Data Mining in Finance: B20.3355 / B90.3355

Title

Data Mining in Finance: Intelligent Information Systems and Computer Intensive Methods for Financial Modeling and Data Analysis

Number

B20.3355 (Information Systems)
B90.3355 (Statistics and Operations Research)

Instructor

Andreas S. Weigend
Associate Professor
Department of Information Systems
Leonard N. Stern School of Business
New York University
aweigend@stern.nyu.edu
Tel: +1 (212) 998-0803
Fax: +1 (212) 995-4228

Credits

3 credits

Two lectures a week: Monday Wednesday 5:30 - 6:50 or 7:00pm - 8:20pm
TA and/or instructor present in computer lab after one class each week

Prerequisite

B90.3302 (Statistical Inference and Regression Analysis) or instructor consent

Description in Bulletin

This course develops links among Information Systems, Statistics, and Finance. It presents the foundations of data mining methods and their computer implementations. It applies them to current problems in finance, including the building and evaluating of trading models and the managing of risk. Assignments are weekly hands-on lab homeworks as well as an in-depth group project carried out in conjunction with a major Wall Street firm.

Home

www.stern.nyu.edu/~aweigend/Teaching/DataMiningFinance
Data Mining in Finance: Intelligent Information Systems and Computer Intensive Methods for Financial Modeling and Data Analysis

The term "data mining" refers to new methods for the intelligent analysis of large data sets. These methods have emerged from several historically disjoint fields, such as applied statistics, information systems, machine learning, data engineering, artificial intelligence, and knowledge discovery. One of the most enticing application areas of these emerging technologies is finance, becoming more amenable to data-driven modeling as large sets of financial data become available.

This course provides the relevant background knowledge, presents the foundations of data mining methods as well as their computer implementations, and applies them to current problems in finance, including the building and evaluating of trading models, and the managing of risk.

Besides the two lectures each week (that introduce the new approaches and discuss their advantages and disadvantages), there are two additional elements to the didactics of this course:

- Weekly hands-on homework assignments that help students gain familiarity with the corresponding tools, and
- A major project that deeply explores one or two of the approaches from class on a real-world problem.

Each project is carried out in conjunction with a major Wall Street firm. This is a unique opportunity for the student to apply one of the concepts introduced in the course to a genuine, current business situation. Besides the in-depth learning and understanding of the technical aspects, this provides insights into the challenges for data mining groups, as well as experience and connections for future jobs. These projects are carried out in teams, improving the students' communication and team skills.

This course, while housed in the Information Systems Department, is cross-functional, addressing three new audiences:

- The Dean sees the course as a key element in the effort to build a highly quantitative track for MBA students interested in financial engineering;
- The Chair of Statistics/Operations Research views it as the capstone course of the new MS in Statistics program with specialization in Financial Engineering;
- It also serves as a bridge to Courant as an elective for their MS in Mathematics in Finance.

This course is taught regularly in Spring semesters.
Introduction
These past few years have seen several dramatic changes at the intersection of the fields of finance, information systems, and statistics:

- The exponential increase in computational power has enabled intelligent information systems that can process and explore massive data sets.
- Not unrelated to these advances in information systems, the amount of data collected has exploded; Goldman Sachs, for example, keeps track of 1.1 million financial time series.
- New computationally intensive methods for finding information in such data, sometimes called "data mining" techniques, have emerged in a number of fields including neural networks, genetic algorithms, artificial intelligence, machine learning, and data engineering.

These emerging concepts from academic research are just becoming of interest to the industry; they are also currently reaching a stage of maturity that brings them into the reach of a graduate course.

This course presents these concepts in a coherent framework and discusses examples of their application in the financial markets. It consists of four parts:

1. Introduction: Statistics and Computation (3 weeks);
2. Building and Evaluating Trading Models (5 weeks);
3. Managing Risk (3 - 4 weeks);
4. Special topics, guest lectures by traders and strategists, and showcase presentation of the group project (2 - 3 weeks).

Background and Need
In recent years the application of data mining, applied statistics, machine learning, artificial intelligence, and knowledge discovery to technical trading and finance has seen exciting results, as witnessed by several conferences. In addition to ideas originating in academia, Wall Street firms such as J.P.Morgan have also contributed methods that are emerging as industry benchmarks for risk transparency and management.

