BUAN 4310 Data Mining and Big Data

Albers School of Business and Economics
Seattle University

1. General information

Instructor: Ben Yeo (Master Yeoda)
Email: byeo@seattleu.edu
Classroom: PIGT 200
Day/Time: Tue & Thu; 1545 – 1750 hrs
Office Hours: Tue & Thu; 1445 – 1545 hrs
Pre-requisite: BUAN 3210

2. Course description

Business analytics involves the use of data driven decision-making techniques. Among which is data mining. Businesses, governments, and non-profit organisations collect and have access to large amounts of data about customers, suppliers, employees, regions, population, transactions, and such. Data mining covers a wide range of techniques to detect actionable patterns in data and generate predictions. Most business analytics applications are geared towards benefitting the company, sometimes at the expense of the individual, community, and society. Our focus in this course is on humane and socially responsible business analytics. In the course, we will work with various data problems. We will examine types of questions that data mining can answer and the appropriate data mining techniques for answering different questions. The emphasis is on understanding the concepts and logic behind a wide range of data mining techniques and their relation to business analytics. By completing the course, students will be better equipped to apply these techniques to address various business problems.

3. Faculty biographical sketch

Yeoda is a faculty at Seattle University. His research interests include innovation- and technology-driven economic growth and social informatics, using both quantitative and qualitative methods, including traditional statistics and data mining. Prior to academia, Yeoda worked in the industry as a Senior Research Analyst in Economics. He received a Ph.D. in Information Science from the College of Information Sciences and Technology at the Pennsylvania State University. Obviously, a Star Wars nerd, he is ☺

4. Learning objectives

The primary course outcome is to provide students with an understanding of data mining techniques and apply them to solve different business problems. Specific learning objectives include:

1. Become familiar with the potential and dangers of data mining in today’s data rich-environment
2. Gain experience using key data mining methods of classification, prediction and exploration
3. Know how to decide when to use which technique
4. Be able to implement major data mining techniques using software
5. Gain the intellectual capital required to provide responsible business analytics services
6. Be conscious of humanistic and societal implications of deploying data mining in organisations
7. Define key terminology in model fitting and statistical learning
8. Pre process data for data mining
9. Compute and interpret descriptive statistics and graphical techniques
10. Build, analyse and interpret results of various supervised and unsupervised learning techniques

5. Course materials

In this course, we will use the R programming environment and language. R is a highly flexible tool, and it is now ubiquitous in modelling across many disciplines. R is not a point and click application like SPSS or Stata; it requires some programming skills, which students should already have a basic understanding. This course uses RStudio the environment for the language. Since the course focuses on the techniques and applications, rather than the R syntax, students may use XLMiner (with a graphical user interface) if they prefer. Access to XLMiner will be discussed in class. Additional materials may be assigned as needed.


OR


6. Evidence students will submit

Grades are based on course deliverables. Specifically, there are 4 graded components for the course. These are given as follows.

- Recaps x8
- Term Project
- Mid Term
- Final exam

Recaps are self-paced homework on the material covered in that corresponding week2. They are designed to keep us up to date with the concepts and force us to revise our notes in preparation for the mid term and exam. These are auto-graded on Canvas. Each recap

1 Students may use this textbook if they prefer to use XLMiner. However, the lectures will be based on R. We will try to cover XLMiner if time permits, but XLMiner is rather straightforward to use.
2 Not every week has an associated Recap assigned. Please check the course schedule.
comprises approximately 10-15 multiple choice and true/false questions, and possibly some other formats. Students can re-do these for a higher grade (or for revision) before the course officially ends on the day and time of the final exam, if the first attempt was completed before the due date.

The term project is problem-based, where students apply the tools covered to develop solutions to solve the given problem(s). The project comprises a proposal, a presentation, and a final report\(^3\). Students should complete these projects in groups of 3-4 members\(^4\) each. All deliverables should be submitted on Canvas. Feedback will be provided to submitted work on Canvas approximately within one week of submission. **All late submissions of any deliverable without an approved extension\(^5\) will receive a 5% penalty for each day late.** The quality of submissions will be measured using corresponding rubrics, available on the assignment itself. However, some aspects may involve subjective evaluation. As much as possible, the rubrics will frame the assessments\(^6\). All deliverables will be evaluated on evidence of learning, depth of analysis, organisation and thoroughness. They should demonstrate the consideration of key issues and critical thinking with respect to the data mining algorithms applied in a business setting. All written deliverables should be professional prepared, typed, and contain a bibliography of cited sources and appropriate footnotes (where applicable). All ideas, quotes and statistics borrowed from another author/source must be cited. If external content (such as facts or data) are used, please be sure to cite the corresponding authorities. Students may use any citation style, formal or otherwise. The objective here is simply to attribute to corresponding resources accordingly.

