ANALISIS PENGARUH METODE OVER SAMPLING DALAM CHURN PREDICTION UNTUK PERUSAHAAN TELEKOMUNIKASI

ZK. Abdurahman Baizal[1], Moch. Arif Bijaksana[2], Angelina Sagita Sastrawan[3]
Telp (022)7564108 ext 2298 Fax (022)7565934
[1]Program Studi Ilmu Komputasi, Fakultas Sains Intitut Teknologi Telkom, Bandung
[2,3]Program Studi Teknik Informatika, Fakultas Teknik Informatika Institut Teknologi Telkom, Bandung
Jl Telekomunikasi, Terusan Buah Batu, Bandung
E-mail: zka@ittelkom.ac.id[1], mab@ittelkom.ac.id[2], angelina.sagita@yahoo.com[3],

ABSTRAK

Churn prediction adalah suatu cara untuk memprediksi pelanggan yang berpotensial untuk churn. Data mining, khususnya klasifikasi tampaknya dapat menjadi salah satu alternatif solusi dalam membuat model churn prediction yang akurat. Namun hasil klasifikasi menjadi tidak akurat disebabkan karena data churn bersifat imbalance. Kelas data menjadi tidak stabil karena data akan lebih condong ke bagian data yang memiliki komposisi data yang lebih besar. Salah satu cara untuk menangani permasalahan ini adalah dengan memodifikasi dataset yang digunakan atau yang lebih dikenal dengan metode oversampling. Analisis yang dilakukan pada penelitian ini adalah mengetahui bagaimana pengaruh metode oversampling yang digunakan terhadap akurasi prediksi data churn dengan melakukan penghitungan akurasi model churn prediction yang dinyatakan dalam bentuk lift curve, top decile dan gini coefficient serta f-measure untuk penghitungan akurasi prediksi data sebagai data yang imbalance. Hasil yang didapat dari penelitian menunjukkan bahwa metode oversampling yang menghasilkan data synthetic belum sesuai diterapkan pada data churn, karena cenderung masih menghasilkan nilai top decile yang kecil. Tetapi secara umum metode oversampling ini mampu meningkatkan akurasi untuk memprediksi data minor. Dengan penerapan metode oversampling, data churn yang memiliki tingkat imbalance yang besar dapat diklasifikasi tanpa mengorbankan data minor yang menjadi fokus penelitian. Metode oversampling yang digunakan juga memiliki hasil evaluasi yang berbeda terhadap dataset sebagai data churn dan sebagai data imbalance.

Kata kunci: churn prediction, imbalance, sampling, akurasi, evaluasi.

1. PENDAHULUAN

Tiap perusahaan memiliki cara tersendiri dalam menawarkan layanan yang berkualitas dengan harga seminimum mungkin. Hal ini diharapkan dapat menarik pelanggan sebanyak mungkin agar pendapatan yang masuk kas perusahaan semakin optimal pula. Hal-hal seperti diataslah yang menyebabkan fenomena *churn* terjadi, dimana para pelanggan cabut akibat tidak puas atas layanan yang diberikan, ataupun memutuskan untuk pindah dari satu *provider* ke *provider* yang lain karena tergiur penawaran fasilitas dan harga yang lebih menarik. Fenomena *churn* ini tentu saja meresahkan, karena jika tidak dicegah dan ditangani akan berakibat pada penurunan *revenue* perusahaan.

Data *churn* bersifat *imbalance class* sehingga kecenderungan kelas data menjadi tidak stabil karena data akan lebih condong ke bagian data yang memiliki komposisi data lebih besar.

Dalam penelitian ini, penyelesaian *imbalance* data akan dilakukan dengan memodifikasi dataset dengan cara menduplikasi data minor atau lebih dikenal dengan metode *oversampling.* Hasil akhirnya adalah mengetahui bagaimana pengaruh metode *oversampling* yang digunakan terhadap akurasi prediksi data *churn* dengan melakukan penghitungan akurasi model *churn prediction* yang dinyatakan dalam bentuk *lift curve* dan *gini coefficient* dan *top decile.*

1.1 Imbalance class

Imbalance class merupakan ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya merepresentasikan jumlah data yang sangat besar (*majority class)* sedangkan kelas yang lainya merepresentasikan jumlah data yang sangat kecil (*minority class*).

