Seminar Nasional Informatika 2010 (semnasIF 2010) *ISSN: 1979-2328*
UPN "Veteran" Yogyakarta, 22 Mei 2010

ANALISIS PENGARUH METODE COMBINE SAMPLING DALAM CHURN PREDICTION UNTUK PERUSAHAAN TELEKOMUNIKASI

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

Abstrak
Churn prediction pada pelanggan telekomunikasi merupakan upaya memprediksi/mengklasifikasi pelanggan jasa telekomunikasi yang berhenti atau berpindah berlangganan dari suatu operator ke operator yang lain. Namun dataset pada kasus churn ini biasanya memiliki kelas yang imbalance dimana jumlah instance suatu kelas (kelas active atau tidak churn atau mayor atau negatif) jauh lebih besar dari jumlah kelas yang lain (kelas churn atau minor atau positif). Akibatnya, kebanyakan classifier cenderung memprediksi kelas mayor dan mengabaikan kelas minor sehingga akurasi kelas minor sangat kecil. Salah satu pendekatan yang dilakukan untuk menangani permasalahan ini adalah dengan memodifikasi distribusi instances dari dataset yang digunakan atau yang lebih dikenal dengan pendekatan sampling-based. Teknik resampling ini meliputi over-sampling, under-sampling, dan combine-sampling. Analisis yang dilakukan pada penelitian ini adalah mengetahui bagaimana pengaruh metode combine sampling 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 combine sampling belum sesuai diterapkan pada data churn, karena cenderung masih menghasilkan nilai top decile yang kecil. Tetapi secara umum metode combine sampling ini mampu meningkatkan akurasi untuk memprediksi data minor. Dengan penerapan metode combine sampling, data churn yang memiliki tingkat imbalance yang besar dapat diklasifikasi tanpa mengorbankan data minor yang menjadi fokus penelitian. Metode combine sampling yang digunakan juga memiliki hasil evaluasi yang berbeda terhadap dataset sebagai data churn dan sebagai data imbalance.

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

1. PENDAHULUAN

Industri penyedia jasa telekomunikasi merupakan industri yang terus berkembang dan selalu dibutuhkan masyarakat. Dengan semakin banyaknya jumlah perusahaan telekomunikasi baik penyedia layanan GSM (*Global System Mobile*) maupun CDMA (*Code Division Multiple Access*), masing-masing akan saling menerapkan strategi untuk memperebutkan perhatian pelanggan. *Churn* lahir dari fenomena di atas. C*hurn* adalah pemutusan jasa suatu perusahaan oleh pelanggan karena pelanggan tersebut lebih memilih menggunakan layanan jasa perusahaan kompetitor. *Churn* harus diwaspadai oleh perusahaan karena dengan bertambahnya jumlah *churn* akan semakin mengakibatkan penurunan *revenue*.

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 dan mengurangi data mayor. Hasil akhirnya adalah mengetahui bagaimana pengaruh metode *combine sampling* 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*.

2. TINJAUAN PUSTAKA

2.1. 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*).

2.2. 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* [2].

Dalam hal ini, pelanggan yang *churn* dapat dibagi menjadi dua kelompok utama [1], 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.

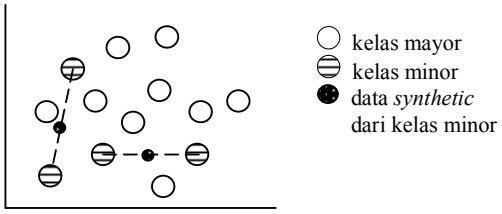
2.3. Metode Combine Sampling

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 [6]. Sedangkan metode *Combine sampling* dilakukan dengan menyeimbangkan jumlah distribusi data dengan meningkatkan jumlah data kelas minor (oversampling) dan mengurangi data mayor (undersampling). Metode yang digunakan dalam penelitian ini adalah SMOTE + Tomek Link. Dalam pengujian, metode Random Combine Sampling pada Clementine dan WEKA sebagai pembanding.

2.3. Smote

Synthetic Minority Oversampling Technique (SMOTE) pertama kali diperkenalkan oleh Nithes V. Chawla [3]. 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 1.