The successful application of these new methods in the financial services industries requires analytical maturity, mathematical skills, and statistical sophistication. NYU with its proximity to Wall Street is the school to offer an education for modern computational finance! Indeed, a Financial Engineering Track started for MBA students

---

1 These conferences include CIFEr (Computational Intelligence in Financial Engineering, since 1995), and NNCM (Neural Networks in the Capital Markets, since 1993; NNCM changed its name to Computational Finance in 1997, see www.stern.nyu.edu/cf99).
at Stern last year--this course is a key elective for this highly quantitative track. And this year, the first group of about 70 students entered the MS in Statistics with Specialization in Financial Engineering program--this course is the capstone course for this unique program.

**Course Objectives**

- Introduce emerging computer intensive data analysis methods, intelligent information systems, data mining techniques, and state-of-the-art results from the fields of financial engineering and computational finance.
- Relate and integrate often disparately experienced pieces of knowledge that the students have been exposed to in earlier courses, and provide the relevant background from other fields such as machine learning, neural networks and genetic algorithms.
- Demonstrate the strengths and weaknesses of each of the new methods on financial data by discussing applications of these techniques to current problems in the financial domain, such as building and evaluating trading models, and managing risk.
- Provide practical experiences with these methods in the form of homework assignments, and demonstrate industrial strength software packages in class.
- Compare the promise of the new methods with current practice on Wall Street through lectures and in-class discussions with several traders and strategists from key financial institutions.
- Examine one or two of the techniques in great depth through a group project in conjunction with a major financial firm. The contacts are easily available through one of the group members (e.g., to a part-time MBA's work place), through connections of the instructor, or the Stern alumni network.

**Target Audience**

The target audience reflects the diversity and interest of students in this emerging field:
1. Stern Master of Science in Statistics with Specialization in Financial Engineering;
2. Stern MBAs, in particular those in the Financial Engineering Track;
3. Courant Masters of Science in Mathematics in Finance.

Full time students complete these programs within two years. The first and third programs require a total of twelve courses. The course described here will be taken by students in their last semester when they have a solid foundation not only in finance, but also in applied probability and statistics, including courses on time series and on multivariate regression. This course is viewed as the capstone course for the first program listed above, and as an important elective for the other two programs. It can also be taken by any Ph.D. student at Stern, and by other students with consent of instructor.
Pre-requisites

This course does not require any low-level programming. It assumes, however, a genuine interest in applied statistics, and a positive learning attitude to computer modeling. The formal prerequisites are coordinated with the MBA Financial Engineering Track and the MS in Statistics with Specialization in Financial Engineering and will be announced at a later stage.

Didactic Approach

Classroom Teaching and Hands-on Assignments

The teaching strategy is based on the fact that hands-on experience greatly facilitates the mastering of theoretical concepts and also helps the student retain significantly more than from lectures alone. Therefore, the course follows a lecture-lab-lecture sequence every week:

1. The instructor introduces the problem to be solved, explains the key aspects of the new method in sufficient detail, and briefly demonstrates the implementation in the classroom;
2. The students apply the method to a real world problem in the computer lab;
3. The lab experiences are reviewed and the general principles extracted, and further examples of successes and failures of this method on financial problems are presented.

The week-by-week schedule is given below.

On-line Material

Any material (course readings, assignments, software and data) that can legally be provided on-line will be made available to the students in the course through the website of the course.

Guest Speakers

The physical and intellectual proximity between Stern and Wall Street gives us access to world-class traders and strategists who are willing to share some of their insights on applying the methods discussed in the course. Towards the end of the semester, the course includes three or four presentations by practitioners from major financial firms, on trading, data mining, and information systems aspects of their work. To deepen these interactions, we provide further opportunities for informal contacts between the students and the guests (e.g., receptions), often leading to valuable continuations including internships and jobs for the students after obtaining their degrees.
Group Projects

Group projects are an important ingredient of the course: they deepen the understanding of the specific method chosen by each group, and they are true real world introductions to working with new data. Based on students' interests, teams of four to five students are formed at the end of the first week. Developing communication skills is central—it is important to learn to collaborate with people from different mind sets, and to experience that everybody benefits from bringing in their different competencies and diverse intellectual backgrounds.

