The Power of Sampling and Stacking for the PAKDD-2007 Cross-Selling Problem

Paulo J.L. Adeodato, NeuroTech Ltd. and Federal University of Pernambuco, Brazil
Germano C. Vasconcelos, NeuroTech Ltd. and Federal University of Pernambuco, Brazil
Adrian L. Arnaud, NeuroTech Ltd. and Federal University of Pernambuco, Brazil
Rodrigo C.L.V. Cunha, NeuroTech Ltd. and Federal University of Pernambuco, Brazil
Domingos S.M.P. Monteiro, NeuroTech Ltd. and Federal University of Pernambuco, Brazil
Rosalvo F. Oliveira Neto, NeuroTech Ltd. and Federal University of Pernambuco, Brazil

ABSTRACT

This article presents an efficient solution for the PAKDD-2007 Competition cross-selling problem. The solution is based on a thorough approach which involves the creation of new input variables, efficient data preparation and transformation, adequate data sampling strategy and a combination of two of the most robust modeling techniques. Due to the complexity imposed by the very small amount of examples in the target class, the approach for model robustness was to produce the median score of the 11 models developed with an adapted version of the 11-fold cross-validation process and the use of a combination of two robust techniques via stacking, the MLP neural network and the n-tuple classifier. Despite the problem complexity, the performance on the prediction data set (unlabeled samples), measured through KS2 and ROC curves was shown to be very effective and finished as the first runner-up solution of the competition.

Keywords: stacking, sampling strategies, multilayer perceptron, n-tuple classifier, cross-selling, data mining, database marketing

INTRODUCTION

The Pacific-Asian Conference on Knowledge Discovery in Databases (PAKDD) has been organizing open competitions on real world problems for stimulating the development of effective business solutions. This year the PAKDD-2007 Competition presented the challenging cross-selling business problem of a consumer finance company for developing a propensity scoring model. The goal was to point out the credit card customers who would be more likely to accept a home loan (mortgage) offer.
The task was to produce a propensity score for each customer in an independent prediction dataset, separated for performance evaluation, indicating how likely he or she would be for taking up a home loan with the company, with higher scores meaning higher propensity.

In this particular problem, the overlap between the portfolio of credit card customers and the base of home loan (mortgage) customers was very small. In the dataset of 40,700 customers made available for modeling, only 700 belonged to both bases and were labeled with the target for having higher scores in the propensity model. The prediction set consisted of 8,000 customers with both the target labels and the proportion of mortgage offer acceptance undisclosed. For all the data (modeling plus prediction sets) employed, there were 40 modeling variables with information collected at the credit card application moment.

These data referred to a random sample of customers who had opened a new credit card with the company within a specific 2-year period but did not have an existing home loan with the company. The target variable has the value 1 if the customer opened a home loan with the company within 12 months after opening the credit card account and the value 0 if otherwise.

Two main challenges were present in this cross-selling problem, the first being the small amount of data from the positive class in the modeling data set, and the second being the unknown proportion of classes in the prediction data set. As a matter of fact, there was 4.375% of the positive class in the prediction set representing more than 2.5 times the 1.72% proportion found in the modeling set. Therefore, the problem solving task involved much more complexity than the simple application of classification techniques, as will be described in the next sections.

This article is organized as follows. Section 2 presents the technical approach taken. Section 3 presents the creation of new modeling variables. Section 4 describes the data preparation. Section 5 presents the 5-fold cross-validation sampling strategy. The modeling system is described in Section 6. The performance evaluation and methodology is described in Section 7 and the results and comparisons are presented in Section 8. Section 9 points out the most relevant aspects of the research carried out.

**TECHNICAL APPROACH**

As mentioned in the previous section, the main challenges of this cross-selling problem were the small amount of data from the positive class in the modeling data set and the unknown proportion of classes in the prediction data set. The employment of an efficient sampling strategy and the choice of robust modeling techniques were the approach taken for dealing with the situation.

Independent of the amount of data available, a complementary idea applied was to use additional information from the application domain based on expert knowledge representation to boost the solution performance. That was accomplished by applying the expertise of domain professionals supported by data analysis.

The work conducted here presents not only the solution developed for the competition, but also its performance evaluation on the prediction data set made available after the competition results were released. The solution development approach was decomposed in stages for extracting the most out of the available modeling data while allowing a reliable estimation of the scoring system quality. This consisted of the following technical tasks:

1. Creation of new input variables
2. Data preparation
3. Stage 1: Modeling with a single technique (MLP) on a single sample with a statistically independent hold-out set for defining the minimum performance expected for the solution
   3.1. Standard division of the data set for MLP network modeling (50% for training, 25% for validating, 25% for testing)
   3.2. MLP training
   3.3. Performance evaluation on the statistically independent test set
LF-LDA: A Supervised Topic Model for Multi-Label Documents Classification
*International Journal of Data Warehousing and Mining* (pp. 18-36).
[www.igi-global.com/article/lf-lda/202996?camid=4v1a](http://www.igi-global.com/article/lf-lda/202996?camid=4v1a)