APPLIED DATA MINING (BIA 6301) SYLLABUS

Spring 2018 SESSION B

|  |  |
| --- | --- |
| COURSE NUMBER | BIA 6301 |
| COURSE TITLE | Applied Data Mining |
| MEETING DAY & TIME  CLASSROOM NUMBER | Wednesday 5:45 – 9:00 PM  Conway 004 |
| INSTRUCTOR | San Cannon |
| CONTACT INFORMATION  OFFICE  OFFICE HOURS | Office: 816-501- 4088  Email: [sandra.cannon@rockhurst.edu](mailto:sandra.cannon@rockhurst.edu)  Conway 315  Wednesday 4:30 – 5:30 PM or by appointment |

REQUIRED TEXTBOOK:

Lantz, Brett (2013). *Machine Learning with R*. 2nd ed. Birmingham, UK: Packt Publishing. ISBN: 978-1-78439-390-8. **Required.**

RECOMMENDED TEXTBOOKS (FOR YOUR PERSONAL LIBRARY):

Han, Jiawei, Micheline Kamber, and Jian Pei (2012). *Data Mining: Concepts and Techniques*. 3rd edition. Waltham, MA: Elsevier. ISBN: 978-0123814791.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). The Elements of Statistical Learning. 2nd edition. New York: Springer. ISBN: 978-0387848570. Free download: <http://statweb.stanford.edu/~tibs/ElemStatLearn/>.

Gareth, James and et al. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer. ISBN: 978-1461471370. A condensed Version of Hastie, Tibshirani, and Friedman (2009)} <http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf>

Lander, Jared P. (2014). *R for Everyone*. Upper Saddle River, NJ: Addison-Wesley. ISBN-13: 978-0-321-88803-7.

ARTICLES

Selected articles will be given as supplemental materials throughout the course.

WEB RESOURCES

Software Carpentry: <https://software-carpentry.org/lessons/>

R-bloggers: <https://www.r-bloggers.com/>

Quick R: <http://www.statmethods.net/>

Stack Overflow: <http://stackoverflow.com/>

Cross Validated: <http://stats.stackexchange.com/>

COURSE DESCRIPTION:The course provides a comprehensive overview of data mining techniques used to realize unseen patterns, including traditional statistical analysis and machine learning techniques. Students will analyze large datasets and develop modeling solutions to support decision making in various domains such as healthcare, finance, security, marketing, and customer relationship management (CRM). Models will include decision trees, clustering, classification, k-means, neural nets, support vector machines, ensemble methods and other supervised and unsupervised predictive models primarily for structured data. Students will also learn how to apply these models into production through business rules and SQL. The primary software tools for this class will be R and Python. Prerequisite: Statistics, knowledge of SQL and programming fundamentals.

EXPANDED COURSE DESCRIPTION: Models and concepts examine during the course include k-nearest neighbor, Naïve Bayes, decision trees, cluster analysis, association rules, and dimension reduction. Students will also learn how to conduct reproducibility research using R markdown.

COURSE OBJECTIVES:

1. To be able recognize business problems that can be solved with data mining applications.
2. To be able to effectively communicate data mining results to a general business audience.
3. To be able to prepare and evaluate basic and some advanced classification models.
4. To be able to prepare and evaluate basic and some advanced clustering models.
5. To be able to prepare and evaluate basic association models.
6. To be effective using the R platform for data mining and other data science applications.
7. To be able to distinguish ways of implementation of data mining models into production.

WEB PAGE: A web page has been set up for this class on Blackboard at <http://courses.rockhurst.edu> . If you have any difficulty accessing this web site please email me at sandra.cannon@rockhurst.edu. Data sets, slides and other materials will be posted on this site.

ATTENDANCE: For a technical course, attendance is of great importance. Therefore, students who are absent more than (1) time, may have their final course grade reduced by at least one full letter. If a student needs to miss more than twice, it is recommended they enroll in this class at a later time.