Typically, the group projects form around a problem and a data set which is brought to the group by one of its members, often a part-time MBA, from his or her workplace. Group projects also emerge from consulting and interactions with affiliate firms to the Stern school. The final write-up of the project is expected in a form adequate for submission to a scientific journal or conference.

End-of-Semester Showcase

During the week of final exams, half a day is set aside for the presentation of the course projects. These presentations take the role of the final examination, and are observed by a board composed of external practitioners and faculty from several departments. These presentations give the students the opportunity to show off what they have learned and applied to a problem they were genuinely interested in. This event is also open to other students as a preview of what they could learn by taking this course, and to sponsoring companies as a glance of this exciting field of data mining and financial engineering at Stern. The project presentations will also be available on the Web.

Related Courses

This course complements several other courses that also include some hands-on experience:

1. **Financial Information Systems** (B20.3350, taught by Bruce Weber, Information Systems Department) that uses derivatives analysis software, and

2. **Statistical Computing with Application to Finance** (developed by Sean Chen, Statistics and Operations Research Department) that focuses on sampling methods for finance and includes a module on iterative sampling methods (Markov Chain Monte Carlos), and a module on Bayesian inference and decision making.

Furthermore, the course **Knowledge Systems in Organizations** (B20.3336, developed by Vasant Dhar, Information Systems Department) gives an overview of knowledge discovery and data mining techniques, focusing on their impact on business and issues regarding their deployment.
**Week-by-Week Schedule**

The following table lists the contents on a week-by-week basis:

<table>
<thead>
<tr>
<th>Week</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction: (i) Computer environment, (ii) Statistics background</td>
</tr>
<tr>
<td>2</td>
<td>Explorative data analysis (Visualization, Sonification, Wavelets)</td>
</tr>
<tr>
<td>3</td>
<td>Model evaluation and testing (Bootstrap, Simulation)</td>
</tr>
<tr>
<td>4</td>
<td>Time series prediction using nonlinear neural networks</td>
</tr>
<tr>
<td>5</td>
<td>Trading strategy development using neuro-fuzzy modeling</td>
</tr>
<tr>
<td>6</td>
<td>Trading rule discovery using genetic algorithms</td>
</tr>
<tr>
<td>7</td>
<td>Nonlinear regime switching models (Hidden Markov models)</td>
</tr>
<tr>
<td>8</td>
<td>GARCH and stochastic volatility (State space models)</td>
</tr>
<tr>
<td>9</td>
<td>Market risk (Value-at-Risk, RiskMetrics)</td>
</tr>
<tr>
<td>10</td>
<td>Credit risk (Nonlinear logistic regression, CreditMetrics)</td>
</tr>
<tr>
<td>11</td>
<td>Transaction risk (Bayes nets and graphical models)</td>
</tr>
<tr>
<td>12</td>
<td>Student projects (groups of five in collaboration with firms), and</td>
</tr>
<tr>
<td>13</td>
<td>Presentations by Wall Street traders and discussion, and</td>
</tr>
<tr>
<td>14</td>
<td>Special topics (pairs trading, execution, ICA, new products, ...)</td>
</tr>
</tbody>
</table>

A detailed description of the material for each week is given below.

**Software**

The two main software platforms are Matlab and S-Plus.

- I thank Sean Curry of The MathWorks, Inc. for his truly enthusiastic and most generous support with Matlab.
- I thank Doug Martin of the University of Washington (Seattle) and Mathsoft, Inc. for discussions, data, code, and also the generous support with S-Plus and S+GARCH.

Several other industrial strength programs will be available: I am grateful to Georg Zimmermann (Siemens Central Research, Munich) for the hands-on tutorial on neuro-fuzzy modeling, and to several groups at J.P.Morgan for providing me with programs and data for this course.
Detailed Week-by-Week Schedule

Part 1: Introduction-Understanding through Visualization and Simulation

Week 1: Tutorials-Managing Backgrounds and Expectations

The first class (of tutorial nature) takes place in the lab and goes through the mechanics of computer use, ranging from the specific setup in the computer labs and the access of the data sets for the assignments, to simple examples of the two main packages used in this course, S-Plus and Matlab.²

In the first week, the students fill out a questionnaire about their expectations, technical backgrounds, and interests. Answers will help the instructor identify problems early and optimally adjust the level of the course. For the group project, students are asked whether they have any specific problems in mind that they would be interested working on.

