CSC 367: INTRODUCTION TO DATA MINING

FINAL PROJECT

Total Points: 45

Rules and guidelines for the group project

The project must be completed by a team of max three students. In-class students are encouraged to include online students in their group. Students looking for group partners can post messages on our discussion forum, or can send me an email. Each team must elect a team leader that will serve as a contact point for the group. The leader will also make sure that the deadlines are met, and deliverables are submitted through D2L.

The project consists in the analysis of a large data set using the data mining techniques presented in class. The possible analytics tasks for the data analytics project include a classification task, or a clustering task (e.g. customer segmentation).

Each team should find a large publicly available dataset. A list of data repositories are included below. As a minimal requirement, the dataset should contain at least ten variables and >500 observations, but I am willing to make exceptions if students provide reasonable motivation to use a smaller dataset.

At the completion of the project, each team will submit a report of their analyses in the form of a paper and will present their results in class during Week 10 and week 11. Online students are exempted from the presentation requirement, but if they work in a team with in-class students, the other students in their team still need to present the results.

The project assignment is organized into various deliverables and milestones, as described below.

FIRST DELIVERABLE: ONE-PAGE PROPOSAL (5 PTS)

Due date Tuesday 10/6/2015 before 11:50am

Each group will identify a dataset/problem they wish to investigate. Each group will submit a one page pdf proposal. Proposal needs to contain the following information:

1. Who is in your group
2. Name of your *team leader* that will make sure deadlines are met, and will be the contact point for the team.
3. Information about the dataset you plan to use, including number of attributes and size of dataset (remember you need at least ten variables and >500 records in dataset – I am willing to grant exceptions if necessary) and where you plan to get it from
4. What is the question(s) of interest? Try to be specific, and define a question that your team can answer with the data. (you may need to change this question and adjust it once you start analyzing the data)

SUBMISSION: Submit one proposal for group. The team leader will be responsible to upload the proposal on the group project dropbox folder on D2L

Data repositories:

There are many publicly available datasets that can be used for your project. Here is a list of sites where you can find large datasets. I'll also post some datasets can students can use.

- UCI repository at http://archive.ics.uci.edu/ml/datasets.html . You can search data by type, and data mining task
- KDnuggets is a great website that contains lots of information of interest to data scientists. It also includes a long list of data repositories: http://www.kdnuggets.com/datasets/index.html
- Datasets used for data analytics competitions at http://www.kaggle.com/
- Health Data Sets at the Center for Disease Control and Prevention website: https://data.cdc.gov/
- Data and Story Library at http://lib.stat.cmu.edu/DASL/
- Chance datasets posted at the University of Dartmouth http://www.dartmouth.edu/~chance/teaching_aids/data.html

SECOND DELIVERABLE: PRELIMINARY ANALYSIS (10 POINTS)

Due date Tuesday 10/27/2015 before 11:59pm

The cleaning and preparation of the data will take a considerable amount of time, but this is an essential task in your analytics work.

1. Apply data cleaning and preprocessing techniques to prepare the data for your analytic task.
2. Are there missing variables or outliers? Decide what you will do with them.
3. After you clean your data, you will conduct an exploratory analysis of your data analyzing frequencies, distributions and associations among the attributes.

The deliverable for this milestone will include a technical document describing

a) preprocessing steps to clean the dataset including
   - Variable transformations and recoding
   - Analysis of outliers
   - analysis of missing values
   - Analysis of associations among variables

b) a summary of the exploratory data analysis. (3-4 pages max)

SUBMISSION: The team leader will be responsible to upload the technical document summarizing the preliminary analysis at the group project dropbox folder on D2L

THIRD DELIVERABLE: CHECKPOINT (8 POINTS)

Due date Tuesday 11/5/2015 before 11:50am

During week nine, I will meet with each group. I'll meet with In-class groups during regular class time. Online students can meet with me through Skype or Google Chat. During this meeting, we will review the steps of your analysis in detail.

For this checkpoint meeting your team should prepare a 1-2 page paper that outlines the steps of your analysis, including the following information:

a) Explain data mining task of your project: is this a classification task or a clustering task?
b) Specify list of attributes that will be used to answer the questions of your project, for

instance at this step you should have identified the response variables, and attributes for a classification task, or a list of the attributes to be used for the clustering task.

c) Discuss work progress on your project, and your next plans, including the data mining methods that have been used to complete task (choose among the techniques presented in class). You need to be specific here, and show me that you have a solid plan in place.
d) Discuss your chosen method for validating/testing your model, including how you will define the testing set and training set, and the validation metrics.
e) Discuss possible issues with your dataset

SUBMISSION: Submit the checkpoint summary paper. The team leader will be responsible to upload the document on the group project dropbox folder on D2L prior to the meeting.

