22-BANA 7046 Data Mining I (2 cr.)
22-BANA 7047 Data Mining II (2 cr.)
Section 002
Spring Semester, 2017-2018

Instructor: Professor Yan YU

Office:	527 Lindner Hall
Office Hours:	TH, 2:15pm-3:15pm and by appointment
Class Time:	TH, 3:30pm-5:20pm, Lindner 221
Email:	Yan.Yu@uc.edu (*preferred*)
Phone:	556-7147
Web Page:	http://www.blackboard.uc.edu

Course materials including syllabus, lecture notes, reading assignments, case, homework, data sets, R and SAS programs, and course handouts will be posted on the course web in blackboard.

Course Objectives: To provide students with a hands-on data analysis experience using various statistical methods and major statistical software (R and SAS) to analyze complex real world data in business and industry.

Course Format: Classes will be provided in three forms: lecture, case study, and project discussion/presentation. In case study, students will be led through practical problems addressed by data analysis techniques. The aim is to provide a detailed view on how to manage complex real world data; how to convert real problems into models so that statistical software can be used appropriately; and how to interpret and diagnose the model fitting. Project discussion and presentation will enable students to share and compare ideas with one another and to receive specific guidance from the instructor.

Recommended Texts:

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2014). *An Introduction to Statistical Learning with Applications in R*. Springer New York, ISBN-13: 978-1-4614-7138-7. (A free eBook is downloadable at http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf)

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.
(A free eBook is downloadable at https://web.stanford.edu/~hastie/ElemStatLearn/)

This is an advanced book to read. The course will cover the less theoretical material in the text. Lectures should be easier to follow than the book. Supplemental readings and notes will be posted.

Zhao, Yanchang (2012), *R and Data Mining: Examples and Case Studies*.
(A free eBook is downloadable at http://www.rdatamining.com/**)**

This book gives hands on examples and case studies for Data Mining using R.

There is little explanation on the contents, which will be mainly addressed through the lecture notes.

Prerequisites: BANA7041 or BANA7038 or equivalent (linear regression at graduate level); Basic Statistical Computing skills.

Data Mining I

Grading:	Exam (In-Class, Close Book)	60%
	Four Group Homework	40%

Data Mining II

Grading:	Group Project and Presentation	30%
	Two Individual Cases	20%
	Quiz	50%

Grade Scale:
Letter grade is assigned with the following scale:
A-range: 90 – 100%; B-range: 76 – 89%; C-range: 64 – 75%
'Plus' and 'Minus' grades will be used. Final letter grade is assigned based on an overall assessment of your grade, intragroup evaluation, and class performance at the instructor's discretion.

Exam, Individual Cases and Honor Policy

Exam is **in-class, close-book and close-notes**. You can bring one page (standard 81/2 x 11 paper) **hand-written** cheat sheet, which will be turned in along the exam. Exam should be the sole work of each student. Anyone cheating or assisting another during an exam will be given a 0 for that exam and possibly a grade of F for the class. College procedures will be followed and the graduate dean will be notified.

Computational quiz involves in-class coding. It is open-book, open-notes, and Google search is allowed. However, any form of peer-to-peer communication is strictly prohibited.

Individual Cases should be treated as a take-home exam, which should be the sole work of each student. You can only discuss with the instructor if you have any question.

Academic Integrity: As with all Lindner College of Business efforts, this course will uphold the highest ethical standards, critical to building character (the C in PACE). Ensuing your integrity is vital and your responsibility. LCB instructors are required to report ANY incident of academic misconduct (e.g., cheating, plagiarism) to the college review process, which could result in severe consequences, including potential dismissal from the college. For further information on Academic Misconduct or related university policies and procedures, please see the UC Code of Conduct (http://www.uc.edu/conduct/Code_of_Conduct.html).

Ethics in Science: The project should be the sole work of the students on the team. None of the work on the project should have been used as part of any other course, independent study project, master's project, or other type of project for academic credit. *Exception*: Slight overlap between this project and another project for academic credit might be permissible if discussed **in advance** with the instructor. Please notify the instructor if any of the course projects is further carried out

to be your MS project. Typically this is possible if you conduct your course project by yourself. Ethics in Science should be strictly followed. See detail through *http://www.files.chem.vt.edu/chem-ed/ethics/index.html* **Absolutely NO copy/paste from others' work. Please quote, refer and give credit to any work that are highly related or inspire your future project. A literature review should be conducted before any claim of "novelty" is made. Action will be taken if any violation is found. Program director will be notified for further action.**

Use of Electronic Devices: Cell phones should be switched off during classes. Laptops should be used only for course related practice. Any form of entertainment (online chatting, watching videos and the like) is prohibited. Violations will result in dismissal from the class.

