11-1-2009

Data Mining as Applied to the Social Sciences

Walter Morris, Dyson School of Arts and Sciences
Pace University

Title of the Project: Data Mining as Applied to the Social Sciences

Cornerstone # 3

Principal Investigator: Dr Walter Morris, Dyson School of Arts and Sciences, Economics Department

Date: 11/11/2009

Original Goals/Progress

To date, most of the grants' funding distributions have been for purchases covering SAS Enterprise Miner 5.3 and XLMiner software and related expenses. I begin the report with a spending summary of these funds.

SAS

I have received a Thinkfinity grant to pursue research in the area of Data Mining as Applied to the Social Sciences. To date, I have been successful in purchasing and installing the SAS Enterprise Miner 5.3 computer add-in package on the Pace operating system (Cost=$2,896). SAS Enterprise Miner 5.3 is an add-in to the current SAS educational software package that includes SAS Basics, SAS Statistics, SAS Operations Research, SAS Econometric Time Series, and SAS Graph, amongst other baseline programs. Currently, Pace has 25 licenses of SAS Enterprise Miner 5.3 that are installed on the Pace computer facilities located in Wilcox Hall. These facilities are available to anyone in the Pace community, students and faculty alike, wishing to utilize them. Quite frankly, this program is state of the art compared to other Data Mining computer software merchandise. In my view, it is far superior to its major competitor, Clementine, which is an SPSS product. Also, I have gained the support of Dyson College by having the Dean agree to an extended renewal of the SAS Enterprise Miner licenses into next year (i.e., July 2010-July 2011) at a cost of $1,239. I am also in the process of establishing contact with the new Pace CIO, Dr Ravi Ravishanker, to determine the type of support DOIT will provide for the successful completion of this project.

As you are well aware, exceptional software is a necessary but not a sufficient condition for the successful implementation of this type endeavor. With that in mind I am actively engaged in improving my knowledge of this computer system since SAS Enterprise Miner is not an easy

program to master. Towards this end, I have successfully completed several SAS seminars on Data Mining—the initial workshop being 'Data Mining Techniques: Theory and Practice' (June 3-5 in NYC---3-day workshop @ cost=$1,080). This workshop provided an in depth overview of Data Mining but had few SAS applications. The workshop was headed by two prominent leaders in the Data Mining field (i.e., Dr. Michael Berry and Dr. Gordon Linhoff, 'Data Mining Techniques', Second Edition, Wiley) and as such provided an excellent overview of topical interest and highlighted some of the most recent advances in the subject.

The second workshop attended was 'Applied Analytics Using SAS Enterprise Miner 5.3' (June 17-19 in NYC--3-day workshop @ cost=$1,080). It offered a distinctive overview of the entire SAS Enterprise Miner product and, as such, was more of a cookbook approach to the programs' utilities with several applications in a wide range of fields. As such it was more SAS intensive and specifically tailored towards resident SAS Enterprise Miner 5.3 programs.

Data Mining mainly uses three predictive techniques to forecast outcomes of interest; Decision and Regression Trees (CART), Logistic Regression Analysis, and Neural Networks. My particular expertise is in the subject of Regression Analysis. Decision Trees and Neural Networks are areas that I have limited experience and knowledge. To partially fill this gap, I have taken two additional workshops in Decision Tree Modeling (July 23-24 in NYC--2-day workshop @ cost= $725) and Neural Network Modeling (October 28-29 in Las Vegas--2-day workshop @ cost= $725).

During the last Neural Network workshop I also attended the largest national Data Mining convention in the country. The convention was held in Las Vegas (October 25-27—3-day Conference registration cost=$1,499 waived by a SAS grant). During the conference I was able to observe what cutting edged researchers are carrying out in this new and exciting subject. The sessions themselves were an interesting blend of general interest presentations combined with more specialized papers given by prominent researchers in the field. Several presentations in both Data and Text Mining were intriguing—one in particular that caught my eye was entitled 'Text Mining to Discover Influential Communications in Social Movements'. Indeed, after attending the conference, I returned with several new avenues of research that I would like to pursue in the future.

At some future date, I will be attending another SAS workshop, 'Exploratory Data Mining with Applications to Life and Social Sciences' (3-day workshop @ cost=$1,080). As the title suggests, this clinic is tailored towards applications in the Life and Social Sciences and as such provides a unique look into Data Mining as applied to these disciplines. Since many Data Mining

applications are currently in the Business area, I believe that this workshop beneficial to the successful completion of my project which focus on the Social Sciences. The workshop emphasizes applications in such diverse fields as Biology (DNA targeting), Economics (Prediction of Mortgage Defaults), Psychology (which is where Neural Networks originated), and Applied Statistics, to mention just a few.

I would like to add that the 'Exploratory Data Mining for the Life and Social Sciences' workshop contains sessions/examples that use Data Mining that might be of casual interest to a broader audience within Pace University. The application I'm referring relates to University student retention. To appreciate the totality of Data Mining, consider for a moment student retention rates that are derived from observable student attributes. Given certain measurable student characteristics, Data Mining is able to rank and assign probabilities to students who are most likely to exit the University. This could be a critical piece of information for a University with low retention. Rather than randomly selecting students who might exit the University, Data Mining modeling gives the analyst a clear set of probabilities identifying students most likely to leave. Given the University knows which students are most likely to exit, early retention strategies can be implemented to retain those individuals. I do believe that Dyson College and Pace University are interested in these questions. (I also consider the current Pace strategy of waiting until a student applies for a transcript is too late—students applying for transcripts have for the most part already decided to leave Pace. Earlier detection flags are needed to identify potentially exiting students and Data Mining provides a methodology to discover those individuals.)