Furthermore, part-time MBA students, usually working in related fields, often have access to specific data sets from their workplace. The learning experience of the group project draws on these real questions and real data from some of their members.

Integrating these real world projects and presenting the final results both at Stern and in the firms is an important ingredient in the last part of this course.

The second class (also of tutorial nature) reviews some concepts from probability and statistics that the students are expected to be familiar with.³ This class also makes the expected background knowledge explicit and points to the pertinent literature that can help refresh some potentially rusty knowledge. At the end of the second class, groups of five students are formed. Care is taken to have in each group at least one student strong in theory, and one student strong in practical computer issues. Furthermore, students who already have a specific problem they want to model are distributed evenly across the groups.

The next two weeks continue with introductory character. Each week is dedicated to one of the two main software packages used in the course. The general approaches and specific commands that will be useful throughout the course are explained in conjunction with some interesting real world data.

Week 2: Exploring Data Through Visualization and Sonification

Visualizing data is a central skill for exploratory data analysis as well as for high-bandwidth communication of features of data sets. The second week uses a cross-sectional data set of equity returns, firm size, and book-to-market value.⁴ In their widely discussed 1992 paper, Fama and French argued that size and book-to-market play a dominant role in explaining cross-sectional differences in expected returns. Re-analyzing

---

² An argument can be made that the education should be streamlined into one single software package that the students have encountered in prior classes, such as Minitab. However, in a course covering such a wide terrain such as this one, it is important to draw on the advantages of different packages. Furthermore, familiarity with several modeling languages is an important asset.

³ These concepts include random variables, distributions, the correlation coefficient and its meaning, linear regression and the error on parameters, the maximum likelihood framework, conditional probabilities, conditional expectations, conditional variances, the Expectation Maximization algorithm.

⁴ The data include all NYSE, AMEX and NASDAQ non-financial firms in the monthly CRSP (Center for Research in Security Prices) tapes for the period July 1963 through December 1990.
their data with Trellis graphics (a method of conditional histograms) uncovers some striking features, including the complete disappearance of the risk premium on size. The background reading is an introduction to S-Plus, and the lab consists of reproducing the Trellis plots and further exploring the data that could lead to an understanding of the economic forces underlying the size effect. Besides gaining some initial familiarity with S-Plus, the students learn to appreciate the power and importance of exploratory data analysis and visualization. If time remains, there will also be a demonstration of various ways of revealing structure and information through rendering time series data in the auditory domain (sonification), as well as examples of multi-scale analysis with wavelets, two examples of modern exploratory data analysis.

Week 3: Understanding Results Through Simulation (Monte Carlo and Bootstrap)
The third week introduces bootstrapping, a computer intensive method to generate distribution of variables of interest, such as the profit of a certain trading strategy. The lab uses both computer generated examples and real world data to gain an appreciation of the inherent uncertainty of results in noisy environments. The synthetic data is generated to show just how easy it is to fool oneself about the profitability of a trading strategy.

The reading includes chapters on model evaluation and testing procedures of a forthcoming book by Blake LeBaron (Economics Department, University of Wisconsin, Madison). The students obtain a deeper understanding of the stochastic nature of financial markets and experience the central importance of absolutely clean methodologies for model testing to avoid data snooping, such as using true out-of-sample data in any performance evaluation. This assignment is carried out in Matlab.

Part 2: Building and Evaluating Prediction and Trading Models
Making predictions and building trading models are central goals for financial institutions. They were the earliest areas of the application of modern machine learning techniques to real world problems.

While the methods in the first part assumed that the model was already known, and that the sole task was to evaluate it, the course now turns to estimating the parameters of a model.

Consisting of five weeks, this second part of the course is the longest individual part. One important ingredient throughout this part is the methodology of turning predictions (e.g., the expected value of a future return and its variance) into actions (e.g., taking a position of a certain size). For students not familiar with statistical decision theory and the utility function framework, the concepts needed for this part of the course will be presented in an extra lecture of tutorial character at the beginning of this part.

