NORTHWESTERN UNIVERSITY | SCHOOL OF CONTINUING STUDIES

PREDICT 410: Predictive Modeling I Syllabus — Fall 2012

Chad R. Bhatti, Ph.D.
chad.bhatti@northwestern.edu

Course Description

This course introduces statistical models as they are used in predictive analytics. The course reviews traditional linear and generalized linear models, including multiple regression and logistic regression. It addresses issues of model specification and model selection, as well as best practices in developing models for management. The course also demonstrates the application of multivariate methods in predictive analytics.

Texts

Allison, P. (2012). *Logistic Regression Using the SAS System* (2nd ed.). Cary, NC: SAS Publishing. [ISBN-13:9781599946412]

Chatterjee, S. & Hadi, A. S. (2012). *Regression Analysis By Example* (5th ed.). New York: Wiley [ISBN-13: 9780470905845]

Cody, R. (2011). *SAS Statistics By Example*. Cary, N.C.: SAS Publishing. [ISBN-13 9781607648000]

Delwiche, L., & Slaughter, S. (2008). *The Little SAS Book: A Primer*. (4th ed.). Cary, NC: SAS Publishing. [ISBN-13: 9781599947259]

Everitt, B.S. & Dunn, G. (2001). *Applied Multivariate Data Analysis*. (2nd ed.). New York: Wiley [ISBN-13: 9780470711170]

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression* (2nd ed.). New York: Wiley [ISBN-13: 9780471356325]

Ratner, B. (2012). *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data* (2nd ed.). New York: CRC Press [ISBN-13: 9781439860915]

All of these texts are considered 'Required' for this course. Material from each of these books is tied directly to the course reading and the course assignments. Any material covered in the course reading and the course assignments is also eligible to be on the final exam. The NU Bookstore will show Ratner (2012) and Delwiche and Slaughter (2008) as 'Recommended', where this is to be interpreted as 'Recommended for Purchase' since ebook versions of these two books will be available from the NU library. However, students should note that only three students can check an ebook out at one time.

I strongly recommend that you purchase all of these books. I consider each of these books required for this course. However, you are free to make your own textbook decisions, keeping in mind that those decisions can influence your course performance.

Software

The Northwestern University School of Continuing Studies provides JMP Pro software at no cost to students in the Predictive Analytics program. Students will be provided with instructions on how to download and install the JMP Pro software, how to set up a user account on the SAS OnDemand server, and how to register for PREDICT 410 on the SAS OnDemand server.

JMP Pro is a SAS Institute Inc. product that is available for both PC/Windows and Mac/OSX environments. We will not use JMP Pro directly as the software for statistical analyses in PREDICT 410. Instead, we will use JMP Pro as a software to allow us to run SAS on the remote SAS OnDemand server, i.e. it is through JMP Pro and the World Wide Web that we gain access to SAS® OnDemand for Academics running on a SAS Server (also referred to as the SAS Cloud).

SAS Institute, Inc. provides on-cost access to SAS® OnDemand for Academics for students registered in classes like this. Each PREDICT class has its own SAS location within SAS® OnDemand for Academics.

See Guide to Gaining Access to SAS for PREDICT 410 under Blackboard > Course Information for instructions on obtaining JMP Pro and setting up your computer to access the SAS Server/Cloud for this course.

Prerequisites

PREDICT 401

Learning Goals

The goals of this course are to:

- Develop statistically sound and robust multiple linear regression models, logistic regression modelsExplain advantages and major issues of modeling tools.
- Determine what type of model is the appropriate statistical tool in analyzing a given business problem.
- Build segmentation using multivariate analysis techniques.
- Interpret modeling results from both a statistical and a managerial viewpoint.
- Create appropriate modeling strategies to solve business problems.
- Apply best practices for implementing predictive analytics and modeling strategy in performance-based organizations.

