INTISARI

ANALISIS PERBANDINGAN *GRANULAR SUPPORT VECTOR MACHINE REPETITIVE UNDERSAMPLING* PADA DATA TIDAK SEIMBANG

(Studi kasus: Prediksi *Lapse* pada Produk Asuransi Kesehatan)

Oleh

Priestha Zara Yenni Wijaya

18/427710/PA/18670

Bangkrutnya asuransi AIG tahun 2008 menunjukkan pentingnya risiko likuiditas bagi perusahaan asuransi, serta risiko *lapse* dan *underwriting* adalah risiko yang berhubungan erat dengan risiko likuiditas, risiko *lapse* sendiri berkontribusi kurang lebih 50% pada kalkulasi risiko *underwriting*. Sehingga penting bagi perusahaan asuransi untuk memprediksi polis yang akan *lapse* pada periode mendatang untuk mempersiapkan strategi pencegahan. Melalukan prediksi, perlu metode klasifikasi yang mampu menangani masalah klasifikasi pada data sangat tidak seimbang, karena perusahaan asuransi yang kondisi keuangannya sehat memiliki data polis *lapse* jauh lebih sedikit dari yang tidak *lapse*. Metode klasifikasi yang digunakan adalah *Support Vector Machine* (SVM), *SVM Random Undersampling* (SVM-RANDU), dan *Granular SVM Repetitive Undersampling* (GSVM-RU). Ketika dibandingkan performa klasifikasinya pada *training* dan *testing dataset*, didapatkan SVM-RANDU sebagai model dengan performa terbaik. Hal ini dapat dilihat pada performa SVM-RANDU pada *training dataset*, yaitu G-mean dan ROC AUC 75.99%, Spesifisitas 76%. Sedangkan, GSVM-RU memiliki G-mean 75.15%, ROC AUC 75.09%, Spesifisitas 72%, dan SVM memiliki G-mean 52.78%, ROC AUC 32.95%, Spesifisitas 12%. Meskipun GSVM-RU memiliki performa yang sedikit lebih buruk, tetap perlu dipertimbangkan sebagai model klasifikasi yang baik karena mengurangi efek kehilangan informasi pada proses *undersampling* dan mendorong kemungkinan menggunakan SVM untuk *data cleaning*.

Kata kunci: data sangat tidak seimbang, GSVM-RU, kehilangan informasi, *data cleaning*

ABSTRACT

COMPARATIVE ANALYSIS OF GRANULAR SUPPORT VECTOR MACHINE REPETITIVE UNDERSAMPLING ON IMBALANCED DATASET

(Case study: Lapse Prediction of Health Insurance Product)

By

Priestha Zara Yenni Wijaya

18/427710/PA/18670

The 2008 AIG insurance crisis shows how substantial liquidity risk is for insurers, also lapse and underwriting risk are risks that are closely related to liquidity risk, where lapse risk contributes approximately 50% to the calculation of underwriting risk. So, insurers need to predict which policies will lapse in the coming period to prepare retention strategies. Prediction needs a classification method that can solve the classification problem in the data is very unbalanced because insurance companies whose financial conditions are healthy have much fewer policies that lapse than the one that doesn't lapse. The classification methods used are Support Vector Machine (SVM), SVM Random Undersampling (SVM-RANDU), and Granular Support Vector Machine Repetitive Undersampling (GSVM-RU). When compared to its classification performance in training and testing datasets, SVM-RANDU was obtained as the best-performing model. This can be seen in the performance of SVM-RANDU in the training dataset, namely G-mean and ROC AUC 75. 99% and Specificity 76%. Meanwhile, the performance of GSVM-RU is G-mean 75. 15%, ROC AUC 75. 09%, specificity 72%. For SVM performance without modification are G-mean 52.78%, ROC AUC 32.95%, and Specificity 12%. Although GSVM-RU performs slightly poorer, it can still be considered a classification model because it reduces the effect of information loss on the undersampling process and encourages the possibility of using SVM for data cleaning.

Keywords: highly imbalanced dataset, GSVM-RU, information loss, data cleaning