The mid term comprises a series of conceptual, factual and some applied questions that students should be able to address within 90 minutes. There are likely to be approximately 60\(^7\) questions (multiple choice, true/false, and/or short answer questions). **Due to the short quarter, make-ups must be completed within a week of the original test date.** However, this requires appropriate documentation, such as medical documentation, for a valid and/or official reason to justify the absence. **Make ups will be different (i.e. different questions) and thus, may not be at the same level of difficulty.** If a student requires special accommodation, he/she must inform the instructor in advance with the appropriate documentation so special arrangements can be made.

---

\(^3\) All project deliverables cannot be graded twice. Students can submit multiple times before the corresponding deadlines, but only the latest submission before the corresponding deadlines will be taken into account.

\(^4\) A peer evaluation of each group member’s contributions is available at the end of the term. Students may choose not to submit a peer evaluation if they think everyone contributed equally. This score will affect the final project score.

\(^5\) Extensions on any deliverable have to be very well justified on reasonably valid reasons and with appropriate documentation. The deliverables are assigned well in advance and any delays should be anticipated. Last minute extensions are not acceptable except under emergency situations.

\(^6\) All grade disputes (if any) must be addressed immediately. No grade changes to the deliverables are allowed 1 week after the grade for the corresponding deliverable is released.

\(^7\) +/- 10%. Subject to minor changes as needed. The format will be clearly communicated in class prior to the test date.
The final exam follows a similar format to the mid term, except it is cumulative. Hence, it is weighted more heavily. Make ups for the final exam require sufficient documentation (such as medical), and must be taken within the same finals week. Make ups for the final exam require sufficient documentation (such as medical), and must be taken within the same finals week. Per the University's policy, INCOMPLETE grades can only be given under extenuating circumstances. This does not include missing the deadline for a final project because of connection issues, or having a common cold on the day of the exam. Make ups will be different and may not be at the same level of difficulty. Likewise, if a student requires special accommodation, he/she must inform the instructor in advance with the appropriate documentation so special arrangements can be made.

6.1 Expectations
Participation and studying outside class times are critical in completing the course. Students are expected to read the assigned materials according to the course schedule.

Attendance is critical in completing the course. Absence from more than two classes can lead to a failing grade. Students should contact the instructor if he/she anticipates missing a class. Students are responsible for reviewing materials posted on Canvas. Students are expected to be present in all classes barring unforeseen circumstances and official/professionally-related absences due to necessity. In addition, students are expected to participate in class discussions to facilitate a healthy exchange of ideas and an effective learning experience for everyone. Class participation comprises (i) class attendance and (ii) substantive contributions to class discussions.

6.2 Criteria for assessment
The final grade will be based on the graded deliverables

<table>
<thead>
<tr>
<th>Recaps (x8)</th>
<th>6%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project:</td>
<td>52%</td>
</tr>
<tr>
<td>• Proposal</td>
<td>13%</td>
</tr>
<tr>
<td>• Presentation</td>
<td>13%</td>
</tr>
<tr>
<td>• Final report</td>
<td>26%</td>
</tr>
<tr>
<td>Mid Term:</td>
<td>20%</td>
</tr>
<tr>
<td>Final exam:</td>
<td>22%</td>
</tr>
<tr>
<td>Total:</td>
<td>100%</td>
</tr>
</tbody>
</table>

6.3 Grade descriptions
Grading will follow the Albers School’s recommended rigour. Students are expected to follow the rubrics very closely.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Point</th>
<th>Grade Range</th>
<th>Grade Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4.0</td>
<td>94% +</td>
<td>The instructor judged the student to have</td>
</tr>
</tbody>
</table>

