Predictive Analytics: MKTG 5220, Section 503 (DL students)

(Tentative Syllabus for Spring 2016)

Professor: Dr. Goutam Chakraborty

Office: 420A SSB, Phone: (405) 744-7644.

Class Materials: Most of the class materials will be distributed via the D2L web site for this class (https://oc.okstate.edu/). If you are a registered student for this class, you should be able to see this course when you log-in to D2L (the site becomes active one-week before classes begin). If you have problems accessing the D2L class site, please call OSU's IT help desk (405-744-4357) or (toll free) 1-877-951-4836). If you have problems viewing video lectures, please call Administrative/Video lecture support provided by the OSU's Distance Learning office (405, 744-4048, email: spearsdistance@okstate.edu).

Teaching Assistants (TAs): Names and emails of TAs will be announced in the first week of class (check lab videos or email). They will be your primary point of contacts for any issues related to this class. *When writing any email to my TAs, please copy to all TAs.*

E-mail: Please use the class discussion bulletin-board via D2L for any general questions, comments, clarifications about any of the class topic. Use the e-mail to my TAs sparingly. *There is no need to copy me with your email to my TAs – if my TAs are unable to answer your question,* they will discuss with me and get back to you. However, if you are not satisfied with TA answers, please do not hesitate to contact me.

Class Discussion via D2L (https://oc.okstate.edu/): We will use this format extensively for communication among students as well as between students and the instructor. This will be a bulletin-board type system with specific folders for different aspects of this course. There will be multiple forums (folders) in this bulletin board. Please check these folders regularly. Please post your questions only in the **appropriate forums**. Please use appropriate subject line in your posting and use threaded discussion whenever possible. Do not ask direct questions about how to solve an assignment (asking for clarification or software help is ok).

Required Text:

Text Mining and Analysis: Practical methods, Examples and case Studies using SAS® by Goutam Chakraborty, Murali Pagolu and Satish Garla, SAS Publishing, 2013, Cary, NC, SAS Institute Inc.

I will also use readings off the web, cases, SAS training materials, chapters from reference books, etc. in this class. I have indicated a number of good books (under reference texts) on this topic that you may find useful. I will announce readings via postings on D2L or via email.

Reference Texts (Optional materials – some of these will be put in reserve at the library)

- Data Mining Techniques for Marketing, Sales and Customer Relationship Management, by Michael J. Berry and Gordon S. Linoff, Wiley Publishing Inc., 2011. (OSU library call number: 658.802 B534d for 2004 version of this book)
- Data Mining, by Witten and Frank, Morgan Kaufman publications, 2011 (OSU library call number: 006.3 W829d 2011).

- Handbook of statistical analysis and data mining applications, by Nisbet, Elder and Miner, Academic Press/Elsevier, 2009, (OSU library call number : 006.312 N724h)
- Principles of Data Mining, by Hand, Mannila, and Smyth. MIT Press 2001. (OSU library call number: 006.3.H236p)
- Data Preparation for Data Mining by Dorian Pyle, Morgan Kauffman publications, 1999. (OSU library call number : 005.74 P996d)

COURSE OBJECTIVES

This course has five major objectives that fit within five of the program learning goals.

Course Objective	Program Learning Goal
Students will be able to engage in analytical reasoning to break problems into their component parts; identify important patterns by analyzing data; and test for assumptions behind models.	• Critical Thinking
Student can apply science and business principles to analyze and interpret data, using analytic and computer-based techniques.	• Critical and Creative Thinking
Students will be able to present written results from their analyses by relating those back to the business issues that demonstrate a mastery of language and mechanics.	• Written Communication
Students will be able to present their results orally using a message that is well organized, concise and quickly understandable by business professionals.	• Oral Communication
Students will be able to use appropriate tools and technologies for data visualization and statistical model building	• Technology Skills

