Final Project CMDA 3654- Guidelines

Fall 2014

1. **Submission**
   Please submit all your project deliverables to the Dropbox and your GitHub CMDA/FINAL PROJECT folder by 1pm on December 8th. All projects for all teams are due at that time. Each team member must submit the project materials to their own Dropbox and GitHub.

2. **Deliverables:**

   1. **PPT Presentation slides**
   2. **Data file (csv etc.).**
   3. **Data manual pdf.**
   4. **Ipython notebook with code and code results, with clear comments throughout.**

3. **Minimum requirements for project development and ipython notebook code:**
   1. **Import data from a csv, txt, or JSON object. Use pandas. Save your data as a pickle.**
   2. **Provide at least one reshaping technique implementation (manipulate your data, merge, pivot, delete columns, delete rows etc.)**
   3. **Provide at least one treatment to missing data.**
   4. **Provide and comment on numerical summaries for at least three variables (at least one categorical and at least one numeric). You can use Python code or R code with pyper.**
   5. **Provide and comment on at least 3 visualizations (using three different types of charts) with either Python code or R code through pyper. PCA visualization can be a one of the three charts.**
   6. **Implement at least one machine learning algorithm with either scikit-learn in Python or R through pyper, or both. If you decide on R, you should code it first separately in an R script and provide the R script as well. Be sure to use training and testing sets and comment on validation.**

4. **Project Presentation Instructions**
    1. **Project teams are posted on Scholar. Work with your team to devise the presentation. Teams 1-5 present on December 8; Teams 6-10 present on December 10th.**
    2. **All members of all teams must be present on both December 8th and December 10th. Absence leads to 10% penalty in project grade for the absent student.**
    3. **Each team member must present on the assigned date. Each team has 7 minutes. If you take more than 8 minutes, a 10% penalty will be applied to the team.**
    4. **The format of the ppt (color, fonts etc.) is each team's choice and must be carefully examined for maximum impact.**
    5. **Slides of ppt:**

Slide 1 Executive Summary (Motivation behind project, goals + results)

Slide 2 Presentation of Data and Data sources (what variables, how they are measured, what type, where they are coming from)

Slide 3 Discussion of data analytics and interpretations (numerical summaries)

Slide 4 Visualizations and discussion

Slide 5 Machine learning algorithm employed, and how it works

Slide 6 ML algorithm results and validation

Slide 7 Conclusions and recommendations for deployment

    6. **Consult Chapter 11 from Practical Data Science with R text for presentation suggestions. Focus on selling your data science project to a project sponsor/CEO/client in the first part (first three slides) and then you can think of your audience as an audience of peer data scientists.**