TECHNOLOGY: In each class session, students will need to bring a laptop capable of downloading and running open-source software and their associated packages. The laptop also needs the capability of accessing the university’s wireless internet. It is also important that the person has administrative privileges for their laptop in order to load and modify programs and packages easily. The laptop may be a Windows, Mac or Linux based system. It is highly recommended that the laptop be running a more recent operating system (i.e. Windows 7 and above or Mac 10.8 or above) and have at least 4GB of RAM. Higher amounts of RAM (8 GB or more) is also recommended.

HOMEWORK ASSIGNMENTS: Assignments will be given to practice the concepts and techniques we learn in class. While I encourage you to communicate with each other, the assignments are meant to be individual work. I should not see identical outputs or copies of code between students. If you borrow code from a website, please provide proper citation. It goes without saying that you should not be copying codes from instructor’s manuals.

I will be using SafeAssign, a Blackboard feature that scans all writing submissions for plagiarism and cheating. If I determine that you have copied from another student, I will report you to the Dean’s Office. If I find that you have “borrowed” code or text from another source without appropriate attribution or citation, I will report that to the Dean’s Office as well. There will be no “second chances.”

**Assignments are due at 5:45 the night of class (so @ 5:45 PM Tuesday). 10% penalty is assessed for each day an assignment is submitted late. I WILL NOT GRADE any assignment that is over a week late.**

COURSE PROJECT: Students will create a data mining project of their own choosing. Data sets from employers are encouraged but not required. Three grading components will be given for course project: 1) a presentation; 2) R markdown file documenting completed work; and 3) feedback comments given to peers’ presentations.

EVALUATION

The student’s final grade will be determined as follows:

* Three assignments: 70 points
* Final Project: 30 pts <= 23 points for presentation, 2 points for documentation, 5 points for feedback on other presentations.

**Total Points: 100 pts.**

|  |  |
| --- | --- |
| Course Grade Percentage | Course Grade Letter |
| 93% and above | A |
| 92.9% - 86% | B+ |
| 85.9% - 80% | B |
| 79.9% - 70% | C |
| 69.9% and below | F |

TENTATIVE CLASS SCHEDULE

|  |  |  |
| --- | --- | --- |
| Week | Topic(s) | Reading(s) |
| 1 (3/14) | Review of Linear & Logistic Regression  K-Nearest Neighbor and Naïve Bayes  **Homework #1 Available** | Chapter 6 (review on your own)  Chapters 3 & 4 |
| 2 (3/21) | Decision Trees | Chapter 5 |
| 3 (3/28) | Data Preprocessing, Data Partitioning & Cross Validation  **Homework #1 due @ 5:45 PM**  **Homework #2 Available** | Han et al, Chapter 3  Gareth et al, pp. 175-186 |
| 4 (4/4) | Cluster Analysis | Chapter 9 |
| 5 (4/11) | Association Rules Mining  **Homework #2 Due @ 5:45 PM**  **Homework #3 Available** | Chapter 8 |
| 6 (4/18) | Model Evaluation and Tuning | Chapter 10 & 11 |
| 7 (4/25) | Curse of Dimensionality  **Homework #3 due @ 5:45 PM** | Gareth et al, pp. 373-385  Handout |
| 8 (5/2) | Final presentations done in class  **Final slides and compiled markdown due SATURDAY, MAY 5 at @ 11:59 PM** |  |

FINAL PROJECT INSTRUCTIONS

**Goal**: To apply multiple (and appropriate) data mining technique(s) to a selected data set.

**Instructions**:

1. Find a dataset you are interested in mining. The dataset can be from work or personal interests. If you are in need of a data set, consider these sites:

