Paper 077-30

The University of Alabama and SAS® Data Mining Certificate Partnership

J. Michael Hardin and Michael D. Conerly,
The University of Alabama, Tuscaloosa, Alabama

ABSTRACT: In March 2002, The University of Alabama (UA) and SAS® developed a partnership whereby students who complete a sequence of four graduate-level courses can receive a joint Data Mining Certificate from UA and SAS®. Recently, the opportunity to earn this certificate has been made available to MBA students as well. This presentation will discuss the data mining curriculum at UA and the advantages of the partnership with SAS®. The market demands for students with this certificate and lessons learned in developing and implementing the program will also be covered.

INTRODUCTION

Recent computing advances have created an increased demand for Business Intelligence professionals. We understand Business Intelligence (BI) to mean activities related to summary, reporting (including OLAP and drill down techniques) and data mining. The demand for BI professionals encompasses industries ranging from financial services to manufacturing to healthcare.

According to Daniels College of Business at the University of Denver, "Though closely tied to statistics and information technology, data mining cuts across all business functions and draws on expertise at every level of an organization … drawing on the expertise of many different disciplines …". In a recent presentation, by Dr. Robert Hannum (2004), Associate Professor at the University of Denver observed that there was a tremendous need for BI professionals and very little training available to meet the demands of industry.

In evaluating job skills required in this emerging area, the Gartner Report (2003) cited SAS® as the clear leader in business intelligence platforms. Additionally, a review of recent job postings indicated that the premiere software required by industry for business intelligence / data mining was SAS®'s Enterprise Miner (Sue Walsh presentation, 2004). As is well known (Walsh, 2004) SAS® is used by

- More than 40,000 business, government & university sites
- 115 different countries
- 90% of Fortune 500; 98 of top 100 of Forbes Super 500.

Clearly SAS® is the vendor of choice in business intelligence.

UA DATA MINING CURRICULUM

After reviewing the demands of the marketplace, we chose business intelligence as a strategic direction for our MS program in Statistics. Further, it was apparent to us that partnering with SAS® and the use of SAS® software in our program would enhance job prospects for students completing our curriculum. Hence, in March 2002, we entered into a partnership with SAS® whereby students completing four graduate level courses would receive a joint UA/SAS® data mining certificate. The courses would utilize leading SAS® analytic and data warehousing technologies as well as SAS® award-winning data mining tool, Enterprise Miner™. By partnering with SAS®, we felt we could more quickly provide a high quality educational experience for our students to meet the growing marketplace demand. These four courses were new courses in the Statistics curriculum at the University of Alabama.

The four courses we developed are:

- ST 521 Statistical Data Management
- ST 522 Advanced Statistical Data Management
- ST 531 Introduction to Data Mining
- ST 532 Advanced Data Mining

In ST 521 we present the fundamentals of SAS® programming, with an emphasis on proper data management techniques. At the conclusion of this course a student should be prepared to take the SAS® base certification exam. We continue the concepts from ST 521 into ST 522 giving the students a thorough knowledge of SAS® MACRO, SAS® SQL and SAS® ODS. An emphasis throughout this course is the application of these tools for data cleaning and preparation and for producing professional quality reports. Additionally, in this course we introduce students to the ideas of relational database design including data modeling and entity relationship diagrams. We also discuss

data models for data warehousing and how to access corporate warehouses. This class concludes with a real case study involving data from multiple sources using many formats and containing many data quality problems. These two courses form a foundation for the reporting and drill down components of the BI task.

Our data mining sequence consists of two courses (ST 531 and ST 532). In the first course, we introduce the student to the concepts of BI in data mining. We also introduce the Enterprise Miner™. A particular theme of this course is the use of predictive models for BI. We provide broad overviews to the modeling techniques of Logistic Regression, Decision Trees and Neural Networks. The concepts of data partitioning, model assessment specifically lift charts and ROC curves are presented. A key objective of this course is to allow the student to observe and participate in the entire data mining process from data acquisition to final model deployment. This objective is met through a real world project that serves as the culmination of this sequence.

The advanced data mining course provides a more in-depth coverage of the technical aspects of each of the modeling tools discussed in the first course. Additionally topics in Statistical Decision Theory and unsupervised learning are covered. The students gain additional experience with the Enterprise Miner including the use of SAS® code nodes, and other advanced aspects of the uses the Enterprise Miner tools.

Students from other disciplines within the college and university are encouraged to participate in these courses as well. To date we have accepted students with graduate standing that had a working knowledge of basic statistics. We have had students from Industrial Engineering, Marketing, Finance, Economics, MBA program as well as Statistics successfully complete the certificate.

EXPERIENCE WITH CURRICULUM

This curriculum has worked well for us so far. However, we have adjusted the sequence of the courses to satisfy the demand that has arisen. Specifically, our original plan was to off the courses over a two year period in the sequence data management followed by data mining in each year, i.e., ST 521, ST 531, ST 522 and ST532. To accommodate students completing the sequence in one year we have adapted to the sequence both introductory courses in the fall semester followed by the advanced courses in the spring semester.