Course Description

According to Wikipedia, “predictive analytics” encompasses a variety of techniques from statistics, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events. The underlying premise of predictive analytics is to turn business data (both numeric and text) into actionable information via building of models. Therefore, this course will focus on learning how to use various data mining (machine learning) and statistical tools such as neural networks, decision trees, and classification and prediction models including logistic regression etc. in the context of common applications in business. Students will be expected to a state-of-the-art enterprise level advanced analytics software (SAS Enterprise Miner) to analyze real-world data and make strategic recommendations for managerial actions. My philosophy in teaching the course is “*you learn by doing,*” that is, you should be prepared to work extensively with SAS Enterprise Miner in analyzing data sets using various techniques such as **neural networks, decision trees, multiple/logistic regression, association rules, sequence detection, ensemble models, text mining, sentiment mining, content categorization,** etc. The course will use lectures, data analysis using state-of-the-art data mining software, discussions, and exercises. All class lectures will be handled via video (video links will be posted on the D2L course site) that you can watch at our own convenience (you will need a high-speed Internet connection to watch the lectures).

Real Office Hours (to talk to me in person)

Monday and Tuesday 8:00-9:00AM, or by appointment (set up via phone or email).

Virtual Office Hours (to get my opinion on any issue related to this class)

Please use the desire to learn (D2L) platform for this purpose. I (and my TAs) will monitor this platform closely and try to answer your questions quickly. I may also set up a SKYPE or GoTo based call-in office hours for DL students so you can talk to me (individually or as a group). Details will be communicated via D2L class site.

Course Prerequisites

Students must complete MKTG5983 (Database Marketing) before taking this class. I will assume that all students enrolled in this class have very good ideas of basic probability/statistics, basic statistical models (such as multiple regressions, ANOVAs) and perhaps some exposure to SAS software before joining this class.

Course Format:

Please note that this course has a unique format (a combination of all video lectures and lab sessions for discussion/activities). Also, the class requirements are **very** different for non-distance learning (non-DL) and distance learning (DL) students.

Lectures: The class sessions (Mondays, 2:30 PM – 4:20 PM) will be used as labs/discussions (see below). A video for each lecture will be streamed over the Internet. The link for each video lecture will be posted on the D2L. It is your responsibility to **watch the lecture video and do appropriate readings/work before coming to the class sessions for lab/discussions**. *All students (non-distance learning or distance learning) will have access to lecture videos*.

Labs:

- *Distance Learning (section 503) students*: ***You do not have to attend labs physically***. You will however be given access to lab videos where non-distance learning students will participate in various activities such as discussion of readings, questions and answers, etc. I expect you to watch these lab videos as they become available.

Finally, as an instructor I retain the right to modify this tentative syllabus based on how the class progresses. If I make changes, I will let you know via D2L and/or email.

Class Requirements for Section 503 (DL students) only

Exams: One midterm (**20% of course grade**) and one comprehensive final exam (**40% of the course grade)**. *The midterm exam is shown on the schedule. The final exam will be due by the Wednesday of the Finals week.*

For some students who may want to work on a comprehensive analytics project on their own using data from their companies (or, from publicly available sources), I am willing to let you do so as **an alternative to taking the comprehensive final exam**. If you pursue this option, an added benefit may be that you could publish a white paper and/or write a joint paper with me for next year's SAS analytics or SAS Global Forum conference. If you choose to work on such a project, then understand the rules as stated at the end of this document. If you are interested in this option, then submit a 2-3 page proposal of your project by February 1, 2016 via the appropriate drop box in D2L. *Your proposal must contain enough details (see end of this document) about the project for me to judge its suitability as an alternative to final exam*. If I accept your proposal, then you will have other interim deliverables (in mid-semester) and your final project report will be due on the Wednesday of the finals week.

Individual Exercises: There will be eleven individual exercises in this class. I will drop your worst exercise. **You must do these individual exercises alone and not seek help from others**. These exercises will primarily reinforce the concepts covered in the lectures. These exercises will count for **20% of the course grade**.

Major Assignment: There will be one major assignment (**worth 20% of the course grade**). The purpose of this assignment is to give you a flavor of all steps involved in doing an analytics project. It will encompass analyzing a comprehensive business case with data.