Most of the hands-on components use either Matlab or S-Plus. When appropriate, comparisons to Excel are presented since most students are familiar with spreadsheets. Presenting conceptually different approaches to a problem clarifies the advantages and disadvantages of the different computational paradigms.

The five weeks of this part range from neural networks and neuro-fuzzy models, over trading rule discovery using genetic algorithms, to nonlinear time series models with hidden states. The details of week 4 through week 8 are given below.
Week 4: Nonlinear time series prediction

This week begins reviewing--from the perspective of data mining--the concepts the students had been exposed to in the prerequisite time series course, as well as presenting some implementations of linear time series models in S-Plus and Matlab. The central part of this week introduces "classical" neural networks for regression, corresponding to a nonlinear autoregressive model if the inputs consist of past values of the time series (tapped delay line). This framework is easily extended to incorporate inputs that are of various quantities which are derived from the time series itself (e.g., exponential moving averages, curvatures, volatility estimates), as well as additional exogenous time series (e.g., interest rates, indices, other markets, news).

In the lab, the students use both linear methods and nonlinear neural networks on the same data sets, in order to understand the strengths and problems of the more flexible neural network approach. One example shows that nonlinear methods can indeed outperform linear methods on financial forecasts, but also that the problem of overfitting can be very serious.

Week 5: Neuro-fuzzy trading model

The second week on neural networks extends the framework to incorporate rules and relationships elicited from some traders. The learning algorithm moves the parameters of these rules and adjusts their weights. In addition to this neuro-fuzzy framework, this week presents various methods to combat overfitting, including pruning (removing parameters that model the noise rather than the signal), cleaning (the combination of cleaning the data and learning the model), and adding adaptive noise to the inputs (to prevent the weights from learning more than there is).

During this week, as an exception, the lab is replaced by a one-day workshop where a practitioner shows the entire process of building a successful trading model, using SENN (Simulation Environment for Neural Networks). The data set of daily data compiled from multiple sources is provided by Georg Zimmermann (Siemens AG, Muenchen). The students are to gain an appreciation of the different phases and iterative nature of the modeling process; however, they are not expected to reproduce the model in detail. Students who cannot attend the full-day workshop (held on a Saturday) complete an assignment on predicting conditional variances and conditional percentiles in addition to the usual conditional mean (in Matlab).

Week 6: Trading rule discovery using genetic algorithms

This week introduces genetic algorithms as a search technique and shows how it can be used to solve problems such as volatility prediction, portfolio optimization, and stock picking. Often, in very complicated search spaces, the genetic algorithm can find interesting niches or relationships among variables. We show how genetic algorithms can be used to create rule-like models for prediction and classification. While these resulting models are less flexible than general neural networks, they are more amenable to interpretation. More importantly, genetic algorithms can produce many alternate models of the phenomenon which are similar in performance. The derived rules can be discussed with experts, and modified or discarded when they seem not to make any sense. The assignment uses futures data from an equity index to build models that forecast volatilities. The subsequent discussion shows how techniques influence the formulation of a problem and the representations chosen for the data and the model.
Week 7: Time series with discrete hidden states

The time series and trading problems so far have all been framed as regression problems. While this has made the computation relatively easy, time specific information is ignored: reshuffling all the input-output pairs of the training set would not have changed the resulting model. In contrast, this and the following week exploit time series specific information by introducing hidden states that are non-local in time.

This week's framework of hidden Markov models uses several discrete states with nonlinear neural networks as sub-models (also called experts or agents). We assume that we know the number of sub-models and their structure; the algorithm estimates the parameters of the sub-models, the probabilities of transition between them, and, for each time step, the probabilities to be in each state.

The assignment first analyzes a toy problem where the hidden state is known from the data generating process. However, it is not used in building the model, but afterwards, in order to verify that the model indeed discovered the hidden states. On the real world data, the students discover that the regimes form according to market volatility, rather than according to other possible dynamic criteria they might have expected (such as trending vs. mean reverting). The assignment is based on an implementation in Matlab by Shanming Shi (J.P.Morgan, New York).