Evaluation

The student's final grade will be determined from a total of 500 possible points as follows:

- Participation 20% (100 possible points, 10 points each session)
- Final Exam 20% (100 possible points, proctored exam)
- Assignments 60% (300 possible points from 8 homework assignments)

Grading Scale

A	= 93–100%	(465–500 points)
A-	= 90–92%	(450–464 points)
B+	= 87–89%	(435–449 points)
B	= 83–86%	(415–434 points)
B-	= 80–82%	(400–414 points)
C+	= 77–79%	(385–399 points)
C	= 73–76%	(365–384 points)
C-	= 70–72%	(350–364 points)
F	= 00–69%	(000–349 points)

Discussion Board Etiquette

The purpose of the discussion boards is to allow students to freely exchange ideas. It is imperative to remain respectful of all viewpoints and positions and, when necessary, agree to respectfully disagree. While active and frequent participation is encouraged, cluttering a discussion board with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in receiving less than full credit. Frequency is not unimportant, but content of the message is paramount. Please remember to cite all sources—when relevant—in order to avoid plagiarism.

Proctored Assessment

There is a proctored assessment requirement for this class. Please see the Assignments section in Blackboard for more information. The final exam must be proctored. Students are encouraged to use the ProctorU software.

Attendance

This course will not meet at a particular time each week. All course goals, session learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in threaded discussion boards is required, graded, and paramount to your success in this class. Please note that any scheduled synchronous or "live" meetings are considered supplemental and optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation.

Late Work

Students must provide written notification of late work 24 hours prior to the deadline. One grace day is allowed for those who provide late work notification. Only one grace day without reduction of points is allowed. A 25% reduction is applied to the grade for every 12 hours late. No negative points are applied.

Learning Groups

Student study groups will be utilized in this course as a means to foster a collaborative learning environment. The study groups are facilitated through the use of Adobe Connect, much in the same manner as our sync sessions are held. An Adobe Connect link will be provided by the instructor via the Blackboard course site.

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit <www.scs.northwestern.edu/student/issues/academic_integrity.cfm>.

Plagiarism is one form of academic dishonesty. Students can familiarize themselves with the definition and examples of plagiarism, by visiting <www.northwestern.edu/uacc/plagiar.html>. A myriad of other sources can be found online.

Some assignments in this course may be required to be submitted through SafeAssign, a plagiarism detection and education tool. You can find an explanation of the tool at <http://wiki.safeassign.com/display/SAFE/How+Does+SafeAssign+Work>. In brief, SafeAssign compares the submitted assignment to millions of documents in large databases. It then generates a report showing the extent to which text within a paper is similar to pre-existing sources. The user can see how or whether the flagged text is appropriately cited. SafeAssign also returns a percentage score, indicating the percentage of the submitted paper that is similar or identical to pre-existing sources. High scores are not necessarily bad, nor do they necessarily indicate plagiarism, since the score does not take into account how or whether material is cited. If a paper consisted of one long quote that was cited appropriately, it would score 100%. This would not be plagiarism, due to the appropriate citation. However, submitting one long quote would probably be a poor paper. Low scores are not necessarily good, nor do they necessarily indicate a lack of plagiarism. If a 50-page paper contained all original material, except for one short quote that was not cited, it might score around 1%. But, not citing a quotation is still plagiarism.

SafeAssign includes an option in which the student can submit a paper and see the resultant report before submitting a final copy to the instructor. This ideally will help students better understand and avoid plagiarism.

Other Processes and Policies

Please refer to your SCS student handbook at <www.scs.northwestern.edu/grad/information/handbook.cfm> for additional course and program processes and policies.

Course Schedule

Important Note: Changes may occur to the syllabus at the instructor's discretion. When changes are made, students will be notified via an announcement in Blackboard.

Note: All courses operate on a Monday to Sunday schedule.

Session 1 – Complete By Sun 10/7

Learning Objectives

After this session, the student will be able to:

- Understand the importance and role of exploratory data analysis in the model building process.
- Understand the difference between statistical inference and predictive modeling.
- Understand how to develop a plan for modeling.
- Perform an exploratory data analysis for a simple linear regression model.