<https://www.kaggle.com/>

<http://www.data.gov/>

<https://data.kcmo.org/>

<https://catalog.data.gov/dataset?groups=local>

<https://archive.ics.uci.edu/ml/datasets.html>

<http://unstats.un.org/unsd/databases.htm>

<http://www.who.int/healthinfo/statistics/en/>

<https://www.cdc.gov/DataStatistics/>

1. Apply **multiple and appropriate** data mining techniques learned from class to examine the data set. This does not mean to apply **ALL** the techniques that we have covered. If you are curious about a data mining technique not discussed in class and want to apply it for the project, please feel free to do so. The goal is to learn and discover.
2. Create a PowerPoint presentation and deliver it to the class during Week 8. Please keep your presentation brief—no more than **seven (7)** minutes.
3. During class, everyone will be expected to provide constructive feedback on the presentations. After class, the instructor will provide the feedback from your classmates to you. The recommendations are anonymous.
4. You may choose to incorporate your classmates’ recommendations to improve the project. Your grade may improve if you choose to make changes; it will not go down if you choose not to.
5. Upload the final slides and R markdown file to Blackboard for grading. Please note if you made changes to the presentation and provide appropriate commentaries in the R markdown file. Some things to consider adding to the markdown file include:
   * What is your research question?
   * What models did you use to mine the data set(s)? Why did you choose these models?
   * What were the findings?

**BIA 6301: Grading Rubric for Course Presentation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Exceeds Expectations**  Clear Comprehension  (Audience clearly understands the presenter’s message.) | **Meets Expectations**  Moderate Comprehension  (Audience understands the message being relayed for the most parts, but there are still some unclear/confusing remarks.) | **Below Expectations**  Confusion  (Lack of clarity in presenting the message to the audience.) | **Fail**  Does not address grading parameter. |
| Identify an area of improvement; problem to be solved; or interesting question to be studied | 4 points | 2-3 points | 1 points | 0 points |
| Apply data mining concepts, models, and tools to make improvements/solve problems/or to study questions of interest | 7 | 4-6 | 1-3 | 0 |
| Effectiveness in delivering the presentation for a GENERAL AUDIENCE including oral presentation and written output (eg. Slides or handouts) | 8 | 5-7 | 2-4 | 0 |
| Make appropriate conclusions | 4 | 2-3 | 1 | 0 |
| Documentation/code submitted in support of presentation | 2 | 2 | 2 | 0 |
| Thoroughness/effectiveness/relevance of comments provided to fellow students | 5 | 3-4 | 1-2 | 0 |

**Rockhurst University and Helzberg School Policies** **and Statements**

**Online Wall Street Journal Subscription**

All Helzberg School of Management students are given an online subscription to the Wall Street Journal.  Using Google Chrome students can sign-up for the online journal at <https://forms.rockhurst.edu/helzberg/wsj> (use your Rockhurst username and password).  Access to your online journal is at [www.wsj.com](http://www.wsj.com).  Should you have questions or problems contact the Helzberg School front desk (C201) at x4090 or email [lynn.ross@rockhurst.edu](mailto:lynn.ross@rockhurst.edu).  The customer service number for the online journal is 800-369-2834 or [onlinejournal@wsj.com](mailto:onlinejournal@wsj.com). Student guides can be found in C201.

**Academic Honesty Policy:**

Plagiarism and cheating will not be tolerated. The Rockhurst University Catalog provides examples of academic dishonesty and outlines the procedures, penalties, and due process accorded students involved in academic dishonesty. All infractions will be immediately referred to the Dean's office.  In your research paper, make sure you provide citations for all ideas and information that are not your own. This includes *copying, or the offering, requesting, receiving or using of unauthorized assistance or information in examinations, texts, reports, computer programs, term papers or other assignments*.

Graduate Policy:

<http://catalog.rockhurst.edu/content.php?catoid=9&navoid=452#Academic_Honesty_Policy>

**Helzberg School Student Complaint Process**

Students who feel that they have a nontrivial complaint, either academic or nonacademic, should contact either Paul Nunez, Director of Undergraduate Advising, or Jonnae Hill, Director of Graduate Advising, for guidance on submission.

