ABSTRACT

Today’s credit card issuers are increasingly offering a broad range of products and services with separate lines of business responsible for different product groups. Too often, the separate lines of business operate independently and information available to one line of business may not be used productively by others. In this study, we examine the potential of using information from customers of multiple products to identify customers most likely to respond to cross-sell product offers. Specifically, we examine the potential for offering home loans to a population of credit card holders by studying individuals who do hold both a credit card and a mortgage with the card issuer. Using real world data provided to the 2007 PAKDD data mining competition, we employ Friedman’s stochastic gradient boosting (MART™, TreeNet® ) for the rapid development of a high performance cross-sell predictive model.

Keywords: boosting; MART; mortgages; targeted marketing; TreeNet; tree ensembles

INTRODUCTION

This report describes our participation in the PAKDD2007 data mining competition. The article is organized as follows. In Section 1 we offer our understanding of the competitive challenge, the data available, and how we framed the modeling objectives. In Section 2 we provide a summary of the key descriptive statistics that provide an initial picture of the nature of the data and its adequacy for modeling purposes. Section 3 describes our modeling methods and reports our results and performance based on the labeled data. Section 4 delves further into the results to examine specific findings at the predictor level. Finally, Section 5 summarizes our results and offers conclusions.

THE MODELING CONTEXT

The data provided for the PAKDD 2007 modeling competition consisted of historical records.
for each of 40,070 customers of a consumer finance company. The records included customer demographics, residential and employment history, income category, summaries of credit card use, and various components of credit bureau reports for the prior 3-, 6-, and 12-month periods. The training data came in the form of a flat file containing 40 modeling variables (columns) for customers who had opened a new credit card account in a 2-year observation window and who did not already have a home loan with the company. Seven hundred of these customers also signed up for a home loan within 12 months after opening their credit card account. The mortgage customers were flagged as “1,” while the remaining customers were flagged as “0.” Only 1.7% percent of these customers opened a home loan account, putting this problem into the category of “rare event modeling.”

The organizers of this competition are to be commended for acquiring a substantial volume of real world customer data for public release. Such data can rarely be acquired without limitation. In this instance, the limitations pertain to the data fields made available, and to the descriptive information characterizing the data. We know that the data were drawn from a financial institution, but we are not given the time period from which the data were drawn, and several valuable fields are provided with partial information only. For example, an important variable CUSTOMER_SEGMENT with 11 categories was supplied without further elaboration on the actual meaning of those categories. We are provided with no information regarding the competitive landscape, the nature of the marketing campaigns, trends in the real estate market, or the overall market share in the country for the institution in question.

These informational limitations severely restrict both the business value of any models developed and our ability to extract real world insight into the workings of the marketplace or into consumer behavior.

Another fine point must be made concerning the actual loan application process. The mere fact that someone is looking for a home loan does not guarantee that the loan will be approved. The 700 loan holders in the learn data are the end result of a multistage process involving both consumer initiative as well as lender’s decision making policies. Knowing the specifics of the process as well as time dimensions may in some cases drastically improve modeling solutions.

The formal description of the competition data was confined to one page. Therefore, we resorted to making plausible assumptions about its nature. Our tests (reported below) establish that the train and prediction sets were very likely drawn from the same customer population and from the same time period.

DATA OVERVIEW

To get a better grasp of the data, we grouped the available predictors into the following illustrative categories (not all variables are listed) shown in Table 1.

Based on our past experience, we expected the bureau and demographic groups to be more predictive than the self-reported application data, which seemed to be incomplete and may be susceptible to misrepresentation.

Table 1. Grouping of predictors

<table>
<thead>
<tr>
<th>Group</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Marital status, dependents, age, employment duration, residence duration and district, occupation, income</td>
</tr>
<tr>
<td>Self-Reported</td>
<td>Checking and savings account indicators, major credit card indicators</td>
</tr>
<tr>
<td>Bureau</td>
<td>Number of bureau inquiries in the last 3, 6 and 12 months for loans, mortgages, and consumer credit; default data</td>
</tr>
</tbody>
</table>
Related Content

Optimization of Mean and Standard Deviation of Multiple Responses Using Patient Rule Induction Method

The Challenges of Data Cleansing with Data Warehouses
Essentiality of Machine Learning Algorithms for Big Data Computation
www.igi-global.com/chapter/essentiality-of-machine-learning-algorithms-for-big-data-computation/150198?camid=4v1a

Efficient Compression and Storage of XML OLAP Cubes
www.igi-global.com/article/efficient-compression-and-storage-of-xml-olap-cubes/129522?camid=4v1a