Class Communication:
We will use Blackboard to communicate. The student is responsible for all communications sent by the instructor using email via Blackboard. Therefore, students must check to see if their accounts have reached maximum capacity or are otherwise not functioning, and to correct this situation. I receive a large quantity of email messages, many of which appear, based on the subject line, to be junk mail or spam. I delete these messages without reading them. To make sure that your message is not accidentally deleted as junk, please include '**BANA7046/7047-002**' in the email subject line. Also, be sure to identify yourself in the message. Otherwise, you may not receive a response.

Submission Information:
Group homework, project proposal, final project, and related presentation power point slides should all be submitted via Blackboard. You shall submit only ONE copy per group. Please include a cover page with your group number and group members.

Due Dates (One Copy per Group):
Homework (4): Due at **10am** on the day listed in the schedule (**1/23, 1/30, 2/6, 2/13**).
Please email your report or power point for HW presentation/discussion at **10am**.
Homework Intragroup Evaluation: **2/13**, **11pm.**
See **INTRAGROUP_EVALUATION_BANA7046.pdf** for details on HW intragroup evaluation.

Project Proposal: **04/03**, **12pm**, Maximum of five typed, double spaced pages with one inch margins and 12-point font. Submit a .docx file with name "**7047-002-Proposal-Group#.docx**" through blackboard.
Find a topic and necessary data; summarize what you plan to do in the project.
Project Proposal Presentation: **04/03, 12pm.** Submit power point slides with name "**7047-002-Proposal-Group#.pptx**" through blackboard.

Project Final Presentation: **04/19, 12pm.** Submit power point slides with name "**7047-002-Project-Group#.pptx**" through blackboard.
Project Final Report and Intragroup Evaluation: **04/22**, **11pm**. Maximum of twenty typed, double spaced pages with one inch margins and 12 point font. Submit a .zip file through blackboard with name "**7047-002-Project-Group#.zip**" with the following files: 1) Final report

in WORD (project-group#.doc; 2) presentation (project-group#.ppt); 3) R/SAS code. (I'll ask for data file if needed.) You do not need to submit a separate hard copy.
See **INTRAGROUP_EVALUATION_BANA7047.pdf** for details on project intragroup evaluation.

Computing Resources:
We will use R primarily. You can download R for free through the URL http://www.r-project.org/. You may use SAS if you insist. SAS license can be purchased from the campus bookstore. You can also access SAS and R in the second floor computer lab (215 and 202). SAS help files are available online.

Group Work Structure of the Course: After the first class, each student will join a work group. A work group will typically consist of **four** students. This work group will be maintained for the length of the semester. Except the individual cases, the work group will cooperate in all work given during the semester including group homework and group project. All members of a group will share grades on any submitted work. All members are to contribute equitably to the shared workload, carrying a fair weight for the burden. Periodically, members of each group will be asked to evaluate the contribution of the other work group peers on the basis of a number of criteria such as intellectual contribution, attendance at group meetings, mentoring and sharing knowledge, writing up the results, presentation, and running relevant SAS and R codes. The peer score will reflect, in some sense, an average over all the assigned work as well as an average of the above criteria. Thus, a student in a work group who may have contributed much on one assignment, may not have contributed the majority of the work on another, yet still such work may be considered by other members to be meritorious "on the average".

Group Project for Data Mining II:

Purpose of project: One goal of the project is to provide you with more experience using data mining tools on practical problems. A second goal is to help you become a self-directed learner; this is the type of learning that you will be doing in the future. It would be most interesting if you have some new methodology learned from this course for some interesting business problems. For some MS-BANA students, this project could serve as serious starting for your MS project.

Project Teams: You should work in a team of **four** students. You should form a team yourself. If you cannot find teammates, let me know and I will help you find a team. If you insist, you may work alone.

Types of projects: Almost any type of project is acceptable. However, I expect that most projects will be either one of or a combination of the following types:

- Applying tools that you have already learned in this course to a data set not used before.

- A study of procedures and software for data mining that are not used in the class with some comparison with what you have learned in the class. For example, topics might include

- A Study of Data Mining with SAS Enterprise Miner
- A Study of Data Management with SQL using R

You will use your own words to describe what you have learned and illustrate with real data examples.

- A report describing one of the data mining tools that are discussed in the textbook but either are not covered in the lectures or will be covered only briefly at the end of the course. Data mining tools that could be studied in your report would be any *except* linear and logistic regression, GAM, decision trees, neural networks, clustering, and association rules. For example, data mining tools that *could* be studied include

  - MARS
  - Random forest
  - Bagging
  - Boosting
  - Support vector machines

  For a project of this type, in addition to the textbook you should use several other books or articles as source materials. It should involve some programming of the data mining tool to study and testing it on a large data set;

- A Monte Carlo simulation of a data mining tool, where large data sets of known structure are simulated and data mining tools are tested to see how well they can detect the known structure.

What is required? Each team must write a project proposal, find the necessary data, carry out the project, and write a project report. You are asked to implement your work both in R and SAS Enterprise Miner.