Course Content

Textbook Reading

Ratner (2012) Chapters 1-2 pp. 1-30
Chatterjee and Hadi (2012) Chapters 1–2 pp. 1–56

SAS References

Cody (2011) Chapters 2-4 pp. 19-68
Cody (2011) Chapter 8 pp. 111-134
Delwiche and Slaughter Chapter 8 pp. 225-260

Online Reading

An Introduction to the SAS System

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 30 Points

Assignment #1: Exploratory Data Analysis for Regression is due Sunday, October 7, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Assignment #4: Problem Set for Ordinary Least Squares Regression is a problem set on ordinary least squares regression. Students should be conscience of this problem set and work on it as they progress through Weeks 1-3.

Sync Session

Saturday, October 6, 2012 at 11:00 a.m. – 1:00 p.m. (central time)

Session 2 – Complete By Sun 10/14

Learning Objectives

After this session, the student will be able to:

- Build a simple linear regression model using PROC REG.
- Understand how to interpret the SAS outputs for a simple linear regression model.
- Perform an analysis of goodness-of-fit to verify the assumptions for simple regression models.

Course Content

Textbook Reading

Chatterjee and Hadi (2012) Chapters 3–5 pp. 57–162

SAS References

Cody (2011) Chapter 8 pp. 111-134
Delwiche and Slaughter Chapter 8 pp. 225-260

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 30 Points

Assignment #2: Single Variable Regression Model is due Sunday, October 14, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Assignment #4: Problem Set for Ordinary Least Squares Regression is a problem set on ordinary least squares regression. Students should be conscience of this problem set and work on it as they progress through Weeks 1-3.

Sync Session

None.

Session 3 – Complete By Sun 10/21

Learning Objectives

After this session, the student will be able to:

- Understand how to perform an exploratory data analysis for a multiple linear regression model.
- Understand how to use automated variable selection techniques as part of the model building process.
- Understand how forward, backward, and stepwise variable selection techniques select optimal subsets of predictor variables.
- Understand the pros and cons of stepwise selection method.
- Develop a multiple linear regression model using SAS.
- Understand how to interpret the SAS output for a multiple linear regression model.
- Perform an analysis of goodness-of-fit for a multiple linear regression model.
- Understand how to compare two multiple linear regression models to decide which one is better.
- Understand how to identify mulitcollinearity and how it affects a fitted regression model.

Course Content

Textbook Reading

Chatterjee and Hadi (2012) Chapter 11 pp. 299–334

Ratner (2012) Chapter 10 pp. 177-194

SAS References

Cody (2011) Chapter 9 pp. 135-162

Online Reading

SAS Support Document for PROC REG

Handouts

Best Practice of Modeling Process in a Business Environment

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 40 Points

Assignment #3: Multiple Regression Model is due Sunday, October 21, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Assignment #4: Problem Set for Ordinary Least Squares Regression is a problem set on ordinary least squares regression. Students should be conscience of this problem set and work on it as they progress through Weeks 1-3.

Sync Session

None.

Session 4 – Complete By Sun 10/28

Learning Objectives

After this session, the student will be able to:

- Understand how to interpret a regression coefficient and its effect on prediction.
- Understand how to assess the importance of a predictor variable.

Course Content

Textbook Reading

Ratner (2012) Chapters 12-13 pp. 213-236

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 50 Points

Assignment #4: Problem Set for Ordinary Least Squares Regression is due Sunday, October 28, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 5 – Complete By Sun 11/4

Learning Objectives

After this session, the student will be able to:

- Understand when to use ordinary least squares regression and when to use logistic regression.
- Understand the statistical differences between ordinary least squares and logistic regression.
- Understand the differences between maximum likelihood estimation (MLE) and ordinary least squares estimation (OLS).
- Understand the likelihood ratio test, Wald test, and score test.
- Understand how to formulate and interpret a binary logistic regression model.
- Understand how to interpret the odds ratio, risk ratio, and risk difference.
- Understand how to assess the goodness-of-fit for a binary logistic regression model.
- Apply the modeling process to conduct a binary logistic regression analysis using SAS and interpret the model results.