Gambar 1. Ilustrasi SMOTE

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

Untuk data *continues* :

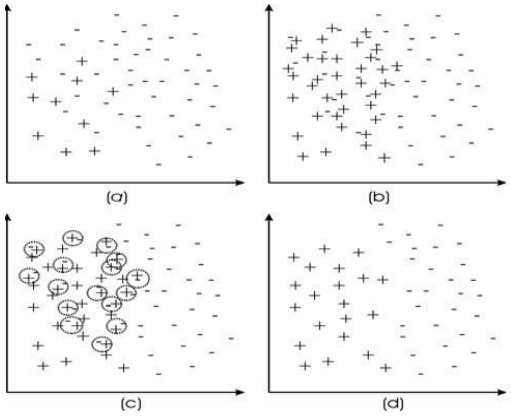
- Dihitung perbedaan untuk setiap atribut antara *minority sample* (*k*) dengan salah satu dari *k nearest neighboornya* (*i*).
- Perbedaan ini dikalikan dengan nilai *random* antara 1 dan 0
- Hasilnya ditambahkan dengan nilai *minority sample*, inilah hasil pembuatan *feature vector* yang baru (*synthetic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

2.4. Smote + Tomek Link

Metode ini merupakan metode kombinasi antara SMOTE dan Tomek Link sebagai metode pembersihan data. Cara kerja Tomek Link adalah dengan menghapus data minor ataupun mayor yang memiliki kesamaan karakteristik. Untuk setiap data, jika satu tetangga yang paling dekat memiliki kelas label yang berbeda dengan data tersebut maka kedua data akan dihapus karena dianggap sebagai *noise* atau *misclassify*.



Gambar 2 : Ilustrasi Smote+Tomek Link

Ilustrasi langkah-langkahnya dapat dilihat pada Gambar 2. Data asli pada Gambar2(a) akan di-*oversampling* dengan metode SMOTE sehingga menghasilkan data dengan karakteristik seperti Gambar2(b). Kemudian di Gambar 2(c) memperlihatkan metode Tomek Link bekerja dengan pengecekan setiap tetangga terdekat untuk tiap data. Apabila ditemukan tetangga yang memiliki kelas label berbeda, maka kedua data itu akan dihapus dari data training sampai menghasilkan data training yang bersih dari *noise* seperti pada Gambar 2(d).

2.5. Parameter Evaluasi Utuk Churn Prediction

2.5.1. Lift curve

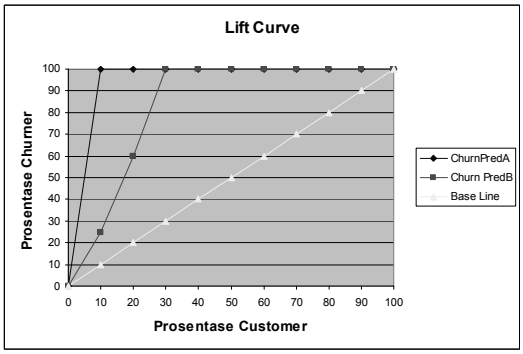
Lift curve adalah alat ukur yang biasa di gunakan di dalam kasus *churn prediction* yang memetakan hasil prediksi dari model *classifier* ke dalam bentuk kurva. Untuk membuat *lift curve*, costumer diurutkan berdasarkan kemungkinan mengalami *churn* dari yang paling tinggi sampai yang paling rendah. 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 3.

2.5.2. 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 [7].

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

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



Gambar 3. *Lift Curve*

2.5.3. 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*.

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

2.6. 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*.

$$Precision = \frac{categories\ found\ and\ correct}{total\ categories\ found} \qquad (3)$$

$$Recall = \frac{categories\ found\ and\ correct}{total\ categories\ correct} \qquad (4)$$

$$F\ measure = \frac{2(precision \times recall)}{(precision + recall)} \qquad (5)$$

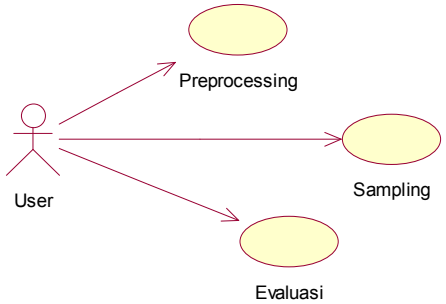
3. GAMBARAN UMUM SISTEM

Sistem yang akan dibangun adalah perangkat lunak yang mengimplementasikan metode *combine sampling* 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.