Week 8: Time series with a continuous hidden state

State space models, the focus of this week, keep the assumption that the dynamics is hidden, but--in contrast to having several discrete hidden states with individual dynamics--they assume that the dynamics is captured by a continuous hidden state. State space models apply to time series processes with observational noise, i.e., noise that is not fed back into the systems but added during the observation.

The assignment uses the state space approach to model volatility, showing the striking advantage of this model class in the presence of observational noise over a standard autoregressive approach.

The state space model is compared to stochastic volatility models and to generalized autoregressive heteroskedastic models, as implemented as S+GARCH. The data consists of eight years of high-frequency foreign exchange data (in Theta-time), kindly provided by Michel Dacorogna (Olsen and Associates, Zuerich).

Part 3: Statistical Modeling in Risk Management

Risk is the degree of uncertainty of future returns. Risk management has become a central concern for financial institutions, since derivatives have become a major component of the markets. The approaches to risk management are becoming increasingly sophisticated in their computational and statistical methodology.

Risk stems from a variety of sources and exposures. The first of the three weeks addresses market risk, the uncertainty of future earnings resulting from changes in market conditions. The second week addresses credit risk, the possible inability of a counterparty to meet its obligations. Both of these weeks draw on methodology developed by J.P. Morgan, released in 1994 and 1997, respectively. The third week discusses transaction risk, using the example of identifying credit card fraud, one of the success stories of data mining techniques for financial problems.

During these three weeks, the students are also carrying out exploratory studies in preparation for their final group projects. This implies that the hands-on homework assignments require less individual work than those in the first half of the semester.
Week 9: Market risk
Market risk is the uncertainty of future earnings due to changes in market conditions, such as prices of the assets, exchange rates, interest rates, volatility, and market liquidity. The students are introduced to FourFifteen, a state-of-the-art tool, named after J.P. Morgan's internal market risk report, which was originally produced daily at 4:15 p.m. FourFifteen is a risk engine and a set of reporting tools. It is based on J.P. Morgan's RiskMetrics framework. Version 2 of FourFifteen is the result of J.P. Morgan teaming up with a software and interface company (The MathWorks, Inc., the company that produces Matlab), and with a data company (Reuters). FourFifteen is an excellent didactic tool, allowing the students to compare the original RiskMetrics variance-covariance approach with historical value-at-risk and Monte Carlo value-at-risk computations. They learn to separate the different contributions to market risk, and gain a deeper understanding of the implications of the choices that have to be made in modeling market risk.

Week 10: Credit risk
The booming global economy in recent years has created a business environment which enticed many institutions to take more credit risk. The proliferation of complex financial instruments has created uncertain and market-sensitive exposures that are more difficult to manage than traditional instruments. Active credit risk management, that includes the generation of consistent risk-based credit limits and rational risk-based capital allocations, is currently among the hardest challenges in the risk area. The course uses J.P. Morgan's CreditMetrics, a portfolio approach to credit risk analysis that treats all elements of the portfolio on an equal basis. This approach also considers the correlations of the changes in credit quality. It gives not only expected losses, but also their uncertainty (i.e., variance), expressed in the same value-at-risk framework as market risk.

In the assignment, students use CreditManager, J.P. Morgan's software implementation of CreditMetrics, to evaluate investment decisions, credit extension and risk-mitigating actions. They learn to identify concentrations of risk within a single portfolio, and to explore various scenarios by investigating marginal risk statistics. They also compare different risk measures (value-at-risk, standard deviation, percentiles), to gain an appreciation of the decisions that need to be made in modeling credit risk, and how these decisions affect the outcome. The data set includes historic probabilities of default and migration (upgrade and downgrade), correlations, recovery rates, and credit spreads.

Week 11: Transaction risk
With the widespread use of credit cards and rise of more complex transactions in electronic commerce, detecting fraudulent transactions automatically has become a major area for the use of data mining techniques. While credit risk is essentially a static problem (or made static by incorporating migration), transaction risk is fundamentally dynamic: whether a transaction is likely to be fraudulent depends on both the usage pattern in the past and recent history.

