7-1-2010

Data Mining as Applied to the SocialSciences

Walter Morris
Dyson College 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: 07/15/10

Original Goals/Progress

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

SAS

I 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 as 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 have established contact with the new Pace CIO, Dr Ravi Ravishanker, and he has agreed to support the continuing Data Mining 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 teaching and research that I would like to pursue in the future.

In November 11-13, I attended 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 this workshop beneficial to the

successful completion of my project which focuses 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 contained 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 likely to exit, early intervention strategies can be implemented to retain those individuals. I do believe that Dyson College in particular and Pace University in general are interested in this issue.

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, exclusively tailored to 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 adapted 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 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 for two semesters. 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 .

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, 2010, 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 have attached a Syllabus for Eco 240)

I'm also pleased to report that because of the Thinkfinity project, a new course was offered this past Spring (2010) semester entitled 'Forecasting Applications with Data Mining' (Eco 396). 10 students were enrolled in the course. I have attached a course Syllabus and a considerable amount of Blackboard material is available upon request. This Blackboard material includes PowerPoint lecture slides, several write-ups concerning the covered course material, and sample computer programs used in class. The selected textbook was Shmueli, Patel, and Bruce 'Data Mining for Business Intelligence', Wiley, 2007.

In addition, a Data Mining segment has been added to my Economics 240 classes (Quantitative Analysis and Forecasting). I have set aside three hours of lecture time during the course of the semester. Since I taught two sections of Eco 240 last Fall semester, and one section in the Spring term, approximately 60 students were impacted.

As previously stated, the SAS and XLMiner software are in place. I collected several applicable data sets from a diverse assortment of disciplines and have made them compatible with the SAS and XLMiner software. Included in these data sets were individual Bank Failures, Private Contributions/Donations to the Veterans Administration, Housing Values in the Boston Metropolitan Area, Book Selections from the Charles Book Club, etc. For example, the Boston Housing data contained environmental information and the class was able to estimate the impact environmental factors exerted on housing values. It goes without saying that the Environmental Studies program would be interested in these findings. Also using data supplied by a national Veterans Administration, the class was able to develop a model identifying repeat contributors to the Veterans; that is, the model identified those individuals most likely to make a second contribution to the Administration. This type of modeling is helpful to non-profits and the Community Outreach department in collaboration with Project Pericles has expressed an interest in my developing a Data Mining class with an AOK1 designation. We have tentatively identified the Westchester chapter of the American Red Cross as a community partner. A full listing of the data is available upon request.

A positive spinoff is this course provided Pace students the opportunity to engage in quality research not generally afforded students at other universities. What was interesting at the Data Mining conference I attended in October is that several Universities now offer advanced/graduate degrees in Data Mining and more are starting to propose 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 for research/teaching purposes.

In addition to this new undergraduate course, several of my colleagues have expressed an interest in Data Mining for own research possibilities. To address these interests, I fully expect to offer a workshop in the future to interested faculty members (Probably something similar to the Geographical Information Systems (GIS) workshop Dr. Dan Farcas offered to faculty this past

Spring semester—which I attended). 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

A significant allocation of time and effort were spent designing the new course in Data Mining (Eco 396). I spent a considerable amount of time on syllabus preparation, data installation, and program accessibility in order to get this course underway. In addition, I'm attempting to write an article highlighting the advantageous of Data Mining for the Social Sciences.

This coming academic year (Fall '10- Spring '11), I've been awarded a full year Sabbatical Leave. During this time I will be researching opportunities for publication of an article in Data Mining. I am particularly interested in Data Mining applications for non-profit corporations. In this regard I have been in contact with 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 existing contributors having a strong likelihood for future or repeat donations. I expect this will lead into a new course in Data Mining with a community partner component that will satisfy the AOK-1 portion of the core curriculum.

I have also developed an ancillary interest in Text Mining and expect to submit a research proposal to Thinkfinity in this area for the upcoming round of research grants. Essentially the techniques of Text Mining are best understood as an extension of Data Mining's standard predictive methods as applied to unstructured text. Considerable attention to the data preparation and handling methods that are required to transform unstructured text into a form in which it can be texted mined. Text Mining is a logical extension of Data Mining in the sense that unstructured text is reviewed and analyzed rather than structured data sets. Indeed, Text Mining has many applications for the Social Sciences.