Special Note: Although as DL students, you are not required to attend labs on campus, I *strongly suggest that you watch the lab videos* as soon as they become available because in those videos I will do demonstrations, discuss questions related to lecture topics, exercise solutions, etc. These lab videos will enhance your learning and also help you in doing exercises/assignments/cases/exam.

Semester Grades: The final grade for this section will be based as follows: 90% or above will result in A, 80% or more will result in B, 70% or above will result in C, 60% or above will result in D. Those getting less than 60% will get an F. I will look at the distribution of the total scores within this section and use any appropriate normalization as needed.

Late Assignments: Assignments or cases must be turned in by the class time on the due date via D2L drop box (not emails). All late assignments (***even 1-minute late***) must be turned in **via the Late Drop Box** and will be *penalized* as follows:

- One late assignment (within 1-hour of due date and time) – *no penalty*
- All other late assignments will carry following penalty structure:
    - Within 1 hour of due date and time – 15% penalty
    - More than 1 hour but less than 24 hours of due date and time – 30% penalty
    - More than 24 hours but less than 48 hours of due date and time – 50% penalty
    - More than 48 of due date and time – will not be graded (no credit)

I enforce this rule because I believe that part of effective functioning in business is the ability to complete projects on time. **Please do not email/call/contact me or my TAs with excuses (however valid they may be) about making exceptions to my late submission policy.**

Note: For all other issues such as add/drop policy, academic integrity etc., I will follow OSU guidelines as posted in the website (http://academicaffairs.okstate.edu/content/resources-faculty-staff)

Project (Alternative to Final Exam) Proposal Details for Section 503 Students Only

Any student who wants to do a project as an alternative to the final exam, must submit a 2-3 page proposal via the appropriate D2L drop box by 11:59 PM US CST on **Feb. 1, 2016**. It will be your responsibility to secure project data and arrange for any necessary permission from your company to share the project report with me. You may use numeric, textual or a combination of both types of data. The primary goal for your project should be developing a predictive model to predict your target variable or, to develop a segmentation model.

The purpose of the proposal document is to help me (1) understand the nature and the scope of your project, (2) judge if it is doable within the semester and (3) if it is suitable as an alternative to the final exam, worth 35% of the course grade. I will read your proposal and let you know my acceptance or rejection within a week. If I accept you proposal, I will let you know about interim (mid-semester) and final (Wednesday of the Finals week) deliverables.

When you write your proposal, include following:

1. A 2-3 lines description of the main idea of your project
2. Describe the business opportunity or problem that your project addresses
3. Describe why anyone (the likely users/clients of your project) should care about this project. Think about what's the best thing that could happen if everything goes according to your plan and how that might benefit the company.
4. What data will you be using for the project?
    a. I need a metadata for your data that describes variable names, variable types, data types, variable values, variable description, etc.
    b. I need a clear statement of number of observations (records) and number of variables in your data. Most predictive analytics algorithm require a large amount of data and unless I am convinced that you have access to such data, I will not accept your proposal.
5. How you will be getting this data?
    a. If you have to get it from your company, do you currently have access to it?
        i. Has the company agreed to let you use the data for your project?
        ii. If you do not have access now, how would you get it?
    b. If you are using data from publicly available sources, then cite the source and answer following questions.
        i. Have you downloaded the data? If not, then when?
6. A discussion of following items:
    a. Your expectation of how much data editing/cleaning will be needed?
    b. What type of analysis you might be doing – exploratory, predictive, segmentation, others
    c. What types of recommendations do you expect to make based on this project

In general, the more details you can provide in your proposal, the more likely it is that I will accept your proposal.

Tentative Schedule of Topics (Spring 2016)

General: The schedule *below is tentative and subject to change* based on the pace of discussion of topics in class. All changes in the schedule will be communicated to you in labs and/or via email or D2L class site. Some of the class readings are mentioned in this schedule. Other readings will be assigned and posted on D2L or announced in the lab. Exercises, assignments, projects, etc. will be assigned and posted on the D2L class site or via email. It is your responsibility to check D2L site every week for changes/announcements with respect to schedule/exercises/assignments etc.