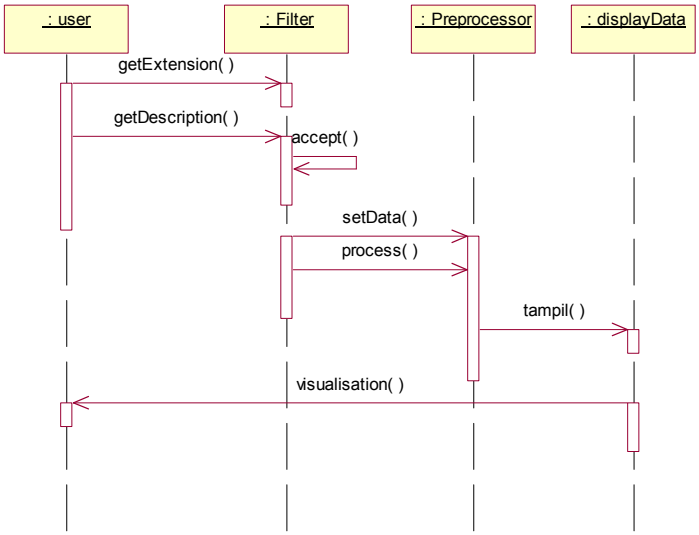
3.1. Perancangan Sistem

3.1.1 Use Case Diagram

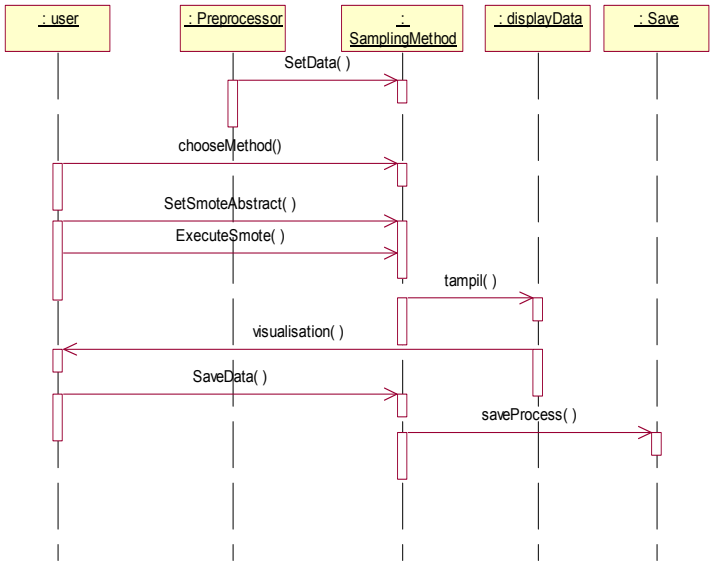
Use case dapat menggambarkan interaksi antara pengguna sistem (*user*) dengan sistem itu sendiri. *Use case* hanya menggambarkan apa yang dilihat *user* terhadap keadaan lingkungan sistem dan bukan menggambarkan bagaimana fungsi yang ada dalam sistem. *Use case* untuk sistem *churn prediction* adalah sebagai berikut :



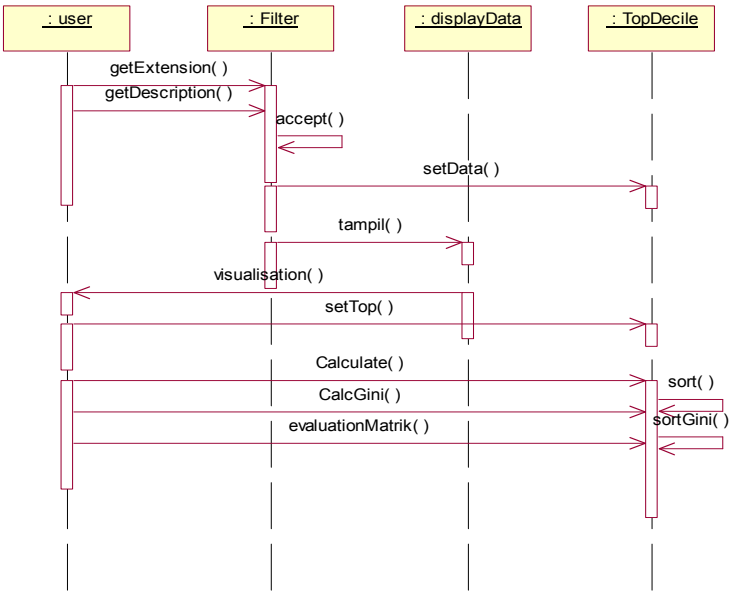
Gambar 4. Use Case Diagram



Gambar 5. Sequence untuk Use Case Preprocessing



Gambar 6. Sequence untuk Use Case *Sampling* Process



Gambar 7. Sequence untuk Use Case Evaluasi

3.2. 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'.