8 Several concepts are related.
9 This will be approximately 20% of the course, which is synonymous with missing a day's work in a week.
10 It is very likely that we will only have 7 recaps due to time constraints.
11 The final score will be modified by peer evaluations.
12 The format will be discussed in class.
accomplished the stated objectives of the course in an OUTSTANDING manner.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-</td>
<td>3.7</td>
<td>90 – 93.99%</td>
</tr>
<tr>
<td>B+</td>
<td>3.3</td>
<td>87 – 89.99%</td>
</tr>
<tr>
<td>B</td>
<td>3.0</td>
<td>83 – 86.99%</td>
</tr>
<tr>
<td>B-</td>
<td>2.7</td>
<td>80 – 82.99%</td>
</tr>
<tr>
<td>C+</td>
<td>2.3</td>
<td>77 – 79.99%</td>
</tr>
<tr>
<td>C</td>
<td>2.0</td>
<td>73 – 76.99%</td>
</tr>
<tr>
<td>C-</td>
<td>1.7</td>
<td>70 – 72.99%</td>
</tr>
<tr>
<td>D</td>
<td>1.0</td>
<td>60 – 69.99%</td>
</tr>
<tr>
<td>F</td>
<td>0.0</td>
<td>Less than 60%</td>
</tr>
</tbody>
</table>

The instructor judged the student to have accomplished the stated objectives of the course in an ACCEPTABLE manner.

The instructor judged the student to have accomplished the stated objectives of the course in a POOR manner; but it is still a passing grade.

The instructor judged the student to have accomplished the stated objectives of the course in a DREADFUL manner. (A grade of D will not fulfill the requirements in a major field of concentration.)

The instructor judged the student NOT to have accomplished the stated objectives of the course.

7. Course schedule

Classes will comprise a mixture of lectures, demos, and labs. Lectures are geared towards explaining the concepts behind the techniques. Demos illustrate how the techniques work. Labs are hands-on activities where students will solve a business problem in class. These are designed to provide students with sufficient knowledge and experience with these techniques. In addition, homework is assigned on most weeks to provide conceptual reviews and hands-on practice.

<table>
<thead>
<tr>
<th>Week Date</th>
<th>Topic</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview and Data Preparation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

13 Class schedules are tentative and are subject to changes as necessary.
14 Additional readings will be assigned as needed.
| Week 1 | Introduction to Data Mining\(^{15}\) | Readings:  
- Galit et al. Ch 2  
Activities:  
- Course overview  
- Conceptual overview  
- Getting into project groups\(^{16}\)  
- Lab 1: Software installation\(^{17}\)  
Homework:  
- Recap 1 (1st attempt due Oct 6\(^{18}\)) |
| Week 2 | Data Pre Processing and Visualisations | Readings:  
- Galit et al. Ch 3  
Activities:  
- Demo: Alita’s World Time Heist  
- Lab 2: Laptop Sales  
Homework:  
- Recap 2 (1st attempt due Oct 6)  
- Term project assigned (check due dates for different parts)\(^{19}\) |

**Supervised Learning**

| Week 3 | K Nearest Neighbours | Readings:  
- Galit et al. Ch 7  
Activities:  
- Demo: Hogwarts Sorting Hat  
- Lab 3: Bank Loans  
Homework:  
- Recap 3 (1st attempt due Oct 13) |

---

\(^{15}\) Since we start on the second class of the week in Week 1, it is very likely that some of this Chapter’s materials will be carried over to the following week.  
\(^{16}\) For those who are unable to find a group of 3-4 students in time, the instructor will randomly assign them.  
\(^{17}\) Please follow the instructions to install R/R Studio (and/or XLMiner if you prefer). Given the time constraint, as we only meet once this week, Lab 1 may be to be completed outside class hours.  
\(^{18}\) Since part of Week 1 will be covered in the following week (see previous footnote), the due date for the first attempt of Recap 1 is Oct 6.  
\(^{19}\) The proposal is due by the end of Week 4. This is to allocate sufficient time for feedback. The presentation will be done in Week 10.
<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Topic</th>
<th>Readings</th>
<th>Activities</th>
<th>Homework</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Demo: Dungeons and Dragons</td>
<td>Assignment 1 (due Oct 27)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lab 4: Used Car Sales</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Oct 22 &amp; Oct 24</td>
<td>Large Data Sets and Unbalanced Data</td>
<td>To be determined</td>
<td>Demo: It's a Big, Big World</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Demo: Every Rose has Its Thorn</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>Mid Term</strong> (Taken on Oct 22 in class)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q&amp;A for term project</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Demo: Every Rose has its Thorn</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lab 5: Airfare prediction</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Nov 5 &amp; Nov 7</td>
<td>Logistic Regression</td>
<td>Galit et al. Ch 10</td>
<td>Demo: Glory! Glory! Man United!</td>
<td>Recap 6 (1st attempt due Nov 10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lab 6: Competitive eBay Auctions</td>
<td></td>
</tr>
</tbody>
</table>