Note about Exercise and Labs: Expect something to do (exercise, assignments, cases, etc.) in each week. In general, exercise/assignment corresponding to lecture topics covered in any week is due on **Monday of the next week by 1159 PM US CST** via appropriate drop box unless mentioned otherwise in the schedule or announced on D2L or communicated during the lab session. So, for example, Individual Exercise 1 mentioned in week 2 is actually due on Monday of week 3. Therefore, I will assume that before doing Individual Exercise 1, you must have reviewed lectures up to week 2 and reviewed lab videos of week 1 and 2.

Week 1 (Week begins Jan. 11):

Video lecture 1: Course and faculty introduction, overview of data mining, a recap of basic statistical concepts, and overview of software access and interface.

Readings: What is data mining? (http://www.twocrows.com/intro-dm.pdf)

***Lab (Jan 11)*:** Discussion of assigned readings and other activities as communicated via D2L or email. Software demonstration by instructor in the lab.

Week 2 (Week begins Jan. 18):

Video lecture 2: Overview of analytical methodology, Overview of data management and integration, Data preparation for data Mining, Overview of initial challenges in data mining, Overview of honest assessment in predictive modeling, Overview and demonstration of RFM Analysis.

Readings: Quick profits with RFM Analysis by Arthur Hughes (http://www.dbmarketing.com/articles/Art149.htm)

***Lab (Jan. 18)*: No lab this week due to MLK holiday.**

Exercise

Complete Individual Exercise 1 and upload your solution by Jan. 25, 1159 PM CST.

Week 3 (Week begins Jan. 25):

Video lecture 3: Exploring and preparing data with summary statistics, plots. Handling missing values, handling transformations and handling extreme values (outliers) in the data.

Readings:

- http://www.togaware.com/datamining/survivor/Exploring_Data.html
- http://en.wikipedia.org/wiki/Exploratory_data_analysis
- http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

***Lab (Jan. 25)*:** Discussion of assigned readings and other activities as communicated via D2L or email.

Exercise

Complete Individual Exercise 2 and upload your solution by Feb. 1, 1159 PM CST.

Week 4 (Week begins Feb. 1):

Video lecture 4: Introduction to predictive modeling via decision trees. Predictive modeling essentials, Understanding how split search works in decision trees, building and pruning decision trees, understanding tree variations with different splitting rules.

Readings:

- http://www.wuss.org/proceedings10/analy/3055_2_ANL-Hobbs.pdf
- http://www.statsoft.com/textbook/classification-trees/?button=1

Lab (Feb. 1): Discussion of assigned readings and other activities as communicated via D2L or email.

Exercise

Complete Individual Exercise 3 and upload your solution by Feb. 8, 1159 PM CST.

Week 5 (Week begins Feb. 8):

Video lecture 5: Introduction to predictive modeling via regression models. Overview of Logistic regression, handling sequential variable selection in regression, optimizing complexity in regression, using nonmetric inputs in regression, accounting for nonlinearities in regression.

Readings:

- Basics of logistic regression: http://faculty.cas.usf.edu/mbrannick/regression/Logistic.html
- Odds ratio : http://en.wikipedia.org/wiki/Odds_ratio
- Regression vs. Decision Trees: http://www.forecastingprinciples.com/paperpdf/exploratory.pdf

Lab (Feb. 8): Discussion of assigned readings and other activities as communicated via D2L or email.

Exercise

Complete Individual Exercise 4 and upload your solution by Feb. 15, 1159 PM CST.

Week 6 (Week begins Feb. 15):

Video lecture 6: Introduction to predictive modeling via Neural Networks. Background of ANN, Understanding how ANN works, Optimizing ANN models, other modeling tools in SAS EM, different types of ANN (MLP vs. RBF) models.

Readings:

- http://www.statsoft.com/textbook/neural-networks/?button=2
- http://en.wikipedia.org/wiki/Artificial_neural_network
- http://ulcar.uml.edu/~iag/CS/Intro-to-ANN.html

Lab (Feb. 15): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise
Complete Individual Exercise 5 and upload your solution by Feb. 22, 1159 PM CST.