4. ANALISIS DAN PENGUJIAN

4.1. Skenario Pengujian Sistem

Sebelumnya dataset akan disampling menggunakan metode sampling yang telah ditentukan. Ketika dilakukan proses SMOTE+Tomek Link terlebih dahulu diatur jumlah *nearest neighbor*-nya adalah 5. Pertimbangannya adalah nilai atribut pada data *synthetic* yang terbentuk dari 5 *nearest neighbor*, tidak akan jauh berbeda dengan nilai atribut data minor acuannya. Jumlah *nearest neighbor* 5 juga merupakan jumlah yang sering digunakan pada percobaan metode yang menerapkan SMOTE, seperti diterangkan pada referensi [3], [4], [7]. Sebagai metode pembanding adalah Random combine sampling pada Clementine dan WEKA.

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	35744	1439	3,87%
27	35665	7520	17,41%
50	35646	13941	28,11%
75	35640	20935	37,01%
100	35640	27935	43,94%

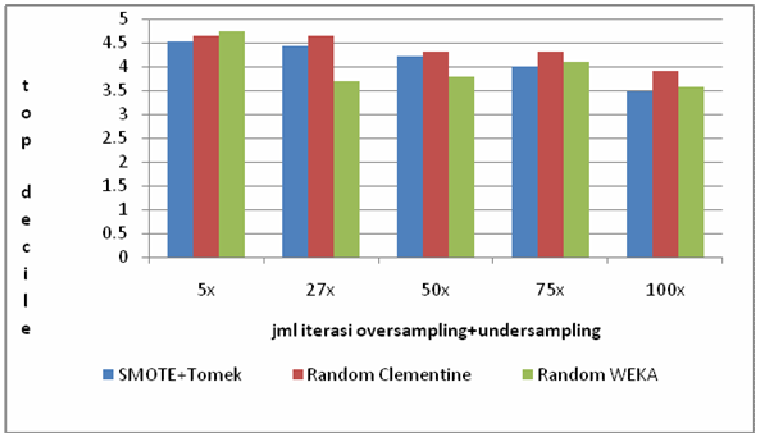
Proses pengujian dengan parameter-parameter yang ditentukan di tabel 1 diatas akan menghasilkan suatu analisa tentang pengaruh perubahan parameter tersebut terhadap hasil akhir akurasi prediksi.

4.2. Pengaruh Tingkat *Imbalance* Dan Metode Combined 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* [5].

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.

Dari hasil pengujian didapatkan, Random Clementine mendapat nilai terbaik di seluruh pengujian. Dari Gambar 4, dapat dilihat bahwa pada beberapa kali pengujian dengan prosentase *oversampling/undersampling* yang semakin besar, nilai *top decile* justru semakin kecil. Ini menunjukkan bahwa semakin banyak *oversampling /undersampling* yang dilakukan untuk memperkecil tingkat *imbalance* data *churn*, justru memperkecil nilai *top decile*.



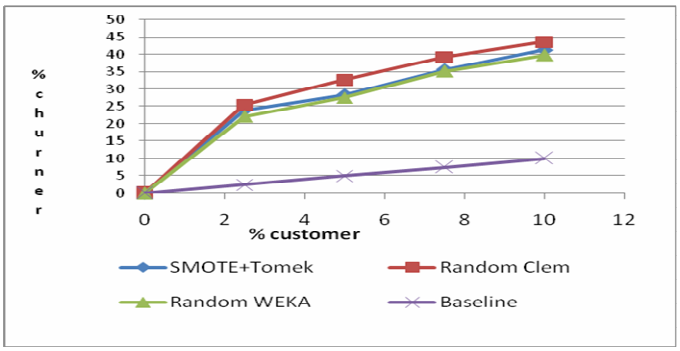
Gambar 8. Hasil Pengujian *Top Decile*

4.3. 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 5.