Economics 240: Quantitative Analysis and Forecasting

Dr. Walter Morris
E-mail wmorris@pace.edu
Phone 773-3308
Office Hours: Tuesday, Thursday 2PM-3:30PM and
Wednesday 4PM-6PM and by appointment

Textbook, Software, and Data Bases

(1) John E. Hanke and Dean W. Wichern, Business Forecasting, Pearson/Prentice Hall, Ninth Edition, 2009.
(2) The MINITAB statistical/forecasting computer program is available free on the Pace Network. Also, it can also be ordered separately on-line for your PC for a six month period (for a fee of approximately $30 directly from MINITAB).
(3) The ECONOMAGIC data base which is supplied on the Pace Network and is free.

Course Objectives

This course is designed to give students a detailed understanding of the mathematical methodologies associated with economic and financial forecasting. Emphasis is placed on five forms of forecasting: i.e., Regression Analysis, Exponential Smoothing, Time Series Decomposition, Autoregressive Integrated Moving Averaging (ARIMA) models, and Data Mining. Students will also be required to demonstrate competence using the statistical/econometric forecasting programs residing in MINITAB. There is also a brief introduction to the Data Mining computer software programs residing in SAS Enterprise Miner 5.3 and XLMiner which is an EXCEL add-in. Data bases, in particular ECONMAGIC, are also explored in detail.

At the conclusion of the course students are expected to:

1. Understand forecasting theories as they relate to current economic/financial issues.
2. To evaluate economic issues from a variety of different forecasting perspectives.
3. To critically evaluate different research methodologies with regards to current forecasting problems.
4. To demonstrate critical analytical and thinking skills as they relate to forecasting issues.
5. Be able to evaluate forecasting issues from both a global as well as a national perspective.

6. To demonstrate competence in several econometric/forecasting computer packages such as MINITAB, SAS, SAS Enterprise Miner 5.3, and XLMiner.
7. To demonstrate competence in handling data bases such as ECONOMAGIC as applied to forecasting issues.

As partial fulfillment towards receiving credit for the course, students must sit for two examinations; a midterm and final. Each of these examinations is worth 25 points. In addition, subject to the instructors' approval, students are required to submit a 5-10 page research paper that demonstrates to the instructor a measure of competence in the forecasting area. The topic of the paper should be in a specific area of forecasting and it is worth an additional 25 points. Homework assignments located in the "Assignments' folder in Blackboard count for 25% of the final mark.

Students are expected to attend and participate in class. Homework is assigned, required and reviewed in class and is an integral part of the course. The homework is found under the '*ASSIGNMENTS*' menu located in your BLACKBOARD account. **All homework must be submitted to your folder on the date and time specified in the 'Assignment' folder. It cannot be emphasized enough that under any circumstances late submissions will not be allowed.** Students are strongly advised that this course requires an intensive amount of preparation that is most readily grasped in a classroom environment. If you miss a class or a computer workshop the covered material is your responsibility.

SUMMARY OF COURSE REQUIREMENTS

Mid-term Examination	25%
Final Examination `	25%
WEB-BASED Homework Assignments	25%
Final Paper	25%

Course Outline

Topic 1: Statistical Review and an Introduction to Forecasting

Chapters 1-3

A brief review of the statistical concepts and graphical techniques used in class. An exploration of data patterns and overview of forecasting techniques. Forecasting through the use of regression analysis is developed. A brief introduction to MINITAB and ECONOMAGIC is covered in this section and data sets are supplied.

Topic 2: Introduction to Time Series Models

Chapters 4-5

Moving Averages and Exponential Smoothing methods are introduced. Decomposing time series data into trends, cycles, and seasonality's are discussed. Advanced demonstrations of MINITAB and ECONOMAGIC are covered.