**Course Assessments:**

Rockhurst University and The Helzberg School are committed to a comprehensive, ongoing assessment process. On occasion students will be expected to participate in aspects of this program so that the institution can document its strengths and identify opportunities for improvement. Student involvement is integral to the University’s success in this endeavor.

**ADA Statement**

RU official policy: “Rockhurst University is committed to providing reasonable accommodations for students with disabilities. Please contact Sandy Waddell, Director, Access Office (Massman Hall Room 7, (816) 501-4689, [sandy.waddell@rockhurst.edu](mailto:sandy.waddell@rockhurst.edu)) to provide documentation and request accommodations. If accommodations have already been approved by the Access Office, please communicate with the instructor of this course regarding these arrangements by the second week of class in order to coordinate receipt of services.”

**Student Resources**

In keeping with our commitment to *cura personalis*, “care of the whole person,” Rockhurst University provides a variety of programs and services beyond those of the Helzberg School of Management that are designed to support you as you pursue your education and navigate challenges, both expected and unexpected.  These include, but are not limited to: the Dean of Students Office, Dining Services, Bookstore, Counseling Center, Career Services, Access/Disability Services, Campus Ministry, Student Health Insurance and Clinic, Student Organizations & Greek Affairs (alumnus advisors), Athletics/Intramurals/Workout spaces, International and Multicultural Student Services, Security, Financial Aid, Student Accounts, Computer Services, The Registrar, The Library, Service Learning, and the Learning Center.  For information on any of these, please visit the University’s website at [www.rockhurst.edu](http://www.rockhurst.edu) – or call/visit Student Development at 816-501-4030 or Massman Hall, Room 1.

**Crisis Management**

The Crisis Management Team for Rockhurst University, out of a concern for the safety and welfare of all community members, urges you to familiarize yourself with Campus Emergency Procedures as well as emergency, evacuation, and shelter signage located within and outside of buildings across campus.  They also request that you sign up for Rock@lerts, the University’s emergency communications system.  All community members must comply with University staff instructions during regularly scheduled fire/tornado drills as well as actual emergencies.  Information regarding the above resources can be found at [www.rockhurst.edu/emergency](http://www.rockhurst.edu/emergency). In addition, the University has introduced a new tool, CampusEye, to send photos and reports — anonymously, if you choose — and check on reports submitted previously to security directly from your smartphone. *To use CampusEye, download the free app for iOS or Android operating systems and sign in using our campus code, 2YPJ.*

**University Communication with Students**

Clear and timely communication allows students to receive information related to policies, programs, events, and other practical matters (i.e. billing, grades, etc.) affiliated with their education at Rockhurst. The University has established several key routes for communication with students including:

* Rockhurst E-mail Account: All Rockhurst students receive a University e-mail account. All e-mail communication from the University is directed exclusively to the Rockhurst electronic mailbox system. Students are expected to access their e-mail account on a regular basis (daily is recommended) in order to stay abreast of important and time-sensitive information. University departments, faculty, and staff will routinely use e-mail to communicate important campus, academic, and extra-curricular/co-curricular information. It is the responsibility of each student to clean their e-mail boxes to allow capacity for incoming messages (i.e. empty deleted items, keep a limited number of sent items, etc.). For further information on your Rockhurst e-mail account, please see Computer Services (Conway Hall 413; x4357; [*www.rockhurst.edu*](http://www.rockhurst.edu/)).
* Addresses and Phone Numbers: Students are required to maintain accurate local, billing, permanent, and emergency contact information so that attempts to communicate by the U.S. Postal System as well as phone will not be impeded. To make changes to your addresses or phone numbers of record, please visit the Rockhurst website ([*www.rockhurst.edu*](http://www.rockhurst.edu/); Rockweb section under Registrar) or contact the University Registrar (Massman Hall 110).

*Updated March 2018*