1.2 Churn prediction

Salah satu kasus nyata dari permasalahan *imbalance* class adalah kasus *churn* pada perusahaan telekomunikasi. Karakteristik dari data *churn* adalah tingkat *imbalance* yang besar, karena pelanggan yang mengalami *churn* jauh lebih sedikit dibandingkan pelanggan yang loyal. Ini mengakibatkan sulitnya membuat pemodelan terhadap data *churn* (Cardell ,2003).

Dalam hal ini, pelanggan yang *churn* dapat dibagi menjadi dua kelompok utama (Batista, 2004), yaitu:

1. *Voluntary churners* / sukarela
   Voluntary churners lebih sukar untuk ditentukan, sebab pada pelanggan jenis ini *churn* terjadi ketika seorang pelanggan membuat keputusan

secara sadar untuk mengakhiri layanan yang digunakan.

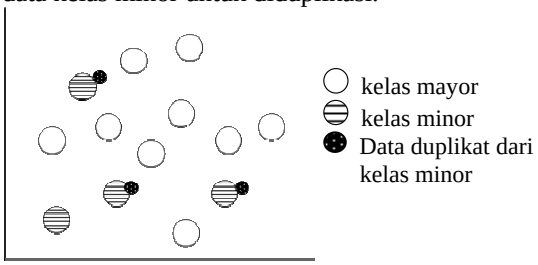
2. *Involuntary churners / tidak sukarela*
   Involuntary churners ini lebih mudah untuk diidentifikasi, seperti pelanggan yang menggunakan jasa ditarik/dicabut dengan sengaja oleh perusahaan tersebut dikarenakan adanya beberapa alasan.

2. METODE SAMPLING DAN OVERSAMPLING

Sampling merupakan bagian dari ilmu statistik yang memfokuskan penelitian terhadap pemilihan data yang dihasilkan dari satu kumpulan populasi data. Metode *sampling* atau yang lebih dikenal dengan *resample* adalah metode umum yang digunakan dalam menyelesaikan permasalahan *imbalance* data. Dengan adanya penerapan *sampling* pada data yang *imbalance*, tingkat *imbalance* semakin kecil dan klasifikasi dapat dilakukan dengan tepat (Laurikkala, 2001). Sedangkan metode *Oversampling* dilakukan dengan menyeimbangkan jumlah distribusi data dengan meningkatkan jumlah data kelas minor. Metode yang digunakan dalam penelitian ini adalah *random oversampling*, SMOTE, dan *borderline* SMOTE.

3. RANDOM OVERSAMPLING

Random oversampling bekerja dengan memilih data kelas minor untuk diduplikasi.



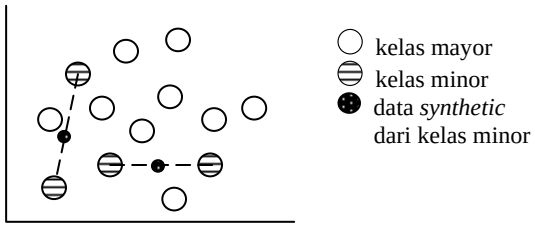
Gambar 1. Ilustrasi *Random Oversampling*

Hasil dari *random oversampling* tidak selalu meningkatkan prediksi kelas minor. Apabila data kelas minor diduplikasi dalam jumlah yang besar, maka akan sulit untuk mengidentifikasi data yang memiliki kemiripan karakteristik namun berada di kelas yang berbeda.

4. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) pertama kali diperkenalkan oleh Nithes V. Chawla (Chawla, 2002). Pendekatan ini bekerja dengan membuat "*synthetic*" data, yaitu data replikasi dari data minor,

Metode SMOTE bekerja dengan mencari *k nearest neighbors* (yaitu ketetanggaan data) untuk setiap data di kelas minor, setelah itu buat *synthetic data* sebanyak prosentase duplikasi yang diinginkan antara data minor dan *k nearest neighbors* yang dipilih secara random. Ilustrasi distribusi data setelah diterapkan metode SMOTE dapat dilihat pada Gambar 2.



Gambar 4. Ilustrasi SMOTE

Pada pembentukan data synthetic yang baru, ada 2 jenis perhitungan terhadap *nearest neighbor*, untuk data nominal dan data *continues*.