Week 7 (Week begins Feb. 22):

Video lecture 7: Model assessment, model implementation and special topics. Assessment and comparison of models via summary metrics, graphs and charts. Adjustments for oversampling, creating and using profit matrices for model selection and scoring of new data. Use of decision tree to consolidate large number of categories into a smaller number and to understand predictions from a Neural net model.

Readings:

- http://homepage.cs.uri.edu/faculty/hamel/pubs/hamel-roc.pdf
- http://gim.unmc.edu/dxtests/ROC1.htm
- http://www.scholarpedia.org/article/Ensemble_learning

Lab (Feb. 22): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise (Special Deadline):
Complete Individual Exercise 6 and upload your solution by Friday, Feb. 26, 1159 PM CST.

Week 8 (Week begins Feb. 29):

All students in sections **503** must make arrangements with the DL office (405, 744-4048, email: spearsdistance@okstate.edu) to complete mid-term exam between Monday Feb. 29th and Wednesday Mar. 2nd of this week. The details of the midterm exam will be communicated to you.

Week 9 (Week begins Mar. 7):

Video lecture 8: Pattern discovery techniques (market segmentation via clustering)

Readings:

- Chapter 6, pp. 111-113 from your text
- http://en.wikipedia.org/wiki/Market_segmentation
- http://en.wikipedia.org/wiki/Cluster_analysis
- Eliminating Response Style Segments in Survey Data via Double Standardization before Clustering – read online)

Lab (Mar. 7): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise (Special Deadline)
Complete Individual Exercise 7 and upload your solution by Mar. 21, 1159 PM CST.

Week 10 (Mar. 14 – Mar. 18): SPRING BREAK WEEK (NO LECTURE or LAB)

Week 11 (Week begins Mar. 21):

Video lecture 9: More on clustering, segmentation and profiling.

Readings:

- Chapter 6, pp. 111-113 from your text
- http://support.sas.com/resources/papers/proceedings13/068-2013.pdf
- http://support.sas.com/resources/papers/proceedings12/200-2012.pdf
- Product Affinity Segmentation That Uses the Doughnut Clustering Approach - read online)

Lab (Mar. 21): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise
Complete Individual Exercise 8 and upload your solution by Mar. 28, 1159 PM CST.

Week 12 (Week begins Mar. 28):

Video lecture 10: Introduction to Text Analytics. How to import Text in SAS EM. Introduction to SAS IR studio. Understanding Text Parsing.

Readings:

- Chapters 1-4 from your text

Lab (Mar. 28): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise
Complete Individual Exercise 9 and upload your solution by Apr. 4, 1159 PM CST.

Week 13 (Week begins Apr. 4):

Video lecture 11: Role of SVD and LSI in Text Clustering. Text topic extraction, role of predictive models in text analytics

Readings:

- Chapters 4-6 from your text

Lab (Apr. 4): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise
Complete Individual Exercise 10 and upload your solution by Apr. 11, 1159 PM CST.

Week 14 (Week begins Apr. 11):

Video lecture 12: Sentiment Analysis and Opinion Mining

Readings:

- Chapter 8 from your text

Lab (Apr. 11): Discussion of assigned readings and other activities as communicated via D2L or email.
Exercise
Complete Individual Exercise 11 and upload your solution by Apr. 18, 1159 PM CST.

Week 15 (Week begins Apr. 18):

Video lecture: None – work on major assignment
Lab (Apr. 18): None - work on major assignment
Exercise: None – work on major assignment

Week 16 (Week begins Apr. 25):

Video lecture: None
Lab (Apr. 25): Finish major assignment.
Group Assignment:
Complete major assignment and upload your solution by Apr. 25, 1159 PM CST.

Week 17 (Week begins May 2):

Final Exam Week:

- All students in sections **503** must make arrangements with the DL office (405, 744-4048, email: spearsdistance@okstate.edu) to complete final exam by Wednesday of the Finals week.
- If I have approved your analytics project as an alternative to final exam, your project report is also due by the Wednesday of the finals week.