Topic 3: Statistical Review and an Introduction to Regression Analysis

Chapter 6

A brief review of the statistical concepts used in class. The important univariate statistics are the t-test and the F-test (i.e., ANOVA). Bi-variate regression is introduced by estimating coefficients using Ordinary Least Squares. Hypothesis testing, confidence intervals, *p*-values, and goodness of fit measures are reviewed. Forecasting through the use of regression analysis is developed. Functional form or non linearity in the variables is covered. Advanced demonstrations of MINITAB and ECONOMAGIC.

Topic 3: Multiple Regressions and Forecasting

Chapters 7

The multiple regression model is estimated and interpreted. Multicollinearity is addressed by investigating its causes, consequences, tests-to-detect, and correction procedures. Dichotomous or Dummy Variables are examined. Point and interval forecasts are developed as are *ex post* and *ex anti* forecasting procedures. Forecasts are evaluated using conventional statistical tests (i.e., root mean square error, Theil's inequality coefficient, etc.) Several economic/financial applications with regard to forecasting are introduced. Advanced demonstrations of MINITAB and ECONOMAGIC.

Topic 4: Time Series Data and Associated Forecasting Problems in Regression Analysis

Chapter 8

Single equation regression problems associated with time series data such as autocorrelation and heteroscedasticity are covered in detail. The approach here is to first identify the cause of the problem, then discuss the consequences, develop a test statistic to detect the problem, and finally come up with a correction procedure to remedy it. Point and interval forecasts are developed as are *ex post* and *ex anti* forecasting procedures. Forecasts are evaluated using conventional statistical tests (i.e., root mean square error, Theil's inequality coefficient, etc.) Several economic/financial applications with regard to forecasting are introduced. Advanced demonstrations of MINITAB and ECONOMAGIC.

Topic 5: Time Series Models

Chapters 4,5, 9

Moving Averages and Exponential Smoothing methods are introduced. Decomposing time series data into trends, cycles, and seasonality's are discussed. Re-visit exponential smoothing and forecasting. Autoregressive Integrated Moving Averaging (ARIMA) models are reviewed. The Box-Jenkins method of model estimation, diagnostic checking, and forecasting are presented and evaluated. Combinations of regression and ARIMA (i.e., transfer functions) are explored. Comprehensive economic/financial models are developed using these methodologies. More advanced demonstrations of MINITAB and ECONOMAGIC are offered.

Topic 6: Data Mining and Forecasting

Handout on Blackboard (Ereserves)

A brief introduction on Data Mining and its applications to forecasting issues. Three techniques of Data Mining are introduced—Classification Trees, Regression Trees, and Neural Networks. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc. A brief introduction to Data Mining computer software XLMiner (an EXCEL add-in) and SAS Enterprise Miner 5.3 are introduced.

Topic 6: Current Issues in Forecasting

Topical issues are discussed and tailored toward specific interests of the student. The topics are selected in accordance with the research interests of the student.

Quality of Homework Assignments

You will receive a grade for each of the Homework assignments and what follows is a brief description of exactly how you are graded. In either case you must submit your assignments on time in order to receive credit.

	A or A-	B+ or B	B- or C+	C or C-	D+ / D / D-
Timeliness	Responses are always posted on time.	Responses are on time - one occasionally missed.	Responses usually on time or occasionally missed	Responses missed or late more than once.	Responses consistently posted late or missed.
Questions	Provides provocative questions on time that help us rethink readings.	Questions adequate but not provocative.	"Surface" questions that provide not critical analysis	Questions off base, inappropriate, or irrelevant for readings.	Doesn't post questions on time.
Thorough	Consistently addresses all parts of questions well.	Usually all parts addressed, some better than others.	Sometimes some parts are missing but what is posted is usually thorough.	Responses to questions are often quite incomplete.	Responses consistently address questions minimally.
Thoughtful Connections	Consistently responses make thoughtful & specific connections to readings. Links theory to practice and across posts, readings, or discussions.	Usually thoughtful connections to readings but often lacking connections made across posts, readings, or class discussions.	Sometimes responses do not demonstrate an under-standing of readings. Connections may be forced or incomplete.	Often responses are lacking in substance and, quite possibly, inaccurate. Connections made may be inaccurate or incoherent.	Responses are consistently done in a hasty fashion with no thoughtful connections made.
Substantive Responses to Classmates	Often reads others' posts and offers informed questions, comments, & connections.	Usually comments on others' posts. Interesting comments but they may not indicate clearly a close read.	Sometimes comments on others' posts. Comments are sometimes off base - not clearly relating to others' posts.	Infrequently comments on others' posts and comments are sometimes inappropriate - unhelpful, not constructive, rude.	Rarely comments on others' posts or, when does, comments are often inappropriate - unhelpful, not constructive, rude.