Bayes nets naturally incorporate prior knowledge in the structure by exploiting conditional independence and modeling as much as possible locally. They are a powerful method for discovering patterns and hidden causes in problems with categorical variables (e.g., merchant type). After introducing Bayes nets, we apply them to predict the probability that a transaction is fraudulent, and we show how to select one of the possible actions in this setting of an asymmetric payoff matrix.
In the lab, the students can choose between three hands-on approaches to the problem of detecting fraudulent transactions on credit cards; their choice tends to be influenced by the method they decided to use in their group project. The first approach uses a Bayes net to reinforce the concepts they had just encountered in class. The second approach uses a genetic algorithm to find relationships likely to indicate fraudulent behavior (see week 6). The third approach uses nonlinear logistic regression expressed as a neural network (see week 4). All approaches will be compared to ordinary logistic regression. This part concludes with the discussion of a business case featuring First Union (Charlotte, NC). This bank, with more than $200 billion in assets uses a neural network in their nonlinear transaction processing. This case study, written by Vasant Dhar at Stern, serves as the transition from risk management to the final part of the course that focuses on the role data mining methods are playing (and are going to be playing) in business and finance.

Part 4: Speakers from the Finance Industry and Projects

The physical and intellectual proximity between Stern and Wall Street gives us access to world-class traders who are willing to share some of their insights on applying the methods discussed in the course. The last three weeks include presentations by guest speakers to help balance the excitement about new academic techniques with an assessment of their usefulness in the real world. Sufficient time for less formal discussions will be included. One of the talks is complemented by a case study from a firm where the students are asked to critique the process and the results and suggest and justify alternative methods. Besides allowing students to obtain further information, these interactions allow the trader to get to know some of the students. Such informal interactions often lead to valuable contacts including internships and jobs for the students after obtaining their degrees. Most of these presentations are scheduled in the last three weeks of the semester.\(^5\)

One or two classes are reserved to present special topics, ranging from execution issues (that most students already encountered in the course Financial Information Systems) to modern techniques such as independent component analysis (ICA) or blind source separation (BSS), applied to financial markets and trading. In the last three weeks, besides attending the lectures and discussions, the students are expected to primarily focus on their group projects. (No further readings, prepared lab sessions, or individual homeworks are assigned any more.) These group projects are an important part of the course: they deepen the understanding of the specific method for each group, and they are true real-world introductions to some of the unforeseen and unforeseeable problems that are an integral part of working with new data.

The time line of the group project is the following. Based on students' interests, groups of five students were formed at the end of the second class. In the third week, one specific project from the past was discussed in detail, highlighting the ingredients that contributed to its success. It is important to give a good example of the scope of the project and to clarify its role in the learning process. By week 4, each group will submit two one-page proposals that includes a description of the data sets. Several iterations with feedback from the instructor and the teaching assistant, as well as learning about new techniques during the course, tends to change the proposal significantly. By week 8 (before spring break), each group is required to converge on one of their two proposals.

\(^5\) A viable alternative is to lighten the load in the second part by eliminating the topic of either week 5 or week 7. This makes space for one external presentation on trading models at the end of the second part.
The selected proposal is finalized in week 9 to a well-focused and sufficiently narrow contract between the group members and the instructor. It is important that the expectations and evaluation criteria are made clear. Weeks 10 and 11 explore the data and obtain simple benchmarks for the subsequent evaluation of the proposed method. In the remaining three weeks, the entire lab time is focused on obtaining results and writing up the project.

During the week of final exams, half a day is set aside for the presentations of the course projects. These case presentations give students the gratification to show off what they have learned and applied to a problem they were genuinely interested in. This event is also open to first-year students as a preview of what they could learn in their next year, and to the sponsoring companies as a glance of this exciting field of data mining and financial engineering at Stern.

Acknowledgments

I thank Sean Chen, Vasant Dhar, Steve Figlewski, Joel Hasbrouck, Ed Melnick, Sridhar Seshadri, Bruce Weber and Norm White for their valuable feedback and help with coordination with their courses. I thank Fei Chen and Balaji Padmanabhan for sharing the students' perspective, and last but certainly not least Caroline Kim for all her help with proofreading, logistics, and everything else. I also thank the participants of the Post-NNCM-96 Workshop on Teaching Computer Intensive Methods for Financial Modeling and Data Analysis (Weigend, 1997) for the open and fruitful discussion at Caltech in November 1996.

Finally, I genuinely welcome any comments, related experiences and suggestions.

Reference