan *NumOfProduct* (rasio). Fitur *Balance* berkontribusi paling rendah dalam membangun model. Fitur-fitur kategori mendominasi dalam membangun model ML.

4. DISKUSI

Berdasarkan penjabaran dan hasil yang didapatkan dalam penelitian ini penulis mengasumsikan jika teknik *oversampling* memberikan dampak positif pada keseimbangan nilai *recall* dan *precision*. Peningkatan *F-score* rata-rata ada pada 1% – 3%, tetapi akurasi mengalami penurunan jika tidak adanya optimasi tambahan pada parameter yang digunakan.

Setelah itu melakukan penilaian kepada setiap *feature* dalam mempengaruhi hasil prediksi. Hasilnya fitur kategorikal lebih berpengaruh daripada fitur rasio. Sehingga perlu diulas lebih lanjut untuk meningkatkan fitur-fitur yang ada dalam melakukan prediksi.

Stephane dan kawan-kawan melakukan prediksi churn modeling menggunakan dataset yang sama tetapi menggunakan semua fitur yang ada dan melakukan *handling imbalance data* menggunakan SMOTE. Hasilnya nilai akurasi *F-score* meningkat 30% menggunakan algoritma *Random Forest* dengan akurasi prediksi 86%. Pada waktu yang sama algoritma *Gradient Boosting* mengalami penurunan ketika melakukan penanggulanagan data yang tidak seimbang (*handling imbalance dataset*). Pada kasusnya penurunan akurasi dari 87% menjadi 84% tetapi nilai *F-score* mengalami peningkatan 25%..

5. KESIMPULAN

Berdasarkan penelitian yang sudah dilakukan pada penanggulangan *imbalance data* menggunakan *Polynomial Fit SMOTE Mesh* dengan algoritma XGBoost yang dioptimasikan menggunakan *Bayesian Optimization*, hasil akurasi meningkat tidak terlalu signifikan tetapi nilai *F-score* mengalami peningkatan yang sangat baik dari 57,79% menjadi 61,17%. Perlu diperhatikan penanggulangan *imbalance data* tanpa melakukan optimasi parameter, model tidak akan optimal dalam melakukan prediksi. Meningkatkan kinerja fitur dapat memberikan dampak positf dalam melakukan permodelan dalam XGBoost. Kedepannya penulis akan melakukan hal mendalam tentang pengaruh teknik oversampling pada tipe data yang digunakan.

DAFTAR PUSTAKA

[1] I. Kaur and J. Kaur, "Customer Churn Analysis and Prediction in Banking Industry using Machine Learning," in *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, IEEE, Nov. 2020, pp. 434–437. doi: 10.1109/PDGC50313.2020.9315761.

[2] P. Chen, N. Liu, and B. Wang, "Evaluation of Customer Behaviour with Machine Learning for Churn Prediction: The Case of Bank Customer Churn in Europe," in *Proceedings of the International Conference on Financial Innovation, FinTech and Information Technology, FFIT 2022, October 28-30, 2022, Shenzhen, China*, EAI, 2023. doi: 10.4108/eai.28-10-2022.2328450.

[3] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Inf Sci (N Y)*, vol. 513, pp. 429–441, Mar. 2020, doi: 10.1016/j.ins.2019.11.004.

[4] G. Kovács, "An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets," *Appl Soft Comput*, vol. 83, p. 105662, Oct. 2019, doi: 10.1016/j.asoc.2019.105662.

[5] M. Maw, S.-C. Haw, and C.-K. Ho, "Utilizing data sampling techniques on algorithmic fairness for customer churn prediction with data imbalance problems,"

F1000Res, vol. 10, p. 988, Jun. 2022, doi: 10.12688/f1000research.72929.2.

[6] M. Rahman and V. Kumar, "Machine Learning Based Customer Churn Prediction In Banking," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, IEEE, Nov. 2020, pp. 1196–1201. doi: 10.1109/ICECA49313.2020.9297529.

[7] A. Muneer, R. Faizan Ali, A. Alghamdi, S. Mohd Taib, A. Almaghthawi, and E. A. A. Ghaleb, "Predicting customers churning in banking industry: A machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, p. 539, Apr. 2022, doi: 10.11591/ijeecs.v26.i1.pp539-549.

[8] H. Al Majzoub, I. Elgedawy, Ö. Akaydın, and M. Köse Ulukök, "HCAB-SMOTE: A Hybrid Clustered Affinitive Borderline SMOTE Approach for Imbalanced Data Binary Classification," *Arab J Sci Eng*, vol. 45, no. 4, pp. 3205–3222, Apr. 2020, doi: 10.1007/s13369-019-04336-1.

[9] T. Wongvorachan, S. He, and O. Bulut, "A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining," *Information*, vol. 14, no. 1, p. 54, Jan. 2023, doi: 10.3390/info14010054.

[10] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int J Distrib Sens Netw*, vol. 18, no. 6, p. 155013292211069, Jun. 2022, doi: 10.1177/15501329221106935.

[11] B. Valarmathi, T. Chellatamilan, H. Mittal, J. Jagrit, and S. Shubham, "Classification of Imbalanced Banking Dataset using Dimensionality Reduction," in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, IEEE, May 2019, pp. 1353–1357. doi: 10.1109/ICCS45141.2019.9065648.

[12] V. Perrone, M. Donini, M. B. Zafar, R. Schmucker, K. Kenthapadi, and C. Archambeau, "Fair Bayesian Optimization," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: ACM, Jul. 2021, pp. 854–863. doi: 10.1145/3461702.3462629.

[13] S. M. Sina Mirabdolbaghi and B. Amiri, "Model Optimization Analysis of Customer Churn Prediction Using Machine Learning Algorithms with Focus on Feature Reductions," *Discrete Dyn Nat Soc*, vol. 2022, pp. 1–20, Jun. 2022, doi: 10.1155/2022/5134356.

[14] S. C. K. Tékouabou, Ștefan C. Gherghina, H. Toulni, P. N. Mata, and J. M. Martins, "Towards Explainable Machine Learning for Bank Churn Prediction Using Data Balancing and Ensemble-Based Methods," *Mathematics*, vol. 10, no. 14, p. 2379, Jul. 2022, doi: 10.3390/math10142379.

[15] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, Feb. 2022, doi: 10.1007/s00607-021-00908-y.

[16] A. Prabha, J. Yadav, A. Rani, and V. Singh, "Design of intelligent diabetes mellitus detection system using hybrid feature selection based XGBoost classifier," *Comput Biol Med*, vol. 136, p. 104664, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104664.