Untuk data *continues* :

- Dihitung perbedaan untuk setiap atribut antara *minority sample* (*k*) dengan salah satu dari *k nearest neighboornya* (*i*).
- Perbedaan ini dikalikan dengan nilai *random* antara 1 dan 0
- Hasilnya ditambahkan dengan nilai *minority sample*, inilah hasil pembuatan *feature vector* yang baru (*synthetic minority class* yang baru (k_1)).

Untuk data *nominal* :

- Diambil voting antara *minority sample* (E_1) dan *nearest neighbor*nya (E_2 dan E_3). Jika tidak ada *majority class*, maka pilihlah nilai atribut pada *minority sample* tersebut.
- Nilai tersebut ditandai menjadi *synthetic minority class* yang baru (E_{smote}).

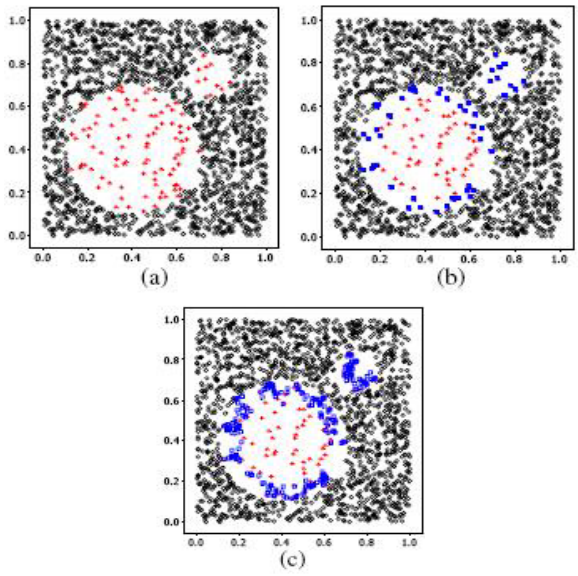
5. BORDERLINE-SMOTE

Borderline SMOTE merupakan metode yang menerapkan metode SMOTE pada data *borderline*, karena dianggap sering misklasifikasi. *Synthetic data* baru akan dibangkitkan sepanjang garis antara kelas minor dan tetangga terdekat yang dipilih sebelumnya (Han, Hui., 2005).

Langkah-langkah pada *borderline* SMOTE adalah :

- Tentukan *k nearest neighbor* untuk tiap data kelas minor
- Periksa apakah data kelas minor masuk himpunan *DANGER* atau tidak. Himpunan *DANGER* adalah himpunan yang berisi data kelas minor dengan mayoritas *nearest neighbor*-nya adalah data kelas mayor.
- Untuk tiap data di himpunan *DANGER*, lakukan proses SMOTE.

Borderline SMOTE ini diilustrasikan dalam gambar 3.



Gambar 3. Ilustrasi Borderline SMOTE

6. PARAMETER EVALUASI UTUK CHURN PREDICTION

6.1 Lift curve

Lift curve adalah alat ukur yang biasa di gunakan di dalam kasus *churn prediction* yang memetakan hasil prediksi dari model *classifier* ke dalam bentuk kurva. Untuk membuat *lift curve*, costumer diurutkan berdasarkan kemungkinan mengalami *churn* dari yang paling tinggi sampai yang paling rendah. Indikasi semakin bagusnya model prediksi adalah pada titik prosentase *customer* yang sama pada *lift curve*, prediksi tersebut mendapatkan prosentase *actual churner* yang lebih besar. Ilustrasi dari *lift curve* dapat dilihat pada gambar 4.

