# CHAPTER 1

# INTRODUCTION

This chapter discusses the rationale in Section 1.1 that explain the background of this study and related problem situation. Theories and concept used to conceptualize this study are discuses in Section 1.2, while Section 1.3 discusses the variable related to the problem and their relationship to the paradigm of this study. The intended problem within this study is explained in Section 1.4. Section 1.5 discusses the proposed approach to solving the intended problem. Besides, this study describes some assumption in Section 1.6, while Section 1.7 describes the scope of works and delimitation. Finally, the contribution of this study is described in Section 1.8.

## 1.1 Rationale

In data mining, especially in classification tasks, data imbalance occurs when the number of patterns of a class is far greater than other classes [2]. Most classification algorithms are trained with the assumption that the class ratio is almost the same. However, in the real classification task, this assumption is often violated. An example is churn prediction. Churn Prediction is a condition where a customer moves from one service to another service or decides not to continue subscribing to a company and what is meant by churn prediction is a prediction to detect whether a customer will churn or not in the future, so the company can take quick steps to prevent this from happening [3]. In churn prediction tasks, the number of customers who churn is much smaller than loyal customers, resulting in an imbalance of data.

Customer churn has become a significant problem and also a challanges for Telecommunication company. It is necessary to evaluate whether the big problems of churn customer and the company's managements will make appropriate strategies to minimize the churn and retaining the customer.

In order to survive in a competitive marketplace, telecommunication companies are turning to data mining technique for churn analysis, with this approach, a company will understand customer behaviour from its own data so the right CRM (Customer Relationship Management) strategies will be implemented as well in order to save its revenue.

The problem that will occur in unbalanced data is that the classifier will be biased towards the negative class and the positive class will be considered noise, so the resulting model performance is very biased towards problems that are more interested in data minority such as churn prediction. In general, in handling data imbalance the most common approach is sampling. Sampling operates at the data level and is widely used to make balanced distribution between classes. The most frequently used method in sampling is the

Oversampling and Undersampling method. The oversampling method aims to increase the number of positive classes to balance the ratio between the two classes. The easiest way to use the oversampling method is to copy positive samples directly. However, this method has shortcomings in time complexity because the amount of data becomes far more due to positive data being copied or overfitting [4]. Therefore, researchers made more advanced algorithms using oversampling bases such as SMOTE [5], ADASYN [6], and Borderline-SMOTE [7]. However, this algorithm still requires a long time in the training process. Almost the same as oversampling, Random Undersampling (RUS), removes random samples from negative patterns to balance the number of samples between classes. However, when the number of positive classes is very low, deleted samples may not represent all negative classes and thus some important information (important samples) will be lost. Usually, the undersampling technique shows better performance than the oversampling technique [8].

## 1.2 Theoretical Framework

This research aim is to handle the imbalanced data in churn prediction. Input of this system is user data from telecommunications companies in Indonesia from a specific product in the telecommunication industry in Indonesia with unbalanced data characteristics. Unbalanced data or Imbalanced data-set problem occurs when one class, usually the one that refers to the concept of interest (positive or minority class), is underrepresented in the data-set and the number of negative (majority) instances outnumbers the amount of positive class instances [3]. The output of this system is a prediction model that can predict whether a customer will churn or not.

The dataset used as input data divided into training data and test data. Training data sets and Testing Dataset contains a proportion of churners that is representative of the actual population to approximate the predictive performance in a real-life situation [8]. This study constructs all variables in the same way for each dataset.

In this study, the authors propose an ensemble learning that combines more effective sampling, bagging and boosting approaches. Ensemble of Undersampling (EUS) is used to change data distribution to handle data imbalances. In making a new subset of data using EUS, data mining techniques are used, it is clustering to select negative samples that will be down sampling in order to handle loss of information from important samples which is the Undersampling issue. For each sample classified in EUS, the authors use Real Adaboost [9]. The output of this classification model is based on a weighted voting system based on the error level of each classifier. To test the proposed method, author apply it to the telecommunications industry, PT Telkom Indonesia.