Other Important Information.

Students must accept the responsibility to be honest and to respect ethical standards in meeting their academic assignments and requirements. Integrity in the academic life requires that students demonstrate intellectual and academic achievement independent of all assistance except that authorized by the instructor. The use of an outside source in any paper, report or submission for academic credit without the appropriate

acknowledgment is *plagiarism*. It is unethical to present as one's own work, the ideas, words or representations of another without the proper indication of the source. *Therefore, it is the student's responsibility to give credit for any quotation, idea or data borrowed from an outside source.*

Students who fail to meet the responsibility for academic integrity subject themselves to sanctions ranging from a reduction in grade or failure in the assignment or course in which the offense occurred to suspension or dismissal from the University. Students penalized for failing to maintain academic integrity who wish to appeal such action may petition the department chair to request a hearing on the matter.

Pace University believes that it is important that students receive appropriate accommodation for any disability. If you have a disability for which you are or may be requesting an academic accommodation, you must register with the Coordinator of Services for Students with Disabilities. Trained professional Counselors will:

-Evaluate your medical documentation;
-Conduct appropriate tests or refer you for same;
-Make recommendations for your plan of accommodation; and
-Contact your professors (with your permission) to arrange for the recommended accommodations.

Your professor is not authorized to provide any accommodation prior to you arranging for same through the Counseling/Personal Development Center. If you have, or believe you have, a disability, be sure to follow the above procedure.

Economics 396: Forecasting and Data Mining

Dr. Walter Morris
E-mail wmorris@pace.edu
Phone 773-3308
Office Hours: Tuesday, Thursday 8:30-10AM and 11:30-12:30PM
and by appointment.

Textbook, Software, and Data Bases

(1) Galit Shmueli, Nitin R Patel, and Peter C Bruce, Data Mining for Business Intellegence, Wiley, 2007.

(2) The XLMiner Data Mining/Forecasting computer program is available free with the textbook. **If XLMiner is not packaged with the text do not purchase the book!**

(3) The ECONOMAGIC data base which is supplied on the Pace Network and is free.

(4) SAS Enterprise Miner 5.3 computer program is provided by Pace free of charge.

Course Objectives

This course is designed to give students a detailed understanding of the mathematical methodologies associated with Data Mining. Emphasis is placed on two forms of Data Mining: Supervised and Unsupervised Data Mining. Students will also be required to demonstrate competence using the Data Mining programs residing in XLMiner. There is also an introduction to the Data Mining computer software programs residing in SAS Enterprise Miner 5.3. Data bases, in particular ECONMAGIC, are also explored in detail.

At the conclusion of the course students are expected to:

1. Understand Data Mining as it relates to current economic/financial issues.
2. To evaluate economic issues from a variety of different Data Mining perspectives.
3. To critically evaluate different research methodologies with regards to current Data Mining problems.
4. To demonstrate critical analytical and thinking skills as they relate to Data Mining issues.
5. Be able to evaluate Data Mining issues from both a global as well as a national perspective.

6. To demonstrate competence in several Data Mining computer packages such as SAS Enterprise Miner 5.3, and XLMiner.
7. To demonstrate competence in handling data bases such as ECONOMAGIC as applied to forecasting issues.