6.2 Top Decile Lift

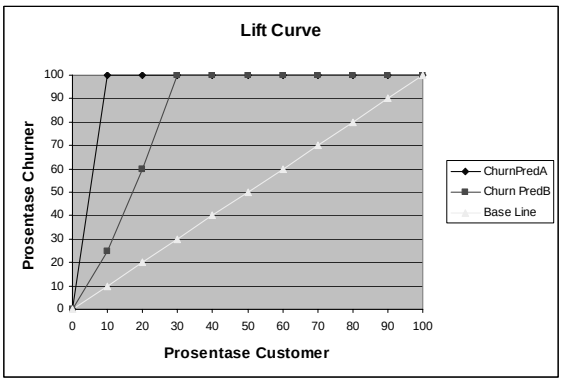
Top decile 10% merupakan akurasi yang lebih memfokuskan pada 10% *riskiest segment* yaitu fokus kepada sekumpulan *customer* sebanyak 10 % dari keseluruhan customer yang memiliki probabilitas *churn* yang paling tinggi. Sehingga dapat diketahui *customer* mana saja yang mempunyai kemungkinan untuk *churn* lebih besar dan suatu perusahaan dapat mengatur strategi untuk customer yang termasuk ke dalam kelompok *riskiest segment*, sehingga dapat dilakukan pencegahan prosentase *churner* yang lebih banyak lagi (Machado ,2007)

$$TopDecile = \frac{\hat{\pi}10\%}{\hat{\pi}} \qquad (1)$$

Keterangan

$\hat{\pi}10\%$: prosentase *churner* yang berada pada *riskiest segment*

$\hat{\pi}$: prosentase *churner* pada keseluruhan *customer*



Gambar 4. Lift Curve

6.3 Gini coefficient

Suatu pemodelan bisa saja hanya baik dalam memprediksi *riskiest segment* namun tidak bagus untuk customer dengan tingkat *churn* rendah (Lemmens, 2006). Untuk mengukur akurasi pada keseluruhan *customer*, maka dapat dilakukan perhitungan *gini coefficient* pada hasil prediksi. Dalam *gini coefficient*, tidak hanya segmen pelanggan tertinggi yang diperhitungkan, namun semua pelanggan yang telah diprediksi, baik *churn* ataupun *loyal*.

$$\text{Gini} = \left(\frac{2}{n}\right)\sum_{i=1}^{n}\left(v_i - \hat{v}_i\right) \qquad (2)$$

7. EVALUASI UNTUK *IMBALANCE CLASS*

Karena *churn* merupakan salah satu kasus *imbalance*, perlu dilakukan pengukuran akurasi *imbalance class*, yaitu penghitungan nilai *recall*, *precision*, dan *f-measure*.

Recall dihitung untuk mengevaluasi seberapa besar *coverage* suatu model dalam memprediksi suatu kelas tertentu. *Precision* dihitung untuk mengevaluasi seberapa baik ketepatan model dapat memprediksi suatu kelas. Dan untuk menetukan hasil prediksi yang paling baik, digunakan nilai *f-measure* yang merupakan kombinasi dari nilai *recall* dan *precision*.

$$\textit{Recall} \quad (r) : \frac{TP}{TP + FN} \qquad (3)$$

$$\textit{Precision} \quad (p) : \frac{TP}{TP + FP} \qquad (4)$$

$$\textit{F-measure} : \frac{2rp}{r + p} \qquad (5)$$

8. GAMBARAN UMUM SISTEM

Sistem yang akan dibangun adalah perangkat lunak yang mengimplementasikan metode *oversampling* dengan perhitungan khusus pada data yang *imbalance*. Proses klasifikasi akan dilakukan oleh *tools* yang telah banyak digunakan oleh perusahaan komunikasi untuk melakukan prediksi

churn, yaitu Clementine dan Weka. Dari hasil klasifikasi, perangkat lunak akan mengukur tingkat akurasi prediksi yang didapat setelah klasifikasi tersebut. Analisa yang akan dilakukan adalah dengan menganalisa pengaruh penerapan *sampling* sebelum klasifikasi, pada hasil prediksi yang dihasilkan oleh *classifier* Clementine 10.1.

9. DATA

Data yang digunakan dalam penelitian ini adalah data pelanggan dari salah satu perusahaan telekomunikasi di Indonesia. Dalam pengujian data perusahaan telekomunikasi yang memiliki jumlah record sebanyak 48384 data dengan 22 atribut dibagi menjadi data training dan data testing, masing-masing 75% dan 25% dari data asli. Tingkat imbalance pada data asli adalah 0,78%, perbandingan jumlah mayor dan minornya sebesar 48009 : 375. Data mayor direpresentasikan dengan nilai 'NO', sedangkan data minor direpresentasikan dengan nilai 'YES'.