LAST DELIVERABLE: PROJECT REPORT AND PRESENTATION (22 POINTS)

Final Report due date Friday 11/20/2015 (Finals week) before 11:59pm

The project presentation (for in-class students) will be scheduled on Tuesday 11/17/2015 at 11:50-1:30pm in our regular classroom. It will be set up as a poster session. Each group will create a poster presentation to illustrate their findings. Posters should present information using graphs, tables and text, in an organized and visually pleasing fashion. More details and guidelines on the poster presentation will be provided later in the course.

Project report

Groups should describe their work in a report consisting of two sections:

1. *A non-technical summary* not longer than one typewritten page, describing the conclusions of your data analysis to a non-technical audience. It should be intelligible to a person who does not know data mining. Suppose you are talking to your boss who does not know statistics or to a friend who is not familiar with statistical terminology. This can be seen as the executive summary/introduction of your report.
2. *A technical report* not exceeding 9 pages, describing your analysis and the data. You can include graphs and output tables, only if you use them in your discussion, but do not exaggerate! This section is intended for a statistically literate audience and must be written in a clear organized fashion. For instance, you can organize the report into subsections such as
    a) Data description
    b) Exploratory analysis of the data. (You can use the paper for deliverable 2)
    c) Data mining technique and analysis of results
    d) Validation and testing
    e) Discussion and Conclusions.

Each group member must submit an evaluation of the work of each group member. This evaluation must be completed by each group member and must be submitted through D2L.

If you have any questions or problems, feel free to consult me at any time.

Grading Rubric for Final Report and presentation (MAX 22 points)

	Excellent	**Good/Fair**	**Poor**
Layout and clarity	Report is clear and neatly organized, with appropriate use of headings and tables to enhance readability. Report is well written with appropriate use of grammar and statistical terminology. (2pts)	Report is mostly clear with parts that are not well organized under sections. Use of appropriate statistical terminology is limited. (1pt)	Report is not clear. The layout is cluttered and not organized in sections. Major editing and revision are required. Errors in spellings, capitalization, punctuation and grammar distract readers. (0pts)
Non technical summary	Summary describes the conclusions clearly, concisely and is written in a language that is appropriate for a non-technical audience (2pts)	Summary describes the conclusions clearly, concisely but contains statistical jargon that is inappropriate to a non-technical audience (1 pt)	Summary is unclear and uses technical terms and language that is inappropriate to a non-technical audience (0 pts)
Technical analysis			
Data Cleaning	Preprocessing steps are adequate. good analysis of outliers and missing values, and overall quality of data (2pts)	Preprocessing steps are incomplete or or is incorrectly interpreted. (1pt)	No preprocessing steps (0 pt)
Exploratory data analysis	Exploratory analysis describes distribution of attributes and identify most important attributes for analysis (2pts)	The analysis of the distribution is incomplete or is incorrectly interpreted. 1pts	The analysis of the distribution is incorrect and there are some major issues with the approach. (0 pt)
Data mining tool	Data mining technique is applied correctly and results are interpreted correctly. (3pts)	There are some minor problems with the data mining approach or interpretation is not adequate (2-1pts)	Major problems with data mining technique and its application 0pt
Diagnostics/validation methods	Results are tested using appropriate methods (2pts)	Model validation methods are incomplete or	Major flaws in validation results 0pts

Discussion & Conclusions	Section includes a clear summary of the findings of your project and a discussion on the limitations of your analysis. 2pts	discussion of results is inadequate 1pt Section is not clearly written or interpretation is incorrect (1pt)	Section is missing (0 pts)
Presentation	Presentation is clear and well organized – excellent use of visual tools (3pts)	Presentation is clear, but not very well organized – poor use of visual tools (2pts)	Presentation is disorganized and unclear – (1pts)
Appendix (optional) – points will be assigned elsewhere.	Well organized – graphs and output are labeled appropriately. (2pts)	Not well organized – graphs and output have missing labels (1pts)	Cluttered and confusing (0pts)
Group Evaluation	Submitted (2pts)	Not submitted (0pts)	