IST707 Data Analytics

HW1: Data Preparation and Association Rule Mining

Due: 11:59pm, Feb 9th, 2020

Homework instructions

- Analyze *employee_attrition.csv* dataset provided. This dataset provides a variety o information about the employees, such as demographics, time on job, etc. and also if they stay with or leave the company (as in binary attribute *"Attrition"* with *No* means stay and *Yes* means leaving).
- Follow CRISP-DM process
  - Data preprocessing, cleaning, transformation: identify potential data quality issues and properly address those issues as part of data preparation.
  - Conduct exploratory data analysis (EDA): derive descriptive statistics and apply data visualization to check for interesting data patterns.
  - Run association rule mining algorithm using default settings as a baseline model.
  - Fine tune the model by experimenting with different algorithm hyper-parameters and discuss how tuning those hyper-parameters could impact the model performance (e.g. overfitting or underfiting).
  - Output and present the most interesting and significant rules which could predict *"Attrition"*; print out the top 5 rules which predict those who stay vs. who leave, respectively.
  - Provide interpretations of the above chosen association rules and also discuss why you consider them interesting and significant.
- Use Rmarkdown (or Jupyter Notebook) to structure your report and submit the html output
  - All the codes and relevant outputs (limit the size of outputs to only include those relevant contents and refrain from printing out excessive amount of irrelvant information or data)
  - Analysis writeup using markdown language (interpretation and discussion of the results with the proper session titles and all the information useful to grade your work)
- Develop a R Shiny (or python Dash) web app to host the analytics process and upload to shinyapps.io
  - Instruction of uploading R app to shinyapps.io (Uploading python app to heroku.com)
  - Include the web app URL in your html submission
  - Provide the appropriate control widgets in the app UI that allows app users to choose different values of model parameters (e.g. support, confidence, length of association rules)
  - Output top associate rules according to users' choices
  - Include one visualization to plot the association rules on 2D space defined by the association rule performance metrics

Grading rubrics

- Rmarkdown/Notebook report (60%)
  - Include all the key data mining steps which are neatly structure in the report with both R codes and relevant otputs (using proper section titles) (40%)
  - In-depth interpretation of the analysis output (20%)
- R shiny/Python Dash app (40%)
  - A functioning web app that meets all the specification as in this instruction (30%)
  - Allow the maximal level of user interactions and return properly formatted analysis results (including visualization) to the browser (10%)

Submission instructions:

- Submit the html report (which includes the shiny/dash app link) together with the rmarkdown (or Jupyter Notebook) to the blackboard
- Deadline: 11:59pm, Sunday, Feb. 9th, 2020
- Late submission policy: Late submission will incur 20% penalty for every additional 24 hours' delay until all points are deducted