10. ANALISIS DAN PENGUJIAN

10.1 SKENARIO PENGUJIAN SISTEM

Pada data akan dilakukan *oversampling*. Ketika dilakukan proses *oversampling*, metode-metode seperti SMOTE, Borderline SMOTE, terlebih dahulu ditentukan jumlah *nearest neighbor*-nya adalah lima. Pertimbangannya adalah nilai atribut pada data sintetis yang terbentuk dari lima *nearest neighbor*, tidak akan jauh berbeda dengan nilai atribut data minor acuannya. Jumlah *nearest neighbor* lima juga merupakan jumlah yang sering digunakan pada percobaan metode yang menerapkan SMOTE, seperti diterangkan pada referensi (Chawla,2002), (Machado, 2007), (Han, Hui., 2005). Sebagai perbandingan dalam melaksanakan pengujian, metode sampling yang digunakan akan meliputi *Random oversampling* pada WEKA, yaitu *resample*, *Random oversampling* pada SPSS Clementine, menggunakan *node balance*, SMOTE, Borderline SMOTE.

Klasifikasi memanfaatkan classifier yang ada pada SPSS Clementine 10.1 yaitu C5.0, dan hasil prediksi yang dihasilkan akan dihitung akurasinya.

Tabel 1. Skenario Distribusi Data

Over *sampling*	Jumlah Mayor	Jumlah Minor	% Imbalance
5	35985	1680	4,46%
27	35985	9072	20,13%
50	35985	16800	31,83%
75	35985	25200	41,19%
100	35985	33600	48,29%

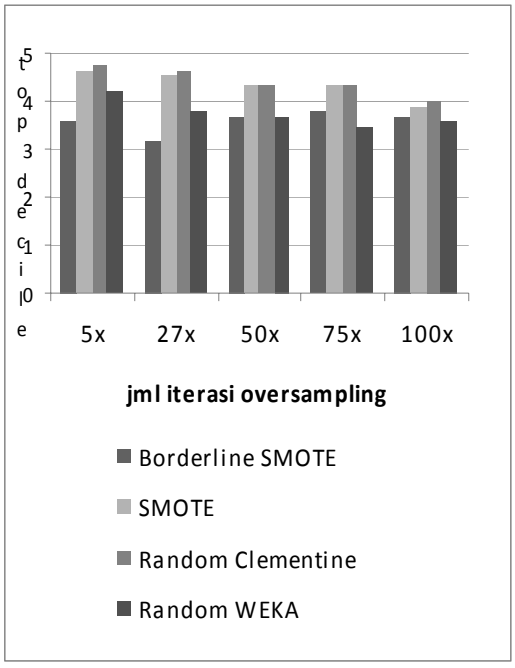
Proses pengujian dengan parameter-parameter yang ditentukan di tabel 1 diatas akan menghasilkan suatu analisa tentang pengaruh perubahan parameter tersebut terhadap hasil akhir akurasi prediksi.

10.2 PENGARUH TINGKAT *IMBALANCE* DAN METODE SAMPLING TERHADAP TOP DECILE 10%

Akurasi difokuskan pada 10% *riskiest segment*. Pertimbangan dalam memilih nilai 10% adalah karena kelompok yang meliputi 10% *customer* dengan tingkat resiko tertinggi merupakan segmentasi ideal bagi perusahaan dalam menerapkan strategi marketing untuk mencegah terjadinya *churn* (Lemmens, 2006).

Perhitungan pada *top decile* sebanding dengan nilai *lift curve* di atas. *Lift curve* menggambarkan tahap-tahap mencapai titik customer 10%, sedangkan *top decile* hanya melihat hasil akhir di titik tersebut. Di bagian ini akan dianalisa pengaruh tingkat imbalance terhadap nilai akurasi *top decile*. Hasil Pengujian ditunjukkan pada gambar 5.

Random Clementine mendapatkan nilai *top decile* terbaik di hampir seluruh pengujian, sementara SMOTE di urutan kedua. Di sini dapat dilihat bahwa, semakin banyak duplikasi data minor dilakukan, maka nilai *top decile* akan semakin kecil, ini dikarenakan semakin banyak duplikasi dilakukan, maka rule klasifikasi juga bertambah, sehingga pada saat pengujian, akan menimbulkan semakin banyak nilai *confidence*, sedikit saja terjadi salah klasifikasi akan menyebabkan keterurutan data menjadi tidak tepat. Sebaliknya, jika nilai *confidence* prediksi sedikit, tidak menjadi suatu masalah yang berarti, karena data yang probailitas *churn*-nya tinggi masih mempunyai peluang besar untuk berada pada *riskiest segment*.