As partial fulfillment towards receiving credit for the course, students must sit for two examinations; a midterm and final. Each of these examinations is worth 25 points. In addition, subject to the instructors' approval, students are required to submit a 5-10 page research paper that demonstrates to the instructor a measure of competence in the forecasting area. The topic of the paper should be in a specific area of forecasting and is worth an additional 25 points. Homework assignments located in the "Assignments' folder in Blackboard count for 25% of the final mark.

Students are expected to attend and participate in class. Homework is assigned, required and reviewed in class and is an integral part of the course. The homework is found under the '*ASSIGNMENTS*' menu located in your BLACKBOARD account. **All homework must be submitted to your folder on the date and time specified in the 'Assignment' folder. It cannot be emphasized enough that under any circumstances late submissions will not be allowed.** Students are strongly advised that this course requires an intensive amount of preparation that is most readily grasped in a classroom environment. If you miss a class or a computer workshop the covered material is your responsibility.

SUMMARY OF COURSE REQUIREMENTS

Mid-term Examination	25%
Final Examination	25%
WEB-BASED Homework Assignments	25%
Final Paper	25%

Course Outline

Topic 1: Statistical Introduction to Data Mining Methodology

Chapters 1-2

A brief review of the statistical concepts and graphical techniques used in class. An exploration of data patterns and overview of Data Mining techniques. Data Mining through both supervised and unsupervised training. Building a Data Mining environment. The virtuous cycle of Data Mining: Training data sets, Validation data sets, and Test data sets.

A brief introduction to XLMiner and Economagic is covered in this section and data sets are supplied. Several economic/financial applications with regard to Data Mining are introduced. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 2: Introduction to Data Exploration and Reduction

Chapters 3-4

Data Preparation and Exploration is developed. Principal Components Analysis (PCA) is introduced as a technique for Data Reduction. Evaluation of Data Mining forecasting performance. Accuracy measures such as lift curves, profit matrix's, classification matrix's, etc. are explored.

Advanced demonstrations of XLMiner and Economagic are covered. SAS Enterprise Miner 5.3 is introduced. Several economic/financial applications with regard to Data Mining are introduced. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 3: Supervised Data Mining: Statistical Review and an Introduction to Predictive Modeling through Multiple Regression Analysis

Chapter 5

Predictive Modeling. A brief review of the statistical concepts used in class. The important univariate statistics are the t-test, Chi-Square test, and the F-test (i.e., ANOVA). Bi-variate regression is introduced by estimating coefficients using Ordinary Least Squares. Hypothesis testing, confidence intervals, *p*-values, and goodness of fit measures are reviewed. The Multiple Regression Model is estimated and interpreted. Dichotomous or Dummy Variables are examined. Binning variables is discussed.

Advanced demonstrations of XLMiner, Economagic, and SAS Enterprise Miner 5.3. Several economic/financial applications with regard to Data Mining are introduced. Case studies. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 4: Supervised Data Mining: Predictive Modeling through Classification Methods

Chapters 6 & 7

Memory based reasoning: k-nearest neighbor. Measurements through Euclidean distances. Classification techniques are introduced; Classification Trees and Regression Trees (CART) are discussed. The problem of overfitting is addressed. Refinements in CART through Pruning. Advantages, weaknesses, and extensions of CART.

Advanced demonstrations of XLMiner, Economagic, and SAS Enterprise Miner 5.3. Several economic/financial applications with regard to Data Mining are introduced. Case studies. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 5: Supervised Data Mining: Predictive Modeling through Logistic Regression

Chapter 8

Inappropriate use of linear probability regression models (LPM) when the dependent variable is dichotomous. Model interpretation and model performance statistics of Logistic Regression models are evaluated.

Advanced demonstrations of XLMiner, Economagic, and SAS Enterprise Miner 5.3. Several economic/financial applications with regard to Data Mining are introduced. Case studies. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 6: Supervised Data Mining: Predictive Modeling through Neural Networks

Chapter 9

Introduction to Neural Networks; Input Layer. Hidden Layer, and Output Layer. Estimation of a basic Neural Network problem. Back Propagation. Data Preparation. Advantages and Weaknesses of the Neural Network Procedure; the 'Black Box' issue. Neural Networks in Time Series Data.