Course Content

Textbook Reading

Hosmer and Lemeshow (2000) Chapters 1-3 pp. 1-90

SAS References

Cody (2011) Chapter 11 pp. 183-204
Allison (2012) Chapters 1-3 pp. 1-108
Delwiche and Slaughter Chapter 3 pp. 77-102

Online Reading

SAS Support Document for PROC LOGISTIC

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 30 Points

Assignment #5: Binary Response Exploratory Data Analysis and a Single Variable Logistic Regression Model is due Sunday, November 4, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Assignment #7: Problem Set for Logistic Regression is a problem set on logistic regression. Students should be conscience of this problem set and work on it as they progress through Weeks 5-7.

Sync Session

None.

Session 6 – Complete By Sun 11/11

Learning Objectives

After this session, the student will be able to:

- Understand how to formulate a multiple logistic regression model.
- Understand how to use dummy variables to code categorical predictor variables.
- Understand how to split a data set into training and testing data sets to use for model validation.
- Understand how to construct and interpret a lift chart (also called a cumulative gains chart) as a means to assess the predictive ability of a response model.

Course Content

Textbook Reading

Hosmer and Lemeshow (2000), Chapter 4-5 pp. 91–202

SAS References

Cody (2011) Chapter 11 pp. 183-204
Allison (2012) Chapters 1-3 pp. 1-108
Delwiche and Slaughter Chapter 3 pp. 77-102

Online Reading

SAS Support Document for PROC LOGISTIC

Handout

Best Practices of Modeling Process in a Business Environment

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 40 Points

Assignment #6: Multiple Logistic Regression Model is due Sunday, November 11, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Assignment #7: Problem Set for Logistic Regression is a problem set on logistic regression. Students should be conscience of this problem set and work on it as they progress through Weeks 5-7.

Sync Session

None

Session 7 – Complete By Sun 11/18

Learning Objectives

After this session, the student will be able to:

- Understand how to combine a logistic regression response model with an ordinary least squares regression model to driver customer response and profit maximization.
- Understand how to construct and interpret a lift chart to assess the predictive accuracy of a response model.

Course Content

Textbook Reading

Ratner (2012) Chapters 8-9 pp. 97–176

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 50 Points

Assignment #7: Problem Set for Logistic Regression is due Sunday, November 18, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 8 – Complete By Sun 11/25

Learning Objectives

After this session, the student will be able to:

- Use factor analysis and principal components analysis as a means of dimension reduction.
- Identify the differences between principal component analysis and common factor analysis models.
- Explain the concept of rotation of factors.
- Describe how to determine the number of factors to extract.
- State the major limitations of factor analysis techniques.
- Run factor analysis using SAS.
- Describe how to determine the number of principal components to extract.
- Run principal components analysis using SAS.

Course Content

Textbook Reading

Chatterjee and Hadi (2012) Chapter 9 pp. 233–258
Everitt and Dunn (2001) Chapter 3 pp. 48-73
Everitt and Dunn (2001) Chapters 12 pp. 271-290

Online Reading

SAS Support Document for PROC FACTOR

SAS Support Document for PROC PRINCOMP

Handout

Best Practice of Modeling Process in a Business Environment

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 0 Points

No Assignment.

Sync Session

None.

Session 9 – Complete By Sun 12/2

Learning Objectives

After this session, the student will be able to:

- Explain why segmentation is necessary in predictive modeling.
- Describe how similarity is measured in cluster analysis.
- Distinguish between the various distance measures.
- Distinguish the differences between hierarchical and non-hierarchical clustering techniques.
- State how to select the number of clusters to be formed.
- State the limitations of cluster analysis.

Course Content

Textbook Reading

Everitt and Dunn (2001) Chapter 6 pp. 125-160

Online Reading

SAS Support Document for PROC CLUSTER

Handout

Introduction to Segmentation Analysis

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 30 Points

Assignment #8: Multivariate Analysis is due Sunday, December 2, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.

Session 10 – Complete By Sun 12/9

Learning Objectives

After this session, the student will be able to:

- No new learning objectives will be introduced.

Course Content

None.

Discussion Board

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. For this session's discussion topic(s), visit the discussion board in Blackboard.

Assignments – 100 Points

Final Exam is due Sunday, December 9, 2012 at 11:55 p.m. (central time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None.