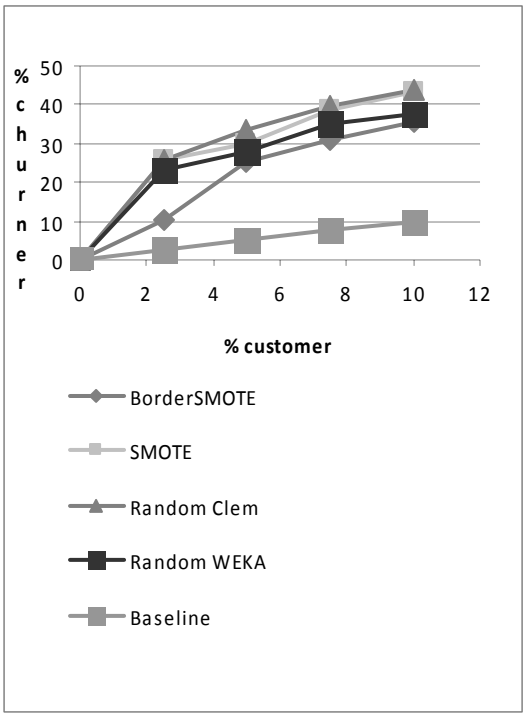


Gambar 5. Hasil Pengujian *Top Decile*

10.3 PENGARUH METODE SAMPLING TERHADAP *LIFT CURVE*

Untuk pengujian terhadap data, akan digambarkan *lift curve* dengan memperhatikan *riskiest segment* sebesar 10% dari keseluruhan *customer*. Akurasi yang ditampilkan dalam bentuk kurva, dapat memudahkan untuk melihat lebih jelas, metode sampling mana yang memiliki tingkat prediksi yang lebih tinggi untuk customer 10%. Pada pengujian ini, digunakan jumlah iterasi 5 kali. Hasil pengujian ditunjukkan pada gambar 6.

Lift curve diatas memperlihatkan bahwa metode *Random Oversampling* yang diterapkan Clementine menempati peringkat tertinggi untuk prediksi churn. Untuk 10% customer teratas, *random oversampling* dapat menebak 44% *actual churner*. Metode sampling di peringkat ke-dua adalah metode SMOTE dengan pencapaian *actual churner* yang hampir mendekati perolehan *random oversampling* Clementine. Hasil ini hampir sama denga pengujian *top decile*.

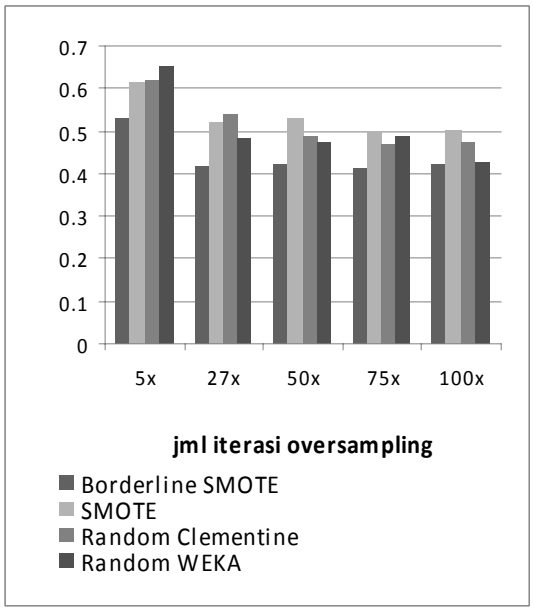


Gambar 6. Hasil Pengujian *Lift Curve*

10.4 PENGARUH METODE SAMPLING TERHADAP PERHITUNGAN *GINI COEFFICIENT*

Seperti yang disebutkan sebelumnya, *gini coefficient* mengukur tingkat akurasi untuk seluruh *customer*. Jika suatu prediksi memiliki nilai *top decile* yang kecil, tidak menutup kemungkinan memiliki nilai *gini coefficient* yang besar. Hal ini disebabkan, *top decile* hanya fokus pada *n*% customer tertinggi, (100-*n*)% lainnya tidak diikutsertakan dalam perhitungan, sedangkan gini coefficient memperhatikan 100% customer yang telah diprediksi. Hasil pengujian ditunjukkan pada gambar 7.