## 1.3 Conceptual Framework/Paradigm

Identify and discuss the variables related to the problem, and present a schematic diagram of the paradigm of the research and discuss the relationship of the elements/variables therein.

Proposed dataset annotated with two class. The class are Churn class and Not Churn Class. Prediction is done by analyzing a customer's data. This data have time-series data characteristis. There are two types of churn for broadband internet customer in PT. Telkom Indonesia,namely CT0 (Change Tariff 0) and APS (Atas Permintaan Sendiri). CT0 is customers who have had their service withdrawn by the company because they did not pay the bill for two consecutive months. APS is also called voluntary churn, which is more difficult to determine,because this occurs when a customer makes a conscious decision to terminate his/her service with the provider [10]. This research uses APS as the customer churn data sets because the number of APS churn is smaller than CT0, and it is more difficult to predict.

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Basic concept of the proposed method is handling unbalanced data and measuring the performance generated based on the results of the classification process in data testing.

## 1.4 Statement of the Problem

Main first, customerS is one of the main sources of corporate earnings that resulted they becoming an important asset to the company. Customers can lead to decrease in revenue of the company.

Secondly, customer churn dataset has an imbalanced data characteristic, where data has one of the sequences with more samples than the other classes. Number of negative instances is higher than the positive instances which reflect imbalanced class distribution.

Third, all basic classifiers assuming that dataset has an equal samples. Classifiers will make biased decision if dataset has an imbalanced data. Since the classification assumes that the data is drawn from the same data distribution, presenting imbalance data to the classifier will produce undesirable results.

Fourth, EUS is used to change data distribution to handle data imbalances. In making a new subset of data using EUS, data mining techniques are used, namely clustering to select negative samples that will be down sampling in order to handle the issue of losing information from important samples (undersampling issues). Real Adaboost is used on every sample classified in EUS.

Based on the some of point problems discussed above, the main issue of this study is the process to solve the imbalanced data on predicting customer churn with three main

approaches, Undersampling with base of clustering, bagging with enhancing Undersampling technique with create of bag, and boosting with Real Adaboost

## 1.5 Objective

Based on the problem statement and rationale above, the objectives of this study are:

1. Create customer churn prediction as Churn Prediction Model on Customer churn data in PT Telkom Indonesia Tbk.
2. To identify the important attributes in churn prediction.
3. Obtain the performances of Modified Ensemble Undersampling-Boost.

## 1.6 Hypotheses

Previous research still presented handling imbalance data in Boosting base approach and RUSBoost. RUSBoost has little effect on the classifier produced [10]. The result from previous research still needs improvement [4][10]. Undersampling makes process is likely to cause the problem of data underrepresentation [4]. A ensemble approach EUS-Boost and undersampling strategy based on clustering (Modified EUS-Boost) to improve more performance on the classification results and time-consumption.

## 1.7 Assumption

The global problem in churn prediction includes the variation of the dataset, the churn prediction accuracy, the main factors causing churn, and the relation with marketing management to determine appropriate strategies to deal with the problem of churn. Continuous researches are needed to address this problem. This study focused on the problem of churn prediction accuracy and using the following assumptions:

1. This research was conducted based on data churn in a specific product of a telecommunication company in Indonesia.
2. This research discusses issues overcome imbalanced data on the prediction of churn in order to produce good performance.
3. This research does not discuss more about the factors which most influence on the churn.

## 1.8 Scope and Delimitation

There are many methods and techniques which can be used to improve the performance of the churn prediction. In order to get more focused analysis on churnprediction performance resulted by the selected method, this research used scopes and delimitations as follows:

1. This study attempts to compare the performance of a churn prediction using Modified Ensemble Undersampling-Boost and churn prediction performance using Undersampling-Boost.
2. The specification of the hardware used in this research is varied, so elapsed time is not taken into the analytical process

## 1.9 Significance of the Study

In this research, proposed methods could increase the accuracy of the churn prediction and perform under label noise and outperform Random Undersampling-Boost.