XLMiner

I have also purchased XLMiner (cost=$199) which is a Data Mining program used as an EXCEL 'add in'. Major advantages of this package are user friendliness (which SAS isn't!) and its integration into the EXCEL spreadsheets products. A major drawback is its capability in handling large data sets and also it has limitations with regards to Data Mining procedures. If I were to review the two products, SAS is more like a BMW 700 series of Data Mining, while XLMiner is the equivalent of a Volkswagen Rabbit. However, ease of use has advantages especially if interest resides in understanding Data Mining and not spending an inordinate amount of time comprehending how to maneuver through complex computer algorithms.

I have taken two 4-week on-line workshops with 'Statistics.com' that relate to Data Mining and more specifically, the XLMiner product. The first 'Introduction to Data Mining' (Sept 11-Oct 15 @ cost= $394) dealt mainly with Data Mining for predictive purposes; that is, assuming that

there is a 'target' or 'dependent' variable to analyze. This workshop emphasized CART, Regression, Neural Network Modeling, k-Nearest Neighbor, and Decision Tree Modeling as applied to XLMIner. The second workshop, 'Data Mining 2: Unsupervised Training' (Oct 16-Nov 15 @ cost= $394), is tailored toward Data Mining issues where no 'target' or 'dependent' variable is offered. Subjects included Cluster Analysis, Association Rules, and Principal Components Analysis. Again the workshop was customized toward the XLMiner programs as well as a few other Data Mining packages.

Research Assistant

I have also hired a Research Assistant, a Mr. Jackson Fallon, who is an undergraduate student majoring in Economics. Jackson and I meet a minimum of twice a week. Jackson has been instrumental in installing data sets into SAS Enterprise Miner 5.3. These data sets are unique and different from traditional data step entries used in other SAS procedures such as SAS/ETS. Jackson has also been studying the use of CART (Classification and Regression Trees) as applied to various applications in Enterprise Miner. Jackson has been an invaluable addition to the project and his research efforts with the grant will continue into the upcoming Spring semester.

Students/Faculty Impacted

What initially sparked my interest in Data Mining is the topic is beginning to turn up in several Economic Forecasting textbooks (i.e., Wilson and Keating, 'Business Forecasting', Sixth Edition, 2009, McGraw-Hill, Chapter 9). Since I teach several Quantitative Forecasting courses to Business, Economics, CS/IS and Math majors on a continuing basis, I felt a need to acquaint myself with the subject and hence my Thinkfinity proposal. As a result, I have included in my Forecasting courses (i.e., Eco 240—Quantitative Analysis and Forecasting, Eco 380--Econometrics, Eco 296A—Advanced Economic Forecasting, and Eco 296M—Forecasting for Non-Profits) several concluding lectures highlighting Data Mining. This impacts over 100 students on a yearly basis.

I'm also pleased to report that I've progressed far enough with the Thinkfinity project that a new course will be offered this coming Spring (2010) semester entitled 'Forecasting Applications with Data Mining' (Eco 396). Much needs to be done over the next couple of weeks with regards to course development. Presently, I'm in the early stages but have selected a textbook (Shmueli, Patel, and Bruce 'Data Mining for Business Intelligence', Wiley, 2007) and written a tentative syllabus. As previously stated, the SAS and XLMiner software are already in place as this

writing. I have collected several applicable data sets from a diverse assortment of disciplines and have made them compatible with the SAS and XLMiner software. This new course should impact approximately 25 students for the upcoming semester.

One positive spinoff is that this course will provide Pace students the opportunity to engage in quality research not generally afforded students at other universities. What was interesting at the Data Mining conference that I attended in Las Vegas is that several Universities now offer advanced/graduate degrees in Data Mining and that more are starting to offer Data Mining courses at the undergraduate level. Many are offering certificates in Data Mining and have forged partnerships with SAS. With Pace using the state of the art SAS Data Mining programs, courses such as 'Forecasting Applications with Data Mining' give Pace students the competitive edge necessary in today's labor market. Indeed, many Lubin Business students, Seidenberg CS/IS students, and Dyson Liberal Arts students all include on their job search resumes knowledge of SAS as a skill enhancement—a skill gained from taking Forecasting courses designed and taught by the Economics Department at Pace. Inclusion of SAS Enterprise Miner would only add to that impressive array of skills.

I would also like to add that since the SAS software is available to the entire Pace community and SAS Enterprise Miner 5.3 is an add-in to this software package, technically the project could touch the entire Pace population that uses SAS.

In addition to this new undergraduate course, several of my colleagues have expressed an interest in Data Mining for own research purposes. To address these interests, I fully expect to offer a workshop in the Spring 2010 semester to interested faculty members. This workshop will feature an overview of Data Mining and look at the SAS Enterprise Miner 5.3 package in particular. It is hoped that shared interests in Data Mining can stimulate other faculty to pursue this rich source of analytical research.

Next Steps

Much of my efforts for the remaining part of this Fall semester are geared towards pulling together the new course in Data Mining (Eco 396). I'll be spending time on syllabus preparation, data installation, and program accessibility in order to get this course underway. In addition, I'll be attending another SAS workshop in 'Exploratory Data Mining for the Life and Social Sciences'.

Next Spring semester, I'll be teaching the new Eco 396 course in 'Forecasting with Data Mining' and continue incorporating Data Mining into the Forecasting related courses I already teach.

Also, I have been in contact several non-profit institutions, such as the Westchester chapter of the American Red Cross, to explore possibilities in these institutions for Data Mining type services. There seems to be an interest in using Data Mining to identify new contributors that have a strong likelihood for future or repeat donations. With regards to publications, to date I have not conducted any research that is publishable quality. I will begin exploring these possibilities now that major purchases for the project in the complimentary areas of physical and human capital have finished.