Abstract

Churn prediction is one of the data mining application to predict the churn customer. Prediction churn is a problem of imbalance class and churn is a minor class that resulted in loss when it is not detected. General data mining classification methods does not take loss of benefit due to an error classified a class to another class, this is less appropriate to the problem of churn which is considering benefit as the primary factor in the handling of churn customers.

At this final project, two cost-sensitive learning algorithm, *Costing* and *CSRoulette* are investigated beside two modifications of cost-sensitive learning algorithm based on the sampling, *UnderCosting* and *UnderCSRoulette*. Model evaluation is done through churn prediction model accuracy which are expressed in total benefit, lift top decile, top decile benefit, lift curve, and gini coefficient. Non-cost-sensitive learning algorithm which is used as a benchmark is *Boosting* with training data which is preprocessed with balancing technique.

Results obtained from the research show that the cost-sensitive learning based sampling method is not always better than non-cost-sensitive learning algorithm in all evaluation parameters. Algorithm based on undersampling: *UnderCosting*, *UnderCSRoulette*, and *Boosting-UnderSampling* are resulted in good performance when cost is low and in high cost resulted in poor performance because low of precision.

Keywords: cost-sensitive learning, sampling, boosting, benefit, cost.