Pada setiap iterasi, tampak SMOTE paling baik. Selain itu, dapat dilihat itearsi paling ail adalah 5 kali. Ini menunjukkan bahwa kita harus berhati-hati dalam melakukan duplikasi data minor. Duplikasi data minor yang berlebihan juga dapat mengakibatkan terjadinya *overfitting*, sehingga hasil dari prediksi juga semakin buruk.



Gambar 7. Hasil Pengujian Gini coefficient

10.5 PENGARUH METODE SAMPLING TERHADAP F-MEASURE

Data *churn* merupakan bagian dari kasus *imbalance*, sehingga perlu dihitung pula akurasinya terhadap *f-measure* yang merupakan kombinasi dari nilai *recall* dan *precision* sebagai evaluasi umum untuk data *imbalance*. Hasil perhitungan terhadap *f-measure* akan ditampilkan pada tabel 2.

Tabel 2. Hasil Pengujian Nilai Akurasi Matrik Evaluasi

METODE	RECALL	PRECISION	F-MEASURE
OVERSAMPLING			
Borderline SMOTE	0.0989	0.0519	0.0555
SMOTE	**0.2273**	**0.0974**	**0.1245**
Random Clementine	0.2252	0.0928	0.1236
Random WEKA	0.2084	0.0957	0.1206

Dari pengujian ini dapat dilihat bahwa SMOTE menghasilkan nilai *f-measure* yang paling baik. Sedangkan *Borderline* SMOTE malah menempati posisi terburuk. *Borderline* SMOTE mempunyai kelemahan, jika data relatif menyebar, maka duplikasi pada borderline justru akan rawan terjadi *overfitting*, karena disekitar *borderline* tentu banyak data kelas mayor.

11. KESIMPULAN

Untuk parameter evaluasi *churn prediction*, seperti *top decile*, *lift curve*, *gini coe*fficient, Random Clementine masih menempati posisi paling baik dibanding yang lain, namun bisa dikatakan hampir berimbang dengan metode SMOTE, karena selisih hasil yang diperoleh tidak terlalu signifikan.

Metode *oversampling* yang menghasilkan data *synthetic* belum sesuai diterapkan pada data *churn*., karena cenderung masih menghasilkan nilai *top decile* yang kecil.

Suatu metode *sampling* yang baik digunakan pada sisi kasus *imbalance*, belum tentu baik jika dilihat dari sisi kasus *churn*, begitu pula sebaliknya.

Performansi model dipengaruhi oleh metode *sampling* dalam menambahkan *knowledge* data *training* yang sesuai terhadap karakteristik data *testing*.

PUSTAKA

Batista, Gustavo E.A.P.A., Prati, Ronaldo C., and Maria Carolina., (2004), *"A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data"*. SIGKDD Explorations 6(1): 20-29

Cardell, Scott., Golovnya, Mikhail., Steinberg, Dan., (2003)., *Churn Modeling for Mobile Telecommunications*. Salford Systems. California.

Chawla, Bowyer, Hall, and Kegelmeyer. (2002) *"SMOTE : Synthetic Minority Oversampling Technique"*. Journal of Artificial Intelligence Research 16. Page 321-357.

Han, Hui., Wang, Wen-Yuan., Mao, Bing-Huan., (2005), *"Borderline-SMOTE A New Over-Sampling Method in Imbalanced Data Sets Learning"*. Beijing. China

Lemmens, Aurelie., Croux, Christophe., (2006).,*"Bagging and Boosting Classification Trees"*. Journal of Marketing Research, 43(2) 276-286.

Laurikkala, Jorma. (2001)*"Improving Identification of Difficult Small Classes by Balancing Class Distribution"*. University of Tampere. Finland..

Machado, Emerson Lopes., Ladeira, Marcelo., (2007) *"Dealing With Rare Cases and Avoiding Overfitting : Combining Cluster Based Oversampling and SMOTE"*. Department of Computer Science. Brazil.