Chapter 21
Visualization of Predictive Modeling for Big Data Using Various Approaches When There Are Rare Events at Differing Levels

Alan Olinsky
Bryant University, USA

John Thomas Quinn
Bryant University, USA

Phyllis Schumacher
Bryant University, USA

ABSTRACT

Many techniques exist for predictive modeling of a bivariate target variable in large data sets. When the target variable represents a rare event with an occurrence in the data set of approximately 10% or less, traditional modeling techniques may fail to identify the rare events. In this chapter, different methods, including oversampling of rare events, undersampling of common events and the Synthetic Minority Over-Sampling Technique are used to improve the prediction outcomes of rare events. The predictive models of decision trees, logistic regression and rule induction are applied with SAS Enterprise Miner (EM) to the revised data. Using a data set of home mortgage applications, misclassification percentages of a target variable with a rare event occurrence rate of 0.8% are obtained by running a multiple comparison node. The percentage is varied from 0.8% up to 50% and the results are compared to see which predictive method worked the best.

DOI: 10.4018/978-1-5225-3142-5.ch021

BACKGROUND

Traditional predictive modeling techniques such as logistic regression and decision trees generally work quite well with a nominal target variable. In addition, the newly developed predictive modeling technique of rule induction available in SAS Enterprise Miner (EM) is an alternative method for predicting a bivariate target variable. This technique has shown promise with a rare event target variable which is a bivariate variable in which the event of interest either occurs (1) or does not occur (0) and the rate of occurrence is approximately 10% or less.

However, with a rare event variable, when the data set is very large and the event of interest is very small, all three of these methods typically fail to detect the rare event. This can be particularly true when the percentage of the rare event is extremely small, as in the example presented here, where the rare event occurs in less than 1% of the observations. In this situation, all observations are predicted as being in the group consisting of the more dominant or common event. For example, in the case of a binary target variable, if the rare event makes up only about 1% of the sample, these methods can be correct 99% of the time by predicting all items as falling in the event with the greater probability. In this way, these methods are correct 99% of the time, but may fail to predict the rare event which is most often the aim of the analysis. Unfortunately, many times these rare events, such as fraud, terrorism, etc., are important to predict.

This problem of imbalanced data has been considered before in applications to a variety of fields such as churn models (Guzman 2015, Lemmens & Croux, 2006, Burez & Van den Poel, 2009), credit scoring (Brown & Mues, 2012), crop insurance fraud (Jin & Little, 2005), cyber threats of network intrusions (Dokas et al., 2002), protein classification (Zhao, Li, Chen & Aihara, 2008), telecommunication equipment failures (Weiss & Hirsh, 2000), landslide prediction (Van Den Eeckhaut, et al., 2006), bankruptcy (Foster & Stine, 2004), cardiac surgery (Yap et al., 2014) and detecting rooftops from overhead imagery (Maloof, 2003). King and Zeng (2001) consider modifications to logistic regression to improve model performance. Chawla, Bowyer, Hall and Kegelmeyer (2002) utilize nine different data sets ranging from diabetes to forest cover. In addition, there have been reviews of imbalanced data including Chawla (2005), Visa and Ralescu (2005), Kotsiantis, Kanellopoulos and Pintelas (2006), Han, Yuan and Liu (2009), Sun, Wong and Kamel (2009), Galar, Fernandez, Barrenechea, Bustince and Herrera (2012) and Elrahman and Abraham (2013).

As mentioned, running such an imbalanced model using a data mining software package would fail to detect these important rare events. To try to correct this problem, there are possible methods that have been suggested. These include, among others, oversampling the rare event undersampling the more prevalent event and Synthetic Minority Over-Sampling Technique (SMOTE).

DATA

In this study, a very large real world dataset is used. The data were publicly provided by the Federal Financial Institutions Examination Council (FFIEC) and are available through the Home Mortgage Disclosure Act (HMDA, https://www.ffiec.gov/hmda/hmdaflat.htm). Specifically, the dataset that was analyzed included mortgage applications submitted to Wells Fargo in 2014. This data set consists of 29 variables with 1.1 million cases. The code sheet for the variables is provided in the first appendix. From these variables, *Action Taken,* the action taken by the bank, was chosen as a natural target vari-