Advanced demonstrations of XLMiner, Economagic, and SAS Enterprise Miner 5.3. Several economic/financial applications with regard to Data Mining are introduced. Case studies. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 7: Unsupervised Data Mining: Cluster Analysis

Clustering detection: k-Means clustering, Wards clustering. Hierachical and Agglomerative Clustering. Dendrograms and Cluster validation.

Advanced demonstrations of XLMiner, Economagic, and SAS Enterprise Miner 5.3. Several economic/financial applications with regard to Data Mining are introduced. Case studies. Examples are drawn from bank fraud, financial forecasting, student retention rates, etc.

Topic 8: Current Issues in Data Mining

Topical issues are discussed and tailored toward specific interests of the student. The topics are selected in accordance with the research interests of the student.

Quality of Homework Assignments

You will receive a grade for each of the Homework assignments and what follows is a brief description of exactly how you are graded. In either case you must submit your assignments on time in order to receive credit.

	A or A-	B+ or B	B- or C+	C or C-	D+ / D / D-
Timeliness	Responses are always posted on time.	Responses are on time - one occasionally missed.	Responses usually on time or occasionally missed	Responses missed or late more than once.	Responses consistently posted late or missed.
Questions	Provides provocative questions on time that help us rethink readings.	Questions adequate but not provocative.	"Surface" questions that provide not critical analysis	Questions off base, inappropriate, or irrelevant for readings.	Doesn't post questions on time.
Thorough	Consistently addresses all parts of questions well.	Usually all parts addressed, some better than others.	Sometimes some parts are missing but what is posted is usually thorough.	Responses to questions are often quite incomplete.	Responses consistently address questions minimally.
Thoughtful Connections	Consistently responses make thoughtful & specific connections to readings. Links theory to practice and across posts, readings, or discussions.	Usually thoughtful connections to readings but often lacking connections made across posts, readings, or class discussions.	Sometimes responses do not demonstrate an under-standing of readings. Connections may be forced or incomplete.	Often responses are lacking in substance and, quite possibly, inaccurate. Connections made may be inaccurate or incoherent.	Responses are consistently done in a hasty fashion with no thoughtful connections made.
Substantive Responses to Classmates	Often reads others' posts and offers informed questions, comments, & connections.	Usually comments on others' posts. Interesting comments but they may not indicate clearly a close read.	Sometimes comments on others' posts. Comments are sometimes off base - not clearly relating to others' posts.	Infrequently comments on others' posts and comments are sometimes inappropriate - unhelpful, not constructive, rude.	Rarely comments on others' posts or, when does, comments are often inappropriate - unhelpful, not constructive, rude.

Other Important Information.

Students must accept the responsibility to be honest and to respect ethical standards in meeting their academic assignments and requirements. Integrity in the academic life requires that students demonstrate intellectual and academic achievement independent of all assistance except that authorized by the instructor. The use of an outside source in any paper, report or submission for academic credit without the appropriate

acknowledgment is *plagiarism*. It is unethical to present as one's own work, the ideas, words or representations of another without the proper indication of the source. *Therefore, it is the student's responsibility to give credit for any quotation, idea or data borrowed from an outside source.*

Students who fail to meet the responsibility for academic integrity subject themselves to sanctions ranging from a reduction in grade or failure in the assignment or course in which the offense occurred to suspension or dismissal from the University. Students penalized for failing to maintain academic integrity who wish to appeal such action may petition the department chair to request a hearing on the matter.

Pace University believes that it is important that students receive appropriate accommodation for any disability. If you have a disability for which you are or may be requesting an academic accommodation, you must register with the Coordinator of Services for Students with Disabilities. Trained professional Counselors will:

-Evaluate your medical documentation;
-Conduct appropriate tests or refer you for same;
-Make recommendations for your plan of accommodation; and
-Contact your professors (with your permission) to arrange for the recommended accommodations.

Your professor is not authorized to provide any accommodation prior to you arranging for same through the Counseling/Personal Development Center. If you have, or believe you have, a disability, be sure to follow the above procedure.