We have not found this to be problematic due to the careful selection of data sets for analysis in the first data mining course. Essentially in the first semester we separate the data management issues from the predictive modeling issues in the first two courses. We have found it is difficult to find appropriate texts and have relied on the Academic Trainer's Program provided by the SAS® Higher Education Consulting group. This program assists faculty by supplying course notes, PowerPoint slides, and data sets. Currently we are using *Step by Step Programming with Base SAS® software* for programming courses and *Applied Data Mining* by Paolo Giudici (2003) for data mining courses. We provide some time in class for completion of assignments. Students spend some time in the computer lab working individually and use e-mail for questions.

We have had seven students to compete the certificate in 2003 and we expect 12 more this year. Many of these students with BI skills are in demand for desirable jobs that pay well. We have placed one student at Key Bank, one at the University of Alabama at Birmingham. Several of our first year certificate students kept their current positions and several others have been recruited but decided to stay in our Ph.D. Program. In recent discussions with firms visiting the campus there has been great interest in our data mining certificate.

LESSONS LEARNED AND CHALLENGES

There are a number of technical issues that we have encountered. A stable lab environment that supports large data sets is vital. A key issue is providing ample space for students to store large data sets and the associated intermediate data sets generated by the EM. When students are working on projects, security issues are important as well as frequent reminders to back up one's work.

Another challenge is overcoming older negative connotations of the term Data Mining (Studenmund, 2001) among some faculty. Often, data mining has been thought of as *fishing* or *data dredging*. We have found it important to emphasize to students and colleagues a thorough understanding of the issues inherent in the antiquated usage of the term *data mining*. We have emphasized terms such as predictive modeling, data partitioning and validation samples for assessing the performance of the estimated models in general settings.

Explaining to students with diverse backgrounds the scope of large, complex data sets and the processing time these require in terms that they can understand is difficult to convey to students without hands-on experience. Obtaining realistic data sets, for case studies, homework assignments and projects is a continual dilemma. We have had positive experiences working with real corporate partners such as our Registrar's Office, where the class project

focused on the freshman retention rate at UA. In our case, the student projects for the Registrar's Office served as a springboard for on-campus collaboration, and have led to the creation of a Retention Task Force in 2005 made up of faculty and staff from many areas of the University of Alabama. This experience takes a student far beyond the classroom setting and prepares them for real jobs.

The data sets used for the KDD cup are another source for interesting data. Students tend to severely under estimate the calamity of messy data. Providing a realistic learning experience is an asset for our program. We have also been able to find a number of corporate partners who use SAS® and Enterprise Miner™ and have been willing to allow us to work with them on real problems. They have found the exchange of information to be mutually satisfactory. Two of our most successful partnerships were with Southtrust Bank and Wise Alloys (see , http://birmingham.bizjournals.com/birmingham/stories/2003/11/03/daily1.html and http://dialog.ua.edu/dialog20030609/datamining20030609.html).

FUTURE DIRECTIONS

The UA/SAS® data mining certificate program has been more successful than we anticipated. In the fall of 2004 the Economics program altered their "applied" MS track to incorporate the 4 course data mining certificate sequence. The Economics faculty believed that this would improve the potential for job placements for their students. The feedback from these students enrolled this year has been extremely positive. This spring, these Economics MS students have attended the Statistics seminars and presented their ST 531 projects in February. Currently we are tapping the students in the second data mining class to create a research group that will focus on credit scoring and financial risk assessment. All of the Economics students and one accounting PhD student have opted to participate.

The success of the program has also caught the attention of the MBA program which is currently being revised. The proposed curriculum includes a data mining concentration and all MBA students will take a new half-semester course on business intelligence. The focus of the revamped MBA program was to increase the technical abilities of these students and to provide them with marketable skills upon graduation.

The partnership with SAS® has been a richly rewarding experience for UA. It has allowed us in only 2½ years to develop a recognized presence in the business intelligence / data mining community and to provide our students with valuable skills in great demand in the current business environment. The partnership with SAS® has been instrumental in developing our certificate program quickly from scratch. Without their support, we would not have the visibility across campus, regionally or nationally that we now enjoy. We anticipate even greater success from this mutually beneficial partnership in the future. The notion that employing data to make useful business decisions is not a fad; but rather an imperative for international business in twenty first century and beyond.

References

Gartner report(2003), "SAS® featured in Gartner business intelligence platform Magic Quadrant" http://www.sas.com/news/preleases/021303/news3.

Giudici, Paolo (2003) *Applied Data Mining*, John Wiley and Sons, New York.

Studenmund, A.H. (2001) *Using Econometrics, a Practical Guide*, fourth edition, Addison Wesley Longman, Inc., Boston.

Step by Step Programming with Base SAS® software (2001) SAS® Publishing.

Walsh, Sue (2004) presentation at the SAS® Data Mining Conference for Higher Education, Cary, NC.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Author Name: J. Michael Hardin
Company: Department of Information Systems, Statistics and Management Science, University of Alabama
Address: BOX 870226
City state ZIP: Tuscaloosa, AL 35487-0226
Work Phone: (205) 348-8901
Email: mhardin@cba.ua.edu
Web: www.cba.ua.edu/~apstat