The report should be at most 20 double spaced pages with one inch margins and 12 point font and should contain:

- Title page with authors and abstract
- Introduction telling what the project is about, what your team has accomplished, and a brief statement of results and conclusions.
- One or more sections describing the project
- Conclusions
- Bibliography

Tables and figures can be interspersed in the text or at the end of the report. All tables and figures should be numbered and referred to by number. The report should not contain raw computer output. Rather, any computer output should be in a table or figure. If necessary, append brief SAS and R codes in the appendix section.

Project Grading: The group project is worth 30% of the course grade. Grades will be based on:

- Interesting application, Creativity.
- How much new materials you have learned.
- Clarity and conciseness of the report.
- Correctness.
- Powerpoint Presentation.
- Intragroup Evaluation.

Project Proposal: Ideally a proposal can serve part of the first **5** pages of your final project report, depending on how much preliminary analysis you have conducted. Proposal should usually include at least 1) Background information of the project. The objective you want to achieve. 2) A detailed data section: discussion of data source and nature of the variables involved in the analysis. 3) Preliminary analysis. That is, some exploratory analysis of the data set, summary, plots, and maybe some kind of linear regression fit to check the feasibility of the problem as well as get a better idea of how this data looks. 4) Proposed work from now till the final project to be turned in. E.g., GLM, GAM, stepwise, CART, NNET, SVM, LDA, clustering, MDS etc. Model comparison, and cross validation. 5) List detailed references if it's suitable. 6) List possible simulation study design if it applies. Of course, the length as well as the content should largely depend on the problem each individual group is facing. A sample proposal is posted on the course web.

Tentative Schedule

Please bring your laptop to all classes.

Date	**Data Mining I**
Week 1 01/09	Overview/Data Mining; Exploratory Data Analysis; KNN
01/11, Lab	Lab: Statistical Software (R); R studio; Random Sampling; Data Exploration and Graphics using R; Exploratory Data Analysis
Week 2 01/16	Review of Linear Regression, Variable Selection; HW Report Group List due
01/18	LASSO Variable Selection; Cross Validation
Week 3 01/23, Lab	Lab: Linear Regression, Variable Selection, and Cross Validation; LASSO Variable Selection HW 1 due
01/25	Generalized Linear Models (e.g. Logistic regression)
Week 4 01/30	Logistic regression; ROC; Cross Validation (cv.glm); HW 2 due
02/01, Lab	Lab: Logistic Regression
Week 5 02/06	Classification and Regression Trees (CART) HW 3 due

02/08, Lab	Lab: Trees
Week 6 02/13	Trees; HW 4 due; HW discussion and Review
02/15	Trees; Case Study; Summary
Week 7 02/20	Exam (1 hour 50 minutes);
02/22	Guest Lecture (former alumni experience sharing)
	Data Mining II
Week 8 02/27	Overview; Generalized Additive Models (GAM); Nonparametric Smoothing
03/01	Neural Network; Case Study; Discriminant Analysis
Week 9 03/06, Lab	Lab: GAM, NNET, LDA, SVM
03/08	Guest Lecture (SPSS Modeler)
Week 10 03/13, 03/15	No Class. Spring Break
Week 11 03/20	Clustering; Project Guideline; Case 1 (individual) due
03/22	Clustering
Week 12 03/27	Association Rules; Case Study
03/29, Lab	Lab: Clustering & Association Rules
Week 13 04/03	Group Project Proposal due, presentation
04/05, Lab	SAS Enterprise Miner
Week 14 04/10	Summary Case 2 (individual) due
04/12	Text Mining
Week 15 04/17	Quiz
04/19	Group Project presentation
4/22, 11pm	Group Project due

Links to Resources

Related Links:

R package: http://cran.r-project.org/

R and Data Mining: http://www.rdatamining.com/
http://www.stat.wmich.edu/wang/R/links.html

R Enterprise
http://www.revolutionanalytics.com/

Getting Started with SAS OnDemand for Academics -
http://support.sas.com/ondemand/getstart.html

Learning to use SAS OnDemand for Academics: Enterprise Miner -
http://support.sas.com/ondemand/learn_em.html

Dunnhumby data: http://www.us.dunnhumby.com/sourcefiles.aspx

Other References:

Williams, Graham (2011), *Data Mining with Rattle and R*, Springer-Verlag New York, ISBN-13: 9781441998897.

Maindonald, J. W. and Braun, J. (2010), *Data Analysis and Graphics Using R – an Example-Based Approach*, ISBN-13: 978-0521762939.

SAS Institute (2003), *Data Mining Using SAS Enterprise Miner: A Case Study Approach, 2nd Ed.*. ASIN: B004R1QK3Y.
http://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf

Georges, Jim (2009), *Applied Analytics Using SAS Enterprise Miner 5.3, Course Notes*

Tan, P., M. Steinbach, V. Kumar, (2004), *Introduction to Data Mining*
http://www-users.cs.umn.edu/~kumar/dmbook/index.php (Computer Science)

Hosmer, D. and Lemeshow, S. (2000), *Applied Logistic Regression, 2nd Ed.*, Wiley, ISBN-13: 978-0471356325.