---

20 There is no specific Lab this week as the topics are shorter and we have a mid term on the first class of the week. We will run through the demos as hands on activities.
| Week 8  | Model Diagnostics\(^{21}\) and Discriminant Analysis (Tentative)\(^{22}\) | Readings:  
- Galit et al. Ch 12  
Activities:  
- Q&A for term project  
- Demo: I Solemnly Swear I am Up to No Good\(^{23}\)  
- Lab 7\(^{24}\): Worker Competence  
Homework:  
- Recap 7\(^{25}\) (1st attempt due Nov 17) |
|---|---|---|
| Week 9  | Clustering 1: Hierarchical Clustering | Readings:  
- Galit et al. Ch 15  
Activities:  
- Demo: Clustering the Star Wars Universe 1  
- Lab 8: Universities and Pharmaceutical Firms  
Homework:  
- Recap 8 (1\(^{st}\) attempt due Nov 26) |
| Week 10 | Clustering 2: Non-Hierarchical Clustering\(^{27}\); and Association (tentative)\(^{28}\) | Readings:  
- Galit et al. Ch 15 and Ch 14\(^{29}\)  
Activities:  
- Demo: Clustering the Star Wars Universe 2  
- Lab 8 (continued): Universities and Pharmaceutical Firms  
- Demo: Course Recommendation for Ron Weasley\(^{30}\) |

\(^{21}\) Time permitting.  
\(^{22}\) Time permitting.  
\(^{23}\) Time permitting.  
\(^{24}\) Time permitting.  
\(^{25}\) This will be unassigned if we do not have time for discriminant analysis.  
\(^{26}\) Nov 28 is a holiday.  
\(^{27}\) Time permitting.  
\(^{28}\) Time permitting.  
\(^{29}\) We are not likely to cover Ch 14 due to time constraints.  
\(^{30}\) Tentative.
Week 11
Dec 3 & Dec 5

<table>
<thead>
<tr>
<th>Presentation³¹</th>
<th>Activities:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Term project presentations³²</td>
</tr>
</tbody>
</table>

Finals
Dec 9 – Dec 13

| Final exam week | Check final exam schedule³³ |

8. Academic resources

- Library and Learning Commons (http://www.seattleu.edu/learningcommons/)
- This includes: Learning Assistance Programs, Research [Library] Services, Writing Center, Math Lab
- Academic Integrity Tutorial (found on Canvas and SU Online)

9. Academic policies

Various Seattle University academic policies can be found at the following URL. Please be sure to review them prior to the course.

https://www.seattleu.edu/redhawk-axis/academic-policies/

- Academic Integrity Policy
- Academic Grading Grievance Policy
- Professional Conduct Policy (only for those professional programs to which it applies)

10. Notice for students concerning disabilities

If you have, or think you may have, a disability (including an ‘invisible disability’ such as a learning disability, a chronic health problem, or a mental health condition) that interferes with your performance as a student in this class, you are encouraged to arrange support services and/or accommodations through Disabilities Services staff located in Loyola 100, (206) 296-5740. Disability-based adjustments to course expectations can be arranged only through this process.

11. Notice on religious accommodations

It is the policy of Seattle University to reasonably accommodate students who, due to the observance of religious holidays, expect to be absent or endure a significant hardship during certain days of their academic course or program. Please see, Policy on Religious Accommodations for Students:

³¹ The format will be discussed in class.
³² All students are expected to be present on both days for the presentations.
³³ Please note the final exam schedule. Family/personal travel plans (and vacation) are not acceptable reasons for missing the final exam. There is sufficient time to plan in advance.
12. Office of institutional equity

Title IX of the Education Amendments of 1972 (Title IX) prohibits discrimination based on sex in educational programs or activities that receive Federal financial assistance. This prohibition includes sexual misconduct, which encompasses sexual harassment and sexual violence. Seattle U remains committed to providing a safe and equitable learning, living, and working environment. Seattle U offers emergency, medical, and other support resources, as well as assistance with safety and support measures, to community members who have experienced or been impacted by sexual misconduct.

Seattle U requires all faculty and staff to notify the University's Title IX Coordinator if they become aware of any incident of sexual misconduct experienced by a student.

For more information, please visit https://www.seattleu.edu/equity/. If you have any questions or concerns, you may also directly contact the Title IX Coordinator in the Office of Institutional Equity (email: oie@seattleu.edu; phone: 206.296.2824) University Resources and Policies