Random Clementine menempati peringkat tertinggi untuk kategori *lift curve* dengan mendapatkan 43,58% aktual *churn*er secara tepat. nilai terbaik ke-dua adalah SMOTE+Tomek Link. Di sini, semua metode menghasilkan nilai yang hampir sama. Secara terurut dari prosentase aktual *churn*er yang didapat pada metode SMOTE+Tomek Link, Random Clementine, dan Random WEKA di data turnament adalah sebesar 41,264, 43,582, dan 39,788.

Metode *random* pada Clementine berhasil menangkap aktual *churner* lebih banyak pada 10% *riskiest segment*, sedangkan metode SMOTE+Tomek Link yang diharapkan lebih baik daripada metode *random* hanya mampu berada di peringkat ke-dua. Analisis yang telah dilakukan terhadap hasil prediksi data untuk tiap metode menghasilkan suatu hipotesa baru bahwa ketika metode *sampling* membentuk data sintetik pada data training, maka rule klasifikasi juga akan semakin bertambah. Hal ini menyebabkan pada saat pengujian, akan menimbulkan semakin banyak nilai *confidence* saat menentukan label kelas data.

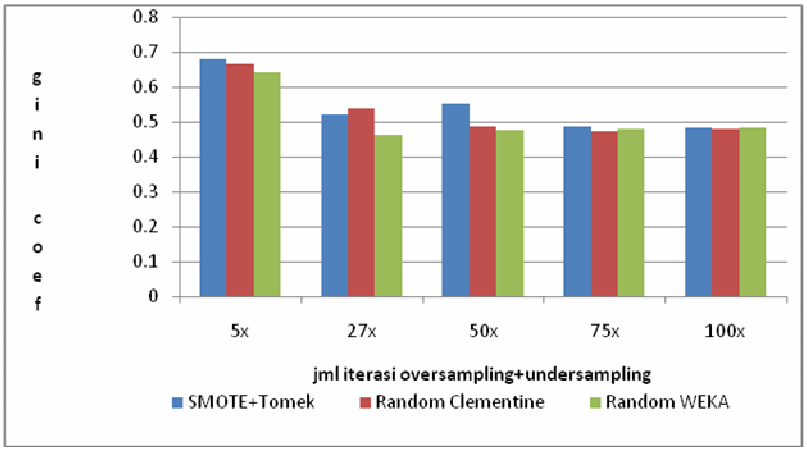


Gambar 9. Hasil Pengujian *Lift Curve*

4.4. 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 6.

Dari hasil pengujian, SMOTE+Tomek Link di peringkat teratas karena mampu menghasilkan nilai *gini coefficient* yang paling tinggi di empat dari lima kali pengujian. Gambar 7 menunjukkan, bahwa kita harus berhati-hati dalam melakukan duplikasi data minor/pengurangan data mayor. Duplikasi data minor/pengurangan data mayor yang berlebihan juga dapat mengakibatkan terjadinya *overfitting*, sehingga hasil dari prediksi juga dapat semakin buruk.



Gambar 10. Hasil Pengujian Gini coefficient

4.5. 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
COMBINE SAMPLING			
SMOTE+Tomek Link	0.0362	**0.0254**	0.0262
Random Clementine	**0.0414**	0.0247	**0.0309**
Random WEKA	0.0365	0.0244	0.0259

Sedangkan untuk kategori kombinasi, random pada Clementine dapat menangkap aktual churner lebih banyak dibandingkan dengan SMOTE+Tomek Link.

Dari pengujian ini dapat dilihat bahwa Random Clementine menghasilkan nilai *f-measure* yang paling baik. SMOTE+Tomek Link menempati posisi kedua. Tetapi nilai recall, precision dan f-mesure tidak berbeda banyak. Jika data relatif menyebar, duplikasi/pengurangan pada metode SMOTE+Tomek Link, akan rawan terjadi *overfitting*.

5. KESIMPULAN

Untuk parameter evaluasi *churn prediction*, seperti *top decile* dan *lift curve*, Random Clementine masih menempati posisi paling baik dibanding yang lain, namun untuk *gini coe*fficient, SMOTE + Tomek Link lebih unggul daripada metode yang lain.

Untuk parameter evaluasi *imbalance class*, seperti *precision, recall*, dan *F-measure*, Random Clementine masih menempati posisi paling baik dibanding yang lain, namun nilai yang dihasilkan tidak terlalu berbeda banyak dengan SMOTE + Tomek Link.

Metode *combine sampling* 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.

6. PUSTAKA

[1] 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

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

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

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

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

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

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