CUNY The City University of New York | School of Professional Studies

DATA 621 – Business Analytics and Data Mining

Final Group Project Requirements

The final course project will be done in groups of 2 or 3 students (depending on your own choice). You may choose your own topics (and teammates) by yourself. The purpose of this project is to explore, analyze and model a real-world data set of your own interest using the regression modeling techniques learned in the course. The real-world data set can be either:

- A data set that you have personally collected (e.g., at your workplace, internship, etc.).
- An open-source data set that you have downloaded from the Internet (e.g., CDC, NIH, NHANES, MEPS, BRFSS, Kaggle, InfoChimps, etc.).
- Textbook exercises are not suitable for course projects.

You will need to develop a problem statement and research question(s) based on the data set that you have obtained. You must survey the state-of-the-art literature and research developments (published in journal papers, not conference papers) dealing with empirical studies, algorithms, or methodologies related to your problem. You will need to achieve one deliverable (a final report).

Deliverable: Final Report (250 points). The final project report should be similar to the technical papers you read in the literature. The report should not exceed 12 single-spaced pages (11pt font). (Appendices do not count in the page limit). Please number the pages. Please upload the report (in PDF format) on Blackboard by **July 21, 2016, 11:59pm EST**. The report should (at least) include the following sections:

- **Abstract:** use 250 words or less to summarize your problem, methodology, and major outcomes.
- **Key words:** select a few key words (up to five) related to your work.
- **Introduction:** describe the background and motivation of your problem.
- **Literature review:** discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the-art. Please do not discuss paper one at a time, instead, identify key characteristics of your topic, and discuss them in a whole. Please cite the relevant papers where appropriate.
- **Methodology:** discuss the key aspects of your problem, data set and regression model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for regression modeling, your process for building regression models, and your model selection.
- **Experimentation and Results:** describe the specifics of what you did (data exploration, data preparation, model building, model selection, model evaluation, etc.), and what you found out (statistical analyses, interpretation and discussion of the results, etc.).
- **Discussion and Conclusions:** conclude your findings, limitations, and suggest areas for future work.
- **References:** be sure to cite all references used in the report (APA format).
- **Appendices:**
    - Supplemental tables and/or figures.
    - R statistical programming code.